Knowledge-Based Systems 59 (2014) 33–47

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Incremental entropy-based clustering on categorical data streams with concept drift

Yanhong Li, Deyu Li *, Suge Wang, Yanhui Zhai

School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China
Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, 030006 Shanxi, China

ARTICLE INFO

Article history:
Received 7 January 2013
Received in revised form 16 January 2014
Accepted 1 February 2014
Available online 7 February 2014

Keywords:
Categorical data stream
Clustering
Data labeling
Concept drift detection
Cluster evolving analysis

ABSTRACT

Clustering on categorical data streams is a relatively new field that has not received as much attention as static data and numerical data streams. One of the main difficulties in categorical data analysis is lacking in an appropriate way to define the similarity or dissimilarity measure on data. In this paper, we propose three dissimilarity measures: a point-cluster dissimilarity measure (based on incremental entropy), a cluster–cluster dissimilarity measure (based on incremental entropy) and a dissimilarity measure between two cluster distributions (based on sample standard deviation). We then propose an integrated framework for clustering categorical data streams with three algorithms: Minimal Dissimilarity Data Labeling (MDDL), Concept Drift Detection (CDD) and Cluster Evolving Analysis (CEA). We also make comparisons with other algorithms on several data streams synthesized from real data sets. Experiments show that the proposed algorithms are more effective in generating clustering results and detecting concept drift.

1. Introduction

Many natural and artificial systems in practical applications such as real-time monitoring, stock market, and credit card fraud detection, continuously generate the temporally ordered, fast changing, massive and potentially infinite data streams. The research on data stream mining is becoming important and meaningful [1,2]. A data stream is defined as a real-time, continuous, ordered (implicitly by arrival time or explicitly by time-stamp) sequence of data items [3]. In recent years, several kinds of data mining researches have been explored for the data stream environment, including the summarization and statistics [4–6], data selection [7], change detection [8,9], sampling [10], data clustering [11–16] and data classification [17–20]. Ditzler and Polikar [21] discussed learning concept drift from imbalanced data. Ghazikhani et al. [22] proposed an online ensemble of classifiers for non-stationary and imbalanced data streams. Lu et al. [23] took the training case-base as an evolving data stream and proposed a new case-base editing method targeting competence enhancement under concept drifting environment.

Clustering is a widely used technique used to identify the cluster structure in an unlabeled data set by objectively organizing data into homogeneous groups and maximizing the within-group-object similarity as well as minimizing the between-group-object similarity [24]. Clustering techniques for data streams are very different from those for static data (i.e., data set that is unchanged in the clustering process), because it is difficult to control the order in which data items arrive, to store an entire data stream, or to scan through it multiple times due to its tremendous volume [1]. Another distinguishing characteristic of data streams is that they are time-varying. Changes in the hidden context can induce more or less radical changes in the target concept, generally known as concept drift [25]. As the concepts behind the data evolve with time, the underlying clusters may also change considerably with time [24]. Performing clustering on the entire time-evolving data not only decreases the quality of clusters but also disregards the expectations of users, who usually require the recent clustering results [26]. Thus, discovery of the concepts hidden in data streams imposes a great challenge upon cluster analysis.

Many researches on clustering data streams in the numerical domain have been reported [11–13,27,15,28–34]. Actually, categorical data streams prevalently exist in real data. In the categorical domain, however, the above algorithm is infeasible because the numerical characteristics of clusters are difficult to

* Corresponding author. Tel.: +86 13303408298; fax: +86 3517018176.
E-mail addresses: liyh@sxu.edu.cn (Y. Li), lidy@sxu.edu.cn (D. Li), wsg@sxu.edu.cn (S. Wang), chai_yanhui@163.com (Y. Zhai).

http://dx.doi.org/10.1016/j.knosys.2014.02.004
0950-7051/

define. Nasraoui et al. [35] presented a strategy to mine evolving user profiles in the Web and designed an algorithm for tracking evolving user profiles based on clustering results. Chen et al. [26] proposed a framework for clustering concept-drifting categorical time-evolving data. In their framework, a kind of cluster representative is defined based on the importance of the combinations of attribute values and an algorithm, named maximal resemblance data labeling, is then proposed to allocate each unlabeled data point into a corresponding appropriate cluster by utilizing cluster representative. In Chen's framework, the reclustering is performed in the current sliding window when quite a large number of outliers are found or quite a large number of clusters are varied in the ratio of data points in the current temporal clustering result obtained by data labeling. However, we claim that the reclustering is not necessary when quite a large number of clusters are varied in the ratio of data points, because every data point in the current sliding window has been properly labeled. By defining the distance between two sliding windows, Cao et al. [36] proposed an algorithm for detecting concept-drifting windows on the categorical time-evolving data. But in his framework, concept-drifting windows are detected based on the distance between adjacent sliding windows. When computing the distance, all data points in each window are regarded as a cluster without taking both cluster distribution and outliers into consideration. Thus, it is desired to devise an efficient method for clustering categorial data streams.

In this paper, we propose an integrated framework for clustering categorical data streams by using sliding window technique and data labeling technique. It consists of three parts: Minimal Dissimilarity Data Labeling (MDDL), Concept Drift Detection (CDD) and Cluster Evolving Analysis (CEA). In this framework, the initial clustering is performed on the first sliding window. MDDL marks an incoming data point in the current sliding window with a proper cluster label by referring to the clustering result of the previous sliding window, and the data points that cannot be exactly marked are regarded as outliers. There are two cases to be considered as concept drift. One case occurs when the outlier ratio in the current window is larger than a given threshold. In this case, a reclustering is performed in the current window. Another case occurs when the cluster distribution in the current window has a larger difference with that in the previous window. CDD is designed to explore the two cases and to find out the concept drift windows. In order to iconically show the cluster evolving process, the representative of a cluster and a dissimilarity measure between two clusters with adjacent time stamps are defined. CEA is designed to analyze the time-evolving trend of clusters at different time stamps. The comparative experiments validate the availability of the proposed framework.

The major contributions of this paper are the following:

- An integrated framework is proposed for clustering categorical data streams by using sliding window technique and data labeling technique.
- An effective data labeling algorithm is developed based on the point-cluster dissimilarity measure.
- The dissimilarity measure between two cluster distributions is employed to detect the concept drift.
- The cluster–cluster dissimilarity measure is employed to analyze the time-evolving trend of data stream.

This paper is set up as follows. In Section 2, the problem of clustering categorical data steams is formulated. In Section 3, a dissimilarity measure between a data point and a cluster is defined by incremental entropy and MDDL algorithm is proposed. In Section 4, a dissimilarity measure between two cluster distributions is defined, and CDD algorithm is designed. In Section 5, the cluster representative is defined, and CEA algorithm is proposed based on the dissimilarity measure between two clusters. Section 6 reports our experimental study on synthetic data sets generated from a few of real raw data sets. Section 7 concludes the paper with some remarks.

2. Problem description

Suppose that a set of categorical data points DS is given, where each data point $\mathbf{x}_i$ is a d-dimensional vector of attribute values, i.e., $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^d)$. Each component x_i^j $(1 \leqslant j \leqslant d)$ takes a value from the domain V_j of the jth attribute. It should be noticed that the data points in DS are ordered. Sliding window is an often-used technique for observing and analyzing a data stream. The size of sliding window usually indicates how large time scale or data granularity will be utilized by analysts to data analysis. When the window size N is given the data set DS is then separated into a series of continuous sliding windows S^t, where the superscript t is the identification number of the sliding window, also called time stamp.

The characteristics of continuation, speediness, order, changing, huge amount of data streams require a fast, real-time response of data analysis method. Data labeling technique is often adopted to improve the efficiency of clustering [26,36]. In our framework, let $C^{t-1} = \left\{c_1^{t-1}, c_2^{t-1}, \ldots, c_{k^{t-1}}^{t-1}\right\}$ be the clustering result of the sliding window S^{t-1}, where c_m^{t-1} $(1 \leqslant m \leqslant k^{t-1})$ is the mth cluster. Utilizing the cluster information of C^{t-1} we mark each data point in S^t with a proper label corresponding to a cluster of C^{t-1}. And the labeling result $C'^t = \left\{c_1'^t, c_2'^t, \ldots, c_{k^{t-1}}'^t, outliers'^t\right\}$ of S^t will be called the temporal clustering result, where $outliers'^t$ is the set of data points in S^t that cannot be marked with any proper cluster label of C^{t-1}.

3. Incremental entropy and data labeling

3.1. Some basic notions of entropy

As a kind of measure of the uncertainty of a random variable [37], Shannon entropy and its variants were widely applied to almost all disciplines such as pattern discovery [38], numerical clustering [39] and categorical data clustering [40–44]. Let x be a discrete random variable taking a finite number of possible values $v_1, v_2, \ldots, v_n$ with probabilities $p_1, p_2, \ldots, p_n$ respectively, such that $p_i \geqslant 0$ $(i = 1, 2, \ldots, n)$, and $\sum_{i=1}^n p_i = 1$. The entropy $H(x)$ of a discrete random variable x is defined by

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{1}$$

Let $X = (x^1, x^2, \ldots, x^d)$ be a discrete random vector, a finite set V_j be the domain of x^j $(1 \leqslant j \leqslant d)$. $p(x^j = v)$ denotes the probability of the event $x^j = v$, where $v \in V_j$. If random variables x^j $(1 \leqslant j \leqslant d)$ are independent, the information entropy $H(X)$ of X is defined as [37]

$$H(X) = \sum_{j=1}^{d} H(x^j) = -\sum_{j=1}^{d} \sum_{v \in V_j} p(x^j = v) \log_2 p(x^j = v). \tag{2}$$

Entropy-based measures can evaluate the orderliness of a given cluster [43]. Also, entropy criterion is especially good for categorical data clustering because of the lack of intuitive distance definition for categorical values.