DOI : https://doi.org/10.56452/7-2-560

ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

Lung Cancer Detection Using Feature Selection and Navie Bayes Classification

Preeti chaudhary

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract

We demonstrate NB and Feature Selection, two highly important algorithms for completing this task due to their simplicity and great performance. The classification accuracy of NB and Feature Selection's class chance estimations on the training data is weighted in NB-Feature Selection. The NB-Feature Selection hybrid classification method, which is presented in this research, greatly beats NB and feature Selection in regards to accuracy in classification. It is built on top of a Bayesian network, which consists of a structural model and a number of conditional probabilities. Data mining has a number of key problems and knowledge discovery is classification, and the objective is to develop a classifier from a set of cases with class labels. Many learning algorithms, including decision trees and Bayesian networks, have been created. There are two groups into which these algorithms can be split: probability-based algorithms and decision boundary-based algorithms. The decision tree is the most popular classification model, and it categorises a test instance by sorting it through the tree from its root node to one of its leaves, and does so by simply voting.

1. INTRODUCTION

As a useful tool for the detection, prognosis, and therapy of cancer, gene expression profiling predicated upon microarray has emerged [1]. The identification of the instructive genes that cause cancer has greatly benefited in recent years by the DNA microarray approach [2,3]. The dimensionality problem, which obstructs the dataset's relevant information and causes computational instability, is the main flaw in microarray data processing [4]. In order to properly analyse cancer microarray data, relevant characteristics (genes) must be chosen or extracted. With the advent of the internet, the world has become a much smaller place. In the meanwhile, new advancements in microarray chip technology enable the simultaneous investigation of hundreds or millions of genes, producing a vast quantity of data.

It is a challenging effort to process it using a regular system with typical processing capability. The MapReduce programming concept and its Hadoop implementation provide a strong foundation for the distributed processing of huge datasets, particularly high dimensional genomic data like microarray data. Doug Cutting created the Hadoop framework in 2008 [10]. A productive way to analyse and store data in enormous amounts of data in a distributed form on big clusters of commodity hardware is offered by Apache Hadoop, an open source programme. It uses a master/slave design for both distributed computing as well as storage, completing two goals at once—massive data storage and quick processing.

By adopting the assumptions, such as the Null hypothesis and Alternative hypothesis, The feature selection strategy employs the statistical evaluations. The traits are either accepted or denied depending on how accurate the hypothesis was. Considering the class labels provided via the K-nearest training instances, the K-Nearest Neighbor classifier offers a straightforward non-parametric approach for classifying the input pattern [13]. To choose the pertinent characteristics in a dataset, an analysis of variance test based on MapReduce has been suggested in this study. Moreover, K-NN based on

MapReduce has been suggested as a classification method for the microarray dataset. in addition to this feature selection approach. These methods have been put into practise to handle different microarray datasets. The algorithms' performance is evaluated using a a Hadoop cluster comprising a standard approach plus four slave (data) nodes.

2. LITERATURE SURVEY

This study introduces a novel graph-theoretical-based cDNA microarray data classification algorithm that may largely circumvent the drawbacks of existing classification approaches. In order to make predictions, the suggested decision rule attempts to mimic a human cognitive procedure by taking into account both the proximity scores' absolute value as well as their corresponding quantities.. Each proximity score that makes up PVs is computed independently of the number of classes that are taken into account while solving the issue, rendering it to an absolute measure of how closely a sample resembles a particular class. This characteristic is particularly helpful for clinical diagnosis since it can reveal crucial diagnostic details like the discovery of genetic resemblances with well-known disorders. It can get around most of the issues with established categorization techniques. Nevertheless, it necessitates the determination of a cutoff point that distinguishes between genes that are significant and those that are not [1].

In order to categorise binary answer variables, this research suggests a novel approach that combines PLS with Ridge penalised logistic regression. The dimension-reduction step is integrated into the classification stage, and it includes a Ridge penalty step plus a PLS step. For the leukaemia, colon, and prostate data sets, the classification rule's predictive ability is demonstrated. The approach is intended to combat the dimensionality curse and solve the issue of a high-dimensional gene expression space. It offers a novel application of partial least squares to binary answer data that appears to have superior characteristics to some of the existing techniques. Future study will focus on the variable as well as model selection themes to discover optimal values for (λ, κ). This makes it possible to combine a regularisation phase with a dimension-reduction stage. Yet this approach is not appropriate [2].

This study created a technique for categorising tumours into distinct diagnostic groups based on their gene expression profiles (ANNs). Small, round blue-cell tumours (SRBCTs), which fall into four different diagnostic categories, were used as a model for training the ANNs. Further blinded samples were examined in order to assess the trained ANN models' capacity to identify SRBCTs. The models were calibrated using linear ANN models, which correctly identified each of the 63 training SRBCTs and shown no signs of over-training. ANN-based pattern recognition algorithms may have trouble identifying causal relationships between the output and the initial input data. To address this issue, we assessed the classification's sensitivity to changes in each gene's expression level and those genes for which reliable measurements were made across all samples will be included after applying a strict quality filter. This quality filter generated more reliable prediction models and helped identify 96 genes that are very important to these malignancies. We were able to create a limited set of genes that can accurately categorise our samples into their diagnostic categories and we also identified the genes that contributed to this classification in ranked order [3].

In this study, a unique hybrid method for choosing marker genes from microarray data is proposed. It chooses a collection of the top-ranked informative genes by combining gene ranking and clustering analysis. With only a small number of marker genes, experiments demonstrated extremely strong LOOCV accuracy (100% on ALL/AML leukaemia dataset employing 5 genes, 91.9% on colon tumour dataset employing 3 genes, plus 100% on MLL leukaemia dataset employing 26 genes). Our technique chooses the gene that is most closely associated with the cluster centroid as its representative. Now, we are looking at different strategies that leverage Gene Ontology to direct this selection procedure. It may increase classification precision. To test every potential cluster number, however, in search of the configuration that offers the highest classification performance, would be prohibitively costly [4].

This book describes a technique for automatically classifying text documents that uses symbolic rule induction and decision trees. It employs a quick algorithm for decision tree induction and a fresh technique for transforming a decision tree into a smaller rule set that is nonetheless intrinsically comparable to the actual tree. The system generates rules from the training data, which may

subsequently be used to classify unrelated data that is comparable to the training data. The KitCat system treats categorization vis-à-vis each category as a discrete binary-classification issue, resulting in a different rule set for each category to facilitate multiple categorisation. In order to categorise a document into a single category, the IBM Text Analyzer evaluates the rules. It also assigns a confidence measure to each rule, which represents the estimated in-class likelihood that a document meets the rule. The KitCat method contains a process for converting a conventional rule set into a simplified equivalent one based on an evaluation of the structure of the decision tree, and is supposed to be a strictly concave function that closely mimics classification error [5].

3. PROPOSED SYSTEM

The classifier model for the data on gene expression was created in the current system using a variety of methods. These methods yield accurate categorization, but the outcomes they create are challenging to understand. PLS was suggested as a solution to this problem before classification, however this is unattractive since it is made to handle continuous answers and models without heteroscedasticity problems. Another long-standing issue is a convergent of the iteratively reweighted least squares method. To be able to stabilise the statistical issue as well as to eliminate numerical degeneracy brought on by multicollinearity, penalising the likelihood proposes utilising the Ridge penalised logistic regression to address the high-dimension problem.

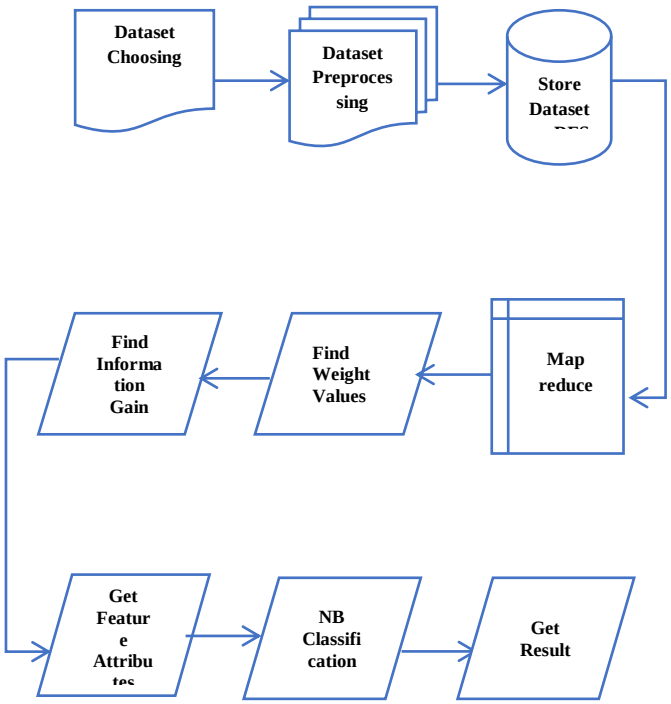

Fig 1: System Architecture

This approach doesn't function as a dimension-reduction strategy, but it seems to perform well with microarray data. The Ridge penalty is deducted from all explanatory variables before they are permitted to enter the regression model. PLS along with main component regression are two examples of similar data structures used in the area of chemometrics (PCR). As it is selected such that the sample correlation among the response as well as a linear combination of the p predictors is highest, PLS is more suited instead of PCR for dimension reduction based prediction.

We introduce our combination approach, known simply as NB-Feature Selection, in this section. Our goal is to scale up NB and Feature Selection's classification accuracy. The class probability estimations of NB and Feature Selection in NB-Feature Selection are weighted based on how accurately they

classified the training data. Let's now review the NB and Feature Selection class probability estimations. Our collaborative NB-Feature Selection method links NB and Feature Selection together. NB-Feature Selection simultaneously trains NB and Feature Selection during the training phase and evaluates each component's classification accuracy using the training data. The following are some of the suggested system's many benefits:

- It is employed to enhance classification performance accuracy.
- It offers more precision.
- It is used to forecast the precision of movement.
- That is adequate.
- The ideal feature selection Decision-making procedures

The following section discusses the many stages involved in implementing the suggested system:

1. Dataset Preprocessing

This technique will convert a csv file to text first, and then delete any unnecessary information from the text file, such as datasets. Because the preprocess replaces (-1) undesired special characters and unneeded data, it greatly aids the pruning process. These trimming methods greatly aid in time and memory conceptualization. After preprocessing, database uploading must be done.

2. Find Feature Selection

Data mining can occasionally lose its usefulness if there is too much data available. It's possible that not all of the columns of data characteristics compiled for creating and testing a model will provide the model with useful information. Some of these could even reduce the model's accuracy and quality. Irrelevant features only amplify the noise in the data and reduce the model's precision. Noise makes the model larger, requiring more time and system resources to develop and score. Moreover, data sets containing a lot of attributes may have clusters of associated attributes. These characteristics can be assessing the same underlying property. They can influence the logic of the algorithm and the model's correctness by being present together in the build data.

3. Find NB

The Bayes Theorem and the concept of predictor independence serve as the foundation of this classification technique. A Naive Bayes classifier, to put it simply, thinks that the presence of one feature in a class has no bearing on the presence of any other features. For example, if a fruit is red, spherical, and around 3 inches in diameter, it may be classified as an apple. Despite the fact that some of these characteristics are reliant upon others or each of these attributes separately raises the probability that this fruit is an apple when present, which is why it is considered to be "Naive".

4. Classification

The NB classification algorithm is being used. This technique analyses using micro-checks and generates two classes, class one with available individuals and class two with unavailable individuals. Just doing a lot of counts is all there is to it. If the NB conditional independence assumption is accurate, a Naive Bayes classifier will converge faster than discriminative approaches such logistic regression, hence less training data is required. However, an NB classifier frequently performs unexpectedly well in practise even when the NB assumption is false. An excellent choice if you want to execute any sort of semi-supervised learning or need something laughably easy to use yet works rather well.

5. RESULTS

In this study, a hybrid classification method called NB-Feature Selection that greatly beats the other two algorithms in terms of classification accuracy is introduced. A variety of learning methods, including decision trees and Bayesian networks, have been created to address classification, Among the key difficulties in data mining and knowledge discovery.

A test instance is classified using the Decision Tree, a popular classification model, by being sorted through the tree from its root node to one of its leaf nodes. Our combined approach, NB-Feature Selection, chains NB and Feature Selection and evaluates each technique's classification accuracy using

DOI : https://doi.org/10.56452/7-2-560

training data. The experimental findings demonstrate that NB - Feature Selection greatly exceeds the other two methods in terms of classification accuracy.

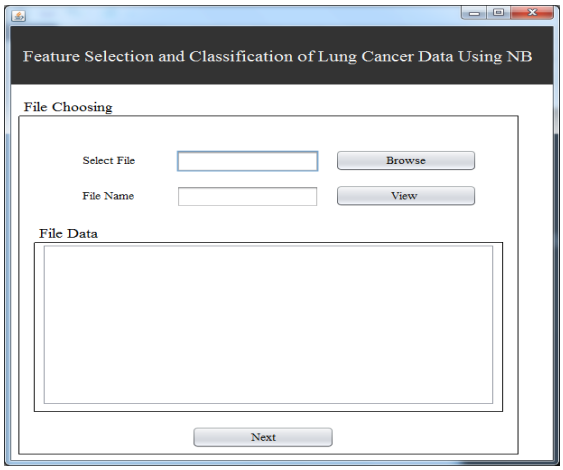

Fig 2: File Choosing

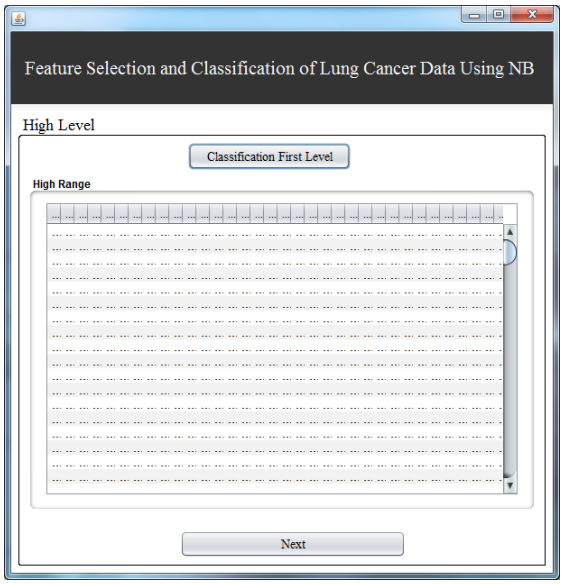

Fig 3: Classificatin First Level

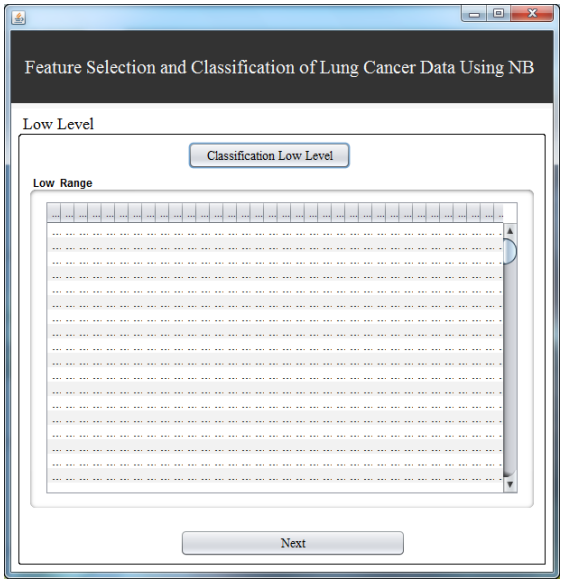

Fig 4: Classification Low Level

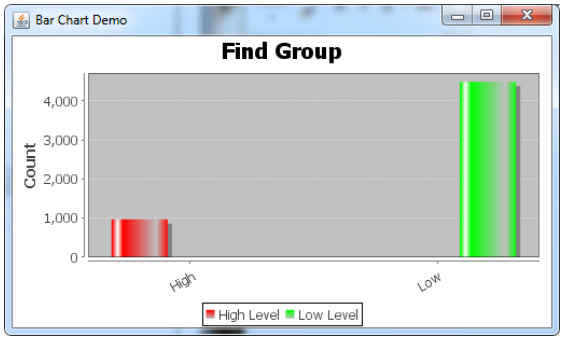

Fig 5: Performance Analysis

6. CONCLUSION

A novel idea, a prediction framework built on hybrid classification approaches, has been put out by the system in this research. This technique finds the precise level. The outcome demonstrates that it is employed to enhance classification performance. It offers more precision. We introduce NB-Feature Selection, a combination classification system based on NB and Feature Selection. The experimental findings demonstrate that NB-Feature Selection greatly outperforms NB and Feature Selection in terms of classification accuracy.

REFERENCE

[1] T.L. Bergemann and L.P. Zhao, "Signal Quality Measurements for cDNA Microarray Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 7, no. 2, pp. 299-308, Mar./Apr. 2010.

[2] A. Benso, S.D. Carlo, and G. Politano, "A cDNA Microarray Gene Expression Data Classifier for Clinical Diagnosis Based on Graph Theory," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 8, no. 3, pp. 577-591, May/June 2011.

[3] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of

Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, pp. 531-537, 1999.

[4] L. Li, "Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the ga/knn Method," Bioinformatics, vol. 17, pp. 1131-1142, 2001.

[5] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J. Am. Statistics Assoc., vol. 97, no. 457, pp. 77-87, 2000.

[6] G. Fort and S.L. Lacroix, "Classification Using Partial Least Squares with Penalized Logistic Regression," Bioinformatics, vol. 21, no. 7, pp. 1104-1111, 2005.

[7] L. Fan, K.L. Poh, and P. Zhou, "A Sequential Feature Extraction Approach for Na€ıve Bayes Classification of Microarray Data," Expert Systems with Applications, vol. 36, no. 6, pp. 9919-9923, 2009.

[8] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," Nature Medicine, vol. 7, pp. 673-679, 2001.

[9] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," Bioinformatics, vol. 16, pp. 906-914, 2000.

[10] A.C. Tan and D. Gilbert, "Ensemble Machine Learning on Gene Expression Data for Cancer Classification," Applied Bioinformatics, vol. 2, pp. 75-83, 2003.

[11] D.E. Johnson, F.J. Oles, T. Zhang, and T. Goetz, "A Decision-TreeBased Symbolic Rule Induction System for Text Categorization," IBM Systems J., vol. 41, no. 3, pp. 1-10, 2002.

[12] J.S.R. Jang, C.T. Sun, and E. Mizutani, Neuro-Fuzzy Type -2 and Soft Computting. Prentice Hall, 1997.

[13] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, and D. Geman, "Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles," Bioinformatics, vol. 21, pp. 3896-3904, 2005.

[14] Y. Yoon, S. Bien, and S. Park, "Microarray Data Classifier Consisting of k-Top-Scoring Rank-Comparison Decision Rules with a Variable Number of Genes," IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Rev., vol. 40, no. 2, pp. 216-226, Mar. 2010.

[15] P. Woolf and Y. Wang, "A Fuzzy Type -2 Logic Approach to Analyzing Gene Expression Data," Physiological Genomics, vol. 3, pp. 9-15, 2000.