# K-means+ Method for Improving Gene Selection for Classification of Microarray Data

Heng Huang, Rong Zhang, Fei Xiong, Fillia Makedon
Computer Science Department
Dartmouth College
Hanover, NH 03755

Li Shen
Computer and Information Science Department
University of Massachusetts Dartmouth
Dartmouth, MA 02747

Bruce Hettleman, Justin Pearlman
Department of Cardiology
Dartmouth Medical School
Lebanon, NH 03756

## Abstract

*Microarray gene expression techniques have recently made it possible to offer phenotype classification of many diseases. One problem in this analysis is that each sample is represented by quite a large number of genes, and many of them are insignificant or redundant to clarify the disease problem. The previous work has shown that selecting informative genes from microarray data can improve the accuracy of classification. Clustering methods have been successfully applied to group similar genes and select informative genes from them to avoid redundancy and extract biological information from them. A problem with these approaches is that the number of clusters must be given and it is time-consuming to try all possible numbers for clusters. In this paper, a heuristic, called K-means+, is used to address the number of clusters dependency and degeneracy problems. The result of our experiments shows that K-means+ method can automatically partition genes into a reasonable number of clusters and then the informative genes are selected from clusters.*

## 1. Introduction

In gene expression data analysis, recently studies have shown that selection of a small subset of genes from broad patterns of gene expression data is useful for improving classification accuracy. People hope to find genes that reflect as many different aspects as possible between different samples. Typically, genes selection methods are based on gene ranking, but their results still include many highly correlated genes.

A number of methods have been developed to deal with gene selection. Jaeger *et al.* [6] proposed to use Fuzzy C-means method to cluster genes and select informative genes from these groups to avoid redundancy. In a similar approach, Hanczar *et al.* [5] used K-means clustering to select ”prototype genes”. A problem with these two approaches is that the number of clusters must be given and it is time-consuming to try all possible numbers for clusters during the experiments.

A K-means+ method that was recently introduced to intrusion detection analysis [4, 3] has been shown to efficiently partition the large data set. In this work, we applied a K-means+ method to automatically find the similar genes and group them together without predefining the number of clusters.

## 2. Method

K-means, an unsupervised learning algorithm, has been used to form clusters of genes in gene expression data analysis. One of the known drawbacks of the K-means method is that the number of clusters must be given and the optimum number of clusters is also unknown. Considering the statistics nature of clusters, we use a heuristic K-means method to efficiently cluster the genes.

### 2.1. Statistics nature

**Central Limit Theorem:** Let $X_1, X_2, X_N$ be a set of N independent random variates and each $X_i$ have an arbitrary probability distribution $P(x_1, ..., x_N)$ with mean $\mu_i$ and a finite variance $\sigma_i^2$. Then the normal form variate

$$X_{norm} \equiv \frac{\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu_i}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}}$$

has a limiting cumulative distribution function which approaches a normal distribution. [1]

The distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal. Thus, the Central Limit theorem is the foundation for many statistical procedures, including the K-means+ method.

**Chebyshev's Theorem:** The fraction of any data set lying within k standard deviations of the mean is at least $1 - 1/k^2$, where $k =$ a number greater than 1.

Chebysehev's theorem allows you to understand how the value of a standard deviation can be applied to any data set. The empirical rule gives more precise information about a data set than the Chebyshev's Theorem, however it only applies to a data set that has symmetrical mound shaped distributions.

**Empirical Rule:** for any normal distribution,

1. $68\%$ of the objects lie within one standard deviation of the mean.
2. $99.7\%$ of the objects lie within three standard deviations of the mean.
3. $99.99994\%$ of the objects lie within five standard deviations of the mean.

If we define a sphere with radius $= 5\sigma$, by Chebyshev' theorem, at least $96\%$ of objects lie inside. This is a lower bound for any data distribution. When the number of objects in cluster is large ($> 30$) [7], we can assume the objects are approximately in a normal distribution by Central Limit theorem. From the Empirical Rule, $99.99994\%$ of objects with normal distribution stay within the sphere with radius of $5\sigma$.

### 2.2. Heuristic K-means clustering method

K-means algorithm is performed first on a random number (between 1 and n). The outlier is defined and removed by a " Confident Area" [4] of each cluster, where the "Confident Area" is a circle area with a radius of five standard deviation of cluster. As the result, the removed outlier is assigned as the centroid of a new cluster.

Meanwhile we merge the adjacent clusters whose overlap is over the threshold $d = 1.414(\sigma_1 + \sigma_2)$ [3]. Distance $d$ represents the Pearson's correlation coefficient between clusters of genes.

After creating the groups of genes, we define the informative gene as the mean of the gene expression in each cluster [5].

## 3. Results

The public lung cancer data set is used to test our method [2]. Fig.1 shows K-means+ method partitions the lung cancer dataset into 150 to 200 clusters. On average, the final number of clusters is about 176. It is close to the optimal value in the previous work [5]. Based on this number, we can easily find the informative genes by using the method in section 2.2.

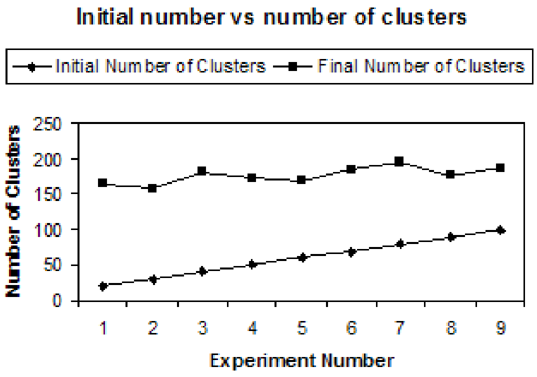

**Figure 1. Initial number vs final number of clusters**

## 4. Conclusions

In this paper, we apply a K-means+ algorithm that considers the statistics nature about the gene clusters and then informative genes are selected from the mean of each cluster. The method is tested the public lung cancer data. The results witness to the successful automatically partition of the gene expression data.

## References

[1] http://mathworld.wolfram.com/centrallimittheorem.html.

[2] A. Bhattacharjee, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *PNAS.*, 98(24):13790–13795, 2001.

[3] Y. Guan, A. Ghorbani, and N. Belacel. K-means+: An autonomous clustering algorithm. *In submission*.

[4] Y. Guan, A. Ghorbani, and N. Belacel. Y-means: A clustering method for intrusion detection. In *Proceedings of Canadian Conference on Electrical and Computer Engineering*, pages 4–7, 2003.

[5] B. Hanczar, M. Courtine, A. Benis, C. Hennegar, K. Clement, and J.-D. Zucker. Improving classification of microarray data using prototype-based feature selection. *SIGKDD Explorations*, 5(2):23–30, 2003.

[6] J. Jaeger, R. Sengupta, and W. L. Ruzzo. Improved gene selection for classification of microarrays. In *Proceedings of Pacific Symposium on Biocomputing*, pages 53–64, 2003.

[7] R. Walpole. *Elementary Statistical Concepts, second ed.* Mamillan, New York, 1983.