International Journal of Computational Intelligence and Informatics, Vol. 5: No. 2, September 2015

Optimal Data Prediction and Classification Applicable for Intelligent Heart Disease Diagnosis System

K. Jayavani
Department of Computer Applications
Sri Vijay College of Arts and Science
Dharmapuri, Tamilnadu, India
vanigopinath@gmail.com

G. M. KadharNawaz
Department of Computer Applications
Sona College of Technology
Salem, Tamilnadu, India
nawazse@yahoo.co.in

***Abstract*- In the past years, medical data mining has become a popular data mining subject. Researchers have proposed several tools and several methodologies for developing effective medical expert systems. Diagnosing heart diseases is one of the much need topics and many researchers have tried to develop intelligent medical expert systems to help the physicians. In this paper, we proposed a novel, particle separable optimization Algorithm, which is derived from particle swarm optimization and efficient nearest algorithm. Efficient Nearest neighbor (ENN) is very simple, most popular, highly efficient and effective algorithm for pattern recognition. ENN is a straight forward classifier, where samples are classified based on the attributes class of their nearest neighbor. Medical data bases are high volume in nature. If the data set contains redundant and irrelevant attributes, classification accuracy will be degraded. The main objective of this approach is to extract rules for diagnosis the existence or in existence of heart disease in a patient. The PSO composed of three steps, first to extract the feature from the database using pso, second one is classify the feature based on efficient nearest neighbor search, finally the optimization level measured from polynomial and Fourier series regression method. In addition classification accuracy improved by this method compared with conventional PSO.**

Keywords- Particle Swarm Optimization, Efficient Nearest Neighbor, Polynomial, Fourier series, Particle Separable Optimization.

I. INTRODUCTION

Every human body and its physiological processes show some symptoms of a diseased condition. The proposed model in this paper used for identification of the heart diseases using heart sounds. The signal of heart sound carries important physiological and pathological information, it's about the general state of contractile activity of the cardiovascular system. The heart murmurs caused by turbulent blood flow and the incomplete opening or closing of the valves, could be heard clearly sounding like whistling, swishing or humming [10]. The feature selection process can be considered as a problem of global combinatorial optimization in machine learning and statistics. Feature selection, also known as variable selection, feature reduction, and attribute selection or variable subset selection. Therefore, a good feature selection is which speeds up the processing rate, predictive accuracy, and avoids incomprehensibility. Several methods have been previously used to perform feature selection on training and testing data, for example genetic algorithms, catfish binary PSO, improved binary PSO, and support vector machine(SVM) [1]–[3]. In this paper, binary particle swarm optimization (BPSO) is used to implement feature selection, and the K-nearest neighbor (KNN) method with leave-one-out-cross-validations of great significance [1]. The "gold standard" method for the diagnosis of CAD, which is widely used, is coronary angiography (CA). However, CA is a costly and invasive procedure and needs technology and high-level technical experience; therefore, it cannot be used to screen large populations or close follow up of treatment [2]. Hence, in the clinical setting, for the detection of CAD, other noninvasive methods are being used. The most important of those include exercise electrocardiogram (ECG) [3] testing, single photon emission computed tomography (SPECT or scintigraphy) [4], and stress echocardiography (ECHO), while multislice spiral computerized tomography (MSCT) or electron-beam computerized tomography (EBCT) and coronary magnetic resonance angiography (CMRA) are also being now used [2].

While many people with heart disease have symptoms such as angina, fatigue, and chest pain, many people have no symptoms until a heart attack happens. According to the American Heart Association (AHA), CAD is one of the most important killers of American men and women, reported as the cause for more than one of every five deaths in 2001 [5].

There are many risk factors related to CAD. Some factors such as family history, gender, and age cannot be controlled. However, other risk factors that are associated with lifestyle can often be controlled [6]. For example, physical inactivity, high cholesterol, high blood pressure, and smoking are all considered as risk factors for this disease that can be modified and even, in some cases, eliminated by modifying daily life routines and taking medication. Early changes in lifestyle can significantly prevent diabetes and obesity. The large number of factors that have to be analyzed for diagnosing CAD makes the physician's work even more difficult. In general, physicians make decisions by evaluating the existing test results of the patients. The earlier diagnoses made on other patients with the same condition are also considered by the physicians. These complicated procedures are not easy to perform when considering many factors that the physician has to evaluate. So, the decision about presence or absence of the disease depends on the physician's experience and skill to compare his patient with his previous ones. This procedure is a challenging task regarding the large number of factors that has to be considered. In this complex stage, the doctor may need an accurate tool that lists his earlier decisions about patient having the same (or close to same) factors [7].

During the past decades, the level of interest in the use of data mining and artificial intelligent tools in medical fields and the provision of healthcare has undergone a significant increase. Several sections of the researches in this area are related to developing the diagnostic tools that are used to help physicians in a diagnosis. As an advanced data mining technique, PSO has been applied to many tasks in medicine.

In PSO, particles are available to be adjusted by the learning process. In the research area of rule extraction and pattern recognition, this approach has been widely used. In this paper, we have applied PSO and fuzzy logic with a proposed boosting algorithm for the diagnosis of coronary artery disease.

The boosting mechanism adapts the distribution of training instances in a way that the previously misclassified or uncovered instances are further considered by the PSO algorithm.

The Cleveland, Hungarian, Switzerland, and VA Long Beach data sets, which are taken from Data Mining Repository of University of California, Irvine (UCI), have been used for testing this method [8].

The results show that this method can classify these data sets with acceptable accuracy or even better than the results achieved by previous works. This method is also superior to other methods in terms of interpretability.

This template, modified in MS Word 2007 and above for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout JournalPublication. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. RELATED WORK

K nearest neighbor(KNN) is a simple algorithm, which stores all cases and classify new cases based on similarity measure.KNN algorithm also called as 1) case based reasoning 2) k nearest neighbor 3)example based reasoning 4)instance based learning 5) memory based reasoning 6) lazy learning [4].KNN algorithms have been used since1970 in many applications like statistical estimation and pattern recognition etc.KNN is a non parametric classification method which is broadly classified into two types1) structure less NN techniques 2) structure based NN techniques. In structure less NN techniques whole data is classified into training and test sample data. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, k-tree, axis tree, nearest future line and central line [5].Nearest neighbor classification is used mainly when all the attributes are continuous .Simple K nearest neighbor algorithm is given below.

Steps 1) find the K training instances which are closest to Unknown instance

Step2) pick the most commonly occurring classification for these K instances

There are various ways of measuring the similarity between two instances with n attribute values. Every measure has the following three requirements. Let dist (A, B) be the distance between two points A, B then

1) dist (A, B).__DQG_GLVW_$_%_ __LII_$ %_

2) dist (A, B) = dist (B, A)

3) dist (A, C)._GLVW_$_%___GLVW_%_&_

Property 3 is called as "Triangle in equality". It states that the shortest distance between any two points is a straight line. Most common distance measures used is Euclidean distance .For continuous variables Z score standardization and min max normalization are used [6].KNN is used in many applications such as

1) Classification and interpretation

2) Problem solving

3) Function learning and teaching and training.

KNN suffers from the following drawbacks:

1) Low efficiency

2) Dependency on the selection of good values for k. Further research is required to improve the accuracy of KNN with good values of K.

A. Initialization

By default, particle swarm creates particles at random uniformly within bounds. If there is an unbounded component, particle swarm creates particles with a random uniform distribution from –1000 to 1000. If you have only one bound, particle swarm shifts the creation to have the bound as an endpoint, and a creation interval 2000 wide. Particle *i* has position x (*i)*, which is a row vector with *nvars* elements. Control the span of the initial swarm using the InitialSwarmSpan option.

Similarly, particle swarm creates initial particle velocities *v* uniformly within the range [*-r, r*], where *r* is the vector of initial *ranges*. The range of component *i* is the ub (*i*) - lb (*i*), but for unbounded or semi-unbounded components the range is the InitialSwarmSpan option.

Particle swarm evaluates the objective function at all particles. It records the current position *p*(i) of each particle i. In subsequent iterations, *p* (*i*) will be the location of the best objective function that particle *i* has found. And b is the best overall particles: *b* = min (fun (*p* (i))). d is the location such that *b* = fun (*d*). Particle swarm initializes the neighborhood size *N* to minNeighborhoodSize = max (1, floor (Swarm Size * minFractionNeighbors)). Particle swarm initializes the inertia *W* = max (Inertia Range), or if Inertia Range is negative, it sets *W* = min(Inertia Range). Particle swarm initializes the stall counter *c* = 0. For convenience of notation, set the variable *y1* = Self Adjustment, and *y2* = Social Adjustment, where Self Adjustment and Social Adjustment are options.

B. Algorithm

The algorithm updates the swarm as follows. For particle *i*, which is at position *x(i):*

1. Choose a random subset *S* of *N* particles other than *i*.
2. Find fbest(*S*), the best objective function among the neighbors, and g(*S*), the position of the neighbor with the best objective function.
3. For *u1* and *u2* uniformly (0,1) distributed random vectors of length *nvars*, update the velocity

   $$V = W * v + y1 * u1. * (p - x) + y2 * u2. * (g - x). \quad (1)$$

   This update uses a weighted sum of:

   - The previous velocity v
   - The difference between the current position and the best position the particle has seen *p-x*
   - The difference between the current position and the best position in the current neighborhood *g-x*

4. Update the position $x = x + v$.
5. Enforce the bounds. If any component of *x* is outside a bound, set it equal to that bound.
6. Evaluate the objective function $f = fun(x)$.
7. If $f < fun(p)$, then set $p = x$. This step ensures p has the best position the particle has seen.
8. If $f < b$, then set $b = f$ and $d = x$. This step ensures b has the best objective function in the swarm, and *d* has the best location.
9. If, in the previous step, the best function value was lowered, then set flag = *true*. Otherwise, flag = *false*. The value of flag is used in the next step.
10. Update the neighborhood. If flag = *true*:

- Set *c = max (0, c-1).*
- Set *N* to *min Neighborhood Size.*
- If *c* < 2, then set *W = 2 * W.*
- If *c* > 5, then set *W = W / 2*.
- Ensure that *W* is in the bounds of the Inertia Range option. If flag = *false*:
- Set c = c + 1.
- Set *N* = min (*N + min NeighborhoodSize, SwarmSize*).

C. Optimization

An optimization problem exists when a *decision maker* seeks to maximize or minimize a function value (e.g. profit or cost) and has control of at least some independent variables affecting that function value. If the formula mapping inputs to output is understood and relatively straightforward, the decision maker may optimize inputs mathematically. But if the formula is unknown or mathematically complicated, an optimization algorithm would be a better approach.

The number of inputs or *decision variables* to be optimized constitutes the problem dimensionality, n. The higher the dimensionality, the more complicated the optimization problem becomes. Each position assumed as particles fly through the n-dimensional *search space* is a *decision vector* or candidate solution to the optimization problem. Ideally, the decision maker seeks a *global optimizer*, which is a decision vector producing a level of optimization that cannot be outperformed by any feasible alternative. It is often easier, however, to find a *local optimizer*, which is a relatively well-performing solution whose degree of optimization cannot be outperformed by any other candidate solution in the same vicinity of the n-dimensional *search space*.

D. Particle Swarm Optimizaiton

Particle Swarm Optimization (PSO) is a member of the Swarm Intelligence family of population-based optimizers. It was presented by James Kennedy and Russell Eberhart in 1995: one paper that year focused on the now popular global best model (Gbest PSO) [1] and another on the local or neighborhood best variant of the algorithm (Lbest PSO) [2]. Each particle follows simple *position and velocity update equations*; yet as particles interact, the collective behavior becomes elaborate enough for the swarm to solve complicated optimization problems.

1) Initialization of the Swarm

Each dimension of each particle's initial position vector is randomly selected from a uniform distribution of feasible values (i.e. all feasible values have an equal probability of selection). Velocity vectors are similarly initialized from a uniform distribution on [-vmax, vmax] per dimension using a reasonable maximum velocity.

III. PROPOSED ALGORITHM

Our proposed approach Particle separable optimization derived from KNN and particle swarm optimization to increase the classification accuracy of heart disease data set. We used Efficient Nearest Neighbor search as a goodness measure to prune redundant and irrelevant attributes, and to rank the attributes which contribute more towards classification. Least ranked attributes are removed, and classification algorithm is built based on evaluated attributes. This classifier is trained to classify heart disease dataset as either healthy or sick. Our proposed algorithm consists of two parts

1) First part deals with evaluating attributes using PSO

2) Part two deals with building classifier and measuring accuracy of the classifier

A. Particle separable algorithm

1. Load Heart Database X*n*
2. Normalize data X*n* € [0,1]
3. Preprocessing the Data sets

   X*n1* = {} Xn1 € Xn
4. Apply ENN Feature Subset
5. Attributes are ranked based on their value
6. Classify the data using ENN search

7. improve accuracy using nonconvex neighbour search
8. optimizing using particle separable optimization
9. repeat step 3 & 4
10. get optimization

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IJCII, CI, PU, pu, ci, and cs do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. ENN Algorithm

1. Get Heart set *Di = {D1, D2, D3…. Dn}*
2. Create Training Data set *Tr = {T1, T2, T3 …. Trn}*
3. Get the Test data set *Te = {Te1, Te2, ..Ten}*
4. Compare *Tr &Te*
5. Select *K* value
5. Get similarity
6. Get dissimilarity

OPTIMAL ACCURACY

α (A B) = P (B | A)

INTEREST (A B) = P (A, B) / P (A) P (B)

CONVICTION (A B) = P (A) P (B) / P (A, B)

$I = \{I_1, I2, \ldots I_M\}$

$T_j \varepsilon \{I_1, I1^-\}*\{I_2, I2^-\}*\ldots.*\{I_M.I_M^-\}$

$X^2 = E[X] = 1/n*(I1+I2+..In\}$

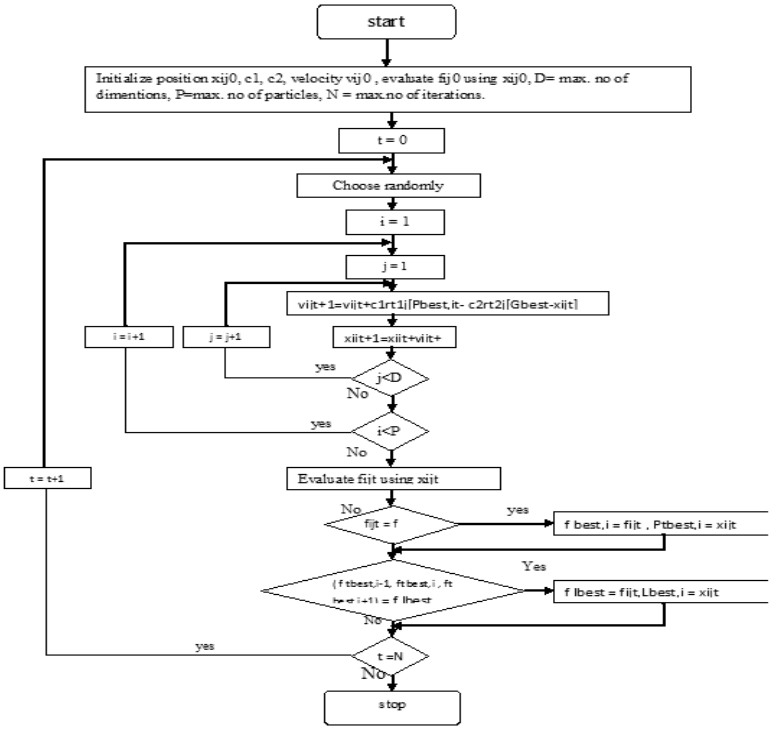

Figure 1. Algorithm flow for PSO

IV. RESULTS AND DISCUSSION

The data set can be divided into two sets namely training data set and test set for experiment. There are five sub-divisions are in the original data set. Four sub-divisions are used for the experiment. Mean and standard deviations are found for training data set and test data set.ENN search give classification accuracy by suitable fitness value of data and find optimization value find the polynomial and Fourier series regression method.

A. Data set

This database contains 13 attributes extracted from a larger set of 75[14].

1) Attribute Information:

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholesterol in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- old peak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by fluoroscopy
- then: 3 = normal; 6 = fixed defect; 7 = reversible defect

2) Attributes types

Real: 1, 4,5,8,10,12

Ordered: 11,

Binary: 2, 6, 9

Nominal: 7,3,13

3) Variable to be predicted

Absence (1) or presence (2) of heart disease

4) Cost Matrix

	abse	*pres*
absence	0	1
presence	5	0

Where the rows represent the true values and the columns the predicted with 270 observation

Linear model Poly2:

$$F(x) = p1 * x \,\hat{}\, 2 + p2 * x + p3 \quad (2)$$

where *x* is normalized by mean 67.78 and std 7.026

Coefficients (with 95% confidence bounds):

p1 = -0.4142 (-0.6159, -0.2125)

p2 = 0.3413 (-0.1357, 0.8183)

p3 = 67.61 (67.13, 68.09)

Goodness of fit:

SSE: 827.4

R-square: 0.1774

Adjusted R-square: 0.1648

RMSE: 2.523

Linear model Poly3:

$$F(x) = p1 * x \wedge 3 + p2 * x \wedge 2 + p3 * x + p4 \quad (3)$$

Coefficients (with 95% confidence bounds):

p1 = 0.0002147 (-0.0006626, 0.001092)

p2 = -0.04909 (-0.2066, 0.1084)

p3 = 3.703 (-5.5, 12.91)

p4 = -24 (-198.9, 150.9)

Goodness of fit:

SSE: 4.392e+005

R-square: 0.03296

Adjusted R-square: 0.01047 RMSE: 58.35

TABLE I. DATA CLASSIFICATION

age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_rate	exercice_angina	disease
43	asympt	140	f	normal	135	yes	positive
39	atyp_angina	120	f	normal	160	yes	negative
39	non_anginal	160	t	normal	160	no	negative
42	non_anginal	160	f	normal	146	no	negative
49	asympt	140	f	normal	130	no	negative
50	asympt	140	f	normal	135	no	negative
59	asympt	140	t	left_vent_hyper	119	yes	positive
54	asympt	200	f	normal	142	yes	positive
59	asympt	130	f	normal	125	no	positive
56	asympt	170	f	st_t_wave_abno	122	yes	positive
52	non_anginal	140	f	st_t_wave_abno	170	no	negative
60	asympt	100	f	normal	125	no	positive
55	atyp_angina	160	t	normal	143	yes	positive
57	atyp_angina	140	t	normal	140	no	negative
38	asympt	110	f	normal	166	no	positive
60	non_anginal	120	f	left_vent_hyper	135	no	negative
55	atyp_angina	140	f	normal	150	no	negative
50	asympt	140	f	st_t_wave_abno	140	yes	positive
48	asympt	106	t	normal	110	no	positive
39	atyp_angina	190	f	normal	106	no	negative
66	asympt	140	f	normal	94	yes	positive
56	asympt	155	t	normal	150	yes	positive
44	asympt	135	f	normal	135	no	positive
43	asympt	120	f	normal	120	yes	positive
54	asympt	140	f	normal	118	yes	positive
52	atyp_angina	140	f	normal	138	yes	negative
48	asympt	120	f	normal	115	no	positive
51	non_anginal	135	f	normal	150	no	positive

General model Fourier1:

$$F(x) = a0 + a1 * \cos(x * w) + b1 * \sin(x * w) \quad (4)$$

Coefficients (with 95% confidence bounds):

a0 = 67.9 (66.65, 69.14)

a1 = -1.044 (-5.415, 3.327)

b1 = 0.2246 (-18.66, 19.11)

w = 0.9364 (0.669, 1.204)

Goodness of fit: SSE: 4.491e+005

R-square: 0.01123 Adjusted R-square: -0.01177

RMSE: 59

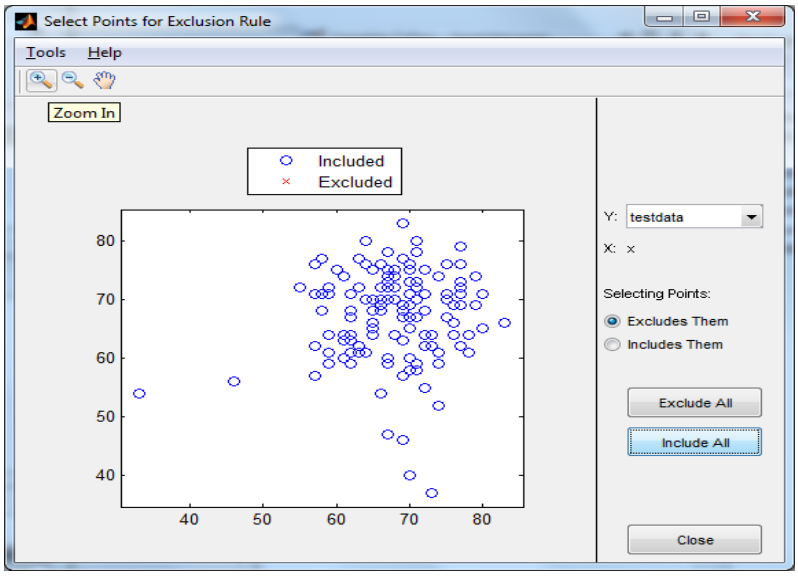

Figure 2. Feature Collection of heart disease dataset

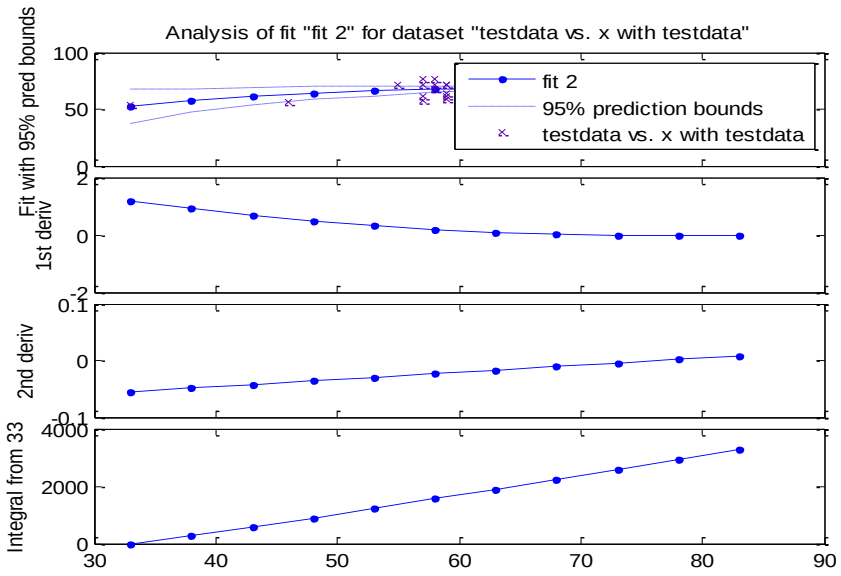

Figure 3. Feature Selection of heart disease dataset

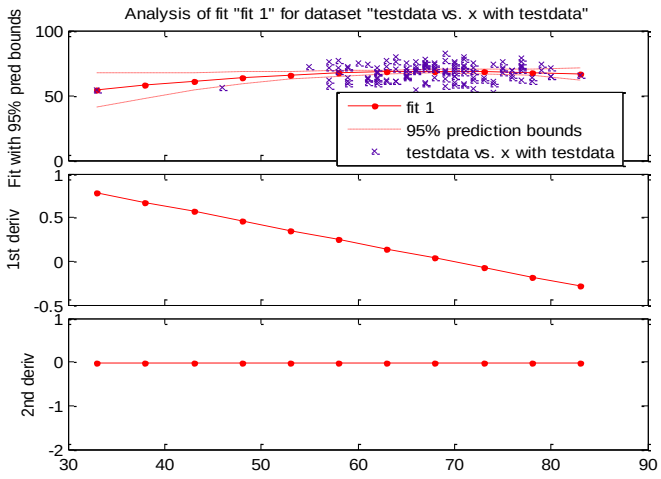

Figure 4. Objective fitness selection data

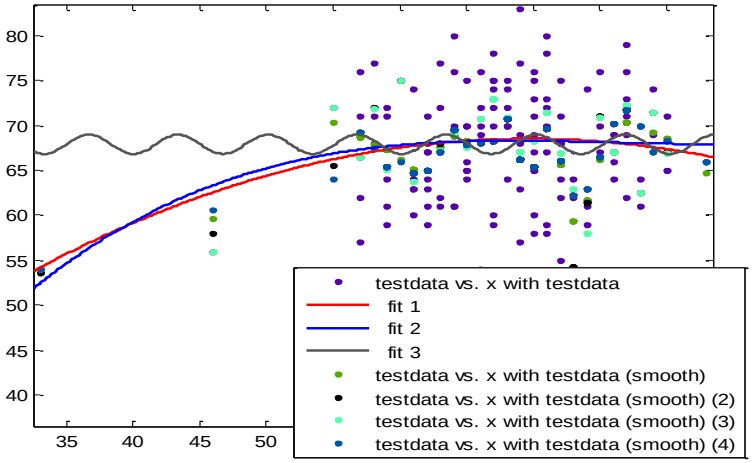

Figure 5. Optimal feature selection in dataset

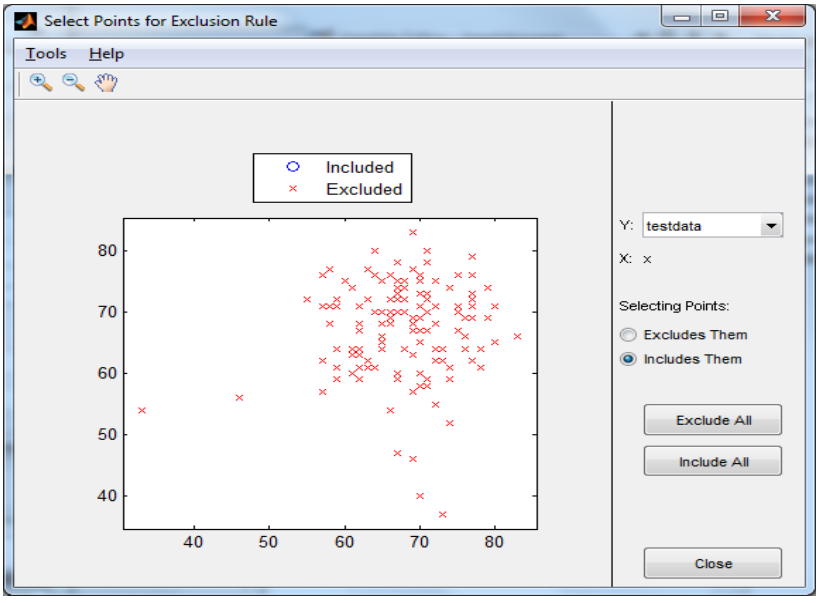

Figure 6. Feature classifications in Test pattern

V. CONCLUSION

Around 18 million people, 7 % Indians are affected by heart disease. Heart disease is mostly affected the person under the age of 65. This paper is based on the heart disease diagnosis of patients. Heart disease is a prevailing disease nowadays. Now due to increasing expenses of heart disease, there was a need to develop a new system which can predict heart diseases in an easy and cheaper way. The diagnosis is based on ENN and PSO. The proposed system feature extraction and classification and optimization characteristics were implemented in MATLAB tool. Results shows classification accuracy increased 15% with conventional methods. The optimization improved existing particle swarm optimization algorithm. Finally proposed method gives best accuracy for diagnose the Heart disease.

REFERENCES

[1] N.Ghadri Hedeshi and M.Saniee Abadeh, "Coronary Artery Disease Detection Using a Fuzzy-Boosting PSO Approach", Hindawi- Computational Intelligence and Neuroscience, 2014.

[2] Michael W.Berry et.al, "Lecture notes in data mining",World Scientific(2006) Michael W.Berry et.al,"Lecture notes in data mining",World Scientific, 2006.

[3] E.Fix and J.Hodges, "Dicriminatory analysis ,non parametric discrimination:consistency properties", Technical report 4,USA,School of aviation medicine Randolph field texas, 1951.

[4] D.E Goldberg, "Genetic algorithm in search .optimization and machine learning"Addison wesley, 1989.

[5] Nitin Bhatia , vandana, "Survey on nearest neighbor techniques", IJCSIS,Vol. 80, No. 2, 2010.

[6] Max bramer, "Principles of data mining", Springer, 2007.

[7] S.N Sivanandam, S.N Deepa, "Introduction to genetic algorithms", Springer, 2008.

[8] MA.Jabbar, B.L Deekshatulu, Priti chandra,"Heart disease prediction system using associative classification and genetic algorithm", pp. 183-192, Elsevier, 2012.

[9] MA.Jabbar, B.L Deekshatulu, Priti chandra, "An evolutionary algorithm for heart disease prediction"CCIS,PP 378-389 , Springer, 2012.

[10] MA.Jabbar, B.L Deekshatulu, Priti chandra, "Prediction of Risk Score for Heart Disease using Associative classification and Hybrid Feature Subset Selection", In .Conf ISDA, pp. 628-634, IEEE, 2013.

[11] MA.Jabbar, B.L Deekshatulu,Priti chandra, "Knowledge discovery from mining association rules for heart disease prediction", pp. 45-53, Vol. 41, No. 2, JATIT, 2013.

[12] MA.Jabbar, B.L deekshatulu,priti chandra, "classification of heart disease using ANN and feature subset selection", GJCST, Vol. 13,issue 3, version1.0 pp. 15-25, 2013.