RESEARCH ARTICLE OPEN ACCESS

An Amalgam Data Mining Miniature to Envision Coronary Artery Disease

Shobana.K [1], Sasikala.M [2]
Department of Computer Science
KG College of Arts and Science
Coimbatore -India

ABSTRACT

Coronary artery disease (CAD) is leads to cardiac arrest and hold functioning of the heart causing heart attack. Prediction of CAD using invasive method is expensive. Hence a non-invasive model to predict CAD is proposed. In an existing CAD prediction method, a novel hybrid data mining model is used. The process of mining is performed on the data set that is collected. Risk factor identification is done using correlation based feature subset (CFS) selection in addition to particle swam optimization (PSO) search pattern and K-means clustering algorithms. Multinomial logistic regression (MLR), multi-layer perception (MLP), C4.5 and fuzzy unordered rule induction algorithm (FURIA) are then used to model CAD cases. The premature convergence of PSO sometimes leads to provide less accuracy. In order to solve these issues in the proposed system performs the prediction of CAD utilizing the BAT algorithm instead of PSO. The experimental results will prove that BAT based feature selection performs better than PSO based feature selection. The obtained accuracy and misclassification rate of CAD enables the proposed technique to enhance the overall prediction of CAD, thus improving the accuracy of the proposed model in predicting CAD.

Keywords:- Coronary artery disease, K-means, fuzzy unordered rule induction algorithm, BAT, Multinomial logistic regression, multi-layer perception, C4.5 **.**

I. INTRODUCTION

Data mining [1] is the process of searching hidden pattern and gain knowledge or information in the dataset from the analysis of data from different perspective. The dataset may be structured or unstructured the main goal of data mining is to apply computational process in the dataset to extract the useful information from the data and convert into an understandable format. Data mining is processed based on five major elements are Extract, transform, and load transaction data.

Data collection is the process of collecting data from different sources either homogeneous or heterogeneous sources. The data collection is more important step in the data mining process because this data are processed in further steps in data mining technique. Data pre-processing [2] is an important task in the data mining. The data from the real world entities may contain the missing values, inconsistent, incomplete or contain some errors. Feature selection is a process to reduce the dimensionality of data by selecting relevant features in the dataset. It varies from the feature extraction feature extraction creates new features from the task of original features whereas in the feature selection return selected number of features from the dataset which are more relevant or strongly correlated.

Correlation based Feature Selection [3] is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy. CFS is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features.

Clustering [4] is the process of grouping the objects based on inter cluster and intra cluster distance between the objects. The main aim of clustering algorithm is to minimize the inter cluster distance between the objects in the same cluster group and to maximize the intra cluster distance between the objects in different cluster groups.

Data classification is the process of arranging data into groups for its most valuable and efficient use. A well developed data classification system constructs essential data simple to find and retrieve. This can be of particular importance for compliance, legal discovery and risk management. Written techniques, rules and procedures for data classification should characterize what categories and criteria the group will use to categorize data and state the roles and responsibilities of employees within the group regarding data management. Once a data-classification method has been created, security standards that give suitable handling practices for individual category and storage standards that describe the data's life cycle requirements should be addressed.

II. CLASSIFICATION TECHNIQUES

A. Linear classifier

A linear classifier identifies the class by building a classification decision which is depends on the value of a sequential combination of the characteristics. Here the object characteristics is defined as feature values and are usually offered to the machine in a vector called a feature vector. Such classifiers classifies the data effectively for practical problems such as document classification, and more commonly for problems with many attributes having accuracy levels comparable to non-linear classifiers whereas taking less time to train and use. Linear classifier uses the linear model to classify the data. Linear model split input vectors into classes through linear decision boundaries.

B. Support vector machine

Support vector machine [5] is a supervised machine learning model which analyse data for classification and regression analysis. This machine learning algorithm split the data into training data and testing data. It constructs a hyper plane in an infinite dimensional space. A good classification is achieved by generating best hyper plane with largest distance between training data points and goal of classification is to choose the hyper plane which has maximum distance from the data point and nearest data point. This is used to decide the input data point belongs to which class which is known as maximum margin hyper plane.

The generalization error of the classifier is reduced by maximum margin hyper plane. The novel problem may be declared in a finite dimensional space, it frequently happens that the sets to discriminate are not sequentially separable in that space. For this motivation, this machine learning was offered that the original finite-dimensional space be recorded into a much higher-dimensional space, probably making the partition easier in that space. To maintain the computational load reasonable, the mappings used by SVM schemes are designed to guarantee that dot products may be calculated easily in provisions of the variables in the original space, by explaining them in terms of a kernel function chosen to ensemble the problem. The hyperplanes in the higher-dimensional space are explained as the set of points and in the set of points dot product is maintained as constant with the vector.

C. K- nearest neighbor

K nearest neighbour classification technique [6] is a non parametric method. It is a simple algorithm that accumulates all available cases and classifies new cases based on a similarity measure. The output of K nearest neighbour is a class membership. As per the name of the technique the classification is based on majority vote of its neighbours, with the object being given to the class most common among its k nearest neighbours and it can be helpful to give weight to the involvement of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones. It is sensitive to the local structure of the data. In the classification phase, *k* is a user-defined constant, and an unlabeled vector in the dataset is classified by giving the label which is most repeated among the *k* training samples nearest to that query point.

D. Decision trees

Decision tree develops classification model in a tree like structure. In this classification technique huge dataset is efficiently handled by splitting the dataset into smaller subsets while splitting the dataset the decision is incrementally developed. In tree structure the leaves hold class labels and branches hold the combination of features. Decision tree called as regression tree when the target variables assigned continuous values. It is mostly used for decision making.

Each interior node in the decision tree communicates to one of the input variables; there are boundaries to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables characterized by the path from the root to the leaf. A tree can be trained by splitting the input dataset into subsets based on an attribute value test. This process is repetitive on each derived subset in a recursive manner called recursive partitioning. The recursion is finished when the subset at a node has all the same value of the target variable. This method of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far away the most common approach for learning decision trees from data.

E. Neural networks

Artificial neural networks [7] are relatively simple electronic networks of neurons which are based on the neural structure of the brain. They handle records one at a time, and learn by evaluating their classification of the record with the well-known actual classification of the record. The errors from the initial classification of the first record is given into the network, and used to alter the networks algorithm for further iterations.

In this technique neurons are arranged in input layer, hidden layer and output layer. In the input layer contains the record values which are input to the net layer. Followed by input layer hidden layer contains more than one neural network. In the neural network adaptive weight are considered as connection strength among the neurons which are stimulated during prediction and training of data. The output layer where there is only one node is created for each class in the dataset.

III. METHODOLOGY AND ANALYSIS

Existing Scenario

In existing system, presented a novel hybrid method based on data mining techniques for CAD diagnosis. The main objective of this method is to predict CAD with high accuracy. Then the dimension of dataset was reduced by correlation based feature subset (CFS) and Particle Swarm Optimization (PSO). Then the selected features were clustered using K means algorithm based on the selected features. Supervised learning algorithms such as multi-layer perceptron (MLP), multinomial logistic regression (MLR), fuzzy unordered rule induction algorithm (FURIA) and C4.5 are then used to model CAD cases.

Disadvantages

- PSO based feature selection leads premature convergence that leads low accuracy
- PSO is easy to fall into local optimum in high-dimensional space.
- PSO has low convergence rate.

Proposed Scenario

The proposed system is developed for selecting more relevant features and reduced the dimensionality of coronary artery disease data from huge volume of dataset. In order to predict the CAD, data mining methods are utilized in this paper. To improve the performance of feature selection BAT algorithm is proposed to predict CAD more effectively. Also, the performance of accuracy and misclassification rate is also improved by BAT algorithm.

Advantages

- The rate of accuracy has comparatively improved.
- It achieves prominent results in this scenario.
- The complexity of the system is immensely reduced.
- It improves the system performance.

Dimensionality reduction with BAT algorithm

The collected data contains vast number of features it leads the dimension of data is large. In order to reduce the dimension of data and select the more relevant features from the CAD dataset BAT algorithm is proposed. Such technique has been developed to behave as a band of bats tracking prey/foods using their capability of echolocation. This algorithm is processed based on the following rules:

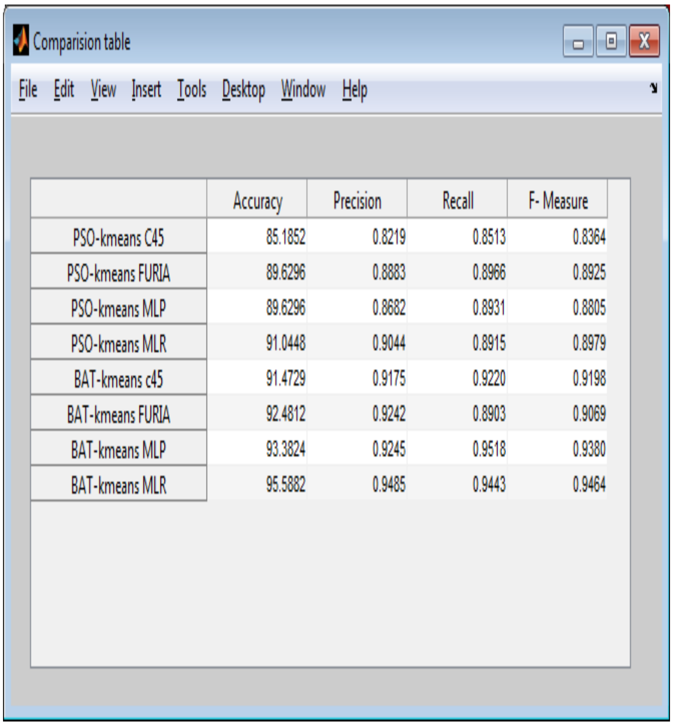
- All bats use echolocation to sense distance, and they also "know" the difference between food/prey and background barriers in some magical way
- A bat b_i fly randomly with velocity v_i at position x_i with a fixed frequency f_{min}, varying wavelength λ and loudness A_0 to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission $r \in [0, 1]$, depending on the proximity of their target
- Although the loudness can vary in many ways, we assume that the loudness varies from a large (positive) A_0 to a minimum constant value A_{min}.

Algorithm:

```
Objective function ($x$)
Initialize the bat population $x_i$ and $v_i$, $i$ = 1, 2, ...,$m$.
Define pulse frequency $f_i$ at $x_i$, ∀ $i$ = 1, 2, . . . , $m$.
Initialize pulse rates $r_i$ and the loudness $A_i$, $i$ = 1, 2, . . . , $m$.
Step 1: while t<T
Step 2:     for each bat $b_i$, do
Step 3:         generate new solution using
                    $f_i = f_{min} + (f_{min} - f_{max})\beta$
                    $v_i^j(t) = v_i^j(t-1) + (x^j - x^j(t-1))f_i$
                    $x^j(t) = x_i^j(t-1) + v_i^j(t)$
```

Where β denotes a randomly generated number within the interval 0 to 1. $x^j(t)$ denotes the value of decision variable j for bat i at time step t. The results of f_i is used to control the pace and range of the movement of the bats. $x^j(t)$ represents the current global best location (solution) for decision variable j, which is achieved comparing all the solutions provided by the m bats.

```
Step 4:         if rand>$r_i$, then
Step 5:         Select a solution among the best solutions.
Step 6:             Generate a local solution around the best solution.
Step 7:         if rand<$A_i$ and f($x_i$)<f(x), then
Step 8:             Accept the new solutions.
Step 9:             Increase $r_i$ and reduce $A_i$
Step 12:        Rank the bats and find the current best x.
```


Comparision table
File Edit View Insert Tools Desktop Window Help

	Accuracy	Precision	Recall	F- Measure
PSO-kmeans C45	85.1852	0.8219	0.8513	0.8364
PSO-kmeans FURIA	89.6296	0.8883	0.8966	0.8925
PSO-kmeans MLP	89.6296	0.8682	0.8931	0.8805
PSO-kmeans MLR	91.0448	0.9044	0.8915	0.8979
BAT-kmeans c45	91.4729	0.9175	0.9220	0.9198
BAT-kmeans FURIA	92.4812	0.9242	0.8903	0.9069
BAT-kmeans MLP	93.3824	0.9245	0.9518	0.9380
BAT-kmeans MLR	95.5882	0.9485	0.9443	0.9464

Fig 1. Comparison table of PSO and BAT on core factors including accuracy, precision, recall and F-measure

Firstly, the initial position x_i, velocity v_i and frequency f_i are initialized for each bat b_i. For each time step t, being T the maximum number of iterations, the movement of the virtual bats is given by updating their velocity and position. These selected features are clustered using k-means technique.

For the identification of CAD with the selected relevant features, MLP, MLR, FURIA and C4.5 models are used. MLP is popular ANN architecture. Artificial Neural network (ANN) is used as learning algorithm which is more suitable to tackle complex tasks and problems. A typical architecture of neural based learning machine consists of an input layer, one or more hidden layer and an output layer.

Each hidden layer extracts the more complex representation for the previous layer such that the last hidden layer would have representation meant to discriminate between the samples of different classes. MLR is an extension of extension of logistic regression with ridge estimator and it is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable and it utilize the maximum likelihood estimation to evaluate the probability of categorical membership.

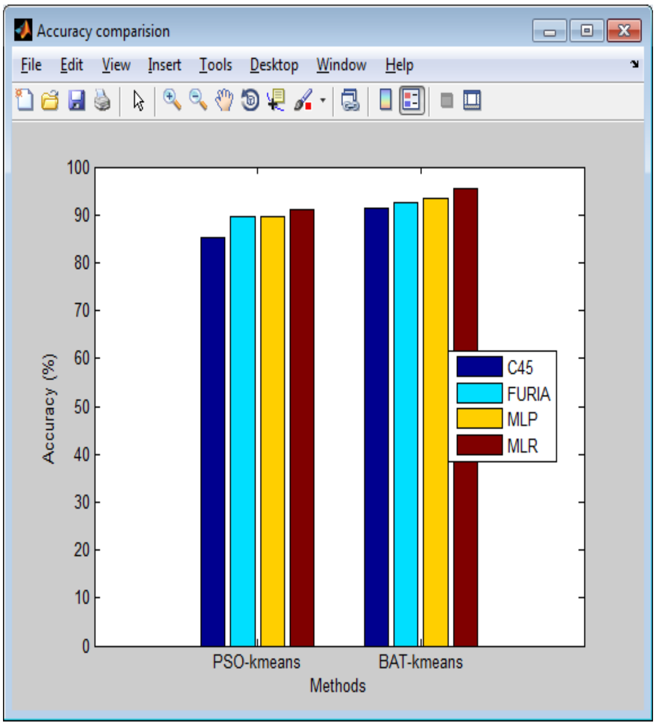


Fig 2. Comparitive improvement in accuracy of BAT

This algorithm is more efficient than the PSO technique. Based on the behaviour of the bats, a new and interesting meta-heuristic optimization technique called Bat Algorithm was proposed.

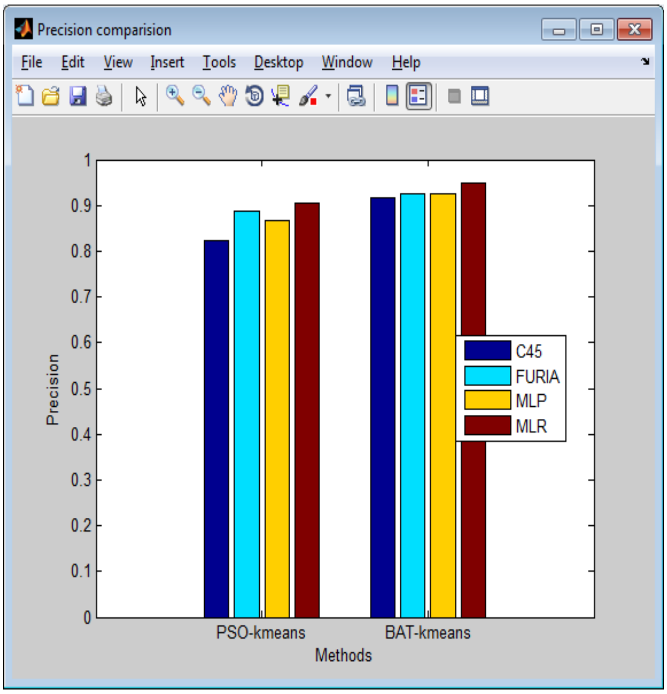


Fig 3. Precision of BAT is more compared to PSO

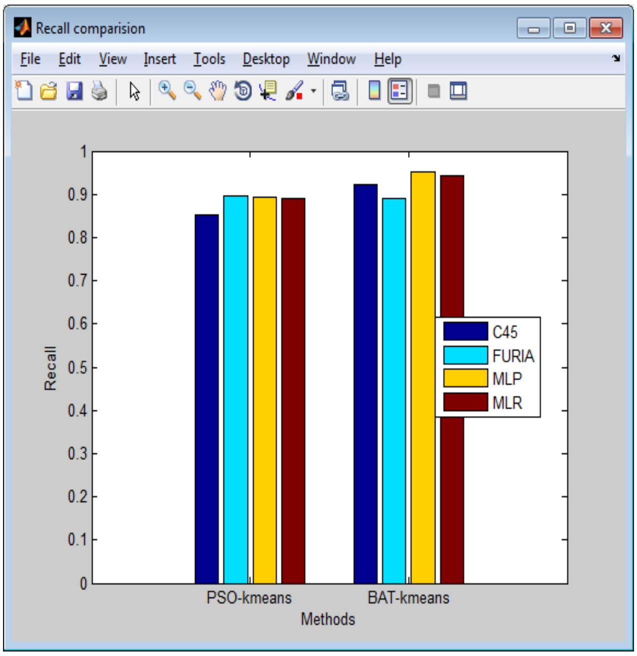
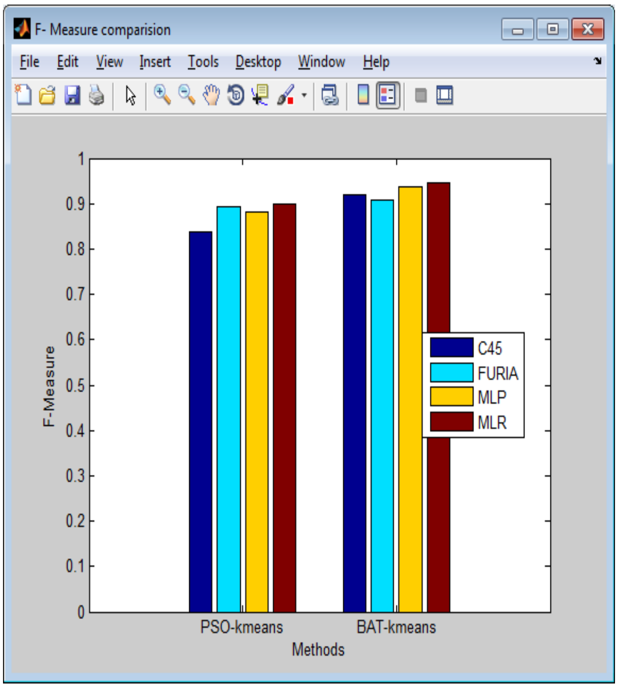


Fig 4. The recall factor of BAT is comparatively good compared to PSO

FURIA is an improvement of RIPPER algorithm. It is used to create fuzzy rules in order to model the decision boundaries. Fuzzy rules are obtained through replacing intervals with fuzzy intervals using trapezoidal membership function combined with the sophisticated rule induction techniques employed by the original RIPPER algorithm.



IV. CONCLUSION

The prediction of coronary artery disease is improved by introducing a new efficient feature selection method is called BAT. This feature selection is based on the behaviour of bat, in which the number of bats is initialized and each bat updates their position and velocity value according to their objective function. Each bat randomly select different number of features and different features and using the global best and local best solution finally get an optimized number of features. These selected features are clustered using k means algorithm. Then finally models are constructed using Multi layer perceptron, Multinomial logistic regression model, Fuzzy unordered rule induction algorithm, and C4.5. These models are constructed using training data and these models are used in testing data in order to predict the CAD disease. The experiment result clearly demonstrates the proposed technique has high accuracy and less misclassification rate for the prediction of CAD disease.

REFERENCES

[1] Hardin, J. M., & Chhieng, D. C. (2007). Data mining and clinical decision support systems. In Clinical Decision Support Systems (pp. 44-63). Springer New York.

[2] Aneyrao, T. A., & Fadnavis, R. A. (2016, February). Analysis for data preprocessing to prevent direct discrimination in data mining. In Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on (pp. 1-4). IEEE.

[3] Li, J. (2008, August). Feature selection based on correlation between fuzzy features and optimal fuzzy-valued feature subset selection. In Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP'08 International Conference on (pp. 775-778). IEEE.

[4] Kumar, D., Bezdek, J. C., Palaniswami, M., Rajasegarar, S., Leckie, C., & Havens, T. C. (2015). A hybrid approach to clustering in big data.

[5] Shao, Z., Zhang, L., Zhou, X., & Ding, L. (2014). A novel hierarchical semisupervised SVM for classification of hyperspectral images. IEEE Geoscience and Remote Sensing Letters, 11(9), 1609-1613.

[6] Manjusha, M., & Harikumar, R. (2016, March). Performance analysis of KNN classifier and K-means clustering for robust classification of epilepsy from EEG signals. In Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on (pp. 2412-2416). IEEE.

[7] Nowicki, R. K., Scherer, R., & Rutkowski, L. (2016, August). Novel rough neural network for classification with missing data. In Methods and Models in Automation and Robotics (MMAR), 2016 21st International Conference on(pp. 820-825). IEEE.

[8] Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K., & Michalis, L. K. (2008). Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Transactions on Information Technology in Biomedicine, 12(4), 447-458.

[9] Acharya, U. R., Faust, O., Sree, V., Swapna, G., Martis, R. J., Kadri, N. A., & Suri, J. S. (2014). Linear and nonlinear analysis of normal and CAD-affected heart rate signals. Computer methods and programs in biomedicine, 113(1), 55-68.

[10] Giri, D., Acharya, U. R., Martis, R. J., Sree, S. V., Lim, T. C., Ahamed, T., & Suri, J. S. (2013). Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowledge-Based Systems, 37, 274-282.