INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2277 - 2281
RECEIVED : 25.07.2017 PUBLISHED : 19.08.2017
Aug 19, 2017

Prognosis of Lung Cancer Using Data Mining Techniques

[1]C. Saranya, M.Phil, Research Scholar, Dr.M.G.R.Chockalingam Arts College, Arni
[2]K. R. Dillirani, Associate Professor, Department of Computer Science, Dr.M.G.R.Chockalingam Arts College.

Abstract

Data mining is the extraction of hidden predictive information and unknown data, patterns, relationships and knowledge by exploring the large data sets which are difficult to find and detect with traditional statistical methods. Data mining it is powerful technology which will discover most important information from the data warehouse of the organizations. It is a very crucial step that collectively examine large amount of routinely data. To find latest patterns in healthcare industry, there exist various interactive and scalable data mining methods. Data mining is a quantitative approach which is user friendly in reading reports and reducing errors and controls the quality more uniformly. Important task of data mining is data pre-processing. Data mining tools are used for decision making. Prediction and classification techniques are used in which classification technique predicts the unknown values with respect to generated model. An assortment of data mining techniques can be applied to find associations and regularities in data, extract knowledge in the forms of rules and predict the value of the dependent variables.

Keywords: Data mining, Error, Generated model.

1. INTRODUCTION:

In this thesis we propose an ensemble approach for feature selection, where multiple feature selection techniques are combined to yield more robust and stable results. Ensemble of multiple feature ranking techniques is performed in two steps. The first step involves creating a set of different feature selectors, each providing its sorted order of features, while the second step aggregates the results of all feature ranking techniques. The ensemble method used in our study is frequency count which is accompanied by mean to resolve any frequency count collision.Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships from large amounts of data stored in databases, data warehouses, or other information repositories. Data mining has two approaches. The first approach tries to produce an overall summary of a set of data to identify and describe main features. The second approach, pattern detection, seeks to identify small unusual patterns of behavior. The data mining analysis tasks typically fall into the following categories: data summarization, segmentation, classification, prediction, dependency analysis. One of the models is CRISP-DM [1]. It is a De Facto standard for industry. The CRISP- DM project began in mid-1997 to define and validate an industry and tool-neutral data mining process model. The six steps developed in this model are: business understanding, data understanding, data preparation, modeling, evaluation and deployment. Business understanding is the phase of understanding objectives and requirements of a project. Data understanding is the phase of becoming familiar with the data like identifying data quality problems, discover first insights into data. Data preparation phase describes the entire activities essential in constructing a final dataset from raw data. In the modeling phase

various modeling techniques are selected and applied to the model. Evaluation is the phase where the project is thoroughly evaluated before the final deployment. Deployment is the phase where the knowledge discovered will be organized and presented in a way a client can use. Therefore, it grasp various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological "taxonomies", to medical "syndromes" and genetic "genotypes" to manufacturing "group technology"— the problem is identical: forming categories of entities and assigning individuals to the proper groups within it. Following are the various clustering algorithms used in healthcare.

2. RELATED WORK

This section provides a brief coverage of the works performed in the area of ensemble feature ranking. These works assess how an ensemble of feature ranking techniques can improve robustness, performance and diversity. Feature ranking is a process of selecting the most relevant features from a large set of features. It is considered as one of the most critical problems researchers face today in data mining and machine learning. The main focus of ensemble feature ranking approach is on improving classification performance through the combination of feature ranking techniques. Very limited research exists on ensemble feature ranking. Early studies on ensemble of feature ranking techniques were performed by Rokach et al. [28]. The experiments in this study are performed to check whether ensemble of feature subsets improve classification accuracy over individual rankers. The experiments are performed on datasets obtained from UCI machine learning repository. Five different feature selection algorithms were used to generate 10 ensembles. The combining methods used for ensemble are: majority voting, take-it-all, smaller is heavier. The ensembles were evaluated using C4.5 classification model. The experimental results have shown that ensemble method performed better than individual feature rankers.

Olsson and Oard [30] performed a study on combining feature selectors for text classification. The experiments were performed on two sets containing 23, 149 documents and 200,000 documents from RCV1-v2. The documents were combined using document frequency thresholding, information gain and the chi-square feature selection methods. The combination methods used are highest rank, lowest rank and average rank combination. The documents were classified using k-nearest neighbours with k=100. The evaluation criteria used for this study was R-precision.

Wilker et al. [6] performed a study using six standard and eleven threshold based filter based feature ranking techniques. In this study six ensemble approaches were considered based on standard and threshold based filters. In addition, four other ensemble approaches were developed based on their robustness to class noise. This study used seven datasets from different domain applications, with different dimensions and different level of class imbalance. This work was evaluated on binary classification datasets. The experimental results showed that ensemble robustness can be predicated from the knowledge of individual components.

Cancer is potentially fatal disease. Detecting cancer is still challenging for the doctors in the field of medicine. Even now the actual reason and complete cure of cancer is not invented. Detection of cancer in

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2277 - 2281
RECEIVED : 25.07.2017 PUBLISHED : 19.08.2017
Aug 19, 2017

earlier stage is curable. In this work we have developed a system called data mining based cancer prediction system. The main aim of this model is to provide the earlier warning to the users and it is also cost and time saving benefit to the user. It predicts three specific cancer risks. Specifically, Cancer prediction system estimates the risk of the breast, skin, and lung cancers by examining a number of userprovided genetic and non-genetic factors. This system is validated by comparing its predicted results with the patient's prior medical record and also this is analyzed using weka system.

3. PROCESS

Data mining is the process of automatically collecting large volumes of data with the objective of finding hidden patterns and analyzing the relationships between numerous types of data to develop predictive models. we use the classification techniques. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Dataset used in this model should be more precise and accurate in order to improve the predictive accuracy of data mining algorithms. Which is collected may have missing (or) irrelevant attributes.

Name of the Algorithm	Time Taken to build the decision tree
Naive Bayes	0.01 seconds
Decision Table	0.05 seconds
J48	0.03 seconds

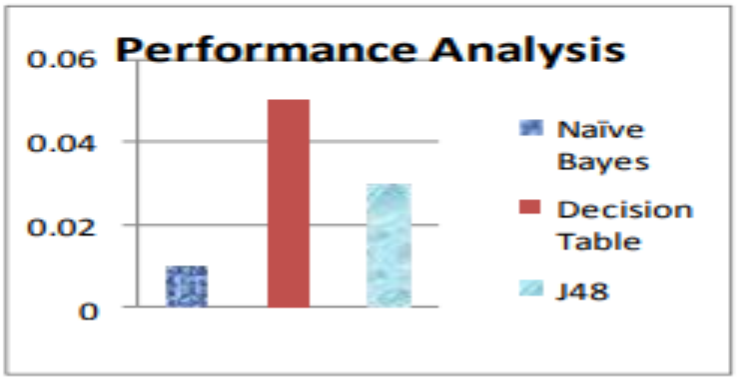

Fig.1. Analysis

Data mining can be used to help predict future patient behavior and to improve treatment programs. By identifying high-risk patients, clinicians can better manage the care of patients today so that they do not become the problems of tomorrow. For instance, early stage lung and oral cancers are very hard to diagnose by conventional means; genomic signatures can be used to provide more timely and perhaps more accurate diagnosis [2]. Data mining refers to extracting or "mining" knowledge from large amounts of data. It is an increasingly popular field that uses statistical, visualization, machine learning, and other data manipulation and knowledge extraction techniques aimed at gaining an insight into the relationships and patterns hidden

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2277 - 2281
RECEIVED : 25.07.2017 PUBLISHED : 19.08.2017
Aug 19, 2017

in the data [3]. In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, and the World-Wide Web. Advanced database systems include object-oriented, object-relational databases, and specific application-oriented databases, such as biological database, spatial databases, time-series databases, text databases, and multimedia databases

4. ANALYSIS

We also perform ANOVA test on the AUC performance metric. ANOVA is an acronym for Analysis of Variance. It is defined as a procedure for assigning sample variance to different sources and making a decision if the variation is within or among different population groups. Samples are described in terms of variation around group means and variation of group means around an overall mean. If variations within groups are small relative to variations between groups, a difference in group means may be inferred. Hypothesis Tests are used to quantify decisions.

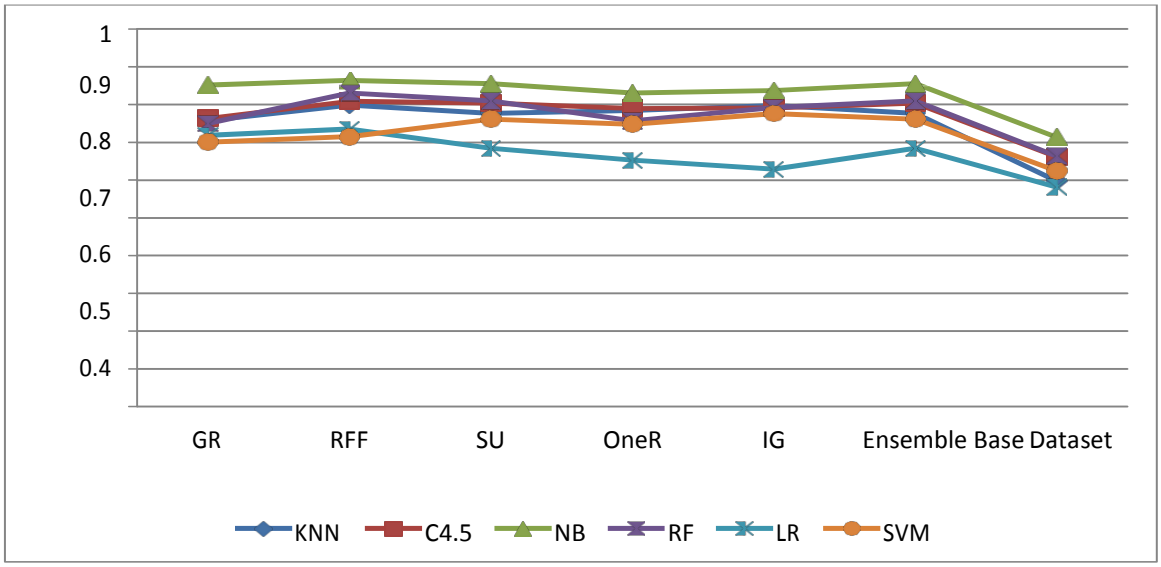

Fig.2. Waveform output

In this notation parameters with two subscripts, such as $(\alpha\beta)_{ij.}$, represent the interaction effect of two factors. The parameter $(\alpha\beta\gamma)_{ijk}$ represents the three-way interaction. An ANOVA model can have the full set of parameters or any subset, but conventionally it does not include complex interaction terms unless it also includes all simpler terms for those factors.

CONCLUSION

In this thesis, we have reviewed feature selection and explained the basic concept of different feature selection methods: filter, wrapper and hybrid model. We reviewed four filter based feature ranking techniques and one wrapper based feature ranking technique. They are information gain, gain ratio, symmetrical uncertainty, reliefF and oneRattribute evaluation. We examined classification models that are built using various classification techniques such as naïve bayes, k-nearest neighbor, Bayesian Network, support vector machine, Logistic Regression (J48) and decision trees. We took a brief review of the

evaluation criteria used to evaluate the classification models. We have also introduced ensemble methods for feature ranking technique that can help build stable and robust classification models.

REFERENCES

[1] J. Jackson, "Data Mining: A Conceptual Overview", Communications of the Association for Information Systems (Volume 8, 2002), pages 267- 296.

[2] H. Wang, T. M. Khoshgoftaar, K. Gao, "Ensemble feature selection technique for software quality classification", Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering, Redwood City, San Francisco Bay, CA, USA, July 1 - July 3, 2010, pages 215-220.

[3] H. Wang, T. M. Khoshgoftaar, A. Napolitano, "A comparative study of ensemble feature selection techniques for software defect prediction", Proceedings of the Ninth International Conference on Machine Learning and Applications, Washington, DC, USA, December 12-14, 2010, pages 135-140.

[4] H. Liu, H. Motoda, R. Setiono, Z. Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining", JMLR: Workshop and Conference Proceedings 2010, Volume: 4, Publisher: Citeseer, pages 4-13.

[5] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation- Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Leaning, ICML-03, Washington, D.C., August, 2003, pages 856-863.

[6] W. Altidor, T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, "Ensemble feature ranking methods for data intensive computing applications", Handbook of data intensive computing, Springer Science + Business media, LLC 2011, pages 349 -376.