Journal of Theoretical and Applied Information Technology
31st January 2014. Vol. 59 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FEATURE SELECTION BASED ON THE CLASSIFIER MODELS: PERFORMANCE ISSUES IN THE PRE-DIAGNOSIS OF LUNG CANCER

[1] K.BALACHANDRAN, [2] DR. R.ANITHA

[1] Research scholar & Associate Professor, Computer Science and Engineering Department,
Christ University, Bangalore, Karnataka, India
[2] Director, MCA, K.S.Rangasamy College of Technology, Tiruchengodu, Tamil Nadu, India
E-mail: [1] balachander63@gmail.com , [2] aniraniraj@rediffmail.com

ABSTRACT

Dimensionality reduction is generally carried out to reduce the complexity of the computations in the large data set environment by removing redundant or de-pendent attributes. For the Lung cancer disease prediction, in the pre-diagnosis stage, symptoms and risk factors are the main information carriers. Large number of symptoms and risk attributes poses major challenge in the computation. Here in this study an attempt is made to compare the performance of the attribute selection models prior and after applying the classifier models. A total of 16 classifier models are preferred based on relevancy of the models with respect to the data types chosen, which are based on statistical, rule based, logic based and artificial neural network approaches. Feature set selection and ranking of attributes are done based on individual models. Based on the confusion matrix parameters the models prediction outcomes are found out in the supervisory training mode. The Confusion matrix of the models before and after dimensionality reduction is computed. Models are compared based on weighted Reader Operator Characteristics. Normalized weights are assigned based for the result of individual models and predictive model is developed. Predictive models performance is studied with target under supervised classifier model and it is observed that it is tallying with the expected outcome.

Keywords: *Lung Cancer, Pre-Diagnosis, Data Mining, Artificial Neural Network, Classifier, Feature Selections*

1. INTRODUCTION

Lung cancer is one of the leading cause of sufferings and death in the modern world. The increase in the incidence rate and mortality is mainly due to the delay in the diagnosis and inadequacy of timely treatment. Various parameters are identified as cancer causing agents, carcinogens which affect the patients with the varying degrees. In this study Lung cancer influencing factors (attributes) are chosen based on the domain expert's knowledge. This study is tried, to aim at finding the appropriate factors to be considered for predicting Lung cancer during pre-diagnosis of the disease. The process of carcinogenicity presents a major challenge to scientists and provides limited tools for its control. Indian health services are also not adequately equipped with facilities and expertise for management of cancers. [4]. Some of the parameters that are taken for the study includes smoking, alcohol consumption, weight loss, age, family history etc., Though there are many factors attributable for the cause of Lung cancer, the extent with which each factor is contributing, to different individuals is very unpredictable. More than 70 factors are reported as possible Lung cancer causing symptoms and risk factors.

According to WebMD the official US government website nearly one fourth of all people with lung cancer have no symptoms when the cancer is diagnosed. These cancers are usually identified incidentally when a chest x-ray is performed for another reason. The other three fourths of people develop some symptoms. The symptoms are due to direct effects of the primary tumor, or due to effects of metastatic tumors in other parts of the body; or to malignant disturbances of hormones, blood, or other systems [14]. While lung cancer survival rates overall are generally poor, lung cancer survival rates vary by patient and tumor characteristics For lung cancer, stage had the most prognosis, but other factors such as grade, age, sex, and histologic type also played a role[15]. According to Doctor Barry Bloom, dean of the Harvard School of Public Health, the cancer data show that 50% of cancers could be averted with a proper diet, no smoking and other personal choices.

And it's not very expensive. We can reduce the risks, and then hope that cures will eventually be found for the remaining cancers. Our emphasis to do the most is to prevent cancer—that's a lot cheaper and a lot less painful. Even if we are born with defective genes, we may be able to avoid cancer by minimizing the environmental and life-style conditions that can initiate and promote cancer [16].

Some of the approaches that are attempted earlier to design a pre diagnosis system for different type of malignancy detection are based on:

i. Logic based approach
ii. Rule based approach
iii. Knowledge based Expert system approach
iv. Statistical approach
v. Artificial neural network based approach [supervised learning, semi supervised learning & machine learning]
vi. Genetic Algorithm approach

Data mining approaches helps in the knowledge discovery from the data. Data mining lies at the interface of statistics, database technology, pattern recognition, ma-chine learning, data visualization, and expert systems. Data sources can have records with missing values for one or more variables, and outliers could obstruct some of the patterns. A wide range of methods are available to deal with missing values and outliers. Some of the processes of data mining includes pre-processing, classification, clustering, attribute selection, model development etc. One of the useful measure is Confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix.

2. RELATED WORK

When too many factors are to be processed the number of degrees of freedom, in-creases the complexity of computation.

Expert system approaches are attempted earlier for oncology protocol management study. These models are basically designed with, either a simple rule based approach or logic based approaches. But with too many dimensions and degrees of freedom, and also the complexity involved in representing them into a numerical or logical scale poses multiple challenges.

When huge amount of data has to be handled and maintained in the databases for processing, Mazurowski et al. suggested a methodology wherein, the total number of examples stored in the system can be reduced to only 2–4% of the original database without a decrease in the diagnostic performance and algorithms based random mutation hill climbing provides the best balance between the diagnostic performance and computational efficiency [1]. Breast cancer detection, Association Rules and Artificial Neural Network techniques have been used by Murat Karabatak et al. They developed an automatic diagnosis system for detecting breast cancer based on association rules (AR) and neural network (NN)[2]

Artificial Neural network based supervised learning method are used earlier for pre-diction in health care[5] and Some modifications in the standard Neural network model for parameter estimation using compensatory neural network model is suggested by M.Sinha et al[6]

Ta-Cheng Chen used GA-based mining approach to discover the useful decision rules automatically from the breast cancer database. By using their proposed GA based approach, the significant predictors with the corresponding equality/inequality and threshold values are decided simultaneously, so as to generate the decision rules [7]. Machine learning approaches have been used for the early detection and screening of the gastric and Oesophageal cancers [8]. For cancer therapy consultation system, Curtis P.Laanglotz et al suggested automated assistance based a computer program called ONYX that combines decision-theoretic and artificial intelligence approaches to planning [17].

Wlodzislaw Duch et al., suggested the measures to be taken care of for extraction and use of logical rules for data understanding. They have also suggested the advantage of fuzzy logic in the soft evaluation of probabilities of different classes, instead of binary yes or no crisp logic answers. According to them fuzzification of the input values may give the same probabilities as the Monte Carlo procedure performed for input vectors distributed around measured values. Thus, simple interpretation of crisp logical rules is preserved, accuracy is improved by using additional parameters for estimation of measurement uncertainties, and gradient procedures, instead of costly global minimization may be used [1].

3. METHODOLOGY

3 .1 Scheme of experimental setup

Based on the domain experts' advice 74 parameters are chosen and data is collected from the confirmed lung and other type of cancer patients. A total of 41 instances have been taken for the field study and the data is classified into three class outputs. [Lung cancer, other cancer, No cancer]. The data is pre-processed, transformed and normalized.

Brief details of the classifiers used in the study:

a. Bayes net classifier: is a directed acyclic graph (DAG) with a conditional probability distribution
b. Complement Naïve Bayes classifier: CNB is related to the one-versus-all-but-one technique that is frequently used in multi-label classification, where each example may have more than one label.
c. Discriminative Multinomial Naïve Bayes Text: It is a combined generative and discriminative classifier.
d. Naïve Bayes classifier: is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions
e. Naïve Bayes Multinomial: Similar to Naïve Bayes classifier with the additional incorporation of frequency information
f. Naïve Bayes Multinomial Updateable:
g. Naïve Bayes updateable
h. Logistic: Class for building and using a multinomial logistic regression model with a ridge estimator.
i. Multi-Layer Perceptron: a feed-forward neural network with one or more layers between input and output layer, trained with back-propagation algorithm. A Classifier that uses back propagation to classify instances
j. Radial Basis Function network: is a real-valued function whose value depends only on the distance from the origin. 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.
k. Simple Logistic: classifier for building linear logistic regression models. The optimal number of Logic Boost iterations is performed and cross-validated, which leads to automatic attribute selection.
l. Sequential Minimal Optimization (SMO): SMO breaks down a large QP[linearly constrained optimization problem with a quadratic objective function is called a quadratic program .
m. Classification via clustering: A user defined cluster algorithm built with the training data presented to the meta-classifier (after the class attribute got re-moved, of course) and then the mapping between classes and clusters is determined. This mapping is then used for predicting class labels of unseen instances.
n. Classification via regression: Class for doing classification using regression method
o. Conjunctive Rule based: A rule consists of antecedents "AND"ed together and the consequent (class value) for the classification/regression.
p. Decision table: Class for building and using a simple decision table majority classifier

Brief details of the attribute selectors used in the study: Correlation Feature selection Subset Evaluation

Correlation feature Selection (CFS) evaluates subset of features. Merit (M) of the features subset (A) consisting features (e) can be expressed as

$$M_{ae}=e*R_{cf}/[e+e(e-1)*R_{ff}] \quad (1)$$

R_{cf} refers average value of feature classification correlation

R_{ff} average value of feature-feature correlation

Some of the performance parameters used in the study are:

Sensitivity: TP rate: True positive rate = True Positives/(True positives +False Negatives)

Specificity = True Negatives/(True Negatives + False Positives)

FP rate: False positive rate = 1 − specificity = False Positives / (False Positives + True Negatives)

Precision: (also called positive predictive value) is the fraction of retrieved instances that are relevant

Recall (also known as sensitivity): is the fraction of relevant instances that are retrieved

F-measure: It is the harmonic mean between recall and precision

ROC area: Reader Operator Characteristics

3.2 Algorithm

Step 1: Parameter formulation

Attributes for the symptoms [S] and risk factors [R] are carefully chosen based on domain experts' advice and a set [P] is formed.

$$P = \{x \mid x \in R \cup S\} \quad (3)$$

Step 2:

Data collection:

Confirmed malignant [Lung, Associated organs {Esophagus, Mouth, Head & Neck etc.,] patients data and confirmed benign cases [though some patients have been asked to undergo clinical tests based on the likelihood of having cancer] are collected from Hospital. Total of 41 patients data are collected.

Table 1: Data Collection Details

Sl. No.	Cancer type	Number of patients
1	Lung	13
2	Other	22
3	Benign [No-cancer]	06

Patient's data is represented as set D = {d1,d2..dn} where d_i = {p_{i1}, p_{i2}… p_{im}}

Step 3:

Data is pre-processed by following steps

i. Filling missing value.

ii. Converting numerical continuous variables into discrete variables using multi-level threshold function

iii. Converting fuzzy input to crisp input by de-fuzzier

iv. Filtering using Multi filter approach

v. Normalizing the data

Step 4:

Pre-processed data is put into set of classifiers. For each classifier model Confusion matrix [True positive, True Negative and ROC area] is obtained

Step 5:

All attribute data and the each classifier output is guided into the following Attribute selection criteria for feature selection process.

Correlation Feature selection Subset Evaluation

Confusion matrix is obtained similar to the step 4 and based on the performance optimal attribute selection criteria is selected.

Step 6:

Model development:

Based on the result find out the difference between ROC area of after and before Dimensionality reduction. If the difference is negative then don't consider the classifier model. For all positive difference model compute the weighted ROC area as the confidence parameter c. Sum all the confidence parameter and normalize it to 'weight' factor ω. This factor ω is representing the relative weight of the particular classifier model. ℝi is the outcome of each classifiers prediction. Model outcome is given by

$$O= \sum_{i=1}^{n} \omega i \, \mathbb{R}_i \qquad (3)$$

4. EXPERIMENT

The collected data is pre-processed based on the following criteria. The experiment is conducted on the collected data of 41 instances with 74 parameters. The data is pre-processed using the following pre-processing algorithm.

Pre-processing algorithm (PPA):

If the attributes sample data is binary [True / False] then it is represented in numeric form as [1,0].

Else if the data is discrete multi valued then it is normalized to the scale of 0-1.

Else if the data is continuous value data using Center of Gravity (COG) method converted in to normalized crisp values.

Else if the data is a missing value based on expectation-maximization suitable value has been replaced.

Pre-processed data is then processed using weka[Waikato Environment for Knowledge Analysis] tools[9] [10],[11], [12]. This tool is machine learning software tool developed in Java at University of Waikato, New Zealand.

The classification output is obtained under cross validation approach of 10 folds. The confusion matrix obtained is processed for each chosen classifiers and the result is tabulated in the following tables 2 and 3.

Table 2: Classifiers Performance With All Attributes

Classifiers	Sensitivity	False Positive rate	Precision	F-Measure	ROC area
Bayes net	0.634	0.339	0.542	0.584	0.612
Complement Naïve Bayes classifier	0.756	0.181	0.655	0.699	0.788
DMNB	0.805	0.175	0.829	0.770	0.897
Naïve Bayes classifier	0.683	0.278	0.708	0.679	0.709
Naïve Bayes Multinomial	0.732	0.185	0.648	0.686	0.732
Naïve Bayes Multinomial Updateable	0.756	0.181	0.655	0.699	0.868
Naïve Bayes updateable	0.683	0.278	0.708	0.679	0.709
Logistic	0.732	0.125	0.791	0.750	0.908
Multi-Layer Perceptron	0.732	0.101	0.818	0.754	0.908
Radial Basis Function network	0.659	0.337	0.657	0.623	0.671
Simple Logistic	0.707	0.192	0.715	0.711	0.895
Sequential	0.756	0.145	0.793	0.764	0.811

Minimal Optimization (SMO)					
Classification via clustering	0.692	0.306	0.586	0.635	0.689
Classification via regression	0.634	0.250	0.631	0.626	0.787
Conjunctive Rule based	0.537	0.435	0.455	0.489	0.590
Decision Table	0.707	0.298	0.645	0.656	0.680

Table 3: Classifiers Performance After Feature Selection Process

Classifiers	Sensitivity	FP rate	Precision	F-Measure	ROC area
Bayes net	0.659	0.311	0.659	0.607	0.637
Complement Naïve Bayes classifier	0.780	0.170	0.672	0.719	0.805
DMNB	0.805	0.158	0.690	0.741	0.947
Naïve Bayes classifier	0.902	0.065	0.909	0.905	0.918
Naïve Bayes Multinomial	0.829	0.181	0.712	0.765	0.929
Naïve Bayes Multinomial Updateable	0.829	0.181	0.712	0.765	0.921
Naïve bayes updateable	0.902	0.065	0.909	0.905	0.918
Logistic	0.805	0.072	0.851	0.813	0.882
Multi-Layer Perceptron	0.902	0.024	0.929	0.905	0.965
Radial Basis Function network	0.902	0.048	0.912	0.904	0.956
Simple Logistic	0.854	0.071	0.870	0.853	0.957
Sequential Minimal Optimization(SMO)	0.829	0.116	0.823	0.823	0.880
Classification via clustering	0.683	0.283	0.583	0.629	0.700
Classification via regression	0.780	0.138	0.788	0.779	0.889
Conjunctive Rule based	0.537	0.418	0.457	0.493	0.575
Decision Table	0.634	0.325	0.617	0.616	0.662

After feature selection process number of attributes are reduced to 10, and in general, most of the models there is an improvement in the performance after applying feature selection [Table [4].

Based on the above results classification model is created using the normalized generated values. Using this model testing is done again on the instances.

5. RESULTS AND DISCUSSION

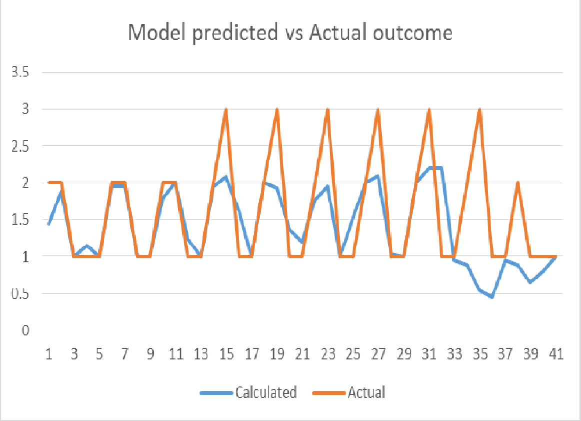
The result of the observation is tabulated in Table 1-Comparison by CFS. The result shows that though the Discriminative Multinomial Naïve Bayes (DMNB) method and other methods like Naïve Bayes Multinomial(NBM), Naïve Bayes Multinomial Up-dateable (NBMU), Sequential Minimal Optimization(SMO) results are better than the Multi-Layer Perceptron(MLP) before applying Dimensionality reduction technique, the MLP, NBU, NB & RBF classifier methods gives much improved performance after DM. Rule based and Decision tree methods does not show much variation in TP before and after DM.

Table 4. Comparison Of Performances By CFS

Classifier Name	TP before DM by CFS	TP After DM by CFS	Improvement %
Conjunctive Rule based (CR)	53.66	53.66	0.00
Decision Tree (DT)	63.41	63.42	0.02
Bayes Net (BN)	63.41	65.85	3.85
Classification via Clustering (CVC)	65.85	68.3	3.72
Classification via Regression (CVR)	63.41	78.05	23.09
Compliment Naïve Bayes (BNB)	75.61	78.05	3.23
Logistic(L)	73.17	80.49	10.00
Discriminative Multinomial Naïve Bayes (DMNB)	80.49	80.49	0.00
Naïve Bayes Multinomial(NBM)	73.17	82.93	13.34
Naïve Bayes Multinomial Updateable (NBMU)	75.61	82.93	9.68
Sequential Minimal Optimization(SMO)	75.61	82.93	9.68
Simple Logistic(SL)	70.73	85.37	20.70

Radial Basis Function network (RBF)	65.85	90.24	37.04
Naïve Bayes (NB)	68.29	90.24	32.14
Naïve Bayes updateable (NBU)	68.29	90.24	32.14
Multi-Layer Perceptron(MLP)	73.17	90.24	23.33

The following Figure-1 depicts the True Positive performance of different classifier models before and after applying dimensionality reduction.



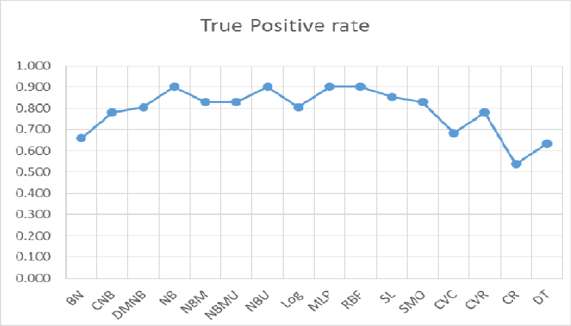


Figure 2 Predicted outcome with actual outcome

Model predicted outcome has plotted againt the actual outcome. It is observed that the model outcome closely matches with actual outcome.[Figure 2]. The normalied weight chart of the classifiers is plotted in figure [3].

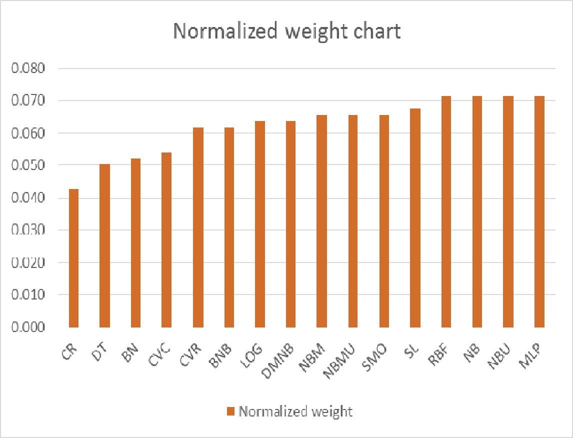


Figure 3 Normalized weight graph

6. CONCLUSIONS

Based on the above study it is observed that during the pre- diagnosis stage for the prediction of Lung cancer, data mining of multi-pronged classifier approach is performing better than any individual classifier approaches. In handling sensitive issues of determining and to draw the conclusion of presence or absence of the disease, it is better to augment the finding using multi directional approach. By combining the effort of different classifiers sensitivity, weighted ensemble model is developed. This model performance is closely matches with the actual outcome. This study also bring out the artificial neural network based models performance is superior to the simple statistical or rule based models.

REFRENCES:

[1] Maciej A. Mazurowskia, Piotr A. Habasa, Jacek M. Zuradaa et al.,. "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance." Neural Network, 2008: 429-434.

[2] 2. Murat Karabatak, M. Cevdet Ince. "An expert system for detection of breast cancer based on association rules and neural network." Expert Systems with Ap-plications, 2008.

[3] Marko Bohanec a, Blaz Zupan , Vladislav Rajkovic. "Applications of qualitative multi-attribute decision models in health care." International Journal of Medical Informatics 58–59 (International Journal of Medical

Informatics 58–59 (2000) 191–205), 2000: 191–205.

[4] ICMR. Cancer Research in ICMR Achievements in Nineties. Periodic Registry re-port, Bangalore: ICMR, 2006.

[5] Alex L. P. Tay, Member, IEEE, Jacek M. Zurada, Fellow, IEEE, Lai-Ping Wong, and Jian Xu. "The Hierarchical Fast Learning Artificial Neural Network (HieF-LANN)—An Autonomous Platform for Hierarchical Neural Network Construc-tion." IEEE Transactions on Neural Networks, VOL. 18, NO. 6,, 2007: 1645-1657.

[6] M Sinha, P K Kalra and K Kumar. "Parameter estimation using compensatory neural networks." Sadhana, Vol. 25, Part 2, 2000: 193-203.

[7] Ta-Cheng Chen, Tung-Chou Hsu. "A GAs based approach for mining breast cancer." Expert Systems with Applications 30, 2006: 674–681.

[8] W.Z. Liu a, A.P. White , M.T. Hallissey , J.W.L. Fielding. "Machine learning techniques in early screening for gastric and oesophageal cancer." Artificial Intelligence in Medicine 8, 1996: 327-341.

[9] Ware, Malcolm. weka source. May 17, 2012. http://weka.sourceforge.net/doc/weka/classifiers/functions/MultilayerPerceptron.html (accessed May 17, 2012).

[10] .weka.classifiers.functions, Package. Weka. n.d. http://weka.sourceforge.net/doc/weka/classifiers/functions/package-summary.html (accessed May 17, 2012).

[11] WŁODZISŁAW DUCH, RUDY SETIONO. "Computational Intelligence Methods for Rule-Based Data Understanding." PROCEEDINGS OF THE IEEE, VOL. 92, NO. 5,. IEEE, 2004. 771-805.

[12] Yanfeng Hou, Jacek M. Zurada, Waldemar Karwowski, William S. Marras, and Kermit Davis. "Identification of Key Variables Using Fuzzy Average With Fuzzy Cluster Distribution." IEEE Transactions on Fuzzy systems, VOL. 15, NO. 4,, 2007: 673-685.

[13] Mahalanobis, Prasanta Chandra (1936). "On the generalised distance in statistics". Proceedings of the National Institute of Sciences of India 2 (1): 49–55. Retrieved 2012-05-03.

[14] eMedicineHealth, WebMD Medical Reference from. WebMD. May 15, 2009. file:///D:/AnnaPhD/Downloads/Lung%20Cancer.htm (accessed July 01, 2009).

[15] Eisner, Lynn A. Gloeckler Ries and Milton P. "Cancer of the Lung- Chapter 9." SEER Survival Monograph. SEER, 2010.

[16] Mary K. Obenshain, MAT. "Application of Data Mining Techniques to Healthcare Data." Statistics for Hospital Epidemiology Vol 25 No.8, 2004: 690-695.

[17] L. Curtis P. Langotz, " Therapy Planning Architecture That Combines Deci-sion Theory and Artificial Intelligence Techniques," Computers and Bio-Medical Research, vol. 20, pp. 279-303, 1987.

[18] R. S. M. Wlodiszlaw W Duch, "Computational Intelligence Methods for Rule-Based Data Understanding," Proceedings of the IEEE, vol. VOL. 92, no. NO. 5, pp. 772-805, 2004