Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy

Varun Kumar.M
Assistant professor
School of Information Technology & Engineering
Vellore Institute of Technology
Vellore

Vijay Sharathi.V
M.S(Software Engineering)
School of Information Technology & Engineering
Vellore Institute of Technology
Vellore

Gayathri Devi.B.R
M.S(Software Engineering)
School of Information Technology & Engineering
Vellore Institute of Technology
Vellore

ABSTRACT
Data mining techniques are widely used in classification and prediction in the field of bioinformatics. This even helps in identifying the relationships and patterns in the data which helps in construction of prediction model. Classification and prediction model supports medical diagnosis which helps in improving the quality of patients. Noisy features are identified and eliminated by chi-square attribute evaluation which may further improve the classification accuracy of support vector machine. Hepatitis patients are those who need continuous special medical treatment to reduce mortality rate. Machine learning technologies are used for classification and prediction for Hepatitis patients.

General Terms
Data mining algorithm such as Support Vector Machine which is the most widely used algorithm for classification and prediction. Chi-square attribute evaluation is used to assign weight to the attributes thereby improving the classification accuracy.

Keywords
Support Vector Machine (SVM), Chi-Square attribute evaluation, Feature selection.

1. INTRODUCTION
Life Prognosis of hepatitis is quite a challenging task in early stage due to various interdependent features. A model can be developed which can be used in prediction of life prognosis of hepatitis disease. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques have been extensively used in bioinformatics to analyze biomedical data. Data mining algorithms can be used efficiently in prediction and classification of inter-related data. The objective of this analysis is to classify and scaling the accuracy of hepatitis.
RapidMiner is the most widely used data mining tool which support huge amount of data mining algorithms for classification and regression. Support Vector Machine (SVM) is the most widely used data mining algorithm in the field of medicine. Feature selection techniques are used to obtain dominant set of attributes.

2. LITRATURE SURVEY
In previous methods [1] which are used for prediction of hepatitis, wrapper method is used for feature selection in which attributes are removed based on trial and error method. Life Prognosis of hepatitis patients can be predicted by using classifier such as Support Vector Machine. In support vector machine, dataset is classified into training and testing data. Support vector machine analyze the training data and makes prediction on testing data. Predictive accuracy of classifiers can be enhanced by applying the techniques of feature selection. In this paper, Wrapper methods were incorporated to remove noise features before classification. After removal of noisy attributes, accuracy of the algorithm was further increased. SVM algorithm provides more and improved accuracy with the 10 attributes identified using wrapper method using a data mining tool called weka. Data mining concepts and techniques [2] provide us the how to preprocess data and handle with missing values. Preprocessing is important because the data collected in real world is incomplete, noisy and inconsistent. Preprocessing stages include data cleaning, data integration, data transformation and data reduction. Data cleaning is carried out for missing values and Noisy data. Data integration combines data from multiple sources into a coherent data store. Inconsistencies in attribute may result in redundancies and these redundancies can be detected by correlation analysis. Data Transformation involves smoothing, aggregation, generalization, normalization and attributes construction. Data reduction includes attribute subset selection, dimensionality reduction and discretization. [3] Feature selection is more significant for data mining algorithms for variety of reasons such as generalization performance, running time requirements and constraints. This method is based upon finding those features which minimize bounds on the leave-one-out error. Subsets of features are selected for preserving and improving the discriminative ability of a classifier. Feature selection for SVM is computationally feasible for high dimensional datasets. [4]Wrapper methods embed the model hypothesis search within the feature subset search. The feature subset selection conducts a search for good subset using induction algorithm such as ID3 and decision tree as a part of evaluation function. The accuracy of the induced classifiers is estimated using accuracy estimation techniques. To achieve the best possible performance with a learning algorithm on a particular training set, a feature subset selection method considers the interaction between the algorithm and the training dataset.

We proposed an empirical approach in which attributes are weighted by CHI-Square attribute evaluation and accuracy of algorithm is scaled and accuracy is improved when compared to previous results. In this approach, attributes are eliminated by evaluating their weights which thereby reduces time utilized in identifying a perfect combination based on trial and error method. Various (see Figure 2) stages in proposed method includes,

3. DATA PREPROCESSING

Dataset used in the prediction model should be more precise and accurate in order to improve the predictive accuracy of data mining algorithms. Dataset which is collected may have missing (or) irrelevant attributes. These are to be handled efficiently to obtain the optimal outcome from the data mining process.

3.1 Attribute Identification

Dataset collected from UC Irvine machine learning repository which consists of 155 instances and 19 attributes with the class stating the life prognosis yes (or) no. The dataset consist of 14 nominal attribute and 6 multi-valued attributes. The attributes which are identified are

Table 1.Attributes in dataset

. Attributes	Value
Class	die (1), live (2)
Age	numerical value
Sex	male (1), female (2)
Steroid	no (1), yes (2)
Antivirals	no (1), yes (2)
Fatigue	no (1), yes (2)
Malaise	no (1), yes (2)
Anorexia	no (1), yes (2)
Liver Big	no (1), yes (2)
Liver Firm	no (1), yes (2)
Spleen Palpable	no (1), yes (2)
Spiders	no (1), yes (2)
Ascites	no (1), yes (2)
Varices	no (1), yes (2)
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alk Phosphate	33, 80, 120, 160, 200, 250
SGOT	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	no (1), yes (2)

3.2 Data cleaning and feature selection

Dataset which is collected from UCI repository may have missing values and redundant attributes. Missing values can be handled either by removing the instances or replacing them by mean, average, maximum (or) minimum. Removing the instances may further reduce the amount of data, thereby reducing the quality in prediction. Hence these missing values are replaced by zero which doesn't much affect the quality of data.

Feature selection can be done by giving weights to the attributes. Attributes are weighted by PCA, SVM attribute evaluation and Chi-Square attribute evaluation using RapidMiner. In this paper, Chi-Square attribute evaluation is used since it works well with efficient data mining algorithm such as Support Vector Machine.

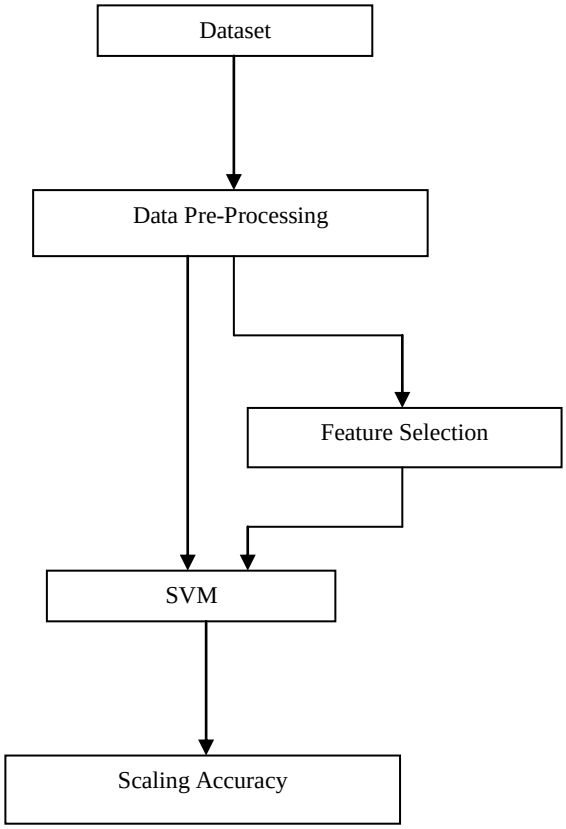

Fig 2: Architecture of proposed method

4. SUPPORT VECTOR MACHINE

Support vector machine is the most widely used in bio-informatics since it minimizes the expected error rate rather than reducing the classification error rate. SVM algorithm predicts well even if the testing data is entirely different from training data. SVM attempts to determine a plane that will have smallest generalization error, among the infinite number of planes. Support vector machine chooses the plane that maximizes the margin separating two classes. Wider is the gap smaller is the generalization error.

5. CHI-SQUARE ATTRIBUTE EVALUATION

Attributes are weighted by Chi-Square attribute evaluation which helps in identification of irrelevant attributes. These evaluated attributes are given to data mining algorithm which helps in classification and prediction of life prognosis of hepatitis disease. Chi-Square attribute evaluation is aimed at finding the optimal feature subset that enhances the classification accuracy of Support Vector Machine (SVM). Support vector machine with backward search approach leads to improvement of feature selection and enhances the classification accuracy.

Table 2. Chi-Square attribute evaluation

Attributes	Weight
Steroid	0.0
Antivirals	0.0
Anorexia	0.0
Liver Big	0.0
Liver Firm	0.0
Spleen Palpable	0.0
Alk Phosphate	0.0
SGOT	0.0
Sex	0.147
Age	0.245
Fatigue	0.458
Varices	0.518
Malaise	0.540
Histology	0.559
Spiders	0.705
Ascites	0.891
Protime	0.965
Bilirubin	0.986
Albumin	1.0

6. CLASSIFICATION ACCURACY

After removal of each attribute based on weight (see Table 2), set of features are given to Support Vector Machine which results in higher accuracy. Dominant set of features are obtained from these set of attributes which results in higher classification accuracy (see Figure 1). These minimal set of attributes are collected from patients and these attributes are enough to predict the life prognosis of hepatitis.

Table 3. Performance of classifier

Algorithm	Without Feature Selection	With Feature Selection
SVM	79.33%	83.12%

Dominant set of attributes which results in higher predictive accuracy (see Table 3).

Table 4. Comparative study of previous works

The name of Article	Classification Accuracy	
	Without Feature selection	With Feature selection
Proposed Method	79.33%	83.12%
Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method.	70.91 %	74.55%

Accuracy of the algorithm is based on the data preprocessing methods and management of missing values. Feature selection aids in improving the accuracy of classifier (see table 4).

7. CONCLUSION AND FUTURE WORK

Data mining techniques are widely used in interpretation of patterns and relationships in huge amount of data. In this research work, we propose a feature selection approach for classification and prediction of life prognosis of Hepatitis disease. The performance of this approach is analyzed by computing the classification accuracy of Support Vector Machine with and without feature selection which helps in reducing the number of clinical measures by 60%. Previous work and research reduces the attributes from 19 to 10 to attain maximum predictive accuracy. In our proposed model, attributes are further reduced to 8 with an improved accuracy of 83.12%. This prediction model helps the medical practitioner in effective decision making process with fewer attributes. These kind of model helps in reducing the human efforts in analyzing the clinical measures. Further, the dataset should be increased in order to attain the prediction with higher accuracy and this model can be further improvised which aid in improving the precautions and treatment to the patients with more optimal and accurate results. The constructed prediction model can also be used for identification of onset of severe diseases such as diabetes, heart disease, etc. Apart from bioinformatics, our prediction model can be used in data which involves prediction and classification even if data have missing or irrelevant values. Further, this constructed model can be extended to handle multi-valued classes which can be used in predicting the high and low values of stock market.

8. REFERENCES

[1] Roslina, A.H. and Noraziah, A "Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method", Seventh International Conference on Fuzzy Systems and knowledge Discovery (FSKD 2010), 978-1-4244-5934-6/10, 2010 IEEE.

[2] Jiawei Han and Micheline Kamber. "Data Mining: Concepts and Techniques",Data Preprocessing, Third Edition, 2011

[3] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V., " Feature Selection For SVMs", Advances in Neural Information processing Systems, MIT Press 2001, pg 668- 674.

[4] Ron Kohavai and George H. John., "Wrappers for feature subset selection" , Artificial Intelligence

[5] Hepatitis dataset, UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/ Irvine. CA University of California, School of information technology and computer science.

[6] Rapid – I 2011, Interactive Design. Product: RapidMiner, http://rapid-i.com/content/view/281/225/lang,en

[7] Mamdouh Refaat. "Data Preparation for Data Mining using SAS (The Morgan Kaufmann Series in Data Management Systems)", 2006

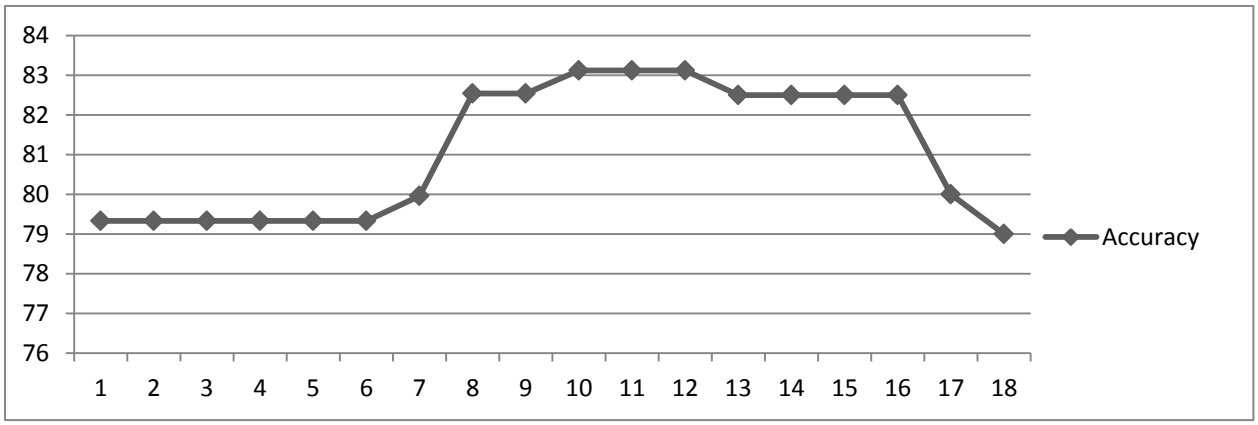

Fig 1: Classification of accuracy by removal of each attributes by weight