International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.3, No.6, June 2014 DOI:10.15693/ijaist/2014.v3i6.76-80

Predicting Progression Of Primary Stage Cancer To Secondary Stage Using Decision Tree Algorithm

V.Mary Kiruba Rani [1], Prof.M.Safish Mary[2]
[1]M.Phil Student, Department of computer science, St.Xavier's College, palayamkottai.
[2]Assistant Professor, Department of Computer Science, St.Xavier's College, palayamkottai.
marykirubarani@gmail.com[1],marysafish@gmail.com[2]

ABSTRACT

In case of genomic research, there are multiple late mounting technologies to hoard and reclaim data. Various techniques act as bridge between experimental and real life usage. Progression of primary stage cancer to secondary stage gives a statistical report about the instance of breaking of cancer which spreads in next region. Data mining classification and decision tree algorithms are useful in this field. They are also used in finding out the relations between the cancer stages. Data mining techniques are used in cancer research to mine practical results. In this work, the biological data from cancer patients is collected and analyze the data to reach purely important conclusion. To do this analysis, decision tree algorithm is used in predicting the future reoccurrence of the disease. This method determines primary and secondary structure of cancer patients by means of classification technique. To evaluate projected technique with existing algorithm, cancer is taken as the case study and results are calculated with help of the collected database. The result of proposed technique indicates various uses of cancer prediction with better classification and predicts the duration between cancer occurrence stages.

KEYWORDS*: Decision tree algorithm, primary stage, re-occurrence of cancer, classification*

1. INTRODUCTION

Data mining is providing a vast knowledge in predicting disease, based on various symptoms. Data deposited from various experiments being conducted through cancer centers and academic institutions has contributed greatly in establishing relationship and conclude useful results [1]. Typically, a biological data set may consist of dozens of observations but with thousands or even tens of thousands of genes. Classifying cancer types using this very high ratio between number of variables and number of clarification is a delicate process. As a result, dimensionality reduction and in particular feature selection techniques may be very useful. Finding relevant characteristic among a huge quantity could gain in much specific and efficient treatment.

During the last few decades it is observed that nearly 20% of human are affected by cancer disease. In this collection 7% of the affected persons have the possibilities for recurrence of cancer tumors. Several cancer outcomes have improved during the last decade with development of more effective diagnostic techniques and improvements in treatment methodologies. A key factor in this trend is the early detection and accurate diagnosis of this disease. Hence automated medical diagnostic decision support systems have become an established component of medical technology. The main concept of the medical technology is an inductive engine that finds the decision characteristics of the diseases and can then be used to diagnose future patients with uncertain disease states. Data mining and decision tree algorithms are in great demand to be incorporated as an integral part of cancer research. Regular patterns in data provide essential results. These results help in predicting future behavior or condition of components in cancer research. Classification [2] has been used in many algorithm that commonly include decision tree learning, nearest neighbour, naive Bayesian classification and neural networks which are said to be supervised learning. Data is set into groups of pre-labelled data. Learning schemes be trained through training data and efficiency of this system is tested by using test data

In this paper, the technique uses the following cancer types for prediction. Major categories of cancer are, Leukemia, Lymphoma, Lung Cancer, CNS Cancer.

This paper is organized into following sections. Section 2 deals with the related work in the field of cancer diagnosis and prediction. Section 3 evaluates the core idea which deals with the concepts and implementation done through WEKA and other software for predicting the reoccurrence duration between primary stage and secondary stage of cancer. Section 4 discusses results on parameters like accuracy, root mean squared error. Section 5 finally discusses the conclusion and future scope for proposed technique.

2. RELATED WORK

Kohbalan Moorthy [3] proposes a technique which possibly says that microarray technology allows obtaining genetic information from cancer patients and it can be computed and classified through computational software. It is built by finding the best split of the training data at each node. Random Forest is particularly of interest in problems with a large number of independent variables. Fears of over-fitting are allayed because each decision tree is based on a subset of variables and validation is done with a running count as the model is being developed. Here a proposed technique known as gene range selection based on random forest method is used. This method allows accurate classification of tumor types. The result gives Lymphoma 0.5%, breast 0.6%, colon 0.8%, leukemia 0.9%, lung 0.9% full accuracy. This method lacks support from pathway databases in case of poor description of disease related data.

Amir Hussein kayvantoo [4] proves an application that essentially utilizes information from various databases of lung tumor disease. It considers the possibility of responsible

gene to have better expressed in tissues, suffering that particular disorder. Notable thing with this system is that the algorithm predicts lung cancer tumor types for certain years which increases when applied on datasets created by attribute weighing model. It provides best accuracy of 82% than X-Validation methods.

G. Romeo [5] uses an algorithm which considers the knowledge of another disease causing gene in that chromosomal area or genetic internals of genes related to disease. The main principle of this algorithm is to classify the types of cancer by using bagging and boosting. Top N values are chosen for classification. Adaptive boosting (AdaB) is another aggregate classification tree method that bases its predictions on the aggregate prediction of its member trees. AdaB was implemented with the ada package for R. At the heart of AdaB is the idea that a combination of weak predictors can be combined to make a strong predictor. The algorithm compares with the following models: Accuracy, Kappa, F-Measure, RMS error and prove his algorithm is more beneficial and practically significance yielding accurate results in an outcome of solving various challenges posed during cancer classification.

Bootlick [6] early study was notable for considering the percentage of sampled tissue that was cancerous by length. In the same year, Presti [7] considered the percentage (or fraction) of cancerous tissue by number of sampled regions, and considered the sum of the percentages by length of each region. These early studies created general interest in the inclusion of cancerous tissue proportion as a predictive variable, and as an extension, the count of the number of positive samples also became a popular method to gauge extent of cancer in biopsies. Others have considered maximum cancer proportion at any region, average cancer proportion across all regions, absolute length/volume of cancer, and more.

John O.Schorge, proposed utilization of the CA as an ovarian cancer serum has improved cancer detection rates during last few years. This study was conducted to determine if the commercially available classification algorithm offers better results towards cancer type classification. Biomarker Pattern Software (BPS) which is based on classification and regression tree (CART) would be effective in discriminating ovarian cancer from diseases and healthy controls [8]. Serum protein mars spectrum profiles from 139 patients with either ovarian cancer, benign diseases or healthy women were analyzed using the BPS software. A decision tree using five protein peaks resulted in an accuracy of 81%in cross validation analysis and 80% in blinded set of samples in differentiating the ovarian cancer from the control group.

The proposed system for predicting primary to secondary stage of cancer for samples is proven to be efficient upon many algorithms. This integrated framework gives details about the occurrence of cancer tumor types for number of years. Genetic mapping method has shown to be objective function to be optimized by the system. The neural network perceptron concept is introduced to select tumor types within threshold value. Root mean square measure will provide the error values between actual and predicted values, and it is found to be efficient for regression based model.

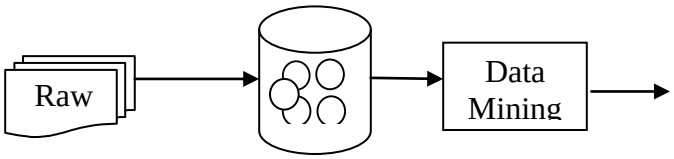
3. METHODOLOGY/TOOL

(A) CLASSIFICATION

Classification technique is implemented to train and test the data. Thus c4.5 [8] classifier is used. It falls in tree based classifiers domain and is an extension of ID3 algorithm to handle numeric classes. Attributes are discretized using suitable discretization technique. Training model is built based on information gain calculation.

C4.5 has **features** such as handling missing values, categorization of continuous attributes, pruning of decision trees, rule derivation and others. C4.5 [9] constructs a very big tree by considering all attribute values and finalizes the decision rule by pruning.

Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.



Feature Selection Classification

Classification Model

(B) DECISION TREE ALGORITHM

Decision tree Algorithm is put into operation for deciding the functionality of given dataset. It uses a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The decision tree is constructed by means of IF-THEN rule.

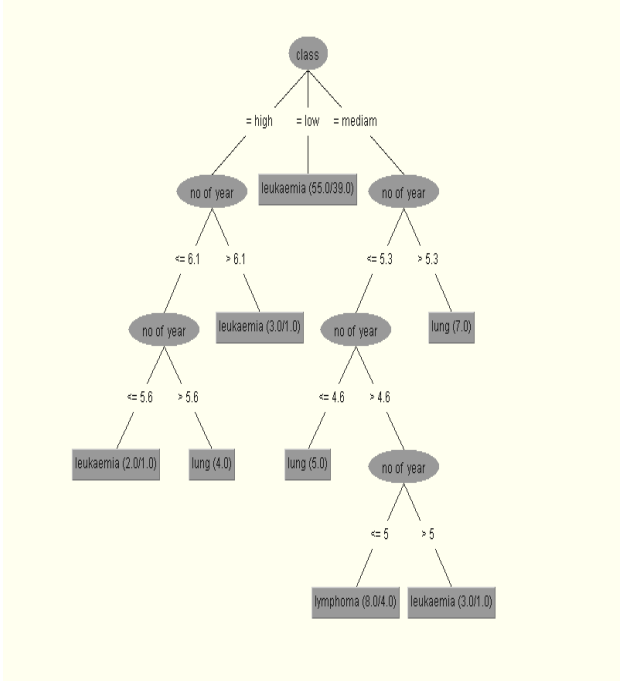
The structure of decision tree involves the following foot steps

Construction part: By using the training data set, the initial tree is constructed. Using splitting criteria it is divided into two or more sub-partitions, until stopping criteria is met.

Pruning phase: The pruning phase removes one or lower branches for best performance .Since in constructing phase may not result in best possible set of rules due to over fitting,

pruning is done. Finally processing the pruning phase is performed to improve understandability.

Tree structure is



The algorithm for decision tree learning Algorithm

Decision Tree Algorithm

Decision Tree Algorithm

Input: A data set, S

Output: A decision tree

IF the entire occurrence has the same valuefor goal attribute then return a decision tree with same value.

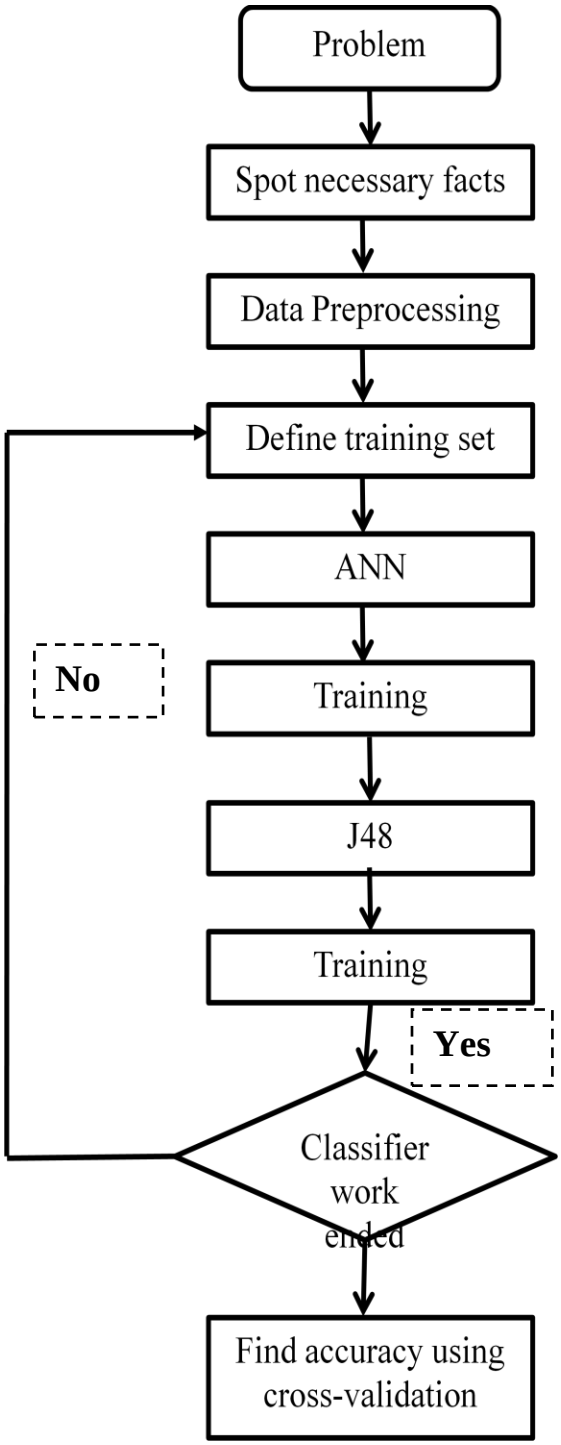
Else

For all attributes and select an attribute with the low, median and high values and create a node for that attribute.

Make a branch from this node for everv value of the attribute

TOOL:

To perform the prediction, data mining WEKA tool is used to analyze the results. The following chart representation speaks out the procedure of the research.



DATASET

The dataset are shown in the following table which has name of cancer type number of samples and number of classes.

Cancer Type	Samples	Classes
Leukaemia	98	2
Lung	180	2
Lymphoma	50	2
CNS	50	2

Table 3.1 Main characteristics of the cancer dataset used in this research

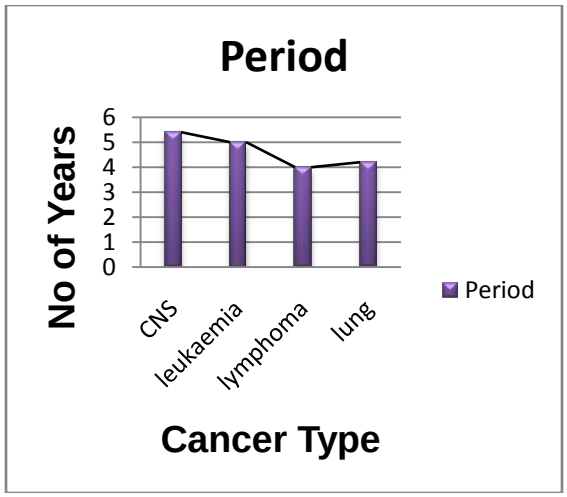
4. OUTCOME & CONCLUSION

THEORETICAL BASIS

A characteristic of the common methods of predicting cancer stages is that it treats all regions of the cancer equally. Information about the quantity and stage of cancerous tissue remains, but information on the spatial distribution of the cancer is lost when the variable depth is reduced through the common methods. In order to avoid this loss of spatial data, the original dataset for this study is coded in such a way as to contain all of the data in its rawest form, including its spatial information. Performing prediction on this raw data will produce better accuracy over all information.

Data mining and decision tree algorithms are in great demand to be incorporated as an integral part of cancer research. Regular patterns in data provide essential results. These results help in predicting future behavior or condition of components in cancer research. Many algorithms have been used for cancer data classification and the most commonly used are decision tree learning, nearest neighbour, naive Bayesian classification and neural networks which are said to be supervised learning. Data is set into groups of pre-labelled data. Learning schemes be trained through training data and efficiency of this system is tested by using test data.

By using the proposed technique the occurrence of cancer over four stages (Leukaemia, Lung, Lymphoma, CNS) comes to conclusion that within a period of minimum 4-5 years, the patient is again supposed to get affected by cancer in the same or different area. The detailed description of the cancer dataset is presented in the Table 3.1, where the number of cancer samples from patients and the main reference of the data are listed. Chart Representation of cancer stages are shown below



The complete analysis for the selected cancer was processed and the error rates also found. The data are classified by cancer names and using the attributes occurrence of cancer from primary to secondary stage is observed. The following table shows the

Cancer Type	Leukaemia	Lung	Lymphoma	CNS
Random Forest	0.9%	0.9%	0.5%	0.8%
Our Model	0.61%	0.68%	0.64%	0.58%
Occurrence of year	5years	4.2years	4 years	5.3years

From the results obtained,one can conclude that the minimum error rate is less compared to the other models used for prediction. With the selected cancer types even though the minimum error rate is 0.603%, there might be useful indication in increasing the overall accuracy.

From this analysis, we could deduce that the suitable range for occurrence of cancer from primary to secondary stage was at 4-5 years range, as most of the dataset shown better or higher accuracy in this range.

5. FUTURE SCOPE

The proposed result illustrates hopeful results thus encouraging research interests so as to improve it further. The system proposed is thought of as a framework because its components can always be replaced. Even better accuracy using more robust base classifiers can be used. Thus this method is supple to conversion and has the possibility to predict whether the recurrence of cancer stage will increase in future or not.

REFERENCES

[1] K.Raza, "Application of Data Mining in Bioinformatics", Indian Journal of Computer Science and Engineering, vol. 1, no. 2, pp.114-118, 2010.

[2] V. Krishnaiah, " Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques ", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1), 39 - 45 , 2013.

[3] Kohbalan Moorthy," Multiple Gene Sets for Cancer Classification Using Gene Range Selection Based on Random Forest, Artificial Intelligence & Bioinformatics Research Group, 142-146, 2011.

[4] Amir," extraction and integration of human disease-related information from web-based genetic databases," Nucleic Acids Research, vol. 33, no. Web-Server-Issue, pp. 758–761, 2005.

[5] S. Rossi, D. Masotti, C. Nardini, E. Bonora, G. Romeo, E. Macii, L. Benini, and S. Volinia, "TOM: a web-based integrated approach for identification of candidate disease genes," Nucleic Acids Research, vol. 34, no.Web-Server-Issue, pp. 285–292, 2006.

[6] Bootlick, "An efficient statistical model based classification algorithm for classifying cancer gene expression data with minimal gene subsets," 2009.

[7] Presti, J," Prostate biopsy: How many cores are enough?" Urologic Oncology: International Journal of Radiation Oncology Biology", 21, 135-140, 2003.

[8] Han, J. and M. Kamber,"Data Mining and knowledge discovery", Kluwer Academic Publishers-Plenum Publishers, Volume 15, Issue 1, pp 55-86,2007

[9] K. Polat and S. Güne, A novel hybrid intelligent method based on C4. 5 decision tree classifier and oneagainst- all approach for multi-class classification problems, Expert Systems with Applications, vol. 36, pp. 1587-1592, 2009.

Authors Profile

V.Mary Kiruba Rani received the **MCA** degree from the St.Xavier's College, Palayamkottai, Tirunelveli, India, in 2013.Currently doing **M.Phil.** in Computer science, from St.Xavier's College, Palayamkottai,Tirunelveli, India. Her research interest includes Data mining, fuzzy logic, Communication networks

M. Safish Mary received her MPhil degree in Computer Science from Manonmaniam Sundaranar University, Tirunelveli, India. She is presently working as Assistant professor at St. Xavier's College (Autonomous), Palayamkottai, India. She has a vast teaching experience of about 16 years and research experience of about 8 years. Her research interests include neural network, data mining and soft computing..