2015 International Conference on Computational Science and Computational Intelligence

GO-based gene expression cluster validation

Thanh Le
Department of Radiology & Biomedical Imaging, University of California, San Francisco, USA
lntmail@yahoo.com

Abstract—Fuzzy C-Means (FCM) algorithm has been widely used in cluster analysis of gene expression data. It can converge rapidly and provide more information regarding relationships within the data thanks to the usage of fuzzy sets to represent the degrees of cluster membership for every data point. However, FCM has a shortcoming in that it requires a priori specification of cluster number and cluster validation. Most cluster validity indices are based on the data themselves and may not be applicable for gene expression data. In this paper, we propose a Bayesian method using Gene Ontology (GO) annotations for gene expression cluster validation. We show that our method outperforms popular validity indices on gene expression datasets.

Keywords—fuzzy c-means; Bayesian model; cluster validation; gene ontology; semantic similarity

I. INTRODUCTION

Cluster analysis groups data points based on their similar properties and can help to discover patterns and correlations in large datasets. Fuzzy C-Means (FCM, Bezdek 1981) is a popular clustering algorithm based on partitioning approach with fuzzy cluster boundaries and fuzzy sets that associate each data point with one or more clusters. A good FCM's clustering solution maximizes both the compactness of data points within a cluster and the discrimination between clusters. An advantage of FCM is that it converges rapidly, however, like most partitioning clustering algorithms, it depends strongly on the initial parameters and requires estimation of the number of clusters. For some initial values, FCM may converge to a global optimum, but for others, it may get stuck in a local optimum. In addition, during the clustering process, the optimization of the compactness and separation of a fuzzy partition may be inconsistent with the optimal number of clusters in the dataset. For these reasons, final clustering results require validation to assess how good the fuzzy partition is, if better fuzzy partitions exist, and, when not known a priori, the optimal number of clusters in the dataset.

Several cluster validity index functions have been proposed. Bezdek [1] measured performance using partition entropy and the overlap of adjacent clusters. Fukuyama and Sugeno [2] combined the FCM objective function with the separation factor, while Xie and Beni [3], integrated the Bezdek index [1] with the cluster separation factor. Rezaee et al. [4] combined the compactness and separation factors, and Pakhira et al. [5] combined the same two factors where the separation factor was normalized. Rezaee [6] proposed a new cluster index in which the two factors are normalized across the range of possible numbers of clusters. Recently, Le et al. [8] proposed a Bayesian method for fuzzy cluster validation where the possibility model of a fuzzy partition is approximated by a probability one which is used to compute the goodness-of-fit of the fuzzy partition against the data.

Common drawback of these methods is that they are solely based on the internal properties of the data. When applied in gene expression data analysis, they do not allow incorporation of prior biological knowledge, such as Gene Ontology (GO), making their results less biology relevant. In this paper, we describe fzBGO, a Bayesian cluster validation method that applies GO annotations in validating fuzzy cluster partition. Instead of computing the compactness and separation factors, fzBGO utilizes GO based semantic similarity measure to generate a biological meaning fuzzy partition (BMFP) and construct the probabilistic model of BMFP. A log-likelihood estimator is then applied to measure the model goodness-of-fit. By using both the possibility and probability models to represent the data distributions, fzBGO is appropriate for not only artificial data where the data distributions usually follow a standard model, but also for real-world datasets, particularly gene expression data, that lack a standard distribution. The rest of the paper is organized as follows: in section II, we describe the principle methods and our model for the cluster validation problem. In section III, we demonstrate the performance of fzBGO on gene expression datasets. In the last section, we summarize our conclusions and potential improvements for future work.

II. METHODS

A. Fuzzy C-Means algorithm (FCM)

Cluster analysis decomposes a set of objects into clusters based on dissimilarity. In analysis of gene expression microarray datasets, we require the clustering to allow a single gene to belong to more than one cluster, because one gene may participate in multiple biological processes. FCM was chosen for this work because it provides both an effective mechanism for cluster validation methods and allows genes to belong to multiple clusters.

Given a dataset X = $\{x_i \in R^p, i=1\ldots n\}$, where n, n>0, is the number of data points, and p, p>0, is the dimension of the data space of X. Let c, $c \in N$ and $2 \le c \le n$, be the number of clusters in X. Denote V = $\{v_k \in R^p, k=1\ldots c\}$ as the set of center points of c clusters in the fuzzy partition; U = $\{u_{ki} \in [0,1], i = 1\ldots n, k = 1\ldots c\}$ as the partition matrix, where u_{ki} is the membership degree of the data point x_i to the k^th cluster, and

$$\sum_{k=1}^{c} u_{ki} = 1,\ i = 1\ldots n \cdot \qquad (1)$$

FCM divides X into c clusters by minimizing the objective function [1, 8],

$$J_m(U,V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki}^{m} d^2(x_i, v_k) \to \min, \qquad (2)$$

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.34

where d(.) is a distance function, defined using Euclidean norm,

$$d^2(x,y)=\sum_{i=1}^{p}\left(x^i-y^i\right)^2, \tag{3}$$

and m, 1≤m<∞, the fuzzifier factor that defines the degree of fuzziness in membership functions.

Minimizing J_m in (2) with respect to (1), we obtain an estimated model of U and V as:

$$v_k=\sum_{i=1}^{n}u_{ki}^{m}x_i\Big/\sum_{i=1}^{n}u_{ki}^{m}, \tag{4}$$

$$u_{ki}=\left(\frac{1}{\left\|x_i-v_k\right\|^2}\right)^{\frac{1}{1-m}}\Bigg/\sum_{j=1}^{c}\left(\frac{1}{\left\|x_i-v_j\right\|^2}\right)^{\frac{1}{1-m}}. \tag{5}$$

FCM uses an iteration process to estimate the solution of (4) and (5). This process is iterated until convergent where, given T, $T>0;\ \forall t>T$,

$$\exists\varepsilon_u>0:\left\|U_{t+1}-U_t\right\|=\max_{k,i}\left\{\left\|u_{ki}(t+1)-u_{ki}(t)\right\|\right\}<\varepsilon_u, \tag{6}$$

or,

$$\exists\varepsilon_v>0:\left\|V_{t+1}-V_t\right\|=\max_{k}\left\{\left\|v_k(t+1)-v_k(t)\right\|\right\}<\varepsilon_v. \tag{7}$$

FCM can converge rapidly, and provides soft partitions applicable to many real-world applications. However, it is unable to determine the optimal number of clusters as well as to validate the clustering solution against the data.

B. *Cluster validity indices*

To validate a fuzzy partition, traditional cluster validity indices use two criteria, (i) *compactness*, which measures the closeness of cluster elements typically using the variance. Because variance indicates how different the members are, a low value of variance is an indicator of closeness, and (ii) *separation*, which computes the "distance" between two different clusters, e.g., the distance between representative objects of two clusters. This measure has been widely used due to its computational efficiency and its effectiveness for hyper sphere-shaped clusters.

1) *PC index*

Partition coefficient (PC) index was proposed by Bezdek [1] as in (8). It indicates the average relative amount of shared membership between pairs of fuzzy subsets in U, by combining into a single number, the average content of pairs of fuzzy algebraic products. The index values range from [1/c, 1].

$$V_{PC}=\frac{1}{n}\sum_{k=1}^{c}\sum_{i=1}^{n}u_{ki}^{2}. \tag{8}$$

An optimal cluster number c can be found by solving,

$$V_{PC}(c_{opt})=\max_{2\le c\le n}\{V_{PC}(c)\}.$$

2) *FS index*

Fukuyama-Sugeno (FS) validity index was proposed by Fukuyama and Sugeno [2] as

$$V_{FS}=J-\sum_{k=1}^{c}\sum_{i=1}^{n}u_{ki}^{m}\left\|v_k-\bar{v}\right\|^2, \tag{9}$$

where, $\bar{v}=\sum_{k=1}^{c}v_k\big/c$. An optimal number of clusters can be found by solving V_{FS}→min.

3) *XB index*

XB validity index was proposed by Xie and Beni as in (10). The numerator indicates the compactness of the fuzzy partition, while the denominator indicates the strength of the separation between clusters. A good partition produces a small value for the compactness, and well-separated $\{v_i\}$ will produce a high value for the separation. An optimal c therefore is found by solving V_{XB}→min.

$$V_{XB}=\frac{\sum_{k=1}^{c}\sum_{i=1}^{n}u_{ki}^{m}\times\left\|x_i-v_k\right\|^2}{n\times\min_{k,l}\left\|v_k-v_l\right\|^2}. \tag{10}$$

4) *CWB index*

Compose Within and Between scattering (CWB) validity index was proposed by Rezaee et al. [4].

$$V_{CWB}=\alpha Scat(c)+Dis(c), \tag{11}$$

where α is a weighting factor equal to $Dis(c_{max})$.

The average scatter function, Scat(.), is defined as

$$Scat(c)=\frac{\sum_{k=1}^{c}\left\|\sigma(v_k)\right\|}{c\times\left\|\sigma(X)\right\|}, \tag{12}$$

where $\left\|x\right\|=\left(x^Tx\right)^{1/2}$, $\sigma(X)=\frac{1}{n}\sum_{i=1}^{n}\left\|x_i-\bar{x}\right\|^2$, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

The cluster distance function, Dis(.), is defined as

$$Dis(c)=\frac{D_{max}}{D_{min}}\sum_{k=1}^{c}\left(\sum_{l=1}^{c}\left\|v_k-v_l\right\|\right)^{-1}, \tag{13}$$

where $D_{min}=\min_{k,l}\|v_k-v_l\|$ and $D_{max}=\max_{k,l}\|v_k-v_l\|$. The Scat function indicates the average of the scattering variation within the clusters. A small value for this term indicates a compact partition. The Dis function indicates the total scattering separation between the clusters, it is influenced by the geometry of the cluster centroids and increases with the number of clusters. An optimal number of clusters c is found by solving V_{CWB}→min.

5) PBMF index

PBMF validity index is a fuzzy version of the PBM validity index proposed by Pakhira, Bandyopadhyay and Maulik [5] as

$$V_{PBMF} = \left(\frac{1}{c} \frac{E_1}{J} D_c \right)^2, \tag{14}$$

$$E_1 = \sum_{i=1}^{n} u_{1i} \left\| x_i - \bar{x} \right\|, \tag{15}$$

where $D_c = \max_{k,l}\|v_k - v_l\|$. The value of V_{PBMF} decreases as the number of clusters c increases. An optimal number of clusters can be found by solving V_{PBMF}→max.

6) BR index

The validity index of Rezaee B. (BR) [6] uses both the compactness and separation criteria normalized across clustering partitions using possible numbers of clusters in a given range. The index is defined as

$$V_{BR} = \frac{Sep(c)}{\max\limits_{c}\{Sep(c)\}} + \frac{J(c)}{\max\limits_{c}\{J(c)\}}, \tag{16}$$

where $Sep(c) = \frac{2}{c(c-1)} \sum_{k \neq l}^{c} S_{rel}(v_k, v_l).$

The similarity of two fuzzy sets is defined as

$$S_{rel}(v_k, v_l) = \sum_{i=1}^{n} S(x_i : v_k, v_l) \times h(x_i), \tag{17}$$

where $S(x_i : v_k, v_l) = \min(u_{ki}, u_{li})$, $h(x_i) = -\sum_{k=1}^{c} u_{ki} \times \log_a(u_{ki}) \cdot$

Because V_{BR} is a sum of compactness and separation factors, the smaller it is, the better the fuzzy partition is. An optimal number of clusters c therefore can be found by solving V_{BR}→ min.

C. GO-based semantic similarity measures

1) Gene Ontology (GO)

The Gene Ontology (GO) [16] is a hierarchy of biological terms using a controlled vocabulary that includes three independent ontologies for biological process (BP), molecular function (MF) and cellular component (CC). Standardized terms known as GO terms describe roles of genes and gene products in any organism. GO terms are related to each other in the form of parent-child relationships. A gene product can have one or more molecular functions, participating in one or more biological processes, and can be associated with one or more cellular components [17]. As a way to share knowledge about functionalities of genes, GO itself does not contain gene products of any organism. Rather, expert curators specialized in different organisms annotate biological roles of gene products using GO annotations. Each GO annotation is assigned with an evidence code that indicates the type of evidences supporting the annotation (TABLE I.).

2) Semantic similarity

GO is structured as directed acyclic graphs (DAGs) in which the terms form nodes, and the two kinds of semantic relationships, *"is-a"* and *"part-of"*, form edges [19]. "is-a" is a simple class-subclass relation, where A is-a B means that A is a subclass of B. 'part-of ' is a partial ownership relation; C part-of D means that whenever C is present, it is always a part of D, but C need not always be present. The structure of DAG allows assigning a metric to a set of terms based on the likeliness of their meaning content which is used to measure semantic similarity between terms. Multiple GO based semantic similarity measures have been recently developed [18, 19], and are increasingly used to evaluate the relationships between proteins in protein-protein interactions, or co-regulated genes in gene expression data analysis. Among of the existing semantic similarity measurement methods, that of Resnik is most appropriate for gene expression data analysis because it is strong correlated with gene sequence similarities and gene expression profiles [18]. However, Wang et al. [19] had shown that the Resnik's method has a drawback in that it uses only the information content derived from annotation statistics which is not suitable for measuring semantic similarity of GO terms. We therefore propose to use Wang's method for GO semantic similarity measurement in this work.

3) GO term semantic similarity [19]

For each term A, a semantic value, S(A), is computed as in

$$SV(A) = \sum_{t \in T_A} S_A(t), \tag{18}$$

where T_A is a set of terms including the term A and its ancestors, and $S_A(.)$ is the sematic value regarding the term A, defined as:

$$S_A(t) = \begin{cases} 1, t \equiv A \\ \max\{w_t^u \times S_A(u), u \in ChildrenOf(t)\}, t \neq A \end{cases}, \tag{19}$$

where w_t^u is the semantic contribution factor for edge connecting term t with its child, term u. The semantic similarity between two terms, A and B, is defined as

$$S_{GO}(A,B) = \frac{\sum\limits_{t \in T_A \cap T_B} \left(S_A(t) + S_B(t)\right)}{SV(A) + SV(B)} \tag{20}$$

4) GO term set semantic similarity

Given the two sets of GO terms, G1 and G2, the semantic similarity between G1 and G2 is defined as

$$Sim(G1,G2) = \frac{\sum\limits_{g1 \in G1} Sim(g1,G2) + \sum\limits_{g2 \in G2} Sim(g2,G1)}{|G1| + |G2|}, \tag{21}$$

where Sim(t,T) is the similarity between the term t and the term set T, defined as

$$\mathrm{Sim}(t,T) = \max_{u \in T}\{S_{GO}(t,u)\} \cdot \tag{22}$$

D. *Bayesian cluster validation method using GO (fzBGO)*

fzBGO validates a clustering solution by generating a biological meaning fuzzy partition (BMFP) model using a GO based semantic similarity measure. The model is approximated by a probabilistic one which is then used with a log likelihood estimator to measure the goodness-of-fit of the clustering solution against the data.

1) *BMFP model of clustering solution*

Given a crisp clustering solution θ^c, $\theta^c = \{M, V\}$ where V represents the cluster centers and M, $M=\{m_{ki}\}$; $m_{ki} \in \{0,1\}$; k=1…c; i=1…n, is the crisp c-partition matrix representing the membership of the data points to c clusters. If the given clustering solution is a fuzzy one, say θ^f, then we can apply the defuzzification method [9] to generate θ^c from θ^f.

To generate BMFP model of θ^c, θ^c is first mapped onto GO terms space. Each data object x_i, i=1…n, is corresponding with a vector of GO annotations x_i^{GO}, and a vector of degree of belief (DOB), x_i^{CF}, defined as in TABLE I. , where x_{ij}^{CF} is the DOB of the term x_{ij}^{GO} annotated to x_i. GO annotations for each cluster v_k, v_k^{GO}, k=1…c, is determined using GO annotations of its members:

$$v_k^{GO} = \bigcup_{x \in v_k} x^{GO}, \tag{23}$$

where the DOB, v_{kt}^{CF}, of an annotation $v_{kt}^{GO} \in \left|v_k^{GO}\right|$ is defined as:

$$v_{kt}^{CF} = \mathrm{mean}\left\{x_{.j}^{CF}, x_{.} \in v_k, x_{.j}^{GO} = v_{kt}^{GO}\right\} \tag{24}$$

We now can apply (21) and (24) in lieu of (20) to compute the semantic similarity between cluster v_k and data point x_i, $\mathrm{Sim}'(v_k, x_i)$, using their GO annotation sets. The GO based distance between v_k and x_i is defined as:

$$d_{GO}^2(v_k, x_i) = d^2\left(1, \mathrm{Sim}'(v_k, x_i)\right) \tag{25}$$

The BMFP model, θ^b, of θ^c is then generated by using GO based distance, d_{GO}, in the model of (4) and (5) with θ^c.

TABLE I. DEGREES OF BELIEF OF GO ANNOTATION EVIDENCE

Evidence code	Degree of belief
EXP	1.0
IDA, IPI, TAS	0.9
IMP, IGI, IEP	0.7
ISS, ISO, ISA, ISM	0.4
IGC	0.2
IBA, IBD, IKR, IRD, RCA	0.3
NAS, IC, ND, NR	0.0
IEA	0.1

2) *BMFP model validation*

We use the method of Le et al. [8] for BMFP validation. Let θ, θ = {U,V} be the BMFP model. The likelihood of the model against the data is measured as

$$\begin{aligned} L(\theta \mid X) &= L(U,V \mid X) \\ &= \prod_{i=1}^{n} P(x_i \mid U,V) = \prod_{i=1}^{n}\sum_{k=1}^{c} P(v_k) \times P(x_i \mid v_k). \end{aligned} \tag{26}$$

The log likelihood estimator is then computed as

$$\log(L) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{c} P(v_k) \times P(x_i \mid v_k)\right) \to \max \cdot \tag{27}$$

Equation (27) is used as a validation measure of BMFP model. Because θ is a possibility based model, a possibility to probability transformation method [7] is used to generate the probabilistic model of θ. For each cluster v_k, k=1…c, a probability distribution $\{p_{ki}\}_{i=1\ldots n}$ is derived from the possibility distribution $\{u_{ki}\}_{i=1\ldots n}$ [8]. Then, the following statistics at v_k are computed:

$$\sigma_k = \sum_{i=1}^{n} p_{ki} \times d_{GO}^2(x_i, v_k), \tag{28}$$

$$P(x_i \mid v_k) = \left((2\pi)^{1/n} \times \sigma_k \times e^{\frac{\|x_i - v_k\|^2}{2\sigma_k^2}}\right)^{-1}, \tag{29}$$

$$P_s(v_k) = \sum_{i=1}^{n} P(x_i \mid v_k) \Bigg/ \sum_{i=1}^{n}\sum_{q=1}^{c} P(x_i \mid v_q), \tag{30}$$

where $P(x_i|v_k)$ indicates the conditional probability of x_i given v_k, i=1…n, k=1…c; σ_k and $P_s(v_k)$ are the variance and posterior probability of v_k respectively. Computation of $P_s(v_k)$, as in (30), is similar to that of [21] where $P(v_k)$ is approximated by 1/c. The prior probability of v_k, $P_p(v_k)$, k=1…c, can be estimated using prior probabilities of the GO terms, T, used in the dataset. For a given term t, t∈T,

$$P_p(t) = \frac{\mathrm{freq}(t)}{\mathrm{freq}(T)}, \tag{31}$$

$$P(t \mid v_k) = \frac{\underset{u \in v_k^{GO}, u=t}{\mathrm{sum}}\ (v_{ku}^{CF})}{\mathrm{sum}(v_k^{CF})}, \tag{32}$$

where $P_p(t)$ is the prior probability of the term t, $P(t|v_k)$ is the conditional probability of term t given v_k. The prior probability of v_k, k=1…c, is computed as:

$$\begin{aligned} P_p(v_k) &= \sum_{t \in v_k^{GO}} P(t, v_k) \times P_p(t) \\ &= \sum_{t \in v_k^{GO}} P(t \mid v_k) \times P(v_k) \times P_p(t). \end{aligned} \tag{33}$$

3) ***fzBGO*** *algorithm*

- Input
  X = $\{xi\}_{i=1\ldots n}$: Gene expression dataset
  [cmin, cmax] : Range of cluster numbers to validate
- Output
  C_{opt} : Optimal number of clusters
  G_{opt} : GO annotations on clustering solution

<u>*Steps*</u>

1) Set $C_{opt} = c_{min}$
2) For each value of c in $[c_{min}, c_{max}]$
3) Generate a crisp cluster solution θ^c of c clusters
4) Generate BMFP model θ of θ^c using Section II.D.1
5) Evaluate θ using Section II.D.2
6) If the current clustering solution is better than the current optimal one: set $C_{opt} = c$, $G_{opt}=V^{GO}$
7) Return $\{C_{opt}, G_{opt}\}$

III. Experimental results

To evaluate the performance of fzBGO, we used three gene expression datasets: Yeast [12], Yeast-MIPS [13, 14] and RCNS [10]. These datasets contain classification information, useful for comparing cluster validation methods. We compared performance of fzBGO with fzBLE [8] and six cluster validity indices: PC, FS, XB, CWR, PBMF and BR [1-6]. We used GOSemSim [20], an R package, to compute semantic similarity between GO terms, and applied in fzBGO for term set distance measure. The GO annotation knowledge bases for Yeast [22] and RAT [23] were applied on the Yeast, Yeast-MIPS and RCNS datasets respectively. While only GO-BP ontology was used for Yeast and Yeast-MIPS datasets, both the CC and BP ontologies were applied for RCNS dataset.

For each dataset, we ran the standard FCM algorithm five times with the fuzzifier factor, m, set to 1.17 [11] and the partition matrix initialized randomly. In each case, the best fuzzy partition was then selected to run fzBGO and other cluster validation methods to search for the optimal number of clusters between 2-13 for Yeast dataset, or 2-10 for the other datasets, and to compare this with the known number of clusters. We repeated the experiment 20 times and averaged the performance of each method.

A. Yeast and Yeast-MIPS datasets

The yeast cell cycle data showed expression levels of approximately 6000 genes across two cell cycles comprising 17 time points [12]. By visual inspection of the raw data, Cho et al. [12] identified 420 genes that show significant variation. From the subset of 420 genes, Yeung et al. [13] selected 384 genes that achieved peak values in only one phase, and obtained five standard categories by grouping into one category genes that peaked during the same phase. Among the 384 selected genes, Tavazoie et al. [14], through a search of the protein sequence database, MIPS [15], found 237 genes that can be grouped into 4 functional categories: DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins. The functional annotations show the cluster structure in the dataset.

TABLE II. shows the algorithm performance on the Yeast dataset which consists five groups of genes that peaked in the same phase. Only fzBGO and fzBLE correctly identified the number of clusters. Among the other methods, PC and CWB were the two methods having better results.

For the Yeast-MIPS dataset (TABLE III.), fzBGO and fzBLE again were the only two methods that successfully detected the number of clusters. While fzBLE works based on gene expression levels, fzBGO works based on GO biological process (BP) annotations. This result also shows that, under the biological functionality context, GO-BP annotations are strong correlated with gene-gene differential co-expression patterns. In other words, we may utilize either gene-gene co-expression patterns in gene expression data or GO-BP annotations to search for genes of similar function.

B. RCNS (Rat Central Nervous System) dataset

The RCNS dataset was obtained by reverse transcription-coupled PCR designed to study the expression levels of 112 genes over nine time points during rat central nervous system development [10]. Wen et al. [10] classified these genes into six groups based on expression patterns. Four of which are composed of biologically functionally related genes. These four classes are external criterion in this dataset.

TABLE II. ALGORITHM PERFORMANCE ON THE YEAST DATASET

#c	fzBGO	PC	FS	XB	CWB	PBMF	BR	fzBLE
2	-4832.7	0.93	-85.1	**0.21**	8.37	**1.21**	2.00	-2289.8
3	-4832.5	0.94	-157.3	0.22	4.76	0.69	1.05	-2296.5
4	-4832.8	**0.95**	-191.8	0.22	**4.06**	0.56	0.72	-2305.3
5	**-4831.5**	0.91	-187.1	1.05	13.68	0.41	0.67	**-2289.3**
6	-4831.9	0.90	-196.7	0.99	13.86	0.31	0.62	-2296.3
7	-4833.2	0.88	-198.3	1.06	15.49	0.24	0.57	-2296.6
8	-4833.6	0.86	-201.8	1.10	16.96	0.21	0.51	-2299.4
9	-4832.4	0.85	-205.1	1.23	20.25	0.17	0.48	-2299.4
10	-4832.2	0.84	-208.6	1.20	20.78	0.15	0.45	-2302.8
11	-4832.6	0.83	-209.4	1.17	21.15	0.13	0.43	-2300.3
12	-4832.7	0.83	-213.5	1.23	23.04	0.12	0.40	-2307.6
13	-4832.2	0.83	**-215.2**	1.30	25.41	0.10	**0.39**	-2310.8

TABLE III. ALGORITHM PERFORMANCE ON THE YEAST-MIPS DATASET

#c	fzBGO	PC	FS	XB	CWB	PBMF	BR	fzBLE
2	-2288.4	0.90	25.43	0.35	16.76	0.72	2.00	-1316.5
3	-2286.8	**0.91**	-32.85	**0.30**	**10.16**	**0.80**	1.25	-1317.4
4	**-2283.9**	0.82	-39.49	2.53	39.84	0.54	1.32	**-1304.0**
5	-2285.1	0.83	-54.50	2.43	35.00	0.36	0.96	-1308.7
6	-2286.3	0.82	-59.89	2.35	35.45	0.27	0.83	-1310.0
7	-2286.8	0.81	-65.49	2.36	38.88	0.24	0.73	-1315.4
8	-2287.5	0.80	-67.68	2.50	43.95	0.20	0.67	-1315.2
9	-2288.3	0.81	-72.32	2.29	41.21	0.17	0.61	-1321.2
10	-2289.0	0.82	**-74.79**	2.04	37.62	0.14	**0.56**	-1324.2

We first ran fzBGO using GO-CC annotations and compared with the other methods. The results are shown in TABLE IV. Only fzBGO and fzBLE identified six clusters in the dataset, corresponding to the six waves, two of which are invariant, in the analysis results of Wen et al. [10]. We then reran fzBGO using GO-BP annotations and compared with the other methods. Results are shown in TABLE V. fzBGO identified four clusters in the dataset that are corresponding to the four stages in the rat central nervous system developmental process. This result of fzBGO again shows that GO-BP annotations are strong correlated with gene-gene differential co-expression patterns. It also shows that GO-BP annotations are useful in creating effective external criteria for cluster analysis of gene expression data.

TABLE IV. ALGORITHM PERFORMANCE ON THE RCNS DATASET[a]

#c	fzBGO	PC	FS	XB	CWB	PBMF	BR	fzBLE
2	-970.34	**0.99**	-568.8	**0.06**	5.51	4.21	1.11	-580.1
3	-970.46	0.94	-487.6	0.49	**4.13**	**4.28**	1.66	-564.2
4	-969.91	0.95	-430.5	0.93	6.12	3.37	1.32	-561.0
5	-969.67	0.89	-397.1	1.30	9.48	2.61	1.17	-561.7
6	**-969.66**	0.87	-300.7	2.52	20.65	1.95	1.10	**-553.0**
7	-970.15	0.87	-468.3	2.14	21.02	2.87	0.79	-556.3
8	-969.79	0.89	-462.1	1.73	20.01	2.53	0.59	-555.4
9	-970.30	0.89	-512.4	1.62	22.48	2.60	0.50	-558.9
10	-970.49	0.89	**-644.2**	1.19	21.99	3.50	**0.39**	-565.8

[a] fzBGO was run using GO cellular component (CC) annotations

TABLE V. ALGORITHM PERFORMANCE ON THE RCNS DATASET[b]

#c	fzBGO	PC	FS	XB	CWB	PBMF	BR	fzBLE
2	-1374.0	**1.00**	-568.8	**0.06**	5.51	4.21	1.11	-580.1
3	-1374.0	0.94	-487.6	0.49	**4.13**	**4.28**	1.66	-564.2
4	**-1373.8**	0.91	-430.5	0.93	6.12	3.37	1.32	-561.0
5	-1374.2	0.89	-397.1	1.30	9.48	2.61	1.17	-561.7
6	-1374.5	0.87	-300.7	2.52	20.65	1.95	1.10	**-552.9**
7	-1374.5	0.87	-468.3	2.14	21.02	2.87	0.79	-556.3
8	-1374.8	0.89	-462.1	1.73	20.01	2.53	0.59	-555.4
9	-1375.2	0.89	-512.4	1.62	22.48	2.60	0.50	-558. 9
10	-1375.4	0.89	**-644.2**	1.19	21.99	3.50	**0.39**	-565.8

[b] fzBGO was run using GO biological process (BP) annotations

IV. CONCLUSIONS

We have presented fzBGO, a novel method to evaluate results of cluster analysis using a standard FCM algorithm or a crisp clustering algorithm. fzBGO is novel in that it uses statistical models with Gene Ontology based semantic similarity to describe the data distributions and to validate the clustering results. By using GO annotations as external validation criteria, fzBGO successfully identifies the correct number of clusters in gene expression datasets. The results have also shown that fzBGO performs effectively on biological datasets with both internal and external criteria. In future work, we will integrate this method with our former cluster validation method, fzBLE [8], and apply into optimization algorithms to develop new clustering algorithms that can effectively support cluster analysis on gene expression data.

REFERENCES

[1] J.C. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms”, Plenum Press, NewYork, 1981.

[2] Y. Fukuyama and M. Sugeno, “A new method of choosing the number of clusters for the fuzzy c-means method”, In Proc. Fifth Fuzzy Systems Symp., 1989, pp. 247-250.

[3] X.L. Xie and G. Beni, “A validity measure for fuzzy clustering”, IEEE Trans. Pattern Anal. Mach. Intell, Vol. 13, pp. 841–847, 1991.

[4] M.R. Rezaee, B.P.F. Lelieveldt and J.H.C. Reiber, “A new cluster validity index for the fuzzy c-mean”, Pattern Recognition Lett, Vol. 19, pp. 237–246, 1998.

[5] M.K. Pakhira, S. Bandyopadhyay and U. Maulik, “Validity index for crisp and fuzzy clusters”, Pattern Recognition, Vol. 37, pp. 481–501, 2004.

[6] B. Rezaee, “A cluster validity index for fuzzy clustering”, Fuzzy Sets and Systems, Vol. 161, pp. 3014-3025, 2010.

[7] M.C. Florea, A.L. Jousselme, D. Grenier and E. Bosse, “Approximation techniques for the transformation of fuzzy sets into random sets”, Fuzzy Sets and Systems, Vol. 159, pp. 270–288, 2008.

[8] T. Le and K.J. Gardiner, “A validation method for fuzzy clustering of gene expression data,” In Proc. Intl’ Conf. on Bioinformatics & Computational Biology (BIOCOMP'11) , 2011, Vol. 1, pp. 23-29.

[9] T. Le, T. Altman and K. J. Gardiner. “A Probability Based Defuzzification Method for Fuzzy Cluster Partition,” Proc. Intl' Conf. on Artificial Intelligence (ICAI'12), 2012, Vol. 2, pp. 1038-1043.

[10] X. Wen, S. Fuhrman, G.S. Michaels, G.S. Carr, D.B. Smith, J.L. Barker and R. Somogyi, “Large scale temporal gene expression mapping of central nervous system development”. In Proc. of the National Academy of Science USA, 1998, Vol. 95, pp. 334-339.

[11] D. Dembele and P. Kastner, “Fuzzy C-Means method for clustering microarray data”, Bioinformatics, Vol. 19, pp. 973-980, 2003.

[12] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart and R.W. Davis, “A genome-wide transcriptional analysis of the mitotic cell cycle”, Mole Cell, Vol. 2, pp. 65-73, 1998.

[13] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery and W.L. Ruzzo, “Model based clustering and data transformations for gene expression data”, Bioinformatics, Vol. 17, pp. 977-987, 2001.

[14] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, “Systematic determination of genetic network architecture,” Nat. Genet. Vol. 22, pp. 281–285, 1999.

[15] H. W. Mewes, J. Hani, F. Pfeiffer and D. Frishman, “MIPS: A database for protein sequences and complete genomes”, Nucleic Acids Research, Vol. 26, pp. 33-37, 1998.

[16] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler et al., “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” Nat Genet, 2000, Vol. 25(1), pp. 25-29.

[17] L. Tari, C. Baral and S. Kim, “Fuzzy c-means clustering with prior biological knowledge,” J. Biomed Inform., Vol. 42, pp. 74-81, 2009.

[18] T. Xu, L.F. Du, and Y. Zhou, “Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data,” BMC Bioinformatics, Vol. 9, pp. 472-481, 2008.

[19] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu and C.F. Chen, “A new method to measure the semantic similarity of GO terms,” Bioinformatics, Vol. 23, pp. 1274-1281, 2007.

[20] G. Yu, F. Li, Y. Qin et al., “GOSemSim: an r package for measuring semantic similarity among GO terms and gene products,” Bioinformatics, Vol. 26, pp. 976-983, 2010.

[21] S. Chandra, S. Kumar. and C.V. Jawahar, “Learning Hierarchical Bag of Words Using Naive Bayes Clustering,” Computer Vision - Lecture Notes in Computer Science, Vol. 7724, pp. 382-395, 2012.

[22] Saccharomyces Genome Database, 20 December, 2012, http://www.yeastgenome.org/download-data/curation

[23] RAT Gene Ontology Annotations 104 Released, 6 February, 2013 http://www.ebi.ac.uk/GOA/.