PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

A Seminar Report

On

Data Mining Techniques for Gene Expression Analysis

SUBMITTED BY

NAME: Mrugakshi Chidrawar
ROLL NO: 3152
CLASS: TE-1

GUIDED BY
PROF. M. S. Wakode

COMPUTER ENGINEERING DEPARTMENT

Academic Year: 2018-19

PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

CERTIFICATE

This is to certify that Ms. ***Mrugakshi Chidrawar*** , *R*oll No**.** **3152** a student of T.E. (Computer Engineering Department) Batch 2018-2019, has satisfactorily completed a seminar report on "**Data Mining Techniques for Gene Expression Analysis**" under the guidance of Prof. M. S. Wakode towards the partial fulfillment of the third year Computer Engineering Semester II of Pune University.

Prof. M. S. Wakode
Internal Guide

Dr. R.B.Ingle
Head of Department,
Computer Engineering

Date:

Place:

Abstract:

Classification of gene expression data has been a booming topic in the recent years. This can help develop efficient methods in the field of bioinformatics to be used for cancer diagnosis and further, it's treatment. This seminar gives a critical review of existing data mining techniques being practiced in the field of analysis of gene expression data. Due to the complexity of the underlying biological processes, the mining of gene expression data for the purpose of gene function prediction becomes very difficult. A fuzzy mining technique is explained to overcome these difficulties. This can effectively capture disparateness in expression data for discovery of patterns by transforming quantitative expression values into linguistic terms, such as highly or lowly expressed. It makes use of a fuzzy measure to determine if interesting association patterns exist between the linguistic gene expression levels. It can be used to expose hidden patterns to accurately predict and classify gene functions.

Keywords*: Gene expression, Data mining, Bioinformatics, Classification, DNA, Fuzzy set theory*

Data Mining Techniques for Gene Expression Analysis

Contents

List of Figures

List of Tables

1. INTRODUCTION

One of the important goals in the post-genomic era is to discover the functions of genes. There has been a growing interest in discovering the functions performed by different genes. There are many methodologies for performing gene expression profiling on transcripts, and through their use scientists have been generating vast amounts of experimental data. Turning the raw experimental data into meaningful biological observation requires a number of processing steps: to remove noise, t identify the "true" expression value, normalize the data, compare it to reference data and to extract patterns. The traditional approaches for gene profiling includes DNA Microarray and SAGE. The approach discussed here for analysing gene expressions is the through the utilisation of various data mining techniques to classify gene expressions.

1.1 Motivation and History

The Human Genome Project (HGP) was an international research project, launched in 1990, with the goal of determining the sequence of nucleotide base pairs that make up human DNA, and of identifying and mapping all the genes from a both physical and functional standpoint. This provided the complete set of data for studying gene makeup and gene regulation.

Various new methodologies for identifying and mapping gene functions were discovered.The older methods like RNA blotting, reporter gene, fluorescent in situ hybridisation, etc. were slow and limited with respected to the size of data they could process simultaneously. In particular, microarrays are considered a breakthrough technology in biology, facilitating the quantitative study of thousands of genes simultaneously from a single sample of cells.

The advent of computational methods to perform classification of genes based on gene expression data proved to have achieved amazing success in modeling large-scale data recently. This not only helped in processing large amount of data simultaneously, but it also help increase the accuracy of classifying genes through machine learning.

1.2 Literature Survey:

1.2.1 Supervised and Unsupervised learning

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here, the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

- Clustering is a Machine Learning technique which is used to form groups of data points. Data points that belong to the same group exhibit similar patterns whereas, data points belonging to different groups show highly dissimilar properties.

Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing.

- A classification problem is a supervised learning problem where the output variable is a category, such as “Red” or “blue” or “disease” and “no disease” .

1.3.2 Raw Data Handling

Recent technologies in gene expression analysis have made it possible to simultaneously monitor the expression pattern of thousands of genes. One difficulty in identifying differentially expressed genes is that experimental measurements of expression levels include variations resulting from noise, systematic error and biological variation. To overcome these difficulties, methodologies for normalizing, scaling and difference finding for gene expression data are applied.

A typical approach is as follows:

1) Define noise
2) Perform normalization
3) Adjust data through scaling
4) Compare data from experiments to identify differences
5) Perform Data mining analysis

1.3.3 Expression Data Sources

- One of the most comprehensive collection of gene expression microarray data can be found at Gene Expression Omnibus (http://www.ncbi.nih.gov/geo/) and is maintained by National Center for Biotechnology Information (NCBI).
- SAGANET (http://saganet.org/resources/data.htm) is a central site for collecting and organizing SAGE data on cancer tissues.

1.3.4 Fuzzy c-means clustering

In hard clustering methods like hierarchical and k-means clustering techniques, data is divided into distinct clusters, where each data element belongs to exactly one cluster, and so the outcome of the clustering may be incorrect, many times. The problems occuring in hard clustering methods could be solved by the fuzzy clustering technique. Among fuzzy based clustering, Fuzzy C-Means (FCM) was the most suitable for microarray gene expression data.

Fuzzy logic deals with the uncertainties arising from noisy and inexact data, which are quite common place in expression data and also to make the patterns discovered easily interpretable by human users.

This algorithm works by assigning membership to each data point corresponding to each cluster center based on the distance between the cluster center and the data point. The nearer the data point is to the cluster center, more is its membership towards the particular cluster. Clearly, summation of membership of each data point should be equal to one.

1.4 Applications

The most common use of gene expression analysis is to compare expression levels of one or more genes from different samples.

Some of the common comparisons include:

- Normal vs Disease
- Mutant vs Wild-type
- Before and After treatment

As access to high throughput technology improves and data analysis programs continue to manage the ever growing gene expression data, the applications of gene expression analysis can only improve further.

1.5 Challenges

1.5.1 Needs large dataset

As every deep learning problem this problem also requires large amount of dataset for more accuracy. But fortunately gene expression data is publicly available on various websites and can be used as training data.

1.5.2 Requires preprocessing of data

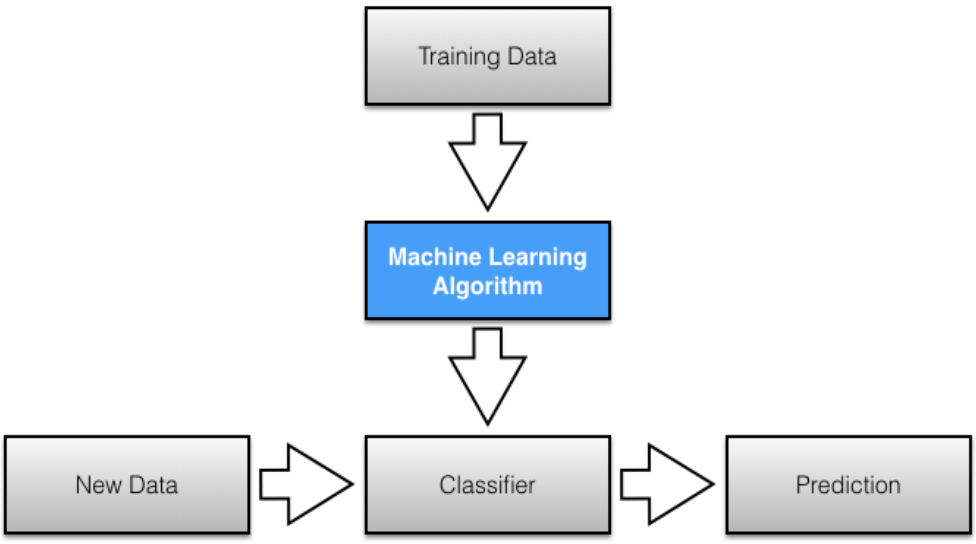
The preprocessing of gene expression data, before carrying out a supervised classification study, is required. Sometimes, due to technological problems or mishandling of the microarrays, expression values for some genes cannot accurately be measured, in which case the problem of missing data arises. Classification methods do not generally have the capability or provision to handle missing data. Therefore, the missing values need to be filled in or imputed with some reasonable estimates before proceeding with the classification study. Not doing so results in discarding genes (rows) with missing entries in the gene expression data matrix, therefore precious training data may be seriously reduced and consequently its ability to represent the investigated biological situations/conditions may diminish.

Another key preprocessing step is the normalization or the method by which expression levels are made comparable.

1.5.3 Needs Good GPUs for training

Again, as every deep learning problem this problems also requires good quality GPUs for training of model large gene expression dataset.

2. CLASSIFICATION TECHNIQUES FOR GENE EXPRESSION ANALYSIS



2.1 k-NN

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition since the 1970's.

A data point is classified by a majority vote of its neighbors, with the data point being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

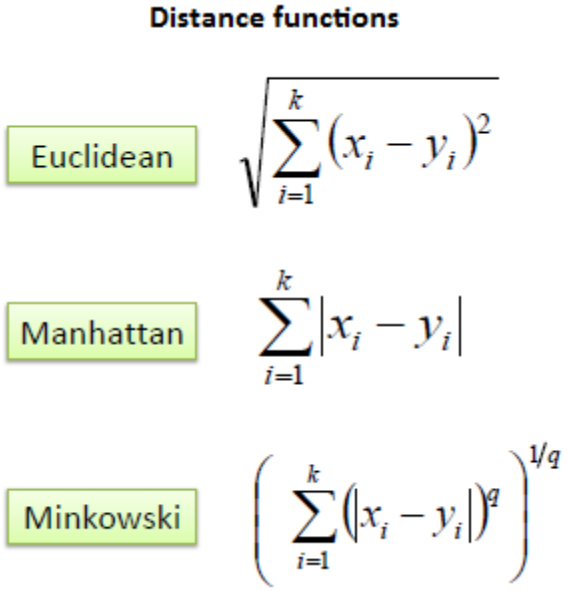


Figure 1: Various types of Distances in k-NN

2.2 SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space, this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

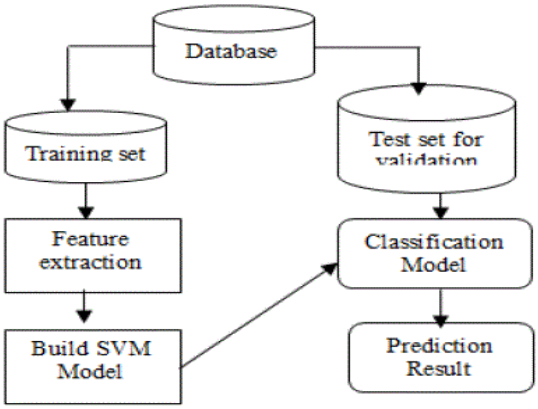


Fig 2.Testing and training of images

Figure 2: Flowchart of SVM

2.3 Decision Tree

Decision Tree learning algorithm generates decision trees from the training data to solve classification and regression problem. In Decision Tree algorithm, the best attribute means the attribute which has the most information gain.

Algorithm:

1) Start with a training data set which we'll call S. It should have attributes and classification.
2) Determine the best attribute in the dataset.
3) Split S into subset that contains the possible values for the best attribute.
4) Make decision tree node that contains the best attribute.
5) Recursively generate new decision trees by using the subset of data created from step 3 until a stage is reached where you cannot classify the data further. Represent the class as leaf node.

2.4 Random Forest

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random Forest adds additional randomness to the model, while growing

the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

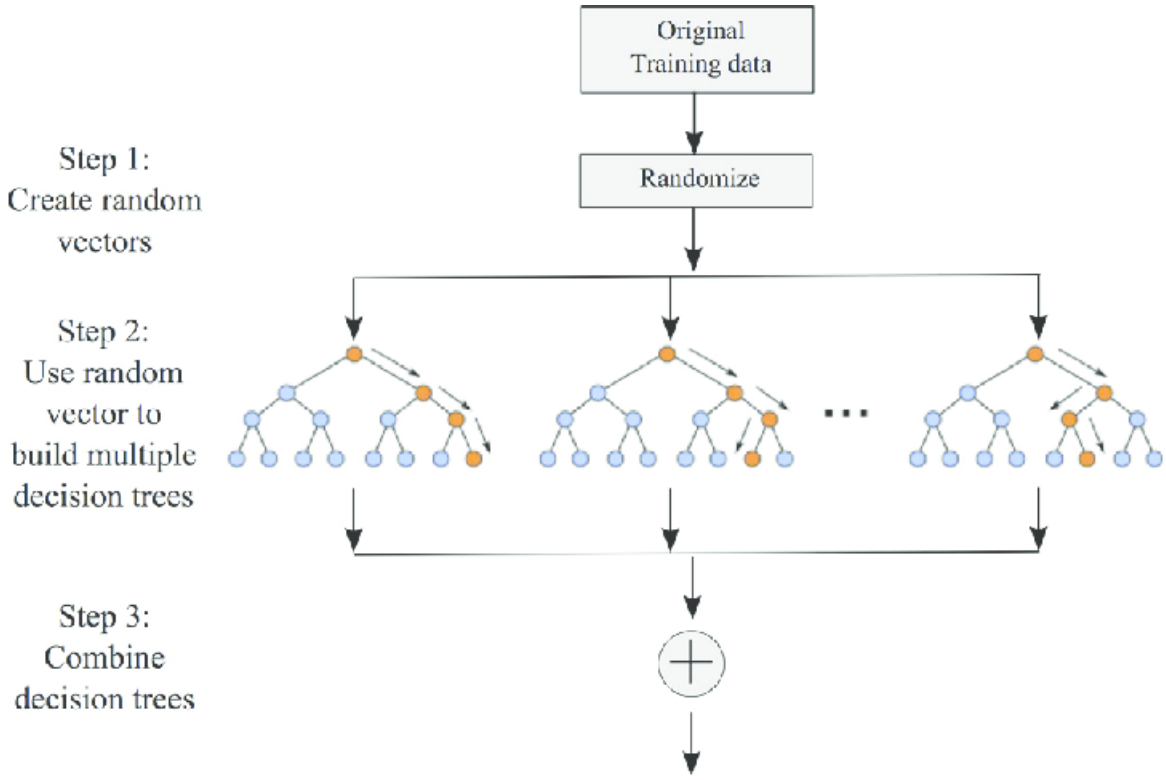


Figure 3: Demonstration of Random Forest Classifier

2.5 Fuzzy C-means

FCM adopts fuzzy partitions to make each given value of data input between 0 and 1 in order to determine the degree of its belonging to a group. It allows a piece of data to belong to two or more clusters.

In the context of gene expression analysis, it helps depict a more natural representation of genes, since genes are usually involved in multiple functions.

Let xi be a vector of values for data point gi.

1) Initialize membership U(0) = [uij] for data point gi of cluster clj by random

2) At the k-th step, compute the fuzzy centroid C(k) = [cj] for j = 1, .., nc, where nc is the number of clusters, using

$$c_j = \frac{\sum_{i=1}^{n} (u_{ij})^m x_i}{\sum_{i=1}^{n} (u_{ij})^m}$$

,where m is the fuzzy parameter and n is the number of data points.

1) Update the fuzzy membership U(k) = [uij], using

$$u_{ij} = \frac{\left(\frac{1}{\|x_i - c_j\|} \right)^{\frac{1}{(m-1)}}}{\sum_{j=1}^{n_c} \left(\frac{1}{\|x_i - c_j\|} \right)^{\frac{1}{(m-1)}}}$$

2) If | | U(k) – U(k-1)| | < e , then STOP, else return to step 2.
3) Determine membership cutoff:

- For each data point gi, assign gi to cluster clj if uij of U(k) > a

3. DISCUSSION ON IMPLEMENTATION RESULTS

The k-NN model, SVC (Support Vector Classifier) model, Decision Tree and Random Forest classifiers were implemented.

The following Table shows the accuracy scores of the 4 implemented classification model.

CLASSIFICATION MODEL	ACCURACY SCORE
k-NN	0.5
SVC	0.42
Decision Tree	0.33
Random Forest	0.458

Table 1: Comparison of Accuracy scores of different classification models

The following are the prediction error graphs for the above mentioned classification models.

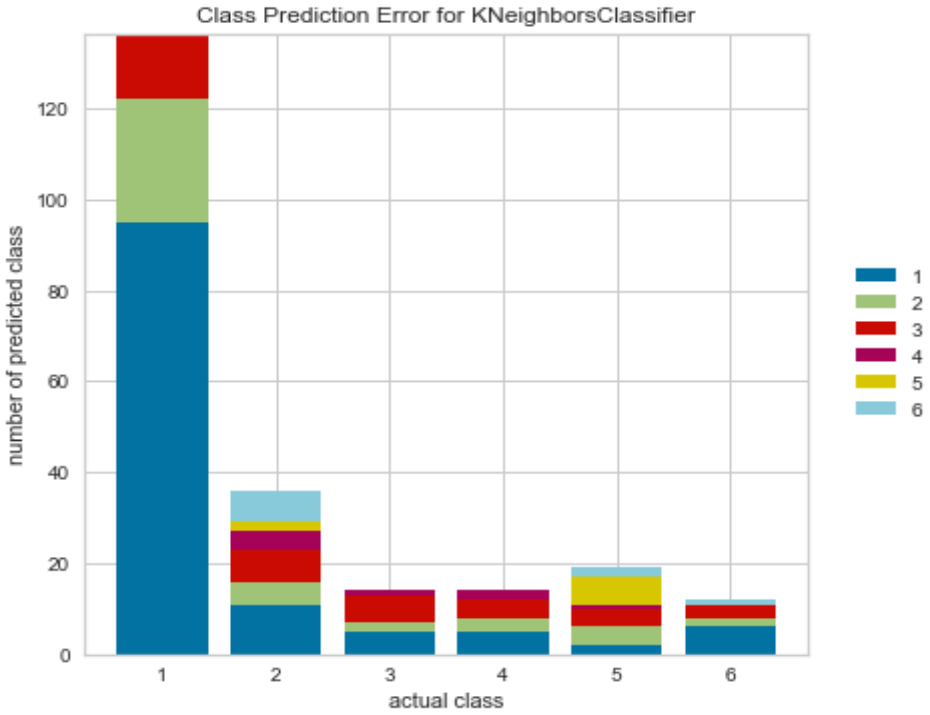


Figure 4: Class prediction error for KNeighborsClassifier

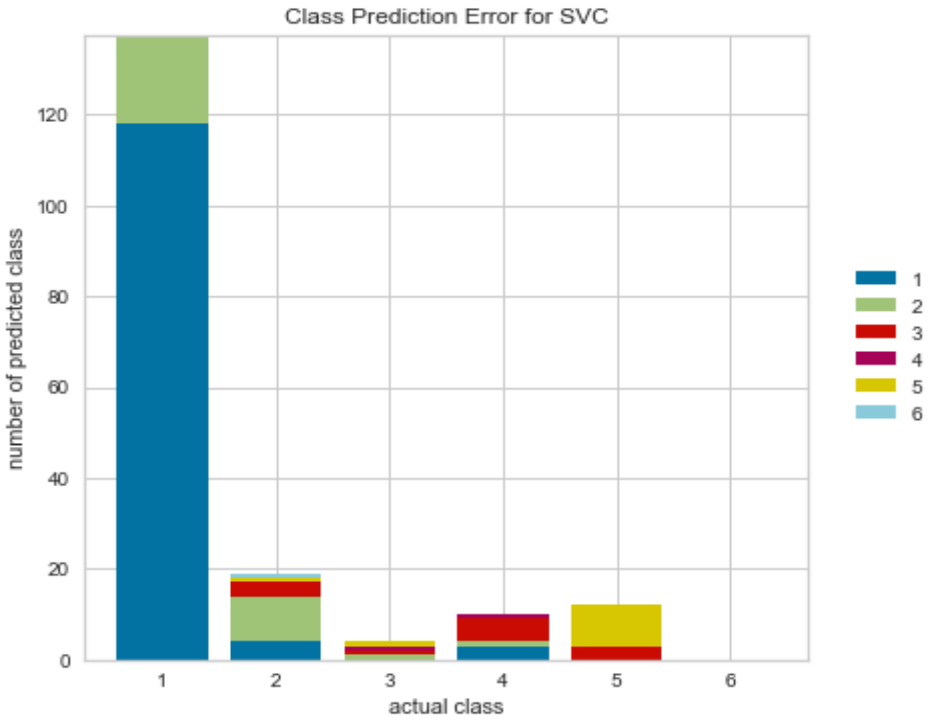


Figure 5: Class prediction error for SVC

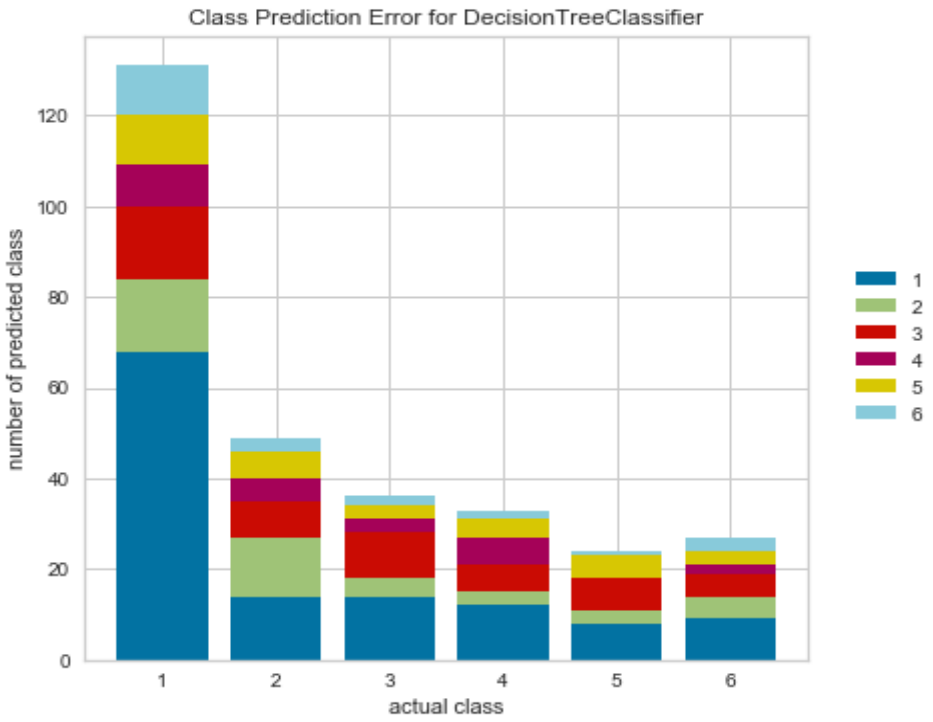


Figure 6: Class prediction error for DecisionTreeClassifier

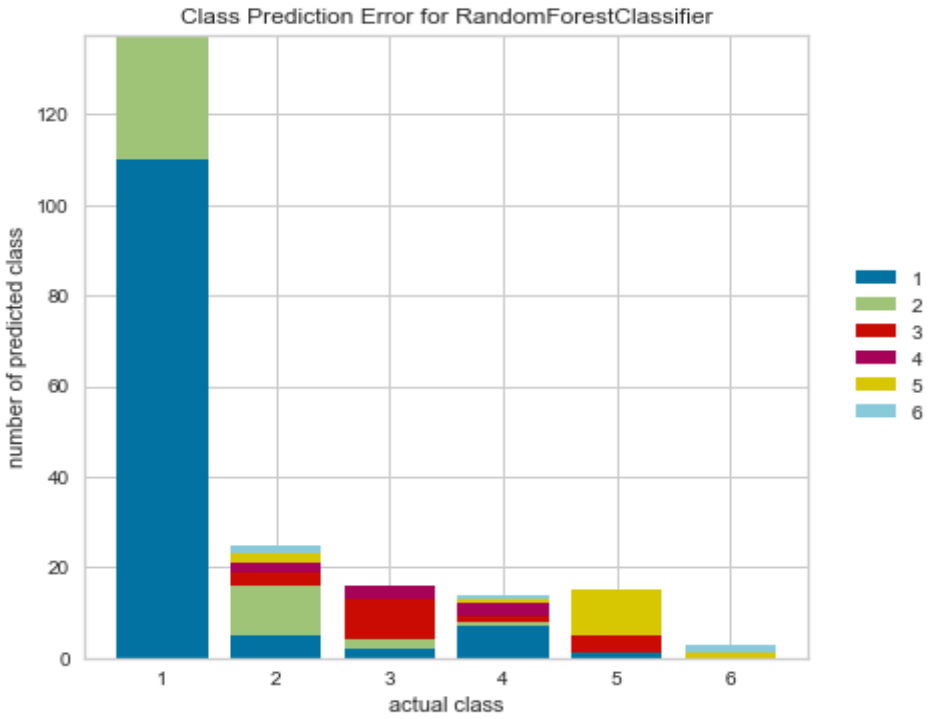


Figure 7: Class prediction error for RandomForestClassifier

4. CONCLUSION AND FUTURE ENHANCEMENTS

4.1 Conclusion

Following this comparison, we observed among numerous studies that various data mining techniques are currently in use in order to develop an appropriate methodology for effective gene expression data classification.

We can also conclude that fuzzy c-means clustering algorithm will classify genes more naturally into 2 or more clusters and therefore increase the accuracy of gene function prediction.

4.2 Future Enhancements

In the bioinformatics domain, existing methodologies for gene expression analysis are an ongoing topic focusing on disease diagnosis. Addressing this intrinsic problem, various approaches have been developed.

- Training on large dataset can improve accuracy and model can be generalized.
- Further investigation in term of gene data optimization and elimination of all redundant data can improve performance.

References

[1] P. C. H. Ma and K. C. C. Chan, "Incremental Fuzzy Mining of Gene Expression Data for Gene Function Prediction," in IEEE Transactions on Biomedical Engineering, vol. 58, no. 5, pp. 1246-1252, May 2011. doi: 10.1109/TBME.2010.2047724

[2] M. Aouf, L. Liyanage and S. Hansen, "Critical Review of Data Mining Techniques for Gene Expression Analysis," 2008 4th International Conference on Information and Automation for Sustainability, Colombo, 2008, pp. 367-371. doi: 10.1109/ICIAFS.2008.4783954

[3] Edgar, Ron et al. “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository” Nucleic acids research vol. 30,1 (2002): 207-10.

[4] Panda S., Sahu S., Jena P., Chattopadhyay S. (2012) Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study. In: Wyld D., Zizka J., Nagamalai D. (eds) Advances in Computer Science, Engineering & Applications. Advances in Intelligent and Soft Computing, vol 166. Springer, Berlin, Heidelberg

[5] Androulakis IP, Yang E, Almon RR. Analysis of time-series gene expression data: methods, challenges, and opportunities. Annu Rev Biomed Eng. 2007;9:205-28. PubMed PMID: 17341157; PubMed Central PMCID: PMC4181347.

[6] L. Tari, C. Baral, and S. Kim, “Fuzzy c-means clustering with prior biological knowledge,” J. Biomed. Inf., vol. 42, no. 1, pp. 74–81, 2009.

[7] Chattopadhyay, Subhagata & amp; Pratihar, Dilip & amp; Sarkar, Sukanta. (2011). A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms. Computing and Informatics. 30. 701-720.

[8] Michael T. Zimmermann, “The Importance of Biologic Knowledge and Gene Expression Context for Genomic Data Interpretation” , (2018) Frontiers in Genetics. 9.670. 10.3389/fgene.2018.00670.

[9] Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., & Altschuler, S. J. (2002). “Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters” . Nature Genetics, 31(3), 255–265. doi:10.1038/ng906

APPENDIX – D

Log Book

Roll No. **:- 3152**

Name of the Student **:- Mrugakshi Chidrawar**

Name of the Guide **:- Prof.M.S.Wakode**

Seminar Title **:- Data Mining Techniques for Gene Expression Analysis**

Sr. No.	Date	Details of Discussion/ Remarks	Signature of guide / Seminar Incharge
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			
9.			
10.			

Student Signature **Guide Signature**