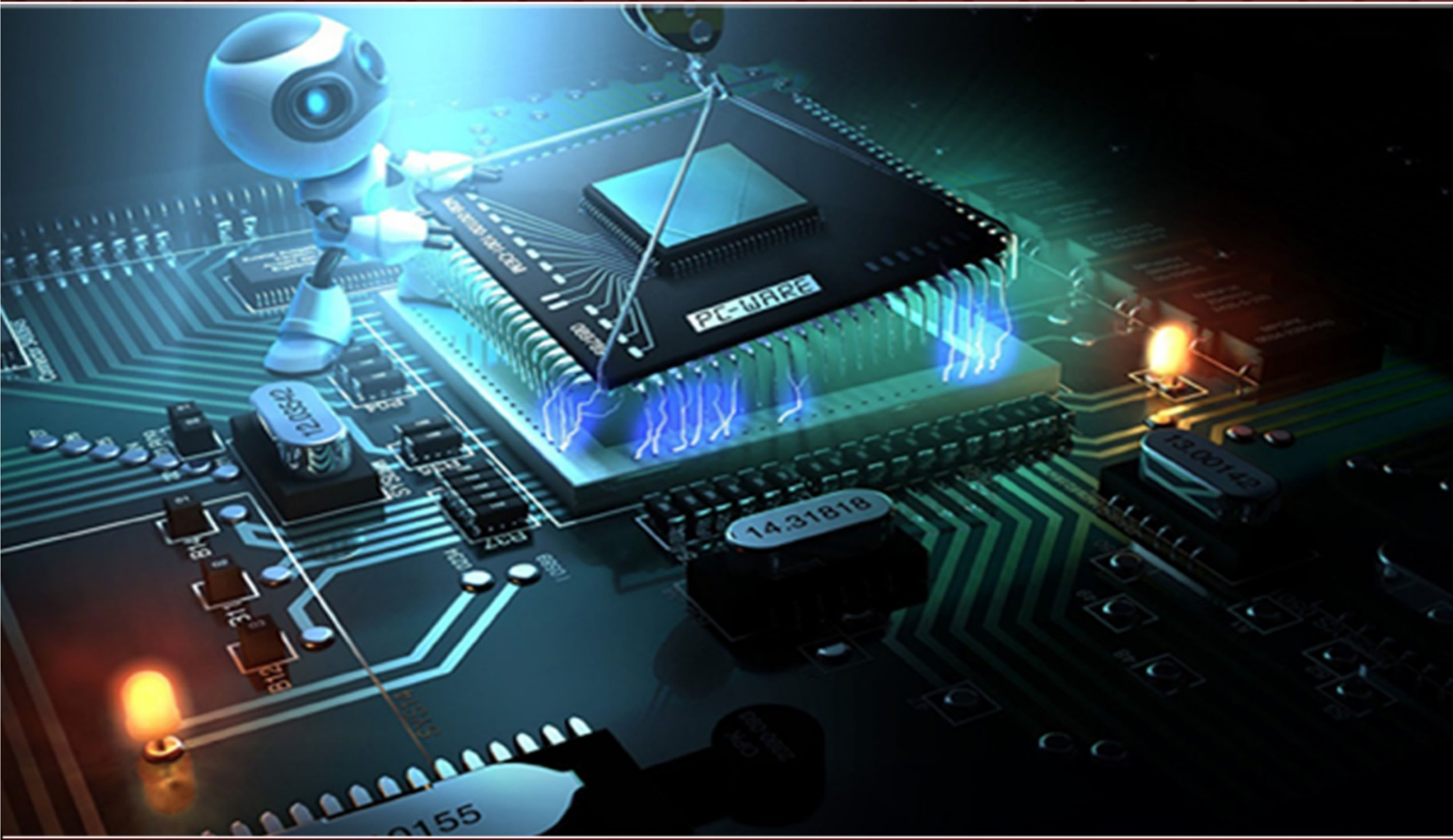
ISSN: 2054-7390 Volume 2. Issue 5 - October 2014

Transactions on Machine Learning and Artificial Intelligence



Society for Science and Education
UNITED KINGDOM

TABLE OF CONTENTS

EDITORIAL ADVISORY BOARD

DISCLAIMER

All the contributions are published in good faith and intentions to promote and encourage research activities around the globe. The contributions are property of their respective authors/owners and the journal is not responsible for any content that hurts someone's views or feelings etc.

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 5
ISSN: 2054 - 7390

Semantic Web Improved with Fuzziness added in Weighted Score

Jyoti Gautam and Ela Kumar
School of Information and Communication Technology, Gautam Buddha University, Greater NOIDA, INDIA;
jyotig@jssaten.ac.in; ela_kumar@gbu.ac.in

ABSTRACT

A lot of improvement has gone in the area of information retrieval. But, still improvements can be done. Social networking giants like Facebook, LinkedIn, CiteULike have taken a new role. There is a huge data collection from these sites. A lot of work is going on to convert this data into information. As we are aware that term weighting has a significant role in text classification. Many techniques of text classification are based on the term frequency (tf) and inverse document frequency (idf) for representing importance of terms and computing weights in classifying a text document. In this paper, we are extending the queries by “keyword+tags” instead of keywords only. In addition to this, we have developed a new ranking algorithm which utilizes semantic tags to enhance the already existing semantic web by using the weighted score. The data for the tags has been obtained through CiteUlike. Here, we have manually added fuzziness in the weighted score for the purpose of improving the algorithm.

Keywords: Text classification; Semantic Web with weighted idf feature; Expanded query; Fuzzy Semantic Web; Fuzzy Ranking Algorithm.

1 Introduction

Result sets have less relevance in response to the queries given. To improve the results, a lot of research is happening in the direction of semantic web. There are lots of social networking sites and a lot of data is produced by them. But, there is urgent requirement of converting this data into meaningful information.

This paper utilizes the semantic tag information with web page. We have obtained this information from CiteULike [19] (Research Paper Recommender and online Tagging System). Semantic tag’s information is added along with the query for the purpose of disambiguation. Then, by matching the semantic description between the query and web page, user’s query intent can be well understood.

Research on folksonomies is growing at a very fast rate in spite of the various difficulties encountered. We have sites like Delicious [18], Facebook, LinkedIn and CiteUlike etc. which are producing this data. Social tagging, also known as social annotation or collaborative tagging is one of the major characteristics of Web 2.0.Users annotate resources with free-form tags in social-tagging systems The

DOI: 10.14738/tmlai.25.333
Publication Date: 10th September, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.333

resources can be of any type, such as Web pages (e.g., delicious), videos (e.g., YouTube), photographs (e.g., Flickr), academic papers (e.g., CiteULIke), and so on .

In this paper, the following approach has been adopted. We have tried to use the metadata available in the form of user feedback from CiteUlike.

1. A new approach has been developed. A ranking algorithm based on semantic tags which adds the fuzziness in the weighted score, is proposed and the data is obtained through CiteULike.
2. The query was expanded. The idea was to use "keyword + tags" instead of keywords only.

   The data for the tags was obtained through CiteUlike.
3. The performance analysis of the approach was done with Google by using normalized DCG.

2 The Existing Ranking Methods

Tf-idf, term frequency-inverse document frequency, often used as a weighting factor in information retrieval, is a numerical statistic which reflects how important a word is to a document in a corpus. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others (Wikipedia) [17]. Various variations of the tf–idf weighting scheme are often used by search engines. Search engines use these weighted measures as a central tool in scoring and ranking a document's relevance given a user query. The tf-idf is improved by many literatures.The literature [9] provides an improved approach named tf.idf.IG to remedy this defect by Information Gain from Information Theory.

The authors [12] explored the technique of Social Annotations for the Semantic Web. These annotations are manually made by normal web users without a predefined formal ontology. The evaluation of the approach shows that the method can effectively discover semantically related web bookmarks that current social bookmark service cannot discover easily.

[2] The authors use six tag metrics to understand the characteristics of a social bookmarking system. Possible design heuristics was suggested to implement a social bookmarking system for Cite Seer using the metrics. The authors Cilibrasi and Vitanyi in 2007 [1] described a technique for calculating the Google similarity distance. Jin, Lin and Lin [6] proposed the architecture of a semantic search engine and an improved algorithm based on TFIDF algorithm. The algorithm considers crawling of static web pages. A personalized search framework was proposed by Shenliang, Shenghua and Fei [10] .It utilizes folksonomy for personalized search.

[5] The other method of basic TFIDF model uses supervised term weighting approach. The model uses class information to compute weighting of the terms.

The authors [14] proposed a trust-network based fuzzy knowledge sharing in semantic web in the year 2009. In this paper, we propose a trust network based fuzzy knowledge sharing model in semantic web. According to this model, any agent can broadcast its knowledge queries through the trust network, and aggregate different opinions according to the trustworthiness of each agent echoed.

Zhao and Zhang [15] proposed a new viewpoint on how to improve the quality of information retrieval. The queries are extended by "keywords+tags" instead of keywords only. A new tag based ranking algorithm (OSEARCH) was proposed and the results obtained were also compared with Google by several evaluation methods.

The authors [8] focussed on search engine personalization and developed several concept-based user profiling methods that are based on both positive and negative preferences. The proposed methods were evaluated against the previously proposed personalized query clustering method.

Another supervised term weighting method, proposed by the authors, [13] provides an improved tf-idf-ci model to compute weighting of the terms. The method uses intra and inner class information.

The paper proposed by Yoo [11] suggests a hybrid query processing method for the effective retrieval of personalized information on the semantic web. When individual requirements change, the current method of query processing requires additional reasoning for knowledge to support personalization.

The authors [4] proposed the method of relevance feedback between hypertext and semantic web search. The paper proposed investigates the possibility of using semantic web data to improve hypertext web search.

The authors [16] proposed an effective pattern discovery method for text mining. The paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.
The paper [7] proposes searching and ranking method of relevant resources by user intention on the semantic web.

This paper proposed by the authors [3] proposes a framework for a tag-based Academic Information Sharing and Recommender System which shares information such as question papers, assignments, tutorials and quizzes on a specific area.

3 User Query Intent and Storage of Tags

3.1 Metadata information in the Web Pages and Expansion of the Query

Semantic Web implies the content, meaning or the metadata which is related to the web pages. This metadata information is hidden in the web pages. Different websites are working upon it since a long time. Sites like Delicious, CiteUlike, Flickr, LinkedIn etc. allow different users to create their accounts. After creating the accounts, metadata can be added to the different web pages. This metadata conveys the content of the website as interpreted by different users.

The purpose of the search engine is to return optimal results. So, here we have tried to return the optimal results by modifying the ranking algorithm and expanding the query. We have added semantic information with the query. We have expanded the query by “ keyword+tags” instead of the keywords only.

So, the idea is to utilize this semantic tag information. Here, we are proposing the development of a new algorithm based on semantic tags and fuzziness is added in the tags to add a new dimension to the algorithm.

3.2 Storage of Semantic Tags on Web Pages

Multiple tags are associated with a web page, because the pages always contain multi information. These tags carry the semantic information or metadata along with them.

We are storing the tags from CiteUlike. A popular website in academia is CiteULike (www.CiteULike.org). CiteUlike is a free service for managing and discovering scholarly references.

- Easily store references you find online
- Discover new articles and resources
- Automated article recommendations
- Share references with your peers
- Find out who's reading what you are reading
- Store and search your PDF's

CiteULike has a filing system based on tags. Tags provide an open, quick and user-defined classification model that can produce interesting new categorizations.

Additionally, it is also capable to:

- 'tag' papers into categories.
- Add your own comments on papers.
- Allow others to see your library

The semantic tags are obtained from CiteUlike. The URLs along with their tags are stored in a local database. Each URL is opened in CiteUlike and the tags with their numeric values are stored in the database. We add tags' values in the MYSQL database. The data was retrieved from April, 2012 to June, 2013 from CiteUlike for the 10 queries. A total of 500 URLs were opened in CiteUlike and the database was created.

4 A new optimized ranking algorithm

4.1 A New Optimized Ranking Algorithm

Initially, when users want to submit a query, they will not only give the query in the form of keywords, they will also expand the query by adding some semantic information along with the query. Afterwards, the algorithm compares the inputted tags in query with the semantic information on the web pages in order to provide the user with better results [15].

Accordingly, the user query can be expressed as:

Query = {keyword1, keyword2,…, tag1, tag2,…}

In the above formulation, keyword1, keyword2 is the main query keyword.Tag1; tag2 is the semantic information which we are adding to expand the query. For example, Query = {books, artificial intelligence) represents that the user wants to find information relating to books on artificial intelligence.

Similarly, Query = {research papers, statistics}

Represents that the user wants research papers in the field of statistics.

Once, the query is submitted, the system creates a vector of all the user tags.

V_usrt = {user_tag1, user_tag2,...}

Once the query is submitted to the search engine, the engine returns an initial result page list. The vector of all the tags on the result pages is recorded.

V_rest = {r_tag1, r_tag2,...}

V_rest = {r_tag1, r_tag2,...}

Where, r_tag1, r_tag2 represent semantic tags on result pages.

The similarity is calculated between the two tag vectors, and recorded as a Tg_score.

Then, the final score of the web page is:

$$\text{TotalScore} = \text{google\_score} + \text{Tg\_score} * \text{IDF score} \quad (1)$$

$$\text{Score} = \text{Tg\_score} * \text{IDFscore} \quad (2)$$

Re – rank the google results according to this score.
Here, google score represents the original google results score when the query is applied.

$$\text{Google score} = (p-q+1)/p. \quad (3)$$

Here, p represents the total no. of documents, which is 50 in the experiment; q represents the location of the document on search engine's result list. So, google score for the 6th result is (50- 6 + 1) / 50 = 0.9.

In the Equation (1), Tg_score is calculated by matching the tags of the user with the tags of the result page. The match between the two vectors is based on the following factors:

- The similarity between the user tag vector and web page tag vector. The high value is obtained by high similarity between the two vectors.
- The other factor being the weight of the tags on the result pages. Weight refers to the frequency of the tags in the result pages which match with the tags of the user.

Tg_score is defined as given below based on the factors considered:

$$\text{Tg\_score} = \frac{\sum_{i=1}^{|\text{V\_usrt}|} \sum_{k=1}^{|\text{V\_rest}|} (freq(\text{V\_rest[i]}) * \text{sim}(\text{V\_usrt[i]}), \text{V\_rest[k]}))}{\sum_{k=1}^{|\text{V\_rest}|} freq(\text{V\_rest[k]})} \quad (4)$$

In the above equation, freq (tag) represents the frequency or weight of the particular tag on the result page. $\text{sim}(\text{V\_usrt[i]}), \text{V\_rest[k]})$ Represents the similarity between the user tag vector $\text{V_usrt[i]}$and the result page tag vector V_rest[k] and similarity is defined as given below:

$\text{sim}(\text{V\_usrt[i]}), \text{V\_rest[k]})$

= 1, V_usrt[i] and V_rest[k] have the same root,

= 1, V_usrt[i] and V_rest[k] have the same meaning,

= 0, V_usrt[i] and V_rest[k] does not have a semantic relation, (5)

= 0.5, even if half of the V_usrt[i] tag resembles with the V_rest[k]tag.

The similarity between the user tag vector and the result page tag vector can have fuzzy values also. The fuzzy values range from 0 to 1. For example,

Let us take the example of the query {books, artificial intelligence}. For a single query, we are storing the tags for the first 50 Google results. We are storing tag weights for the tags with maximum values. For the tag agent, we have the weight as 3 and the fuzzy value that we are assigning is 0.1. Similarly for tags artificial, intelligence and systems, we have weights as 2, 2, 1 and the fuzzy values assigned by us are 0.5, 0.5 and 0.1. So, we get the weighted score (Tg_score) as (3*0.1+2*0.5+2*0.5+1*0.1)/ (3+2+2+1), which is equal to 0.3.

Google score for the first link is 1= ((50-1+1)/50).

The fuzzy values are assigned to tags, keeping in mind the relevance of that tag with the query. The assignment of fuzzy values is done by a group of users. The assignment is done manually by collecting users of that particular domain. Next in the equation (1) is the IDF score multiplied by Tg_score. We know from the TFIDF algorithm.

Given a document collection D, a word w, and an individual document d Є D, we calculate

$$w_{d=}f_{w,d*}\log(|D|/f_{w,D}) \tag{6}$$

Where $f_{w,d}$ equals the number of times w appears in d, |D| is the size of the corpus, and $f_{w,D}$ equals the number of documents in which w appears in D. Words with high w_d imply that w is an important word in d but not common in D.
Here, if the above equation is analyzed properly, we see that if we replace words with tags, this equation (6) can be used in the context of semantic web. So, $f_{w,d}$ has already been considered as the Tg_score. Now remains the log (|D|/$f_{w,D}$), (which is IDF score). Here, for each query, we have taken the 50 Google results. So, for a particular query, D is 50 and $f_{w,D}$ equals the number of documents which contain the tags.

Now, why we have included this IDF score?

Suppose that Tg_score is large and fw,D score is small. Then log (|D|/$f_{w,D)}$ will be rather large, and so in Equation. (1), the score will be large. This is the case we are most interested in, since this makes the score large in which we are interested. Here, we are calculating $f_{w,D}$ taking into consideration all the urls , which contain tags.

In the above equation Equation. (5), we have manually added fuzziness to the tags. The database is created using MYSQL.

For example, consider the query {books, web mining}.The query contains the tags - data mining, personalisation, personalization, web-personalisation, web- personalization. The weights of the tags are 3, 5, 6, 3, 3. The fuzzy values assigned are 0.9, 0.2, 0.2, 0.2, and 0.2. So, the Tg_score is 0.305 and Google score for the first link is 1. Since out of 50, 31 urls contain tags, so that the IDF score is .207608. Finally, the total score of the first link is 1.063321.

5 Experiments and Analysis

The experiments are performed as follows:

- Initially, submit the query to Google, and obtain the original Google search results.
- Now, submit the Google search results to CiteUlike to obtain the relevant tags.
- Re-rank the search results according to our algorithm.
- Compare the Google results with our algorithm.

5.1 Data Set

5.1.1 Query Set

Initially, we determine the queries which we input to the search engine. We determine a total of ten queries. The queries are a combination of keywords and tags. These queries are submitted to Google. We have chosen academic domain as CiteUlike provides tags for the academic database only

5.1.2 Result Set

Now, submit each query to Google and record the first 50 results. This way, the result set of 10 queries become 500 results.

5.1.3 Results Tag Set

Now, we submit the 500 results to CiteUlike and the resulting tag vector is recorded. We obtain lots of tag values for a result.

We have chosen the following queries.

Q1. {Books, artificial intelligence}	Q6. {Research papers, software engineering}
Q2. {Books, data mining}	Q7. {Pdf, genetic algorithm}
Q3. {Research papers, data mining}	Q8. {Research papers, statistics}
Q4. {Pdf, information retrieval}	Q9. {Books, web mining}
Q5. {Research papers, semantic web}	Q10. {Books, java programming}

For these queries, we compute the values of normalized DCG gains for Google as well as for our algorithm (Fuzzy JEKS algorithm) in Table1.

Table 1. Normalized DCG gains of Google and our fuzzy JEKS algorithm.

Query	**Google Ranking**	**Fuzzy JEKS Algorithm Ranking**
	nDCG(G)	nDCG(A)
Q1	0.980211	0.959274
Q2	0.896716	0.92342
Q3	0.937156	0.926431
Q4	0.979388	0.978542
Q5	0.987652	0.987472
Q6	0.94898	0.948706
Q7	0.98502	0.98638
Q8	0.91282	0.877635
Q9	0.900639	0.929049
Q10	0.943141	0.936474

We obtained normalized DCG values for the 10 queries for our algorithm as well as for Google results. It can be seen that our algorithm acquires higher values of normalized DCG for 3 queries out of 10 queries when compared to Google.

6 Conclusion

We have proposed a new ranking algorithm based on the previous methods only. We have added fuzziness in this new proposed algorithm. Semantic tag of a web page is the metadata information associated with it and depicts a lot about the information associated with it. Here, we have added fuzzy values with the weights of the tag. The fuzzy values have been associated with the tags by looking at the relevance of the tag with the query.

We have proposed the new algorithm using the already existing semantic web algorithm which basically calculates the weighted score of the tags. We have added fuzziness with the weighted score to improve the semantic web. In experiments, we have collected the data from Citeulike and implemented the above algorithm. The relevant fuzzy scores to the different web links have been given by a group of users. Comparing with Google search results, we find that Fuzzy JEKS algorithm acquires better ranking results for 3 queries out of a total of 10 queries. Our algorithm acquires higher values of normalized DCG for 3 queries out of a total of ten queries when compared to Google.

In the future work, we will further improve the algorithm. We will consider combining with the search engines user logs, and mining out information repeated to user's query, such as the click information, the browse information and so on. We can enhance the algorithm by adding these effects.

REFERENCES

[1]. Cilibrasi, R.L., Vitanyi, P.M.B., *The Google similarity distance*. Knowledge and Data Engineering, IEEE Transactions on, 2007. 19: p. 370-383.

[2]. *Farooq, U., Kannampallil, T.G., Song, Y., Evaluating Tagging Behaviour in Social Bookmarking Systems: Metrics and design heuristics, Supporting Group Work, 2007. Proceedings. The international ACM Conference on, 2007: p. 351-360.*

[3]. *Gautam, J., Kumar, E., An Improved Framework for Tag-Based Academic Information Sharing and Recommender System, Proceedings. World Congress on Engineering, 2012. 2: p. 845-850. (IAENG, London).*

[4]. Halpin, H., Lavrenko, V., *Relevance feedback between hypertext and Semantic Web search*, Journal of Web Semantics, 2011. 9: p. 474-489.

[5]. *Jiang, H., Hu, X., Li, P., Wang S., An improved method of term weighting for text classification, Intelligent Computing and Intelligent Systems, International Conference on, 2009.1:p. 294-298, IEEE Press.*

[6]. *Jin Y., Lin Z., Lin H., The Research of Search Engine Based on Semantic Web, Intelligent Information Technology Application Workshops (IITAW), 2008. Proceedings. International Symposium on,2008: p. 360-363, IEEE Press.*

[7]. Lee, M., Kim, W., Park, S., Searching *and ranking method of relevant resources by user intention on the Semantic Web,* Expert Systems with Applications,2012. 39: p. 4111- 4121.

[8]. Leung, K.W.T., Lee, D. L., Deriving *concept-based user profiles from search engine logs*. Knowledge and Data Engineering, IEEE Transactions on, 2010. 22: p. 969-982.

[9]. S. Lu, X. Li, S. Bai , S. Wang., *An improved approach to weighting terms in text*. Journal of Chinese Information Processing, 2000. 14: p. 8-13.

[10]. *Shenliang, X., Shenghua, B., Fei, B., Exploring Folksonomy for Personalized Search, Research and Development in information retrieval, the 31st annual international ACM SIGIR conference on,2008, p. 155-162. ACM, USA*

[11]. Yoo, D., *Hybrid Query Processing for Personalized Information Retrieval on the Semantic Web*, Knowledge-Based Systems,2012. 27: p.211-218.

[12]. *Wu, X., Zhang, L., Yu Y., Exploring Social Annotations for the Semantic Web, World Wide Web (WWW 06), the 15th International Conference on, 2006, p. 417-426, ACM Press, USA.*

[13]. *Zhanguo, M., Jing, F., Liang, C., Xiangyi H., Yanqin, S., An improved approach to terms weighting in text classification, Computer and Management, the International Conference on, 2011, p. 1-4, IEEE Press.*

[14]. *Zhang, C., Yan, M., Trust Network Based Fuzzy Knowledge Sharing, Computational Intelligence and Software Engineering, 2009, Proceedings, 2009 International Conference on, p. 1-5, IEEE Press.*

[15]. *Zhao, C., Zhang, Z., A New Keywords Method to Improve Web Search, High Performance Computing and Communications, 2010, Proceedings, 2010 International Conference on, p. 477-484, IEEE Press*

[16]. Zhong, N., Li, Y., Wu. S.T, *Effective Pattern Discovery for Text Mining*. Knowledge and Data Engineering, IEEE Transactions on, 2012. 24: p. 30-44.

[17]. “tf-idf,” Wikipedia, http://en.wikipedia.org/wiki/(accessed June 2013).

[18]. Keep, share, and discover the best of the Web using *Delicious*, the world's leading social bookmarking service. http:// delicious.com/.

[19]. Search, organize, and share scholarly papers. Indexes over 2 million articles. http://www.citeulike.org/ (accessed april, 2012 to july 2012).

SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

VOLUME 2, ISSUE 5
ISSN: 2054 - 7390

Ultra-fast Lithium cell charging architecture for Mission Critical Applications

[1]Arif Sari, [2]Behnam Rahnama and [3]Ersin Caglar,
[1]*School of Applied Sciences, Department of Management Information, European University of Lefke, Cyprus*
[2]*Department of Computer Engineering, Okan University, Istanbul, Turkey*
[3]*Department of Computer Engineering, European University of Lefke, Cyprus;*
asari@ieee.org ; rahnama@ieee.org ; erscaglar@gmail.com

ABSTRACT

This research presents design and implementation of the ultra-fast parallel lithium charging architecture with an embedded algorithm using an active PWM charge pump supported by a hybrid control mechanism consisting of Temperature, humidity and current sensors. The new architecture guaranties the ultra-fast parallel charging cycles of lithium cells without lifespan reduction due to possible overheat side effects in mission critical applications.

Keywords: Lithium Cell, Fast Charging, PWM Charge Pump, Mission Critical Application.

1 Introduction

Lithium batteries are used everywhere in our daily life from cellular phones to handheld devices, notebooks and tablet PCs. Chargers basically protect cells by limiting the charge cycle from over-current and over-voltage. Such chargers normally are used in living environmental temperature and humidity. On the other hand, mission critical applications require dedicated and specifically designed architectures to overcome challenging temperature and humidity levels.

In such situation, charging circuitry and cells are kept in a temperature controlled area while exterior climate change would not affect charging performance. However, providing such a safe protective cover is not always feasible everywhere. Wireless Sensor Network nodes might be placed in any climate and the usual expectation is their durability. However, the limited space does not allow implementation of a perfect protection case to keep the cells temperature in a steady state. Use of ultra-fast charge pump circuitry is vital in satellites and spacecraft missions while the temperature exceeds our living limits toward extremities.

In addition, charging circuitry must be fault tolerant and self-recoverable after any power failure. This can be managed by applying watchdog timers and various timing stages in a cascade control to make sure the reliability of the system.

The core implementation is to capture the voltage level and passing current at each cell, battery temperature and surrounding humidity. These raw data are processed and a fast phase and frequency correct PWM generator compare match value is updated.

DOI: 10.14738/tmlai.25.430
Publication Date: 10th September, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.430

Using hardware timer allows the microcontroller to jump to different power saving stages and utilizes less energy while controlling the charging circuitry.

The PWM frequency for switching the parallel charging circuit per channel is fixed on 1KHz. The duty cycle varies from 0% for total de-charge cycle to 100% for total charging cycle according to the aforesaid measuring variables.

Next section covers a survey on various lithium charge management integrated circuits and similar challenges in the literature. Then we provide our proposed novel and simple to implement but yet efficient charging algorithm and circuit for mission critical applications. Finally, this research paper is concluded by findings and detailed prospective study.

2 Related Work

Due to growing interest in flexible energy storage devices, scientists in South Korea proposed a method to tackle with charging duration of lithium ion rechargeable batteries [1]. The proposed approach is related with the use of cathode material, standard lithium manganese oxide (LMO), soaked in a solution containing graphite. Recharging process of the battery proposed through a network of conductive traces which runs throughout the cathode graphite-soaked LMO. A lithium ion battery consists of several parts where the carbonized graphite networks recharge all parts of the battery simultaneously. Researchers' speed-up the recharging process and decreased the amount of time to recharge lithium ion batteries through a specific method. An electrolyte and graphite anode is used to package the cathode and all energy holding particles of the battery start recharging simultaneously by this way which leads to speed up of charging lithium ion batteries. In the literature, another group of researchers demonstrated a lightweight, thin and flexible full lithium ion battery with high-rate performance and energy density [2]. The proposed battery model was loaded with $Li_4Ti_5O_{12}$ (Lithium Titanate – LTO) and $LiFePO_4$ (Lithium Iron Phosphate – LFE). The Lithium Titanate is a prospective material for anodes of Li-ion batteries. The LTO leads Li-ion batteries to expand into different applications such as electrical cars and hybrid vehicles. The usage of LTO in Li-ion batteries dramatically increases the battery lifetime and provides wide range of operation possibilities under different operation temperatures such as -30°C – 60°C [3].

The charging of Li-ion batteries requires specific attention in order to prevent overcharging and thermal runway problems where these problems may lead to damage in Li-ion batteries. Extended charging of Li-ion batteries may lead to a change in forms of battery when it is above 4.30V. This change can be observed on the batteries as changes in plating of metallic lithium on the anodes, while the cathode material becomes an oxidizing agent, loses stability and produces carbon dioxide (CO_2) [4]. As it is mentioned before, Li-ion batteries consist of cells and cell pressure can rise due to overcharging. The usage of LTO in Li-ion batteries decreases the amount of charging while preventing battery against damage due to overcharging.

LFE, is a similar component in terms of their usage in high power applications likewise Hybrid vehicles. The usage of LFE cells in Li-ion batteries leads non-explosive and long battery life cycle for them however their energy density is lower than normal Li-Ion cells. These anode materials, LFE and LTO, are used to improve performance and durability of the Li-ion batteries by the researchers. The LTO is introduced for sodium ion-batteries to extend lifetime and provide fast charging facilities [4].

In order to meet the modern society needs for flexible and durable energy storage devices, another research has conducted to propose a new structure of thin and flexible Li-ion battery design [5]. The goal of proposed design was to reach to high energy density with mechanical flexibility. The Li-ion battery materials are coated with thin carbon nanotube (CNT) films. The flexibility is provided through CNT design where it acts as a current collector for both anode and cathode. Since the actual performance of rechargeable Li-ion batteries depends on the thermodynamics and kinetics of the electrochemical reactions involved in the components of the cells such as reductant, oxidant, electrolyte and separators, the usage of nanotechniques and nanomaterials becomes an essence of designing rechargeable Li-ion batteries. The usage of different nanotechniques and nanomaterials such as LiCoO2, $LiFePO_4$, $LiMn_2O_4$, with $Li_4Ti_5O_{12}$ and $LiMn_2O_4$ aims to increase lifetime, durability and energy density of rechargeable Li-ion batteries without affecting safety and reliability issues. Variety of researches conducted on implementation of different nanomaterials on development of rechargeable Li-ion batteries where the researches are spread to hybrid electric vehicles, clean energy and aerospace fields [6].

Researchers focused on rechargeable Li-ion batteries toward the electron transfer on key materials and components of Li-ion batteries on molecular, atomic or ionic transport for to solve long cycle, extended energy diversification and reliability issues [7]. Variety of functional materials which are used to design rechargeable Li-ion cells such as lead acid, nickel-metal hydride and sodium-sulfur batteries are reviewed. However, implementation of fast charging mechanisms or architectures are not proposed or examined to expose the impact of nanomaterials on cell design and safety, reliability and lifetime issues of rechargeable Li-ion batteries.

3 Proposed Ultra-Fast Charging Architecture

3.1 Materials

Lithium cells are sensitive to over-voltage and under-voltage cycles. Thresholds for minimum and maximum voltage level as well as charging cycle start and end latency may be controlled by local charging integrated circuit. A widely used example of such chips is S-8254A. Several lithium charge management ICs also support a PTC or an NTC resistor on cell to cutoff charging in case the cell temperature exceeds a certain threshold. MCP73844 series of Microchip battery management chips can be considered as a well-known IC in this category. However, all aforementioned ICs are designed for living environmental temperature range.

Humidity might not play an important role as usually charging circuitry can be isolated in a closed plastic like packaging such as laptop chargers. The Environmental humidity does not have considerable effect on charger. However, temperature is the main factor we must take into account. Maxim DS18S20 High-Precision 1-Wire Digital Thermometer was used in test platform to catch temperature changes from -55°C to +125°C degrees.

This is known that the extreme temperature levels on earth are -88°C at Vostok Station in Antarctica to +58°C in the Libyan Desert. Of course, the entire design must be changed in case a WSN node powered by a lithium based solar charger is used at minimum temperature proximity.

Our designed platform so called “Adamus” consists of four blocks; first, Charging circuitry, second, Voltage, Current, temperature and humidity measurement sensors, third, microcontroller and Darlington switching gates and fourth, data logger over UART serial interface transferred over Serial to

USB bridge (AS8303CDS). Other peripherals including LGM128128A LCD screen and buzzer help to better understanding of test case during runtime.

3.2 Designed Hardware

Parallel charging circuit is powered by FTD2017A on four separate channels on extension boards. As mentioned earlier, DS18S20 is used to catch the cell temperature. In addition, a combined temp/humidity sensor DHT11 sensor checks the charger circuit conditions. Voltage per cell is captured by ADC interface available at microcontroller. A semi electromagnetic transducer IC (ACS712ELC-05B) as separate board converts the current flow to voltage level and captured separately per each channel. Atmel xmega series of microcontrollers are more desired due to extreme low power consumption. However, this test board is managed by Atmega32, the older version but yet practical from laboratory purposes. ULN2803 are used to provide four coupled n-channel switching for all four parallel cells. Finally, captured data is transferred over the serial link to the host computer for logging purposes. A future amendment to the project is to include a microSD card onboard to save captured data without necessarily transferring them over serial link. Following figure presents the Adamus project PCB development board.

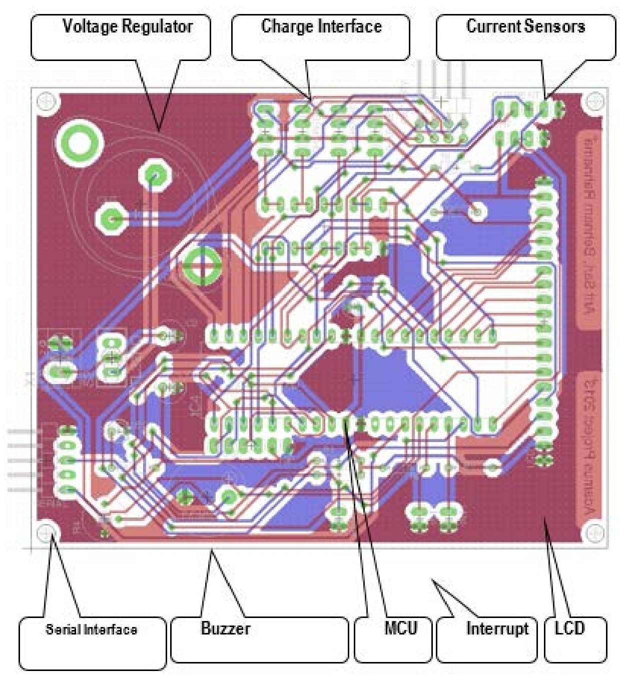


Figure.1 The charge Management Printed Circuit Board

3.3 Charging Algorithm

The charging algorithm basically considers three measurement variables namely, Cell Voltage (VCi), Charge Current (ICi), Cell temperature (TC). In addition, Board Temperature (TB) and Humidity (HB) are taken into account for board management purposes. As the result, compare value (PCi) in phase and frequency correct PWM mode is changed per each channel. The PWM duty cycle for each channel is defined as PDi. The modularity of design helps in parallelization of charging circuitry for more number of cells.

The program captures aforementioned values and transfers them to the host at each second in addition to reset the watchdog timer value. Accordingly, read timeout on the host device is considered as two seconds. Watchdog timer is set to restart the program flow at each two second in case of system inactivity unless, total charging cycle is accomplished.

```
ON 1 second TIMER event{
      READ ADC as average of 10 samples per channel
      READ DHT11 and DS18S20 as TB, HB and TC
      CALCULATE VCi as ADCi * 5 / 256
      CALCULATE ICi as (ADCi * 5 / 256)/0.185
      if MSB=0xff then LSB is TWO'S COMPLEMENT of LSB
      CALCULATE TC as LSB * 0.5
      FOR ALL channels i (0 to 3){
              IF (VCi < MaxVoltThreshold)
                       PDi = PDi + (100 – Pdi)/2
              ELSE{
                      PDi = 0
                      End Mission for Channel i
              }
              IF (ICi > MaxCurrThreshold)
                      PDi = PDi – (PDi / 10)
              IF (TC > MaxTempThreshold OR < MinTempThreshold)
              PDi = PDi / 2
              IF (HB > MaxHumThreshold){
                      PDi = 0
                      STANDBY
              }
              IF (TB > MaxEnvThreshold OR TB > MinEnvThreshold)
              PDi = PDi * 2 / 3
      }
      CREATE Data Packet and SEND it to Host Computer
      WAIT for ACKNOWLEDGE
    IF host returns ERR message RESEND the packet OR DISREGARD it according to the policy
}
```

The algorithm executes the same event at each second. ADC values as well as sensors are read and corresponding variables take values. Each channel is controlled by a separate PWM signal. The charge pump starts logarithmically and feeds cells. If the charge current flow is more than preferred threshold it reduces the charge activity by 10 percent, per sampling iteration until it stabilizes the current flow. Cell temperature decreases the charge pump activity of corresponding cell logarithmically.

On the other hand, board temperature affects the PWM with a lower ratio than cell temperature. An extraordinary level of humidity protects the charging mechanism by changing its state to standby.

Data Packet is generated and is sent to the host computer waiting for the acknowledgement. If transmitted CRC would not match with the calculated one at the host controller, an ERR message is returned. This can be considered to repeat the previous message immediately or ignore it due to the nature of the application.

3.4 Communication Protocol

Captured temperature and humidity values from DHT11 sensor is used to protect the charging board against environmental failures. Therefore, TB and HB values are not transferred to the host computer. The remaining values are transferred in 12 bytes data packet format as follows:

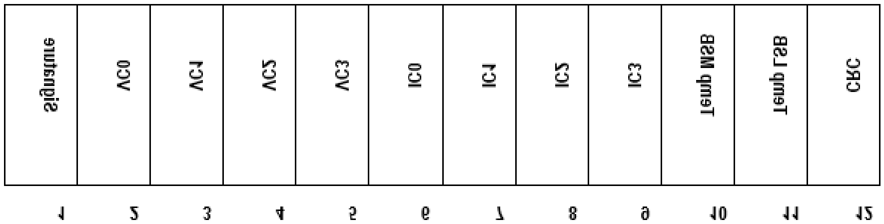


Figure.2 Transferred Data packet Structure

Signature byte is the fixed 0XFF value as the start byte for data packet transfer. VCi and ICi values are 8-bit resolution result of ADC module presenting 195mv per unit and 185mv/A. TC is transferred according to DS18S20 data register standard. It takes 750ms at most to read the temperature value. The host emulator software reads the data packet and records processed information in a spreadsheet file.

In case of any OS interruption to the host program, it searches for the next received packet starting by packet signature. CRC is calculated at host to ensure the proper data transfer from MCU. An Acknowledge message is sent to MCU upon appropriate data transfer while an error message is sent in case of data corruption. Due to the real-time nature of the mission critical charging application, corrupted packet might not be retransferred to the host. Finally, MCU sends a packet with all capturing values as 0XFF meaning the process termination.

Following figure presents the Adamus emulator while listing the available host serial ports to initiate data retrieval.

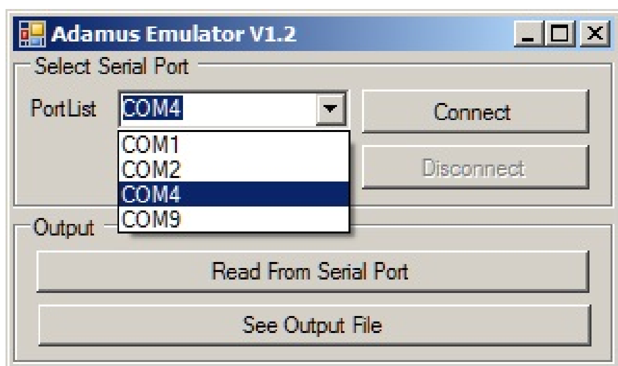


Figure 3. Adamus Emulator Software Interface

Other required commands can be easily defined and being transferred to the host device or reversely to the MCU by changing the initial signature and packet structure.

4 Conclusion & Future Work

This research presents the design and implementation of an embedded ultra-fast parallel lithium charging cell with consideration of extreme cell and board temperature and humidity conditions. Switching charge pump varies using variable PWM values per each channel. The interface is easily connected to the host device for data logging purposes.

As the future work, we wish to examine various types of lithium cells including Li-MnO2, Li-SOCl2, Li-I2, Li-CuS, Li-FeS2, Li-NiCoO2, Li/Al-MnO2, and Li/Al-V2O5 in laboratory while simulating extreme high and low temperature and humidity conditions. Surrounding pressure might be also taken into account in case mission critical applications such as charging circuitry for satellites are being examined as a case study.

ACKNOWLEDGMENTS

Authors would like to thank the European University of Lefke for providing a wide range of research facilities at Robotics and Chemistry laboratories.

REFERENCES

[1]. S. Lee, Y. Cho, HK. Song, KT. Lee, J. Cho (2012). Angewandte Chemie, "Carbon-Coated Single-Crystal LiMn2O4 Nanoparticle Clusters as Cathode Material for High-Energy and High-Power Lithium-Ion Batteries", Vol. 124, Issue 35, pp. 8878-8882, DOI: 10.1002/ange.201203581.

[2]. N. Li, Z. Chen, W. Ren, F. Li, H.-M. Cheng (2012). *Proceedings of the National Academy of Sciences,* "*Flexible Graphene-based lithium ion batteries with ultrafast charge and discharge rates"*, *Vol. 109*, Issue *43*, *17360*.

[3]. HW. Lu, W. Zeng, YS. Li, ZW. Fu (2007). Journal of Power Sources, "Fabrication and Electrochemical properties of three-dimensional net architectures of anatase TiO_2 and spinel Li4Ti5O12 nanofibers", Vol. 164, Issue 2, pp. 874-879,

[4]. DOI: 10.1016/j.jpowsour.2006.11.009

[5]. Z. Liang, HL. Pan, YS. Hu, L. Hong, C. Li-Quan (2012). China Physics B. "Spinel lithium titanate ($Li_4Ti_5O_{12}$) as a novel anode material for room-temperature sodium-ion battery", Vol. 12, No. 2, p. 028201-4. DOI: 10.1088/1674-1056/21/2/028201.

[6]. L. Hu, H. Wu, F. La Manita, Y. Yang, Y. Cui, (2010). American Chemical Society Nano. "Thin, Flexible Secondary Li-ion Paper Batteries", Vol 26, No.4, p.5843-8, DOI: 10.1021/nn1018158.

[7]. HK. Liu, GX. Wang, Z. Guo, J. Wang, K. Konstantiov (2006). Journal of Nanoscience and Nanotechnology, "Nanomaterials for lithium-ion rechargeable batteries", Vol 6, No.1, pp. 1-15.

[8]. F. Cheng, J. Liang, Z. Tao, J. Chen (2011). Advanced Materials, "Functional Materials for rechargeable batteries", Vol 23., No.15, p.1695-715, DOI: 10.1002/adma.201003587

Society for Science and Education United Kingdom

TMLAI Transactions on Machine Learning and Artificial Intelligence

Volume 2, Issue 5
ISSN: 2054 - 7390

Proxy Framework for Database Extension: Database-Independent Function Extension

Pinaet Phoonsasrakun[1], Alexander Adam[2], Attaporn Khaesawad[1] and Seree Chinodom[1]

[1]Department of Information Technology, Burapha University Sakaeo Campus, 27160 Sakaeo, Thailand;
[2]Development Division, dimensio Informatics GmbH, 09111 Chemnitz, Germany
pinaet@buu.ac.th; alad@dimensio-informatics.com; attaporn@buu.ac.th; seree@buu.ac.th

ABSTRACT

This paper presents a proxy framework, which allows introducing various functionalities in an existing database system and application environment. By using a proxy not interfering with the application and with the database are necessary. An abstraction layer is introduced to make functional modules, database independent, and illustrates its use in two examples – index integration and a result cache. Finally, an overview of the results that have been achieved is given.

Keywords: Database Extension, Indexing, Proxy, Transparent.

1 Introduction

Database systems are designed to support a large number of applications. Where this support is not given, they were extended. For example, SQL continuously developed and the manufacturers add even non-standard extensions to it. Some examples of this are the geodetic extensions of IBM DB2 [1] and Oracle [2]. In addition to these built-in extensions that are available for the total number available, manufacturers allow the user-specific expansion of their systems. This includes, for a long time, user-defined data types and functions [3, 4]. Applications, these possibilities do not go far enough that even does data storage, i.e. how the data is to be stored and how the indexing has to take place on them, have the right to decide [5, 6].

The possibilities enumerated here are used partially by different groups of database users. On the one hand, application vendor draw their benefits from user-defined types and functions. On the other hand with skilful distribution of data, indexes, and other parameters, database administrators try to get the most out of the individual systems.

Besides, both sides can run into problems. For example, an application developer who wants to keep his application portable cannot make use of all capabilities of the database systems straight away, since these often are not standardized. On the other hand an administrator must not carry out changes in a scheme which provides an application without compromising support and warranty claims. In the end, the access to the database system can simply be refused for him also since it is perhaps a purchased service under constraints and the administrative rights are missing.

DOI: 10.14738/tmlai.25.498
Publication Date: 10th September, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.498

In application and database system environment, this is illustrated in Figure 1. In addition, the registered proxy framework will be presented in this paper. It complements this environment on the network level for new features and it is transparent to the application and the database system.

The following are some works that affect the function of the proxy frameworks, presented (Section 2) and its functions implemented (Section 3.1). An example of index integration (Section 3.2) and a result cache (Section 3.3), illustrates their use and finally some achieved results presented (Section 4).

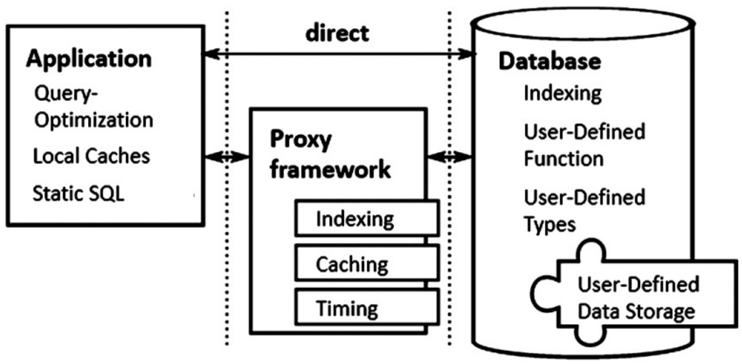


Figure 1: Parts of the existing environment from application and database system that has been extended by our proxy framework. In the proxy framework, some of the potential function modules are registered.

2 Related Work

As mentioned in the introduction, there are primarily two challenges. First, there must be a way to introduce functionalities into multiple database system environments; second, it should be transparently done outside of the database system or application.

Independence of the database system used is offered at the application level through various interfaces, JDBC [7], ODBC [8], OLE DB [9] and ADO (.NET) [10] are highlighted here. Common to all is a way for the application-defined interface. This is, if necessary, shown or passed on directly on a database-specific interface. Such structure (see Figure 2) is always possible if the underlying layer (e.g. Oracle OCI [11]) supports all the features of the overlying layer (ODBC-API).

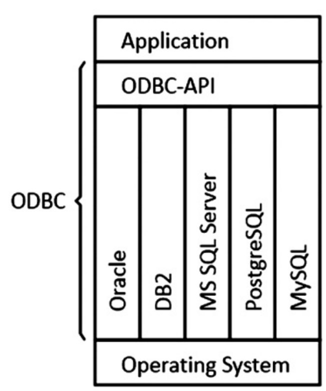


Figure 2: Structure of a system of application, ODBC abstraction layer and database-specific drivers

Furthermore it shall be all about the integration as a proxy in this paper. Also in this area, there are already preliminary works. The concept of a proxy is not new and is designed for database systems, used inter alia by the following developments:

1. *The TDS Protocol Analyzer is a tool that works on the network protocol of tabular data stream (TDS) [12] of Microsoft SQL Server. With this tool, it is possible to intercept the data packets which are sent to record, to analyze, to execute statistical analyses and to monitor vulnerabilities on them. [13]*
2. *GreenSQL is an open source proxy, which is to increase security of MySQL and PostgreSQL, as well as TDS in a commercial version. It scans the communication between the clients and the database system for suspicious requests that might be suitable to exploit security vulnerabilities. Furthermore, it can act as a firewall. [14]*
3. *By means of the Security Testing Framework, invalid packets are generated syntactically for the DRDA protocol used in IBM DB2 (Distributed Relational Database Architecture) [15] to test the robustness of the implementation. [16]*

What all these aforementioned systems have in common is that they work only for selected databases and limited to a few manufacturers. Their area of application is focused on monitoring, testing and access control. Right here, the proxy framework is to apply and offer a possibility, regardless of the database system used, to integrate easily extensible functionality.

A way, how the time performance of the database system can be influenced positively, was presented in [17]. Besides, the index ICIx [18] was requested directly from the application and then the results were used for the expansion of the queries for the database system. This approach is possible only for an application developer.

3 Concept

3.1 Proxy Architecture

Based on the requirements set regardless of the database system used and to be integrated in the network, in the following, the architecture of the proxy framework is presented.

The environment, into which the proxy inserts itself, consists of one or more clients and a database system. The proxy is used in this environment, by all connections which formerly led directly to the database system, are now redirected through the proxy. This then establishes the connection to the database system. Its principle function is to forward data packets received from the client to the database system. Figure 3 illustrates this scenario again.

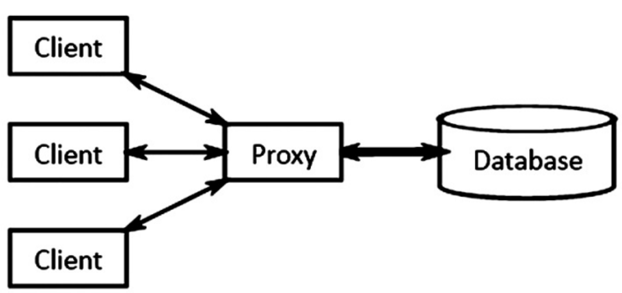


Figure 3: Environment, in which a proxy has been integrated between a database and its clients

Up to this point a pre-control can be already realized by connections (e.g. load balancing, IP access control). Nevertheless, for a further integration of functions, this is not yet enough. A content analysis of the packets to be forwarded is required for this. Client and database system communicate using a protocol, which varies from database system to database system. The proxy framework presented here deals with the database systems IBM DB2, Microsoft SQL Server, and Oracle.

In examining, the database protocols reveal that the sequence of events in the communication between client and the database system consists of the following main steps:

1. *prepare*
2. *bind*
3. *execute*
4. *fetch*

Unlike on the application level, the operations just listed can be combined into composite operations. For multiple bind calls, only a single bind command, which collectedly contains the individual variables, is set off, for example, in Oracle OCI. It is also possible that *prepare*, *bind* and *execute* together are sent in one command. These composite operations can be separated into their individual operations, and then further processed as individual operations.

Based on this observation, the framework provides a set of event-driven callback functions. These exist for each communication direction client → database system (*on_... _request*), as well as for the reverse direction (*on_... _reply*). The following is an overview of functions offered to the user of the proxy framework presented and how they are among other things. All functions have a lot of input parameters that they cannot change and a lot of return parameters, in which any changed values are entered.

- ***on_connect****_...*: these functions receive the connection parameters as input. With them it is possible to modify the connection, for example, to stop or to redirect to a different database than the actual target database.
- ***on_prepare****_...*: this function prepares requests to the database. The database initially receives the requests and checks the syntactic correctness. At this point, the request can be checked or changed. The change will be important for integrating index in the following. The database responds with a request identifier about the further steps to process.
- ***on_bind****_...*: At *prepare* requests with parameters can be provided. These must be bound prior to execution, thus occupied with values. Within this callback function, the parameters can be changed, or, completely held back in the event of a change of the request.
- ***on_exec****_...*: Finally it is communicated to the database that everything needed for the execution is present. The call to this callback function is the only time at which the state of the request is defined with its possible binding parameters, because up to this point, parameters can still be bound or all requests will be discarded.
- ***on_fetch*** *_...*: After the request has been processed in the database, the data can be collected. The client sends a request identifier and then receives the corresponding request and data response. This callback function makes it possible to create a new result set. This is fixed neither on the size, nor on the type of the original result set.

- ***on_free_...***: Ultimately the data structures that have been created in the database system, as well as in the client, must be cleaned up. All data provided by means of the proxy frameworks for such request can then be released.

These functions are part of an abstract interface. By this if a derivative is provided, then this derivative is called *module*. Modules provide the methods with which the real work is carried out. They are similar to applications that use the ODBC interface, through the callback functions described above, regardless of the database being used. Without loaded module, a standard module is used, which copies the data of the input parameters onto the output parameters. A self-written module inherits all of the methods of these standard modules and needs to overload only the methods that it needed for its function. In the following, two modules are presented, based on this interface, a module for index integration and a result cache module.

3.2 Index Integration

The fundamental integration of an external index in a database was discussed in detail in [19] and [17]. The idea is to provide the database system with the tuple identifiers (TIDs) of the requested records in the query. Database systems can then access the records very quickly and take into account such limitations on the queries in the optimizer by using these TIDs.

Adding the TIDs can be done by using a temporary table that is added to a composite operation for the remaining request, or an IN list which takes over the function of the temporary table, for example. The answer of the index is exactly these TIDs. In an example from [19] as the simple request:

```
SELECT * FROM work WHERE salary > 2000
```

The primary key id (as a TID) extends the request:

```
SELECT * FROM work WHERE salary > 2000
AND id IN (4, 18)
```

Alternatively, a temporary table can be used to circumvent the length restrictions in the database:

```
SELECT * FROM work WHERE salary > 2000
AND id IN (SELECT * FROM tmp_tid_table)
```

Methods for *prepare*, *bind* and *execute* must be overloaded. Then the processing for the case of the temporary table is as follows:

1. ***on_prepare_request****: At this point, it can be decided whether the request is to be answered by the index. This query statement can be made solely on the basis of the tables, views, and attributes. If the index is to be used, the query statement is extended by the composite with the temporary table and then passed on. By the analysis of the request, the statically binding parameters are already defined.*
2. ***on_prepare_reply****: The database system returns the request identifier q or an error back as a response. In case of failure, the subsequent method calls are to check this and may make the editing on "pass through".*

3. ***on_bind_request**: If the request q should be processed through the index, calls are set to bind the parameters for dynamically binding parameters. This is important for an index exactly when the binding parameters for the index request are relevant.*
4. ***on_execute_request**: At this point, the request is fully specified. The request for the index is compiled and sent to it after that. The results of the index request are transferred into the temporary table in the database and finally the actual execute of the client is redirected.*

If the use of an IN list is aimed at, some changes arise. For example, the client's original request must be completely suppressed. All binding operations collected must be saved and finally run when *execute* is called. This is insofar difficult, since calling the event-driven methods suppresses *prepare*, also no response from the database releases. There are only two ways to deal with this situation:

1. *Sending a placeholder request whose return is used internally as an identifier for the further operation on the request to be actually indexed.*
2. *An answer from the proxy is sent directly to the client from whom the database has never become aware.*

In both cases, the actual processing is performed in *execute*. The entire operational sequence of communication for the case of the IN list is again presented in Figure 4.

Furthermore, it is possible, depending on the type of the index, to save the complete database query. In this case, the *fetch* method must also be overloaded. It will suppress all messages to the database system and finally the *fetch* generates the results suitably for the client's request from the response of the index.

3.3 Result Cache

Index integration is not the only application of this framework. It is also possible to implement a result cache. Some limitations arise, because integration into the database is not intended. There is no chance to know about data modification if it is done without the proxy. The cache is therefore relevant only for static data. Because queries that ask for the existence of tables, the type of some data records and so on can be systematically viewed as constant, they pose a broad application spectrum.

The cache is configured to specific requests. The result is stored during the first execution and played the same for other identical requests. Even requests with binding parameters can be cached in this way, if for each combination of parameter values, a separate cache entry is created.

For putting into action, the methods for *prepare*, *bind*, *execute* and *fetch* needs:

1. ***on_prepare_requrest**: It is first checked whether it is a request to be cached. If it is the first execution of the request, then a cache entry is created and the request is forwarded to the database. In a repeated execution must be maintained on the possibly binding parameters. The request is therefore retained.*
2. ***on_prepare_reply**: To save the request identifier for later recognition.*
3. ***on_bind_request**: To save binding parameters and hold back.*
4. ***on_bind_reply:** To signal success.*

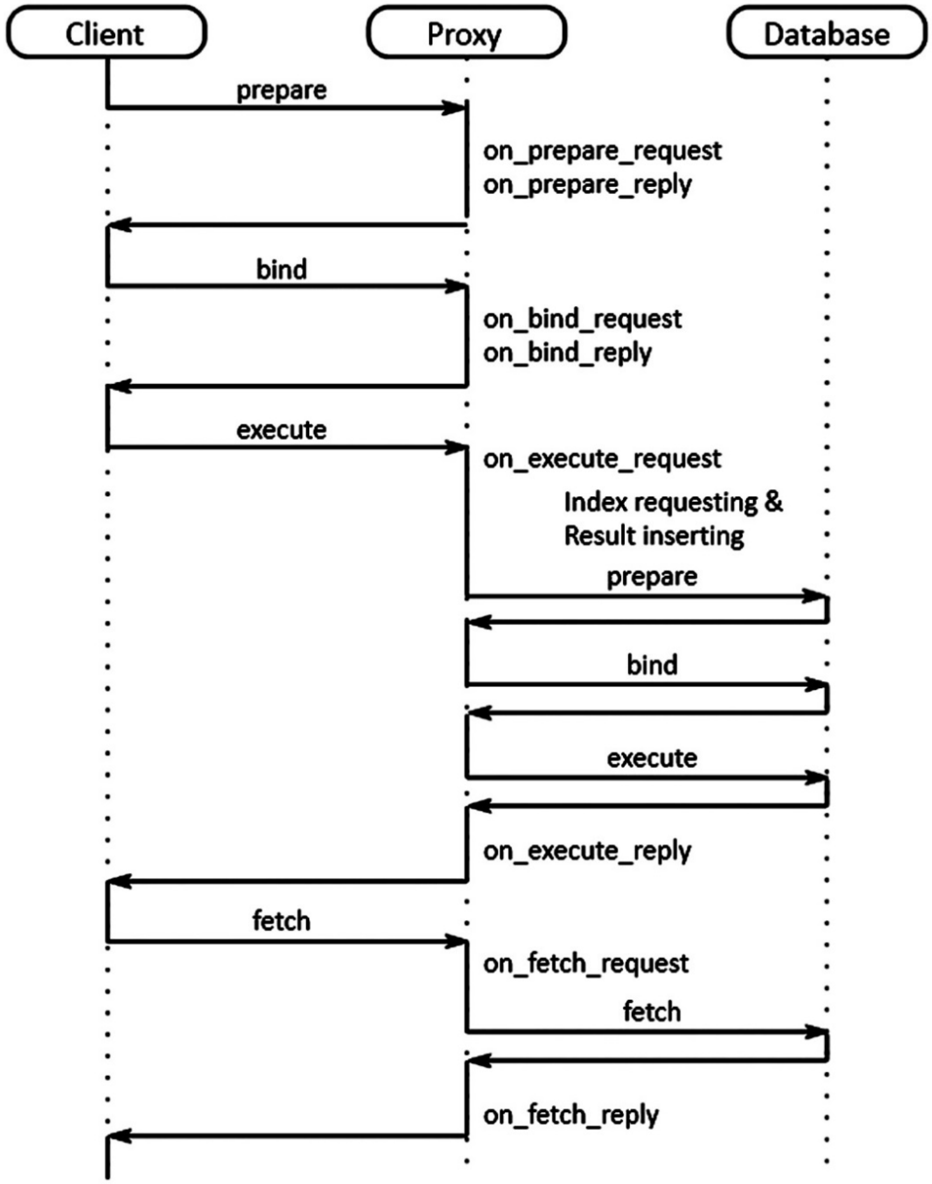


Figure 4: Flow of communications for the index integration using IN-list

1. ***on_execute_request***: First it is checked whether the request including the attached parameters is already in the cache. If this is the case, the request is suppressed to the database. In the case of a new combination of request and parameters, the database is requested.
2. ***on_execute_reply***: The result of the database is returned if requests are not in the cache, otherwise the result from the cache.
3. ***on_fetch_request***: For new requests, the database is queried.
4. ***on_fetch_reply***: For new requests, the results are recorded in the cache. Otherwise it is answered from the cache.

4 Results

The indexing just described has been implemented for the proxy framework. The index usage and the achieved gains are dependent on many factors, such as the database and the query. Thus, in a well-known example [20] an increase in speed by the use of the indexing module for the proxy framework of 32.1 seconds achieved at 0.05 second. The proxy framework required on average 479 microseconds for its part of the processing.

In addition to the time needed by the index itself to answer the requests is already covered in Figure 3, that the times for the network communication come into play twice, because the proxy is interposed. The packages those were previously able to go directly to the database system the way "Client → Database System", must now cover two routes. A scenario with which the proxy framework changed a request statically was tested. The difference in response time at the client with and without proxy was at best 5 milliseconds, the time was still averaged 15 milliseconds.

The processing times within the proxy framework (without modules caused by maturities) are however a few microseconds in the field and play in the face of network latencies (200-300 microseconds measured using **`ping(8)`**) does not matter.

Oracle TNS [21] is an exception. Since no documentation is available for TNS, this protocol had to be previously analyzed. On this occasion, there are several places in the protocol where requests within a TNS packet can be found. This results in an increased workload for this step.

An overview of the times that are needed for certain database protocols within the framework is presented in Figure 5 in the first part of the graph. For this test, requests with predicate part being longer and longer were used. The system used consisted of an Intel Core i5 660 3.33 GHz (two physical CPU cores with hyper threading), 4 GB DDR-1333 RAM under Debian 6.0 (Squeeze). At the beginning, the time required for the parser is not significant and it outweighs the time spent on the protocol.

It is clearly to recognize the increased effort for TNS in the first part of the graph. The relatively strong increase of the graph for DRDA is explained by the fact that the protocol is designed to be very inclusive, which however ensures appropriate overhead time for "normal" scenarios corresponding in demand of computation time.

Furthermore, we have changed result sets as a proof of application possibilities by means of the proxy frameworks. It was found that could not only insert the same data types, but also, depending on the client used, could create completely new result sets and types. At Clients, which evaluate the result at run time, completely different data could be transmitted, than were requested. However, if the expected data type is not checked, this is not possible. In this way, the simple request "**`SELECT 2 FROM dual`**", the result is "2", we were able to return a completely new table in sqlplus [22], which consisted of several attributes per tuple.

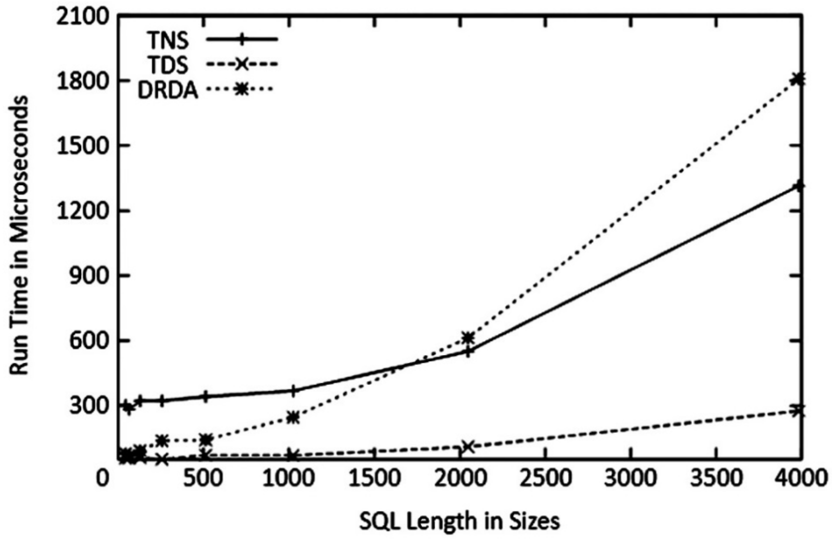


Figure 5: Processing times within the proxy framework for various database protocols and various sizes of the contained SQL queries

In addition, a function was implemented for the time analysis which measures the time of the first *prepare* up to the last response on *fetch*. If there is no access to an evaluation functionality of the database system, this is very useful. In our case, it is used to determine index candidates.

Already in section 3.1, the restriction is introduced by encryption, to which the proxy framework is subject. The encryption is end-to-end and cannot be simply bypassed. If this is desired, it must each implement an end and starting point for the encryption and they will therefore be repealed for the modules within the proxies in the proxy framework.

5 Outlook

Beside the presented proxy framework, we deal with other ways to integrate functionalities into different database systems. This also applies that a change of the database system or application is to be avoided. For integration on library level on client side, the event-driven approach is not good as there exists a way more elegant one. So no placeholder query is needed for the IN-list variant of the index integration e.g., as the return code from *prepare* can signal the successful processing. A standardization of these two so-called integration points is the goal of our further work.

As already described in [19], the challenge also positions itself with this integration variant to recognize data-changing operations, if not the whole data traffic is handled via the proxy framework described here. On this issue, which is not only specific to the proxy framework, we will discuss in the future.

REFERENCES

[1]. IBM Deutschland GmbH. DB2 Spatial Extender und Geodetic Data Management Feature – Benutzer- und Referenzhandbuch, July 2006.

[2]. C. Murray. Oracle Spatial Developers Guide, 11g Release 2 (11.2). Oracle, Dec. 2009.

[3]. IBM. SQL Reference, Volume 1. IBM Corporation, Nov. 2009.

[4]. B. Rich. Oracle Database Reference, 11g Release 2 (11.2), Sept. 2011.

[5]. K. Stolze and T. Steinbach. DB2 Index Extensions by example and in detail, IBM Developer works DB2 library. Dec. 2003.

[6]. E. Belden, T. Chorma, D. Das, Y. Hu, S. Kotsovolos, G. Lee, R. Leyderman, S. Mavris, V. Moore, M. Morsi, C. Murray, D. Raphaely, H. Slattery, S. Sundara, and A. Yoaz. Oracle Database Data Cartridge Developers Guide, 11g Release 2 (11.2). Oracle, July 2009.

[7]. R. Menon. Expert Oracle JDBC Programming. Apress, 2005.

[8]. International Standard for Database Language SQL -Part 3: Call Level Interface, 1995.

[9]. J. Blakeley. In Compcon '97. Proceedings, IEEE, title=Universal data access with OLE DB, pages 2–7, 1997.

[10]. D. Sceppa. Microsoft Ado. Net (Core Reference). Microsoft Press, 2002.

[11]. J. Melnick. Oracle Call Interface Programmer's Guide, 11g Release 2 (11.2). Oracle, Oct. 2009.

[12]. Microsoft Corporation. Tabular Data Stream Protocol, Jan. 2013.

[13]. L. Guo and H. Wu. Design and implementation of TDS protocol analyzer. In Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on, pages 633–636, 2009.

[14]. http://www.greensql.com, 2013.

[15]. The Open Group. DRDA V5 Vol. 1: Distributed Relational Database Architecture, Aug. 2011.

[16]. M. Aboelfotoh, T. Dean, and R. Mayor. An empirical evaluation of a language-based security testing technique. In Proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research, pages 112–121. ACM, 2009.

[17]. S. Leuoth, A. Adam, and W. Benn. Profit of extending standard relational databases with the Intelligent Cluster Index (ICIx). In ICARCV, pages 1198–1205. IEEE, 2010.

[18]. O. Görlitz. Inhaltsorientierte Indexierung auf Basis künstlicher neuronaler Netze. PhD thesis, 2005.

[19]. Adam, S. Leuoth, and W. Benn. Minimal-Invasive Indexintegration – Transparente Datenbankbeschleunigung. In I. Schmitt, S. Saretz, and M. Zierenberg, editors, Grundlagen von Datenbanken, volume 850 of CEUR Workshop Proceedings, pages 83–87. CEUR-WS.org, 2012.

[20]. Council, Transaction Processing Performance. TPC-H benchmark specification. Published at http://www.tcp.org/hspec.html, 2008.

[21]. D. Litchfield. The Oracle Hacker's Handbook: Hacking and Defending Oracle. John Wiley & Sons, Inc., New York, NY, USA, 2007.

[22]. S. Watt. SQL*Plus – Users Guilde and Reference, Release 10.2. Oracle, June 2005.

SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

VOLUME 2, ISSUE 5
ISSN: 2054 - 7390

Adaptive Speed Control System Based on Interval Type-2 Fuzzy Logic

Yousif I. Al Mashhadany
Electrical Engineering Department, College of Engineering, University of Anbar, Baghdad, Iraq
yousif.almashhadany@uoanbar.edu.iq

ABSTRACT

This paper presents the design and simulation of adaptive control system based on an Interval Type-2 Fuzzy Logic Controller (IT2FLC) with multilevel inverter (MLI) for control of power system, and the command line programming for constructing, editing, and simulating the IT2FLC. The modification was done by considering new membership functions (MFs) for the environment of an Interval Type-2 Fuzzy Logic System (ITFLS). Two other controllers besides the designed IT2FLC, comparing their outputs: a PID Controller and a Type-1 Fuzzy Logic Controller (T1FLC). Various design phases for the fuzzy system, from initial description to final implementation. In this work, the case study is consider for brushless DC (BLDC) motor that is driven by multilevel inverter based on discrete three-phase pulse width modulation (DPWM) generator that forced-commuted the IGBT's three-level converters using three bridges to vectored outputs 12- pulses with three levels. Using DPWM with a three-level inverter solves the problem of harmonic distortions and low electromagnetic interference. The final design system was capable to solve the overshoot problem of BLDC motors and responded better in transient and steady states and was more reliability.

Keywords: Interval Type-2 Fuzzy Logic Controller (IT2FLC), discrete pulse width modulation (DPWM), brushless DC (BLDC) motor.

1 Introduction

Brushless DC (BLDC) motor is a permanent-magnet synchronous machine with a six-transistor inverter whose on/off switching is determined by the motor's rotor position. It has neither brush nor commutator. With no possible brush/commutation failure, its torque-speed characteristic is similar to that of a permanent-magnet conventional DC motor. The BLDC motor is becoming more popular in high-performance, variable-speed drives; it needs relatively little maintenance and has low inertia, large power-to-volume ratio, is friction-less, and has less noise than does a conventional permanent-magnet DC servo motor of the same output rating. The advantages, however, are costly, and the controller of a BLDC motor is more complex than that of a conventional motor. Also, to satisfactorily operate a BLDC motor, good armature current response is necessary [1-4].

BLDC motors have higher power density than do other motors (e.g., induction motors) because there is no rotor copper loss and no commutation. Their structure is compact and robust. These contribute to their popularity in efficiency-critical applications or where commutation-induced spikes are unwelcome. Commutation necessitates the use of an inverter and a rotor position sensor, which however, can increase drive cost and machine size and reduce reliability and noise immunity [5-8].

DOI: 10.14738/tmlai.25.500
Publication Date: 05th October, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.500

Existing current-control techniques include vector control, predictive control, dead-beat control, and direct torque control; each with advantages and limits. Classical controllers are subject to variation in electrical machine parameters such as armature resistance. Intelligent fuzzy logic (FL) has been used often in controller design. The advantage of fuzzy control methods is their non-requirement for precision, unattainable in a dynamic model [9-12].

The control aim of BLDC motors (which have uncertain parameters that affect performance) is to force speed and/or current into following the reference trajectories. The problem can be alleviated through a Proportional Integral (PI) controller, which is simple to implement and common in BLDC motor control. Robust PID/PI controllers for minimum overshoot response of BLDC drives have been introduced into various applications [13-15].

This paper presents the design of Interval Fuzzy Type-2 (IT2FLC). The IT2FLC aims to solve the main problem of overshoot in BLDC motor through quick achievement of the rated speed without any overshoot, and for the motor to respond the same as first-order system response. Section 2 describes the mathematical model of the BLDC motor and its conventional controller problem. Section 3 presents the modelling of a control design that uses the IT2FLC scheme. Section 4 presents a block diagram of the design of the control system and the simulation results. Section 6 offers conclusions.

2 Mathematical Model of BLDC Motor

A BLDC motor is a rotating electric machine with a classic 3-phase stator similar to that of an induction motor. Its rotor mounts permanent magnets (see Figure 1). The magnet rotates, whereas the conductors are stationary. This motor equals a reverse-DC commutator motor. DC commutator motors have their current polarity altered by commutator and brushes, whereas BLDC motors do so through power transistors that switches synchronously with rotor position. Often, BLDC motors must therefore incorporate either internal or external position sensors that discern the actual rotor position, or use sensor-less detection [16-18].

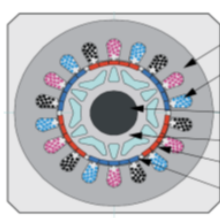


Figure 1: BLDC Motor, cross-sectioned [17]

Figure 2 is a block diagram of a BLDC motor drive. Assuming equal stator resistances of all of the windings and constant self-inductance and mutual inductance, the voltage equation of the three phases can be expressed as in Equation (1) [18], neglecting the magnets, the high-resistivity stainless-steel retaining sleeves, and the rotor-induced currents, and not modelling the damper windings [19-21].

$$\begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} = \begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix} + \begin{bmatrix} L_s - M & 0 & 0 \\ 0 & L_s - M & 0 \\ 0 & 0 & L_s - M \end{bmatrix} \frac{d}{dt} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \quad (1)$$

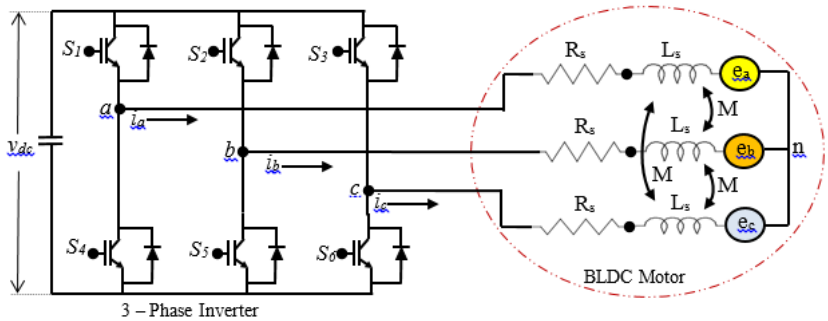


Figure 2: BLDC motor drive [20]

Here v_a, v_b, and v_c denote the phase voltages, R_s the stator resistance, i_a, i_b, and i_c the phase currents, L_s the stator inductance, M the mutual inductance, and $L=L_s–M$. The back-EMFs of the phase are e_a, e_b, and e_c. The mechanical angular velocity is w_m. Figure 3 shows that injecting a square-wave phase current into the part that has the magnitude of the back-EMFs fixed will reduce the torque ripple and stabilise control. The work presented in this paper substituted an MIFT2 controller for the current and speed controllers of a BLDC motor to obtain the deadbeat response desired in high-performance applications.

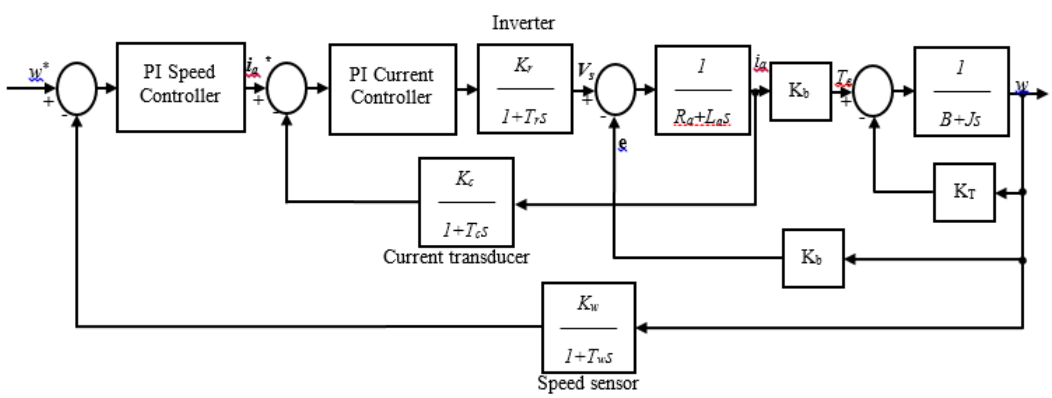


Figure 3: classical controller diagram for BLDC motor [23]

The plant transfer function varies with the operating conditions. In all the old techniques, maintaining the desired performance requires appropriate changes to the controller parameters. Figure 3 is the block diagram of a BLDC motor drive system with a conventional controller. The system uses two controllers: one in the inner loop (for current control), another in the outer loop (for speed control). Both controllers are replaced by one intelligent controller that does not require tuning - which increases response accuracy and overcomes the problem of in-operation-tuning of the controller parameters [22-27].

3 Modelling and Design Of IT2FLC System

Figure 5 shows the structure of IT2 direct reasoning with IT2 fuzzy inputs. An Fuzzy Inference System (FIS) is a rule-based system that uses FL instead of Boolean logic. Its basic structure has four components (see Figure 4) [28-30].

Characterising a T-2 fuzzy set is not as easy as characterising a T-1 fuzzy set. A T-2 fuzzy set, denoted by $A^{\sim}$, is characterised by a T-2 MF $\mu_{A^{\sim}}(s,u)$ where $x \in X$ and $u \in J_x \subseteq [0,1]$, i.e.,:

$$A^{\sim} = \{((x,u), \mu_{A^{\sim}}(x,u)) \mid \forall u \in J_x \subseteq [0,1]\} \tag{2}$$

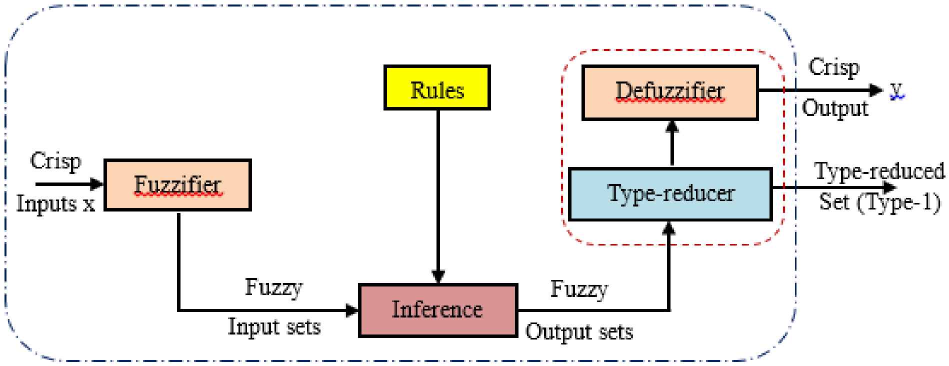


Figure 4: Structure of IT-2 FIS

The domain of a secondary MF is called the primary membership of x, where $J_x \subseteq [0,1] \forall u \in X$. If X and J_x are both discrete, then Equation (3) can be expressed as

$$\left.\begin{aligned} A^{\sim} &= \sum_{x\in X} \frac{\left[\sum_{u\in J_x} f_x(u)/u\right]}{x} = \sum_{i=1}^{N} \frac{\left[\sum_{u\in J_{x_i}} f_{x_i}(u)/u\right]}{x_i} \\ &= \frac{\left[\sum_{K=1}^{M_1} fx_1(u_{1k})/u_{1k}\right]}{x_1} + \dots + \frac{\left[\sum_{K=1}^{M_N} fx_N(u_{Nk})/u_{Nk}\right]}{x_N} \end{aligned}\right\} \tag{3}$$

Uncertainty $A^{\sim}$ in the primary memberships of a T-2 fuzzy set consists of a bounded region (FOU):

$$FOU(A^{\sim}) = \underset{x\in X}{U} J_x \tag{4}$$

The concept of FOU, associated with the concepts of lower and upper MFs, allows easy characterising of T-2 fuzzy sets. FOU models uncertainties in the shape and position of a T-1 fuzzy set. T-2 Gaussian MF is obtained by blurring a T-1 Gaussian MF with mean m_k and standard deviation σ_k. Consider the case of a Gaussian primary MF with a fixed mean m_k and an uncertain standard deviation that takes on values in $[\sigma_{k1}, \sigma_{k2}]$, i.e.,

$$\mu_A(x) = \exp\left[-\frac{1}{2}\left(\frac{x-m_k}{\sigma_k}\right)^2\right]; \quad \sigma_k \in [\sigma_{k1}, \sigma_{k2}] \tag{5}$$

Different membership curves for each of the two σ_k (σ_{k1}, σ_{k2}) values can be calculated. The uniform shading for the FOU again denotes interval sets for secondary MFs and represents the entire T-2 fuzzy set $\mu_A(x, u)$ interval. The FOU can be described in terms of upper and lower MFs. The modified MFs are bounds for the FOU of an IT2 fuzzy set $A^{\sim}$. The upper MF is associated with the upper bound of the FOU ($A^{\sim}$) and is denoted $\overline{\mu}_{A^{\sim}}(x), \forall_x \in X$.

$$\overline{\mu}_{\tilde{A}}(x) \equiv \overline{FOU(\tilde{A})} \quad \forall_x \in X \quad and \quad \underline{\mu}_{\tilde{A}}(x) \equiv \overline{FOU(\tilde{A})} \quad \forall_x \in X \tag{6}$$

Because the domain of a secondary MF has been constrained in (1.0) to be contained in [0, 1], the lower and the upper MFs always exist. Note that J_x is an interval set:

$$J_x = \{(x,u) : u \in \underline{[\underline{\mu}_{\tilde{A}}(x), \overline{\mu}_{\tilde{A}}(x)]}\} \tag{7}$$

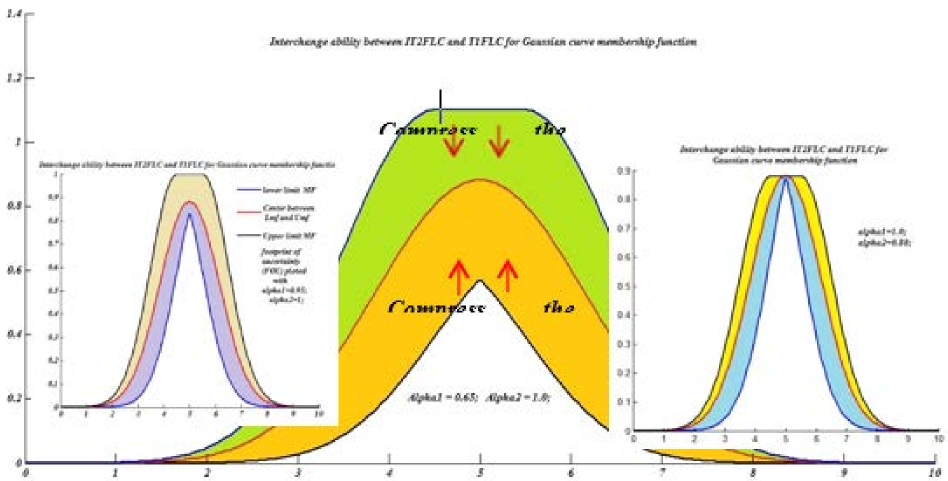


Figure 5: modified MF that used with IT2FLC

This work used as a tuning start point the embedded IT2-FIS located in the middle of each FOU. Once the MIT2-FSs had been converted to IT2-FIS, the MF parameters were tuned in Matlab, to calculate and plot the upper and the lower limits of the change. Figure 5 presents the limits of the modified MF that used with IT2FLC in this work (see Figure 5).

Unlike conventional control, which is based on a plant's mathematical model, FLC usually embeds the intuition and experience of a human operator and sometimes those of designers and researchers. When controlling a plant, a skilled human operator's aim is to manipulate the process input (i.e., controller output) based on e and Δe, as fast as possible and with the least error. The variable of a fuzzy controller is u(t). The MFs of the input and output variables must be considered once the fuzzy controller inputs and outputs are chosen. This paper defines all MFs for conventional fuzzy controller inputs (e and Δe), and the controller output, on a common normalised domain [-1, 1]. Symmetric triangles (except the two MFs at the extreme ends) were used, with an equal base and 50% overlap with the neighbouring MFs. This choice is the most natural and unbiased for MFs. Figure 6 shows the seven MFs: MN (Most Negative), NB (Negative Big), NM (Negative Medium), NS (Negative Small), ZE (Zero Error), PS (Positive Small), MP (Most Positive), PM (Positive Medium), and PB (Positive Big).

The rule base was designed next. If the inputs had 9 MFs, the corresponding rules are 9^2=81. The MF was calculated and plotted for two types of fuzzy controller (type-1, interval type-2). The area under the curve of the IT2FLC MF was wide during transient period but narrow at the envelope end - the modification eliminated the overshoot from the speed response (which also became more stable at steady state).

For T-1 fuzzy controller input voltage, the control law is:

$$u = \frac{\sum_{j=1}^{M} f_j u_j}{\sum_{j=1}^{M} f_j}; \qquad \text{where}: \quad f_j(x) = \mu_{A_{j1}}(x_1) * \ldots\ldots * \mu_{A_{jn}}(x_n) \tag{8}$$

For T2-FLC input voltage,

$$U(F_1,\ldots F_M) = \int_{f_1} \ldots \int_{f_M} \tau_{j=1}^{M} \mu_{F_j}(f_j) \Bigg/ \frac{\sum_{j=1}^{M} f_j u_j}{\sum_{j=1}^{M} f_j}; \tag{9}$$

For traditional IT2-FLC input voltage,

$$U(F_1,\ldots F_M) = \int_{f^1 \in [\underline{f}^1, \overline{f}^1]} \ldots \int_{f^M \in [\underline{f}^M, \overline{f}^M]} 1 \Bigg/ \frac{\sum_{j=1}^{M} f_j u_j}{\sum_{j=1}^{M} f_j}; \tag{10}$$

And for IT2-FLC, according to the modified MF:

$$U(F_1,\ldots F_M) = \int_{f^1 \in [\underline{f}^1, \overline{f}^1]} \ldots \int_{f^M \in [\underline{f}^M, \overline{f}^M]} 1 \Bigg/ \frac{\sum_{x=1}^{M} f_x J_x}{\sum_{x=1}^{M} f_x}; \qquad \text{where}: J_x = \{(x,u) : u \in [\underline{\mu}_{\tilde{A}}(x), \overline{\mu}_{\tilde{A}}(x)]\} \tag{11}$$

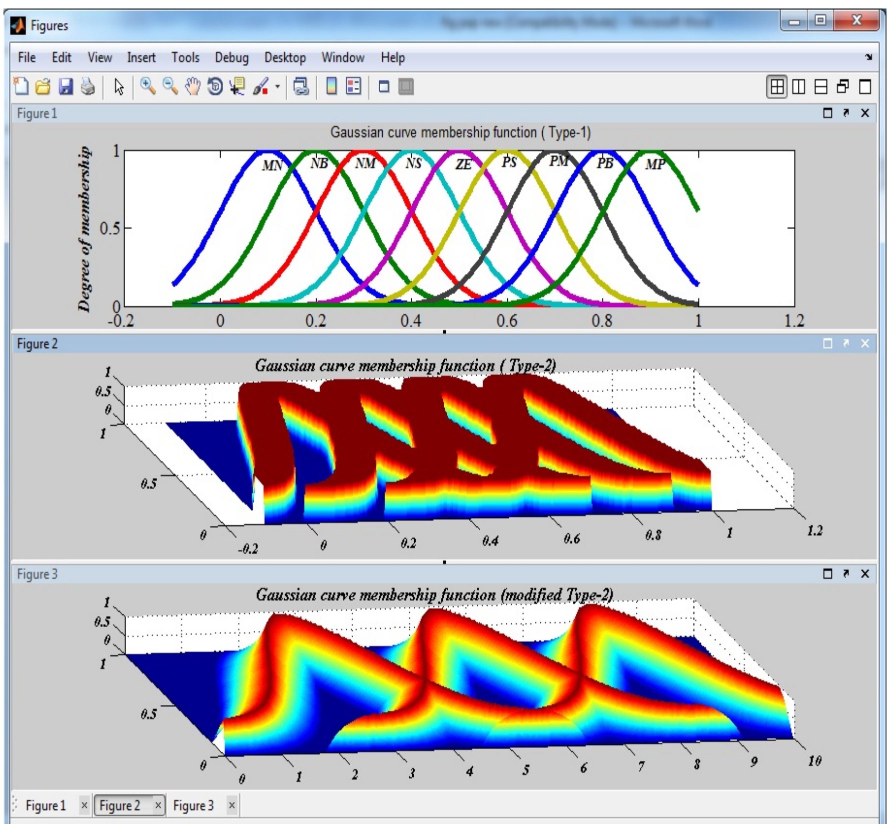


Figure 6: Gaussian curve MFs of various fuzzy controller types

4 Design Simulation and Analysis

The simulation of the system design was done by using Matlab / Simulink Ver. 2013a, this simulation starts with described and simulation for the three-phase multilevel inverter (MLI) that is fed to the BLDC motor is shown in Figure 7. This model represents modeling a 50 kW, 380 V, 50 Hz, three-phase, three-level inverter. The IGBT inverter uses the discrete three-phase pulse width modulation (DPWM) technique (8 kHz carrier frequency) to convert DC power from a +/− Vdc source to V AC, 50 Hz. The inverter feeds a 50 kW resistive load through a three-phase transformer. L–C filters are used at the converter output to filter out harmonic frequencies generated mainly around multiples of 8 kHz switching frequency. The 12-inverter pulses required by the inverter are generated by the discrete three-phase PWM generator. The system operates in open loop at a constant modulation index. The inverter is built with individual IGBTs and diodes.

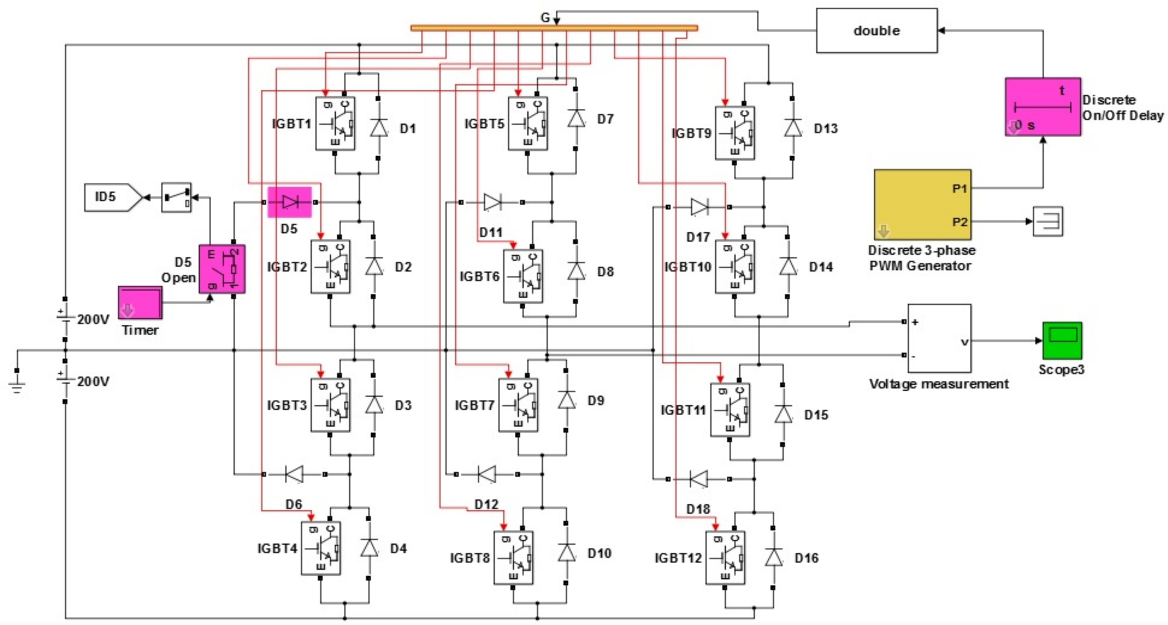


Figure 7: Three-phase three-level inverter with DPWM

In a three-level voltage-sourced converter (VSC) using ideal switches, the two pairs of pulses sent to each arm could be complementary. For example, for phase A, IGBT1 is complementary of IGBT3 and IGBT2 is complementary of IGBT4. However, in practical VSCs, the turnoff of semiconductor switches is delayed because of the storage effect. Therefore, a time delay of a few microseconds (storage time + safety margin) is required to allow complete extinction of the IGBT that is switched off before switching on the other IGBT.

The proposed model of multilevel technique synthesizes the AC output terminal voltage with low harmonic distortion, thus reducing the filter requirements. In particular, MLIs are emerging as a visible alternative for high-power, medium-voltage applications. One of the significant advantages of multilevel configuration is the harmonic reduction in the output waveform without increasing the switching frequency or decreasing the inverter power output. In this simulation used same numerical values for system design to make compression, table 1 presents the numerical values.

Table 1. The numerical values for system design.

parameter	value	parameter	value
Stator phase resistance Rs (ohm):	2.8750	Torque Constant (N.m / A_peak)	1.4
Stator phase inductance Ls (H) :	8.5e-3	Back EMF flat area (degrees)	120
Flux linkage established by magnets (V.s)	0.175	Inertia, viscous damping, pole pairs, static friction [J(kg.m^2) F(N.m.s) p() Tf(N.m)]:	[0.8, 1e-3 1e-3, 4]
Voltage Constant (V_peak L-L / krpm)	146.6	Initial conditions [wm(rad/s) thetam(deg) ia,ib(A)]	[0,0,0,0]
The parameters of PI controller are: Proportional K_P= 25, Integral k_I=0.13.			

The output voltage waveform of an MLI is composed of the number of levels of voltages, typically obtained from capacitor voltage sources starting from three levels, the number of levels can increase until the output is a pure sinusoidal. The output of the simulation for system design for MLI model described in Figure 7 and used with BLDC motor is shown in Figure 8. The output of this inverter based on DPWM with 12 pulses sequences, Figure 9 explains the sample from the sequences used for operation MLI Otherwise, a short circuit could result on the DC bus.

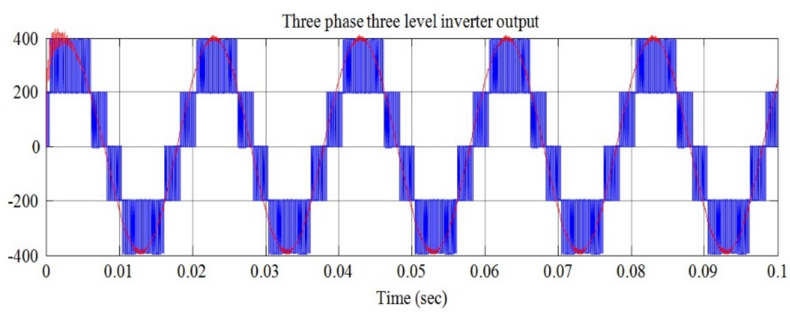


Figure 8: Output of three-phase three-level inverter with DPWM.

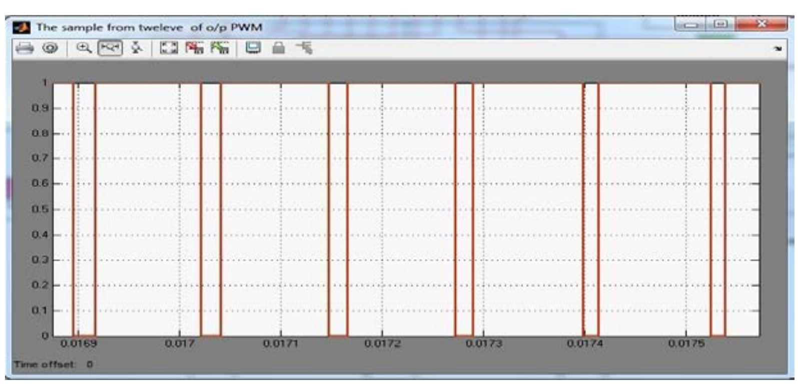


Figure 9: The sample from output of the DPWM

To examine the feasibility and validity of the proposed controller, IT2FLC was applied to the speed control of a BLDC motor. The simulation was applied with three types of controllers and compared between them according to the settling times, rising time and Max. Overshoots of the responses. Performance indices for the controller feasibility were absolute error and root mean square error:

$$AE(k) = |y(k) - y_d(k)| \tag{12}$$

$$RMS(k) = \sqrt{(y(k) - y_d(k))} \tag{13}$$

Where: *y(k)* represents the output of the plant while $y_d(k)$ represents the desired output.

- *Simulation of the classical controller*

Figure 3 shows the control system having two loops: current control loop, which controls motor torque, and speed control loop, which adjusts motor speed. The model can be used in various researches, but identifying most parameters of the current controller and speed controller needs fuzzy controller Type 1 or Type 2. Figure 9 gives the structure and simulation of a classical controller (PI) and fuzzy controller Type 1. The BLDC motor was modelled and analysed on Matlab/Simulink.

The inverter gate signals are produced by decoding the motor Hall Effect signal. The inverter three-phase outputs are applied to the PMSM-block stator windings. The load torque applied to the shaft of the machine is set to 0 and then (at t=0.1s) stepped to its nominal value (11Nm). The inner loop synchronises the inverter gate signals and electromotive force (*emf*). The outer loop controls the motor speed through variation in the DC bus voltage. With this design and taking the best results among trials with many parameter sets, the final PI parameter values post tuning were K_i=0.13 and K_p=25. Figure 9 gives the simulation results of the classical controller, whose most advantageous responses were quick settling and a good steady-state error of 0.99% o/p w_m=2982 (see Figure 9).

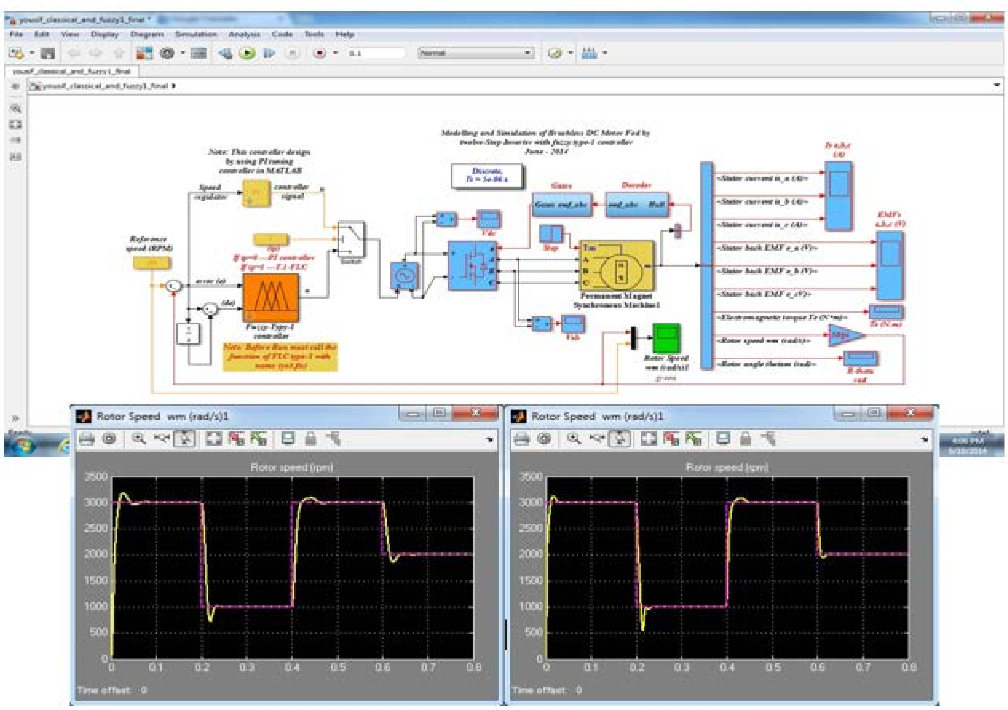


Figure 10: Simulation of classical and Fuzzy type-1controller of BLDC motor

For simplicity and accuracy, the analysis of the BLDC motor assumed an unsaturated BLDC motor, equal stator resistances, constant self-inductance, and constant mutual inductance, ideal semiconductor devices of the inverter, negligible iron losses, and equal back-EMF waveforms for all of the phases. The assumptions were based on the

Simulation of Fuzzy (IT2FLC, T1FLC) controller for the BLDC motor

Figure 10 is the block diagram for T1FLC and IT2FLC with two inputs (e_1, e_2) and one output (u). The error signal is calculated by subtracting the reference speed from the actual rotor speed; the same for all other types of FLC:

$$e_1(k) = w_r(k) - w_m(k) \tag{15}$$

with e_1 being the error, w_r the reference speed, and w_m the actual motor speed. The change in error (e_2(k)) is calculated by Eqn. (16), with e_1(k-1) being the preceding error value.

$$e_2(k) = e_1(k) - e_1(k-1) \tag{16}$$

The FLC system defines two normalisation parameters (e_{1N}, e_{2N}, for input) and one de-normalisation parameter (u_N, for output). In normalisation, the input values are scaled between (-1, +1). In de-normalisation, the output value of the fuzzy controller is converted to a value that depends on the terminal control element. The fuzzy values obtained from the fuzzy inference mechanism have to be defuzzified into a crisp output value (u). A Gaussian fuzzy MF is thus defined for each input value and each output value by the nine clusters.

The Modified Matlab Fuzzy Toolbox depended on T-2 Fuzzy Logic Toolbox. The IT2FLC was simulated in Matlab GUI, which enabled calculation and design of the main FIS structure of the T-1 fuzzy controller. The FIS model was also calculated and designed through the Generalised Fuzzy System (GFS) algorithm. Describe the design and simulation of the two controller types (classical, T1FLC).

The first window introduces the idea behind the IT2FLC controller and allows the simulation windows of all the designs to be called. The next window allows each of the controller types to be called. The third window runs the classical controller; it enables analysis of the BLDC motor mathematical model and tuning of the PID controller.

The fourth window allows running of T1FLC and IT2FLC processes. The fifth window presents the MIT2FLC design. The modified MF is advantageous mainly through its adjustable parameters; also, it is not severely limited (allowing the controller to achieve the desired response without any overshoots). The wide-ranging MF base overcomes the initial-operation (during transient response) problem of the BLDC motor.

The second button in the fifth GUI (see Figure 10) allows converting the TIFLC to IT2FLC before modifying it into MIT2FLC. The window enables calling of Type-1 rules and then changing them into IT2FLC by calculating the upper and the lower FOU for each MF. The third button enables plotting of the surface error of the new design.

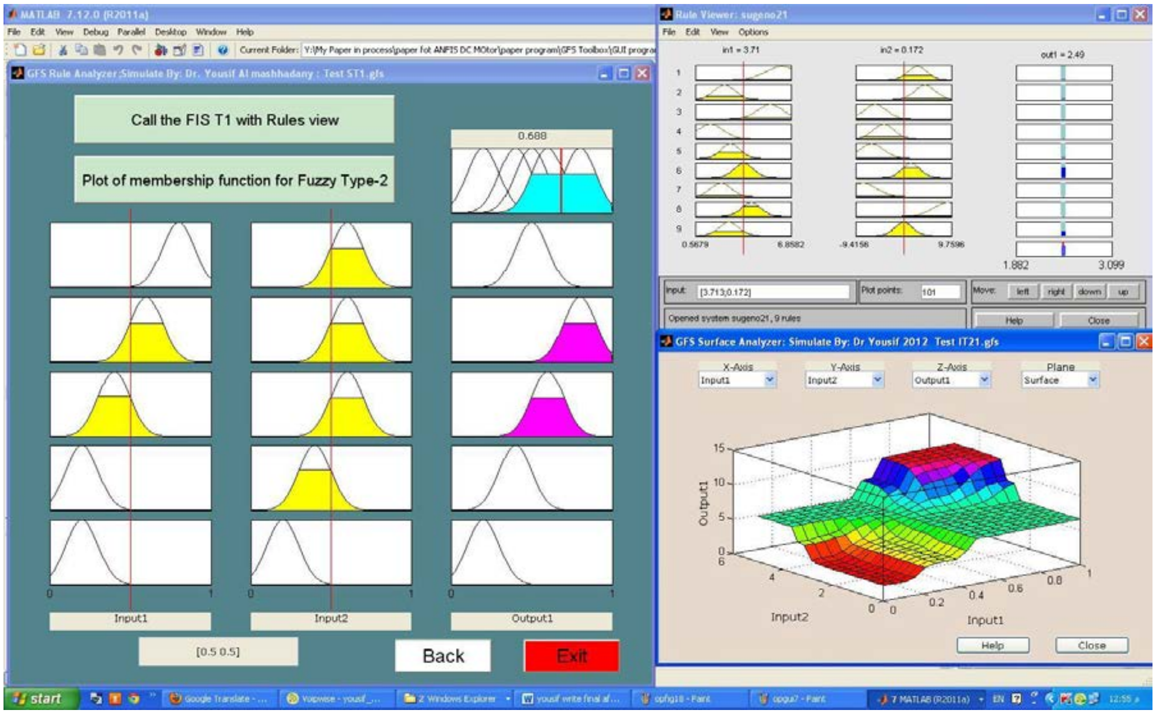


Figure 11: The GFS, the rule analyzer, the viewer

The fourth button reveals the final design of the IT2FLC, whose simulation block diagram (used to select the switching technique for the two controllers in the same plant) is as of Figure 11. Figure 12 gives the design outputs: four outputs for the two controllers, each subjected to the same simulation conditions of the BLDC motor. The IT2FLC showed better response than did the other controllers; its overshoot value was the smallest (overcoming the problem of overshooting), and the settling time was optimal.

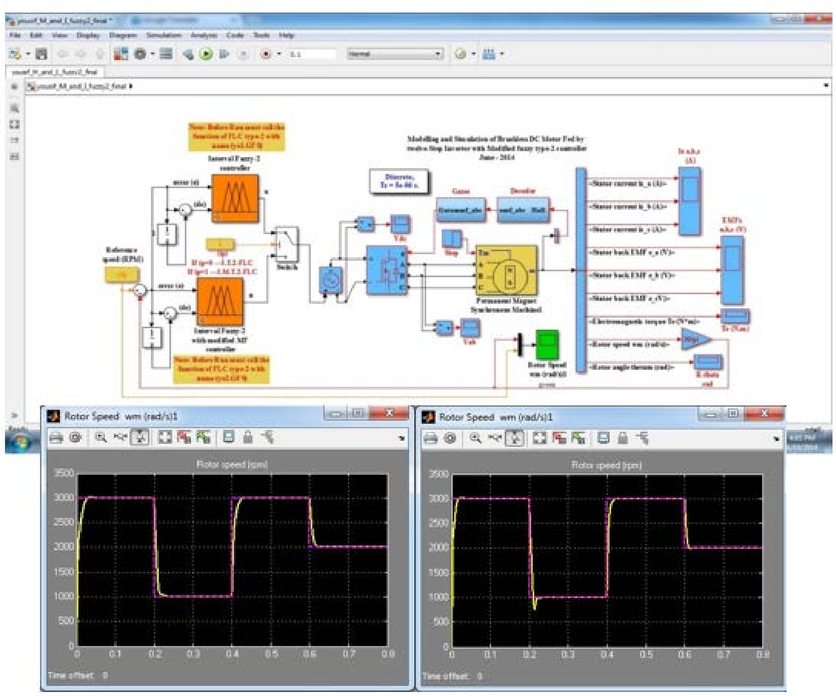


Figure 12: Final simulation of the IT2FLC and IT2FLC with modified MF function

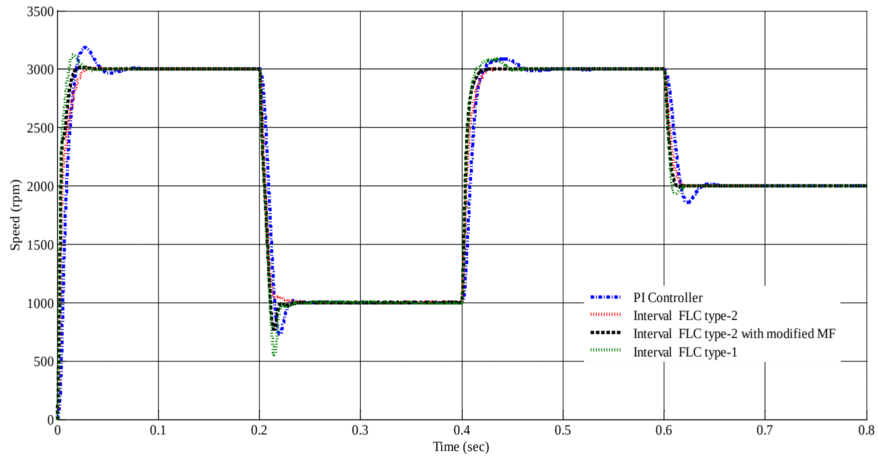


Figure 13: Compression of the controller outputs of the BLDC motor

5 Conclusions

The time series results show the ability of adaptive control system as direct control for a power system through an interval type-2 fuzzy controller. They also compared the results of the three controller types. The design and implementation performed first by using tuning classical controller (PI), than the IT1FLS Toolbox is potentially important for research on interval type-2 fuzzy logic, as the proposed Toolbox design model for the IT2 fuzzy controller is for solving complex problems in various applications. Our future work is to improve the IT2FLS Toolbox with the power-system-toolbox-based GUI for all types of controllers. For now,

- *The IT2FLS has shown high performance that outperforms all of the other controllers in terms of speed response and minimum overshoot.*
- *The GUI designed here has shown high functionality and ease of use (the easiest method by far) for the simulation of the IT2FLC for a BLDC motor.*
- *Future work is planned for a full toolbox that is able to simulate any machine controller.*

ACKNOWLEDGEMENT

Special thanks to the University of Anbar Iraq / Renewable Energy Research Center for supporting this work through Grant No. RERC-PT28.

REFERENCES

[1] A.S. Oshaba, E.S. Ali, “Swarming Speed Control for DC Permanent Magnet Motor Drive via Pulse Width Modulation Technique and DC/DC Converter”, Research Journal of Applied Sciences, Engineering and Technology, Vol. 5, I.18, ISSN: 4576-4583, 2013

[2] S. Dubey, S. K. Srivastava, “Speed Control of Dc Motor Using Fuzzy Logic Controller”, IJBSTR, ISSN 2320 – 6020, Vol. 1, I. 7, 2013

[3] C.Ganesh, M.Prabhu, M.Rajalakshmi, G.Sumathi, V. Bhola, S.K.Patnaik, "ANN Based PID Controlled Brushless DC drive System", Proc. of Int. Conf. on Advances in Electrical & Electronics, © 2011 ACEEE, DOI: 02.AEE.2011.02.83, 2011

[4] M. Rajshekar, V. G. Swamy, T.A. Kumar, ""Modeling and \simulation of discontinues current mode inverter fee permanent magnet synchronous motor derive", Journal of Theoretical and Applied Information Technology, JATIT & LLS. All rights reserved. 2011, pp. 64-94

[5] H. ZENK1, A. S. Akpinar, "Dynamic Performance Comparison of Cúk Converter with DC Motor Driving and Using PI, PID, Fuzzy Logic Types Controllers", Universal Journal of Electrical and Electronic Engineering, Vol. 2, I.2, pp: 90-96, 2014

[6] K.H. Chua, W.P. Hew, C.P. Ooi, C.Y. Foo, K.S. Lai, "A Comparative analysis of PI, fuzzy logic and ANFIS speed control of permanent synchronous motor", ELEKTROPIKA: Int. J. of Electrical, Electronic Engineering and Technology, Vol. 1, , 2011, pp 10-22.

[7] A. Ismail; B. O. Faruk, "A type-2 fuzzy logic controller design for buck and boost DC-DC converters", journal of intelligent manufacturing, Vol. 23, Issue. 4, 2012, pp. 1023 - 1034

[8] C. W. Chen, "Modeling, control, and stability analysis for time-delay TLP systems using the fuzzy Lyapunov method", journal of neural computing and applications, Vol. 20, Isuue. 4, 2011, pp. 527-534

[9] P. A. Adewuyi, "DC Motor Speed Control: A Case between PID Controller and Fuzzy Logic Controller", International Journal Of Multidisciplinary Sciences And Engineering, Vol. 4, No. 4, 2013

[10] S. Saini, A. Kumar, "Speed Control of Separately Excited D.C Motor using Self Tuned ANFIS Techniques", International Journal of Computer Science and Technology, IJCST Vol. 3, Issue 1, Jan. - March 2012, pp. 500-504.

[11] B. A. Omar, A. Y. Haikal, F. F. Areed, "An Adaptive Neuro-Fuzzy Speed Controller for a Separately excited DC Motor", International Journal of Computer Applications, Vol. 39– No.9, February 2012 ,pp. 0975 – 8887

[12] V.M. Varatharaju, B. Mathur, U. Dhayakumar, "Adaptive Controllers for Permanent Magnet Brushless DC Motor Drive System using Adaptive-Network-based Fuzzy Interference System", American Journal of Applied Sciences 8 (8), ISSN 1546-9239, 2011, pp. 810-815

[13] M. Pratama, S. Rajab, E. M. Joo, "Extended Approach of ANFIS in Cascade Control", International Journal of Computer and Electrical Engineering, Vol. 3, No. 4, August 2011, pp. 572-576.

[14] J. Chakravorty, R. Sharma, "Fuzzy Logic Based Method of Speed Control of DC Motor", International Journal of Emerging Technology and Advanced Engineering, Vol. 3, I. 4, 2013

[15] M. R. Dastranj, E. Ebrahimi, N. Changizi, E. Sameni, "Control DC Motorspeed with Adaptive Neuro-Fuzzy control (ANFIS)", Australian Journal of Basic and Applied Sciences, 5(10): ISSN 1991-8178, 2011, pp. 1499-1504

[16] V. M. Varatharaju, B. L. Mathur, " Adaptive neuro-fuzzy controller for hysteresis current controlled PMBLDC motor", International Journal of Advances in Engineering & Technology, IJAET ISSN: 2231-1963, 212 Vol. 1, Issue 4, Sept 2011, pp. 212-223

[17] L. Leottau, M. Melgarejo, "An Embedded Type-2 Fuzzy Controller for a Mobile Robot Application", chapter 18, Recent Advances in Mobile Robotics, 2011, pp.365-384

[18] A.A Thorat, S. Yadav, S.S.Patil, "Implementation of Fuzzy Logic System for DC Motor Speed Control using Microcontroller", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 3, I. 2, pp.950-956, 2013

[19] T. Dereli , A. Baykasoglu , K. Altun , A. Durmusoglu , I. B. Turksen, "Industrial applications of type-2 fuzzy sets and systems: A concise review", Elsevier, Journal of Computers in Industry, 62, 2011, pp.125–137

[20] X. Ling, Y. Zhang, "Operations on Triangle Type-2 Fuzzy Sets" Elsevier, Journal of Advanced in Control Engineering and Information Science,12, 2011, pp. 3346-3350

[21] M. Tushir, S. Srivastava, "Type-2 fuzzy logic controller Implementation for tracking control of DC motor", International Journal of Computer Network and Security (IJCNS) ISSN : 0975-8283,Vol. 3 No. 1, 2011, pp. 34-41

[22] R. C. Chourasia, M. Kumar, "Speed Control of S.E.D.C. Motor by Using PI and Fuzzy Logic Controller", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Vol.3, I.2, 2013

[23] M. Manceur, N. Essounbouli, A. Hamzaoui, "Higher Order Sliding Fuzzy Type-2 Interval Control for SISO Uncertain Nonlinear Systems", IEEE International Conference on Fuzzy Systems June 27-30, 2011, Taipei, Taiwan, pp. 1388 – 1396

[24] Y. Maldonado, O. Castillo, P. Melin, "Optimal Design of Type-2 Fuzzy Controllers with a Multiple Objective Genetic Algorithm for FPGA Implementation", 978-1-61284-968-3/11/$26.00 ©2011 IEEE xxx

[25] P. C. Chen, C. W Chen, W. L.Chiang, "Linear Matrix Inequality Conditions of Nonlinear systems by Genetic Algorithm-based H-infinity Adaptive Fuzzy Sliding Mode Controller", Vol. 17, Isuue. 4, 2011, pp. 163-173

[26] C.S. Chen, "Supervisory adaptive tracking control of robot manipulators using interval type-2 TSK fuzzy logic system", IET Control Theory & Applications, Volume 5, Issue 15, 13 October 2011, p. 1796 – 1807

[27] X. Du, H. Ying, "Derivation and Analysis of the Analytical Structures of the Interval Type-2 Fuzzy-PI and PD Controllers", IEEE transactions on fuzzy system, Vol, 18, No. 4, Aug. 2010, pp. 802-814

[28] M. Biglarbegian, W. W. Melek, J. M. Mendel, "Design of Novel Interval Type-2 Fuzzy Controllers for Modular and Reconfigurable Robots: Theory and Experiments", IEEE transactions on industrial electronics, Vol. 58, No. 4, April 2011, pp. 1371-1384

[29] M. Manceur, N. Essounbouli, A. Hamzaoui, "Higher Order Sliding Fuzzy Type-2 Interval Control for SISO Uncertain Nonlinear Systems", IEEE International Conference on Fuzzy Systems, June 27-30, 2011, Taipei, Taiwan, pp. 1388-1396

[30] E. Kayacan, O. Kaynak, R. Abiyev, J. Tørresen, M. Høvin, K. Glette, "Design of an Adaptive Interval Type-2 Fuzzy Logic Controller for the Position Control of a Servo System with an Intelligent Sensor", WCCI fuzzy IEEE World Congress on Computational Intelligence July, 18-23, Barcelona, Spain, 2010, pp. 1125-1132

Society for Science and Education United Kingdom
TMLAI Transactions on Machine Learning and Artificial Intelligence
Volume 2, Issue 5
ISSN: 2054 - 7390

Modeling the Financial Market with Multiple Prices

Zhao Danting, Tan Tingfeng, Jin Zhongwei, Xuan Huichao, Li Xian and Wang Qing-Guo
Department of Electrical and Computer Engineering, National University of Singapore, Singapore, 117576.
sendy87716@163.com; a0123784@nus.edu.sg; jinzhongwei@nus.edu.sg; xuanhuichao@nus.edu.sg; lixian@nus.edu.sg; elewqg@nus.edu.sg

ABSTRACT

An effective financial market trading decision is usually dependent on superior forecasting. Forex market as the largest financial market is chosen in this study. The main objective of this paper is to explore the forecasting performance of the proposed multiple-price model which integrates close, low and high price information, by using Artificial Neural Network (ANN). The architecture of the network and the related algorithms are described. The effects due to different choices of preprocessing methods, combinations of input variables and different time intervals of forecasting are examined. By using the best multiple-price model, trading strategies with high and low prices are developed as well. The results have shown that in terms of both absolute values and trends of the prices, forecasting accuracy has improved compared with single price model. This is especially so for low and high prices whose directional accuracies are much higher. The trading performance is also proven to have much better total return than buy & hold strategy, and trading with high price has the best overall performance considering both return and risk.

Keywords: *Foreign Exchange Rate, Artificial Neural Network, Trading, Technical Indicator, Fundamental Indicator*

1 Introduction

Forecasting future value of financial market time series has been extensively explored and developed. Foreign Exchange market as the largest and most liquid financial market [1] is studied in this paper. Studies making use of Hurst Exponent [2], a measure of the bias in fractional Brownian motion, have indicated that most of the financial markets are not random walk and are not highly efficient, thus forecasting of financial indices is possible. Technical and fundamental approaches are two of the most popular disciplines in financial market forecasting. As a result, in this paper, a combination of fundamental and technical factors will be taking into consideration in financial market forecasting.

Artificial Neural Network (ANN) known as universal function approximators can map any nonlinear function [3]. It is a non-parametric model which is data driven. So, no parametric modeling assumptions need to be made and the entire data set can be utilized. Furthermore, neural networks are less sensitive to error term assumptions and they can tolerate noise, chaotic components, and heavy tails which are common characteristics in financial market, better than most other methods [4].

DOI: 10.14738/tmlai.25.446
Publication Date: 14th October, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.446

With regards to existing studies on forex market forecasting using ANN, only the closing currency exchange rate is used. Apart from the closing price, the high and low prices can play important roles in making investment decisions. More importantly, High, Low and Close prices should all have effects on one and another. Forecasting accuracy might improve in the multiple-price model. Therefore, the major objective of this paper is to explore the accuracy and trading performance of financial forecasting using models with integration of multiple prices and other inputs. Also, different choices of preprocessing methods on the raw data, combinations of input variables with technical and fundamental indicators, as well as different time intervals of forecasting may affect forecasting performance of the multiple-price model and will be investigated in this paper.

2 Artificial Neural Network

The Back propagation (BP) multilayer feed forward neural network is the most widely used type of neural networks in financial time series forecasting [5]. It has been shown that if sufficient number of hidden neurons is used, standard Back propagation Neural network using an arbitrary transfer function can approximate almost any measurable function in a very precise manner [6].

2.1 The Proposed Multiple-price Model (MPM) Structure

The standard ANN structure for financial time series forecasting only has one output which represents the price when the market closes, and it is called Single-price Model (SPM) in this paper. On the other hand, in the proposed Multiple-price Model (MPM), there will be three outputs which are for close, low and high prices at time instant t+1 respectively. Figure 1 demonstrates the structure of the MPM.

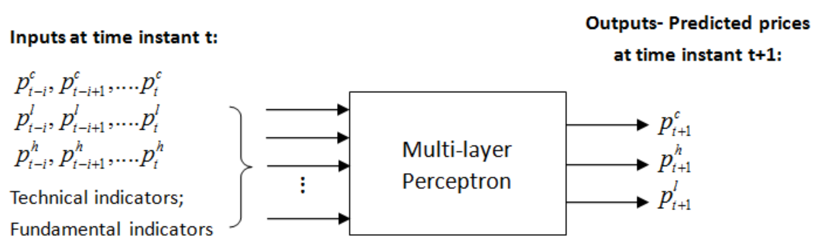


Fig. 1. Multiple-price model structure

2.2 Multilayer Perceptron Architecture

The graphical representation of an MLP with one hidden layer is shown in Figure 2. There are two types of links in the network, Synaptic link and Activation link [7]. Synaptic links connect the neurons from the previous layer to the neurons in the next layer, associated with the weight $w^{(l)}_{ji}$ - weight from unit i in layer (*l*-1) to unit j in layer (*l*). Input signal from unit j in layer (*l*) to neurons in layer (*l*+1) is denoted as $x^{(l)}_j$, and it can also be viewed as the output signal of unit j in layer (*l*). The bias terms at layer (*l*) which are always equal to one are noted as $x^{(l)}_0$. The input signals passing through the synaptic links will then be summed together with their associated weights at the junction. The activation link connects the summing junction and the respective output neuron. Therefore, the outputs from the hidden layer and output neurons are:

$$x_j^{(1)} = \varphi^{(1)}\left(\sum_{i=0}^{n_0} w_{ji}^{(1)} x_i^{(0)}\right), \ where \ j = 1, 2, ..., n_1 \quad (1)$$

$$x_j^{(2)} = \varphi^{(2)}\left(\sum_{i=0}^{n_1} w_{ji}^{(2)} x_i^{(1)}\right), \ where \ j = 1 \ for \ \text{Single-price Model}, \ and \ j = 1, 2, 3 \ for \ \text{Multiple-price Model}. \quad (2)$$

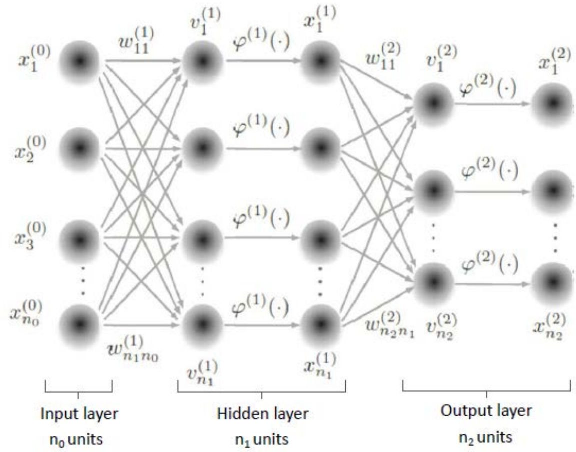


Fig. 2. Three layer MLP architecture

2.3 Data Division and Normalization

In Neural network applications, it is a common practice to split all the available sample data into training, validation and testing sets. The training set is used to adjust the weights of neural network model during the training process. The validation set is used in the training process to monitor the performance of the network after each update of the weights and stops the training when the performance gets worse, which is a way to avoid the over-fitting problem. The testing set is used to evaluate the generalization ability of the trained network and it is evaluated via the performance measures. According to the work of Yao and Tan [8], the training, validation and testing data sets shall be set 70%, 20% and 10% of the collected sample data respectively. In this project, this data division method will be adopted.

Data normalization is one of the most important and basic data preprocessing steps when use ANN. Different inputs could have different ranges. The transfer function in the neurons might become saturated if it is presented with a very large input. To avoid this situation, the input and target data has to be normalized. The normalization process will map all the input and target data into [-1 1] using the maximum and minimum of the particular data set. In other words, each of the data set will be normalized independently of the other sets.

2.4 Training

The ANN is trained with a training set of the form as follows:

$$S = \{(X_1, Y_1), \ (X_2, Y_2), \ (X_3, Y_3), \ ... \ , (X_N, Y_N)\}, \quad where \ X_k = [x_1, x_2, ..., x_m]^T, Y_k = [p] \text{ or } [p^c, p^h, p^l] \quad (3)$$

X_k is the input data vector, which consists of m elements. Y_k is the desired output vector when X_k is applied as the input, and it has one element for Single-price Model and three elements for Multiple-price Model. N is the total number of data pairs in the training set. During the training, each data pair is presented to the network in turn by setting $X^{(0)}(k) = X_k$, where $X^{(0)}(k) = [x_1^{(0)}(k),...,x_m^{(0)}(k)]^T$ is the input to the MLP network. So the number of input layer neuron n_0 equals to m. The energy function for network at iteration k is defined as:

$$E = \sum_{k=1}^{k=N}[E(k)] = \sum_{k=1}^{k=N}[\frac{1}{2}\sum_{j=1}^{n_L}(y_j - x_j^{(L)}(k))^2],$$
$$\text{where L = 2 in the case of 3 layer MLP} \tag{4}$$

The aim of the training process is to minimize the energy function E at the output layer by adjusting the weights systematically. Here we will adopt the batch training mode, which will only update the weights after all the N data pairs have been presented to the network, and N iterations is called 1 epoch. The energy function depends on only the weights $w_{ji}^{(1)}$ and $w_{ji}^{(2)}$. The standard Back propagation algorithm by Rumelhart et al. uses the steepest gradient descent technique to minimize the energy function E at each epoch. The weight adjust rule is defined as:

$$\Delta w_{ji}^{l} = -\eta(\frac{\partial E}{\partial w_{ji}^{(l)}}) \tag{5}$$

where $\eta \in (0,1]$ is the learning rate. The learning rate is critical, because it determines how fast the function converges. The training process repeats, epoch by epoch. The training process normally will stop if one of the following criteria is met: 1) the maximum number of training epoch has been reached; 2) the error E has dropped to some minimum value; or 3) the performance for the validation set gets worse for a certain number of consecutive updates.

2.5 Performance Measure

With this network, predictions are made with the input data in the test set. The network's performance is measured by Root Mean Square Error (RMSE) and Directional Accuracy (DA) methods. Both RMSE and DA will be used to measure the prediction performance for Close, Low and High prices separately. Taking Close price as an example (Low and High prices' RMSE and DA are calculated similarly), the formulas of RMSE and DA for Close price are as shown in equation (6) and (7). In the equations, N is the number of iterations, p_{t+1}^c is

the actual price at day t+1. $\hat{p}_{t+1}^c$ is the predicted price at day t+1, i.e. the output from the network.

$$RMSE^c = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_{t+1}^c - \hat{p}_{t+1}^c)^2} \tag{6}$$

$$DA^c = \frac{1}{N}\sum_{i=1}^{N}a_i, \text{where } a_i = \begin{cases} 1 & if (p_{t+1}^c - p_t^c) \times (\hat{p}_{t+1}^c - p_t^c) > 0 \\ 0 & otherwise \end{cases} \tag{7}$$

RMSE is calculated based on the difference of the actual price and the predicted price. It shall be used only when the aim of the network performance is to minimize the squared error. In the financial context, it is essential to predict the trend of the currency rate too. Hence in this study, both the RMSE and directional accuracy will be used to evaluate the network performance. Directional Accuracy is the percentage of correct predictions in terms of direction.

3 Simulation Study with Multiple-price Model

The fundamental objective of the simulation is to check the capability of the Multiple-price model compared to the Single-price model. On top of this fundamental objective, the model will be varied from different aspects, namely number of hidden layer neurons, currency pairs, preprocessing methods, time interval of forecasting and most importantly different input variable combinations. By comparing their performances, the best model will finally be chosen.

3.1 Overview of Simulation Study

Bearing in mind the above objectives, all simulations will be carried out for both Single-price and Multiple-price models, and the number of hidden layer neurons n_1 is varied from 4 to 20 with step of 4. The reason for simulating each model with different numbers of hidden layer neuron is to avoid the bias due to the fact that ANN with a certain number of hidden layer neuron normally performs differently when the input changes.

The number of input neurons n_0 , depends on different combination of inputs to be applied to the MLP. In each different model, the basic input variables are the multiple time-delayed close, low and high prices. In particular, they include 5 time-delayed close price data at day t $\{p_t^c, p_{t-1}^c, p_{t-2}^c, p_{t-3}^c, p_{t-4}^c\}$, and similar format for low and high price inputs. These input data will be used as part of the input variables to predict close, low and high prices at t+1. We will investigate the usefulness of the multiple-price model compared to the conventional single-price model where the number of output neurons n_2 is set as 1 for the single price model, and 3 for the multiple-price model.

Other than the variations of SPM versus MPM and number of hidden layer neurons, in each of the simulations, the model will be modified in certain aspects.

First, different currency pairs will be simulated. This is to select the most predictable currency pair to be studied in this project. Three currency pairs are studied in this first simulation, USD/SGD, AUD/USD and GBP/USD. In the first simulation, the most basic forecasting model with purely time-delayed inputs using first differencing and normalization preprocessing methods will be adopted.

Second, the simulation model will vary in terms of the preprocessing method after choosing the best currency pair. Thus, simulation model with first differencing ($p_i - p_{i-1}$), natural Log transformation ($\log(p_i)$), or ratio transformation (p_i / p_{i-1}), or natural Log on ratio transformation ($\log(p_i / p_{i-1})$), or price-1 ($p_i - 1$) transformation or no transformation will be carried out. Next, the simulation model is varied in terms of the input combinations. Instead of the purely time-delayed inputs, different combinations of technical indicator and fundamental indicator variables will be fed into the network. Lastly, due to the fact that the daily data might contain relatively high level of noise compared to the weekly data, the time interval of forecasting will be varied from daily to weekly in the simulation to investigate network performance.

3.2 Simulation Results and Discussions

The simulation results for different simulation tests are presented in Table 1 to 5.

Table 1. GBP/USD Averaged Results with Different Currency Pairs for Daily Time-delayed Inputs with First Differencing Preprocessing

Currency Pair	Testing RMSE Close	Testing RMSE Low	Testing RMSE High	Average	Testing DA Close	Testing DA Low	Testing DA High	Average
Multiple-Price Model								
USD/SGD	0.0125	0.0119	0.0273	0.0172	0.5158	0.6188	0.6194	0.5847
AUD/USD	0.0083	0.0061	0.0051	0.0065	0.5018	0.6314	0.6590	0.5974
GBP/USD	0.0100	0.0072	0.0065	0.0079	0.5223	0.6819	0.6802	0.6281
Single-Price Model								
USD/SGD	0.0053	0.0046	0.0366	0.0155	0.5109	0.5316	0.5297	0.5241
AUD/USD	0.0082	0.0079	0.0064	0.0075	0.5271	0.4952	0.5591	0.5271
GBP/USD	0.0101	0.0096	0.0088	0.0095	0.5214	0.5545	0.5410	0.5390

Table 2. GBP/USD Averaged Results for Daily Time-delayed Inputs with Different Preprocessing Methods

Preprocessing Method	Testing RMSE Close	Testing RMSE Low	Testing RMSE High	Average	Testing DA Close	Testing DA Low	Testing DA High	Average
Multiple-Price Model								
First Differencing	0.0100	0.0072	0.0065	0.0079	0.5223	0.6819	0.6802	0.6281
Natural Log	0.0101	0.0071	0.0063	0.0078	0.5199	0.6850	0.6928	0.6325
Ratio	0.0101	0.0073	0.0065	0.0080	0.5151	0.6735	0.6615	0.6167
Logged Ratio	0.0102	0.0074	0.0066	0.0080	0.5253	0.6632	0.6729	0.6205
Price-1	0.0101	0.0071	0.0064	0.0079	0.5271	0.6837	0.6928	0.6345
No Transformation	0.0100	0.0071	0.0063	0.0078	0.5199	0.6813	0.6873	0.6295
Single-Price Model								
First Differencing	0.0101	0.0096	0.0088	0.0095	0.5214	0.5545	0.5410	0.5390
Natural Log	0.0102	0.0098	0.0088	0.0096	0.5227	0.540	0.5541	0.5407
Ratio	0.0101	0.0097	0.0088	0.0095	0.5263	0.5311	0.5450	0.5341
Logged Ratio	0.0101	0.0097	0.0089	0.0095	0.4592	0.4121	0.4060	0.4258
Price-1	0.0102	0.0088	0.0098	0.0096	0.5396	0.5456	0.5402	0.5418
No Transformation	0.0102	0.0098	0.0088	0.0096	0.5354	0.5396	0.5474	0.5408

Table 3. GBP/USD Averaged Results for Different Technical Indicators and Daily Time-delayed Input with Natural Log Preprocessing

Technical Indicators	Testing RMSE Close	Testing RMSE Low	Testing RMSE High	Average	Testing DA Close	Testing DA Low	Testing DA High	Average
Multiple-Price Model								
None*	0.0101	0.0071	0.0063	0.0078	0.5199	0.6850	0.6928	0.6325
Moving Averages	0.0101	0.0071	0.0064	0.0079	0.5187	0.6711	0.6934	0.6277
Natural Log MAs	0.0102	0.0071	0.0064	0.0079	0.5246	0.6869	0.7018	0.6378
MA+MACD+RSI+SO	0.0102	0.0072	0.0066	0.0080	0.5185	0.6916	0.6880	0.6327
Single-Price Model								
None	0.0102	0.0098	0.0088	0.0096	0.5227	0.5400	0.5541	0.5407
Moving Averages	0.0102	0.0098	0.0089	0.0096	0.5166	0.5359	0.5565	0.5363
Natural Log MAs	0.0102	0.0097	0.0089	0.0096	0.5239	0.5305	0.5504	0.5349
MA+MACD+RSI+SO	0.0104	0.0087	0.0079	0.0090	0.5118	0.6194	0.6199	0.5837

Notes: * "None" represents the network model with purely time-delayed inputs

Table 4. GBP/USD Averaged Results for Different Fundamental Indicators and Daily Time-delayed Input with Natural Log Preprocessing

Fundamental Indicators	Close	Low	High	Average	Close	Low	High	Average
	Testing RMSE				Testing DA			
Multiple-Price Model								
None*	0.0101	0.0071	0.0063	0.0078	0.5199	0.6850	0.6928	0.6325
Stock Exchange Index	0.0100	0.0070	0.0063	0.0078	0.5211	0.6873	0.7012	0.6365
Gold Price	0.0105	0.0093	0.0071	0.0090	0.5163	0.6723	0.6934	0.6273
Interest Rate	0.0101	0.0070	0.0064	0.0078	0.5211	0.6958	0.6904	0.6358
All Three Indicators**	0.0107	0.0093	0.0079	0.0093	0.5217	0.6862	0.6970	0.6349
All without Gold Price	0.0102	0.0072	0.0065	0.0080	0.5257	0.6873	0.6987	0.6372
Single-Price Model								
None	0.0102	0.0098	0.0088	0.0096	0.5227	0.5400	0.5541	0.5407
Stock Exchange Index	0.0102	0.0097	0.0089	0.0096	0.5281	0.5486	0.5547	0.5438
Gold Price	0.0123	0.0131	0.0108	0.0120	0.5366	0.5390	0.5492	0.5416
Interest Rate	0.0104	0.0098	0.0091	0.0098	0.5390	0.5517	0.5511	0.5472
All Three Indicators	0.0125	0.0139	0.0135	0.0133	0.5299	0.5619	0.5680	0.5533

Notes: * "None" represents the network model with purely time-delayed inputs
** "All Three Indicators" represents the network model with stock exchange index, gold price, interest rate and time-delayed inputs.

Table 5. GBP/USD Averaged Results for Daily Time-delayed Input with Natural Log Preprocessing Using Different Time Interval of Forecasting

Time Inerval	Close	Low	High	Average	Close	Low	High	Average
	Testing RMSE				Testing DA			
Multiple-Price Model								
None*	0.0101	0.0071	0.0063	0.0078	0.5199	0.6850	0.6928	0.6325
Weekly Forecasting**	0.0139	0.0115	0.0088	0.0114	0.6243	0.7091	0.6667	0.6667
Daily with Same Amount	0.0093	0.0052	0.0063	0.0070	0.5394	0.7333	0.6611	0.6446
Single-Price Model								
None	0.0102	0.0098	0.0088	0.0096	0.5227	0.5400	0.5541	0.5407
Weekly Forecasting	0.0222	0.0206	0.0181	0.0203	0.6500	0.6038	0.6188	0.6242
Daily with Same Amount	0.0097	0.0077	0.0069	0.0081	0.5243	0.5375	0.5625	0.5414

Notes: * "None" represents the network model originally carried out with data from Jan 1998 to Nov 2010 with daily forecasting.
** "Weekly Forecasting" represents the network model with data from Jan 1998 to Nov 2010 with weekly Forecasting.
*** "Daily with Same Amount of Data" represents the network model with data from Apr 2008 to Nov 2010 with daily forecasting, which has same number of data sets as model **.

Referring to Table 1, three currency pairs' performances are examined. The USD/SGD has the worst RMSE and DA performances in both MPM and SPM. One possible reason is that SGD is not an actively traded currency in the forex market, the higher uncertainties make it more unpredictable, hence the simulation results are not as good as AUD/USD and GBP/USD, which are two of the most actively traded currency pairs. Between AUD/USD and GBP/USD, the later has higher low and high price directional accuracy. Since trend forecasting is crucial in forex trading and to better assess the predictive ability of the neural network model, GBP/USD is chosen for the rest of the simulations. Next, the effects of the MPM and SPM are investigated. By looking into each of the simulation test results presented in the above tables, we find that the forecasting performance for close price is similar between MPM and SPM in terms of RMSE and DA. However, forecasting performance is largely improved by using MPM. By considering the relatively better performance models in the above results, the average RMSE values for low and high prices of MPM are around 0.0072 and 0.0064, which perform 25.8% and 27.3% better than the relevant average RMSE values of SPM, which are around 0.0097 and 0.0088. The average DA values for low and high prices of MPM are above 0.68, which are 30.8% higher than SPM with values around 0.52. Therefore, it is proven that the proposed Multiple-price model performs better than Single-price model in terms of both the absolute value (RMSE) and direction (DA) prediction.

By looking at the different preprocessing methods, the RMSE values of all five simulations are very similar. Thus we can infer that the choice of preprocessing method has minimal effect on the prediction capability in terms of the absolute value. However, preprocessing method has significant effect on the prediction capability in terms of the trend, which is reflected by DA, the average DA of Close, Low and High prices in the MPM using Price-1 transformation is 0.6345, which is the highest among all the preprocessing methods. Moreover, Price-1, Natural Log and No Transformation methods have better performance than the other three methods. Our motivation to use the three methods, First Differencing, Ratio and Logged Ratio preprocessing is to highlight the relationship between two consecutive days' prices, by either feeding the price difference or ratio into the network instead of using price directly. This might, however, provide less price information than using no transformation or natural Log transformation. Furthermore, 5 time-delayed transformed price data are used as the NN input for all different models. Thus, in the price-1, natural Log or no transformation models, the relationship between consecutive days' prices can be easily learned by the network using these time-delayed data. Therefore, the price-1, natural Log and no transformation models which provide information for both single day price and consecutive days' price relationship, are expected to have better performance than the other three preprocessing transformation techniques.

The effects of using technical indicators, which include moving averages (MA10, MA20, MA30, MA60, MA120), moving average convergence divergence (MACD), relative strength index with period of 14 (RSI_{14}) and stochastic oscillator with period 14 (SO_{14}), are examined as well. By looking into the MPM results section in Table 3, we can find out the effect on MPM prediction performance after adding technical indicator inputs compared to the purely time-delayed model. In terms of RMSE, the performances are very similar for all four input combinations. Thus, different type of input variable combinations has little effect on the absolute value prediction accuracy. In terms of DA, the model with natural Log moving averages input has average DA value of 0.6378 which is 0.84% higher than the purely time-delayed inputs model with average DA value, 0.6325. The model with all technical indicators also performs slightly better than the purely time-delayed input model. Thus we can conclude that after adding the technical indicator moving averages, the network's direction prediction performance improves.

Moreover, the effects on MPM prediction performance after adding fundamental indicator inputs compared to the purely time-delayed model are examined. In terms of RMSE, the performances are very similar for input combinations without Gold price. Thus, adding fundamental indicator – Stock Exchange Index and Interest Rate, has little effect on the absolute value prediction accuracy. However, the model with gold price as well as all three indicators inputs performs exceptionally worse, which indicates gold price is not a good input for GBP/USD exchange rate forecasting. In terms of DA, the model with stock exchange rate has average DA value of 0.6365 which is 0.63% higher than the purely time-delayed inputs model with average DA value, 0.6325. Without using gold price, the models with fundamental indicators perform slightly better than the purely time-delayed input model. Thus, we can conclude that after adding fundamental indicators such as stock exchange index and interest rate, the network's direction prediction performance improves.

Finally, by testing over different time intervals and time lengths, it is found that the weekly forecasting model performs 62.9% and 150.6% worse than the daily model with same amount data which has RMSE value of 0.0070 and 0.0081 for MPM and SPM correspondingly. Therefore, the weekly forecasting model performs much worse than the daily forecasting model in terms of absolute value, i.e. RMSE. However, the weekly forecasting model improves the network performance on close price prediction's directional accuracy significantly, which might be because there is less noise in the weekly data than the daily data. When comparing daily model with different total time length, for both RMSE and DA, the performance of the model with data over shorter time length is improved, which indicates the dated data may contain less information of the current market exchange rate compared to more recent data.

3.3 Study of Multiple-price Model

In order to investigate further in the high directional accuracy in low and high prices prediction in the MPM, this simulation is carried out for a multiple-price model without close price time series. It means the historical low and high prices information are fed into the network, and the output will be the predicted high and low prices in the next time period. The simulation result is shown in Table 6.

Table 6. GBP/USD High and Low Prices' Averaged Results of Multiple Price Model with and without Close Price Time Series Information

	Low	High		Low	High	
Multiple Price Model	Testing RMSE		Average	Testing DA		Average
With Three Prices' Information	0.0071	0.0063	0.0067	0.685	0.6928	0.6889
With Only Low and High Prices' Information	0.0090	0.0082	0.0086	0.5446	0.5687	0.5566

After excluding close price time series in the multiple-price model, the prediction results of the low and high prices in terms of both RMSE and DA values are much worse than the multiple-price model with all three prices. Moreover, without close price time series, the result of the multiple-price model becomes similar with the single-price model. Therefore, it implies the important role of close price time series information played in the forecasting of low and high prices, which means that the high and low prices are highly related with the close price information.

4 Testing for Model Performance

Stemming from the fact that the multiple-price model shows a high degree of directional accuracy in the prediction of the movement of the high and low GBP/USD exchange rates on daily or weekly basis, a trading strategy utilizing the high and low prices is desired. Moreover, after examining the different variations on the neural network performance in the previous sections, the multiple-price model with 5 natural logged daily time-delayed prices and natural logged moving averages inputs with 20 hidden layer neurons has relatively higher training directional accuracy than other models. Therefore, this model with the highest training performance is fed with the testing data and whose output is used for the trading.

In this simulation, a normal ask/bid spread of 3 pips will be used. It means whenever there is a "buy and sell" or "sell and buy" transaction carried out, the return will be deducted by the transaction cost 3 pips, which is 0.0003.

On day t, BUY signal is triggered if (day t+1)'s predicted high price is greater than today's close price for more than 60 pips. The choice of 60 pips is due to the absolute value accuracy of the multiple-price model is around 0.0060 which is 60 pips, therefore, a signal larger than 60 pips might give us more meaningful BUY signal.

On day t+1, SELL signal is triggered if (day t+1)'s actual high price is higher than (day t+1)'s predicted high price. Otherwise, a STOP signal will be triggered if (day t+1)'s actual low price is more than 50 pips lower than the lowest of 3 previous days' low prices, and then the position will be closed out by selling all GBP at the STOP price. Mathematically, STOP price is calculated as:

$$\text{StopPrice} = \min(p_t^l, p_{t-1}^l, p_{t-2}^l) - 50\,pips \quad (8)$$

The stop signal check rule is to ensure no big loss is incurred when the market price falls dramatically within a day, which may often be caused by big disasters or big events such as financial crisis. Whenever a SELL decision is triggered, the return will also be deducted by the bid/ask spread or transaction cost which is 3 pips. When both the SELL and STOP signals are not triggered on day t+1, the position will be hold and carried over to the next day.

The decisions will be continuously carried out on every day throughout the whole testing period, which consists of 331 days in this study's simulation. Similar trading decisions are carried out for trading strategy using low price. Their trading results together with using the normal buy & hold trading strategy is shown in Table 7.

Table 7 Trading Performance of Three Trading Strategy Measured by Trading Performance Indicators

Trading Performance Indicators	Trading with High Price	Trading with Low Price	Buy & Hold Strategy
Total trades	173	141	1
% Positive trades	71.10%	74.47%	-
Best trade*	0.0345	0.0641	0.0292
Worst trade*	(0.0830)	(0.1143)	(0.0316)
Max drawdown*	(0.2037)	(0.2793)	(0.2093)
Mean return per trade*	0.00120	0.00097	(0.00005)
Total Return	0.2062	0.1371	(0.0162)
Annualized Return	0.2297	0.1522	(0.0177)
Standard Deviation	0.0187	0.0300	0.0101
Annualized Standard Deviation	0.0203	0.0325	0.0109
Sharpe ratio	11.23	4.63	(1.77)

Notes: * For Buy & Hold strategy, as only one trade is carried out (i.e. buy on the first day and sell on the last day), the indicators in the table are calculated for profit and loss of the balance on each day throughout the trading period.

Buy & Hold strategy has negative total return, which is obviously the worst strategy in terms of trading return. Trading with high price has slightly lower positive trade percentage, however; it has higher mean return as well as total return than trading with low price. More specifically, its annualized return is 50.9% higher than trading with low price strategy. Therefore, in terms of trading total return, trading with high price has the best performance.

Risk is another important factor for trading decisions. When we look at the range of return, which is simply best trade minus worst trade. Trading with low price has the highest range of return, while buy & hold strategy has the smallest. Moreover, as the standard deviation of returns is commonly used as a measure of investment risk, by comparing three strategies' standard deviation, we can find that trading with low price has the highest risk, which buy & hold has the lowest, which matches with the observations using return range.

Sharpe ratio is a return-risk measure, which is intuitive for a risk-adjusted performance measure. More precisely, it measures the reward, in terms of mean excess return, per unit of risk. For positive sharp ratio, the larger the value, the smaller the risk associated, which indicates, the trading strategy with high price has better risk-adjusted performance. For negative sharp ratio, increasing risk results in numerically larger value; however, one limitation about Sharpe ratio is that it is difficult to compare between positive and negative value. Here, after we examine returns, risk as well as risk-adjusted performance, it's reasonable to conclude that the trading strategy with high price has the best performance.

5 Conclusion

The major objective of the study is to examine whether the proposed multiple-price model has better prediction capability than the traditional single price model. And all simulations are carried out for both models to examine the effects. The simulation results have clearly shown that the multiple-price model has better performance than the single-price model in terms of both absolute error and directional accuracy based on the two performance measure indicators, RMSE and DA. Different variations on the proposed Multiple-price Model have been investigated. By utilizing the high forecasting performance in low and high prices, trading strategies have also been carried out, and are proven to have better returns than buy & hold strategy.

ACKNOWLEDGMENT

The author thanks Professor Wang Qing-Guo from the Electrical and Computer Engineering department of National University of Singapore for his support and valuable advice. Without him, this project would not have been possible. The author also thanks Qin Qin from the Electrical and Computer Engineering department of National University of Singapore for his contributions to this project.

REFERENCES

[1]. J.Yao, C.L. Tan, A case study on using neural networks to perform technical forecasting of forex, Neurocomputing 34 (2000) 79-98

[2]. E.E. Peters, Chaos and Order in the Capital markets: A New View of Cycles, Prices, and Market Volatility, Wiley, New York, 1991.

[3]. R.R. Trippi and E. Turban, eds. Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real- World Performance, (Probus, Chicago, 1993).

[4]. H. White, Learning in neural networks: A statistical perspective, Neural Computat. 4 (1989) 42.5-464.

[5]. E.E. Peters, Chaos and Order in the Capital markets: A New View of Cycles, Prices, and Market Volatility, Wiley, New York, 1991.

[6]. R. Hecht-Nielsen, *Neurocomputing,* (Addison Wesley, Menlo Park, CA, 1989).

[7]. M. Adya, F. Collopy, How e!ective are neural networks at forecasting and prediction? A review and evaluation, J. Forecasting, 17 (1998) 481-495.

[8]. J. T. Yao and C. L. Tan, A case study on using neural networks to perform technical forecasting of forex, *Neurocomputing* 34 (2000) 79–98.

Society for Science and Education United Kingdom
TMLAI Transactions on Machine Learning and Artificial Intelligence
Volume 2, Issue 5
ISSN: 2054 - 7390

The automation of teaching processes based on knowledge processing

Stefan Svetsky, Oliver Moravcik
Faculty of Materials Sciences and Technology, Slovak University of Technology in Bratislava, Slovak Republic;
stefan.svetsky@stuba.sk; oliver.moravcik@stuba.sk

ABSTRACT

Technology-enhanced learning as one of the EU research priorities is focused on "how information and communication technologies can be used to support learning and teaching". However, such “definition” is too much general, so, mostly technology-driven approaches are prevailing in the practice, which do not take enough in consideration didactic aspects of knowledge processing, and that teaching processes are related to mental processes of individuals. In addition, there are many open questions, especially “what is knowledge”, “what is knowledge representation”. An interdisciplinary definition of knowledge is missing, despite the fact that teaching processes are knowledge based. Within a long-term participatory action research on TEL when teaching bachelors, a strategy of automation of teaching processes was evolved. This seem to be a crucial point for solving any personalized computer support of teacher and students. Because these processes are primarily uncertain, or unstructured, it was found, that to make these processes better computerizable, a unification both teaching and informatics processes is needed. In this context, the knowledge processing is based on an idea of “virtual knowledge unit” (as a part of patent application, 2014). For this purpose, an in-house software has been developed that enables individuals to perform a “batch knowledge processing paradigm” in order to process a large amount of knowledge in natural language on their personal computers, university’s cloud and servers. This paper deals with a specific approach to automation of teaching processes based on the knowledge processing.

Keywords: Technology enhanced learning, Knowledge, Knowledge processing, Automation of teaching processes, Database technology.

1 Introduction

Research on Technology-enhanced Learning (TEL) belonged to the priorities under the European Union's Seventh Framework Programme (FP7). According to European research, TEL investigates "how information and communication technologies can be used to support learning, teaching, and the development of competence throughout one’s life" [1]. For example, the latest calls was focused on Educational Data Mining and Learning Analytics issues.

A participatory action research on TEL (PAR) was at the beginning mostly empirical. It was based on a simply idea to support by computer whatever is needed in classrooms when teaching bachelors students. The PAR focus was especially on personalized support of teacher as a key player within

DOI: 10.14738/tmlai.25.568
Publication Date: 17th October, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.568

teaching processes, and step by step was transformed on systematic research on TEL. This research can be characterized by the developing phases: (i) e-Learning), (ii) technology - driven approach, (iii) educational (didactic) - driven approach, and (iv) the well balanced combined technology - educational driven approach (more details can be found in [2-5] . During this period the faculty participated as well on research EU FP7 project proposals KEPLER (2007, focus was on system of keywords and ontology linking WEB-libraries with teaching processes), L3Pulse (2013, focus was on how to link knowledge flow within teaching processes to learning analytics). Actually, within PanEULangNet proposal for the first Horizon 2020 call ICT 17 (the cracking language barriers) the focus was on “embedding a human factor into automated machine translation system”.

To better understand this paper, it should be mentioned that when comparing the mentioned empirical research approach to personalized computer support with existing approaches, these key problems were found:

- such research approach is considered on American and European universities as “participatory action research”, i.e. when a teacher explores technology for his own targets and teaching activities [6], however , moreover, in our case the researcher design his own technology (he writes programming codes for database applications)
- reviewers, evaluators or opponents consider the developed “batch information and knowledge paradigm” (as a TEL’s method), the in-house software *BIKE /WPad (as TEL’s tool), and created TEL-system for a complex issue of several areas of Computer Science (e.g., knowledge management, soft computing, natural language processing, human computer interaction, text mining,…) - so, the BIKE seems to be an all-in-one software (something like an “empty knowledge based system”)
- BIKE - Batch Information and Knowledge Editor/Environment; WPad-Writing Pad
- The research approach is different from state-of-the-art in the view of the abstraction of knowledge, knowledge representation, and knowledge processing in connection to support mental processes of teacher (individuals) within teaching processes. This resulted in formulation of a virtual knowledge unit which could be understandable both by humans and computers without a need to design machine readable schemas and use knowledge representation languages (e.g. RDFs, OWL). This idea is a part of application submitted on Slovak Patent Office (4/2014).

In this context, the actual PAR on TEL is understood as the “automation of knowledge based processes “, i.e. it considers teaching for teaching and learning processes, and the knowledge should be processed as a parameter of the teaching process. In addition, this approach, must challenge not only to issues as how to solve didactics aspects but how to structure domain content, teaching processes and sub-processes as well, in order to make these processes better computerizable (despite the fact that these are not enough described, structured, or standardized). One should be always aware that any education system is based on transition and transfer of domain knowledge into the brain of students, thus, within teaching processes, and directed by a teacher. Moreover, in comparison with automation of dead technical systems, there is no guaranty whether student (individuals) will have understood a content of curriculum properly (although the automation “works”). Due to this complexity of automation, and support of cognitive processes respectively, the PAR research challenges to many terminological problems both from educational (see any didactic text book, e.g. [7,8]) or informatics point of view (it does not exist tailored Computer Science discipline).

2 Terminological Challenges of Research on Tel

The global or institutional challenges regarding a role of digital technology in supporting teaching processes are comprehensively described in [9]. The more detailed information on TEL as a research field can be found in in [1] (EU ICT research policy), in [6,10] specific books, research reports of networks of excellences [11,12] , as well in the published papers of authors of this paper [2-5], in which another appropriated literature is recommended (e.g. [13-16]. In this literature or in the contemporary scientific literature, it is often criticized that technology - driven approach dominates too much in state-of-the-art, which does not take enough in consideration didactic aspects of knowledge processing, and that teaching processes are related to mental processes of individuals (teachers, students, researchers), or knowledge workers in general. This results in fact that TEL research outcomes into educational practice are often considered for questionable.

For example, in [17] the TPACK framework (Technology, Pedagogy and Content Knowledge) is recommended as “a useful tool for the consideration of the interaction of technology with content and with pedagogy". This model should be helpful in overcoming “perceived barriers to the effective integration of ICT in teaching and learning environments".

In [18] a more education specific technological approach is required, because (i) “HE TEL/IT community has been just ineffective at delivering real improvements in education - some of the key reasons for this failure are embedded in the terminology itself", (ii) "people are transfixed by generic, off-the-shelf technology, developed by other people for other purposes -and the importance (even the possibility) of education-specific technology is ignored", and (iii) "... European Commission, which has funded a series of large academic research projects into TEL, explicitly recognized that there was a general “absence of evidence” that the projects had achieved lasting impact". The conclusion is: “TEL is a poorly conceived acronym, a new approach is needed that focuses on ‘education-specific technologies’, and we should leave ‘TEL’ behind and talk instead about ‘education technology’.

The complexity of approaches to computer support of teaching can be also explained on another example from the University of Oxford. The course eLearning was recently renamed to Learning and Technology with this justification: "More recently, terms such as Technology Enhanced Learning and Educational Technologies have been popular, but the first is somewhat loaded in meaning, and the second too technology-focused" [19].

If one is a teacher, researcher - beginner, even expert, he may be confused about this existing terminological chaos in scientific literature. Moreover, this concerns surprisingly as well the term “knowledge”, despite the fact that teaching, or related cognitive and mental processes of humans are knowledge based. Thus, although the “knowledge” is an important interdisciplinary issue, and frequently used term, it has different meaning in different research fields. This has an negative impact on the practice because experts of fields as are education, knowledge management, psychology, philosophy, artificial intelligence, Semantic WEB, including teachers and students do not understand each other.

One could continue by discussing another terminological issues related to the data, information, knowledge, big data, linked data, and so on. However, this paper is not focused on the analyzing of complexity of terminology. This, was mentioned mainly due to a better understanding that any automation of teaching processes, which is based on the knowledge processing, requires researchers to model an interdisciplinary acceptable abstraction of knowledge, or knowledge representation for the purpose of personalized computer support. If any practical abstraction was at the disposal - thus, the virtual knowledge unit in our case, it would enable to solve issues “how to support”, respectively “how to solve programmatically” the infinity of knowledge flow within the infinity of teaching processes and sequences of sub-processes that are daily performed by teachers (including any knowledge workers).

In this context, there is a certain analogy between the presented personalized automation based on knowledge processing (the focus is on building tools for helping teachers in their work rather than replacing them) and a vision of Nielsson’s “habile systems” within the field of Artificial Intelligence (he argues: "achieving real human-level artificial intelligence would necessarily imply that most of the tasks that humans perform for pay could be automated") [20].

3 The Methodology of Automation of Teaching Process

Teachers, students, or humans in general, all need to work with knowledge in their natural language. This requires humans and computers to be able read the same texts. In this view, the PAR on TEL resulted in the interdisciplinary acceptable formulation of knowledge and knowledge representation (the mentioned universal virtual knowledge unit), the developing in-house all-in-one software BIKE / WPad, and in implementation of a novel paradigm of batch knowledge processing, which enables teacher, or individuals to work out a large amount of information and knowledge, including the building up of supporting virtual space and the overall TEL system (the principles are described in [3]). This research can be characterized by the following issues and categories.

3.1 TEL tools

“all-in-one” database applications BIKE/WritingPad, BIKE 2/WPad, virtual learning environment /virtual learning space on university’s servers, communication channels (internet php/mysql-application), virtual application of technical calculations (for courses of study “Chemistry”, and “Background of environmental protection” -this is a php-solution generated by the BIKE), including various tutorials and tests.

Note: all-in-one means, that it works both as desktop, and internet application, and covers tens of areas of computer science.

3.2 Databases platforms and languages

FoxPro for Windows 2.6a, Visual FoxPro 9, MySQL, DB2 (IBM); foxpro, html, C++, php, SQL.

3.3 Research outcomes

Applications into teaching bachelors: a bulk-design of teaching and slf-study materials; writing semester works; embedding teaching curriculum into communication channels for several courses of study; modelling of calculations, teaching tutorials and tests for PAR purposes; multi-language support ; testing of audio technologies (TTS, Speech Recognition); personalized support of individual teachers; modelling of multi-medial support (resulted in a anti-plagiarism methodology); modelling of feedback and communication; modelling of methodology of teaching programming languages (a cooperation between WPad and source forge C++ application (http://www.bloodshed.net/dev/devcpp.html).

3.4 Computer Science and IT research areas (projects topic)

these are mentioned without a strictly classification: database technologies, programming languages, knowledge processing, knowledge repositories, knowledge management system, combined data-/ text-/ WEB-mining, virtual learning environments,, learning analytics, soft computing; technology enhanced learning, blended learning, active learning, cooperative learning, e-learning. For example the BIKE 2 consist of around two thousands programming codes and hundreds of items of user menu (in other words the programming code was written in the way that they cover useful functions or elements of the above mentioned research areas. Therefore, it works as the personalized all-in-one software).

3.5 Methodology pillars

- The modelling **unification** of content, teaching processes, and didactic approaches (communication for a feedback teacher-students) to be computerizable.
- The modelling automation of informatics processes, which are performed by teachers and students on desktop computers, faculty's cloud and servers.
- The design of set of applications, which integrate these modelled humans and computer activities, and enhance teaching processes in classrooms, including self-study.
- The strategic approach that the computer support of teaching processes is basically the automation of these processes based on knowledge (flow) processing, which is controlled by the teacher, and performed in the natural language.
- The teacher (to be sustainable) needs the complex supporting system that enlarged his social memory, and skills related to his mental (cognitive) processes, thus, the computer works as teacher's partner, "mind-ware", or "external virtual chip".

This approach illustrated schematically the Figure 1.

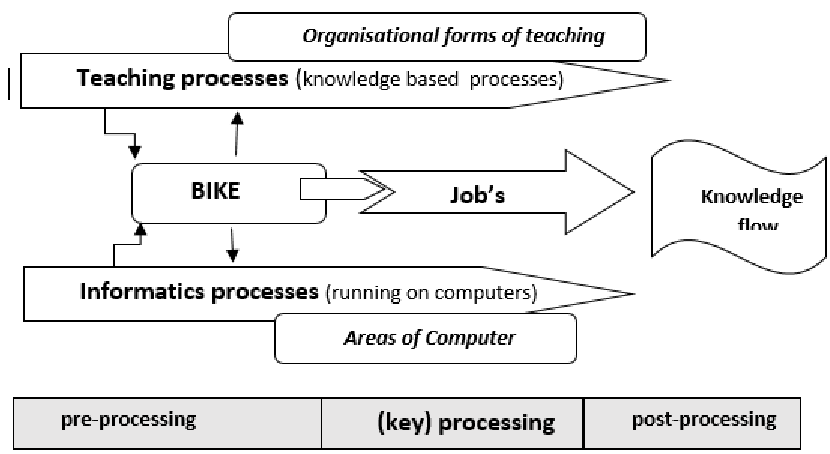


Figure 1: Schema of automation of teaching processes (knowledge flow processing)

4 The examples of Applications in Teaching Bachelors

To model the content and processes, the writing semester works resulted to be ideal. For this purpose, the BIKE (WPad) cooperates with the browser OPERA (version 9.27), which has two browser-specific features “Notes” and “Sessions”. These enabled students (individuals) (i) to open many windows in one step including scrolling between them, (ii) to mark with mouse a text on the computer screen, and save it as the Notes (*.adr file) , (iii) to save many open windows with the surfed internet pages into the Session (*.win file). The functions Notes and Sessions work as retrieving tools “for dummies”, i.e. enabled students - also with very low informatics skills, to create and catalog own libraries consisting of the visited WEB-pages, and teacher to collect the *.adr and *.win files from computers in the classroom into one cooperative study material, including possibility sending these files via e-mail.

This was used for training of multilingual retrievals when writing semester works (a simulation of working with literature), because the WPad (installed on computers in the classrooms) has in the user menu item for batch internet retrieving. This means, that students clicked on the item and simply wrote some keywords within WPad (or via using F5-key). After some seconds OPERA opened a sat of windows with search results, which were made up by Google, Bing, Yahoo and IxQuick search engines. Activities are commented on communication channels (php/mysql application), which can be used for teacher’s instructions, information exchange or also for uploading/downloading files to/from the faculty’s server. These activities illustrate Figures 2-4.

Figure 2: Screenshots of the batch internet retrieving (left: keywords, right: search engine results into OPERA)

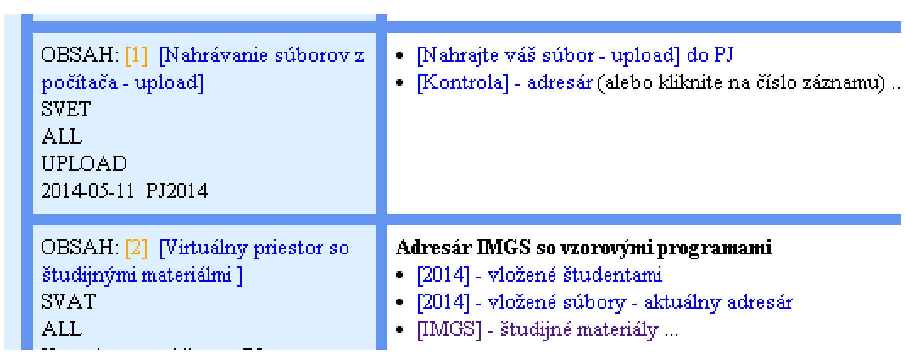


Figure 3: Screenshot of the communication channel with teacher’s instructions and information exchange (e.g. after clicking on the [2014]), students see the study materials, which they constructed via the shared internet retrieving)

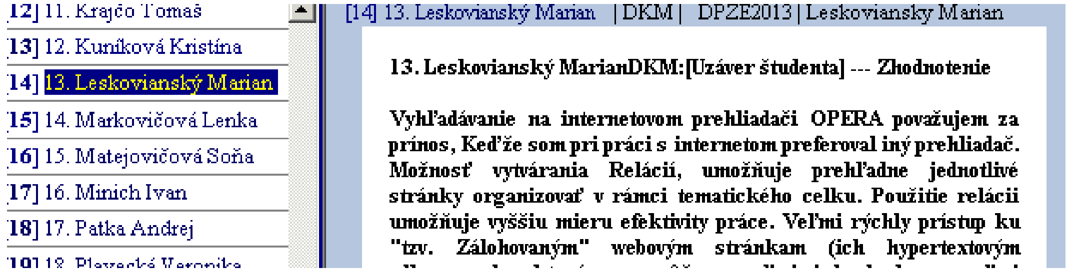


Figure 4: Screenshot of html-file with collected opinions of students related to effectiveness of OPERA

Figure 5 illustrates a schema of modelling learning analytics issues in the classroom with computers. There were indicated and analyzed: (i) log-files from internet domain, (ii) communication channels (faculty's server), (iii) off-line desktop computers (records in WPad, and evidence files). This process enable teacher to model an interlinking of learning analytics directly with the curriculum and tacit knowledge of students because WPad-tables consist of a set of linked virtual knowledge units.

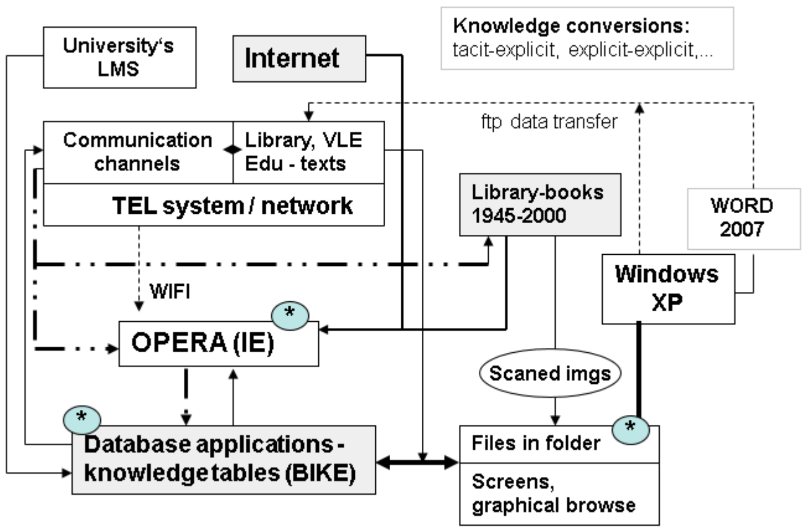


Figure 5: Schema of modelling learning analytics in the classroom - interlinking process knowledge with data

5 The examples of teacher's activities support

The computer support of teacher's activities consist of two levels

- The automation of common activities within the teaching process - this means, that the teacher selects appropriated items from the user menu of BIKE/WPad, performs batch knowledge processing paradigm, and utilizes the TEL system with libraries, study materials, tutorials, informatics tools.
- For sophisticated applications – this requires (i) modelling teaching processes, (ii) modelling informatics processes, and (iii) writing programming codes for database applications.

The modelling teaching processes basically means, that any teaching process must be analyzed for sequences of processes steps (to be more certain, structured, and computerizable).

The modelling informatics processes means, that teacher's activities running on the computer must be also investigated in the view of repeatability of steps and a need to solve adaptability of BIKE/WPad with

Windows, browsers or existing software and hardware. This is a basis for writing sequences of programming codes, something like triggers when using relational databases.

The writing programming codes for database applications represents the core of automation of teaching processes. Because if the teacher defines steps of any teacher process, i.e. his structure and has at his disposal a portfolio of programming tools, then he can start with semi-automatic and automatic activities, writing a set programming codes and testing within teaching. Maybe one of the most important issue is to divide activities on default activities an optional activities. Just here is the added value of human factor, i.e. the synergic effect of personalized computer support is achieved by the cooperation between teacher and computer, and this is based on the fact that teacher knows preferably what he needs to solve within his teaching, how often, at what time, to what extent, how fast, how often, etc. If the computer support was based on dominancy of machine – driven approach (a standard approach for Artificial Intelligence), these priorities does not know to formulate any machine. In addition, in the state of the art specific machine readable schemas and tailored languages are prevailing for knowledge processing. As was mentioned above, the needed effect is achieved in our case by natural language and that the knowledge abstraction enable both humans and machines to read the knowledge. Thus, then the knowledge can be processed within the teaching process through a synergic interoperability between teacher and computer in automatic or semi-automatic way. However, the more detailed description is beyond this paper, and it is still a question of the next progress because the PAR on TEL is based on implementing applications in the practice firstly. The next step is a back-analyze, "way it works", or "what fields of Computer Sciences cover the solution", or "what type of organizational form of learning is solved up".

Figures 6-7 illustrate the automation of common teacher's activities.

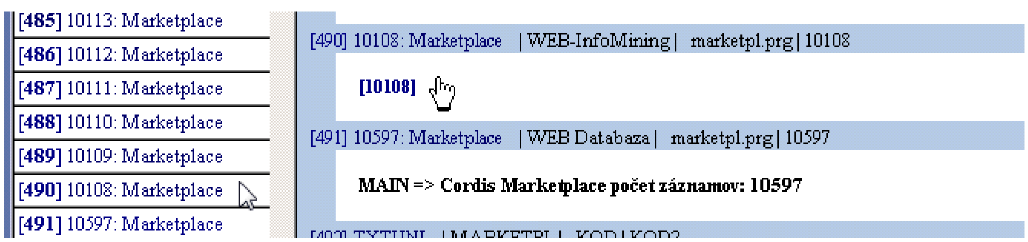


Figure 6: Screenshot of solving bilingual navigation via WEB-mining from the European database Marketplace

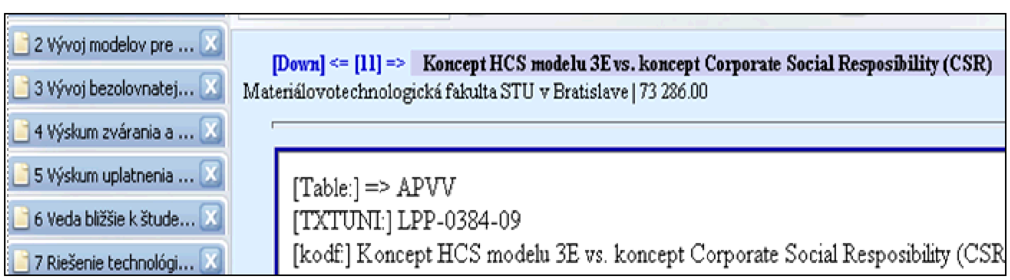


Figure 7: Screenshot of final output from national database of founded projects (the aim was to found data and how many faculty's projects were founded during previous years from thousands of accepted projects)

Figures 8-11 illustrate solving the more sophisticated applications - the modelling creation of personal bilingual corpora, as a set of screenshot captured within working out of the European Parliament Proceedings Parallel Corpus 1996-2011. The aim was to extract any text that is affected to bilateral relations Hungarian/Romania-Slovakia, and to test automated translation of one selected sentence into 24 European languages by using Google translator.

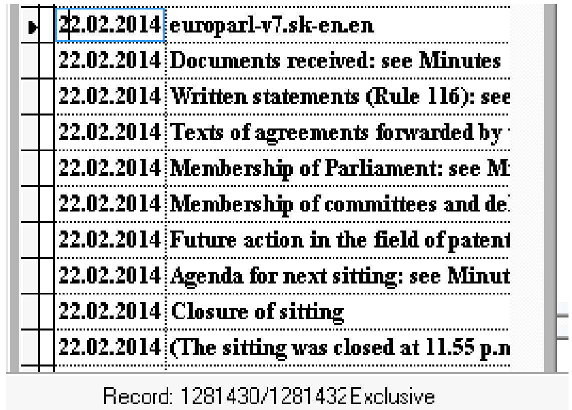


Figure 8: Screenshot of BIKE – work environment with 1 281 430 sentences of the Corpus (English sentences)

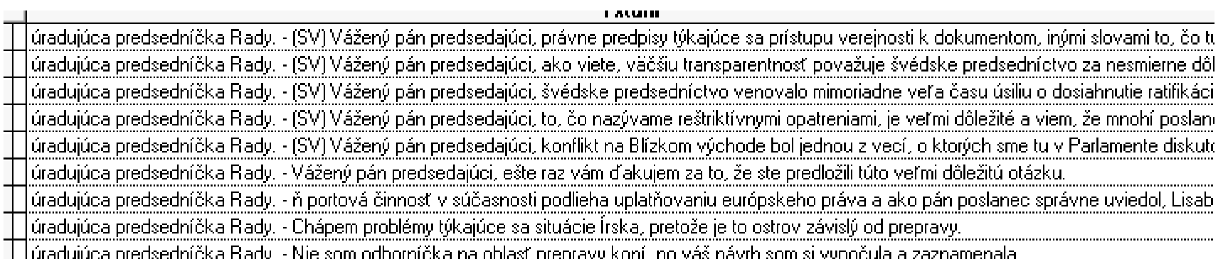


Figure 9: Screenshot of BIKE – work environment with 1 281 430 sentences of the Corpus (Slovak sentences)

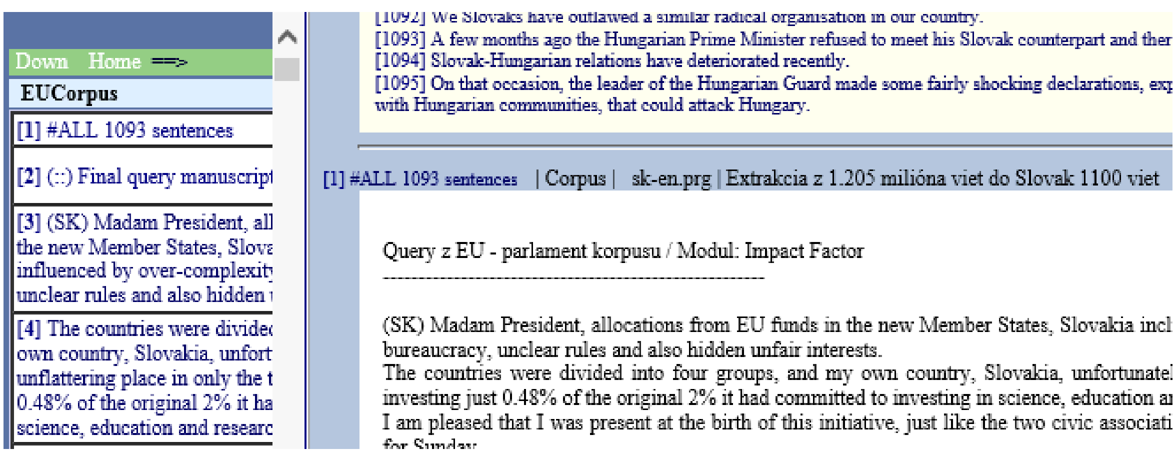


Figure 10: Screenshot of query obtaining by processing with BIKE – it was found 1093 appropriated sentences

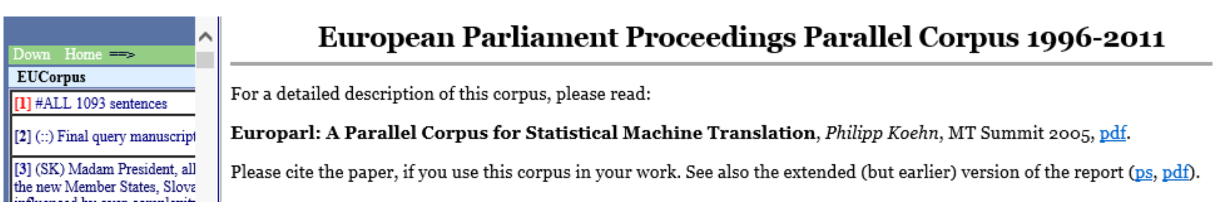


Figure 11: Screenshot of browsing within the query in html - format

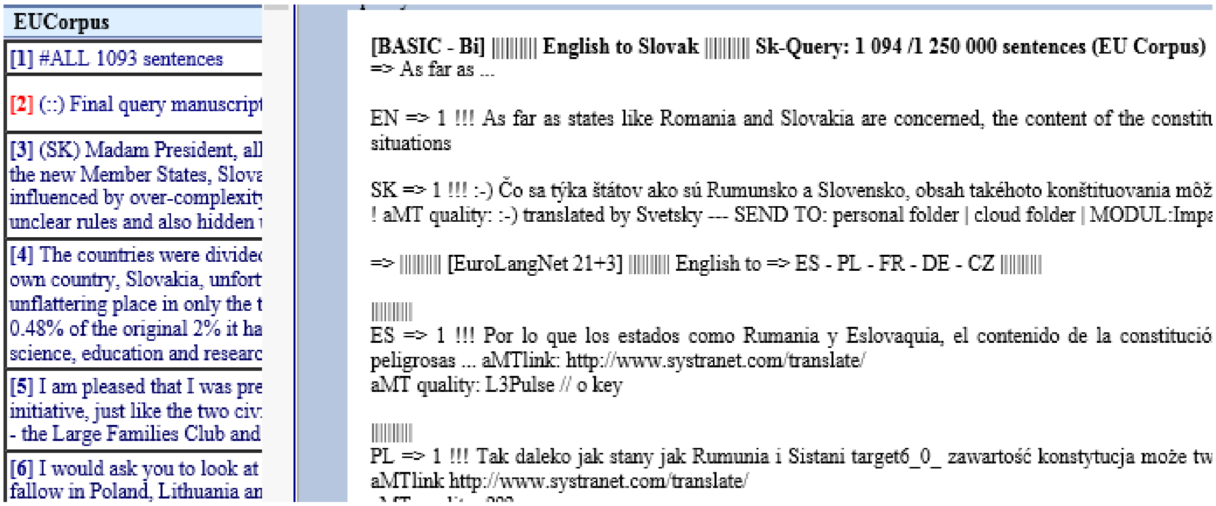


Figure 12: Screenshot of template for automated translation into EU 24 languages via online translation services

6 Conclusion

In this paper, the overall approach and methodology of automation of teaching processes based on knowledge processing was presented. The practice of the participatory action research on technology-enhanced learning showed that it is not crucial whether the computer support is focused on a certain organizational form of teaching, or a field of Computer Science, because this resulted mostly to the indirect mono-purpose support or technology-specific solutions within state-of-the-art. For educational practice, it is more important to accept teaching processes as knowledge based, and the computer support understand as an issue of automation of these processes where the knowledge is a basic parameter of the process. In this content, the paper has presented basic pillars of automation of teaching processes developed within the research on TEL, especially (i) an universal knowledge representation in the form of virtual knowledge unit (as the knowledge abstraction of real world), and (ii) in-house developed software BIKE (it runs only on the author's computer) / WPad (a standalone platform for knowledge processing and exchange on desktop computers or clouds), for performing (iii) the batch knowledge processing paradigm, which enables individuals to process "big" educational data. The paper further emphasized a need of modification and unification of teaching content and processes, in parallel with informatics processes running on computers in order to create applications, which are needed for supporting a set of teacher's job activities and processes. This was illustrated on the several examples of applications implemented in engineering teaching bachelor's students. Because the described approach resulted into a vision of "human centered computer intelligence", the actual research activities are focused as well on the modelling creativity support of individuals (solutions when writing English papers, text of songs, or multilingual issues in the field of Human Technology Languages- e.g. a PanEULangNet- project was proposed in the ICT 17 call of Horizon 2020 in this year).

ACKNOWLEDGEMENT

This work was supported within the funded project KEGA No. 047STU-4/2012: The building online teaching room for education in the field of machining of complex shaped components.

REFERENCES

[1]. TeLearn-European research on technology-enhanced learning. http://cordis.europa.eu/fp7/ict/telearn-digicult/telearn_en.html

[2]. Svetsky, S, et al., The Educational - Driven Approach for Technology Enhanced Learning. World Congress on Engineering and Computer Science on, 2012. Proceedings IAENG & IET: p. 290-296.

[3]. Svetsky, S. The practical aspect of knowledge construction and automation of teaching processes within technology-enhanced learning and eLearning. Habilitation thesis. Slovak University of Technology; 2012.

[4]. Svetsky, S. et al. The Particular Approach for Personalised Knowledge Processing. In: Advances in Intelligent and Soft Computing. Springer-Verlag Berlin Heidelberg, 2012, pp. 937-946.

[5]. S. Svetsky - O. Moravcik: The practice of Research on Technology - enhanced Learning. In: International Journal of Computer and Information Technology. Vol. 2, Iss. 4 (2013), s. 701-705.

[6]. Goodman Paul S., et al. Technology Enhanced Learning: Opportunities for Change. Laurence Erlbaum Associates, Mahwah, NJ, USA; 2002.

[7]. Skalková, J. The general Didactics (Obecná didaktika). Grada Publishing. a.s. 2007 (ISBN 978-80-247-1821-7)

[8]. Cube, F. v.: Die kybernetische - informationstheoretishe Didaktik. Westermanns Pädagogishe Beiträge, 32, 1980, č. 3, s. 120-124.

[9]. Saljö, R. Digital tools and challenges to institutional traditions of learning: technologies, social memory and the performative nature of learning. Journal of Computer Assisted Learning. Vol.26, pp. 53–64; 2010.

[10]. N. Balacheff, S. Ludvigsen, T. Jong, A. Lazonder, S. Barnes (Eds.). Technology - Enhanced Learning. Principles and Products. Springer, 2009, XXVI, 326 p.

[11]. R. Sutherland, S. Eagle, M. Joubert. A Vision and Strategy for Technology Enhanced Learning:Report from the STELLAR Network of Excellence. STELLAR, 2012. [cit. 2014-07-15] Available at http://www.teleurope.eu/pg/file/read/152343/a-vision-and-strategy-for-technology-enhanced-learning-report-from-the-stellar-network-of-excellence.

[12]. C. L. P. Nápoles, L., Montandon (Eds). D 1.1.2 Emerging technologies landscape: Report on Field Research results. Public report final 7-30/4/2013-M7 audience. HoTEL, 2013. [cit. 2014-07-15] Available at http://hotel-project.eu/content/d112-emerging-technologies-landscape-report.

[13]. A. Stolar. META-EDUCATION. Journal of Technology and Information Education. 2009, Olomouc - EU, Univerzita Palackého, Ročník 1, Číslo 1, s. 7 – 18.

[14]. P. Dillenbourg,. Integrating technologies into educational ecosystems. Distance Education, Routledge - Taylor & Francis Group (2008), Vol. 29, No. 2, p.127-134.

[15]. Derry, J. Technology - Enhanced Learning: A Question of Knowledge. Journal of Philosophy of Education, vol. 42, pp. 505–519; 2008.

[16]. D. Laurillard, "Technology Enhanced Learning as a Tool for Pedagogical Innovation," Journal of Philosophy of Education, vol. 42: pp. 521-533, 2008.

[17]. Kinchin, I. (2012). Avoiding technology-enhanced non-learning. British Journal of Educational Technology, 43(2), E43-E48.

[18]. Crispin Weston. The problem with “Technology Enhanced Learning”, http://edtechnow.net/2012/12/05/tel/, Ed Tech Now, 2012.

[19]. University of Oxford. SSD/MT10/MScEd, http://www.ox.ac.uk/media/global/wwwoxacuk/localsites/gazette/documents/examregs/oldsocsci/13-01-11_MSc_in_Education.pdf

[20]. Nielsson. J.N., Human - Level Artificial Intelligence? Be serious! AI Magazine 25 anniversary Issue. Winter 2005, p. 68 - 75.

SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

VOLUME 2, ISSUE 5
ISSN: 2054 - 7390

Classifying an Object using Class Differentiators

[1]Seunghyun Im, [2]Li-Shiang Tsay
[1]Computer Science Department, University of Pittsburgh at Johnstown, USA;
[2] Department of Computer Systems Technology, North Carolina A&T University, USA;
sim@pitt.edu; ltsay@ncat.edu

ABSTRACT

This paper discusses a supervised classification method. The method classifies an object using class differentiators. The class differentiators are the smallest set of values in a class that effectively distinguish one class from the others. The class membership is determined by the degree of homology between the test object and the class differentiators. Unlike many rule based classifiers, the algorithm proposed in this paper does not require input parameters and always produces the same results from the same data set. The algorithm is designed to work with categorical data, and is particularly useful when the quantification of the data is infeasible. We present an experimental result to show the validity of the algorithm.

Keywords: Classification algorithm, Supervised Classifier, Categorical Data, Reduct, Rough Set.

1 Introduction

We present a supervised classification method for categorical data. The classification problem in this paper focuses on the prediction of the class of a test object. We assume that the training data set has condition and decision attributes. The class labels of the training data set are determined by the discrete values in the decision attribute. The class label of the test object is uncertain. We use the training data to predict the unknown class label of the test object. The prediction can be performed by measuring the similarity (or distance) between objects. The data type has great influence on the similarity measure. For categorical data, the method can be divided into two types (1) discrete similarity measure and (2) approximated similarity measure. In general, the similarity between two attribute values is either 1 or 0 in discrete similarity measure; 1 for the same symbols and 0 for the different symbols between two objects. Therefore, the amount and frequency of the overlap between the training data and the newly observed object determine how much the object is close to a class. The classifiers based on the classical Naive Bayes [3] or Classification Rule [4] fall into this category. On the other hand, the methods derived from the approximated similarity measure convert the categorical data to numeric data in order to measure the degree of similarities between non-identical symbols. (e.g. the distance between red and blue is 0.7). Many algorithms have been proposed to address the conversion problem [5]. Although they are fundamentally useful, the quantification of categorical data is still challenging due to the fact that distances are often too arbitrary or it requires an intensive phase of data preprocessing, such as discretization of data [6] or construction of ontology [7][8]. The quantification is even more difficult when data are sanitized for privacy and regulatory requirements [9][10] because we do not know the meaning of the symbols. This is the case with many dataset from medical area. The algorithm presented

DOI: 10.14738/tmlai.25.594
Publication Date: 17th October, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.594

in this paper falls into the discrete similarity measure category. We developed this algorithm for the classification task that quantification is not possible.

The rule based classifier is one of the most widely studied supervised classifier for categorical data. A number of algorithms have been proposed [11]. These algorithms generate if-then style rules, and use the antecedent to measure the similarity and the consequent to predict the class label. Despite being very useful in many applications, these methods require quality threshold values, e.g. minimum support and confidence value for rule extraction. The determination of the optimal threshold value requires a deep understanding of the data and often need help from the domain expert. Another potential problem is that the result of the classification varies depending on the input parameters because the changes in these parameters result in different set of rules. The proposed algorithm does not require input parameters. This is an advantage for the classification task that requires consistent result or for non-technical users who are not familiar with the notion of the threshold value. Naive Bayes classifier is another well-known classification algorithm. It is simple and easy to use. We believe our method is as simple as the classical Naive Bayes classifier, yet it does not need any type of pseudo-count to solve the zero frequency problem.

The proposed algorithm is inspired by the concept of object reduct [2]. An object reduct is the minimal set of values of an object which differentiates the object from other objects. We extended the concept to detect patterns characterizing each class, and use them to find the most promising class label for a newly observed object. We present our method in Chapter 2, an experimental result in Chapter 3, and the conclusion in Chapter 4.

Table 1: Example of Information System.

	B	C	E	D
x_1	b_1	c_1	e_1	d_1
x_2	b_2	c_1	e_2	d_2
x_3	b_2	c_3		d_2
x_4	b_1	c_1	e_1	d_2

2 Algorithm

2.1 Basic Notations

We will use the following notations for dataset. By an information system [1] we mean a triple $S = (X,A,V)$, where

$X = \{x_1, x_2, ..., x_i\}$ *is a finite set of objects,*

$A = \{a_1, a_2, ..., a_j\}$ *is a finite set of attributes*

$V = \{v_1, v_2, ..., v_k\}$ *is a finite set of attribute values.*

The information system in Table 1 has 4 objects $X = \{x_1, x_2, x_3, x_4\}$ and 4 attributes $A = \{B,C,E,D\}$. The attributes are partitioned into two types: condition and decision. The condition attributes are B, C, and E, and the decision attribute is D. They are written as,

$A_C = \{B, C, E\}$ *condition attribute*

$A_D = \{D\}$ *decision attribute*

The following notations are used for attribute values.

$B(x_1) = b_1$ b_1 *is the value of B in* x_1

$V_D = \{d_1, d_2\}$ *attribute values for attribute D*

We assume that the classification is guided by the decision values. Therefore, a class is a set of objects in *S* with the same decision value, which is written as,

$X_d = \{x_k \in X; D(x_k) = d\}$ *Class defined by decision value d*

In Table 1, $X_{d1} = \{x_1\}$ and $X_{d2} = \{x_2, x_3, x_4\}$. We use an information system that has two decision values (d_1 and d_2) for simplicity of presentation, but the algorithm directly carries over to the general case where $|V_D| \geq 2$.

2.2 Method Description

2.2.1 Differentiators

The *differentiator* used in this paper is a derivation of reduct [2] in rough set theory [1]. An o*bject-object differentiator,* denoted as $\Delta_{OO}(x',x'')$, is the disjunction of the attribute values in object *x'* that distinguishes *x'* from the other object *x''*. For example, in Table 1, we can easily distinguish x_1 from x_2 with b_1 of *B*. i.e., a person wearing a red shirt (b_1) from another person wearing a blue shirt (b_2). We do not need other attribute values to distinguish x_1 from x_2. $\Delta_{OO}(x_1, x_2)$ is b_1 *or* e_1, and it is written as ($b_1 \vee e_1$). Although we have two attribute values, one is enough to discern x_1. An object may contain null values. We use two methods for handling the null value: (1) if an attribute value of *x''* is null the corresponding attribute value in *x'* becomes an element of $\Delta_{OO}(x', x'')$ because the null value in *x''* is different from the value in *x'*. (2) if an attribute value in *x'* is null we do not add the null to $\Delta_{OO}(x', x'')$ because a null is not a specific value that can distinguish *x'* from other objects.

An *object-class differentiator,* denoted as $\Delta_{OC}(x', X_{d''})$, is *the shortest terms* that distinguish *x'* from $X_{d''}$ (e.g. x_1 and $X_{d2} = \{x_2, x_3, x_4\}$). The shortest terms are acquired by finding the prime implicants [12] of all object-object differentiators (Δ_{OO}) between *x' and* $\{x \in X_{d''}\}$. A prime implicant is a product term that cannot be subsumed by any other product term. To obtain the prime implicant set, we put all Δ_{OO}s into conjunctive normal form (CNF) and transform it to disjunctive normal forms (DNF). $\Delta_{OO} = \emptyset$ when two objects are identical. In this case, we do not include the empty set as a conjunct of the CNF. The object-class differentiator (Δ_{OC}) consisting of at least one empty Δ_{OO} is called approximated object-class differentiator. Otherwise, it becomes a precise object-class differentiator. When *x'* is compared with all other objects in $\{X - x'\}$ it is simply written as $\Delta_{OC}(x')$. Intuitively, $\Delta_{OC}(x')$ is the attribute values in *x'* that can most effectively (not necessarily precisely) differentiate *x'* from the other objects in *S*.

A *class-class differentiator*, denoted as $\Delta_{CC}(X_{d'}, X_{d''})$, is the disjunction of Δ_{OC}s. It is the set of unique terms in $X_{d'}$ in relation to $X_{d''}$. A $\Delta_{CC}(X_{d'}, X_{d''})$ is categorized as an approximated class-class differentiator if it has one or more approximated Δ_{OC}. Otherwise, it becomes a precise class-class differentiator. Some terms may appear more than once in a Δ_{CC}. The frequency of a term *t* is defined as the number of times *t* is a subset of *x* in X_d

$$f(t) = \sum_{i=1}^{n} \begin{cases} 1, & t \subseteq x_i : x_i \in X_d \\ 0, & otherwise \end{cases}$$

When $X_{d'}$ of $\Delta_{CC}(X_{d'}, X_{d''})$ is $X - X_{d''}$ (e.g. there are only 2 class labels), the class differentiator of $X_{d'}$ is simply written as $\Delta_{CC}(X_{d'})$. Let (t_i, f_i) be a term and its frequency. Then, $\Delta_{CC}(X_{d'})$ is,

$$\Delta_{CC}(X_{d'}) = \{(t_1, f_1), (t_2, f_2),..., (t_n, f_n)\}$$

We use $\Delta_{CC}(X_d)$ to classify a test object.

2.2.2 Classifying a test object

Suppose that we want to classify the object x_{new}. The class label of x_{new} is determined by its unknown decision value. We calculate the value by measuring how similar x_{new} is to the terms in the class differentiators. For example, the decision value of x_{new} is d_1 if x_{new} is closer to the terms of $\Delta_{CC}(X_{d1})$ than those of $\Delta_{CC}(X_{d2})$. The idea is that $\Delta_{CC}(X_{d1})$ is the distinctive pattern of values that characterizes the class. If the same pattern is found in x_{new}, x_{new} is most likely having the same characterestics of $\Delta_{CC}(X_{d1})$. Since the example problem has two class differentiators, we measure the similarity between the terms of x_{new} and the terms of $\Delta_{CC}(X_{d'})$ and $\Delta_{CC}(X_{d''})$ respectively. Then, we compare the degree of similarity to choose the right decision value. The degree of similarity, expressed as a weight, is calculated by the ratio between the sum of the frequencies of the terms in $\Delta_{CC}(X_d)$ and the number of the objects in X_d. Let $\omega^{x_{new}}(d)$ be the weight of the unknown decision value of x_{new}. Then,

$$\omega^{x_{new}}(d) = \frac{\sum f(t_i) : t_i \subseteq x_{new}, t_i \in \Delta_{CC}(X_d)}{|X_d|}$$

Next, we compare the weights of all decision values. The object x_{new} belongs to a class that has the highest $\omega^{x_{new}}$.

$$x_{new} \in X_d : \omega^{x_{new}}(d) = max(\omega^{x_{new}}(d_i)), i = 1 \dots |V_D|$$

When two or more decision values have the equal weight, we randomly select one. We show an example in the next section.

2.3 Sample Problem

We will use the information system in Table 2. $A_C = \{B,C,E\}$. $A_D = \{D\}$. The dashed line indicates the division between two classes. $X_{d1} = \{x_1, x_2, x_3, x_4\}$ and $X_{d2} = \{x_5, x_6, x_7\}$. x_{new} is the object to be classified.

Table 2: Information System S and an object x_{new}

	B	C	E	D
x_1	b_2	c_1	e_1	d_1
x_2	b_2	c_3	e_1	d_1
x_3	b_1	c_1		d_1
x_4	b_1	c_3	e_1	d_1
x_5	b_1	c_1	e_1	d_2
x_6	b_1	c_1	e_1	d_2
x_7	b_2	c_2	e_2	d_2
x_{new}	b_2	c_3	e_1	

We first build a discernibility matrix from Table 2 in order to obtain $\Delta_{CC}(X_{d1})$. Table 3 is the discernibility matrix that the elements in each cell are the *object-object differentiators.* For example, b_2 is the attribute value that discerns x_1 from x_5 in Table 2. Then, b_2 is placed in the cell between x_1 and x_5 in Table 3. Either c_1 or e_1 can distinguish x_1 from x_7 (*or* is denoted as $\vee$ sign), and they are placed between x_1 and x_7 in Table 3.

Table 3: Discernibility matrix for X_{d1}

	x_1	x_2	x_3	x_4
x_5	b_2	$b_2 \vee c_3$		c_3
x_6	b_2	$b_2 \vee c_3$		c_3
x_7	$c_1 \vee e_1$	$c_3 \vee e_1$	$b_1 \vee c_1$	$b_1 \vee c_3 \vee e_1$

The conjunction of all $\Delta_{OO}(x_1)$s is $(b_2) \wedge (b_2) \wedge (c_1 \vee e_1)$. We transform it to a DNF to find $\Delta_{OC}(x_1)$. That is,

$\Delta_{OC}(x_1)= (b_2) \vee (b_2) \vee (c_1 \wedge e_1) = (b_2 \wedge e_1) \vee (b_2 \wedge c_1)$

We also compute Δ_{OC} for x_2, x_3, x_4.

$\Delta_{OC}(x_2)= (b_2 \wedge c_3) \vee (b_2 \wedge c_3) \vee (c_3 \wedge e_1) = (c_3) \vee (b_2 \wedge e_1)$

$\Delta_{OC}(x_3)= (b_1 \wedge c_1)$

$\Delta_{OC}(x_4)= (c_3) \wedge (c_3) \wedge (b_1 \wedge c_3 \wedge e_1) = (c_3)$

Next, we calculate the frequency of the terms in Δ_{OO}s to build $\Delta_{CC}(d_1)$. For instance, the frequency of c_3 is 2 because it is in $\Delta_{OC}(x_2)$ and $\Delta_{OC}(x_4)$. We can find $b_2 \wedge e_1$ twice in $\Delta_{OC}(x_1)$ and $\Delta_{OC}(x_2)$. Its frequency is 2. Thus, $\Delta_{CC}(d_1)$ *is,*

$\Delta_{CC}(d_1) = \{(c_3,2), (b_2 \wedge e_1,2), (b_2 \wedge c_1,1), (b_1,1), (c_1,1)\}$

The class-class differentiator for d_2, $\Delta_{CC}(d_2)$, can be obtained from the same matrix because there are only 2 decision values in *S* and the discernibility matrix is symmetric. Using the same method we acquire,

$\Delta_{CC}(d_2) = \{(b_1 \wedge c_1 \wedge e_1,2), (c_2,1), (e_1,1)\}$

Table 4 shows the terms, frequencies, and class labels of $\Delta_{CC}(d_1)$ and $\Delta_{CC}(d_2)$. We calculate the weight of the class label of x_{new} using Table 4. A term in Table 4 is counted to compute the weight if the term is a subset of x_{new}. We can find that term #1 = $\{c_3\}$ and term #2 = $\{b_2 , e_1\}$ are the subsets of $\{b_2,c_3,e_1\}$. Their frequencies are 2 and 2 respectively. There are 4 objects in X_{d1}. Therefore, the weights of d_1 for x_{new} is,

$$\omega^{x_{new}}(d_1) = \frac{2+2}{4}$$

In the same way, term #8 is a subset of x_{new}, and we use its frequency value 1 and the number of terms in X_{d2} to calculate the weight of d_2.

$$\omega^{x_{new}}(d_2) = \frac{1}{3} = 0.33$$

We classify x_{new} to X_{d1} because $\omega^{x_{new}}(d_1) > \omega^{x_{new}}(d_2)$.

Table 4. Class differentiator (ΔCC) for S

term #	*term*	*frequency*	*class label*
1	c_3	2	d_1
2	$b_2 \cdot e_1$	2	
3	$b_2 \cdot c_1$	1	
4	b_1	1	
5	c_1	1	
6	$b_1 \cdot c_1 \cdot e_1$	2	d_2
7	c_2	1	
8	e_1	1	

3 Implementation and experiment

We implemented the algorithm in Python programming language and tested it using the data in Table 5. The data set has information about stolen cars. The training data has 10 objects $\{x_1,...,x_{10}\}$, three condition attributes {*color, type, origin*}, and a decision attribute {*stolen*}. It is divided into two classes by *'yes'* and *'no'*. All attributes are categorical.

Table 5. Stolen Car

object	***color***	***type***	***origin***	***stolen***
x_1	*red*	*sports*	*domestic*	*yes*
x_2	*red*	*sports*	*domestic*	*no*
x_3	*red*	*sports*	*domestic*	*yes*
x_4	*yellow*	*sports*	*domestic*	*no*
x_5	*yellow*	*sports*	*imported*	*yes*
x_6	*yellow*	*suv*	*imported*	*no*
x_7	*yellow*	*suv*	*imported*	*yes*
x_8	*yellow*	*suv*	*domestic*	*no*
x_9	*red*	*suv*	*imported*	*no*
x_{10}	*red*	*sports*	*imported*	*yes*

We conducted two experiments. In the first, we compared our algorithm to Naive Bayes classifier by running two algorithms with several different test objects. For example, both algorithms classified x_{new1} = {*red, suv, domestic*} to *'no'* as shown in Table 6. However, Naive Bayes classifier failed to classify another object x_{new2} = {*blue, suv, domestic*} to a class due to the zero probability problem created by the attribute value *blue*. We need to use some type of data modification, such as Laplace correction [3] to solve this problem. On the other hand, our algorithm successfully classified x_{new2} using the class differentiators listed below.

```
Class-Class Differentiator

Decision Value : Yes (5), Hit 0/5

(red ∧ sports)              2

(domestic ∧ red)            2

(imported ∧ sports)         2

(imported ∧ yellow)         1
```

```
Decision Value : No (5), Hit 3/5
(domestic)                    1
(domestic ∧ yellow)           2
(suv)                         1
(domestic ∧ suv)              1
(red ∧ suv)                   1
```

In Table 5, the attribute value set {*suv, domestic*} is found only in x_8, and it works as a class differentiator of X_{no}. This term is a subset of x_{new2} = {*blue, suv, domestic*}, and is used to predict the decision value of x_{new2} to '*no*'.

Table 6. Classification of x_{new1} and x_{new2}.

Algorithm	x_{new1}	x_{new2}
Naive Bayes	*yes : 0.037*	*yes : 0.0*
	no : 0.069	*no : 0.0*
Class Differentiator	*yes : 0.4*	*yes : 0.0*
	no : 0.8	***no : 0.6***

We also compared our algorithm to a rule based classifier. Although the details vary, most rule based classifiers generate a set of if-then rules to predict the unknown data [11]. As described earlier, we often need to run a rule extraction algorithm several times to obtain the rule set that can classify an object. In this experiment, we extracted two sets of rules using ERID [14] and ran Chase [13] algorithm to classify x_{new2} = {*blue, suv, domestic*}. As shown in Table 7, we could classify x_{new2} to X_{no} using ERID rule set #1 (that is generated with support = 3 and confidence = 0.75). However, no decision value was predicted with another rule set (rule set #2 with support = 2 and confidence = 0.8) because the terms in rule set #2 did not have a matching attribute value. The proposed algorithm does not have this problem because the class differentiators are not dependent on input parameters such as threshold values.

```
Rules generated by ERID
[Rule set #1] min support 3, confidence 0.75
suv→No                        3, 0.75
red,sports→Yes                3, 0.75
[Rule set #2] min support 2, confidence 0.8
yellow,domestic→No            2, 1.0
sports,imported→Yes           2, 1.0
```

Table 7. Classification of x_{new2} using Chase

	sup =3, conf = 0.75	*sup = 2, conf = 0.8*
Rule Based Classification	*yes : 0.0*	*yes : 0.0*
	no : 1.0	*no : 0.0*

4 Conclusion

This paper discussed a method for the supervised classification of categorical data. The algorithm uses class differentiators to find out the class label of a test object. The class differentiators are the set of attribute values in the training data set that distinguishes one class from the others. The proposed algorithm measures the similarity between the test object and the class differentiators to determine the class membership. The comparison of our algorithm with a rule based classifier and Naive Bayes classifier shows that the proposed algorithm produces more consistent results and less prone to the zero probability problem.

REFERENCES

[1]. Pawlak, Z., "Rough sets", International Journal of Computing and Information Sciences, 11(5), pp. 341-356, 1982

[2]. Pawlak, Z., Skowron, A., "Rough sets and Boolean reasoning" Information Sciences, 177(1), pp. 41–73, 2007

[3]. Duda, R., Hart, P. and Stork, D., "Pattern classification", Wiley, New York, 2nd edition, 2001

[4]. Tan, P., Steinbach, M. and Kumar, V., "Introduction to data mining", 1st edition, Pearson Addison Wesley, Boston, 2005

[5]. Boriah, S., Ch, ola, V. and Kumar, V., "Similarity measures for categorical data: A comparative evaluation", In: SIAM Data Mining Conference, pp.243-254, 2008

[6]. Dougherty, J., Kohavi, R., Sahami, M., "Supervised and unsupervised discretization of continuous features", In: Twelfth International Conference on Machine Learning. pp.194-202, 1995

[7]. Guarino, N., "Formal ontology, conceptual analysis and knowledge representation", International journal of human-computer studies, 43(5), pp.625-640, 1995

[8]. Xiaodan Zhang, Liping Jing, Xiaohua Hu, Michael Ng, and Xiaohua Zhou. "A comparative study of ontology based term similarity measures on PubMed document clustering", In: 12th international conference on Database systems for advanced applications, pp.115-126, 2007

[9]. Brickell, J. and Shmatikov, V., "The cost of privacy: destruction of data-mining utility in anonymized data publishing" In: ACM SIGKDD international conference, pp.70-78, 2008

[10]. Im, S., "Privacy aware data management and chase", Fundamenta Informaticae, 78(4), pp.507-524, 2007

[11]. Charu C. Aggarwal, and ChengXiang Zhai., "A Survey of Text Classification Algorithms", Mining Text Data, Springer, pp 43-76, 2012

[12]. Posthoff, Ch. and Steinbach, B., "Logic Functions and Equations - Binary Models for Computer Science", Springer, Dordrecht, 2004

[13]. Dardzinska, A. and Ras, Z., "Rule-based Chase algorithm for partially incomplete information systems", In: Second International Workshop on Active Mining, pp.42-51, 2003

[14]. Dardzinska, A. and Ras, Z.W., "Extracting Rules from Incomplete Decision Systems: System ERID" In: Foundations and Novel Approaches in Data Mining, Studies in Computational Intelligence, Vol. 9, Springer, pp.143-154, 2006

SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 2, ISSUE 5
ISSN: 2054 - 7390

A Novel Approach to Distributed Multi-Class SVM

Aruna Govada, Shree Ranjani, Aditi Viswanathan and S.K. Sahay
Department of CS&IS, BITS Pilani K. K. Birla Goa Campus, Zuarinagar, Goa, PIN-403726, India.
shreeranjaniss@gmail.com; aditivin@gmail.com; garuna@goa.bits-pilani.ac.in

ABSTRACT

With data sizes constantly expanding, and with classical machine learning algorithms that analyze such data requiring larger and larger amounts of computation time and storage space, the need to distribute computation and memory requirements among several computers has become apparent. Although substantial work has been done in developing distributed binary SVM algorithms and multi-class SVM algorithms individually, the field of multi-class distributed SVMs remains largely unexplored. This research proposes a novel algorithm that implements the Support Vector Machine over a multi-class dataset and is efficient in a distributed environment (here, Hadoop). The idea is to divide the dataset into half recursively and thus compute the optimal Support Vector Machine for this half during the training phase, much like a divide and conquer approach. While testing, this structure has been effectively exploited to significantly reduce the prediction time. Our algorithm has shown better computation time during the prediction phase than the traditional sequential SVM methods (One vs. One, One vs. Rest) and out-performs them as the size of the dataset grows. This approach also classifies the data with higher accuracy than the traditional multi-class algorithms.

Keywords: Distributed algorithm, Support Vector Machine, Machine learning, Map reduce, Multi class

1 Introduction and Related Work

In the machine learning world, SVMs offer one of the most accurate results. SVMs are accurate because of their high generalization property to classify unknown examples. Yet SVM algorithms have been largely restricted to simple 2-class (binary) classification problems. However, numerous practical applications involve multi-class classifications - like identifying the galaxy that a star belongs to, remote sensing applications, etc. Some of the most used multi-class SVM approaches include One vs One, One vs Rest, DAG and Error correcting codes (all of which have their own drawbacks and are not as efficient as binary SVM algorithms).

In One vs. Rest classification, the n-class problem is converted into n 2-class sub problems with one positive class and (n-1) negative classes. In One vs. Rest classification, the n-class problem is converted into n(n-1)/2 two-class problems. Krebel [1] showed that by this formulation, unclassifiable regions reduce, but still they remain. To solve the problem of unclassifiable regions, Taylor et al. [2] proposed decision-tree based pairwise classification Graph. Pontil et al. [3] proposed to use rules of a tennis tournament to solve unclassified regions. Kiksirikul et al. [4] proposed the same method and called it Adaptive Directed Acyclic Graph. A comparison of these approaches [5] suggest the usefulness of One vs

DOI: 10.14738/tmlai.25.562
Publication Date: 30th October, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.562

One in terms of accuracy and computation and this is why we have chosen to compare our approach with this.

For both binary SVMs as well as multi class SVMs, in the recent years, handling large datasets has become an arduous task. Data Scientists are overwhelmed with the amount of data and the need for excessive data pre-processing that this explosion has caused. Given that data handling has become tough, data mining – the process of discovering new patterns from large data datasets – is a herculean task. This has given rise to scientists developing distributed parallel algorithms to meet the scalability and performance requirements for big data. Computation time and computation complexity (which involves solving the quadratic optimization problem) has been a limiting factor for SVMs especially for large data sets. To overcome this, many parallel and distributed SVMs were proposed. Initially most of the parallel SVM was based on MPI programming model. Moving from the MPI programming model based parallel SVM, parallelization has been achieved through the MapReduce Framework now. Fox [6] developed parallel SVM based on iterative MapReduce model Twister. A parallelization scheme was proposed where the kernel matrix is approximated by a block-diagonal approach [7]. Further improvements to parallel SVM implementations like Cascade SVMs [8] have been proposed which heavily reduce the communication overhead among the computers. In this method, dataset is split into parts in feature space. Non-support vectors of each sub dataset are filtered and only support vectors are transmitted. Collobert et al. [9] proposed a new parallel SVM training and classification algorithm that each subset of a dataset is trained with SVM and then the classifiers are combined into a final single classifier function. Lu et al. [10] proposed a connected network based distributed support vector machine algorithm. In this method, the dataset is split into roughly equal part for each computer in a network then, support vectors are exchanged among these computers. Sun et al. proposed a novel method for parallelized SVM based on MapReduce technique. This method is based on the cascade SVM model. Their approach is based on iterative MapReduce model Twister which is different from our implementation which is a recursive MapReduce algorithm. Ferhat et al. [11] proposed a novel MapReduce based binary SVM training method in which the whole training dataset is distributed over data nodes of cloud computing system using Hadoop streaming and MRjob python library. Despite such extensive work on multi class SVMs as well as distributed binary SVMs, the arena of multi class distributed SVMs has remained largely unexplored. In this paper, we propose a novel algorithm for distributed multi class SVMs and have compared our results with the most popular multi class SVM approaches (One vs. One and One vs. Rest)

2 Proposed Framework

The proposed algorithm is based on binary tree kind of structure created during the training phase. Our algorithm aims to reduce the total number of SVMs required to classify a data point, thus enabling better efficiency during run-time of the model that was built out of our algorithm. While One vs. One, One vs. Rest and DAGSVM classify using $\frac{n(n-1)}{2}$, n, and $\frac{n(n-1)}{2}$ SVMs respectively, we use $log_2(n+1)$ SVMs to classify the data point at run-time. One possible structure that can be obtained in depicted in Figure II. It is critical to choose the most appropriate combination to obtain the most optimal case while testing for a new sample data point. For this, we have separated the training stage into 2 significant phases where the first stage (Training) is devoted to compute all possible support vectors and the second stage (Cross Validation) evaluates all of them and returns the best division.

2.1 Training

Given N classes, we partition the entire dataset into 2 halves each containing [n/2] classes using support vectors. This is done neglecting the differences among the classes on one side. Without loss of generality, one half has been assigned as positive class and the other negative class. To generate support vectors, Atbrox's [12] method for parallel machine learning has been used which gave us the mapper and reducer implementation for binary classification. Atbrox's method implements incremental SVM algorithm for binary classification as described below:

The SVM classifier solves the following problem of finding w,y i.e. the coefficients of the support vector formulated as

$$(w, y) = (I/\mu + E^T E)^{-1} E^T D e$$

Where I – identity matrix

μ - parameter >0

E = [A –e]

D – Diagonal matrix with plus ones or minus ones

To classify a test sample with feature vector x, following equation is used.

$$sgn\,(x^T w - y) = \begin{cases} 1, & x \in A+ \\ -1, & x \in A- \end{cases}$$

Where A+ and A- denote the positive and negative classes respectively.

Mappers and reducers have been used to parallelize the calculation of E^TE and E^TDe and Figure I. depicts a brief outline of the algorithm which explains the function of each mapper and reducer used in this approach.

As at any point, binary classification is performed where each class represents many, mappers and reducers from Atbrox have been modified to suit our purpose. A single run of the training stage is as follows:

- Divide the dataset into 2 regions using $({}^nC_{n/2})$ / 2 planes where 'n' stands for the number of classes in the dataset. This figure is arrived based on the intuition that a plane divides the data points into roughly half the number of classes on each side. i.e choose n/2 out of n and and to avoid repeated counting , the number of possible combinations was divided by 2.
- For all possible combinations support vectors are formed.

2.2 Cross Validation

This stage of training primarily involves identifying the best plane from the possible options obtained from previous stage.

A single run of the second stage is as follows.

- For each of the partitions thus obtained, accuracy with which each plane divides is calculated using the classification accuracy metric ((true positives + true negatives)/total samples).

- Mappers split the task of obtaining the confusion matrix (The matrix which contains true positives, true negatives, false positives, false negatives). Reducers assimilate the values in the confusion matrix from each node and compute the classification accuracy metric. This metric is used to identify the best split and store the 2 separated lists of classes for further computation.
- At the end of this second stage, we obtain a set of positive and negative classes along with their corresponding accuracy calculation.

Both the stages are repeated until the number of classes in the positive and negative become one, which effectively means that the dataset has been successfully divided into all N classes.

2.3 Testing

The classification of a test sample starts at the root of the tree. At each node of the binary tree a decision is being made about the assignment of the input pattern into one of the two possible groups obtained after the training phase. Each of these groups may contain multiple classes. This is repeated recursively downward the tree until the sample reaches a leaf node that represents the class it has been assigned to (Figure II). Any test sample will go through a maximum of $log2N$ SVMs during the test phase.

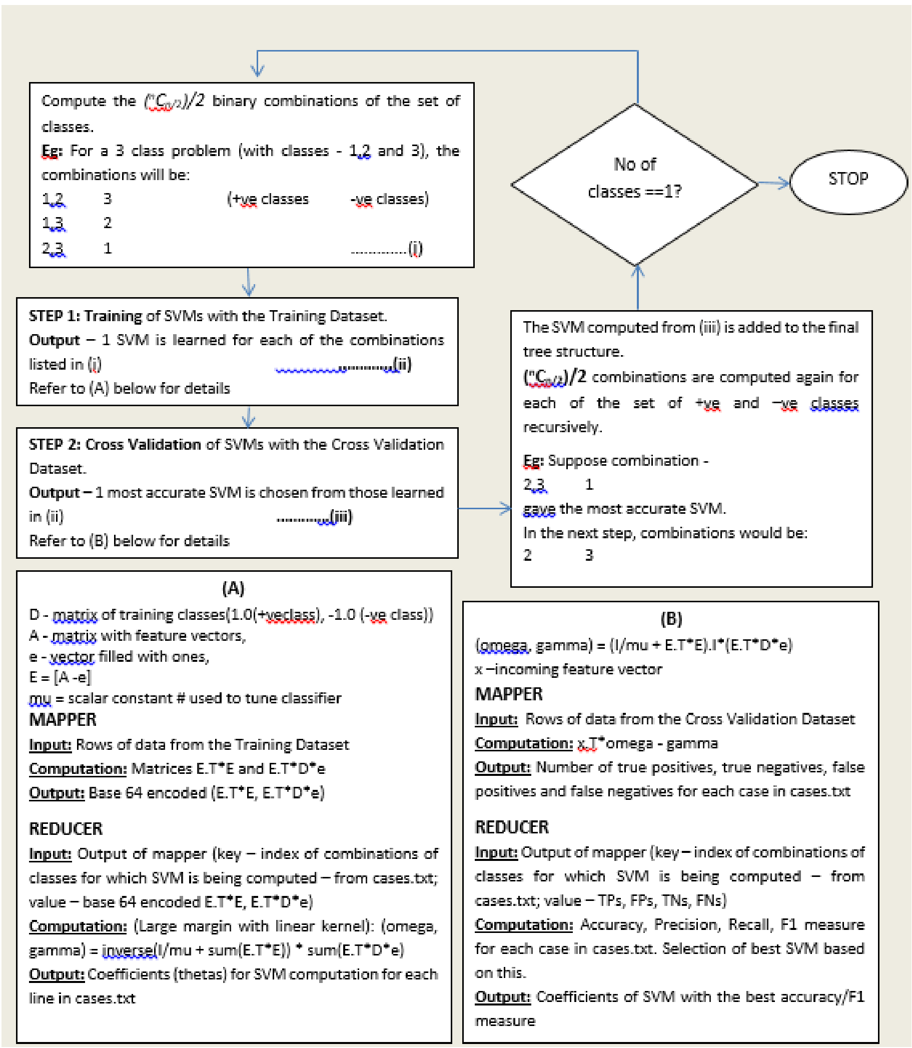


Figure I. A brief outline of the algorithm

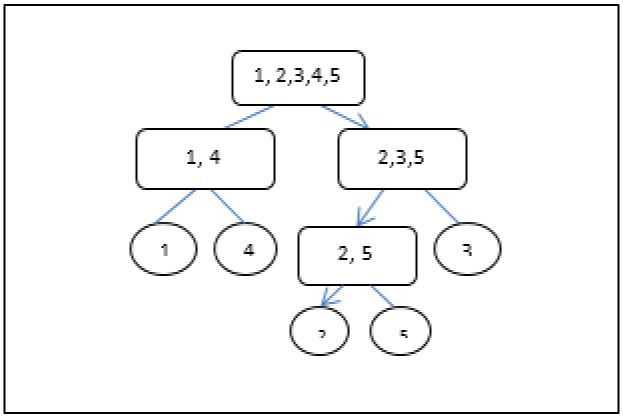


Figure II: This structure is created during the training phase. A test sample belonging to class 2 will follow the path depicted by the arrows

3 Experiments and Results

For all of these experiments, we have used a 3 node cluster to measure the metrics of our approach, and Python's Scikit-learn library for One vs. One and One vs. Rest). Datasets used for experimentation are described below and the sources for those are indicated in references.

3.1 Datasets used

Dataset name	**SDSS[15]**	**Iris[16]**	**Mfeat[17]**
# Training samples	40000	150	1500
# Testing samples	10000	50	500
# Features	6	3	6 (mor) 47 (zer) 64 (kar)
# Classes	3	4	10

Figure III: Datasets used

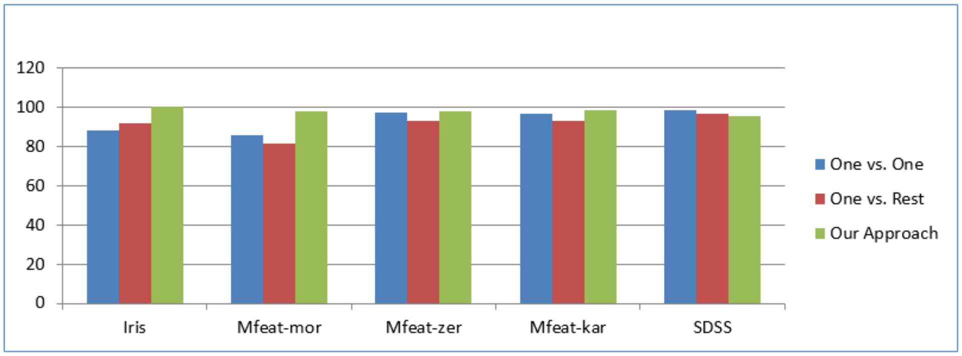
3.2 Accuracy (Figure IV)

We have measured the accuracy of the algorithm using the following formula on the testing samples

$$Accuracy = \frac{TPs + TNs}{Total\ Samples} \quad (1)$$

Our approach gives better accuracies in all the datasets except the SDSS dataset. SDSS is a skewed dataset, so accuracy is not the best performance metric to use during cross-validation. We will have to use metrics other than the accuracy (such as precision, recall and F1 measure) to select the best SVM here

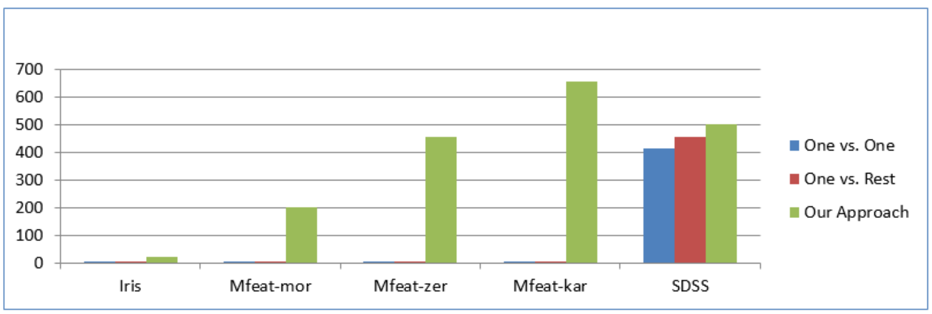
Figure IV: Comparison of Accuracy (%)



3.3 Training Time (Figure V.)

While the single-machine implementations are more efficient for the smaller datasets, in the SDSS dataset we see that our training time is comparable to the single-machine implementations due to the large data size of SDSS. We can thus show that distribution of the computation gets more beneficial as the data size increases.

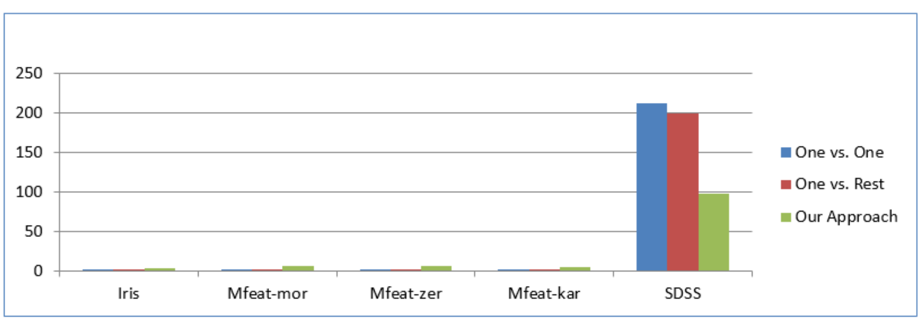
Figure V: Comparison of Training Time (seconds)



3.4 Testing Time (Figure VI.)

We show a significant reduction (53.7%) in testing time for the SDSS dataset, a result of the distributed approach working hand-in-hand with the decision tree based algorithm.

Figure IV: Comparison of Training Time (seconds)



4 Conclusion and Future work

In this research, we have proposed a novel distributed multi class SVM algorithm in which instead of extending binary SVMs or all-together methods, the idea is to divide the dataset into half at any point of time and obtain the visual distribution during the training phase. While testing, this structure has been effectively exploited and hence saving huge amount of testing time. This approach has been found to excel as data size increases which caters to our needs of handling big data.

In the future, we hope to enhance this algorithm by doing the following:

- Implementing a distributed Gaussian Kernel (we are currently using a linear kernel)
- Optimizing the algorithm for skewed datasets by using performance metrics such as the F1 measure (instead of Accuracy that we use currently)
- Running the algorithm with data sizes of about 20-30 GB with very large clusters (which we haven't been able to do so far for lack of resources)
- Comparing this algorithm with other multi-class Machine Learning techniques (non-SVM)

REFERENCES

[1]. U. H.G. Krebel Pairwise classification and support vector machines. In B.Scholkopf, C.J.C. Burges, and A.J.Smola, editors, Advances in Kernel methods: Support Vector Learning, pages 255-268.The MIT Press,Cambridge,MA,1999

[2]. J.C.Platt, N.Cristianni and J.Shawe-Taylor.Large margin DAGs for multiclass classification. In S.A.Solla, T.K.Leen, and K.R.Muller, editors Advances in Neural Information processing systems 12,pages 547-553.The MIT Press,2000.

[3]. M. Pontil and A.Verri. Support Vector machines for 3-d object recognition IEEE Transactions On Pattern Analysis and Machine Intelligence,20(6) : 637-646, 1998.

[4]. B.Kijsirikul and N. Ussikavul Multiclass support vector machines using adaptive directed acyclic graph. In Proceedings of International Joint Conference on Neural Networks (IJCNN 2002), pages 980-985, 2002.

[5]. Mahesh Pal, Multiclass Approached for Support Vector Machine Based Land Cover Classification. NIT Kurukshetra

[6]. Zhanquan Sun1, Geoffrey Fox2 .Study on Parallel SVM Based on MapReduce Key Laboratory for Computer Network of Shandong Province, Shandong Computer Science Center, School of Informatics and Computing, Pervasive Technology Institute, Indiana University Bloomington

[7]. J X Dong, A Krzyzak, C Y Suen. "A fast Parallel Optimization for Training Support Vector Machine." Proceedings of 3rd International Conference on Machine Learning and Data Mining, 2003: 96-105.

[8]. Hans Peter Graf, Eric Cosatto, Leon Bottou, Igor Durdanovic, Vladimir Vapnik Parallel Support Vector Machines: The Cascade SVM .NEC Laboratories.

[9]. R. Collobert, Y. Bengio, S. Bengio, A Parallel Mixture of SVMs for Very Large Scale Problems. in Neural Information Processing Systems, Vol. 17, MIT Press, 2004.

[10]. Lu Y, Roychowdhury V, Vandenberghe L. Distributed Parallel Support Vector Machines in Strongly Connected Networks. IEEE Transactions on Neural Networks 2008; 7: pp. 1167 - 1178

[11]. Ferhat Özgür Çatak1, Mehmet Erdal Balaban A MapReduce based distributed SVM algorithm for binary classification National Research Institute of Electronics and Cryptology, TUBITAK, Turkey

[12]. http://atbrox.com/2010/02/08/parallel-machine-learning-for-hadoopmapreduce-a-python-example/

[13]. C C Chang, C J Lin. LIBSVM: a library for support vector machines ACM Transactions on Intelligent Systems and Technology, 2011, 27(2): 1-27.

[14]. Edward Y. Chang¤, Kaihua Zhu, HaoWang, Hongjie Bai, Jian Li, Zhihuan Qiu, & Hang Cui PSVM: Parallelizing Support Vector Machines on Distributed Computers,Google Research, Beijing, China

[15]. http://skyserver.sdss.org/public/en/

[16]. https://archive.ics.uci.edu/ml/datasets/Iris

[17]. https://archive.ics.uci.edu/ml/datasets/Multiple+Features

Society for Science and Education United Kingdom
TMLAI Transactions on Machine Learning and Artificial Intelligence
Volume 2, Issue 5
ISSN: 2054 - 7390

Approaches to solve cell formation, machine layout and cell layout problem: A Review

J. N. Patel[1], S. V. Patel[2]
[1]*Chhotubhai Gopalbhai Patel Institute of Technology, Uka Tarsadia University, Surat, Gujarat, India*
[2] *Computer Science Department, Veer Narmad South Gujarat University, Surat, Gujarat, India*
ce6_jigar@yahoo.co.in; patelsv@gmail.com

ABSTRACT

Cell formation, machine layout and cell layout problems in cellular manufacturing system are NP-Complete optimization problems. Good cell formation & layout design in cellular manufacturing system is achieved by finding optimum or near-optimum solutions of these problems, which substantially reduces manufacturing cost and time. Many approaches have been advocated by researchers to obtain better cell formation & layout design. An attempt has been done in this paper to review such approaches based on heuristics, meta-heuristics, hybrid methods and exact solution methods developed by past researchers to solve these problems. The main objective of this review paper is to find out the effective and efficient approaches by comparing them based on performance criteria, their benefits and drawbacks in solving cellular manufacturing system problems and find out future research scope in this area.

Keywords: Genetic Algorithm, Cell Formation Problem, Machine Layout Problem, Cell Layout Problem.

1 Introduction

Frederick Taylor introduced group technology as a manufacturing philosophy that improves productivity by grouping parts with similar characteristics into part-families and creating production cells with a group of dissimilar machines. Cellular manufacturing uses the principles of group technology to achieve higher production efficiency compared to traditional manufacturing so as to create factory and shop floor layout design. Cellular manufacturing can be implemented in four stages as per the following.

1. Cell formation: Grouping parts into part families and corresponding machines into machine cells by using the parts production process.
2. Intra-cell layout: Layout of machines within each cell.
3. Inter-cell layout: Layout of cells within the factory or shop floor.
4. Scheduling: Scheduling of jobs in each cell.

Advantages of cellular manufacturing compared to traditional manufacturing by efficient layout design reported in the literature are as per the following.

1. Reduced production lead time
2. Reduced setup time
3. Reduced work-in-process
4. Reduced material handling cost

DOI: 10.14738/tmlai.25.600
Publication Date: 30th October, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.600

5. Reduced flow distance of material
6. Reduced lot sizes
7. Reduced throughput times
8. Reduced tooling cost
9. Reduced labour cost
10. Reduced production equipment cost
11. Improved machine utilization
12. Simplified process planning
13. Better worker morale
14. Improved quality

Cell formation, machine layout and cell layout problem in cellular manufacturing system are known to be NP-Complete optimization problems. Exact solution methods are inefficient for solving these large-sized NP-Complete problems. So researchers have developed heuristics, meta-heuristics and hybrid methods to solve these large-sized NP-Complete problems efficiently. Appropriate performance criteria to be selected by past researchers to compare performance of their approach or approaches with other existing approaches to solve cell formation problem are grouping efficiency [1, 2], grouping efficacy [3], comparison with optimum solution [4], machine utilization [1], number of inter-cellular moves [5], bond energy [6], percentage of exceptional elements [7], number of non-dominated solutions [8], quality metrics [8], hyper area ratio metric [8], relative metric [8], spacing metric [8], cell flow index [9], overall flow index [9], average cell flow index [9]. Appropriate performance criteria to be selected by past researchers to compare performance of their approach or approaches with other existing approaches to solve machine layout problem are percentage of variation with optimum solution [10, 11], total distance travelled by all parts [12], price of a layout [13], cell flow index [9], average cell flow index [9]. Appropriate performance criteria to be selected by past researchers to compare performance of their approach or approaches with other existing approaches to solve cell layout problem are total inter-cell material handling distances [14], percentage of deviation with best-known solution [14], overall flow index [9].

This paper discusses comparative analysis of approaches to solve these problems done by researchers available in past. Better approach to be found based on average performance in past papers is also suggested in this paper.

The rest of the paper is organized as follows: Section 2 describes cell formation, machine layout and cell layout problem in cellular manufacturing system. Section 3 presents review & comparison of approaches such as heuristics, meta-heuristics, hybrid methods and exact solution methods taken from literature to solve cell formation, machine layout and cell layout problem in cellular manufacturing system. Analytical review of approaches is presented in Section 4. Conclusion and future scope is mentioned in Section 5.

2 Cell formation, machine layout and cell layout problem

Cell formation problem: the process of grouping parts with similar design features or processing requirements into part families and the corresponding machines into machine cells so as to maximize grouping efficiency and grouping efficacy.

Machine layout problem: consider m machines and n parts are used during the manufacturing of a various kinds of products in Cellular Manufacturing. Various parts are processed by a machine and a single machine may often be utilized during the manufacturing of a various kinds of products. Objective is to decide an optimal layout scheme of all m machines in a way that minimizes total traveling distance of all n parts, so substantially reduces total manufacturing costs for manufacturing industries, subject to no overlap and no duplication of machine constraints.

Cell layout problem: layout of cells within the factory or shop floor in a way that minimizes the inter-cell movement of various parts.

3 Approaches to solve cell formation, machine layout and cell layout problem

3.1 Heuristics

Heuristics use domain-specific knowledge to solve large-sized NP-Complete problems efficiently. They may or may not produce optimum solution of the problem. But good heuristics may produce near optimum solution of the problem.

3.1.1 Effective non-iterative sorting method

An effective non-iterative sorting method was developed to solve facility layout problem in [15]. This method is capable of obtaining solutions that are not too far from the optimum and also provides a base for improvement methods. It needs to be refined further such that it becomes capable to solve facility layout problem having different shapes and sizes of facilities. It also needs to be extended to find out optimum solution of facility layout problem.

3.1.2 Interactive hierarchical design approach

An interactive hierarchical design approach was developed to solve cellular layout problems in [16]. Numerous example problems were taken from the literature to test this approach. The comparison demonstrated that this approach finds out generally more efficient layouts.

3.1.3 Heuristic algorithm

A heuristic algorithm was developed in [11] to solve machine layout problem in cellular manufacturing system. The solutions generated by heuristic algorithm were compared with the optimum solutions & from comparison it was found that average percentage of variation of results generated by heuristic algorithm with optimum is 4.78% in evaluated cases. So this heuristic algorithm is able to find out the very good solution, outperforms solution techniques described in [15] in all five cases taken from [15] and outperforms solution technique described in [16] in three cases out of five cases taken from [15].

3.1.4 Novel construction-cum-improvement heuristic

A novel construction-cum-improvement heuristic was developed in [12] to solve layout formation problem of type QAP. The novel construction-cum-improvement heuristic is capable of obtaining sub-optimal solution of the test problem and outperforms other solution techniques described in [15], [16] and [11] except in period 1. For period 1, an interactive hierarchical design approach developed in [16] outperforms other solution techniques described in [15], [11] and [12]. A robust layout procedure was also recommended for dynamic environment in [12], which uses a layout constructed from an expected demand situation or expected flow matrix.

3.1.5 Linear assignment algorithm

A linear assignment algorithm was developed in [7] to create machine cells and part families in cellular manufacturing systems. The bond energy, percentage of exceptional elements, the machine utilization, grouping efficiency, grouping efficacy were used as performance criteria & many existing data sets were taken from literature to compare this linear assignment algorithm with other methods taken from the literature. The results generated by linear assignment algorithm were found almost the same as, equivalent to, or better than the best-known results. The results generated by present linear assignment algorithm were found identical to the best-known results in terms of the five performance criteria in 11 out of the 20 test problems taken from the literature. While in the remaining 9 test problems, the ratio of better performance criteria between the present results and the best-known results taken from the literature was found as 5:4. The results generated by linear assignment algorithm were found dominantly superior in terms of percentage of exceptional elements and bond energy.

The results generated by linear assignment algorithm were compared with the results generated by two commonly used heuristics, ROC [17] and GRAFICS [18]. The results generated by linear assignment algorithm were found extremely better than the results generated by ROC in terms of all five performance criteria. The results generated by linear assignment algorithm were found same as generated by GRAFICS in seven out of the 17 test problems. In the remaining 10 test problems, the ratio of better performance criteria between the linear assignment algorithms and GRAFICS was found as 8:2. This linear assignment algorithm found better solutions compared to both heuristics in terms of machine utilization, grouping efficiency and grouping efficacy.

3.1.6 Similarity coefficient methods

The performance comparison of 20 well-known similarity coefficient methods was presented in [19] to solve 214 cell formation problems which are either taken from the literature or created intentionally. Nine performance measures were used to assess the goodness of solutions. Based on comparison, the performance of Jaccard, Sorenson, and Sokal and Sneath 2 coefficient method was found as best [19]. Jaccard was found as most stable similarity coefficient method [19]. Hamann, Simple matching, Rogers and Tanimoto, and Sokal and Sneath coefficient method was found as inefficient under all conditions, thus not recommendable to solve cell formation problems [19].

3.1.7 Reduced integer programming

Reduced Integer Programming, Ant Colony Optimization & Simulated Annealing were developed in [13] to solve machine layout problem. It was observed in [13] that it is very difficult to find out very good solutions of machine layout problem using RIP because RIP puts the machines in a hexagonal graph and does not consider machine sizes into account. The solution quality of ACO is up to 12% better than RIP. Running time of ACO is high compared to RIP. But running time of ACO can be reduced by implementing parallel ACO running on multiple processors. So it was recommended in [13] to choose ACO over RIP to solve machine layout problem. Solution quality and running time of ACO found better compared to SA on small input size of machine layout problem with 8 to 12 machines. Solution quality and running time of SA exceeded by far compared to ACO when number of machines considered in a problem is 25. It is possible to improve the solution quality of ACO by increasing number of iterations and ants [13]. But the running time of ACO is proportional to both the amount of ants and the number of iterations; this would increase the running time of ACO significantly [13].

When solving the flexible machine layout problem, the Silver-meal lot size algorithm and brute force method of finding the right time to change the layout were expressed in [13]. A limitation to these methods is that future demands are not considered when finding solutions. So a new approach to solve this problem needs to be developed [13].

3.1.8 SAW & TOPSIS methods

Multiple attribute decision making (MADM) concept was used in [20] & cell formation, machine layout and cell layout problem were solved in cellular manufacturing system using a two-stage method in [20]. Initial solution was obtained by TOPSIS and then it was improved by SAW (TOPSIS-IMP-SAW) and TOPSIS (TOPSIS-IMP-TOPSIS) in [20]. The results of this method were compared with well-known approaches available in literature. These comparisons showed that this method offers good solutions for the CMS problem. The grouping efficiency and the grouping efficacy were used as a measure to compare the goodness of methods. "TOPSIS-IMP-SAW" method is better than the "TOPSIS-IMP-TOPSIS" method to solve small scale problems [20]. "TOPSIS-IMP-TOPSIS" method is better than the "TOPSIS-IMP-SAW" method to solve large scale problems [20].

3.2 Meta-heuristics

Meta-heuristics is a framework of heuristics used to update a set of solutions during a search. A number of points at once are searched by meta-heuristics rather than a single point to find out the solution of the problem. They perform more powerful search because they have mechanism which escapes them from trapping into a local minimum.

3.2.1 Genetic algorithm

Two Genetic algorithms were developed in [21] to solve single-row machine layout problems & multi-row machine layout problems in cellular manufacturing system. Genetic-based approaches are able to find out good solutions in reasonably short computational time compared to using enumerative approaches. The advantage of genetic algorithm [21] is that the solution of machine layout problem can be generated using a minimum amount of data giving a benefit in a situation where cost occurred per distance of movement is not known.

A nonlinear integer model of cell formation problem in dynamic condition was first developed in [4] and then solved by Genetic Algorithm, Simulated Annealing and Tabu Search. Simulated Annealing found the better near-optimal solutions in shorter average computational times than Genetic Algorithm and Tabu Search in most of the test problems. It was reported in [4] that if the dimension of machine x part matrix is greater than 10 x 10 and the number of cells is greater than 3 and the number of periods is greater than 2, then Lingo 6 cannot find any solution experimentally for the presented dynamic model. It was also reported in [4] that in general, by improving and developing the GA operations, the chance of finding optimal solution will be increased, because these operations can also be used for generating neighboring solutions in Simulated Annealing, Tabu Search and in any other similar meta-heuristic approach.

In [10], a model was developed for the machine layout problem in cellular manufacturing system with the objective of minimizing total distance travelled by all parts & Genetic Algorithm was implemented to solve the problem. The computational results from [10] show that the GA is capable of obtaining near optimal (99%) solution of the test problem and outperforms other solution techniques described in [15], [16] and [11]. Also GA convergence analysis was performed in [10] to find out the number of generations in which GA converges to a near optimum or optimum solutions for five different periods. All the machine locations are considered to be of equal size in [10]. This work may be extended by considering the machine locations of unequal size.

A genetic algorithm was proposed in [22] to solve the real-sized cell formation problem. Numerical examples show that genetic algorithm is efficient and effective in searching for optimal solutions. 22 problems were selected from the literature and grouping efficacy was used as a measure to test and compare the performance of genetic algorithm with ZODIAC method [23], GRAFICS [24], GATSP-Genetic algorithm [25], GA-Genetic algorithm [26], EA-evolutionary algorithm [27], SA-simulated annealing [28]. The grouping efficacy of the solution found by genetic algorithm [22] is either better than that of other methods or it is equal to the best one in all 22 benchmark problems. In six benchmark problems, the grouping efficacy of the solution found by genetic algorithm [22] is better than that of all other methods. The best solutions for these six benchmark problems are reported for the first time in [22]. In eleven benchmark problems, the solution found by genetic algorithm [22] is as good as the best available solution in the literature. In five benchmark problems, the grouping efficacy of the solution found by all the methods are reported same in [22].

3.2.2 Grouping genetic algorithm

A Grouping Genetic Algorithm (GGA), a special class of genetic algorithms, heavily changed to suit the structure of grouping problems was developed in [29] to solve the cell formation problem. The fundamental advantage of this GGA is that it is able to solve the problem of large input size thus an engineer can use it as a powerful tool to decide a best plant layout from a number of different plant layout options in a short computation times. GGA is applicable to solve industrial problems and it is not trapped in local optima like heuristics [29].

A cell formation-grouping genetic algorithm was developed in [30] to solve cell formation problem. Grouping efficiency & grouping efficacy were used as a measure to compare a grouping genetic algorithm with other methods taken from literature. Cell formation-grouping genetic algorithm outperforms ZODIAC [23] on five of the six data sets, and matches its performance on the sixth when grouping efficiency was taken as a comparison measure. Cell formation-grouping genetic algorithm found the solutions with 17% average improvement compared to ZODIAC. Cell formation-grouping genetic algorithm found the solutions with average improvement of 43% compared to ZODIAC, 2.85% compared to GRAFICS [24] and 1.5% compared to algorithm [31] when grouping efficacy was taken as a measure of effective of solutions. When grouping efficacy was taken as a measure of effective of solutions, Genetic Algorithm [32] outperformed cell formation-grouping genetic algorithm by 8.1% on one data set and by 0.3% on another out of five test data sets. For other three test data sets, Genetic Algorithm matched its performance with cell formation-grouping genetic algorithm. Genetic Algorithm slightly outperformed cell formation-grouping genetic algorithm on two test data sets out of five test data sets, but it requires more number of generation to find out the solutions compared to cell formation-grouping genetic algorithm.

An adapted grouping genetic algorithm was developed in [5] to solve a generalized cell formation problem in cellular manufacturing system. Four test problems were taken from [33] to test this adapted grouping genetic algorithm. Both adapted grouping genetic algorithm and adapted simulated annealing-based heuristic [33] have found the solution with same number of inter-cellular moves for first three test problems. Adapted grouping genetic algorithm has found the solution with less number of inter-cellular moves compared to adapted simulated annealing-based heuristic for fourth test problem. Improvement compared to the previous works was demonstrated by adapted grouping genetic algorithm & it was found fast and efficient with all types of data in term of alternative process plans and alternative routings.

3.2.3 Modified genetic algorithm

Modified Genetic Algorithm based novel approach was developed in [34] to solve cell formation problem in cellular manufacturing system. Machine utilization, grouping efficiency and grouping efficacy was used as an aggregate performance measure of clustering to evaluate the quality of solution found by the Modified Genetic Algorithm for cell formation. Modified genetic algorithm finds out the better quality solution in most cases compared to rank order clustering [17], original ART1 [35], modified ART1 [36]. Modified genetic algorithm finds out the best results to solve cell-formations problem for most simulated examples, or at least same results as found by modified ART1. Modified Genetic Algorithm shows the ability of generalization to solve cell formation problem [34].

3.2.4 Hierarchical genetic algorithm

Hierarchical genetic algorithm was developed in [37] to solve cell formation problem, intracellular machine layout design problem & cell layout design problem concurrently in cellular manufacturing system. The results generated by hierarchical genetic algorithm were evaluated by taking the six test problems from [38]. The results generated by hierarchical genetic algorithm were compared with two existing methods (TOPSIS-IMP-SAW (TIS) and TOPSIS-IMP-TOPSIS (TIT)) that were proposed in [38]. Comparison of results in terms of grouping efficiency and grouping efficacy showed that hierarchical genetic algorithm outperforms two existing approaches TIS and TIT.

3.2.5 Simulated annealing

In [39], a mathematical programming model for the cell formation problem with multiple identical machines, which minimizes the inter-cellular flow, was presented and a simulated annealing algorithm was implemented to solve the problem. A set of problems taken from [40] to test simulated annealing algorithm. A real world case was also solved by simulated annealing algorithm. The efficiency of the model was specified by the computational results, even for large sized problems.

A nonlinear integer model of cell formation problem in dynamic condition was first developed in [4] and then solved by Simulated Annealing is described in 3.2.1. Simulated annealing was developed in [41] to solve cell formation problem in cellular manufacturing system is described in 3.3.3. Simulated Annealing was developed in [13] to solve machine layout problem is described in 3.1.5.

3.2.6 Ant colony optimization

In [14], a mathematical model for inter-cell layout problem was developed & an ant algorithm was developed to solve the problem. The performance of ant algorithm was compared to the facility layout algorithms such as H63 [42], HC63-66 [42], CRAFT [42] and Bubble Search [43] as well as other existing ant colony implementations for QAP such as FANT [44], HAS-QAP [45], MMAS-QAP2-opt [46], and ANTS

[47] algorithms. The performance of ant algorithm was also compared with GH Method [48]. Ant algorithm significantly outperforms the facility layout algorithms & it is effective and efficient as compared to other existing ant algorithms.

Ant Colony Optimization was developed in [13] to solve machine layout problem is described in 3.1.5.

3.2.7 Multi-objective scatter search

A multi-objective scatter search was developed in [49] to solve a dynamic cell formation problem in cellular manufacturing system. Two well-known multi-objective genetic algorithms namely SPEA-II and NSGA-II were implemented in [49] according to their description found from the literature & they were compared with a multi-objective scatter search based on some comparison metrics and statistical approach. The performance of a multi-objective scatter search was found superior compared to two genetic algorithms.

3.2.8 Fuzzy relational data clustering algorithm

A fuzzy relational data clustering algorithm was developed in [50] to solve cell formation problem in cellular manufacturing system. This fuzzy relational data clustering algorithm was found better than mixed-variable fuzzy clustering approach [51] according to the grouping efficiency measure. The fuzzy relational data clustering algorithm is still good to solve cell formation problem where the machine/part matrix has only general numeric data. In total, the fuzzy relational data clustering algorithm presents a realistic solution methodology to solve cell formation problem based on group technology concept, especially when the machine/part matrix has a mixed-variable type with symbolic and fuzzy data.

3.2.9 Evolutionary algorithm

Enhanced evolutionary algorithm was developed in [52] to solve cell formation and layout problems together, based on sequence data. This approach capable of producing high quality solutions was based on enhanced group chromosome scheme, group crossover operator, group mutation operator, and a chromosome repair mechanism. It was noted in [52] that increasing the number of cells and/or machines may require more iterations before convergence to a good solution; the solution space was not affected by the number of parts. Parallel mechanism of evolutionary algorithm provides the algorithm robustness and effectiveness over a variety of ill-structured input matrices [52]. Evolutionary algorithm was found more acceptable compared to other heuristics available in the literature.

3.2.10 Tabu Search

A nonlinear integer model of cell formation problem in dynamic condition was first developed in [4] and then solved by Tabu Search is described in 3.2.1.

3.3 Hybrid Methods

Hybrid methods are formed by combining two methods to solve the problem. An individual method without hybridizing with other method is not able to provide high quality solution in some cases. So, main purpose of hybrid methods is to find out high quality solution of the problem.

3.3.1 Multi-objective design methodology

The application of recently developed multi-objective intra-cell layout & inter-cell layout designs methodologies in a cellular manufacturing was presented in [53] which addressed real time problems from a dynamic food manufacturing and packaging company in Australia. Methodology was developed by hybridizing Non-linear Goal Programming with Simulated Annealing. Layout designs generated by this model were compared with the company existing layout designs. Benefits offered by this model compared to company existing layout design were: safe working environment, 30% reduced material handling cost, half a million dollars reduction in purchasing bar-coding machines, reduced number of lift-trucks needed, increased employees efficiency and reduced waste.

3.3.2 New approach hybridizing local search heuristic with genetic algorithm

A new approach which joins a local search heuristic with a genetic algorithm was developed in [54] to solve cell formation problem in cellular manufacturing system. A set of problems were taken from the literature to test this new approach. The performance of this new approach was found remarkably well. This new approach found solutions that are at least as good as the ones found previously by other methods in the literature. This new approach has improved the previous solutions for 57% of the problems, in some cases by as much as 12%.

3.3.3 Hybridizing genetic algorithm with large-scale optimization techniques

In [55], a complete model which links several known problems in that it joins the cell formation problem, the machine allocation problem, and the part routing problem was presented for designing a cellular manufacturing system & a hybrid solution methodology which combines genetic algorithm and large-scale optimization techniques was developed to solve large-scale capacitated cell formation problems with multiple routings. A limited computational experiment was conducted which compares the solution quality of this hybrid solution methodology with existing solution methods available in the literature concerning smaller problems taken from the literature. Hybrid solution methodology was found capable to find solutions that are at least as good as solutions given by existing methods available in the literature. A more extensive computational study to assess the practicability and the performance of full-scale approach [55] to solve large-scale problems was conducted. The final solutions found show an improvement of 18–40% (average 28%) over the corresponding reference solutions in regards to the objective function. An opportunity to increase the sampling of the space of solutions without much increasing the CPU time per evaluation was achieved by using local improvement strategy together with other mutation operators. This may increase the change of improving the solution quality by opening up the opportunity to explore more solutions [55].

3.3.4 Evolutionary algorithm created by hybridizing standard genetic algorithm with local search

An evolutionary algorithm that improves the efficiency of the standard genetic algorithm by hybridizing it with a local search around some of the solutions it visits and simulated annealing were developed in [41] to solve cell formation problem in cellular manufacturing system. Cell formation problem instances were taken from the literature to assess the performance of both algorithms. Comparison of results of both algorithms with the results of five other algorithms from the literature was presented. Evolutionary method [41] outperformed other Evolutionary algorithms from the literature in solving 8 out of 36 instances of cell formation problem while in solving 26 instances, the solutions generated by Evolutionary method were found same as the best previously known solutions. Not only the best previously known solutions, but also better solutions than previously best known solutions for various problem instances were found by simulated annealing.

3.3.5 Hybrid method created by hybridizing steady state genetic algorithm with local search algorithm

Local Search Algorithm (LSA) and Hybrid Method (HM) were developed in [56] to solve the cell formation problem. Each offspring solution generated with a steady state genetic algorithm was improved by LSA and thus both LSA & steady state genetic algorithm formed a hybrid method. 35 benchmark problems were selected and solved using LSA & HM. Average percentage of variation of results generated by LSA with respect to the best known solutions was 1% in evaluated 35 benchmark problems. Average running time of LSA to solve 35 benchmark problems was reported as 0.64 seconds. Percentage of variation of results generated by HM with respect to the best-known solutions was 0% for 31 evaluated benchmark problems out of 35 selected benchmark problems. HM improved the best-known solution of other three benchmark problems & found 0.01% bad solution compared to the best-known solution of the last benchmark problem.

3.3.6 Hybrid meta-heuristic algorithm created by hybridizing genetic algorithm with variable neighborhood search

A linear fractional programming model for cell formation problem with the objective of maximizing the grouping efficacy was developed in [57] in case the number of cells was not pre-determined. Two test problems were selected from literature to compare the proposed model with the five previous methods which have solved these problems. Comparison showed that this linear fractional programming model outperforms all other methods.

A hybrid meta-heuristic algorithm was developed in [57] in which genetic algorithm and variable neighborhood search (GA-VNS) were combined. 35 test problems from the literature were selected to compare the quality of the solutions of this hybrid meta-heuristic algorithm with 17 previous well-known methods using the grouping efficacy measure. Comparison showed that GA-VNS is a better algorithm for solving the cell formation problem as compared to other methods.

3.4 Exact solution methods

Exact solution methods find out the optimum solution of the problem. But they fail to find out the solution of large-sized NP-Complete optimization problems in a practical length of time. They are used to solve small-sized & medium-sized NP-Complete optimization problems optimally.

3.4.1 Branch-and-bound approach

A new nonlinear mixed-integer programming model was presented in [58] for the facility layout problem in a two-dimensional area with the objective of minimizing the total distance traveled by the material in the shop floor, a technique was used to linearize this model & branch-and-bound approach was developed to optimally solve the proposed mathematical programming model. Comparison of results showed that the total distance traveled by the products was reduced about 41.8% for small-sized and about 44.8% for medium-sized problems by this model as compared to the process layouts for the example problems. Large-sized combinatorial optimization problems can’t be solved using the exact algorithms in a reasonable time [58]. This limitation is also applicable to the branch-and-bound approach because it is an exact method. This model can also be used to determine the processing route of products in an existing machine layout system with the objective of minimizing the total distance traveled by the material in the shop floor.

3.4.2 Dinkelbach algorithm

In [59], the cell formation problem was first converted into an equivalent binary linear fractional programming problem & then solved using Dinkelbach algorithm where the CPLEX 12.2 Optimizer was utilized to solve the binary linear programming problem at each iteration. Dinkelbach algorithm solved 27 out of the 35 benchmarked problems optimally. Remaining 8 problems out of the 35 benchmarked problems, Dinkelbach algorithm prematurely stopped before getting optimal solution because of memory limit. For these 8 problems average grouping efficacy of best known solutions was found as 0.78% better than grouping efficacy of the solutions generated by Dinkelbach algorithm. Running time of Dinkelbach algorithm was reported high compared to meta-heuristics.

4 Analytical review of approaches

Table 1 presents problem-wise classification table of approaches along with their benefits and drawbacks. To solve each of the three problems namely cell formation, machine layout and cell layout problem, approaches are classified into one of four categories namely heuristics, meta-heuristics, hybrid methods and exact solution methods. Classification also shows whether approaches have solved these three problems individually or jointly.

Table 1: Problem-wise classification table of approaches

Problems	**Cell Formation Problem**	**Machine Layout Problem**	**Cell Layout Problem**
Approaches which solve cell formation, machine layout & cell layout problem individually are further categorized as per the following.			
Heuristics	ZODIAC Method, GRAFICS Method, Rank Order Clustering, Minimum Spanning Trees—Clustering Algorithm, CASE Algorithm, Hierarchical Clustering, Linear Assignment Algorithm, Similarity Coefficient Methods	Yaman's Spiral 1, Yaman's Spiral 2, Tang's Approach, Heuristic Algorithm, Novel Construction-cum-improvement Heuristic, Reduced Integer Programming,	H63, HC63-66, CRAFT, Bubble Search,
Benefits	They can solve large-sized NP-Complete Problems efficiently in such situations where exact solution methods fail to find out the solution of NP-Complete Problems in a practical length of time.		
Drawbacks	They may not find out optimum solution and sometimes some heuristics may find out solution that is very far from optimum solution.		
Meta-heuristics	Genetic Algorithm, Cell Formation - Grouping Genetic Algorithm, Modified Genetic Algorithm, Evolutionary Algorithm, Adapted Grouping Genetic Algorithm, Enhanced Grouping Genetic Algorithm, Simulated Annealing, Adapted Simulated Annealing-based Heuristic, Tabu Search, Ant Colony Optimization, Multi-objective Scatter Search, Greedy Randomized Adaptive Search Procedure, Water Flow-like Algorithm, Differential Evolution	Genetic Algorithm, Ant Colony Optimization, Simulated Annealing	Ant Colony Optimization, Genetic Hybrid Algorithm

	Algorithm, Original ART1, Modified ART1, Fuzzy ART, Fuzzy Relational Data Clustering Algorithm, Mixed-variable Fuzzy Clustering Approach		
Benefits	They perform more powerful search compared to heuristics because they have mechanism which escapes them from trapping into a local minimum & number of points at once are searched by them rather than a single point.		
Drawbacks	Compared to heuristics, they consume more time to find out the solution of large-sized NP-Complete Problems.		
Hybrid Methods	New Approach hybridizing Local Search Heuristic with Genetic Algorithm, Simulated Annealing with Variable Neighborhood, Hybrid Meta-heuristic Algorithm which combines Genetic Algorithm with Variable Neighborhood Search, Hybrid Heuristic Algorithm employing both the Boltzmann Function from Simulated Annealing and the Mutation Operator from the Genetic Algorithm, Hybrid Approach which combines Genetic Algorithm with Large-Scale Optimization Techniques	Hybrid Method which combines Non-linear Goal Programming and Simulated Annealing	Hybrid Method which combines Non-linear Goal Programming and Simulated Annealing
Benefits	They find out high quality solution compared to individual heuristics & individual meta-heuristics without hybridizing.		
Drawbacks	Design & Implementation cost of hybrid methods are high because they created by combining two or more methods. Compared to individual heuristics & individual meta-heuristics without hybridizing, they consume more time to find out the solution.		
Exact Solution Methods	Dinkelbach Algorithm	Branch-and-Bound Approach	
Benefits	They find out optimum solution of the problem.		
Drawbacks	Compared to heuristics, meta-heuristics & hybrid methods, they consume much more time to find out the solution. They may fail to find out the solution of large-sized NP-Complete Problems due to hardware limitation.		
Approaches which solve cell formation & machine layout problem jointly are further categorized as per the following.			
Heuristics	CLASS Algorithm		
Approaches which solve cell formation, machine layout & cell layout problem jointly are further categorized as per the following.			
Heuristics	TOPSIS-IMP-SAW, TOPSIS-IMP-TOPSIS		
Meta-heuristics	Hierarchical Genetic Algorithm, Enhanced Evolutionary Algorithm		

It can be clear from the above discussions that most of the researches have solved only cell formation problem whereas many researchers have solved only machine layout problem without first solving cell formation problem. Only few researches have solved cell layout problem & these three problems completely. In order to generate complete cellular manufacturing system design, three problems namely cell formation, machine layout and cell layout problem should be efficiently solved sequentially. Some researchers have solved these problems concurrently by implementing single approach. While solving two or more problems concurrently requires single approach to be implemented but it also wastes computational time & may not find out the good quality solution of machine layout & cell layout problem because cell formation, machine layout and cell layout are sequential step not concurrent in the design of cellular manufacturing system. It makes no sense to start solving machine layout & cell layout problem without first solving cell formation problem completely.

Genetic algorithm based approaches are given more attention as compared to other approaches to solve cell formation, machine layout and cell layout problem. Here selection of appropriate component of GA is of prime importance. Some researchers have selected the roulette wheel selection which sometimes causes the premature convergence and thus genetic algorithm is not able to find out the global optimum solution. In our opinion use of Rank selection may avoid the premature convergence and leads find out the global optimum solution.

Analysis should be performed on the various mutation rate values to find out the suitable mutation rate value for which genetic algorithm has more chances to find out good quality solution or best solution. If suitable mutation rate value is utilized in subsequent implementation of genetic algorithm then it will improve the performance of resulting genetic algorithm.

Population size & maximum generation number of genetic algorithm depends on nature & complexity of problem. Appropriate value of population size & maximum generation number improves the performance of genetic algorithm. So analysis should be performed to find out how population size & maximum generation number are related to the nature & complexity of problem. Analysis is also needed to be performed to find out whether larger population size with less number of generations or smaller population size with large number of generations can improve the performance of genetic algorithm.

5 Conclusion and Future Research Scope

In order to achieve higher production efficiency using cellular manufacturing concept compared to traditional manufacturing, Researchers have solved cell formation, machine layout and cell layout problem in cellular manufacturing system by using various approaches. The paper discussed most of these approaches along with existing performance criteria and compared them. Approaches have been classified into one of four categories namely heuristics, meta-heuristics, hybrid methods and exact solution methods and their benefits and drawbacks are reported in this paper. Genetic algorithm based approaches were widely used by researchers to solve cellular manufacturing system problems. Performance of genetic algorithm can still be improved by selecting appropriate components and proper value of parameters of genetic algorithm.

Future research scope in this area is to develop approaches which can improve best-known solutions of cell formation, machine layout and cell layout problem reported in the literature in order to further minimize total manufacturing costs and time of manufacturing industries. Further work in the area can ultimately optimize the performance of industrial manufacturing processes.

REFERENCES

[1]. M. P. Chandrasekharan, & R. Rajagopalan, "An ideal seed non-hierarchical clustering algorithm for cellular manufacturing", International Journal of Production Research, vol. 24(2), 1986, pp. 451–463.

[2]. M. P. Chandrasekharan, & R. Rajagopalan, "MODROC: an extension to rank order clustering for group technology", International Journal of Production Research, vol. 24(5), 1986, pp. 1221–1233.

[3]. Kumar, C. S., & Chandrasekharan, M. P., "Grouping efficacy: A quantitative criterion for goodness of block diagonal forms of binary matrices in group technology", International Journal of Production Research, vol. 28(2), 1990, pp. 233–243.

[4]. Tavakkoli-Moghaddam R., Aryanezhad M. B., Safaei N., & Azaron A., "Solving a dynamic cell formation problem using metaheuristics", Applied Mathematics and Computation, vol. 170(2), 2005, pp. 761-780.

[5]. Vin E., Delchambre A. & Francq P., "A generalized cell formation problem solved by an adapted GGA", 8th International Conference of Modeling and Simulation - MOSIM'10, May 10-12, 2010.

[6]. McCormick Jr., W. T., Schweitzer, P. J., & White, T. W., "Problem decomposition and data reorganization by a clustering technique", Operations Research, vol. 20(5), 1972, pp. 993–1009.

[7]. Wang J., "Formation of machine cells and part families in cellular manufacturing systems using a linear assignment algorithm", Automatica, vol. 39(9), 2003, pp. 1607-1615.

[8]. Collette Y, & Siarry P., "Multi-objective optimization: principles and case studies", New York: Springer, 2003.

[9]. Mahdavi, I. and Mahadevan, B., "CLASS: An algorithm for cellular manufacturing system and layout design using sequence data", Robotics and Computer-Integrated Manufacturing, vol. 24(3), 2008, pp. 488-497.

[10]. V. Anbumakar and G. Prabhaharan, "Machine assignment in cellular manufacturing layout using genetic algorithm", The International Journal of Applied Management and Technology, vol. 4(1), 2006, pp. 165-179.

[11]. Chan W. M., Chan C. Y. & Ip W. H., "A heuristic algorithm for machine assignment in cellular layout", Computers & industrial Engineering, vol. 44, 2002, pp. 49-73.

[12]. I. B. Hunagund & M. Pillai V, "Development of a heuristic for layout formation and design of robust layout under dynamic demand", Proceedings of the International Conference on Digital Factory, ICDF 2008, August 11-13, 2008 Organized, pp. 1398-1405.

[13]. Andersen R., "Solution methods to the machine layout problem", Doctoral dissertation, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2006.

[14]. Solimanpur M., Vrat P. & Shankar R., "Ant colony optimization algorithm to the inter-cell layout problem in cellular manufacturing", European Journal of Operational Research, vol. 157(3), 2004, pp. 592-606.

[15]. Yaman.R, Gethin.D.T. and Clarke.M.J., "An effective sorting method for facility layout constructions", International Journal of Production Research, vol. 31(2), 1993, pp. 413-427.

[16]. Tang.C. and Abdel-Malek.L.L., "A frame work for hierarchical interactive generation of cellular layout", International Journal of Production Research, vol. 34(8), 1996, pp. 2133-2162.

[17]. J. R. King, "Machine-component grouping in production flow analysis: an approach using a rank order clustering algorithm", International Journal of Production Research, vol. 18 (2), 1980, pp. 213–232.

[18]. Srinivasan, G., Narendran, T. T., & Mahadevan, B., "An assignment model for the part-families problem in group technology", International Journal of Production Research, vol. 28(1), 1990, pp. 145–152.

[19]. Yin Y. & Yasuda K., "Similarity coefficient methods applied to the cell formation problem: a comparative investigation", Computers & Industrial Engineering, vol. 48(3), 2005, pp. 471-489.

[20]. Ahi A., Aryanezhad M. B., Ashtiani B., & Makui A., "A novel approach to determine cell formation, intracellular machine layout and cell layout in the CMS problem based on TOPSIS method", Computers & Operations Research, vol. 36(5), 2009, pp. 1478-1496.

[21]. Morad N., "Genetic Algorithms Optimization for the Machine Layout Problem", International Journal of the Computer, the Internet and Management, vol. 8(1), 2000.

[22]. Mahdavi I., Paydar M. M., Solimanpur M. & Heidarzade A., "Genetic algorithm approach for solving a cell formation problem in cellular manufacturing", Expert Systems with Applications, vol. 36(3), 2009, pp. 6598-6604.

[23]. Chandrasekharan, M. P., & Rajagopalan, R., "ZODIAC: An algorithm for concurrent formation of part-families and machine-cells", International Journal of Production Research, vol. 25(6), 1987, pp. 835–850.

[24]. Srinivasan, G., & Narendran, T. T., "GRAFICS-A nonhierarchical clustering algorithm for group technology", International Journal of Production Research, vol. 29(3), 1991, pp. 463–478.

[25]. Cheng, C., Gupta, Y., Lee, W., & Wong, K., "A TSP-based heuristic for forming machine groups and part families", International Journal of Production Research, vol. 36, 1998, pp. 1325–1337.

[26]. Onwubolu, G. C., & Mutingi, M., "A genetic algorithm approach to cellular manufacturing systems", Computers and Industrial Engineering, vol. 39(1), 2001, pp. 125–144.

[27]. Goncalves, J., & Resende, M., "An evolutionary algorithm for manufacturing cell formation", Computers and Industrial Engineering, vol. 47, 2004, pp. 247–273.

[28]. Wu, T. H., Chang, C. C., & Chung, S. H., "A simulated annealing algorithm for manufacturing cell formation problems", Expert Systems with Applications, vol. 34, 2008, pp. 1609–1617.

[29]. De Lit P., Falkenauer E., & Delchambre A., "Grouping genetic algorithms: an efficient method to solve the cell formation problem", Mathematics and Computers in simulation, vol. 51(3), 2000, pp. 257-271.

[30]. Brown E. C., & Sumichrast R. T., "CF-GGA: a grouping genetic algorithm for the cell formation problem", International Journal of Production Research, vol. 39(16), 2001, pp. 3651-3669.

[31]. SRINIVASAN, G., "A clustering algorithm for machine cell formation in group technology using minimum spanning trees", International Journal of Production Research, vol. 32, 1994, pp. 2149-2158.

[32]. JOINES, J., CULBRETH, C. T. and KING, R. E., "Manufacturing cell design: an integer programming model employing genetic algorithms", IEE Transactions, vol. 28, 1996, pp. 69-85.

[33]. Sofianopoulou, S., "Manufacturing cell design with alternative process plans and/or replicate machines", International Journal of Production research, vol. 37(3), 1999, pp. 707-720.

[34]. Car Z. & Mikac T., “Evolutionary approach for solving cell-formation problem in cell manufacturing”, Advanced Engineering Informatics, vol. 20(3), 2006, pp. 227-232.

[35]. S. J. Chen, & C. S. Cheng, “A neural network-based cell formation algorithm in cellular manufacturing”, The International Journal of Production Research, vol. 33 (2), 1995, pp. 293–318.

[36]. Mikac T, Jurkovic´ M, & Pekic´ Z., “A neural network approach in cellular manufacturing”, in: Proceedings of the 5th International Conference on Advanced Manufacturing Systems and Technology, AMST ’99, 1999, pp. 521–528.

[37]. K. Chandrasekar, P. Venkumar, “Genetic algorithm approach for integrating cell formation with machine layout and cell layout”, International Journal of Operational Research, vol. 16(2), 2013, pp. 155-171.

[38]. Ahi, A., Aryanezhad, M. B., Ashtiani, B. and Makui, A., “A novel approach to determine cell formation, intracellular machine layout and cell layout in the CMS problem based on TOPSIS method”, Computers and Operations Research, vol. 36, 2009, pp. 1478–1496.

[39]. Xambre A. R., & Vilarinho P. M., “A simulated annealing approach for manufacturing cell formation with multiple identical machines”, European Journal of Operational Research, vol. 151(2), 2003, pp. 434-446.

[40]. Wu, N., & Salvendy, G., “An efficient heuristic for the design of cellular manufacturing systems with multiple identical machines”, International Journal of Production Research, vol. 37(15), 1999, pp. 3519–3540.

[41]. Pailla A., Trindade A. R., Parada V. & Ochi L. S., “A numerical comparison between simulated annealing and evolutionary approaches to the cell formation problem”, Expert Systems with Applications, vol. 37(7), 2010, pp. 5476-5483.

[42]. Nugent, C. E., Vollman, T. E., & Ruml, J., “An experimental comparison of techniques for the assignment of facilities to locations”, Operations Research, vol. 16(1), 1968, pp. 150-173.

[43]. Wang, S., & Sarker, B. R., “Locating cells with bottleneck machines in cellular manufacturing systems”, International Journal of Production Research, vol. 40 (2), 2002, pp. 403-424.

[44]. Taillard, E. D., & Gambardella, L. M., “Adaptive memories for the quadratic assignment problem”, Technical Report IDSIA-87-97, IDSIA, Lugano, Switzerland, 1997.

[45]. L. M. Gambardella, E. Taillard, and M. Dorigo, “Ant colonies for QAP,” IDSIA, Lugano, Switzerland, Technical Report IDSIA 97-4, 1997.

[46]. Stutzle, T., & Hoos, H. H., “MAX-MIN ant system”, Future Generation Computer Systems, vol. 16(8), 2000, pp. 889-914.

[47]. Maniezzo, V., & Colorni, A., “The ant system applied to the quadratic assignment problem”, IEEE Transactions on Knowledge and Data Engineering, vol. 11 (5), 1999, pp. 769-778.

[48]. C. Fleurent, J. A. Ferland, “Genetic hybrids for the quadratic assignment problem”, in: P.M. Pardalos, H. Wolkowicz (Eds.), Quadratic Assignment and Related Problems, DIMACS Series on Discrete Mathematics and Theoretical Computer Science, vol. 16, American Mathematical Society, Providence, RI, 1994, pp. 173–187.

[49]. Bajestani M. A., Rabbani M., Rahimi-Vahed A. R. & Baharian Khoshkhou G., "A multi-objective scatter search for a dynamic cell formation problem", Computers & operations research, vol. 36(3), 2009, pp. 777-794.

[50]. Hung W. L., Yang M. S. & Lee E. S., "Cell formation using fuzzy relational clustering algorithm", Mathematical and computer Modelling, vol. 53(9), 2011, pp. 1776-1787.

[51]. M. S. Yang, W. L. Hung, & F. C. Cheng, "Mixed-variable fuzzy clustering approach to part family and machine cell formation for GT applications", International Journal of Production Economics, vol. 103(1), 2006, pp. 185–198.

[52]. Mutingi M., Mbohwa C. T., Mhlanga S., & Goriwondo W., "Integrated Cellular Manufacturing System Design: an Evolutionary Algorithm Approach", Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management, July 3 – 6, 2012, pp. 254-264.

[53]. Bazargan-Lari.M., "Layout designs in cellular manufacturing", European Journal of Operational Research, vol. 112, 1999, pp. 258-272.

[54]. Gonçalves J. F. & Resende M. G., "An evolutionary algorithm for manufacturing cell formation", Computers & Industrial Engineering, vol. 47(2), 2004, pp. 247-273.

[55]. Nsakanda A. L., Diaby M. & Price W. L., "Hybrid genetic approach for solving large-scale capacitated cell formation problems with multiple routings", European Journal of Operational Research, vol. 171(3), 2006, pp. 1051-1070.

[56]. Elbenani B., Ferland J. A., & Bellemare J., "Genetic algorithm and large neighbourhood search to solve the cell formation problem", Expert Systems with Applications, vol. 39(3), 2012, pp. 2408-2414.

[57]. Paydar M. M., & Saidi-Mehrabad M., "A hybrid genetic-variable neighborhood search algorithm for the cell formation problem based on grouping efficacy", Computers & Operations Research, vol. 40(4), 2013, pp. 980-990.

[58]. Maghsud Solimanpur and Amir Jafari, "Optimal solution for the two-dimensional facility layout problem using a branch-and-bound algorithm", Computers & Industrial Engineering, vol. 55(3), 2008, pp. 606–619.

[59]. Elbenani, B., & Ferland, J. A., "Cell Formation Problem Solved Exactly with the Dinkelbach Algorithm", Publication CIRRELT-2012-07, University of Montreal, Canada, 2012.

SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 2, ISSUE 5
ISSN: 2054 - 7390

Behavioral Modeling of Typical Non-Ideal Analog to Digital Converter Using MATLAB

[1]Sayed Mostafa Taheri and [2]Bahareh Mohammadi
[1] Signal Processing and Communications Group, Electronics Engineering Department
Barthawa Institute of Technology, Mashhad, Iran;
[2] Communication and Information Systems Group, Electronics Engineering Department
Barthawa Institute of Technology, Mashhad, Iran;
[1]s.m.taheri@ieee.org; [2]b.mohammadi@agb.prof.ir

ABSTRACT

Analog-to-digital converters (ADC) are one the most fundamental and crucial parts of an electronic device which needs to convert analog inputs into the digital format. Nowadays, you can find some type of ADCs in almost every communication device. Consequently, simulating an ADC plays a challenging and very fundamental role while aiming to simulate real-world instruments and designed systems. Despite the fundamental significance of this task, a general simulation model for typical non-ideal ADCs have not been presented yet. In this paper, we consider the most important non-ideality parameters of a typical ADC and initiate simulation models in the Simulink environment of MATLAB, as the most versatile widely-used simulation software for engineering tasks. For this purpose, we first reconsider the non-ideality parameters of a typical ADC from general perspective, aiming to bring a unified view in hand. Afterwards, the potential models for the non-ideality parameters are initiated separately, and then, are combined together to make the whole model. By creating a MATLAB Block of the whole general-topology model at the end, and defining the non-ideality parameters as variable inputs, we simulate the behavior of a typical ADC using practical non-ideality parameters data, and compare with an ideal ADC. This additionally enables a user interface for quick input of non-ideality parameters of any practical ADC for future clients. Despite the pure generality of the model, simulation results demonstrate acceptable outcomes proving the applicability of this model in wide range of engineering simulation tasks, and an appreciable step towards higher precision general simulations of typical ADCs.

Keywords: Analog to digital converter, non-ideal ADC, behavioral modeling and simulation of ADC, MATLAB Simulink, typical modeling of non-ideal ADC.

1 Introduction

Analog to digital conversion has been a tremendously active field in digital signal processing and communications realm. Due to its wide vast range of applications, as well as high demand for up-to-date and capable analog-to-digital converters (ADC or A/D) to be used in digital communication and electronic systems satisfying today's requirements, it has always been an attractive and ongoing field of study and research. Efforts have been put to design and implement high precision, fast, and reliable ADCs over the past decades. There have been variety of different types of ADCs being proposed for

DOI: 10.14738/tmlai.25.567
Publication Date: 3th November, 2014
URL: http://dx.doi.org/10.14738/tmlai.25.567

different applications having various characteristics and features. As a general concept, ADCs are known to be non-linear non-ideal instruments as a result of being comprised of non-linear circuits. Therefore, the non-ideality parameters of ADCs usually imply some constraints to be looked after while employing them in real-world system designs. From this perspective, it is quit crucial to be able to predict and handle these non-idealities while designing electronic or communications systems. To do this, many researches have been done to study, improve, and overcome the non-ideality parameters of different types of ADCs; [1]–[15] are some examples. These studies could be generally divided into two main categories. The first category consist of researches which are focused on infrastructure design and improvements of ADCs, i.e., circuit-level studies. The second class contains the works aiming to predict or simulate the performance of different types of ADCs based on their infrastructure characteristics and features, usually in MATLAB Simulink or Spice. [2], [3], [6]–[10], [12]–[14] are some examples for this category. Generally speaking, the first category has attracted more interest than the other over the past decade. However, with the development of capable computer systems, the need for more works in the second category has been raised. Given these facts, simulation, as a powerful, cost-effective, and time-saving tool plays an unsurpassed role to address these issues. Nevertheless, despite the fundamental importance of simulation tasks they has not been truly employed to predict and evaluate the performance of ADCs, typically. Performing high precision, fast, and cost-efficient simulations is therefore a real need in the present circumstances. Nonetheless, due to the generally complicated circuitry characteristics of these converters, modeling their non-ideality parameters is a challenging task. As a result, previous works are mainly focused on modeling and simulation of *some* of the particular non-ideality parameters of *specific types* of ADCs. However, the employability of these works are presumably restricted as a consequence of being proposed for only specific types of ADCs, and can not be reused for others. Moreover, in real world situations and practical applications, we are sometimes confronted with circumstances wherein the type and infrastructure properties of the ADC is fully unknown. Presumably, these sort of methods will not be employable in such circumstances. In this paper, we propose considering an ADC from a general point of view with its main characteristics and features, and try to initiate models which are appropriate for the most influential nonideality parameters of a typical ADC. As a result, the generality of this work is quite vaster than previous researches, being capable to be applied on and employed for generally any analog to digital converter without having to look at its infrastructure properties. Additionally, it makes it competent to be employed in practical situations where the type and circuitry features of the ADC is not known. The initiated models are described with appropriate details in Simulink environment, and eventually implemented for a typical ADC using practical non-ideality parameters data. The simulation results are presented to show the effect of the non-ideality parameters being modelled. Simulation results support the potential of the initiated models for the future higher precision general simulations of typical ADCs. The proposed Simulink model is designed in such a way to be reproducible and employable for fast and cost-effective handy simulation of a typical non-ideal ADC repeatedly.

The rest of the paper is organised as follows. In Section 2 we reconsider the required preliminaries and bring a general view of the most important non-ideality parameters of a typical ADC into hand. In Section 3 we commence initiating appropriate models for the non-ideality parameters of a typical ADC discussed in Section 2 in Simulink environment. In Section 4, by creating a Simulink Block called typical non-ideal digital converter (TNDC), we perform the simulation task using real-world values and compare some results with an ideal ADC. Section 5 presents the limitations and describes some directions for future research, and eventually in Section 6, the conclusion of the work is presented.

2 Preliminaries and Unified View

Although there are various types of ADCs having different structures, nevertheless, the non-ideality parameters of a typical ADC can be generally classified into two main categories: static specifications and dynamics. The transfer function of an ideal ADC has been shown in Figure. 1 [16].

Static specifications of an ADC causing inaccuracy in conversion can be completely described by the following errors [16]:

- Finite resolution
- Offset error
- Differential nonlinearity (DNL)
- Integral nonlinearity (INL)

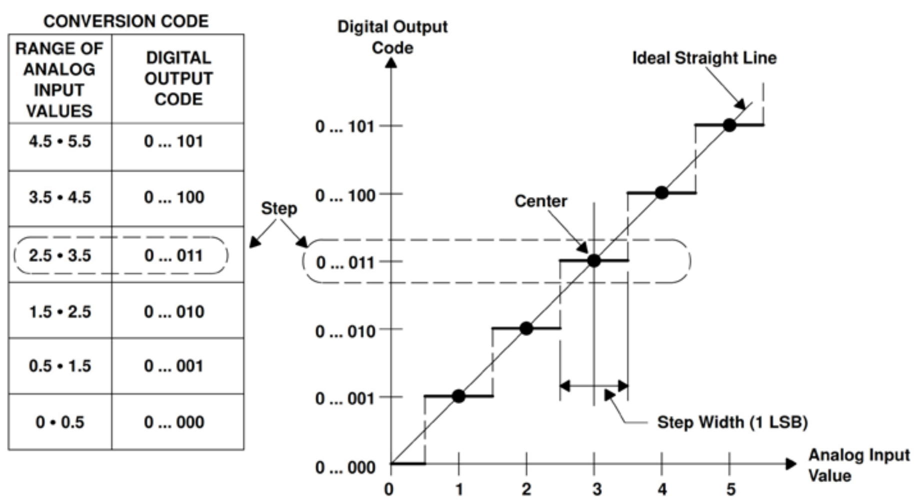


Figure 1: General transfer function of an ideal ADC [16]**.**

Followings can be categorised as the dynamic specifications of a non-ideal ADC:

- Signal-to-noise ratio (SNR)
- Total harmonic distortion (THD)
- Signal-to-noise-and-distortion-ratio (SNDR)
- Effective number of bits (ENOB)
- Spurious-free dynamic range (SFDR)

While we do not aim to demonstrate extensive explanations of the ADC parameters in this article, however, reminding the fundamental concepts from a unified point of view can be handful for the next section. We assume these classified expressions along with the included figures can remind the reader about more in-depth details.

2.1 Static Non-Ideality

The static non-ideality parameters of an ADC can be stated in terms of both the least significant bit (LSB) and percentage of the full scale range $\%FSR$, which will be defined as follows. These could easily be converted to each other, as will be presented in Equation (39).

2.1.1 Finite Resolution

The resolution of an ADC is usually the number of bits which shows the maximum number of distinct levels available in conversion process [17]. In other words, the number of bits generally determines the resolution of the data acquisition system. A/D resolution can be mathematically expressed as follows:

$$V_{LSB} = V_{FSR}/(2^n - 1) \quad (1)$$

where V is the maximum input voltage and n is the number of conversion bits. LSB is the least significant bit at low output, and FSR is the full scale range, i.e. maximum input voltage at 2^n output, which is unreachable. In other words, there are 2^n possible digital output codes for an n-bit converter which each value is an equal fraction of the total input voltage range. Equation (1) is for a binary converter. However, it can be rewritten for decimal ADCs as follows:

$$V_{LSB} = V_{FSR}/(10^D) \quad (2)$$

wherein D represents the number of decimal digits. In this paper, we work with binary A/Ds due to their vast popularity.

2.1.2 Offset Error

The offset error is usually defined as the difference between the nominal and actual offset points on the analog output value vs. digital output code diagram of the ADC (Figure. 2)[16]. The nominal offset point for an ADC is the midstep value for which the digital output code is zero. This error has an equal influence on all codes by the same amount [16].

2.1.3 Gain Error

The gain error is usually defined as the difference between the nominal and actual gain points on the analog input value vs. digital output code transfer function of the ADC (Fig. 2) after the offset error is rectified to zero [16]. For an ADC, the nominal gain point is the midstep value for which the digital output code is at full range. In essence, this non-ideality error points out a change in the slope of the actual and ideal transfer functions, and thus matches the same percentage error at each step [16].

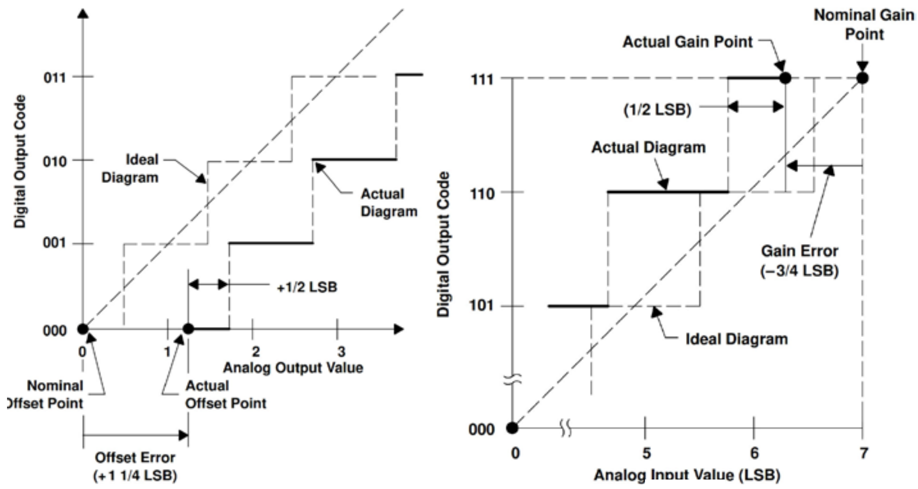


Figure 2. 3-bit converter: (a) Offset Error; (b) Gain Error.

2.1.4 Differential Nonlinearity Error

The differential nonlinearity error for an ADC, as shown in Figure. 3, is the dissimilarity between each actual steps widths and the ideal value of 1 LSB widths [16]. Hence, if the steps widths are exactly 1 LSB, in that case the differential nonlinearity error is exactly zero. Nonetheless, in cases where the DNL exceeds the ideal 1 LSB width, there are possibility of encountering variety of non-linear behaviours, including the A/D turns to be nonmonotonic. This signifies that the magnitude of the output could become smaller in spite of increment in the magnitude of the input. Some other possibilities of non-linear behaviour could include having missing codes i.e., one or more of the possible 2^n binary values are never seen output. The following equation could therefore be used to describe or evaluate the DNL error for each digital (D) code mathematically:

$$DNLm = \frac{V_m^Q - V_{m-1}^Q - V_{LSB}}{V_{LSB}}. \qquad (3)$$

The overall DNL is usually referred to as the maximum of individual DNLs:

$$DNL = \max(DNL_m), \qquad m \in [0, 2^n - 1]. \qquad (4)$$

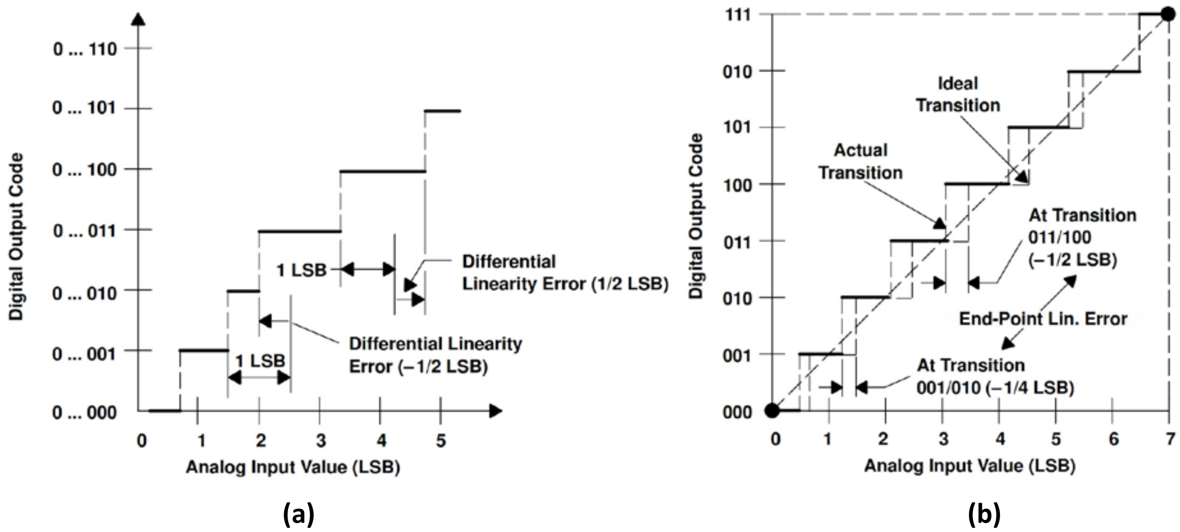


Figure 3. 3-bit converter: (a) Differential nonlinearity error; (b) Integral nonlinearity error [16]. Offset and gain errors are set to zero.

2.1.5 Integral Nonlinearity Error

The integral nonlinearity error is referred to the deviation in values of the actual transfer function from an ideal straight line. Figure. 3 shows this concept [16]. This straight line can be drawn so as to minimize the deviations, or it can also be a line fitted between the end points of the transfer function while the gain and offset errors have been corrected to zero e.g. by trimming. For a typical ADC these deviations are measured at the transition points from one step to the other. In fact, the summation of the individual differential nonlinearities (DNLs) from the bottom up to a particular step i.e.,k, determines the value of the integral nonlinearity at that step. The following equation describes this fact:

$$INL_k = \sum_{i=1}^{i=k} DNL_i, \tag{5}$$

or

$$INL_m = \sum_{m=0}^{2^n-1} DNL_m, \tag{6}$$

and the overall INL is defined as follows:

$$INL = \max(INL_m), \qquad m \in [0, 2^n - 1]. \tag{7}$$

2.2 Dynamic Non-Ideality

2.2.1 SNR

SNR refers to the power of signal in comparison to the power of *permanent* noise, simply written as:

$$SNR = \frac{P_s}{P_n}. \tag{8}$$

This, while converted to *dB*, follows the subsequent equation:

$$SNR_{dB} = 10\log\left(\frac{P_s}{P_n}\right) = 6.02N + 1.76 \quad dB \tag{9}$$

wherein N denotes the converter's total number of bits (maximum resolution). The noise mentioned in above equations are usually counted as quantization noise (error) which is inevitable even in ideal ADCs. This obeys the following formula:

$$V_{Q,rms} = \frac{V_{LSB}}{\sqrt{12}} = \frac{V_{FSR}}{2^N\sqrt{12}} \tag{10}$$

in which Q stands for quantization and rms is the root mean square; and equation (9)can be concluded from (10) [18], [19]. As a result, the SNR defined above is the ideal signal to noise ratio in any N-Bit A/D converter.

2.2.2 THD

Due to the non-linear characteristics of converter circuits' elements, other harmonics of the input signal also appear at the output. In other words, the multiplies of the main signal's frequency also appear at the output and is present along with the main frequency (i.e., nf_{sig} or nf_{in}). Assume having a single-tone signal at the input, this property is usually described by the following equations:

$$x_i = a_i\sin(2\pi f_i t + \varphi) \tag{11}$$

$$y_o = \sum_{k=1}^{m} \zeta_k x_i^k = \zeta_1 x_i + \zeta_2 x_i^2 + \cdots$$

$$= \zeta_1 a_i\sin(2\pi f_i t + \varphi) + \zeta_2 a_i\sin^2(2\pi f_i t + \varphi) +$$

$$= \xi_1\sin(2\pi f_i t + \varphi) + \frac{\xi_2}{2}(1 - \cos(4\pi f_i t + \varphi)) + \ldots, \tag{12}$$

in which a_i, ζ_k and ξ_k are the amplitude coefficients, and x_i and y_o represent the input and output signal to the non-linear system, respectively. As can be seen, second, third and so forth multiplies of the f_i appears at the output unwontedly. Similarly, equations (11) and (12) could be modified to the following forms in cases of working with multi-tone signals:

$$x'_i = \sum_{i=1}^{n} a_i\sin(2\pi f_i t + \varphi) \tag{13}$$

and hence,

$$y_o = \sum_{k=1}^{m} \zeta_k x_i^k = \sum_{k=1}^{m}\sum_{i=1}^{n} \zeta_k a_i^k \sin^k(2\pi f_i t + \varphi)$$
$$= \sum_{k=1}^{m}\sum_{i=1}^{n} \zeta_k a_i^k \left(\sum_{j=1}^{p} \sin(2\pi(kf_i \pm k'f_i)t + \psi)\right) \tag{14}$$

In analog to digital conversion realm, similar to the second form mentioned above is usually called intermodulation distortion (IMD), which is the presence of distortions with additive frequencies at the output in addition to the correct multiples of the main frequency. In this paper we work with a tone sinusoid. After simplifications, equations (12) can be rewritten as:

$$y_o = HD_1 \sin(2\pi f_i t + \varphi_1) + HD_2 \sin(4\pi f_i t + \varphi_2) +$$

$$= \sum_{i=1}^{n} HD_i \sin(2\pi n f_i t + \varphi_i). \tag{15}$$

Generally, the amplitude of the distortion harmonics above the fourth (i.e. HD_4) is very small and negligible, i.e. $HD_i \approx 0$ for $i > 4$. Consequently, the total harmonic distortion (THD) can be calculated by calculating the distortion power as follows:

$$P_{dist} = \sum_{i=2}^{\infty} (\frac{HD_i}{\sqrt{2}})^2 = \sum_{i=2}^{\infty} \frac{HD_i^2}{2}, \tag{16}$$

and eventually:

$$THD = \frac{\sqrt{\sum_{i=2}^{\infty} HD_i^2}}{V(f_{sig})}. \tag{17}$$

Given the fact explained above, equation (17) reduces to the following:

$$THD = \frac{\sqrt{\sum_{i=2}^{4} HD_i^2}}{V(f_{sig})}, \tag{18}$$

where V represents the voltage or amplitude of the signal. THD is usually specified in units of dBc (dB to carrier, i.e. f_i).

2.2.3 SNDR

Signal to noise and distortion ratio (SNDR) is usually referred to the fact that, the SNR contains the inevitable quantization noise only, as explained before. However, there are other types of noises involved, i.e. thermal noise. As a result, the SNDR parameter takes into account the quantization noise, thermal noise, and the total harmonic distortion. The following equations demonstrate this fact:

$$SNDR_{dB} = 10\log\left(\frac{P_{signal}}{P_{noise+dist.}}\right) \tag{19}$$

$$SNDR = 10\log\left(\frac{V_{sig}^2}{V_Q^2+V_{n,th}^2+V_{dist}^2}\right). \tag{20}$$

Therefore, we can easily conclude that the following inequality is always true:

$$SNDR_{dB} \leq SNR_{dB}. \tag{21}$$

It is worth mentioning that SNDR is sometimes recalled as signal to noise and distortion ratio (SINAD) as well.

2.2.4 SFDR

The spurious-free dynamic range (SFDR) refers to the strongest harmonic's (f_{spur}) power to the main signal's power. It is therefore usually given in units of dBFS (dB full scale) or dBc (dB to carrier). The following equation describes this:

$$SFDR_{dB} = 10\log\left(\frac{V_{f_{sig}}^2}{V_{f_{spur}}^2}\right). \tag{22}$$

Commonly, the ADC's manufacturer specifies the harmonic which has the highest power e.g. HD_i. Figure. 4 shows this concept.

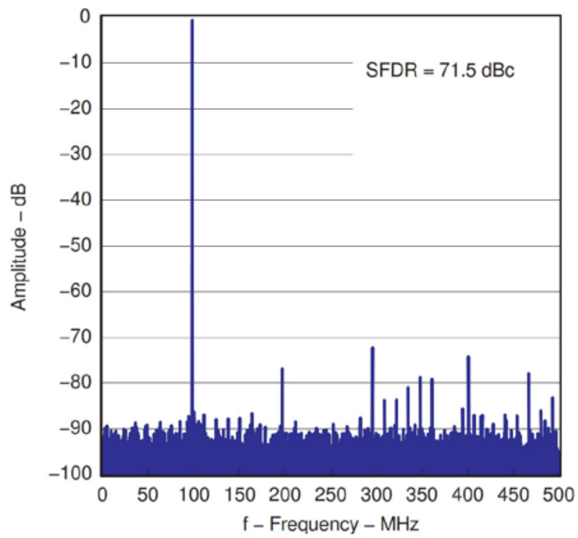


Figure 4: Spurious-free dynamic range (SFDR) [20].

For a pure sine wave input, SFDR can be seen as the ratio of the amplitude of the averaged Fourier transform (FT) value at f_{sig} or f_i, to the averaged FT amplitude of the largest magnitude harmonic or spurious signal component being detected in the full Nyquist band i.e., $\max\{V_{avg}[f_h] \cup V_{avg}[f_{spur}]\}$. The following equation describes this [1]:

$$SFDR_{dB} = 20\log\left(\frac{|V_{avg}(f_{sig})|}{\max_{f_{spur} \cap f_h}\left(|V_{avg}(f_{spur})| \cup |V_{avg}(f_h)|\right)}\right) \tag{23}$$

in which avg denotes the averaged value and h stands for harmonic.

2.2.5 ENOB

Effective number of bits (ENOB) shows the concept associated with the SNDR explained in Section 2.2.3. Recall equation (9) where N represents the ADC's total number of bits. However, based on inequality (21)the *practical* signal to noise ratio is always less than the *nominal* one in equation (9). Thus, if we replace the total number of bits in (9) with the effective number of bits, we reach the value specified in (20), hence:

$$ENOB = \frac{(SNDR - 1.76)_{dB}}{6.02}.$$

3 Non-Ideality Parameters and Modeling

In this section, we initiate considering the most important and influential nonideality parameters of a typical ADC having general point of view in mind. This will assure that the research is not identical to previous efforts taking the infrastructure of ADCs into account, which thus restricted the generality of the proposed models for other types of ADCs. As a result, in this paper the initiated models are not produced by discussing circuit-level characteristics of a specific type of ADC. To achieve this main objective, we ought to consider the nonideality parameters of a practical ADC, looking from a widespread perspective. To do this, we propose the initiated models for these non-ideality parameters recalling the unified view discussed in Section 2. It is worth noting that the original analog input signal could be generally defined to be the sum of ℓ signals, $x_{in,\ell}(t)$, having the following format:

$$x_{in,\ell}(t) = a_{\ell,0} + \sum_{k=1}^{+\infty} a_{\ell,k} \sin(2\pi f_\ell kt + \psi_{\ell,k}), \tag{24}$$

in which the coefficients of the respective Fourier series are used. Consequently, the total input signal can be written as follows [14]:

$$x_{in}(t) = \sum_{\ell=0}^{\ell-1} x_\ell(t). \tag{25}$$

In this paper we aim at working with a tone conventionally. Nevertheless, the proposed model has the flexibility to be employed for the above definitions as well.

3.1 Total Harmonic Distortion

There are two ways of modeling the THD. In ideal situations, we assume that the input signal is known, or specified by the client manually. Therefore, the parameters of the input signal such as frequency, amplitude and phase are fully known. Having this, we can easily model the THD by defining the input frequency (and other parameters if applicable) as a variable, converting the HD_i data from datasheet to pure amplitudes, define a new wave with the frequency equal to if_{sig} and add them up together. Figure. 5 shows the model diagram for the second, third, and fourth distortion harmonics.

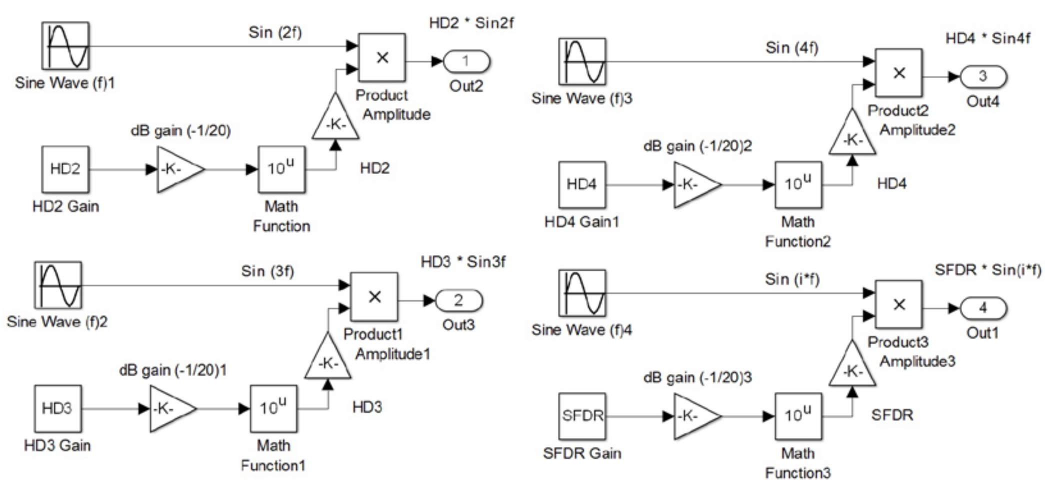


Figure 5: model (for $HD_i \approx 0$ for $i \pm 5$), and SFDR model ($f_{spur} = if_{sig}$). More harmonics can be added if needed.

We can obviously add more distortion harmonics to the above model as needed, and create a subsystem for ease of use. Nevertheless, as the second case, we can assume that the full information about the input signal might not be available. In this situation, we must use the input signal itself to generate the distortion harmonics. This is possible using sine and cosine expansions:

$$\sin^2\Theta = \frac{1-\cos 2\Theta}{2} \tag{26}$$

$$\cos^2\Theta = \frac{1+\cos 2\Theta}{2} \tag{27}$$

Other powers of the sine and cosine can be generated using the above main formulas. The cosine in (26) can be generated from the input signal by applying the Hilbert transform. We should note that for simulation purpose, x samples delay equal to the half of the Hilbert transform order should be made. It is also worth noting that shifting the input phase by $\pi/2$ will make some problems for the simulation and should not be always used. Another method could be usage of input derivatives, however, this can also make some problems, specially if the input signal is sampled or discrete in time. Hence, the output after this step is as follows:

$$x_{out,THD}(t) = x_{in}(t)|_{f=f_{sig}} + \sum_{i=2}^{p} HD_i x_{in}(t)|_{f_i=if_{sig}} \tag{28}$$

in which $x_{out,THD}$ denotes the output after the THD modeling step.

3.2 Spurious-Free Dynamic Range

The SFDR can be easily modeled by considering the information in Section 2.2.4. To do this, while applying the THD method discussed in Section 3.1 the SFDR of the ADC can also be constructed taking the information in ADC manufacturer's datasheet into account. For instance, the SFDR for the Texas Instruments ADS5400 A/D at f_{in} = $1200MHz$ is $66dBc$, and this occurs at the second harmonic (i.e., the spur's amplitude is HD_2). Hence, by generating a harmonic at the spur's frequency and applying the relevant amplitude given, the SFDR parameter is modelled (Figure. 5). Needless to say, that the relevant building blocks must be chosen according to the SFDR's given unit, to convert it back to pure amplitude usable. The output of this stage can be written as follows:

$$x_{out,SFDR}(t) = x_{out,THD}(t) + SFDRx_{in}(t)|_{f=if_{sig}}. \tag{29}$$

3.3 Offset Error

The offset value, δ, of the ADC which is entered by the user can be added after the THD subsystem. This will affect the output according to the offset error's definition. The output after this stage can be presented as follows:

$$x_{out,Off}(t) = x_{out,SFDR}(t) + \delta. \tag{30}$$

It is again notable that the relevant units across the different building blocks must be unified first, and dissimilar units to be converted to the appropriate ones before performing the simulation task.

3.4 Gain Error

The output of the last part discussed in Section 3.3 could be multiplied by the gain error (converted from %FS (%full-scale) to pure amplitude), η_G, and be added up with the previous signal to model the gain error effect. The output after this stage is as follows:

$$x_{out,G}(t) = x_{out,Off}(t) + \eta_G x_{out,Off}(t). \tag{31}$$

3.5 Aperture Delay

Aperture delay, which is the delay time between the rising edge of the input sampling clock and the actual time at which the sampling starts [20], can be modelled by placing a delay line right after the sampling clock input of the sample-and-hold (S/H) block. This will delay the actual starting point of the sampling task while the simulation time has already been started. Fig. 6 describes this method.

It is worth noting that, to model the aperture delay non-ideality of ADC we ought to employ the S/H block rather than using the simple zero-order-hold (ZOH).

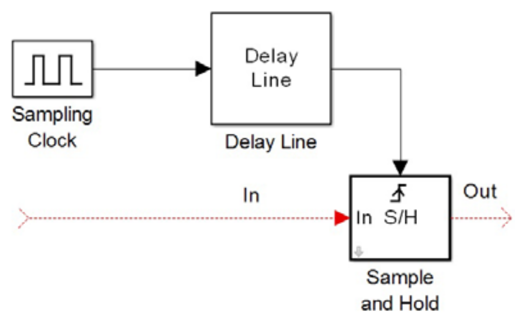


Figure 6: General aperture delay model. Maximum allowable aperture jitter for a given frequency f_i and finite resolution N is $t_j = 1/\pi f_{in} \times 2^{N+1}$.

3.6 Differential and Integral Non-Linearity

To model the DNL and INL, we initiate using a completely uniform random source to be added to the output node of the gain error, described in Section 3.4. Using this will assure having the maximum randomness for the aim of moving the ideal transfer function and changing the transition code length for the aim of modeling INL and DNL, respectively. The range of this uniformly random source equals to DNL, and the mean is equal to the INL value. We should also define these parameters as variables to be entered by the user. We do this using the mask property in Simulink. Therefore, the general random source to be added is:

$$Uniform\left\{\left(\frac{INL-DNL}{(2^n)-1} \times V_{FS}\right) : \left(\frac{INL+DNL}{(2^n)-1} \times V_{FS}\right)\right\}. \tag{32}$$

Another method can also be presented if we know the internal structure of the A/D. A third order power series is used to model the INL and DNL. Although this method's nature is far from the intuitive understandable ideas for modeling the INL and DNL like the previous method, it usually provides good accuracy, given the fact that it is similar to the conventional circuit non-ideality modeling. The elementary idea of this method was first proposed in a project at the University of California Berkeley under supervision of Professor Bernhard Boser [21]. Figure. 7 shows this method.

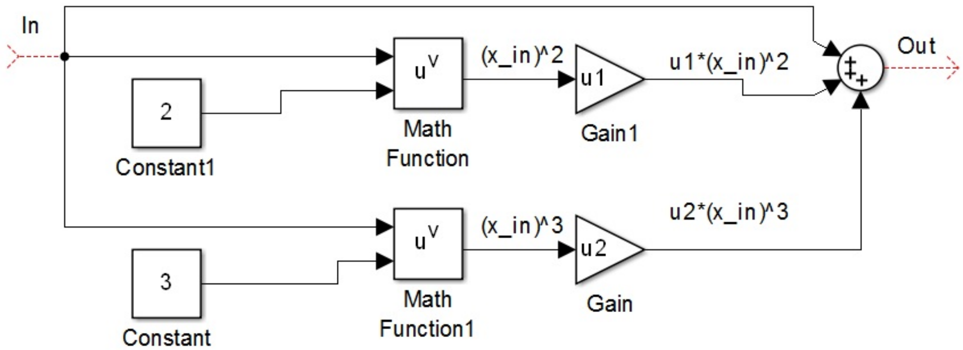


Figure 7: Second method for modeling DNL&INL.

$$\Gamma_{out} = \Gamma_{in} + \mu_1 \Gamma_{in}^2 + \mu_2 \Gamma_{in}^3. \quad (33)$$

It should be noted that we are restricted to choose either of the two positive or negative values for both the DNL and INL, and cannot afford producing a method which can entirely model the range from minimum negative to the maximum positive, as we are looking at the ADC from general perspective and do not intend to involve circuit or infrastructure level properties. The output after this stage will hence be:

$$x_{out,(D\&I)NL}(t) = x_{out,G}(t) + \rho|_{Uniform\{\alpha:\beta\}} \quad (34)$$

if we use the first method, and:

$$x_{out,(D\&I)NL}(t) = x_{out,G}(t) + \Gamma_{out}|_{\Gamma_{in}=x_{out,G}(t)} \quad (35)$$

if the second method is employed. ρ has the uniform distribution described in (32).

3.7 Other Non-Idealities Modeling

Recall the equations (19), (20), (21), and (9), and the discussions in Sections (2.2.3) and (2.2.1). Paying in-depth attention to those, we can propose an intuitive working way to achieve modeling other non-idealities of the ADC, helping to skip and overcome the theoretical difficulties. We explained that the difference between the theoretical (unreachable) signal to noise ration (known as SNR in ADC realm, or more accurately SNR_{ideal}) is caused by other types of noises except the unavoidable quantization noise. Those include thermal noise, distortions and many other things. Although some of these are partly predictable knowing the circuit level information of the specific type of ADC, nevertheless, it is hidden to us from systematic point of view. So far, we have modeled some non-idealities including the total distortion caused by the ADC's non-linear nature. We can now assume that any other non-ideality causing the SNDR to be lower than the theoretical SNR, (21), can be modeled by adding a white Gaussian noise having the power coming from an external port. As the mean of the Gaussian noise is zero, the power evaluation block will determine the variance. To do this, we need to go through the following steps:

1) We calculate the ideal SNR using (9). To do this, we only need the ADC's total number of bits, given by the user.
2) Sometimes in practical and more commonly in industrial ADCs, there is very small and negligible difference between the SINAD value and the 'practical' SNR, which expectedly should be the same. To take this effect into account as well, we add a uniform random source δ ranging between zero and the $\max(SNR_{ptactical} - SINAD)$ to the SINAD building block. This presumably owns a very small value and will have negligible effect on simulation results. The SINAD ($\sim$ ENOB) and the practical SNR comes from the ADC's datasheet and is entered by the user. Likewise before, by defining variable parameters at each building block of the simulation model using the mask property of the Simulink, we can use them in other blocks. Figure. 8 demonstrates the first two steps.

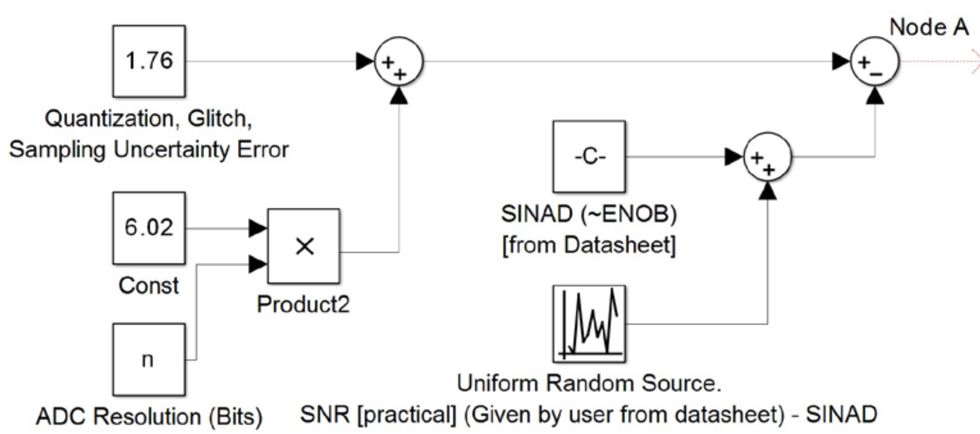


Figure 8: First steps of modeling other non-idealities of ADC

$$NodeA = (1.76 + 6.02N) - (SINAD + \vartheta)$$

$$\vartheta|_{Uniform(0:SNR_{practical} - SINAD)} \quad (36)$$

3) At this step, we need to compute two power quantities: the total input power (i.e., $P(x_{in}) = P_{in}$) and the power after the INL and DNL modeling block (i.e., $P(x_{out,(D\&I)NL}(t))$, point P2 in Figure. 9). The reason is, from the total difference between the SNR_{ideal} and the SINAD values, we have already modeled some non-idealities till point P2. Therefore, before adding a Gaussian noise to the model to compensate the rest, we should calculate the effect of the previous modelled parts on the noise level (i.e., how much noise has virtually been added from the beginning till point P2), subtract it from the $SNR_{ideal} - SINAD$, and then add the remaining unconsidered noise to the model to modify and construct the correct difference between the

SNR_{ideal} and SINAD in our simulation. To do these, we place a power evaluation block at the beginning of the model to determine the input power i.e., P_{in}. It is notable that we cannot use the mathematical formulations to calculate the total input power, as we have considered the input to be unknown in general. To use the evaluated data, the output is connected to a "Data Store Write" block accompanied by a "Data Store Memory". Similarly, the total power at point P2 can be evaluated. The difference between the total input signal power and the power at point P2 would therefore be the power of added noise, i.e. n'. Hence:

$$P_{n'} = P_{in,signal} - P2 \qquad (W), \tag{37}$$

$P_{n'}$ must be converted to dB.

4) The output of the above (which is in dB) must be converted back to amplitude to be usable by data store and memory blocks, and then be entered as the input variance of the white Gaussian noise needed. Figure. 9 presents the above procedure.

After doing the above procedures, we can now place an additive white Gaussian noise (AWGN) to the model which its variance comes from external port (i.e., the Data Store Read block) (Figure. 8).

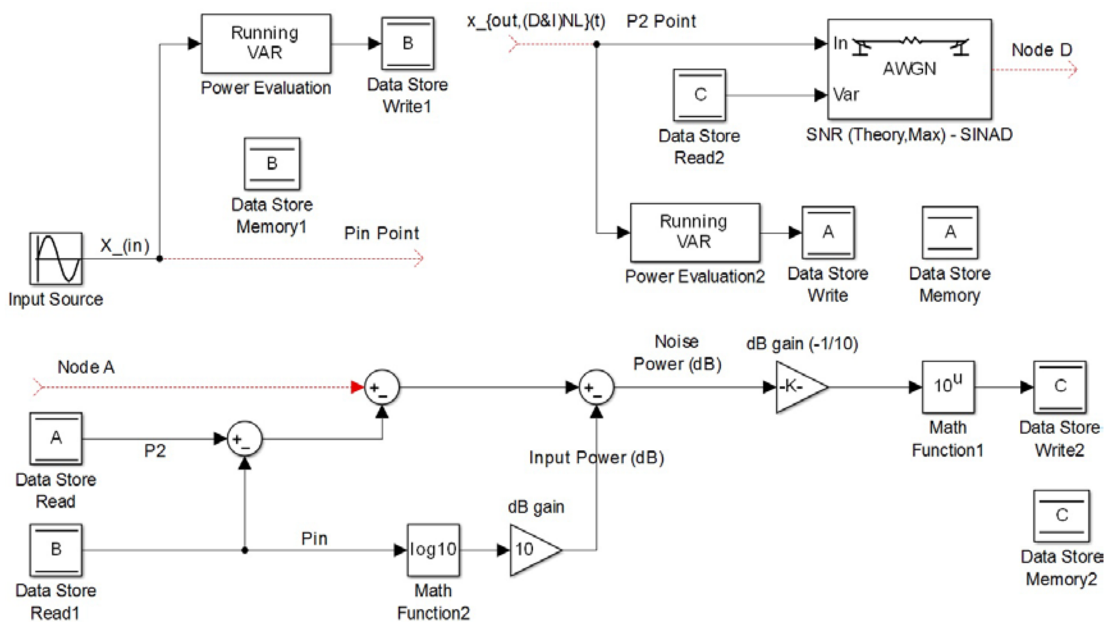


Figure 8: First steps of modeling other non-idealities of ADC.

4 Simulation Results

In this section, by placing all the aforementioned building blocks in a serial fashion, we have completed the whole simulation model having the non-ideal ADC in hand. By creating a Simulink subsystem we can make a handy block to be used each time we aim to simulate a new ADC having different non-ideality parameters. For this purpose, we define the input values of the subsystem (TNDC Block) to be the same as the values needed for each of the previously considered blocks. In other words, by doing this, the non-ideal ADC parameters will be entered by the user each time. To do this, we can define the non-ideality parameters' values at each individual block discussed previously as a double integer. In this case, at each individual block, we will enter the respective defined parameters names instead of a constant integer. This values are specified by the ADC's manufacturer and is findable in their respective datasheet. In addition, some points should be taken into consideration while putting the individual blocks in serial to each other. The output of each building block which is the input to the next block must have same characteristics to be mathematically acceptable. For instance, the output of one block could not be in LSB units while entering a block which works with pure amplitudes. The following formula is useful for conversion:

$$n\ LSB = x\ \%FSR$$

$$n \times \frac{V_{FSR}}{2^N - 1} \times 100 = x\%\ V_{FSR}$$

$$n_{LSB} \times \frac{100}{2^N - 1} = \%x_{FSR}. \tag{38}$$

Table 1: The overall initiated model for simulating a typical non-ideal analog to digital converter in MATLAB Simulink

NON-IDEALITY PARAMETERS (MOSTLY FOR TEXAS INSTRUMENTS ADS5400)	VALUE
Input amplitude ($\boldsymbol{a}$)	1
Input frequency (rad/sec) ($\boldsymbol{f_i}$ or $\boldsymbol{f_{sig}}$)	$2\pi 125e6$
Input phase (rad) ($\boldsymbol{\varphi_i}$)	0
ADC Resolution (Bits) ($\boldsymbol{N}$)	12
SFDR (dBFS or dBc)	66
SINAD (~ENOB) (dB)	57.5
SNR (Practical, from datasheet)(dB)	57.6
Offset Error (V)	$2.5e-3$
Gain Error (%FS)	0.05
Differential Non-Linearity (DNL) (LSB)	0.7
Integral Non-Linearity (INL) (LSB)	2
ADC Sampling Time ($\sim f_s$)	$1/1200e6$
i-th Harmonic Distortion (HD_i)(dBc)	$HD_2 = SFDR = 66$ $HD_3 = 70$
Min input voltage at logical low output (V)	-1
Max input voltage at 2^N output (V)	$+1$

By entering the output of the above serial to S/H and quantizer blocks (or simply ZOH and quantizer if aperture delay or clock duty cycle is not aimed to be modeled) we can have the simulation task started. In this paper, we will have the following parameters to be entered by the client before the simulation process starts: Input amplitude, Input frequency (rad/sec), Input phase (rad), Input (sine) sample time (if applicable; i.e., sinusoids sampled with ultra high frequency), ADC resolution (Bits), SFDR (dBFS or dBc), SINAD (~ENOB)(dB), SNR (dB), Offset (V), Gain Error (%FS), DNL (LSB), INL (LSB), ADC sampling time (plus sampling clock duty cycle, if applicable), HD2, HD3, HD4, HD5 (in case that a particular HD_i has considerable amplitude, the user can simply change the mask variable in THD modeling section to have the frequency equal to if_{sig} and enter the relevant amplitude to model that specific harmonic too), Min input voltage at logical low output (V), and Max input voltage at 2^n output (V). Practical data listed in Table. 1 is used for the simulation (i.e., values will be entered by the user based on his own data) tasks.

A strong advantage of the proposed system is, that the employed data in Table. 1 can be easily changed by the user according to different practical ADCs, owing to the structure of the model. The overall employed model which is used for simulation task is depicted in Figure. 9.

Needless to mention that, the basic parameters of the ADC apart from the non-ideality parameters can be modelled using internal parameters of individual building blocks being employed in this model. For instance, the ADC number of bits or minimum and maximum voltage levels can be entered in the idealized quantizer block. In case of using ordinary quantizer block in Simulink, we should enter the quantization interval instead. The simulation is performed for a 12-bit non-ideal ADC having practical parameters' values listed in Table 1, and for 1024 data samples. The following figures demonstrate the simulation results and output waveforms. The output spectrums can be obtained by applying appropriate fast fourier transforms.

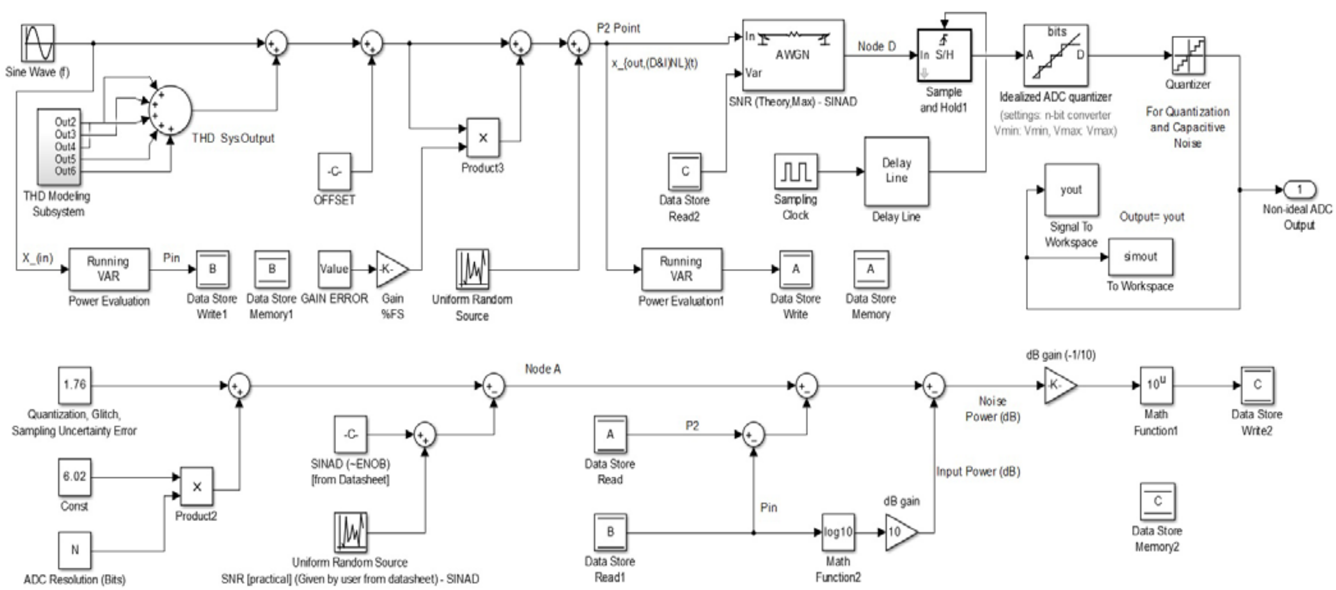


Figure 9: The overall initiated model for simulating a typical non-ideal analog to digital converter in MATLAB Simulink.

Since the distorting harmonics have very low amplitude comparing to the amplitude of the main frequency, their effect on the signal cannot be easily seen in time domain. However, Figure. 11 shows that the proposed model in Section 3.1 perfectly carries out the THD modeling. The simulation results can be exported to the MATLAB Workspace to be used for mathematical manipulations, as well as being accessible for MATLAB functions and/or M-Files. You can see the ideal ADC's results being compared to the non-ideal ADC modelled in this paper.

At this stage, it is important to remember that although the proposed TNDC simulation Block was a fundamental step forward towards creating typical ADC models and paving the way for handy and inexpensive simulation tasks, it cannot be compared to manufacturers' experimental results, due to the following facts and obstacles, which some of them could to be addressed by future works:

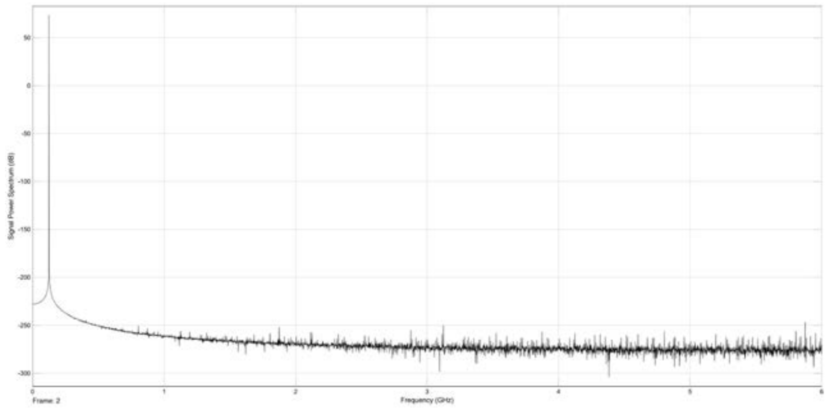


Figure 10: Power spectral after adding offset and gain errors.

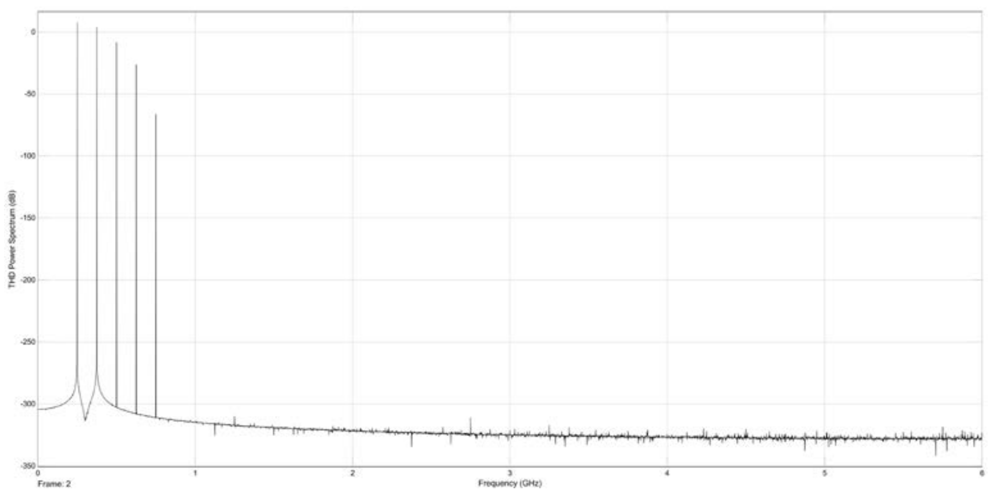


Figure 11: THD power spectral.

1. Most manufacturers use their own set of tools for extracting characteristics and experimental plots. Those are deeply linked with the infrastructure of each particular ADC.

2. The results in practical ADCs' datasheets are obtained by practical test experiments as mentioned above, and not performing simulation tasks. Consequently, you can find different results based on different local temperatures and so forth. Evidently, there are many affective issues (including physical circumstances) that cannot be taken into account in simulation tasks, especially when we do not emphasize on infrastructure and circuitry features of each particular converter.

3. Practical testing diagrams of ADCs always contain many other linear and non-linear elements which does affect the obtained experimental results, however, cannot be modelled here.

4. There are other non-ideality parameters apart from those modelled in this paper, which although are not much important, nevertheless, they do affect the simulation results somehow.

5. Above all the aforementioned points, there are some other non-ideality parameters that can be neither correctly measured nor being involved in simulation procedure, e.g. packaging non-idealities. Once again, we should emphasize that having a handy TNDC Block which can acceptably simulate typical ADCs according to their fundamental non-ideality parameters, can be notably helpful while simulating complicated projects. Employing the model and running the simulation repeatedly does not increase any overhead cost while offering a valuable vision about what's going on.

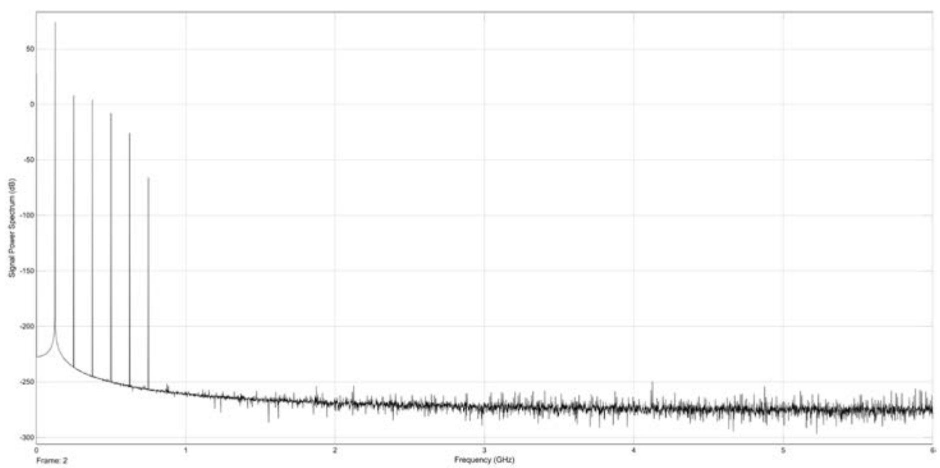
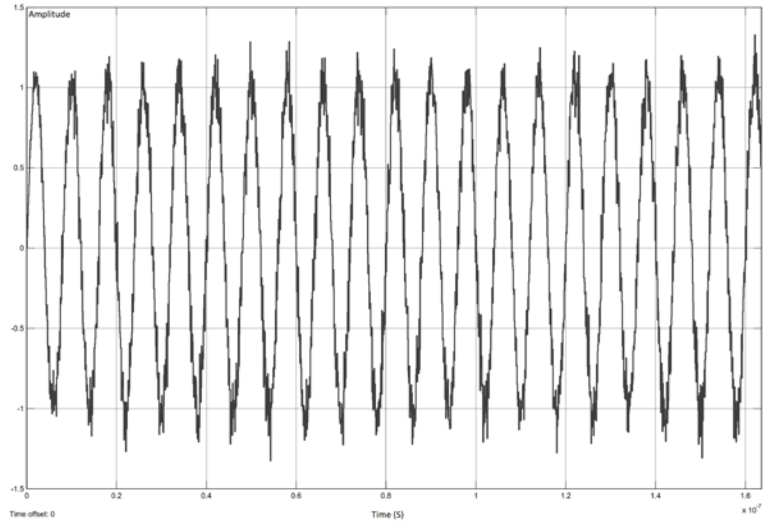
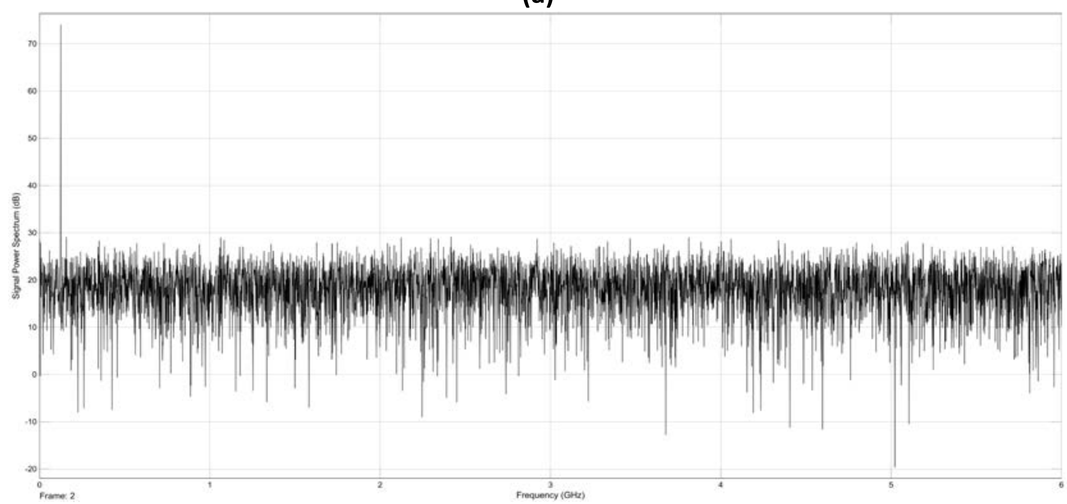


Figure 12: Power spectral after modeling offset and gain errors with THD.



(a)



(b)

Figure 13: Signal after AWGN block: (a) time-domain; (b) Power spectral.

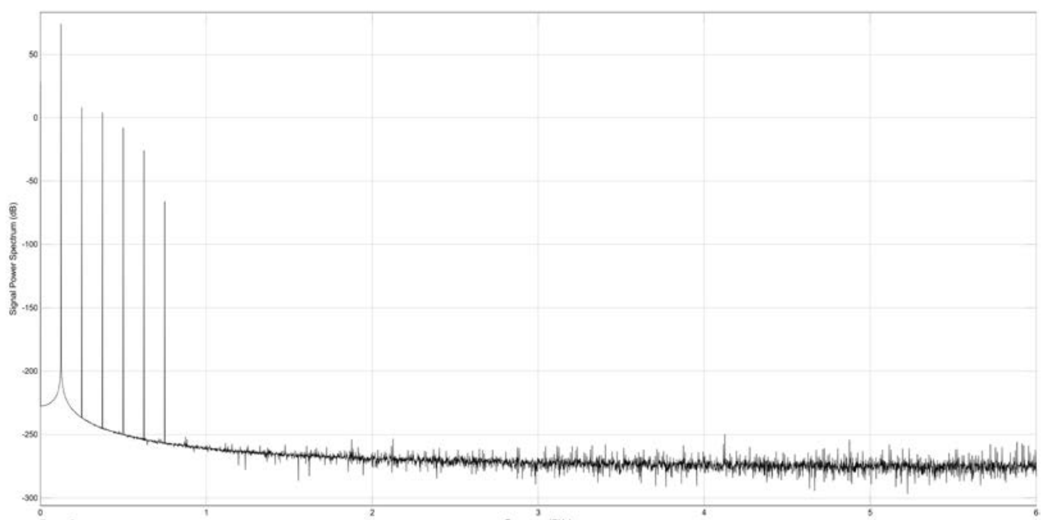
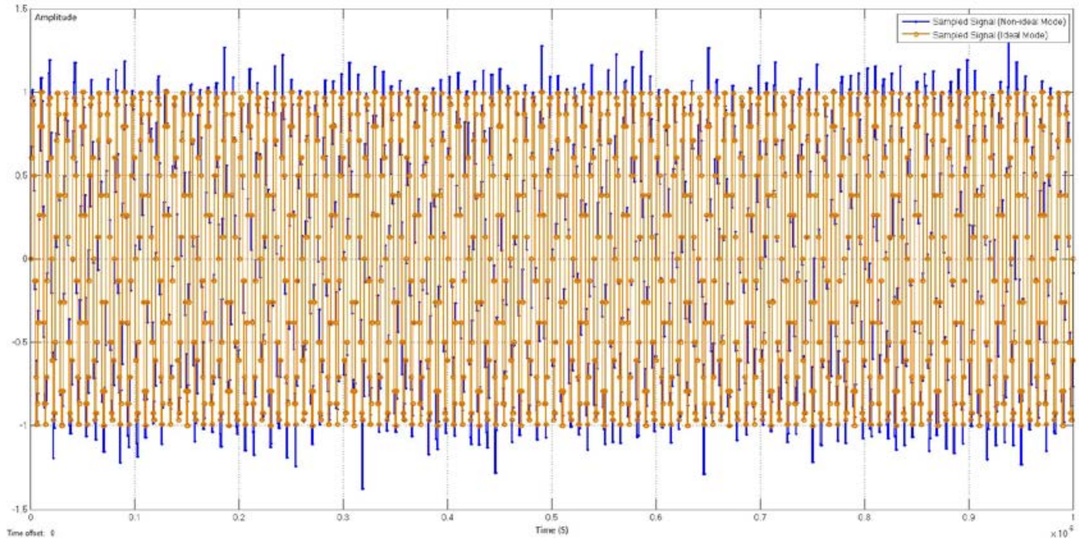
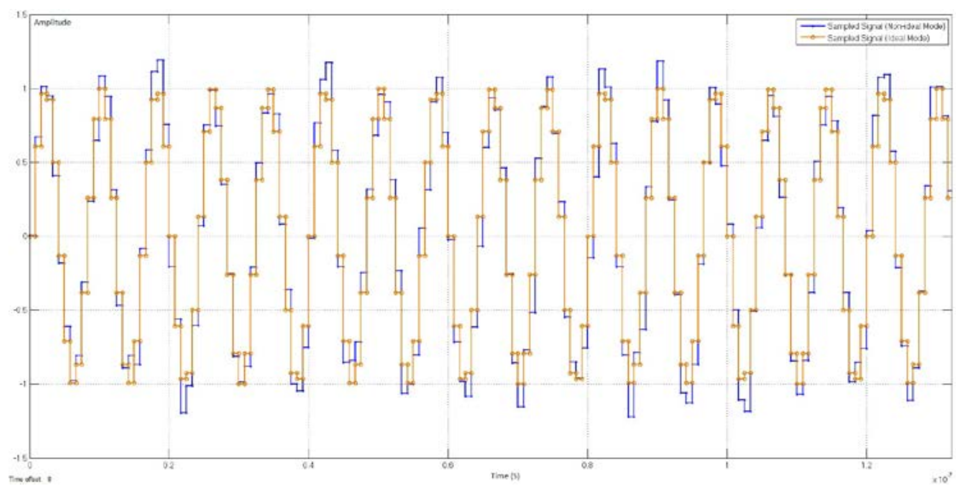


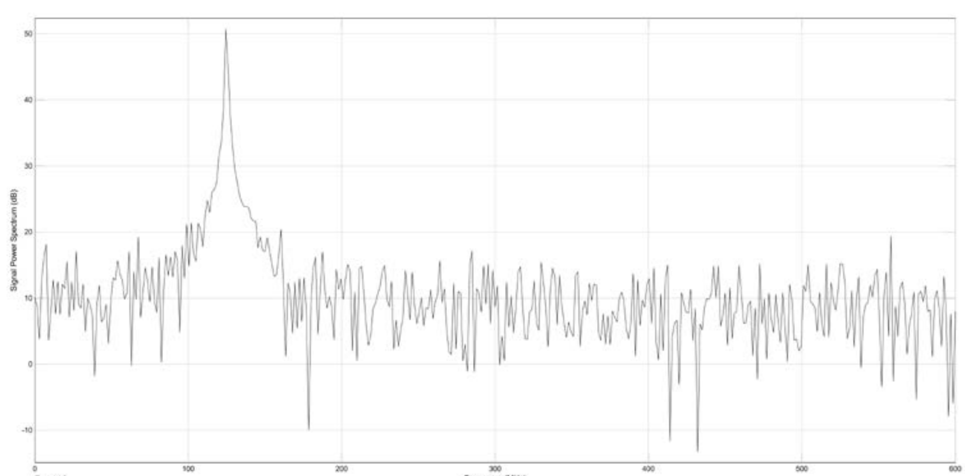
Figure 14: Power spectral after modeling offset and gain errors with THD.



(a)



(b)

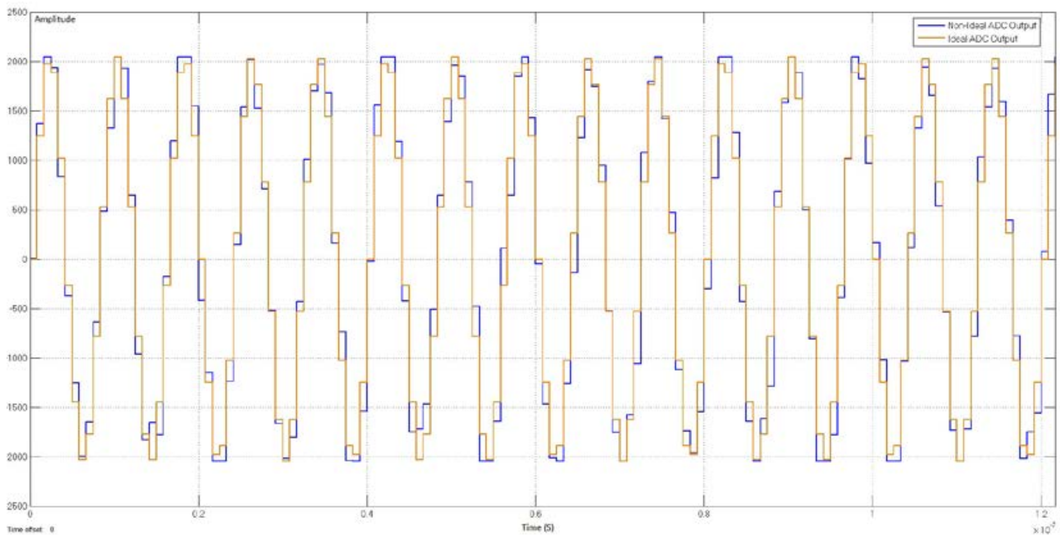


(c)

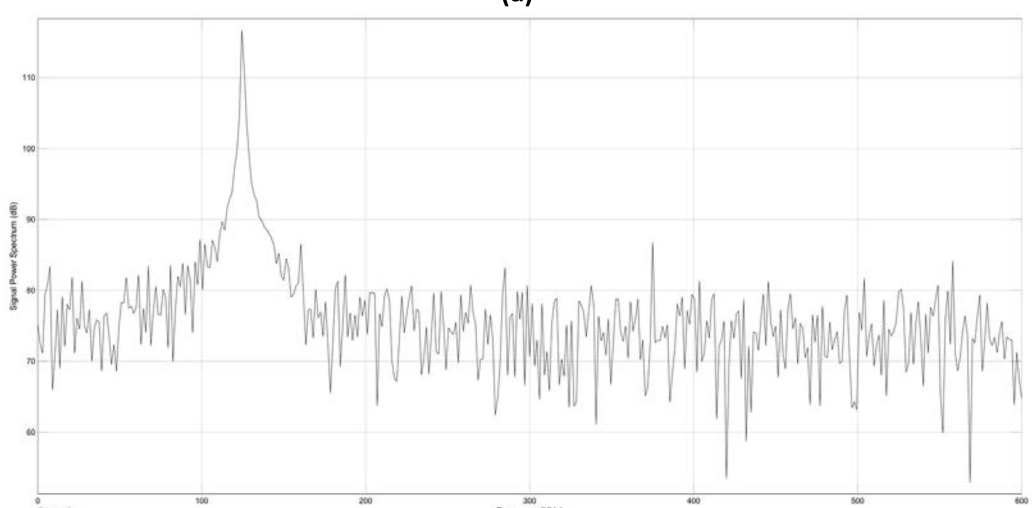
Figure 15: Sampled Signal: (a),(b) time-domain; (c) Power spectral.

5 Limitations and Future Research

As mentioned earlier, although this paper presented some first steps towards simulating practical analog to digital converters, it is essential to be aware of the need to considerable more in-depth works to achieve more realistic assumptions and implementations. Considering a non-ideal ADC to be a typical black box has always been a challenge in simulation and implementation of real-world projects, and was the main challenge of this paper too. However, this obstacle has to be somehow overcome in order to pave the way for more handy and inexpensive simulations of bigger projects. In addition to the points mentioned at the end of Section 4 to be addressed, considering an ADC as a typical converting block having some non-ideality parameters as its inputs has some drawbacks inevitably, as a result of not contemplating the actual infrastructure of each ADC itself. Therefore, some more advanced research should be carried out to consider the relations of non-ideality parameters with the infrastructural features of ADCs.



(a)



(b)

Figure 16: Signal after quantization (digital signal): (a) time-domain; (b) Power spectral.

Moreover, a unified version of these parameters and models based on different types of ADCs can notably direct to simulate typical ADCs more realistically. The above suggestions could be pointed as some directions, needless to say, there are a lot more to do than what described.

6 Conclusion

In this paper, we initiated a complete MATLAB Simulink model for a typical ADC without restricting it to a specific type. Considering an ADC as a black box while being competent for simulating non-ideality parameters was a notable challenge. By presenting an initiative model for most important non-ideality parameters of an ADC, in general, we proceeded an appreciable step towards more in-depth simulations of typical ADCs in future. This work can potentially lead to quick, handy and inexpensive rough simulation of ADCs in bigger projects, where splitting the whole project's diagram into parts can play an effective role in simplifying the overall simulation tasks, and considerably decreasing complicatedness. By placing individual building blocks in serial fashion, and defining the non-ideality parameters as parametric variables using Simulink mask feature, we created the whole Simulink model, namely, typical non-ideal digital converter (TNDC) Block, which takes the non-ideality parameters' values from the client before the simulation process starts. By showing the output of each of important TNDC blocks' nodes, we demonstrated the performance of the proposed method. At the end, we employed the TNDC Block to simulate a typical ADC using practical values for its non-ideality parameters. The results prove the validity of the proposed models according to the non-ideality parameters' definitions.

ACKNOWLEDGEMENT

This work was supported by the Research Committee of Barthawa Institute of Technology. The authors would also like to thank respected reviewers for their comments.

REFERENCES

[1]. IEEE-SA_Standards_Board, “IEEE standard for terminology and test methods for analog-to-digital converters,” 2000.

[2]. H. Zare-Hoseini, I. Kale, and O. Shoaei, “Modeling of switched-capacitor delta-sigma modulators in Simulink,” *IEEE Transactions on Instrumentation and Measurements*, vol. 54, no. 4, pp. 1646–1654, 2005.

[3]. P. Malcovati, S. Brigati, F. Francesconi, F. Maloberti, P. Cusinato, and A. Baschirotto, “Behavioral modeling of switched-capacitor sigma-delta modulators,” *IEEE Transactions on Circuits & Systems I Fundam. Theory Appl.*, vol. 50, no. 3, pp. 352–364, 2003.

[4]. G. Li, Y. M. Tousi, A. Hassibi, and E. Afshari, “Delay-line-based analog-to-digital converters,” *IEEE Transactions on Circuits & Systems II Express Briefs*, vol. 56, pp. 464–468, 2009.

[5]. H. S. Lee and C. G. Sodini, “Analog-to-digital converters: Digitizing the analog world,” *Proceeding of IEEE*, vol. 96, pp. 323–334, 2008.

[6]. S. Brigati, F. Francesconi, P. Malcovati, D. Tonietto, A. Baschirotto, and F. Maloberti, “Modeling sigma-delta modulator non-idealities in Simulink,” *1999 IEEE Int. Symp. Circuits Syst. VLSI*, vol. 2, 1999.

[7]. A. Fornasari, P. Malcovati, and F. Maloberti, “Improved modeling of sigma-delta modulator non-idealities in Simulink,” in *Proceedings - IEEE International Symposium on Circuits and Systems*, 2005, pp. 5982–5985.

[8]. S. Jaykar, P. Palsodkar, and P. Dakhole, “Modeling of sigma-delta modulator non-idealities with two step quantization in MATLAB/Simulink,” in *Proceedings - 2011 International Conference on Computational Intelligence and Communication Systems, CICN 2011*, 2011, pp. 532–536.

[9]. V. Kledrowetz and J. Haze, “Analysis of non-ideal effects of pipelined ADC by using MATLAB - Simulink,” *Adv. Sensors, Signals Mater. Anal.*, pp. 85–88.

[10]. D. P. Jayker, Shashant, Palsodkar Prachi, “Modeling of sigma-delta modulator non-idealities in MATLAB/Simulink,” in *Proceedings - 2011 International Conference on Communication Systems and Network Technologies, CSNT 2011*, 2011, pp. 525–530.

[11]. M. Koe and J. Z. J. Zhang, “Understanding the effect of circuit non-idealities on sigma-delta modulator,” *Proc. 2002 IEEE Int. Work. Behav. Model. Simulation, 2002. BMAS 2002.*, 2002.

[12]. J. M. Lei, X. W. Dai, X. C. Zou, and Z. G. Zou, “Modeling non-idealities of sigma delta ADC in Simulink,” in *2008 International Conference on Communications, Circuits and Systems Proceedings, ICCCAS 2008*, 2008, pp. 1040–1043.

[13]. G. Qingbo, J. Xinzhang, and T. Hualian, “Co-simulation of pipeline ADC using simulink and PSpice,” in *Proceedings - 4th International Conference on Intelligent Computation Technology and Automation, ICICTA 2011*, 2011, vol. 2, pp. 487–490.

[14]. C. Vogel and H. Koeppl, “Behavioral modeling of time-interleaved ADCs using MATLAB,” in *Proceedings of the AUSTROCHIP 2003*, 2003, no. October, pp. 45–48.

[15]. M. Webb and H. U. A. Tang, “System-level simulation for continuous-time delta-sigma modulator in MATLAB Simulink,” in *Proceedings of the 5th WSEAS Int. Conf. on Circuits, Systems, Electronics, Control & Signal Processing, Dallas, USA.*, 2006, pp. 236–241.

[16]. T. I. Inc, “Understanding data converters, SLAA013,” *Mix. Prod. Appl. Reports*, pp. 1–22, 1999.

[17]. B.-B. Texas Instruments (Inc., “Principles of data acquisition and conversion,” *BURR-BROWN Appl. Bull.*, no. 602, pp. 1–6, 1994.

[18]. M. J. M. Pelgrom, *Analog-to-digital conversion*. Springer New York Press, 2012.

[19]. I. Glover and P. Grant, *Digital Communications*. McGrew Hill Publications, 2009.

[20]. T. I. Inc, “ADS 5400, 12-Bit , 1-GSPS analog-to-digital converter,” *Texas Instruments Inc.*, no. March, 2010.

[21].] E. Ng and M. Bohsali, “Multifrequency cell impedence measurement,” University of California, Berkeley, USA., 2010.