IJDI-ERET

INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH AND EMERGING TECHNOLOGIES
Vol. 11, No. 1, 2022
www.ijdieret.in

(Research Article)

Dimensionality Reduction for Microarray Data: An Analytical Survey

Ipsita Paul[1*]

[1*]*Department of Computer Science and Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar, Orissa, INDIA*

Abstract

Recent studies in bio-informatics and bio-medical data include microarray technology to gain insight about organisms. While researchers still study on microarray data the real challenge remains as to interpret the huge dimensions that is the vast features with a very low sample space. This paper aims to reduce the enormous dimension of the microarray dataset using feature selection and classification methods. A variety of datasets with binary and multi-classes were taken for experimental analysis. Different classification algorithms were deployed to the reduced datasets and a high classification accuracy was achieved. In this paper, comparative study on classifiers using proper visualization tools have also been discussed.

Keywords: Dimensionality reduction, Microarray data, Classification, Machine Learning Algorithms.

1. Introduction

In the recent years, microarray data analysis has gained huge importance in machine learning. Generally, the bio-informatics data which includes microarray data are in raw form. Due to which, we need to pre-process the data to apply machine learning algorithms. Analysis and interpretation of DNA Microarray datasets has led to the development of an active area of research in machine learning and bioinformatics. It has been established from the past literature that very few genes from these enormous number of genes available in a DNA are actually required for classification. This makes feature selection (i.e., removing redundant and irrelevant features) a challenging task, specifically for microarray data. Effective feature selection improves the classification accuracy by removing a large number of irrelevant genes [7].Dealing with microarray data is usually difficult, since it is a structured data which is characterized by very few samples with numerous features. To remove the irrelevant characteristics and finally classifying the dataset pose serious challenge for researchers in the field of machine learning. We also have to consider the likelihood of "false positives", which can occur during our model creation for prediction or during classification of relevant features/genes [16].

In this paper, various feature selection techniques are used which address the “Curse of Dimensionality” to some extent. So, along with reducing the data size, the running time is also effectively reduced. After feature selection is performed, the classification accuracy was measured using certain classifier models. Effective feature selection has improved the classification accuracy by removing a large number of irrelevant features. As a part of result analysis, satisfactory accuracy score was achieved after feature selection which has been discussed later. Also, a comparative study has been conducted related to the performance analysis and accuracy score of the microarray data classification. For easy performance comparison among the classifiers, a visualization tool has also been used in this paper.

2. Methodology

This paper aims at preparing the raw data preliminarily by removing noise and redundant features and further applying proper dimension reduction techniques to reduce the number of features. Finally, the performance of classifiers were noted and compared as an analytical review.

2.1 Preparing the data (the early stage): The raw data with which we started, may contain irregularities (e.g.; duplicate entry, some missing values etc.) for processing using machine learning tools. Duplicates in the dataset were checked thoroughly using machine learning techniques and removed the same, if found. For missing values, they were replaced with the column average, so that the mean value of the variable remains the same. Next the features with very low self-variance (quasi-constant valued features) were removed, which behaves almost like a constant and can least affect our analysis. For this, self-variance threshold level was kept as low as 0.01. Correlation is a statistical method used to assess a

[*]*Corresponding Author: e-mail: ipsitapaulparna@gmail.com, Tel-+91-7205308557*
ISSN 2320-7590 (Print) 2583-3863 (Online)

possible linear association between two continuous variables. When the correlation coefficient is positive, it means that the gene expression profiles behave similarly. The larger the correlation coefficient, the stronger the relationship. When the correlation coefficient is one, it means that the gene expression profiles are identical. For a correlation between variables x and y, the formula for calculating the sample Pearson's correlation coefficient **r**, is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

Finally to remove the mutually correlated features with high values, Pearson's co-relation method was used with threshold value for mutual correlation co-efficient among the features >0.9.

2.2 Feature Selection Techniques: Feature selection is the process of reducing the number of features (i.e.; input variables) to reduce the computational cost of modelling and, in some cases, to improve the performance of the model. Feature selection methods can be used to identify and remove unnecessary, irrelevant and redundant attributes from dataset that do not contribute to a predictive model. There are two main types of feature selection techniques: supervised and unsupervised, and supervised methods may be further divided into wrapper, filter and intrinsic sub-sections.

Supervised Feature selection methods first evaluate the relationship between input variable(s) and the target variable and selects the most appropriate or removes the least significant input variable(s). Anova, which stands for Analysis of Variance, is one of the supervised algorithms. The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of ANOVA, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples. The Anova test is performed by comparing two types of variation, the variation between the sample means, as well as the variation within each of the samples. The below mentioned formula in table 1 represents one-way Anova test statistics:

Table 1. One Way Anova

	Sum of Squares	Degrees of Freedom	Mean Square (MS)
Within samples	SSW $= \sum_{j=1}^{k}\sum_{j=1}^{l}(x - \bar{x}_j)^2)$	k-1	MSW $= \frac{SSW}{k-1}$
Between samples	SSB $= \sum_{j=1}^{k}(\bar{x}_j - \bar{x})^2$	n-k	MSB $= \frac{SSB}{n-k}$
Total	SST $= \sum_{j=1}^{n}(\bar{x}_j - \bar{x})^2$	n-1	

Exhaustive Feature selection was also used to reduce the dimension of the dataset thereby finding every possible feature combination and utilizing the best feature subset. RFE is also a supervised method, which is specifically a wrapper method that uses cross-validation. Wrapper feature selection methods create many subsets with different combination of input features and select those features which gives best result according to a performance metric. RFE is a greedy optimization algorithm which tries to find the best performing feature subset using backward elimination. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration and constructs the next model with the left features until all the features are exhausted. Finally, it ranks the features based on the order of their elimination. Below is a flowchart of the RFE technique in figure 1.

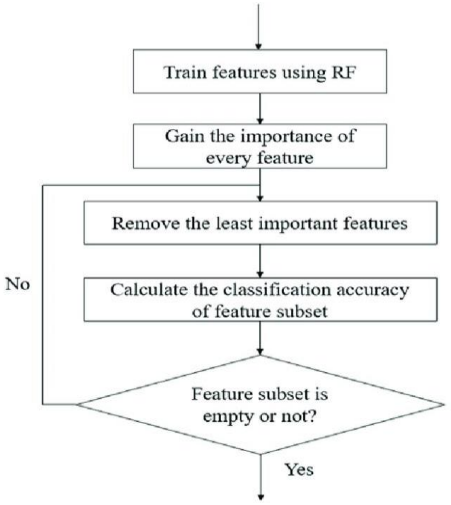

Figure 1. Flowchart of RFE

Unsupervised feature selection techniques ignores the target variable, such as methods that remove redundant variables using correlation. PCA is an unsupervised method, which is a most widely used tool in exploratory data analysis and in machine learning for predictive models. It is the process of computing the principal components which creates a visualization of data that minimizes residual variance (in the least square sense) and maximizes the variance of the projection co-ordinates. PCA has various steps which are given below in figure 2.

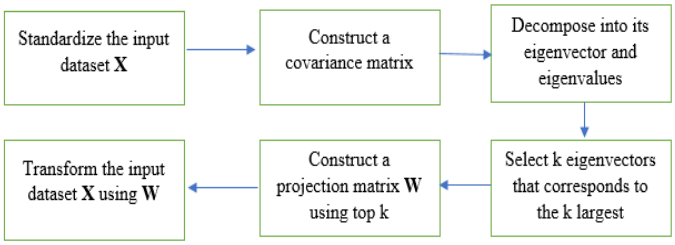

Figure 2. Stepwise PCA

We can reconstruct the original data (X) from the reduced (k) number of variables using the formula:

$$\boldsymbol{X} = \sum_{k=1}^{n} \boldsymbol{W}_k \, \boldsymbol{C}_k^T$$

Where $\boldsymbol{W}$ are the loadings and $\boldsymbol{C}$ are the components, **T** at superscript indicates Transpose vector.

2.3 Classification Techniques: After performing feature selection, the classification accuracy of the reduced datasets were calculated using some well-known classifiers, namely, Logistic Regression, k-Nearest Neighbour (KNN), Random Forest and Decision Tree. The datasets were put into these models, where we got the accuracy and f1 scores. These models have been chosen in such a manner that they could address both the multi-class classification and the binary class classification.

Logistic Regression basic model is described by:

$Logiti = b0 + b1X1 + b2X2 + \cdots + bnXn$

Where *Logiti* represents the logit transformation used with the dependent variable of the sample *i*, *Xn* represents the *n*th attribute, and *bn* represents its corresponding coefficient. The higher the absolute value of the coefficient, the higher the influence of the corresponding attribute for the class membership decision. For classification using KNN, a class label assigned to the majority of K Nearest Neighbours from the training dataset is considered as a predicted class for the new data point. KNN is a very versatile algorithm since it can also be used for regression and searching problems as well. Decision Tree as a classifier tool is versatile, so it was applied across both small and large raining datasets. It was used for multi-dimensional analysis with multiple classes. Recursive partitioning was deployed on the basis of feature values. Last but not the least, Random Forest classifier was used effectively on these large datasets since the method creates an unbiased estimate as the forest progresses.

3. Result and Analysis

In this paper, various ways were used to address the massive dimensions of the Microarray datasets. Effectively the dimensions of the data could be reduced, which were required classification. Microarray datasets like Colon Tumor, Breast Cancer, MLL, Leukemia and SRBCT were chosen for analysis purpose. Some of these datasets have a binary target class and some with multi class targets.

After initial data preparing, some well-known dimension reduction techniques like PCA, ANOVA, Exhaustive feature selection and RFE were used.

Now, lets have a look at the reduced datasets in the below table (table no. 2).The below table gives the minimum number of features for each dataset that will be considered for classification algorithms.

Table 2. Analysis of dimension reduction

Dataset	Initial no. of features	Reduced no. of features
Breast Cancer	24481	137
Colon Tumor	2000	133
MLL	12582	526
SRBCT	2308	352
Leukemia	7129	878

Now, classification methods as discussed earlier in this paper were used on the reduced datasets. Further the Classifier performance was analysed using accuracy score and f1 score.

Accuracy score is the measure of the classification accuracy, which is the ratio of number of correct predictions to the total number of predictions made. The value is significant in deciding whether the classification has been done exactly as per the prescribed collected data. The range of accuracy score is [0, 1].

This score can be mathematically calculated as:
Accuracy score = (TP+TN) / (TP+TN+FP+FN)

We also have measured the f1 score along with the Accuracy score. f1 score is a measure about how precise our classifier is as well as how robust it is. The range for f1 score is [0, 1]. It tries to find a balance between precision and recall. Greater the f1 score, better is the performance of our classifier model.

The f1 score is the harmonic mean of precision and recall and can be expressed as:

F1 score = 2 * (precision * recall) / (precision + recall)
where, Precision is the ratio of true positive to actual results and Recall is the ratio of true positive to predicted results. A more clear understanding of the same can be obtained from the confusion matrix (figure 3).

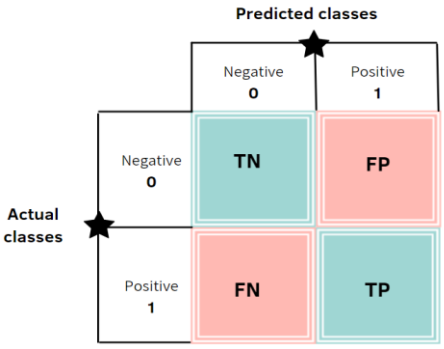

Figure 3. Confusion Matrix

Thus, the respective accuracy scores and f1 scores (highest value among all classifiers) for the datasets are discussed in below table (table no. 3)

Table 3. Performance analysis of classification

Dataset	Accuracy score	F1 score
Breast Cancer	0.8527	0.8701
Colon Tumor	0.8827	0.9045
MLL	0.9602	0.9732
SRBCT	0.8951	0.9545
Leukemia	0.8554	0.8145

Now, furthermore, for an easy visualization purpose and to better understand the classifier performance; Receiver operating characteristic (ROC) curves were used. It detects the true positives (model has correctly classified the instances) by neglecting the false positives. The area under the curve, AUC value ranges from 0 to 1. A sample AUC-ROC is given below in figure 4.

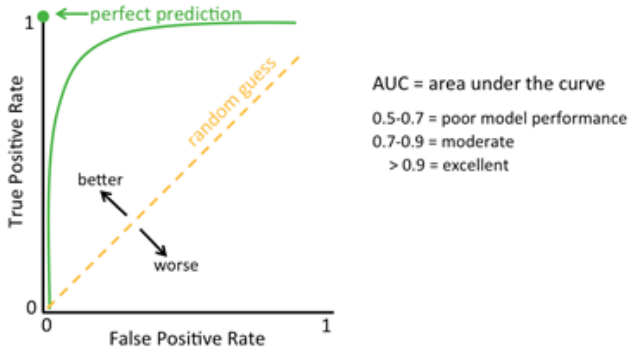

Figure 4. ROC curve

Please note that the classifier can't predict where the AUC value is less than 0.5. For comparison among any given classifiers, the curve can choose the best one. The maximum AUC-ROC score obtained against the corresponding classifier is given in table 4.

Table 4. AUC-ROC score for classifiers

Dataset	AUC-ROC score (max)	Corresponding Classifier
Breast Cancer	0.82	Random Forest
Colon Tumor	0.78	Logistic Regression
MLL	0.90	Decision Tree
SRBCT	0.87	KNN
Leukemia	0.81	Random Forest

4. Conclusions

In this paper, an attempt has been made to enhance the performance of a classification algorithm by effectively performing dimension reduction of the microarray datasets. Both multiclass and binary class classification were performed and accuracy scores were compared. The experimental results achieved were satisfactory for classification tasks.

References

1. Moshood A. Hambali, Tinuke O. Oladele , Kayode S. Adewole. Microarray cancer feature selection: Review, challenges and research directions. International Journal of Cognitive Computing in Engineering (2020), 78-97.
2. Ricardo Ocampo, Marco A. de Luna, Roberto Vega, Gildardo Sanchez-Ante, Luis E. Falcon-Morales, and Humberto Sossa. Pattern Analysis in DNA Microarray Data through PCA-Based Gene Selection. Conference paper: Springer International Publishing Switzerland (2014), 532–539.
3. Amit Bhola and Arvind Kumar Tiwari. MACHINE LEARNING BASED APPROACHES FOR CANCER CLASSIFICATION USING GENE EXPRESSION DATA. Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.3/4, December 2015.
4. I.J. Information Engineering and Electronic Business, 2012, 2, 43-50 Published Online April 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2012.02.07 by Sujata Dash, Bichitrananda Patra and B.K. Tripathy "A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set".
5. Venkatesh and Thangaraj, "Investigation of Micro Array Gene Expression Using Linear Vector Quantization for Cancer", International Journal on Computer Science and Engineering, 2010, Vol. 02, No. 06, pp. 2114-2116.
6. Dev, Jayashree, et al. "A Classification Technique for Microarray Gene Expression Data using PSO-FLANN." International Journal on Computer Science and Engineering 4.9 (2012): 1534.
7. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: Theory and applications. Neurocomputing 70(1-3), 489–501 (2006).
8. Park, D., Jung, E.-Y., Lee, S.-H., Lim, J.: A composite gene selection for dna microarray data analysis. Multimedia Tools and Applications, 1–11 (2013).

9. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F., Mewes, H.W.: Gene selection from microarray data for cancer classification-a machine learning approach. Computational Biology and Chemistry 29(1), 37–46 (2005).
10. Aditi Nautiyal, Amit Kumar Mishra, A Study on Future Trends of Data Mining for Prediction of Cancer, Conference on Recent Innovations in Emerging Technology & Science, April 6-7, 2018.
11. Remeseiro, Beatriz & Bolón-Canedo, Verónica. (2019). A review of feature selection methods in medical applications. Computers in Biology and Medicine.112.103375.10.1016/j.compbiomed.2019.103375.
12. Tsamardinos, I., Aliferis, C.F.: Towards principled feature selection: Relevancy, filters and wrappers. In: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (2003)
13. M. Kumar and S. K. Rath, Microarray Data Classification using Fuzzy *k*-Nearest Neighbor, In: International Conference on Contemporary Computing and Informatics (IC3I), *IEEE*, pp. 1032–1038, (2014).
14. A. T. Islam, B.-S. Jeong, A. G. Bari, C.-G. Lim and S.-H. Jeon, Mapreduce based Parallel Gene Selection Method, *Applied Intelligence*, pp. 1–10, (2014).
15. S. Wang, I. Pandis, D. Johnson, I. Emam, F. Guitton, A. Oehmichen and Y. Guo, Optimising Parallel *r* Correlation Matrix Calculations on Gene Expression Data using Mapreduce, *BMC Bioinformatics*, vol. 15(1), pp. 351, (2014).
16. Y. K. Jain and S. K. Bhandare, Min Max Normalization based data Perturbation Method for Privacy Protection, *International Journal of Computer & Communication Technology (IJCCT)*, vol. 2(8), pp. 45–50, (2011).
17. M. Kumar and S. Kumar Rath, Classification of Microarray Data using Kernel Fuzzy Inference System, International Scholarly Research Notices 2014 (Article ID 769159), pp. 18, (2014).
18. D. Borthakur, The Hadoop Distributed File System: Architecture and Design, Hadoop Project Website, vol. 11, pp. 21, (2007).
19. A. C. Murthy, V. K. Vavilapalli, D. Eadline, J. Niemiec and J. Markham, Apache Hadoop YARN: Moving Beyond MapReduce and Batch Processing with Apache Hadoop 2, Pearson Education, (2013).
20. J. Dean and S. Ghemawat, Mapreduce: Simplified Data Processing on Large Clusters, *Communications of the ACM*, vol. 51(1), pp. 107–113, (2008).

Biographical notes

Ipsita Paul has received M.Tech in CSE from Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar, Orissa. She is an Assistant Professor in department of Computer Science and Engineering, KIIT, Bhubaneswar, Orissa. Her research area includes Machine Learning, Data mining, Deep learning.