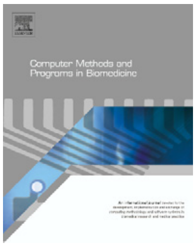

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

An attribute weight assignment and particle swarm optimization algorithm for medical database classifications

Pei-Chann Chang[a,*], *Jyun-Jie Lin*[a], *Chen-Hao Liu*[b]

[a] *Department of Information Management, Yuan Ze University, Taoyuan 32026, Taiwan, ROC*
[b] *Department of Information Management, Kai-Nan University, Taoyuan 33857, Taiwan, ROC*

ARTICLE INFO

Article history:
Received 8 May 2010
Received in revised form
6 September 2010
Accepted 6 December 2010

Keywords:
Hybrid intelligence
Case base reasoning
Medical decision making
Particle swarm optimization

ABSTRACT

In this research, a hybrid model is developed by integrating a case-based reasoning approach and a particle swarm optimization model for medical data classification. Two data sets from UCI Machine Learning Repository, i.e., Liver Disorders Data Set and Breast Cancer Wisconsin (Diagnosis), are employed for benchmark test. Initially a case-based reasoning method is applied to preprocess the data set thus a weight vector for each feature is derived. A particle swarm optimization model is then applied to construct a decision-making system for diseases identified. The PSO algorithm starts by partitioning the data set into a relatively large number of clusters to reduce the effects of initial conditions and then reducing the number of clusters into two. The average forecasting accuracy for breast cancer of CBRPSO model is 97.4% and for liver disorders is 76.8%. The proposed case-based particle swarm optimization model is able to produce more accurate and comprehensible results for medical experts in medical diagnosis.

1. Introduction

Recently, the incorporation of computational intelligence in medical diagnosis is a new tendency and with a large number of medical applications. Many of the medical diagnosis procedures can be categorized as intelligent data classification tasks. These classification procedures can be divided into two types, with regard to the number of categories that each time is classified. The first classification type separates the data between only two classes (known as binary classification or two-class task), and the second type classifies the data between more than two classes (multi-class task). For example, there are methods for intelligent classification that handle efficiently the two-class task such as the Ada Boost and the support vector machines. Any multi-class problem can be substituted by more than one or two-class problems. Such an approach is to build independent classification rules for each of the classes and then run these competitive rules simultaneously [1].

In Medical area, many researchers have tried to use different methods to improve the accuracy of data classification. 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, and Tabu Search) and data mining tools (neural network and decision tree) have been applied in this area and have derived significant results as described in refs. [2,3]. Chang et al. [4] employ hybrid heuristics in breast cancer classification and achieve more than 90% accuracy rate. Other researchers as in refs. [5,6] also apply hybrid neural network and Boolean

[*] Corresponding author. Tel.: +886 936101320; fax: +886 34638884.
E-mail address: iepchang@saturn.yzu.edu.tw (P.-C. Chang).

0169-2607/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2010.12.004

rules to derive results which exhibiting good performance and reduced number of rules with relevant input variables. However, in liver disorders classification [7–9], the classification accuracy of current methods is still low and insignificant enough to be adopted in practical applications.

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. Neural networks [5–6] and linear programming models [10] have been reported and these models almost obtain high classification accuracy rate. However, their decision process is essentially a black box, with no explanation as to how they were attained. Hybrid heuristic methods like GA or neural networks combining with fuzzy rules will handle this problem caused by black box approaches, but they still cannot identify which input factors are more significant than the others.

In this research, our contribution is to develop a hybrid model for medical data classification in two medical domains: breast cancer diagnosis and the classification of a liver disorder diagnosis. This hybrid model combines the soft computing techniques including a case based approach and a particle swarm optimization tool evolved by genetic algorithms. The proposed model is able to classify the breast cancer and liver disorder data more precisely and offer medical doctor a better information platform during the diagnosis of a breast cancer or liver disorder patient. The medical data classification is conducted by extracting and analyzing available data with an appropriate clustering procedure and a PSO model. The clustering procedure collect groups of similar data in term of medical data profile (breast cancer diagnosis or the classification of a liver disorder diagnosis) and weight of each input variables. Therefore, the importance of each factor (input variable) will be identified through this procedure. In addition, the evolved PSO tends to discover hidden knowledge between these input variables with each cluster and the classification of breast cancer or liver disorder diagnosis.

This paper is organized as follows. A literature review for medical classification problems is introduced in Section 2. Section 3 detailing the method and algorithm of a case based PSO model for medical classification problems, while Section 4 presents the experimental results. Section 5 provides the conclusions and future directions of researches.

2. Literature review

In general, machine learning techniques applied in computer-aided medical diagnosis can be categorized into two classes, i.e., symbolic or connectionist learning techniques. Symbolic learning techniques such as rule induction are usually regarded as comprehensible techniques because the learned knowledge is expressed in forms such as production rules that are easy to be understood by the user. Rule induction has already been widely applied in medical diagnosis [11–13]. On the other hand, most connectionist learning techniques such as artificial neural networks are regarded as incomprehensible techniques because the learned knowledge is concealed in a lot of connections and is not transparent to the user. Although artificial neural networks have already been tried in several medical tasks [14], they have not yet been widely accepted in medicine [15]. Fortunately, during the last decade much work has addressed the issue of improving the comprehensibility of artificial neural networks [16,17] and some results have already been applied to medical tasks [5,13,18]. Rule induction for medical diagnosis is more easily accepted if the diagnostic process can be checked by the doctor and adequately explained to the patients.

Recently, the hybrid CBR techniques have been widely applied in various applications including manufacturing planning, fault diagnosis, knowledge modeling and management, and medical diagnosis and application. Hui et al. [19] integrated NN, CBR, and rule-based reasoning to support customer service activities, such as decision support and machine fault diagnosis in a manufacturing environment. Liao [20] integrated a CBR method with a multi-layer perception for the automatic identification of failure mechanisms in the entire failure analysis process. Yang et al. [21] integrated CBR with an ART-Kohonen NN to enhance fault diagnosis of electric motors. Hua Tan et al. [22] integrated CBR and the fuzzy ARTMAP NN to support managers in making timely and optimal manufacturing technology investment decisions. Saridakis et al. [23] 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 and co-workers [24] introduced a case-based classifier system to solve the automatic diagnosis of Mammary Biopsy Images. Hsu et al. [25] 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. [26] applied a modified Kohonen mapping combined with a CBR evaluation criterion to predict early arthritis, including rheumatoid arthritis and spondyloarthropathy. Ahn et al. [27] 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. [28] 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 et al. [29] presented a case-indexing method of CBR which utilizes SOM for the prediction of corporate bond rating. Li et al. [30] introduced a feature-based similarity measure to deal with financial distress prediction (e.g., bankruptcy prediction) in China. Chang et al. [31] integrated the SOM and CBR for sales forecasts of newly released books. Chang et al. [32] evolved a CBR system with genetic algorithm for wholesaler returning book forecasting. Chun et al. [33] devised a regression CBR for financial forecasting, which applies different weights to independent variables before finding similar cases. Kumar et al. [34] presented a comprehensive review of the works utilizing NN and CBR to solve the bankruptcy prediction problems faced by banks.

Many other data mining techniques exist for medical classification [35] and for time series identification [36]. Among these methods, a wide variety of statistical models, such as linear discriminate analysis or logistic regression perform well on a large number of applications [37]. However, the classification accuracy of these models is often limited when

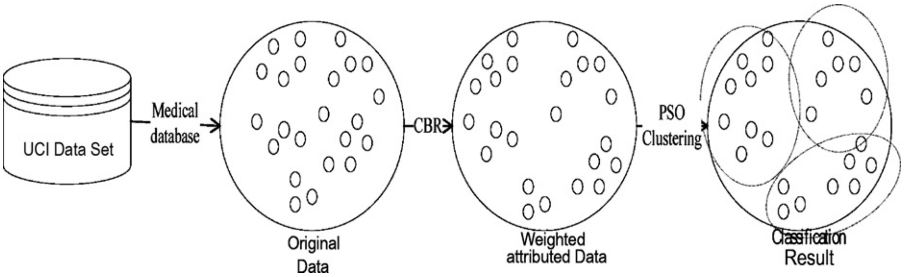

Fig. 1 – A concept model for CBRPSO.

the relationships of the input/output dataset are complex and/or non-linear [38]. In such situation, which are frequently found in real world problems, machine learning methods are more suitable for building simple and interpretable pattern classification models [36]. The most common models are [39–40]: Bayesian networks [41], neural networks, rough sets [42], decision trees [43] and genetic algorithm classifiers [44].

In this research, we focus on two basic methods to cluster (case base weighted cluster algorithm) data and to classify (PSO algorithm) the diseases. The algorithm starts off by partitioning the data set into a relatively large number of clusters in order to reduce the effects of initial conditions and then reducing the number of clusters into two. These tools are effective not only to find and describe patterns in data in order to make prediction but also to build an explicit representation of the knowledge.

3. Research method: a case-based particle swarm optimization model

PSO was originally developed by Eberhart and Kennedy in 1995 [45], and was inspired by the social behavior of a flock of birds. In the PSO algorithm, the birds in a flock are symbolically represented as particles. These particles can be considered as simple agents “flying” through a problem space. A particle’s location in the multi-dimensional problem space represents one solution for the problem. When a particle moves to a new location, a different problem solution is generated. This solution is evaluated by a fitness function that provides a quantitative value of the solution’s utility.

The velocity and direction of each particle moving along each dimension of the problem space will be altered with each generation of movement. In combination, the particle’s personal experience, p_{id} and its neighbors’ experience, p_{gd} influence the movement of each particle through a problem space. The random values $rand_1$ and $rand_2$ are used for the sake of completeness, that is, to make sure that particles explore a wide search space before converging around the optimal solution. The values of c_1 and c_2 control the weight balance of p_{id} and p_{gd} in deciding the particle’s next movement velocity. At every iteration, the particle’s new location is computed by adding the particle’s current velocity, v_{id}, to its location, x_{id}.

In PSO, each particle is treated as a point (solution) in the D-dimensional problem space. The ith particle is represented as $X_I = (x_{i1}, x_{i2}, \cdots, x_{in})$. The best previous position of the ith particle is recorded and represented as $P_I = (p_{i1}, p_{i2}, \cdots, p_{in})$. The index of the best particle among all the particles in the population is represented by the symbol g. The rate of the position change (velocity) for particle i is represented as $V_I = (v_{i1}, v_{i2}, \cdots, v_{in})$. The particles are manipulated according to the Eq. (1) and (2):

$$v_{id} = w \times v_{id} + c_1 \times rand_1() \times (p_{id} - x_{id}) + c_2 \times rand_2() \times (p_{gd} - x_{id}) \tag{1}$$

$$x_{id} = x_{id} + v_{id} \tag{2}$$

where w denotes the inertia weight factor; p_{id} is the location of the particle that experiences the best fitness value; p_{gd} is the location of the particles that experience a global best fitness value; c_1 and c_2 are constants and are known as acceleration coefficients; d denotes the dimension of the problem space; $rand_1$, $rand_2$ are random values in the range of (0, 1).

A novel model is developed by an attribute weight assignment and a clustering-based algorithm for the medical data classification problems. This cluster-based model integrates the particle swarm optimization approach (PSO), and genetic algorithms (GA) to construct a medical classification system based on medical database. The data clustering technique will calculate the weight of each input feature and GA is then applied to evolve the weights in PSO. The concept of CBRPSO is shown in Fig. 1 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) outputting the classification results. The details of each step are further explained in the following sections.

3.1. *The selection of medical databases*

Two medical data sets including liver disorders and Breast Cancer Wisconsin are selected from UCI database. Liver disorder database was support by BUPA Medical Research Ltd. Breast Cancer Wisconsin was support by Dr. William H. Wolberg et al. All these data can be found in UCI machine library database.

The liver disorders database includes 6 indices, i.e., mcv, alkphos, sgpt, sgot, gammagt and drinks. There is a total of 345 data and the first 5 features are all blood tests thought to be sensitive to liver disorders arise from excessive alcohol consumption. Each line in the data file constitutes the record

Table 1 – Basic indices for Liver Disoders.

Indices	Descriptions
mcv	Mean corpuscular volume
alkphos	Alkaline phosphotase
sgpt	Alamine aminotransferase
sgot	Aspartate aminotransferase
gammagt	Gamma-glutamyl transpeptidase
drinks	Number of half-pint equivalents of alcoholic beverages drunk per day

Table 2 – Basic indices for Wisconsin Diagnostic Breast Cancer.

Indices	Descriptions
Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter area	
Smoothness	Local variation in radius lengths
Compactness	Perimeter^2/area − 1.0
Concavity	Severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	
Fractal dimension	“Coastline approximation” – 1

of a single male individual. The basic indices for liver disorders are described in Table 1. This dataset donate by Richard S. Forsyth et al. in 1990-05-15.

Another data set applied in this research is Wisconsin Diagnostic Breast Cancer (WDBC) (Diagnostic). Those data set samples arrive periodically as Dr. Wolberg reports in his clinical cases. This Data set includes 32 features and 569 data sets. Basic Indices for Wisconsin Diagnostic Breast Cancer are described in Table 2. The input variables including the mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is mean radius, field 13 is radius SE, and field 23 is worst radius.

In this research, we will apply these two medical data sets to test the effectiveness and efficiency of the proposed method. In the next section, we will first describe the detail procedure of the weighted clustering method.

3.2. A weighted clustering method

A medical case library from UCI medical library is applied to develop the weighted distance metric. A similarity measure is described in the following.

First assume a medical case library (ML) equal to ML = (e_1, $e_2 \cdots e_N$). Each case in the library can be identified by an index of corresponding features. In addition, each Medical case has an associated action to be made for its current performance and the action is either a positive or negative decision (to judge a patient’s symptom). To be exact, we use a collection of features $F_j(j = 1 \ldots n)$ to represent the cases and a variable V to denote the action. The ith case e_j in the library can be represented as a $n+1$-dimensional vector, i.e. $e_i = (x_{i1}, x_{i2}, \ldots\ldots, x_{in}, y_i)$. Where x_j corresponds to the value of feature $F_j(j = 1 \ldots n)$ and y_i corresponds to the action $i = 1 \ldots n$ to be taken and it will be defined later. Suppose that for each j ($1 \le j \le n$) a weight wj ($w_j \in [0, 1]$) has been assigned to the jth feature to indicate the importance of the feature, then for any pair of cases e_p and e_q in the library, a weighted distance metric can be defined as Eq. (3).

$$d_{pq}^{(w)} = d^{(w)}(e_p, e_q) = \left(\sum_{j=1}^{n} w_j^2 (x_{pj} - x_{qj})^2\right)^{1/2} = \left(\sum_{j=1}^{n} w_j^2 x_j^2\right)^{1/2} \tag{3}$$

where $x_j^2 = (x_{pj} - x_{qj})^2$ when all the weights are equal to 1 the distance metric defined above coincides with the Euclidean measure, denote by $d_{pq}^{(1)}$.

By using the weighted distance defined in Eq. (3), a similarity measure between two cases, $SM_{pq}^{(w)}$, can be defined as follows:

$$SM_{pq}^{(w)} = \frac{1}{1 + \alpha d_{pq}^{(w)}} \tag{4}$$

where α is a positive parameter. When all weighs take value one the similarity measure is denoted by $SM_{pq}^{(1)}$.

After introducing the weighted distance metric and the similarity measure, the weighted clustering methodology is further described in the following:

In this step, the gradient method is applied to find the weighted value from medical data set indices and a feature evaluation function is defined [46]. The smaller is the evaluation value, the better are the corresponding features. Thus we would like to find the weights such that the evaluation function attains its minimum. The detail processes can be described as follows:

Step 1: Select the parameter α and the learning rate η.
Step 2: Initialize w_j with random values in [0,1].
Step 3: Compute Δw_j for each j using Eq. (5).

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j}, \quad w_j \in [0, 1] \tag{5}$$

and E is defined in Eq. (6).

$$E(w) = \frac{2 \times \left[\sum_{pq}\sum_{(q<p)} \left(SM_{pq}^{(w)}(1 - SM_{pq}^{1}) + SM_{pq}^{1}(1 - SM_{pq}^{(w)})\right)\right]}{N(N-1)} \tag{6}$$

where N is the number of cases in the ML base.

Step 1: Update w_j with $w_j + \Delta w_j$ for each j.
Step 2: Repeat step 3 and step 4 until convergence, i.e., until the value of E becomes less than or equal to a given threshold or until the number of iterations exceeds a certain predefined number.
Step 3: Using the final w_j as the weighted feature for PSO-clustering.

3.3. Evolving PSO forecasting model

This research combines genetic algorithms, attribute weights assignment approach and particle swarm optimization model

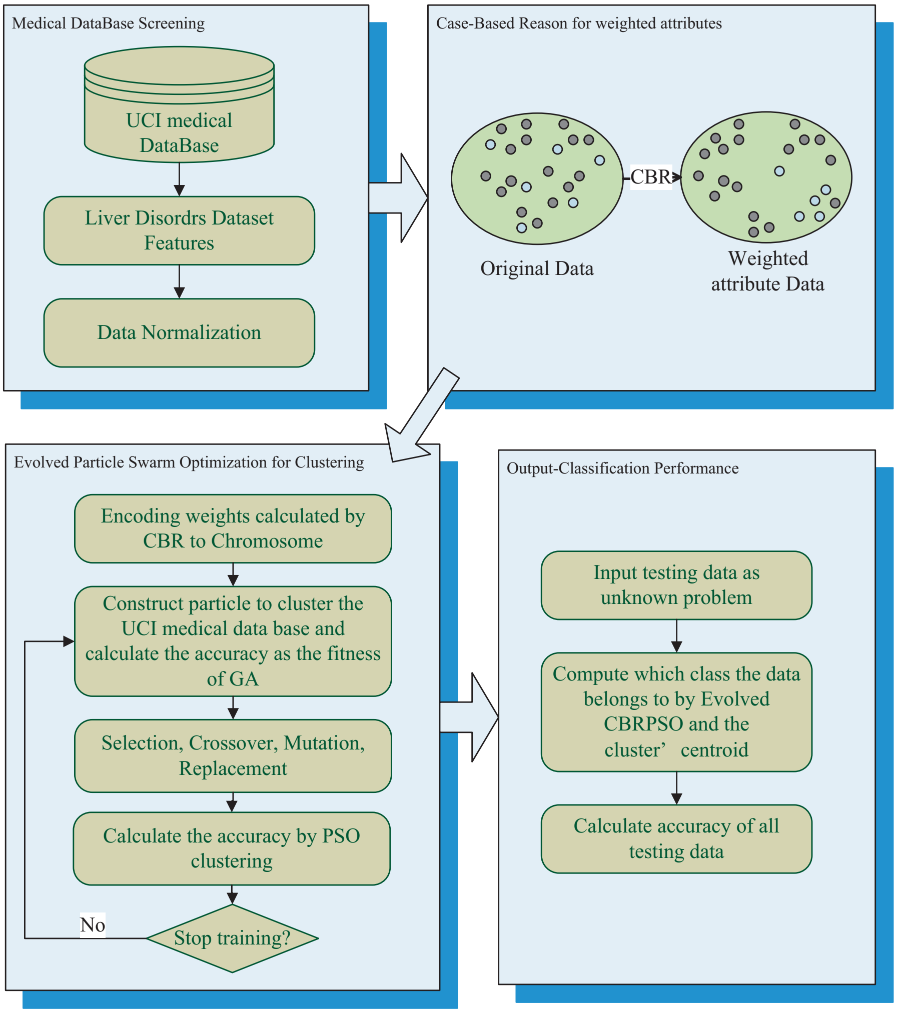
Medical DataBase Screening
UCI medical DataBase
Liver Disordrs Dataset Features
Data Normalization
Case-Based Reason for weighted attributes
CBR
Original Data
Weighted attribute Data
Evolved Particle Swarm Optimization for Clustering
Encoding weights calculated by CBR to Chromosome
Construct particle to cluster the UCI medical data base and calculate the accuracy as the fitness of GA
Selection, Crossover, Mutation, Replacement
Calculate the accuracy by PSO clustering
No
Stop training?
Output-Classification Performance
Input testing data as unknown problem
Compute which class the data belongs to by Evolved CBRPSO and the cluster' centroid
Calculate accuracy of all testing data

Fig. 2 – A framework of GAPSO model.

to develop a classification model for the prediction of medical symptom decisions. The framework of GA-CBRPSO is depicted as Fig. 2.

The whole clustering behavior of the PSO-clustering algorithm can be classed into two stages: a global searching stage and a local refining stage. At the initial iterations, based on the PSO algorithm's particle velocity updating Eq. (2), the particle's initial velocity v_{id}, the two randomly generated values ($rand_1$, $rand_2$) at each iteration and the inertia weight factor w provide the necessary diversity to the particle swarm by changing the momentum of particles to avoid the stagnation of particles at the local optima. Multiple particles parallel searching, using multiple different solutions at a time, can explore more area in the problem space. The initial iterations can be classified as the global searching stage. After several iterations, the particle's velocity will gradually reduce and the particle's explore area will shrink while the particle will approach the optimal solution. The global searching stage gradually changes to the local refining stage. By selecting different parameters in the PSO algorithm, we can control the shift time from the global searching stage to the local refining stage. The later the particle shift from the global searching stage to local refining stage, greater the possibility that it can find the global optimal solution.

3.4. 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

$$\min_{P_K \in \Gamma_K} \sum_{i=1}^{K} \sum_{X_l \in C_i} dist(X_l, \bar{X}), \quad \bar{X} = \text{centroid of each cluster } C_i \tag{7}$$

$$dist(X_l, \bar{X}) = \sqrt{\sum_{i=1}^{n} (x_{li} - \bar{x}_i)^2} \tag{8}$$

$$X_l = (x_{l1}, x_{l2}, \ldots, x_{ln}), \quad \bar{X} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$$

where n is the number of attributes.

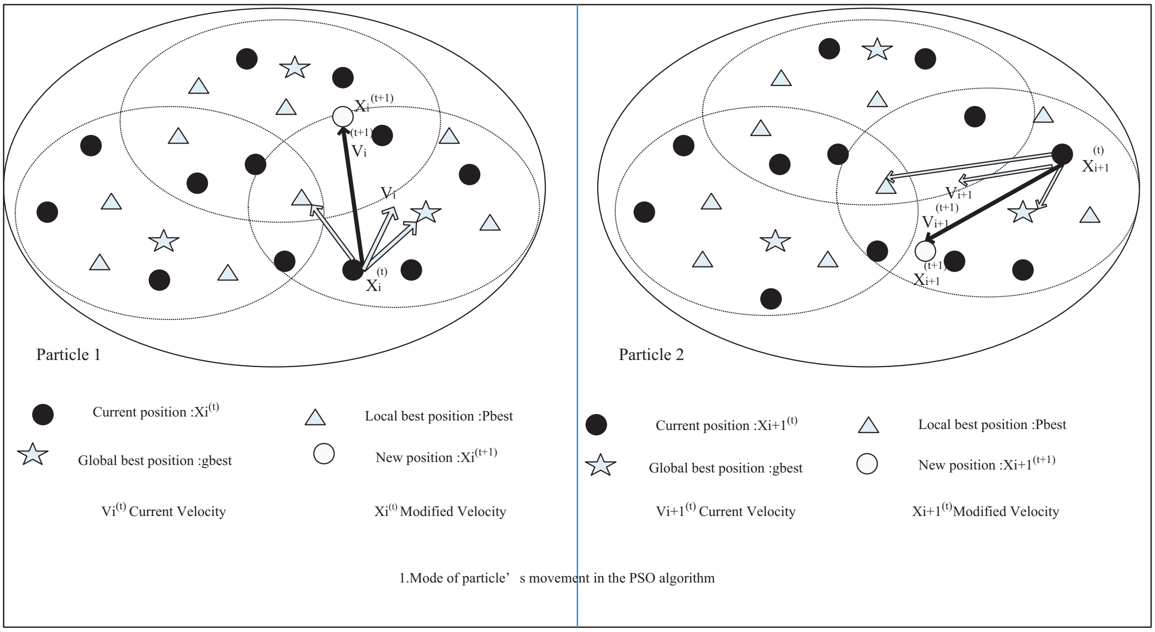

Fig. 3 – An example of two particles in PSO-clustering.

The following diagram is the pseudo code of PSO-clustering algorithm. Fig. 3 shows the concept about the pseudo code and Fig. 4 shows the example about distance measurement used in one particle while K is set as five.

```
Input: Medical Data Set $K$: 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 $= \min_{P_K \in \Gamma_K} \sum_{i=1}^{K} \sum_{x_l \in C_i} dist(X_l, \bar{X})$
    $v_{id} = w \times v_{id} + c_1 \times rand_1() \times (p_{id} - x_{id})$
      $+ c_2 \times rand_2() \times (p_{gd} - x_{id})$
    $x_{id} = x_{id} + v_{id}$
    Update $p_{id}$
  End
  Update $p_{gd}$
End
```

Because of overlapping situation always happens in real case, this study adopts a stepwise cluster reduce algorithm to avoid the overlapping and increase the accuracy. For example, in the steps of clustering the liver disorder data, we first separate 3345 data into 20 clusters, and then we assign the 20 centroids of each cluster as new data points and PSO cluster will cluster the 20 data points into final result, the two kinds of diagnostic result. The pseudo code is in the following diagram and Fig. 5 shows an example.

```
$N$: data points
$K$: number of classes (the same with number of cluster'
  centroids)
$M$: temporary centroids ($M > K$, for initial)
Procedure Stepwise_Centroids_PSO_Clustering
  $M$: = PSO_Clustering($N$, |$M$|);
  Reassign $M$ as data points ($N$: = $M$);
  Reduce number of $M$ to $M$'
  Recursive execute Stepwise_Centroids_PSO_Clustering 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
```

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

$$dist(X_l, \bar{X}) = \sqrt{\sum_{i=1}^{M} w_i \times (x_{li} - \bar{x}_i)^2}, \qquad X_l = (x_{l1}, x_{l2}, \ldots, x_{lM}) \text{ and } \bar{X} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_M) \tag{9}$$

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.

3.5. *Hybrid PSO-clustering with CBR*

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

```
    Do
        Compute $\Delta w_j = -\eta \frac{\partial E}{\partial w_j}$; // formula (5)
        Update $w_j := w_j + \Delta w_j$
  While not convergent;
  Assign each attribute $j$ has its own weight;
End;
```

As described in Section 2, the CBR algorithm calculates weights of each attributes; hence the pseudo code of CBR-PSO-clustering can be modified as following two diagrams.

```
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 data set
    d: dimension of each data (number of attributes)
Input: data: Medical Data Set
    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 data set.
  For each particle
    Objective function $= \min_{P_M \in \Gamma_M} \sum_{i=1}^{K} \sum_{x_l \in C_i} \sqrt{\sum_{j}^{d} w_j (x_l - C_i)^2}$
    $v_{id} = w \times v_{id} + c_1 \times rand_1() \times (p_{id} - x_{id})$
        $+ c_2 \times rand_2() \times (p_{gd} - x_{id})$
    $x_{id} = x_{id} + v_{id}$
    Update $p_{id}$
  End
  Update $p_{gd}$
End
```

3.6. *Hybrid PSO-clustering with CBR and evolved by GA*

In this study, we adopted GA to evolve the weights calculate from CBR to find out the better weight of each attribute. In experimental result, the binary encoding is used to covert weights into an individual chromosome.

```
Procedure Evolved_Weighted_PSO_Clustering
  Weights_Calculate_by_CBR();
    Binary encoding weights to chromosome
    Randomly generating the rest chromosomes in population
    While not terminated
      For each individual
        Stepwise_Centroids_PSO_Clustering_with_CBR(data, K, weights);
      End;
      Selection();
      Crossover();
      Mutation();
  End;
End;
```

Since GA is a population-base heuristic algorithm, we used the weights from CBR as one of the initial solution in population and randomly generated the rest chromosomes in population in order to find the global optimization.

Merwe's research [47] indicates that utilizing the PSO algorithm's optimal ability, if given enough time, the PSO-clustering algorithm could generate more compact clustering results from the low dimensional dataset than the traditional *K*-means clustering algorithm. However, when clustering large document datasets, the slow shift from the global searching stage to the local refining stage causes the PSO-clustering algorithm to require many more iterations to converge to the optima in the refining stage than the *K*-means algorithm requiring. Although the PSO algorithm is inherently parallel and can be implemented using parallel hardware, such as a computer cluster, the computation requirement for clustering large document dataset is still high. In our experiments, it needs more than 500 iterations for the PSO algorithm to converge to the optimal result for a database that includes 800 medical data. The *K*-means algorithm only requires 10–20 iterations.

Although the PSO algorithm generates much better clustering result than the *K*-means algorithm does, in terms of execution time, the *K*-means algorithm is more efficient for large datasets. For this reason, we present a hybrid PSO approach that uses *K*-means algorithm to replace the refining stage in the PSO algorithm. In the hybrid PSO algorithm, the algorithm includes two modules, the PSO module and the *K*-means module. The global searching stage and local refine stage are accomplished by those two modules, respectively. In the initial stage, the PSO module is executed for a short period (50–100 iterations) to discover the vicinity of the optimal solution by a global search and at the same time to avoid consuming high computation. The result from the PSO module is used as the initial seed of the *K*-means module. The *K*-means algorithm will be applied for refining and generating the final result. The whole approach can be summarized as:

(1) Start the PSO-clustering process until the maximum number of iterations is exceeded.
(2) Inherit clustering result from PSO as the initial centroid vectors of *K*-means module.

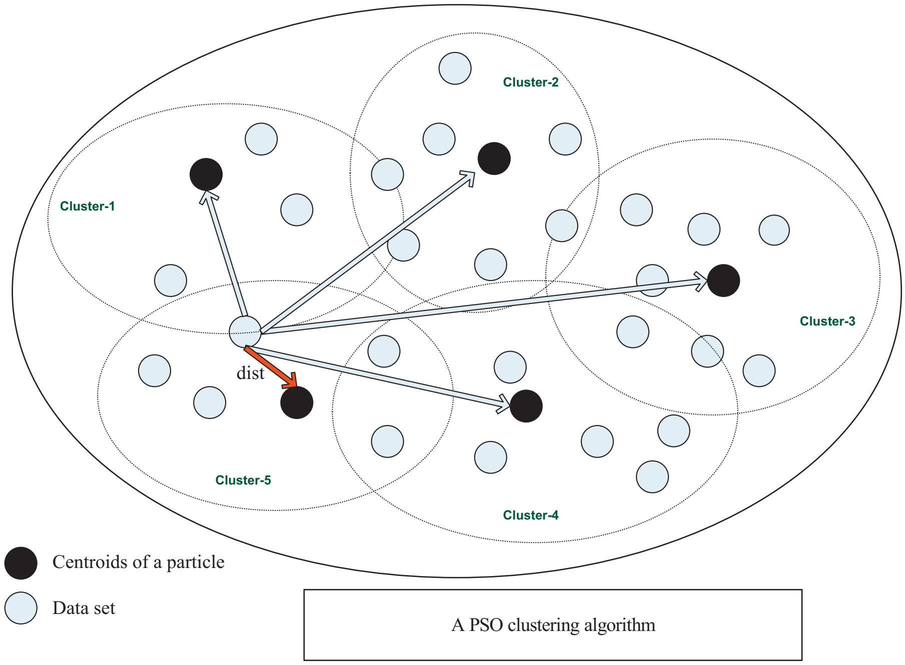

Fig. 4 – The example about distance measurement in PSO-clustering while $K = 5$.

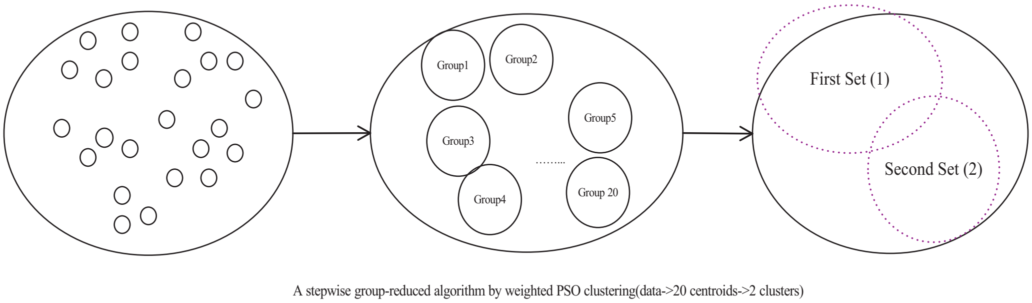
Group1
Group2
Group3
Group4
Group5
Group 20
First Set (1)
Second Set (2)
A stepwise group-reduced algorithm by weighted PSO clustering(data->20 centroids->2 clusters)

Fig. 5 – A step by step PSO-clustering examples (take liver disorder database for example).

(3) Start K-means process until maximum number of iterations is reached.

4. The judgment of output value

This research mainly applies evolutional case-based PSO-clustering algorithm to classify the potential medical symptom by comparing the results from a medical dataset library.

5. Experimental results

According to the criteria listed in Tables 1 and 2, there are two medical database applied to test the efficiency of the proposed model. In addition, the proposed model will be compared with other models developed earlier in the literatures. Another purpose in our research to choose the different kinds of medical database to show that the proposed model have a robust performance even under different type of medical database. Detailed procedures of CBRPSO applied in these two different medical databases are explained in the following sections.

Table 3 – Best parameter setting from experimental designs.

Parameter setting	Liver disorders	Wisconsin Diagnostic Breast Cancer
α	0.6	0.6
Learning rate	0.05	0.05
Weighted stage running times	5000	5000

5.1. *A weighted clustering method*

Experimental design is applied to decide the best parameter setting. After the experimental tests, the parameter setting is shown in Table 3.

For taking liver disorder data set as an example, this research finds if the attributes weighted as shown in Table 4, the best accuracy calculated can achieve 78.2%. In comparison, the results are compared with the original PSO

Table 4 – The final calculated weights of each attribute of Liver Disorder.

Factor	mcv	alkphos	sgpt	sgot	gammagt	Drinks
Weight	0.67	0.12	0.52	0.51	0.65	0.98

Table 5 – Accuracy comparisons of different PSO approaches used in this research.

Medical database	PSO	CBR-PSO	GA-CBRPSO
	Average diagnosis accuracy (500 running times)		
Liver disorders			
Best	71.5%	73.9%	78.2%
Average	62.5%	68.4%	76.8%
Lowest	50.4%	52.1%	54.2%
Wisconsin Diagnostic Breast Cancer			
Best	92.4%	93.8%	97.9%
Average	90.3%	92.6%	97.4%
Lowest	88.4%	90.8%	96.3%

Table 6 – Accuracy comparisons of different forecasting models in two different medical database.

Medical database	SVM	KNN	Naïve Bayes	FDT	CBR-PSO
	Average diagnosis accuracy ratio (500 running times)				
Liver disorders					
Best	77.6%	73.7%	70.2%	68.3%	78.2%
Average	69.3%	61.1%	59.2%	60.1%	76.8%
Lowest	63.2%	54.8%	58.1%	58.7%	54.2%
Wisconsin Diagnostic Breast Cancer					
Best	98.1%	96.9%	91.4%	90.2%	97.9%
Average	93.2%	89.8%	87.6%	86.2%	97.4%
Lowest	81.3%	80.1%	85.3%	78.9%	96.3%

Table 7 – Accuracy rate comparisons of CBPSO with other approaches from previous researches in Liver Disorder medical database.

Author (year)	Method	Classification
Accuracy (average)		
Pham et al. (2000) [48]	RULES-4	55.90%
Cheung (2001) [8]	C4.5	65.59%
Cheung (2001) [8]	Naïve Bayes	63.39%
Cheung (2001) [8]	BNND	61.83%
Cheung (2001) [8]	BNNF	61.42%
Van Gestel et al. (2002) [49]	SVM with GP	69.70%
Lee et al. (2001a) [9]	SSVM	70.33%
Lee et al. (2001b) [50]	RSVM	74.86%
Yalçın et al. (2003) [51]	MLP	73.05%
Yalçın et al. (2003) [51]	PNN	42.03%
Yalçın et al. (2003) [51]	GRNN	65.55%
Our methods	CBR + PSO	76.81%

without weighted approach and the other classification approaches.

5.2. *Best parameters Setting for genetic algorithms*

Genetic algorithms are applied to evolve the weights of each attributes (factors) in this research. The important factors in GA, population size, number of generation, crossover rate and mutation rate are 20, 100, 0.9 and 0.1, respectively. The parameters are determined by experiments and the test benchmarks chosen in this research are adopted the same setting.

5.3. *Comparisons of different models*

After setting up the parameters of the experiments, we take the output of CBRPSO and compare with the output from traditional classification tools. In our experimental study, Medical database sets (liver disorder and Wisconsin Diagnostic Breast Cancer) are weighted by the Case Based method first. The result is then classified by Evolving PSO. As shown in Table 5, 62.5% classification accuracy in liver disorder and 90.3% in Breast Cancer is obtained from standard PSO while with the pre-weighted processing in data clustering, 76.8% (liver disorders) and 97.4% (Wisconsin Diagnostic Breast Cancer) classification accuracy is obtained. In our research, KNN, Naïve Bayes, SVM were chosen to be compared with our method and the results are listed in Table 6. The CBRPSO approach performs much better than traditional classification tools in liver disorder database, but in Wisconsin database, CBRPSO shows more similar than traditional SVM method. 75% of the data were 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 7 and 8, the CBRPSO is also compared with other approaches developed in the literature to show the effectiveness of our approach.

6. Discussions

As the results listed in Table 5 to Table 8, CBPSO outperforms other traditional methods. The reasons including, first, the attributes can have difference weights before applying the clustering algorithm. The weights can be calculated by the proposed CBR approaches and then evolving by GA algorithm for deciding the best weight for each attribute. Second, stepwise based PSO can overcome the overlapping situation of data set. Therefore, the accuracy can be improved when compared with original PSO-clustering algorithm as shown in Table 5 and other traditional algorithm.

Table 8 – Accuracy rate comparisons of CBPSO with other approaches from previous researches in Wisconsin Diagnostic Breast Cancer medical database.

Author (year)	Method	Classification accuracy (average)
Quinlan (1996) [52]	C4.5	94.74%
Hamilton et al. (1996) [53]	RIAC	94.99%
Ster et al. (1996) [54]	LDA	96.80%
Bennett et al. (1997) [7]	SVM	97.20%
Nauck et al. (1999) [55]	NEFCLASS	95.06%
Pena-Reyes et al. (1999) [3]	Fuzzy-GA1	97.36%
Goodman et al. (2002) [56]	Optimized-LVQ	96.70%
Goodman et al. (2002) [56]	Big-LVQ	96.80%
Goodman et al. (2002) [56]	AIRS	97.20%
Abonyi et al. (2003)	Supervised fuzzy clustering	95.57%
Our methods	CBR + PSO	97.41%

7. Conclusions

A considerable amount of research has been conducted to study the behavior of a series of medical symptoms. However, the researcher is more interested in finding potential weights of disease factors. Therefore, we take a different approach by applying a CBRPSO-clustering approach to diagnose the Potential illness symptoms diagnosis. Next, a weighted clustering method is adopted, therefore, these data react to the detection prominently. The GA is applied to evolve the weights of each factor in order to derive the best weights of each attribute. Through a series of experimental tests, the CBRPSO outperforms other approaches with an average accuracy rate around 97.4% in breast cancer and 76.8% in liver disorder, respectively. It is the highest among the literature published up to present. This model can be further applied in classification of other medical disease database to help researcher or doctors to make better decision in medical diagnosis. Our experimental results illustrate that using this CBRPSO algorithm can generate higher compact clustering than using either the PSO or the *K*-means alone.

In the future, the proposed system can be further investigated by incorporating other soft computing techniques or a better Data Mining forecasting model. They are listed as follows:

1. A different classification model: There are numerous forecasting models other than clustering model exists in the academic area. It is worth a while to study the behavior of these models when applied in prediction of the illness symptom detection. Different forecasting models such as support vector regression machine are a possible candidate models for improving the accuracy of the performance.
2. Different data transformation method: The non-linear transformation like Gaussian transformation may lead to a better performance result.

REFERENCES

[1] C.C. Bojarczuk, H.S. Lopes, A.A. Freitas, Genetic programming for knowledge discovery in chest-pain diagnosis: exploring a promising data mining approach, IEEE Engineering in Medicine and Biology Magazine 19 (2000) 38–44.

[2] C.A. Pena-Reyes, M. Sipper, Evolving fuzzy rules for breast cancer diagnosis, in: Proceedings of 1998 International Symposium on Nonlinear Theory and Applications (NOLTA'98), vol. 2, 1998, pp. 369–372.

[3] C.A. Peña-Reyes, M. Sipper, A fuzzy-genetic approach to breast cancer diagnosis, Artificial Intelligence in Medicine 17 (1999) 131–155.

[4] X. Chang, J.H. Lilly, Evolutionary design of a fuzzy classifier from data, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34 (2004) 1894–1906.

[5] R. Setiono, Extracting rules from pruned neural networks for breast cancer diagnosis, Artificial Intelligence in Medicine 8 (1996) 37–51.

[6] R. Setiono, H. Liu, Symbolic representation of neural networks, Computer 29 (1996) 71–77.

[7] K.P. Bennett, J.A. Blue, A support vector machine approach to decision trees, Math Report (1997).

[8] N. Cheung, Machine learning techniques for medical analysis, Machine Learning Techniques for Medical Analysis (2001).

[9] Y.J. Lee, O.L. Mangasarian, SSVM: a smooth support vector machine for classification, Computational Optimization and Applications 20 (2001) 5–22.

[10] K.P. Bennett, O.L. Mangasarian, Neural network training via linear programming, Advances in Optimization and Parallel Computing (1992) 56–67.

[11] D. Cosic, S. Loncaric, Rule-based labeling of CT head image, Lecture Notes in Artificial Intelligence 1211 (1999) 453–456.

[12] W. Duch, R. Adamczak, K. Grabczewski, G. Zal, Y. Hayashi, Fuzzy and crisp logical rule extraction methods in application to medical data, Computational Intelligence and Applications 23 (2000) 593–616.

[13] Y. Hayashi, R. Setiono, K. Yoshida, A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders, Artificial Intelligence in Medicine 20 (2000) 205–216.

[14] P.J.G. Lisboa, E.C. Ifeachor, P.S. Szczepaniak, Artificial Neural Networks in Biomedicine, Springer-Verlag, London, Berlin, Heidelberg, 2000.

[15] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artificial Intelligence in Medicine 23 (2001) 89–109.

[16] R Andrews, J. Diederich, A.B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-Based Systems 8 (1995) 373–389.

[17] A.B. Tickle, R. Andrews, M. Golea, J. Diederich, The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks, IEEE Transactions on Neural Networks 9 (1998) 1057–1068.

[18] R. Setiono, Generating concise and accurate classification rules for breast cancer diagnosis, Artificial Intelligence in Medicine 18 (2000) 205–219.

[19] S.C. Hui, G. Jha, Data mining for customer service support, Information and Management 38 (2000) 1–13.

[20] T.W. Liao, An investigation of a hybrid CBR method for failure mechanisms identification, Engineering Applications of Artificial Intelligence 17 (2004) 123–134.

[21] B.S. Yang, T. Han, Y.S. Kim, Integration of ART-Kohonen neural network and case-based reasoning for intelligent

fault diagnosis, Expert Systems with Applications 26 (2004) 387–395.

[22] K. Hua Tan, C. Peng Lim, K. Platts, H. Shen Koay, An intelligent decision support system for manufacturing technology investments, International Journal of Production Economics 104 (2006) 179–190.

[23] K.M. Saridakis, A.J. Dentsoras, Case-DeSC: a system for case-based design with soft computing techniques, Expert Systems with Applications 32 (2007) 641–657.

[24] J.M. Garrell, I. Guiu, E. Golobardes, I. Ribeĭ, E. Bernadoĭ, I. Mansilla, X. Llora¸ I. Fab̀rega, Automatic diagnosis with genetic algorithms and case-based reasoning, Artificial Intelligence in Engineering 13 (1999) 367–372.

[25] C.C. Hsu, C.S. Ho, A new hybrid case-based architecture for medical diagnosis, Information Sciences 166 (2004) 231–247.

[26] B. Wyns, et al., Prediction of arthritis using a modified Kohonen mapping and case based reasoning, Engineering Applications of Artificial Intelligence 17 (2004) 205–211.

[27] H. Ahn, K.j. Kim, Global optimization of case-based reasoning for breast cytology diagnosis, Expert Systems with Applications 36 (2009) 724–734.

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

[29] K.S. Kim, 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 21 (2001) 147–156.

[30] H. Li, J. Sun, B.L. Sun, Financial distress prediction based on OR-CBR in the principle of *k*-nearest neighbors, Expert Systems with Applications 36 (2009) 643–659.

[31] P.C. Chang, 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 29 (2005) 183–192.

[32] P.C. Chang, C.Y. Lai, K.R. Lai, A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting, Decision Support Systems 42 (2006) 1715–1729.

[33] S.H. Chun, Y.J. Park, A new hybrid data mining technique using a regression case based reasoning: application to financial forecasting, Expert Systems with Applications 31 (2006) 329–336.

[34] P. Ravi Kumar, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review, European Journal of Operational Research 180 (2007) 1–28.

[35] R.S. Youssif, C.N. Purdy, Combining genetic algorithms and neural networks to build a signal pattern classifier, Neurocomputing 61 (2004) 39–56.

[36] K.C. Lee, S.B. Oh, An intelligent approach to time series identification by a neural network-driven decision tree classifier, Decision Support Systems 17 (1996) 183–197.

[37] J. De Andrés, M. Landajo, P. Lorca, Forecasting business profitability by using classification techniques: a comparative analysis based on a Spanish case, European Journal of Operational Research 167 (2005) 518–542.

[38] E.I. Altman, G. Marco, F. Varetto, Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience), Journal of Banking and Finance 18 (1994) 505–529.

[39] T. Kervahut, J.Y. Potvin, An interactive-graphic environment for automatic generation of decision trees, Decision Support Systems 18 (1996) 117–134.

[40] M. Setnes, U. Kaymak, Fuzzy modeling of client preference from large data sets: an application to target selection in direct marketing, IEEE Transactions on Fuzzy Systems 9 (2001) 153–163.

[41] A.K. Thompson, An analysis of bayesian classifiers, in: Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92), 1992, pp. 223–228.

[42] S. Greco, B. Matarazzo, R. Slowinski, Rough sets theory for multicriteria decision analysis, European Journal of Operational Research 129 (2001) 1–47.

[43] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees, Wadsworth Intl (1984).

[44] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, 1989.

[45] J Kennedy, R. Eberhart, Particle Swarm Optimization (1995) 1942–1948.

[46] S.C.K. Shiu, C.H. Sun, X.Z. Wang, D.S. Yeung, Maintaining case based reasoning systems using fuzzy decision trees, Lecture Notes in Artificial Intelligence 1898 (2000) 285–296.

[47] D.W. van der Merwe, A.P. Engelbrecht, Data clustering using particle swarm optimization, Congress on Evolutionary Computation (2003) 215–220.

[48] D.T. Pham, S.S. Dimov, Z. Salem, Technique for selecting examples in inductive learning, Esit 2000 European Symposium on Intelligent Techniques (2000) 119–127.

[49] T. Van Gestel, et al., Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and Kernel Fisher discriminant analysis, Neural Computation 14 (2002) 1115–1147.

[50] Y.J Lee, O.L. Mangasarian, RSVM: reduced support vector machines, Proceedings of the First SIAM International Conference on Data Mining (2001).

[51] M. Yalçýn, T. Yildirim, Karaciðer bozukluklarinin yapay sinir aðlarý ile tesphisi, Biomedical Mühendisliý i Ulusal Toplantisi (BIYOMUT 2003) (2003) 293–297.

[52] J.R. Quinlan, Improved use of continuous attributes in C4.5, Journal of Artificial Intelligence Research 4 (1996) 77–90.

[53] H.J. Hamilton, N. Shan, N. Cercone, RIAC: A rule induction algorithm based on approximate classification, Technical Report CS 96 (1996).

[54] B. Ster, A. Dobnikar, Neural networks in medical diagnosis: Comparison with other methods, in: Proceedings of International Conference EANN'96, 1996, pp. 427–430.

[55] D. Nauck, R. Kruse, Obtaining interpretable fuzzy classification rules from medical data, Artificial Intelligence in Medicine 16 (1999) 149–169.

[56] D.E. Goodman Jr., L.C. Boggess, A.B. Watkins, Artificial Immune System Classification of Multiple-Class Problems (2002) 179–184.