ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Method for Determining Optimum Number of Clusters for Clustering Gene Expression Cancer Dataset

P.Prabhu
Assistant Professor in Information Technology,
Directorate of Distance Education,
Alagappa University
Karaikudi, Tamilnadu, INDIA.
pprabhu70@gmail.com

Abstract: Clustering in gene expression data sets is a challenging problem. Different algorithms for clustering of genes have been proposed in the literature. Most of the partition based algorithms like k-means and k-medoids depend on the number of clusters as input parameter. This paper introduced method for determining the optimum number of clusters in a partition simply by examining various cluster validity measures for different values of numbers of clusters.

Keywords: Clustering, gene expression, external indexes, internal indexes

I. INTRODUCTION

Cluster analysis organises data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups [1]. It has many applications in different areas of computer sciences such as computational biology, machine learning, data mining and pattern recognition. There are various clustering algorithms are proposed in the literature like k-means, k-medoids, fuzzy C-Means etc., The usage of cluster algorithms is the problem of determining the number of classes existing in a dataset. Most clustering algorithms are parameterized approaches, with the target number of clusters k as the most frequent input parameter. We present experiments that compare the method for determining the number of clusters that is derived from cluster validate measures. Clustering validation, which evaluates the goodness of clustering results [13], has long been recognized as one of the vital issues essential to the success of clustering applications [1]. External clustering validation and internal clustering validation are the two main categories of clustering validation. The main difference is whether or not external information is used for clustering validation.

The performance of validation techniques usually depends on the data set or the cluster algorithm used to partition the data. In addition, the distance metric applied prior to clustering has proven a relevant factor for the final cluster solution and may also influence the cluster validity success to determine the optimum number of clusters. Cluster validation is performed by multiple simulations on a dataset varying the distance and clustering technique as well as the number of clusters k.

II. CLUSTERING ALGORITHMS

There are various clustering algorithms are proposed in the literature like k-means, k-medoids, fuzzy C-Means etc., The k-means and PAM clustering is discussed here.

A. K-Means Clustering

K- Means clustering algorithm was developed by J. MacQueen and then by J. A. Hartigan and M. A.Wong around. K-means is the simplest and most popular classical clustering method that is easy to implement. K-means clustering is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. It is also called centroid method. The K-means method uses the Euclidean distance measure, which appears to work well with compact clusters.

The K-means method may be described by the following steps.

Algorithm:

Step 1. Select the number of clusters. Let this number be k.
Step 2. Select k seeds as centroids of the k clusters. The seeds may be selected randomly.
Step 3. Computer the Euclidean distance of each object in the dataset from each of the centroids
Step 4. Allocate each object the cluster it is nearest to based on the distances computed in the step 3.
Step 5. Compute the centroids of the clusters by computing the means of the attribute values of the objects in each cluster.
Step 6. Check if the stopping criterion has been met. If yes go to step 7, else go to Step 3
Step 7. One may decide to stop at this stage or to split a cluster or combine two clusters heuristically until a Stopping criterion is met.

B. Partitioning Around Medoids (PAM)

Unfortunately, K-means clustering is sensitive to the outliers and a set of objects closest to a centroid may be empty, in which case centroids cannot be updated. For this reason, K-medoids clustering are sometimes used, where representative objects called medoids are considered instead of centroids. Because it uses the most centrally located object in a cluster, it is less sensitive to outliers compared with the K-means clustering. Among many algorithms for K-medoids clustering, Partitioning around Medoids (PAM) proposed by Kaufman and Rousseeuw (1990) is known to be most powerful.

III. CLUSTER VALIDITY MEASURES

There are two types of cluster validation to determine the optimum number of groups from a dataset, one way is to use external validation indexes for which a priori knowledge of dataset information is required, but it is hard to say if they can be used in real problems (usually, real problems do not have prior information of the dataset in question). Another way is to use internal validity indexes which do not require a priori information from dataset.

In the literature we can find different external and internal indexes, each approach has clear scope, in this paper we present a method to determine the optimum number of clusters using validity measures. We used K-means and PAM clustering algorithms to generate clusters.

A. External Validity Measures

External validation measures, which use external information present in the data (when true labels are known).The Rand Measure, Adjusted Random measure, Merkin and Huber measure are discussed in this paper.

a. Rand Measure

The Rand index or Rand measure in statistics, and in particular in data clustering, is a measure of the similarity between two data clusterings[15].

Given a set of n elements and two partitions of S to compare, and the following is defined:

i. the number of pairs of elements in S that are in the same set in X and in the same set in Y
ii. The number of pairs of elements in S that are in different sets in X and in different sets in Y
iii. The number of pairs of elements in S that are in the same set in X and in different sets in Y
iv. The number of pairs of elements in S that are in different sets in X and in the same set in Y

The Rand index, R, is:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \qquad (1)$$

Intuitively, a + b can be considered as the number of agreements between X and Y and c + d as the number of disagreements between X and Y. The adjusted-for-chance form of the Rand index is the adjusted Rand index. The Rand index lies between 0 and 1.When the two partitions agree perfectly, the Rand index is 1.

b. Adjusted Random Index

A problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value (say zero). The adjusted Rand index proposed by Hubert and Arabie 1985 assumes the generalized hyper geometric distribution as the model of randomness, i.e., the U and V partitions are picked at random such that the number of objects in the classes and clusters are fixed.

c. Mirkin Metric

This coefficient assumes null value for identical clustering's and positive values otherwise. It corresponds to the Hamming distance between the binary vector representations of each partition. It provides an alternative adjusted form of Rand index. However, unlike Hubert and Arabie's adjusted Rand (Hubert, 1985) it doesn't provide a correction for chance agreement.

B. Internal Validity Measures

Unlike external validation measures, internal validation measures only rely on information in the data (i.e., when true labels are unknown).The Silhouette index, Davies-Bouldin index, Calinski-Harabasz index and Krzanowski-lai index are discussed in this paper.

a. Silhouette Index (SI)

The Silhouette index (SI) [11] validates the clustering performance based on the pair wise difference of between and within-cluster distances. In addition, the optimal cluster number is determined by maximizing the value of this index.

b. Davies-Bouldin Index (DB)

The Davies-Bouldin index (DB) [2] is calculated as follows. For each cluster C, the similarities between C and all other clusters are computed, and the highest value is assigned to C as its cluster similarity. Then the DB index can be obtained by averaging all the cluster similarities. The smaller the index is, the better the clustering results. By minimizing this index, clusters are the most distinct from each other, and therefore achieve the best partition.

c. Calinski-Harabasz Index (CH)

The Calinski-Harabasz index (CH) [12] evaluates the cluster validity based on the average between- and within cluster sum of squares. Index I (I) [13] measures separation based on the maximum distance between cluster centers, and measures compactness based on the sum of distances between objects and their cluster center.

d. Krzanowski-lai Index (KL)

This index developed by Krzanowski and Lai (1988) by following the general approach of Marriott (1971) is formulated as follows;

$$diff_k = (k-1)^{2/p} trW_{k-1} - n^{2/p} trW_k \qquad (2)$$

Where k is the number of clusters, W is the pooled within-group covariance matrix for any given partition of the sample, tr (W) is the sum of squares and p denotes the number of features in the dataset. According to this formula, a stopping criterion shown as following formula is developed.

$$C_k = |diff_k| / |diff_{k+1}| \qquad (3)$$

The optimum value of k is the value maximizes Ck. Krzanowski and Lai (1988) mentioned that if the particular data set is inappropriate for the sum of squares objective function, then this criterion does not yield optimum results. Secondly, the frequent occurrence of multiple local maxima of Ck should be checked for unusual features, so they advised that the results should never be accepted uncritically but should always be examined for their meaningfulness.

IV. METHODOLOGY

The steps for determining the optimum number of clusters using different validity measures can be expressed in the Algorithm;

Algorithm

A. Load the dataset, data file: rows - data points, Columns – dimensions
B. Initialize the number of cluster as 2.
C. Calculate dis-similarity/distance matrix of a data set using Person similarity when true labels are known. The Pearson between two vectors x and y is,

$$\Sigma_j(x_j\text{-avg}(x))*(y_j\text{-avg}(y)) \, / \, (\Sigma_j(x_j\text{-avg}(x))\ {}^2\Sigma_j(y_j - \text{avg}(y))^2)^{1/2} \qquad (4)$$

Where the summation over j are over entries where both x and y have values, and avg(x) is the average value of the vector x.

D. Calculate dis-similarity/distance matrix of a data set using Euclidean similarity when true labels are unknown.
E. Pearson similarity [-1, 1] is normalized to Pearson distance [0, 1]
F. Run PAM or K-means clustering algorithm with initial number of clusters.
G. Calculated Clusters are validated with various distance measures.
H. Repeat the above steps with different cluster values to determine the optimal number of clusters.

Input

Data (nrow, numdim) – Data with n number of genes and n samples

Output

RI - Random Index values
ARI - Adjusted Random Index values
MI - Mirkin Index values
HI - Hubert Index values
SI - Silhouette index values
DB - Davies-Bouldin index values
CH - Caliniski-Harabasz index values
KL - Krzanowski-Lai index values

Different datasets may produce different results. Here we used Gene Expression Cancer datasets for our experiment.

V. GENE EXPRESSION DATASET

Different Gene Expression cancer datasets were used to find the number of clusters using various indices. The Leukemia and Colon cancer datasets are used in this experiment. The Leukemia data set is a collection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral blood) samples of 7129 genes. It contains an initial training set composed of 47 samples of Acute Lymphoblastic Leukemia and 25 samples of Acute Myeloblastic Leukemia. The Colon cancer dataset is a collection of gene expression measurements from 62 Colon biopsy samples of. It contains 22 normal and 40 Colon cancer samples of 2000 genes.

VI. RESULTS AND DISCUSSION

The results from various simulations show that this algorithm determines optimum number of clusters for leukemia data set and Colon cancer dataset.

When leukemia data set is used with known class label, the result of clustering using Partitioned around Medoids (PAM) is compared with input data using validity indices. The result of internal indices identified number of clusters are shown in fig.1.The result of External indices correctly identified the number of clusters as 3 is shown in Fig.2.

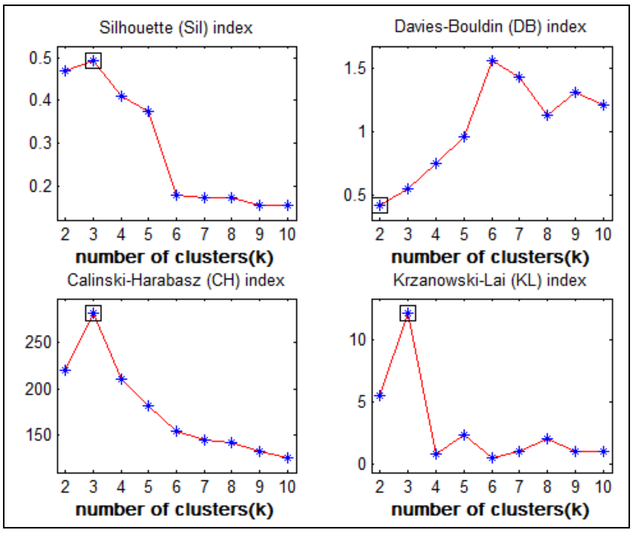


Figure.1 Leaukemia Dataset with Internal indices

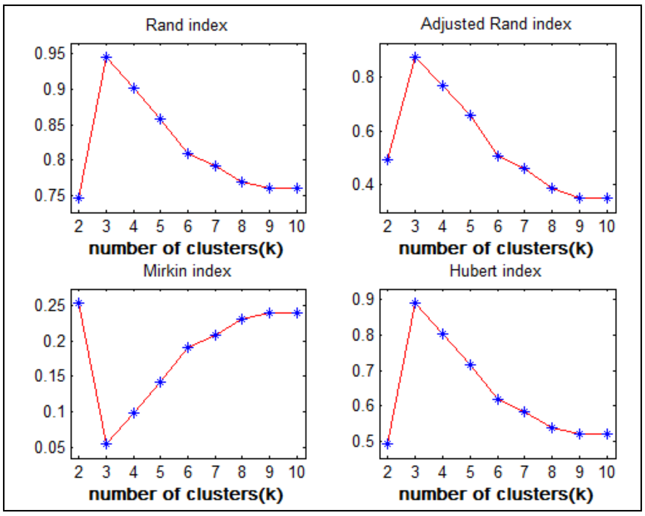
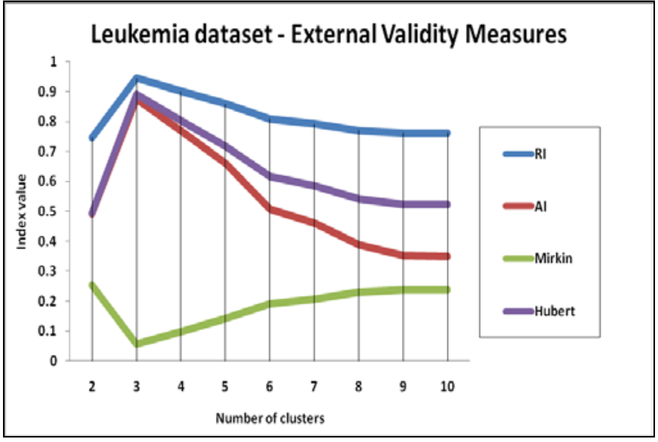


Figure.2 Leaukemia Dataset with External indices

The following Graph.1 shows that, the correct numbers of clusters were identified as 3 for Leukemia Dataset using External validity indices.



Graph1. Leukemia Dataset using External indices

When Colon cancer dataset is used with unknown label, the results of two cluster algorithms are compared using various validity indices. The following Fig.3 shows most of the internal validity indices give optimum number of cluster as 2.

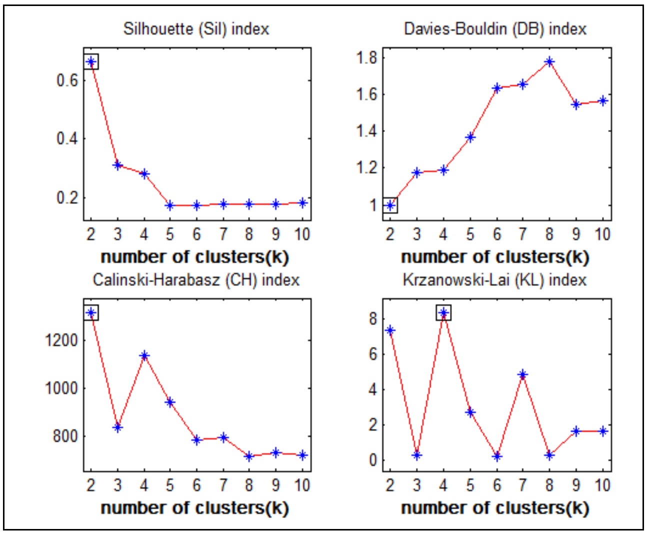


Figure.3 Colon Cancer Dataset with Internal Indices

The following Fig.4 shows the number of clusters identified by various External validity indices using colon dataset.

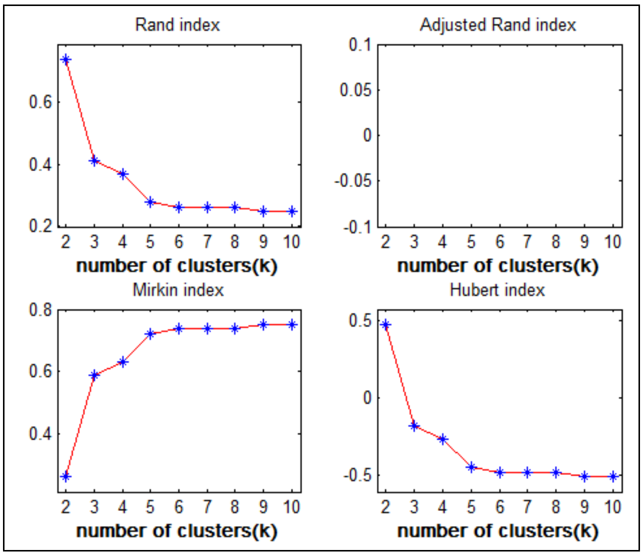


Figure.4 Colon Dataset with External indices

VII. CONCLUSION AND FUTURE WORK

The results presented in this paper confirm that measuring cluster validity can be a valuable approach to determine the number of clusters of any partition. In this paper we presented a method for determining optimum number of clusters using validity measures for Leukemia and Colon dataset. This method finds optimum number of cluster efficiently. Other validity measures and clustering algorithms may be used with different datasets to test the correctness of determining the number of clusters as a future work.

VIII. REFERENCES

[1] A. K. Jain and R. C. Dubes, Algorithms for clustering data. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[2] D. Davies and D. Bouldin, “A cluster separation measure”, IEEE PAMI, vol. 1, no. 2, pp. 224–227, 1979.

[3] Hubert, L. and Arabie, P. (1985) Comparing partitions. Journal of Classification, 193–218.

[4] Jiawei Han, Micheline Kamber, “Data Mining concepts and Techniques”, Morgan Kaufmann Publishers, San Fracisco, CA, USA.

[5] Leonard Kaufman, Peter J. Rousseeuw Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons 1990.

[6] Margaret H.Dunham, Data Mining Introductory and Advanced Topics, Pearson Education in SouthAsia.

[7] Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, IEEE Press & John Wiley, November 2002.

[8] Michael J.A.Berry Gordon Linoff, Mastering Data Mining” John wiley & sons ptd, Ltd, Singapore 2001.

[9] P.Prabhu and N.Anbazhagan, “Improving the performance of k-means clustering for high dimensional dataset’, International Journal of Computer Science and Engineering”, Vol 3. No.6. Pg 2317-2322 June 2011.

[10] P.Prabhu,’ Discovery of Novel Patterns in Animal Dataset using Hierarchical Techniques’, Indian Streams Research Journal, Vol I, Issue V, [June 2011] Information Technology.

[11] P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, no. 1, pp. 53–65, 1987.

[12] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” Comm. in Statistics, vol. 3, no. 1, pp. 1–27, 1974.

[13] U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,” IEEE PAMI, vol. 24, pp. 1650–1654, 2002.

[14] Yeung K.Y, Haynor D.R, Ruzzo W.L. Validating clustering for gene expression data. Bioinformatics.2001.

[15]http://en.wikipedia.org/wiki/Rand_index