ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 5, Issue 4, April 2016

Adaptive Preprocessing And Prediction Using SVM Classifier

Malathi B, Ramya K, Sangavi S

Abstract— ***Now a days, healthcare sector uses different data mining functionalities to predict diseases and efficient healthcare services. Data mining is the process of mining information from other repositories such as data warehouse, massive datasets etc. The proposed system performs decoupling of preprocessing and prediction. In preprocessing phase, redundant data are removed by using filtering method and missing values are filled by mean method in medical dataset (Heart disease dataset). The main objective of this paper is to predict the type of heart disease and their treatment using two classification algorithms. The two classifiers are SVM classifier and Naïve Bayes classifier. The Performance (such as accuracy and execution time) of these classifier are compared and then from the experimental results it is clear that SVM classifier is better than Naïve Bayes for predict the heart disease.***

Index Terms— Classification, Heart disease function test, Naïve Bayes (NB), Preprocessing, Support Vector Machine (SVM).

I. Introduction

Data mining is a process of extract knowledge from an existing data or other repositories and transform it into a human understandable format for further use. The processes involved in data mining are preprocessing, classification, clustering and association [3]. Preprocessing is most important step when considering raw data. In this research work, mean method and filtering method is used for preprocessing. The classification techniques in data mining are much popular in medical diagnosis and disease prediction [1]. In healthcare sector researchers faces many challenging task to classify the diseases from massive medical databases. Nowadays data mining became more essential in health sector. In this paper, Support Vector Machine (SVM) and Naïve Bayes (NB) classifier algorithms are used for heart disease prediction. Using these classifiers the research work predicts the heart disease such as type of angina heart disease.

II. EXISTING SYSTEM

In existing system, decoupling of adaptive preprocessing and prediction had been implemented over a chemical dataset .The system uses chemical dataset which was taken from chemical production industry (e.g., sensor reading). The preprocessing technique use in this system is Principal Component Analysis (PCA). PCA identifies the pattern between non-linearly separable data using eigenvectors and eigenvalues. For classification, Naïve Bayes classifier is used and adapt to the changing environment over time (i.e., concept drift) [3].In this system, the preprocessor may need feedback from prediction to decide upon adapting or retrain itself. Disadvantages in this system are Principal Component Analysis (PCA) is not applicable for linearly separable data and also dataset loses its interpretability.

III. PROPOSED SYSTEM

This research work involves decoupling of preprocessing and prediction. Decoupling means preprocessing and prediction processes are separately done.

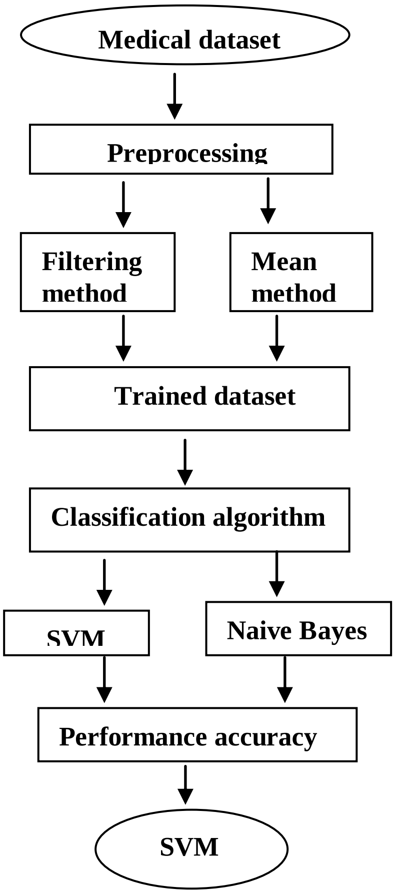

Figure - 1: System Architecture

A.MEDICAL DATASET

An Indian Heart Patient Dataset (IHPD) is collected from UCI repository. This dataset has five hundred instances and fifteen attributes. Attributes are Name, Patient id, Age, Gender, Chol (serum cholestoral), Cp (Chest pain type), restecg (resting electrocardiographic results), thalach (maximum heart rate achieved), exang (exercise include angina), oldpeak (ST depression include by exercise relative to test), slope (slope of the peak exercise ST segment), Ca (number of major vessels coloured by flourosopy), thal.

B. PREPROCESSING

Preprocessing can be done by several methods. Real world data are generally inconsistent, noisy and incomplete data. In this system filtering method is used for data reduction (removing the repeated data), mean method is used for incomplete data (i.e., missing values) and also remove the inconsistent data (impossible data combination).

1) MEAN METHOD EQUATION:

$$\mathbf{MA_n} = \frac{\sum_{i=1}^{n}}{n}$$

By using the above formula missing values are identified and filled in according place. Then the trained data is given as an input to the predictor to classify the heart disease using classification algorithm.

C. CLASSIFICATION ALGORITHM

Classification is one of the most important techniques in data mining process. In this research work, two classifier algorithms are used to predict heart disease. They are Support Vector Machine and Naive Bayes classifier.

1) SUPPORT VECTOR MACHINE: Support Vector Machine was first found by Vapnik in 1979. It was again recommended by Vapnik in 1995 for regression and classification [8]. SVMs are set of related supervised learning methods used for classification and regression. The SVM is the advanced technology with maximum classification algorithms impacted in statistical learning theory. SVM methods are used in classification of linear and non-linear data. For a non-linearly separable data, it transforms the original training data into higher dimension using non-linear mapping. Using this new dimension it searches for linear optimal separating hyperplane. Data from two classes can be separated by hyperplane with an appropriate nonlinear mapping to a sufficiently high dimension. Using support vectors and margins the SVM finds these hyperplane. SVM implements the classification task by maximizing the margin classifies both class while minimizing the classification errors. The basic idea is shown in figure - 2. The data points are identified as being positive or negative, and the problem is to find a hyper-plane that separates the data points by a maximal margin.

The figure - 2 shows the 2-dimensional case where the data points are linearly separable. The mathematics of the problem to be solved is the following:

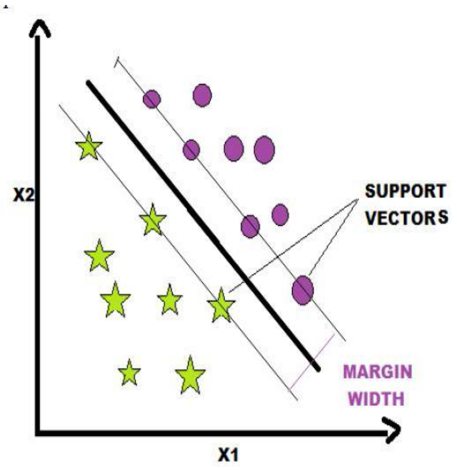

Figure - 2: Support Vector Machine

$$\left.\begin{array}{l} \min\limits_{\vec{w},b} \frac{1}{2}\|w\| \\ \textbf{s.t} \quad y_i = +1 \Rightarrow \vec{w}.\vec{x}_i + b \geq +1 \\ \qquad y_i = -1 \Rightarrow \vec{w}.\vec{x}_i + b \geq -1 \\ \textbf{s.t} \quad y_i(\vec{x}_i.\vec{w} + b) \geq 1, \ \forall i \end{array}\right\} \qquad \textbf{(1)}$$

The identification of the each data point xi is yi, which can take a value of +1 or -1 (representing positive or negative respectively). The solution hyper-plane is the following:

$$u = \vec{w} \bullet \vec{x} + b \qquad \textbf{(2)}$$

The scalar b is also termed the bias. A standard method to solve this problem is to apply the theory of Lagrange to convert it to a dual Lagrangian problem. The dual problem is the following:

$$\left.\begin{array}{l} \min_\alpha \Psi(\vec{\alpha}) = \min_\alpha \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j (\vec{x}_i . \vec{x}_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i \\ \sum_{i=1}^{N}\alpha_i y_i = 0 \qquad \textbf{(3)} \\ \alpha_i \geq 0, \forall i \end{array}\right\}$$

The variables αi are the Lagrangian multipliers for corresponding data point xi.

2) NAÏVE BAYES

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumption. A more descriptive term for the underlying probability model would be the self-determining feature model. The Naive Bayes classifier performs reasonably well even if the underlying assumption is not true.

The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets.

Bayes theorem provides a way of calculating the posterior probability, P (c|x), from P(c), P(x), and P (x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence [9].

Likelihood Class prior probability

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Posterior probability Predictor prior probability

$$P(C|X) = P(X_1|C) * P(X_2|C) * \dots * P(X_n|C) * P(C)$$

$P(C|X)$ is the posterior probability of class (target) given predictor(attribute).
$P(C)$ is the prior probability of class.
$P(X|C)$ is the likelihood which is the probability of predictor given class.
$P(X)$ is the prior probability of predictor.

IV. CONCLUSION

Preprocessing and Classification is a major data mining technique which is used in healthcare sectors for medical diagnosis and disease prediction. This research work used two classification algorithm namely Support Vector Machine (SVM) and Naïve Bayes classifier for heart disease prediction. Comparisons of these algorithms are done based on performance factors such as accuracy and execution time. From the results, this work concludes SVM classifier is more suitable to predict heart disease because of its highest classification accuracy. While comparing the execution time Naïve Bayes classifier needs minimum execution time.

ACKNOWLEDGEMENT

We are thankful and would like to express our sincere gratitude for our Head of the Department **Mr.S.Venkatasuburamanian** and our supervisor **Mr.P.Dineshkumar** for the guidance, support and continuous encouragement in making this project possible. Their guidance from initial to final level enabled us to achieve our objective of final year project. Our sincere thanks to all the lectures who helped us in many ways, gave valuable advises and made our journey easy.

REFERENCES

[1] Jyoti Soni Ujma Ansari, Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887)Volume 17– No.8, March 2011.

[2] Dr. S. Vijayarani, Mr.S.Dhayanand,Liver Disease Prediction using SVM and Naïve Bayes Algorithms,International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015

[3] Ketan Desale, Roshani Ade, Comparative Study of Pre-processing Techniques for Classifying Streaming Data, International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3294-3297.

[4] Vikas Chaurasia, et al, Carib.j.SciTech,2013,Vol.1,208-217,Early Prediction of Heart Diseases Using Data Mining Techniques.

[5] P.Radhabai Mrs, M.Priya Packialatha, Dr.G.Geetha, Scenario Based Adaptive Preprocessing for Stream Data using SVM Classifier, International Journal of Emerging Technology in Computer Science & Electronics (ijetcse)issn: 0976-1353 volume 13 issue 1 –march 2015.

[6] Chaitrali S,Dangare Sulabha S, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012 44.

AUTHORS

B.Malathi,
B.E., Computer Science and Engineering,
Saranathan College Of Engineering.

K.Ramya,
B.E., Computer Science and Engineering,
Saranathan College Of Engineering.

S.Sangavi
B.E.,Compuer Science And Engineering,
Saranathan College Of Engineering.