Research Article

Hepatitis Disease Diagnosis Using Hybrid Case Based Reasoning and Particle Swarm Optimization

Mehdi Neshat,[1] Mehdi Sargolzaei,[2] Adel Nadjaran Toosi,[3] and Azra Masoumi[1]

[1] *Department of Computer Science, Shirvan Branch, Islamic Azad University, Shirvan 91738, Iran*
[2] *Department of Computer Engineering, Shirvan Branch, Islamic Azad University, Shirvan 92457, Iran*
[3] *Department of Computer Science and Software Engineering, Shirvan Branch, Islamic Azad University, Shirvan 92174, Iran*

Correspondence should be addressed to Mehdi Neshat, neshat_mehdi@ieee.org

Correct diagnosis of a disease is one of the most important problems in medicine. Hepatitis disease is one of the most dangerous diseases that affect millions of people every year and take man's life. In this paper, the combination of two methods of PSO and CBR (case-based reasoning) has been used to diagnose hepatitis disease. First, a case-based reasoning method is workable to preprocess the data set therefore a weight vector for every one feature is extracted. A particle swarm optimization model is then practical to assemble a decision-making system based on the selected features and diseases recognized. Many researchers have tried to have a more accurate diagnosis of the disease through the use of various methods. The data used has been taken from the site UCI called hepatitis disease. This database has 155 records and 19 fields. This method was compared with five other classification methods and given the results of the proposed method (CBR-PSO), better results were achieved. The proposed method could diagnose hepatitis disease with the accuracy of 93.25%.

1. Introduction

Hepatitis refers to inflammation of the liver parenchyma and can be created for various reasons. Some of them are contagious and some of them are not. Among the factors creating hepatitis, it can be referred to the excess in alcohol consumption, the effects of some medications, and infection with bacteria and also viruses. Viral hepatitis results in liver infection. The cause of the viral hepatitis disease is a virus, and initially, it can appear like a cold. But unlike a common cold, due to liver failure and difficulty in treatment, chronic "C" hepatitis disease can threaten the patient's life. Most of those suffering hepatitis kinds C and B have no symptoms. Some of these patients show symptoms of viral infection in nature such as, fatigue, stomach ache, muscle pain, and nausea, and loss of appetite. But symptoms of liver failure occur in advanced cases including swelling of the abdomen and limbs, jaundice, and digestive bleedings. More than 3% percent of the individuals are infected with the virus in Iran.

A lot of researchers have recently used computational intelligence in diagnosing different diseases. All these intelligent techniques can only help the physician's diagnosis as an assistant and all have a small amount of error. Among these methods, neural networks are most widely used. Different kinds of neural networks with various specifications have been used in diagnosing diseases [1]. A lot of researches have been done through neural networks and fuzzy system for diagnosis of B hepatitis disease [2, 3].

Methods with better classification accuracy will provide more sufficient information to identify the potential patients and to improve the diagnosis accuracy. Meta-heuristic algorithms (like genetic algorithms, particle swarm optimizations, fish swarm optimization, and Tabu Search) and data mining tools (neural network and decision tree) have been applied in this area. Aside from other traditional classification problems, medical data classifications are further applied in disease diagnosis. Therefore, patients or doctors not only need to know the answer (classification result), they also need to know the symptoms that derive this answer. As for other clinical diagnosis problems, classification systems have been used for hepatitis disease diagnosis problem. When the studies in the literature related with this classification application are examined, it can be seen that a great variety of methods were used which reached high

TABLE 1: Classification accuracies for hepatitis disease classification problem.

Authors	Method	Classification accuracy (%)	Reference
Duch and Grudzinski	Weighted 9 NN	92.9	[19]
	18 NN, stand, manhattan	90.2	
	15 NN, stand, Euclidean	89.0	
Duch et al.	FSM with rotational	89.7	[20]
	FSM without rotational	88.5	
	RBF(tooldiag)	79	
	MLP+BP(tooldiag)	77.4	
Ster and Dobnikar	LDA	86.4	[21]
	Naïve Bayes and semi NB	86.3	
	QAD	85.8	
	1-NN	85.3	
	ASR	85	
	Fisher discriminant analysis	84.5	
	LVQ	83.2	
	CART(decision tree)	82.7	
	MLP with BP	82.1	
	ASI	82.0	
	LFC	81.9	
Jankowski	IncNET	86.0	[22]
Ozyilmaz and Yildirim	MLP	74.37	[23]
	RBF	83.75	
	GRNN	80.0	
Bascil and Temurtas	MLNN with Levenberg Marquardt	91.87	[24]

Classification accuracies using the dataset taken from UCI machine learning repository.

Table 1 is a review of different methods for diagnosis of hepatitis disease. Different methods of neural networks and their combination with other methods have achieved good results.

Liao [4] investigated of a hybrid CBR method for failure mechanisms identification. Yang et al. [5] integrated CBR with an ART-Kohonen NN to enhance fault diagnosis of electric motors. Hua Tan et al. [6] integrated CBR and the fuzzy ARTMAP NN to support managers in making timely and optimal manufacturing technology investment decisions. Saridakis and Dentsoras [7] introduced a case-based design with a soft computing system to evaluate the parametric design of an oscillating conveyor. Hybrid CBR has also been used in the medical planning and application areas. Guiu et al. [8] introduced a case-based classifier system to solve the automatic diagnosis of Mammary Biopsy Images. Hsu and Ho [9] combined the CBR, NN, fuzzy theory, and induction theory together to facilitate multiple-disease diagnosis and the learning of new adaptation knowledge. Wyns et al. [10] applied a modified Kohonen mapping combined with a CBR evaluation criterion to predict early arthritis, including rheumatoid arthritis and spondyloarthropathy. Ahn and Kim [11] combined the CBR with genetic algorithms to evaluate cytological features derived from a digital scan of

breast fine needle aspirate (FNA) slides. Panchal et al. [12] use CBR and wave of swarm (WOS) derived from PSO to detect ground water potential. In addition, hybrid CBRs have been used in the financial forecasting areas. Kim and Han [13] presented a case-indexing method of CBR which utilizes SOM for the prediction of corporate bond rating. Li et al. [14] introduced a feature-based similarity measure to deal with financial distress prediction (e.g., bankruptcy prediction) in China. Chang and Lai [15] integrated the SOM and CBR for sales forecasts of newly released books. Chang et al. [16] evolved a CBR system with genetic algorithm for wholesaler returning book forecasting. Chun and Park [17] devised a regression CBR for financial forecasting, which applies different weights to independent variables before finding similar cases. Kumar and Ravi [18] presented a comprehensive review of the works utilizing NN and CBR to solve the bankruptcy prediction problems faced by banks.

Following, the data used are explained (Section 2). In Section 3, the methods used for combination of CBR and PSO are stated. In Section 4, the experimental research and finally discussion and conclusion will be dealt with.

2. Data

This hepatitis disease dataset requires determination of whether patients with hepatitis will either live or die. It was

donated by Jozef Stefan Institute, Yugoslavia. The used data source in this study was taken from UCI machine learning repository. The purpose of the dataset is to predict the presence or absence of hepatitis disease given the results of various medical tests carried out on a patient. This database contains 19 attributes, which have been extracted from a larger set of 155. Hepatitis dataset contains 155 samples belonging to two different classes (32 "die" cases, 123 "live" cases). There are 19 attributes, 13 binary, and 6 attributes with 6–8 discrete values. Attributes of symptoms that is obtained from patient are as follows (UCI Machine Learning Repository):

(1) Age: 10, 20, 30, 40, 50, 60, 70, 80

(2) Sex: male, female

(3) Steroid: no, yes

(4) Antivirals: no, yes

(5) Fatigue: no, yes

(6) Malaise: no, yes

(7) Anorexia: no, yes

(8) Liver Big: no, yes

(9) Liver Firm: no, yes

(10) Spleen Palpable: no, yes

(11) Spiders: no, yes

(12) Ascites: no, yes

(13) Varices: no, yes

(14) Bilirubin: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00

(15) Alk phosphate: 33, 80, 120, 160, 200, 250

(16) Sgot: 13, 100, 200, 300, 400, 500

(17) Albumin: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0

(18) Protime: 10, 20, 30, 40, 50, 60, 70, 80, 90

(19) Histology: no, yes.

3. Method

In this research, the combination methods of CBR (Case base weighted cluster algorithm) for clustering and PSO for classifying have been used. This algorithm first partitions the data in relatively large number of clusters. Then, primary conditions are used for reduction of the number of clusters into 2 (two main groups, healthy individuals and patients) [25].

3.1. PSO Clustering. As K means algorithm, the number of clusters has to be decided first. For classification problem, suppose we have N kinds of classes and in PSO-clustering, we try to find K clusters corresponding to N classes. For

traditional PSO-clustering problem, the objective function is defined as [26]

$$\min_{P_K \in \Gamma_K} \sum_{i=1}^{K} \sum_{X_l \in C_i} \text{dist}\left(X_l, \overline{X}\right), \qquad (1)$$

$$\overline{X} = \text{Centroid of each cluster } C_i,$$

$$\text{dist}\left(X_l, \overline{X}\right) = \sqrt{\sum_{i=1}^{n} (x_{li} - \overline{x}_i)^2}, \quad X_l = (x_{l1}, x_{l2}, \ldots, x_{ln}),$$

$$\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n), \qquad (2)$$

where n is the number of attributes.

The following diagram is the pseudocode of PSO-clustering algorithm. Figure 1 shows the concept about the pseudocode and Figure 2 shows the example about distance measurement used in one particle while K is set as five.

PSO Clustering.

Input: hepatitis disease dataset K: number of classes

Output: Classification Result (the location of K centroids)

Procedure PSO Clustering (data, K)

Generate P solutions (particles);

each solution has its own K centroids selected randomly from data set.

For each particle

Objective function = $\min_{P_K \in \Gamma_K} \sum_{i=1}^{K} \sum_{X_l \in C_i} \text{dist}(X_l, \overline{X})$,

$v_{id} = w \cdot v_{id} + c_1 \cdot rand\,1 \cdot (p_{id} - x_{id}) + c_2 \cdot rand\,2() \cdot (p_{gd} - x_{id})$

$x_{id} = x_{id} + v_{id}$

Update p_{id}

End

Update p_{gd}

End.

In this study, we apply the weights of each attribute and the Euclidean distance in objective function which can be modified as following:

$$\text{dist}\left(X_l, \overline{X}\right) = \sqrt{\sum_{i=1}^{m} w_i \cdot (x_{li} - \overline{x}_i)^2},$$

$$X_l = (x_{l1}, x_{l2}, \ldots, x_{lm}), \quad \overline{X} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_m). \qquad (3)$$

In this study, the weights of each attribute will be calculate by a case-base reasoning algorithm, the detail description will be in the next part.
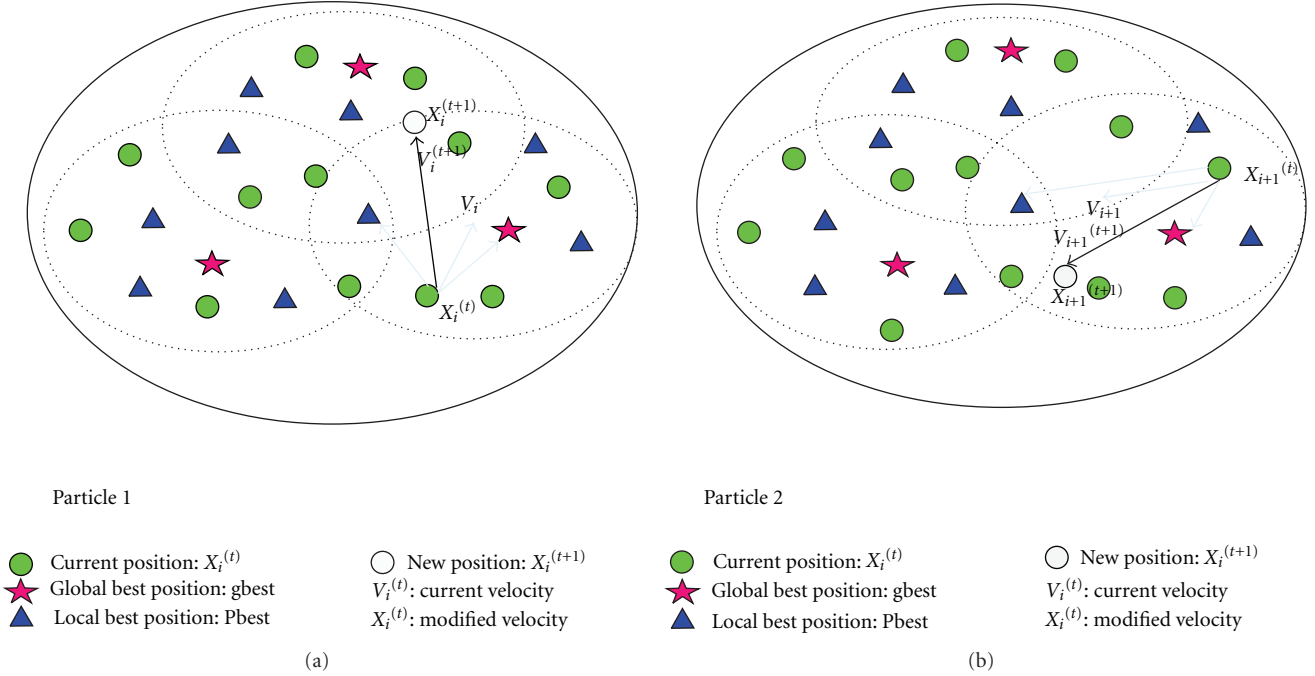
Particle 1

● Current position: $X_i^{(t)}$ ○ New position: $X_i^{(t+1)}$
★ Global best position: gbest $V_i^{(t)}$: current velocity
▲ Local best position: Pbest $X_i^{(t)}$: modified velocity

(a)

Particle 2

● Current position: $X_i^{(t)}$ ○ New position: $X_i^{(t+1)}$
★ Global best position: gbest $V_i^{(t)}$: current velocity
▲ Local best position: Pbest $X_i^{(t)}$: modified velocity

(b)

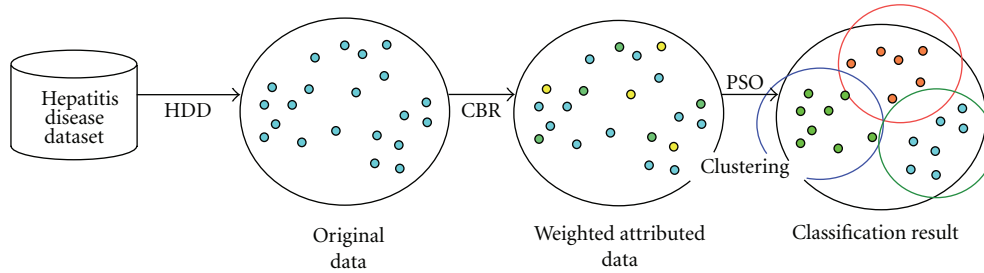FIGURE 1: An example of two particles in PSO clustering.

FIGURE 2: A concept model for CBRPSO.

3.2. Hybrid PSO Clustering with CBR

Procedure Weights Calculated by CBR [25].

Initialize weight of each attributes j in each data with random values in $[0, 1]$;

Do

Compute $\Delta w_j = -\eta \dfrac{\partial E}{\partial w_j}$; //formula (5)

Update $w_j = w_j + \Delta w_j$

While not convergent

Assign each attribute j has its own weight;

End.

The concept of CBRPSO is shown in Figure 2 and it can be divided into four major steps. They are

(1) Screening medical database from UCI data set;

(2) Using CBR to find the weighted feature value from indices;

(3) Establishing PSO classification model; and finally

(4) out-putting the classification results.

The CBR algorithm calculates weights of each attributes; hence the pseudocode of CBR-PSO clustering can be modified as following two diagrams [25].

Input:

N: data points

K: number of classes (the same with number of cluster' centroids)

M: temporary centroids ($M > K$, for initial)

W: weights calculated by CBR

Procedure Stepwise Centroids PSO Clustering with CBR

M: =Weighted PSO Clustering (N, $|M|$, W);

Reassign M as data points (N: =M);

Reduce number of M to M'

Recursive execute

Stepwise Centroids PSO Clustering with CBR until M' equals to K;

//means Re-cluster the M data points into M' clusters, if M' equals to K,

then final result is found

Return K centroids;

End;

Var:

j: attribute of dataset

d: dimension of each data (number of attributes)

Input: data: hepatitis disease dataset

K: number of classes

Output: Classification Result (the location of K centroids)

Procedure Weighted PSO Clustering (data, K, weights)

Generate P solutions (particles); //each solution has its own

K centroids selected randomly from dataset.

For each particle

Objective function =
$\min_{P_K \in \Gamma_K} \sum_{i=1}^{K} \sum_{X_l \in C_i} \sqrt{\sum_{j=1}^{d} w_i \cdot (x_{li} - c_i)^2}$

$v_{\text{id}} = w \cdot v_{\text{id}} + c_1 \cdot rand\,1 \cdot (p_{\text{id}} - x_{\text{id}}) + c_2 \cdot rand\,2()$
$\quad \cdot (p_{\text{gd}} - x_{\text{id}})$

$x_{\text{id}} = x_{\text{id}} + v_{\text{id}}$

Update p_{id}

End

Update p_{gd}

End.

4. Experimental Results

According to Section 2, the data used in this research have been taken from UCI. This database has 19 fields and 155 samples. In addition, the results of this method are compared with other modern methods. The method CBR-PSO has been also widely used in other medical data and has had good performance.

Table 2: Accuracy comparisons of different PSO approaches used in this research.

Hepatitis disease	PSO	CBR PSO
Best	89.46%	**94.58**%
Average	82.66%	92.83%
Lowest	75.35%	77.16%

*Average diagnosis accuracy (500 running times).

In Table 2, the efficiency of CBR-PSO method was compared with PSO method. CBR-PSO method could diagnose hepatitis disease in the best state with the accuracy of 94.58%, but PSO method could diagnose this disease in the best state with the accuracy of 89.46%. The overall function of CBR-PSO method is better in relation to PSO method in the average state and has higher efficiency.

In order to investigate the function of CBR-PSO method better, it was compared also with methods of KNN, Naïve Bayes, SVM, and FDT. Table 3 shows the comparison of this method with four important methods of classification.

Given Table 3, the best results have been achieved through CBR-PSO method, and SVM method could diagnose hepatitis disease in the best state with the accuracy of 90.31%. The methods of NB, KNN, and FDT received the third to fifth grade, respectively. Various methods have been investigated for diagnosis of hepatitis disease, and each has advantages and disadvantages. Among these, CBR-PSO method could obtain the best results.

75% of the data wea randomly chosen for training while 25% of these data is chosen for testing for these models with a total number of 500 execution times. In addition, as shown in Tables 2 and 3, the CBRPSO is also compared with other approaches developed in the literature to show the effectiveness of our approach.

5. Conclusion

Diagnosing disorders and diseases is one of the most difficult physician's responsibilities. An incorrect diagnosis can endanger a man's life and cause his death. In this regard, the use of different methods of artificial intelligence and expert system has become common and it is tried to minimize the error amount of these methods. In this paper, the combination of two methods of PSO and CBR has been used to diagnose the dangerous hepatitis disease. First through the use of CBR method, a preprocessing is done on the data considered, and the weight of the effect of each field in diagnosis is extracted, and then clustering is done through PSO method. PSO is responsible for determining being patient or not, being patient of each record and specifies to which class each record belongs. CBR-PSO method was compared with different methods, such as, FDT, KNN, SVM, PSO and Naïve Bays and could diagnose hepatitis disease in the best state with the accuracy of %94.58. This method has had better

TABLE 3: Accuracy comparisons of different forecasting models in hepatitis disease.

Hepatitis disease	KNN	Naïve Bayes	SVM	FDT	CBR-PSO
Best	89.86%	86.35%	90.31%	78.15%	**94.58**%
Average	83.45%	82.05%	86.92%	75.39%	92.83%
Lowest	70.29%	66.94%	65.22%	61.49%	77.16%

function in comparison with different methods. The combination of this method and fuzzy logic and its use in medical data will be among future the authors' works researches.

References

[1] P. J. G. Lisboa, E. C. Ifeachor, and P. S. Szczepaniak, *Artificial Neural Networks in Biomedicine*, Springer, London, UK, 2000.

[2] M. Neshat and M. Yaghobi, "FESHDD: fuzzy expert system for hepatitis B diseases diagnosis," in *Proceedings of the 5th International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control (ICSCCW '09)*, Cyprus, September 2009.

[3] M. Neshat and M. Yaghobi, "Designing a fuzzy expert system of diagnosing the hepatitis B intensity rate and comparing it with adaptive neural network fuzzy system," in *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco, Calif, USA, October 2009.

[4] T. W. Liao, "An investigation of a hybrid CBR method for failure mechanisms identification," *Engineering Applications of Artificial Intelligence*, vol. 17, no. 1, pp. 123–134, 2004.

[5] B. S. Yang, T. Han, and Y. S. Kim, "Integration of ART-Kohonen neural network and case-based reasoning for intelligent fault diagnosis," *Expert Systems with Applications*, vol. 26, no. 3, pp. 387–395, 2004.

[6] K. Hua Tan, C. Peng Lim, K. Platts, and H. Shen Koay, "An intelligent decision support system for manufacturing technology investments," *International Journal of Production Economics*, vol. 104, no. 1, pp. 179–190, 2006.

[7] K. M. Saridakis and A. J. Dentsoras, "Case-DeSC: a system for case-based design with soft computing techniques," *Expert Systems with Applications*, vol. 32, no. 2, pp. 641–657, 2007.

[8] J. M. Garrell I Guiu, E. Golobardes I Ribé, E. Bernadó I Mansilla, and X. Llorà I Fàbrega, "Automatic diagnosis with genetic algorithms and case-based reasoning," *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp. 367–372, 1999.

[9] C. C. Hsu and C. S. Ho, "A new hybrid case-based architecture for medical diagnosis," *Information Sciences*, vol. 166, no. 1–4, pp. 231–247, 2004.

[10] B. Wyns, L. Boullart, S. Sette, D. Baeten, I. Hoffman, and F. De Keyser, "Prediction of arthritis using a modified Kohonen mapping and case based reasoning," *Engineering Applications of Artificial Intelligence*, vol. 17, no. 2, pp. 205–211, 2004.

[11] H. Ahn and K. J. Kim, "Global optimization of case-based reasoning for breast cytology diagnosis," *Expert Systems with Applications*, vol. 36, no. 1, pp. 724–734, 2009.

[12] V. K. Panchal, H. Kundra, and N. Kaur, "A novel approach of waves of Swarm with case based reasoning to detect ground water potential," *Journal of Technology and Engineering Sciences*, vol. 1, pp. 3–8, 2009.

[13] K. S. Kim and I. Han, "The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases," *Expert Systems with Applications*, vol. 21, no. 3, pp. 147–156, 2001.

[14] H. Li, J. Sun, and B. L. Sun, "Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors," *Expert Systems with Applications*, vol. 36, no. 1, pp. 643–659, 2009.

[15] P. C. Chang and C. Y. Lai, "A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting," *Expert Systems with Applications*, vol. 29, no. 1, pp. 183–192, 2005.

[16] P. C. Chang, C. Y. Lai, and K. R. Lai, "A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting," *Decision Support Systems*, vol. 42, no. 3, pp. 1715–1729, 2006.

[17] S. H. Chun and Y. J. Park, "A new hybrid data mining technique using a regression case based reasoning: application to financial forecasting," *Expert Systems with Applications*, vol. 31, no. 2, pp. 329–336, 2006.

[18] P. Ravi Kumar and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review," *European Journal of Operational Research*, vol. 180, no. 1, pp. 1–28, 2007.

[19] W. Duch and K. Grudzinski, "Ensembles of similarity-based models," in *Proceedings of the Intelligent Information Systems*, 2001.

[20] W. Duch, R. Adamczak, and G. H. F. Diercksen, *Neural Networks from Similarity Based Perspective*, Department of Computer Methods, Nicholas Copernicus University, 2000.

[21] B. Ster and A. Dobnikar, "Neural networks in medical diagnosis: comparison with other methods," in *Proceedings of the International Conference EANN*, pp. 427–430, 1996.

[22] N. Jankowski, "Approximation and classification in medicine with IncNet neural networks," in *Proceedings of the Workshop on Machine Learning in Medical Applications*, pp. 53–58, Greece, 1999.

[23] L. Ozyilmaz and T. Yildirim, "Artificial neural networks for diagnosis of hepatitis disease," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '03)*, vol. 1, pp. 586–589, Portland, Ore, USA, 2003.

[24] M. S. Bascil and F. Temurtas, "A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt training algorithm," *Journal of Medical Systems*, vol. 35, no. 3, pp. 433–436, 2011.

[25] P.-C. Chang, J.-J. Lin, and C.-H. Liu, "An attribute weight assignment and particle swarm optimization algorithm for medical database classifications," *Computer Methods and Programs in Biomedicine*. In press.

[26] D. W. van der Merwe and A. P. Engelbrecht, "Data clustering using particle swarm optimization," in *Proceedings of the Congress on Evolutionary Computation*, pp. 215–220, 2003.