# 



[Expert Systems with Applications 39 (2012) 11657–11665](http://dx.doi.org/10.1016/j.eswa.2012.04.036)

Contents lists available at [SciVerse ScienceDirect](http://www.sciencedirect.com/science/journal/09574174)

Expert Systems with Applications

journal homepage: [www.elsevier.com/locate/eswa](http://www.elsevier.com/locate/eswa)



A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease

S. Muthukaruppan [a](#_bookmark0),[⇑](#_bookmark2), M.J. Er [b](#_bookmark1)

a *School of Mechanical and Aerospace Engineering, Nanyang Technological University, N3, 50 Nanyang Avenue, Singapore 639798, Singapore*

b *School of Electrical and Electronic Engineering, Nanyang Technological University, S1, 50 Nanyang Avenue, Singapore 639798, Singapore*

## a r t i c l e i n f o

*Keywords:*

Particle swarm optimization Decision tree

Coronary artery disease If–then rules Membership function Expert system

## a b s t r a c t

This paper presents a particle swarm optimization (PSO)-based fuzzy expert system for the diagnosis of coronary artery disease (CAD). The designed system is based on the Cleveland and Hungarian Heart Dis-

ease datasets. Since the datasets consist of many input attributes, decision tree (DT) was used to unravel the attributes that contribute towards the diagnosis. The output of the DT was converted into crisp if– then rules and then transformed into fuzzy rule base. PSO was employed to tune the fuzzy membership functions (MFs). Having applied the optimized MFs, the generated fuzzy expert system has yielded 93.27% classiﬁcation accuracy. The major advantage of this approach is the ability to interpret the deci- sions made from the created fuzzy expert system, when compared with other approaches.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Cardiovascular diseases are a group of disorders of the heart and its blood vessels, including coronary artery disease (CAD), cerebro-

vascular disease, peripheral arterial disease, rheumatic heart dis-

ease, congenital heart disease, pulmonary embolism. Among those mentioned above, CAD is the most common type of cardio- vascular disease and accounts for over 600,000 deaths each year in the European Union ([Wilson et al., 1998](#_bookmark40)). It is the largest killer disease of American males and females, which caused about one of every six deaths in the United States (US) in 2006. The estimated direct and indirect cost of CAD in the US in 2010 is $177.1 billion ([American Heart Association, 2010](#_bookmark31)). In the United Kingdom, CAD caused more than 120,000 deaths in 2001 ([British Heart Founda-](#_bookmark36) [tion, 2003](#_bookmark36)). Worldwide, coronary artery disease is becoming pan- demic as developing countries experience the epidemiological transition from famine to degenerative disease. Moreover, CAD tends to affect the younger population and thus could negatively affect the productivity and workforce ([Omran, 1979](#_bookmark20)). The World

Health Organization (WHO) estimates 11.1 million deaths from

CAD in 2020.

In CAD, changes in one or more of the coronary arteries cause inadequate blood ﬂow to the heart, which results in the develop- ment of atherosclerotic plaques within the walls of coronary arter- ies, narrowing of the lumen of the coronary artery, and subsequently, occlusion, and thus leading to myocardial infarction

\* Corresponding author. Tel.: +65 93519126.

*E-mail address:* [muthukar1@e.ntu.edu.sg](mailto:muthukar1@e.ntu.edu.sg) (S. Muthukaruppan).

(MI) or sudden death. The identiﬁcation of cardiovascular risk

factors, preventive measures, its diagnosis and early treatment are

of great importance to lessen the cardiac morbidity and mortality.

Several computer aided diagnosis methodologies have been proposed in the literature for the diagnosis of CAD. More speciﬁ- cally, the use of approaches like artiﬁcial neural networks ([Patil](#_bookmark24) [& Kumaraswamy, 2009; Resul, Ibrahim, & Abdulkadir, 2009; Ture,](#_bookmark24) [Kurt, & Kurum, 2008](#_bookmark24)), Naïve Bayes ([Tu et al., 2009](#_bookmark41)), support vector machines ([Andreeva, 2006](#_bookmark32)), decision trees ([Palaniappan and](#_bookmark22) [Awang, 2008](#_bookmark22)) have been previously reported.

Even though these approaches produce good classiﬁcation accu- racy, the interpretation of results is hard. They are popularly known as ‘‘Black Box’’ method since they focus only on the classi- ﬁcation accuracy. Although rule based classiﬁer systems, reported in [Tsipouras et al. (2008)](#_bookmark41) and [Adeli and Neshat (2010)](#_bookmark29) produces interpretable rules, they lack the robustness in the missing data.

Expert systems are a branch of artiﬁcial intelligence which solve the problems at the level of a human expert making use of special- ized knowledge represented by a set of rules. The fuzziness and

imprecision, which is inherited in the biomedical problems can

be treated incorporating fuzzy logic. A fuzzy expert system (FES) is simply an expert system which include set of fuzzy rules and membership functions, i.e., knowledge acquisition, considered to be the most important issue in the design of fuzzy expert system. In general, knowledge could be obtained from the experts in the particular area. When there is an increase in possible number of rules of the FES, experts ﬁnd it difﬁcult to deﬁne the complete rule set. Also the performance of the system can also be increased by

tuning of membership function using optimization algorithms ([Ganesh Kumar et al., 2012](#_bookmark12)).

***acer***

*2018-05-17 22:13:29*

--------------------------------------------

objective

***acer***

*2018-05-17 22:14:18*

--------------------------------------------

note

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved. <http://dx.doi.org/10.1016/j.eswa.2012.04.036>

This paper proposes a new particle swarm optimization (PSO)- based fuzzy expert system that involves four stages. In the ﬁrst

stage, the missing data are imputed using nearest neighbour hot

deck imputation, while in the second stage, decision tree induction

and set of rules is extracted from it. In the third stage, the crisp

rules are transformed into fuzzy rule base using fuzzy membership

functions. Finally, in the fourth stage, the fuzzy membership func-

tions are tuned by PSO. The fuzzy model with the optimized parameters results in the ﬁnal FES. Since the generated FES is based on the set of rules, they are able to provide interpretations for their decisions. The use of decision tree in the ﬁrst stage has the advan- tage of discovering new knowledge and is considered to be a very effective technique for the classiﬁcation tasks ([Pedrycz & Sosnow-](#_bookmark25) [ski, 2005; Quinlan, 1996](#_bookmark25)). Furthermore, the development of the FES from the set of rules and tuning of the MFs, improve the accuracy. The incorporation of fuzzy logic deals with the uncertain situation and fuzziness which are inherent in biomedical classiﬁcation prob-

lems ([Tsoukalas and Uhrig, 1997](#_bookmark41)).

[and Wohlin, 2004; Yu, 2005; Zhang et al., 2005](#_bookmark13)). In this method, for each record that contains the missing values, the most similar record is found from the same data set and the missing values are imputed from that record. If the most similar record also con- tains the missing value for the same attribute, it is neglected, and the next closest record is found. This process is repeated until all the missing values of the entire database are imputed. There are several ways to ﬁnd the most similar record to the record with the missing values ([Rubin, 1987](#_bookmark34)).

In this research, Heterogeneous Euclidean Overlap Metric

(HEOM) distance function is presented, which uses the overlap method for categorical attributes and a normalized euclidean dis- tance for numeric attributes ([Farhangfar et al., 2004](#_bookmark12)). This HEOM distance eliminates the effects of arbitrary ordering of the categor- ical attributes ([Jerez et al., 2010](#_bookmark12)). The HEOM distance between two input vectors *x* and *y* is given as follows:

### sﬃXﬃﬃ*m*ﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃﬃ2**ﬃ**

The paper is organized as follows. Section 2 brieﬂy introduces the datasets employed, decision tree algorithm. In Section 3, fuzzy

HEOMð*x*; *y*Þ¼

*a*¼1

*da*ð*xa*; *ya*Þ

inference system and optimization of the fuzzy parameters are presented. Details of the simulations conducted and the results are reported in Section 4. Section 6 discusses about the concluding remarks.

***acer***

*2018-05-17 22:17:39*

--------------------------------------------

note

1. 2.Decision tree algorithm
   1. *Datasets*

The system is designed based on the Hungarian Institute of Cardiology, Budapest and the Cleveland Clinic Foundation datasets ([Newman, Hettich, Blake, & Merz, 1998](#_bookmark21)). These datasets are the part of collection of databases at the University of California, Irvine. It provides 597 records in total. The database contains 76 attri- butes. However, all the published experiments refer to 13 of them as inputs and 1 attribute as a result. The input variables are age,

blood pressure, serum cholesterol, maximum heart rate, sex, chest

pain type, fasting blood sugar, resting ECG, exercise induced

angina, old peak, slope, ﬂuoroscopy and thallium scan. The output

variable is the angiographic status.

* 1. *Missing data imputation*

Many real world databases contain missing values arising due to many reasons such as data entry procedures, incorrect measure- ments and malfunction of the equipments or essential information has not been collected from the sources. With this incompleteness, it is difﬁcult to generate useful knowledge from data, since many machine learning algorithms can only work with complete data. The simplest way of dealing with this missing data is to discard the record that contains missing values. This method is applicable only when the data contain relatively small number of missing values. However, when a large proportion of missing values is pres- ent, this approach may lead to wrong conclusions ([Brown & Kros,](#_bookmark38) [2003](#_bookmark38)).

A possible approach to deal with this problem is to carry out data imputation that is deﬁned as the process in which the missing data are estimated by appropriately computed values. One advan- tage of data imputation is that the missing value treatment is inde-

pendent of the machine learning algorithm employed. This enables the user to select the most appropriate data imputation method for their application. Different approaches have been discussed in [Grzymala-Busse and Hu (2001) and Su et al. (2008)](#_bookmark12).

In this paper, nearest neighbour hot deck imputation has been employed owing to its superiority over other methods ([Jonsson](#_bookmark13)

where *da*(*x*, *y*) is the distance between two values *x* and *y* of a given attribute ‘*a*’ and is given as follows:

### 8 1; if *x* or *y* is unknown; else

*da*ð*x*; *y*Þ¼

*overlap*ð*x*; *y*Þ; if *a* is nominal; else

><

*rn diffa*ð*x*; *y*Þ

>:

The overlap function assigns a value of 0 if both the categorical val- ues are the same; otherwise the value is 1.The range normalized dif- ference function is given as follows:

*rn diffa*ð*x*; *y*Þ¼ j*x* — *y*j=ð*maxa* — min*a*Þ

where *maxa* and *mina* are the observed maximum and minimum values in the attribute *a*. The above deﬁnition for *da* returns a value in the range of 0–1 whether the input is categorical or numeric ([Wilson & Martinez, 1997](#_bookmark40)).

* 1. *Decision tree induction*

The decision tree algorithm is one of the most widely used data mining algorithms. It is an induction learning algorithm based on the training data, which has the advantages of simplicity, transpar- ency and ability to extract decision rules. A decision tree is a clas- siﬁer that can be expressed as a recursive partition of the instance space ([Rokach & Maimon, 2008; Weihong et al., 2006](#_bookmark33)). The learn- ing system of a typical decision tree adopts a top–down strategy which ensures a simple tree but not necessarily the simplest tree will be found. The decision tree consists of nodes having only one incoming edge. A node with outgoing branches is referred to as a ‘‘test’’ node while all other nodes are called ‘‘terminal nodes’’. Each test node splits the instance space into two or more sub- spaces according to the attribute values. Each terminal node is assigned to one class that represents the most appropriate target value. Every path to the terminal node in the decision tree repre- sents a classiﬁcation rule. The key to construct an efﬁcient decision tree is to select good splitting criteria. Gini diversity index is cho- sen as splitting criterion. The Gini impurity measure *d*(*t*) at node *t* is calculated as follows:

*i*ð*t*Þ¼ 1 — *S*

P

where *S* (the impurity criteria) = *p*2(*j*|*t*), for *j* = 0, 1, 2,.. . *k*. *k* de- notes the number of classes existing in that node and *p*(*j*|*t*) corre- sponds to the relative frequency of class *j* in node *t*. The Gini diversity index of a node attains its maximum value when all the classes in the node occur with equal probability and is minimal

Fig. 1. Impact of reduced error pruning on accuracy.

when the node contains only one class ([Breiman et al., 1984; Sun &](#_bookmark35) [Clark, 2009](#_bookmark35)).

Generally speaking, a decision tree that is not complex is pref- erable since tree complexity has a serious effect on its accuracy. Tree complexity is controlled by the employed pruning method. Decision-tree pruning is the main task which simpliﬁes the deci- sion tree by discarding one or more parts of the tree (sub trees) and replacing them with the terminal nodes. The reduced error pruning (REP) method has been implemented in this study. This method, proposed by [Quinlan (1987)](#_bookmark27) is a conceptually simple and understandable method in decision tree pruning. For every sub-tree *T* with no terminal nodes in the original decision tree, the change in the misclassiﬁcation error over the test set is exam- ined. Misclassiﬁcation errors would occur if this sub-tree is re- placed by the most frequent class. If the error rate of the new tree would be equal to or smaller than that of the original decision tree, the *T* is replaced by that most frequent class. This process con- tinues until any extra pruning would drastically decrease the accu- racy. The main advantage of this method is its linear computational complexity since each node in the tree is visited only once to deter- mine the opportunity of pruning it ([Patil et al., 2010](#_bookmark23)).

IF (T.S ≤ 4.5 and F.S > 0.5 and C.P ≤ 3.5)

IF (T.S ≤ 4.5 and F.S > 0.5 and C.P > 3.5) IF (T.S ≤ 4.5 and F.S ≤ 0.5)

THEN

THEN

`CAD

normal

THEN

normal

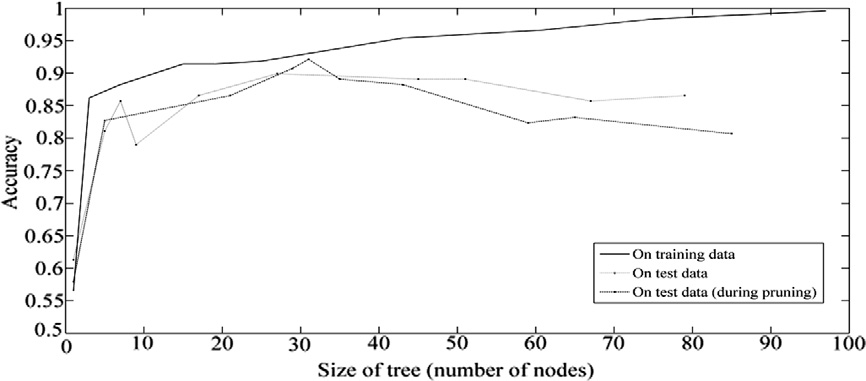
The problem with REP is its bias towards over pruning when the test set is much smaller than the training set but becomes less appropriate when the number of cases in the test set increases. The performance of REP is found to be better in terms of accuracy and size when compared with other methods ([Esposito et al.,](#_bookmark12) [1997](#_bookmark12)).

The impact of REP on the accuracy of the decision tree is illus- trated in [Fig. 1](#_bookmark3). When pruning begins, the tree is at its maximum size and lowest accuracy over the test data. As pruning proceeds, the number of nodes is reduced and accuracy over the test data increases.

The crisp set of rules from the decision tree is shown in [Fig. 2](#_bookmark4).

1. Fuzzy inference system
   1. *Development of a fuzzy model*

A fuzzy model is based on three basic aspects: the fuzziﬁcation process, fuzzy inference system (FIS) and defuzziﬁcation process. Different combinations of the realization of the aforementioned



|  |  |  |
| --- | --- | --- |
| IF (T.S > 4.5 and Slope ≤ 1.5 and CP > 3.5 and Chol ≤ 240.5 and F.S ≤ 0.5) | THEN | normal |
| IF (T.S > 4.5 and Slope ≤ 1.5 and CP > 3.5 and Chol ≤ 240.5 and F.S > 0.5) | THEN | CAD |
| IF (T.S > 4.5 and Slope > 1.5 and CP ≤ 3.5 and ECG ≤ 1.5 and HR ≤ 188) | THEN | CAD |
| IF (T.S > 4.5 and Slope > 1.5 and CP ≤ 3.5 and ECG ≤ 1.5 and HR > 188) | THEN | normal |
| IF (T.S > 4.5 and Slope > 1.5 and CP ≤ 3.5 and ECG > 1.5 and F.S ≤ 0.5) | THEN | normal |
| IF (T.S > 4.5 and Slope > 1.5 and CP ≤ 3.5 and ECG > 1.5 and F.S > 0.5) | THEN | CAD |
| IF (T.S > 4.5 and Slope ≤ 1.5 and CP ≤ 3.5 and BP ≤ 182 and O.P ≤ 2.4) | THEN | normal |
| IF (T.S > 4.5 and Slope ≤ 1.5 and CP ≤ 3.5 and BP ≤ 182 and O.P > 2.4) | THEN | CAD |
| IF (T.S > 4.5 and Slope ≤ 1.5 and CP > 3.5 and Chol > 240.5) | THEN | CAD |
| IF (T.S > 4.5 and Slope ≤ 1.5 and CP ≤ 3.5 and BP > 182) | THEN | CAD |
| IF (T.S > 4.5 and Slope > 1.5 and CP > 3.5) | THEN | CAD |

T.S = Thallium Scan, CP = Chest Pain type, BP = Blood Pressure, O.P = Old Peak, HR = Heart Rate, F.S =

Fluoroscopy, ECG = Resting ECG, Chol = Serum Cholesterol

Fig. 2. Indicative crisp rules.

Table 3

Heart Disease

Data

Missing Data

Imputation

Fuzzy Inference

System

Class Label

Classiﬁcation of serum cholesterol.

Input ﬁeld Range Fuzzy set

Serum cholesterol 0–198 Low 188–250 Medium

Membership Function

Optimized

Knowledge Base

217–307 High

281–681 Very high

(From Decision Tree)

Fuzzy Rule Base

Table 1

Classiﬁcation of age.

|  |  |  |
| --- | --- | --- |
| Input ﬁeld | Range | Fuzzy set |
| Age | 0–34  30–35  40–58 | Low Medium Old |
|  | 52–77 | Very old |

Fig. 3. Proposed fuzzy expert system.

Table 4

Classiﬁcation of maximum heart rate.

Input ﬁeld Range Fuzzy set

Maximum heart rate 0–141 Low 111–194 Medium

153–353 High

Table 2

Classiﬁcation of blood pressure.

|  |  |  |
| --- | --- | --- |
| Input ﬁeld | Range | Fuzzy set |
| Blood pressure | 0-134 | Low |
|  | 126-154 | Medium |
|  | 142-172 | High |
|  | 154-354 | Very high |

aspects yield different fuzzy models. The design process of a fuzzy model consists of determining the following tasks:

1. The input and output variables;
2. The fuzzy membership functions for each variable;
3. The fuzzy rules and (4) The parameters in (2) and (3)

Tasks (1), (2) and (3) are associated with the structure of the fuzzy model while task (4) is related to the tuning of the parame- ters of the model ([Nawa et al., 1999](#_bookmark19)). [Fig. 3](#_bookmark5) illustrates the various components of the proposed fuzzy expert system.

Mamdani fuzzy inference system has been employed and for the fuzziﬁcation process, the crisp set of rules is transformed into a fuzzy model using a triangular membership function. The mem- bership deﬁnition for a triangular membership function is given as follows:

8**>**

0 *x* < *a*

ð Þ¼ <

Triangle *x* : *a*; *b*; *c* ð*x* — *a*Þ=ð*b* — *a*Þ *a* ≤ *x* ≤ *b*

ð*c* — *x*Þ=ð*c* — *b*Þ *b* ≤ *x* ≤ *c*

**>**:

0 *x* > *c*

This can also be represented as follows:

Triangleð*x* : *a*; *b*; *c*Þ¼ maxðminð*x* — *a*Þ=ð*b* — *a*Þ; ð*c* — *x*Þ=ð*c* — *d*Þ; 0ÞÞ

where the parameters a and b denote the lower and upper bounds respectively while *c* locates the peak of the triangle.

Thus crisp rules can be represented by fuzzy rules with triangu- lar membership function. Basically, there are two reasons for trans- forming crisp rule-base into a fuzzy rule-base:

1. Fuzzy rule-base in comparison with the crisp rule-base con- tain additional information about the certainty degree of the

classiﬁer decision.

1. Fuzzy systems can easily deﬁne non-axis parallel decision boundaries, while the crisp rule base system approximates in the step wise manner ([Abonyi and Roubos, 2000](#_bookmark28)).

For the defuzzication process, center of gravity (COG) is em- ployed, which is the most commonly used and capable of produc- ing very accurate results ([Shi et al., 1999](#_bookmark41)).

As mentioned above, the important section of the fuzzy model is the fuzzy membership functions of each attribute. For this fuzzy model, there are 13 input variables and one output variable. They are given as follows.

Age: This input variable has been divided into 4 fuzzy sets namely ‘‘Young’’, ‘‘Middle’’, ‘‘Old’’ and ‘‘Very old’’. Membership functions of these fuzzy sets are triangular. These fuzzy sets are shown in [Table 1](#_bookmark5).

Blood pressure: This input variable is divided into four fuzzy sets. They are ‘‘Low’’, ‘‘Medium’’, ‘‘High’’ and ‘‘Very high’’. Member- ship functions of these fuzzy sets are triangular. They are shown in [Table 2](#_bookmark6).

Serum cholesterol: This input variable has four fuzzy sets (Low, Medium, High and Very high). Membership functions of these fuzzy sets are triangular. They are shown in [Table 3](#_bookmark7).

Maximum heart rate: In this ﬁeld, there are three fuzzy sets (Low, Medium and High). They are triangular membership func- tions. In [Table 4](#_bookmark7), it has been shown.

Sex: This input ﬁeld has two values (0 and 1) which correspond to female and male respectively.

Chest pain type: This input ﬁeld supports four chest pain types.

Each chest pain type is a fuzzy set. In this ﬁeld, fuzzy sets do not have overlap. Chest Pain types with their values have been shown below.

1 = Typical Angina 2 = Atypical Angina

3 = Non-anginal pain 4 = Asymptomatic

Fasting blood sugar: This input ﬁeld has two values (1 and 0) which correspond to the amount of blood sugar higher than 120 and lesser than 120 respectively.

Resting ECG: In this ﬁeld, there are three values (0, 1 and 2) which correspond to normal, ST-T abnormality and left ventricular hypertrophy.

Exercise Induced Angina: This input ﬁeld has just two values (0 and 1).When the exercise induced angina is present, it corre- sponds to 1. Otherwise, it corresponds to 0.

Old Peak: This input ﬁeld has 2 fuzzy sets (low and high). These fuzzy sets have been shown in [Table 5](#_bookmark8) with their ranges.

Slope: This input ﬁeld has three values (0, 1 and 2) which cor- respond to upsloping, ﬂat and downsloping respectively.

Fluoroscopy: This input ﬁeld has four values (0, 1, 2 and 3) which correspond to the number of blood vessels coloured by ﬂuoroscopy.

Table 5

Classiﬁcation of old peak.

Table 6

PSO parameters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Input ﬁeld | Range | Fuzzy set |  | Parameter | Value |
| Old peak | 0–4.2 | Low |  | *C*1 | 0.5 |
|  | 2.55–7 | High |  | *C*2 | 0.8 |
|  | | | | Population size | 50 |
| *Wmin* | 0.1 |
| *Wmax* | 0.9 |
| Maximum generation | 300 |

µ



C’

k

C

L’

R’

***acer***

*2018-05-17 22:24:11*

--------------------------------------------

note

L R

Fig. 4. Fuzzy membership parameters.

Thallium scan: This ﬁeld has three values (3, 6 and 7) which correspond to normal, ﬁxed defect and reversible defect.

The output variable is the angiographic disease status. It has two values (0 and 1) which correspond to healthy and heart dis- ease conditions.

* 1. *Fuzzy membership optimization*

Formation of desired fuzzy rule base is the crucial step in designing the fuzzy system. In general, the rules and membership function are generated by experts in the particular area, since the deﬁnition of these is generally affected by subjective decisions. While fuzzy rules are relatively easy to derive by them, the MFs are difﬁcult to obtain. Tuning of MFs is a time consuming process. From the foregoing discussion, it can be observed that fuzzy system is characterized by its membership function and the perfor- mance of the system is determined by the type and parameters of MFs. Despite their importance, there are no observational methods

available for determining them.

The fuzzy systems can be formulated as a space search problem, where each point in the space corresponds to a rule set and MFs. This makes evolutionary algorithms such as Genetic Algorithms

(GAs), particle swarm optimization (PSO), better choices for

searching these spaces ([Cordon et al., 2001, 2004; Herrera et al.,](#_bookmark12) [1998; Park et al., 1994](#_bookmark12)).

Though PSO is similar to GAs, the main difference between them is that PSO does not have genetic operators such as crossover and mutation. Particles in PSO have memory which is important to

the algorithm. When compared to GAs, the advantages of PSO are

simplicity in implementation and fewer parameters to adjust

([Jones, 2005](#_bookmark14)).

PSO refers to a relatively new family of evolutionary algorithms which can be used to ﬁnd optimal solutions to the numeric prob- lem using a population of particles. The PSO technique was devel- oped by [Kennedy and Eberhart (1995)](#_bookmark17) as a population based on stochastic optimization strategy. It is inspired by social behaviour of ﬂocking birds, school of ﬁsh, swarm of bees and even human so- cial behaviour. A performance evaluation parameter called ﬁtness value is used to guide the search process of the particles.

The particle which is associated with the ﬁtness value seems to be the leader and each particle keeps track of its coordinates in the

search space. This ﬁtness value is stored and is called as pbest (personal best). Another ‘‘best’’ value that is tracked by the swarm is the best value, obtained so far by any particle in the neighbours of the particle (local best lbest). When a particle takes all the par-

ticles in the populations as its topological neighbours, the best va-

lue is gbest. The values of pbest and gbest are updated for every time instant inﬂuences the interactions between the particles and the search process ([Li and Engelbrecht, 2007; Paquet and](#_bookmark18) [Engelbrecht, 2003](#_bookmark18)).

For each fuzzy membership function, there are three parame- ters as shown in ([Fig. 4](#_bookmark8)): C (centre), L (left) and R (right) corre- sponds to the original membership function, where C0 , L0 and R0 refers to the centre, left and right of the adjusted membership function.

For the adjustment of membership functions the following equations are deﬁned:

*C*0 ¼ ð*C* þ *ki*Þ— *wi L*0 ¼ ð*L* þ *ki*Þ— *wi R*0 ¼ ð*R* þ *ki*Þ— *wi*

Being *ki* and *wi* adjustment coefﬁcients, *ki* makes each membership function move the membership function left or right with no distor- tion in the form. The membership function shrinks or expands through the parameter *wi*. These parameters take any integer either positive or negative value. PSO with inertia weight will be used to ﬁnd the optimum values for *ki* and *wi* for the membership functions. [Table 6](#_bookmark9) shows the PSO parameter set for tuning the fuzzy mem-

bership functions.

The membership functions of the attributes such as age, blood pressure, maximum heart rate, old peak and serum cholesterol have been shown in [Fig. 5](#_bookmark10).

1. Results and discussion

After tuning the deﬁned MFs and generating the fuzzy rule base, the fuzzy toolbox available in MATLAB 7 was used for building FIS. Rule viewer of the generated FIS is shown in [Fig. 6](#_bookmark11).

The rules were obtained from a training data set of 478 in- stances for which 278 instances were healthy and 200 instances were heart disease condition.

For testing the built fuzzy model a portion of the data (119 in- stances) called testing data was used. Among the 119 instances for testing, 74 instances were healthy and the remaining 45 instances were heart disease condition. Using the test set, the created FIS was evaluated and its performance was given as confusion matrix (C.M.). [Table 7](#_bookmark15) shows the C.M. for the test set. The entries of the

C.M. are given as follows

¼

### C:M: TP FP FN TN

where TP, TN, FP and FN are the number of true positives, true neg- atives, false positives, and false negatives respectively. TP: disease status predicted as healthy when it actually is healthy. TN: disease

11662 *S. Muthukaruppan, M.J. Er / Expert Systems with Applications 39 (2012) 11657–11665*

status predicted as heart disease when it actually is heart disease condition. FP: disease status predicted as healthy when it actually is heart disease condition. FN: disease status predicted as heart dis- ease condition when it actually is healthy.

The number of correctly classiﬁed instances is shown in the diagonal elements in the C.M. In the ﬁrst row, the ﬁrst element shows the number of instances belonging to healthy and classiﬁed by FIS as healthy condition. The second element in the second row shows the instances belonging to heart disease and classiﬁed by FIS as heart disease condition. Sensitivity, speciﬁcity and accuracy based on the C.M. is given as follows

### Sensitivity ¼ ðTP=ðTP þ FNÞÞ ¼ 93:2% Specificity ¼ ðTN=ðFP þ TNÞÞ ¼ 93:3% Accuracy ¼ ðTP=ðTP þ FN þ FP þ TNÞÞ ¼ 93:3%

Sensitivity is thus a measure of accuracy of FIS of healthy condition instances, and speciﬁcity is a measure of accuracy of FIS of heart disease condition instances. Accuracy is the number of correctly classiﬁed instances divided by the total number of classiﬁcations.

The proposed particle swarm optimization (PSO)-based fuzzy expert system results have been compared with the similar

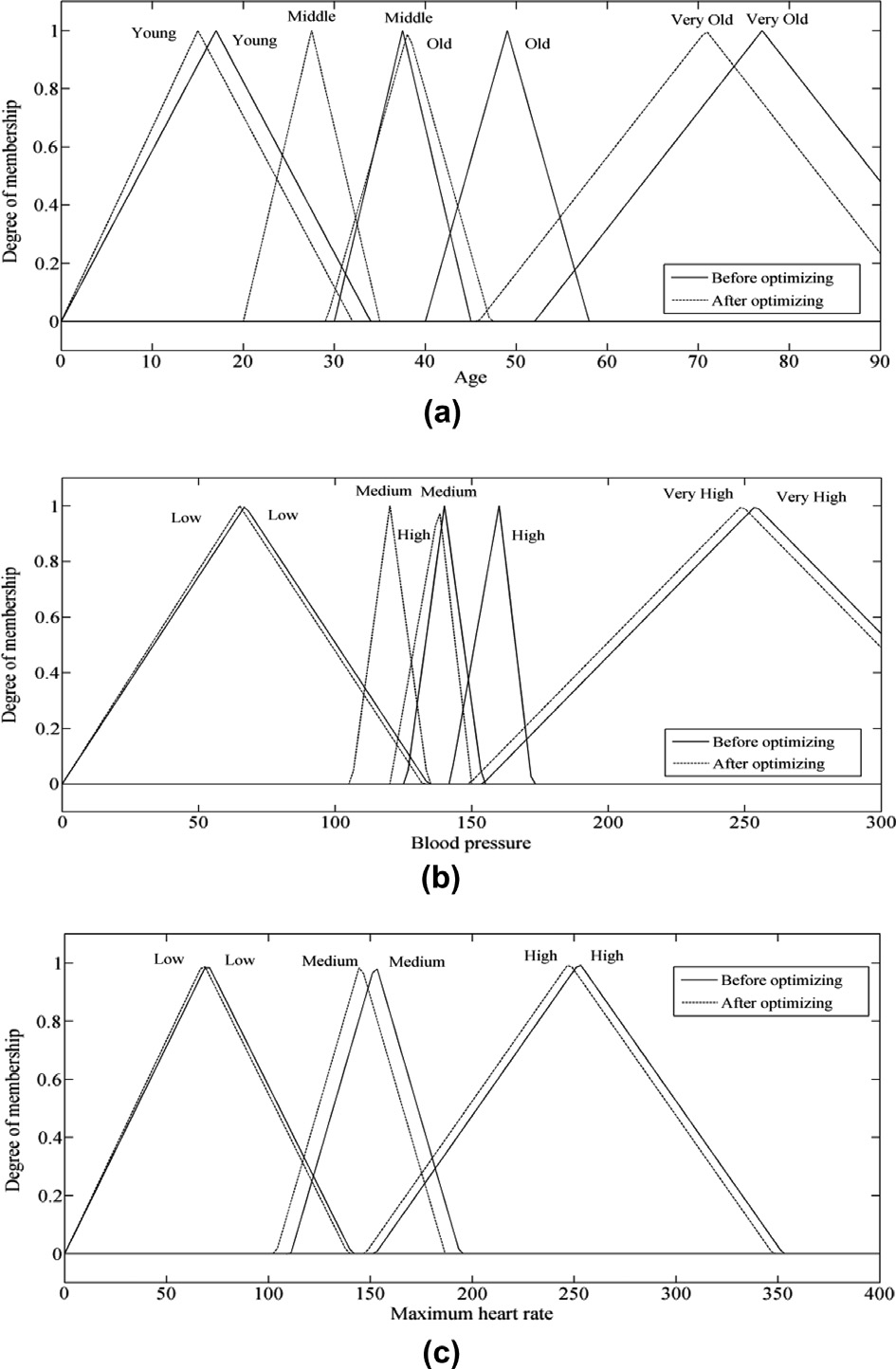


Fig. 5. Membership functions of (a) age (b) blood pressure (c) maximum heart rate (d) old peak and (e) serum cholesterol.

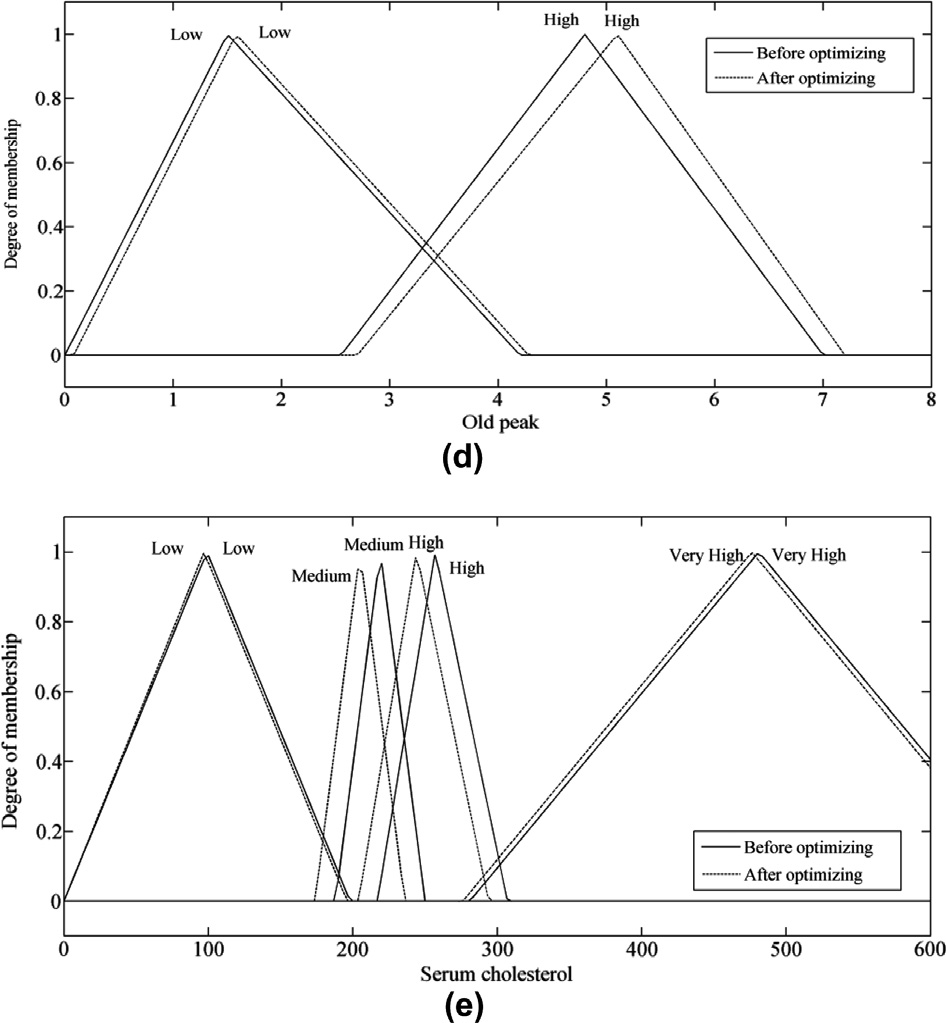


Fig. 5 (*continued*)

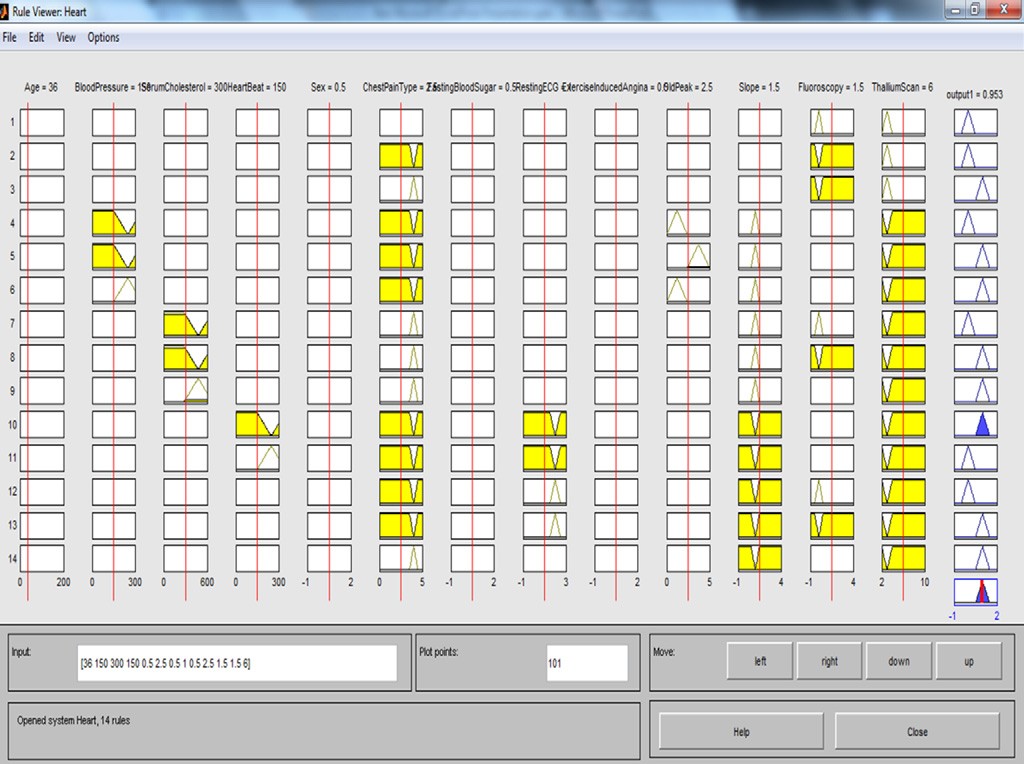


Fig. 6. Rule viewer for one of the test data.

11664 *S. Muthukaruppan, M.J. Er / Expert Systems with Applications 39 (2012) 11657–11665*

Table 7

Confusion matrix for test data.

|  |  |  |
| --- | --- | --- |
| Predicted | Actual |  |
|  | Normal | Heart disease |
| Normal | 69 | 3 |
| Heart disease | 5 | 42 |

Table 8

Comparison of the proposed system outcome with the similar researches.

Author Method Accuracy

|  |  |  |
| --- | --- | --- |
| ([Cheung, 2001](#_bookmark12)) | BNNF | 80.96 |
| ([Cheung, 2001](#_bookmark12)) | C 4.5 | 81.11 |

([Cheung, 2001](#_bookmark12)) BNND 81.11

([Cheung, 2001](#_bookmark12)) Naïve Bayes 81.48

Cheung, N. (2001). Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queenland.

Cordon, O. et al. (2001). Generating the knowledge base of a fuzzy rule-based system by the genetic learning of data base. *IEEE Transactions on Fuzzy Systems*, 667–674.

Cordon, O. et al. (2004). Ten years of genetic fuzzy systems: Current framework and new trends. *Fuzzy Sets and Systems, 141*, 5–31.

Esposito, F. et al. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*, 476–491. Farhangfar, A. et al. (2004). Experimental analysis of methods for imputation of missing values in databases. In *Presented at the intelligent computing: theory and*

*applications II, Orlando, FL, USA*, (pp. 172–182).

Ganesh Kumar, P. et al. (2012). Design of fuzzy expert system for microarray data classiﬁcation using a novel genetic swarm algorithm. *Expert Systems with Applications, 39*, 1811–1821.

Grzymala-Busse, J.W., & Hu, M. (2001). A comparison of several approaches to missing attribute values in data mining. In *Revised papers from the second international conference on rough sets and current trends in computing, London* (pp. 378–385).

([Ster and Dobnikar, 1996](#_bookmark41)) Fisher discriminant

Analysis

84.2

Herrera, F. et al. (1998). A learning process for fuzzy control rules using genetic algorithms. *Fuzzy Sets and Systems*, 143–158.

([Ster & Dobnikar, 1996](#_bookmark41)) LDA 84.5

([Ster & Dobnikar, 1996](#_bookmark41)) Naïve Bayes 82.5–

83.4

Jankowski, N., & Kadirkamanathan, V. (1997). Statistical control of rbf-like networks for classiﬁcation. In *Proc. 7th international conference on artiﬁcial, neural networks* (pp. 385–390).

([Kahramanli & Allahverdi, 2008](#_bookmark16)) Hybrid neural network

system

([Polat et al., 2007](#_bookmark26)) Fuzzy-AIRS-Knn based system

86.8

87.00

Jerez, J. M. et al. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artiﬁcial Intelligence in Medicine, 50*, 105–115.

Jones, K.O. (2005). Comparison of genetic algorithm and particle swarm

([Resul et al., 2009](#_bookmark30)) Neural network ensembles 89.01 ([Jankowski and Kadirkamanathan,](#_bookmark12) IncNet 90.00

[1997](#_bookmark12))

([Senthil Kumar, 2011](#_bookmark37)) ANFIS 91.18

optimization. In *Proc. international conference on computer systems and*

*technologies* (pp. IIIA1–6).

Jonsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. In *Proc. software metrics, 10th international, symposium* (pp. 108–118).

([Senthil Kumar, 2012](#_bookmark39)) Fuzzy resolution

mechanism

Proposed system PSO based fuzzy expert system

91.83%

93.27%

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart disease. *Expert Systems with Applications, 35*, 82–89.

Kennedy, J., & Eberhart, R.C. (1995). Particle swarm optimization. In *Proc. 1995 IEEE international conference on neural networks, Australia* (pp. 1942–1948).

Li, X., & Engelbrecht, A.P. (2007). Particle swarm optimization: An introduction and its recent developments. In *Presented at the 2007 GECCO conference companion**on Genetic and evolutionary computation, London* (pp. 3391–3414).

researches ([Table 8](#_bookmark15)) which reveals that the proposed method achieves the higher accuracy.

1. Conclusions

In this study, a fuzzy expert system based on particle swarm optimization (PSO) was developed in Matlab’s Simulink in order to classify heart disease and healthy condition. With this proposed approach, 93.27% correct classiﬁcation on the test set could be achieved. The discovery of the signiﬁcant attributes and fuzzy rules was achieved using the decision tree algorithm. The importance of discovering signiﬁcant and relevant fuzzy rules without the aid of the experts opens the possibility of knowledge discovery. The main advantages of the FES as a knowledge acquisition tool are the fol- lowing: (1) a small number of rules are obtained (2) the obtained rules can be easily interpreted. These results imply promising re- search areas employing decision trees and fuzzy expert system in several classiﬁcation problems.

References

Abonyi, J., & Roubos, H. (2000). Structure identiﬁcation of fuzzy classiﬁers. In *5th online world conference on soft computing in industrial applications (WSCS)*.

Adeli, A., & Neshat, M. (2010). A fuzzy expert system for heart disease diagnosis. In *Proc. international multiconference of engineering and computer scientists* (pp. 134–139).

American Heart Association, Heart Disease and Stroke Statistics – 2010 Update at a Glance. Dallas: American Heart Association (2010).

Andreeva, P. (2006). Data modelling and speciﬁc rule generation via data mining techniques. In *Proc. international conference on computer systems and technologies* (pp. IIIA.17–23).

Breiman, L. et al. (1984). *Classiﬁcation and regression trees*. Belmont, CA: Wadsworth International Group.

British Heart Foundation, Coronary Heart Disease Statistics – 2003 edition. London: British Heart Foundation (2003).

Brown, M. I., & Kros, J. F. (2003). Data mining and the impact of missing data.

*Industrial Management and Data Systems, 103*, 611–621.

Nawa, N. E. et al. (1999). Study on the discovery of relevant fuzzy rules using pseudobacterial genetic algorithm. *IEEE Transactions on Industrial Electronics, 46*, 1080–1089.

Newman, D. J., Hettich, S., Blake, C. L. S., & Merz, C. J. (1998). *UCI repository of*

*machine learning database*. Irvine, CA: University of California.

Omran, A. (1979). Changing patterns of health and disease during the process of national development. In G.L. Albrecht, P.C. Higgins (Eds.), *Health, illness, and medicine: a reader in medical sociology*, Chicago: Rand McNally.

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system

using data mining techniques. In *Proc. 2008 IEEE/ACS international conference on computer systems and applications* (pp. 108–115).

Paquet, U., & Engelbrecht, A.P. (2003). Training support vector machines with particle swarms. In *Proc. international joint conference on neural networks (IJCNN) conference* (pp. 1593–1598).

Park, D. et al. (1994). Genetic-based new fuzzy reasoning models with application to fuzzy control. *IEEE Transactions on Systems, Man and Cybernetics, 24*, 39–47. Patil, D. D. et al. (2010). Evaluation of decision tree pruning algorithms for complexity and classiﬁcation accuracy. *International Journal of Computer*

*Applications, 11*, 23–30.

Intelligent and effective heart attack

Patil, S. B., & Kumaraswamy, Y. S. (2009).

prediction system using data mining and artiﬁcial neural network. *European Journal of Scientiﬁc Research, 31*, 642–656.

Pedrycz, W., & Sosnowski, Z. A. (2005). Genetically optimized fuzzy decision trees. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 35*, 633–651.

Polat, K. et al. (2007). Automatic detection of heart disease using an artiﬁcial immune recognition system (AIRS) with fuzzy resource allocation mechanism and K-Nn (nearest neighbour) based weighting preprocessing. *Expert Systems with Applications, 32*, 625–631.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man– Machine Studies, 27*, 221–234.

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

Resul, D., Ibrahim, T., & Abdulkadir, S. (2009). Effective diagnosis of heart disease

through neural networks ensembles. *Expert Systems with Applications, 36*, 7675–7680.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees theory and applications*. Singapore: World Scientiﬁc Publishing.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys: J*. New York: Wiley & Sons.

Senthil Kumar, A.V. (2011). Adaptive neuro-fuzzy inference system for heart disease diagnosis. In *Proc. international conference on information system, computer engineering and application*.

Senthil Kumar, A.V. (2012). Diagnosis of heart disease using fuzzy resolution mechanism. *Journal of Artiﬁcial Intelligence*, pp. 1–9, [http://dx.doi.org/10.3923/](http://dx.doi.org/10.3923/jai.2012) [jai.2012](http://dx.doi.org/10.3923/jai.2012).

Shi, Y. et al. (1999). Implementation of evolutionary fuzzy systems. *IEEE Transactions on Fuzzy Systems, 7*, 109–119.

Ster, B., & Dobnikar, A. (1996). Neural networks in medical diagnosis: Comparison with other methods. In *Proc. international conference on engineering applications of, neural networks* (pp. 427–430).

Su, X. et al. (2008). Using imputation techniques to help learn accurate classiﬁers. In *Proc. IEEE international conference on tools with artiﬁcial intelligence* (Vol. 01, pp. 437–444).

Sun, Y. C., & Clark, O. G. (2009). Implementing an intuitive reasoner with a large weather database. *International Journal of Information and Mathematical Sciences, 5*, 85–102.

Tsipouras, M. G. et al. (2008). Automated diagnosis of coronary artery disease based

on data mining and fuzzy modeling. *IEEE Transactions on Information technology in biomedicine, 12*, 447–458.

Tsoukalas, L. H., & Uhrig, R. E. (1997). *Fuzzy and neural approaches in engineering*.

John Wiley & Sons.

Tu, M.C. et al. (2009). A comparative study of medical data classiﬁcation methods

based on decision tree and bagging algorithms. In *Proc. IEEE 8th international conference on dependable, autonomic and secure, computing* (pp. 183–187).

Ture, M., Kurt, I., & Kurum, A. T. (2008). Comparing performances of logistic

regression, classiﬁcation and regression tree, and neural networks for

predicting coronary artery disease. *Expert Systems with Applications, 34*, 366–374.

Weihong, W. et al. (2006). A preliminary study on constructing decision tree with gene expression programming. In *Proc. 1st international conference on innovative computing, information and control* (pp. 222–225).

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions.

*Journal of Artiﬁcial Intelligence Research (JAIR), 6*, 1–34.

Wilson, P. W. F. et al. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation, 97*, 1837–1847.

Yu, J. (2005). General C-means clustering model. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1197–1211.

Zhang, S. et al. (2005). ‘‘Missing is Useful’’: missing values in cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering, 17*, 1689–1693.