Neurocomputing 120 (2013) 590–596

Contents lists available at ScienceDirect

Neurocomputing

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

An improved k-prototypes clustering algorithm for mixed numeric and categorical data

Jinchao Ji [a,b], Tian Bai [a], Chunguang Zhou [a], Chao Ma [a], Zhe Wang [a,*]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] School of Computer Science and Information Technology, Northeast Normal University, Changchun 130024, China

ARTICLE INFO

Article history:
Received 2 February 2012
Received in revised form
4 April 2013
Accepted 7 April 2013
Communicated by T. Heskes
Available online 25 May 2013

Keywords:
Clustering
Data mining
Mixed data
Dissimilarity measure
Attribute significance

ABSTRACT

Data objects with mixed numeric and categorical attributes are commonly encountered in real world. The k-prototypes algorithm is one of the principal algorithms for clustering this type of data objects. In this paper, we propose an improved k-prototypes algorithm to cluster mixed data. In our method, we first introduce the concept of the distribution centroid for representing the prototype of categorical attributes in a cluster. Then we combine both mean with distribution centroid to represent the prototype of the cluster with mixed attributes, and thus propose a new measure to calculate the dissimilarity between data objects and prototypes of clusters. This measure takes into account the significance of different attributes towards the clustering process. Finally, we present our algorithm for clustering mixed data, and the performance of our method is demonstrated by a series of experiments on four real-world datasets in comparison with that of traditional clustering algorithms.

1. Introduction

The rapid development of information technology has leaded to the generation of data at an exponential rate in a wide range of areas, such as business, society, engineering and healthcare. The emergence of large-scale data requires new techniques that can extract useful information and knowledge from them. Data mining is a suitable technique to satisfy this need [1–3].

In data mining, clustering is an important technique. The clustering algorithms are widely used in image processing, customer segmentation, gene expression analysis [4], and text documents analysis [5] etc. The aim of clustering is to divide a set of data objects into clusters such that data objects in the same cluster are more similar to each other than those in other clusters [6–9]. In real world, data sets usually contain both numeric and categorical attributes [10,11]. However, most existing clustering algorithms assume all attributes are either numeric or categorical, examples of which include the k-means [12], k-modes [13], fuzzy k-modes [14], TGCA [15], COOLCAT [16], and G-ANMI algorithms [17]. When mixed data are encountered, most of them usually exploit transformation approaches to convert one type of the attributes to the other and then apply traditional single-type clustering algorithms [3]. However, in most cases, transformation scheme may result in loss of information, leading to undesired clustering outcomes [11].

Up till now there has been some research carried out for directly dealing with mixed data. Cheessman and Stutz proposed the approach AutoClass [18], which exploited mixture model and Bayesian method to deal with mixed data. Li and Biswas presented the Similarity-Based Agglomerative Clustering (SBAC) [19], which is a hierarchical agglomerative algorithm. The SBAC algorithm adopts the similarity measure defined by Goodall [20] to evaluate similarities among data objects. Hsu and Chen gave a Clustering Algorithm based on the Variance and Entropy (CAVE) [3] for clustering mixed data. The CAVE algorithm needs to build a distance hierarchy for every categorical attribute, and the determination of the distance hierarchy requires domain expertise. Hsu et al. extended self-organizing map to analysis mixed data [11]. In their method, the distance hierarchy can be automatically constructed by using the values of class attributes. Chatzis proposed the KL-FCM-GM algorithm [21], which is based on the assumption that data deriving from clusters are in the Gaussian form and designed for the Gauss-Multinomial distributed data. David and Averbuch presented SpectralCAT [22] to deal with mixed data. In the SpectralCAT system, the numeric values were transformed to the categorical ones. Huang presented a k-prototypes algorithm [23], which integrated the k-means with k-modes methods to partition mixed data. Bezdek et al. proposed the fuzzy k-prototypes [24] which took into account the fuzzy nature of data objects. Zheng et al. [25] presented an evolutionary k-prototypes algorithm (EKP) by introducing an evolutionary algorithm framework.

* Corresponding author. Tel.: +86 431 85166477; fax: +86 431 85155352.
E-mail addresses: jinchao0374@163.com (J. Ji), dayton915@gmail.com (T. Bai), cgzhou@jlu.edu.cn (C. Zhou), billmach0913@gmail.com (C. Ma), wzj0431@gmail.com (Z. Wang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.04.011

In this paper, we present an improved k-prototypes algorithm for clustering mixed data. In our method, we first propose a new concept named the distribution centroid to represent the prototype of categorical attributes in a cluster. Then we integrate the mean with distribution centroid to represent the prototype of a cluster with mixed attributes, and propose a new dissimilarity measure, which takes account of the significance of each attribute, to evaluate the dissimilarity between data objects and prototypes. Finally we propose our algorithm for clustering mixed data.

The rest of this paper is organized as follows: we first introduce some notations used throughout this paper in Section 2. In Section 3, the conventional k-prototypes algorithm is described in detail. This is followed by the presentation of our method in Section 4. In Section 5, we present the experimental evidence that demonstrates the advantages of the proposed algorithm. Following this, an application to catalog marketing is presented in Section 6. Finally in Section 7 we draw conclusions of this paper.

2. Notations

Let $X=\{X_1,X_2,\ldots,X_n\}$ denote a set of n data objects to be clustered and X_i $(1\le i\le n)$ be a data object represented by m attributes $A_1,A_2,\ldots,A_m$. Each attribute A_j describes a domain of values denoted by $Dom(A_j)$ [14]. The domains of attributes associated with mixed data are numeric and categorical domains respectively. The numeric domain is represented by continuous values, and the categorical domain is represented by a finite set without any natural ordering (such as gender, color), which usually denoted by $Dom(A_j)=\{a_j^1,a_j^2,\ldots,a_j^t\}$, where t is the number of category values of the categorical attribute A_j. A data object X_i is logically represented as a conjunction of attribute-value pairs

$$[A_1=x_{i1}]\wedge[A_2=x_{i2}]\wedge\ldots\wedge[A_j=x_{ij}]\wedge\ldots\wedge[A_m=x_{im}],$$

where $x_{ij}\in Dom(A_j)$ for $1\le j\le m$. Without ambiguity, we represent X_i as a vector $[x_{i1},x_{i2},\ldots,x_{im}]$. We assume that every data object has exactly m attributes for the datasets considered in this paper.

3. The k-prototypes algorithm

This algorithm was proposed by Huang in [11]. The aim of the k-prototypes algorithm is to group the dataset X into k clusters by minimizing the cost function as given below:

$$E(U,Q)=\sum_{l=1}^{k}\sum_{i=1}^{n}u_{il}d(x_i,Q_l). \tag{3.1}$$

here Q_l is the prototype of the cluster l; $u_{il}(0\le u_{il}\le 1)$ is an element of the partition matrix $U_{n\times k}$; and $d(x_i,Q_l)$ is the dissimilarity measure which is given as:

$$d(x_i,Q_l)=\sum_{j=1}^{m}d(x_{ij},q_{lj}), \tag{3.2}$$

$$d(x_{ij},q_{lj})=\begin{cases}(x_{ij}-q_{lj})^2 \text{ if the } j\text{th attribute is the numeric attribute,}\\ \mu_l\delta(x_{ij},q_{lj}) \text{ if the } j\text{th attribute is the categorical attribute,}\end{cases} \tag{3.3}$$

where $\delta(p,q)=0$ for $p=q$, and $\delta(p,q)=1$ for $p\neq q$; μ_l is a weight for categorical attributes in the cluster l. When x_{ij} is a value of the numeric attribute, q_{ij} is the mean of the jth numeric attribute in the cluster l; when x_{ij} is the value of categorical attribute, q_{ij} is the mode of the jth categorical attribute in the cluster l. The process of the k-prototypes algorithm is described as follows:

Step 1. Randomly choose k data objects from the dataset X as the initial prototypes of clusters.

Step 2. For each data object in X, assign it to the cluster whose prototype is the nearest one to this data object in terms of Eq. (3.2). Following each assignation, update the prototype of cluster.

Step 3. Recalculate the similarity of data objects against the current prototypes after all data objects have been assigned to a cluster. If a data object is discovered that its nearest prototype belongs to another cluster rather than the current one, reassign this data object to that cluster and update the prototypes of both clusters.

Step 4. After a full circle test of X, if no data objects have changed clusters, end the algorithm; or else repeat Step3.

4. The proposed algorithm

The motivation of our algorithm is on one hand to present an effective representation for the categorical attribute part in a mixed prototype since the mean is good enough for the numeric attribute part, and on the other hand to consider the significance of different attributes towards the clustering process. In this section, we first present the idea of distribution centroid, which considers the former, and Huang's strategy of evaluating significance of attributes, which takes into account the latter. Then we propose our algorithm, which considers both of the aforementioned issues.

4.1. The distribution centroid

The fuzzy centroid was proposed in [26]. Kim et al. demonstrated that the fuzzy centroid was suitable for representing the center of a cluster for categorical attributes in the fuzzy scenario. Inspired by the idea about the fuzzy centroid, we define the distribution centroid to represent the center of a cluster for categorical attributes in the hard scenario. For $Dom(A_j)=\{a_j^1,a_j^2,\ldots,a_j^t\}$, the distribution centroid of a cluster l, denoted as C_l', is given by:

$$C_l'=\{c_{l}',c_{l2}',\ldots c_{lj}',\ldots,c_{lm}'\}, \tag{4.1.1}$$

where

$$c_{lj}'=\{\{a_j^1,\omega_{lj}^1\},\{a_j^2,\omega_{lj}^2\},\cdots\{a_j^\kappa,\omega_{lj}^\kappa\},\cdots,\{a_j^t,\omega_{lj}^t\}\}. \tag{4.1.2}$$

In the above,

$$\omega_{lj}^\kappa=\sum_{i=1}^{n}\eta(x_{ij}), \tag{4.1.3}$$

where

$$\eta(x_{ij})=\begin{cases}\frac{u_{il}}{\sum_{i=1}^{n}u_{il}}, & \text{if } x_{ij}=a_j^\kappa,\\ 0, & \text{if } x_{ij}\neq a_j^\kappa.\end{cases} \tag{4.1.4}$$

here

$$u_{il}=\begin{cases}1, \text{ if the data object } x_i \text{ appeared in the cluster } l,\\ 0, \text{ if the data object } x_i \text{ did not appear in the cluster } l.\end{cases} \tag{4.1.5}$$

In Eq. (4.1.2) ω_{lj}^κ is subject to

$$\begin{cases}0\le\omega_{lj}^\kappa\le 1, \quad 1\le\kappa\le t,\\ \sum_{\kappa=1}^{t}\omega_{lj}^\kappa=1, \quad 1\le j\le m.\end{cases} \tag{4.1.6}$$

From Eqs.(4.1.1)–(4.1.6), we can see that the distribution centroid records the frequency of occurrences for each value of the categorical attribute in a cluster. It therefore captures the