Signal Processing 126 (2016) 111–116

Contents lists available at ScienceDirect

Signal Processing

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

Data stream clustering based on Fuzzy C-Mean algorithm and entropy theory

Baoju Zhang, Shan Qin, Wei Wang*, Dan Wang, Lei Xue

College of Electronic and communication engineering, Tianjin Normal University, Tianjin, China

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form
8 October 2015
Accepted 15 October 2015
Available online 3 November 2015

Keywords:
Fuzzy C-Means
Clustering
Entropy theory
Concept drift detection

ABSTRACT

In data stream clustering studies, majority of methods are traditional hard clustering, the literatures of fuzzy clustering in clustering are few. In this paper, the fuzzy clustering algorithm is used to research data stream clustering, and the clustering results can truly reflect the actual relationship between objects and classes. It overcomes the either-or shortcoming of hard clustering. This paper presents a new method to detect concept drift. The membership degree of fuzzy clustering is used to calculate the information entropy of data, and according to the entropy to detect concept drift. The experimental results show that the detection of concept drift based on the entropy theory is effective and sensitive.

1. Introduction

In recent years, there are various sources for generating data streams, such as data from sensor networks, data generated by web click stream and data stream from internet traffic data transfer, nowadays, data stream become an important source of data. The data model flows into our life, people are facing the data rich, little knowledge situation. In order to seek the solution, all countries invest more manpower and material resources to develop the technology of data mining of data stream.

Data stream [1] is potential infinite, with uncertain arriving speed and can be scanned one pass. The processing of data stream has to implement within a limited memory space and a strict time constraint. Due to this, an efficient data stream mining algorithms must satisfy a more strict demand. S.Guha proposed Stream algorithm based on k-means [2], using centroid and submanifold to represent clustering. Stream algorithms process data in batches, and deal with the number of data points limited by memory size every time. This can't meet the large amount of data and fast distribution changes of the data flow. Alex [3] proposed NG and the SOM model based on a single pass (one pass). Chen et al. [4] use a tree structure in data stream clustering, to some extent that can overcome the problem of concept drift of data flow. Aboalsamh et al. [5] put forward a method of using incremental learning data streams classification model. Aggarwal, J.H proposed CLU Stream algorithm [6]. It is the first time to put forward the data stream as a process, and the process is changing over time, rather than regard it as a whole for clustering analysis. But the main clustering method is hard clustering. This algorithm divides each object to be identified strictly to a certain class. And, in fact, most of the objects are not strict attributes, they exist intermediately. They are suitable for fuzzy clustering. Fuzzy clustering on categorical data streams is a relatively new field.

In 1965, Zadeh proposed the fuzzy set theory, fuzzy theory was born. In the second year, Bellman and Kalaba and Zadeh, the first time put forward the fuzzy set theory combined with clustering analysis problems. Ruspini first expresses and systematically studies the fuzzy clustering. Now many researchers are giving importance on it to find an efficient algorithm in data stream mining. ZHANG Bo used fuzzy clustering method in the data flow model [7]. It has obtained the good result of clustering, but there is no in-depth study. CAI fully understands fuzzy clustering in data mining [8].

In the social production and life practice, a certain type of problem is the concept of the data contained may

* Corresponding author.

http://dx.doi.org/10.1016/j.sigpro.2015.10.014
0165-1684/

change with time. In other words, the concept that we try to learn from the data, is constantly evolving. For example, the customer buying interest change with time, network access patterns change with different users, the user actual behavior changes with its registered location [9]. The common characteristics of these questions are: the data constantly produced and flows, there is no end in the data flow, and the concept of data stream contains could change at any time. Concept drift [10] requires the learning system can detect the concept drift as early as possible, and adjustment to adapt to their concept drift, to keep the precision judgment of the subsequent data as far as possible. There are three types of concept drift, feature change, class change and both of them change. In this paper, entropy is introduced to detect the concept drift with feature change [11].

2. Methods

2.1. Fuzzy C-Means clustering algorithm

Fuzzy C-Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is developed by Dunn in 1973 and improved by Bezdek [12] in 1981. Fuzzy clustering generates a fuzzy partition based on the idea of partial membership expressed by the degree of membership of each object in a given cluster and Fuzzy C-Means is one of the most common fuzzy clustering techniques in data mining.

In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely too just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available. The FCM algorithm attempts to partition a finite collection of elements into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of cluster centers and a partition matrix where each element tells the degree to which element belongs to cluster . Any point x has a set of coefficients giving the degree of being in the kth cluster $u_k(x)$. With fuzzy C-Means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \quad m \in [1, \infty) \tag{1}$$

m is the a weighted index. The FCM aims to minimize an objective function:

$$J(U, c_1, \ldots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^m ||x_i - c_j||^2. \tag{2}$$

where:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/m-1}} \tag{3}$$

The necessary condition for the minimum of the objective function is

$$\bar{J}(U, c_1, \ldots, c_c, \lambda_1, \ldots, \lambda_n) = J(U, c_1, \ldots, c_c) + \sum_{j=1}^{n} \lambda \left(\sum_{i=1}^{c} u_{ij} - 1 \right)$$

$$= \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^m d_{ij}^2 + \sum_{j=1}^{n} \lambda i \left(\sum_{i=1}^{c} u_{ij} - 1 \right) \tag{4}$$

The FCM algorithm determines the cluster centroid and the membership matrix through iterations using the following steps:

Step 1. Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$.
Step 2. At k-step: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$,and $c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$.
Step 3. Update $U^{(k)}$, $U^{(k+1)}$,and $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/m-1}} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||xi - Vj||}{||xi - Vk||} \right)^{2/m-1}}$

Step 4. If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$ then stop; otherwise return to Step 2.

Unsupervised and always converges are the main advantages of FCM. Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes.

The output of the system has a set of coefficients giving the degree of being in the kth cluster $u_k(x)$ of any point x. The values of the membership will be used to calculate the entropy of new data to detect concept drift.

2.2. Degree of membership

In order to understand the meaning of the membership degree, we begin with the definition of membership.

Definition 1. For any set X, a membership function on X is any function from X to the real unit interval $[0, 1]$. Membership functions on Xrepresents fuzzy subsets of X. The membership function which represents a fuzzy set A is usually denoted by μ_A. For an element x of X, the value μ_A is called the *membership degree* of x in the fuzzy set A. The membership degree μ_A quantifies the grade of membership of the element x to the fuzzy set A. The value 0 means that x is not a member of the fuzzy set; the value 1 means that x is fully a member of the fuzzy set. The values between 0 and 1 characterize fuzzy members, which belong to the fuzzy set only partially [13].

The values between 0 and 1 are membership degree, which means the degree to belong to the Clustering center (Fig. 1).