DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4182

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AN ENSEMBLE MODEL WITH FEATURE SELECTION TECHNIQUE FOR CLASSIFICATION OF LUNG CANCER DISEASE

A. K. Shrivas
Department of IT
Dr. C. V. Raman University
Bilaspur, India

Shatruhan Prasad
Department of IT
Dr. C. V. Raman University
Bilaspur, India

Abstract: Cancer is very serious and dangerous disease facing by many people in the world. Lung cancer is one of the most dangerous cancer types which directly affected to the human life. This disease can spread worldwide by uncontrolled cell growth in the tissues of the lung. An identification and classification of lung cancer is very necessary to diagnosis of lung cancer disease. In this paper we have analyzed the lung cancer prediction using data mining based classification algorithm such as J48, LMT, REP Tree, CART, Bayes Net, Naïve Bayes, SVM and its ensemble model. The proposed ensemble of Naïve Bayes and LMT gives better classification accuracy compare to others. We have also applied the various ranking based feature selection techniques on proposed ensemble model and achieved better classification accuracy with less number of features and less computational time.

Keywords: Waikato Environment for Knowledge Analysis (WEKA), Classification, Feature Selection Technique (FST), Lung Cancer, Cross Validation.

1. INTRODUCTION

Diagnosis of health condition is very critical issues and necessary to prevent the disease. There are various doctors and researchers have done in the field of lung cancer diagnosis and achieved better solution to find out better results. According to the World Health Organization (WHO) [1] approximately 14 million new cases and 8.2 million cancers related death in 2012. According to American Lung Association, deaths due to Lung cancer is more than the next most common cancers combined (colon, breast and pancreatic). An estimated 158,040 Americans are expected to die from lung cancer in 2015, accounting for approximately 27 percent of all cancer deaths. The number of deaths due to lung cancer has increased approximately 3.5% between 1999 and 2012 from 152,156 to 157,499. Tobacco and smoking are common causes of lung cancer that is facing by most of people.

There are various researchers have worked to identify and classification of lung cancer disease. P. Naresh, et al. [2] have suggested SVM classifier with RBF kernel function and compared with ANN and KNN classifiers was made on lung CT scan images of stage I and stage II. The proposed SVM is robust robust classifier. S. P. Tidke, et al. [3] have suggested SVM algorithm to diagnosis of lung cancer disease and achieved satisfactory accuracy. R. Kohad et al. [1] have used classifiers like SVM and ANN and proposed ACO_ANN. The proposed algorithm gives better accuracy which higher than the accuracy compare to others. T. Christopher et al. [6] have used various classification techniques to analysis and classify the lung cancer disease in WEKA environment. The Naïve Bayes algorithm gives a better performance over the other classification techniques.

P. Nithya et al. [5] have used data mining classification techniques such as neural network and SVM for identification and classification of lung cancer in X-ray chest film. N.V. Ramana Murty et al. [7] have suggested various classification algorithms such as Naive Bayesian, RBF Neural Network, Multilayer Perceptron, Decision Tree and C4.5 (J48) algorithm for lung cancer classification. Naïve Bayesian gives better classification accuracy compare to others.

2. PROPOSED APPROACH

The main objective of this research work is to develop the robust and computationally efficient model for classification of lung cancer disease. Figure 1 shows that architecture of proposed system which consist data set, individuals classifiers, ensemble classifiers, feature selection techniques and performance measures. We have used lung cancer dataset collected from UCI repository and divided the data set with 10-fold cross validation for training and testing. We have applied the dataset into various data mining based classification techniques to analysis and classification of lung cancer disease. To develop robust model, we have ensemble two or more classifiers. We have selected best classifier and applied the various raking based feature selection techniques to computationally efficient model. Finally developed model is robust and computationally efficient for classification of lung cancer disease.

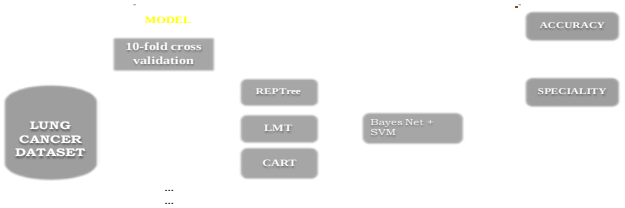

Figure 1. Architecture of proposed system

© 2015-19, IJARCS All Rights Reserved

A. Lung Cancer Dataset

Lung cancer dataset taken from UCI repository[12]. This data set consist 56 features and 1 classes having small cell lung cancer, Non small cell lung cancer and lung carcinoid tumor. All 56 predictive features are nominal, taking on integer values 0-3, and contain also missing value.

B. Classification Technique

Classification technique is one of the important data mining application and supervised learning technique. Classification process involved into two steps: training and testing. Training dataset is used to train the classifiers and testing data set is used to testing the trained classifiers. Finally testing accuracy is final accuracy of model to check the robustness of model. In this research work we have various classification techniques as explained below:

- **C4.5**

C4.5 [8] is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. In building a decision tree, we can deal with training sets that have records with unknown attributes values by evaluating the gain, or the gain ratio, for an attribute values are available. We can classify the records that have unknown attribute value by estimating the probability of the various possible results. Unlike, CART, which generates a binary decision tree, C4.5 produces tree with variable branches per node. When a discrete variable is chosen as the splitting attribute in C4.5, there will be one branch for each value of the attribute.

- **Classification and Regression Technique (CART)**

CART [8] is one of the popular methods of building decision tree in the machine learning community. CART builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. It uses the gini index for determining the best split. The initial split produces the nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine the entire input field to find the candidate splitters. If no split can be found then significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of overfitting.

- **REP Tree**

REP Tree [11] builds a decision or regression tree using information gain/variance reduction and prunes it using reduced-error pruning. Optimized for speed, it only sorts values for numeric attributes once. It deals with missing values by splitting instances into pieces, as C4.5 does. We can set the minimum number of instances per leaf, maximum tree depth (useful when boosting trees), minimum proportion of training set variance for a split (numeric classes only), and number of folds for pruning.

- **Logistic Model Tree (LMT)**

A logistic model tree [4] basically consists of a standard decision tree structure with logistic regression functions at the leaves, much like a model tree is a regression tree with regression functions at the leaves. As in ordinary decision trees, a test on one of the attributes is associated with every inner node. For a nominal (enumerated) attribute with *k* values, the node has *k* child nodes, and instances are sorted down one of the *k* branches depending on their value of the attribute. For numeric attributes, the node has two child nodes and the test consists of comparing the attribute value to a threshold: an instance is sorted down the left branch if its value for that attribute is smaller than the threshold and sorted down the right branch otherwise.

- **Bayes Net and Naïve Bayes**

Bayesian classifiers [9] are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Classification algorithms have found a simple Bayesian classifier known as the Naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

- **Support Vector Machine (SVM)**

Support Vector Machines [9] is a promising new method for the classification of both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. Although the training time of even the fastest SVMs can be extremely slow, they are highly accurate, owing to their ability to model complex nonlinear decision boundaries. They are much less prone to overfitting than other methods. The support vectors found also provide a compact description of the learned model.

C. Ensemble Model

An ensemble model is new approach to combines the two or more classifier to improve the performance of model. An ensemble model [10] combines the output of several classifier produced by weak learner into a single composite classification. It can be used to reduce the error of any weak learning algorithm. The purpose of combining all these classifier together is to build a hybrid model which will improve classification accuracy as compared to each individual classifier. This research work used voting scheme to combine the classifiers. In this scheme, combining [9] the decisions of different models means amalgamating the various outputs into a single prediction. The simplest way to do this in the case of classification is to take a vote (perhaps a weighted vote), in the case of numeric prediction, it is to calculate the average weighted average).

© 2015-19, IJARCS All Rights Reserved

D. Feature Selection

Feature selection [14] is optimization technique to remove the irrelevant feature space original feature space with the goal of obtaining a subset of features that describes properly the given problem with a minimum degradation of performance. Feature selection is not only improving the accuracy of models also decrease the computational time. In this research work we have used ranking based feature selection techniques like Info Gain, Gain Ratio, Chi Squared, ReliefF and Symmetric Uncertainty to rank the features from high important to low important.

3. RESULT AND DISCUSSION

In this research work, we have used WEKA data mining software [13] to analysis and classification of lung cancer disease using developed classifier. We have used lung cancer data set with 10-fold cross validation and applied various individuals and ensemble data mining techniques for classification of lung cancer. The main motive of ensemble model is to achieve better performance compare to its individual's model. Table 1 shows that accuracy of individuals and ensemble models for classification lung cancer disease. We have ensemble REP Tree and CART , REP Tree, CART and SVM , Bayes Net and SVM , Bayes Net, SVM and CART and ensemble Naïve Bayes and LMT as shown in table 1. An ensemble of Naïve Bayes and LMT achieved better classification accuracy as 65.62% among individuals and ensemble models.

Table 1. Accuracy of models with 10 folds cross validation

Model Name	Accuracy
C4.5	40.62 %
LMT	62.50%
REP Tree	50 .00%
CART	50.00 %
Bayes Net	50.00 %
Naïve Bayes	62.50 %
SVM	50.00 %
REP Tree + CART	**53.12 %**
REP Tree + CART+ SVM	**53.12 %**
Bayes Net + SVM	**53.12 %**
Bayes Net + SVM +CART	**59.37 %**
Naïve Bayes + LMT	**65.62 %**

Feature selection is very important technique to achieve better classification accuracy and decrease computational time. In this research work, we have used four ranking based feature selection techniques and applied the reduced feature subset in best ensemble model. Table 2 shows that ranking of top 14 features with different feature selection techniques. Table 3 shows that accuracy of best ensemble model (Naïve Bayes + LMT) with top 14 feature subset. We have compared the classification accuracy of best ensemble model with different feature subset from 14 to 3.Our proposed ensemble model gives 78.12% of accuracy in case of Gain Ratio and Symmetric Uncertainty with 5 features. Similarly, Info Gain gives 78.12% of accuracy with 7 features while Chi Squared gives 75% of accuracy with 6 feature set. Finally our proposed model gives best 81.25% of accuracy with 12 features in case of ReliefF FST

Table 2. Ranking of top best features with different feature selection techniques

FST	Ranking of features from high to low important
Gain Ratio	19,20,37,6,14,33,23,56,8,21,4,7,3,22
Info Gain	20,19,6,14,23,37,56,33,7,3,4,21,22,8
Chi Squared	19,20,6,14,23,37,33,56,8,21,4,7,3,22
ReliefF	20,19,23,6,56,8,14,27,37,2,33,40,13,53
Symmetric Uncertainty	20,19,6,14,37,23,33,56,8,21,4,7,3,22

Table 3. Accuracy of best ensemble model (Naïve Bayes + LMT) with top 14 feature subset

No. of Features	Gain Ratio	Info Gain	Chi Squared	ReliefF	Symmetric Uncertainty
14	68.75 %	65.62 %	68.75%	75.00%	68.75%
13	68.75 %	65.62 %	68.75%	75.00%	68.75%
12	65.62 %	68.75 %	65.62%	**81.25%**	65.62%
11	68.75 %	65.62%	68.75%	75.00%	68.75%
10	68.75 %	59.37%	68.75%	75.00%	68.75%
9	68.75 %	71.87%	68.75%	71.87%	68.75%
8	71.875 %	71.87%	71.87%	71.87%	71.87%

© 2015-19, IJARCS All Rights Reserved

7	71.87 %	**78.12%**	71.87%	71.87%	71.87%
6	71.87 %	75 %	**75%**	75.00%	75.00%
5	**78.12 %**	71.87 %	71.87%	75.00%	**78.12%**
3	68.75 %	71.87%	71.87%	68.75%	71.87%

4. CONCLUSION

In medical science, identification and diagnosis of lung cancer disease is very challenging task. In this research work, we have design and developed the robust and computationally efficient model for classification of lung cancer disease. We have used data mining based classification techniques to analysis and develop the robust model. We have also used ranking based feature selection techniques to rank the features based on its important and select the relevant feature from original feature space. These feature selection techniques are helpful to computationally increase the performance of proposed model. Finally our proposed ensemble of Naïve Bayes and LMT model with ReliefF FST gives satisfactory result for classification of lung cancer disease.

5. REFERENCES

[1]. R. Kohad and V. Ahire , "Application of Machine Learning Techniques for the Diagnosis of Lung Cancer with ANT Colony Optimization", International Journal of Computer Applications,Vol. 113. No. 18, pp. 34-41, 2015.

[2]. P. Naresh and R. Shettar , "Early Detection of Lung Cancer using Neural Network Techniques", International Journal of Engineering Research and Applications, Vol. 4, Issue 8, pp. 78-83, 2014.

[3]. S. P. Tidke and V. A. Chakkarwar, "Classification of Lung Tumor using SVM", International Journal Of Computational Engineering Research ,Vol. 2, Issue 5, pp. 1254-1257, 2012.

[4]. N. LandwehrMark and M. Hall and E. Frank , "Logistic Model Tree"s, Kluwer Academic Publishers, 2006.

[5]. P. Nithya, B.Umamaheswari and R. Deepa , "Detection of Lung Cancer Using Data Mining Classification Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 7, pp. 1060-1062 , 2015.

[6]. T. Christopher and J. Jamera banu , "Study of Classification Algorithm for Lung Cancer Prediction", International Journal of Innovative Science, Engineering & Technology, Vol. 3, Issue 2,pp. 42-49, 2016.

[7]. N.V. Ramana Murty and M.S. Prasad Babu, "A Critical Study of Classification Algorithms for Lung Cancer Disease Detection and Diagnosis", International Journal of Computational Intelligence Research, Vol.13, No. 5, pp. 1041-1048,2017.

[8]. A. K. Pujari, "Data Mining Techniques", Universities Press (India) Private Limited, 4th ed., ISBN: 81-7371-380-4,2001.

[9]. J. Han, M. Kamber and J. Pei , "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[10]. M. Pal, "Ensemble Learning with Decision Tree for Remote Sensing Classification", World Academy of Science, Engineering and Technology. Vol. 36, pp. 258-260, 2007.

[11]. H. W. Ian and F. Eibe, "Data Mining Practical Machine Learning Tools and Techniques", Morgan Kaufmann, San, 2nd ed., 2005.

[12]. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html]. (Browing date: 23-03-2017).

[13]. WEKA Data Mining Tools: http://www.cs.waikato,ac.nz/~ml/weka/ (Browsing date: Mar. 2017).

J. Wang, "Data Mining: opportunities and challenge", Idea Group, USA, 2003.

© 2015-19, IJARCS All Rights Reserved