Automatic identification of Diabetes Diseases using a Modified Artificial Immune Recognition System2 (MAIRS2)

Meryem SAIDI
Biomedical Engineering Laboratory
Tlemcen University - Algeria
Email: miryem.saidi@gmail.com

Mohamed Amine CHIKH
Biomedical Engineering Laboratory
Tlemcen University - Algeria
Email:mea_chihk@mail.univ-tlemcen.dz

Nesma SETTOUTI
Biomedical Engineering Laboratory
Tlemcen University - Algeria
Email:nesma.settouti@gmail.com

***Abstract*—The use of expert systems and artificial intelligence techniques in disease diagnosis has been increasing gradually. Artificial Immune Recognition System 2 (AIRS2) is one of the methods used in medical classification problems. In this paper, we used a Modified AIRS2 (MAIRS2) where we replace the K-nearest neighbors algorithm with the fuzzy K-nearest neighbors to improve the diagnostic accuracy of diabetes diseases. The diabetes disease dataset used in our work is retrieved from UCI machine learning repository. The performances of the AIRS2 and MAIRS2 are evaluated regarding classification accuracy, sensitivity and specificity values. The highest classification accuracy obtained when applying the AIRS2 and AIRS2 with fuzzy K-nn algorithm using 10-fold cross-validation was, respectively 82.69% and 89.10%.**

***Index Terms*—Pima Indians diabetes data set, diagnosis, AIRS, fuzzy k- nearest neighbors.**

I. Introduction

Diabetes is a chronic illness that requires continuing medical care and patient self-management education to prevent acute complications and to reduce the risk of long-term complications. People develop diabetes because the pancreas does not make enough insulin or because the cells do not use insulin properly, or both [1].
A medical diagnosis is a classification process. Using the computer science to perform this classification is becoming more frequent. There is a great variety of methods related to classification and diagnosis of diabetes disease in literature. In [2], a principal component analysis and adaptive neuro-fuzzy inference were used for diagnosing Pima Indian diabetes. They have reported 89.47% classification accuracy. Purnami & al [3] , obtained 93.2% classification accuracies using a new smooth support vector machine and its applications in diabetes disease diagnosis. In [4],Sahan & al. used Attribute Weighted Artificial Immune System with 10-fold cross validation method; they obtained a classification accuracy of 75.87%. Salami & al [5] accomplished 80.00% and 80.65% using CVNN-based CAR and RVNN- based AR. In [6], Exarchos & al. used Automated creation of transparent fuzzy models based on decision trees and obtained a classification accuracy of 75.91%. Ganji & al [7] used a fuzzy Ant Colony Optimization; they have reported 79.48% classification accuracy. In [8], Jayalakshmi & al used the ANN method for diagnosing diabetes, using the Pima Indian diabetes dataset without missing data and obtained 68.56% classification accuracy.

Artificial immune recognition system has shown an effective performance on several problems such as medical classification problems. In [9], Polat & al. used AIRS and an AIRS with fuzzy resource allocation, the classification accuracy obtained using 10-fold cross-validation was, respectively, 98.53% and 99.00% for classification of WBCD; 79.22% and 84.42% for classification of the Pima Indians diabetes data set; and 100% and 92.86% for classification of the ECG arrhythmia data set. Polat, Sahan, Kodaz, and Gunes [10] classified Breast cancer and liver disorders using artificial immune recognition system with fuzzy resource allocation mechanism and 10-fold cross-validation they have reported 98.51% for breast cancer and 83.36% accuracy for the Liver Disorders dataset.

In this work we have proposed a new approach called MAIRS2 to recognize diabetes disease. In the first phase, we used the AIRS2 learning algorithm to reduce the size of the diabetes dataset; we obtained a reduced database named Memory Cells Pool. Since classification is performed in a k-nearest neighbor approach, whose classification time is dependent upon the number of data points used for classifying a previously unseen data item, any reduction in the overall number of evolved memory cells is useful for the algorithm. In the second phase, we apply the fuzzy k-nearest neighbor to overcome the limitations of the k-nn classifier by assigning a class membership to each patient. We evaluate the performances of our MAIRS2 algorithm by using the Pima Indians diabetes dataset.

This paper is organized as follows: Section 2 introduces the used methods: artificial immune recognition systems and fuzzy-Knn algorithms. The results obtained in applications are given in Section 3. In Section 4, we conclude the paper.

II. THEORY

A. Natural and artificial immune system

The Immune System (IS) is a complex of cells, molecules and organs that represent an identification mechanism capable of perceiving and combating dysfunction from our own cells (infectious self) and the action of exogenous infectious microorganisms (infectious nonself) [11].

The vertebrate immune system is particularly interesting due to its several computational capabilities, like: recognition, feature extraction, learning, memory, distributed detection, self-regulation, metadynamics, and immune network [12].

Artificial Immune Systems is the collective name for a number of algorithms inspired by the human immune system. AIS emerged in the 1990s as a new computational research area inspired by theoretical immunology and observed immune functions, principles and models like: clonal selection theory, negative selection theory, positive selection theory and immune network theory. Artificial Immune Systems (AIS) are being used in many applications such as anomaly detection, pattern recognition, data mining, computer security, adaptive control and fault detection [13], [14], [15].

B. Artificial Immune Recognition System 2

Artificial Immune Recognition System (AIRS2), is a supervised learning algorithm that has shown significant success on broad range of classification problems. The immune mechanisms used by AIRS2 are resource competition, clonal selection, affinity maturation and memory cell formation. The terms and concepts used in AIRS are [16], [17]:

- Artificial Recognition Ball (ARB): also known as a B-Cell. It consists of an antibody, a count of the number of resources held by the cell, and the current stimulation value of the cell.
- Candidate Memory Cell: the antibody of an ARB, of the same class as the training antigen, which was the most stimulated after exposure to the given antigen.
- Resources: a parameter which limits the number of ARBs allowed in the system. Each ARB is allocated a number of resources based on its stimulation value and the clonal rate.

C. The AIRS2 algorithm

This algorithm is composed of four main stages [9], [18], [17] :

1) **Initialization**: Normalize all items in the data set such that the Euclidean distance between the feature vectors of any two items is in the range of [0, 1]. Create a random base called the memory pool (M) from training data. Antigenic Presentation: for each antigenic pattern do:
2) **Memory cell identification and ARB generation**: Clone and mutate the highest affinity memory cell and add them to the set of ARBs (P).
3) **Competition for resources and development of a candidate memory cell**: Process each ARBs through the resource allocation mechanism. This will result in some ARB death, and ultimately controls the population. Calculate the average stimulation for each ARB. Clone and mutate a randomly selected subset of the ARBs left in P based in proportion to their stimulation level. While the average stimulation value of each ARB of the same class as the antigen is less than a given stimulation threshold repeat step 3.
4) **Memory cell introduction** : Select the highest affinity ARB from the last antigenic interaction. If the affinity of this ARB with the antigenic pattern is better than that of the previously identified best memory cell mc then add the candidate (mc-candidate) to memory set M. additionally, if the affinity of mcmatch and mc-candidate is below the affinity threshold, and then remove mcmatch from M.

Cycle: Repeat steps 2, 3, 4 until all antigenic patterns have been presented.

The classification is performed in a k-nearest neighbor approach. K-nearest-neighbor (kNN) is a classification algorithm and one of the most important methods in nonparametric algorithm. Each memory cell is iteratively presented with each data item for stimulation. The system's classification of a data item is determined by using a majority vote of the outputs of the k most stimulated memory cells.

There are some problems with K-nn algorithms. One of these problems is that normally each of the neighbors is considered equally important in the assignment of the class label of the input vector. Another problem is that when an input vector is assigned to a class, it does not determine the strength of membership in this class [19].

D. Fuzzy K-nn

Keller et al. [19], propose fuzzy K-nn classifier to overcome the limitations of the K-nn classifier. The fuzzy K-nn algorithm assigns class membership to a sample vector rather than assigning the data to a particular class. The basis of the algorithm is to assign membership as a function of the vector's distance from its k-nearest neighbors and those neighbors' membership in the possible classes.

There are different methods to assign membership for the labeled data. A crisp labeling method is to assign each labeled sample complete membership in its known class and zero membership in all other classes. A fuzzy method is to assign membership to the labeled samples according to a k-nn rule. The following equation 1 assigns class membership to the k-nn of each sample x (say x in class i):

$$u_i(x) = \begin{cases} 0.51 + \left(\frac{n_j}{k}\right) \times 0.49 & \text{if} \quad i = j \\ \left(\frac{n_j}{k}\right) \times 0.49 & \text{if} \quad i \neq j \end{cases} \tag{1}$$

Where n_j is the number of neighbors amongst the k closest labeled reference patterns which are labeled in class j.

The Fuzzy-KNN algorithm is as follows [19], [20]:

Algorithm 1 Fuzzy-KNN Algorithm

Input: x of unknown classification.
  Set k, $1 \leq k \leq n$
  $i = 1$
  while k-nn to x found **do**
    Compute distance from x to x_i
    if $i \leq k$ **then**
      Include x_i in the set of $k - nn$
    else
      if x_i closer to x than any previous nearest neighbor **then**
        Delete the farthest of the $k - nn$
        Include x_i in the set of $k - nn$
      end if
    end if
    $i = i + 1$.
  end while
  $i = 1$
  while x assigned membership in all classes **do**
    Compute 2

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij} \left(\frac{1}{\|x - x_j\|^{\frac{2}{(m-1)}}} \right)}{\sum_{j=1}^{k} \frac{1}{\|x - x_j\|^{\frac{2}{(m-1)}}}} \quad (2)$$

    $i = i + 1$.
  end while

Where u_{ij} is the membership in the i^{th} class of the j^{th} vector of labeled sample set, the parameter m determines how heavily the distance is weighted when calculating the class membership and $\|x - x_j\|$ is the distance between x and its j^{th} nearest neighbor x_j. The pattern x is assigned to the class given by :$argmax_{i=1}^{nc}(u_i(x))$.

III. Experimentation and results

A. Pima Indians diabetes data set

The proposed method has been tested using the public Pima Indian Diabetes dataset of National Institute of Diabetes and Digestive and Kidney Diseases. Pima Indians of Arizona have the highest prevalence and incidence of diabetes Type 2 of any population in the world. This dataset contains information on various medical measurements on 768 individuals, 500 of these samples belong to persons with no diabetes problem while the remaining 268 sample are of persons with diabetes. Noting that there are some records with missing data. After removing these cases with unreasonable physical data, the total number of cases is 392 where 262 are normal cases and 130 are diabetes cases. The class information contained in this data set is given by 0 for healthy persons and by 1 for diabetic patients. The number of attributes in samples is 8.

B. Results and discussion

In this study, diabetes disease diagnosis was conducted using a hybrid medical decision support system based on MAIRS2 (Modified AIRS2). The experimental study was performed with 392 sample (262 healthy, 130 diabetics) using 10-fold cross validation method.

The experiments were performed to generate the optimum number of memory cells that will be used for classification purposes using fuzzy K-nn algorithm. We have performed several trials in order to obtain the optimum memory cells set that give highest classification accuracy. The best classification accuracy (89.10%) was obtained with a data reduction of 28.57%.
The relation between data reduction, classification accuracy, sensitivity and specificity values is shown in Table 1, 2 and 3. Table 1 presents the results obtained using the AIRS2 learning algorithm with the k-nn method. Tables 2 and 3, present the results obtained using the MAIRS2 with respectively the crisp and the fuzzy labeling methods.

MC set	Data reduction	Accuracy	Sensitivity	Specificity
132	66.32%	69.14%	58.97%	74.40%
280	28.57%	82.69%	73.47%	87.55%
272	30.61%	81.16%	73.46%	84.35%
248	36.73%	81.62%	70.79%	88.25%

TABLE 1
THE CLASSIFICATION ACCURACIES, SENSITIVITY AND SPECIFICITY VALUES FOR THE AIRS2

In Tables 1, AIRS2 showed an accuracy of 82.69% using 280 memory cells and a data reduction of 28.57%, whereas in Table 2 and 3, MAIRS2 with the crisp labeling method and MAIRS2 with the fuzzy labeling method obtained respectively 89.10% and 84.51% using the same memory cells pool. With 248 memory cells, we have a data reduction of 36.73% and an accuracy of 81.62%, 86.26% and 84.96% using respectively AIRS2, MAIRS2 with crisp labeling method and MAIRS2 with fuzzy labeling method.

MC set	Data reduction	Accuracy	Sensitivity	Specificity
365	6.88%	91.40%	85.85%	94.90%
132	66.32%	72.24%	61.19%	76.29%
280	28.57%	89.10%	85.18%	91.50%
282	30.61%	85.56%	85.56%	85.44%
248	36.73%	86.26%	76.78%	91.89%

TABLE 2
THE CLASSIFICATION ACCURACIES, SENSITIVITY AND SPECIFICITY VALUES FOR THE MAIRS2 (CRISP METHOD).

The medical decision support system present in literature for classification of Pima Indian Diabetes dataset use the complete dataset, Table 4 presents the obtained classification accuracies of these classifiers. It can be shown from these results that the obtained classification accuracy by combination of AIRS2 and fuzzy-Knn classifier for diabetes disease was among the best classifier report from literature. Whereas our method couldn't reach the highest classification accuracy for the problem, it has been over-performed to another AIS algorithm: AWAIS, AIRS

MC set	Data reduction	Accuracy	Sensitivity	Specificity
365	6.88%	86.75%	75.50%	93.50%
132	66.32%	73.78%	61.23%	81.36%
280	28.57%	84.51%	74.57%	89.90%
282	30.61%	86.76%	81.67%	89.46%
248	36.73%	84.96%	73.12%	90.94%

TABLE 3
THE CLASSIFICATION ACCURACIES, SENSITIVITY AND SPECIFICITY VALUES FOR THE MAIRS2 (FUZZY METHOD).

and AIRS with fuzzy resource allocation mechanism which obtained respectively 75.87%, 79.22%, 84.42% classification accuracy.

Methods	Accuracy
smooth SVM	93.20%
MAIRS2	89.10%
AIRS with fuzzy resource allocation mechanism	84.42%
AIRS2	82.69%
CVNN-based CAR	81.00%
fuzzy Ant Colony Optimization	79.48%
AIRS	79.22%
AWAIS	75.87%
ANN	68.56%

TABLE 4
THE CLASSIFICATION ACCURACIES, SENSITIVITY AND SPECIFICITY VALUES FOR THE MAIRS2 (FUZZY METHOD).

IV. CONCLUSION

Classification systems that are used in clinical diagnosis allow medical data to be examined in a shorter time and in more detail. In the research reported in this paper, MAIRS2 was applied to recognize diabetes disease.We noted that an increase in the number of memory cells led to a better recognition rate but in other hand, it diminishes the degree of data reduction. In this work, we have achieved a good tradeoff between classification accuracy and data reduction. Thus the combination of AIRS2 and fuzzy K-nn led us to a new system (MAIRS2) that performed better than classical AIRS2 implemented in this paper and others AISs algorithms cited in literature. Finally, we hope that the use of others classification techniques and affinity measure methods such as Manhattan distance increase the recognition rate of the diabetes disease.

REFERENCES

[1] *National Diabetes Information Clearinghouse (NDIC). http://diabetes.niddk.nih.gov/dm/pubs/diagnosis/*, Std.

[2] K. Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digital Signal Processing*, p. 702710, 2007.

[3] S. Purnami, A. Embong, J. Zain, and S. Rahayu, "A new smooth support vector machine and its applications in diabetes disease diagnosis," *Journal of Computer Science*, p. 10031008, 2009.

[4] S. Sahan, K. Polat, H. Kodaz, and S. Gunes, "The medical applications of attribute weighted artificial immune system (awais): Diagnosis of heart and diabetes diseases," in *ICARIS*, 2005, p. 456 468.

[5] M. Salami, A. Shafie, and A. Aibinu, "Application of modeling techniques to diabetes diagnosis," in *IEEE EMBS Conference on Biomedical Engineering & Sciences*, 2010.

[6] T. Exarchos, D. Fotiadis, and M. Tsipouras, "Automated creation of transparent fuzzy models based on decision trees application to diabetes diagnosis," in *30th Annual International IEEE EMBS Conference*, 2008.

[7] M. Ganji and M. Abadeh, "Using fuzzy ant colony optimization for diagnosis of diabetes disease," *IEEE*, 2010.

[8] T. Jayalakshmi and A. Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks," in *International Conference on Data Storage and Data Engineering*, 2010.

[9] K. Polat and S. Gunes, "An improved approach to medical data sets classification: artificial immune recognition system with fuzzy resource allocation mechanism," *Expert Systems*, pp. 252–270, 2007.

[10] K. Polat, S. Sahan, H. Kodaz, and S. Gunes, "Breast cancer and liver disorders classification using artificial immune recognition system (airs) with performance evaluation by fuzzy resource allocation mechanism," *Expert Systems with Application*, p. 172183, 2007.

[11] L. De Castro and F. Von Zuben, "Artificial immune systems: part i basic theory and applications," DCA 01/99, Tech. Rep., 1999.

[12] J. Timmis, T. Knight, L. De Castro, and E. Hart, "An overview of artificial immune systems," in *Computation in Cells and Tissues: Perspectives and Tools for Thought*, M. H. R. Paton, H. Bolouri and J. Parish, Eds. Natural Computation Series, 2004, pp. 51–86.

[13] J. Greensmith and U. Whitbrook, A.and Aickelin, *Artificial Immune Systems, Handbook of Metaheuristics*, 2010, ch. 13, pp. 421–448.

[14] L. De Castro and F. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 239–251, 2002.

[15] U. Aickelin and D. Dasgupta, *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*, 2005, ch. ARTIFICIAL IMMUNE SYSTEMS, pp. 375–399.

[16] A. Watkins and J. Timmis, "Artificial immune recognition system (airs): Revisions and refinements," in *1st International Conference on Artificial Immune Systems (ICARIS2002)*, 2002.

[17] ——, "Artificial immune recognition system (airs): An immune-inspired supervised learning algorithm," *Genetic Programming and Evolvable Machines*, p. 291317, 2004.

[18] A. Watkins and L. Boggess, "A new classifier based on resource limited artificial immune systems," in *IEEE World Congress on Computational Intelligence*, 2002, pp. 1546–1551.

[19] J. Keller, M. Gray, and J. J. Givens, "A fuzzy k-nearest neighbor algorithm," *Syst. Man Cybern*, p. 580585, 1985.

[20] A. Sengur, "An expert system based on principal component analysis, artificial immune system and fuzzy k-nn for diagnosis of valvular heart diseases." *Computers in Biology and Medicine*, p. 329 338, 2008.