Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining

Mehak Naib
Mtech(Software system) Student
Department of Computer Science & Engineering
Guru Nanak Dev University, Amritsar

Amit Chhabra
Assistant Professor
Department of Computer Science & Engineering
Guru Nanak Dev University, Amritsar

ABSTRACT
Data mining has been widely adopted in recent years in many fields, especially in the medical field. This paper highlights the prediction of unknown primary tumors in the dataset. The multiclass classifier with Random forest is used for classification of multiclass dataset as it gives much higher accuracy than binary classifiers. SMOTE method for this imbalanced dataset with Randomize technique is applied during preprocessing for reducing the biasness among classes. These all evaluations and results are carried out using WEKA 3.6.10 as a data mining tool.

General Terms
Data mining, Tumors, Classifiers

Keywords
SMOTE, WEKA, Primary tumor, Multiclass classifier, Random forest

1. INTRODUCTION
Data mining plays an important role in the medical field by predicting various diseases [1]. This paper deals with one of the major health problems to which each country is dealing with, it is the tumor or cancer. The organs and tissues of the body are made up of tiny building blocks called cells. Cancer is a disease of these cells. Cancer can sometimes spread from where it first started to grow (primary cancer) to form cancers in other parts of the body (secondary cancers). The primary tumor is generally the easiest to remove. It is very important to find it as it may grow and may arise some other linked tumors called secondary tumors. When a secondary cancer is diagnosed, but even after tests have been carried out, doctors can't tell where the cancer first started. The primary cancer is unknown [2] [9]. Data mining has tried to explore the primary tumor dataset to uncover previously unknown patterns.
The paper is organized as follows. In Section 2, problem statement is given. In Section 3, discusses some recent review of similar work in the data mining field. Section 4 discusses experiments and results with which the primary-tumor dataset is mined. Section 5 concludes the paper.

2. PROBLEM STATEMENT
The problem is to predict the primary tumor classes from large amount of data without any biasness among classes. This has been done by doing a comparative study of classification algorithm Random forest as a base classifier in multiclass classifier approach along with oversampling technique SMOTE and Random forest as a binary classifier on various parameters using primary tumor dataset containing 18 attributes and 339 instances.

3. RELATED WORK
Several studies have been reported on applying machine learning techniques for survivability analysis and predictive analysis. In this section, numbers of papers are studied in related to data mining contribution in medical field.
Data mining techniques such as clustering, classification, regression, association rule mining, CART (Classification and Regression Tree) are widely used in healthcare domain [5]. The main focus of this paper [5] is to analyze data mining techniques required for medical data mining, especially to discover locally frequent diseases such as heart ailments, lung cancer, and breast cancer and so on. It also throws light into the importance of locally frequent patterns and the mining techniques used for the purpose.
Authors have studied that Cancer is one of the dreadful diseases in the world claiming majority of lives. There is a study of blood cancer and its symptoms and staging. Then Linear regression algorithm is applied and 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.[6]
Author discussed data mining techniques, approaches and many different researches which are ongoing and helpful to medical diagnosis of disease. The study revealed that depending on the type of dataset used each model differs in their performance. If the dataset consists of unlabelled classes or features, then the clustering model better suits for pattern recognition among the several algorithms, k-means algorithm adopted by researches due to its simplicity. [8]
Data mining plays an important role in predicting diseases in the health care industry [12]. Authors reviewed the research papers which mainly concentrated on predicting heart disease, Diabetes and Breast cancer. There was study of NaïveBayes, K-NN, and Decision List algorithm for analyzing the heart disease dataset. Tanagra data mining tool is used for classifying the dataset. The classified dataset is evaluated using 10 fold crossvalidation and the results are compared. This observation is performed using training data 3000 instances with 14 different attributes. The dataset is divided into two testing and training i.e. 70% of data is used for training and 30 % is used for testing. The study to investigate and compare seven different classification algorithms, namely, Naive Bayes, Naive Bayes updatable, FT Tree, KStar, J48, LMT, and Neural network for analyzing Hepatitis prognostic data has been presented [3]. The study concludes that the Naive Bayes classification performance is better than other classification techniques for hepatitis dataset.
A prototype model for the breast cancer as well as heart disease prediction using data mining techniques is discussed in [2]. The data used is the Public-Use Data available on the

web, consisting of 909 records for heart disease and 699 for breast cancer. Two decision tree algorithms C4.5 and the C5.0 have been used on these datasets for the prediction and performance of both algorithms is compared.

Authors studied the application of data mining processes, particularly classification, in analyzing the students' data by studying the main attributes that may affect the student performance in courses. For this purpose, the CRISP framework for data mining is used for mining student related academic data. The classification rule generation process is based on the decision tree as a classification method where the generated rules are studied and evaluated [10].

Authors concluded the hybrid approach of CART classifier with feature selection and bagging approach of analyzing various breast cancer datasets. Experiments are conducted in WEKA and results are compared with & without 10-folds cross validation [7].

4. PRIMARY-TUMOR PREDICTION: RESULTS AND ANALYSIS

There exist many research papers include mining of the datasets that applies different algorithms with different statistics. Hence, with the advent of improved and modified prediction techniques, there is a need for an analyst to know which algorithm performs best for a particular type of dataset. Following steps are used in the Experiment and analysis:

I. Dataset Collected

Primary-tumor.arff dataset is selected for this work. Dataset contains 18 attributes, one class attribute and 339 instances. It contains total 22 classes of primary tumor. The rest of the attributes indicate the areas from where primary tumors start. The dataset has been collected from the University Medical Centre, Institute of Oncology, Ljuljana, Yugoslavia. [4]

Attribute Information:
@attribute age {<30, 30-59,>=60}
@attribute sex {male, female}
@attribute histological-type {epidermoid, adeno, anaplastic}
@attribute degree-of-differ {well, fairly, poorly}
@attribute bone {yes, no}
@attribute bone-marrow {yes, no}
@attribute lung {yes, no}
@attribute pleura {yes, no}
@attribute liver {yes, no}
@attribute peritoneum {yes, no}
@attribute brain {yes, no}
@attribute skin {yes, no}
@attribute neck {yes, no}
@attribute supraclavicular {yes, no}
@attribute axillar {yes, no}
@attribute mediastinum {yes, no}
@attribute abdominal {yes, no}
@attribute class {lung, 'head and neck', esophagus, thyroid, stomach,' duoden and sm.int', colon, rectum, anus, 'salivary glands', pancreas, gallbladder, liver, kidney, bladder, testis, prostate, ovary, 'corpus uteri', 'cervix uteri', vagina, breast} (class is location of tumor).

II. Data preprocessing

1. Remove useless variables using unsupervised filter.
2. Remove Missing values using unsupervised *filter.*
3. Due to multiclass dataset, it has non-uniform class distribution among instances; hence Synthetic Minority Oversampling Technique (SMOTE) algorithm of supervised technique (Figure 1) for resampling the imbalanced dataset with Randomize algorithm is applied to remove biasness towards majority classes like lung, gallbladder.etc in predictions during testing. It has been applied over 7 times to increase instances from 339 to 354 then randomize the whole instances. The instances of minority classes are added in this loop of 7 times applying SMOTE technique, example of those classes are anus, duoden and sm.int, testis, ovary.etc.

Table 1: Variables group in the dataset

Variable group	Variable type	Variable examples of attributes
Multivariate Data	Primary tumor	Histologic-type, Degree of Difference.
Demographic	Patient Related	Age, Gender, lung, axillar.
Class Variable	Location of Primary tumor	Lung, Head and neck, Thyroid, stomach, Pancreas etc.

III. Perform Classification

Various classifiers are applied to the dataset using WEKA. The results show that if there are at least two classes, then multiclass classifier with Random forest algorithm can provide better results with accuracy 85.7% than binary classifiers such as Random Forest with accuracy 84.5%.

As the Random Forest algorithm works best among all classification algorithms, so this algorithm has been chosen for applying with multiclass classifier to improve accuracy and precision. The comparison among all algorithms is shown in table 2 which shows that for the multiclass datasets multiclass classifier with random forest containing 10 random trees works best among all algorithms. As shown in the table, Accuracy of 85.7% with ROC area is 0.997 has been achieved by multiclass classifier with random forest. Figure 3 shows the comparison of algorithms based on Kappa statistics.

Graphically figure 2 shows the performance of Random forest with multiclass classifier achieves a high accuracy rate of 85.7 % among all classifiers. Figure 4 shows ROC area of classifiers where multiclass classifier achieves high ROC area.

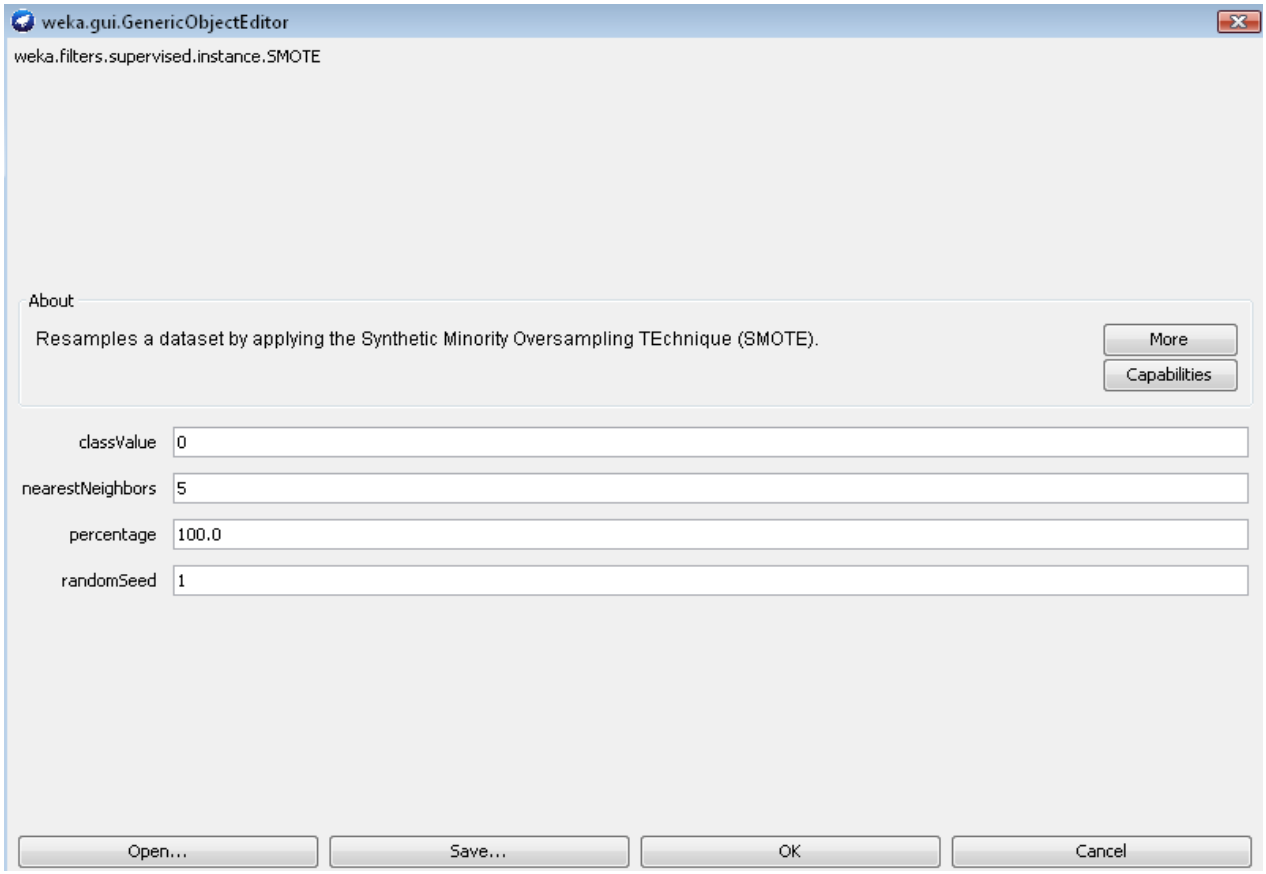

Figure 1: SMOTE technique in WEKA

Table 2: Comparison of Classification Algorithms

Algorithm	Accuracy (%)	F-Measure	Kappa Statistics	ROC Area	Precision
Naïve bayes	54.23	0.512	0.489	0.892	0.537
Logistic	65.50	0.637	0.616	0.944	0.645
Multilayer perception	68.90	0.672	0.653	0.949	0.70
Bagging	62.42	0.597	0.573	0.954	0.64
PART	65.53	0.644	0.615	0.946	0.652
J48	64.40	0.619	0.597	0.940	0.622
Random Forest	84.20	0.846	0.828	0.992	0.853
Multiclass Classifier with Random forest	85.70	0.856	0.840	0.997	0.883

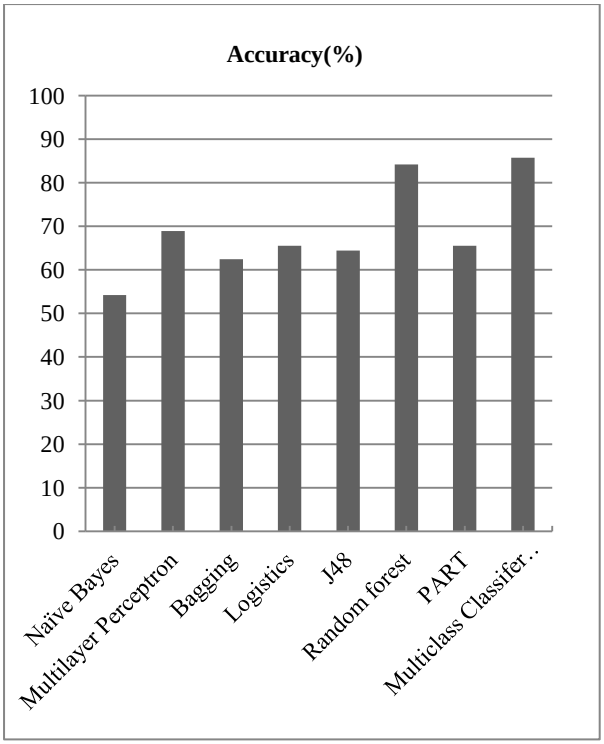

Figure 2: Graph showing Accuracy of classifiers

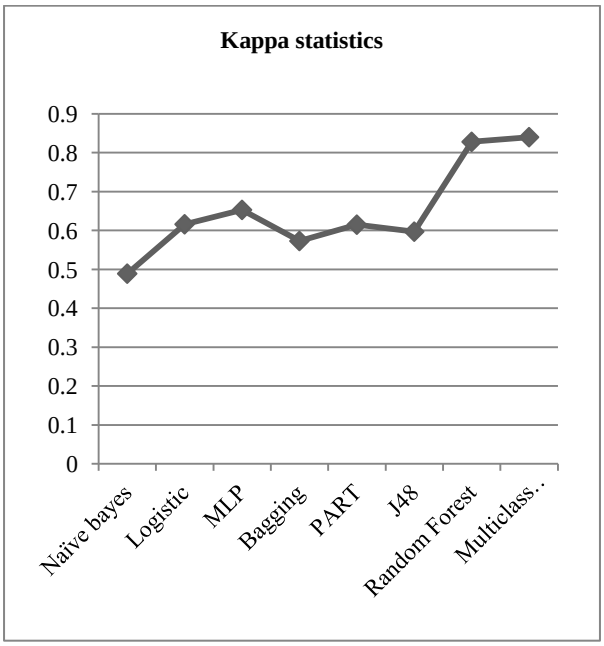

Figure 3: Graph showing Kappa Statistics of classifiers

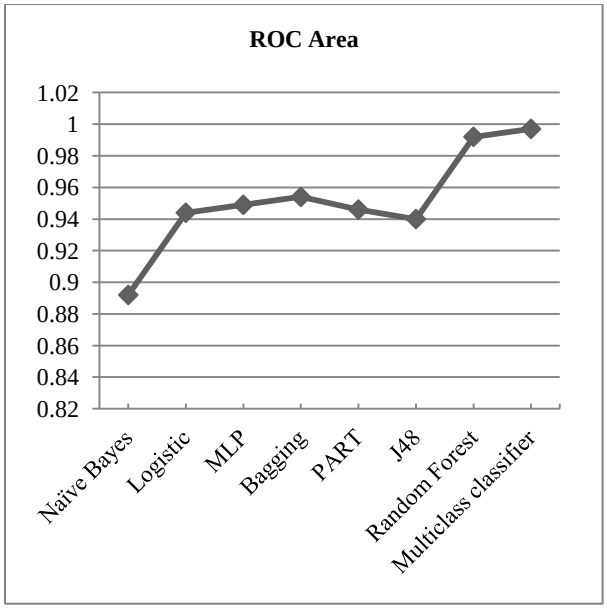

Figure 4: Graph showing ROC Area of classifiers

IV. Perform testing

Testing is performed on the selected model in step 2 on the test dataset by taking some number of instances to predict the primary tumor of each instance. Figure 5 shows the predictions of primary tumor classes.

Classifier output

inst#,	actual,	predicted,
1	?	6:duoden a
2	?	1:lung
3	?	22:breast
4	?	18:ovary
5	?	18:ovary
6	?	11:pancrea
7	?	14:kidney
8	?	1:lung
9	?	17:prostat
10	?	18:ovary
11	?	1:lung
12	?	2:head and
13	?	4:thyroid
14	?	11:pancrea
15	?	1:lung

Figure 5: Output predictions on the Test dataset in WEKA data mining tool

5. CONCLUSION

In this study the problem of predicting the primary tumor is discussed. The main focus is on using different algorithms and a combination of several targets attribute for primary tumor prediction using data mining techniques. Depending upon the dataset particular data mining algorithm is selected and mostly it is found that with at most two classes, binary classifiers are accurate whereas with increasing number of class multiclass classifier with binary classifier like Random forest is more accurate. It is also found that with SMOTE technique improves the results of the selected classifier with better accuracy. In future the work can be expanded and enhanced for the automation of primary tumor prediction.

6. ACKNOWLEDGEMENTS

This research paper is made possible through the help and support from everyone, including: parents, experts, family, friends, they were a great source of support and encouragement; we thank them all and wish them all the best in their lives.

6. REFERENCES

[1] Ba-Alwi FM. and Hintaya M. 2013. "Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach", International Journal of Scientific & Engineering Research, vol. 4, Issue 8, ISSN 2229-5518.

[2] Bhuvaneswari N. and Yamuna S. 2013. "Information extraction of predicting blood cancer", International Journal of Computer Science, Vol. 1, Issue 4.

[3] Damtew A. 2011. "Designing a predictive model for heart disease detection using data mining techniques", A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics.

[4] Dataset collected, [http://tunedit.org/repo/UCI] accessed 2013.

[5] Khaleel M., Pradham S. and Dash G.N. 2013. "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, Vol. 3, Issue 8, ISSN: 2277 128X.

[6] Khan M., Qamar S. and Massin L. 2012. " A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining", International Journal of Applied Engineering Research, Vol.7, No.11, ISSN 0973-4562.

[7] Lavanya D. and Rani K. 2012. "Ensemble decision tree classifier for breast cancer data", International Journal of Information Technology Convergence and Services (IJITCS), Vol.2, No.1.

[8] Lokanayaki K. and Malathi A. 2013., "Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis", International Journal of Computer Applications, Vol. 77, No.5.

[9] Primary Tumor, [http://www.wisegeek.com] accessed 23 march 2014.

[10] Radaideh Q., Shawakfa E. and Najjar M. 2006, "Mining Student Data Using Decision Trees" The International Arab Journal of Information Technology.

[11] Ramamohan Y., Vasantharao K., Chakravarti C.and Ratnam A.S.K. 2012. "A Study of Data Mining Tools in Knowledge Discovery Process", International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue-3, ISSN: 2231-2307.

[12] Vijiyarani S. and Sudha S. 2013. "Disease Prediction in Data Mining Technique – A Survey", International Journal of Computer Applications & Information Technology, Vol. II, Issue I, ISSN: 2278-7720.