MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 6, Issue 1, January-June 2021

Microarray Data Pre-processing for Feature Selection and Classification

Ipsita Paul
Faculty Associate
Department of Computer Science and Engineering
Kalinga Institute of Industrial Technology, Bhubaneswar, Orissa, India
Corresponding Author's Email id: *ipsita.paulfcs@kiit.ac.in*
DOI: https://doi.org/10.47531/JoDMKE.6.1.2021

Abstract

Microarray technology is one of the most challenging areas where researchers study genetics and organisms. Microarray data are a subset of biomedical data that have huge dimensions. The real challenge that such microarray data pose is its enormous features with a very low sample space. This paper presents various pre-processing techniques on microarray data for Feature Selection and Classification. Several microarray datasets were taken for experimental analysis, where both binary and multi-class classification was performed. Extensively large numbers of features were successfully reduced while achieving good classification accuracy.

Keywords*: - Microarray data, data pre-processing, Principal Component Analysis (PCA), Analysis of Variance (ANOVA), Recursive feature elimination (RFE), Feature Selection, Dimension Reduction, Machine Learning Algorithms.*

INTRODUCTION

In 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 this, we need to pre-process the data to apply machine learning algorithms. Analysis and interpretation of DNA Microarray datasets have 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

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

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 classification accuracy by removing a large number of irrelevant genes [7]. Dealing with microarray data is usually difficult since it is structured data that is characterized by very few samples with numerous features. To remove the irrelevant characteristics and finally classifying the dataset pose a 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 the classification of relevant features/genes [16].

In this paper, various data cleaning and 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 classification accuracy by removing a large number of irrelevant features. As a part of result analysis, a 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 before and after pre-processing.

METHODOLOGY

The work is mainly divided into three segments:

1. Pre-processing the data to remove redundant or similar features.
2. Applying proper feature selection/elimination techniques to reduce the number of features.
3. Check the efficiency of the reduced dataset using classification tools.

Various data cleaning techniques have been used, followed by systematic reduction of the data dimension. We have used well-known feature selection techniques to effectively eliminate the irrelevant features.

Pre-processing the data: 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

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

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, the self-variance threshold level was kept as low as 0.01. Correlation is a statistical method used to assess a 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 a threshold value for mutual correlation co-efficient among the features >0.9.

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 the 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 the wrapper, filter and intrinsic sub-sections.

Supervised Feature selection methods first evaluate the relationship between input variable(s) and the target variable and select the most appropriate or remove 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.

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

Table 1

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$	$df_w = k-1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$	$df_b = n-k$	$MSB = \frac{SSB}{df_b}$	
Total	$SST = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$	$df_t = n-1$		

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 above-mentioned formula in **table 1** represents one-way Anova test statistics.

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 give the best result according to a performance metric. RFE is a greedy optimization algorithm that 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**.

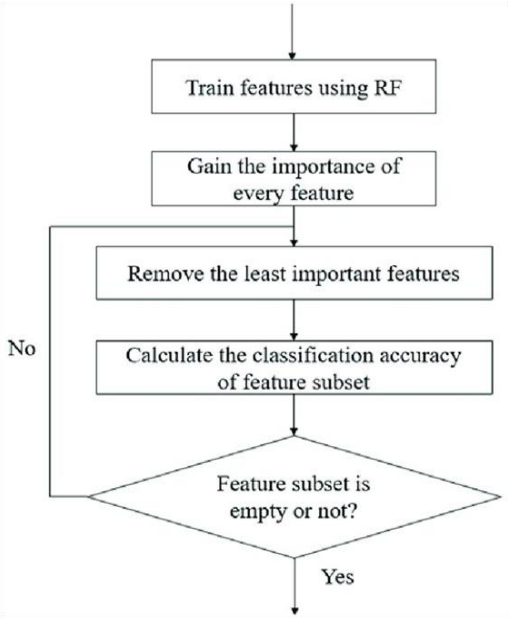


Figure 1

Unsupervised feature selection techniques ignore the target variable, such as methods that remove redundant variables using correlation. PCA is an unsupervised method, which is the most widely used tool in exploratory data analysis and in machine learning for predictive models. It is the process of computing the principal

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

components which creates a visualization of data that minimizes residual variance (in the least square sense) and maximizes the variance of the projection coordinates. PCA has various steps, which are given below in **figure 2**.

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 **W** are the loadings and **C** are the components, **T** at superscript indicates Transpose vector.

Classification Techniques: After performing feature selection, the classification accuracy of the pre-processed datasets was calculated using our two well-known classifiers, namely, Logistic Regression and k-Nearest Neighbour (KNN). The datasets were put into these models, where we got the accuracy and f1 scores. These two 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 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.

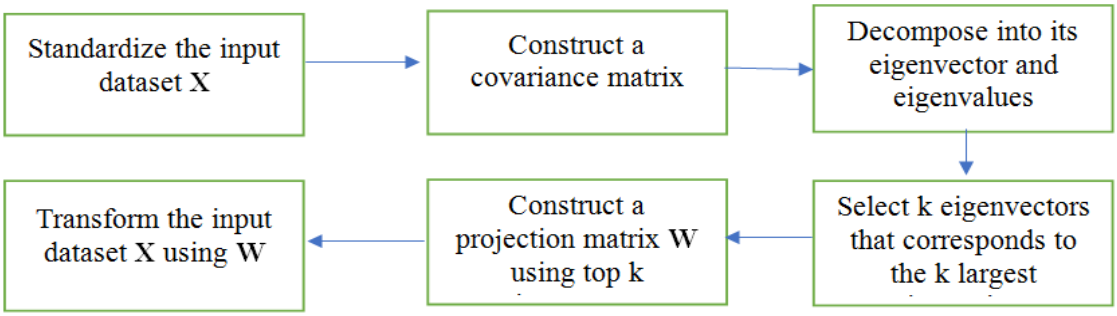


Figure 2

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

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 for feature selection and 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. The irrelevant features were removed after pre-processing of the dataset using certain techniques; the brief study is represented in the below **table 2**.

Certain assumptions were made prior to data pre-processing; as in, the threshold level for selecting features with low self-variance is below 0.01, and the threshold value for mutual co-relation co-efficient among the features is above 0.9. Considering the above mentioned criterions, the results have been generated. The final number of features that was obtained after pre-processing the data has been mentioned in the last column, which shows that how effectively a huge number of features were reduced.

After pre-processing, some well-known feature selection techniques like PCA, ANOVA and RFE were used. So, after applying feature selection, we can again reduce the dimensions of our chosen datasets, the analysis of which is provided below in **table 3**.

Table 2.

	No. of target classes	**Initial No. of features**	**No. of duplicate rows**	**No. of missing values**	**No. of features with low variance**	**No. of highly co-related features**	**Final No. of features after pre-processing**
SRBCT	4	2308	0	0	23	11	2274
Leukemia	3	7129	0	0	0	98	7031
MLL	3	12582	0	0	0	476	12106
Breast Cancer	2	24481	0	0	2132	11509	10840
Colon Tumor	2	2000	0	0	0	765	1235

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

Table 3.

	Initial No. of features	No. of features for PCA	No. of features for ANOVA	No. of features for RFE
SRBCT	2308	352	630	1137
Leukemia	7129	1485	878	3515
MLL	12582	526	1963	6053
Breast Cancer	24481	137	1052	5420
Colon Tumor	2000	406	133	618

Please note that for PCA, the number of selected features vary on each run of the feature selection program. Also, the Accuracy score and F1 score also varies on each run.

Classification performance analysis of the datasets using the Conventional classification models (like Logistic Regression and K-Nearest Neighbour) is shown in **tables 4 and 5**. The accuracy scores and f1 scores for the same are mentioned, which are discussed below.

Accuracy score is the measure of classification accuracy. 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:

$$Accuracy\ Score = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions\ made} = \frac{TP + TN}{TP + TN + FP + FN}$$

The **Accuracy Score** of various datasets after applying the feature selection technique (like PCA, ANOVA and RFE) is mentioned below in **table 4**. The datasets that have been chosen were: Colon Tumor, Breast cancer, MLL, Leukemia and SRBCT.

We also have measured the f1 score along with the Accuracy score. f1 score is a measure of 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. The greater the f1 score, the better is the performance of our classifier model.

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

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}$$

Precision and Recall is found from the confusion matrix (**figure 3**).

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

Table 4.

	SRBCT	Leukemia	MLL	Breast Cancer	Colon Tumor
PCA	0.7826	0.8036	0.8121	0.7883	0.8827
ANOVA	0.8858	0.8554	0.9536	0.8103	0.8153
RFE	0.8951	0.8540	0.9602	0.8527	0.8041

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

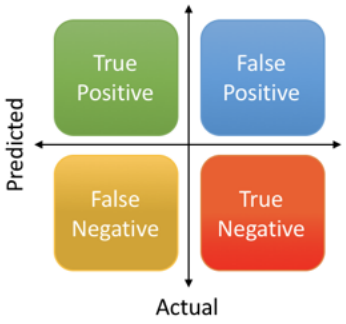


Figure 3

Table 5.

	SRBCT	Leukemia	MLL	Breast Cancer	Colon Tumor
PCA	0.7809	0.8089	0.8404	0.8055	0.9045
ANOVA	0.9545	0.8145	0.9732	0.8534	0.8469
RFE	0.9348	0.8063	0.9679	0.8701	0.8463

F1 score of various datasets after applying the feature selection technique (like PCA, ANOVA and RFE) is mentioned below in table 5. The datasets that have been chosen were: Colon Tumor, Breast cancer, MLL, Leukemia and SRBCT.

CONCLUSION

In this paper, an attempt has been made to enhance the performance of a classification algorithm by effectively performing data pre-processing prior to feature selection. Both multiclass and binary class classification were performed to measure the effectiveness of data pre-processing. 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.

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

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

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved

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, TheHadoop 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).

Page 36-45 © MANTECH PUBLICATIONS 2021. All Rights Reserved