Kernel Based Rough-Fuzzy C-Means

Rohan Bhargava and Balakrushna Tripathy

School of Computing Science and Engineering, VIT University
Vellore 632014, Tamil Nadu, India
{rb.bhargava,tripathybk}@gmail.com

Abstract. Data clustering has found its usefulness in various fields. Algorithms are mostly developed using euclidean distance. But it has several drawbacks which maybe rectified by using kernel distance formula. In this paper, we propose a kernel based rough-fuzzy C-Means (KRFCM) algorithm and use modified version of the performance indexes (DB and D) obtained by replacing the distance function with kernel function. We provide a comparative analysis of RFCM with KRFCM by computing their DB and D index values. The analysis is based upon both numerical as well as image datasets. The results establish that the proposed algotihtm outperforms the existing one.

Keywords: Clustering, Kernel, DB Index, Dunn Index, Rough-Fuzzy C-Means.

1 Introduction

Cluster is a collection of data elements that are similar to each other but dissimilar to elements in other clusters. Cluster analysis is a key tool in the field of data analysis. Clustering techniques have their use in areas like analysis of statistical data, pattern recognition, image analysis, information retrieval, bioinformatics and data mining. Clustering algorithms partition data into a certain number of groups or so called clusters. There is no set of predefined rules to determine the correctness of clustering. Hence many variations can be made to any single algorithm to develop a new algorithm. An iterative technique of partitioning a dataset into K-clusters was introduced by MacQueen in 1967 [1]. Applying this concepts of fuzzy sets Ruspini [2] first proposed the fuzzy clustering algorithm, which was later modified and generalized by Dunn [3] and Bezdek [4] respectively. Similarly using the concept of rough sets P. Lingras proposed the rough k-means clustering algorithm[5]. Further developments led to the proposal of rough set based kernel k-means algorithm by Zhou et al. [6] and Tripathy et al. [7]-[8].

Distance between objects can be calculated in many ways, the euclidean distance based clustering is easy to implement and hence most commonly used. It has two drawbacks, firstly the final results are dependent on the initial centers and secondly it can only find linearly separable cluster. Kernel based clustering helps in rectifying the second problem as it produces nonlinear separating hyper surfaces among clusters [9] . Kernel functions are used to transform the data in

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 148–155, 2013.

the image plane into a feature plane of higher dimension known as kernel space. Nonlinear mapping functions used transforms the nonlinear seperation problem in the image plane into a linear serperation problem in kernel space facilitating clustering in feature space. Mercer's theorem [10] can be used to calculate the distance between the pixel feature values in kernel space without knowing the transformation function.

It was pointed out by Dubois and Prade [11] that rough and fuzzy sets complement each other. In fact the hybrid model of rough fuzzy and fuzzy rough sets provide a better model for representing imperfect data. In fuzzy set theory we have definite formulae for the computation of membership values. Thus the hybrid algorithms takes care of both features by providing membership values to elements as well as modeling vagueness in data through the boundary concept. The concepts of lower and upper approximations in rough set deals with uncertainity and vagueness, whereas the concept of membership function in fuzzy set helps in enhancing and evaluating overlapping clusters.

In this paper we implement and further modify the Rough-Fuzzy C-Means given by Maji et al. [12] to propose a new hybrid kernel based algorithm. We show the comparison between the two using numeric datasets and image datasets. The paper contains 5 sections. Section 2 provides the basic information about the euclidean and kernel distance functions. Section 3 gives a detailed explanation on the proposed kernel based Rough-Fuzzy C-Means algorithm. Section 4 is where the evaluation results are discussed. Finally the paper is concluded in section 5.

2 Types of Distance Functions

Euclidean Distance. The euclidean distance $d(x, y)$ between any two objects x and y in any n-dimensional plane can be found using

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}. \tag{1}$$

where, $x_1, x_2 \ldots x_n$ and $y_1, y_2 \ldots y_n$ are attributes of x and y respectively.

Kernel Distance. If x is an object then $\phi(x)$ is the transformation of x in high dimensional feature space where the inner product space is defined by $K(x, y) = \langle \phi(x), \phi(y) \rangle$. In this paper we use the Gaussian kernel function.

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right). \tag{2}$$

Where, $\sigma^2 = \sum_{k=1}^{N} \|x_k - \overline{x}\|^2 / N$ with $\overline{x} = \sum_{k=1}^{N} x_k / N$.

N is total number of data objects [9]. According to Phillips et al. [13] kernel distance function $D(x, y)$ in the generalized form is $D(x, y) = K(x, x) + K(y, y) - 2K(x, y)$ and on applying the property of similarity (i.e., $K(x, x) = 1$) it can be further reduced to (3).

$$D(x, y) = 2(1 - K(x, y)). \tag{3}$$

3 Kernel Based Rough-Fuzzy C-Means (K-RFCM)

Rough Fuzzy C-Means was proposed by P. Maji et al.[12] and S. Mitra et al. [14]; it combines the concepts of rough set theory and fuzzy set theory. The concepts of lower and upper approximations in rough set deals with uncertainty, vagueness and incompleteness whereas the concept of membership function in fuzzy set helps in enhancing and evaluating overlapping clusters. We follow the same concept and replace all euclidean distance functions with kernel distance function given in (3). According to rough set theory if $x_j \in \underline{B}U_i$ then object x_j is contained completely in cluster U_i and if $x_j \in BN(U_i)$ then object x_j belongs to cluster U_i and also belongs to the boundary of another cluster. Hence, according to fuzzy set theory the objects in boundary of clusters will have different membership values for the concerned clusters. Hence, membership values of objects in lower approximation are $\mu_{ij} = 1$ while for those in boundary region are the same as that in FCM. The steps followed in this algorithm are given below

1. Assign initial means v_i for c clusters.
2. Compute μ_{ik} using

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{D_{ik}}{D_{jk}})^{\frac{2}{m-1}}} \tag{4}$$

3. Let μ_{ik} and μ_{jk} be the maximum and next to maximum membership values of object x_k to cluster centroids v_i and v_j.
   If $\mu_{ik} - \mu_{jk} < \delta$ then
   $x_k \in \overline{B}U_i$ and $x_k \in \overline{B}U_j$ and x_k cannot be a member of any lower approximation.
   Else $x_k \in \underline{B}U_i$.
   where, delta is given by

$$\delta = \frac{1}{N} \sum_{i=1}^{N} (\mu_{ik} - \mu_{jk}) \tag{5}$$

4. Calculate new cluster means by using

$$V_i = \begin{cases} w_{low} \times A + w_{up} \times B & if\ |\underline{B}U_i| \neq \phi \text{ and } |BN(U_i)| \neq \phi; \\ B & if\ |\underline{B}U_i| = \phi \text{ and } |BN(U_i)| \neq \phi; \\ A & ELSE \end{cases} \tag{6}$$

$$\text{where, } A = \frac{\sum_{x_k \in \underline{B}U_i} x_k}{|\underline{B}U_i|} \text{ and } B = \frac{\sum_{x_k \in \overline{B}U_i - \underline{B}U_i} \mu_{ik}^m x_k}{\sum_{x_k \in \overline{B}U_i - \underline{B}U_i} \mu_{ik}^m}$$

5. Repeat from step 2 until termination condition is met of until there is no more assignment of objects

DB and Dunn Index. The Davis-Bouldin (DB) and Dunn (D) indexes [15] are two of the basic performance indexes. They help in evaluating the efficiency of clustering. The results are depend upon the number of clusters required. The DB index is defined as the ratio of sum of within-cluster distance to between-cluster distance. It is formulated as

$$DB = \frac{1}{c}\sum_{i=1}^{c} max_{i\neq j}\left\{\frac{S(U_i)+S(U_j)}{d(U_i,U_j)}\right\} \qquad \text{for } 1<i,j<c \tag{7}$$

A good clustering procedure should give value of DB index as low as possible. Similar to the DB index the D index is used for the identification of clusters that are compact and separated. It is computed as

$$Dunn = min_i\left\{min_{j\neq i}\left\{\frac{d(U_i,U_j)}{max_l S(U_l)}\right\}\right\} \qquad \text{for } 1<i,j,l<c \tag{8}$$

Greater value for the D index proves to be more efficient The within cluster distance denoted be $S(U_i)$ is given as

$$S(U_i) = \begin{cases} w_{low}\times C + w_{up}\times D & if\ |\underline{B}U_i| \neq \phi \text{ and } |BN(U_i)| \neq \phi; \\ D & if\ |\underline{B}U_i| = \phi \text{ and } |BN(U_i)| \neq \phi; \\ C & ELSE \end{cases} \tag{9}$$

$$\text{where, } C = \frac{\sum_{x_k\in\underline{B}U_i} {D_{ik}}^2}{|\underline{B}U_i|} \text{ and } D = \frac{\sum_{x_k\in\overline{B}U_i-\underline{B}U_i} \mu_{ik}^m {D_{ik}}^2}{\sum_{x_k\in\overline{B}U_i-\underline{B}U_i} \mu_{ik}^m}$$

4 Evaluation

The evaluation has been done in 2 parts. Firstly using a few real datasets and then on image datasets. We compare the results from RFCM algorithm and the proposed K-RFCM algorithm. Each datset is compared on the basis of DB and Dunn index.

4.1 Numerical Dataset: Iris and Soybean

Table 1 shows the cluster centers that are formed after applying each algorithm on the iris dataset consisting of 50 elements having 4 attributes each. Evaluation has been performed for c = 2,3 and 4. The initial centers are taken to be the first c elements of the dataset. We see that the results of both algorithms have very minute difference. Further analysis has been done for $c = 2$ in Table 2, it shows which all elements lie in the lower and boundary regions of each cluster. Again esults are approximately the same except for some values that have been highlighted in the table. These values are responsible for the minute difference observed in Table1.

Table 3 and 4 provides a comparison of the methods RFCM and KRFCM based upon the computations of DB and D index on the iris data set and the

Table 1. Cluster Center Values on Iris Dataset ($c = 2$,3 & 4)

No. of Clusters		RFCM	K-RFCM
2	Center 1	5.1639; 4.3754; 3.3974; 2.6100	5.1531; 4.3497; 3.3873; 2.6047
	Center 2	4.8498; 3.9619; 3.1373; 2.9359	4.8373; 3.9638; 3.1324; 2.3930
3	Center 1	5.1571; 4.3643; 3.3929; 2.6036	5.1308; 4.3538; 3.3897; 2.6019
	Center 2	4.8927; 3.9783; 3.1430; 2.3980	4.8935; 3.9788; 3.1431; 2.3981
	Center 3	4.5961; 3.9278; 3.0708; 2.3600	4.5984; 3.9296; 3.0719; 2.3608
4	Center 1	5.1571; 4.3643; 3.3929; 2.6036	5.1308; 4.3538; 3.3897; 2.6019
	Center 2	4.9150; 3.9850; 3.1400; 2.4059	4.9146; 3.9990; 3.1573; 2.4089
	Center 3	4.7134; 3.9620; 3.0776; 2.3589	4.7157; 3.9645; 3.0791; 2.3602
	Center 4	4.6128; 3.8615; 3.0740; 2.3561	4.6528; 3.9027; 3.1164; 2.3880

Table 2. Lower and Boundary Elements for Iris Dataset ($c = 2$)

Cluster Center	Lower		Boundary	
	RFCM	K-RFCM	RFCM	K-RFCM
Center 1	0, 4, 7, 10, **16**, 17, 19, 21, 27, 28, 36, 39, 40, 46, 48	0, 4, 7, 10, 17, 19, 21, **26**, 27, 28, **31**, 36, 39, 40, 46, 48	**2**, 5, 6, 8, 11, 13, 14, 15, 18, 20, 22, 23, 24, **26**, 29, **31**, 32, 33, 35, 38, 41, 42, 43, 44, 47, 49	5, 6, 8, 11, 13, 14, 15, **16**, 18, 20, 22, 23, 24, 29, 32, 33, 35, 38, 41, 42, 43, 44, 47, 49
Center 2	1, 3, 9, 12, 25, 30, 34, 37, 45	1, **2**, 3, 9, 12, 25, 30, 34, 37, 45	**2**, 5, 6, 8, 11, 13, 14, 15, 18, 20, 22, 23, 24, **26**, 29, **31**, 32, 33, 35, 38, 41, 42, 43, 44, 47, 49	5, 6, 8, 11, 13, 14, 15, **16**, 18, 20, 22, 23, 24, 29, 32, 33, 35, 38, 41, 42, 43, 44, 47, 49

Table 3. DB and Dunn Indexes for Iris Dataset

No. of Clusters	DB Index		Dunn Index	
	RFCM	K-RFCM	RFCM	K-RFCM
2	115.8219	15.6873	0.0159	0.1275
3	157.5086	20.7628	0.0093	0.0743
4	309.0443	84.4577	0.0042	0.01365

soybean dataset. In Table 3 it is clear that the values for DB index in K-RFCM are far lower than those of RFCM and the values for Dunn index are larger in the former algorithm. Hence stating that while results are similar the performance of K-RCCM is better than RFCM. Looking over at Table 4 the soybean data set consists of 37 elements having 35 attributes, the DB index values are lower for K-RFCM and Dunn index values are also low for the same. Though larger values for Dunn index were expected but we predict the low values are due to the large number of attributes involved in the dataset.

Table 4. DB and Dunn Indexes for Soybean Dataset

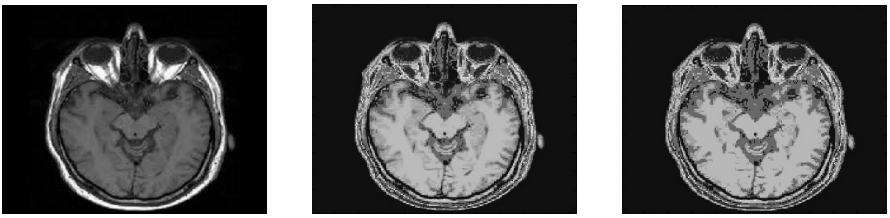
No. of Clusters	DB Index		Dunn Index	
	RFCM	K-RFCM	RFCM	K-RFCM
2	624.7929	119.7455	0.0029	0.0162
3	1074.2871	116.8911	9.4638	0.0096
4	1737.3794	192.9677	6.4533	0.0059

Table 5. DB and Dunn Indexes for Image datasets

No. of Clusters	DB Index		Dunn Index	
	RFCM	K-RFCM	RFCM	K-RFCM
Brain	4.0469	0.0594	0.1060	10.3673
Cell	0.0561	negligible	11.4630	very large
Iris	4.7198	0.1643	0.1643	5.4695
Penny	9.2726	0.2369	0.0916	4.9417

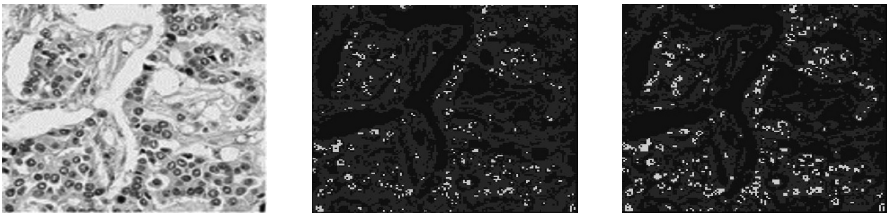
4.2 Image Dataset

We have processed a number of images using both the algorithms to obtain a resultant image and as well as the DB and Dunn index values. Table 5 gives us the overview of the DB and Dunn index values obtained for each image. In all cases we observe that we achieve our desired results. There is a significant drop in DB index and significant increase in Dunn index. Comparing each image indivdually. Fig. 1c is sharp and has more clarity in outlines and finite details than that of Fig 1b. The difference between Fig 2b. and Fig 2c. is highly evident. Cell can be correctly identified in Fig. 2c. Finally, there is no noticeable change seen in Fig 3b. and Fig 3c.



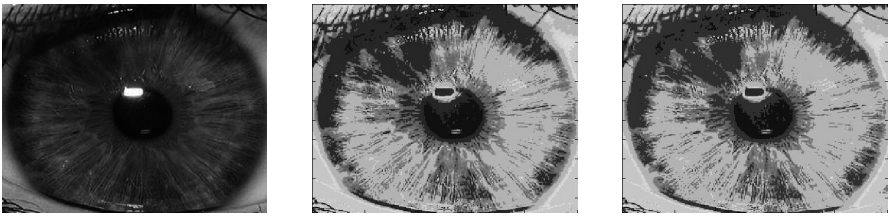
1a : Original 1b : RFCM Version 1c : K-RFCM Version

Fig. 1. Brain Image



2a : Original 2b : RFCM Version 2c : K-RFCM Version

Fig. 2. Cell Image



3a : Original 3b : RFCM Version 3c : K-RFCM Version

Fig. 3. Iris Image

5 Conclusion

This paper focuses on improving the performance of the existing RFCM algorithm by using kernel function instead of euclidean distance. Hence, developing a new hybrid kernel based algorithm. Also, two of the most widely used performance indexes have been modified using kernel distance function for the evaluation of kernel based algorithms. Comparison between RFCM and proposed K-RFCM has been done on a wide variety of datasets to obtain favourable results. From the obtained results we can conclude that the proposed algorithm clearly outperforms the existing algorithm on the basis of performance and yields equivalent or better outputs in image dataset. The DB and D index introduced in this paper can also be applied on kernel based algorithms using rough sets and fuzzy sets individually to compare their performances.

References

1. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press (1967)
2. Ruspini, E.H.: A new approach to clustering. Information and Control 15(1), 22–32 (1969)
3. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, 32–57 (1973)
4. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers (1981)
5. Lingras, P., West, J.: Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems 23(1), 5–16 (2004)
6. Zhou, T., Zhang, Y., Lu, H., Deng, F., Wang, F.: Rough Cluster Algorithm Based on Kernel Function. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 172–179. Springer, Heidelberg (2008)
7. Tripathy, B.K., Ghosh, A., Panda, G.K.: Kernel based K-means clustering using rough set. In: IEEE 2012 International Conference on Computer Communication and Informatics, ICCCI (2012)
8. Tripathy, B.K., Ghosh, A., Panda, G.K.: Adaptive K-Means Clustering to Handle Heterogeneous Data Using Basic Rough Set Theory. In: Meghanathan, N., Chaki, N., Nagamalai, D. (eds.) CCSIT 2012, Part I. LNICST, vol. 84, pp. 193–202. Springer, Heidelberg (2012)
9. Yang, M.S., Tsai, H.S.: A Gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction. Pattern Recognition Letters 29(12), 1713–1725 (2008)
10. Zhang, D.Q., Chen, S.C.: Kernel Based Fuzzy and Possibilistic C-means Clustering. In: Proc. the International Conference Artificial Neural Network, Turkey, pp. 122–125 (2003)
11. Dubois, D., Prade, H.: Rough Fuzzy sets model. International Journal of General Systems 46(1), 191–208 (1990)
12. Maji, P., Pal, S.K.: RFCM: A Hybrid Clustering Algorithm using rough and fuzzy sets. Fundamenta Informaticae 80(4), 475–496 (2007)
13. Phillips, J.M., Venkatasubramanian, S.: A gentle introduction to the kernel distance, arXiv preprint arXiv:1103.1625 (2011)
14. Mitra, S., Banka, H., Pedrycz, W.: Rough-Fuzzy Collaborative Clustering. IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics 36(4), 795–805 (2006)
15. Bezdek, J.C., Pal, N.R.: Some new indexes for cluster validity. IEEE Transaction on System, Man and Cybernetics, Part B: Cybernetics 28, 301–315 (1998)