Applications of Data Mining and Medical Imaging for Disease Diagnosis

Ramandeep Singh
Associate Professor
School of Computer Science and Engineering
Lovely Professional University, India

Abstract- -Primary tumor alludes to a tumor or mass that is developing in the area where malignant growth started. For example, if a patient is determined to have stomach malignancy, the primary tumor would be found in the stomach itself instead of somewhere else in the body. Data mining has wide reception as of late in numerous ventures, to a great extent as a result of the capacity of mining methods to quickly yield answers to business inquiries in a brief timeframe and the accessibility of huge amounts of information to misuse. The study uses primary tumor imbalanced dataset as input for several predictive data mining methods. The objective of the current study is to determine the ability of these data mining methods to predict the location of primary tumor. Several variables are known to be predictive of primary tumor. These are based on the patient's risk variable location and associated basic parameters. This study analyses the patient's risk factor variables from the primary tumor dataset to monitor output to develop accurate and comprehensible models for the prediction of Primary tumor. This study aims to find predictors of various primary tumors using primary tumor dataset.

Keywords—
Cloud Computing, Data Deduplication, Encryption, Security.

1 INTRODUCTION

A primary tumor is known as tumor or mass that is developing in the area where malignant growth started. The organs and tissues of the body are comprised of small structure squares known as cells. Malignant growth is an ailment of these cells. Disease can once in a while spread from where it initially began to develop (essential malignant growth) to shape tumors in different body parts (optional tumors). For example, if a patient is determined to have stomach malignant growth, the essential tumor would be found in the stomach itself instead of somewhere else in the body. The essential tumor is commonly the most straightforward to evacuate. It is imperative to discover it else it might develop and may emerge some other connected tumors. Now and again malignant growth cells split away from the first (essential) disease. It may spread to parts of body through the circulatory system or lymphatic framework[1][2]. The lymphatic framework is a piece of the safe framework . Lymph hubs (organs) are a piece of this framework. These are present all through the body and are associated together by a system of modest cylinders (conduits) that convey a liquid called lymph. The essential malignancy is obscure (Erin J. Slope, 2013). This project has been undertaken to assess features of a large automatically collected database of Primary tumor and to determine whether data mining analyses is able to find the primary tumors from attributes and instances of dataset. The predicted rules that are found capable in identifying the increased risk of secondary cancers from primary tumors are effective measures and therapies that can be conducted in the initial stage.

Paper organization

The rest of study is organized as follows Section 2, addresses review of similar work, Section 3 addressres research methodology. Experimental setup and its results are discussed in Section 4. Section 5, concludes the paper.

2 Review of Related Work

The wide range of study has been done about utilizing AI strategies for survivability examination and expectation investigation. Several papers are studied in context of data mining commitment in clinical domain.

Data mining methods, for example, grouping, arrangement, relapse, affiliation rule mining, CART (Classification and Regression Tree) are broadly utilized in social insurance area (Mohammed Abdul Khaleel, 2013).The principle focal point of this paper is to investigate information digging strategies required for clinical information mining particularly to find locally visit maladies, for example, heart illnesses, lung malignant growth, and bosom disease, etc. N.Sudha Bhuvaneswari (2013) examined that Cancer is one of the horrible infections on the planet guaranteeing larger part of lives.(K.Lokanayaki, 2013) talks about assortment of information mining systems, approaches and various explores which are continuous and accommodating to clinical conclusion of disease.P.Gayathri (2012) inferred that The Health care industry uses information mining Techniques and discovers the data which is covered up in the informational index. Rajneet Kaur (2013) reasoned that location of lung disease in its beginning times. Right now the utilization a few methods are basic to the assignment of clinical picture mining, Lung Field Segmentation, Data Processing, Feature Extraction, Classification utilizing neural system and SVMs.

3 Research Methodology

The study uses primary tumor imbalanced dataset as input for several predictive data mining methods. The main goal of this paper is to evaluate the ability of these data mining methods to predict the location of primary tumor. Several variables are known to be predictive of primary tumor. These are based on the patient's risk variable location and associated basic parameters. This study analyses the patient's risk factor variables from the primary tumor dataset to monitor output to develop accurate and comprehensible models for the prediction of Primary tumor. Other goals of this study include the assessment of feature selection method and modifying parameters of the models for their ability to reduce the complexity and potentially increase the accuracy of predictive models developed for this dataset[3]. An attempt is also made to determine which predictive methods are most suited to this application. Since Vote classifier combining KNN and Random forest tend to be more comprehensible and more accurate. This study follows the CRISP-DM methodology of data mining. The main steps of the methodology as they are used in this study are briefly described below.

3.1 Data Acquisition

The data regarding tumor domain is acquired from the University Medical Centre, Institute of Oncology, Yugoslavia. The common primary tumor that is predicted from among 339 instances including men and women.

Number of Instances: 339

3.2 Data Selection

The next task in the research methodology is the selection of target and input variables according to the data mining methodology (CRISP-DM).

3.2.1 Target Variable Selection

Target variable for this dataset was selected after analyzing the dataset. As it is predefined in the primary tumor.arff dataset. Total target/class variables in the dataset is 22.

3.2.2 Variable Selection

Selection of variables is covered in more detail in Chapter 4, Data Exploration. Suitable variables must have clinical significance be available for most. The variables chosen on the basis of cancer predicting variables that contributes a lot. Variable like histologic type having labels epidermoid, adeno, anaplastic

3.2.3 Pre-processing tasks

The data pre-processing tasks addressed in the remainder includes Imputation of missing values , removal of useless attributes, apply resampling technique without replacement and sample size percent of 1000.0 and data reduction using feature selection methods or filter methods (Cfs, Wrapper, Consistency, Classifier and ReliefF)[4][5].

3.3 Data modelling

Data modelling was conducted on the datasets described above with the predictive accuracy of models based on the dataset[6]. Several preprocessing methods and modelling algorithms are used and model performance is evaluated using several measures,

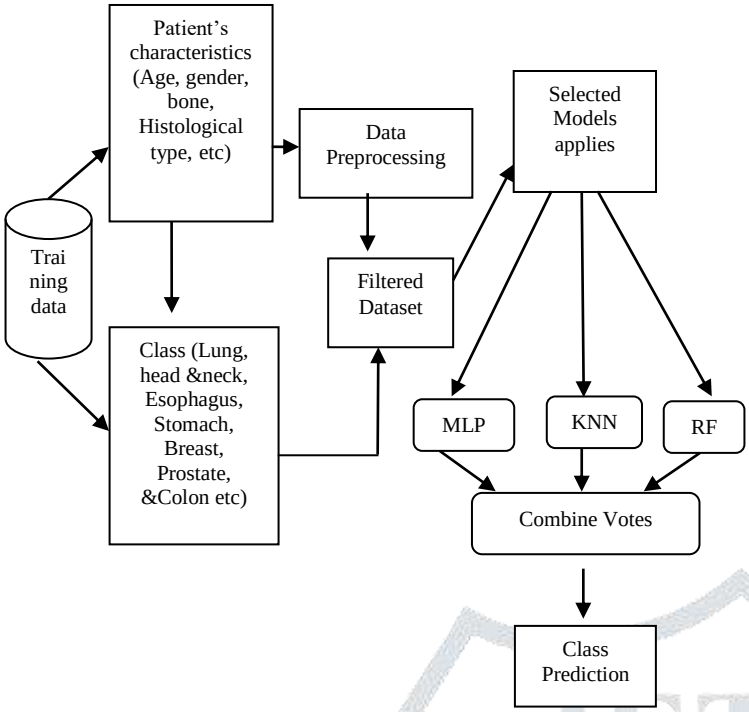

Figure 1: Research Methodology

4. Experimental Setup and Results

The Primary tumor dataset is mined for predicting the primary tumors in the human body. It is on the basis of selected standard dataset Primary–tumor.arff on which the methods to be used are chosen from each of the main categories of predictive data mining algorithms (instance-based, tree-based, rule-based, neural network-based and probability based)[7][8]. The methods used are:

Model	Accuracy	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Multilayer Perceptron without Resampling and non-uniform class distribution	41.00 %	0.41	0.063	0.408	0.407	0.775
Multilayer Perceptron with Resampling and uniform class distribution .	83.45 %	0.835	0.008	0.835	0.831	0.954

Table 1 Description of Performance Parameters for Simulation

- Naïve Bayes (NB)
- Multilayer perceptron (MP)
- Nearest Neighbour (KNN)
- Rule Based (PART)
- Decision tree analysis (J48, Random Forest)

Experiment was conducted for this study and two scenarios were considered, one containing the total 339 instances with imbalanced class distribution, second containing the 3390 instances with Random oversampling having uniform class distribution along with attribute selector[4].

The trials was directed with 10-Fold Cross Validation that was embraced for haphazardly inspecting the preparation and test beds. While playing out the examinations the parameter value is fixed default for every calculation with the exception of Multilayer Perceptron classifier where the parameter "preparing time" which had default esteem "500" was changed to "100" and "Learning rate" from "0.3" to "0.1" and "energy" from "0.2" to "0.4" for quicker the exhibition of classifier in the second model of the main trial. The absolute first model of each analysis has considered default settings of classifiers. The exhibitions of the models right now assessed utilizing the standard measurements of exactness, accuracy, review and F-measure which were determined utilizing the prescient characterization table, known as Confusion Matrix. ROC territory was likewise used to analyze the exhibitions of the classifiers.

Experiment : Model Building Using Multilayer Perceptron classifier

The first simulation was devised to determine the performance of Neural Network using Back Propagation that is Multilayer Perceptron classifier in predicting Primary tumors and to investigate the effect of random oversampling. In this experiment two scenarios were considered.

First case: The algorithm was run on dataset containing 339 instances with 18 attributes with 10-folds cross validation.

Second case:

1. The algorithm was run on a 3390 instances with attribute selector and Random oversampling with uniform class distribution with parameter "biastouniformclass" changes from "0.0" to "1.0" and "samplesizepercent" from "100.0" to"1000.0".

2. Classifier conatins Parameter "training time" reduces to "100" from "500" and "Learning Rate" from "0.3" to "0.1" and momentum from "0.2" to "0.4" for faster the execution. The model built with multilayer perceptron with all attributes accurately characterized (anticipated the right result) 139 (41.0%) cases while 200(58.91%) occasions of the 339 cases were arranged inaccurately with 10-crease cross approval. The general precision pace of the model isn't effective, yet it is essential to consider the TP Rate

(Sensitivity), and TN Rate (Specificity), to see the presentation of the model for each class. This outcome gave the model a TP Rate of 0.41. With respect to Precision score of the model, exactness of 40.08% it is a most noticeably awful model in recovering applicable qualities for every class. With F-Measure estimation of 0.407, it very well may be inferred that the Precision and the Recall of the model are essentially adjusted.

5 CONCLUSION AND FUTURE PERSPECTIVES

In this study, point was to structure a prescient model for primary tumor identification utilizing information mining procedures from Primary tumor Report dataset that is fit for upgrading the unwavering quality of threatening tumors conclusion utilizing recognition of its essential birthplace. Information gathered by International Cardiovascular Hospital from the year 2008 to 2011 containing 339 examples was chosen and preprocessed for this examination. For futuristic purpose, the main goal is to play out extra examinations with more instances and calculations to enhance the characterization precision and to construct a framework that can anticipate explicit essential Tumors types.

References

[1] Dave Smith, "Data Mining in the Clinical Research Environment", Source: http://www.sas.com, UK, 2007.

[2] Y. Ramamohan et. Al "A Study of Data Mining Tools in Knowledge Discovery Process", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[3] Chandrika Kamath, "The Scientific Data Mining Process", Chapter 4.

[4] Nevine M. Labib et. al, "World Academy of Science, Engineering and Technology" 2007.

[5] Source: From website " http://www.zentut.com/data-mining/data-mining-applications/," Copyright © 2013 by ZenTut Website.

[6] Mohammed Abdul Khaleel et.al , "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013, ISSN: 2277 128X.

[7] N.Sudha Bhuvaneswari et.al , "Information extraction of predicting blood cancer", IJCS International Journal of Computer Science, Volume 1, Issue 4, September 2013.

[8] K.Lokanayaki et.al, "Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis", International Journal of Computer Applications (0975 – 8887) Volume 77 – No.5, September 2013.