Survival Prediction of Colon Cancer using Ensemble Data Mining Based on SMOTE

Karthikaa. R
Department of Computer Science, Sri Krishna College of arts and science, Coimbatore, Tamil Nadu, India.

Sandhya. S
Assistant professor, Sri Krishna college of arts and science, Coimbatore, Tamil Nadu, India.

Abstract – The Colon Cancer is an aggressive and well known disease that affects people all around the world. I am going to analyze the colon cancer data with the aim of developing accurate survival prediction models for colon cancer. The preprocessing steps such as data cleaning, integration, transformation and reduction results in removal of several attributes from the dataset. I will also adopt synthetic minority over-sampling technique (SMOTE) to balance the survival and non-survival classes. The feature selection is done using Correlation Feature Selection (CFS) and Information Gain Ratio. Various several supervised classification methods will be applied which includes basic classifiers and meta classifiers. In my experiments, ensemble voting algorithm is used to find the top three performing classifiers to result in the best prediction performance in terms of prediction accuracy and area under the ROC curve. We evaluate multiple classification schemes to estimate the risk of mortality after 1 year, 2 years and 5 years of diagnosis, on a subset of attributes which will be acquired after the data cleanup process, attribute selection techniques, and SMOTE balanced set of attributes. Moreover, I will demonstrate the importance of balancing the classes of the data set to yield better results.

Index Terms – SMOTE Algorithm, Colon Cancer, Ensemble, Machine Learning.

1. INTRODUCTION

Colon and rectum cancers rank among the top cancer types worldwide. Early diagnosis, increase the chances of survival and eliminating the disease by early treatment. Colorectal cancer (CRC) is cancer that develops in either the colon or the rectum. It is a major cause of condition of being diseased and mortality throughout the world. CRC accounts for over 9% of all cancers worldwide and affects about 5% of the U.S. Up to 150,000 new cases per year. According to the American cancer society, an estimated 136,830 Americans were diagnosed with colorectal cancer, including 71,830 males and 65,000 females per year. In contrast, the incidence of CRC in Thailand is 13.67% of all cancers for men and 7.40% for women, with over 8,000 new cases per year.

Although firm scientific evidence for the prevention of CRC is available, researchers continue to look for the causes of CRC as well as ways to prevent and cure the disease. In present medical studies, identifying risk factors for CRC and prediction models are normally based on multivariate statistical analysis. Big data in the healthcare system, however, contains hidden knowledge, which is impossible to discover by using conventional approaches. Data mining, therefore, is more appropriate for medical studies.

To predict survival of colon cancer patients we use supervised classification methods, at the end of 1 year, 2 years and 5 years of diagnosis. The classification schemes in our proposed system consist of 5 basic classifiers and 3 meta classifiers. Compared to basic classifiers we carried experiments with several classifiers to find that many meta classifiers used with decision trees and functions can give better results. These results can be improved by adopting SMOTE (Synthetic Minority Oversampling Technique) to balance the survival and non-survival classes, and by combining the resulting prediction probabilities from several classifiers using an ensemble-voting scheme.

Predicting of tumor cells into malignant Classification is done, classifying the secondary structure of proteins into alpha-helixes, beta-sheets or random coils, Categorizing news stories as finance, weather, entertainment, sports, etc. Association rule mining is used to solve the problem of how to search efficiently for those dependencies. Single & Multidimensional Association Rules are used to solve the problem.

2. REVIEW OF LITERATURE

The increase in availability of electronic medical records leads to interest in mining medical data. Data mining research has been published on private hospital data and publicly available data such as American College of Surgeons National Surgical Quality Improvement Program. Data mining applications have been developed on difference types of cancer. Reda Al-Bahrani, Ankit Agrawal, and Alok Choudhary explored survivability prediction of colon cancer patients using Neural Networks [1]. This model is experimented with multiple neural network structures and found that a network with 5 hidden layers produces best results. A meta-transformer with a base algorithm of extra trees is used to improve accuracy and avoid over-fitting. NN models are built using the training and validation sets and results are reported after running the testing

set against the learned models. Survival of lung cancer on from Surveillance, Epidemiology, and End Results (SEER) program data has been studied by Chen ET al.18. Agrawal ET [2]analyzes SEER lung cancer patients and provides an outcome calculator using ensemble voting techniques for survival and conditional survival20.

We studied that there is clear imbalance in the two classes of colon survival across most of the years. The shorter periods have higher imbalance than others.

Survival	Survived	Not Survived
1 year	135,372	42,114
2 years	125,689	45,242
5 years	76,124	72,410

Studies were also conducted on treating colon cancer survivability prediction as a Classification Problem by Ana Silva, Tiago Oliveira, Jose Neves, and Paulo Novais [3]. With another one with 18 features indicated by a physician this model was compared. Using 18 features the results show that the performance of the six-feature model is close to that of the model using , which indicates that the first may be a good compromise between usability and performance. The Rapid Miner software was chosen to develop the prediction model. The feature selection was determined using the Optimize Selection operator of Rapid Miner. Breast Cancer Data Classification Using Neural Network Approach of MLP Algorithm by A.Kathija, S. Shajun Nisha and Dr .M .Mohamed Sathik M.Phil. [4]. This model uses ZeroR classifier as baseline classifier which improves performance measurement. Breast or lung cancers covered as much as Data mining applications and studies of colorectal cancer.

3. METHODOLOGY

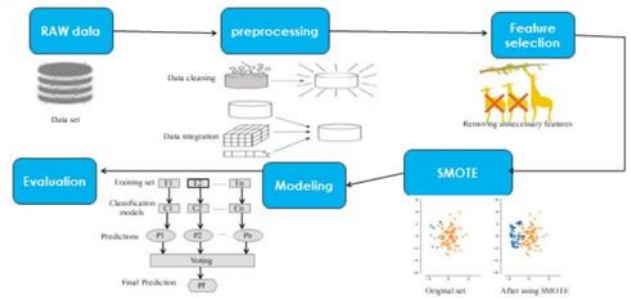

Fig.1: Prediction System Flow

This section provides description of the data set, data preprocessing, predictive modeling and performance evaluation. Understand and clean the data is the most important step. Classification algorithms had been proposed by many researchers in the field of classification of application and investigated in breast cancer data using decision tree algorithms. To predict classification of breast cancer data they used algorithms and find the most suitable one for predicting cancer.

3.1. Dataset

The new data enabled us to experiment with a newer set of features, and get better predictive accuracy using ensemble methods. If a patient survived after the cutoff date, but passed away afterwards their status at the cutoff date is the one reported i.e. the patient is reported alive. If a patient was diagnosed past the cutoff date their record is excluded from the data release. In different periods We analyze in our study; as a result each period has a different end date.

X	X1	X2	X3
1	1	0	0
2	0	1	0
3	0	0	1
1	1	0	0

Fig.2 : Categorical feature transformation

Many features in the data set are categorical features such as sex, birth place, and stage.

Some features are numerical such as tumor size, number of nodes, etc. Categorical features were transformed using a one-hot scheme to overcome this. Each categorical feature was transformed to integers and mapped to sparse matrices where every column corresponds to a category of a feature (see Table. 2). Numerical features were normalized to improve performance of estimators. To make such features look more like normal distributions standard normalization was applied to numerical features to make such features look more like normal distributions.

$$z = \frac{x-\mu}{\sigma}$$

3.2. Data Preprocessing

Data pre-processing refers to the tasks needed to convert the raw data into input data and is an important step in data mining. It is common that outliers exist in real world datasets. Outlier values might arise from fraudulent behavior, human error, instrument error or simply through natural deviations in populations.

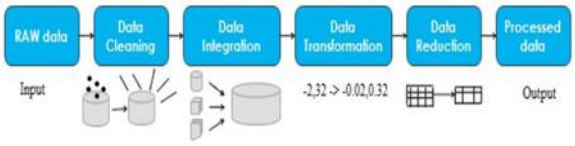

Fig.3: Data Preprocessing

Thus the preprocessing includes data cleaning, data integration, data transformation and data reduction. Also, in this stage to format the raw data to appropriate values the following conversions/calculations were performed on the datasets.

- Convert evidently numeric attributes to nominal, e.g., marital status, sex
- Convert Size of Tumor to cm. E.g. code 100 is equivalent to 10.0 cm
- Calculate the survival time in months (numeric) of YYMM format
- Extract data records for the period of interest
- Extract data records that are related to the cancer in study
- Extract all attributes that may indicate that vital status of the patient.

Finally, the attributes are removed that do not vary at all or that vary too much. attributes that exceed a maximum variance threshold Constant attributes are removed e.g. 99%.

3.3. Predictive Modelling

To construct predictive models for cancer-specific survival, on the preprocessed data supervised classification methods are employed. The two straightforward steps of this stage are:

a) Separate the data into training and testing sets or use cross-validation

b) Conducting experiments using the different classification schemes

3.4. Evaluation

In this stage from the predictive modeling stage the models were compared/ with respect to different metrics. These metrics include:

Positive Predictive Value: also known as precision is the ratio of true positives to both false positives and true positives combined, and is calculated as follows

$$\text{Precision}= \frac{FP}{TP + TP}$$

Negative Predictive Value: is the ratio of true negative to both false negative and true negative combined, and is calculated as follows

$$\text{NPV}= \frac{TN}{TN + FN}$$

Sensitivity: is the part of positive labeled examples in the dataset that are classified as positive.

$$\text{Sensitivity}= \frac{TP}{TN + FN}$$

Specificity: is the part of negative labeled examples in the dataset that are classified as negative.

$$\text{Specificity}= \frac{TN}{TN + FP}$$

Area under ROC Curve: is a calculation of the area under a curve after plotting the false positive rate versus the true positive rate

$$\text{AROC} = Z \int_0^1 TP/P \, d\, FP/N = 1/PN \int_0^1 TP \, dFP$$

4. FEATURE SELECTION

Feature selection is the process of removing unnecessary features.

. This includes,

- Simplification of models are used to make them easier to interpret
- Less training times.
- To avoid the curse of dimensionality.
- Enhanced generalization by reducing over fitting.

4.1. Correlation Feature Selection

The CFS measure is used to remove the duplicate attributes in the dataset.

The following equation provides the advantage of a feature subset S consisting of k features.

$$\text{Merit}_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}.$$

The CFS criterion is defined as

$$\text{CFS} = \max_{S_k} \left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1 f_2} + \cdots + r_{f_i f_j} + \cdots + r_{f_k f_1})}} \right]$$

5. SMOTE

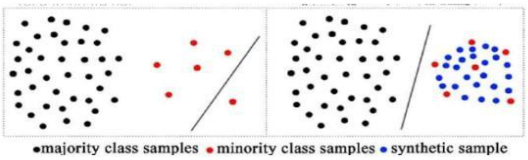

Fig.3: Smote Algorithm

SMOTE (Synthetic minority oversampling technique) algorithm overcomes imbalances by generating artificial data. It is also a type of oversampling technique. The SMOTE algorithm generates artificial examples by oversampling the minority class and introducing new artificial patient records. In

regards to synthetic data generation, SMOTE (Synthetic minority oversampling technique) is a powerful and widely used method.

5.1. Selected Attributes*:*

This model consists of 13 attributes, which we obtained after running SMOTE to balance the two class instances of survived and non-survived patients

1. EOD-Extension:

Documented extension of tumor removed from the primary site.

2. SEER modified AJCC:

The modified version stages cases that would be unstated under strict AJCC staging rules

3. Birth Place:

Place of birth is encoded for the patients under diagnosis.

4. EOD-Lymph Node:

Recode for highest specific lymph node chain that is involved by the tumor.

5. Regional Nodes Positive:

Records the precise variety of regional lymph nodes examined.

6. RX Summ-Surg Prim Site:

Describes a surgical procedure removes and/or destroys tissue of the primary site performed as part of the initial work-up or first course of therapy.

7. Histologic Type ICD-O-3:

Describes the microscopic composition of cells and/or tissue for a particular primary.

8. Reason for no surgery:

Documents the reason that surgery wasn't performed on the website site.

9. Age at diagnosis:

For this cancer represents the age of the patient at diagnosis.

10. Diagnostic Confirmation:

Records the simplest method used to confirm the presence of the cancer being reportable. The data item isn't restricted to the confirmation at the time of diagnosis; it's the simplest method of confirmation during the entire course of the disease.

11. EOD-Tumor Size:

Records the biggest dimension of the first tumor in millimeters.

12. Behavior (92-00) ICD-O-2:

Behavior codes of the cancer.

13. Primary Site:

Identifies the positioning during which the first tumor originated.

6. CLASSIFICATION SCHEMES

The classification schemes utilized in our experiments are of 2 types: basic classifiers, and meta classifiers. The basic classifiers contains of trees, functions, and statistical methods. The meta classifiers are utilized to boost these basic classifiers and improve their performance.

Classification is one of the data mining methodologies utilized to predict and classify the predetermined data for the specific class. Classification is a method used to extract models describing vital data classes or to predict the future data. Also there are several classification algorithms offered in literature but decision tree is the most typically used due to its simple of implementation and easier to grasp compared to other classification algorithms.

Classification is two step processes:

(i) Learning or training step wherever data is analyzed by a classification algorithm.

(ii) Testing step wherever data is used for classification and to estimate the accuracy of the Classification.

6.1. Basic classifiers

J48 Decision tree:

Decision tree is a flow chart like tree structure, wherever each internal node denotes a test on an attribute, every branch represents an outcome of the test, and each leaf node holds a class label. The decision tree classifier has two phases:Growth phase or Build phase and Pruning phase

The tree is made in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belong to each partition. The pruning phase handles the problem of over fitting the data in the decision tree. It removes the noise and outliers. The accuracy of the classification will increase during this part.It uses Gain Ratio as an attribute selection measure to make a decision tree.

Step1: In case the instances belong to the same class the tree denotes a leaf so the leaf is returned by labeling with the same class.

Step 2: The potential information is calculated for each attribute, given by a test on the attribute. Gain in information is taken into account that may result from a test on the attribute.

Step 3: The best attribute is found on the basis of present criterion and that attribute selected for branching will be measured.

Random Forest:

In 2001 Leo Breiman and Adele Cutler was developed Random forest algorithm, is an ensemble classifier that consists of many decision tree and outputs the class that is the mode of the class's output by individual trees. Random Forests grows many classification trees without pruning.

The Random Forest classifier consists of several decision trees. The final class of an instance in a Random Forest is assigned by outputting the class which is the mode of the outputs of individual trees, which can produce robust and correct classification, and skill to handle a really sizable amount of input variables.

Step1: Let M be the number of training cases, and let N be the number of variables in the classifier. Choose m as input variables, to be used to determine the decision at a node of the tree; m should be much less than N.

Step2: Recurse a training set for this tree by choosing N times with replacement from all M available training cases (take a bootstrap sample). Rest of the cases will be estimated as error of the tree by predicting their classes.

Step3: For every node within the tree, randomly select m variables, which should be based on the decision at that node.

Step4: Calculate the best split based on these m variables in the training set. The value of n remains to be constant during forest growing. Random forest is sensitive to the value of n.

Step5: Each tree is grown to the largest extent possible, into many classification trees without pruning, in constructing a normal tree classifier.

Reduced Error Pruning Tree:

The REPTree, a fast decision tree learner, has been proposed by Quinlan.It uses a decision or regression tree and creates multiple trees in different iterations using information gained or variance for selecting the best attribute. Then, it applies a greedy algorithm and reduced error pruning (with back-fitting) algorithm.

In the REPTree algorithm, numerical values are sorted only in one round for each numerical attribute.The sorting process is for determining split points of numeric attributes. The tree is built in a greedy fashion, with the best attribute chosen at each point according to information gain.

Logistic Regression:

Logistic Regression is associate approach to learning functions of the form

$$f : X \; !Y, \text{ or } P(YjX)$$

where Y is known as discrete-valued, and X = hX1 ::: Xni is known as any vector containing discrete or continuous variables. Logistic Regression assumes a parametric form for the distribution P(YjX), then directly estimates its parameters from the training data.Logistic Regression is used for prediction of the probability of occurrence of an event by fitting data to a sigmoidal S-formed logistical curve. Logistic regression is often used with ridge estimators to improve the parameter estimates and to reduce the error made by further predictions.

If there are k classes for n instances with m attributes, the parameter matrix B to be calculated are going to be an m*(k-1) matrix.

Alternating Decision Tree(ADT)

The ADTree is taken into account as another semantic for representing decision trees. In the ADTree, each decision node is replaced by two nodes: a prediction node and splitter node.The decision tree in ADTree algorithm is identical while the prediction node is associated with a real valued number. The classification in ADTree that's related with the path isn't the label of the leaf. Instead, it's the sign of sum of the prediction along the path.This is totally different from binary classification trees such as CART or C4.5 which an instance follows only one path through the tree.

Step1: if (precondition)

Step 2: if (condition)

Step3: return score one

Step4: else

Step5: return score two

Step6: end if

Step7: else

Step8: return 0

Step9: end if

6.2. Meta classifiers

A) Bagging: Bagging is a meta-algorithm to improve the stability of classification and regression algorithms by reducing variance.

B) AdaBoost: AdaBoost is a commonly used ensemble technique for boosting a nominal class classifier. In general, boosting can be used to significantly cut back the error of any weak learning algorithms.

C) Random Subspace: The Random Subspace classifier constructs a decision tree based classifier consisting of multiple trees, which are constructed systematically by pseudo-randomly selecting subsets of features, trying to achieve a balance between over fitting and achieving most accuracy.

6.3. Ensemble voting

Ensemble ways are techniques that make multiple models and then combine them to produce improved results. This method produces more accurate solution than a single model. Thus the prediction of basic classifiers and meta classifiers combined to produce a stronger final prediction set.

7. CONCLUSION

In this proposed system, we use multiple classification schemes such as basic and meta classifiers to construct models for survival prediction for colon cancer patients. The survived and non-survived classes are balanced due to the usage of SMOTE algorithm. Ensemble voting is used which increases the prediction accuracies when compared to other methodologies. The area under ROC curve and positive prediction value is increased since the ensemble voting chooses the top performing classifiers. Thus the highest prediction accuracies were obtained for the 1-year, 2-year and 5-year colon cancer survival prediction using the ensemble voting classification scheme.

REFERENCES

[1] Survivability Prediction of Colon Cancer Patients Using Neural Networks, Authors: Reda Al-Bahrani1, Ankit Agrawal, and Alok Choudhary, Health Informatics Journal, May 2017.

[2] Lung cancer survival prediction using ensemble data mining on SEER data Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi and Alok Choudhary Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, Scientific Programming 20 (2012).

[3] Treating Colon Cancer Survivability Prediction as a Classification Problem Ana Silva, Tiago Oliveira, Jose Neves, and Paulo Novais, Department of Informatics, University of Minho, Braga, Portugal, Advances in Distributed Computing and Artificial Intelligence Journal, Vol.5 N.1 (2016).

[4] Breast Cancer Data Classification Using Neural Network Approach of MLP Algorithm by A.Kathija, S. Shajun Nisha and Dr .M .Mohamed Sathik M.Phil. (PG Scholar), Prof & Head, Principal, Department of Computer Science, Sadakathullah Appa College, Tamil Nadu, India Volume 4(3), June 2017.

[5] A Study on Classification Algorithms for Predicting Colon Cancer using Gene Tissue Parameters , Aditya Tekur1, Prerna Jain, Department of Information Technology, SRM Institute of Science and Technology, Chennai, India, International Journal of Pure and Applied Mathematics, Volume119 No. 18 2018, 2147-2166.

[6] Spiculated Lesion Detection in Digital Mammogram Based on Artificial Neural Network Ensemble Ning Li, Huajie Zhou, Jinjiang Ling, and Zhihua Zhou National Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangshu 210093, China, pp. 790795, 2005.

[7] Classification Algorithm Based Analysis of Breast Cancer Data B.Padmapriya1, T.Velmurugan2 1Research Scholar, Bharathiyar University, Coimbatore, Tamil Nadu, India, 2Associate Professor, PG.and Research Dept. of Computer Science, D.G. VaishnavColllege, Chennai-600106, India.

[8] Evaluation of Decision Tree Classifiers on Tumor Datasets, G. Sujatha1, Dr. K. Usha Rani2 1Assistant Professor, Master of Computer Applications Rao & Naidu Engineering College, Ongole Andhra Pradesh, India Associate Professor Department of Computer Science Sri Padmavati Mahila Viswavidyalayam (Women's University), Tirupati Andhra Pradesh, India.

[9] Correlation Based Feature Selection (Cfs) Technique To Predict Student Perfromance, Mital Doshi 1, Dr.Setu K Chaturvedi, Ph.D 2 1Mtech. Research Scholar Technocrats Institute of Technology Bhopal, India Professor & HOD (Dept. of CSE) Technocrats Institute of Technology Bhopal, India.

[10] Gene Expression Data Analysis Using Data Mining Algorithms For Colon Cancer Archana Mishra1, Rachna Devi, Sachin Shrivastava, M.Tech Student, Assistant Professor, Department of Computer Science & Engineering, SCET, Palwal (India).

[11] Spiculated Lesion Detection in Digital Mammogram Based on Artificial Neural Network Ensemble, Ning Li, Huajie Zhou, Jinjiang Ling, and Zhihua Zhou National Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangshu 210093, China.

[12] Classification algorithm in Data mining: An Overview S.Neelamegam, Dr.E.Ramaraj.phil Scholar, Department of Computer Science and Engineering.