IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 8, October 2014.
www.ijiset.com
ISSN 2348 – 7968

A Literature Review of Predicting Cancer Disease Using Modified ID3 Algorithm

Mr.A.Deivendran[1], Ms.K.Yemuna Rane M.Sc., M.Phil[2].,

[1]M.Phil Research Scholar, Dept of Computer Science, Kongunadu Arts and Science college, Coimbatore-29

[2]Assistant Professor, Dept of Computer Applications, Kongunadu Arts and Science college, Coimbatore-29

ABSTRACT

Data mining techniques have been used in medical research for many years and have been known to be effective. Cancer has become the leading cause of death worldwide. Cancer is one of the dreadful diseases in the world claiming majority of lives. Though cancer research is generally clinical and biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where data mining techniques have to be applied. Implementing accuracy in this classification by finding out the ratio of cancer in male versus female cases and also which areas record highest cancer rate and whether habits, diets, education, marital status, living area etc., which play important roles in cancer pattern. The proposed system enforces the two unusual data imputation process in a task and the aim is to conclude the probability of finding missing data in blood cancer and occurrence of blood cancer using improved ID3 algorithm. Cancer is one of the deadliest diseases found among many people across the world. This research aims at helping the medical practitioners to diagnose the patients at the early stage which can reduce the number of deaths.

Keywords— ID3, classification, decision learning, Common Disease, Prediction.

INTRODUCTION

To find out the missing values, sometimes prediction may use to fill the data. Prediction should be highly accurate. Execute a Prediction of cancer disease using with modified ID3 algorithm. The modifiedID3 algorithm will compare the current spatial database with the normal database from the input database. From the data set, an operational database will be created for the cancer patients and a database for normal patients. This database will be individual and many numbers of practical data are available. The result will be recovered from the dataset. Whether

ISSN 2348 – 7968

the patient will affect from cancer or not, also their infection ration percentage can be find out. Along with the lost values in the database during the time of data migrating. And also to analyzing the occurrences of cancer patients using data density clustering.

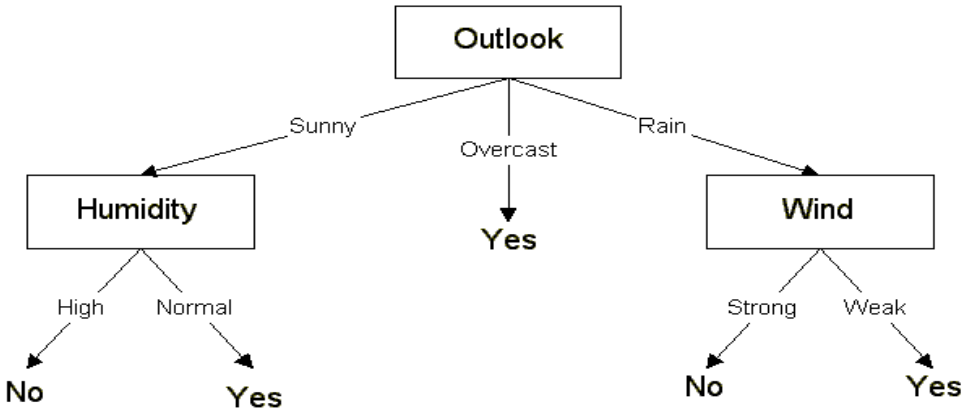

Figure. Decision Tree

LITERATURE SURVEY

The current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The goal of this study is to identify the most well performing data mining algorithms used on medical databases. The following algorithms have been identified: Decision Trees, Support Vector Machine, Artificial neural networks and their Multilayer Perception model, Naïve Bayes, Fuzzy Rules. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results more effective.

Performance of Decision tree and Logistic regression in cancer prediction

Performances of classification techniques were compared in order to predict the presence of the patients getting a blood cancer. A retrospective analysis was performed in 303 subjects. We compared the performance of logistic regression(LR), decision trees(DTs), and Artificial neural networks (ANNs). The variables were medical profiles are age, Sex, Chest Pain Type, Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum Heart Rate, Induced

ISSN 2348 – 7968

Angina, Ole Peak, Slope, Number Colored Vessels, Thal and Concept Class. We have created the model using logistic regression classifiers, artificial neural networks and decision trees that they are often used for classification problems. Performances of classification techniques were compared using lift chart and error rates. In the result, artificial neural networks have the greatest area between the model curve and the baseline curve. The error rates are 0.22, 0.198, 0.21, respectively for logistic regression, artificial neural networks and decision trees. The neural networks exhibited sensitivity of 81.1%, specificity of 78.7% and accuracy of 80.2%, while the decision tree provided the prediction performance with a sensitivity, specificity and accuracy of 81.7%, 76.0% and 79.3%. And the logistic regression provided the prediction performance with a sensitivity, specificity and accuracy of 81.2%,73.1% and 77.7% Artificial neural networks have the least of error rate and has the highest accuracy, therefore Artificial neural networks is the best technique to classify in this data set.

Performance Comparative Study

Logistic Regression (LR) is a well known classification method in the field of statistical learning. It allows probabilistic classification and shows promising results on several benchmark problems. Logistic regression enables us to investigate the relationship between a categorical outcome and a set of explanatory variables. Artificial Neural Networks (ANNs) are popularly used as universal non-linear inference models and have gained extensive popularity in recent years. Research activities are considerable and literature is growing. The goal of this research work is to compare the performance of logistic regression and neural network models on publicly available medical datasets. The evaluation process of the model is as follows. The logistic regression and neural network methods with sensitivity analysis have been evaluated for the effectiveness of the classification. The classification accuracy is used to measure the performance of both the models. From the experimental results it is confirmed that the neural network model with sensitivity analysis model gives more efficient result.

Result

ID3 algorithm derived 100 different association rules that were ranked and ordered with several accuracy level values. The rule with the highest accuracy had a value of 0.99498 and the one with lowest accuracy was observed as 0.9733. This accuracy term denotes metrics

ISSN 2348 – 7968

for ranking the association rules by means of confidence, which is the proportion of the examples covered by the premise that are also covered by the consequent ones among these rules, some of them had one condition or attribute with a resultant condition where some others had two or more combined condition that corresponds to a specific condition.

The sensitivity analysis provides information about the relative importance of the input variables in predicting the output field. In the process of performing sensitivity analysis, the ANN learning is disabled so that the network weights are not affected. The basic idea is that the inputs to the network are perturbed slightly, and the corresponding change in the output is reported as a percentage change in the output. The first input is varied between its mean plus (or minus) a user-defined number of standard deviations, while all other inputs are fixed at their respective means. The network output is computed and recorded as the percent change above and below the mean of that output channel. This process is repeated for each and every input variable. As an outcome of this process, a report (usually a column plot) is generated, which summarizes the variation of each out-put with respect to the variation in each input. The sensitivity analysis performed for this research project and presented in a graphical format in Fig. 4,lists the input variables by their relative importance(from most important to least important). The value shown for each input variable is a measure of its relative importance, with 0 representing a variable that has no effect on the prediction and 1.0 representing a field that completely dominates the prediction.

The sensitivity analysis results are based on the 10 different ANN models developed for the 10 data folds. After each of the 10 training, the network weights are frozen (testing stage) and the cause and effect relationship between the independent variables and the dependent variables are investigated as per the above-mentioned procedure. The aggregated results are summarized and presented as a column plot in Fig. The x-axis represents the input variables and the y -axis represents the percent change on the output variables, while the input variables (one at a time) are perturbed gradually around their mean with the magnitude of 1 standard deviation.

ISSN 2348 – 7968

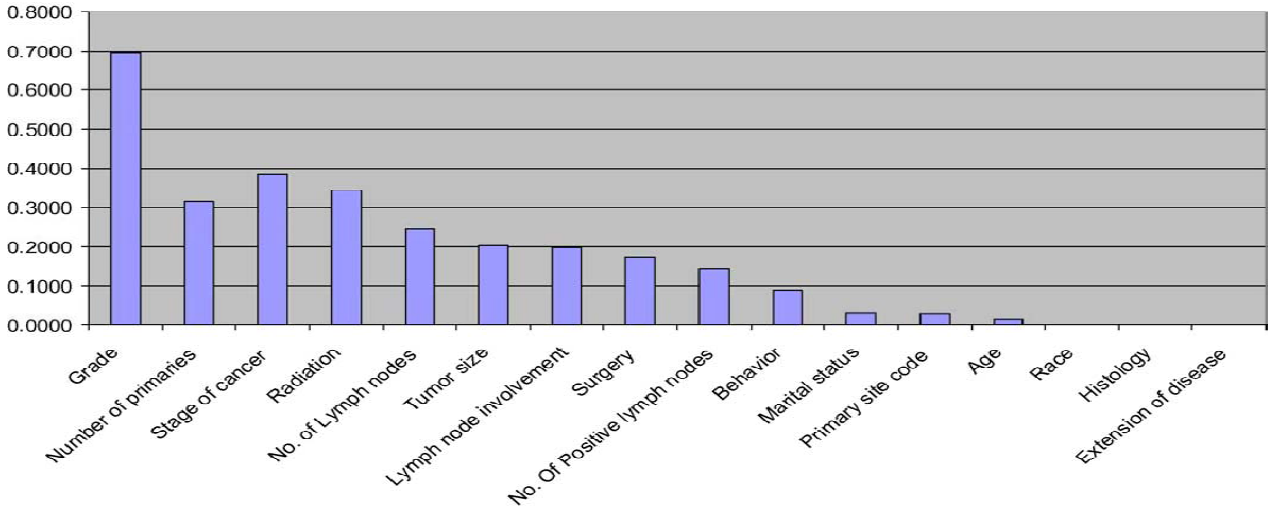

Figure. Sensitivity analysis graph.

The second most important variable in the pre-diction of survival is Stage of Cancer. This variable is concerned with the degree to which the cancer has spread, from In situ (noninvasive) to Distant (spread other areas). This is a chronological factor based on the amount of time the disease is present. The longer the disease is present the more chance that it will spread to other areas and the worse the prognosis.

Survey Report

In the general hospital wards (GHWs), a collection of features of a patient are repeatedly measured at the bed-side. Such continuous measuring generates a high-dimensional data stream for each patient. Most indicators are typically collected manually by a nurse, at a granularity of only a handful of readings per day. Errors are frequent due to reading or input mistakes. The values of some features are recorded only once within an hour. Other features are recorded only for a subset of the patients. Hence, the overall high dimensional data space is sparse. This makes our dataset a very irregular time-series. It is obvious that they have multi-scale gaps: different vital signs have different time gaps. Even more, a vital sign may not have the same reading gap.

A number of scoring systems exist that use medical knowledge for various medical conditions. For example, the effectiveness of Several Community-Acquire Pneumonia (SCAP) and Pneumonia Severity Index (PSI) in predicting outcomes in patients with pneumonia is evaluated in Similarly, outcomes in patients with renal failures may be predicted using the Acute

ISSN 2348 – 7968

Physiology Score (12 physiologic variables), Chronic Health Score (organ dysfunction), and APACHE score. However, these algorithms are best for specialized hospital units for specific visits. In contrast, the detection of clinical deterioration on general hospital units requires more general algorithms. For example, the Modified Early Warning Score (MEWS) uses systolic blood pressure, pulse rate, temperature, respiratory rate, age and BMI to predict clinical deterioration. These physiological and demographic parameters may be collected at bedside making MEWS suitable for a general hospital an alternative to algorithms that rely on medical knowledge is adapting standard machine learning techniques. This approach has two important advantages over traditional rule based algorithms. First, it allows us to consider a large number of parameters during prediction of patients' outcomes. Second, since they do not use a small set of rules to predict outcomes, it is possible to improve accuracy. Machine learning techniques such as decision trees, neural networks and logistic regression have been used to identify clinical deterioration. In integrate heterogeneous data (neuroimages, demographic, and genetic measures) is used for Alzheimer's disease (AD) prediction based on a kernel method. A support vector machine (SVM) classifier with radial basis kernels and an ensemble of templates are used to localize the tumor position in Also, in a hyper-graph based learning algorithm is proposed to integrate micro array gene expressions and protein-protein interactions for cancer outcome prediction and bio-marker identification.

CONCLUSION

Preventing clinical deterioration of hospital patients is a leading research in U.S. and every year a lot of money is spent in such area. The proposed system has developed as a predictive system for patients that can provide early warning of deterioration. This is an important advance, representing a significant opportunity to intervene prior to clinical deterioration. In this a ID3 technique introduced to capture the changes in the blood. Mean while, the system can handle the missing data so that the vicious who do not have all the parameters can still be classified. Here a pilot feasibility study has conducted by using a combination of logistic regression, bucket bootstrap aggregating for addressing over fitting, and exploratory under sampling for addressing class imbalance. Along with this combination can significantly improve the prediction accuracy for all performance metrics, over other major methods. Further, in the real-time system, the EMA smoothing also may apply to tackle volatility of data inputs and model outputs.

ISSN 2348 – 7968

REFERENCE

1. T.Sakthimurugan ―An Effective Retrieval of Medical Records using Data Mining Techniques‖ International Journal of Pharmaceutical Science and Health Care, Issue 2, Volume 2 (April 2012).

2.QuKaishe, Cheng Wenli, Wang Junwang. Improved Algorithm Based on ID3[J]. Computer Engineering and Applications. 39(25): 104107,2003.

3.Elma kolce(cela),Neki Frasheri "A Literature Review of Data Mining Techniques used in Healthcare Databases". ICT Innovations 2012 Web Proceedings - Poster Session ISSN 1857-7288.

4.D.S.Kumar, G.Sathyadevi, S.Sivanesh Decision(2011) ."Support System for Medical Diagnosis Using Data Mining ".

5.N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik."Data Mining Machine Learning Approaches and Medical Diagnose Systems ". A Survey Algorithm".Global Journal of Computer Science and Technology-Page 38,Vol-10,Ver-1.0.

6.E.Jiawei Hen and Micheline Kamber(2006) "DataMining Concepts and Techniques ".CA:Elsevier Inc,SanFranciso,2009.

7.D.Rubben,Jr.Canals(2009) "DataMining in Healthcare :Current Applications and Issues".