Incorporating Fuzziness to CLARANS

Sampreeti Ghosh and Sushmita Mitra

Center for Soft Computing, Indian Statistical Institute, Kolkata 700108, India
{sampreeti_t,sushmita}@isical.ac.in

Abstract. In this paper we propose a way of handling fuzziness while mining large data. Clustering Large Applications based on RANdomized Search (CLARANS) is enhanced to incorporate the fuzzy component. A new scalable approximation to the maximum number of neighbours, explored at a node, is developed. The goodness of the generated clusters is evaluated in terms of validity indices. Experimental results on various data sets is run to converge to the optimal number of partitions.

Keywords: Data mining, CLARANS, medoid, fuzzy sets, clustering.

1 Introduction

Clustering LARge Applications (CLARA) [1] incorporates sampling in the framework of PAM [1] to make it scalable for handling large data. Again, the results here remain dependent on the quality of the sampling process undertaken. An efficient variation of CLARA is Clustering Large Applications based on RANdomized Search (CLARANS) [2], which partitions large data. It is found to outperform both PAM and CLARA [3] in terms of accuracy and computational complexity, and can handle outliers. However, since all the three clustering algorithms are designed to generate crisp clusters, they fare poorly when modeling overlapping clusters.

Fuzzy sets [4] constitute the oldest and most reported soft computing paradigm [5]. These provide soft decision by taking into account characteristics like tractability, robustness, low cost, etc., and have close resemblance to human decision making. They are well-suited to modeling different forms of uncertainties and ambiguities, often encountered in real life. The concept of fuzzy membership μ, lying in $[0, 1]$, allows the simultaneous finite belongingness of a pattern to two or more overlapping clusters.

In this article we propose a novel fuzzy clustering algorithm, Fuzzy CLARANS (FCLARANS), that performs efficiently on large data. It incorporates the concept of fuzzy membership onto the framework of CLARANS for manoeuvering uncertainty in the context of data mining. Cluster validity indices, like Davies-Bouldin (DB) [6] and Xie-Beni (XB) [7], are used to evaluate the goodness of the generated partitions. Note that, unlike DB, the index XB incorporates fuzzy membership in its computations. The clustering algorithms compared here are hard c -means (HCM), fuzzy c - means (FCM), fuzzy c - medoid (FCMd), and CLARANS.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 122–127, 2009.

The performance on various data sets demonstrates that the proposed Fuzzy CLARANS always converges to the lowest value for both the indices DB and XB at an optimal number of clusters. The cost function of CLARANS is fuzzified using membership values. A new scalable approximation is developed to compute the maximum number of neighbours being explored at each node. It also helps to eliminate user-defined parameters in the expression.

The rest of the paper is organized as follows. Section 2 describes the preliminaries, like algorithms CLARANS and the clustering validity indices DB and XB. Then we move to the proposed Fuzzy CLARANS in Section 3. The experimental results are presented in Section 4. Finally, Section 5 concludes the article.

2 Preliminaries

In this section we describe some of the basic concepts like algorithms CLARANS [2],and clustering validity indices.

2.1 CLARANS

Large datasets require the application of scalable algorithms. CLARANS [2] draws a sample of the large data, with some randomness, at each stage of the search. Each cluster is represented by its medoid. Multiple scans of the database are required by the algorithm. Here the clustering process searches through a graph G, where node v^q is represented by a set of c medoids (or centroids) $\{m_1^q, \ldots, m_c^q\}$. Two nodes are termed as neighbors if they differ by only one medoid, and are connected by an edge. More formally, two nodes $v^1 = \{m_1^1, \ldots, m_c^1\}$ and $v^2 = \{m_1^2, \ldots, m_c^2\}$ are termed neighbors if and only if the cardinality of the intersection of v^1 and v^2 is given as $card(v^1 \bigcap v^2) = c - 1$. Hence each node in the graph has $c * (N - c)$ neighbors. For each node v^q we assign a cost function

$$J_c^q = \sum_{x_j \varepsilon U_i} \sum_{i=1}^{c} d_{ji}^q, \tag{1}$$

where d_{ji}^q denotes the dissimilarity measure of the jth object x_j from the ith cluster medoid m_i^q in the qth node. The aim is to determine that set of c-medoids $\{m_1^0, \ldots, m_c^0\}$ at node v^0, for which the corresponding cost is the minimum as compared to all other nodes in the tree.

Note that *the maximum number of neighbors* is computed as

$$neigh = p\% \text{ of } \{c * (N - c)\}, \tag{2}$$

with p being provided as input by the user. Typically, $1.25 \leq p \leq 1.5$ [2].

2.2 Validity Indices

There exist validity indices [8] to evaluate the goodness of clustering, corresponding to a given value of c. Two of the commonly used measures include the Davies-Bouldin (DB) [6] and the Xie-Beni (XB) [7] indices.

The DB index is a function of the ratio of sum of within-cluster distance to between-cluster separation. It is expressed as

$$DB = \frac{1}{c}\sum_{i=1}^{c} max_{k\neq i}\frac{diam(U_i)+diam(U_j)}{d'(U_i,U_j)}, \tag{3}$$

where the diameter of cluster U_i is $diam(U_i) = \frac{1}{|U|_i}\sum_{x_j\in U_i}||x_j - m_i||^2$. The inter-cluster distance between cluster pair U_i, U_j is expressed as $d'(U_i,U_j) = ||m_i - m_j||^2$. DB is minimized when searching for the optimal number of clusters c_0.

The XB index is defined as

$$XB = \frac{\sum_{j=1}^{N}\sum_{i=1}^{c}\mu_{ij}^{m'}d_{ji}}{N * \min_{i,j} d'(U_i,U_j)^2}, \tag{4}$$

where μ_{ij} is the membership of pattern x_j to cluster U_i. Minimization of XB is indicative of better clustering, particularly in case of fuzzy data. Note that for crisp clustering the membership component μ_{ij} boils down to zero or one.

3 Fuzzy CLARANS

In this section we describe the proposed algorithm Fuzzy CLARANS (FCLARANS). Here fuzzy membership is incorporated in the framework of CLARANS. This enables appropriate modeling of ambiguity among overlapping clusters. A pattern is allowed finite, non-zero membership $\mu_{ij} \in [0,1]$ to two or more partitions. The distance component is weighted by the corresponding membership value, analogously to FCM and FCMd.

The hybridization allows the modeling of uncertainty in the domain of large data. Although the computational complexity is higher than that of CLARANS, yet the performance is found to be superior for the optimal partitioning, as evaluated in terms of clustering validity indices. It is interesting to observe that fuzzy clustering in FCLARANS boils down to the crisp version in CLARANS, when $\mu_{ij} \in \{0,1\}$.

The cost of a node, as defined in eqn. (1), is now modified to

$$J_{fc}^{q} = \sum_{j=1}^{N}\sum_{i=1}^{c}(\mu_{ij}^{q})^{m'}d_{ji}^{q}. \tag{5}$$

We chose $m' = 2$ [9] after several experiments. The membership at node v^q is computed as $\mu_{ij}^{q} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ji}^{q}}{d_{jk}^{q}}\right)^{\frac{2}{m'-1}}}$.

We observed that the value of $neigh = p\%$ of $\{c*(N-c)\}$ [as in CLARANS, eqn. (2)] turns up to be very high for large data having $N \geq 10,000$. This increases the computational burden. We, therefore, propose a new scalable approximation expressed as

$$neigh = c^2\log_2(N-c). \tag{6}$$

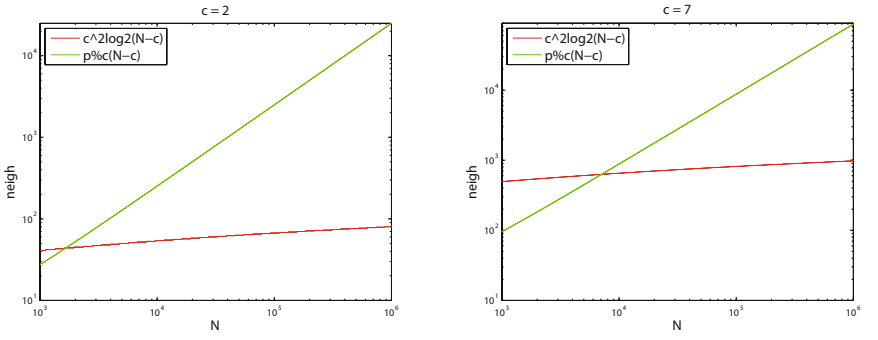

Fig. 1. Variation of *neigh* (by the two expressions) with N, intersecting at N_{cross}, for cluster numbers (a) $c = 2$ and (b) $c = 7$

A study of its behaviour is provided in Fig. 1, with respect to the previous expression [eqn. (2)]. We are, as a result, able to eliminate the user-defined parameter p while also reducing computational time of the algorithm. The new eqn. (6) has been employed in this article for experiments involving large datasets ($N \geq 10,000$). Note Figs. 1(a)-(b),the expressions for *neigh* [by eqns. (2) and (6)] intersect each other.

4 Experimental Results

The performance of the algorithm Fuzzy CLARANS was tested on various data sets. The goodness of the partitions was evaluated in terms of cluster validity indices DB and XB. Comparative study was made with related clustering algorithms like HCM, FCM, FCMd, and CLARANS.

We used a synthetic dataset (*Set1*), and three real datasets *Magic gamma*, *Shuttle* and *Forest Cover*. Average results were reported over five runs.

4.1 Set 1

The data contains three clusters, each with 100 randomly generated patterns. The two-dimensional scatter plot of Fig. 2 depicts the patterns lying within circles of unit radius, each having different centers. A lot of overlapping is artificially introduced.

Table 1 establishes that DB and XB are minimum for Fuzzy CLARANS for the correct number of three partitions. Although FCM generates the globally least value for XB, yet the partitioning is incorrect at $c = 5$. On the other hand, FCMd also provides best result in terms of both the indices. However, algorithm FCLARANS is able to model the overlapping partitions in a better manner.

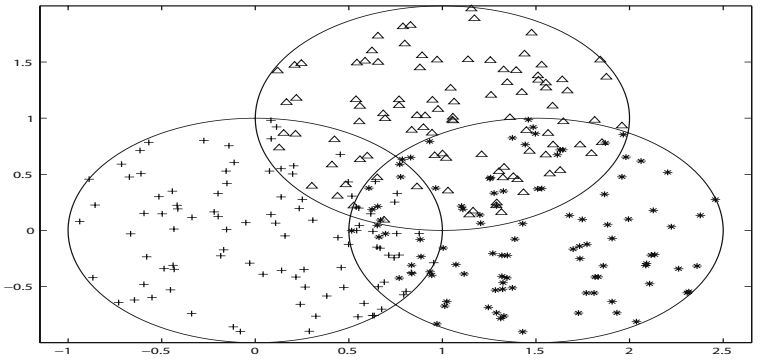

Fig. 2. Data *Set1*

Table 1. Average comparative performance on synthetic data

Algorithm	c=2, *neigh*=12	c=3, *neigh*=18	c=4, *neigh*=24	c=5, *neigh*=30
	DB, XB	DB, XB	DB, XB	DB, XB
HCM	0.70, 0.35	0.51, 0.28	0.49, 0.28	**0.45, 0.26**
FCM	0.76, 0.27	0.54, 0.20	0.56, 0.18	**0.48, 0.14**
FCMd	1.36, 0.40	**1.24, 0.31**	1.60, 0.32	1.64, 0.37
CLARANS	0.68, **0.34**	0.75, 0.42	0.83, 0.55	**0.55**, 0.42
FCLARANS	0.72, 0.26	**0.72, 0.23**	0.85, 0.28	0.76, 0.31

Table 2. Performance of FCLARANS on large datasets

Clusters	2	3	4	5	6	7	8
Dataset	DB,XB (*neigh*)	DB,XB (*neigh*)	DB,XB (*neigh*)	DB,XB (*neigh*)	DB,XB (*neigh*)	DB,XB (*neigh*)	DB,XB (*neigh*)
Magic gamma	**0.40, 1.44** (57)	0.70, 2.17 (128)	0.99, 3.14 (228)	1.10, 2.45 (356)	0.82, 1.69 (512)	0.81, 1.73 (697)	0.83, 2.04 (910)
Shuttle	1.61, 4.23 (64)	1.11, 2.71 (143)	1.84, 4.31 (254)	2.60, 4.97 (396)	2.51, 4.41 (570)	**1.10, 2.63** (776)	2.34, 4.11 (1013)
Forest cover	0.05, 1.73 (77)	0.07, 2.07 (173)	0.10, 2.34 (307)	0.15, 3.31 (479)	0.09, 1.60 (690)	0.08,**1.43** (939)	0.14, 2.54 (1226)

4.2 Large Data

The large data were taken from the *UCI Machine Learning Repository.*The *magic gamma* telescope data 2004 is made up of 19,020 instances, with ten features and two classes (gamma signal and hadron background). The *Shuttle* data (statlog version) consists of 58,000 measurements corresponding to seven classes, *viz.*, Rad flow, Fpv close, Fpv open, High, Bypass, Bpv close, Bpv open. There are nine numerical attributes. The *Forest cover* data consists of 5,81,012 instances. There are 10 numeric-valued attributes, with seven kinds of forest cover

corresponding to spruce/fir, lodgepole pine, ponderosa pine, cottonwood/willow, aspen, douglas-fir and krummholz.

Table 2 provides the clustering results with FCLARANS on these three datasets. We observe that the minimum values for DB and XB always correspond to the correct number of clusters.

5 Conclusions

We have developed a new algorithm Fuzzy CLARANS, by incorporating fuzziness in CLARANS while clustering of large data. The cost function is weighted by fuzzy membership. A scalable approximation to the maximum number of neighbours, explored at a node, has been designed. It helps in reducing the computational time for large data, while eliminating the need for user-defined parameters. The goodness of the generated clusters has been evaluated in terms of the Davies Bouldin and Xie Beni validity indices. Results demonstrate the superiority of Fuzzy CLARANS in modeling overlaps, particularly in large data. The algorithm is found to converge to the optimal number of partitions.

Acknowledgements

The authors gratefully acknowledge Mr. Shivnath Shaw for his programming support and helpful discussions.

References

1. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York (1990)
2. Ng, R.T., Han, J.: Clarans: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering 14, 1003–1016 (2002)
3. Mitra, S., Acharya, T.: Data Mining: Multimedia, Soft Computing, and Bioinformatics. John Wiley, New York (2003)
4. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
5. Fuzzy logic, neural networks, and soft computing. Communications of the ACM 37, 77–84 (1994)
6. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227 (1979)
7. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 841–847 (1991)
8. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, NJ (1988)
9. Yu, J.: General c-means clustering model. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1197–1211 (2005)