E-CLONALG: A classifier based on Clonal Selection Algorithm

[1]Sanghamitra Dash, [2]Rabindra Kishore Mishra, [3]Arijit Panigrahy and [4]Rama Krushna Das
[1,2]Department of Electronic Sciences, Berhampur University, Berhampur, Odisha, India;
[3] Indian Institute of Technology, Kharagpur, India
[4]National Informatics Centre, Berhampur, Odisha, India;
smdash71@gmail.com; rkm.es@ buodisha.edu.in; arijit.pani@gmail.com; ramdash@yahoo.com

ABSTRACT

This paper proposes an improved version of CLONALG, Clone Selection Algorithm based on Artificial Immune System(AIS), that matches with the conventional classifiers in terms of accuracy tested on the same data sets. Clonal Selection Algorithm is an artificial immune system model. Instead of randomly selecting antibodies, it is proposed to take k memory pools consisting of all the learning cases. Also, an array averaged over the pools is created and is considered for cloning. Instead of using the best clone and calculating the similarity measure and comparing with the original cell, here, k best clones were selected, the average similarity measure was evaluated and noise was filtered. This process enhances the accuracy from 76.9 percentage to 94.2 percentage, ahead of the conventional classification methods.

Keywords: Accuracy; Antibody; Artificial Immune System; classification; classifier; CLONALG; CSCA; cloning; intrusion detection; machine learning

1 Introduction

Artificial immune system (AIS), one of the rapidly emerging areas of artificial intelligence, has been inspired from Biological immune system [1]. The main function of the Biological immune system is to shield our body from antigens. Our immune system has an efficient pattern recognition capability that differentiates between antigen and the body cells. Immune systems have many features such as uniqueness, autonomous, recognition of foreigners, distributed detection, and the noise tolerance [1]. AIS is related with abstracting the structure and function of biological immune system in order to solve various problems. Clonal Selection Algorithm is an AIS model [2]. It has been used mostly for pattern matching and optimization. It has also been tested for classification purposes but generally the results have not been satisfactory. The widely used models in the field of AIS are immune networks, clonal selection, negative selection and danger model [3].

After a brief explanation of AIS in the following sections, this paper develops an enhanced CLONALG (E-CLONALG) algorithm and implements it for standard classification. We are proposing a modified version of the Clone Selection Algorithm (CLONALG) for improving accuracy on classification. The enhanced CLONAL-G algorithm increases the accuracy to a great extent and can be used for various other purposes with a little modification such as pattern recognition, intrusion detection, filtering the noise etc.

2 Artificial Immune System

2.1   Immune Network Theory

This was proposed by Jerne (1974) to explain the learning properties of the Biological immune system. The logic behind this theory is that any lymphocyte receptor within an organism can be recognized by a subset of the total receptor repertoire. Immune networks are often referred to as an idiotypic network [4]. This theory also explains the immunological behaviors such as tolerance and memory emerge. The model was known as Artificial Immune Network (AIN) which was later known as Artificial Immune NEtwork (AINE).

2.2   Negative Selection

Negative selection algorithms were inspired by the mechanism in Thymus that produce a set of mature T-cells capable of binding only non-self-antigens. It was proposed by Forrest et al (1994). The starting point of this algorithm is to produce a set of self-strings, S, that defines the normal state of the system [5]. The task then is to generate a set of detectors, D, that only recognize the complement of S. These can be applied on a test data set to classify them as self or non-self [6] [7].

2.3   Danger Model

A new immunological model, suggested by Polly Matzinger (1994), that states which the immune system does not distinguish between self and non-self. Rather, it discriminates dangerous and safe by recognition [8].

2.4   Clonal Selection

The clonal selection theory was proposed by Burnet (1959). It is the process of antigen recognition, cell proliferation and differentiation into memory cell [9]. A clonal selection algorithm named CLONALG, was proposed by Castro and Zuben (2002), for learning and optimization. When applied to pattern matching, a set of patterns, S, to be matched are considered to be antigens [10]. The task of CLONALG is to then produce a set of memory antibodies, M, that match the members in S [1].

3 Clonal Selection Algorithm (CLONALG)

The Clonal Selection Algorithm, in inspired by the following elements of the clonal selection theory [11]: -

- o Maintenance of a specific memory set
- o Selection and cloning of most stimulated antibodies
- o Death of non-stimulated antibodies
- o Affinity maturation (mutation)
- o Re-selection of clones proportional to affinity with antigen
- o Generation maintenance of diversity

The algorithm is outlined below:

3.1   Initialization

The first step of the CLONALG technique is initialization, which involves preparing an antibody pool of fixed size N. This pool is then partitioned into two components, a memory antibody section m that eventually becomes representative of the algorithms solution and a remaining antibody pool r used for introducing additional diversity into the system.

3.2 Loop

The algorithm then proceeds by executing a number of iterations of exposing the system to all known antigens. A single round of exposure or iteration is referred to as a generation. The number of generations G the system executes is user configurable, though the system can use a problem specific stopping condition.

1. *Select Antigen:* A single antigen is selected at random without replacement (for the current generation) from the pool of antigens.
2. *Exposure*: The system is exposed to the selected antigen. Affinity values are calculated for all antibodies against the antigen. Affinity is a measure of similarity, and is problem dependent. It is common to use Hamming distance.
3. *Selection:* A set of n antibodies are selected from the entire antibody pool that has the highest affinity for the antigen.
4. *Cloning:* The set of selected antibodies are then cloned in proportion to their affinity (rank based).
5. *Affinity Maturation (mutation):* The clones (set of duplicate antigens) are then subjected to an affinity maturation process to better match the antigen in question. Here, the degree of maturation is inversely proportional to their parent's affinity (rank based), meaning that the greater the affinity, the lower the mutation.
6. *Clone Exposure:* The clone is then exposed to the antigen, and affinity measures are calculated.
7. *Candidature:* The antibody or antibodies with the highest affinity in the clone are then selected as candidate memory antibodies for placement into m. If the affinity of a candidate memory cell is higher than that of the highest stimulated antigen from the memory pool m, then it replaces said antigen. Group replacements occur in a similar, but batched manner.
8. *Replacement*: Finally, the d individuals in the remaining r antigen pool with the lowest affinity are replaced with new random antibodies.
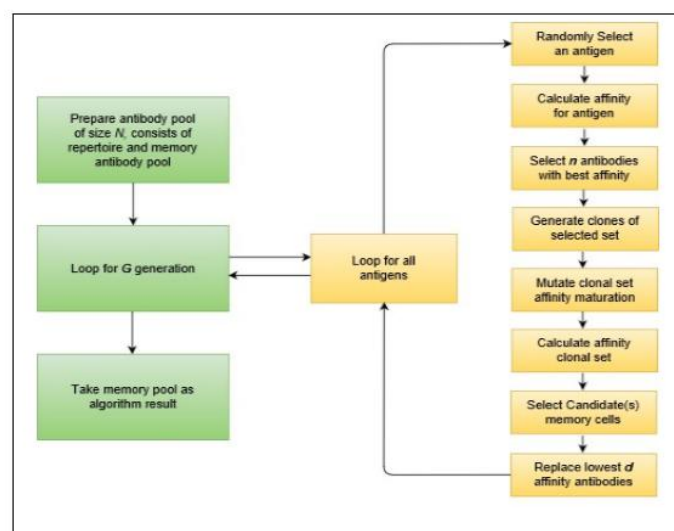

Figure 1. Overview of CLONALG

3.3   Finish

After the completion of the training regime, the memory 'm' component of the antigen pool is then taken as the solution of the algorithms. Depending on the problem domain, the solution may be a single best individual antigen or the collection of all antigens in the pool [12].

4   Enhanced Clonalg Algorithm (E-CLONALG)

The proposed algorithm is based on CLONALG algorithm. We have improved the algorithm in the initialization as well as training phase. We have introduced a new parameter in the antibody class known as 'ClassF'. In CLONALG, the memory pool was formed by picking up random antibodies. As a result, all the learning instances were not used. In E-CLONALG, we formed 'k' memory pools such that all the learning instances are accommodated in the memory pool [14]. Also this memory pool was visualized as a combination of 'c' memory pools where 'c' is the number of classifications. Thus, we considered the total learning set for classification instead of randomly picking a few instances [15]. We call the memory pool as 'M1'.

$$k = ceil\left(\frac{xSize\ of\ Learning\ set}{ySize\ of\ memory\ pool}\right) \tag{1}$$

In the learning phase, after the affinity of the memory cells were found with the antigens, we created another memory pool such that the affinity value of antibodies is the average of the memory pool created initially. We call the average memory pool as 'M2'. Then, both 'M1' and 'M2' was used for cloning. After that, the memory cells with maximum similarity are added to antibodies. Instead of one best clone, we selected 'k' clone cells in order to encourage local search. Also, we calculated the average similar measure between the cloned cells and the training set to filter out the noise and ensure that the generated cells are representative of the class that corresponds to them [15].

The algorithm has been outlined below: -

4.1   Initialization

The first step of the CLONALG technique is initialization, which involves preparing an antibody pool of fixed size N. This pool is then partitioned into two components, a memory antibody section 'm' that eventually becomes representative classes in the memory pool are denoted by 'Ab$_{mi}$' where 'i' is an integer.

4.2   Loop

The algorithm then proceeds by executing a number of iterations of exposing the system to all known antigens. A single round of exposure or iteration is referred to as a generation. The number of generations 'G' the system executes is user configurable, though the system can use a problem specific stopping condition.

1. *Select Antigen:* A single antigen is selected at random   without replacement (for the current generation) from the pool of antigens and its class is determined.
2. *Exposure:* The system is exposed to the selected antigen. Affinity values are calculated for all antibodies against the antigen. Affinity is a measure of similarity, and is problem dependent. We have used Hamming distance for our purpose.
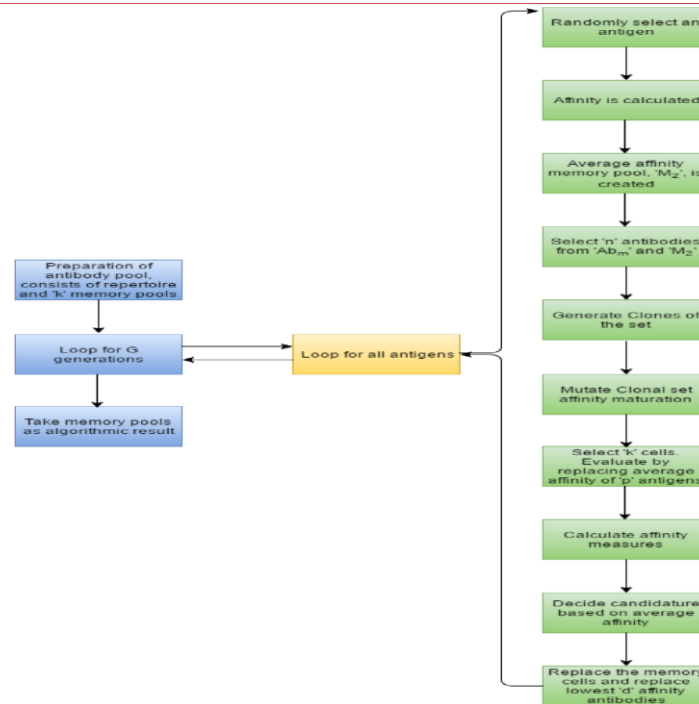
Figure 2. Overview of E-CLONALG

1. *Selection:* Another memory pool is created that consists the antibodies whose affinity values are the average of the 'k' memory cells. A set of n antibodies are selected from the entire antibody pool consisting of the 'Ab$_m$' as well as 'M$_2$' that have the highest affinity with the antigen.

2. *Cloning:* The set of selected antibodies are then cloned in proportion to their affinity (rank based).

3. *Affinity Maturation (mutation):* The clones (set of duplicate antigens) are then subjected to an affinity maturation process to better match the antigen in question. Here, the degree of maturation is inversely proportional to their parent's affinity (rank based), meaning that the greater the affinity, the lower the mutation.

4. *Clone Exposure*: The clone is then exposed to the antigen, and affinity measures are calculated.

5. *Candidature:* Instead of picking only one cloned cell as in the original CLONALG algorithm, 'k' highest affinity cloned cells of the same class are picked to promote local search. These 'k' cells are again evaluated by replacing their affinities with average affinity of 'p' nearest antigens belonging to the same classification. If the average affinity calculated above is lower than at least two antigens of different classifications then the cloned cell is no longer a candidate. If only one antigen of different classification has higher affinity than one of the k cloned cells, then this higher affinity antigen cell is removed from the training data set.

6. *Replacement:* Finally, if the highest affinity clone among the k filtered clones is higher then the minimum affinity value of the same class of the existing memory pool, then it replaces the minimum affinity memory cell i.e. 'Ab$_{mi}$'. Finally, replace 'd' lowest affinity antibodies from the repertoire of 'Ab$_r$'

4.3   Finish

After the completion of the training regime, the memory m component of the antigen pool is then taken as the solution.

5 Experiment

For our purpose, we used the 'Adult Data Set' available in UCI Library to test the classification accuracy of various algorithms and compare them with our algorithm. The dataset has 2 sets, one for learning purpose and other for testing, consisting of 48842 instances. The data set has 14 parameters and 2 classes [16]. The algorithm was tested using WEKA Tool [17][18] and Compared methods such as Bayes, J48, CLONALG etc. The findings are enlisted in the table below.

From the findings, in the Table 1 below, it can be easily seen that E-CLONALG has a better classification accuracy than other conventional algorithms. This variation of CLONALG uses a new technique in the initialization as well as training (candidature) phase with an aim of improving the classification accuracy.

Table 1. Various Classification Algorithm And Their Accuracy.

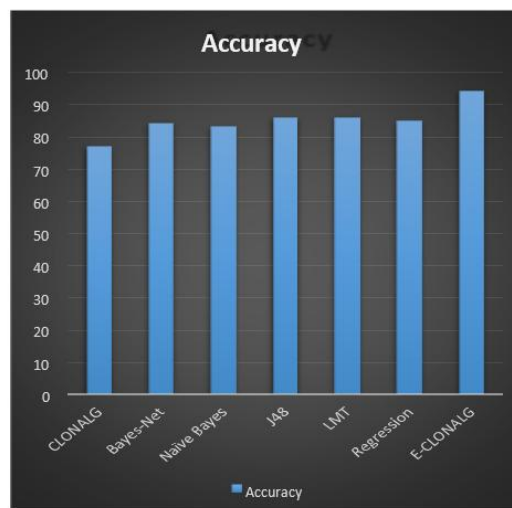| Name of the Algorithm	Accuracy (%)
CLONALG	76.9
Bayes-Net	84.2
Naïve Bayes	83.2
J48	85.8
Regression	84.7
LMT	85.9
E-CLONALG (proposed)	94.2

Figure. 3. Comparision of accuracy of Classification Algorithms

Figure 3 compares the accuracy of the various conventional classification algorithms with CLONALG and E-CLONALG algorithms. As we can observe, the accuracy of E-CLONALG is better than other conventional algorithms by 10%. The accuracy is calculated using

$$Accuracy = \left(\frac{Number\ of\ Correct\ Predictions}{Number\ of\ Instances\ in\ the\ Test\ Set} \right) \qquad (2)$$

Figure 4 below compares the kappa statistics of the various conventional classification algorithms with CLONALG and ECLONALG algorithms. Cohen's Kappa coefficient is a measure which indicates the interrater agreement for qualitative terms. The equation for calculating kappa statistic is [19]

$$k = \frac{p_0 - p_e}{1 - p_e} \qquad (3)$$

p_o = relative observed agreement among raters i.e. total number cases for which the prediction matches with the actual value.

p_e = hypothetical probability of chance agreement, i.e. Probability that the prediction will match with the actual value.
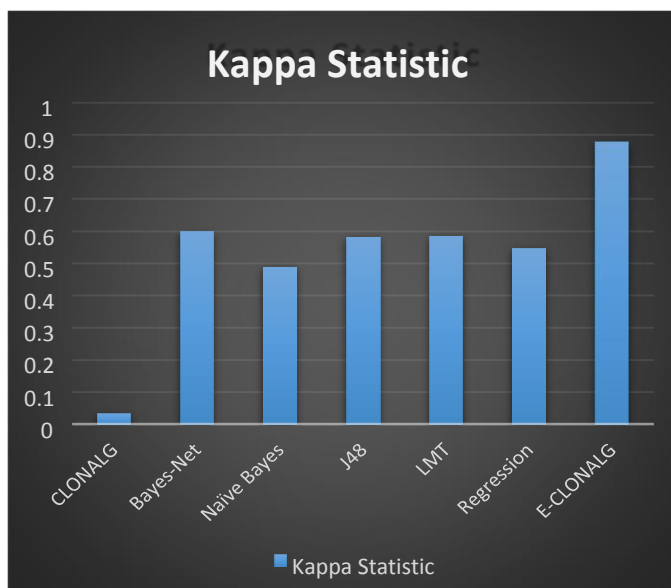

Figure. 4. Comparision of kappa Statistic of Classification Algorithms

The greater the value of kappa, the greater similarity between the actual value and the predicted value. The theoretical maximum of kappa is 1. In our case, the value of kappa for ECLONALG is 0.876 which is higher than all other algorithms.

6 Conclusion

A new form of CLONALG algorithm, E-CLONALG, has been proposed in this paper. The E-CLONALG algorithm seems to be an alternative method for classification. Even though it has not been tested across all the available data sets, but still the results indicate that it can be an alternative method for classification. It also shows the algorithms based on Clonal Selection Principle can also be used for classification.

A lot of further work is possible in this area. This algorithm can be modified to be used for intrusion detection, pattern recognition, noise removal etc. Moreover, the complexity of this algorithm can be reduced to make the classification task faster. Generalization of the algorithm can be done to deduce that E-CLONALG can be used as a classifier alongside conventional classifiers as well.

From the findings, it can be easily seen that E-CLONALG has better classification accuracy than other conventional algorithms. E-CLONALG uses a new technique in the initialization as well as training (candidature) phase with an aim of improving the classification accuracy.

REFERENCES

[1].   J.R. Al-Enezi, M.F. Abbod , S. Alsharhan : 'Artificial Immune Systems- Models, Algorithms And Applications', International Journal of Research and Reviews in Applied Sciences, 3 (2), May (2010) 118-131

[2].   R. Murugesan, K. Sivasakthi Balan : 'Positive Selection based Modified Clonal Selection Algorithm for Solving Job Shop Scheduling Problem', Applied Mathematical Sciences, Vol. 6,2012, no. 46, 2255-2271

[3].   Ramsha Rizwan, Farrukh Aslam Khan, Haider Abbas, Sajjad Hussain Chauhdary: 'Anomaly Detection in Wireless Sensor Networks Using Immune- Based Bioinspired Mechanism', International Journal of Distributed Sensor Networks Vol. (2015)

[4].   AISWeb – The Online Home of Artficial Immune Systems (http://www.artificial-immune-systems.org/algorithms.shtml)

[5].   K.Parthasarathy, 'Clonal selection method for immunity based intrustion, detection systems', Project Report (2014) , 1-19

[6].   Junyuan Shen, Jidong Wang, Hao Ai: 'An Improved Artificial Immune System- Based Network Intrusion Detection by Using Rough Set', Communication and Network, 2012,4,41-47.

[7].   Amira Sayed A. Aziz, mostafa A. Salama, Aboul ella Hassanien, sanna El-Ola Hanafi: 'Artificial Immune System Inspired Instrusion detection System Using Genetic Alorithm', Informatica 36 (2012) 347-357

[8].   Julie Greensmith, Uwe Aickelin, Steve Cayzer: 'Introducing Dendritic Cells as a Novel Immune- Inspired Algorithm for Anomaly Detection', International Conference on Artificial Immune Systems, ICARIS (2005), 14th -17th August 2005, Banff, Alberta, Canada.

[9].   Ezgi Deniz Ulker, Sadik Ulker.: 'Comparison Study for Clonal Selection Algorithm and Genetic Algorithm', Int. J. of Computer. Science & Information Technology Vol 4, No.4. August (2012) 107–118

[10].   Ilhan Aydin, Mehmet Karakose, Erhan Akin, 'Generation of classification rules using artificial immune system for fault diagnosism', IEEE Conference on Systems Man and Cybernetics (SMC), pp 343-349, (2010)

[11].   Vincenzo Cutello, Giuseppe Narzisi, Giuseppe Nicosia, Mario Pavone : 'A Comparative Case Study Using Effecctive Mutation Potentials, C. Jacob et al. (Eds.)': ICARIS (2005), LNCS 3627, pp. 13-28,200525: 1967-1978

[12].   Jason Brownlee, 'Clonal Seleciton theory and ClonalG:The Clonal Selection Classification Algorithm', Technical Report No, 2-02, January (2005)

[13]. Linquan Xie, Ying Wang, Liping Chen, Guangxue Yue: 'An Anomaly Detection Method Based on Fuzzy C-means Clustering Algorithm', Second International Symposium on Networking and Network Security (ISNNS,10) Jinggangshan, P.R. China, 2-4, 2010, pp.089-092

[14]. Ryma Daoudi, Khalifa Djemal, Abdelkader Benyettou: 'Cells Clonal Selection for Breast Cancer Classification', International MultiConference on Systems, Signals & Devices (SSD0, Hammamet, Tunisia, Mach 18-21,2013

[15]. Anurag Sharma, D. Sharma, 'Clonal Selection Algorithm for Classification', Lecture Notes in Computer Science, Vol 6825, pp 361-370 (2011)

[16]. UCI Machine Learning Repository –Adult Data Set (https://archive.ics.uci.edu/ml/datasets/Adult)

[17]. Weka: Data Mining Software in Java, (http://www.cs.waikato.ac.nz/~ml/weka/)

[18]. Github, (https://github.com/arijitiit)

[19]. Wikipedia Kappa stat (https://en.wikipedia.org/wiki/Cohen%27s_kappa)

[20]. Wikipedia RMSE stat (https://en.wikipedia.org/wiki/Root-meansquare_deviation)