SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 6
ISSN 2054-7390

Absorption Spectra Analysis using Modified Self-Organizing Feature Maps

[1]Thomas Bryant, [2]Jessica Koppen and [3]Mohamed Zohdy
[1]Department of Computer Science and Engineering, Oakland University, United States;
[2]Department of Chemistry, Oakland University, United States;
[3]Department of Electrical and Computer Engineering, Oakland University, United States
tcbryant@oakland.edu; jvkoppen@oakland.edu; zohdyma@oakland.edu

ABSTRACT

This research demonstrates an application of a modified self-organizing feature map (SOFM) algorithm to analyze and discover the quality of chemical absorption spectrum data. By forming an *NxN* neural array from input features and varying the essential parameters of the algorithm, map recognition quality is increased at the expense of more computation. The features of this SOFM are based on absorption intensity variations with excitation wavelength. SOFMs are used to discern pattern similarities and differences between spectral data. A context tree allows individual features, or key numbers in the data, to be input and classifies the vector to the type of data that is most similar. This research also use the self-organizing map to enhance as well as visualize resultant classification efficiency through the use of watershed transformation.

Keywords: Self-Organizing Feature Map, Unsupervised Learning, Energies, Intensities, Chemical Absorption

1 Introduction

UV/Vis spectroscopy is a routinely used tool to probe the electronic energy levels of atoms and molecules. Changing the levels by an electron, if symmetry allowed, is associated with emission or absorption of a photon of W/Vis light.

The interest in the interactions between gold nanoparticles and light dates back to Michael Faraday's observation of ruby-red colored colloidal solutions of gold. The property that gives the gold nanoparticles their ability to absorb visible light is known as surface plasmon, i.e., the collective oscillation of conduction electrons. As the size of the nanoparticle decreases and the electrons become confined, the metallic band structure becomes discretized. In other words, the discrete level structure emerges from continuous bands. This process can be explained by a quantum mechanical model of a particle in a box except for the "box" being spherical in shape. At this point a nanoparticle becomes a molecule with defined energy level structure. This species is referred to as a nanocluster (or sometimes a quantum dot).

Energy levels of gold nanoclusters have been probed by UV spectroscopy, which is uniquely suited to discern the electronic level structure [1]. UV spectroscopy shows that the plasmons disappear and the

DOI: 10.14738/tmlai.36.1155
Publication Date: 11th November, 2015
URL: http://dx.doi.org/10.14738/tmlai.36.1155

UV spectrum displays distinct, molecular-like transitions. The first study to demonstrate these distinct molecule-like transitions involved photodissociation spectroscopy of Aun+-Xe clusters by Collings et al. [2]. A recent study by Lecoultre et al. recorded the high resolution UV spectra of small Aun nanoclusters (n=1-5, 7-9) in low-temperature Ne matrices [3]. These spectra reveal a number of sharp electronic transitions which vary with the cluster size. The theoretical calculations of energy levels that give rise to these electronic transitions are necessary for a number of reasons including:

(i) assignments of excited states
(ii) determination of optical gaps
(iii) finding positions of the lowest excited state, etc.

Theoretical predictions of excited states rely on wave function methods among which the most trustworthy is Equation-of-Motion Coupled Cluster with single and double excitations (EOM-CCSD) [4]. This method is reliable but scales prohibitively with the number of electrons (as N^7 where N is the number of electrons) and therefore cannot be used for large molecules. For this reason methods based on linear response time-dependent density functional theory (TD-DFT) which scale as N^4-N^5 are gaining in popularity [4]. However, these methods show a strong dependence on the density functional used, that is often hard to rationalize, in that that seemingly "better" functionals may results in worse predictions of excited states. In gold clusters one is faced with additional obstacles in excited states calculations: large number of electrons and the fact that valence d electrons form large number of molecular orbitals with nearly identical energies (practically a d-band).

In assessing the performance of functionals used in TD-DFT one typically chooses a priori fixed number of transitions to excited states and then performs a statistical analysis on the basis of comparison with either more accurate method such as EOM-CCSD or experiments. Each excitation is characterized by the excitation energy (in eV) and oscillator strength f which is related to the intensity of the transition. A transition may be allowed or forbidden by symmetry. In addition, some symmetry-allowed transitions may show zero f.

In gold, the approach based on assignment and comparison is impossible to follow for all functionals because TD-DFT leads, at least for some functionals, to too many transitions with nonzero oscillator intensity. This effect may be compared to noise. It is not clear which element of density functional theory causes this failure. It is crucial at this point to identify the functionals that have this problem and the ones that do not.

In this paper, we use the UV spectrum of the cluster Au to assess the performance of 9 popularly used density functionals in the prediction of this spectrum. The corresponding EOM-CCSD results for transition energies and associated oscillator strengths, f, serve as the reference spectrum. TD-DFT calculations were performed for 200 excited states, each for the following list of functionals:

Table 1. The list of functional, each representing its own data set

B3PLYP
CAM-B3LYP
LC-LDA
LC-WPBE
LDA
PBE
TPSS
WB97X

These functional are listed in alphabetical order. They were obtained through computational chemistry study.

Each data set also includes excitation energies, which is on the x-axis and oscillator strengths (intensities) which is on the y-axis.

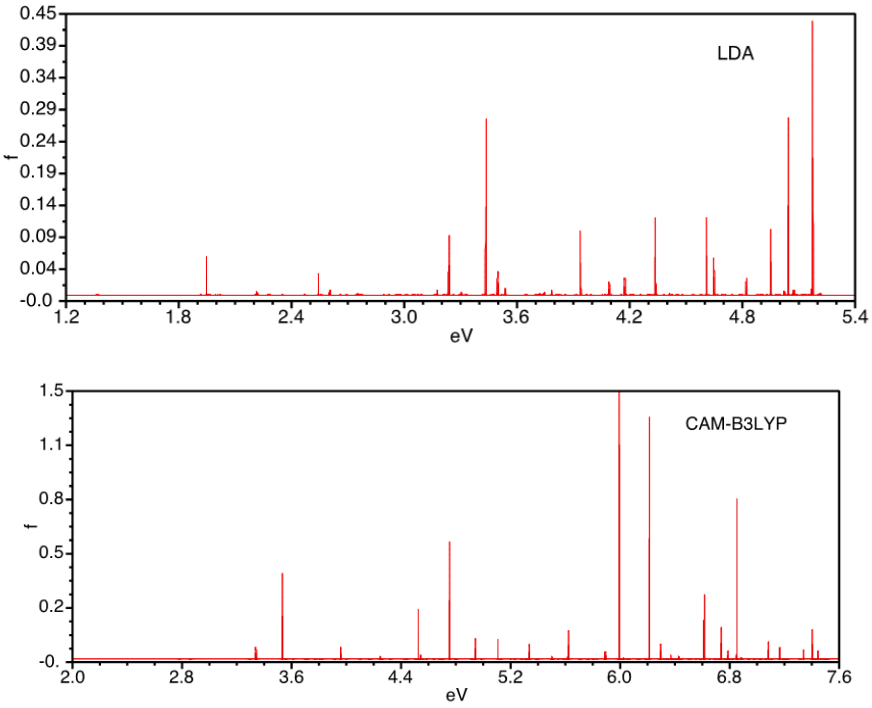

Figure 1. Calculated TD-DFT spectra of Au8 with two different functional (a) LDA; (b) CAM-B3LYP. Note different scales for transition energies and intensities.

The difficulties with assignment and statistical analysis spectra resulting from different functionals is illustrated in Fig. 1 (a) and (b) showing the LDA and CAM-B3LYP spectra, respectively. The spectra differ in three major ways:

(i) number of active transitions (i.e. with nonzero f),
(ii) intensities with CAM-B3LYP resulting with about three-fold higher intensities,
(iii) energy range of the first 200 transitions: in LDA this range is (1.37 eV – 5.2 eV) whereas is CAM-B3LYP the range is (2.79 eV - 7.52 eV). Clearly, the task of evaluating these spectra requires novel methods.

The method presented in this paper can qualitatively distinguish between bad and good spectrum. It is based on the following ideas. A typical result of the application of this method is presented in Fig. 2 below.

By extracting suitable features from the spectra, the unsupervised learning neural network, or modified self-organizing feature map (SOFM) is able to produce the discriminant clustering that occurs and also visualize it. Cluster analysis in terms of the SOFM is the assignment of observations into subsets, or clusters, so that observations within the same densely populated region are compact and separate. Signals can be mathematically represented in time domain, frequency domain, and better still joint time-frequency domain. The main focus of this study will deal with the frequency domain. The key

features extracted from the frequency domain are the resonances observed, which are obtained by analyzing the peaks and frequencies of the spectra.

In the last step of the feature extraction process, features are entered as "input". The preprocessed data is divided into quadrants and then used to obtain the salient features. Once the "input" is given, the top node of a context feature tree changes to the color of the spectrum it most closely resembles. This comparison is based solely on the information obtained from the inputs. The matrix of the magnitudes from the joint time-frequency diagram was relatively small here; the largest matrix was 200 rows by 2 columns; all of the data had similar sizes except for the training data which was 28 rows by 2 columns. As a result, the entire matrix could not be used efficiently when comparing against new data. Thus, an existing algorithm was modified to select features for the self-organizing feature map (SOFM).

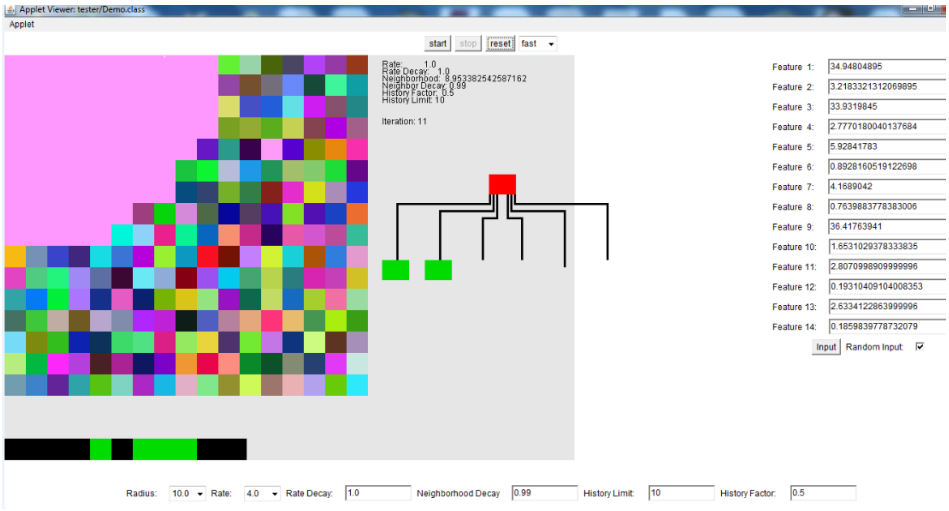

Figure 2: The self-organizing map being trained to recognize clusters

The image above shows the self-organizing map, the context tree, and the features that represent input. Preprocessed data is divided into quadrants and then used to obtain the salient features; the optimal amount is fourteen. Distinct features are obtained through the frequency domain, and then the program learns the patterns hidden in the data through clustering the SOFM [5]. To classify the spectra, the winning node is calculated using the Euclidean distance formula to determine which node is closest. The output is visualized and displayed on a two-dimensional map, and the vector is colored according to the spectra that has the closest node. It visualizes spectra as similar colors that act as competing agents. The "winners" are the vectors that weigh the least [6].

The map was initialized with random weights once the features were obtained. Nodes were created to form map anchors with these pseudo-random entities. The algorithm (mentioned later) reads the data and compares it to the random numbers. This determines which pseudo-random number is closest to using the earlier-mentioned Euclidean distance formula. This stage of updating the weights to ideally adapt the network's behavior is known as the training phase [6].

Before the SOFM algorithm begins to cluster the nodes, it accepts ten parameters: the neighborhood radius, learning rate, rate decay, neighborhood decay, history limit, and history factor. The default values are best depending on the type of data, but the user can specify any or all of the values to obtain

more effective clustering. The colors of the previous several iterations are displayed underneath the map, therefore increasing the accuracy of the clustering [5].

2 Feature Extraction Algorithm

To obtain features from frequency domain, spectrum data was examined. Then, all peaks and resonance frequencies are analyzed to determine key features. W_i, or the frequency, is obtained from the x-axis and M_i, or intensity (or peak), is obtained from the y-axis.

These features are input by the user through fourteen text boxes that are used to manipulate the data. Once the "input" is given, the top node of a context tree changes to the color of the most closely resembled data. This comparison is based on the information obtained from the input file; it does not affect the SOFM.

3 The Learning Algorithm

3.1 Training Stage

The primary goal of SOFM is to transform input patterns, of arbitrary dimensionality, into the responses of one or two-dimensional output arrays of neurons. These three steps are required [6]:

(i) an array of neurons that compute simple functions of the incoming inputs;
(ii) a mechanism for selecting the neuron (or neurons) with the largest produced output called winners;
(iii) an adaptive mechanism that updates the weights of the selected neuron and its neighbors.

This algorithm is used to train SOFM by Kohonen [5] and summarized as follows:

Initially, the weight vectors of the map's neurons are randomized. The SOFM takes an input vector and then traverses the map by calculating the Euclidean distance:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_1 - p_1)^2} \quad (1)$$

where p, q are two vectors being compared, other distances such as p-norm may also be used.

The similarities between the input vectors and neuron weight vector correspond to the distance between them. The algorithm then tracks the smallest distance (this node is called a winner, best matching unit, or BMU). The neighborhood function has the BMU as center node along with its neighborhood making them even more similar to presented input vector by using the equation below, as referenced from a previous paper [6]:

$$W_i(t+1) = W_i(t) + \alpha(t) + \beta(t)[k(t) - W_i(t)] \quad (2)$$

Repeat t, which is a positive constant and iteration and repeat from 2 while $t < \lambda$:

t denotes the current iteration

λ is the limit on iterations of time
W_v is the current weight vector of neuron v
dist(t) is the target input data vector
$\alpha(t)$ is learning rate due to time
$\beta(t)$ is the neighborhood function depending on distance from BMU

The above parameters determine the performance efficiency of the SOFM algorithm. The set of input vectors will each have three components, corresponding to a color space. The target output will correspond to a color space and settle on the RGB output map. Weight vectors convergence to the target in finite number of iterations. The current iteration must always be less than the limit on time iteration. A history limit is also embedded, which shows vectors that have been input previously in color format, and a history factor to update previous winning nodes has been implemented to decrease the time necessary to complete the training phase.

The neurons in the SOFM are initialized with small random weights. These initial weights determine the learning speed of the map. It will take more iterations depending on how large the learning rate is. However, the colors will not group together seamlessly on the visual display if the weights are too small. The next step included clustering and classifying the "ideal", or training data using the learning algorithm. This data trained a 17x17 neural network.

4 Simulation Results

4.1 Spectra Selection

The data was extracted into a feature vector of input length fourteen and shown on a context tree. After testing of the parameters, they were fine-tuned using sequential input, a radius of 10.0, a learning rate of 4.0, a rate decay of 1.0, a neighborhood decay of 0.99, a history limit of 10.0, and a history factor of 0.5. The neighborhood value has a two parameters of value: size and a distance function to influence.

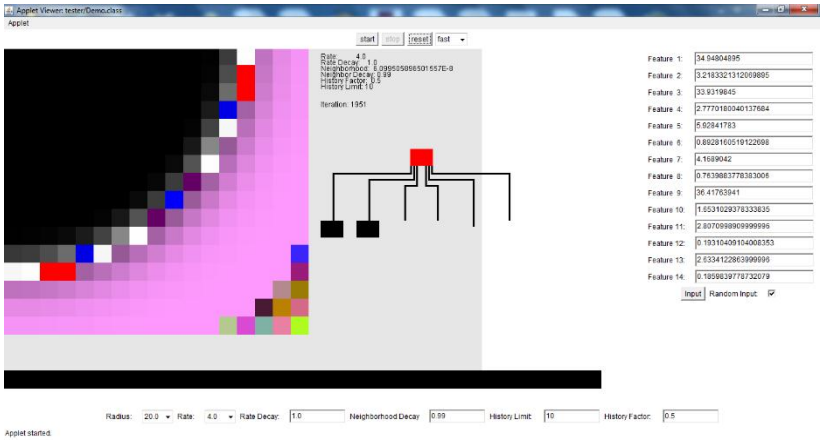

Figure. 3: Training data in comparison to the best absorption spectrum

The self-organizing map above shows training data, or ideal data, compared to the best spectrum, wB97X. The ideal data and spectrum data both have clearly defined clusters.

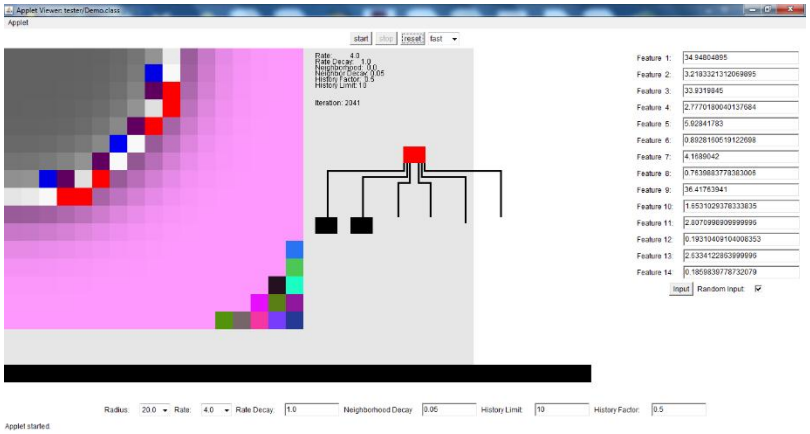

Figure. 4: Absorption spectra B2PLYP and wB97X used in correlation

The above figure shows the best pair of spectrum data, B2PLYP and wB97X. These two spectra have been entered into the self-organizing map in conjunction. The clear purple and silver clustering indicates that both clusters were clearly defined here as well.

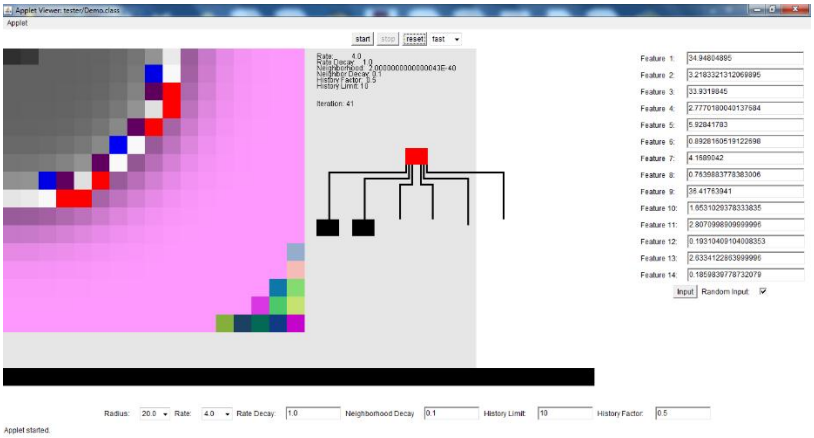

Figure 5. Absorption spectra B2PLYP and wB97X used to compare both energies and intensities

The self-organizing map in Figure 5 above shows the best pair of spectra grouped together with both energies and intensities. Both of the spectra clustered in a very high quality once again, with a few colors (blue, white, dark purple, red) serving as transition colors from purple to silver.

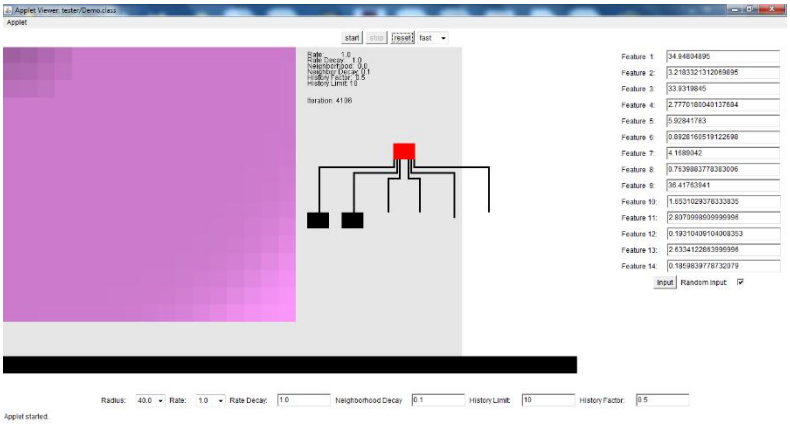

Figure. 6: SOFM correlation of spectra wB97X and LDA

The above image showed the best spectra (wB97X), grouped with the worst (LDA), using both energies and intensities. The large amount of purple indicates that the clusters conglomerated together or that one spectra completely "won" over the other.

4.2 Range of Absorption Spectra

Table 2. List of functionals listed from the ones containing the best to worst spectra

wB97X
B2PLYP
CAMB-3LYP
LCLDA
LCWPBE
B3PLYP
TPSS
PBEPBE
LDA

This table above lists the absorption spectra from the best spectrum to the worst spectrum. They were also determined to have the highest to lowest numbers when the fourteen features were averaged.

Of the spectra that were used in the relevant chemistry testing, wB97X was revealed to be the best, while TPSS was revealed to be the worst. This coincides with the results the self-organizing map found since wB97X had the lowest average of features by number, and while PBE had the highest, it was not used in the chemistry testing, and TPSS, which was used, was found to be the second worst (it had the second highest median of the eight spectra when they were reduced to fourteen features). In addition, it was found that the better they cluster, the better the spectra was. When comparing wB97X and the training data, they clustered almost seamlessly.

5 Conclusion

The SOFM (self-organizing feature map) has successfully classified UV chemical spectra inputs according to properly selected features and accurately clustered and visualized the data presented. It also provided a context for input features that served as training data with an associated context tree diagram and displayed the color that most resembled the spectra. After testing, the colors of the SOFM clustered best with sequential input and the following parameters: a neighborhood radius of 10.0, a learning rate of 4.0, a rate decay of 1.0, a neighborhood decay of 0.99, a history limit of 10.0, and a history factor of 0.5. SOFM performance was determined by the degree to which the colors clustered together and the number of squares of random color were left on the SOFM.

The history function that serves as one of the algorithm parameters improves the performance during the training phase substantially. Correlation usage increases the success rate of identifying unknown input by 15-20%. Blended distance extending Euclidean not only has a significant effect in improvement of computational time in our self-organizing feature maps, but it can also be used in other applications such as recognizing speech patterns, genetics, and music analysis. We finally added watershed transform as a good preprocessor to enhance the fidelity of visualizing SOFM concepts.

REFERENCES

[1] T. G. Schaaff, R. L. Whetten, "Giant gold-glutathione cluster compounds: Intense optical activity in metal-based transitions," J. Phys. Chem. B, 104, 2630-2641 (2000).

[2] A. Collings, K. Athanassenas, D. Lacombe, D. M. Rayner, and P. A. Hackett, "Optical absorption spectra of Au_7, Au_9, Au_{11}, and Au_{13}, and their cations: Gold clusters with 6, 7, 8, 9, 10, 11, 12, and 13 *s*-electrons," J. Chem. Phys. 101, 3506 (1994).

[3] S. Lecoultre, A. Rydlo, C. Felix, J. Buttet, S. Gilb, and W. Harbich, "UV-visible absorption of small gold clusters in neon: Au-n (n=1-5 and 7-9), " J. Chem. Phys. 134, 074302 (2011).

[4] A. I. Krylov, Annu. "Equation-of-motion coupled-cluster methods for open-shell and electronically excited species: The Hitchhiker's guide to Fock space ," Rev. Phys. Chem. 59, 433 (2008).

[5] T. Kohonen, *Self-Organizing Maps.* Springer, 2001, BerlinWhetten RL.

[6] J. Vesanto and Alhoniemi, E. "Clustering of the self-Organizing Map." IEEE Transactions on Neural Networks, vol. 11, No. 3, May 2000. Neural Networks Res. Centre, Helsinki Univ. of Technol., Esposo, Finland.

[7] Bradley, Matthew, Kay Jantharasorn, Keith Jones, and Dr. M. Zohdy. "Self Organized Neural Networks Applied to Animal Communications." REU Report, 2008.

[8] Su, Mu-Chun and Chung, Hsiao-Te. "Fast Self-Organizing Feature Map Algorithm." IEEE Transactions on Neural Networks. Vol. 11, no. 3, May 2000, pp. 721-733.

[9] T. Bryant, M. Hodges, and M. Zohdy. "Modified Self-Organizing Maps for Engine Health Diagnostics", International Journal of Computing and Information Technology. Vol. 3, no. 2, March 2014.

[10] T. Bryant and M. Zohdy, "Noise Signal identification by Modified Self-Organizing Maps", International Journal of Computing and Information Technology. vol. 2, no. 6, November 2013.