AJIT-e: Online Academic Journal of Information Technology
2014 Summer/Yaz – Cilt/Vol: 5 - Sayı/Num: 16
DOI: 10.5824/1309-1581.2014.3.001.x
http://www.ajit-e.org/?menu=pages&p=details_of_article&id=117

Comparison Of K Nearest Neighbours And Regression Tree Classifiers Used With Clonal Selection Algorithm to Diagnose Haematological Diseases

Burcu ÇARKLI YAVUZ, *Sakarya University, Information Systems Engineering Department, bcarkli@sakarya.edu.tr*

Tuba KARAGÜL YILDIZ, *Sakarya University, Computer Engineering Department, tkaragul@sakarya.edu.tr*

Nilüfer YURTAY, *Sakarya University, Computer Engineering Department, nyurtay@sakarya.edu.tr*

Ziynet PAMUK, *Sakarya University, Electrical and Electronics Engineering Department, ziynet@sakarya.edu.tr*

ABSTRACT *The aim of this study is to develop a method to improve the classification performance by haematological parameters. In classification problems it has been seen that kNN classifier is often used with the clonal selection algorithm. In this study unlike other studies Gini algorithm is performed instead of kNN classification algorithm and higher success rate is obtained. According to the World Health Organisation's data nearly 10% of women in the world are anaemia. Anaemia is a disease that disrupts life quality and results in serious effects if not cured. Iron deficiency anaemia is the most common type of anaemia and women suffers this disease comparatively to men. Therefore, in this study, anaemia was preferred as a sample application. It is expected to reach successful results in diagnosis of other diseases by looking at haematological parameters with the proposed method. At the end of the study success ratios of different methods are compared by Receiver Operating Characteristics analysis method. While accuracy in memory-based classification is found as 96%, accuracy in regression tree method classification is 98.73%. Using Gini algorithm instead of kNN a higher success ratio is achieved so CSA surpassed ANN's success ratio.*

Keywords: *Artificial Neural Network, Clonal Selection Algorithm, Computer Aided Diagnosis, K-Nearest Neighbours, Regression Trees.*

Introduction

Anaemia is defined as lack of red blood cells or usually less haemoglobin ratio in blood. Human body needs iron to synthesis haemoglobin and make them to carry oxygen molecules. Inadequate iron intake results in iron deficiency anaemia. Women having iron deficiency anaemia suffer symptoms like weakness overstrain, pale skin, shortness of breath, irregular heartbeat, and these symptoms disrupt life quality.

Due to menstrual blood loss and less intake of iron with nutrients, women are much more affected than men from iron deficiency anaemia (Hillman et al., 2012).

According to a study carried out in 2012, 8.84% of world population suffers from moderate anaemia. This ratio is 9.93% among women. If advanced and mild anaemia problems are added, this ratio would increase dramatically (Vos et al., 2012). Anaemia diagnosed patient ratio is 5.3% in Turkey in 2012 and, 9.5% in women in country wide (Turkish Statistical Institute, 2013). Therefore, anaemia was preferred as a sample application area.

With the development of computer technology, studies have been carried out and high success ratios have been achieved by heuristic approaches such as Artificial Neural Network (ANN) and Genetic Algorithms (GA).

Artificial Immune Systems (AIS) is emerged in 1990s as a new system which integrates biology-based calculation methods as ANN and Artificial Life (AL) (Nasraoui et al., 2002).

AIS also achieved successful results in medicine. Kodaz et al.(2009) introduced a successful study that uses AIS to diagnose thyroid, and reach a success ratio of 95.9%. Polat et al. (2007) also studied on thyroid diagnose with a fuzzy-based AIS. According to their classification success ratio is realized as 85%. Latifoğlu et al. (2007) used AIS in atherosclerosis diagnose and reach an accuracy of 99.29%. Şahan et al.(2007) also reached a high accuracy ratio of 99.14% in breast cancer diagnosis. Kihel & Benyettou (2011) used AIS in Parkinson disease diagnose and achieved successful results. Şengür (2008) achieved 95.9% successful results in mitral valve diseases. Er et al. (2012) introduced a method that can be used in diagnosis of pneumonia, tuberculosis, asthma and lung cancer by AIS and achieved over 90% success ratio for each disease. Bozkurt et a.l (2014) diagnosed diabetes mellus with an accuracy of 68.8% by using AIS.

As understood from the given studies success ratio of AIS algorithm is relatively high. Therefore studies in this area are increasing continuously.

Masala et al. (2013) reached an accuracy ratio of 92.3% by using ANN and kNN in their classification study on diagnose of Thalassemia – a significant type of anaemia.

One of the AIS algorithms, CSA is used to train data in this study to diagnose iron deficiency anaemia in women, separate classifications made and compared by kNN and regression tree methods.

Characteristics of AIS such as recognizing and terminating foreign cells, having memory attracts attention of non-medical researchers, and several successful studies were held by them. Some of these non-medical research areas are computer security (Mohammad & Zitar, 2011; Sobh &Mostafa, 2011), optimization problems (De Castro & Von Zuben, 2002, 2002), workshop scheduling (Tjornfelt-Jensen & Hansen, 1999; Hart et al., 1998), production line control (Mori et al., 1997), autonomous robot system (Jun et al., 1999).

Materials and Methods

Data Set

In this study, whole blood analysis results of 2600 women patient were obtained from Zonguldak City Hospital in 2010. Haematological parameters of laboratory data are given in Table 1 (Özaslan & Delibaşı, 2012).

Table 1: Haematological Parameters Of Iron Deficiency Anaemia

Parameter	Description	Values
RBC	Red Blood Cells	4,5-6
HGB	Haemoglobin	12-16
HCT	Haematocrit	36-48
MCV	Mean Corpuscular Volume	80-100
MCH	Mean Corpuscular Haemoglobin	27-34
MCHT	Mean Corpuscular Haemoglobin Concentration	31-37

Doctors take advantage of the parameters given in Table 1 to diagnose iron deficiency anaemia. In this study, to diagnose anaemia 6 input parameters RBC, HGB, HCT, MCV, MCH, MCHT were used and to identify the class output parameters “anaemia” and “healthy” were used. Distribution of data used for training and testing is given in Table 2.

Table 2. Preliminary data set

Set	Anaemia	Healthy	Total
TRAIN	406	1594	2000
TEST	122	478	600
ALL	528	2072	2600

Artificial Immune System (AIS)

In recent years a significant increase in the number of researches on biology-based systems has been witnessed. Studies in this area have in common that they fulfil vital functions of the human system is used as a source of inspiration.

Biological Immune System, including complex cells, molecules and organs, is a system that officiate several duties such as pattern recognition, learning, memory management, creating diversity, generalization, recognition, and optimization. Calculation techniques based on immune principles targets not only to understand the present system but also solve a lot of engineering problems (De Castro & Von Zuben, 1999).

AIS are modelled as well as other nature-inspired computing techniques from natural life and their ability to adapt.

AIS exhibits a similar approach on the attribute of human bodies' ability and recognize microbes. AIS implements lymphocyte activities, natural antibody production, pre-immunization, selection, tolerance, memory simultaneously. AIS also produces the results by using antibody, antigen, affinity and threshold notions (Garrett, 2005).

AIS indicate a framed manner. Layered framework for AIS is given in Figure 1.

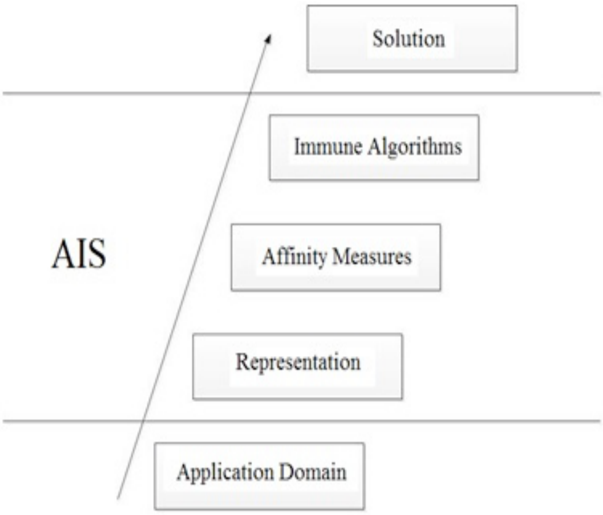

Figure 1. General framework for AIS

First layer of the system is an application domain. Application domain of the system in this study is diagnosis of iron deficiency anaemia in women. In representation layer an appropriate represent is selected for the components that are in the application domain. Hamming or Euclid distances are used to calculate affinity measures according to represents' status is real or binary. Immune Algorithms layer includes algorithms that determine system behaviour. CSA and AIS are the theories utilised for immune system modelling (De Castro & Timmis, 2002). The most common used AIS algorithms are the algorithms inspired from Clonal Selection Theory, Negative Selection Theory and Immune Network Theory. Biological pattern of Clonal Selection Theory used in this study is given in Figure 2.

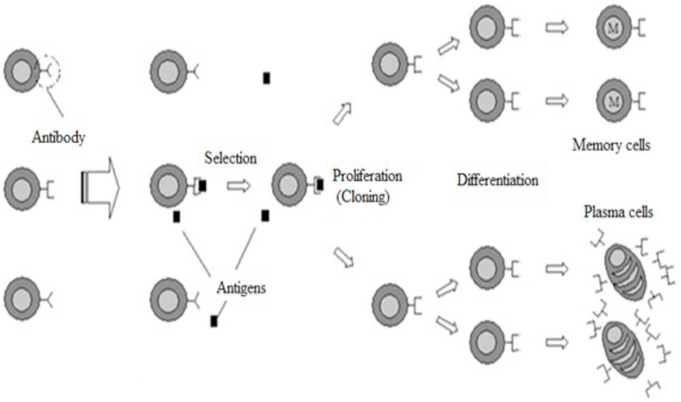

Figure 2. Biological pattern of Clonal Selection Theory (De Castro & Von Zuben, 2002)

Looking at the similarity of antibodies to antigens namely B cells produced by bone marrow, ones having the highest similarity are selected and reproduced. These new replicated cells are subjected to mutation process so differentiation of the cells is provided. Compliance with these newly formed cells to antigens is controlled, cells with high antigen sensitivity turn into memory cells, cells with less antigen sensitivity turn into plasma cells. The ability of immune system to recognize different antigens is increased through these updates.

Suggested System

A system based on Clonal Selection Algorithm is suggested in this study. General frame of the system is given as follows in Figure 3.

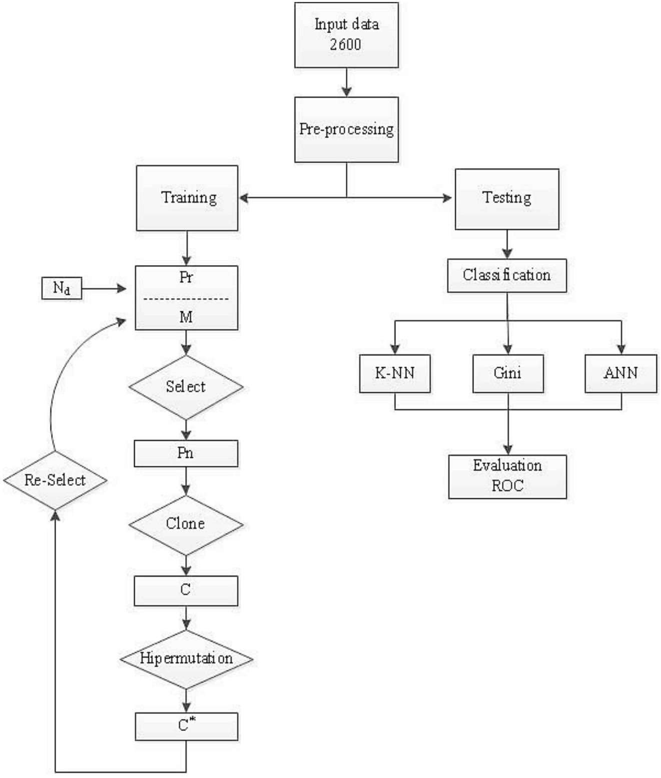

Figure 3. Block diagram of the suggested system

As mentioned in Figure (3) blood parameters of 2600 women are used in this study. These parameters are pulled into -1 and +1 values namely standardized by z-score transformation. A CSA is used to train the data in the training phase (De Castro & Von Zuben, 2002).Steps of CSA are presented below.

Step l. Generate a set of (P) of candidate solutions composed by addition of a part of memory cells (M) to the remaining population (Pr) (P=Pr+M)

Step 2. Select the n best individuals of the population (Pn) based on affinity measure. Euclid distance Formula given in Equation (1) is used to select the best individual.

$$D = \sqrt{\sum_{i=1}^{L} (Ab_i - Ag_i)^2} \qquad (1)$$

In Equation (1), Ab indicates antibody, Ag indicates antigen, L indicates number of attributes, D indicates Euclid distance.

Step 3. Reproduce - Clone these n best individuals of the population, increasing a temporary population of clones (C) (the clone size is an increasing function of the affinity with the antigen)

Step 4. Submit the population of clones to a hyper mutation scheme, generate a maturated antibody population (C*) (the hyper mutation is proportional to the affinity of the antibody with the antigen). Mutation ratio is selected as 4.

Step 5. Reselect the improved individuals from C* to compose the memory set M (Some members of P set can be replaced by other improved members of C*)

Step 6. Replace the lower affinity antibodies of the population by novel ones. Nd individuals replace with M memory cells.

At the end of the training held by CSA number of data rose to 3336 healthy and 685 anaemia diseased from primal number of 2000.

Classifiers

Classification methods are often used in medical decision support systems and medical diagnosis. In testing phase of this study, classifiers are selected as kNN – standard classifier of CSA, and Gini algorithm – one of regression tree methods, are selected. In addition, to measure the success of the proposed method, the classification is also made with FFN (feed-forward network) and PNN (probabilistic neural network)-two of ANN methods. Using these classification methods, class of a new observation value is decided, and compared with previously known class values.

K-Nearest Neighbours Algorithm (kNN)

kNN method is a distance based algorithm and steps given below are applied in this method:

Step 1. K parameter is selected. This parameter is the number of the neighbours which are nearest to the point given (k=60 for this study).

Step 2. Distance between the point and the all remaining points are calculated separately. Euclid distance formula is used to determine the distance. Distances between every other values of 600 test set individuals and 4021 observation values, trained by CSA, are calculated.

Step 3. Lines are arranged due to calculated calculations and the lowest k's are selected.

Step 4. Categories of the selected lines are determined and the most repeated class value is selected.

Step 5. The most repeated class value is accepted as the class of the new observation value (Özkan, 2013).

Gini Algorithm As a Regression Tree Method

Gini algorithm, based on the principle of characteristic values' which are divided into two parts as left and right, is a classification technique that generates a regression tree. According to the Gini algorithm; every other characteristic value are grouped dual. By this way, the resulting left and right values are grouped corresponding to the classes. Values given in Equation (2) and (3) are calculated.

$$Gini_{left} = 1 - \sum_{i=1}^{k}\left(\frac{L_i}{|T_{left}|}\right)^2 \tag{2}$$

$$Gini_{right} = 1 - \sum_{i=1}^{k}\left(\frac{R_i}{|T_{right}|}\right)^2 \tag{3}$$

Terms used in Equation (2) and (3) are given below.

k	: Number of classes
T	: Samples on a node
|Tleft|	: Number of samples on the left side
|Tright|	: Number of samples on the right side
Li	: Number of samples on the left side in category i
Ri	: Number of samples on the right side in category i

For every j characteristic, n as the number of elements in training set, Equation (4) is calculated.

$$Gini_j \quad \frac{1}{n}\left(|T_{left}|Gini_{left} + |T_{right}|Gini_{right}\right) \tag{4}$$

With the selection of the lowest valued Ginij which is calculated for every j characteristic, separation realized on this characteristic. Later process starts again from the beginning step (Özkan, 2013).

In this study, regression tree generated by the Gini algorithm is given in Figure 4.

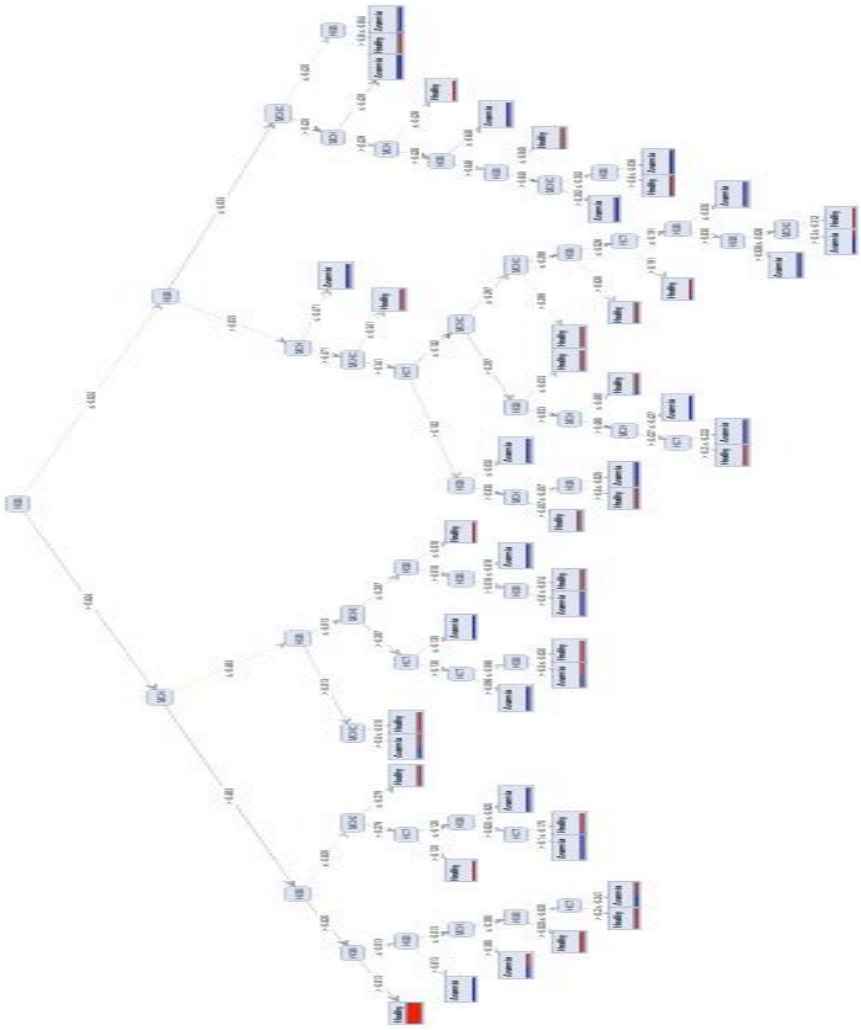

Figure 4. Regression tree generated by the Gini algorithm

Cross validation method is used to determine the training and test data to classify with Gini algorithm. The algorithm has been run for different fold numbers and, the best performance has been reached with 6 folds. Data are divided to 6 fold and so classifying process repeated 6 times. Accuracy ratio is accepted as average of 6 times repeated classifications' accuracy ratios.

Artificial Neural Network Methods

Feed Forward Networks (FFN)

Feed-forward networks can be used for any type of input to output problems. A feed-forward network with one hidden layer and enough neurons in the hidden layers can fit any finite input-output mapping problem.

There is a series of layers in feed-forward networks. The first layer has a connection between the network input and the last layer produces the network's output. Each subsequent layer has a connection from the previous layer (MATLAB Documentation Neural Network Toolbox Help, Feedforward neural network, 2014).

Probabilistic Neural Networks (PNN)

In classification problems, probabilistic neural networks can be used. When an input is presented, the first layer computes distances from the input vector to the training input vectors and generates a vector that its elements indicate how close the input is to a training input. The second layer gathers these contributions for each class of inputs to generate as its net output a vector of probabilities. Finally, a complete transfer function on the output of the second layer picks the maximum of these possibilities, and generates a 1 for that class and a 0 for the remaining classes (MATLAB Documentation Neural Network Toolbox Help, Probabilistic neural networks, 2014).

Results

After training process with CSA, distinct classification processes are realized in the test phase with kNN and Gini algorithms. Additionally the method we propose also has been compared with ANN, recognized as a successful classifier. ROC analysis technique is used to measure the level of performances of classifiers (Weinstein & Obuchowski, 2005; Fawcett, 2006).

Used commonly in medical decision making, ROC analysis have been recently used in machine learning and data mining techniques, too. The common feature of the methods used in Classification processes is to try to create balance between accuracy (elimination ability of the false positives) and sensitivity (identification ability of the true positives). Accuracy, Sensitivity and Specificity values used for ROC analysis are calculated by the use of the formulas given in formulas (5), (6) and (7).

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{5}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{6}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{7}$$

Expressions used in the formulas;

TP (True Positive): The number of “Anaemia”s, who are really “Anaemia” in the real results, at the end of classifier result.

TN (True Negative): The number of "Healthy"s, who are really "Healthy" in the real results, at the end of classifier result.

FP (False Positive): The number of "Anaemia"s, who are really "Healthy" in the real results, at the end of classifier result.

FN (False Negative): The number of "Healthy"s, who are really "Anaemia" in the real results, at the end of classifier result.

Results obtained from the classifications made with kNN, Gini and ANN methods are indicated in Table 3. And comparisons of all these methods are given as a graphical in Figure 5.

Table 3. Comparison of used classifiers

Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)
CSA-kNN	96,00	86,89	98,33
CSA-Gini	98,73	96,35	99,25
FFN	98,50	96,72	98,95
PNN	97,49	94,26	98,32

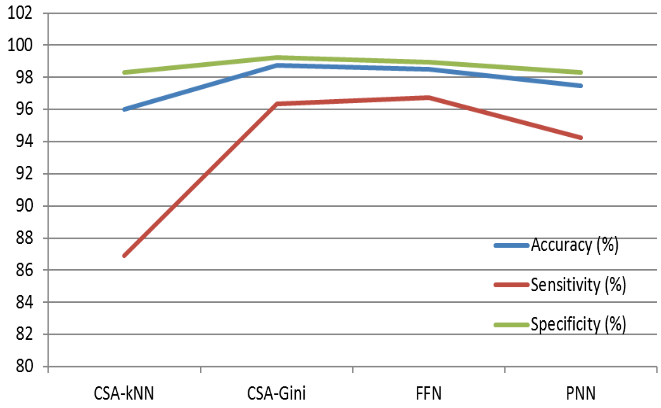

Figure 5. Graphical results of the classification

When ROC results of kNN and Gini classifications are compared, it is seen that classification made with Gini algorithm is more successful than the classification made with kNN for this study. The algorithm made with Gini is also surpassed ANN classifiers-accepted as successful method.

The ROC curve indicates the successes of two classifiers are given in Figure 6.

ROC curve is used to evaluate the balance between accuracy and sensitivity. The area left under the ROC curve is defined as the ROC score. ROC curve plotted due to changing classification threshold values of true positives as a function of false positives. If ROC score is 1 (one) that means, "positives separated from negatives in an excellent way". If ROC score is 0 (zero) that means "no positives is found". As seen in Figure 6, the area left under ROC curve plotted for Gini algorithm is larger than the area left under the ROC curve plotted for kNN algorithm. This result indicates that classification made with Gini algorithm is more successful than the classification made with kNN.

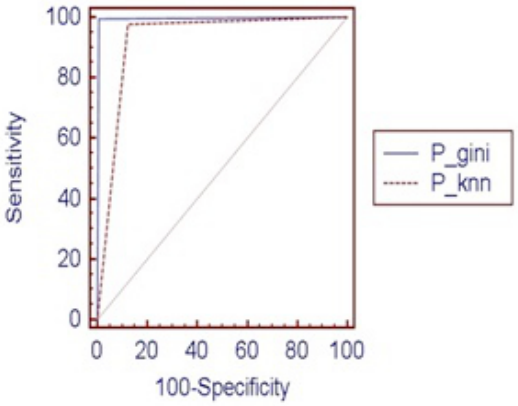

Figure 6. ROC curve for kNN and Gini classifiers (MedCalc, version 12.5)

Conclusion

AIS algorithms experienced in several engineering and medical applications and successful results are achieved. In this study clonal selection algorithm is used for iron deficiency anaemia diagnosis and succeeds well. This method is quite accomplished for medical diagnosis studies, and shall be used for different disease diagnosis studies. Also it is expected to yield good results in another diagnosis made by looking at blood values.

Clonal selection algorithm is often used with kNN classifiers. Unlike the previous studies held, this study makes classification and compares Gini algorithm that makes classification based on regression tree along with clonal selection algorithm, with kNN method. Use of Gini algorithm added a different dimension to this study. Accuracy of classification made with kNN is 96% while accuracy of classification made with Gini is 98.73%. Also accuracy of classification made with FFN is 98.50% and PNN is 97.49%. It is obviously seen in the test results that using Gini algorithm with clonal selection algorithm results in a more successful classification than in both kNN and ANN algorithms.

Acknowledgements

We thank Zonguldak State Hospital chief physician and Information Processing Centre personnel for allowing us to use their data.

References

Bozkurt, M. R., Yurtay, N., Yilmaz, Z., & Sertkaya, C. (2014). Comparison of different methods for determining diabetes. Turkish Journal of Electrical Engineering & Computer Sciences,22(4), 1044–1055. doi:10.3906/elk-1209-82

De Castro, L. N., & Timmis, J. (2002). Artificial Immune Systems: A Novel Paradigm to Pattern Recognition. InUniversity of Paisley (pp. 67–84). Springer Verlag, University of Paisley, UK.

De Castro, L. N., & Von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. IEEE Transactions on Evolutionary Computation,6(3), 239–251. doi:10.1109/TEVC.2002.1011539

De Castro, L. N., & Von Zuben, F. J. (2002). The Clonal Selection Algorithm with Engineering Applications. InIn GECCO 2002 - Workshop Proceedings (pp. 36–37). Morgan Kaufmann.

De Castro, L. N. & Von Zuben, F. J. (1999). Artificial Immune Systems: Part I-Basic Theory and Applications. Technical Report, TR-DCA 01/99.

Er, O., Yumusak, N., & Temurtas, F. (2012). Diagnosis of chest diseases using artificial immune system. Expert Systems with Applications, 39(2), 1862–1868. doi:10.1016/j.eswa.2011.08.064

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters (vol 27, pp 861-874).

Garrett, S. M. (2005). How Do We Evaluate Artificial Immune Systems? Evol. Comput., 13(2), 145–177. doi:10.1162/1063656054088512

Hart, E., Ross, P., & Nelson, J. (1998). Producing robust schedules via an artificial immune system. IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (pp. 464–469). doi:10.1109/ICEC.1998.699852

Hillman, R. S., Ault, K. A., Leporrier, M. & Rinder, H. M. (2012). Hematology in Clinical Practice. McGraw-Hill Publisher.

Jun, J.-H., Lee, D.-W., & Sim, K.-B. (1999). Realization of cooperative strategies and swarm behavior in distributed autonomous robotic systems using artificial immune system. IEEE International Conference on Systems, Man, and Cybernetics. IEEE SMC '99 Conference Proceedings (Vol. 6, pp. 614–619). doi:10.1109/ICSMC.1999.816622

Kihel, B. K., & Benyettou, M. (2011). Parkinson's Disease Recognition Using Artificial Immune System. Journal of Software Engineering and Applications, 04(07), 391–395. doi:10.4236/jsea.2011.47045

Kodaz, H., Özşen, S., Arslan, A., & Güneş, S. (2009). Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease. Expert Systems with Applications, 36(2, Part 2), 3086–3092. doi:10.1016/j.eswa.2008.01.026

Latifo◎lu, F., Kodaz, H., Kara, S., & Güneş, S. (2007). Medical application of Artificial Immune Recognition System (AIRS): Diagnosis of atherosclerosis from carotid artery Doppler signals. Computers in Biology and Medicine,37(8), 1092–1099. doi:10.1016/j.compbiomed.2006.09.009

Masala, G. L., Golosio, B., Cutzu, R., & Pola, R. (2013). A two-layered classifier based on the radial basis function for the screening of thalassaemia. Computers in Biology and Medicine,43(11), 1724–1731. doi:10.1016/j.compbiomed.2013.08.020

MATLAB® Documentation Neural Network Toolbox Help. (2014). Feedforward neural network, Release 2014a, The MathWorks Inc.

MATLAB® Documentation Neural Network Toolbox Help. (2014). Probabilistic neural networks, Release 2014a, The MathWorks Inc.

MedCalc for Windows, version 12.5 MedCalc Software, Ostend, Belgium.

Mohammad, A. H., & Zitar, R. A. (2011). Application of genetic optimized artificial immune system and neural networks in spam detection. Applied Soft Computing, 11(4), 3827–3845. doi:10.1016/j.asoc.2011.02.021

Mori, K., Tsukiyama, M., & Fukuda, T. (1997). Artificial immunity based management system for a semiconductor production line. IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation (Vol. 1, pp. 851–855 vol.1). doi:10.1109/ICSMC.1997.626207

Nasraoui, O., Dasgupta, D., & Gonzalez, F. (2002). The Promise and Challenges of Artificial Immune System Based Web Usage Mining. Workshop on Web Analytics at Second SIAM, International Conference on Data mining (SDM). Arlington.

Özaslan, E. & Delibaşı, T. 2012. Tusem Dahiliye. Ankara: Tusem Publisher.

Özkan, Y. (2013). Veri Madenciliği Yöntemleri (2nd ed.). Istanbul: Papatya Publisher.

Polat, K., Şahan, S., & Güneş, S. (2007). A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis. Expert Systems with Applications, 32(4), 1141–1147. doi:10.1016/j.eswa.2006.02.007

Sengur, A. (2008). An expert system based on principal component analysis, artificial immune system and fuzzy -NN for diagnosis of valvular heart diseases. Computers in Biology and Medicine, 38(3), 329–338. doi:10.1016/j.compbiomed.2007.11.004

Sobh, T. S., & Mostafa, W. M. (2011). A cooperative immunological approach for detecting network anomaly. Applied Soft Computing,11(1), 1275–1283. doi:10.1016/j.asoc.2010.03.004

Şahan, S., Polat, K., Kodaz, H., & Güneş, S. (2007). A new hybrid method based on fuzzy-artificial immune system and -nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine, 37(3), 415–423. doi:10.1016/j.compbiomed.2006.05.003

Tjornfelt-Jensen, M., & Hansen, T. K. (1999). Robust solutions to job shop problems. In Proceedings of the 1999 Congress on Evolutionary Computation. CEC 99 (Vol. 2, p. -1144 Vol. 2). doi:10.1109/CEC.1999.782551

Turkish Statistical Institute. (2013). Health Survey 2012. Ankara: Printing Division.

Vos, T., Flaxman, A. D., Naghavi, M., Lozano, R., Michaud, C., Ezzati, M., … Murray, C. J. (2012). Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet, 380(9859), 2163–2196. doi:10.1016/S0140-6736(12)61729-2

Weinstein, S., Obuchowski, N. A. & Lieber, M. L.(2005). Fundamentals of Clinical Research for Radiologists. American Journal of Roentgenology (vol 184 , pp 14 -19).