# Analysis of Microarray Gene Expression Data using KNN Classification

[1]R. Vengateshkumar, [2]S. Alagukumar, [3]R. Lawrance
*[1]Research Scholar, [2]Assistant Professor, [3]Director*
*[1]Research & Development centre,*
*Bharathiar University,Coimbatore, Tamil Nadu, India.*
*[2,3]Department of Computer Applications,*
*Ayya Nadar Janaki Ammal College,Sivakasi, Tamil Nadu, India*
*vengatesh.kumar@gmail.com,alagukumarmca@gmail.com,lawrancer@yahoo.com*

**Abstract -** Classification techniques play the vital role in the computational biology. Classifications of data have been successfully applied in various applications. One of the major challenges in microarray gene expression data analysis is the prediction of prognosis, especially in cancer gene expression profiles is to determine the genes. Supervised machine learning techniques are used with microarray datasets to build classification models that improve the diagnostic of cancer diseases. Classification is a promising approach in data mining to construct classification systems on microarray gene expression data. In this paper the KNN classification technique used to classifying and predicting the microarray gene expression data.

**Keywords** - Data mining, KNN Classification, Gene Expression Data.

## 1.INTRODUCTION

Cancer research is an interesting research area in the field of medicine. Classification is momentously necessary for cancer diagnosis and treatment. Microarray technologies provide a powerful tool by which the expression patterns of thousands of genes can be monitored simultaneously whose application range from cancer diagnosis to drug response[1,2,3]. Gene expression is the conversion of the DNA sequences into mRNA sequences by transcription then translated into amino acid sequences called proteins. The expression level is associated with the corresponding protein made under different conditions. Microarray experiments produced large volume of data. Microarray data presents the main challenge that is high density of data. The data collected from a microarray experiments is commonly in the form of an M x N matrix of expression level, where M represents columns(genes) and N represents rows(samples)[14,15]. Classification aims to define an abstract model of a set of classes, called classifier, which is built from the training data set. The classifier is then used to appropriately classify new unknown class label gene expression data. Different approaches have been applied to build accurate classifiers, such as, decision tree, naïve Bayesian classification, random forest and support vector machine [4,5,6,7].

The structure of the paper is organized as follows. Section 2 reviewed literature the previous works in this field. Section3 reviewed methodology of the classification techniques on microarray gene expression data profiling. In section 4 it has been presented that the experimental results and discussion. Finally, Conclusion and Future work are explained in section 5.

## 2.REVIEW OF THE LITERATURE

Data mining techniques have become a popular research tool for biological data to identify and exploit patterns and relationships among large number of variables, and to predict the outcome of a disease using the historical datasets [4]. Some of the classification methods have been reviewed in the related literature.

Pique-Regi, R.,*et al.*, [8] have proposed a sequential Diagonal linear discriminant analysis technique that combines attribute selection and classification. Each iteration, one gene is sequentially added and the linear discriminant recomputed using diagonal covariance matrix.

Arevalillo, Jorge M.*et al.*, [9] have proposed an approach using uses the quadratic discriminant analysis for identifying weak marginal/strong bivariate interactions and method is applied both to synthetic data and to a public domain microarray data. When applied to gene expression data, it leads to pairs of genes which are not univariant differentially expressed but exhibit subtle patterns of bivariate.

Ye, J., *et al.*, [11] have proposed a dimension reduction and feature extraction scheme, called Uncorrelated Linear Discriminant Analysis on gene expression data. ULDA employs the Generalized Singular Value Decomposition method to handle the data and the features that it produces in the transformed space are uncorrelated, which makes it attractive for gene expression data.

Huang, D., *et al.*, [12] have compared classification performance of linear discriminant analysis and its modification methods was evaluated by applying linear discriminant analysis, shrinkage centroid regularized discriminant analysis , shrinkage linear discriminant analysis and shrinkage diagonal discriminant analysis on cancer gene expression data.

## 3. METHODOLOGY

Classification technique is a vital role in microarray experiments, for purposes of classifying biological samples and prediction using microarray gene expression data. The microarray gene expression dataset can be form of an M x N matrix D of expression values, where the row represents samples X={ x1,x2,x3… xn }and column represents genes G={g1,g2,g3…, gn}, category column represents the actual class of the sample, An illustration of microarray breast cancer 2 gene expression data shown in Table 1. The gene expression data, usually contains large amount of data [1,2], therefore data mining techniques are used to extract useful knowledge.

Table 1. The gene expression data

| **Samples** | **Attributes(genes)** | | | |
|---|---|---|---|---|
| | **Gene 1** | **Gene2** | **…** | **Class** |
| 1 | G(1,1) | G(1,2) | … | |
| 2 | G(2,1) | G(2,2) | … | cancer |
| 3 | G(3,1) | G(3,2) | … | cancer |
| … | … | … | … | cancer |
| … | … | … | … | cancer |
| n | G(n,1) | G(n,2) | … | normal |

### A. *Classification Techniques*

Classification is a data mining techniques which assigns an object to one of several predefined categories based on the attributes of the object [13]. The input dataset termed as the training data set, which contains the number of predefined labels each having a number of attributes. The attributes are either continuous or categorical. The main aims to use the training data set to build a model, that model can be used to classify unknown label data set [13]. Figure 2 represent the KNN classification algorithm.

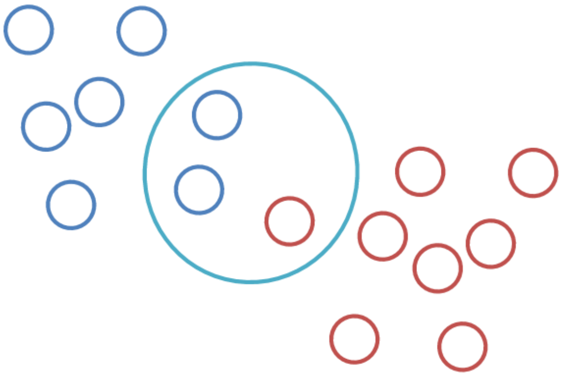


Fig. 2. K-Nearest Neighbor algorithm example

### B. *KNN Classification*

K-nearest neighbor classifier is one of the introductory supervised classifier, which every data science learner should be aware of. Fix & Hodges [10] proposed K-nearest neighbor classifier algorithm in the year of 1951 for performing pattern classification task. This classifier is called as KNN Classifier. The simple version of the K-nearest neighbor classifier algorithms is to predict the target label by finding the nearest neighbor class. The closest class will be identified using the distance measures like Euclidean distance.

### C. *K-nearest neighbor (KNN) algorithm*

Let ($X_i$, $C_i$) where i = 1, 2……., n be data points. $X_i$ denotes feature values & $C_i$ denotes labels for $X_i$ for each i.
Assuming the number of classes as ‘c’ $_{Ci} \in \{1, 2, 3, \ldots\ldots, c\}$ for all values of i
Let x be a point for which label is not known, and we would like to find the label class using k-nearest neighbor algorithms.

### D. *KNN Algorithm Pseudocode*

```
a.  Calculate "d(x, x_i)" i =1, 2, ….., n; where d denotes the Euclidean distance between the points.
b.  Arrange the calculated n Euclidean distances in non-decreasing order.
c.  Let k be a +ve integer, take the first k distances from this sorted list.
d.  Find those k-points corresponding to these k-distances.
e.  Let k_i denotes the number of points belonging to the i^th class among k points i.e. k ≥ 0
f.  If k_i >k_j ∀ i ≠ j then put x in class i.
```

## 4. RESULT AND DISCUSSTION

In this section it has been presented the experimental result of KNN classification on gene expression. The sample microarray gene expression data related to breast cancer 2 dataset has been applied on KNN classification model. Breast cancer 2 data set downloaded from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1379. Data set is available in online public [3].

In this paper, the 17 samples and 5 gene expression conditions are utilized for analyzing KNN classification and result has been obtained. Table 2 represents the gene expression data with class label.

Table 2: Microarray Data with Class Label

| | G1 | G2 | .. | G5 | Label |
|---|---|---|---|---|---|
| S1 | 1 | 6.51 | .. | 1.56 | cancer |
| S2 | 3.01 | 5.96 | .. | 1.91 | normal |
| S3 | 0.68 | 8.6 | .. | 5.33 | cancer |
| S4 | 1.54 | 6.03 | .. | 2.43 | cancer |
| S5 | 2.26 | 6.97 | .. | 5.27 | cancer |
| S6 | 0.16 | 7.52 | .. | 3.57 | cancer |
| S7 | 0.85 | 7.95 | .. | 2.91 | cancer |
| S8 | -1.09 | 7.17 | .. | 2.45 | cancer |
| S9 | 0.73 | 7.27 | .. | 2.07 | normal |
| S10 | 0.55 | 6.51 | .. | 2.58 | normal |
| S11 | 1.05 | 7.68 | .. | 3.62 | cancer |
| S12 | 2.51 | 7.19 | .. | 3.89 | normal |
| S13 | 2.04 | 6.22 | .. | 3.51 | normal |
| S14 | 1.36 | 8.52 | .. | 3.72 | cancer |
| S15 | 0.9 | 7.49 | .. | 0.73 | normal |
| S16 | 3.38 | 6.66 | .. | 1.4 | normal |
| S17 | -1.82 | 8.58 | | 2.59 | cancer |

Table 3: Training Data

Table 3 represents the training gene expression data. This is used to generate the classifier model.

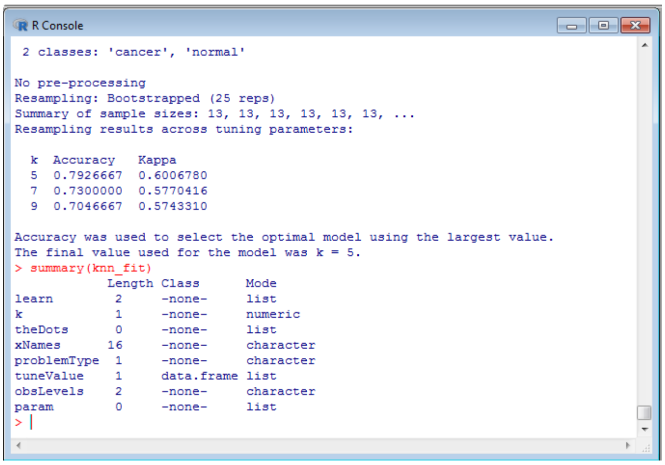
| | G1 | G2 | .. | G5 | Label |
|---|---|---|---|---|---|
| S2 | 3.01 | 5.96 | .. | 1.91 | normal |
| S4 | 1.54 | 6.03 | .. | 2.43 | cancer |
| S6 | 0.16 | 7.52 | .. | 3.57 | cancer |

| | | | | | |
|---|---|---|---|---|---|
| S7 | 0.85 | 7.95 | .. | 2.91 | cancer |
| S8 | -1.09 | 7.17 | .. | 2.45 | cancer |
| S9 | 0.73 | 7.27 | .. | 2.07 | normal |
| S10 | 0.55 | 6.51 | .. | 2.58 | normal |
| S11 | 1.05 | 7.68 | .. | 3.62 | cancer |
| S12 | 2.51 | 7.19 | .. | 3.89 | normal |
| S13 | 2.04 | 6.22 | .. | 3.51 | normal |
| S15 | 0.9 | 7.49 | .. | 0.73 | normal |
| S16 | 3.38 | 6.66 | .. | 1.4 | normal |
| S17 | -1.82 | 8.58 | .. | 2.59 | cancer |

Table 4 represents the test gene expression data. This is used to generate the evaluate classifier model.

Table 4: Test Data

| | G1 | G2 | .. | G5 | Label |
|---|---|---|---|---|---|
| S1 | 1 | 6.51 | .. | 1.56 | cancer |
| S3 | 0.68 | 8.6 | .. | 5.33 | cancer |
| S5 | 2.26 | 6.97 | .. | 5.27 | cancer |
| S14 | 1.36 | 8.52 | .. | 3.72 | cancer |



```
 2 classes: 'cancer', 'normal'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 13, 13, 13, 13, 13, 13, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.7926667  0.6006780
  7  0.7300000  0.5770416
  9  0.7046667  0.5743310

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
> summary(knn_fit)
            Length Class      Mode
learn        2     -none-     list
k            1     -none-     numeric
theDots      0     -none-     list
xNames      16     -none-     character
problemType  1     -none-     character
tuneValue    1     data.frame list
obsLevels    2     -none-     character
param        0     -none-     list
> |
```

Fig. 2. Accuracy of K-Nearest Neighbor algorithm

Figure 2 represent the accuracy of KNN classifier model. It's showing Accuracy metrics result for different k value. From the results, it automatically selects best k-value. Here, KNN training model is choosing k = 9 as its final value. When k value 5, the classifier model gives the accuracy is 79%, when k value is 7, then the accuracy is 73% and finally, when the k value is 9, then the accuracy is 70%. The figure 3 depicts the selecting the k value by the KNN classification algorithm.

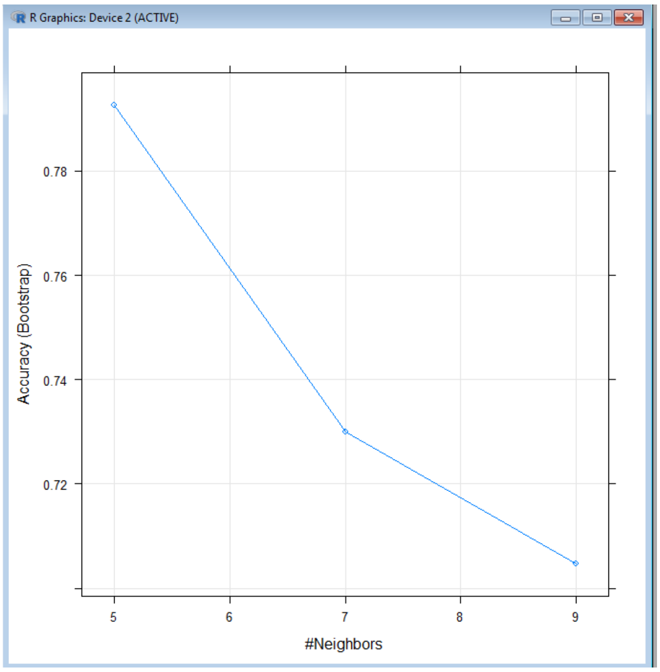


Fig. 3. Graph for selecting k-value

## 5. CONCLUSION

In this paper, it has been reviewed that KNN classification on gene expression data. From this analysis it has been studied that interesting issues related to classification and predicting results on gene expression data. It has been studied that, the KNN classification provides accuracy based on the k value, so in future the gene expression data will be applied on rule based classification for the best accuracy.

## REFERENCES

[1] Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., ... & Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature, 433(7027), 769.

[2] Arma R, Marcos IL, Taboada V, Ucar E, Irantzu B, Fullaondo A, Pedro L, Zubiaga A. Microarray analysis of autoimmune diseases by machine learning procedures. IEEE Trans Inform Biomed 2009;13(3):341–50.

[3] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1379.

[4] Lakshmi, G. Mohana, and K. Mythili. "Survey of gene-expression-based cancer subtypes prediction." International Journal 3.3 (2014).

[5] Dudoit, Sandrine, and Jane Fridlyand. "Classification in microarray experiments." Statistical analysis of gene expression microarray data 1 (2003): 93-158.

[6] Al Snousy, Mohmad Badr, et al. "Suite of decision tree-based classification algorithms on cancer gene expression data." Egyptian Informatics Journal 12.2 (2011): 73-82.

[7] Wang, Xiaosheng, and Osamu Gotoh. "A robust gene selection method for microarray-based cancer classification." Cancer informatics 9 (2010): 15.

[8] Pique-Regi, Roger, Antonio Ortega, and Shahab Asgharzadeh. "Sequential diagonal linear discriminant analysis (seqdlda) for microarray classification and gene identification." Computational Systems Bioinformatics

Conference, 2005. Workshops and Poster Abstracts. IEEE. IEEE, 2005.

[9] Arevalillo, Jorge M., and Hilario Navarro. "A new method for identifying bivariate differential expression in high dimensional microarray data using quadratic discriminant analysis." BMC bioinformatics 12.Suppl 12 (2011): S6.

[10] E. Fix, J.L. Hodges, Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[11] Ye, Jieping, et al. "Using uncorrelated discriminant analysis for tissue classification with gene expression data." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 1.4 (2004): 181-190.

[12] Huang, Desheng, et al. "Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data." Journal of Experimental & Clinical Cancer Research 28.1 (2009): 149.

[13] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.

[14] Alagukumar, S., and R. Lawrance. "A selective analysis of microarray data using association rule mining." Procedia Computer Science 47 (2015): 3-12.

[15] Vengateshkumar, R., S. Alagukumar, and R. Lawrance. "Boolean Association Rule Mining on Microarray Cancer Gene Expression Data using Gene Expression Intervals."