Volume-6,Issue-5 (Dec-16)
ISSN (O) :- 2349-3585

ANTICIPATING HEART DISEASE USING C4.5 CLASSIFICATION AUGMENTED WITH FEATURE SELECTION

Ms. P. Deepika[1], Ms. S. Saranya[2], Ms. S. Sasikala[3], Dr. S. Jansi[4], Ms. A. Kiruthika[5]
[1,3,5]Asst.Prof.,PG & Research Department of Computer Science,
[2]Asst.Prof., Department of Computer Applications,
[4]Asst.Prof., Department of IT/CT,
[1,2,3,4,5] Hindusthan College of Arts and Science,Coimbatore, India

ABSTRACT: Now a days, heart disease are the dangerous problem which cause death in and around countries. Several researches and technologies were implemented to diagnosis and treatment for heart disease. Decision Trees are simple predictive model which map input attributes to a target value using simple conditional rules. Decision tree give a direct and spontaneous way for obtaining the classification of a new instance from a set of simple rules. The removal of irrelevant or redundant attributes could benefit us in decisions making and scrutinize data efficiently. Frequent selection is one of the effective methods in data preprocessing for data mining. This paper shows the Heart disease prediction using various Decision Tree methods with Feature Selection technique.

KEYWORDS: Data mining, Classification, Decision Tree, ID3, C4.5, Feature Selection.

INTRODUCTION

Data mining is growing in various applications widely like medical diagnosis, credit card fraud detection and intrusion detection etc. Data mining is not a specific to one type of media or data. The main goal of using data mining is extract information from large dataset. Data mining has several methods like Clustering Classification, Regression and Association Rules to extract information for further usage. Classification is an effective data mining technique with wide-ranging applications to classify the various kinds of data used in nearly every field in our life. The classification is used to manage data, sometimes tree modeling of data to make predictions about new data. Classification is used to classify the item bestowing to the features of the item with respect to the predefined set of class.

HEART DISEASE

Heart Diseases remain the major cause of death for the last two decades. Recently Computer Technology and machine learning technique to develop software to assist doctors in making decision of Heart Disease in early stage. In Biomedical Field, Data mining plays effective role for prediction of disease. In Biomedical diagnosis, the information provided by the patients may include unneeded and interrelated symptoms and signs especially when the patient suffers from more than one types of disease of the same category. Data mining with intellectual algorithms can be used to tackle the said problematic of prediction in health dataset involving numerous inputs. The term Heart Disease refers to the disease of Heart & Blood vessel system within it.

TYPES OF HEART DISEASE

Heart means “Cardio”. Heart disease is a huge term that includes all types of disease affecting different components of Heart. The types of Heart disease are

1. Coronary Heart Disease

It is a condition in which plaque deposits block the Coronary Blood vessels leading to a reduced supply of blood and oxygen to the heart.

2. Angina pectoris

It is a chest pain that occurs due to insufficient supply of blood to the heart. This chest pain is intervals ranging for few seconds or minutes.

3. Congestive Heart Failure

It is a condition where the heart cannot pump enough blood to the rest of the body. This is known as heart failure.

4. Cardiomyopathy

It is the weakening of heart muscle or a change in the structure of the muscle due to inadequate heart pumping.

5. Congenital Heart Disease

It is the formation of the abnormal heart due to a defect in the structure of the heart or its functioning.

6. Arrhymias

It refers to the abnormal beat of the heart. The Heart beat can slow, fast or irregular.

7. Myocarditis

It is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart.

DECISION TREE

Decision trees are simple predictive model which map input attributes to a target value using simple conditional rules. It gives direct and intuitive way for obtaining the classification of a new instance from a set of simple rules. Decision Tree approach is a powerful method in classification problems. There are two steps in this technique. First one is building a tree and second one is applying the tree to the dataset. There are many decision tree algorithms are used. That are ID3,C4.5, CART (Classification and regression Tree) and J48 etc. Most Decision Tree algorithms are based on a greedy top-down recursive partitioning strategy for tree growth. They use dissimilar variants of impurity measures like: Information Gain, Gain Ratio and Distance based measures to choose an input attribute to be related with an internal node.

FEATURE SELECTION

Feature selection plays an foremost role in Data mining. Feature Selection is one of the most significant and widely used techniques for data preprocessing in Data mining. In the Classification task, the main aspire of feature selection is to reduce the number of attributes used in classification while maintaining acceptable classification accuracy. Feature Selection primarily affects the training phase of classification. After generating features, instead of processing data with the whole features to the learning algorithm directly, feature selection for classification will perform feature selection to select a subset of features and then process the data with the selected features to the learning algorithm.

Feature selection helps to identify the fields that are most important in predicting a certain outcome. There are three steps involved in Feature Selection. That are

□ Screening

□ Ranking

□ Selecting

Screening:

Screening is used to remove the unimportant and problematic predictors and records or cases, such as predictors to many missing values or predictors with too much or too little variant to be useful.

Ranking:

Ranking is used to sort remaining predictors and assigns the ranks based on importance of predictors.

Selection:

Selection is used to identify the subset of features by preserving only the most important predictors and filtering or excluding all other predictors.

FEATURE SELECTION METHODS

There are two types of Feature Selection methods. That are

□ Forward Feature Selection Method

□ Backward Feature Selection Method

FORWARD FEATURE SELECTION METHOD

The Forward Feature Selection method begins by evaluating every feature subsets which consist of only one input attribute.

BACKWARD FEATURE SELECTION METHOD

This method start with all the variables and remove them one by one, at each step removing the one that decreases the error the most, until any further removal increases the error extensively.

CATEGORIES OF FEATURE SELECTION

Features can be characterized into two types.

□ **RELEVANT**

□ **IRRELEVANT**

STEPS OF FEATURE SELECTION

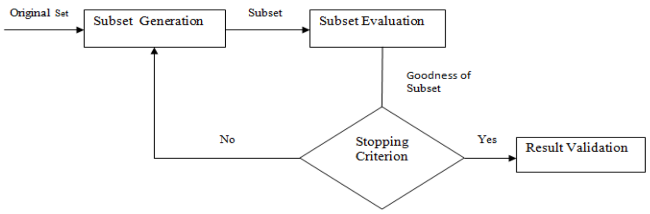
Feature selection process have four basic steps. That are

□ Subset Generation

□ Subset Evaluation

□ Stopping Criterion

□ Result Validation



ATTRIBUTES SELECTION METHODS

There are two methods involving in the selection of the attributes from the data set. That are Filter method and Wrapper method

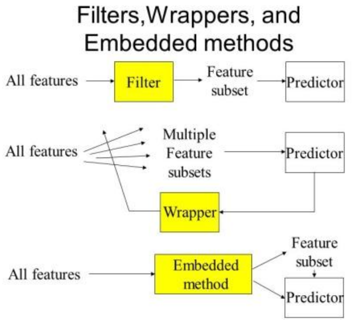


Figure 1: Feature Selection

FILTER METHOD:

In filter method, the attribute selection method is independent and the data mining algorithms are applied to the selected attributes. Advantages of filter techniques are that they easily scale to high dimensional dataset and the computation is simple and fast.

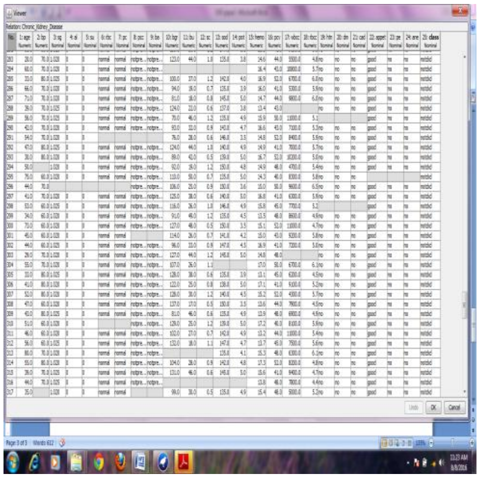
WRAPPER METHOD:

In wrapper method, the attribute selection uses the result of data mining algorithm to determine how good a given attribute set is. The major advantage of wrapper method is that the quality of an attribute is directly measured by the performance of the data mining algorithm applied to that attribute subset. This wrapper method includes the interaction among feature subset search and model selection, and the ability to catch into account feature dependencies.

FOCUS ON THE SURVEY

Mahdi Ismailia et al. proposed the feature selection with the Decision Tree construction. They proposed an algorithm that is used to arrange attributes based on their importance using two independent criteria. Then, they organize the attribute can be used as input one simple and influential algorithm for constructing Decision tree. In their proposed method they had two phases. First phase is used for attribute ranking algorithm(ARA). Using this algorithm, the attributes are ranked and fed as input for the second phase. The ARA includes two parts such as class distance ratio and an attribute-class correlation measure. Next Second phase is used to generate rules called Top-Down Induction of Decision Trees(TDIDT). From this the evaluation parameters such as size of trees, number of leaves, error rate, Recall and precision are computed. They showed the result by producing the smaller tree compared with other algorithms such as J48 and BFTree. And also they suggested that if we want to improve the accuracy by using this method, we can repeat the two phase of their proposed algorithm instead of using TDIDT method. S.Saravanakumar et al. proposed a Frequent Feature Selection Method For Effective Heart Disease Prediction. This method comes from the use of the fuzzy measure and the relevant nonlinear integral. They used WEKA tool for their experiments. And also they used data set of 1000 records with 8 attributes is used. The combination of heart attack parameters for normal and risk level along with their values and weight ages are mentioned. From the result, the lesser value (0.1) of weightage comprises the normal level and the higher values other than 0.1 comprise the higher risk level.

T.John Peter et. al. proposed a development method of novel feature selection framework for heart disease prediction. They proposed a feature selection method for improving the performance of classification methods. Their experiment was conducted on dataset of health care domain. They found that CFS and FILTER SUBSET EVALUATION reduces more number of irrelevant and redundant attributes thereby increases the performance of classifiers. In addition the new feature selection namely CFS and BT was proposed. This algorithm gives better accuracy for NB and KNN classifiers. They conclude that CFS and Bayes theorem feature selector is best suitable for heart disease data prediction. Porf.K.Rajeswari et al. analyzed the approach of feature selection for classification and also presented a innovative approach for the feature selection via using association and correlation mechanism. The aim of their work is to select the correlated features or attributes of the medical dataset so that patient need not to go for various tests and in future it is used for organizing the clinical decision support system which is supportive for decision making of disease prediction. By removing irrelevant attributes the performance of the classification can be improved and also cost of classification may gets reduced. They examined two different methods of feature selection. They also exhibited that a novel approaches for feature selection using correlation and by generating association rules.



Viewer

Relation: Chronic_Kidney_Disease

No. | 1: age Numeric | 2: bp Numeric | 3: sg Nominal | 4: al Nominal | 5: su Nominal | 6: rbc Nominal | 7: pc Nominal | 8: pcc Nominal | 9: ba Nominal | 10: bgr Numeric | 11: bu Numeric | 12: sc Numeric | 13: sod Numeric | 14: pot Numeric | 15: hemo Numeric | 16: pcv Numeric | 17: wbcc Numeric | 18: rbcc Numeric | 19: htn Nominal | 20: dm Nominal | 21: cad Nominal | 22: appet Nominal | 23: pe Nominal | 24: ane Nominal | 25: class Nominal

Figure 2: Features File

Results

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===

=== Summary ===

```
Correctly Classified Instances        396               99      %
Incorrectly Classified Instances        4                1      %
Kappa statistic                      0.9788
Mean absolute error                    0.0182
Root mean squared error                  0.0902
Relative absolute error                3.8914 %
Root relative squared error             18.6237 %
Total Number of Instances               400
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area   Class
 0.984     0         1        0.984    0.992      1       HD
 1        0.016     0.974     1        0.987      1       notHD
Weighted Avg.    0.99      0.006      0.99      0.99      0.99       1
=== Confusion Matrix ===
  a   b   <-- classified as
 246   4 |   a = HD
   0 150 |   b = notHD
```

Feature selection improves the accuracy of the decision tree prediction as witnessed from the results.

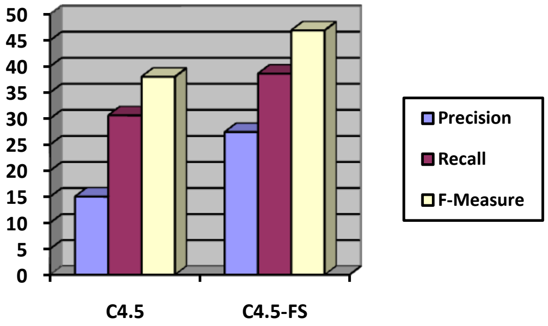


Figure 3: Performance Comparison of c4.5 – with and without feature selection

CONCLUSION

Heart Disease is one of important disease which is dangerous for human life. Researchers have been inspecting applying dissimilar data mining techniques to help health care specialists in the diagnosis of heart disease patients. The readiness of huge amount of data leads to the need of influential data analysis tool to extract useful knowledge. Diagnosis Disease is one of the main applications where data mining tools are showing efficacious results. Feature Selection is a term commonly used in data mining tools and techniques available for dropping inputs to a convenient size of processing and analysis. Feature selection techniques has extensive variety of applications in countless areas like data mining, digital image processing etc. Feature Selection is used to select the attributes which gives better results than using all the attributes in the data set.

REFERENCES

[1] Mahdi Esmaeili, Fazekas Gabor, " Feature Selection as an Improved Step for Decision Tree Construction" 2009 International Conference on Machine Learning and Computing, IPCSIT Vol.3(2011), IACSIT Press, Singapore.

[2] S.Saravanakumar, S.Rinesh, " Effective Heart Disease Prediction using Frequent Feature Selection Method"

International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue, March 2014.

[3] T.John Peter , K. Somasundaram , "Study and Development of Novel Feature Selection Framework for Heart Disease Prediction" International Journal of Scientific and Research Publications, Volume 2, Issue 10, October 2012.

[4] Sunita Beniwal, Jitender Arora, "Classification and Feature Selection Techniques in Data Mining", International Journal of Engineering Research & Technology, Volume 1, Issue 6, August 2012.

[5] L.Ladha, T.Deepa, " Feature Selection Methods and Algorithms" International Journal of Computer Science and Engineering, Volume 3, No.5, May 2011.

[6] Prof. K. Rajeswari, Dr. V. Vaithiyanathan and Shailaja V. Pede," Feature Selection for Classification in Medical Data Mining" International Journal of Emerging Trends & Technology in Computer Science, Volume 2, Issue 2, March-April 2013.

[7] Ritu Ganda, Vijay Chahar, " A Comparative Study on Feature Selection Using Data Mining Tools" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.

[8] YongSeog kim, W. Nick Stree't and Filippo Menczer, University of Lowa, USA, " Feature Selection in Data Mining".

[9] Pushpalata Pujari, Jyoti Bala Gupta," Improving Classification Accuracy by Using Feature Selection and Ensemble Model" International Journal of Soft Computing and Engineering, Volume 2, Issue 2, May 2012.

[10] R.Nithya, B.Santh," Mammogram Classification using Different Feature Selection Method" Journal of Theoritical and Applied Information Technology, Volume 33, Issue 2, November 2011

Deepika qualified M.Sc.,M.Phil., hass been working as an Asssistant Professor foe the past 3 years.

Saranya qualified MCA.,M.Phil. hass been working as an Asssistant Professor foe the past 7 years.

Sasikala qualified M.Sc.,MCA.,M.Phil.,PGDPM & IR, Ph.d hass been working as an Asssistant Professor foe the past 10 years. She has 71 publications and h-index value 5.

Dr.Jansi qualified MCA.,M.Phil.,Ph.d hass been working as an Asssistant Professor foe the past 1 year. Her Research has spanned a large number of disciplines like image analysis and soft computing.

Kiruthika qualified M.Sc., M.Phil., hass been working as an Asssistant Professor foe the past 3 years.