31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A New Validity Measure for a Correlation-Based Fuzzy C-means Clustering Algorithm

Mingrui Zhang, Wei Zhang, Hugues Sicotte and Ping Yang

Abstract—**One of the major challenges in unsupervised clustering is the lack of consistent means for assessing the quality of clusters. In this paper, we evaluate several validity measures in fuzzy clustering and develop a new measure for a fuzzy c-means algorithm which uses a Pearson correlation in its distance metrics. The measure is designed with within-cluster sum of square, and makes use of fuzzy memberships. In comparing to the existing fuzzy partition coefficient and a fuzzy validity index, this new measure performs consistently across six microarray datasets. The newly developed measure could be used to assess the validity of fuzzy clusters produced by a correlation-based fuzzy c-means clustering algorithm.**

I. Introduction

IN clustering microarray data, a fuzzy clustering algorithm assigns a gene with degrees of memberships to multiple clusters. A fuzzy membership is a value between 0 and 1 with one indicating a complete association to a cluster [1]. During clustering, the algorithm minimizes an objective function. Our recent study has shown a fuzzy c-means (FCM) algorithm with a correlation-based objective function outperformed the algorithm using Euclidean distance metrics [2]. In microarray experiments, gene expression profiles may correlate positively or negatively. They should be clustered into one group regardless of their expression values. An FCM algorithm equipped with a Euclidean distance metrics may fail to cluster those genes. The finding is also consistent with another application of fuzzy clustering algorithm to yeast microarray data [3].

One of the major challenges in unsupervised clustering, especially in fuzzy clustering, is the lack of consistent means for assessing the quality of clusters. Several validity indexes, such as Consensus Clustering [4], Figure of Merit (FOM) [5], Gap Statistics and Model Explorer [6], have been proposed for microarray studies. However, they are designed for crisp clustering algorithms, specifically, Hierarchical and K-means clustering algorithms [7]. These indexes are not appropriate for assessing fuzzy clusters. They also show severe limitations, such as high computational demand and the lack of precision, in processing large dataset.

In this paper, we develop a new measure for an FCM algorithm which uses Pearson correlation in its objective function. The rest of the paper is organized into five sections. We introduce the microarray data sets to be used in Section II, give a brief review on a correlation-based FCM algorithm in Section III. Then, we present several validity measures, and propose a new one in Section IV. We will test the validity measures on six microarray datasets and compare them in Section V. We conclude in Section VI.

II. Data sets

A summary of the microarray data sets is given in TABLE I. The first four data sets are obtained from [7], the Yeast II dataset is from [5], and the lung cancer data is from University of Michigan [8]. According to the published results, number of clusters is five, three, three and eight for Yeast I, leukemia, lymphoma and NCI60 datasets, respectively. Dataset Yeast II consists of 4,373 genes. It contains genomic expression data of wild-type *S. cerevisiae* responding to zinc starvation, phosphate limitation, DNA-damaging agents and a variety of stressful environmental changes [3]. This data is normalized, background-corrected log2 value of the red/green ratios measured on the DNA microarrays [9]. The lung cancer dataset of 7,129 genes has 86 lung tumors and 10 normal lung samples [8]. Quartile normalization was performed using dChip; gene expression values were transformed to log2 based values(http://biosun1.harvard.edu/complab/dchip/).

TABLE I
MICROARRAY DATA SETS

Data Sets	# Genes	# Experiments	# Clusters
Yeast I	698	72	5
Leukemia	100	38	3
Lymphoma	100	80	3
CI60	200	57	8
Yeast II	4373	93	Unknown
Lung	7129	96	Unknown

III. Correlation-based FCM algorithm

Given a set of N genes with their expressions as $X=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N\}^T$, each $\boldsymbol{x}_i=[x_{i1}, x_{i2}, \ldots x_{ip}]$ is a p-dimensional vector which represents a gene i with its p experiments or conditions. A cluster is represented by a centroid $\boldsymbol{c}_j=[c_{j1}, c_{j2}, \ldots c_{jp}]$, the "averaged" gene expressions for all genes in the cluster j and for p experiments.

Manuscript received April 6, 2009. This work was supported in part by the HealthForce Minnesota.

M. Zhang is with Computer Science Department, Winona State University, Winona, MN 55987 USA (phone: 507-245-2980; e-mail: mzhang@winona.edu).

W. Zhang is with Computer Science department, Winona State University, Winona, MN 55987 USA.

H. Sicotte is with the Department of Health Science Research, Mayo Clinic Rochester, Rochester, MN 55905 USA, (e-mail: Sicotte.Hugues@mayo.edu).

P. Yang is with the Department of Health Science Research, Mayo Clinic Rochester, Rochester, MN 55905 USA, (e-mail: Yang.Ping@mayo.edu).

A. The Algorithm

In clustering genes, the correlation-based FCM algorithm computes the fuzzy centroids $\boldsymbol{c}_j$ and memberships u_{ij} at each iteration t as

$$C_j^{(t)} = \frac{\sum_{i=1}^{N} \left(u_{ij}^{(t-1)}\right)^m x_i}{\sum_{i=1}^{N} \left(u_{ij}^{(t-1)}\right)^m}; j=1,2,\ldots C \qquad (1)$$

and

$$u_{ij}^{(t)} = \left(\sum_{k=1}^{C} \left[d_{ij}^{(t)} \Big/ d_{ik}^{(t)} \right]^{\frac{2}{(m-1)}} \right)^{-1} \qquad (2)$$

Where, C is the number of clusters. A fuzzy membership u_{ij} represents the membership of gene i to a cluster $\boldsymbol{c}_j$, it satisfies a constraint $\sum_{j=1}^{C} u_{ij} = 1$.

A Pearson correlation coefficient is used in distance metrics d_{ij} which measure the difference between two genes or between a gene and a cluster. In microarray experiments, gene expression profiles may correlate positively or negatively. Correlated genes should be clustered into one group regardless of their expression values. The distance metrics is defined as $d_{ij} = 1 - \rho^2_{X_i, C_j}$, where ρ_{X_i, C_j} is the Pearson correlation between a gene $\boldsymbol{x}_i$ and a cluster $\boldsymbol{c}_j$. It would be 0 if they are highly correlated, either positively or negatively.

B. Correlation vs. Euclidean Distance Metrics

The believability of a cluster could be estimated by the genes sharing a known functional annotation in the cluster. Hypergeometric probability distribution can be used to compute a probability that an observed enrichment of a functional category comes from randomly selected genes [10].

$$\text{p-value} = \left(\sum_{i=m_k}^{M_k} \binom{M_k}{i} \binom{N_k - M_k}{n_k - i} \right) \Big/ \binom{N_k}{n_k} \qquad (3)$$

Where,

N_k, n_k: Number of genes in a fuzzy cluster before and after membership threshold.

M_k, m_k: Number of genes in a cluster assigned to a functional category before and after membership threshold.

In comparing the correlation-based FCM algorithm to a classic FCM algorithm, both algorithms were applied to Yeast II microarray dataset. For each cluster, we computed the p-value of enrichment by m_k genes to a functional GO category. Only the clusters with p-value<0.001 are listed in the Table II; some clusters (C_k) have multiple functional enrichments. Both algorithms are randomly initialized and seeded with 100 clusters. The fuzziness index m is set to *1*.2, and the threshold of fuzzy membership set to 0.15.

The correlation-based FCM algorithm produces more biologically meaningful clusters. There are 25 enriched groups of genes in 15 clusters, and 12 unique functional categories being significantly enriched (p-values < 0.05) for the correlation-based versus 13/8/9 for Euclidean distance-based. Furthermore, the correlation-based is able to identify groups of genes with much lower p-value (< 10^{-4}), again, indicating that the algorithm yields clusters with more biological meanings.

TABLE II
ENRICHMENTS OF FUNCTIONAL CATEGORIES FOR YEAST II

Ck	Functional Category	Mk	nk	mk	p-Value
7	Response Abiotic	215	89	10	9.88E-05
16	Metabolism	2238	316	138	8.41E-11
	Cellular Physiological	2903	316	140	0.0004021
19	Conjugation	67	70	8	7.82E-08
	Sexual Reprod	68	70	8	8.80E-08
	Reprod Physiological	117	70	8	5.84E-06
	Adhesion	15	70	3	0.0002475
23	Response Stress	263	39	14	5.88E-12
24	Response Abiotic	215	46	8	2.08E-05
	Response Stress	263	46	7	0.0005704
32	Localization	723	116	27	1.77E-06
33	Metabolism	2238	268	100	0.0001388
37	Metabolism	2238	217	83	0.0001971
43	Localization	723	319	45	0.0008646
56	Metabolism	2238	62	29	0.0007002
81	Response Abiotic	215	41	6	0.0006131

IV. VALIDITY MEASURES FOR FUZZY CLUSTERING

Despite being an unsupervised clustering method, fuzzy clustering requires a fair amount of user supervision to perform the clustering and to interpret the results. Particularly, the user is expected to choose the number of clusters for an FCM algorithm. Cluster validity measures are meant for validating a partition of the data, and they can also be used to help user to choose number of clusters among data. In this section, we first review several measures, and then design one for correlation-based FCM algorithm.

A. Validity Measures

Most of available measures are designed for crisp clustering and only a few of them are intended for use in fuzzy partition. In addition, they are all based on a Euclidean distance metrics. Let D_j be the compactness of a cluster C_j,

$$D_j = \sum_{i \in C_j} d(x_i, C_j)^2 = \sum_{i \in C_j} \left\| x_i - C_j \right\|^2 . \qquad (4)$$

The conventional within-cluster sum of square (WCSS) is thus the total of D_j,

$$WCSS(k) = \sum_{j=1}^{k} D_j . \qquad (5)$$

In Equation 4, a Euclidean distance, $\left\| x_i - C_j \right\|^2$ is used to measure the overall compactness of clusters. The WCSS is intended for validating a hard partition.

In fuzzy clustering, a partition coefficient F was initially designed by Bezdek [1],

$$F = \frac{1}{N} \sum_{j=1}^{C} \sum_{i=1}^{N} \left(\mu_{ij}\right)^m . \qquad (6)$$

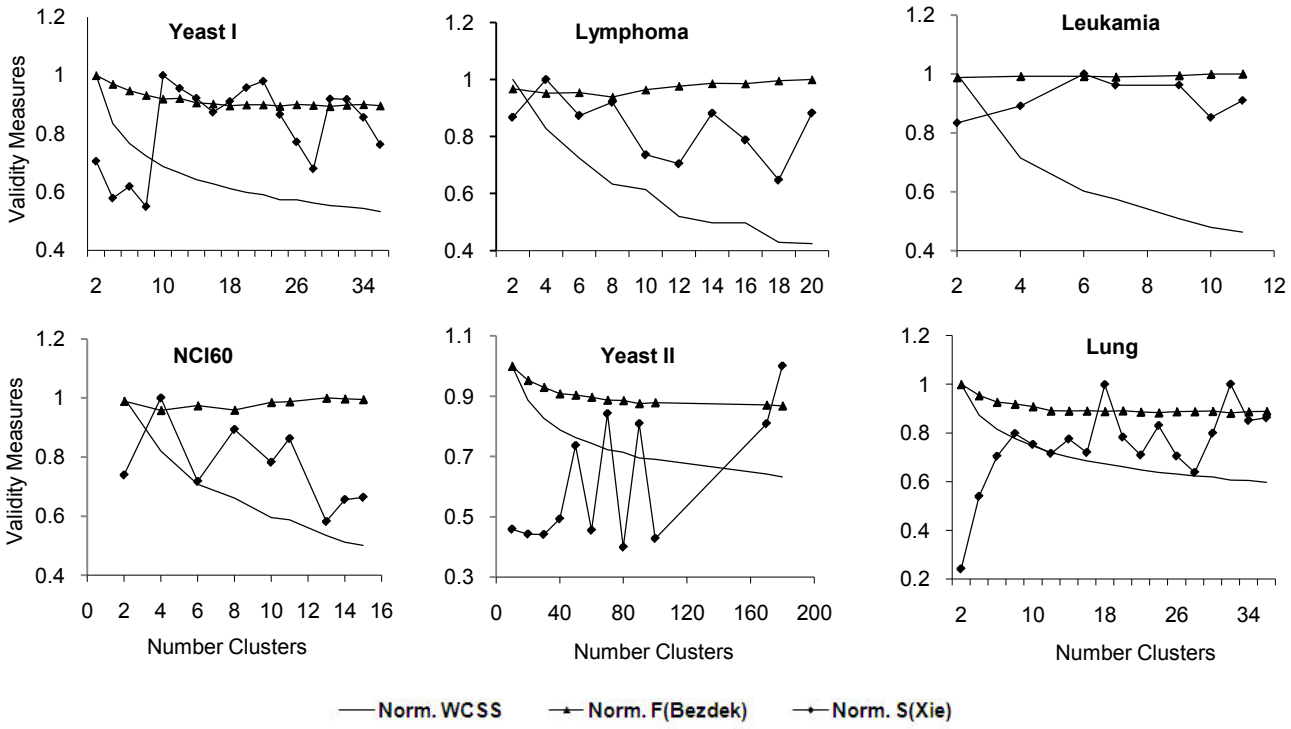

Fig. Validity measures calculated on six microarray data sets.

The coefficient measures the amount of overlap between fuzzy clusters. Its disadvantages are the lack of direct connection to a geometric property and its monotonic decreasing tendency with the number of clusters C.

To overcome these disadvantages, Xie and Beni proposed a measure to take compactness within clusters and separations between clusters into one index [11],

$$S = \frac{\sum_{j}^{C}\sum_{i}^{N} \mu_{ij}^{m} \left\| C_j - x_i \right\|^2}{N \min_{i,j} \left\| C_i - C_j \right\|^2} . \quad (7)$$

B. Validity Measure for Correlation-Based FCM

Both WCSS and the validity index S are based on a Euclidean distance metrics, they need to be adapted for a correlation-based FCM algorithm. In its application to microarray data, the validity measure S has shown dramatic variations. Though the introduction of an *ad hoc* "punishing" function has been discussed in [11], the form of this function itself is still unclear at this moment. We determined it was unsuitable for a correlation-based algorithm. A new validity measure is designed by introducing fuzzy membership into WCSS. This new measure is

$$D_j^{'} = \sum_i \mu_{ij}^m d\left(x_i, C_j\right) = \sum_i \mu_{ij}^m \left[1 - corr\left(x_i, C_j\right)^2\right] , \quad (8)$$

$$fWCSS(k) = \sum_{j=1}^{k} D_j^{'} . \quad (9)$$

The term $1 - corr\left(x_i, C_j\right)^2$ is a Pearson correlation based distance metrics, which is the same one used in the algorithm. The *fWCSS(k)* measures the overall compactness of fuzzy clusters, and the amount of overlap between fuzzy clusters with the inclusion of fuzzy membership. The new validity measure uses the same distance metrics being used in the correlation-based FCM algorithm.

V. EXPERIMENTS

The correlation-based FCM algorithm is applied to six microarray data sets (TABLE I). The Yeast II and the lung cancer data sets are used for explorative purpose.

A. Comparison to Other Measures

We have considered three validity measures for the correlation-based FCM algorithm: the partition coefficient F, Xie and Beni's validity index S, and a newly developed *fWCSS*. Fuzzy clustering was performed on six microarray datasets: Yeast I, lymphoma, leukemia, NCI60, Yeast II and lung cancer. To make a fair comparison, the FCM clustering algorithm is randomly initialized, and the fuzziness index m is set to 1.2 according to [10].

Each chart in Fig. illustrates the changes of three validity measures with number of clusters. For the sake of comparison, a validity measure is normalized by its maximum. Results show the validity measure S is unsuitable for clustering microarray data, it varied dramatically on all data sets. In addition, its value increases with number of fuzzy clusters on Yeast II and Lung data sets, both with more than 4000 genes. As for the partition coefficient F, it changes little

on lymphoma, leukemia and NCI60 datasets, which makes it hard to use for our algorithm. Clearly, the *fWCSS* measure performs consistently across six datasets, and outperforms the other two.

It is worthy of note that the sizes of lymphoma, leukemia and NCI60 datasets are relatively small, less than 300, in term of number of genes. As less overlapped clusters of genes are expected in those datasets, the partition coefficient *F* may not perform well. Like many validity measures, the newly developed measure tends to decrease monotonically with number of clusters. It is also recommended to assess the clusters in terms of the enrichment to a functional category (see TABLE II) based on a hypergeometric probability distribution.

In addition to cluster validation, one may use the measure to determine appropriate number of clusters in a data set. To do it, a threshold could be set beforehand and the number of clusters is identified when the change of a validity measure is below that threshold.

B. Lung Cancer Profiling

The correlation-based FCM clustering algorithm is also applied to lung cancer microarray data for explorative purpose. The dataset of 7,129 genes has 86 lung tumors and 10 normal lung samples. Before the algorithm is applied, the validity measure *fWCSS* is computed on the dataset. A close examination of *fWCSS* revealed 40 as the number of clusters for the data set. After fuzzy clustering, *p*-values are computed for each functional category in a cluster.

TABLE III
ENRICHMENT OF FUNCTIONAL CATEGORIES FOR LUNG DATA

Ck	Functional Category	nk	mk	p-Value
16	Immune Response	932	47	6.09E-16
	Defense Response	932	35	5.47E-11
	Response To External Stimulus	932	33	1.44E-08
17	Immune Response	847	54	8.30E-23
	Defense Response	847	32	3.33E-10
	Response To Biotic Stimulus	847	20	1.09E-07
18	Cell Adhesion	356	21	5.88E-10
24	Cell Cycle	862	43	1.10E-13
	Cell Cycle Process	862	38	2.52E-12
31	Macromolecule Localization	1714	38	1.10E-07
	Establishment Protein Localization	1714	34	3.33E-07
34	Biosynthetic Process	1259	82	1.48E-22

As an unsupervised clustering algorithm is often used for exploratory purposes, a true validation of clustering results requires expensive laboratory experiments [12]. An alternate way is to compute the *p*-values based on hypergeometric probability distribution. TABLE III listed the ones with p-value $< 10^{-6}$. This second level of validation identifies six significant (p-value $< 10^{-4}$) cluster centroids with one or more functionally enriched categories.

VI. CONCLUSION

Several validity measures have been evaluated for use with a correlation-based FCM algorithm. They include WCSS, a partition coefficient, and Xie and Beni's validity index. Results show Xie and Beni's validity index changes dramatically on six microarray data sets, and increases with number of fuzzy clusters on two large microarray data sets. It is of little use in clustering microarray data. As for the partition coefficient, its values barely change on small data sets such as lymphoma, leukemia and NCI60. It is suspected the poor performance of this measure is caused by the small number of genes in the datasets.

A newly developed validity measure performs better than both existing ones in term of consistency and robustness. In our opinion, a purely statistical validity measure by itself is not sufficient to make a judgment on the validity of a fuzzy partition. We recommend on using it with a domain specific validation approach, such as the enrichment to functional categories based on hypergeometric probability distribution.

REFERENCES

[1] J. C. Bezdek, *Pattern recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 198.

[2] M. Zhang, T. Therneau, M. A. McKenzie, P. Li, and P. Yang, "A Fuzzy C-Means Algorithm Using a Correlation Metrics and Gene Ontology," in *The 19th International Conference on Pattern Recognition*, Tampa, Florida, USA, 2008.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *PNAS,* vol. 95, pp. 14863-14868, December 8, 1998 1998.

[4] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning,* vol. 52, pp. 91-118, 2003.

[5] 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," *Mol. Cell,* vol. 2, pp. 65-73, 1998.

[6] K. Shedden, J. M. G. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, A. C. Chang, C. Q. Zhu, D. Strumpf, S. Hanash, F. A. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V. Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V. E. Seshan, M. Meyerson, R. Kuick, K. K. Dobbin, T. Lively, J. W. Jacobson, and D. G. Beer, "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nat Med,* vol. 14, pp. 822-827, 2008.

[7] R. Giancarlo, D. Scaturro, and F. Utro, "Computational cluster validation for microarray data analysis: experimental assessment of Clest, Consensus Clustering, Figure of Merit, Gap Statistics nd Model Explorer," *BMC Bioinformatics,* vol. 9, 2008.

[8] D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nat Med,* vol. 8, pp. 816-824, 2002 2002.

[9] A. Gasch and M. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology,* vol. 3, pp. research0059.1 - research0059.22, 2002.

[10] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics,* vol. 19, pp. 973-980, May 22, 2003 2003.

[11] X. L. Xie and G. NBeni, "A Validity Measure for Fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 13, pp. 841-847, 1991.

[12] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE Transactions on Computational Biology and Bioinformatics,* vol. 1, pp. 24-45, 2004.