Identification of certain cancer-mediating genes using Gaussian fuzzy cluster validity index

ANUPAM GHOSH[1,*] and RAJAT K DE[2]

[1]*Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India*

[2]*Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India*

**Corresponding author (Email, anupam.ghosh@rediffmail.com)*

In this article, we have used an index, called Gaussian fuzzy index (GFI), recently developed by the authors, based on the notion of fuzzy set theory, for validating the clusters obtained by a clustering algorithm applied on cancer gene expression data. GFI is then used for the identification of genes that have altered quite significantly from normal state to carcinogenic state with respect to their mRNA expression patterns. The effectiveness of the methodology has been demonstrated on three gene expression cancer datasets dealing with human lung, colon and leukemia. The performance of GFI is compared with 19 exiting cluster validity indices. The results are appropriately validated biologically and statistically. In this context, we have used biochemical pathways, *p*-value statistics of GO attributes, *t*-test and *z*-score for the validation of the results. It has been reported that GFI is capable of identifying high-quality enriched clusters of genes, and thereby is able to select more cancer-mediating genes.

[Ghosh A and De RK 2015 Identification of certain cancer-mediating genes using Gaussian fuzzy cluster validity index. *J. Biosci.* **40** 741–754] **DOI** 10.1007/s12038-015-9557-x

1. Introduction

The huge amount of data, mainly in the field of molecular biology, is being generated with the advent of high-throughput technology. In order to mine interesting information from this biological data resource, there is increased interest in developing and using data exploration techniques. Clustering is a tool, in this regard, to find natural groups of similar data pattern. Under this scenario, since there is no predefined class label or group information, it is always an issue in finding appropriate measures for determining similarity among the samples, number of clusters to be obtained and cluster shapes. In other words, the quality of the clusters obtained by an algorithm needs to be adjudicated or validated. This quality determines the purity of clusters. Thus, cluster validation is a major and challenging task (Bezdek 1974).

There exist several cluster validity indices in the literature (Deborah *et al.* 2010). Some of them are Dunn index (DI) (Dunn 1974), Davis–Bouldin index (DBI) (Davies and Bouldin 1979), Silhouette index (SLI) (Rousseeuw 1987), C-index (CI) (Hubert and Schultz 1976), Goodman–Kruskal index (GKI) (Goodman and Kruskal 1954), Isolation index (II) (Pauwels and Frederix 1999) and Alternative Dunn Index (ADI) (Trauwaert 1988). The performances of all the existing indices in the domain of image segmentation have been experimentally evaluated and compared on several test images under noisy conditions of varying degrees (Yun and Brereton 2005), and in the domain of 3D MRI images (Bensaid *et al.* 1996).

All these indices have a common objective for finding a good estimate of the number of clusters so that each of the clusters is compact and well separated from others (Wu and Yang 2005). If a dataset contains some noisy points, it can be

Keywords. Biochemical pathways; clustering; fuzzy sets; gene expression; GO-attributes

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/oct2015/supp/Ghosh.pdf*

Published online: 29 September 2015

visualized that cluster validity indices will take each noisy point into a singleton cluster (Bensaid *et al.* 1996). It is to be noted that the disadvantage of the above indices is that they lack the connection to the geometrical structure of the data (Bezdek 1974; Trauwaert 1988).

It is already established that current high-throughput technology has a significant impact on genomic and post-genomic studies including gene 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 cancer. One of the major challenges associated with cancer is the identification of cancer-mediating genes.

Fuzzy set theory was introduced by Zadeh in 1965 (Zadeh 1965) with an objective to provide a formal setting for incomplete and gradual information, as expressed by people in natural language. There is a very long tradition of philosophical interest in modelling ambiguity and imprecision of knowledge (Zadeh 1997). Imprecision of knowledge, along with some others including inexactness, vagueness and uncertainty, has been conceived, modelled and analysed in various ways (Zadeh 1972; Bandler and Kohout 1980).

Incorporation of fuzzy set theory enables one to deal with uncertainties, vagueness, and incompleteness in different tasks of designing an intelligent system, arising from deficiency in information, as in case of biological datasets, in an efficient manner. Apart from designing methods for classification, clustering, feature selection and/or extraction, fuzzy set theory has been applied to formulate several cluster validity indices. They include Partition Coefficient Index (PCI) (for the data, one may refer Bezdek (1974) and Trauwaert (1988)), Classification Entropy Index (CEI) (Bezdek 1974), Partition Index (SCI) (Bensaid *et al.* 1996), Separation Index (SI) (Bensaid *et al.* 1996), Xie and Beni's Index (XBI) (Xie and Beni 1991), Fukuyama and Sugeno Index (FSI) (Fukuyama and Sugeno 1989), Fuzzy Hypervolume Index (FHVI) (Gath and Geva 1989), Dave's modification of the PC index (MPCI) (Dave 1996), Partition Coefficient and Exponential Separation Index (PCAESI) (Wu and Yang 2005), Index Based on Akaike's information criterion (AICI) (Akaike 1979), Compose Within and Between scattering Index (CWBI) (Yun and Brereton 2005), and PBMF-Index (PBMFI) (Pakhira *et al.* 2005). 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 (Tripathy *et al.* 2012).

In the present study, we propose a novel cluster validity index Gaussian fuzzy index (GFI) in fuzzy set theoretic framework. The index GFI was formulated in such a way that its minimization led to minimization of fuzzy intra-cluster distance and maximization of fuzzy inter-cluster distance. Thus, smaller the value of GFI, better is the quality of the clusters. Then GFI was applied to identify disease mediating genes. This was performed by clustering microarray gene expression data and evaluating the quality of the clusters using the proposed cluster validity index, called Gaussian fuzzy index (GFI).

The effectiveness of GFI has been demonstrated on three human cancers (lung [Beer *et al.* 2002], colon [Alon *et al.* 1999], leukemia [Gutierrez *et al.* 2007]) in finding some possible genes mediating these cancers. An initial set of results for lung cancer has been published in Ghosh and De (2013). Moreover, we have demonstrated superior capability of GFI, in identifying genes mediating these cancers, through an extensive comparative study of GFI with 19 existing validity indices like Dunn index (DI) (Dunn 1974), Davis–Bouldin index (DBI) (Davies and Bouldin 1979), Silhouette index (SLI) (Rousseeuw 1987), C-index (CI) (Hubert and Schultz 1976), Goodman–Kruskal index (GKI) (Goodman and Kruskal 1954), Isolation index (II) (Pauwels and Frederix 1999), Partition Coefficient Index (PCI) (Bezdek 1974; Trauwaert 1988), Classification Entropy Index (CEI) (Bezdek 1974), Partition Index (SCI) (Bensaid *et al.* 1996), Separation Index (SI) (Bensaid *et al.* 1996), Xie and Beni's Index (XBI) (Xie and Beni 1991), Fukuyama and Sugeno Index (FSI) (Fukuyama and Sugeno 1989), Fuzzy Hypervolume Index (FHVI) (Gath and Geva 1989), Alternative Dunn Index (ADI) (Trauwaert 1988), Dave's modification of the PC index (MPCI) (Dave 1996), Partition Coefficient and Exponential Separation Index (PCAESI) (Wu and Yang 2005), Index Based on Akaike's information criterion (AICI) (Akaike 1979), Compose Within and Between scattering Index (CWBI) (Yun and Brereton 2005) and PBMF-Index (PBMFI) (Pakhira *et al.* 2005) (table 2). Here, we used two clustering algorithms, viz., *k*-means (Dubes and Jain 1988) and fuzzy *c*-means (FCM) (Bezdek 1981) with Euclidean distance as similarity measure. The results are appropriately validated using biochemical pathways, *p*-value statistics of enriched attributes, *t*-test and *z*-score.

Thus the comparative performance of the cluster validity indices to identify good and meaningful clusters, has been evaluated internally through identification of disease (cancer) mediating genes. The external evaluation of these set indices has been made through consulting pathway database, and using well known parameters like *p*-value, *t*-test and *z*-score. In an earlier investigation, it has been shown that both the forms of evaluation are comparable (Ghosh *et al.* 2013)

The article is organized as follows. Although GFI has been proposed in (Ghosh and De 2013), we again describe it thoroughly, for the sake of the readers, in Appendix A.1, which describes GFI, while Appendix A.2 narrates the way we have used cluster validity indices, and compared their capabilities, to find disease mediating genes. Section 2 describes extensively the experimental results. This section has several subsections. Section 2.1 describes the gene expression data briefly. Sections 2.2–2.4 provide comparative results using pathway database, *p*-value and *z*-score, respectively. Comparative results based on all the three together, are provided in Section 2.5.

Finally in Section 2.7, we provide lists of some possible disease mediating genes obtained in the above subsections, along with their statistical validation using *t*-test in Section 2.8. Section 3 concludes the article.

2. Results

2.1 *Description of the datasets*

Here, we have considered three gene expression datasets related to lung cancer, colon cancer and leukemia gene expression patterns. A brief description of the datasets is given below.

>Human lung expression data: Human lung gene expression data is obtained by oligonucleotide microarray experiments for Ann Arbor tumours and normal lung samples (Beer *et al.* 2002). In this dataset, there are 7129 genes (more specifically, Affymetrix probe-sets) for 86 lung tumour and 10 normal lung samples. The gene expression profiles represent 86 primary lung adenocarcinoma, including 67 stage I and 19 stage III tumours, as well as 10 neoplastic lung samples. More details on this dataset can be found in Beer *et al.* (2002). Database web link for this data is *http://ncbi.nlm.nih.gov/projects/geo/*.

Human colon expression data: Human colon expression data (Alon *et al.* 1999) consists of 18 tumour and 18 normal samples. In this dataset, samples of colon adenocarcinoma and paired normal tissue extracted from the same patient were obtained by the Cooperative Human Tissue Network. The tissue was snap-frozen in liquid nitrogen within 20–30 min of harvesting and stored thereafter at −80°C. mRNA was extracted from the bulk tissue samples and hybridized to the array using standard procedure. The dataset consists of 6600 genes and expressed sequence tags (ESTs). The data can be obtained at *http://microarray.princeton.edu/oncology/*.

Human lymphocytes and plasma cell expression data: Human lymphocytes and plasma cell expression data (Gutierrez *et al.* 2007) has been used for analysis of B lymphocytes (BL) and plasma cells (PC) extracted from patients with Waldenstrom's macroglobulinemia (WM) and B-lymphoproliferative disorder (BLPD). The dataset consists of 22283 genes with 56 samples. Among them, there are 13 normal samples (8 normal B lymphocytes and 5 normal plasma cells) and 43 diseased (20 Waldenstrom's macroglobulinemia, 11 chronic lymphocytic leukemia and 12 multiple myeloma) samples. The data can be obtained at *http://ncbi.nlm.nih.gov/projects/geo/*.

2.2 *Comparative results using pathway database*

In bio-system database of NCBI (*http://www.ncbi.nlm.nih.gov/Database*), we have found some cancer specific pathways for non-small-cell lung cancer, small cell lung cancer, colorectal cancer, and chronic and acute myeloid leukemia related pathways. These pathways are involved in apoptosis or related function in human lung, colon, and lymphocyte and plasma cells. 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 has correctly identified some possible genes mediating a cancer. Higher the number of such match, better is the cluster validity index.

Using *k*-means algorithm on human lung expression data, we have found that the best result produced by GFI corresponds to $k = 10$. This result produces $|K_S - K_{GFI}| = 0$ for GFI. We have also got maximum scores of S (equation 8, in the appendix) for $k = 10$ using *k*-means ($S_{10} = 91.74\%$). Figure 1 depicts that the best *k*-value generated by the scoring method on the pathway database is equal to the best *k*-value selected by GFI and DBI cluster validity indices. Thus, GFI performs the best along with DBI for *k*-means algorithm on lung expression dataset (Ghosh and De 2013). It is to be noted that the other validity indices have generated their best values between $k = 8$ and $k = 12$. Similarly, applying fuzzy *c*-means, the best result generated by GFI is $c = 13$. From figure 1, it is clearly seen that GFI generates the best result for $c = 13$, which is very close to the result generated by the II, CEI, XBI, FHVI, MPCI, CWBI and PBMFI. It is also to be noted that, for fuzzy *c*-means algorithm, all the 20 validity indices have shown their best results between $c = 12$ and $c = 15$.

We have done similar experiments on colon cancer and leukemia datasets. For colon expression data, the best indices have been found to be GFI, 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, and $c = 4$ for fuzzy c-means. From figure 1, it is clearly observed that the high quality clusters will be generated between $k = 3$ and $k = 4$ for *k*-means and between $c = 2$ and $c = 4$ for fuzzy *c*-means.

Likewise, for leukemia dataset considered here, the best indices have been found to be GFI, DBI, SLI, CI, GKI, CWBI, CEI, XBI, MPCI, AICI and PBMFI. The best *k*-values for these algorithms have been found to be $k = 9$ for k-means and $c = 10$ for fuzzy c-means. From figure 1, it is clearly observed that the high quality clusters will be generated between $k = 8$ and $k = 10$ for *k*-means, and between $c = 9$ and $c = 12$ for fuzzy *c*-means. Thus, we can say that our proposed validity index (GFI) is capable of identifying the best clusters from these gene expression datasets.

2.3 *Comparative results using functional enrichment*

For gene expression data analysis, *p*-value is used to check reliability of a clustering solution. *p*-value indicates whether an observed level of annotations for a group of genes is significant within the context of annotation for all the genes in a reference set of genes. Here the objective is to find a set of possible genes that mediate the development of a cancer. In other words, this set of genes should be responsible for specific function(s). Abnormal behaviour of this set of

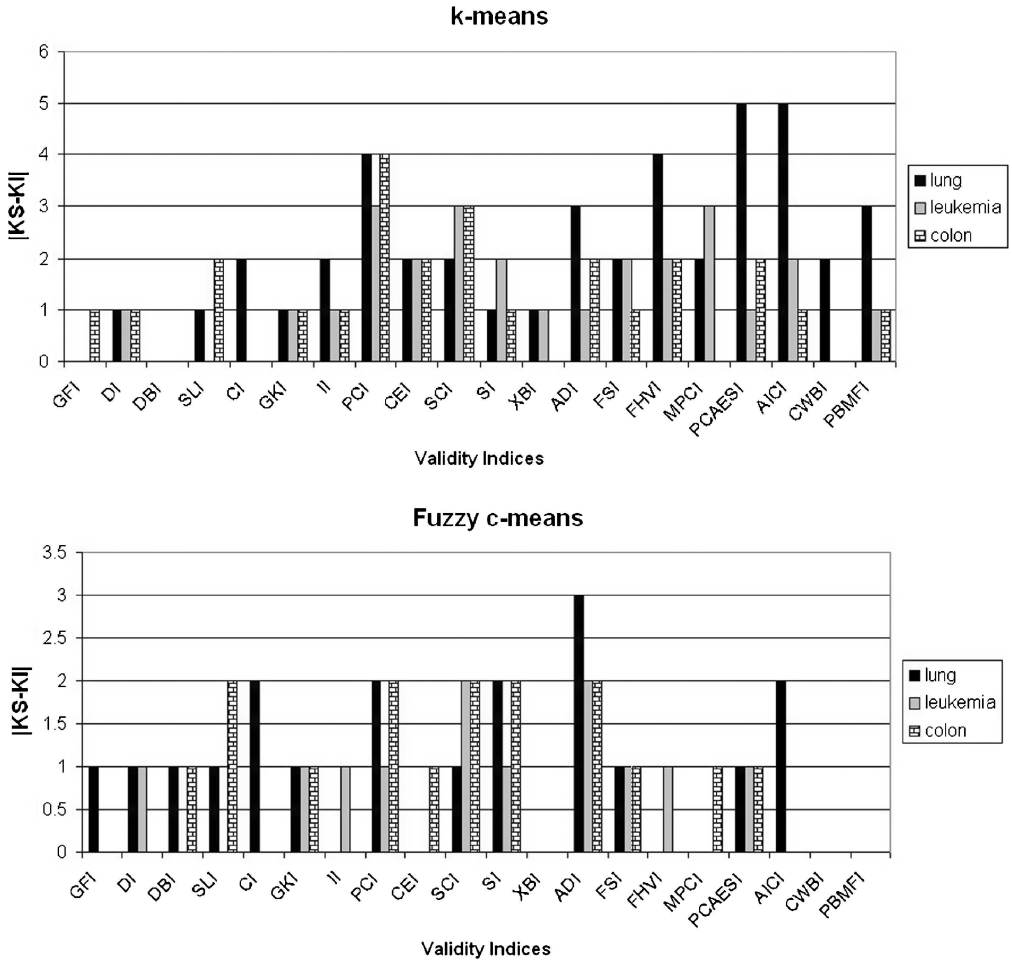

Figure 1. Comparative values of $|K_S - K_I|$ for different cluster validity indices using pathway database for *k*-means and fuzzy *c*-means clustering algorithms on different cancer datasets.

genes, or abnormal behaviour of the corresponding function(s) may lead to a cancer. That is, specific function(s) or functional category is (are) associated or enriched with this set of genes. A specific functional category is said to enriched if the corresponding *p*-value is less than a predefined threshold. A low *p*-value indicates that the genes belonging to specific functional category are biologically significant. In the present study, only functional categories with *p*-value $\leq 5\times10^{-5}$ have been considered as enriched.

We have computed the number of enriched attributes of all the altered gene sets for all the three cancer datasets. The enrichment of each GO category for these altered gene sets has been calculated by its *p*-value. Higher number of enriched attributes for a set of altered genes indicates that they belong to the same functional categories. In other words, this group of genes performs the same set of functions. That is, if one of the genes from the pool is responsible for cancer then the other genes may have a strong influence in mediating the disease.

For human lung expression data, *k*-means and fuzzy c-means 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 and fuzzy *c*-means algorithms have been found to be 507 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*.

Using *k*-means algorithm on lung expression data for $k = 10$, GFI and DBI have shown the best values, which is correctly validated by the enriched attributes (maximum value 507 for $k = 10$). From figure 2, the minimum values of $|K_E - K_I|$ (i.e. $|K_E - K_I| = 0$) have been found for GFI along with DBI. 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 GFI and DBI. The best *k*-value obtained by the *p*-value statistics of enriched attributes has differed by 1, 2, 3 from the best *k*-

value selected by the remaining cluster validity indices. Thus, we can conclude from figure 2 that our proposed index GFI along with DBI perform the best over the other existing indices for *k*-means algorithm on lung expression data.

Likewise, we have done similar experiments on colon cancer and leukemia. For colon expression data, the best indices have been found to be GFI along with DI, CI, II, XBI, FHVI, AICI, CWBI, PBMFI. The best *k*-values for the aforesaid indices have been found to be $k = 4$ for *k*-means, and $c = 4$ for fuzzy *c*-means. For leukemia dataset considered here, the best indices have been to be GFI along with DI, CI, II, XBI, FHVI, AICI, CWBI and PBMFI. The best *k*-values for these indices have been found to be $k = 9$ for k-means, and $c = 10$ for fuzzy *c*-means.

Thus, we can say that our proposed index GFI is capable of identifying the high quality enriched clusters of genes with appropriate adjustment of *k*/*c*-values on the gene expression datasets. From figure 2, it is clearly observed that functional enrichment is also able to identify the high quality biologically enriched clusters of genes and to select the cluster validity index from a group of such indices. List of results of enriched attributes (GO attributes) for lung, colon and leukemia are mentioned in (supplementary material).

2.4 *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 dataset 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 (Gibbons and Ro 2002). For good clustering algorithm with appropriate *k*/*c*-value, there should be some common attributes, depicting particular functions, of genes in a cluster.

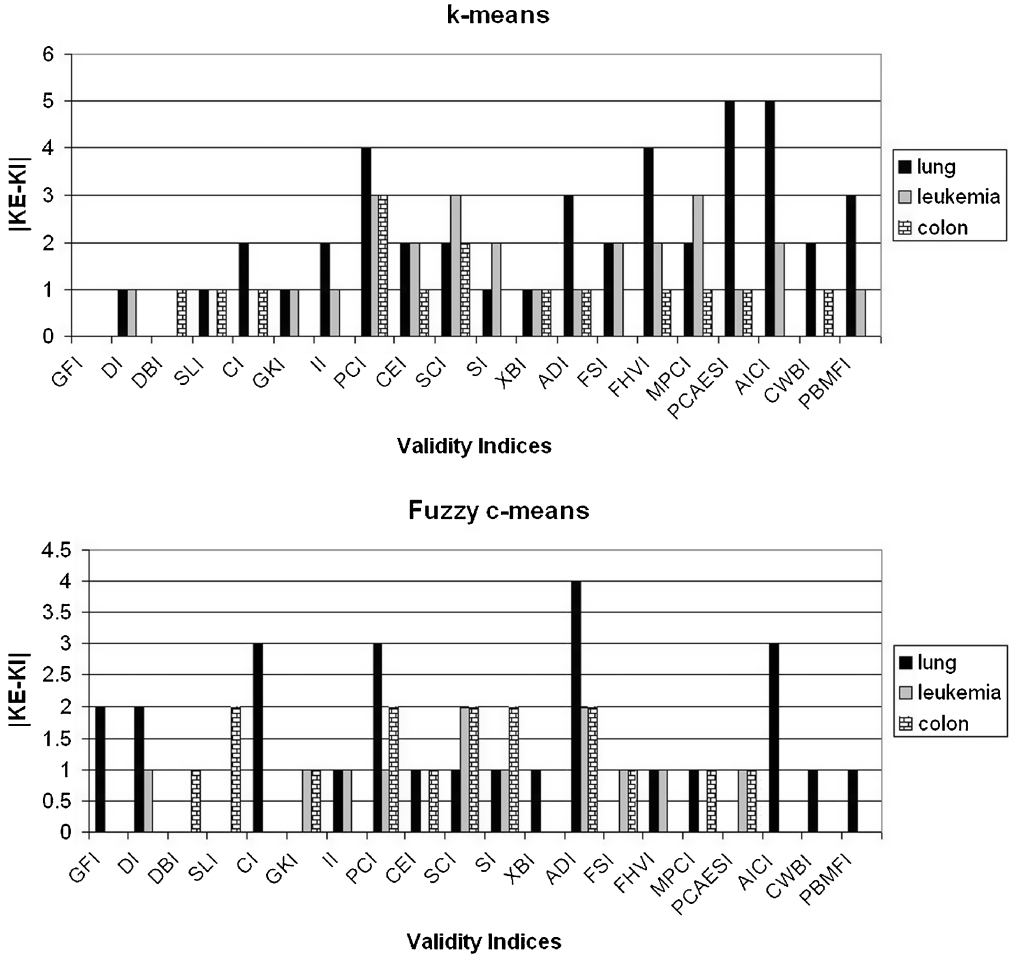

Figure 2. Comparative $|K_E - K_I|$ for different cluster validity indices using *p*-value statistics of enriched attributes for *k*-means and fuzzy *c*-means clustering algorithms on different cancer datasets.

z-score is based on mutual information between a clustering result and gene annotation data. It (see, for instance, Gibbons and Ro (2002)) 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 (Gibbons and Ro 2002).

Applying *k*-means algorithm on lung expression data, it has been found that GFI performs the best along with SI, DBI, MPCI, CWBI than the other indices. From figure 3, the minimum values (=0) of $|K_Z - K_{I_1}|$ has been found for these indices. Likewise, using fuzzy *c*-means on lung expression data, we have found that GFI along with CI, AICI and DI perform the best compared to the other validity indices (figure 3) considered here. Similarly for *k*-means and fuzzy *c*-means,

GFI, DI, GKI, II, SI, FSI, AICI and PBMFI perform the best compared to the other validity indices for colon expression data (figure 3). For leukemia dataset, we have found that GFI, XBI, PCAESI, DI, GFI and II perform the best compared to the other validity indices (figure 3).

2.5 *Comparative results using pathway database, functional enrichment and z-score altogether*

Finally, we have considered the validation using pathway database, *p*-value statistics and *z*-score altogether. Using *k*-

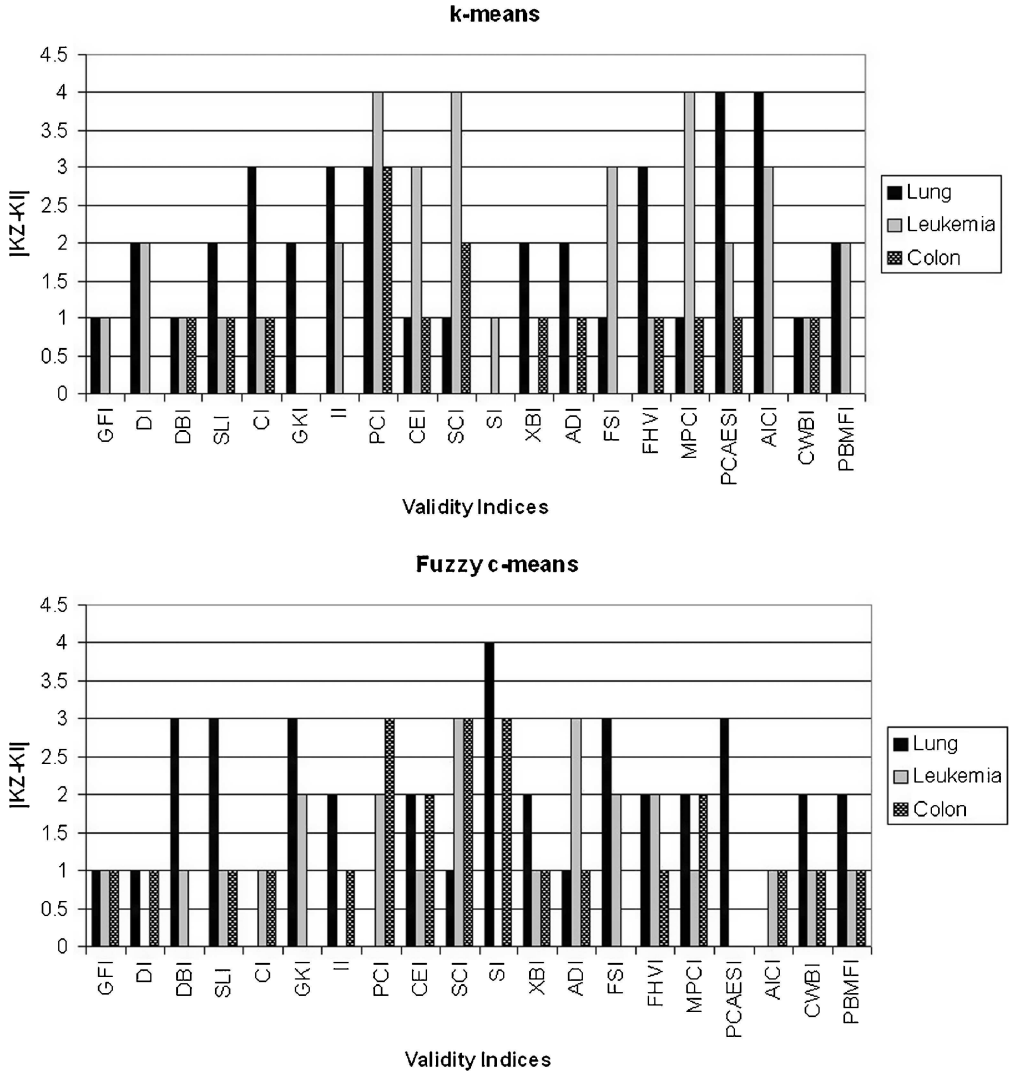

Figure 3. Comparative values of $|K_Z - K_I|$ for different cluster validity indices using *z*-score for *k*-means and fuzzy *c*-means clustering algorithms on different cancer datasets.

means algorithm on lung expression data, it has been found that GFI performs the best along with DBI than the other indices. From figure 4, the minimum values (=0) of $|K_S-K_I|+|K_E-K_I|+|K_Z-K_I|$ has been found for GFI and DBI. Likewise, using fuzzy *c*-means on lung expression data, we have found that GFI along with DBI, CI, II, CEI, SCI, SI, PCAESI,CWBI, SLI, GKI, XBI, FSI, FHVI, MPCI and PBMFI performed the best compared to the other validity indices (figure 4) considered here. Similarly for *k*-means and fuzzy *c*-means, GFI along with DI, DBI, CI, SLI, GKI, II, CEI, ADI, SI, XBI, FHVI FSI, MPCI, AICI, CWBI and PBMFI perform the best compared to the other validity indices for colon expression data (figure 4). For leukemia dataset, GFI, DBI, SLI, GKI, CI, CWBI, CEI, XBI, MPCI, AICI and PBMFI performed the best compared to the other validity indices (figure 4).

2.6 *Justification through expression profile plots*

Here we consider some genes that are among the most significant top genes of our results. The expression values of these genes have changed significantly from normal samples to diseased samples. We have provided only the expression profile plots of some important genes in lung adenocarcinoma (figure 5), human colon expression data (figure 6), human lymphocytes and plasma cell expression data (figure 7).

2.7 *Selection of some possible genes mediating certain cancers*

Here we report the genes in the altered gene sets whose expression values have deviated from normal to disease states of

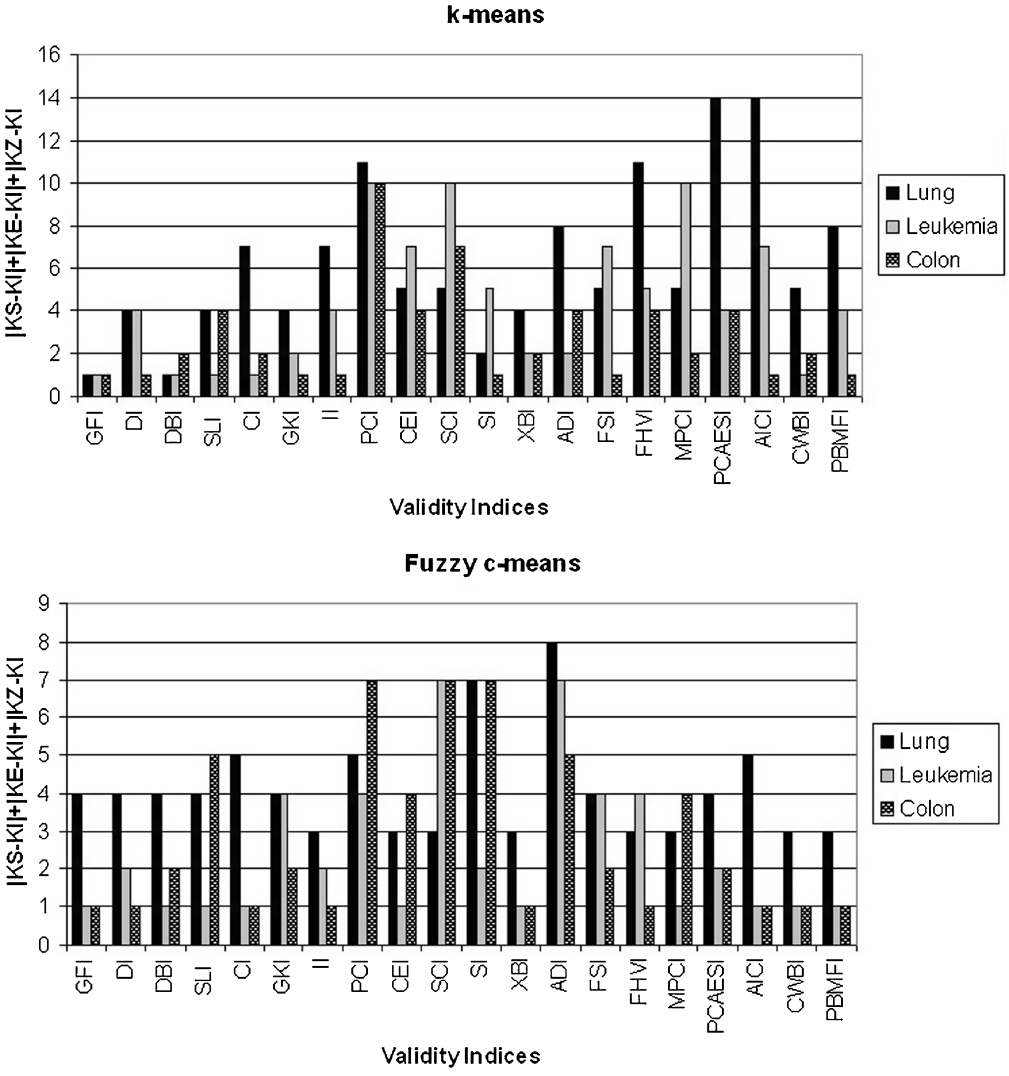

Figure 4. Comparative values of $|K_S-K_I|+|K_E-K_I|+|K_Z-K_I|$ for different cluster validity indices for *k*-means and fuzzy *c*-means clustering algorithms on different cancer datasets using the combined effect of pathway database, functional enrichment and *z*-score.

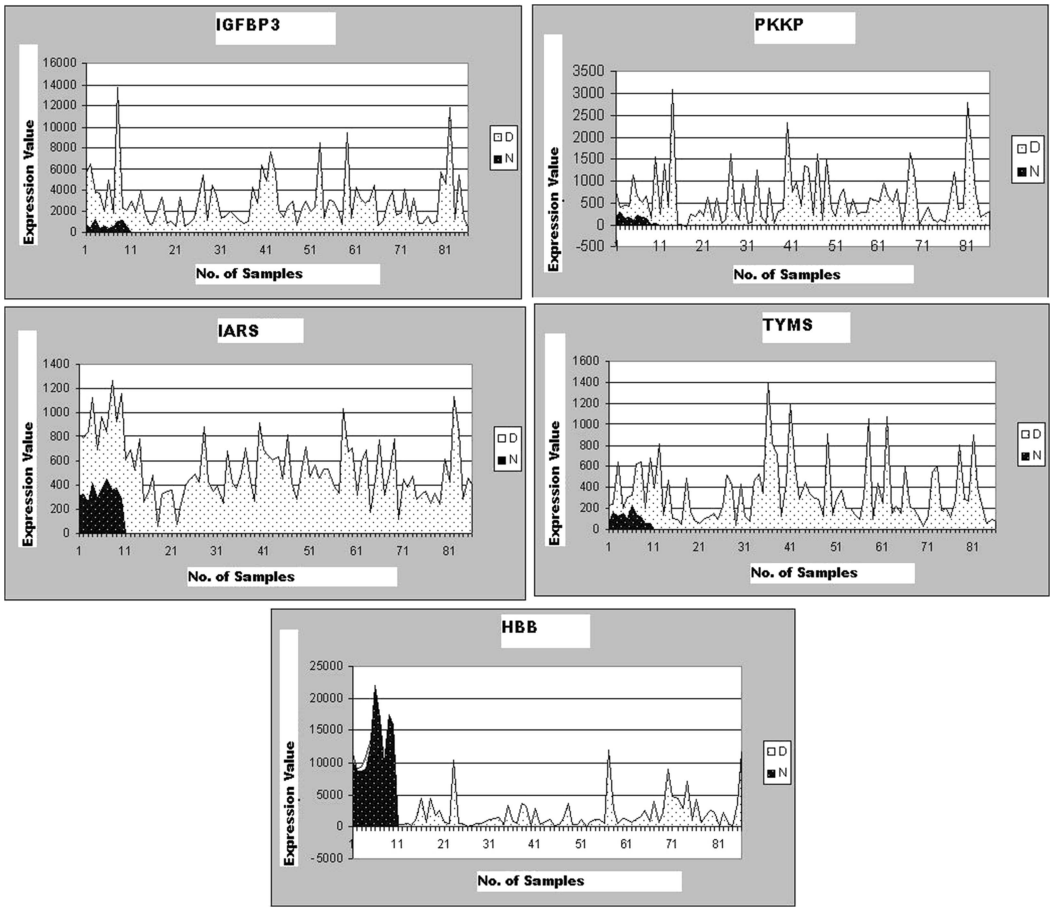

Figure 5. Expression profiles of some altered genes (IGFBP3, PFKP, IARS, TYMS, HBB) of Lung Adenocarcinoma data. 'N' represents the normal samples and 'D' indicates diseased samples.

human lung, leukemia and colon cancer datasets. Using 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. Likewise, from human colon expression, GFI has selected 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. Similarly, altered 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 GFI from human leukemia dataset. Moreover, we can say that the aforesaid altered genes have a significant role in the development of the lung cancer, leukemia and colon cancer. In other words, we can make a remark that the above mentioned genes may have a strong influence in mediating the certain cancers. It is interesting to note that the index GFI has been able to identify more cancer-mediating genes which have been supported biologically and statistically. Thus, we can draw a conclusion that GFI is able to identify biologically more significant genes than the other cluster validity indices.

2.8 *Statistical validation using t-test*

In order to validate the results statistically, we have applied *t*-test on the altered gene set identified by GFI on each dataset. For human lung expression data, we have identified some important genes like CALCA (4.02), PFKP (5.78), TYMS (3.98), IGFBP3 (6.98), IARS (5.98), HBB (7.08), HLA-B (5.42), SFTPA2 (6.89), and TNF (4.23). The number in the bracket indicates *t*-value corresponding to the gene. The *t*-values of these genes exceeds the value for $P = 0.001$. It indicates that these genes are highly significant (99.9% level of significance). Similarly, genes like IGHG3 (2.67), PRKACA (2.89), SORT1 (2.76), MEN1

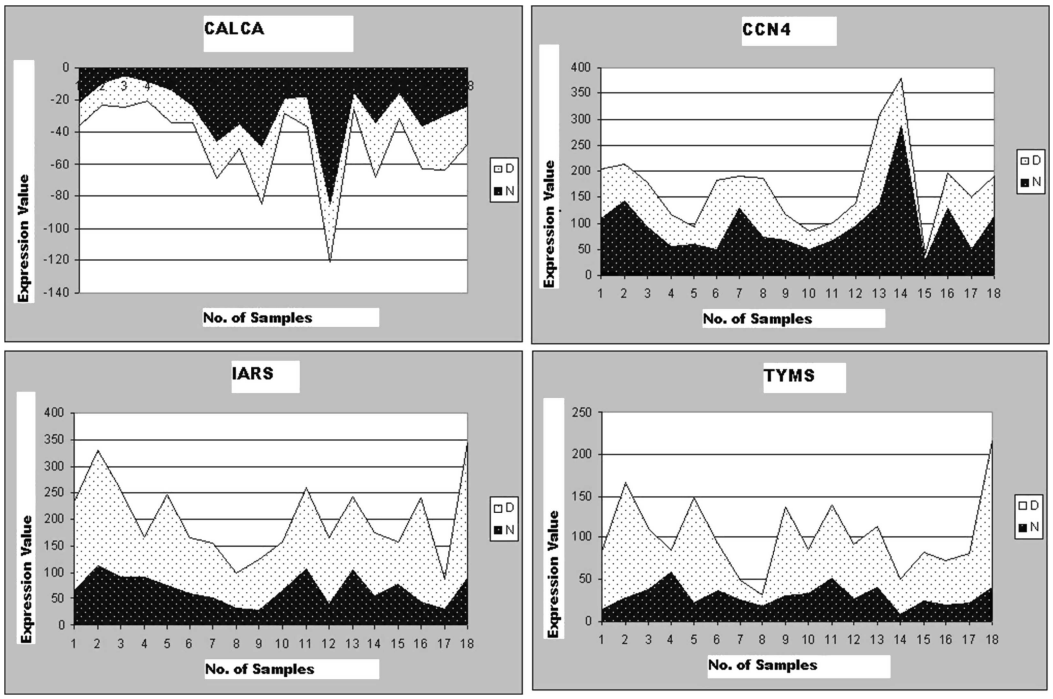

Figure 6. Expression profiles of some altered genes (CALCA, CCN4, IARS, TYMS) of Colon Cancer data. 'N' represents the normal samples and 'D' indicates diseased samples.

(3.15), SFTPA1 (2.92) and IGHM (3.25) exceeds the *t*-value for $P = 0.01$. This means that these genes are significant at the level of 99%. Likewise, RPLP0 (2.12), SMCIL1 (2.07), MGP (2.31), RNASE1 (2.43), SFTPC (2.37), and HLA-DRA (2.27) genes are important at the level of 95% significance. We have performed *t*-test for the altered genes identified by GFI for other two datasets namely colon expression and leukemia datasets. The results are shown in table 1.

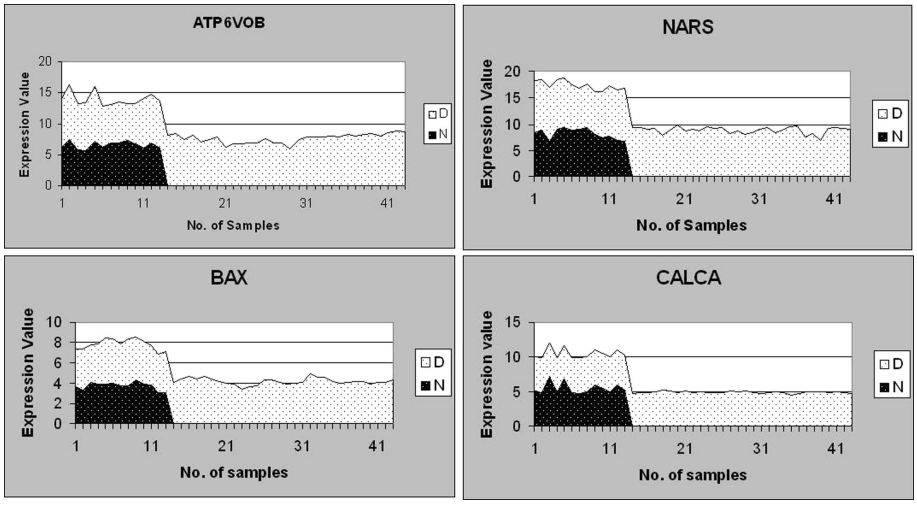

Figure 7. Expression profiles of some altered genes (ATP6VoB, NARS, BAX, CALCA) of Leukemia data. 'N' represents the normal samples and 'D' indicates diseased samples.

Table 1. Some of the significant genes and their level of significance for certain human cancer datasets resulted by GFI

Dataset	Level of significance	Genes
Lung	99.9%	CALCA, PFKP, TYMS, IGFBP3, IARS, HBB, HLA-B,SFTPA2, TNF
	99%	IGHG3, PRKACA, SORT1, SFTPA1, MEN1, IGHM
	95%	RPLP0, SMCIL1, SFTPC, HLA-DRA, MGP, RNASE1
Colon	99.9%	microtubule-associated protein 2 (MAP2), thymidylate syntase (EC 2.1.1.45) (TYMS), phosphofructokinase, platelet type (PFKP), Calcitonin (CALCA), major histocompatibility complex enhancer-binding protein (HLA-B), isoleucyl-tRNA synthetase (IARS), hemoglobin beta chain (HBB), insulin-like growth factor-binding protein-6 (IGFBP6), tumour ecrosis factor (TNF)
	99%	avin-containing monooxygenase form II (FMO2), colon carcinoma kinase-4 (CCK4), methylthioadenosine hosphorylase(MTAP)
	95%	pepsinogen C (PGC), cytochrome P450 4F2 (CYP4F2), platelet derived growth factor receptor alpha (PDGFRA), vasoactive intestinal peptide (VIP)
Leukemia	99.9%	BAX, PFKP, TYMS, NARS, BAT1, BCR, HBB, HAL, IGFBP3,CALCA, HLA-B, IARS, BRCA1
	99%	GDI2, FNTA, SDHC, KRAS, IGF1, H3F3A
	95%	CDKN2A

The results are validated using *t*-test.

3. Conclusions

In this article, a cluster validity index, called Gaussian Fuzzy Index (GFI), has been used to identify certain cancer-mediating genes. This index has recently been developed by the authors (Ghosh and De 2013). GFI involves the average fuzzy intra-cluster distances over all the clusters, and inter-cluster distances between pairs of clusters. GFI has been developed 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 namely *k*-means and Fuzzy *c*-means on three human cancer datasets, i.e. lung, colon and leukemia. Hence, we have made an analysis for identifying the important genes from a gene expression data. This concept leads to predict some possible cancer-mediating genes for certain human cancers. The results have been appropriately validated using biochemical pathways, *p*-value statistics of enriched attributes, *t*-test and using *z*-score. We have also implemented 19 different cluster validity indices to demonstrate superior capability of GFI, in identifying cancer-mediating genes, over the others. It has been shown that GFI is capable of identifying the high quality enriched clusters and finding out more number of cancer-mediating genes.

Appendix

A. Methodology

Although, GFI has already been developed in Ghosh and De (2013), we again describe it here, for the sake of the readers, along with the methodology for identification of disease mediating genes. Thus, this part actually repeats the methodology part of Ghosh and De (2013). Let us consider a set of samples $U = \{x_k \mid k=1,2,\ldots,n\}$ that are distributed in l clusters $C_1, C_2, \ldots, C_l$. These clusters have been obtained by a clustering algorithm.

A.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

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

and E defined by

$$E = \frac{1}{l} \sum_{k=1}^{l} \frac{1}{|C_k|} \sum_{x_{p \in C_k}} \mu_k(\boldsymbol{x}_p) \quad (3)$$

The term $\mu_k(\boldsymbol{c}_j)$ represents the membership value indicating the degree of belongingness of the center of j^{th} cluster C_j to k^{th} 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

$$\mu_k(\boldsymbol{c}_j) = \exp\left(-\frac{\left|\left|\boldsymbol{c}_j - \boldsymbol{c}_k\right|\right|^2}{L^2}\right) \quad (4)$$

Here $\boldsymbol{c}_k$ and $\boldsymbol{c}_j$ are the k^{th} and j^{th} 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

$$L = \max_{\substack{x_p, x \\ p' \in U \\ p \neq p'}} \left|\left|\boldsymbol{x}_p - \boldsymbol{x}_{p'}\right|\right| \quad (5)$$

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

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

The term σ_k is the diameter of k^{th} cluster C_k, and is defined as

$$\sigma_k = \max_{\boldsymbol{x}_p, \boldsymbol{x}_{p' \in C_k}} \left|\left|\boldsymbol{x}_p - \boldsymbol{x}_{p'}\right|\right| \quad (7)$$

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.

A.2 Comparative study of cluster validity indices and selection of possible disease mediating genes: The performance of GFI is compared with 19 cluster validity indices. This comparison leads to demonstrating the capability of identifying a set of good clusters and thereby selecting some possible disease mediating genes. For this comparative study, we consider the following work flow.

Step I: **Generation of clusters**: A clustering algorithm 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 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,..,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_i^N and C_j^D be i^{th} and j^{th} 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_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,…,j,…,k.Without loss of generality, we renumber the cluster C_j^D as C_i^D so that C_i^N corresponds to C_i^D.

Step III.2: **Identifying altered gene clusters**: For both normal and diseased states of data, we get k clusters, i.e., $C_1^N, C_2^N, \ldots, C_k^N$ for normal state, and similarly for diseased state, the corresponding clusters are $C_1^D, C_2^D, \ldots, C_k^D$. 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_i^N and C_j^D where $i \neq j$.Thus, we can write an altered gene set $A_i = \cup_{j=1, j\neq 1}^{k} (C_i^N \cap C_j^D)$ 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**: 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 l_1,l_2, …, l_{k-1} , l_k respectively. Thus, the score for S_k is defined as

$$S_k = \frac{1}{k} * \sum_{i=1}^{k} \frac{l_i}{|A_i|} * 100\% \quad (8)$$

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 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

$$E_k = \sum_{i=1}^{k} e_i \quad (9)$$

Higher the value of E_k, better is the chance of having common functions of the altered genes. Thus the genes together may be responsible for mediating a cancer.

Step III.5: ***z*-score:** It is based on mutual information between a clustering result 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-value (or c-value) for whichS_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

$$|K_S-K_{I_1}| + |K_E-K_{I_1}| + |K_Z-K_{I_1}| < |K_S-K_{I_2}| + |K_E-K_{I_2}| + |K_Z-K_{I_2}| \quad (10)$$

The performance of GFI has been compared extensively with 19 indices (given in table 2).

Table 2. Various cluster validity indices and the underlying notion

Cluster-Validity Index	Underlying notion	References
Dunn index (DI)	Maximization of the intercluster distances and minimization the intracluster distances. A higher Dunn index indicates better clustering. One of the drawbacks of using this, is the computational cost as the number of clusters and dimensionality of the data increase.	Dunn 1974
Davis-Bouldin index (DBI)	It is the Ratio of the sum of within-cluster scatter to between-cluster separation.	Davies and Bouldin 1979
Silhouette index (SLI)	It is based on comparison of its tightness and separation. The largest overall average silhouette indicates the best clustering (number of cluster). Therefore, the number of cluster with maximum overall average silhouette width is taken as the optimal number of the clusters.	Rousseeuw 1987
C-index (CI)	It is based on distances over all pairs of patterns from the same cluster. Hence a small value of CI indicates a good clustering.	Hubert and Schultz 1976

Table 2 (continued)

Cluster-Validity Index	Underlying notion	References
Goodman Kruskal index (GKI)	The Large values of GKI are associated with a good partition. Thus, the number of clusters that maximize the GKI index is taken as the optimal number of clusters, n. A good partition is one with many concordant and few disconcordant quadruples.	Goodman and Kruskal 1954
Isolation index (II)	The technique is based on assertion that neighboring instances in feature space often occur in the same natural cluster. A high value for this measure indicates well-separated clusters.	Pauwels and Frederix 1999
Partition Coefficient Index (PCI)	Partition Coefficient measures the amount of "overlapping" between cluster. It is based on extent of overlapping between cluster. The disadvantage of PC is lack of direct connection to some property of the data themselves. The optimal number of cluster is at the maximum value.	Bezdek 1974; Trauwaert 1988
Classification Entropy Index (CEI)	It is based on Fuzzyness of the cluster partition. The values of index close to the upper bound indicates absence of any clustering structure in the dataset or inability of the algorithm to extract it.	Bezdek 1974
Partition Index (SCI)	It is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized t hrough division by the fuzzy cardinality of each cluster.	Bensaid *et al.* 1996
Separation Index (SI)	Separation Index uses a minimum-distance separation for partition validity.	Bensaid *et al.* 1996
Xie and Beni's Index (XBI)	It represents quantification of the ratio of the total variation within clusters and the separation of clusters. Small values of XBI are expected for compact and well-separated clusters.	Xie and Beni 1991
Fukuyama and Sugeno Index (FSI)	For compact and well-separated clusters we expect small values for the index.	Fukuyama and Sugeno 1989
Fuzzy Hypervolume Index (FHVI)	The index is based on fuzzy covariance of the partition. A fuzzy partition can be expected to have a low index value if the partition is tight. An extremum for this index would ideally indicate a good partition.	Gath and Geva 1989
Alternative Dunn Index (ADI)	The aim of modifying the original Dunn's index was that the calculation becomes more simple, when the dissimilarity function between two clusters is rated in value from beneath by the triangle non equality.	Trauwaert 1988
Dave's modification of the PC index (MPCI)	It reduces the monotonic evolution tendency with cluster number. The index is equivalent to the non-fuzziness index	Dave 1996
Partition Coefficient and Exponential Separation Index (PCAESI)	The index is based on normalized partition coefficient and an exponential separation. The small or negative value of the index indicates that cluster i is not a well-identified cluster.	Wu and Yang 2005
Index Based on Akaikes information criterion (AICI)	It includes noise level, number of degrees of freedom, maximum number of cluster. The smaller the index value is, the better the clustering performance for the dataset.	Akaike 1979
Compose Within and Between scattering Index (CWBI)	The index is based on combination of average scattering for clusters with the distance functional. The index cannot handle properly arbitrary shaped clusters.	Yun and Brereton 2005
PBMF-Index (PBMFI)	It is based on fuzzy membership with optimum value for cluster center (avoidance of monotonicity).	Pakhira *et al.* 2005

References

Akaike H 1979 A Bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika* **66** 237–242

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ 1999 Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96** 6745–6750

Bandler W and Kohout LJ 1980 Fuzzy power sets and fuzzy implication operators. *Fuzzy Sets Syst.* **4** 13–30

Beer GD *et al*. 2002 Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8** 816–823

Bensaid AM, Hall LO, Bezdek J, Clarke LP, Silbiger ML, Arrington JA and Murtagh RF 1996 Validity-guided (re) clustering with applications to image segmentation. *IEEE Trans. Fuzzy Syst.* **4** 112–123

Bezdek JC 1974 On clustering validation techniques. *J. Cybernet.* **17** 58–73

Bezdek J 1981 *Pattern recognition with fuzzy objective function algorithms* (New York: Plenum Press)

Dave RN 1996 Validating fuzzy partition obtained through *c*-shells clustering. *Pattern Recogn. Lett.* **17** 613–623

Davies DL and Bouldin DW 1979 A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1** 224–227

Deborah LJ, Baskaran R and Kannan A 2010 A survey on internal validity measure for cluster validation. *IJCSES.* **1** 85–102

Dubes RC and Jain AK 1988 *Algorithms for clustering data* (Prentice Hall)

Dunn JC 1974 Well separated clusters and optimal fuzzy partitions. *J. Cybern.* **4** 95–104

Fukuyama Y and Sugeno M 1989 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

Gath I and Geva AB 1989 Unsupervised optimal fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 773–781

Ghosh A and De RK 2013 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; in *Pattern Recognition and Machine Intelligence (Proc. PReMI 2013), Kolkata, India, LNCS 8251 Proceedings of the 5th International Conference on Pattern Recognition and Machine Intelligence (PReMI 2013), India* (eds) P Maji, A Ghosh, MN Murty, K Ghosh and SK Pal, pp 680–687

Ghosh A, Dhara BC and De RK 2013 Comparative analysis of cluster validity indices in identifying some possible genes mediating certain cancers. *Mol. Inf.* **32** 347–354

Gibbons FD and Ro FP 2002 Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* **12** 1574–1581

Goodman L and Kruskal W 1954 Measures of associations for cross-validations. *J. Am. Stat. Assoc.* **49** 732–764

Gutierrez NC, Ocio EM, delas Rivas J, Maiso P, Delgado M, Ferminan E, Arcos MJ, Sanchez ML, *et al*. 2007 Gene expression profiling of B lymphocytes and plasma cells from Waldenstroms macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals. *Leukemia.* **21** 541–549

Hubert L and Schultz J 1976 Quadratic assignment as a general data-analysis strategy. *Br. J. Math. Stat. Psychol.* **29** 190–241

Pakhira M, Bandyopadhyay S and Maulik U 2005 A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets Syst.* **155** 191–214

Pauwels EJ and Frederix G 1999 Finding salient regions in images: nonparametric clustering for image segmentation and grouping. *Comput. Vis. Image Underst.* **75** 73–85

Rousseeuw PJ 1987 A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65

Trauwaert E 1988 On the meaning of Dunn's partition coefficient for fuzzy clusters. *Fuzzy Sets Syst.* **25** 217–242

Tripathy BC, Sen M and Nath S 2012 I-convergence in probabilistic *n*-normed space. *Soft. Comput.* **16** 1021–1027

Wu K and Yang M 2005 A cluster validity index for fuzzy clustering. *Pattern Recogn. Lett.* **26** 1275–1291

Xie XL and Beni GA 1991 Validity measure for fuzzy clustering. *IEEE Trans. PAMI.* **3** 841–846

Yun XU and Brereton GR 2005 A comparative study of cluster validation indices applied to genotyping data. *Chemom. Intell. Lab. Syst.* **78** 30–40

Zadeh LA 1965 Fuzzy sets. *Inf. Control.* **8** 338–353

Zadeh LA 1972 A fuzzy-set-theoretic interpretation of linguistic hedges. *J. Cybern.* **2** 4–34

Zadeh LA 1997 Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst.* **90** 111–127