Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2014) 22: 1241 – 1254
© TÜBİTAK
doi:10.3906/elk-1210-62

Comparison of AIS and fuzzy c-means clustering methods on the classification of breast cancer and diabetes datasets

Seral ÖZŞEN, Rahime CEYLAN*
Department of Electrical and Electronics Engineering, Selçuk University, Konya, Turkey

Received: 15.10.2012 • **Accepted:** 28.01.2013 • **Published Online:** 15.08.2014 • **Printed:** 12.09.2014

Abstract: Data reduction is an indispensable part of pattern classification processes in many cases. If the number of samples is excessive, sample reduction or data reduction algorithms can be used for an effective processing time and reliable successive results. Many methods have been used for data reduction. Fuzzy c-means is one of these methods and it is widely used in such applications as clustering algorithms. In this study, we applied a different clustering algorithm, an artificial immune system (AIS), for the data reduction process. We realized the performance evaluation experiments on the standard Chainlink and Iris datasets, while the main application was conducted using the Wisconsin Breast Cancer and Pima Indian datasets, which were taken from the University of California, Irvine Machine Learning Repository. For these datasets, the performance of the AIS in the data reduction process was compared with the fuzzy c-means clustering algorithm, in which a multilayer perceptron artificial neural network was used as a classifier after the data reduction processes. The obtained results show that the maximum classification accuracies were obtained as 73.96% for the Pima Indian Diabetes dataset and 97.80% for the Wisconsin Breast Cancer dataset with the AIS and the compression rates were 80% and 40% for these results. For fuzzy c-means clustering, however, the aforementioned accuracies were obtained as 63% and 86.69% for the Pima Indian Diabetes and Wisconsin Breast Cancer datasets, respectively. Moreover, the compression rates for these results for fuzzy c-means were 90% and 70%. When the mean classification accuracy values over the experimented compression rates were taken into consideration, the AIS reached a mean classification accuracy of 70.07% for the Pima Indian Diabetes dataset, while 47.64% was obtained by fuzzy c-means for this dataset. For the Wisconsin Breast Cancer dataset, however, the mean classification accuracies of the AIS and fuzzy c-means methods were recorded as 94.90% and 75.43%, respectively.

Key words: Artificial immune systems, artificial neural networks, fuzzy c-means clustering, breast cancer dataset, diabetes dataset

1. Introduction

With today's improving technology, data recording opportunities are also expanding and providing lots of ways for information to flow. For large datasets, data mining techniques are affected in 3 ways: computing time, predictive or descriptive accuracy, and representation of the data mining model. Thus, some preliminary data preprocessing steps should be conducted before mining the data. Several approaches can be taken into consideration for data reduction, for example, random sampling of the current dataset. Clustering is another alternative to reduce the number of samples, by taking only a cluster-representative sample for all of the samples in a cluster. In general, clustering algorithms can be categorized in 2 popular approaches [1]: hierarchical clustering and iterative square-error partitional clustering. In the former group, data are displayed in a tree

*Correspondence: rpektatli@selcuk.edu.tr or dendrogram structure. While these kinds of techniques have the advantage of demonstrative easiness, the application of these methods can be cumbersome for large datasets. Hence, when the dataset is crowded, partitional clustering methods should be preferred. These schemes aim at finding high density regions in the data space using a square-error criterion.

Fuzzy c-means (FCM) is a well-known unsupervised partitional clustering algorithm. It was suggested by Dunn in 1973 and improved by Bezdek in 1981. According to this algorithm, 1 pattern in the given pattern set can belong to 2 or more clusters and each pattern belongs to clusters with different membership degrees in a [0 1] interval [2].

Artificial immune systems (AISs), which were formed in the 1990s, have recently been applied in lots of machine learning problems, such as classification, clustering, optimization, and anomaly detection. AISs take metaphors from the natural immune system as an inspiration source and combine them with state-of-the-art machine learning preliminaries. The immune system consists of several cells, tissues, and cytokines, functioning in different parts of the body. It is such a complex system that some mechanisms are still not uncovered. Thus, when looked at as a biological modelling source, it is a fertile inspiration basis but also hard to model. If some mechanisms in the immune system are seen in their most basic form, immune units recognize threats or intruders and act depending on the type of threat. In other words, the immune system conducts a clustering process before letting their units take action. From this point of view, researchers have modelled the immune system as a clustering algorithm in many ways. In his study, Timmis proposed a new data analysis technique, the artificial immune network, for clustering and classification problems [3]. Timmis and Neal developed an unsupervised algorithm for classification and clustering problems [4]. Their system consists of a set of B-cells, which are connected in a network structure and developed based on the clonal selection principle in the natural immune system. De Castro and Von Zuben used also clonal selection in controlling the dynamics of immune system units during the presentation of input patterns [5]. Vishwambhar et al. used innate immunity in their clustering method and found that the obtained results were comparable with existing models [6]. In their study, Graff and Engelbrecht compared their network-based AIS clustering model for clustering nonstationary data [7]. A population based AIS for clustering was proposed by Ahmad and Narayanan by inspiring humoral-mediated immunity [8]. Effective clustering results were obtained by the authors. By utilizing the AIS, Chiu and Lin developed an immunity-based ant clustering algorithm in their study [9]. Their scheme can automatically find the cluster centroids as well as the number of clusters. In another study [10], the authors developed a fuzzy AIS clustering algorithm and they found that this method has a better performance than k-means clustering algorithm. Similar applications were realized as AIS clustering applications [11–13]. They all tried to form an effective AIS clustering algorithm but few of them compared the AIS as a clustering method with other well-known methods. In this study, we intend to prove that for data reduction, the AIS can be a more effective clustering method over another well-known effective clustering method, FCM.

In this study, the application of an AIS system as a clustering method was conducted in the data reduction stage of a data mining process. The AIS was used to reduce the number of input data by clustering similar data in one group and taking the cluster representative instead of these data. After clustering and taking cluster representatives instead of the samples in that cluster, an artificial neural network (ANN) was used to classify the reduced data. The performance analyses of the system were conducted on the Chainlink dataset (which is an artificially generated dataset), Iris dataset, Wisconsin Breast Cancer dataset, and Pima Indian Diabetes dataset classification problems. For Wisconsin Breast Cancer and Pima Indian Diabetes datasets, FCM was also used as a clustering algorithm to compare the results with the AIS. Several classification accuracies were obtained with several data compression rates. The experimented compression rates were approximately 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, and 98%. The AIS and FCM methods were applied to reach these compression rates and optimum ANN parameters were found by experimentation for each rate. When the results for the 50% compression rate were evaluated, the AIS reached classification accuracies of 65% and 97.22% for the Pima Indian Diabetes and Wisconsin Breast Cancer datasets, respectively. When the results of the FCM for the Pima Indian Diabetes and Wisconsin Breast Cancer datasets were taken into consideration, the accuracies for the 50% compression rate were recorded as 44% and 81.11%. These results show that, like other state-of-the-art methods, the AIS can be also used as an effective data reduction scheme.

In Section 2, background information about the AIS, FCM, and applied system is given, in addition to the used datasets and experimental setup. In Section 3, application results are given in graphical and tabulated forms, and then, in Section 4, these results are discussed. Finally, the paper ends with the conclusion in Section 5.

2. Materials and methods

2.1. AISs and the AIS method used

An immune system can extract information from the infectious agents and make it available for future use in cases of reinfection by the same or a similar agent. The adaptive and memory mechanisms in a natural immune system have made researchers take notice of this field. Many studies including successive results have been conducted for a few decades [14].

Lymphocytes are the main cells in the immune response and they divide into 2 classes, as T and B lymphocytes (cells). B cells are more important than T cells in modelling. When activated by signals pointing to the existence of a foreign element in the body, B cells secrete their antibodies (Abs), which are located on their surfaces as bounded in normal conditions. It is these Abs that take critical roles in the elimination of foreign elements.

Generally, to develop a system inspired by a biological system, a representation scheme should be employed to model the biological units. Thus far in AISs, a shape-space representation scheme [15] has been used, mostly for representation of the system units, which have been generally chosen to be B cells or their secreted Abs in the immune system. This model says that each cell is represented as a point in an L-dimensional shape space and has a recognition region determined by its recognition radius. The set of features that characterize a molecule is called its generalized shape. Mathematically, the generalized shape of a molecule (m), either an antibody or an antigen (a foreign molecule that entered the body), can be represented by a set of coordinates $m =< m_1, m_2, ...m_L >$, which can be regarded as a point in an L-dimensional real-valued shape-space $(m \in S^L)$.

We used the supervised AIS for the clustering purpose, which uses a distance criterion while calculating the distance between the system units (Antibody-Ab) and input data (Antigen-Ag).

Euclidean distances of the Abs to the presented data (Ag) were calculated in the following way:

$$D = \sqrt{\sum_{k=1}^{L} (Ab_{j,k} - Ag_{i,k})^2}. \quad (1)$$

Here, $Ab_{i,k}$ and $Ag_{i,k}$ are the kth attribute of the jth Antibody-Ab_j and ith Antigen-Ag_i, respectively.

The training procedure of the learning algorithm is as follows:

(1) For each Ag_i do: (i: *1,...N*)

(1.1) Determine the class of Ag_i. Call memory *Abs* of that class and calculate the distances between Ag_i and this memory *Abs* with Eq. (1).

(1.2) If the minimum distance among the calculated distances above is less than a threshold value named as suppression value (*supp*), then return to step 1.

(1.3) Form a memory *Ab* for Ag_i:

At each iteration do:

(1.3.1) Make a random *Ab* population with *Ab* = *[Ab_mem; Ab_rand]* and calculate the distances of these *Ab*s to Ag_i.

(1.3.2) Select m nearest *Abs* to Ag_i; clon and mutate these *Abs* (*Ab_mutate*).

(1.3.3) Keep the m nearest *Ab*s in the *Ab_mutate* population to Ag_i as an *Ab_mem* temporary memory population.

(1.3.4) Define the nearest *Ab* to Ag_i as *Ab_cand*, candidate memory *Ab* for Ag_i, and stop the iterative process if the distance of *Ab_cand* to Ag_i is less than a threshold value named as a stopping criterion (*sc*).

(1.3.5) Concatenate *Ab_cand* as a new memory *Ab* to the memory matrix of the class of Ag_i.

(1.4) Stop training.

The mutation mechanism in the algorithm was chosen to be the *hypermutation* mechanism used in many AIS algorithms:

$$Ab_{j,k\prime} = Ab_{j.k} \pm D_{j,i} * (Ab_{j.k}) \tag{2}$$

Here, $Ab_{j.k\prime}$ is the new value and $Ab_{j.k}$ is the old value of the kth attribute of the jth *Ab*. $D_{j,i}$ stands for the distance between Ag_i and Ab_j.

After the above training procedure, a memory *Ab*s is formed to cluster the data in their recognition area. During the data reduction phase of the whole classification system, the input samples are presented to this memory *Ab*s and they are used as cluster representatives of the data in their recognition area. Thus, instead of using many data in one region for classification, their cluster representative will be given to the classification system.

2.2. FCM clustering

Partition-based clustering techniques divide patterns into clusters according to the number of input parameters. Partitional clustering algorithms are accomplished in determining the center-based cluster. In this paper, the FCM clustering algorithm was utilized for pattern reduction. The results are presented in the following sections.

The fuzzy clustering algorithm attempts to separate the data into fuzzy partitions that cover with one another. In other words, 2 or more clusters can include the same pattern in fuzzy clustering. For this reason, the membership degree of the data is determined for each cluster. These membership degrees are defined in the interval [0, 1]. In a formal manner, the application of FCM clustering to original data, $X = \{x_1, x_2, ..., x_M\} \subset R^l$, is the separation of c-cluster partitions to the vectors in X. In demonstration of X, M demonstrates the number of data vectors and l the dimension of each data vector. The *c-partition* of X generates sets of $(c \cdot M)\{u_{ik}\}$ membership values that can be easily organized as a $(c \times M)$ matrix $U = [u_{ik}]$. The basic idea of the fuzzy clustering algorithm is to find the optimum membership matrix U. Generally, the weighted within-groups sum of squared errors J_m is utilized as the objective function for the fuzzy clustering algorithm [16]:

$$\min\left\{J_m(U,V,X) = \sum_{k=1}^{M}\sum_{i=1}^{c}(u_{ik})^m \left\|x_k - v_i\right\|_A^2\right\}, \tag{3}$$

where

$$U \in M_{fcn} = \left\{U \in \Re^{cM} \left| \begin{array}{l} 0 \le u_{ik} \le 1 \;\; \forall \;\; ik\& \; \forall k, \;\; u_{ik} > 0 \;\; \exists \;\; i \\ 0 < \sum_{k=1}^{M} u_{ik} > \eta \;\; \forall \;\; i \;\; \& \quad \sum_{i=1}^{\mathrm{c}} u_{ik} = 1 \;\; \forall \;\; k \end{array}\right.\right\}.$$

In Eq. (3), the vector of (unknown) cluster centers is represented as $V = \{v_1, v_2, ..., v_c\}$, and $|x|_A = (x^T Ax)^{1/2}$ is an interior product form. However, in Eq. (3), A is an $l \times l$ positive absolute matrix and A determines the cluster's shape. If A is chosen as the identity matrix, the shape of the clusters leads to the Euclidean distance and eventually, to spherical clusters.

The following 4 steps are executed for implementation of the FCM algorithm [16,17]:

Step 1) Determine the initial parameters: The number of clusters (c), weighting exponent (m), iteration limit (t), termination criterion ($\varepsilon > 0$), and norm for error $|V_t - V_{t-1}|$.

Step 2) Assume the initial values of the cluster centers:

$$V_0 = \{v_{1,0}, v_{2,0}, ..., v_{c,0}\} \subset \Re^{cl}.$$

Step 3) Calculate the membership degree and cluster centers for the $t = 1$ iteration:

$$u_{ik,t} = \left[\sum_{j=1}^{c}\left(\frac{\left\|X_k - V_{i,t-1}\right\|_A}{\left\|X_k - V_{j,t-1}\right\|_A}\right)^{2/m-1}\right]^{-1} \tag{4}$$

and

$$V_{i,t} = \frac{\sum\limits_{k=1}^{M}(u_{ik,t})^m x_k}{\sum\limits_{k=1}^{M}(u_{ik,t})^m}. \tag{5}$$

Step 4) Calculate error $= |V_t - V_{t-1}|$. If the error $\le \varepsilon$, stop the algorithm; otherwise, go on to step (3) for the next t.

2.3. Applied clustering system and experimental setup

A block diagram of the whole classification system can be seen in Figure 1, where it is seen that the applied AIS clustering method was used as a data reduction preprocessing scheme. To compare its performance with another well-known clustering method (FCM), the data reduction scheme was changed to FCM and the results were compared for the same data compression rates.

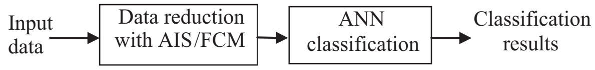

Figure 1. Block diagram of the experimental setup used in the performance evaluation.

The performance of the AIS as a data compression method was analyzed on some benchmark data and its results were compared with the FCM method on the Pima Indian Diabetes and Wisconsin Breast Cancer datasets. Thus, our experimental setup can be organized in 2 branches:

2.3.1. Performance evaluation on the benchmark data

The experiments were done on 2 well-known benchmark datasets: the artificially generated Chainlink dataset and the Iris dataset [18]. The Chainlink dataset consists of 1000 data points with 3 features, where 500 data are in one class and the remaining 500 are in the other. The dataset is shown in Figure 2. In our applications, the number of data was reduced using the AIS and an ANN was trained with this reduced number of data. This used ANN structure was the multilayer perceptron (MLP). Here, they were *Ab*s formed after the AIS was run, which were regarded as reduced input data to the ANN. After training the optimum ANN with that different reduced data, the test classification ratio of the ANN was calculated on the test data, which was the original 1000 point-Chainlink data.

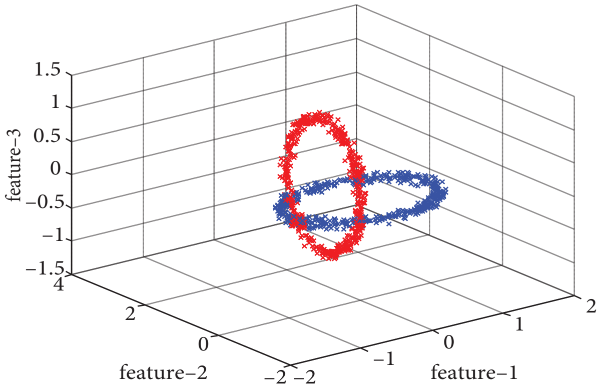

Figure 2. The Chainlink dataset.

The same analysis procedure was conducted for the Iris dataset. The obtained results are given in the results section.

2.3.2. Comparison with FCM clustering on real classification problems

Aside from analyzing the performance of the AIS, we also compared the performance of the AIS as a data reduction method on 2 real-world classification problems. They are diabetes disease and breast cancer classification problems. The related datasets were taken from the University of California, Irvine Machine Learning Repository [18]. In these experimentations however, partitioning of the training and testing data was conducted in a different way. First, the training and testing data were determined and then the training data were reduced with the AIS and FCM methods. In this step, the data were reduced so that approximate compression rates of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, and 98% were obtained. Next, optimum ANN architectures with optimum parameters were trained with these reduced training data and the test data were presented to these trained ANNs to calculate the classification rates. At the same time, these were done 2 times because a 2-fold cross validation method was utilized during the experiments. Let us explain this experimental procedure in a more detailed way for the used datasets separately.

The Pima Indian Diabetes dataset contains 768 samples taken from healthy and unhealthy persons. Of these samples, 500 belong to persons with no diabetes problem, while the remaining 286 samples are of persons with diabetes. The class information contained in this dataset is given by 2 for healthy persons and by 1 for diabetic patients. The number of attributes in the samples is 8. For the purpose of experimentation on this dataset, we first divided the dataset into the 2 folds, as shown in Table 1.

Table 1. Separation of the train-test data for 2-fold cross validation for the Pima Indian Diabetes dataset.

	Fold 1		Fold 2		Total
Class 1	First 250 data for training	Second 250 data for testing	Second 250 data for training	First 250 data for testing	500
Class 2	First 134 data for training	Second 134 data data for testing	Second 134 data for training	First 134 data for testing	268
Total	384	384	384	384	768

By applying the above data preparation procedure, we applied the AIS and FCM to the training data in each fold and reduced the number of training data to some numbers, such that the experimented compression rates were obtained. ANNs were trained many times and optimum parameters were determined giving the highest classification accuracy for each compression rate and data reduction method. By taking the mean value of the test classification accuracies of the 2 folds, performances of the AIS and FCM methods were obtained and compared for different compression rates.

The same procedure was also applied for the Wisconsin Breast Cancer dataset. This dataset consists of 683 data with 9 features. Of these data, 444 belong to subjects with no disease and the remaining 239 are of patients with breast cancer. The training-test partitioning for the 2-fold application is shown in Table 2 for this dataset.

Table 2. Separation of the train-test data for the 2-fold cross validation for the Wisconsin Breast Cancer dataset.

	Fold 1		Fold 2		Total
Class 1	First 222 data for training	Second 222 data for testing	Second 222 data for training	First 222 data for testing	444
Class 2	First 120 data for training	Second 119 data for testing	Second 119 data for training	First 120 data for testing	239
Total	342	341	341	342	683

The obtained results for these 2 problems are presented in the next section.

3. Results and discussion

As stated in the materials and methods section, experiments with the AIS data reduction system were conducted in 2 folds: performance evaluation experiments and real dataset applications with comparisons. Thus, the results are also given in the following subsections under these titles.

3.1. Results for the performance evaluation on the benchmark data

In this category of the experimental study, the Chainlink data and Iris data were compressed using the AIS and the best ANN structure was trained with these reduced data. Next, the trained ANNs with the best parameters were tested on the whole dataset to give test classification rates. The used parameters for the optimum ANNs that give the maximum test classification accuracies are given in Table 3 for each tested compression rate, in addition to the obtained maximum test classification accuracies.

Table 3. Optimum ANN parameters giving the highest classification accuracy for each of the tested compression rates in the Chainlink problem.

				Optimum parameters for the ANN			
Compression rate (%)-CR	Number of data in class 1	Number of data in class 2	Total number of reduced data-N_r	Hidden node number	Learning rate	Momentum constant	Classification accuracy (%)-CA
99	5	5	10	3	0.20	0.50	88.60
95	25	25	50	34	1.00	0.20	96.40
90	50	51	101	15	5.00	0.90	93.90
80	100	100	200	16	1.00	0.40	93.50
70	152	148	300	26	0.20	0.50	91.30
60	201	199	400	26	4.40	0.10	93.40
50	248	253	501	42	0.20	0.90	96.00
40	303	295	598	46	1.20	0.90	96.90
30	353	348	701	31	1.00	0.90	99.70
20	399	403	802	46	2.00	0.40	99.40
10	449	450	899	46	5.00	0.40	99.40

Here, the compression rates were calculated with the aid of Eq. (6), while the test classification accuracies were found using Eq. (7):

$$CR(\%) = \frac{(N_{t-}N_r) * 100}{N_t} \tag{6}$$

Here, CR is the compression rate in percentage form, N_t is the total number of data to be reduced, and N_r is the reduced number of data.

$$CA(\%) = \frac{N_c * 100}{N_t} \tag{7}$$

Here, CA is the classification accuracy in percentage form, N_t is the total number of data, and N_c is the number of correctly classified data. When the results in Table 3, which are also given in graphical form in Figure 3, were evaluated, it was seen that even for high compression rates, satisfactorily high classification rates were obtained. The maximum classification rate was also taken for a compression rate of 30%.

These results were good when seen from the AIS aspect of the situation, but it should be stated here that the Chainlink data classification is not a difficult problem to solve. Thus, we conducted performance evaluation experiments on another benchmark problem, the IRIS classification problem. The IRIS dataset consists of 150 data with 4 features of the IRIS plant. It has 3 classes and each class has 50 data in the dataset. Again, the number of data in each class was reduced to some value and an ANN with optimum parameters was trained with this training data. The test was also done on the whole dataset. The results for different compression rates are given in Table 4 and Figure 4.

As seen from Table 4 and Figure 4, the results seem to be good, even for a low number of training data. Meanwhile, when the successes of other methods in IRIS classification are scrutinized, the classification accuracies about 98% s are very good for this dataset in that many good performing methods obtained this accuracy. This could be also regarded as success for AIS clustering in the data reduction process, but unless conducting a comparison with a state-of-the-art data reduction method, it is not possible to state conclusively that the AIS is good in those kinds of problems. Thus, additional experimentation was realized on 2 well-known real world classification problems: the Pima Indian Diabetes and Wisconsin Breast Cancer classifications. The results of the AIS data reduction stage were also compared with a widely used FCM clustering method, which is a well-known successful scheme used in data reduction problems.

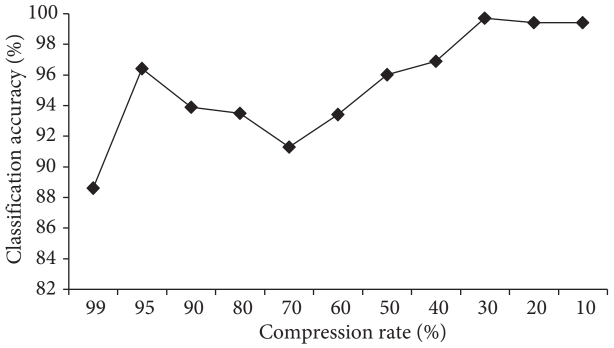

Figure 3. Change of classification accuracy obtained by the AIS clustering scheme with regard to the compression rate for the Chainlink dataset.

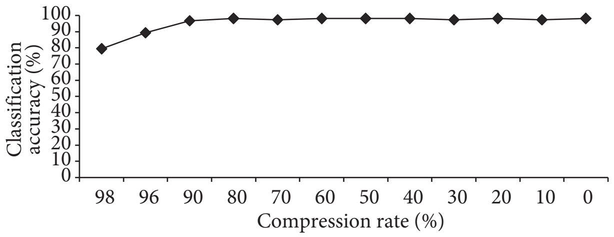

Figure 4. Change of classification accuracy obtained by the AIS clustering scheme with regard to the compression rate for the IRIS dataset.

Table 4. Optimum ANN parameters giving the highest classification accuracy for each of the tested compression rates in the IRIS problem.

Compression rate (%)-CR	Number of data in class 1	Number of data in class 2	Number of data in class 3	Number of reduced data-N_r	Optimum parameters for the ANN			Classification accuracy (%)
					Hidden node number	Learning rate	Momentum constant	
98	1	1	1	3	16	3.4	0.1	79.33
96	2	2	2	6	15	5	0.5	89.33
90	5	5	5	15	5	0.2	0.5	96.67
80	10	10	10	30	8	5	0.5	98.00
70	15	15	15	45	15	5	0.5	97.33
60	20	20	20	60	16	4	0.5	98.00
50	25	25	25	75	1	5	0.5	98.00
40	30	30	30	90	21	5	0.5	98.00
30	35	35	35	105	10	5	0.5	97.33
20	40	40	40	120	31	1	0.5	98.00
10	45	45	45	135	26	5	0.5	97.33

3.2. Results on real-world classification problems and comparison with FCM clustering

The experimental procedure used for this branch of applications is slightly different. For the Pima Indian Diabetes and Wisconsin Breast Cancer datasets, the data were divided 2-fold: 1 fold for training and 1 fold for testing. Training and testing procedures were conducted 2 times in which training and test folds were changed.

The division procedure used for the Pima Indian Diabetes dataset is given in Table 1. Training of the ANN was conducted 2 times; 1 for each fold, and optimum parameters giving the highest classification accuracies were found. The optimum ANN parameters found for the reduced training data used are shown in Table 5 for the Pima Indian Diabetes classification problem.

Table 5. The optimum ANN parameters found for each fold and for each value of CR during applications with the Pima Indian Diabetes data (hn: number of hidden nodes, lr: learning rate, and mc: momentum constant).

Number of data in class 1	Number of data in class 2	Total number of training data	CR (%)	Optimum ANN parameters for fold 1			Optimum ANN parameters for fold 2		
				hn	lr	Mc	hn	lr	mc
5	2	7	98.18	19	5.00	1.00	1	2.40	0.90
12	7	19	95.05	16	0.80	0.50	3	5.50	0.20
25	14	39	89.84	1	4.20	0.50	1	4.00	0.50
50	26	76	80.20	1	2.00	0.50	1	5.00	0.80
75	40	115	70.50	2	2.50	1.00	1	7.80	1.00
100	53	153	60.15	2	1.50	1.00	2	6.50	1.00
125	67	192	50.00	2	6.50	1.00	8	4.90	1.00
151	80	231	40.10	1	8.00	0.10	8	5.00	1.00
175	94	269	29.94	2	5.00	1.00	2	4.50	1.00
200	107	307	20.05	2	5.00	1.00	2	5.10	1.00
225	120	345	10.15	2	4.70	1.00	8	4.90	1.00
250	134	384	0.00	8	4.00	0.50	8	5.00	1.00

The obtained classification accuracies for these CR values are shown in Table 6 for FCM and the AIS.

Table 6. Test classification results obtained with FCM and the AIS for different values of CR in the Pima Indian Diabetes classification problem.

CR (%)	Fold 1		Fold 2		Mean CA of the 2 folds	
	CA of FCM (%)	CA of AIS (%)	CA of FCM (%)	CA of AIS (%)	CA of FCM (%)	CA of AIS (%)
98.18	27.60	72.13	38.30	69.27	32.95	70.70
95.05	41.40	70.83	57.50	69.79	49.45	70.31
89.84	70.05	69.01	56.00	76.04	63.03	72.53
80.20	70.30	72.40	55.46	75.52	62.88	73.96
70.50	30.21	70.83	54.40	70.57	42.31	70.70
60.15	25.78	66.14	52.30	70.31	39.04	68.23
50.00	37.76	65.62	50.26	65.62	44.01	65.62
40.10	70.30	65.62	50.26	66.40	60.28	66.01
29.94	39.30	68.75	51.30	70.05	45.30	69.40
20.05	38.02	69.01	36.72	68.48	37.37	68.75
10.15	39.80	68.75	50.00	70.83	44.90	69.79
0.00	50.52	79.17	50.00	70.65	50.26	74.91

Here, in the comparison, the same ANN parameters were utilized for a given CR in the training process for FCM and the AIS. As can be seen from Table 6, the AIS has an evident over-performing result as a data reducing method for the classification of the Pima Indian Diabetes classification problem. To see this difference more clearly, the change of classification accuracy with respect to the compression rate is given graphically in Figure 5.

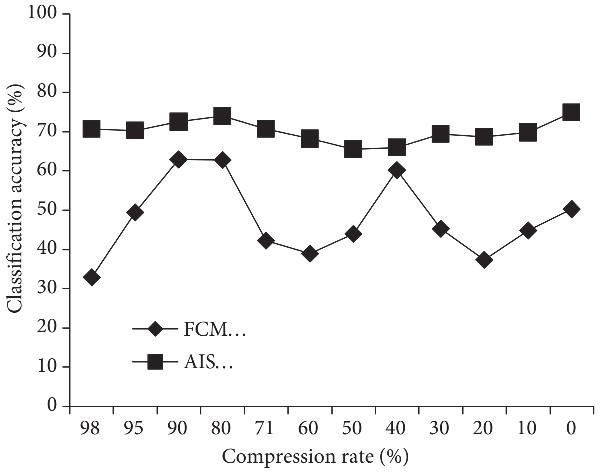

Figure 5. Change of classification accuracy with regard to the compression rate for the Pima Indian Diabetes dataset.

When FCM was applied in the data reduction stage, the classification accuracy of the classifier changed between 32.95% and 63.03%, where especially for the compression rate of 80.20%, which is a relatively high rate, the AIS has performed very well by reaching a classification accuracy of 73.96%. However, the same was not valid for FCM. For that compression rate, the accuracy with FCM was 62.88%. As seen from Figure 5, there are points in which FCM's classification accuracy was close to that of the AIS, but in general it can be said that the AIS performed better than FCM for every compression rate.

When it comes to the application of the AIS and FCM to the Wisconsin Breast Cancer classification, similar results were seen. The experimental setup was the same as for the Pima Indian Diabetes data and the data division procedure for the experiments is summarized in Table 2. The procedure was the same as for the Pima Indian Diabetes data and the obtained optimum ANN parameters are given in Table 7 for different CR values. The obtained classification accuracies with FCM and the AIS for the CR values and ANN parameters given in Table 7 are shown in Table 8.

Table 7. The optimum ANN parameters found for each fold and for each value of CR during the applications with the Wisconsin Breast Cancer data (hn: number of hidden nodes, lr: learning rate, and mc: momentum constant).

Number of data in class 1	Number of data in class 2	Total number of training data	CR (%)	Optimum ANN parameters for fold 1			Optimum ANN parameters for fold 2		
				hn	lr	Mc	hn	lr	mc
4	2	6	98.24	8	0.20	1.00	14	5.00	0.90
11	6	17	95.02	2	5.00	0.50	3	3.60	0.50
21	13	34	90.05	1	5.00	1.00	2	5.00	0.10
44	24	68	80.11	6	5.00	0.20	1	5.00	1.00
66	36	102	70.17	3	5.00	0.50	9	5.00	0.50
86	51	137	59.94	1	5.00	0.50	2	5.00	0.50
111	60	171	50.00	1	5.00	0.50	9	5.00	0.50
134	72	206	40.05	7	5.00	0.50	3	5.00	0.50
152	85	237	30.11	7	5.00	0.80	3	5.00	0.50
176	97	273	20.17	9	5.00	0.70	3	5.00	0.50
200	108	308	9.94	9	5.00	0.70	3	5.00	0.30
222	120	342	0.00	9	5.00	0.70	3	5.00	0.50

Table 8. Test classification results obtained with FCM and the AIS for different values of CR in the Wisconsin Breast Cancer classification problem.

CR (%)	Fold 1		Fold 2		Mean CA of the 2 folds	
	CA of FCM (%)	CA of AIS (%)	CA of FCM (%)	CA of AIS (%)	CA of FCM (%)	CA of AIS (%)
98.24	80.35	81.52	78.36	81.87	79.36	81.70
95.02	91.49	89.15	88.59	85.08	90.04	87.12
90.05	78.88	94.72	81.57	93.27	80.23	94.00
80.11	52.19	97.08	83.62	95.61	67.91	96.35
70.17	92.96	97.95	80.41	95.61	86.69	96.78
59.94	79.18	97.95	81.58	96.20	80.38	97.08
50.00	81.23	98.53	80.99	95.90	81.11	97.22
40.05	81.52	98.83	87.13	96.77	84.33	97.80
30.11	84.46	98.53	86.26	96.77	85.36	97.65
20.17	18.47	98.53	76.90	97.06	47.69	97.80
9.94	48.68	98.53	76.90	97.06	62.79	97.80
0.00	41.35	98.53	77.19	96.48	59.27	97.51

If the results given in Table 8 are evaluated, it can be said that for many of the CR values the AIS performed better than FCM with respect to the classification accuracies. However, at this time, the difference between the AIS and FCM is not as definite as in the Pima Indian Diabetes dataset. To see the difference in a clearer way, the test classification accuracies of FCM and the AIS were drawn with respect to the changing compression rates in Figure 6.

The Figure 6 shows that for the compression rates of 98% and 95%, the classification results of the AIS and FCM were not so different; furthermore, they were very similar. However, for the remaining compression rate values, the difference increased in favor of the AIS. For compression rates of about 50% s, the classification accuracies were in the order of 97% s and this is a very good classification accuracy for this classification problem.

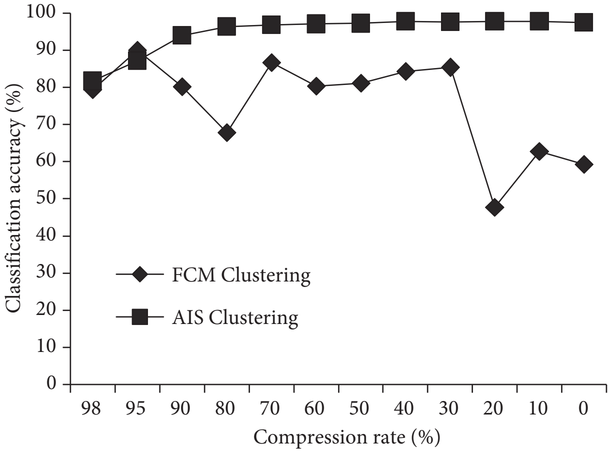

Figure 6. Change of classification accuracy with regard to the compression rate for the Wisconsin Breast Cancer dataset.

4. Conclusion

Many aspects affect the overall performance of a classification process. Preprocessing steps such as feature reduction or data reduction are among these. Thus, selecting the correct preprocessing scheme is as important as defining the correct classification method. In this study, we used an AIS as a sample reduction process before the ANN, which is used as classifier. We produced some Abs in response to the presented data and used this Abs to give the classifier as training data.

To evaluate the performance of the AIS system as a data reduction method, we used a classification structure, which consisted of the data reduction stage of the AIS and the classifier stage of the ANN. We applied this structure to the Chainlink dataset, which is an artificially generated dataset, and the IRIS dataset, which is a well-known benchmark dataset. The results were very promising. The AIS-ANN classifier successfully classified the Chainlink dataset, with a classification accuracy of 99.70% at a 30% compression rate. The result for the IRIS dataset was also good, in that a classification accuracy of 98.00% was obtained, even at an 80% compression rate. Although these results were very good for our method, they do not carry any meaning unless a comparison with a state-of-the-art method is conducted. Thus, we selected the FCM clustering algorithm as a data reduction method in the used ANN-based classification structure. We compared the results of the AIS-ANN classifier with the results of the FCM-ANN classifier on 2 real-world classification problems: the Pima Indian Diabetes and Wisconsin Breast Cancer classification problems. The applications were conducted using the 2-fold cross validation method. In general, the AIS-ANN configuration reached higher classification accuracies for each configuration. For the Pima Indian Diabetes dataset, the maximum classification accuracy when data reduction was applied was seen as 73.96% for a compression rate of 80.30%. However, the FCM-ANN method reached a maximum classification accuracy of 63.03% for an 89.84% compression rate. These findings were also similar for the Wisconsin Breast Cancer dataset. The AIS-ANN method reached a 97.80% classification accuracy for the compression rates of 40.05%, 20.17%, and 9.94%. The FCM-ANN method, on the other hand, reached a maximum classification accuracy of 90.04% for a 95.02% compression rate.

The big differences between the results of the Pima Indian Diabetes and Wisconsin Breast Cancer classifications were caused from the nature of the problems. When the other studies conducted on these datasets were analyzed, it was seen that high classification accuracies could be achieved for the Wisconsin Breast Cancer dataset, while the Indian Diabetes Pima dataset is a difficult classification problem. As can be deduced from the obtained results, the AIS has proven itself as a successful data reduction method in the applied datasets. This success can be attributed to the data transforming feature of the AIS method used. The AIS produces new data in the form of Abs in response to the presented data and the nature of this data transforming made the AIS successful in the experimented datasets. Further work will be realized in this category of the data mining area with AIS-based methods.

Acknowledgment

This work is supported by the Coordinatorship of the Selçuk University's Scientific Research Projects Grant.

References

[1] M. Kantardzic, Data Mining: Concepts, Models, Methods and Algorithms, New York, Wiley, 2002.

[2] P. Dulyakarn, Y. Rangsanseri, "Fuzzy C-means clustering using spatial information with application to remote sensing", Proceedings of the 22nd Asian Conference on Remote Sensing, 2001.

[3] J. Timmis, "Artificial immune systems: a novel data analysis technique inspired by the immune network theory", PhD Dissertation, Department of Computer Science, University of Wales, U.K., 2000.

[4] J. Timmis, M. Neal, "A resource limited artificial immune system for data analysis", Knowledge Based Systems, Vol. 14, pp.121–130, 2001.

[5] L.N. De Castro, F.J. Von Zuben, "An evolutionary immune network for data clustering", Proceedings of the IEEE Brazilian Symposium on Artificial Neural Networks, pp. 84–89, 2000.

[6] P. Vishwambhar, D. Praven, M. Prabhat, "Data clustering with artificial innate immune system adding probabilistic behaviour", International Journal of Data Mining and Emerging Technologies, Vol. 1, pp.77–84, 2011.

[7] A.J. Graff, A.P. Engelnrecht, "Clustering data in an uncertain environment using an artificial immune system", Pattern Recognition Letters, Vol. 32, pp. 342–351, 2011.

[8] W. Ahmad, A. Narayanan, "Population based artificial immune system clustering algorithm", Proceedings of the 10th International Conference on Artificial Immune Systems, pp. 348–360, 2011.

[9] C.Y. Chiu, C.H. Lin, "Cluster analysis based on artificial immune system and ant algorithm", Proceedings of the 3rd Annual Conference on Natural Computation, Vol. 3, pp. 647–650, 2007.

[10] Z. Liu, X. Jin, R. Bie, X. Gao, "FAISC: a fuzzy artificial immune system clustering algorithm", Proceedings of the 3rd Annual Conference on Natural Computation, Vol. 3, pp. 657–661, 2007.

[11] T. Liu, Y. Zhau, H. Zhifeng, "A new clustering algorithm based on artificial immune system", Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, pp. 347–351, 2008.

[12] L. Lan, R. Qiao-Mei, "Implementation of clustering algorithm using artificial immune system", Proceedings of the 1st International Workshop on Database Technology and Applications, pp. 275–278, 2009.

[13] A. Secker, M.N. Davies, A.A. Freitas, J. Timmis, E. Clark, D.R. Flower, "An artificial immune system for clustering amino acids in the context of protein function classification", Journal of Mathematical Modelling and Algorithms, Vol. 8, pp. 103–123, 2009.

[14] D. Dasgupta, Artificial Immune Systems: A Bibliography. Technical Report, No: CS-07-004, USA, 2007.

[15] A.S. Perelson, G.F. Oster, "Theoretical studies of clonal selection: minimal antibody repertuarie size and reliability of self-nonself discrimination", Journal of Theoretical Biology, Vol. 81, pp. 645–670, 1979.

[16] J.S.R. Jang, C.T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing, Prentice Hall, USA, 1997.

[17] D. Li, H. Gu, L. Zhang, "A fuzzy c-means clustering algorithm based on nearest neighbour intervals for incomplete data", Expert Systems with Applications, Vol. 37, pp. 6942–6947, 2010.

[18] http://www.ifs.tuwien.ac.at/dm/somtoolbox/datasets.html, Last accessed: 1 June 2011).