

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Paper / Case Study

Available online at: www.ijarcsms.com

A Survey Analysis on Gene Expression based Cancer in Medical Mining

R. Ramapraha¹SNS College of Engineering
Coimbatore - India**V. Gowri²**Assistant Professor
Department of CSE
SNS College of Engineering
Coimbatore - India

Abstract: *Cancer is a diseases caused by uncontrolled division of cell growth in human part of the body. Cells divided and growth as tumor and spread it in near the parts of the body. A tumor is a neoplasm, abnormal mass of tissue that may be solid or fluid-filled. Microarray technology is used to study expression many gene at once. Microarray technology used to identification of new gene. Microarray leads to incomplete coverage that lead to false normal results and the ability to test only unbalanced rearrangements. Gene expression profiling is the query expression of thousands of genes simultaneously is used to classify a tumor. Supervised classifier is used to classify only training data prediction is difficult in supervised classifier. Semi supervised classifier is work with both labeled and unlabeled data in gene expression. Transductive support vector machine (TSVM) is used to identify the gene marker and improve the accuracy of classifier. TSVM is applied in low density separation and find the objective function. Gene selection method is an integral preprocessing step in cancer.*

Keywords: *cancer, microarray technology, SVM, TSVM.*

I. INTRODUCTION

Cancer is one of the main area in medical field. The identification of separate genes is of fundamental and practical interest. Research medicine top ranking gene recent discoveries in cancer research to be explored. Gene prediction or gene finding refers to the method of identifying the regions of genomic DNA that encode genes. An mRNA sequence, it is inconsequential to derive a unique genomic DNA sequence from which it had to have been transcribed into protein. Identification of gene is a major problem biometrics. Microarray technology is used to identify and simultaneous monitor thousands of genes and huge amount of molecular information is used to extract to find common pattern with group of samples. Gene is located in chromosomes and has a different number of chromosomes. Chromosomes found in nucleus of a cell.

K-Mean clustering is one of the simplest unsupervised learning algorithm that solved cluster problem. k-means clustering aims to partition n interpretation into k clusters in which each observation belongs to the cluster with the nearest mean allocation as a prototype of the cluster. K-mean is simple and easy to classify the data set through certain number of cluster with fixed priori.

Support vector machine is an supervised learning model with learning algorithm.SVM is to analyze data and used for classification In SVM we give set of input data and predict the input and produce two output classes and make it as binary linear classification. SVM is efficient in non-linear classification.

Semi supervised learning is class of a machine learning that is used for both labelled and unlabeled data. Generate the learning and estimate distributed data points to each class. More normal learning problems may also be viewed as instances of semi-supervised learning. Semi supervised learning have some feature model are: It Can be used for classification and regression tasks, Kernel methods to project data into alternate dimensional spaces.

Feature selection is the process of selecting a subset of relevant features for use in model construction. Technique is that the data contains many irrelevant features. The use of feature selection in analysing DNA microarrays has many thousands of features and a few have hundreds of samples. Benefits of feature selection improved model interpretability.

II. LITERATURE REVIEW

2.1 K- Mean Clustering

The main objective of K- Mean Clustering is group the object with similar one and group the dissimilar object in to different cluster. It classify pre defined number of cluster which given by user. K- mean always have K cluster [1]. Cluster is non- hierarchical and they do not have overlap [4]. K- mean clustering have an arbitrarily choose number of desired cluster data point from set of data point as an initial centroids[10]. Centroid are a two dimensional region.

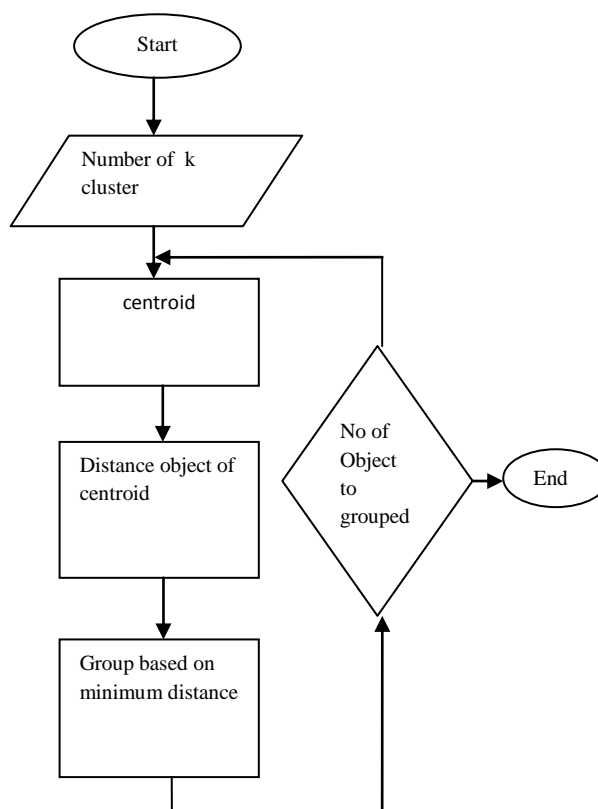


Fig : K- Mean Clustering

The drawback of K- Mean algorithm is quality of final clustering is highly depend upon arbitrarily of initial centroid. K- Mean Cluster only works with spherical sphere shape.

2.2 Support Vector Machine (SVM)

Support Vector Machine, a supervised learning technique it perform well in biological area of analysis and evaluating microarray expression data. SVM is correctly separate entities in appropriate classes and also identify instances which established classification is not support by the data [1]. SVM is not unique among the classification but it show efficient. SVM is work with high dimensional data. SVM found dimensionality reduction and radically improve classification performance [5]. When we use SVM classification they separate set of binary labeled trained data with hyper plane with maximally distant from them. A major goal of SVM is improve speed in training and tested data [9].SVM is more feasible option for large data set .A main drawback of SVM is it extremely slow.

2.3 Quick Branch and Bound(QBB)

QBB is have two phase the first phase is LVF (Las Vegas Filter) and second phase is ABB (Automatic branch and bound). Branch and bound is consisting of regular list of candidate solution where large subset of candidate are discard by using upper and lower bound of quantity begin optimized[6]. Feature selection is efficient technique in dimensionality reduction and used to find an optimal solution of relevant feature the overall accuracy is increase while data size is reduced. LVF is a randomized algorithm that always gives correct result but sometime no answer. LVF is adopting inconsistency rate as evaluation measure. The work focus on inconsistent measure which a feature subset is inconsistent if there exists at least two instances with same feature value with different class label. ABB is a complete search the work focus and start with full set of feature and remove one feature at time. LVF is large size when ABB perform will under in time constrain [7]. LVF is small size when ABB perform may not able to reduce feature. ABB take longer time to find minimal size of feature subset. QBB is more efficient both in average time and number of selected feature although it very fast and accurate.

2.4 Induction Method

Induction method is directly deal with problem of attribute selection and especially ones that operate on logical expression. Logical expression determine the path of program logic that takes conditional looping statement[2]. Eliminate the irrelevant one is a central problem in machine learning before induction method moved with training data. In Induction method they used filter method and wrapper method to find relevant feature. In filter method for relevant gene they filter out irrelevant gene before induction process. In filter method focus the algorithm start with empty feature set and carried out Breath First search until find minimal combination feature to predict pure class. Filter method is usually fast. Wrapper method have forward selection and backward elimination .forward selection start with empty feature set and add feature at each step and backward elimination start with full feature set and discard feature at each step. Wrapper method generate candidate set and run induction algorithm on the training data. It have better selection accuracy than separate measure select among the alternative feature. The main drawbacks of Induction method is computational cost and less optimistic.

2.5 Low Density Separation(LDS)

Low Density Separation is derived graph based distance the graph derived from data such that nodes are the data points and edges are placed between nodes that are nearest neighbour distance. Many semi supervised operate with nearest neighbour graph they usually they do not require data point. The output function varies between connected nodes [3]. Optimized TSVM is an objective function is perfect semi supervised algorithm is powerful regularization and directly implement in cluster. LDS is more accurate[8]. The main drawbacks of objective function is non convex that is difficult to minimize. Searching low density boundary is difficult the task of TSVM algorithm can be used by changing the data representation.

III. CONCLUSION

In this paper we discuss algorithm, K- mean clustering algorithm is minimize the sum of square distance between two point and the cluster center.SVM always separate hyper planes. SVM is evaluating with microarray data and correctly separate the class and mapping data to high dimensional feature space.SVM improve the accuracy of the classifier. QBB algorithm is more robust while set cover. QBB is used find optimal solution of various optimization problems. Induction method is deal with attribute selection and starting point of space which rotate determine the direct search and use stepwise selection or elimination is consider for both add and remove feature at each decision point. Low Density Separation is derived graph based distance and searching Low density boundary is difficult and it improve the computational efficiency and minimized the test error. Finally, improve the accuracy of gene selection and get better performance of individual correlation.

References

1. Bennett K.P and Demiriz .A, “Semi-supervised support vector machines, in Proc. Adv. Neural Inform. Process Syst., 1998, vol. 10, pp. 368–374.
2. Blum. A and Langley. P, “Selection of relevant features and examples in machine learning,” Artif. Intell., vol. 97, no. 1/2, pp. 245–271, 1997
3. Chapelle. O and Zien. A, “Semi-supervised classification by low-density separation,” in Proc. 10th Int. Works. Artif. Intell. Stat., 2005, pp. 57–64
4. Chandrasekhar. T, Thagavel. K and Elayaraja. E, “ Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data”, IJCS,vol 8,Issue 6, No 3,Nov 2011
5. Cristianini,N. and Shawe-Taylor,J. (2000) An Introduction Support Vector Machine . Cambridge University Press, Cambridge, www.support-vector.net
6. Dash. M, Liu .H, Feature selection methods for classification, Intelligent Data Analysis: An Internat. J. 1 (3)(1997).
7. Dash.M and Liu. H, “Consistency based search in feature selection, Artif.Intell” vol. 151, pp. 155–176, 2003
8. Soucy. P, Mineau. G. W, A simple feature selection method for text classification, in: Proceedings of IJCAI-01, Seattle, WA, 2001, pp. 897–903
9. Terrence s. Furey, nello Cristianini, Nigel Duff, David W. Bednarski, Michel Schummer and David Haussler, “Support Vector Machine classification of cancer tissue samples using microarray expression data,Bioinformatics vol 16 ,no 10,2000
10. Zhang,L., Zhou.W., Velculescu.V., Kern,S., Hruban,R., Hamilton,S., Vogelstein,B. And Kinzler,K. (1997) Gene expression Profiles in normal and cancer ,Science,276,1268-1272