# Rough Set Attribute Reduction with Support Vector Machine for Prediction ofHeart Disease

[1]Ravindar Mogili, [2]Dr. G.Narsimha, [3]Mahesh Nagamalla

[1,3]Associate Professor, [2]Professor
[1,3]Department of Computer Science and Engineering, Jyothishmathi Institute of Technology & Science, Karimnagar, Telangana State, India
[2] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, Telangana State, India

**Abstract.**Analysis of health data helps the doctors in decision making. As the continuously addition of daily generated data, analysis becomes more complicated. Besides this it also requires more memory space and computation time. The analysis process made simple by normalizing huge data, i.e., removing the redundant and useless data for analysis. The Rough Set Attribute Reduction (RSAR) calculates reducts of full set attributes and generate minimal attributes set. This is given as input to the Support Vector Machine. The SVM can analyze diagnosis model of health data and predicts the class of data item, i.e., disease stage. The objective of this paper is to improve classification accuracy of the health data by combining RSAR and SVM.

**Keywords:** Medical data, Rough set, Attribute reduction, Support vector Machine, Classification, Heart disease

## 1 Introduction

Data mining is the science of extracting and refining useful information from large databases. It is the process of searching concealed information that can be transformed into knowledge, thus could be used for strategic decision making. Application of data mining techniques in health care has become increasingly popular since it offers benefits to Doctors, patients and healthcare organizations.

The heart purifies and pumps blood into blood vessels and these vessels deliver the purified blood to all over the body and also carry impure blood to the heart. Heart diseases commonly known as cardiovascular diseases (CVD) occur due to abnormal blood flow from the heart. As per WHO reports, maximum of all human deaths around the globe are due to chronic disease like heart Diseases. The early stage detection of chronic disease is more important since the risk factor increases as the detection is delayed. Therefore, detecting cardiac abnormalities at the early stage can reduce the mortality rate. But, the diagnosis of the heart disease by a medical specialist is a challenging task due to requirement of prior knowledge with good skills. This task can be simplified with the help of machine learning methods. The machine learning methods are used to develop prediction model by analyzing existing health data. Different classification algorithms have been used with a number of attributes for prediction of class. One of the most popular machine learning techniques is support vector machines (SVM) [1][2].When a new data item is given as input to the classification model, it predicts the class of data item, i.e., stage of disease. The objective of this paper is to improve the prediction accuracy using SVM by preprocessing health datasets with Rough Set Attribute Reduction (RSAR) [3].

The basic organization of the paper is arranged in 5 sections as follows: Section 2 presents the overview of the Support Vector Machines and also describes about Rough Set Attribute Reduction, Section 3 describes the proposed hybrid model for classification, Section 4 presents the experiment and results and the conclusions are given in section 5.

## 2 Review of Literature and Concepts

Aa Evanthia et al. [4] investigated heart disease prediction by using three machinelearning strategies namely neural network, SVM, regression trees and claimed that SVM is generating better performance than other two models.Srinivas Konda [5] proposed a rough-fuzzy classifier to predict the heart disease failure by combining rough set theory with the fuzzy set to enhance the prediction performance. Das et al.[6] proposed Neural Networks ensemble model by combining the posterior probabilities from multiple models such as Naive Bayes, MLP, C4.5, AIRS, etc.Tan et al. [7] proposed a hybrid model to predict heart failure disease using Support Vector Machine and Genetic Algorithm.Mai Shouman et al.[8] proposed a model by using support vector machine for predicting heart disease and generated accuracy of 84.1%.Heon Gyu Lee et al. [9] used Bayesian classification, associative classifier, classification based on multiple association rule (CMAR), C4.5 (DT) and SVM for predicting coronary artery diseases. Accuracyof SVM, CMAR, C4.5 were 90%, 80% and 78% respectively. SVM showed the best performance.

Besides the availability of huge health data, every day large amount of health data is generated. All the parties associated with health sector such as patients, doctors, health organizations are benefited, if the health data is analyzed and fruitful information is generated. But due to availability of huge data, analysis becomes more complicated [10]. In order to simplify the complexity of analysis, data reductions methods such as rough set theory is used. Later analysis is done by SVM and a classification model is developed. The overview of SVM and attribute reduction using rough set is discussed in the below.

### 2.1 Support vector machines

Support vector machines (SVM) handles classification by defining a separating hyper-plane. Let $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), . . . , (\vec{x}_m, y_m)\}$be the training data set, where $\vec{x}_i$ is a vector with 'n' attributes in space R and $y_i$ is its corresponding output denotes the class labels either +1 or -1. The job of SVM is to linear separate two classes by hyper-plane

$$\vec{w} . \vec{x} + b = 0 \qquad \textbf{(1)}$$

where $\vec{w}$ is a weight vector, $\vec{x}$ is input vector and b is a bias constant.

When the two classes are linearly separable, there may exist an infinite number of linearly separable hyper-planes. SVM attempts to choose one among them such that the distance from the hyper-plane to the nearest points in either class is maximized. These nearest points in the two classes are called as support vectors. As the hyper-plane is separating two classes, data points belongs to one of the class fall above the hyper-plane as $\vec{w} . \vec{x}_i + b \geq +1$, with output $y_i = +1$and other class points fall below the hyper-plane given as $\vec{w} . \vec{x}_i + b \leq -1$, with output $y_i = -1$. So, the decision rule to find the class to which $\vec{x}$ belong can be written as

$$y = f(\vec{x}) = \text{sign}\,(\vec{w} . \vec{x} + b) \qquad \textbf{(2)}$$

The generalized equation for both the classes can be written as

$$y_i(\vec{w} . \vec{x}_i + b) \geq 1 \qquad \textbf{(3)}$$

If the two classes are non-linearly separable, SVM uses φ( ) functions to transform the data of two classes from original space R into a much higher-dimensional space to makelinear separation is possible in that space. However, a perfect linear separation may not possible in most of the cases; a restricted minimum number of misclassifications data points should be allowed. As a result error rate increases. To control error rate, a slack variable ξ is introduced such that $\xi_i \geq 0$. $\xi_i$ is the perpendicular distance from the correct boundary margin of hyper-plane to the misclassified point $\vec{x}_i$ .Therefore equation (3) can be modified as

$$y_i(\vec{w} . \varphi(\vec{x}_i) + b) \geq 1 - \xi_i \qquad \textbf{(4)}$$

The objective of SVM is to maximize the width of hyper-plane separating the two classes and minimize the misclassification errors. The bargain between hyper-plane width and misclassification error is controlled by a user-defined constant 'C'. The objective function is

$$\text{Min}(\vec{w}, \xi) = \frac{1}{2} ||\vec{w}||2 + C \sum_{i=1}^{s} \xi_i \qquad \textbf{(5)}$$

The value of $\vec{w}$ and b are calculated by solving the objective function (5) subject to the equation (4) as per Lagrange multiplier, it can be written as

$$L = \frac{1}{2} ||\vec{w}||2 + C \sum_{i=1}^{s} \xi_i - \sum_{i=1}^{s} \alpha_i [y_i(\vec{w} . \varphi(\vec{x}_i) + b) - 1 + \xi_i] \qquad \textbf{(6)}$$

where 's' is the number of support vector points and $\alpha_i$ are constants .

This equation is also known as the primal function. Differentiating equation (6 ) w.r.t w, b and $\xi_i$ we get

$$\begin{gathered} \frac{\partial L}{\partial w} = 0, \quad \vec{w} - \sum_{i=1}^{s} \alpha_i y_i \vec{w} . \varphi(\vec{x}_i) = 0 \Rightarrow \vec{w} = \sum_{i=1}^{s} \alpha_i y_i \varphi(\vec{x}_i) \\ \frac{\partial L}{\partial b} = 0, \quad \sum_{i=1}^{s} \alpha_i y_i = 0 \qquad \textbf{(7)} \\ \frac{\partial L}{\partial \xi} = 0, \quad \sum_{i=1}^{s} \alpha_i y_i = 0 \end{gathered}$$

Now equation (2) can be re-written by substituting$\vec{w}$value from the equation (7).This is given by

$$y = f(\vec{x}) = \text{sign}\,(\sum_{i=1}^{s} \alpha_i y_i \varphi(\vec{x}_i) . \varphi(\vec{x}) + b) \qquad \textbf{(8)}$$

Instead of computing the dot product of $\varphi(\vec{x}_i)$ and $\varphi(\vec{x})$ explicitly by transforming into higher dimensional space, the same can be computed implicitly by using kernel function $K(\vec{x}_i, \vec{x})$. i.e., $K(\vec{x}_i, \vec{x}) = \varphi(\vec{x}_i).\varphi(\vec{x})$.There exist more kernel trick functions. We should choose the best suitable one for our model by cross validation. The percentage of prediction can be improved by selecting the best suitable kernel trick function. Some of the best kernel trick functions are:

Linear : $K(\vec{x}_i, \vec{x}) = \vec{x}_i . \vec{x}$ ,
dth-Degree polynomial: $K(\vec{x}_i, \vec{x}) = (1 + \vec{x}_i , \vec{x})^d$ , **(9)**
Radial basis (RBF) : $K(\vec{x}_i, \vec{x}) = \exp(-\gamma\, ||\vec{x}_i - \vec{x}||^2)$,

Sigmoid :L $K(\vec{x}_i, \vec{x}) = \tanh(\kappa 1\ \vec{x}_i, \vec{x} + \kappa\ 2\ )$.

Therefore, the decision rule to predict the class of $\vec{x}$ is given by

$$y = \text{sign}\ (\textstyle\sum_{i=1}^{s} \alpha_i\, y_i\, K(\vec{x}_i, \vec{x}\,) + b) \qquad (10)$$

### 2.2 Multi-Class SVM

Originally, SVMs handles only two-class classification problems. However, multi-class classification can be implemented by developing multiple two-class classifiers [11]. This can be done in many ways. The popular two are:

**One vs All Approach:** If adataset need tobe classified with given N classes, then Ntwo-class SVM classifiers should be created such that each classifier is trained to separate one class data items from the remaining N-1 classes data items. That is, one of the class data items are treated as one group and the remaining N-1 class data items treated as other group. So it can be reduced into two-class classification problem. In the same way by treating every class as one group and remaining N-1 classes as other group, we can develop N two-class SVM classifiers. During the testing, the class of data vector is determined by giving it as input to all N SVMs and the perpendicular distance from boundary of the linear separating hyper-planeto the data point is calculated. The final output is the class that belongs to the SVM with the largest margin. This approach is also known asthe winner-take-all approach.

**One vs One Approach:** Another strategy of achieving N class classification using the SVM is to build a set of one-versus-one classifiers. By considering every class data items as one group and every other class data items as second group, we can develop N ( N-1)/2 two-class SVM classifiers.During the testing, the class of data vector is determined by giving it as input to all N ( N-1)/2 classifiers and to choose the class that is selected by the most of the classifiers.

### 2.3 Rough Set Attribute Reduction

The aim of the Rough set attribute reduction (RSAR) is to decease the computation time and to improve classification accuracy [12][13]. Sometimes the significance of few attributes in classification may be zero. They may not have any role in class prediction. Such attributes increases computation time, memory usage and also makes the classification process complex. RSAR reads a dataset, and removes attributes that are irrelevant in categorizing the item. It also helps in identifying what attributes of a problem are most important.

The rough set theory is a mathematic tool proposed by Z. Pawlak, to deal with redundant, noise, fuzzy, uncertain and incomplete data [14]. It does not require any priori information related to the data set. In the rough set approach, the reduction of object description is usually done by the reduct. A reduct is a minimal set of attributes that preserve total data item properties. By applying normalization techniques, a set of candidate keys are generated. One of the candidate keys which gives class of data item is chosen as reduct. Thus a minimal set of attributes is generated.

## 3 Hybrid Prediction Model

The proposed hybrid model for the classification on medical data comprises of two phases: (1) The Rough set Attribute Reduction (RSAR) and (2) The Support Vector Machine (SVM). In the first phase, the Rough set attribute reduction reduces the dimensionality of given data set in a way such that without losing the generality of final result. When a data set with ‘n’ attributes is given as input, RSAR generates a data set with ‘m’ attributes (m < n) by removing few of the irreverent attributes or redundant attributes. The reduced data set which is output of RSAR is divided into training set and testing set. In the second phase,the SVM is used to predict the class of data item. The training data set is given as input to the SVM and it is trained to predict the correct class under supervised learning method. Once the training is completed, data items from testing set is given as input and accuracy of classification is calculated.The training and testing of SVM is carried out using WEKA [15].

## 4 Experiment and Results

The Cleveland Heart disease data set used in the experiment is collected from UCI machine repositories[16]. WEKA is used to implement the experiments. In the first phase, the data setis given input to LIBSVM [17] and their classification accuracy belongs to each type of SVM is calculated. In the second phase the Cleveland Heart disease data sets is given as input to RSAR and data set with reduced attributes is generated as output. These reduced attributes data set is given as input to LIBSVM and their corresponding classification accuracy are calculated. Linear SVM uses the kernel trick function as $K(\vec{x}_i,\vec{x}\ ) = \vec{x}_i.\vec{x}$, Quadratic SVM and Cubic SVM uses the Polynomial kernel formula as $K(\vec{x}_i,\vec{x}\ ) = (1 + \vec{x}_i\ ,\vec{x}\ )^{d}$ with degree ‘d’=3 and ‘d’=4 respectively, Gaussian SVM uses the RBF formula as $K(\vec{x}_i, \vec{x}) = \exp(-\gamma\ \|\ \vec{x}_i - \vec{x}\|^2\ )$. Ensemble SVM methodis a machine learning technique that combines multiple SVM base models to produce optimal results. This ensemble model produce better prediction result compared to any of the single SVM model used in it. We implemented ensemble SVM by combing three models such as linear SVM, Quadratic SVM and Gaussian RBF SVM. Each test instance is supplied to every model in the ensemble method to produce a prediction result with one vote and the final prediction output is that the one that receives more than half of the votes. If none of the prediction results get more than half of the votes, we considered the prediction result of Gaussian RBF SVM as the final result.The results are summarized as follows.

**Table 1.** Classification accuracy on Cleveland Heart Data

| Data Set | Classification Accuracy before attributes reduction | Classification Accuracy after attributes reduction |
|---|---|---|
| Linear SVM | 72.59% | 79.33% |
| Cubic SVM | 69.63% | 76.51% |
| Quadratic SVM | 79.36% | 84.36% |
| Gaussian RBF SVM | 82.97% | 87.67% |
| Ensemble SVM | 84.78% | 92.75% |

The comparison of classification accuracy results overCleveland Heart disease dataset using LIBSVM and proposed hybrid model of RSAR + SVM are shown below as bar charts. From the graphs it is evident that the proposed method performs very well.The cubic SVM produced least prediction result and Ensemble SVM produced highest prediction result as compared to other models. It also evident that prediction accuracy increase after attribute reduction.

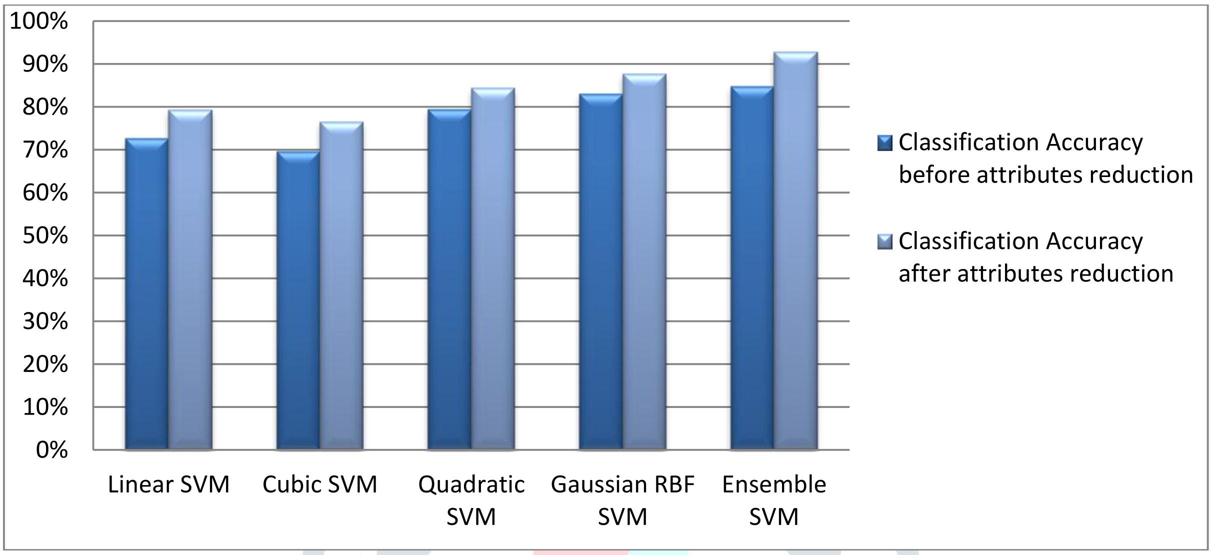

*Figure 1:* Comparison of Classification accuracy on Medical Data

## 5 Conclusion

During the identification of the disease, patient needs to undergo so many investigations. To understand the intensity of disease, these investigations are needed to be repeated frequently. It is a costly affair to the patients. Our work is reducing the investigations without affecting the prediction accuracy of the disease. It is a known fact that every disease has a set of symptoms (attributes). But all the attributes may not play an important role in identification of the disease. These attributes can be eliminated during the classification process. TheRSAR removes attributes that are irrelevant in classifying the item. The data is preprocessed by RSAR and the output is given to SVM to classify the data item.It is observed that attribute reduction on data set using rough set improved the classification accuracy as compared to data sets without attributes reduction. Further the proposed model also decreased computation time and memory usage of classification.

## References

1. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.
2. Chapelle, Olivier. "Training a support vector machine in the primal." Neural computation 19.5 (2007): 1155-1178.
3. Yao, Yiyu, and Yan Zhao. "Attribute reduction in decision-theoretic rough set models." Information sciences 178.17 (2008): 3356-3373.
4. Evanthia E. Tripoliti, Theofilos G. Papadopoulos, Georgia S. Karanasiou, Katerina K. Naka, Dimitrios I. Fotiadis, Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques, Computational and Structural Biotechnology Journal, Volume 15, 2017, Pages 26-47.
5. Srinivas, K., G. Raghavendra Rao, and A. Govardhan. "Rough-Fuzzy classifier: A system to predict the heart disease by blending two different set theories." Arabian Journal for Science and Engineering 39.4 (2014): 2857-2868.
6. Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36.4 (2009): 7675-7680.
7. Tan, Kay Chen, et al. "A hybrid evolutionary algorithm for attribute selection in data mining." Expert Systems with Applications 36.4 (2009): 8616-8630.
8. [a5] Mai Shouman, Tim Turner, Rob Stocker, " Using data mining techniques in heart disease diagnosis and treatment", IEEE Japan-Egypt Conference on Electronics, Communications and Computers, 2012.
9. [a6] Heon Gyu Lee, Ki yong Noh and Keun Ho Ryu, " Mining Biosignal Data: Coronary Artery Disease diagnosis Using Linear and Nonlinear Features of HRV", Springer-Verlag Berlin Heidelberg 2007.
10. Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) 2.02 (2010): 250-255.
11. Duan, Kai-Bo, and S. Sathiya Keerthi. "Which is the best multiclass SVM method? An empirical study." Multiple classifier systems. Springer Berlin Heidelberg, 2005. 278-285.
12. Chen, Hui-Ling, et al. "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis." Expert Systems with Applications 38.7 (2011): 9014-9022.
13. Chen, Rung-Ching, et al. "Using rough set and support vector machine for network intrusion detection system." Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference on. IEEE, 2009.
14. Pawlak, Zdzislaw, et al. "Rough sets." Communications of the ACM 38.11 (1995): 88-95.
15. Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.
16. Blake, Catherine, and Christopher J. Merz. "{UCI} Repository of machine learning databases." (1998).
17. Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 (2011): 27.