Immune Based Clustering for Medical Diagnostic Systems

Noha Abu-Zeid, Rasha Kashif and Osama Mohamed Badawy
Collage of Computing and Information Technology
Arab Academy for Science, Technology & Maritime Transport
Alexandria, Egypt
noha_abuzeid@yahoo.com, obadawy2@gmail.com, rfkashef@gmail.com

***Abstract*—It has recently been shown that Artificial Immune Systems (AIS) can be successfully implemented as biologically inspired systems. The Artificial Immune Network (AIN) is one of the computationally intelligent systems that are inspired by the processes of the immune system. Many algorithms are used to exploit the immune system's characteristics of learning and memory. The aiNet is one of such AIS algorithms with excellent performance on elementary clustering tasks. This paper proposes the use of two clustering techniques (DBSCAN and K-MEANS) in combination with aiNet algorithm for medical diagnosis. Two standard data sets are used to evaluate the clustering performance. The results indicate that both clustering techniques produced closely related outcomes under the usage of aiNet algorithm. However, K-means showed higher accuracy and better percentage of total clustering than DBSCAN for both data sets.**

Keywords-aiNET; DBSCAN; K-means; AIS; Artificial Immune Network; Clustering

I. INTRODUCTION

The main function of a biological immune system is to defend the body from foreign substances known as antigens. It has immense capability to recognize and distinguish between foreign cells entering the body (non-self or antigen) and the body cells (self) [1]. Inspired by biological immune systems, Artificial Immune Systems (AIS) have emerged during the last decade.

Thus, AIS could be defined as computational systems derived from theoretical immunology, observed immune functions, principles, and mechanisms in order to solve problems [2]. AIS have exceptional characteristics of pattern recognition, data analysis, and machine learning. It has been established that AIS performs well in various engineering applications and it is used as a main structure in hybrid systems [17]. The field of AIS has emerged as a hot field of research recently and attracted a growing number of computer scientists and they have proposed several different computer immune models [3-5]. AIS was incited by many researchers to design and build immune-based models for a variety of application domains. These models are still under intensive investigation. Negative selection, clonal selection algorithms, Danger theory, and artificial immune networks are four major and most popular algorithms of the Artificial Immune Systems. There are many successful applications based on these algorithms in computer security, anomaly and virus detection, optimization, machine learning, pattern recognition, and clustering [18-19]. Recently, many artificial immune models were used in combination with other AI techniques such as Fuzzy systems or biological inspired Neural Networks [6]. one of the widely used models of AIS is the Immune Network Theory, it is based on the fact that parts of the immune system are able to recognize other parts of the system [7]. Jerne introduced the immune network theory to model the relation between immune cells and molecules. This theory state that a network of B-cells occurs due to the ability of paratopes, located on B-cells, to match against idiotopes on other B-cells. The binding between idiotopes and paratopes has the effect of stimulating the B-cells. This is because the paratopes on B-cells react to the idiotopes on similar B-cells, as it would to an antigen [7]. There are several algorithms implementing this theory. This class of algorithms focus on the network graph structures involved where antibodies (or antibody producing cells) represent the nodes and the training algorithm involves growing or pruning edges between the nodes based on affinity (similarity in the problems representation space). Immune network algorithms have been used in clustering, data visualization, control, and optimization domains [7-8].

Disease diagnosing process is a complicated process that depends largely on the physician's knowledge, experience and ability to evaluate the patient's current test results and analyze risk factors that might be causation of illness. Therefore, a need for systems to assist physicians in making accurate and fast decisions has arisen [9].

The main focus of this study is to propose a system based on the Artificial Immune Network model in order to assist in the process of diseases diagnosis. There are many algorithms implemented in the Artificial Immune Network model. One of the widely used algorithms that incorporate clonal selection and hypermutation mechanisms is the aiNet (Artificial Immune Network) algorithm. aiNet is described as an AIS approach to data clustering, and the main algorithm used in this study. The paper is organized as follows. Section II gives some basics about the proposed system and then an overview of each component. The experimental results are explained in section III. The results discussion in section IV and the conclusion and future directions are addressed in section V.

II. PROPOSED SYSTEM

The main framework is shown in Fig.1. The feature selection phase is utilized in order to minimize data set dimensionality. The aiNet algorithm has two stages: in the first stage it produces a group of memory cells that represents compressed data representation of the data sets.

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.42

And then in the second stage it automatically detects clusters from the compressed data using Minimum Spanning Tree (MST) [16]. The original aiNet algorithm used in this study was adopted from de Castro and Von Zuben published paper [4], and we will use both DBSCAN and K-MEANS for clustering the memory cells rather than (MST) used by de Castro and Von Zuben (2000).

A. Experimental Datasets

In this paper, two of the UCI Machine Learning Repository data sets [10] are used, Iris data set and Group5 of Wisconsin Breast Cancer database. The Iris dataset [11] consists of 150 samples and five features: Sepal Length (in cm), Sepal Width (in cm), Petal Length (in cm), Petal Width (in cm), and Class. The data is distributed over 3 classes (Iris Setosa, Iris Versicolour, and Iris Virginica), 50 samples in each class. The (WBC-G5) [12] dataset obtained from university of Wisconsin-Madison Hospital, the data was delivered to UCI repository in groups. Group5 (delivered in August 1990) of the data is used here. It consists of 48 samples and 11 features defined as Sample code number (id number), Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class. The data is distributed over 2 classes: 36 Benign, 12 Malignant.

B. Feature Selection

In order to reduce the dimension of data set, it was necessary to represent the data in a concise format/model. Usually, the data represented by L-dimension vector. The dimension L is the number of the total features of the data set. L tends to be large even with a small set of data. Instead of using all the features, the best features (l) can be selected for clustering, where l << L. This can lead to significant savings of computer resources and processing time. It is called feature selection because each column in the data set is considered as a feature. There are many feature selection techniques, which will be tested in this study and the technique that will provide best outcome will be chosen

C. aiNet algorithm

In general, aiNet is an edge-weighted graph, not necessarily fully connected, composed of a set of nodes called cells, and sets of node pairs called edges [4], [13-14]. The aiNet algorithm treats each data point as an antigen. The algorithm develops a population of antibodies based on the immune network theory, affinity maturation and clonal selection [4] and [13]. After that these antibodies outline a network, which can represent the antigens in a compressed way [4] and [13].

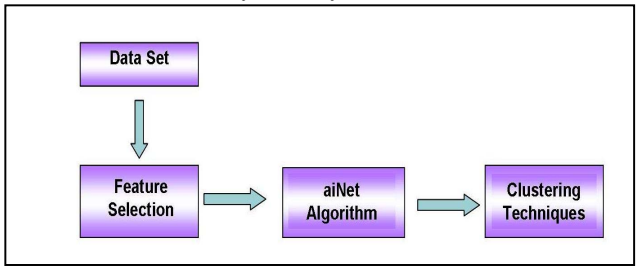

Figure 1. Frame work of proposed system.

The algorithm is iterated for each Antigen (Ag) where the affinity between Antibody (Ab) and Antigen (Ag) is calculated with each Ab in the repertoire. The highest affinity cells are cloned and mutated then exposed to Ag and the affinity is calculated. The highest affinity clones are placed in a separate pool M, and the clones with lowest affinity are eliminated form the pool (clonal suppression). Then after the entire antigens are exposed to the system the antibodies with distance less than the suppression threshold (σ) are eliminated from M (Network suppression) [4]. The aiNet Algorithm is presented in Fig.2.

Algorithm: aiNET (T, N, σ, ε):
Input: Antigen pool (T), Antibody pool size (N), suppression threshold (σ), affinity threshold (ε)
Output: Memory cells pool (M)
Initialization: Highest affinity antibodies pool (P)={}, Highest affinity clones pool (C)={}.
Begin
for (j<T) do
Select random antigen (Ag_j).
While (i<N) do
Calculate affinity (Ab_i , Ag_j).
If affinity (Ab_i , Ag_j) >
affinity (Ab_{i+1} , Ag_j) then
$Pi \leftarrow Abi$
End if
End of wihle
While (i<P) do
Clone (Ab_i).
Mutate (Ab_c).
Calculate affinity (Ab_c , Ag).
If affinity (Ab_{ci} , Ag_j) >
affinity (Ab_{ci+1} , Ag_j) then
$C_i \leftarrow Ab_{ci}$
End if
End of wihle
While (i<C) do
Calculate affinity (Ab_{ci} , Ag).
If affinity (Ab_{ci} , Ag_j)< ε then
Eliminate (Ab_{ci})
End if
Clonal Suppression:

While (i<C) do

Calculate affinity (Ab_{ci}, Ab_{ci+1})

If affinity (Ab_{ci}, Ab_{ci+1}) < σ

Eliminate Ab_{ci}

End if

$M \leftarrow C$

End of while

End for

Network suppression:

While (i<M) do

Calculate affinity (Ab_i, Ab_{i+1})

If affinity (Ab_i, Ab_{i+1}) < σ

Eliminate Ab_i

End if

End of while

End

Figure 2. The aiNet Algorithm.

D. Clustering techniques

The rationale of using the aiNet for data set clustering is that it is capable of reducing data redundancy and obtaining a compressed representation of data [13]. Therefore, more interrelated clusters are generated by the aiNet and thus resulting in better clustering of data set. The clustering techniques used in this paper are DBSCAN (density-based spatial clustering of applications with noise) and K-MEANS. K-MEANS is a partitional algorithm, and is one of the most commonly used clustering methods as it is quite easy to understand and implement. On the other hand, DBSCAN is a density-based clustering method and help in noise minimization [15].

III. Experimental Analysis

After applying the aiNet algorithm to both datasets, the IRIS and WBC-G5, the output of each is clustered once with DBSCAN and once again with K-MEANS using WEKA3.6.2 platform [17].Table 1 presents used parameters of each clustering technique. For the DBSCAN Epsilon (ε) is the distance threshold between the center of cluster and a given point to belong to that cluster. Minimum Points (Min.Pt) is the minimum number of points required to form a cluster. On the other hand the Seed parameter of K-Means algorithm is the initial random cluster centers. Cluster No. is the number of clusters to be generated, and the Iteration parameter is the maximum number of iterations allowed to reach the best cluster centroid.

In this paper, the results of each clustering techniques (DBSCAN and K-MEANS) were evaluated. Table 2 displays the results for DBSCAN and K-MEANS with the usage of the aiNet algorithm. Accuracy was used as a quality measure to evaluate the clustering performance. Accuracy is defined as the percentage of correctly clustered data. Table 2 shows the clustering results using DBSCAN and K-means. It also shows that the clustering results are closely similar using the two different clustering techniques. However, K-means showed higher accuracy and better percentage of total clustering than DBSCAN for both data sets.

IV. Discussions

This study focuses on presenting a merged approach to use clustering techniques (DBSCAN and K-MEANS) with aiNet algorithm on data sets obtained from UCI Machine Learning Repository database. The main goal of this study was to find the best clustering technique to be combined with aiNet algorithm. This consequently will aid in achieving better diagnostic system based on the Artificial Immune Network model. We found that there is a slight difference between the two clustering techniques used (DBSCAN and K-MEANS). As K-means showed higher accuracy and better percentage of total clustering than DBSCAN for both data sets. This might be debatable depending on the parameters chosen to execute each technique. On the other hand the use of the aiNet algorithm provided a compressed image of the original data sets and then automatically allowed the two clustering techniques to reach the desired clustering outcome.

Therefore, we believe that choosing different settings or parameters might contribute to this difference between the two clustering techniques. This approach was tested with the 150 samples from the Iris data set and the 48 samples from Wisconsin Breast Cancer (group 5) data sets. The results showed that the image obtained via the aiNet results in more compact clusters, and thus is capable of obtaining better clustering results. The results showed that integrated clustering techniques (DBSCAN and K-MEANS) have similar accuracy and thus similar results were reached. Finally both DBSCAN and K-MEANS appear to be good for large-sized data sets that contain data redundancy and noise.

TABLE I. Clustering Techniques Parameters

	DBSCAN		**K-MEANS**		
Dataset	ε	***Min.Pt.***	***Iteration***	***Cluster No.***	***Seed***
IRIS	0.9	3	500	3	10
WBC-G5	0.9	4	500	2	10

TABLE II. Clustering Results using DBSCAN and K-MEANS

Data set	**No. Clusters**	**Clustering Technique**	**% Clustering from total**	**% Accuracy**
Iris	3	K-means	100 %	100%
Iris	2	DBSCAN	100 %	75 %
WBC-G5	2	K-means	100%	100%
WBC-G	2	DBSCAN	76%	76%

V. CONCLUSION AND FUTURE DIRECTIONS

Both DBSCAN and K-MEANS showed promising clustering capabilities under aiNet algorithm. However, K-means seems to have higher accuracy and better percentage of total clustering than DBSCAN for both data sets.

Our aim for the future is to use vaccines. Vaccination is another inspired notion of the immune system. Vaccines will be extracted from dataset (antigens) and injected into the aiNet algorithm which will result in memory cells that represents the whole decision space. Thus the clustering results will improve. Feature selection and reduction will be used to reduce the data set dimensionality before exposing it to the aiNet algorithm in order to reduce the complexity of the algorithm.

ACKNOWLEDGMENT

Authors would like to thank Arab Academy for Science, Technology and Maritime Transport library for providing some needed references.

REFERENCES

[1] L. N. de Castro and F. J. Von Zube, " Artificial Immune Systems: Part I -Basic Theory and Applications, " in Technical Report – RT DCA 01/99,1999, pp. 95.

[2] L.N. de Castro, J. Timmis, "Artificial Immune Systems: A New Computational Intelligence Approach, " in Springer, 2002, pp. 57–58

[3] Dasgupta, D (ed), "An Overview of Artificial Immune Systems and Their Applications," in Artificial Immune Systems and Their Applications, Berlin: Springer-Verlag, 1998, pp.3-21.

[4] L.N. De Castro and F.J. Von Zuben, "An Evolutionary Immune Network for Data Clustering,"in Proc. of the IEEE SBRN, 2000, pp. 84-89.

[5] J.R. Al-Enezi, M.F. Abbod and S. Alsharhan, "Artificial Immune Systems- Models, Al-gorithms and Applications, " in IJRRAS, 2010, pp. 118-131.

[6] H. Izadinia, F. Sadeghi and M.M. Ebadzadeh, "A novel multi-epitopic immune network model hybridized with neural theory and fuzzy concept," in Neural Netw, 2009, pp. 633-641.

[7] N.K. Jerne, " Towards a network theory of the immune system," in Ann. Immunol. (Inst. Pasteur), 125C , 1974, pp. 373–389.

[8] J. Timmis, M. Neal and J.Hunt, "An artificial immune system for data analysis," in Biosystems, 2000, pp. 143-150.

[9] K. Polat, K. Tosun, S and S. Gunes, "Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted preprocessing, " in Pattern Recognition, vol 39, 2006, 2186 2193.

[10] UCI Machine Learning Repository Database: (http://archive.ics.uci.edu/ml/).

[11] Iris data set (http://archive.ics.uci.edu/ml/datasets/Iris)

[12] Group5 of Wisconsin Breast Cancer Database (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+ %28Diagnostic%29)

[13] L.N de Castro and F.J. Von Zuben, "aiNet: An Artificial Immune Network for Data Analysis", in Data Mining: A Heuristic Approach, H. A. Abbass, R. A. Sarker, and C. S. Newton (eds.), Idea Group Publishing, USA, Chapter XII, 2001, pp. 231-259.

[14] N. Tang and V. Rao Vemuri, "An Artificial Immune System Approach to Document Clustering," in Proceedings of 20th ACM Symposium on Applied Computing (SAC2005), Evolutionary Computing and Optimization (EC) Track, 2005, pp. 918 - 922.

[15] M. Ester, H-P. Kriegel, J¨org Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226-231

[16] W. Ahmad, A. Narayanan, "Principles and Methods of Artificial Immune System Vaccination of Learning Systems", ICARIS 2011, LNCS 6825, pp. 268–281, 2011.

[17] WEKA, University of Waikato (http://www.cs.waikato.ac.nz/ml/weka)

[18] H. Ge, X. Yan, "A Modified Artificial Immune Network for Feature Extracting", ICSI 2011, Part I, LNCS 6728, pp. 408–415, 2011.

[19] D. Dasgupta, S. Yu, F. Nino, "Recent Advances in Artificial Immune Systems: Models and Applications", Applied Soft Computing 11 (2011) 1574–1587