Gaussian Fuzzy Index (GFI) for Cluster Validation: Identification of High Quality Biologically Enriched Clusters of Genes and Selection of Some Possible Genes Mediating Lung Cancer

Anupam Ghosh[1] and Rajat K. De[2]

[1] Department of Computer Science and Engineering,
Netaji Subhash Engineering College, Kolkata, India
anupam.ghosh@rediffmail.com
[2] Department of Machine Intelligence Unit, Indian Statistical Institute,
Kolkata, India
rajat@isical.ac.in

Abstract. In this article, we propose an index, called Gaussian Fuzzy-index (GFI), based on the notion of fuzzy set theory, for validating the clusters obtained by a clustering algorithm. This index is then used to identify some genes that have altered quite significantly from normal stage to diseased stage with respect to their expression patterns. Thus we can predict some possible disease mediating genes from microarray gene expression data. The methodology has been demonstrated on the gene expression data set dealing with human lung cancer. The performance of GFI is compared with 8 existing cluster validity indices. The results are appropriately validated using biochemical pathways. We have also implemented different cluster validity indices to demonstrate superior capability of GFI over the others.

1 Introduction

With the advent of high throughput technology, huge amount of data in various fields, including the field of molecular biology, is being generated/updated. In order to extract useful information from this huge data resource, people are finding interest in developing and using data exploration techniques. Clustering is a tool, in this regard, in finding natural groups of similar data, under unsupervision. The quality of the clusters obtained by an algorithm needs to be adjudicated or validated. Thus, cluster validation is major and challenging task [1]. There exist several cluster validity indices in literature [2]. Some of them are Dunn index (DI) [3], Davis-Bouldin index (DBI) [4], Silhouette index (SLI) [5], C-index (CI) [6], Goodman-Kruskal index (GKI) [7], Isolation index (II) [8] and Alternative Dunn Index (ADI) [9]. It is already established that current high throughput technology has a significant impact on genomic and post-genomic studies including gene

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 680–687, 2013.

identification, disease diagnosis, drug discovery and toxicological research. For instance, the accurate identification of genes is essential for a successful diagnosis and treatment of a disease like lung cancer. One of the major challenges associated with lung cancer is the identification of disease mediating genes.

Incorporation of fuzzy set theory enables one to deal with uncertainties in different tasks of designing an intelligent system, arising from deficiency (e.g., vagueness, incompleteness) in information, in an efficient manner [13–15]. Fuzzy set theory has been applied to formulate several cluster validity indices like Partition Coefficient Index (PCI) (of the data, one may refer [1, 9]), Classification Entropy Index (CEI) [1], Partition Index (SCI) [11], Separation Index (SI) [11], Xie and Beni's Index (XBI) [16], Fukuyama and Sugeno Index (FSI) [17], Fuzzy Hypervolume Index (FHVI) [18], Dave's modification of the PC index (MPCI) [19], Partition Coefficient and Exponential Separation Index (PCAESI) [12], Index Based on Akaikes information criterion (AICI) [20], Compose Within and Between scattering Index (CWBI) [10], PBMF-Index (PBMFI) [21]. However, there is no instance of using cluster validity index, to our knowledge, which has been applied to the problem of finding disease mediating genes. The importance of the notion of fuzzy sets has been realized and successfully applied in almost all the branches of science and technology.

Thus the present paper proposed a novel cluster validity index, called Gaussian fuzzy index (GFI) in fuzzy set theoretic framework. The index involves measuring the average fuzzy intra-cluster distances over all the clusters, and inter-cluster distances between pairs of clusters. The smaller the value of GFI, better is the quality of the clusters. The effectiveness of GFI has been demonstrated on the human lung cancer in finding some possible genes mediating the disease [22]. Moreover, we have demonstrated superior capability of GFI, in identifying genes mediating the lung cancer, through an extensive comparative study of GFI with 8 existing validity indices namely C-index (CI) [6], Partition Coefficient Index (PCI) [1, 9], Partition Index (SCI) [11], Separation Index (SI) [11], Fukuyama and Sugeno Index (FSI) [17], Alternative Dunn Index (ADI) [9], Partition Coefficient and Exponential Separation Index (PCAESI) [12], Index Based on Akaikes information criterion (AICI) [20] on two clustering algorithms, viz., k-means [23] and fuzzy c-means (FCM) [24] with Euclidean distance as similarity measure. The results are appropriately validated using biochemical pathways. Moreover, we have considered simple pattern recognition problems to visualize the effectiveness of GFI. Thus the comparative performance of the cluster validity indices to identify good and meaningful clusters, has been evaluated internally through identification of disease (lung cancer) mediating genes. The external evaluation of the indices has been made through consulting pathway database and it has been shown that both the forms of evaluation are comparable.

2 Methodology

Let us consider a set of samples $U = \{\mathbf{x}_k | k = 1, 2, \ldots, n\}$ that are distributed in l clusters $C_1, C_2, \ldots, C_l$. These clusters have been obtained by a clustering algorithm.

2.1 Gaussian Fuzzy Index (GFI) for Cluster Validation

We now define a cluster validity index, called Gaussian Fuzzy Index, that will demonstrate the goodness of the results obtained by a clustering algorithm. Gaussian Fuzzy Index (GFI) is defined as

$$GFI = \frac{E'}{1+E} \tag{1}$$

where E' is given by

$$E' = \frac{2}{l(l-1)} \sum_{\substack{k,j=1 \\ k \neq j}}^{l} \mu_k(\mathbf{c}_j) \tag{2}$$

and E defined by

$$E = \frac{1}{l} \sum_{k=1}^{l} \frac{1}{|C_k|} \sum_{\mathbf{x}_p \in C_k} \mu_k(\mathbf{x}_p) \tag{3}$$

The term $\mu_k(\mathbf{c}_j)$ represents the membership value indicating the degree of belongingness of the center of jth cluster C_j to kth cluster C_k, and l stands for the number of resulting clusters. The membership function we have considered here is of Gaussian type, and is defined as

$$\mu_k(\mathbf{c}_j) = \exp(-\frac{||\mathbf{c}_j - \mathbf{c}_k||^2}{L^2}) \tag{4}$$

Here $\mathbf{c}_k$ and $\mathbf{c}_j$ are the kth and jth cluster centers respectively. The term L indicates the maximum distance between two objects in the set U (i.e., set of all the data objects). Thus L is represented by

$$L = \max_{\substack{\mathbf{x}_p, \mathbf{x}_{p'} \in U \\ p \neq p'}} ||\mathbf{x}_p - \mathbf{x}_{p'}|| \tag{5}$$

It is to be mentioned here that the elements are chosen from normed linear space. Similarly, $\mu_k(\mathbf{x}_p)$, the membership value of pth sample $\mathbf{x}_p$ to kth cluster C_k, is defined as

$$\begin{aligned} \mu_k(\mathbf{x}_p) &= \exp(\frac{-||\mathbf{x}_p - \mathbf{c}_k||^2}{\sigma_k^2}), \text{ where } \mathbf{x}_p \in C_k \\ &= 0, \qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{6}$$

The term σ_k is the diameter of kth cluster C_k, and is defined as

$$\sigma_k = \max_{\mathbf{x}_p, \mathbf{x}_{p'} \in C_k} ||\mathbf{x}_p - \mathbf{x}_{p'}|| \tag{7}$$

We say that a set of clusters to be good if the inter-cluster distances are large and intra-cluster distances are small. Here, E (in Equation (3)) represents the average fuzzy intra-cluster distance over all the clusters. The value of E lies in [0,1]. $E = 0$ represents the highest average fuzzy intra-cluster distance over all the clusters. It is to be mentioned that since E can be zero, we have added 1 in the denominator of Equation (1). On the other hand, the lowest average fuzzy

intra-cluster distance over all the clusters is obtained at $E = 1$. Likewise, E' (in Equation (2)) represents the average fuzzy distance among the cluster centers or average fuzzy inter-cluster distance. As in the case of E, E' lies in [0,1]. $E' = 0$ indicates the highest fuzzy inter-cluster distance over all the pairs of clusters. On the other hand, the lowest average fuzzy inter-cluster distance over all the pairs of clusters corresponds to $E' = 1$. Thus, a set of clusters is said to be good if the value of GFI is minimum. In other words, lower the value of GFI, better is the set of clusters obtained by an algorithm.

2.2 Comparative Study of Cluster Validity Indices and Selection of Possible Disease Mediating Genes

The performance of GFI is compared with 8 cluster validity indices. For this comparative study, we consider the following work flow.

- *Step I*: **Generation of clusters** – A clustering algorithm $\mathcal{C}$ is applied on a gene expression data with the different number (k for k-means and c for fuzzy c-means) of clusters as its input. Here we have considered these numbers ranging from 2 to 20. It is to be noted that the gene expression profiles for normal and diseased states are considered separately, and the number of clusters to be generated in the diseased state is kept equal to that for normal state.
- *Step II*: **Selection of the best k-value (or c-value) using a cluster validity index** – Among these 8 k-values (or c-values), the best k-value (or c-value) has been selected based on a cluster validity index.
- *Step III*: **For each k-value (or c-value) and for the clustering algorithm $\mathcal{C}$, the following steps are performed.** It is to be mentioned here that we have considered $k = 2, 3, \ldots 20$, in Step I, for each clustering algorithm. In this step (Step III), we consider the same k-values as in Step I.
  - *Step III.1*: **Determining corresponding clusters** – Clusters obtained in Step I using the clustering algorithm $\mathcal{C}$ for a k-value (or c-value) for both normal and diseased states need to be matched. Let C_i^N and C_j^D be ith and jth clusters, obtained by the clustering algorithm $\mathcal{C}$ for a k-value (or c-value), for normal and diseased states respectively. We say that the cluster C_i^N, for normal state, corresponds to cluster C_j^D, for diseased state, if $|(C_i^N \cap C_j^D)|$ is maximum over $j = 1, 2, \ldots, j, \ldots, k$.
  - *Step III.2*: **Identifying altered gene clusters** – We call a gene to be an altered gene if the gene is in C_i^N and C_j^D, where $i \neq j$. Thus, we can write an altered gene set $A_i = \cup_{j=1\ j\neq i}^{k} (C_i^N \cap C_j^D)$, for C_i^N. Thus, altered gene sets or altered clusters (*i.e.*, $A_1, A_2, \ldots A_{k-1}, A_k$) are generated from k normal clusters.
  - *Step III.3*: **Scoring an altered gene set** – Let the number of matched genes in altered gene sets $A_1, A_2, \ldots A_{k-1}, A_k$ be $l_1, l_2, \ldots, l_{k-1}, l_k$, respectively. Thus, the score for S_k is defined as

$$S_k = \tfrac{1}{k} \times \sum_{i=1}^{k} \frac{l_i}{|A_i|} \times 100\% \tag{8}$$

Higher the value of S_k, better is the matching. In other words, if S_k, for a clustering algorithm and cluster validity index, is high, the index is highly capable of identifying genes mediating the lung cancer provided the said clustering algorithm is used.

– *Step IV*: **Determining the best k-value (or c-value) and selection of some possible genes mediating lung cancer**– Let the k-value (or c-value) for which S_k score are maximum be K_S. Let the best k-value (or c-value) obtained by a cluster validity index I be K_I. A cluster validity index performs the best if and only if $|K_S - K_I| = 0$.

Hence, we say that a cluster validity index I_1 is better than I_2 if

$$|K_S - K_{I_1}| < |K_S - K_{I_2}| \tag{9}$$

3 Results

The effectiveness of GFI has been demonstrated on human lung cancer gene expression dataset. The performance of GFI has been compared extensively with 8 indices. Moreover, we have provided results on two pattern recognition problems to ease of depiction of the effectiveness of GFI.

3.1 Description of the Dataset

Human lung gene expression data is obtained by oligonucleotide microarray experiments for Ann Arbor tumors and normal lung samples [22]. In this data set, there are 7129 genes (more specifically, Affymetrix probe-sets) for 86 lung tumor and 10 normal lung samples. The gene expression profiles represent 86 primary lung adenocarcinoma, including 67 stage I and 19 stage III tumors, as well as 10 neoplastic lung samples.

3.2 Comparative Results Using Pathway Database

A biological pathway corresponds to a specific function or a group of functions. Such a pathway is actually a cascade of reactions involving proteins and metabolites. The proteins are again end product of genes. For example, if a set of proteins (genes) are involved in a pathway and the pathway is responsible for apoptosis (the process by which cells die) related function, then malfunctioning of the pathway may lead to inhibit apoptosis. Thus there may be an overgrowth of cells leading to lung cancer. In NCBI, we have got pathway related information from bio-system database[1]. Here we have found some lung cancer specific pathways like non-small cell lung cancer, small cell lung cancer. We have identified the genes (proteins) involved in these pathways. If the genes (i.e., corresponding proteins) in the altered gene sets are involved in such a pathway, we say that GFI

[1] `http://www.ncbi.nlm.nih.gov/Database`

has correctly identified some possible genes mediating the lung cancer. Higher the number of such match, better is cluster validity index.

For lung expression data, we have considered $k = 2$ to $k = 20$ for k-means and $c = 2$ to $c = 20$ for fuzzy c-means clustering algorithms. Using k-means algorithm, we have found that the best results of GFI corresponds to $k = 10$. This result produces $|K_S - K_{GFI}| = 0$ for GFI. On the other hand, applying k-means algorithm, the best results have been obtained for $k = 8$ by CI; $k = 11$ by SI; $k = 12$ by SCI and FSI; $k = 13$ by ADI; $k = 14$ by PCI; $k = 15$ by PCAESI and AICI. We have also got maximum scores of S (Equation (8)) for $k = 10$ using k-means ($S_{10} = 91.74\%$). Figure 1 depicts that GFI performs the best with respect to 8 cluster validity indices for k-means algorithm on lung expression dataset. It is to be noted that the other validity indices have generated their best values between $k = 8$ and $k = 12$. Hence, we can say that for the lung expression data considered here, the high quality clusters have been generated by k-means algorithm between $k = 8$ and $k = 12$.

Similarly, applying fuzzy c-means, the best result generated by GFI is $c = 13$, such that $|K_S - K_{GFI}| = 1$. On the other hand, the best results for $c = 10$ by PCI; $c = 11$ by ADI; $c = 12$ by CI and AICI; $c = 13$ by SCI; $c = 15$ by PCAESI and FSI; $c = 16$ by SI. The maximum score ($S_{14} = 93.13\%$) has been generated for $c = 14$. From Figure 1, it is clearly seen that our proposed validity index (GFI) generates the best result for $c = 13$, which is very close to the result generated by the above indices. It is to be noted that, for fuzzy c-means algorithm, all the 8 validity indices have shown their best results between $c = 12$ and $c = 15$.

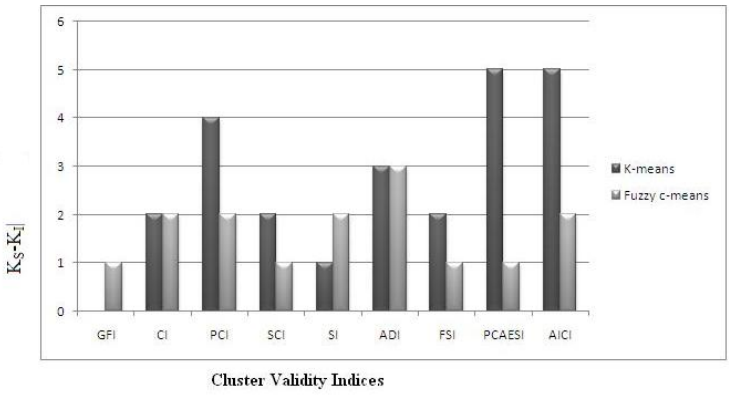

Fig. 1. Comparative values of $|K_S - K_I|$ for different cluster validity indices using pathway database for k-means and fuzzy c-means clustering algorithms on lung cancer dataset

3.3 Selection of Some Possible Genes Mediating Lung Cancer

Here we report the genes in the altered gene sets whose expression values have deviated from normal to disease states of human lung cancer dataset. From human lung expression data, the proposed index (GFI) has identified the genes (from altered gene set) like EGFR, TNF, TNFSF11, RIMS2, KRAS, HLA-G, TP53, VEGFA, IL6, CDKN2A, STAT3, CDH1, TGFB1, IL10, IL8, PTEN, MYC, IGFBP3, TNFSF10, CASP3, CD44, IGF1R. Moreover, we can say that the aforesaid altered genes have a significant role in the development of the aforesaid types of lung adenocarcinoma. In other words, we can say the aforesaid genes may have a strong influence in mediating the lung cancer. It is interesting to note that the proposed index GFI has been able to identify more responsible genes (for mediating the lung cancer) supported by a wide range of earlier investigations.

4 Conclusions

In this article, we have developed a cluster validity index, called Gaussian Fuzzy Index (GFI), using the notion of fuzzy set theory. The proposed index involves the average fuzzy intra-cluster distances over all the clusters, and inter-cluster distances between pairs of clusters. The index GFI is formulated in such a way that its minimization leads to minimization of fuzzy intra cluster distance and maximization of fuzzy inter cluster distance. The smaller the value of GFI signifies the better quality of the clusters.

The effectiveness of the index GFI has been demonstrated using two clustering algorithms on human lung cancer dataset [22]. Moreover, we have made an extensive experiment for identifying the important genes from a gene expression data. This concept leads to predict some possible disease mediating genes for certain human lung cancer. The results have appropriately been validated using biochemical pathways. We have also implemented different cluster validity indices to demonstrate superior capability of the GFI over the others.

References

1. Bezdek, J.C.: On clustering validation techniques. J. Cybernet. 17, 58–73 (1974)
2. Deborah, L.J., Baskaran, R., Kannan, A.: A survey on internal validity measure for cluster validation. IJCSES 1, 85–102 (2010)
3. Dunn, J.C.: Well separated clusters and optimal fuzzy partitions. J. Cybern. 4, 95–104 (1974)
4. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. 1, 224–227 (1979)
5. Rousseeuw, P.J.: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987)
6. Hubert, L., Schultz, J.: Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie 29, 190–241 (1976)

7. Goodman, L., Kruskal, W.: Measures of associations for cross-validations. J. Am. Stat. Assoc. 49, 732–764 (1954)
8. Pauwels, E.J., Frederix, G.: Finding salient regions in images: nonparametric clustering for image segmentation and grouping. Computer Vision and Image Understanding 75, 73–85 (1999)
9. Trauwaert, E.: On the meaning of dunn's partition coefficient for fuzzy clusters. Fuzzy Sets Systems 25, 217–242 (1988)
10. Yun, X.U., Brereton, G.R.: A comparative study of cluster validation indices applied to genotyping data. Chemometrics and Intelligent Laboratory Systems 78, 30–40 (2005)
11. Bensaid, A.M., Hall, L.O., Bezdek, J., Clarke, L.P., Silbiger, M.L., Arrington, J.A., Murtagh, R.F.: Validity-guided (re) clustering with applications to imige segmentation. IEEE Transactions on Fuzzy Systems 4, 112–123 (1996)
12. Wu, K., Yang, M.: A cluster validity index for fuzzy clustering. Pattern Recognition Lett. 26, 1275–1291 (2005)
13. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
14. Zadeh, L.A.: A fuzzy-set-theoretic interpretation of linguistic hedges. Journal of Cybernetics 2, 4–34 (1972)
15. Bandler, W., Kohout, L.J.: Fuzzy power sets and fuzzy implication operators. Fuzzy Sets and Systems 4, 13–30 (1980)
16. Xie, X.L., Beni, G.A.: Validity measure for fuzzy clustering. IEEE Trans. PAMI 3, 841–846 (1991)
17. Fukuyama, Y., Sugeno, M.: A new method of choosing the number of clusters for the fuzzy c-means method. In: Proceeding of fifth Fuzzy Syst. Symp., pp. 247–250 (1989)
18. Gath, I., Geva, A.B.: Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Machine Intell. 11, 773–781 (1989)
19. Dave, R.N.: Validating fuzzy partition obtained through c-shells clustering. Pattern Recognition Lett. 17, 613–623 (1996)
20. Akaike, H.: A bayesian extension of the minimum aic procedure of autoregressive model fitting. Biometrika 66, 237–242 (1979)
21. Pakhira, M., Bandyopadhyay, S., Maulik, U.: A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. Fuzzy Sets and Systems 155, 191–214 (2005)
22. Beer, G.D., et al.: Gene-expression profilespredict survival of patients with lung adenocarcinoma. Nature Medicine 8, 816–823 (2002)
23. Dubes, R.C., Jain, A.K.: Algorithms for clustering data. Prentice Hall (1988)
24. Bezdek, J.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
25. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics 7, 179–188 (1936)
26. Gibbons, F.D., Roth, F.P.: Judging the quality of gene expression-based clustering methods using gene annotation. Genome Research 12, 1574–1581 (2002)