Decision Theoretic Evaluation of Rough Fuzzy Clustering

[1]**Revathy Subramanion;** and [2]**Parvathavarthini Balasubramanion**

[1]Sathyabama University, Chennai, India
[2]St.Joseph's College of Engineering, Anna University, Chennai, India

ID # (2785)
Received: 25/ 11/ 2013
In-revised: 04/07/2014
Correspondent Author:
Revathy Subramanion
E. mail: ramesh.revathy@gmail.com

KEYWORDS

clustering, rough set theory, cluster validity index, decision theory, fuzzy clustering, risk measure

ABSTRACT

Clustering is the process of organizing dissimilar objects into natural groups in such a way objects in the same group is more similar than objects in the different groups. Since we know clustering is an unsupervised learning problem, typical clustering algorithms not achieving its end to handle uncertainty that exists in the real life experience. Though fuzzy clustering handles incompleteness and vagueness in the data set efficiently, it is highly descriptive than hard clustering algorithm. Rough clustering algorithm is the popular soft clustering technique which uses rough set to handle uncertainty. In Rough Fuzzy clustering, each cluster is represented by centroid, crisp lower approximation and fuzzy boundary. Clustering undergoes sequence of partitions where cluster evaluation is the final step in clustering process. Efficient clustering structure can be obtained through validity measures. Various validity measures have been proposed to evaluate rough fuzzy clustering. Since those measures are Geometric measures, this paper proposes decision theoretic measure for validating rough fuzzy clustering structure.

قرار التقييم النظري للتجميع الضبابي الاستقرابي

[1]**ريفاثي سوبرامانيون**، و [2]**بارفاثافرثيني بالاسوبرامانيون**

[1]جامعة ساثيابما، تشيناي، الهند
[2]كلية سان جوزيف للهندسة، جامعة آنا، تشيناي، الهند

رقم المسودة: (2785)
تاريخ استلام المسودة: 2013/11/25
تاريخ المسودة المُعَدَلة: 2014/07/04
الباحث المُرَاسِل: ريفاثي سوبرامانيون
بريد الكتروني:
ramesh.revathy@gmail.com

الكلمات الدالة

التجميع العنقودي، نظرية مجموعة الاستقراب، مؤشر صلاحية العنقود، نظرية القرار، التجميع العنقودي الضبابي، معيار الخطر

المُستلخص

إن التجميع العنقودي هو عملية تنظيم لأشياء مختلفة في مجموعات طبيعية بطريقة تبدو فيها الأشياء من نفس المجموعة أكثر تشابهًا من غيرها في المجموعات المختلفة. وحيث نعلم أن التجميع العنقودي مشكلة تعلم لم تخضع لإشراف، فإن خوارزميات التجميع العنقودي النموذجية لا تحقق المراد منها لإدارة الريبة الموجودة في تجارب الحياة الحقيقية. مع ذلك، فإن التجميع العنقودي الضبابي يتناول عدم اكتمال وغموض مجموعات البيانات بكفاءة، إذ يتميز بالوصفية الشديدة أكثرمن خوارزمية التجميع العنقودي المتشدد. إن خوارزمية التجميع العنقودي الاستقرابي طريقة تجميع مرنة وشهيرة تستخدم مجموعات استقراب لإدارة الريبة. وفي التجميع العنقودي الضبابي الاستقرابي، كل عنقود يمثله نقطة وسطى وتقريب سفلي متموج، وحد ضبابي. يخضع التجميع العنقودي إلى سلسلة من الحواجز حيث يكون تقييم العنقود هو الخطوة الأخيرة في عملية التجميع العنقودي. يمكن الحصول على هيكل عنقودي كفء مع خلال إجراءات الصلاحية. طرحت العديد من معايير الصلاحية إمكانية تقييم التجميع العنقودي الضبابي الاستقرابي. وحيث إن هذه المعايير هندسية، يتناول هذا البحث قرار بشأن المعيار النظري لصلاحية هيكل التجميع العنقودي الضبابي الاستقرابي.

Introduction

Cluster analysis (Mirkin, 1996) organizes a set of unlabeled objects into subsets, such that the objects belonging to the same group of cluster are more similar than those belonging to different group of clusters. It is a main task of exploratory data mining and a common technique for statistical data analysis, widely applied in many areas such as pattern recognition, Marketing, information retrieval, bioinformatics and so on.

In real life, a data point should belong to multiple clusters. This problem can be brought to a solution based on different soft computing approaches. For instance, Fuzzy *c*-means (FCM) algorithm (Bezdek, 1981) uses Fuzzy set representation of clusters. As an important approach for uncertain and vague data analysis, the theory of Rough sets was also incorporated in HCM framework to develop the Rough *c*-means (RCM) algorithm (Lingras, 2004). Like HCM, RCM can be classified into the partitional clustering methods where RCM can also assign a data point to more than one clusters, but uses different restrictions. Representing cluster in RCM includes lower and upper approximations. Based on the distance between a datapoint and cluster centroid, each data point may be assigned to the lower approximation of a certain cluster (and hence in the upper approximation of this cluster) or the boundary areas of two or more clusters (and hence in the upper approximations of the clusters). (Mitra, *et al,* 2006) proposed a Rough Fuzzy *c*-means clustering algorithm (RFCM) with fuzzy lower approximations and fuzzy boundaries and (Maji and Pal,2007) proposed Rough Fuzzy Possibilistic *c*-means clustering algorithm (RFPCM).

To determine the number of clusters, good cluster validity checking method is helpful. Clustering validity evaluation is concerned with “assessing the validity of clustering that has been obtained from the application of clustering procedure”. In general, a cluster validation checking method includes a measure of cluster quality and the optimal number of clusters for some kind of clustering algorithms.

Cluster Validity Index (CVI) is a helpful criterion used to assess the quality of clustering and are based on external criteria, internal criteria and relative criteria (Gardon, 1999). For example, some indices are based on considering the compactness within each cluster or the separation between clusters (Bouguessa, *et al,* 2006), some on the information entropy (Liang, *et al,* 2012). Xie–Beni Index and its derivations need to assign a membership u_{it} for an object x_t to a cluster, (Beni,1991). On the other side, the decision-theoretic rough set (DTRS) model introduced by (Yao, *et al*,2007) has been verified to be helpful in providing a better understanding of clustering, which inspires us to determine the number of clusters through the DTRS model.. This paper proposes the new cluster validity index for rough fuzzy clustering which utilizes this DTRS model.

Materials and Methods

(1) Related Work

(1.1) Rough Fuzzy C Means Clustering

This allows one to incorporate fuzzy membership value u_{ik} of a sample X_k to a cluster mean V_i relative to all other means $V_j \in j = i$, instead of the absolute individual distance d_{ik} from the centroid. The major steps of the algorithm are provided below.

(i) Assign initial means V_i for the *c* clusters.

(ii) Compute u_{ik} by (3) for *c* clusters and *N* data objects.

(iii) Assign each data object (pattern) X_k to the lower approximation $\underline{B}U_i$ or upper approximation $\overline{B}U_i$, $\overline{B}U_i$ of cluster pairs U_i and U_j by computing the difference in its membership $u_{ik} - u_{jk}$ to cluster centroid pairs V_i and V_i .

(iv) Let u_{ik} be maximum and u_{jk} be the next to maximum.

If $u_{ik} - u_{jk}$ is less than some threshold, then $X_k \in \overline{B}U_i$ and $X_k \in \overline{B}U_i$ and X_k cannot be a member of any lower approximation, else $X_k \in \underline{B}U_i$ such that membership u_{ik} is maximum over the *c* clusters.

(v) Compute new mean for each cluster U_i, incorporating (2) and (3) into (4), as in (9), shown at the bottom of the page.

(vi) Repeat Steps 2)–5) until convergence, i.e., there are no more new assignments.

We use $w_{up} = 1 - w_{low}$, $0.5 < w_{low} < 1$, $m = 2$, and $0 < threshold < 0.5$

$$v_i = \begin{cases} w_{low}\dfrac{\sum_{x_k\in \underline{B}U_i} u_{ik}^m X_k}{\sum_{X_k\in \underline{B}U_i} u_{ik}^m} + w_{up}\dfrac{\sum_{x_k\in(\overline{B}U_i-\underline{B}U_i)} u_{ik}^m X_k}{\sum_{x_k\in(\overline{B}U_i-\underline{B}U_i)} u_{ik}^m}, \quad if\ \underline{B}U_i \neq \emptyset \cap \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ \dfrac{\sum_{X_k\in(\overline{B}U_i-\underline{B}U_i)} u_{ik}^m X_k}{\sum_{X_k\in(\overline{B}U_i-\underline{B}U_i)} u_{ik}^m}, if\ \underline{B}U_i = \emptyset \cap \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ \dfrac{\sum_{X_k\in \underline{B}U_i} u_{ik}^m X_k}{\sum_{X_k\in \underline{B}U_i} u_{ik}^m}, otherwise \end{cases} \quad (1)$$

(1.2) Cluster Validity

Although cluster evaluation is the final for clustering process, yet in order to obtain good clustering structure, typical objective functions formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity which uses cluster validity indices to evaluate the resulting clusters. Following are some of the cluster validity indices used to evaluate rough fuzzy clustering structure.

(1.3)Rough Fuzzy Cluster Quality measures

(i) α Index: α index $= \frac{1}{m1}\sum_{i=1}^{c}\frac{wA_i}{wA_i+\tilde{w}B_i}$ (2)

where $A_i = \sum_{x_j\in \underline{A}(\beta_i)} (\mu_{ij})^{m1} = |\underline{A}(\beta_i)|$;and

$B_i = \sum_{x_j\in B(\beta_i)} (\mu_{ij})^{m1}$ (2)

μ_{ij} constitutes the probabilistic memberships of object x_j in cluster B_i. The parameters w and $\tilde{w}$ correspond to the relative importance of lower and boundary region. The α index provides the average accuracy of c clusters. It is the average of the ratio of the number of elements in lower approximation to that in upper approximation of each cluster. In effect, it captures the average degree of completeness of knowledge about all clusters. A good clustering procedure should make all objects as similar to their centroids as possible.The α index increases with increase in similarity within a cluster. Therefore, for a given data set and c value, the higher the similarity values within the clusters, the higher would be the α value. Thus similarity value is directly proportional to the α value. The value of α also increases with c.

(ii) η Index: $\tilde{n}$ η index = 1-α =1-$\frac{1}{c}\sum_{i=1}^{c}\frac{wA_i}{wA_i+\tilde{w}B_i}$ (3)

The η index corresponds to the average roughness of c clusters and is described by subtracting the average accuracy α from 1, where A_i and B_i are given by Equation 3. Note that the lower the value of η, the better is the overall clusters approximations. Thus η is inversely proportional to the overall cluster approximations. Also, 0 ≤η≤ 1. Basically, η index states the average degree of incompleteness of knowledge about all clusters.

(iii) α^* α*Index: α*= $\frac{C}{D}$; where C = $\sum_{i=1}^{c} wA_i$; and D = $\sum_{i=1}^{c}\{wA_i + \tilde{w}B_i\}$ (4)

The α* index constitutes the accuracy of approximation of all clusters. It captures the exactness of approximate clustering. The value of α* being as high as possible makes it a best clustering procedure. The exactness of approximate clustering is maximized by α* index.

(iv) γ Index: $\gamma = \frac{R}{S}$; where R = $\sum_{i=1}^{c}|\underline{A}(\beta_i)|$; and S = $|U|$ = n (5)

It is the proportion of the total number of objects in lower approximations of all clusters to the cardinality of the universe of discourse U. The index primarily symbolizes the quality of approximation of clustering algorithm.

(2) Proposed Rough Fuzzy Cluster Quality Index

(2.1) Risk for Assigning Objects under Clusters

For rough fuzzy clustering (RC), an object $\vec{x_i}$ may belong to more than one cluster. Moreover, each cluster $\vec{c_i}$ is represented by its lower approximation $\underline{apr}(\vec{c_i})$ and upper approximation $\overline{apr}(\vec{c_i})$. There also exists the boundary region bnd$(\vec{c_i}) = \overline{apr}(\vec{c_i}) - \underline{apr}(\vec{c_i})$.

Let C={c1,c2,c3} Then b={{c1},{c2},{c3},{c1,c2},......} B={b1,b2,b3..........bj}

For rough fuzzy clustering, let $b_j(RC,\vec{x_i})$ be the action that assigns the object $\vec{x_i}$ to sub clusters. The loss function for $\vec{x_i}$ can be expressed as follows:

$\lambda_{\vec{x_i}}(b_j(CS,\vec{x_i})|\vec{x_i})$=0, if $\vec{c_i} \in b_j(RC,\vec{x_i})$ (6)

$\lambda_{\vec{x_i}}(b_j(CS,\vec{x_i})|\vec{x_i})$=1, if $\vec{c_i} \notin b_j(RC,\vec{x_i})$

The risk associated with the assignment will then be given as $R(b_j(CS,\vec{x_i})|\vec{x_i})$.It can be calculated by multiplying loss and probability. Risk measure is defined as,

$R(b_j(CS,\vec{x_i})|\vec{x_i}) = \sum_{i=1..k,c_i\notin b_j(RC,\vec{x_i})} \lambda_{\vec{x_i}}(b_j(CS,\vec{x_i})|\vec{x_i})^*$ $P(\vec{c_i}|\vec{x_i})$ (7)

In Rough Fuzzy Clustering $P(C_i|X_l)$ can be calculated using the following equation

$P(\vec{c_i}|\vec{x_i}) = \frac{P(Ci)P(Xl|Ci)}{P(Xl)}$ (8)

Where

P(Ci) $= \frac{\sum_{j=1}^{n}\mu ij}{\sum^{c}\sum^{n}\mu ij}$ (9)

$P(X_l)$ $=\sum_{i=1}^{c} P(Ci)\,P\left(\frac{dij}{ci}\right)$ (10)

$P(X_l)$ $=\sum_{i=1}^{c} P(Ci)\,\mu ij$ (11)

(2.2)Risk for Lower Approximation

For rough clustering, let R(RC, $\underline{apr}(\vec{c_i})$)be the risk for a lower approximation.

R(RC, $\underline{apr}(\vec{c_i})$) $=\sum_{\vec{x}_l\notin \underline{apr}(\vec{c_i})} R(b_j(CS,\vec{x_i})|\vec{x_i})$ (12)

(2.3)Risk for Upper Approximation

For rough clustering, let R(RC, $\underline{apr}(\vec{c_i})$) be the risk for a upper approximation.

(RC,$\overline{apr}(\vec{c_i})$)= $\sum_{\vec{x}_l\notin \overline{apr}(\vec{c_i})} R(b_j(CS,\vec{x_i})|\vec{x_i})$ (13)

(2.4)Risk for Boundary Area

For rough clustering, let R(RC, bnd($\vec{c_i}$))be the risk for the boundary area of $\vec{c_i}\vec{c_i}$.

$$R(RC, bnd(\vec{c_i})) = \sum_{\vec{x}_l \in bnd(\vec{c_i})} R\left(b_j(RC, \vec{x_l}) | \vec{x_l}\right) \qquad (14)$$

(3)Proposed Algorithm

Input : Rough Fuzzy clusters (Obtained thro Rough Fuzzy C Means Clustering).

Output: Best Rough Fuzzy Clustering Structure

1. repeat/, 2. for i=2 to n-1/, 3. for each C_i do /, 4.for each object X_l of Lower Approximation of C_i /, 5. calculate risk $R(b_j(C,X_l))$/, 6. end/, 7. calculate R(RC, $\underline{apr}(\vec{c_i})$)/. 8. for each object X_l of Upper Approximation of C_i /, 9. calculate risk $R(b_j(C,X_l))$/, 10. end/, 11 calculate R(RC, $\overline{apr}(\vec{c_i})$)/, 12. for each object X_l of Boundary of Ci/, 13. calculate risk $R(b_j(C,X_l))$ /,14. end/, 15. calculate R(RC, bnd($\vec{c_i}$)($\vec{c_i}$))/,16. Calculate Risk R(RFC,Ci) for Clustering Structure Ci/, 17.end/, 18. for c=2 to n-1/, 19.compare risk for each clustering structure(CS) /, 20. end/, and, 21. Best Clustering Structure(c) =Clustering structure with minimum risk.

Results and Discussion

(1)Syntactic Data

Let us illustrate the proposed risk measure for two different clustering schemes, rough clustering and rough fuzzy clustering, with the following example. The following data set consists of 10 objects described by two features, table 1.

Table 1: 2- Dimension Data Set

Object (2 Dimension)	
1.7	1.7
2.1	1.8
1.6	2.1
3.5	2.7
3.5	5.1
3.1	5.2
3.3	4.7
7.7	4.6
7.8	5.2
8.2	4.7

For rough clustering, we set k = 3 and wlow= 0:8. We also adjust the threshold to obtain the results presented in figure 1. In the figure, the small circle represents the lower approximation and large circle represents the upper approximation of each cluster. Since fourth object is closer to both c1 and c2, it belongs to upper approximation of both the clusters.

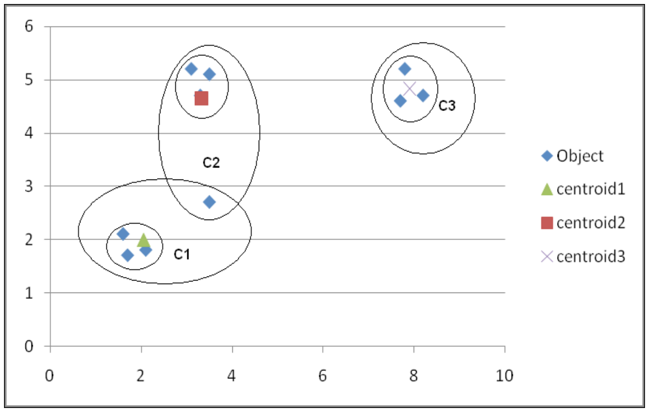

Figure 1: Rough Clustering

For rough fuzzy clustering, we set k = 3 and wlow= 0:8. We also adjust the threshold to obtain the results presented in figure 2. In the figure, the small circle represents the lower approximation and large circle represents the upper approximation of each cluster. Since fourth object is closer to both c1 and c2, it belongs to upper approximation of both the clusters. According to rough fuzzy clustering fourth object is having certain membership value in both the clusters.

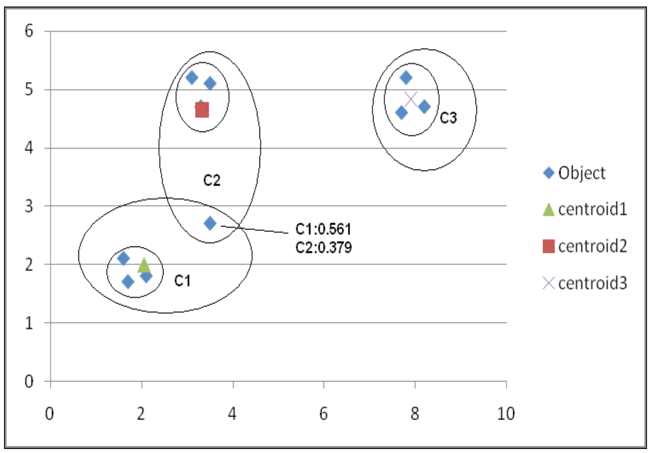

Figure 2: Rough Fuzzy Clustering

Table 2 shows centroid and risk for each cluster under rough clustering. Hence risk for clustering structure can be calculated by adding risk of all the clusters.

Table 2: Risk for Rough Clustering

Object (2 Dimension)		Risk	Centroid	Cluster
1.7	1.7	00.44	2.05,1.99	C1
2.1	1.8	0.474		
1.6	2.1	00.47		
3.5	2.7	0.409	3.3,4.65	C1,C2
3.5	5.1	0.519		C2
3.1	5.2	0.504		
3.3	4.7	0.553		
7.7	4.6	0.393	7.9,4.83	C3
7.8	5.2	0.374		
8.2	4.7	0.354		

Table 3 shows centroid and risk for each cluster under rough fuzzy clustering. Hence risk for clustering structure can be calculated by adding risk of all the clusters.

Table 3: Risk for Rough Fuzzy Clustering

Object (2 Dimension)		Risk	Centroid	Cluster
1.7	1.7	0.451	2.05,1.99	C1
2.1	1.8	0.477		
1.6	2.1	0.483		
3.5	2.7	0.327	3.3,4.65	C1,C2
3.5	5.1	0.535		C2
3.1	5.2	0.528		
3.3	4.7	0.553		
7.7	4.6	0.393	7.9,4.83	C3
7.8	5.2	0.378		
8.2	4.7	0.357		

Figure 3 shows risk for rough clustering with respect to the number of clusters. It shows minimum risk value the cluster number c=3.Hence we can visually verify the results.

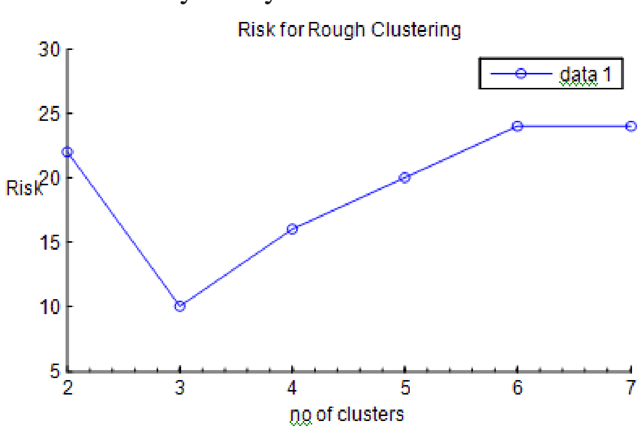

Figure 3: Risk for Rough Clustering with Respect to Number of Clusters

Figure 4 shows risk for rough fuzzy clustering with respect to the number of clusters. It shows minimum risk value the cluster number c=3.It shows that rough fuzzy clustering provides minimum risk value than rough clustering

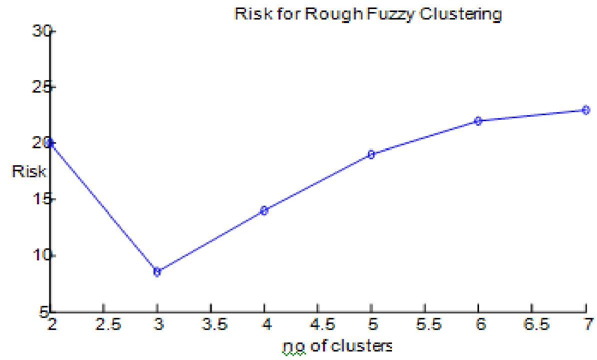

Figure 4: Risk for Rough Clustering with Respect to Number of Clusters

Figure 5 and Figure 6 shows change in risk with threshold for rough and rough fuzzy clustering. Rough Clustering gives minimum risk value for the threshold 1.5 and rough fuzzy clustering gives minimum risk value for the threshold 1.4.

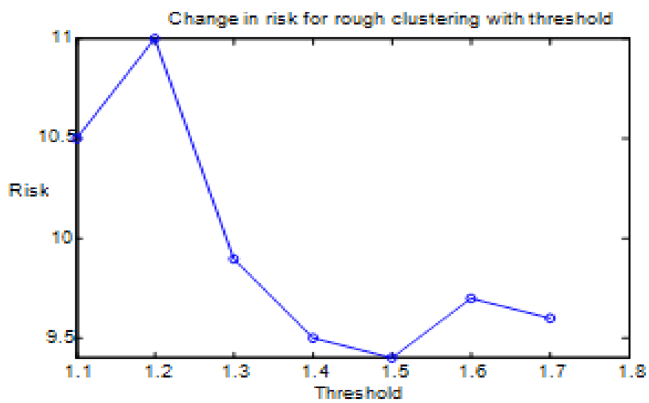

Figure 5: Risk Vs Threshold

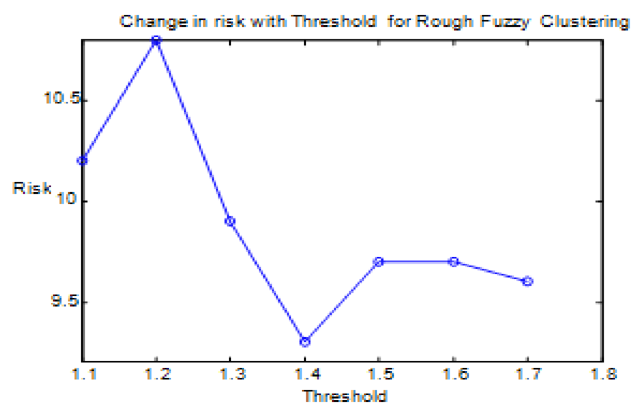

Figure 6: Risk Vs Threshold

(2) Wisconsin Breast Cancer Data

The previous section used synthetic data that was designed to highlight and test salient features of the proposed risk based measure. In this section, we use a standard real-world data set .The testing for such a standard data set makes it possible to compare the proposed approach with some of the previous clustering results. Wisconsin breast cancer databases were obtained from the University of Wisconsin Hospitals. This data set contains 699 instances that fall into two classes: benign (458 instances) and malignant (241 instances). Each instance is represented by nine attributes, all of which are scaled to a 1:10 range. The variation in risk for different values of threshold is shown in figure7 and figure 8. The risk seems to decline from threshold value of 1.1 to 1.7. However, there is a sharp drop in risk when the threshold is reduced from 1.3 to 1.4. Therefore, Threshold of 1.4 can again be used as an appropriate value.

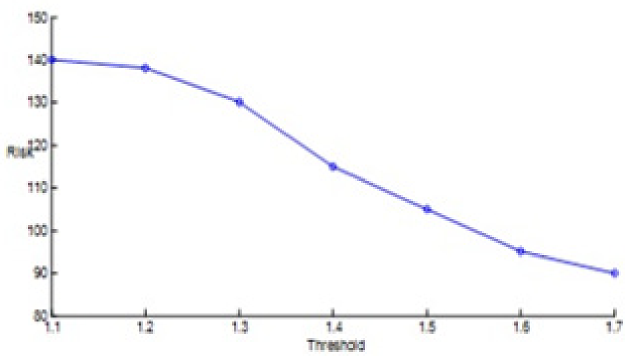

Figure 7: Breast Cancer Data: Rough Clustering

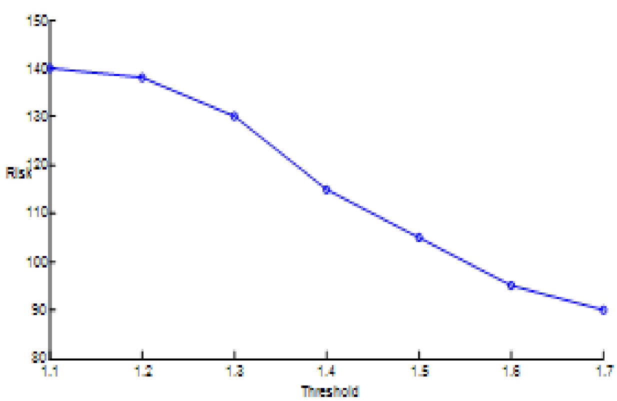

Figure 8: Breast Cancer Data: Rough Fuzzy Clustering

It can be seen from both figure 9 and figure10, that the risk of clustering is minimum for two clusters and then continuously rises. Since we want to group the objects into two categories: benign and malignant, this risk measure provides appropriate number of clusters.

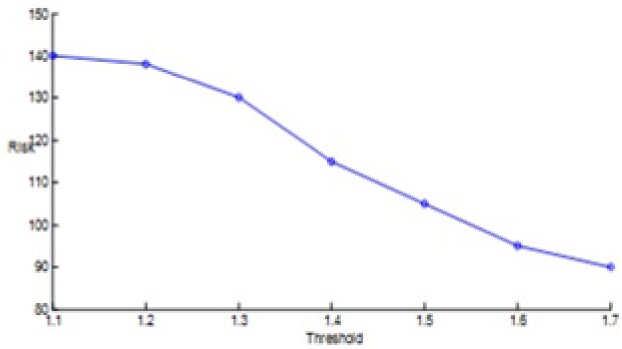

Figure 9: Breast Cancer Data: Rough Clustering

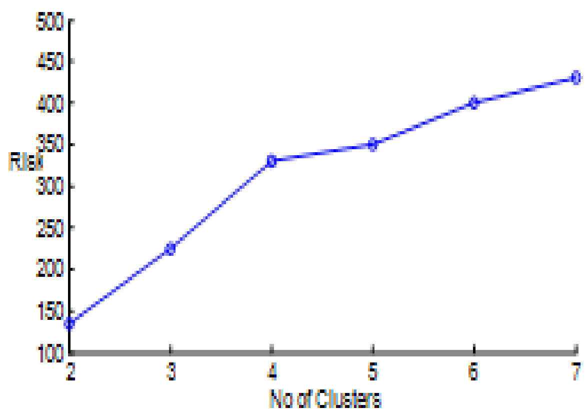

Figure 10: Breast Cancer Data: Rough Fuzzy Clustering

Table 4 shows risk for rough and rough fuzzy clustering for both (when c=2 & when c=3) synthetic data set and breast cancer data set.

Table 4: Risk Values When c=2 & When c=3

Clustering Scheme	Synthetic Data Set		Breast Cancer Data set	
	When c=2	When c=3	When c=2	When c=3
Rough	5.32	4.49	252.67	140.51
Rough Fuzzy	5.21	4.48	225.79	134.64

Conclusion

This paper proposed a rough fuzzy cluster validity index based on decision theory. The proposal uses a risk measure to construct the validity index. Therefore, the cluster quality is evaluated by considering the total risk of categorizing all the objects. Such a decision-theoretic representation of cluster quality may be more useful in business-oriented data mining than traditional geometry-based cluster quality measures. Rough fuzzy clustering provides lesser risk than rough clustering.

References

Bezdek JC (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, NewYork, USA.
Available at: http://www.abebooks.com/servlet/BookDetailsPL?bi=7843722459

Bouguessa M; Wang S; and **Sun H** (2006) An Objective Approach to Cluster Validation. *Pattern Recognition Letters*, **27** (13): 1419-1430.
Available at: http://www.dl.acm.org/citation.cfm?id=1161717

Gordon AD (1999) *Classification: Monographs on Statistics and Applied Probability*, 2nd ed. Chapman & Hall/CRC, London, UK, pp1-272.
Available at: http://www.amazon.com/Classification-Edition-Monographs-Statistics-Probability/dp/1584880139

Halkidi M; Batistakls Y; and **Vazirgiannis M** (2002) Clustering Validity Checking Methods: Part ii. *ACM SIGMOD Record,* **31** (3): 19-27.
Available at: http://www.dl.acm.org/citation.cfm?doid=601858.601862

Liang JY; Zhao XW; Li DY; Cao FY; and **Dang CY** (2012) Determining the Number of Clusters using Information Entropy For Mixed Data, *Pattern Recognition*, **45** (6): 2251–2265.
Available at: http://www.sciencedirect.com/science/article/pii/S0031320311005188

Lingras PC (2004) Interval Set Clustering of Web-users with Rough K-Means. *Journal of Intelligent Information Systems,* **23** (1): 5–16.
Available at: http://www.link.springer.com/article/10.1023/B:JIIS.0000029668.88665.1a#page-1

Maji P; Pal SK (2007) Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices, *IEEE Transactions on Systems, Man and Cybernetics Part B Cybernetics,* **37** (6): 1529-1540.
Available at: http://www.ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4359281

Mirkin B (1996) *Mathematical Classification and Clustering*. Kluwer Academic Publishers, Boston, London, UK, pp1-423.
Available at: http://www.download.springer.com/static/pdf/133/bfm%253A978-1-4613-0457-9%252F1.pdf?auth66=1404932047_2f022989574e7c556d80e416206e2ea6&ext=.pdf

Peters G; Crespo F; and **Weber R** (2013) Soft Clustering–fuzzy and Rough Approaches and their Extensions and Derivatives. *International Journal of Approximate Reasoning*, **54** (2): 307-322.
Available at: http://www.captura.uchile.cl/bitstream/handle/2250/133496/Soft%20clustering%20-%20fuzzy.pdf?sequence=1