Indian Journal of Science and Technology, *Vol 8(2), 120–127, January 2015*
ISSN (Online) : 0974-5645
ISSN (Print) : 0974-6846
DOI : 10.17485/ijst/2015/v8i2/58943

Futuristic Validation Method for Rough Fuzzy Clustering

S. Revathy[1], B. Parvaathavarthini[2] and S. Rajathi[1]

[1]Sathyabama University, Chennai, India; ramesh.revathy@gmail.com, raja_thiit90@yahoo.co.in
[2]St. Joseph's college of Engineering, Anna University, Chennai, India;
parvathavarthini@gmail.com

Abstract

The most challenging issues in cluster analysis are to ratify clustering results to produce optimal number of cluster for a dataset. Customary validity indices are geometry based, and these indices face a vital problem where the resultant index value either increases or decreases as cluster count inflates. This paper exhorts a cluster validity index for rough fuzzy c-means clustering algorithm called rough fuzzy Bayesian like validation method which roots on probabilistic metric. Maximum Bayesian score stipulates optimal number of cluster. The proposed measure is been experimented for synthetic and diverse UCI datasets. This recommended scheme brings out optimal number of cluster for enormous UCI datasets than the prevailing customary validity indices.

Keywords: Bayesian like Validation Method, Cluster Analysis, Probability, Rough Fuzzy C-means and UCI Datasets

1. Introduction

Data mining is the process of withdrawing information and altering it into an intelligible structure for further use. Clustering is an essential approach in data mining where objects are grouped into clusters based on similarities. Clustering algorithms are used to carry out clustering process and are refined once when it is validated. The crucial problem in cluster analysis is to detect the optimal number of cluster by examining standard clustering results.

Numerous validity indices have been proposed to ratify clustering results. Indices such as Bezdek's[2] partition coefficient, partition entropy and Xie-Beni index[6] were proposed to validate fuzzy c-means clustering algorithm. Pratipa Maji[1] et al. introduced $\propto$ index and γ index to validate rough fuzzy c-means clustering algorithm. Sushmita Mitra et al proposed a modified Davies-Bouldin and Dunn validity index for rough fuzzy algorithm[3] to add relative importance to lower and upper approximations. Xie-Beni index focuses on compactness and separations, $\propto$ Index represents the average accuracy of a cluster, γ index represents the average approximation of overall clusters and Davies-Bouldin and Dunn index determines to internal evaluation scheme. But these indices do not perform well on multiclass dataset since it either increases or decreases monotonically. Though the customary validity indices are measured based on inter-intra cluster distance, typical indices fails to provide optimal partition.

Conventional fuzzy c-mean algorithm has utilized this Bayesian score measure, but is not applicable for rough set theory. Hence this paper propagates Bayesian like validation method for rough fuzzy clustering procuring the prototype from fuzzy Bayesian validation method[6] remodeling the scheme with crisp lower and fuzzy boundary concept. The proposed measure experimented for synthetic and tremendous UCI datasets prevails over the downsides of standard validation schemes.

**Author for correspondence*

2. Literature Review

Notations used in this paper are described below in the following table.

Table 1. Notations and description

Notations	Description
N	Total number of data samples
S	Total number of clusters
m	Fuzzifier
M_g	g^{th} cluster
T_h	h^{th} data object
y_{gh}	Distance between T_h and X_g
X_g	Centroid value of g^{th} cluster
$\lvert M_g \rvert$	Cardinality of g^{th} cluster
μ_{gh}	Membership of T_h in X_g
$\underline{R}Mg, \bar{R}Mg$	Lower, Boundary approximation of X_g (g^{th} cluster)
w_{lr}, w_{br}	Relative importance of lower and upper approximation
$P(M_g)$	Probability for g^{th} cluster
$P(T_{gh})$	Probability for a data point which is assigned to g^{th} cluster

2.1 ∝ Index

It is the average ratio of the number of objects in lower approximation to that in upper approximation for each cluster[1]. It is given by

$$\propto = \frac{1}{S}\sum_{g=1}^{S}\frac{w_{lr}A_g}{w_{lr}A_g + w_{br}B_g} \tag{1}$$

Where

$$A_g = \sum_{T_k \in \underline{R}M_g}(\mu_{gh})^m = |\underline{R}M_g| \text{ and } B_g = \sum_{T_k \in (\bar{R}M_g - \underline{R}M_g)}(\mu_{gh})^m \tag{2}$$

μ_{gh} is the membership for g^{th} cluster and h^{th} data point, $1 < m < 3$ represents the fuzzifier and S represents the number of clusters. $\underline{R}M_g$ and $\left(\bar{R}M_g - \underline{R}M_g\right)$ represents the lower and boundary region respectively. The parameters w_{lr} and w_{br} indicates the relative importance for lower and boundary region. The ∝ index value increases as S increases.

2.2 Davies-Bouldin Index

It is the ratio of the sum of within cluster distance to between cluster separations[2]. The DB index can be formulated as follows.

$$DB_{rf} = \frac{1}{S}\sum_{g=1}^{S}\max_{g \neq e}\left\{\frac{P_{rf}(M_g) + P_{rf}(M_e)}{y(M_g, M_e)}\right\} \tag{3}$$

Where

$$P_{rf}(M_g) = \begin{cases} w_{lr}\dfrac{\sum_{T_h \in \underline{R}M_g}\mu_{gh}^m \|T_h - X_g\|^2}{\sum_{T_h \in \underline{R}M_g}\mu_{gh}^m} + w_{br}\dfrac{\sum_{T_h \in (\bar{R}M_g - \underline{R}M_g)}\mu_{gh}^m \|T_h - X_g\|^2}{\sum_{T_h \in (\bar{R}M_g - \underline{R}M_g)}\mu_{gh}^m}, \\ \quad \text{if } \underline{R}M_g \neq \varnothing \cap \bar{R}M_g - \underline{R}M_g \neq \varnothing \\ \dfrac{\sum_{T_h \in (\bar{R}M_g - \underline{R}M_g)}\mu_{gh}^m \|T_h - X_g\|^2}{\sum_{T_h \in (\bar{R}M_g - \underline{R}M_g)}\mu_{gh}^m}, \text{ if } \underline{R}M_g = \varnothing \cap \bar{R}M_g - \underline{R}M_g \neq \varnothing \\ \dfrac{\sum_{T_h \in \underline{R}M_g}\mu_{gh}^m \|T_h - X_g\|^2}{\sum_{T_h \in (\bar{R}M_g - \underline{R}M_g)}\mu_{gh}^m}, \text{ otherswise} \end{cases} \tag{4}$$

$y(M_g, M_e)$ is the distance between g^{th} and e^{th} clusters. X_g represents g^{th} centroid. Lower the DB value better the cluster structure.

2.3 γ Index

It is the ratio of total number of objects in lower approximation of all clusters to the cardinality of the universe of discourse U[3] and is given by

$$\gamma = \frac{A}{B} \tag{5}$$

$$\text{Where } A = \sum_{g=1}^{S}|\underline{R}M_g| \text{ and } B = |U| = N \tag{6}$$

The γ index basically represents the quality of approximation of a clustering algorithm. Choice of rough fuzzy partition is based on maximum index value.

2.4 Fuzzy Bayesian Validation Method

Sung-Bae Cho[6] et al. proposed Bayesian validation method for validating fuzzy c-means clustering algorithm. This validates clustering results based on fuzzy membership grades and selects the optimal cluster number with largest Bayesian score.

$$BS = \frac{\sum_{g=1}^{S} P(M_g \mid T_h)}{S} \tag{7}$$

$$BS = \frac{\sum_{g=1}^{S} \prod_{h=1}^{N_h} \frac{P(M_g)P(T_{gh} \mid M_g)}{P(T_{gh})}}{S}, \tag{8}$$

$$Dh = \{T_{gh} \mid \mu_{gh>\alpha, 1\le h \le N}\} N_h = n(D_g) \tag{9}$$

Cluster probability is given by

$$P(M_g) = \frac{\sum_{g=1,\, \mu_{gh} > \infty}^{S} \mu_{gh}}{\sum_{g=1}^{S} \sum_{h=1}^{n} \mu_{gh}} \tag{10}$$

Probability for a data object to be in a cluster is given by

$$P(T_{gh}) = \sum_{g=1}^{S} P(M_g) P\left(\frac{T_{gh}}{M_g}\right) = \sum_{g=1}^{S} P(M_g)\mu_{gh} \tag{11}$$

The resultant proves Bayesian validation method to be more advantageous to produce optimal cluster partition for enormous UCI datasets over customary validity indices.

3. Rough Fuzzy C-means Algorithm

Rough fuzzy being one of the soft clustering algorithms integrates both crisp lower and fuzzy boundary concept, where fuzzy deals with uncertainty in boundary region and rough deals with overlapping partition. It consists of three parameters such as cluster prototype, lower region and boundary region. Lower approximation is given a full membership value of 1 and upper approximation membership value falls between [0, 1]. The objects in lower region indicate definite inclusion whereas objects in boundary region indicate possible inclusion[4].

Properties such as:

- An object T_h can be a part of at most one lower bound;
- $\Rightarrow T_h \in \overline{RM}_g \Rightarrow T_h \in \underline{RM}_g$; and
- An object T_h cannot be a part of any lower bound
- $\Rightarrow T_h$ belongs to two or more upper bounds.

The parameters such as w_{lr}, w_{br} and n are tunable parameters where $w_{br} = 1 - w_{lr}$ 0.5< w_{lr}<1, and 1<n<3. These tunable parameters have crucial impact on rough and rough fuzzy results. In some cases it affects the convergence of objective function too. For analysis, w_{lr} is set as 0.95 and w_{br} as 0.05. Cluster prototype calculation is based on crisp lower approximation and fuzzy boundary region. This paper implements Rough Fuzzy algorithm according to Pratipa et al[1].

Calculation of new centroid for each cluster X_g is shown below in equation (12).

$$X_g = \begin{cases} w_{lr} \frac{\sum_{T_h \in \underline{RM}_g} \mu_{gh}^m T_h}{\sum_{T_h \in \underline{RM}_g} \mu_{gh}^m} + w_{br} \frac{\sum_{T_h \in (\overline{RM}_g - \underline{RM}_g)} \mu_{gh}^m T_h}{\sum_{T_h \in (\overline{RM}_g - \underline{RM}_g)} \mu_{gh}^m}, \\ \quad if\ \underline{RM}_g \ne \varnothing \cap \overline{RM}_g - \underline{RM}_g \ne \varnothing \\ \frac{\sum_{T_h \in (\overline{RM}_g - \underline{RM}_g)} \mu_{gh}^m T_h}{\sum_{T_h \in (\overline{RM}_g - \underline{RM}_g)} \mu_{gh}^m},\ if\ \underline{RM}_g = \varnothing \cap \overline{RM}_g - \underline{RM}_g \ne \varnothing \\ \qquad \frac{\sum_{T_h \in \underline{RM}_g} \mu_{gh}^m T_h}{\sum_{T_h \in \underline{RM}_g} \mu_{gh}^m}, otherswise \end{cases} \tag{12}$$

Common objective function is given by equation (13).

$$J_{RF} = \begin{cases} W_{lr} \times U_1 + w_{br} \times V_1\ if\ \underline{RM}_g \ne \varnothing, \overline{RM}_g - \underline{RM}_g \ne \varnothing \\ U_1\ if\ \underline{RM}_g \ne \varnothing, \overline{RM}_g - \underline{RM}_g = \varnothing \\ V_1\ if\ \underline{RM}_g = \varnothing, \overline{RM}_g - \overline{RM}_g = \varnothing \end{cases} \tag{13}$$

$$U_1 = \sum_{g=1}^{S} \sum_{T_h \in \underline{RM}_g} (\mu_{gh})^m \left\| T_h - X_g \right\|^2 \tag{14}$$

$$V_1 = \sum_{g=1}^{S} \sum_{T_h \in \overline{RM}_g} (\mu_{gh})^m \left\| T_h - X_g \right\|^2 \tag{15}$$

Reduced objective function:

$$U_1 = \sum_{g=1}^{S} \sum_{T_h \in \underline{RM}_g} \left\| T_h - X_g \right\|^2 \tag{16}$$

3.1 Rough Fuzzy Bayesian Validation Method

Conventional indices such as ∝ index, γ index, ∝*index, Xie-Beni index, Davies-Bouldin index and Dunn index faces difficulty of monotonically increasing or decreasing index values. This may be due to the use of distance metric measure. However, these indices are limited in their capability to provide correct representation for rough fuzzy partition. Thus, the traditional indices lose their ability to provide optimal number of clusters when cluster size approaches to the number of samples. Performance analysis has been done comparing the conventional indices with our proposed index. Results manifests maximum Bayesian score ensures best clustering structure.

Bayesian score is calculated based on the equation given below.

$$BS = w_{lr}\frac{\sum_{g=1,\,T_h \in \underline{R}M_g}^{S} P(M_g \mid T_h)}{S} + w_{br}\frac{\sum_{g=1,\,T_h \in \bar{R}M_g}^{S} P(M_g \mid T_h)}{S} \tag{17}$$

$$BS = w_{lr}\frac{\sum_{g=1,\,T_h \in \underline{R}M_g}^{S} \prod_{h=1}^{N_h} \frac{P(M_g)P(T_{gh} \mid M_g)}{P(T_{gh})}}{S} + w_{br}\frac{\sum_{g=1,\,T_h \in \bar{R}M_g}^{S} \prod_{h=1}^{N_h} \frac{P(M_g)P(T_{gh} \mid M_g)}{P(T_{gh})}}{S} \tag{18}$$

$$T_h = \{T_{gh} \mid \mu_{gh>a,1 \le h \le N}\} N_h = n(T_k)$$

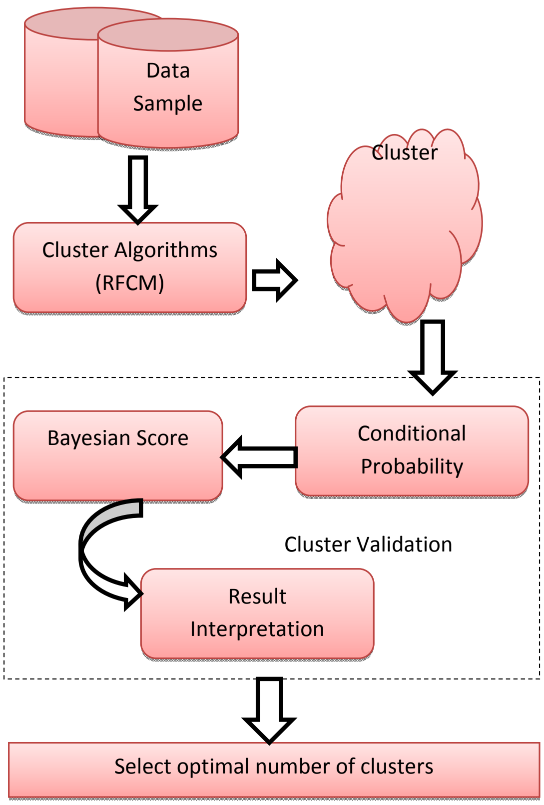

Figure 1. Architectural Diagram.

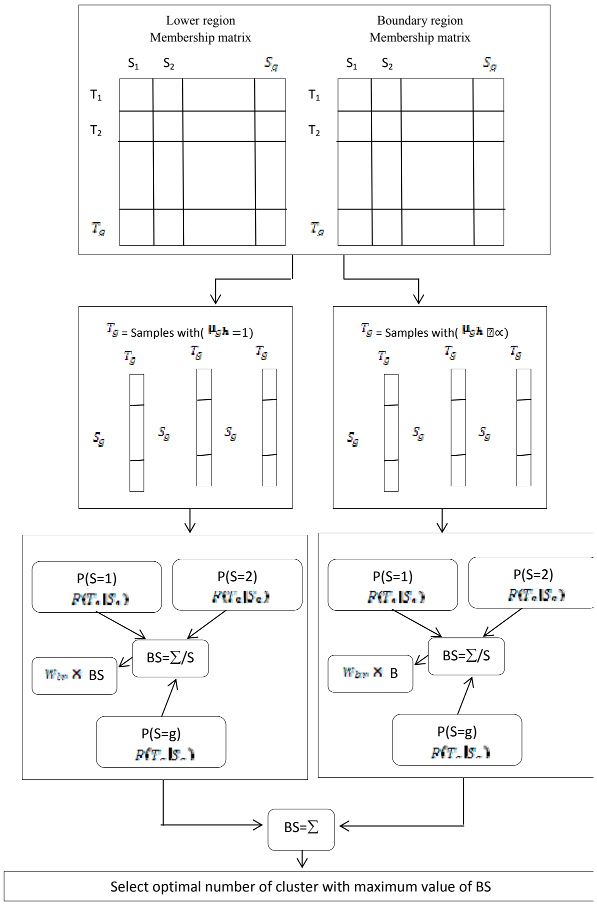

Figure 2. Rough fuzzy Bayesian validation method.

3.1.1 *Algorithm for Computing Bayesian Score for Rough Fuzzy Clustering*

Step 1: Retrieve the final membership values obtained thro Rough Fuzzy Clustering.

Step 2: Compute Bayesian score for lower regions of given clustering structure.

Step 3: Compute Bayesian score for boundary region of given clustering structure.

Step 4: Summate the Bayesian score for lower and boundary region.

3.1.2 *Algorithm for Computing Bayesian Score for Lower Region of Clustering Structure*

Step 1: Retrieve the Membership values

Step 2: Determine the lower region cluster probability by equation (19).

$$P(M_g) = \frac{\sum_{g=1,\,\mu_{gh}=1}^{S} \mu_{gh}}{\sum_{g=1,}^{S} \sum_{h=1}^{N} \mu_{gh}} \tag{19}$$

Step 3: Determine the probability for assigning the elements to Lower region of a particular cluster by equation (20).

$$P(T_{gh}) = \sum_{g=1, T_h \in \underline{R}M_g}^{S} P(M_g)P\left(\frac{T_{gh}}{M_g}\right) = \sum_{g=1, T_h \in \underline{R}M_g}^{S} P(M_g)\mu_{gh} \tag{20}$$

Step 4: Compute the Bayesian score for lower regions of given clustering structure as shown in equation (21).

$$BS = w_{lr} \frac{\sum_{g=1, T_h \in \underline{R}M_g}^{S} \prod_{h=1}^{N_h} \frac{P(M_g)P(T_{gh} \mid M_g)}{P(T_{gh})}}{S} \tag{21}$$

3.1.3 *Algorithm for Computing Bayesian Score for Boundary Region of Clustering Structure*

Step 1: Retrieve Membership values.

Step 2: Determine the boundary region cluster probability as shown in equation (22).

$$P(S_g) = \frac{\sum_{g=1, \mu_{gh} > a}^{S} \mu_{gh}}{\sum_{g=1,}^{S} \sum_{h=1}^{N} \mu_{gh}} \tag{22}$$

Step 3: Determine the probability for assigning elements to the boundary region of a particular cluster as shown in equation (23).

$$P(T_{gh}) = \sum_{g=1, T_h \in \bar{R}M_g}^{S} P(M_g)P\left(\frac{T_{gh}}{M_g}\right) = \sum_{g=1, T_h \in \bar{R}M_g}^{S} P(M_g)\mu_{gh} \tag{23}$$

Step 4: Compute the Bayesian score for boundary regions of given clustering structure by equation (24).

$$BS = w_{br} \frac{\sum_{g=1, T_h \in \bar{R}M_g}^{S} \prod_{h=1}^{N_h} \frac{P(M_g)P(T_{gh} \mid M_g)}{P(T_{gh})}}{S} \tag{24}$$

4. Experimental Results and Discussion

Experiments analyzed in this paper are been implemented in Matlab R2009a. Bayesian like validation method is proposed for rough fuzzy clustering, representing lower and boundary region. Lower region is given a full membership $\mu_{gh} = 1$ and boundary region takes the membership between [0, 1] for values greater than the ∝-cut. The ∝-cut value ranges from 0.1 to 0.6 considering the computed average value to be the Bayesian score. The maximum number of cluster partition is determined by √n.

4.1 Synthetic Dataset

The sample dataset contains 10 instances with two attributes. Optimal number of clusters for this dataset has been identified as three.

4.2 Benchmark Datasets

4.2.1 *Iris Dataset*

It contains 150 samples in four dimensional measurement spaces. Iris consists of two or three clusters because of the substantial overlap of two of the clusters. It consists of four attributes which includes sepal length in cm, sepal width in cm, petal length in cm, petal width in cm. It consists of three classes such as Iris setosa, Iris versicolour, and Iris virginica.

4.2.2 *Seed Dataset*

Seed dataset contain 210 instances with 7 attributes such as 1.area 2.perimeter 3.compactness 4. Length of kernel

Table 2. Synthetic Dataset

Instances	Attributes	
	X1	X2
1	13	13
2	14	14
3	15	15
4	33	33
5	34	34
6	35	35
7	63	63
8	64	64
9	65	65
10	45	45

Table 3. Comparison of Rough Fuzzy validation techniques Synthetic Dataset

Dataset	Number of clusters	Fuzzifier (n)	∝ index	Gamma index	Davies-Bouldin	Bayesian Score
Sample	2	1.2	0.9766	0.7000	0.0612	0.9958
	3	1.2	0.9943	0.9000	0.0254	1

Table 4. Comparison of Rough Fuzzy validation techniques for Iris Dataset

Dataset	Number of clusters	Fuzzifier (n)	∝ index	Gamma index	Davies-Bouldin	Bayesian Score
Iris	2	1.3	0.9968	0.9333	0.1120	0.9780
	3	1.3	0.9882	0.8133	0.1946	0.9792
	4	1.64*	0.9666	0.6267	0.3186	0.9750
	5	1.7	0.9749	0.5800	0.3969	0.9688
	6	2*	0.9795	0.4800	0.3321	0.9696
	7	2.1*	0.8422	0.4800	0.2838	0.9748

* indicates convergence up to some extend for this fuzzifier value

Table 5. Comparison of Rough Fuzzy validation techniques for Seed Dataset

Dataset	Number of clusters	Fuzzifier (n)	∝ index	Gamma index	Davies-Bouldin	Bayesian Score
Seed	2	1.17	0.9978	0.9600	0.1213	0.9676
	3	1.21	0.9937	0.8867	0.2432	0.9781
	4	1.2	0.9878	0.8190	0.3479	0.9692
	5	1.24*	0.9838	0.7905	0.2537	0.9710
	6	1.24*	0.9855	0.7810	0.3163	0.9737
	7	1.25*	0.9848	0.7476	0.3768	0.9701

* indicates convergence up to some extend for this fuzzifier value

Table 6. Comparison of Rough Fuzzy validation techniques for Balloon Dataset

Dataset	Number of clusters	Fuzzifier (n)	∝ index	Gamma index	Davies-Bouldin	Bayesian Score
balloon	2	1.12	0.9519	0.5000	0.5801	0.9833
	3	1.8	0.9579	0.4000	0.3526	0.9751
	4	1.8	0.9568	0.3000	0.3849	0.9793

5. Width of kernel 6. asymmetry coefficient 7. length of kernel groove. All of these parameters were real-valued continuous. It contain three class attribute.

4.2.3 Balloon Dataset

It contains 20 samples with four attributes which are as follows 1.colour 2.size 3.act 4.age. It contains three class attribute which is either inflated or not.

4.2.4 Teaching Assistant Evaluation

Teaching assistant evaluation dataset contains 151 instances with 5 attributes. The attribute information such as 1. whether or not the teaching assistant is a native English speaker 2. course instructor 3. course 4. summer

Table 7. Result for teaching assistant evaluation Dataset

Dataset	Number of clusters	Fuzzifier (n)	∝ index	Gamma index	Davies-Bouldin	Bayesian Score
Teaching Assistant Evaluation	2	1.28	0.9893	0.8079	0.4290	0.9674
	3*	1.2	0.9832	0.7417	0.5852	0.9755
	4*	1.2	0.9700	0.6689	0.6423	0.9717
	5	1.55	0.9699	0.4901	0.3915	0.9695
	6	1.55	0.9794	0.5762	0.3168	0.9713
	7	1.5	0.9781	0.5497	0.3392	0.9722

* indicates convergence up to some extend for this fuzzifier value

Table 8. Comparison of overall Rough Fuzzy Cluster validity indices

Dataset	S^*	∝ index	γ index	Davies-Bouldin index	Bayesian score
Synthetic	3	3	3	3	3
Iris	2 or 3	2	2	2	3
Seed	3	2	2	2	3
Balloon	2	3	4	2	2
Teaching assistant evaluation	3	2	2	6	3

S^* optimal number of clusters according to UCI repository

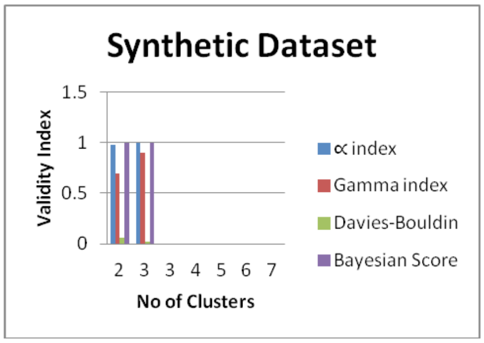

Figure 3. Comparison chart for synthetic dataset.

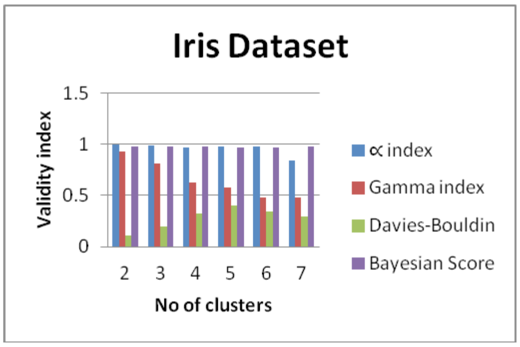

Figure 4. Comparison chart for Iris dataset.

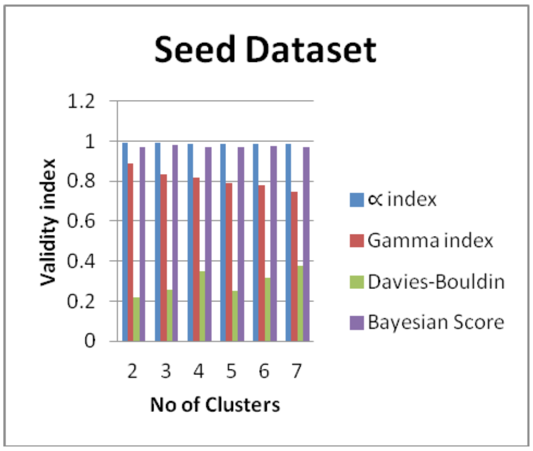

Figure 5. Comparison chart for Seed dataset.

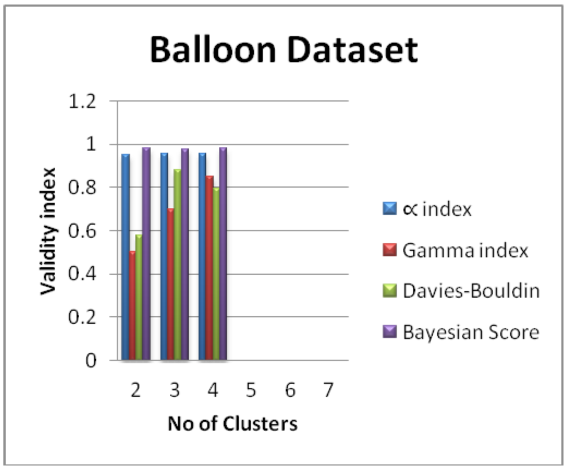

Figure 6. Comparison chart for Balloon dataset.

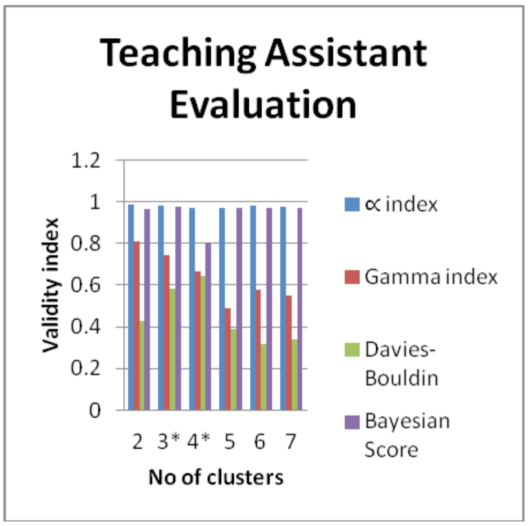

Figure 7. Comparison chart for teaching evaluation dataset.

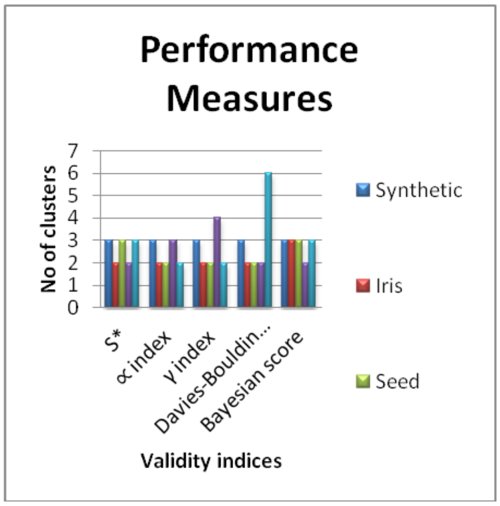

Figure 8. Comparison chart for overall validity indices.

or regular semester 5. class size. It contains three class attributes such as 1. low 2. medium 3. high.

5. Performance Analysis

The performance of this measure is compared with four conventional validity indices and experimented with four well known UCI datasets.

6. Conclusion

Rough fuzzy Clustering is an effective clustering algorithm for cluster analysis. This paper proposes a new validation scheme called rough fuzzy Bayesian like validation method for validating rough fuzzy clustering results. When the cluster number S is not known priori, determining the optimal number of clusters is a tough task. Thus compared to the customary validity indices our proposed index is found to be more reliable and consistent. Hence our proposed index is found to attain its maximum score when the optimal number of cluster is achieved for rough fuzzy partition. Experimental analysis proves our scheme works effectively for tremendous UCI datasets.

7. References

1. Maji P, Pal SK. RFCM: A hybrid clustering algorithm using rough and fuzzy sets. Fundamenta Informaticae. 2007; 80(4):475–96.
2. Wu K-L, Yang M-S. A cluster validity index for fuzzy clustering. Pattern Recognition Letters. 2005; 26(9):1275–91.

3. Mitra S, Banka H, Pedrycz W. Rough fuzzy collaborative clustering. IEEE Transactions on System, Man and Cybernetics. 2006 Aug; 36(4):795–805
4. Maji P, Pal SK. Rough set based generalized fuzzy c-means algorithm and quantitative indices. IEEE Transactions n System, Man and Cybernetics. 2007 Dec; 37(6):1529–40
5. Revathy S, Parvathavarthini B. Comparison of FCM and RFCM on iris dataset using Matlab. International Conference on Computer Networks and Information Technology; 2012; Bangkok.
6. Cho S-B, Yoo S-H. Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data. Pattern Recognition. 2009; 39(12):2405–14.
7. Ramze Rezaee M, Lelieveldt BPF, Reiber JHC. A new cluster validity index for the fuzzy c-mean. Pattern Recogn Lett. 1998; 19(3–4):237–46.
8. Kovacs F, Legany C, Babos A. Cluster validity measurement techniques.
9. Halkidi M, Batistakis Y, Vazhirgiannis M. Cluster validity checking methods: part II. Sigmod Record. 2002; 31(3):19–27.
10. Maulik U, Bandyopadhyay S. Performance evaluation some clustering algorithms and validity indices. IEEE Trans Pattern Anal Mach Intell. 2002 Dec; 24(12):1650–4
11. Li F, Min F, Liu Q. Intra-cluster similarity index based on fuzzy rough sets for fuzzy c-means algorithm.