Applied Soft Computing 13 (2013) 1639–1645

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Clustering large data with uncertainty

Sampreeti Ghosh*,[1], Sushmita Mitra

Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India

ARTICLE INFO

Article history:
Received 8 November 2011
Received in revised form
10 December 2012
Accepted 25 December 2012
Available online 23 January 2013

Keywords:
Data mining
CLARANS
Medoid
Fuzzy sets
Clustering

ABSTRACT

A new algorithm is designed for handling fuzziness while mining large data. A new novel cost function weighted by fuzzy membership, is proposed in the framework of CLARANS. A new scalable approximation to the maximum number of neighbors, explored at each node, is developed; thus reducing the computational time for large data while eliminating the need for user-defined (heuristic) parameters in the existing equation. The goodness of the generated clusters is evaluated in terms of Xie–Beni validity index. Results demonstrate the superiority of the proposed algorithm, over both synthetic and real data sets, in terms of goodness of clustering. It is interesting to note that our algorithm always converges to the globally best values at the optimal number of partitions. Moreover compared to existing fuzzy algorithms, FCLARANS without scanning the whole dataset, searching small number of neighbors, is able to handle the uncertainty due to overlapping nature of the various partitions. This is the main motivation of fuzzification of the algorithm CLARANS.

1. Introduction

A cluster is a collection of data objects which are similar to one another within the same cluster but dissimilar to objects in other clusters [15]. Clustering analysis is a technique for finding natural groups present in the data. In other words, the problem is to partition N patterns (*i.e.*, objects) into c desired clusters with high intra-class similarity and low inter-class similarity by optimizing an objective function. Validity indices [6] are often employed to evaluate the goodness of the generated partitions.

One of the most popular clustering algorithms is c-means, which is also known as hard c-means (HCM) [15]. Its merits lie in fast convergence and small storage requirement. In HCM each cluster is represented by its center of gravity or mean. This need not essentially correspond to an object of the given datasets. Hence HCM is unsuitable for handling non-numeric data. It is also sensitive to the presence of outliers. The objective function minimized is the squared error E of each pattern x_j from the mean (or centroid) m_i of cluster U_i, and is expressed as $E=\sum_{i=1}^{c}\sum_{j=1,x_j\in U_i}^{N}(x_j-m_i)^2$. Other variants include [2,4,10].

Partitioning around medoids (PAM) [8] overcomes some of these difficulties by generating the c most representative objects (or medoids) as the centroids, while minimizing the sum of within-cluster dissimilarity. Medoids are typically more robust to outliers, as compared to means, such that peripheral patterns do not affect them. This scheme is also convenient for representing non-numeric patterns, with particular emphasis to mixed media data. However, due to the high computational complexity, the algorithm is found to be unsuitable for handling large data. Today data is no longer restricted to a few hundred tuples of numeric or character representations only. Therefore the traditional *ad hoc* mixtures of statistical techniques and data management tools are no longer adequate for analyzing this vast collection of mixed data.

Large data can be analyzed using Clustering LARge Applications (CLARA) [8]. Here the concept of sampling is introduced. Though CLARA is robust like PAM its results remain dependent on the quality of the sampling process undertaken. Clustering Large Applications based on RANdomized Search (CLARANS) [13] is more efficient than PAM and CLARA in terms of accuracy, computational complexity and handling of outliers [12]. All these algorithms implicitly assume the existence of disjoint clusters in data, and therefore fare poorly in the presence of overlaps. In practice, the separation of clusters is a fuzzy notion [7].

Fuzzy set [17] is one of the earliest and most widely reported constituent of soft computing [18]. It is able to handle, with greater flexibility, the uncertainties arising from incompleteness or imperfection in information. The uncertainty in clustering of patterns may arise from the overlapping nature of the various partitions. In conventional techniques, it is assumed that a pattern belongs to only one partition. In fuzzy sets, the concept of fuzzy membership μ, lying in [0, 1], allows a pattern to simultaneously belong in more than one partition.

* Corresponding author.
E-mail addresses: sampreeti_t@isical.ac.in (S. Ghosh), sushmita@isical.ac.in (S. Mitra).

[1] This work was carried out when S. Ghosh was attached to Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India.

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.12.036

Fuzzy clustering algorithms like the fuzzy *c*-means (FCM) [3] and fuzzy *c*-medoids (FCMd) [9] have been designed based on HCM and PAM. A robust extension to FCM, termed agglomerative FCM (AFCM) [10], introduces a penalty term to the objective function while determining a consistent clustering. Recently various soft computing methodologies have been applied to handle the different challenges posed by data mining [12], involving large heterogeneous data sets.

A new clustering algorithm Fuzzy CLARANS (FCLARANS) is proposed in this article. The concept of fuzzy membership (like FCM) is incorporated onto the framework of CLARANS for handling uncertainty in the context of mining large data. The goodness of clustering is evaluated using the Xie–Beni (*XB*) cluster validity index [16]. Algorithms like FCM, FCMd and AFCM are used for comparative study.

Fuzzy CLARANS is found to converge to optimal number of clusters, for the lowest value of *XB*. Membership value is incorporated for the fuzzification of the cost function. The maximum number of neighbors, explored at each node, is scalably approximated; thereby eliminating user-defined parameters. The performance over synthetic as well as real data sets, measured in terms of accuracy, cost and validity index, is found to be better in large data sets with less number of computations.

Analysis of variance (ANOVA) [1,5] has also been used for comparative study. The purpose is to test the differences in means (of groups or variables) for statistical significance. This is accomplished by analyzing the variance, that is, by partitioning the total variance into components that are due to true random error and that due to differences between means. The latter variance components are then tested for statistical significance. When significant, the null hypothesis of no differences between means is rejected and the alternative hypothesis that the means (in the population) are different from each other is accepted.

The organization of the rest of the paper is as follows. Section 2 describes the preliminaries, like algorithms FCM, FCMd, and the clustering validity index *XB*. Section 3 explains the incorporation of fuzziness and the scalable approximation in algorithm Fuzzy CLARANS. The experimental results are presented in Section 4, followed by the conclusion in Section 5.

2. Preliminaries

In this section we describe some of the basics concepts like algorithms FCM [3] and FCMd [9], along with a clustering validity index *XB* [16].

2.1. FCM

This is a fuzzification of the *c*-means algorithm [3]. It partitions a set of *N* patterns $\{x_j\}$ into *c* clusters by minimizing the objective function

$$J = \sum_{j=1}^{N}\sum_{i=1}^{c}(\mu_{ij})^{m'}||x_j - m_i||^2, \tag{1}$$

where $1 < m' < \infty$ is the fuzzifier, $\mu_{ij} \in [0, 1]$ is the membership of the *j*th pattern to the *i*th mean m_i, and $||\cdot||$ is the distance norm, such that

$$m_i = \frac{\sum_{j=1}^{N}(\mu_{ij})^{m'}x_j}{\sum_{j=1}^{N}(\mu_{ij})^{m'}} \tag{2}$$

and

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}(d_{ji}/d_{jk})^{2/(m'-1)}}, \tag{3}$$

$\forall i$, with d_{ji} denoting the distance between the *j*th pattern and the *i*th mean, subject to $\sum_{i=1}^{c}\mu_{ik} = 1, \forall k$, and $0 < \sum_{k=1}^{N}\mu_{ik} < N, \forall i$.

2.2. FCMd

The algorithm is a fuzzification of the *c*-medoids algorithm (PAM) and is outlined as follows [9]:

1. Pick the initial medoids m_i, $i = 1, \ldots, c$.
2. **Repeat Steps 3 and 4** until convergence.
3. Compute μ_{ik} for $i = 1, \ldots, c$ and $k = 1, \ldots, N$.
4. Compute new medoids

   $m_i = x_q$,

   where

   $$q = arg \min_{1\le j\le N}\sum_{k=1}^{N}(\mu_{ik})^{m'}||x_j - x_k||^2, \tag{4}$$

   refers to that *j* for which the minimum value of the expression is obtained.

   Note that this boils down to the hard *c*-medoids (PAM) with $\mu_{ik} = 1$, if $i = q$, and to $\mu_{ik} = 0$ otherwise.

2.3. Cluster validity index

The objective of fuzzy clustering is to partition a data set into *c* homogeneous fuzzy clusters. The algorithms typically require the user to pre-specify the number of clusters *c*. This choice can often lead to different clustering partitions. Sometimes the goodness of the clustering is evaluated in terms of a validity index. One such is the Xie–Beni (*XB*) [16] index. It is defined as

$$XB = \frac{\sum_{j=1}^{N}\sum_{i=1}^{c}\mu_{ij}^{m'}d_{ji}}{N * \min_{i,j}d'(U_i, U_j)^2}, \tag{5}$$

where μ_{ij} is the membership of pattern x_j to cluster U_i, d_{ji} denotes the dissimilarity measure of the *j*th object x_j from the *i*th cluster medoid and d' denotes the dissimilarity measure of the *i*th and *j*th cluster medoids. While the numerator indicates the compactness of the fuzzy partitions, the denominator determines the strength of separation between them. 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 clustering of large data

Large data can be suitably handled by algorithms such as CLARANS. However it is unable to handle uncertainty due to overlapping. Fuzzification of CLARANS and introduction of a new scalable approximation allows handling of overlaps, while reducing the computational time for searching neighbors and also eliminating the need for user-defined parameters. We incorporate the concept of fuzzy membership (like FCM) onto the framework of CLARANS for manoeuvering uncertainty in the context of data mining.

3.1. CLARANS

CLARANS [13] considers two parameters *numlocal*, representing the number of iterations (or runs) for the algorithm, and *maxneighbor*, the number of adjacent nodes (set of medoids) in the graph *G* that need to be searched up to convergence. These parameters are provided as input at the beginning. While CLARA compared very