Grenze International Journal of Engineering and Technology, Jan Issue

Preprocessing and Classification Algorithms on Micro Array Data for Cancer Classification

Chaitra.P.C[1] and Dr. R. Saravanakumar[2]
[1-2]Dayananda Sagar Academy of Technology and Management/Computer Science and Engineering, Bangalore, India
Email: Chaitra-cs@dsatm.edu.in, saravanakumar.rsk28@gmail.com

***Abstract*—Cancer is most dangerous disease and most of the deaths are because of this deadly disease. Microarray cancer analysis is currently one of the most popular study topics in machine learning, computational biology. In the case of binary classes, classification and analysis of cancer data play a critical role in diagnosis, distinguishing negative and positive cases and treatment. The curse of dimensionality and a lack of sufficient sample data are the key challenges with microarray cancer datasets. In this paper presents, problems associated with micro array and handling of problems and make datasets ready for analysis. Different classification algorithms explained with pros and cons of each algorithm in detail.**

***Index Terms*— Micro array, Dimensionality reduction, Cancer classification.**

I. INTRODUCTION

Cancer is a broad term that encompasses a comprehensive range of diseases that can affect any section of the human body. Malignant tumours, neoplasms are other terms used. One of the hallmarks of cancer is the rapid emergence of aberrant cells that grow beyond their normal borders, allowing them to infect neighbouring sections of the body and migrate to other organs; this is known as metastasis [1]. Cancer metastases are the leading cause of death.

Internal and external causes can harm the genetic make-up of cells, causing them to grow indefinitely and develop tumours. Internal aspects such as incorrect cell multiplication and destruction of deoxyribonucleic acid are main factors, but environmental aspects such as chemicals in cigarette, radiation, a UV light from the sun, drinking alcohol, improper diet, lack of exercise, and air pollution are significant environmental factors that cause cancer.

A micro array is a high-density array of thousands of individual DNA sequences printed on a glass slide. This enables for the simultaneous monitoring of thousands of genes' expression levels [1]. The matrix from gene expression is provided via DNA microarray technology, allowing for very quick calculations.

Microarray is an appealing study field because it is commonly used to investigate datasets with high dimensionality, which necessitates a lot of memory and processing power.

Microarray technologies are producing fresh data at an incredible rate. It's difficult to assess these data because of their sheer magnitude and the technological intricacy of microarray chip designs.

Due to the expensive and the restricted availability of samples, typical microarray gene expression data has a large feature with very less samples issue. This problem makes it hard for classification algorithms to analyse and learn the data distribution [2].

Grenze ID: 01.GIJET.8.1.112

Once the gene expression data collected, used for analysis. But collected data not suitable for analysis because of noise, missing data with high dimensionality of data. One more problem with micro array data is imbalance data. This has resulted in a decrease in classification accuracy as well as an over-fitting problem. Many existing methods not suitable for these problems. Once all these problems treated, suitable for extract insights from huge amount of data. Next section explains about steps to prepare, analysis and classify micro array data for Classification and performance analysis for better accuracy for proper prediction and proper treatment.

II. Major Steps in Analysing Micro Array Data

Fig. 1 gives steps to follow for analyzing micro array data. In majority of classification approach, after collecting the data from repositories, preprocessing take place for better accuracy. Because of high dimensional data, accuracy may reduce and computation may high. To handle this dimensionality reduction technique used. After these process proper classification method trained and tested using data which processed.

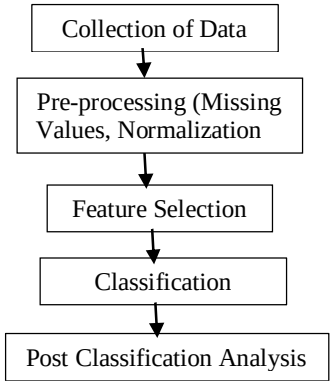

Figure 1: Steps for Analyzing micro array data

A. Collection of Data

The necessary input data for cancer classifications are downloaded from many repositories like Kent Ridge Biomedical Dataset Repository, NCI Designated cancer centers etc.

B. Data Preprocessing (Normalization and missing values)

Features of micro array data has high variance. So it need to be bring to standard format by using proper normalization techniques. In many cases Min-Max method used to do the normalization and The existence of missing values is a prevalent issue in cancer data analysis. In a typical micro array, 1-10% of the data entries are absent, affecting up to 95% of the genes. Missing expression values can be caused by a number of factors like i). weak spots ii). during the hybridization process, there were several technical errors iii). on the slides, there is dust, scratching, and systematic flaws. Many gene data analysis methods need a complete array of gene expression data as input.

There are many methods to treat missing values:

- Do nothing or ignore.
- Substitute missing values with mean/mode/median/ specific values.
- Use some imputation techniques like Knn to impute the values.

C. Feature selection

Microarray data is recognized for having a high dimensionality problem, with thousands of genes in a very less samples in a dataset. Furthermore, the majority of the characteristics are irrelevant to the problem [3]. Thus, reducing dimensionality while ensuring that the selected features still comprise all necessary and useful data could help address the data imbalance issue, while it remains a tough task. Needs to converted into lower dimensional space.

Dimensionality reduction helps to avoid over-fitting, improves accuracy, keeps the model simple. To get strong predictive performance, care should be made in obtaining the smallest feasible set of genes.

Feature selection algorithms are: Filter based, wrapper based, embedded based and hybrid based feature selection.

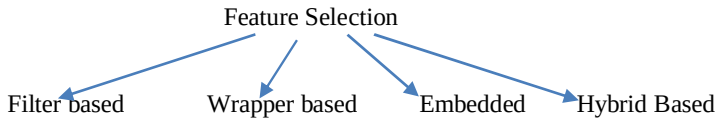

Figure 2: Methods of Feature Selection

Filter algorithm: Individual characteristic features were being used to make the selection. In this considers each features as independent.
Wrapper algorithm: Uses machine learning algorithms to select subset features.
Embedded Approach: When compared to wrapper approaches, this method benefits from interfacing with the classification algorithm and has a reduced computing cost.

D. Imbalance data

Imbalanced classes mean interested positive class examples are less than other class and it is very common in gene expression data. For example, cancer classes likely to be less common than non-cancer classes [4].

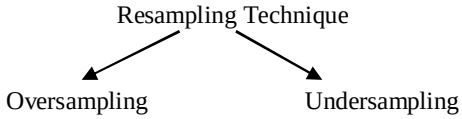

Figure 3: Technique to handle imbalance data

Oversampling is the process of increasing the number of copies of a minority class. When size of the dataset not sufficient then oversampling might be a useful option.
The term "undersampling" refers to the practice of excluding certain observations from the majority class. When dataset size is good, such as millions of rows, undersampling is useful option. However, there is a disadvantage in that by eliminating potentially useful information. This might result in underfitting and poor test set generalization.

E. Classification Algorithms

Once the dataset ready for classification, many supervised and unsupervised algorithms are available to do classification. Before applying algorithm for classification, data split into two sets, training set and testing set. Training set used to train a classifier and test set used to validate the classifiers. In that few algorithms are explained here with advantages and disadvantage of each algorithm [5].

Support Vector Machine
The Support Vector Machine (SVM) is a technique that can handle any amount of data, linear/nonlinear, high-dimensional problems and local minimum problems. SVM frequently using to classify DNA gene expression data, and it has confirmed that, it outperforms all other machine learning approaches [6]. Unbalanced data, on the other hand, will be an issue because SVM treats all samples equally important, resulting in results that are biased in favor of the minority class.
Advantages:

- SVMs have a lot of advantages, one of which is scalability. Because the volume of available micro array data expression increasing intensely, scalability is critical.

Disadvantage

- SVM considers all instances in the equal significance thus the results are partiality for minority class. Imbalanced data will be a problem.
- The resulting model is difficult to grasp and analyses, with varying weights and individual influence.

Ensemble Model
It is the grouping of the predictions of more than one similar/dissimilar classification/ clustering algorithm that are efficient in treating class imbalance related problems. It can have homogeneous / heterogeneous classifiers. Each classifier predicts class label individually [7]. In this techniques prediction take place in two ways. i) Hard voting- majority voting. ii). Soft voting- it calculates weighted probability for every class and final class considered based on maximum probability. In many of the experiments proved that soft voting performs good compare to hard voting.

Advantage:

- Performance: When compared to a single contributing model, an ensemble can make better predictions and produce better results.
- Robustness: The spread or dispersion of the predictions and model performance is reduced by using an ensemble.

Random Forest (RF)

It's a tree-based ensemble learning algorithm. RF Classifier is a group of decision trees can constructed from a various subset of the training data chosen by various technique like bagging, booting etc. It combines the output from various decision trees to determine the test object's final class [8].

The ability to handle a large number of inputs and a speedy learning process are two qualities that make RF an effective algorithm for supervised decision making.

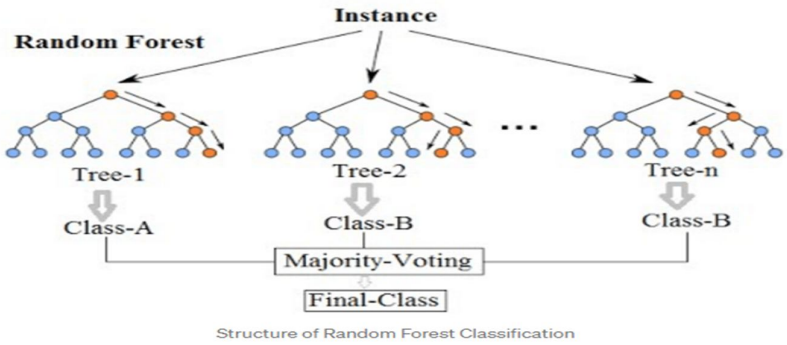

Figure 4: Structure of Random Forest classification

Advantages:

- Can be used when there are many more variables than observations and suits for both binary and multi class problem.
- Can work on both categorical and numerical predictors in a model.
- Interactions between predictor factors are taken into account.
- Returns measures of variable (gene) importance.
- can be used when the number of features larger than the number of observations
- used for multi class classification and returns measures of variable significance.
- When a considerable fraction of the data is missing, this strategy is effective for predicting missing data and maintaining accuracy.

Disadvantages:

- When instances have categorical features which takes more values or levels than others then RF are biased in favors of qualities with more levels.

Convolution Neural Network

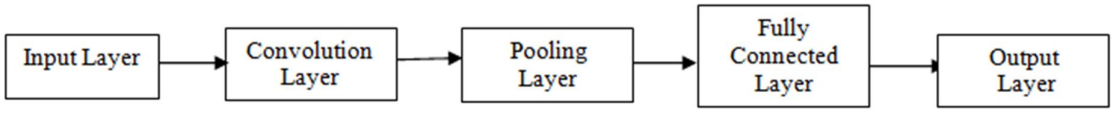

Figure 5 : Different layers of CNN

Convolutional neural network (CNN) is an example of deep learning strategy is representing human brain function in processing data. Convolutional Neural Network (CNN) can extract features from data. CNN simplifies network models by giving weights to single feature mappings, resulting in lower overall weights [9].

Multi-layer perceptron are regularized alternatives of CNNs. Multilayer perceptrons are typically entirely connected networks, meaning that each neuron in one layer is linked to all neurons in the following next layer. These networks "complete connectedness" makes them vulnerable to data overfitting. Fig. 5 gives different layers convolution neural network. It consists of data receiving layer, one or more hidden layers and output giving layer. Each layer contains one or more neurons based on the requirements.

Each neuron connected to all node of next layers and there will be weights associated with each connection. Each node has some threshold and if output generated by that node is above threshold then that node will be activated.
The architecture of the neural network is designed to enforce generic categorization for the samples provided while avoiding over-fitting the classifier to the underlying meta expression set's anomalies.
In CNN architecture, the main layer is the convolutional layer. Convolution is the procedure of iteratively executing a certain function toward the result of another function. Generation of several grids is the responsibility of Max pooling from output from the output of the splitting convolution layer. The full connection layer is a nearly complete CNN that includes 90% of the whole CNN architectural features.
Advantages:

- convolutional neural network is a more scalable approach.

Disadvantages:

- The neural network classifier takes a lot of time and requires a lot of computing power. As a result, iterative DNN training frequently necessitates the use of unique hardware characteristics present in recent GPUs or more complex methods to accelerate computation.
- In learning phase, there is high risk of data overfitting due to the large number of features and these are connected with complicated manner.
- Because of enormous complexity and vast number of parameters, they have large memory and computational requirements.

K Nearest Neighbor (Knn)
KNN is called lazy learner, non-parametric classification technique. In instance based leaning, all data sets need to be stored in memory for analysis. If we divide the entire classification process as leaning phase and testing phase, in learning process only training datasets stored in the memory and there is no computation take place during this phase. When new instance comes for classification then different distance measurements techniques used for calculation of similarity between new instance and the datasets used [10]. After calculation of similarity, top 'k' nearest neighbors used for prediction. To predict the class, maximum occurrence of class among the k nearest neighbors used. For a particular training set, cross-validation determines the number of neighbours utilized (k).
Advantages:

- Better choice to learn non-linear relationship of gene expression data.
- Easy to understand and interpret the data.
- It's an instance based learning and there is no learning phase. During testing phase all computation on data takes place. Hence it is called as lazy learner.
- KNN is easy because only parameter required for analysis is value of number of neighbours used for computation and function which is used for distance calculation like Euclidean, manhattan etc.
- It works only on small amount data which are related to the future example which needs to be tested and computation take place only on small portion of data.

Disadvantages:

- Difficulty in working with high dimensionality data. So need dimensionality reduction techniques to be applied before classification.
- High computation and required high memory for storing entire training data.
- Difficult to work with the data which has high variance because similarity calculation by distance measurements. Hence normalize or feature scaling should be done before classification.
- Sensitive to outliers, noise and missing values

F. Verification and Validation

The performance of the proposed approach is evaluated using receiver operating characteristic(ROC) curve, classification accuracy, recall, log-loss, precision, f-measure and confusion matrix [11].
ROC curve: Most imbalance classification problems predictive models are ROC Curve and precision-recall curve. By merging confusion matrices at all threshold levels, the ROC curve describes the performance. The AUC (Area Under Curve) translates the ROC curve into a numerical representation of a binary classifier's performance.
Confusion matrix: For N Class Problem, it is NxN matrix, which is used to assess the performance of a classification model. In the matrix, row represents actual and columns represents prediction. compares the actual

versus model predictions. Using the values given by matrix the performance of the model calculated [12]. Fig. 6 represents two class problem-Positive and Negative Class.

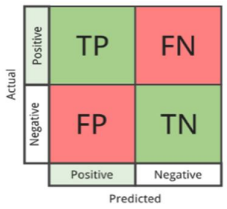

Figure 6: Confusion matrix

TP-Tue Positive, FN- False Negative, FP-False Positive, TN- True Negative

Metric	Formula
True positive rate, recall	$\frac{TP}{TP+FN}$
False positive rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot precision \cdot recall}{precision + recall}$

Figure 7: Performance Metrics for Evaluation

TABLE I: COMPARISON OF CLASSIFICATION ALGORITHM

Algorithms	Advantages	Disadvantage
SVM	• Scalability. • Effective in number of dimensions greater than samples	• Imbalanced data will be a problem. • difficult with varying weights and individual influence
Ensemble Model	• Better predictions and produce better results. • Robustness:	• Computation high
Random Forest (RF)	• Suits for both binary and multi class problem. • Can work on both categorical and numerical • Can be used when the number of features larger than the number of observations • works well with missing values.	• Biased in favors of qualities with more levels. • More computation • More data then more complex structure • suffers interpretability and fails to determine the significance of each variable.
Convolution Neural Network	• Scalable approach. • Doesn't need reprogramming.	• In learning phase, there is high risk of data overfitting • Time requited by learning phase is high • Lack of transparency.
K Nearest Neighbor (Knn)	• Better choice to learn non-linear relationship of gene expression data. • Easy to understand and interpret the data. • It works only on small amount data.	• Difficulty in working with high dimensionality data. • High computation and required high memory. • Difficult to work with the data which has high variance. • Sensitive to outliers, noise and missing values

III. CONCLUSIONS

Micro array technology has the potential to improve the development of rational therapeutic approaches as well as cancer diagnostics and prognosis. The advancement in methods and technology has leaded to gain high-throughput microarray data containing thousands of genes expression profile. This requires selection of relevant

and non-redundant top-ranking genes that contains all necessary information out of complex structured genes in order to process and diagnose diseases. So feature selection is important to reduce the dimensionality. Gene data has high variance and to work on these data, normalization crucial. Many algorithms are there for classification, for better prediction of cancer, for better and proper treatment. For better performance, selection of algorithm is very important based on the datasets and number of classes in classification process. Selection of performance metric also matters based on the type of datasets and problems associated with it. It is observed that applying classification model after pre-processing the data gives more accurate result than non-treated data. In order to get better performance GPU programming model can be used because it maximizes parallel processing. While several algorithms have been developed to distinguish between two disease classes, particularly for classification in the presence of multiple cancer types needs more concentration for better treatment.

REFERENCES

[1] Y. F. Leung and D. Cavalieri, "Fundamentals of cDNA microarray data analysis," Trends in Genetics, vol. 19, no. 11, pp. 649–659, 2003.

[2] RM Parry, W Jones, TH Stokes, JH Phan , RA Moffitt , H Fang , L Shi , A Oberthuer, M Fischer , "K-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction", The Pharmacogenomics Journal (2010)

[3] Ramón Díaz-Uriarte, Sara Alvarez de Andrés, "Gene selection and classification of microarray data using random forest" *BMC Bioinformatics 2006*.

[4] Yuanyu He, Junhai Zhou , Yaping Lin, Tuanfei Zhu, "A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data", Computational Biology and Chemistry.

[5] Chaitra P.C, Dr.R. Saravana Kumar, "A Review of Multi-Class Classification Algorithms," International journal of pure and applied Mathematics.

[6] Sayan Mukherjee , "Classifying Microarray Data Using Support Vector Machines", MIT/Whitehead Institute for Genome Research and Center for Biological and Computational Learning at MIT.

[7] Sajid Nagi • Dhruba Kr. Bhattacharyya, "Classification of microarray cancer data using ensemble approach" Netw Model Anal Health Inform Bioinforma (2013) 2:159–173I.

[8] C. Catal, "Performance evaluation metrics for software fault prediction studies," Acta Polytechnica Hungarica, vol. 9, no. 4, pp. 193–206, 2012

[9] Jinglan Zhang , Amjad . Humaidi, Ayad Al-Dujaili , Ye Duan , Omran Al-Shamma, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions" Alzubaidi et al. J Big Data (2021) 8:53.

[10] RM Parry, W Jones, TH Stokes, JH Phan , RA Moffitt , H Fang , L Shi , A Oberthuer, M Fischer , "K-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction", The Pharmacogenomics Journal (2010)

[11] B. H. Shekar, Guesh Dagnew, "Classification of Multi-class Microarray Cancer Data Using Ensemble Learning Method," Springer Nature Singapore Pte Ltd. 2019

[12] Y. Peng, "A novel ensemble machine learning for robust microarray data classification," Computers in Biology and Medicine, vol. 36, no. 6, pp. 553–573, 2006.

[13] C. Catal, "Performance evaluation metrics for software fault prediction studies," Acta Polytechnica Hungarica, vol. 9, no. 4, pp. 193–206, 2012

[14] https://medium.com/swlh/random-forest-classification-and-its-implementation-d5d840dbead0