Classifiers based Approach for Pre-Diagnosis of Lung Cancer Disease

K. Balachandran
Associate Professor,
Christ University, Bengaluru
Research Scholar,
Anna University of Technology

R. Anitha, Ph. D
Director, MCA,
K.S.Rangasamy College of Technology,
Tiruchengodu

ABSTRACT

Lung cancer disease is one of the dreaded diseases in the developing and developed countries. The pre-diagnosis is an important stage of identifying the target group of persons who can undergo diagnosis stage. Here in this study, prediction of lung cancer is attempted based on symptoms and risk factors. Data collected from the confirmed case of the patients is pre-processed based on multi filter approach. Pre-processed data is then tried with different classifier algorithms. It has been observed that Sequential Minimal Optimization, simple logistic and supervised learning based algorithms resulted in better performance compared to other algorithms. Detailed analysis is done based on Radial Basis function. All these algorithms are tried under cross validation approach.

General Terms

Data mining, classification pre-diagnosis, Lung cancer

Keywords

Lung cancer, Pre-diagnosis, Classification, SVM, SMO, Multi-layer Perceptron, Logistic.

1. INTRODUCTION

Lung cancer is one of the leading cause death in the developed and developing countries. In most of the cases the disease is diagnosed and detected in the advanced stage. Hence, the survival rate of the patients diminishes drastically. The prognosis is poor, with less than 15% of patients surviving 5 years after diagnosis. The poor prognosis is attributable to lack of efficient diagnostic methods for early detection and lack of successful treatment for metastatic disease. [1]. Hence, early detection of the Lung cancer is paramount to the survival rate.

The symptomatic and risk factors responsible and associated with the Lung cancer helps us to identify the target group of people who can be screened for further diagnosis of Lung cancer. As many of these factors and its impacts vary and fuzzy in nature prediction of the disease is complex. Many cellular changes have been reported to be associated with malignant process. Such studies may provide an important lead not only in the philosophy of study cancers, but also for early diagnosis of the disease and prognosis with respect to treatment modalities. It is important to comprehensively study the biological processes at cellular levels, before a logical conclusion on such association can be made. A study is carried out to find appropriate model that can test, based on these factors with statistical, artificial neural network and rule based approaches [2].

2. LITERATURE REVIEW

Simple divide and conquer algorithms for producing decision trees have been implemented in machine learning algorithms. These algorithms have been used as basis of many systems that generate rules. However, general problem with this rule based approaches are, they tend to over fit the training data. Predictive nonparametric classification and approximation methods frequently achieve high accuracy using a large number of numerical parameters in a way that is incomprehensible to humans [3]. Increasing class imbalance in the training dataset generally has a progressively detrimental effect on the classifier’s test performance. [4].

Hierarchical decision models are increasingly used within health care. For practical applications, it is particularly important that these models and supporting decision-making tools allow the structuring ofDomain knowledge and are capable of dealing with qualitative variables and utility functions. [5]. Increasing class imbalance in the training dataset generally has a progressively detrimental effect on the classifier’s test performance. [6].

According to Wai-Ho Au et al., “Unlike decision tree based algorithms, other classification techniques such as logit regression and neural networks can determine a probability for a prediction with its likelihood. However, comparing with decision tree based algorithms, these algorithms do not explicitly express the uncovered patterns in a symbolic, easily understandable form (e.g., if-then rules).” [7]

Recent research in the support vector machines arena allows the handling of large data sets.

3. METHODOLOGY

3.1 Approach

Symptoms and risk factors of Lung and similar type of cancers are collected based on the domain expert’s knowledge 74 attributes are chosen. Attribute data from the patients then collected and pre-processed with multi filter approach and classified into one of the following classes viz. Lung, Other and No cancer. The preprocessed data is then given as input to different classifier algorithms.

Some of the classification methods that are tried in this study are:

- ZeroR relies on the target and ignores all predictors. It simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification method.

- Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. Given a finite set of discrete random variables where each variable Xi may take values from a finite set, denoted by val(Xi). A Bayesian network is an annotated directed acyclic graph (DAG) G that encodes a joint probability distribution over X. Bayesian approach to unsupervised classification describes each class by a likelihood function with some free parameters, and then adds in a few more parameters to describe how those classes are combined. Prior expectations on those parameters combine with the evidence to produce a marginal joint which is used as an evaluation function for classifications in a region. The Bayes' function is based on the conditional probability Distribution function wherein the probability of 'y' occurring given 'x' can be given by

$$P\left(\frac{y}{x}\right) = {P\left({}^{x}/_{y}\right)P(y)} \Big/ {\int P\left({}^{x}/_{y}\right)P(y)\,dy}$$

Equation 1
This approach splits this posterior distribution into a prior distribution P(x) and a likelihood P[x/y]

- A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.
- Sequential Minimizing Optimization (SMO): Breaks the problem into sequence of small problems and then solve iteratively.
- Logistic: Class for building and using a multinomial logistic regression model with a ridge estimator
- Simple Logistic: classifier for building linear logistic regression models. The optimal number of Logic Boost iterations is perform and cross validated, which leads to automatic attribute selection.
- Multi-layer Perceptron: a feed-forward neural network with one or more layers between input and output layer, trained with back-propagation algorithm
- Radial basis Function Network: A Neural network model, the neurons in the hidden layer contain Gaussian transfer functions whose outputs are inversely proportional to the distance from the center of the neuron.
- Sequential Minimal Optimization (SMO) Classifier: is an improved training algorithm for Support Vector Machines. Like other SVM training algorithms, SMO breaks down a large QP problem into a series of smaller QP problems. Unlike other algorithms, SMO utilizes the smallest possible QP problems, which are solved quickly and analytically, generally improving its scaling and computation time significantly. [8]
- Classification-via-clustering: A user defined cluster algorithm built with the training data presented to the meta-classifier (after the class attribute got removed, of course) and then the mapping between classes and clusters is determined. This mapping is then used for predicting class labels of unseen instances.

For each approach, the confusion matrix is obtained and tabulated.

3.2 Algorithm

Step1:

1.1 Let the sets S and R contains the list of symptom and Risk factors. These parameters are chosen based on apriori information and domain expert's opinion.

$$\{s_1, s_2 \dots s_n\} \in S$$
$$\{r_1, r \dots r_m\} \in R$$

Defined a set W contains S and R sets.

$$W = \{i\ |i \in R \cup S\}$$

1.2 S & R data are collected from the confirmed patients of Lung and associated cancer patients. The patient data is represented as

$$P = \{p_1, p_2 \dots p_n\}$$

Some of the variable is S &Rare time continuous variables, like consumption of alcohol represented in a fuzzy set. A membership function $\mu_{A(si)}$ contains all the information contained in the fuzzy set(A).Fuzzy set A in the universe of discourse χ is defined as a set of ordered pairs.

$$A = \{(i, \mu A(i))|\ i\ \in\ \chi\}$$

1.3 For each value i in W:

1.3.1 If the output is logical output the parameter is labeled as { 0,1}

1.3.2 Else If the variable value is discrete multi value function, it is subjected to normalization. Since different variables are measured in different units and with different numerical ranges, a bias is introduced to the process.

1.3.3 Else if the variable is a continuous real value, fed into de-fuzzy classifier to convert into crisp binary input based on the individual parameter threshold. The process

1.3.4 Else if the variable is a missing value, the '0' is assigned to the variable to indicate the absence of value

1.4 Target: For each value of j in set P the output is assigned in either of the classes L, O or N (Lung, Other cancer, No-cancer).

$$D = \{L, O, N\}$$

1.5 Construct the matrix with M with P and W.

Step 2:

M data is then put into following classification tests: $T_a \dots T_j$

1. ZeroR
2. Bayes net
3. Naïve Bayes
4. Logistic
5. Simple Logistic
6. Multi layer perceptron approach
7. Radial Basis Function network
8. Sequential Minimal Optimization

9. ClassificationviaClustering

Step 3:
For each Ti the confusion matrix is found for the data matrix M.

4. RESULTS AND DISCUSSION

The results of the confusion matrix is tabulated for each test cases. These results are obtained in cross validation environment of 10 folds. The data is processed using Weka open source tools, Matlab and SPSS tools.The obtained data is tabulated in Table1. TP rate is the true positive rate, FP rate is False Positive rate, ROC is Reader Operator Characteristics. Weighted average of the three cases [L, N, O] is taken for computing the table. The weighted ROC values are plotted in the figure 1. The relative importance of the parameters along with the normalized percentage is given in the table 2.The P set contained 41 elements. For few W_i values of samples, output classification T isclosely matching with that of the D [L, N] cases. The patients who are classified as N, have been advised based by medical experts to undergo clinical tests, on suspicion that they are likely to have malignancy. Based on further pathological and clinical analysis the patients output is classified as D[N]. Since the parametric values of these samples are almost similar to D[L] cases most of the classification algorithms are not able to resolve these cases clearly..

Table 1. Confusion Matrix

Classifiers	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
ZeroR	0.537	0.537	0.288	0.537	0.375	0.396
Bayes net (BN)	0.634	0.339	0.542	0.634	0.584	0.612
Naïve bayes (NB)	0.683	0.278	0.708	0.683	0.679	0.709
Logistic (Log)	0.732	0.125	0.791	0.732	0.750	0.908
Simple Logistic (SL)	0.707	0.192	0.715	0.707	0.711	0.895
Multi layer perceptron approach(MLP)	0.732	0.101	0.818	0.732	0.754	0.908
Radial Basis Function network(RBF)	0.659	0.337	0.657	0.659	0.623	0.671
Sequential Minimal Optimization (SMO)	0.756	0.145	0.793	0.756	0.764	0.811
Classification-via-Clustering (CVC)	0.692	0.306	0.586	0.692	0.635	0.689

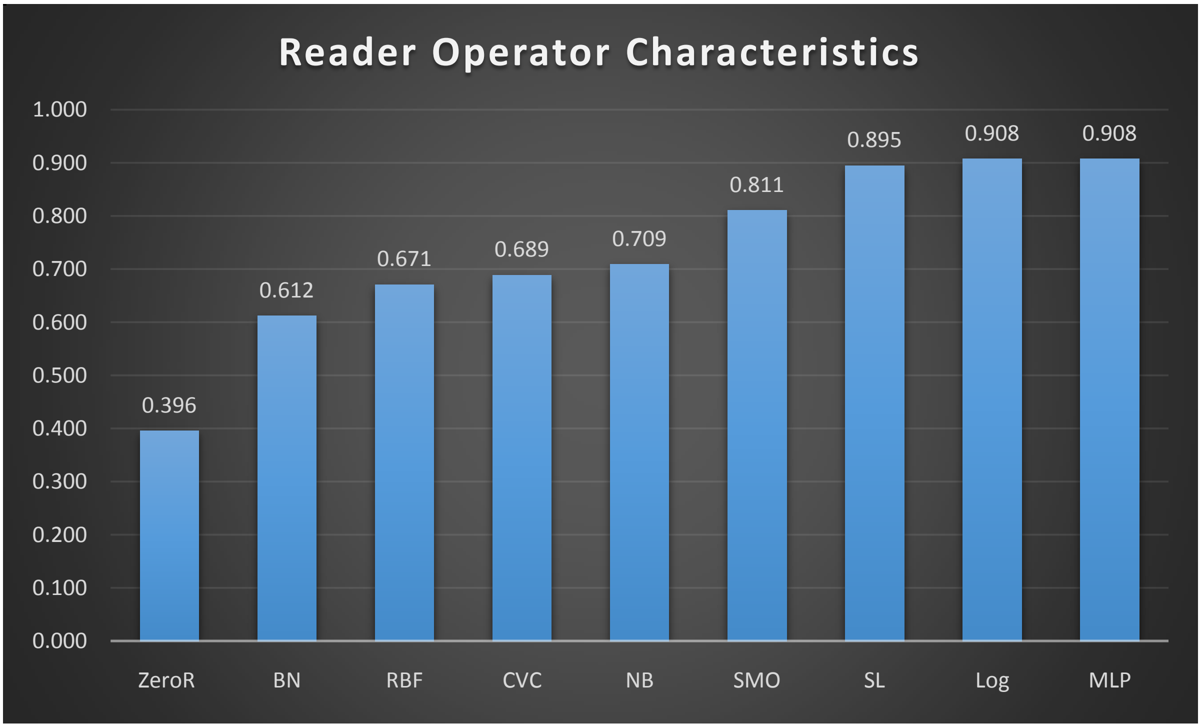

Fig.1: Reader Operator Characteristics

Independent Variable Importance					
Parameter	Importance	Normalized Importance	Parameter	Importance	Normalized Importance
G01	0.016	47.80%	C38	0.015	44.00%
G02	0.017	49.10%	C39	0.015	44.20%
G03	0.014	42.20%	C40	0.016	45.50%
G04	0.018	53.10%	C41	0.01	27.90%
G05	0.019	55.80%	C42	0.01	28.00%
G06	0.005	13.30%	C43	0.01	28.40%
G07	0.005	14.80%	C44	0.015	44.10%
G08	0.012	36.10%	P45	0.012	35.60%
G10	0.012	36.20%	P48	0.012	36.20%
G12	0.012	36.50%	P49	0.009	26.90%
G13	0.014	40.40%	RG51	0.02	57.90%
L14	0.034	100.00%	RL52	0.024	70.60%
L15	0.011	32.50%	RL53	0.018	52.20%
L16	0.019	55.20%	RL54	0.029	85.20%
L17	0.027	80.20%	RL55	0.015	43.60%
L18	0.022	63.00%	RL56	0.005	15.50%
L19	0.025	74.50%	RL57	0.015	43.40%
L20	0.014	40.90%	RL59	0.022	64.70%
L21	0.016	46.10%	RL60	0.03	88.60%
L22	0.023	67.20%	RL62	0.018	51.40%
L23	0.032	93.50%	RL63	0.019	55.10%
L24	0.021	62.40%	RL64	0.008	23.70%
L25	0.025	72.30%	RL65	0.011	31.50%
L28	0.013	39.00%	RL66	0.032	94.30%
L29	0.016	46.10%	RL67	0.007	20.80%
H30	0.016	45.40%	RL68	0.011	32.40%
H31	0.024	69.60%	RL69	0.007	19.30%
H32	0.029	84.70%	RL70	0.006	18.10%
H34	0.012	36.10%	RS73	0.02	57.30%
H35	0.024	71.20%	RS74	0.011	30.80%

Table 2 Importance of the parameters

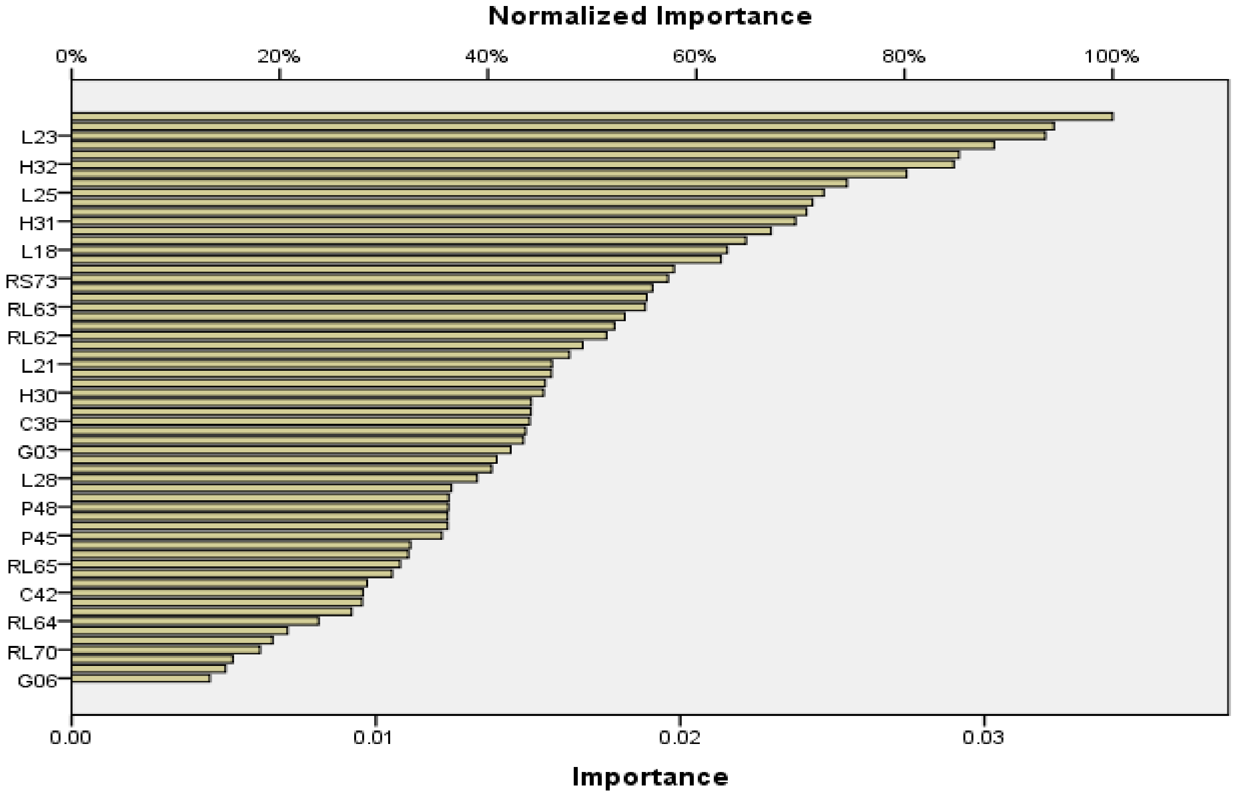

Fig.1: Normalized Importance

5. CONCLUSIONS

The result of the analysis indicates, classification based on many conventional algorithms yielded from 60 % to 85% true classification. In a multi- dimensional framework of heterogeneous types of data handling, simple statistical based or simple rule based models performances appeared to be lower compared to Logistic and Multi-layer perceptron models. Based on level of importance the parameters could be further reduced by feature selection process by removing redundant parameters.

6. ACKNOWLEDGMENTS

Our thanks to Amala Cancer research center, Thrissur who helped in the data collection and Kidwai Institute of Oncology for providing the insight and expert advice, Christ University, Bangalore, and K.S.Rangasamy college of Technology, Tiruchengodu for proving all the facilities to carry out the research .

7. REFERENCES

[1] Fred.R.Hersch, "Early Detection of Lung Cancer: Clinical Perspectives of Recent Advances in Biology and Radiology1," The National Academy of Sciences, October 2000.

[2] ICMR, "Cancer Research in ICMR achievements in Nineties," ICMR, Bangalore, 2008.

[3] D. Jensen, "Large datasets lead to overly complex models: An explanation and a solution," Proceedings 4th International Conference on Knowledge discovery and Data mining, pp. 294-298, 1998.

[4] Maciej A. Mazurowskia, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural Network, pp. 429-434, 2008.

[5] Marko Bohanec, "Applications of qualitative multi-attribute decision models in health care," International Journal of Medical Informatics 58–59, p. 191–205, 2000.

[6] Maciej A. Mazurowskia, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural Networks 21 2008 Special Issue, p. 427–436, 2008.

[7] Wai-Ho Au, "A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 7, NO. 6, pp. 532-545, DECEMBER 2003.

[8] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Microsoft Research, April 21, 1998.

[9] Marko Bohanec a, "Applications of qualitative multi-attribute decision models in health care," International Journal of Medical Informatics 58–59, p. 191–205, 2000.