Biomedical Research 2017; 28 (9): 4154-4158 *ISSN 0970-938X*
www.biomedres.info

Prediction of heart disease using k-nearest neighbor and particle swarm optimization.

Jabbar MA*

Vardhaman College of Engineering, Hyderabad, India

Abstract

Heart disease commonly occurring disease and is the major cause of sudden death nowadays. This disease attacks the persons instantly. Most of the people do not aware of the symptoms of heart disease. Timely attention and proper diagnosis of heart disease will reduce the mortality rate. Medical data mining is to explore hidden pattern from the data sets. Supervised algorithms are used for the early prediction of heart disease. Nearest Neighbor (KNN) is the widely used lazy classification algorithm. KNN is the most popular, effective and efficient algorithm used for pattern recognition. Medical data sets contain a large number of features. The Performance of the classifier will be reduced if the data sets contain noisy features. Feature subset selection is proposed to solve this problem. Feature selection will improve accuracy and reduces the running time. Particle Swarm Optimization (PSO) is an Evolutionary Computation (EC) technique used for feature selection. PSO are computationally inexpensive and converges quickly. This paper investigates to apply KNN and PSO for prediction of heart disease. Experimental results show that the algorithm performs very well with 100% accuracy with PSO as feature selection.

Keywords: Medical data mining, Heart disease, KNN, Feature selection, Particle swam optimization.

Accepted on February 13, 2017

Introduction

Coronary Heart Disease (CHD) is obstruction of the coronary arteries with symptoms such as angina, chest pain, and heart attacks. Arteries supply blood to heart muscle. CHD is a leading cause of death in many countries. In India there are roughly 3 crore heart patients and 2 lakh open heart surgeries are performed every year [1]. CHD is a leading cause of mortality claiming nearly 17.3 million people every year. The reason for this is smoking, high levels of cholesterol, diabetes. Early prediction of heart disease is essential to reduce the mortality rate. Data mining provides a user-oriented approach to extract novel and uncovered patterns in the data set. Data mining is to extract useful knowledge within medical data for medical diagnosis [2]. Data mining is widely applied in the medical domain. Medical data mining is used to infer diagnostic rules and help physicians to make diagnosis process more accurate [3]. K-nearest neighbor is the most widely used lazy classification algorithm as it reduces misclassification error [4]. Feature Subset Selection (FSS) is widely used in data mining and machine learning. FSS is a dimensionality reduction technique use to enhance accuracy [5]. Particle swarm optimization is an effective EC technique used as feature selection [6]. PSO converges quickly and is computationally inexpensive.

This paper investigates by applying KNN and PSO to predict heart disease. PSO is used as feature selection measure. Medical data contains the huge volume of undiscovered data. This data may contain redundant, noisy and irrelevant data. Redundant data may cause classifier to produce less accurate results. PSO as a feature selection measure discards redundant features to improve the accuracy of the classifier. Our proposed method effectively identifies the redundant features compared to other existing features to effectively predict the heart disease. The rest of the paper is organized as follows. Section 2 discusses prior work done by peers and put our work in perspective. Section 3 reviews particle swarm optimization, K-nearest neighbor classifier, and feature subset selection.

In section 4, we will discuss our proposed method of predicting heart disease using PSO and KNN. Detailed discussions on experimental results are presented in section 5. Finally, we conclude in section 6.

Related Work

Data mining is a multidisciplinary field widely used in the clinical field such as prediction of heart disease. Researchers developed various techniques to predict the heart using data mining.

Prediction of heart disease using neural network was proposed by Dangare et al. in [7]. Feature selection is used to predict the disease. Their method obtained an accuracy of 92.5% for 13 features and 100% accuracy with 15 features. There is a 7.5% improvement after discarding 2 features from 15 to 13.

Jabbar et al. proposed a method using associative classification and feature subset selection for risk score of disease [2]. Authors used information gain, symmetrical uncertainty, and genetic algorithm as feature selection measures. Their method obtained an accuracy of 95% with hybrid feature selection. Heart disease data set collected with 11 features for experimental analysis.

Diagnosis of heart disease using fuzzy techniques is proposed [8]. Authors classified the patients based on the characteristics gathered from the therapeutic field. Authors used fuzzy and KNN and achieved an accuracy of 97%.

Fuzzy logic based heart disease detection is proposed in [9]. Authors considered 6 parameters for their experiments. Their approach achieved an accuracy of 92%. This method produces less accuracy compared [8]. Fuzzy with KNN approach produced 97%. Discretization and other filters may improve the performance of the algorithm. In [10] authors proposed prediction of heart disease using genetic neural networks. Experiments were done on American heart association data set. Their approach recorded an accuracy of 96.2%.

Masethe et al. [11] proposed a model using decision tree for heart disease prediction. Authors compared their approach with other classification approaches. RepTree and J48 achieved an accuracy of 99.07%.

Assessment of coronary heart events risk factors was proposed by karaolis et al. [12]. Authors investigated 2 types of risk factors namely modifiable and non-modifiable. 528 samples were collected and data mining analysis was done using C4.5. The Highest accuracy obtained by their model was 75% for PCI and CABG models. Authors used C4.5 classifier without feature selection measures. The accuracy obtained by this approach is less compared with other approaches.

Diagnosis of heart disease using regression trees was proposed by Amir [13]. Authors collected 116 heart sound signals data set and applied regression tree. Their model is proposed to classify Phonocardiogram's (PCG) data. Authors calculated like hood ratio to classify the disease. Their method obtained an accuracy of 99%. In the year 2014 authors [14] proposed a framework to expect the coronary illness using multilayer perceptron. Their method uses 13 clinical elements as input and achieved an accuracy of 98%. Literature mentioned in this related work has not used effective feature selection measures to improve the accuracy. Authors used weak classifiers to predict the disease. In this paper, we integrated PSO with KNN classifier to obtain effective results.

In this paper, we propose a decision support system (DSS) which uses KNN as classifier and PSO as feature subset selection measure for prediction of heart disease.

Theoretical Background

This section reviews basic concepts like PSO, KNN, and feature selection.

Particle swarm optimization

Particle swarm optimization here onwards referred as PSO is an EC based optimization algorithm proposed by Kennedy and Eberhart [15]. PSO is inspired by social behavior such as fish schooling and birds flocking. In PSO population (swarm) are encoded as particles. The Search starts with the random initialization of a population. The whole swarm moves in the search space for best solution by updating the position of each particle. The Position of each particle is done based on the position of its own and also based on neighboring particles. Current position of particle is represented by $X_i=\{x_{i1}, x_{i2}\text{-}x_{iD}\}$, where D is search space dimensionality. The velocity of a particle is represented by $V=(V_{i1},V_{i2}\text{----}V_{iD})$. The velocity of the particle is limited by V_{max} and V_{tid} $[-V_{max}, V_{max}]$. Best previous position and best position obtained are represented by p_{best} and g_{best}. An Optimal solution is searched by PSO by updating position and velocity based on p_{best} and g_{best} [16]. PSO is used as feature selection method due to its advantages like

1. Easy to implement
2. Can converge more quickly
3. Computationally less expensive and easy to implement

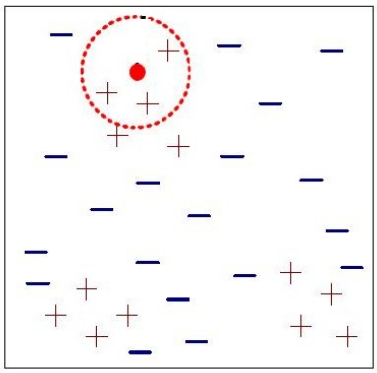

Figure 1. *K-Nearest neighbor classification.*

K-nearest neighbor classifier (KNN)

K-Nearest neighbor (KNN) is a simple, lazy and nonparametric classifier. KNN is preferred when all the features are continuous. KNN is also called as case-based reasoning and has been used in many applications like pattern recognition, statistical estimation. Classification is obtained by identifying the nearest neighbor to determine the class of an unknown sample. KNN is preferred over other classification algorithms due to its high convergence speed and simplicity [17]. Figure 1 show nearest neighbor classification. KNN classification has two stages

1) Find the k number of instances in the dataset that is closest to instance S

2) These k number of instances then vote to determine the class of instance S

The Accuracy of KNN depends on distance metric and *K* value. Various ways of measuring the distance between two instances are cosine, Euclidian distance. To evaluate the new unknown sample, KNN computes its *K* nearest neighbors and assign a class by majority voting.

Feature subset selection

The medical data set contains a large number of features which are redundant and irrelevant in nature. The performance of classifier may reduce if the data set contains this type of features. By removing redundant features accuracy of the classifier is improved and shortens the running time. Feature selection methods are broadly classified as

1. Filter

2. Wrapper

3. Hybrid approaches

Feature selection for large data set is a challenging task. Many search techniques used for feature selection suffers from local optima and high computational cost. Hence a cheap global search algorithm is required to develop a feature selection method. In our proposed approach, we used PSO for feature selection problem.

Proposed Approach

Our proposed method aims to enhance the performance of KNN classifier for disease prediction. Algorithm for our proposed method is shown below as Algorithm 1.

Algorithm 1. *Heart disease prediction using KNN and PSO.*

Step 1: Input: Heart disease data set
Step 2: Output: Classification of data set into patients with heart disease and normal
Step 3: Input the data set
Step 4: Apply pre-processing techniques-Fill in missing values
Step 5: select the features based on values obtained after applying PSO as FSS
Step 6: Discard redundant features (features with low values of PSO)
Step 7: Apply (KNN+IQR) on Predominant features
Step 8: Measure the performance of the KNN+PSO model

Algorithm takes the heart disease dataset and classify whether a person is having heart disease or not. The above algorithm is divided into 2 parts. Part 1 (Line 3-6) performs processing and feature subset selection. This part selects only predominant features for further process. In part 2 (Line 7-8), KNN is applied on pre-processed data set and performance is measured. Feature selection measure PSO is used to select the best features to obtain high accuracy.

Experimental Results

To predict heart disease the dataset containing 270 instances is collected from UCI repository [18]. Information about heart disease data set is shown in Table 1.

WEKA is used as the main package. To find the accuracy, we run 10 cross validation. Specifications of KNN and PSO listed in Tables 2 and 3.

Out of 14 features, PSO search selects 8 features (including class). Remaining 6 features will not be considered for classification of heart disease. These 7 features are predominant features which will enhance the accuracy of the classifier. Table 5 shows the accuracy obtained by our model for heart disease data.

Table 1. *Heart disease data set.*

Data set	Instances	Features
Heart disease	270	14

Table 2. *KNN specifications.*

Sl. no	KNN Specifications
1	KNN=2
2	Cross validation=2
3	NN Search=linear
4	Mean square =false

Features selected by PSO (Dominant features) are listed in Table 4.

Table 3. *PSO specifications.*

Sl. no	Specification
1	Population size: 100
2	Number of generations: 50
3	Report frequency: 50
4	Random seed=1

Out of 14 features, PSO search selects 8 features (including class). Remaining 6 features will not be considered for classification of heart disease. These 7 features are predominant features which will enhance the accuracy of the classifier. Table 5 shows the accuracy obtained by our model for heart disease data.

The accuracy obtained for various values of *K*. We tested four methods to record the accuracy of the classifier. The Interquartile Range (IQR) is a measure of variability. It divides data set into quartiles.

Table 4. *Features selected by PSO.*

Sl.no	Feature name
1	Chest
2	Resting_electrocardiographic_results
3	Maximum_heart_rate_achieved
4	Exercise_induced_angina
5	Old peak
6	Number_of_major_vessels
7	Thal

Table 5. *Accuracy obtained by our model.*

Method	K value				
	K=1	K=2	K=3	K=4	K=5
Before FSS	75.18	77.03	78	78	78.14
After normalization	77.7	78.8	81.1	81.4	81.4
After discretization	79.2	79.25	81.1	81.1	80.3
After PSO+KNN	78.8	81.1	81.4	81.4	81.4
KNN+PSO+IQR	100	100	100	100	100

Q1: In a rank- ordered data set middle value in first half. Q2: Median value in the data set.

Q3: is the "middle" value in the second half of the data set.

$IQR=Q3-Q1 \rightarrow (1)$

Accuracy recorded by our model before feature subset selection is 75.18 for *k=1* and 78.14 for *k=5*. Discretization filter in WEKA has improved the accuracy from 75.18 to 79.2 for *k=1* and 78.14 to 80.3 for *k=5*. IQR filter along with PSO improved the accuracy to 100%. Figure 2 shows accuracy recorded by our model for various values of K. The results obtained by KNN+PSO model reveal that our proposed model will improve the accuracy in good level. Experiments for our proposed approach were conducted on four different data sets. Heart disease-1 and Heart disease-2 are real data sets collected from various hospitals in India.

Results of proposed approach for various data sets are shown in Table 6. Values for values performance parameters are recorded in Table 7. True positive rate and sensitivity are recorded as 100%.

Proposed approach (KNN+PSO) is compared with KNN+GA. Using GA, accuracy is recorded as 77.7%, which is shown in Table 8.

Table 6 and Figure 3 show the comparison of learning accuracy of our model with other models mentioned in the survey.

From the simulation results perceived from Table 9 and Figure 3, our approach achieved enhanced accuracy by considering only predominate features. Our approach helps doctors to suggest the patients for diagnosis test to be taken for disease prediction.

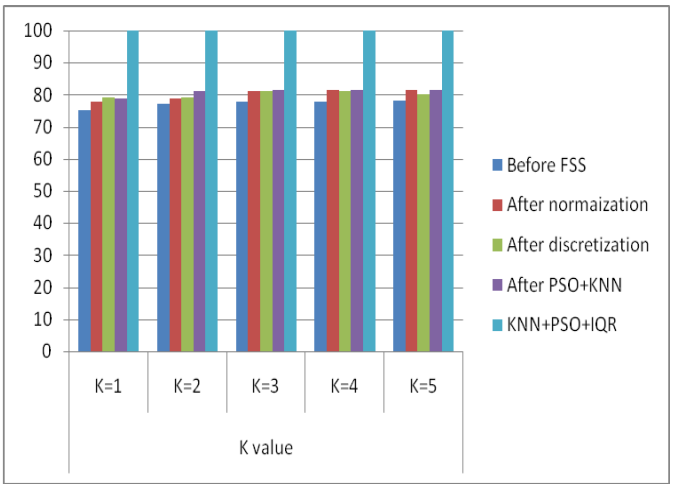

Figure 2. *Accuracy recorded by proposed model for various values of k.*

Table 6. *Accuracy obtained for various data sets.*

Sl. no	Data set	Instances	Attributes	Accuracy
1	Heart disease-1	40	10	97.5
2	Heart disease-2	75	12	100
3	Labour data	57	17	100
4	Soyabean	683	36	100

Table 7. *Values for various parameters. (Heart stalog data set).*

Parameter name	Value
Sensitivity	100%
TP Rate	100%
Accuracy	100%

Table 8. *Accuracy comparison with GA and PSO.*

Data set name	Approach	Accuracy
Heart disease	KNN+GA	77.7
	KNN+PSO	100

Table 9. *Accuracy comparison.*

Sl. no	Method	Accuracy (%)
1	Dangare [7]	92.5
2	Krishnail [8]	97
3	Kumar [9]	92
4	Amin [10]	96.2
5	Masetro [11]	99
6	Sonawale [14]	98
7	Our approach	100

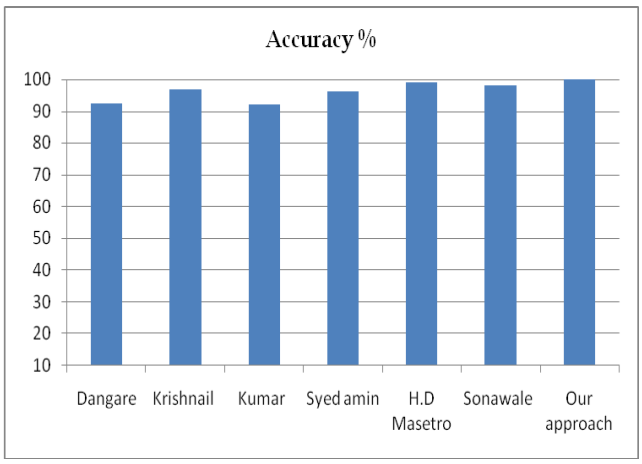

Figure 3. *Accuracy comparison.*

Before feature subset selection accuracy obtained is 75%. PSO search filters the number of features and selects the features which contribute more to the classification. By applying KNN with PSO accuracy improved to 100%. There is almost 25% increase in the accuracy. We tested OUR model for various values of *k*. Our experiment is limited to *k=5* as there is no much increment in the accuracy. Proposed approach is well suitable for multivariate data set. From the Table 9, it is evident that PSO plays a key role in enhancing the accuracy as a feature selection measure.

Conclusion

This paper addressed the prediction of heart disease based on PSO and KNN. Our approach uses KNN as a classifier to reduce the misclassification rate. This paper also investigates PSO based feature selection measure to select a small number of features and to improve the classification performance. The results suggest that proposed approach can significantly improve the learning accuracy. From simulation results, it is concluded that PSO based feature selection is important for classification of heart disease. This model helps the physicians in an efficient prediction of diseases with predominant features. In future, we want to integrate ensemble classifiers with PSO to develop a decision support system for early diagnosis of heart disease and also would like to compare GA and PSO for heart disease set.

References

1. www.neeman_medical.com
2. Jabbar MA, Deekshatulu BL, Priti C. Prediction of risk score for heart disease using associative classification and hybrid feature Selection. IEEE ISDA 2012; 628-634.
3. Jabbar MA, Deekshatulu BL, Priti C. Prediction of heart disease using random forest and feature subset selection. IBICA 2015; 425: 187-193.
4. Seema K, Bomare DS, Vaishnavi N. Heart disease prediction using KNN based handwritten text. AISC 2016; 49-56.
5. Han JK. Data mining concepts and techniques. (2nd Edn) 2009.
6. Tran B, Xue B. SEAL 2014; LNCS8886: 605-617.
7. Dangare A. Data mining approach for prediction of heart disease using neural network. IJCET 2012; 3: 30-40.
8. Krishnaiah V. Diagnosis of heart disease patients using fuzzy classification techniques. ICCCT 2014; 1-7.
9. Kumar. Detection of heart disease using fuzzy logic. IJETT 2013; 4.
10. Syed R, Agarwal B. Genetic neural network based data mining in prediction of heart disease using risk factors. IEEE Conference on ICT 2013.
11. Masethe HD, Masathe MA. Prediction of heart disease using classification algorithm. Wcess 2014.
12. Karaolis. Assessment of risk factor of coronary heart events based on data mining. IEEE Transactions IT Med 2016; 559-566.
13. Amir M. Early diagnosis of heart disease using classification and regression trees. University Caglian Italy 2013; 1-4.
14. Sonwane J. Prediction of heart disease using multilayer perceptron neural network. ICICES 2014; 1-6.
15. Ebehart R. A new optimization using particle Swam theory MHS 2013; 19: 39-43.
16. Cerventateciam XB. Binary PSO for feature selection filter based approach. WCCI 2012; 2012.
17. Jabbar MA, Deekshatulu BL, Priti C. Heart disease classification using nearest neighbor classifier with feature subset selection. Annals Computer Science 2013.
18. UCI machine learning repository, www.uci.org.

*Correspondence to

Jabbar MA

Vardhaman college of Engineering

Hyderabad

India