molecular informatics

DOI: 10.1002/minf.201200142

Comparative Analysis of Cluster Validity Indices in Identifying Some Possible Genes Mediating Certain Cancers

Anupam Ghosh,[a] Bibhas Chandra Dhara,[b] and Rajat K. De*[c]

Abstract: In this article, we compare the performance of 19 cluster validity indices, in identifying some possible genes mediating certain cancers, based on gene expression data. For the purpose of this comparison, we have developed a method. The proposed method involves cluster generation, selection of the best *k*-value or *c*-values, cluster identification, identifying the altered gene cluster, scoring an altered gene cluster and determining the best *k*-value or *c*-value exploring through biological repositories. The effectiveness of the method has been demonstrated on three gene expression data sets dealing with human lung cancer, colon cancer, and leukemia. Here, we have used three clustering algorithms, i.e., *k*-means, PAM and fuzzy *c*-means. We have used biochemical pathways related to these cancers and *p*-value statistics for validating the study.

Keywords: Lung cancer · Colon cancer · Leukemia · Biochemical pathways related to cancers · *p*-Value · Functional enrichment

1 Introduction

Clustering is an unsupervised process in data mining and pattern recognition, and most of the clustering algorithms are very sensitive to their input parameters. Therefore it is very important to evaluate the results of the clustering algorithms. In the clustering processes there are no predefined classes, and therefore, it is difficult to find an appropriate metric for measuring quality of the resulting clusters. Several clustering validity indices have been formulated to cater this issue. The process of evaluating the results of a clustering algorithm is called cluster validity assessment. Two measurement criteria are well accepted for evaluating and selecting an optimal clustering scheme: compactness of individual clusters and separation between a pair of clusters. In an attempt to understand complicated biological systems, a large amount of gene expression data, along with other types of information, are being generated. In this article, we concentrate on gene expression data. This data involves expression level of a large number of genes over various samples/time points/experiments. Thus this type of data itself is huge and is associated with a lot of information. In order reduce the complexity of this type of data and thereby extracting useful information, clustering is widely used. Many clustering algorithms have been used and/or proposed in this regard. However, the majority of these attempts lack critical evaluation of the suitability of the clustering methods and/or proximity indices used, and their results.[1] Given the same data set, different clustering algorithms can generate very different clusters.[2] Thus assessing and interpreting the resulting clusters are as important as generating the clusters. A biologist working with a gene expression dataset faces with the problem of choosing an appropriate clustering algorithm for his or her dataset.[3] In much of the published clustering work on gene expression, the success of the clustering algorithms is accessed by visual inspection of biological knowledge. However, only a few attempts, in which cluster validation has been applied, involve focusing on the evaluation of the validation methodology proposed or used.[3,4] As a consequence, the study on the validation assessment of clusters is comparatively sparse.[5,6] Thus the present article is an attempt in this regard, which deals with comparing the performance of nineteen cluster validity indices in identifying biologically meaningful clusters obtained by three clustering algorithms, i.e., *k*-means, PAM and fuzzy *c*-means. Finally, we have compared the ability of these cluster validity indices in selecting possible genes mediating a disease like cancer. For the study of this comparison, we have developed a method. The method involves cluster generation, selection of the best *k*-value (or *c*-value) by a cluster validity

[a] *A. Ghosh*
Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India

[b] *B. C. Dhara*
Department of Information Technology, Jadavpur University Kolkata, India

[c] *R. K. De*
Machine Intelligence Unit, Indian Statistical Institute Kolkata, India
**e-mail: rajat@isical.ac.in*

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201200142.

index for a clustering algorithm, cluster identification, identifying the altered gene cluster, scoring an altered gene cluster and determining the best *k*-value (or *c*-value) by exploring biological repositories. Here we consider 19 clustering validity indices for their comparison. These indices are provided in Table 1. Further details on these cluster validity indices are provided in Supporting Information. The effectiveness of the methodology has been demonstrated on three human cancer datasets (i.e., lung,[7] colon,[8] leukemia.[9] The methodology uses biochemical pathways and *p*-values for validating the results biologically and statistically.

2 Method for Comparing Cluster Validity Indices

Here we describe the method for comparing performances of nineteen cluster validity indices (Table 1) in selecting possible genes mediating the development of three different cancers, i.e., lung, colon and leukemia, based on microarray gene expression datasets. For this purpose, we have considered three clustering algorithms, i.e., *k*-means, PAM and Fuzzy *c*-means, with Euclidian distance as the similarity measure. The steps of the method, for comparing the performances of the aforesaid cluster validity indices, are described below. The method is depicted in Figure 1.

2.1 Algorithm

- Step I: Generation of clusters using a clustering algorithm. A clustering algorithm C is applied on a gene expression data with the different number (*k* for *k*-means and PAM, 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 for 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 19 *k*-values (or *c*-values), the best *k*-value (or *c*-value) has been selected based on a cluster validity index. Thus we have got 19 best *k*-values (*c*-values) corresponding to 19 cluster validity indices, for a clustering algorithm C. These best *k*-values (or *c*-values) have been selected from gene expression data of normal states. These best *k*-values (or *c*-values) have been obtained by the cluster validity indices, and will be compared with the corresponding best *k*-values obtained in Steps III and IV.
- Step III: For each *k*-value (or *c*-value) and for the clustering algorithm *C*, the following steps are performed. It is to be mentioned here that we have considered $k = 2, 3, .$

Table 1. Various cluster validity indices and the corresponding measurements. Details of these indices are included in the Supporting Information.

Cluster-validity-index	Based on	References
Dunn index (DI)	Maximization of the intercluster distances and minimization the intracluster distances.	[10]
Davis-Bouldin index (DBI)	Ratio of the sum of within-cluster scatter to between-cluster separation	[11]
Silhouette index (SLI)	Comparison of its tightness and separation	[12]
C-index (CI)	Distances over all pairs of patterns from the same cluster	[13]
Goodman-Kruskal index (GKI)	Concordant and disconcordant quadruples	[14]
Isolation index (II)	Assertion that neighboring instances in feature space often occur in the same natural cluster	[15]
Partition coefficient index (PCI)	Extent of overlapping between cluster	[16,17]
Classification entropy index (CEI)	Fuzzyness of the cluster partition	[16]
Partition index (SCI)	Normalization through division by the fuzzy cardinality	[18]
Separation index (SI)	Minimum-distance separation for partition validity	[18]
Xie and Beni's index (XBI)	Quantification of the ratio of the total variation within clusters and the separation of clusters	[19]
Fukuyama and Sugeno index (FSI)	Fuzzy Index	[20]
Fuzzy hypervolume index (FHVI)	Fuzzy covariance of the partition	[21]
Alternative dunn index (ADI)	Dissimilarity function between two clusters rated in value from beneath by the triangle-non equality	[17]
Dave's modification of the PC index (MPCI)	Monotonic evolution tendency with cluster number	[22]
Partition coefficient and exponential separation index (PCAESI)	Normalized partition coefficient and an exponential separation	[23]
Index based on Akaikes in formation criterion (AICI)	Noise level, number of degrees of freedom, maximum number of cluster	[24]
Compose within and Bewween scattering index (CWBI)	Combination of average scattering for clusters with the distance functional	[25]
PBMF-index (PBMFI)	Fuzzy membership with optimum value for cluster center (avoidance of monotonicity)	[26]

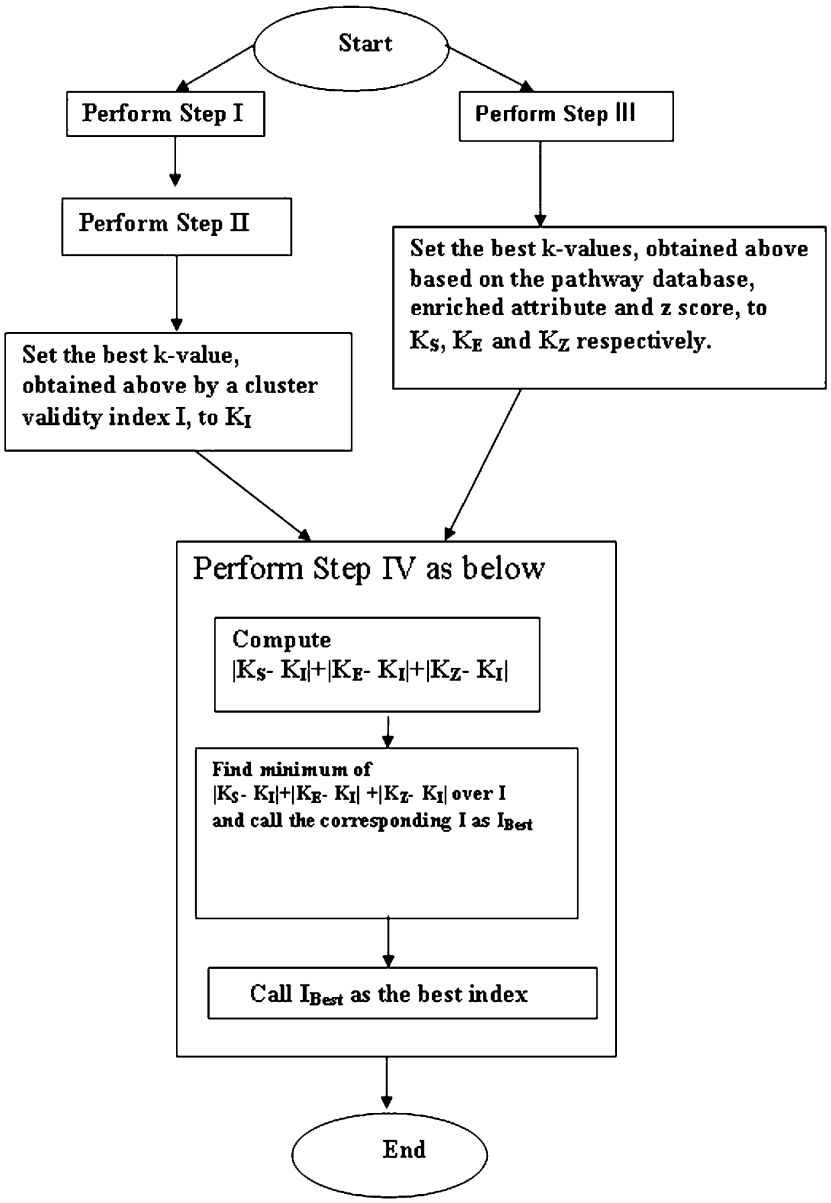

Figure 1. Flow chart for comparing cluster validity indices.

. . . 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 *C* for a *k*-value (or *c*-value) for both normal and diseased states need to be matched. Let C^N_i and C^D_j be ith and jth clusters, obtained by the clustering algorithm *C* for a *k*-value (or *c*-value), for normal and diseased states respectively. We say that the cluster C^N_i, for normal state, corresponds to cluster C^D_j, for diseased state, if $|(C^N_i \cap C^D_j)|$ is maximum over $j = 1, 2, \ldots, j, \ldots, k$. Without loss of generality, we renumber the cluster C^D_j as C^D_i so that C^N_i corresponds to C^D_i. Step III.2: Identifying altered gene clusters. For both normal and diseased states of data, we get *k* clusters, i.e., $C^N_1, C^N_2, \ldots C^N_k$ for normal state, and similarly for diseased state, the corresponding clusters are $C^D_1, C^D_2, \ldots C^D_k$. The clusters of normal state have been compared with the clusters of diseased state to identify the altered gene sets. We call a gene to be an altered gene if the gene is in C^N_i and C^D_j, where $i \neq j$. Thus, we can write an altered gene set $A_i = U^k_{j=1\, j\neq i}(C^N_i \cap C^D_j)$, for C^N_i. 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 – In this step, we compare the altered gene sets with an existing pathway database. If a gene in an altered gene set A_i is also included in a cancer pathway, we call the said gene in A_i to be a matched gene. Here, we generate a score (*S*) for

the altered gene set. Let the number of matched genes in altered gene sets A_1, A_2, . . . A_{k-1}, A_k be I_1, I_2, . . . I_{k-1}, I_k, respectively. Thus, the score for S_k is defined as

$$S_k = 1/k \times \Sigma^k_{i=1}\ I_i/|A_i| \times 100\,\% \quad (1)$$

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 a cancer provided the said clustering algorithm is used.

– Step III.4: Enriched attributes of an altered gene set. In this step, we compute the enriched attributes of the altered gene sets using *p*-value statistics. It is to be noted that only functional categories with *p*-value $\leq 5\times10^{-5}$ have been considered. Here, we compute a count of enriched attributes (E) for genes in an altered set. Let the number of enriched attributes for the matched genes in altered gene sets A_1, A_2, . . . A_{k-1}, A_k be e_1, e_2, . . . e_{k-1}, e_k respectively. Thus, the count for E_k is defined as

$$E_k = \Sigma^k_{i=1}\ e_i \quad (2)$$

If the value of E_k increases, the number of common attributes of the altered genes also increases. In other words, if an altered gene is associated with an attribute related to apoptosis, for example, the other altered genes are expected to be involved in apoptosis. Thus the altered genes together may have a significant role in mediating a cancer.

– Step III.5: *z*-score – This score is based on mutual information between a result obtained by a clustering algorithm, and gene annotation data. The *z*-score indicates relationships between clustering and annotation, relative to a clustering method that randomly assigns genes to clusters. a higher *z*-score indicates a clustering result that is further from a random one. In order to compare the performance of the clustering algorithms, this *z*-score is plotted for clustering results as a function of number of clusters, k, and to determine an optimal value for k.
– Step IV: Determining the best *k*-value (or *c*-value) and selection of some possible genes mediating certain cancers – Let the *k*-values (or *c*-values), for which S_k, E_k and *z*-score are maximum, be K_S, K_E and K_Z respectively. Thus K_S, K_E and K_Z are the best *k*-values (or *c*-values) considering the pathway database and *p*-value statistics of the enriched attributes and *z*-score respectively. Let the best *k*-value (or *c*-value) obtained by a cluster validity index I be K_I. For example, the best *k*-value (or *c*-value) selected by Dunn Index (DI) is denoted as K_{DI}. A cluster validity index performs the best if and only if $|K_S-K_I|=0$, $|K_E-K_I|=0$ and $|K_Z-K_I|=0$. Now, after selecting the best *k*-value (or *c*-value), the genes in the corresponding altered gene sets are selected as possible genes mediating certain cancers.

The best *k*-values (or *c*-values) obtained by different cluster validity indices (Step II) for a clustering algorithm are compared with those obtained in Step IV. We say that a cluster validity index I_1 is better than I_2 if

$$|K_S-K_{I_1}| + |K_E-K_{I_1}| + |K_z-K_{I_1}| < |K_S-K_{I_2}| + |K_E-K_{I_2}| + |K_z-K_{I_2}| \quad (3)$$

3 Results and Discussion

The comparative performance of these 19 cluster validity indices (Table 1) in identifying some possible genes mediating certain cancers, has been demonstrated extensively on three cancer gene expression datasets, i.e., lung[7] (7129 genes with 10 normal lung and 86 tumor samples), colon[8] (6600 genes with 18 normal and 18 tumor samples) and leukemia[9] (22283 genes with 13 normal and 43 diseased samples, GEO-ID: GDS2643). Here we have provided a brief description of the measurements used in the 19 validity indices in Table 1. A brief description of the datasets is provided in the Supporting Information. Finally some possible disease (cancer) mediating genes are selected.

3.1 Comparison Using Pathway Database

In NCBI, we have got pathway related information from bio-system database (http://www.ncbi.nlm.nih.gov/biosystems/). Here, we have found some cancer specific pathways including non-small cell lung cancer, small cell cancer, colorectal cancer, and chronic and acute myeloid leukemia related pathways. These pathways are involved in human lung, colon, and lymphocyte and plasma cells. From the aforesaid pathways, we have identified the genes (proteins) involved in these pathways. Now we consider the altered gene sets in such a way that the genes in these sets match to the genes (proteins) involved in the pathways.

For lung expression data, we have taken $k=2$ to $k=20$ for *k*-means, PAM and fuzzy *c*-means (for fuzzy *c*-means, $c=2$ to $c=20$ instead of k) clustering algorithms. Using *k*-means algorithm, the best results have been obtained for $k=8$ by CI and II; $k=9$ by DI, SLI, GKI and XBI; $k=10$ by DBI; $k=11$ by SI; $k=12$ by CWBI, SCI, MPCI, FSI and CEI; $k=13$ by ADI and PBMFI; $k=14$ by PCI and FHVI; and $k=15$ by PCAESI and AICI. We have also got maximum scores of S_k (Equation 1) for $k=10$ using *k*-means ($S_{10}=91.74\,\%$). For *k*-means algorithm, it has been clearly observed that DBI has shown the best result for $k=10$. The best *k*-value generated by the scoring method of pathway database is equal to the best *k*-value selected by the DBI cluster validity index. Thus, the present method has correctly validated the results generated by DBI. It is to be noted that the

Table 2. Best cluster validity indices in different experiments.

Dataset	Database/Score	Clustering algorithm *k*-means	PAM	Fuzzy c-means
Lung	Pathway	DBI	CI, CEI, PCAESI	II, CEI, XBI, FHVI, MPCI,CWBI, PBMFI
	Enriched attributes	DBI	DBI, II, SI, CWBI	DBI, SLI, GKI, FSI, PCAESI
	z-score	SI,DBI, MPCI, CWBI	DBI	CI, AICI, DI
	Combining pathway, enriched attributes and *z*-score	DBI, CI, II, CEI, SCI, SI, FSI, FHVI, MPCI, PBMFI	DBI, MPCI, FHVI	CWBI,SLI, GKI,CEI,SCI,SI, PCAESI
Colon	Pathway	DBI, CI, SLI, GKI, DI	DI, II,XBI, FHVI, AICI,	DBI,CWBI, PBMFI
	Enriched attributes	DBI, DI, CI, II, CWBI, PBMFI	DBI, DI, CI	FHVI, AICI, CWBI
	z-score	DI, GKI, II, SI,	DBI, II	FSI, AICI, PBMFI
	Combining pathway, enriched attributes and *z*-score	DI, DBI, CI, SLI, GKI, II, CEI, ADI, FSI, MPCI, AICI	DBI, CI, SLI, GKI, II, CEI, ADI	SI, XBI, FHVI, CWBI, PBMFI
Leukemia	Pathway	DBI, SLI, CI, GKI, CWBI, CEI	DBI, AICI,II	DBI, XBI, MPCI, PBMFI
	Enriched attributes	DBI, DI, AICI, CWBI	DI, CI, II	XBI, FHVI, PBMFI
	z-score	XBI, DI, II	DBI, II	DI, SLI, PCAESI
	Combining pathway, enriched attributes and *z*-score	DBI, SLI, CWBI, CEI	DBI, GKI, CI	DBI, XBI, MPCI, AICI, PBMFI

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 $k=12$. Likewise, using PAM the best results have been generated for $k=13$ by DBI; $k=8$ by FHVI and MPCI; $k=9$ by SLI and FSI; $k=10$ by GKI, PCI and XBI; $k=11$ by DI and ADI; $k=12$ by CI, CEI, SCI and PCAESI; $k=13$ by DBI, II, SI and CWBI; and $k=14$ by PBMFI and AICI. The validity indices have generated their best results for $k=9$ to $k=13$, using PAM. For PAM, the maximum score ($S_{12=}$91.83 %) has been generated for $k=12$. It is clearly observed that our method supports the result generated by CI, CEI and PCAESI for $k=12$. The remaining validity indices have generated their best values between $k=9$ and $k=13$. Thus we can say that the high quality clusters will be generated, for the said lung expression data, by PAM between $k=9$ and $k=13$. For fuzzy *c*-means algorithm, the best results have been generated for $c=15$ by DBI; $c=10$ by PCI; $c=11$ by ADI; $c=12$ by CI and AICI; $c=13$ by SCI and DI; $c=14$ by CEI, XBI, FHVI, MPCI, CWBI, PBMFI and II; $c=15$ by PCAESI, FSI, DBI, SLI and GKI; and for $c=16$ by SI. The maximum score ($S_{14}=93.13$ %) has been generated for $c=14$. It is clearly observed that our method supports the result generated by II, CEI, XBI, FHVI, MPCI, CWBI and PBMFI for $c=14$. From the results, it is clearly observed that II has performed better than other validity indices. It is to be noted that, for fuzzy *c*-means algorithm, all the 19 validity indices have shown their best results between $c=12$ and $c=15$. It is clearly observed that for the said lung expression data, the high quality clusters will be generated by FCM algorithm between $c=12$ and $c=15$. We have done similar experiments on the colon cancer and leukemia datasets. For colon expression data, the good indices have been found to be DBI, CI, SLI, GKI, DI, II, XBI, FHVI, AICI, CWBI and PBMFI. The best *k*-values for these algorithms have been found to be $k=3$ for *k*-means, $k=2$ for PAM and $c=4$ for fuzzy *c*-means. From the results, it is clearly observed that the high quality clusters will be generated between $k=3$ and $k=4$ for *k*-means, $k=3$ and $k=5$ for PAM, and $c=2$ and $c=4$ for fuzzy c-means. Likewise, for the leukemia dataset considered here, the good indices have been found to be DBI, SLI, CI, GKI, CWBI, CEI, XBI, MPCI, AICI and PBMFI. The best *k*-values for these algorithms are $k=9$ for *k*-means, $k=9$ for PAM and $c=10$ for fuzzy c-means. It is clearly observed that the high quality clusters have been generated between $k=8$ and $k=10$ for *k*-means, $k=8$ and $k=11$ for PAM and between $c=9$ and $c=12$ for fuzzy *c*-means. Thus, we can say that our proposed method is capable of identifying the best cluster validity index from a set of indices (Table 2). In other words, the method is able to identify the high quality clusters of genes with appropriate adjustment of *k*/*c*-values for gene expression datasets.

3.2 Using Functional Enrichment

In our study, the enrichment of each GO category for a group of genes has been calculated by its *p*-value. A low *p*-value indicates that the genes belonging to the enriched functional categories are biologically significant. Here only functional categories with *p*-value $\leq 5\times10^{-5}$ have been considered. Higher number of enriched attributes for a set of altered genes indicates that the resulting genes are belonging to the same functional categories. In other words, these genes perform similar functions. This means, if one of the genes from the pool is found responsible for a cancer then the other genes may have a strong influence in mediating the disease. In order to demonstrate the ability to identify cancer mediating genes correctly, we have computed the number of enriched attributes of all the altered gene sets for all the three cancer datasets. For human lung expression data, *k*-means, PAM, fuzzy c-means algorithms have generated maximum number of enriched attributes

for $k=10$, $k=13$ and $c=15$, respectively. The maximum number of enriched attributes for *k*-means, PAM and fuzzy *c*-means have been found to be 507, 503 and 465, respectively. From biological point view, higher number of enriched attributes generated by an altered gene set for a specific value of *k/c* using an algorithm signifies that the algorithm is able to find out the biologically enriched clusters for the specified value of *k/c*. For *k*-means algorithm on lung expression data, the result of DBI is supported by that of enriched attributes, for $k=10$. Actually, for $k=10$, DBI has shown the best value that is correctly validated by the enriched attributes (maximum value 507 for $k=10$). From the above results, we can say that the best *k*-value generated by the p-value statistics of enriched attributes is equal to the best *k*-value selected by DBI. In contrast, the best *k*-value generated by the *p*-value statistics of enriched attributes has differed by 1, 2 or 3 with respect to the best *k*-value selected by the remaining cluster validity indices. In addition, we can say that the method has almost validated the results generated by DI, SLI, CI, GKI, II indices biologically. Thus, the proposed algorithm has correctly validated the results generated by DBI to identify the high quality biologically enriched clusters. From results, we can conclude that DBI performs the best with respect to the other indices for *k*-means algorithm on lung expression data. For PAM on lung expression data, the highest number of enriched attributes have been generated for $k=13$. On the other hand, DBI and II have shown their best results for $k=13$. Thus, DBI, II, SI and CWBI perform better than other indices for PAM on lung expression data. It is also to be mentioned that the validity indices have generated their best results from $k=9$ to $k=13$ using PAM. Applying FCM, the highest number of enriched attributes has been generated for $c=15$. Interestingly, it has already been mentioned that DBI, SLI, GKI, FSI and PCAESI generated their best results for $c=15$. From the results, it is clearly visible that DBI, SLI, GKI, FSI and PCAESI are better indices than the others for fuzzy c-means algorithm on lung expression dataset. Likewise, we have done similar experiments on colon cancer and leukemia. For colon expression data, the good indices have been found to be DBI, DI, CI, II, XBI, FHVI, AICI, CWBI and PBMFI. The best *k*-values for these algorithms has been found to be $k=4$ for *k*-means, PAM and $c=4$ for fuzzy c-means. From results, it is clearly observed that the high quality clusters will be generated between $k=3$ and $k=4$ for *k*-means, $k=3$ and $k=5$ for PAM and between $c=2$ and $c=4$ for fuzzy c-means. For the leukemia dataset considered here, the best indices have been found to be DBI, DI, CI, II, XBI, FHVI, AICI, CWBI and PBMFI. The best *k*-values for these algorithms have been found to be $k=9$ for *k*-means and PAM, and $c=10$ for fuzzy c-means. From the results, it is clearly observed that the high quality clusters will be generated between $k=8$ and $k=10$ for *k*-means, $k=8$ and $k=11$ for PAM, and between $c=9$ and $c=12$ for fuzzy c-means. Thus, we can say that our proposed method is capable of identifying the best cluster validity index from a set of indices. In other words, the method is able to identify the high quality clusters of genes with appropriate adjustment of *k/c*-values for gene expression datasets. From the results, it is clearly observed that functional enrichment is also able to identify the high quality biologically enriched cluster of genes and select the best cluster validity index from a group of indices (Table 2).

3.3 Comparative Results Using *z*-Score

While the objective of clustering gene expression patterns is to bring genes of similar function together, we consider that the best method of clustering a particular data set is that which has the strongest tendency to bring genes of similar functions together. The clustering results obtained by an algorithm were evaluated by examining the relationship between the resulting clusters produced and the known attributes of the genes in those clusters. This annotation is made with a controlled vocabulary of gene attributes.[27] For good clustering algorithm with appropriate k/c-value, there should be some common attributes, depicting particular functions, of genes in a cluster. *z*-score is based on mutual information between a clustering result and gene annotation data.[27] It indicates relationships between clustering and annotation, relative to a clustering method that randomly assigns genes to clusters. A higher *z*-score indicates a clustering result that is further from random. In order to compare *k/c* values and/or clustering algorithms, *z*-score is plotted as a function of number of clusters, k, an optimal value for *k/c* is determined.[27] Applying *k*-means algorithm on lung expression data, it has been found that the indices like SI, DBI, MPCI, CWBI have the minimum values ($=0$) of $|K_Z-K_I|$. Likewise, using fuzzy *c*-means on lung expression data, we have found that CI, AICI and DI perform the best compared to the other validity indices considered here. Similarly for *k*-means and fuzzy c-means, DI, GKI, II, SI, FSI, AICI and PBMFI perform the best compared to the other validity indices for colon expression data. For leukemia dataset, we have found that XBI, PCAESI, DI and II perform the best compared to the other validity indices (Table 2).

3.4 Comparative Analysis

Applying *k*-means algorithm on lung expression data, it has been observed that DBI has performed the best with respect to other indices. From Figure 2, the minimum values of $|K_S-K_I|+|K_E-K_I|+|K_Z-K_I|$ have been found for DBI (i.e., $|K_S-K_I|+|K_E-K_I|+|K_Z-K_I|=0$). Likewise, using PAM and fuzzy *c*-means on lung expression data, we have found that DBI, CI, II, CEI, SCI, SI, PCAESI, CWBI, SLI, GKI, XBI, FSI, FHVI, MPCI and PBMFI have performed the best compared to the other validity indices (Figure 2) considered here. Similarly for *k*-means, PAM and fuzzy *c*-means, indices like DI, DBI, CI, SLI, GKI, II, CEI, ADI, SI, XBI, FHVI FSI, MPCI, AICI, CWBI and PBMFI have performed the best compared to

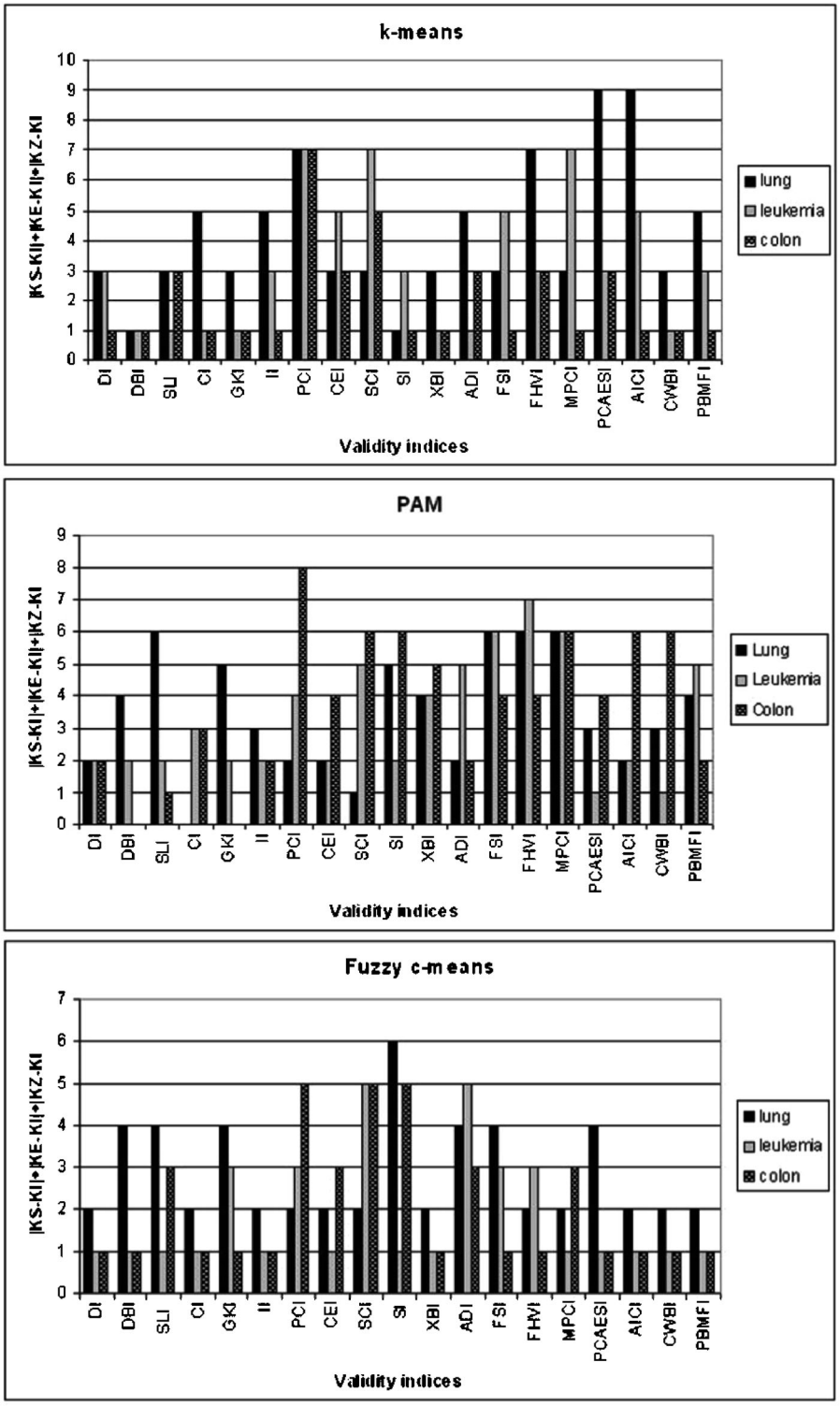

Figure 2. Comparative representation of different cluster validity indices for *k*-means, PAM and Fuzzy *c*-means clustering algorithms on different cancer datasets.

other validity indices for colon expression data (Figure 2). Lastly, for leukemia dataset, we have found that DBI, SLI, GKI, CI, CWBI, CEI, XBI, MPCI, AICI and PBMFI have outperformed the other validity indices (Figure 2). Thus DBI has performed the best compared to the other indices. This may be due to the following reason. Since the clustering algorithms used here consider distances of samples within and between clusters, and DBI depends on the ratio of sum

of within-cluster scatter to that of between-cluster scatter, DBI has performed the best in this case (Table 2).

3.5 Selection of Some Possible Genes Mediating Certain Cancers

In this section, we report the sets of altered genes whose expression values have deviated from normal to disease state of human lung, leukemia and colon cancer datasets. From human lung expression data, the proposed method 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. Likewise, the method has found the genes like MSH2, TP53, VEGFA, PTGS2, AKT1, HIF1A, CDKN1A, EGFR, MMP9, MMP2, MAPK1, TGFB1, NFKB1, IGF1, MMP7, MTHFR, MSH6, STAT3, MAPK14, BAX, CDH1, MAPK3, CDKN2A, JUN, IGF1R, MAPK8, PTEN, MMP13, PIK3CA for human colon expression dataset. Lastly, genes like MLL, ARHGEF12, RUNX1, PML, PBX1 BCR, EGFR, ERBB2, MCL1, TNF, MLLT4, BCL2, KRAS, BRCA2, HLA-DRB1, HLA-G, DEK, PTK2, TP53, VEGFA, IL6, TGFB1, IL8, STAT3, MYC, IGF1, BRAF, LEP have been identified by the proposed method for human leukemia dataset. GO and PUBMED identifiers of these genes have been provided in Tables 3–11 in the Supporting Information. In addition, we can say that the aforesaid altered gene set of different datasets have a significant role in mediating the carcinogenic nature of the human cells. In other words, we can say the aforesaid genes may have a strong influence in mediating the cancers. It is interesting to note that the proposed method has been able to find more responsible genes (for mediating cancer) than that are supported by wide range of earlier investigations. Thus, the methodology developed in this article is able to identify biologically more significant genes. Hence, these results may facilitate the biologists and researchers carrying out the biochemical analysis to do further study on these genes.

4 Conclusions

In this article, we have compared various cluster validity indices in identifying genes mediating certain cancers, using gene expression data. For this purpose, we have developed a method that compares these indices. Finally, some possible genes mediating certain cancers, have been selected. We have considered gene expression data related to lung, colon and leukemia to demonstrate the comparative performance of 19 cluster validity indices. The results have appropriately been validated using biochemical pathways and p-value. It has been found that DBI has performed the best for *k*-means; DBI, CI, CEI and PCAESI for PAM; and DBI, II, CEI, XBI, FHVI, MPCI, CWBI and PBMFI for fuzzy c-means on lung cancer dataset. Similarly, indices like DBI, CI, SLI, GKI, DI, II, XBI, FHVI, AICI, CWBI and PBMFI have performed the best for colon expression data on applying *k*-means, PAM and fuzzy c-means. Lastly, it has been found that indices like DBI, SLI, CI, GKI, CWBI, CEI, XBI, MPCI, AICI and PBMFI have performed the best for leukemia dataset. In other words, DBI has performed the best for all the algorithms as well as the datasets considered here. Thus we can say that Davis-Bouldin index (DBI) has performed the best compared to the other indices considered here, in identifying genes mediating certain cancers.

References

[1] M. Eisen, P. Spellman, P. Brown, D. Botstein, *Proc. Natl. Acad. Sci.* **1998**, *95*, 14863 – 14868.
[2] J. L. DeRisi, V. R. Iyer, P. O. Brown, *Science* **1997**, *278*, 680 – 686.
[3] K. Y. Yeung, D. R. Haynor, W. L. Ruzzo, *Bioinformatics* **2001**, *17*, 309 – 318.
[4] Z. Lubovac, B. Olsson, P. Jonsson, K. Laurio, M. L. Andersson, *Proc. Math. Comput. Biol. Chem.* **2001**, 149 – 155.
[5] I. G. Costa, F. A. T de Carvalho, M. C. P. de Souto, *Proc. Braz. Workshop Bioinform.* **2002**, 88 – 90.
[6] S. Datta, S. Datta , *Bioinformatics* **2003**, *19,* 459 – 466.
[7] G. D. Beer et al. *Nature Med.* **2002**, *8*, 816–823; ftp://ftp.camda.duke.edu/CAMDA03-DATASETS/michigan-publication.zip.
[8] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra et al., *Proc. Natl. Acad. Sci.* **1999**, *96*, 6745–6750; http://microarray.princeton.edu/oncology/.
[9] N. C. Gutierrez, E. M. Ocio, J. delas Rivas, P. Maiso, M. Delgado et al. *Leukaemia* **2007**, *21*, 541–549; http://ncbi.nlm.nih.gov/projects/geo/.
[10] J. C. Dunn, *J. Cybern.* **1974**, *4*, 95 – 104.
[11] D. L. Davies, D. W. Bouldin, *IEEE Trans Pattern Anal. Machine Intell.* **1979**, *1*, 224 – 227.
[12] P. J. Rousseeuw, *J. Comput. Appl. Math.* **1987**, *20*, 53 – 65.
[13] L. Hubert, J. Schultz, *Brit. J. Math. Statist. Psychol.* **1976**, *29*, 190 – 241.
[14] L. Goodman, W. Kruskal, *J. Am. Stat. Assoc.* **1954**, *49*, 732 – 764.
[15] E. J. Pauwels, G. Frederix, *Comput. Vision Image Understanding* **1999**, *75*, 73 – 85.
[16] J. C. Bezdek, *Cybernet* **1974**, *3*, 58 – 73.
[17] E. Trauwaert, Fuzzy Sets, *Fuzzy Sets Syst.* **1988**, *25*, 217 – 242.
[18] A. M. Bensaid, L. O. Hall, J. Bezdek, L. P. Clarke, M. L. Silbiger et al., *IEEE Transact. Fuzzy Syst.* **1996** *4*, 112 – 123.
[19] X. L. Xie, G. A. Beni, *IEEE Trans. PAMI* **1991**, *3,* 841 – 846.
[20] Y. Fukuyama, M. Sugeno , *Proc. 5th Fuzzy Syst. Symp.* **1989**, 247 – 250.
[21] I. Gath, A. B. Geva, *IEEE Trans. Pattern Anal. Machine Intell.* **1989**, *11*, 773 – 781.
[22] R. N. Dave, *Pattern Recognition Lett.* **1996**, *17*, 613 – 623.
[23] K. Wu, M. Yang, *Pattern Recognition Lett*, **2005**, *26*, 1275 – 1291.
[24] H. Akaike, in *Applications of Statistics* (Ed.: P. R. Krishnaiah), North Holland, Amsterdam, **1977**, pp. 27–41.
[25] X. U. Yun, G. R. Brereton, *Chemomet. Intell. Lab. Syst.* **2005**, *78*, 30 – 40.
[26] M. Pakhira, S. Bandhyopadhya, U. Moulik, *Fuzzy Sets Syst.* **2005**, *155*, 191 – 214.
[27] F. D. Gibbons, F. P. Roth, *Genome Res.* **2002**, *12*, 1574 – 1581

Received: November 14, 2012
Accepted: February 25, 2012
Published online: ■■ ■■, 0000

