A validation method for fuzzy clustering of gene expression data

Thanh Le[1], Katheleen J. Gardiner[2]
[1]Department of CSE, University of Colorado Denver, Denver, CO, USA
[2]Department of Pediatrics, University of Colorado Denver, Denver, CO, USA

Abstract - *Clustering is a key process in data mining for revealing structure and patterns in data. Fuzzy C-means (FCM) is a popular algorithm using a partitioning approach for clustering. One advantage of FCM is that it converges rapidly. In addition, using fuzzy sets to represent the degrees of cluster membership of each data point provides more information regarding relationships within the data than do alternative approaches that use crisp clustering. However, a limitation of FCM is that it requires initial specification of the number of clusters and subsequent validation of this number. Here, we propose a Bayesian method for fuzzy clustering validation using the fuzzy partition. We show that this method outperforms popular fuzzy cluster indices on both artificial and real biological datasets.*

Availability: The supplementary documents and the method software are at http://ouray.ucdenver.edu/~tnle/fzble.

Keywords: fuzzy c-means; Bayesian; cluster index

1 Introduction

Cluster analysis groups data points based on their similar properties and can help to discover patterns and correlations in large datasets. Successful clustering maximizes both the compactness of data points within a cluster and the discrimination between clusters. Fuzzy C-Means (FCM, Bezdek 1981) is a popular algorithm that uses a partitioning approach with fuzzy cluster boundaries and fuzzy sets that associate each data point with one or more 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. While for some initial values, FCM may converge to a global optimum, 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. Recently, Rezaee [6] proposed a new cluster index in which the two factors are normalized across the range of possible numbers of clusters.

Here, we propose a fuzzy clustering cluster index that uses the fuzzy partition and the distance matrix between cluster centers and data points. Instead of compactness and separation, our cluster index is based on a Bayesian model and a log-likelihood estimator. With the use of both the possibility model and the probability model to represent the data distribution, our method is appropriate for artificial data where the distribution follows a standard model, as well as for real datasets, in particular, gene expression data, that lack a standard distribution. We show that our method outperforms popular cluster indices on both artificial and biological datasets.

2 Fuzzy C-Means and popular cluster indices

2.1 Fuzzy C-Means algorithm

Fuzzy C-Means (FCM) is an unsupervised clustering algorithm that has been applied successfully to numerous problems involving feature analysis. Its applications include biological data analysis, in particular, gene expression data.

This work was supported by the National Institutes of Health (HD056235), the Linda Crnic Institute (KG) and the Vietnamese Ministry of Education and Training (TL).

Thanh Le is a doctoral student in the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217-3364, USA (email: lntmail@yahoo.com).

Katheleen J. Gardiner is a professor in the Department of Pediatrics; the Intellectual and Developmental Disabilities Research Center; and the Computational Biosciences, Human Medical Genetics and Neuroscience Programs, University of Colorado Denver, Aurora, CO 80045, USA (phone: 303-724-0572; email: katheleen.gardiner@ucdenver.edu).

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

$$\sum_{k=1}^{c} u_{ki} = 1, \ i = 1..n. \tag{1}$$

The clustering problem is to determine the values of c and V such that:

$$J(X \mid U, V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki} \left\| x_i - v_k \right\| \to \min, \tag{2}$$

where $\|x-y\|$ is the distance between the data points x and y in R^p, defined using Euclidean distance as:

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

By using fuzzy sets to assign data points to clusters, FCM allows adjacent clusters to overlap. It thus provides more information on the relationships of data points. In addition, by using a fuzzifier factor, m, $1 \le m < \infty$, in its objective function (4), the clustering model from FCM is more flexible in changing the overlap regions among clusters.

$$J(X \mid U, V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki}^m \left\| x_i - v_k \right\| \to \min, \tag{4}$$

The following is a solution of (4) with respect to (1):

$$v_k = \sum_{i=1}^{n} u_{ki}^m x_i \Big/ \sum_{i=1}^{n} u_{ki}^m, \tag{5}$$

$$u_{ki} = \left(\frac{1}{\left\| x_i - v_k \right\|^2} \right)^{\frac{1}{1-m}} \Bigg/ \sum_{j=1}^{c} \left(\frac{1}{\left\| x_i - v_j \right\|^2} \right)^{\frac{1}{1-m}}. \tag{6}$$

FCM uses an iteration process to estimate the solution of (5) and (6). This process is iterated until convergent where

$\exists \varepsilon_u > 0$, $T > 0$: $\forall t > T$,

$$\left\| U_{t+1} - U_t \right\| = \max_{k,i} \left\{ \left\| u_{ki}(t+1) - u_{ki}(t) \right\| \right\} < \varepsilon_u. \tag{7}$$

Or, $\exists \varepsilon_v > 0$, $T > 0$: $\forall t > T$,

$$\left\| V_{t+1} - V_t \right\| = \max_{k} \left\{ \left\| v_k(t+1) - v_k(t) \right\| \right\} < \varepsilon_v. \tag{8}$$

While FCM can converge quickly, it is unable to determine the optimal number of clusters in the dataset.

2.2 Cluster validation indices

(i) To determine if the fuzzy partition is valid, traditional cluster indices use two criteria, 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)

(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.

2.2.1 PC index

The partition coefficient (PC) index was proposed by Bezdek [1] as in (9). 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].

$$V_{PC} = \frac{1}{n} \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^2. \tag{9}$$

An optimal cluster number c can be found by solving,

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

2.2.2 PE index

The partition entropy (PE) index was proposed by Bezdek [1] as

$$V_{PE} = \frac{1}{n} \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki} \times \log_a(u_{ki}), \tag{10}$$

where a is the base of the logarithm. According to [1], the limitation of the PE can be attributed to its apparent monotonicity and to an extent, to the heuristic nature of the rationale underlying its formulation. An optimal cluster number c can be found by solving $V_{PE} \to \min$.

2.2.3 FS index

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

$$V_{FS} = J - \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m \left\| v_k - \bar{v} \right\|^2, \tag{11}$$

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

2.2.4 XB index

The XB index was proposed by Xie and Beni as in (12). 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 $V_{XB} \rightarrow \min$.

$$V_{XB} = \frac{\sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^{m} \times \|x_i - v_k\|^2}{n \times \min_{k,l}\|v_k - v_l\|^2} \tag{12}$$

2.2.5 CWB index

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

$$V_{CWB} = \alpha \mathrm{Scat}(c) + \mathrm{Dis}(c), \tag{13}$$

where α is a weighting factor equal to $\mathrm{Dis}(c_{max})$. The average scatter is defined as

$$\mathrm{Scat}(c) = \frac{\sum_{k=1}^{c}\|\sigma(v_k)\|}{c \times \|\sigma(X)\|}, \tag{14}$$

where $\|x\| = (x^T x)^{1/2}$, $\sigma(X) = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \bar{x}\|^2$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The Dis function is defined as

$$\mathrm{Dis}(c) = \frac{D_{max}}{D_{min}}\sum_{k=1}^{c}\left(\sum_{i=1}^{n}\|v_k - x_i\|\right)^{-1}, \tag{15}$$

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 $V_{CWB} \rightarrow \min$.

2.2.6 PBMF index

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

$$V_{PBMF} = \left(\frac{1}{c}\frac{E_1}{J}D_c\right)^2, \tag{16}$$

$$E_1 = \sum_{i=1}^{n} u_{1i}\|x_i - \bar{x}\|, \tag{17}$$

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 $V_{PBMF} \rightarrow \max$.

2.2.7 BR index

The cluster 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

$$V_{BR} = \frac{\mathrm{Sep}(c)}{\max_c\{\mathrm{Sep}(c)\}} + \frac{J(c)}{\max_c\{J(c)\}}, \tag{18}$$

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

The similarity $S_{rel}(.)$ of two fuzzy sets is defined as

$$S_{rel}(v_k, v_l) = \sum_{i=1}^{n} S(x_i : v_k, v_l) \times h(x_i), \tag{19}$$

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

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 $V_{BR} \rightarrow \min$.

3 The proposed validation method

3.1 The proposed validation method (fzBLE)

Instead of compactness and separation factors, we propose a validation method (fzBLE) that is based on a log likelihood estimator with a fuzzy based Bayesian model. Each fuzzy clustering solution is modeled with $\theta = \{U, V\}$, where V represents the cluster centers and, U is the partition matrix representing the membership degrees of the data points to the clusters. The likelihood of the clustering model and the data is measured as

$$\begin{aligned} L(\theta \mid X) &= L(U, V \mid X) \\ &= \prod_{i=1}^{n} P(x_i \mid U, V) = \prod_{i=1}^{n}\sum_{k=1}^{c} P(v_k) \times P(x_i \mid v_k). \end{aligned} \tag{20}$$

The log likelihood estimator is then computed as

$$\log(L) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{c} P(v_k) \times P(x_i \mid v_k)\right) \rightarrow \max. \tag{21}$$

An optimal number of clusters is obtained by solving (21).

3.2 Possibility to probability transformation

Because our clustering model is possibility-based, before applying equations (20) and (21), a transformation of possibility to probability is needed. Given a fuzzy clustering model $\theta = \{U, V\}$, according to [7], u_{ki} is the possibility that $v_k = x_i$. If θ is a proper fuzzy partition, then there exists some x^* such that $U_k(x^*) = 1$, k=1..c, and U_k is a normal possibility distribution. Assume P_k is the probability distribution of v_k on X where $p_{k1} \geq p_{k2} \geq p_{k3} \geq \ldots \geq p_{kn}$. We associate with P_k a possibility distribution U_k on X [7] such that u_{ki} is the possibility of x_i where

$$\begin{aligned} u_{kn} &= n \times p_{kn} \\ u_{ki} &= i\left(p_{ki} - p_{k,i+1}\right) + u_{k,i+1},\ i = n-1,\ldots 1. \end{aligned} \tag{22}$$

Reversing (22), we obtain the transformation of a possibility distribution to a probability distribution. Assume that U_k is ordered the same way with P_k on X: $u_{k1} \geq u_{k2} \geq u_{k3} \geq \ldots \geq u_{kn}$.

$$\begin{aligned} p_{kn} &= u_{kn} / n \\ p_{ki} &= p_{k,i+1} + \left(u_{ki} - u_{k,i+1}\right)/ i. \end{aligned} \tag{23}$$

P_k is an approximate probability distribution of v_k on X, and $p_{ki} = P(x_i|v_k)$. If U_k is a normal possibility distribution then $\sum p_{ki} = 1$.

3.3 Data distributions

Using the value of P_k, we can estimate the variance σ_k, the prior probability $P(v_k)$ and the normal distribution of v_k.

$$\sigma_k = \sum_{i=1}^{n} p_{ki} \left\| x_i - v_k \right\|^2, \tag{24}$$

$$P(v_k) = \sum_{i=1}^{n} P(x_i \mid v_k) \Big/ \sum_{l=1}^{c} \sum_{i=1}^{n} P(x_i \mid v_l), \tag{25}$$

$$P_n(x_i \mid v_k) = \left((2\pi)^{1/n} \times \sigma_k \times e^{\frac{\|x_i - v_k\|^2}{2\sigma_k^2}} \right)^{-1}. \tag{26}$$

In real datasets, for a cluster v_k, the data points usually come from different random distributions. Because they cluster in v_k, they tend to follow the normal distribution estimated as in (26). This idea is based on the Central Limit Theorem. We therefore integrate the probabilities computed in (23) and (26) for the probability of the data point x_i given cluster v_k as

$$P^*(x_i \mid v_k) = \max\left\{P(x_i \mid v_k), P_n(x_i \mid v_k)\right\}. \tag{27}$$

Equation (27) better represents the data distribution, particularly in real datasets. The fzBLE method is based on (21) with (25) and (27).

3.4 fzBLE and FCM combination

fzBLE can be used with the standard FCM algorithm to search for the optimal number of clusters for a dataset using a cluster range.

- Input:
  - The data to cluster $X=\{x_i\}$, i=1..n
  - Cluster range $[c_{min}, c_{max}]$
- Output: An optimal fuzzy partition solution,
  - c_{opt}: Optimal number of clusters
  - $V = \{v_i\}$, i =1..c: Cluster centers
  - $U=\{u_{ki}\}$, i=1..n, k=1..c: Partition matrix

Steps

1. Set $c_{opt} = c_{min}$
2. For each value of c in $[c_{min}, c_{max}]$
   - Generate a fuzzy partition using FCM
   - Validate the partition using fzBLE
   - If the current partition is better than the optimal one then, set $c_{opt} = c$
3. Return $\{ c_{opt}, U, V\}$ an optimal solution.

4 Experimental results

To evaluate fzBLE, we generated 84 artificial datasets using the method in [8]. Datasets are distinguished by the dimensions and cluster number, and we generated (3-2+1)*(9-3+1)=14 dataset types. For each type, we generated 6 datasets, for a total of 6*14=84. For real datasets, we used the Iris, Wine and Glass datasets from the UC Irvine Machine Learning Repository [9], and the gene expression datasets, Yeast [13], Yeast-MIPS [14, 15] and RCNS [10]. These datasets contain classification information, useful for comparing cluster indices. We compared performance of fzBLE with the cluster indices from PC, PE, FS, XB, CWR, PBMF and BR [1-6]. The compactness factor (CF) of the FCM algorithm is also recorded in the results.

4.1 Artificial datasets

For each artificial dataset, we ran the standard FCM algorithm five times with m set to 2.0 and the partition matrix initialized randomly. In each case, the best fuzzy partition was then selected to run fzBLE and the other cluster indices to search for the optimal number of clusters between 2-12 and to compare this with the known number of clusters. We repeated the experiment 20 times and averaged the performance of each method. Table 1 shows the fraction of correct predictions. fzBLE and PBMF outperform other approaches, while CF is the least effective.

Table 1

Fraction of correct cluster predictions on artificial datasets

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
3	1.00	0.42	0.42	0.42	0.42	1.00	1.00	0.83	0.00
4	1.00	0.92	0.92	0.92	0.83	1.00	1.00	1.00	0.00
5	1.00	0.75	0.75	0.83	0.75	0.83	1.00	1.00	0.00
6	1.00	0.92	0.83	0.92	0.58	0.58	1.00	0.92	0.00
7	1.00	0.83	0.83	0.83	0.67	0.58	1.00	0.67	0.00
8	1.00	1.00	0.92	1.00	0.92	0.67	1.00	0.83	0.00
9	1.00	0.92	0.67	0.92	0.67	0.33	1.00	0.83	0.00

Table 2

Validation method performance on the Iris dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-763.0965	**0.9554**	**0.0977**	**-10.6467**	**0.0203**	**177.1838**	12.3280	1.1910	0.9420
3	**-762.8034**	0.8522	0.2732	-9.3369	0.1292	213.4392	**17.7131**	1.0382	0.3632
4	-764.8687	0.7616	0.4381	-7.4821	0.2508	613.2656	14.4981	1.1344	0.2665
5	-770.2670	0.6930	0.5703	-8.2331	0.3473	783.4697	13.6101	1.0465	0.1977
6	-773.6223	0.6549	0.6702	-7.3202	0.2805	904.3365	12.3695	1.0612	0.1542
7	-774.4740	0.6155	0.7530	-6.8508	0.2245	1029.7342	11.2850	0.9246	0.1262
8	-774.8463	0.6000	0.8111	-6.9273	0.3546	1635.3593	10.5320	0.8692	0.1072
9	-780.1901	0.5865	0.8556	-6.6474	0.3147	1831.5705	9.9357	0.7653	0.0905
10	-781.7951	0.5765	0.8991	-6.0251	0.2829	2080.3339	9.3580	**0.7076**	**0.0787**

Table 3

Validation method performance on the Wine dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-926.4540	**0.9264**	**0.1235**	-113.0951	**0.1786**	3.9100	**1.3996**	2.0000	61.1350
3	**-924.0916**	0.8977	0.1764	-104.9060	0.2154	**3.2981**	0.9316	1.4199	39.3986
4	-932.8377	0.8607	0.2525	**-139.9144**	0.5295	6.6108	0.6306	1.1983	33.7059
5	-929.6146	0.8225	0.3281	-126.5746	0.5028	6.9001	0.4700	1.0401	28.4741
6	-928.8121	0.8066	0.3669	-118.4715	0.6173	9.2558	0.3706	0.9111	25.3451
7	-930.6451	0.7988	0.3874	-120.3128	0.6465	10.3803	0.2972	0.7629	23.1742
8	-932.0462	0.7993	0.3917	-124.7999	0.6459	11.0836	0.2471	0.6392	21.4411
9	-932.1902	0.7929	0.4120	-122.8396	0.6367	11.8373	0.2100	0.5801	19.9154
10	-935.0478	0.7909	0.4217	-130.9089	0.6270	11.9941	0.1773	**0.5252**	**18.9891**

Table 4

Validation method performance on the Glass dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-1135.6886	**0.8884**	**0.1776**	0.3700	0.7222	6538.9311	0.3732	1.9817	0.5782
3	-1127.6854	0.8386	0.2747	0.1081	0.7817	4410.3006	0.4821	1.5004	0.4150
4	-1119.2457	0.8625	0.2515	-0.0630	0.6917	3266.5876	0.4463	1.0455	0.3354
5	-1123.2826	0.8577	0.2698	-0.1978	**0.6450**	**2878.8912**	0.4610	0.8380	0.2818
6	**-1113.8339**	0.8004	0.3865	-0.2050	1.4944	5001.1752	0.3400	0.8371	0.2430
7	-1116.5724	0.8183	0.3650	-0.2834	1.3802	5109.6082	0.3891	0.6914	0.2214
8	-1127.2626	0.8190	0.3637	-0.3948	1.4904	7172.2250	0.6065	0.5916	0.2108
9	-1117.7484	0.8119	0.3925	-0.3583	1.7503	8148.7667	0.3225	0.5634	0.1887
10	-1122.1585	0.8161	0.3852	-0.4214	1.7821	9439.3785	0.3909	0.4926	0.1758
11	-1121.9848	0.8259	0.3689	-0.4305	1.6260	9826.4211	0.3265	0.4470	0.1704
12	-1135.0453	0.8325	0.3555	-0.5183	1.4213	11318.4879	0.5317	0.3949	0.1591
13	-1138.9462	0.8317	0.3556	**-0.5816**	1.4918	14316.7592	**0.6243**	**0.3544**	**0.1472**

Table 5

Validation method performance on the Yeast dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-2289.8269	0.9275	0.1172	-85.1435	**0.2060**	8.3660	**1.2138**	2.0000	133.0734
3	-2296.4502	0.9419	**0.0983**	-157.2825	0.2099	4.7637	0.6894	1.0470	94.6589
4	-2305.3369	**0.9437**	0.1000	-191.7664	0.2175	**4.0639**	0.5575	0.7240	74.7629
5	**-2289.3070**	0.9087	0.1648	-187.1073	1.0473	13.6838	0.4087	0.6722	65.9119
6	-2296.3098	0.8945	0.1939	-196.6711	0.9932	13.8624	0.3050	0.6170	60.8480
7	-2296.6017	0.8759	0.2299	-198.2858	1.0558	15.4911	0.2434	0.5686	56.1525
8	-2299.4225	0.8634	0.2526	-201.7688	1.0994	16.9644	0.2050	0.5132	51.2865
9	-2299.3653	0.8453	0.2871	-205.1489	1.2340	20.2532	0.1741	0.4819	48.0737
10	-2302.7581	0.8413	0.2992	-208.5687	1.1947	20.7818	0.1512	0.4533	45.9442
11	-2300.3294	0.8325	0.3186	-209.4023	1.1731	21.1525	0.1307	0.4272	43.6600
12	-2307.5701	0.8290	0.3272	-213.4658	1.2245	23.0389	0.1157	0.4040	42.1594
13	-2310.7819	0.8270	0.3354	**-215.2463**	1.3036	25.4062	0.1016	**0.3847**	**40.8654**

Table 6

Validation method performance on the YEAST-MIPS dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-1316.4936	0.9000	0.1625	25.4302	0.3527	16.7630	0.7155	1.9978	81.0848
3	-1317.3751	**0.9092**	**0.1615**	-32.8476	**0.2981**	**10.1546**	**0.8032**	1.2476	58.2557
4	**-1304.0374**	0.8216	0.3252	-39.4858	2.5297	39.8434	0.5400	1.3218	48.6275
5	-1308.6776	0.8279	0.3216	-54.4979	2.4245	34.9963	0.3620	0.9558	41.9671
6	-1309.9191	0.8211	0.3460	-59.8918	2.3511	35.4533	0.2691	0.8291	38.5468
7	-1315.3692	0.8139	0.3654	-65.4866	2.3562	38.8797	0.2423	0.7252	36.0906
8	-1315.1479	0.8062	0.3918	-67.6774	2.4958	43.9502	0.1966	0.6712	34.1387
9	-1321.2280	0.8109	0.3874	-72.3197	2.2854	41.2112	0.1664	0.6072	32.3289
10	-1324.1578	0.8158	0.3847	**-74.7867**	2.0433	37.6154	0.1395	**0.5588**	**30.9686**

Table 7

Validation method performance on the RCNS dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-580.0728	**0.9942**	**0.0121**	-568.7972	**0.0594**	5.5107	4.2087	1.1107	177.8094
3	-564.1986	0.9430	0.0942	-487.6104	0.4877	**4.1309**	**4.2839**	1.6634	117.9632
4	-561.0169	0.9142	0.1470	-430.4863	0.9245	6.1224	3.3723	1.3184	99.1409
5	-561.7420	0.8900	0.1941	-397.0935	1.3006	9.4770	2.6071	1.1669	88.5963
6	**-552.9153**	0.8695	0.2387	-300.6564	2.5231	20.6496	1.9499	1.1026	84.0905
7	-556.2905	0.8707	0.2386	-468.3121	2.1422	21.0187	2.8692	0.7875	57.5159
8	-555.3507	0.8925	0.2078	-462.0673	1.7245	20.0113	2.5323	0.5894	52.0348
9	-558.8686	0.8863	0.2192	-512.4278	1.6208	22.4772	2.6041	0.5019	45.9214
10	-565.8360	0.8847	0.2241	**-644.1451**	1.1897	21.9932	3.4949	**0.3918**	**33.1378**

4.2 Real datasets

The Iris, Wine and Glass datasets contain 3, 3 and 6 clusters, respectively. For the Iris dataset, only fzBLE and PBMF detected the correct number of clusters (Table 2). For the Wine and Glass datasets, only fzBLE and CWB, and only fzBLE, respectively, detected the correct number of clusters (Tables 3 and 4).

4.3 Biological datasets

4.3.1 Yeast

The Yeast dataset [13] reports expression levels of yeast genes throughout two cell cycles at 17 time points spaced at 10-minute intervals. Each of the 384 differentially expressed genes was labeled with one of the five cell cycle phases where their expression changed. We ran the FCM algorithm with m set to 1.17 [12] and used the clustering partition to test all methods as in previous sections. Table 5 shows that only fzBLE detected the correct number of clusters (5) in Yeast dataset.

4.3.2 Yeast-MIPS

The Yeast-MIPS dataset [14] is a subset of the Yeast dataset [14]. It contains 237 genes belonging to four functional categories: DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins [15]. We ran the FCM algorithm using the same parameters as with Yeast dataset. The results in Table 6 show that only fzBLE detected the four clusters in the Yeast-MIPS dataset.

4.3.3 RCNS

The RCNS (Rat Central Nervous System) dataset contains expression levels of 112 genes measured at nine time points during rat central nervous system development [10]. Wen et al. [11] preprocessed the dataset using a normalization method and scaling across adjacent axises to generate a 112x17 dataset so that Euclidean distance can be applied. The FITCH software was used to detect 6 clusters with biological relevance. Dembélé and Kastner [12] used the FCM algorithm varying the number of clusters and reported that 6 is the optimal number. We ran fzBLE and the other cluster indices on the dataset clustering partition found by the standard FCM algorithm using the Euclidean metric for distance measurement. Table 7 shows that again only fzBLE detected the correct number of clusters.

5 Conclusions

We have presented a novel method, fzBLE to evaluate results of fuzzy partitioning by the standard FCM algorithm. fzBLE is novel in that it uses the log likelihood estimator with a Bayesian model and the possibility, rather than the probability, distribution model of the dataset from the fuzzy partition. By using the Central Limit Theorem, fzBLE effectively represents distributions in real datasets. Results have shown that fzBLE performs effectively on both artificial and real datasets. In future work, we will integrate this method with optimization algorithms, to develop new clustering algorithms that can effectively support clustering analysis on real datasets.

6 References

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

[2] Y. Fukuyama, 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, 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, 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, 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, E. Bosse, "Approximation techniques for the transformation of fuzzy sets into random sets", Fuzzy Sets and Systems, Vol. 159, pp. 270–288, 2008.

[8] L. Xu, M.I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures", Neural Computation, Vol. 8, pp. 129-151, 1996.

[9] A. Frank and A. Asuncion, (2010) "Machine Learning Repository", [Online], http://archive.ics.uci.edu/ml.

[10] R. Somogyi, X. Wen, W. Ma, J.L. Barker, "Developmental kinetic of GLAD family mRNAs parallel neurogenesis in the rat", Journal of Neurosciences, Vol. 15, pp. 2575-2591, 1995.

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

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

[13] 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, R.W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle", Mole Cell, Vol. 2, pp. 65-73, 1998.

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

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