Journal of Biomedical Informatics 62 (2016) 48–58

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Information bottleneck based incremental fuzzy clustering for large biomedical data

Yongli Liu *, Xing Wan

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, China

ARTICLE INFO

Article history:
Received 28 September 2015
Revised 24 April 2016
Accepted 30 May 2016
Available online 31 May 2016

Keywords:
Fuzzy clustering
Information bottleneck
Incremental clustering

ABSTRACT

Incremental fuzzy clustering combines advantages of fuzzy clustering and incremental clustering, and therefore is important in classifying large biomedical literature. Conventional algorithms, suffering from data sparsity and high-dimensionality, often fail to produce reasonable results and may even assign all the objects to a single cluster. In this paper, we propose two incremental algorithms based on information bottleneck, *Single-Pass* fuzzy c-means (spFCM-IB) and *Online* fuzzy c-means (oFCM-IB). These two algorithms modify conventional algorithms by considering different weights for each centroid and object and scoring mutual information loss to measure the distance between centroids and objects. spFCM-IB and oFCM-IB are used to group a collection of biomedical text abstracts from Medline database. Experimental results show that clustering performances of our approaches are better than such prominent counterparts as spFCM, spHFCM, oFCM and oHFCM, in terms of accuracy.

1. Introduction

In recent years, medical workers and patients are producing many forms of content such as blogs, wikis and discussion forums via social media websites every day. The amount of biomedical data therefore grows rapidly. In order to find more applicable ways for classifying biomedical literature so that users could find relevant articles easily from a huge amount of biomedical data, clustering, as a significant data mining technique, has been widely studied [1].

Given a set of biomedical articles, clustering tries to detect intrinsic structures so that a list of clusters is generated where inter-cluster similarity is maximized and intra-cluster similarity is minimized. When large biomedical data is clustered, it is a big challenge that all the data is not available at once or too large to be loaded into memory. To handle this challenge, incremental clustering framework has been adopted, which is effective to work with continuous data streams. In the incremental approaches, large data is randomly partitioned into many chunks and incremental clustering is performed with one chunk at a time.

There is already a large body of work that investigates approaches to clustering objects incrementally. Ning et al. [2] thought that the capability of incrementally updating is essential to some applications such as websphere or blogsphere, and extended the standard spectral clustering by incrementally updating the eigen-system. Li et al. [3] studied clustering algorithms on categorical data streams, and then proposed an integrated framework for incrementally clustering categorical data with three algorithms: Minimal Dissimilarity Data Labeling, Concept Drift Detection and Cluster Evolving Analysis. Liu and Ban [4] presented a new clustering algorithm called growing incremental self-organizing neural network (GISONN), which detects clusters by learning data distribution of each cluster. The GISONN method is able to work incrementally and detect arbitrary-shaped clusters without requiring the number of clusters as a prerequisite. In [5], we proposed an incremental method for document clustering based on information bottleneck theory. Peng and Liu [6] presented a novel down-top incremental conceptual hierarchical text clustering approach using CFu-tree (ICHTC-CF) representation, which starts with each item as a separate cluster. Keeping the benefits of conventional clustering, the above algorithms add the capability of incrementally updating, which enables them to update existing clusters incrementally, instead of recomputing all the clusters.

The clustering algorithms discussed previously are classified as hard clustering which puts each object into a single cluster. However, a biomedical article may actually involve multiple medical specialties and subjects, so it could belong to multiple clusters [7]. This suggests the appearance of soft clustering algorithms. Introducing fuzziness to clustering gives us the flexible solutions for soft clustering algorithms. Fuzzy clustering extends

* Corresponding author.
E-mail address: yongli.buaa@gmail.com (Y. Liu).

http://dx.doi.org/10.1016/j.jbi.2016.05.009
1532-0464/

conventional clustering via representing the affiliation of objects to clusters by memberships. Fuzzy *c* means (FCM) [8] is a representative fuzzy clustering algorithm. Based on this approach, there are many varieties [9,10]. Nowadays, many FCM-type clustering algorithms are designed as incremental approaches to support continuous data streams, as most Web datasets are known to be large and high dimensional. Hore et al. introduced two incremental approaches, *Single-Pass* FCM (spFCM) [11] and *Online* FCM (oFCM) [12], which are very classic and important. In spFCM, large data is processed chunk by chunk. The previous chunk is represented by its centroids, which will be regarded as virtual articles and integrated with the newly coming chunk for the next round of clustering. Different from spFCM, two related steps are included in oFCM. The first step classifies each chunk individually, and obtains a set of centroids, on which clustering will be performed in the second step. Although spFCM and oFCM are effective in many applications handling large-scale data, they still have some weaknesses when the dataset is sparse and high dimensional [13]. Mei et al. [13] worked on incremental fuzzy clustering for large document data by considering scalability as well as the ability to deal with sparsity and high-dimensionality, and changed spFCM and oFCM into spHFCM (*Single-Pass* hyperspherical fuzzy *c* means) and oHFCM (*Online* hyperspherical fuzzy *c* means) respectively by adding another step to normalize the centroids after they are updated in each iteration. The spHFCM and oHFCM normalize all the centroids to unit norm after each iteration, and adopt cosine similarity to measure the closeness or distance between a centroid and each object, instead of Euclidean distance.

In this paper, two novel incremental FCM-type clustering methods based on information bottleneck (IB) theory are presented. The two algorithms modify spFCM and oFCM by scoring mutual information loss to measure the distance between centroids and objects, so they are called spFCM-IB and oFCM-IB respectively. The similarity measure based on IB considers different weights for each centroid and object, which is equivalent to the normalization step in spHFCM and oHFCM. Furthermore, many empirical results show that IB based similarity measure is much better than such conventional measures as cosine and Jaccard in clustering documents [5,14,15].

The remainder of this paper is organized as follows. In Section 2, we provide a literature review of FCM, spFCM, oFCM and IB theory. Section 3 introduces in detail the proposed algorithms, spFCM-IB and oFCM-IB. Several comprehensive experiments are performed to evaluate the effectiveness of our methods in Section 4 where experimental settings, performance metrics and results are provided. Finally, we conclude our work.

2. Related work

In this section, we briefly review related FCM-type clustering algorithms and information bottleneck theory, which could help to understand our algorithms introduced in next section. The explanations on the mathematical notations used in this paper are listed in Table 1.

Table 1
List of mathematical notations.

Notation	Description
C, N	Numbers of clusters, objects
u_{ci}	Fuzzy object partitioning membership
v_c	Fuzzy feature partitioning membership
d_{ij}	Relatedness measure between an object and a feature
m	FCM user-defined parameters
w	Weights of centroids and objects

2.1. Fuzzy c Means and Weighted Fuzzy c Means

FCM is a classic fuzzy clustering algorithm, which comes from K-Means, a classic hard clustering algorithm. FCM adds fuzzy logic into K-Means, represents the affiliation of objects to clusters by memberships and is thus regarded as the fuzzy version of K-Means. In FCM, every object has a degree of belonging to clusters, rather than belonging completely to just one cluster.

FCM aims at minimizing the objective function in Eq. (1).

$$J_{FCM} = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^{m} d_E(x_i - v_c) \tag{1}$$

where x_i is the i-th object, v_c is centroid of the c-th cluster, and the m ($m > 1$) determines the level of cluster fuzziness. A large m results in smaller memberships u_{ci} and in the limit $m = 1$, the memberships u_{ci} converge to 0 or 1, which implies the crisp K-Means. $d_E(x_i - v_c)$ is the Euclidean distance between object x_i and centroid v_c, and $d_E(x_i - v_c) = ||x_i - v_c||^2$.

In FCM, objects are equally important, however weighted FCM (wFCM) considers the relative importance of objects. The objective function of wFCM is defined as Eq. (2).

$$J_{wFCM} = \sum_{c=1}^{C}\sum_{i=1}^{N} w_i u_{ci}^{m} d_E(x_i - v_c) \tag{2}$$

where w_i is the importance weight of the i-th object. A large w_i shows the i-th object has high responsibility in centroid estimation. When incremental clustering is performed chunk by chunk, wFCM will be very important, because the centroids of previous chunks will be added into the next chunk as virtual objects that are certainly more important than ordinary objects.

2.2. Single-Pass Fuzzy c Means and Online Fuzzy c Means

The spFCM and oFCM are two incremental fuzzy clustering algorithms designed based on FCM for large data. These two algorithms employ wFCM to consider the relative importance of centroids and objects.

The process of spFCM is shown as Fig. 1. In spFCM, data in the first chunk is divided into c clusters, represented by c centroids, using FCM. Then the c weighted centroids are merged into the next chunk, and clustered again into c new higher weighted centroids together with objects in this chunk. This partial clustering is iterated until all the data has been scanned once.

Different from spFCM, oFCM classifies each chunk of data individually using FCM, and then centroids of all the chunks are collected and grouped by performing clustering again. The oFCM is illustrated as Fig. 2.

2.3. Information bottleneck theory

The information bottleneck method is motivated by Shannon's rate distortion theory. Given a random variable, X, and a distortion measure, $d(x_1, x_2)$, defined on the alphabet of X, we want to classify-quantize the symbols of X such that the average quantization error is less than a given number D.

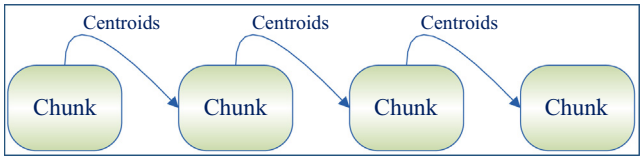

Fig. 1. *Single-Pass* fuzzy *c* means.