A *k*-Means-Like Algorithm for Clustering Categorical Data Using an Information Theoretic-Based Dissimilarity Measure

Thu-Hien Thi Nguyen and Van-Nam Huynh(✉)

School of Knowledge Science, Japan Advanced Institute of Science and Technology,
Nomi, Japan
{nguyen.hien,huynh}@jaist.ac.jp

Abstract. Clustering large datasets is one of the important research problems for many machine learning applications. The k-means is very popular and widely used due to its ease of implementation, linear time complexity in size of the data, and almost surely convergence to local optima. However, working only on numerical data prohibits it from being used for clustering categorical data. In this paper, we aim to introduce an extension of k-means algorithm for clustering categorical data. Basically, we propose a new dissimilarity measure based on an information theoretic definition of similarity that considers the amount of information of two values in the domain set. The definition of cluster centers is generalized using kernel density estimation approach. Then, the new algorithm is proposed by incorporating a feature weighting scheme that automatically measures the contribution of individual attributes for the clusters. In order to demonstrate the performance of the new algorithm, we conduct a series of experiments on real datasets from UCI Machine Learning Repository and compare the obtained results with several previously developed algorithms for clustering categorical data.

Keywords: Cluster analysis · Categorical data clustering · K-means · Dissimilarity measures

1 Introduction

During the last decades, data mining has emerged as a rapidly growing interdisciplinary field, which merges together databases, statistics, machine learning and other related areas in order to extract useful knowledge from data [11]. Cluster analysis or simply clustering is one of fundamental tasks in data mining that aims at grouping a set of data objects into multiple clusters, such that objects within a cluster are similar one another, yet dissimilar to objects in other clusters. Dissimilarities and similarities between objects are assessed based on those attribute values describing the objects and often involve distance measures.

Typically, objects can be considered as vectors in n-dimensional space, where n is the number of features. When objects are described by numerical features,

M. Gyssens and G. Simari (Eds.): FoIKS 2016, LNCS 9616, pp. 115–130, 2016.
DOI: 10.1007/978-3-319-30024-5_7

the distance measure based on geometric concept such as Euclid distance or Manhattan distance can be used to define similarity between objects. However, these geometric distance measures are not applicable for categorical data which contains values, for instance, from gender, locations, etc. Recently, clustering data with categorical attributes have increasingly gained considerable attention [7–10,13,14]. As for categorical data, the comparison measure is most naturally used [13]. However, this metric does not distinguish between the different values taken by the attribute, since we only measure the equality between pair of values, as argued in [18].

In this paper we propose a new extension of the k-means algorithm for clustering categorical data. In particular, as for measuring dissimilarity between categorical objects, we make use of the information theoretic definition of similarity proposed in [20], which is intuitively defined based on the amount of information contained in the statement of commonality between values in the domain set of a categorical attribute. On the other hand, the definition of cluster centers is generalized using the kernel-based density estimates for categorical clusters as similarly considered in [6], instead of using the frequency estimates as originally in [24]. We then develop a new clustering algorithm by incorporating a feature weighting scheme that automatically measures the contribution of individual attributes to formation of the clusters.

The rest of this paper is organized as follows. Section 2 briefly describes the related work. Section 3 first introduces the k-means algorithm, and then presents its existing extensions for clustering categorical data. The proposed method is discussed in Sect. 4, and the experimental results are presented in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related Work

Probably, the k-means clustering [21] is the most well-known approach for clustering data sets with numerical attributes. It is a traditional partitioning based approach which starts with k random centroids and the centroids are updated iteratively by computing the average of the numerical features in each cluster. Each observation or object is assigned to clusters based upon the nearest distance to the means of the clusters. The iteration continues until the assignment is stable, that is, the clusters formed in the current stage are the same as those formed in the previous stage. The k-means is very popular due to its ease of implementation, linear time complexity in size of the data, and almost surely convergence to local optima [25]. However, in real life many data sets are categorical, of which k-means algorithm cannot be directly applied.

In recent years several attempts have been made in order to overcome the numerical-only limitation of k-means algorithm so as to make it applicable to clustering for categorical data, such as k-modes algorithm [14] and k-representative algorithm [24]. Particularly, in the k-modes algorithm [14], the simple matching similarity measure is used to compute distance between categorical objects, and "modes" are used instead of means for cluster centers. The mode of a cluster is a data point, in which the value of each attribute is assigned the most frequent value of the attribute's domain set appearing in the cluster. Furthermore, Huang also combined the k-modes algorithm with k-means algorithm in order to deal with mixed numerical and categorical databases. These extensions allow us to efficiently cluster very large data sets from real world applications. It is worth, however, noting that a cluster can have more than one mode and the performance of k-mode algorithm depends strongly on the selection of modes during the clustering process. In an attempt to overcome this drawback, San et al. [24] introduced a new notion of "cluster centers" called representatives for categorical objects. In particular, the representative of a cluster is defined making use of the distributions of categorical values appearing in clusters. Then, the dissimilarity between a categorical object and the representative of a cluster is easily defined based on relative frequencies of categorical values within the cluster and the simple matching measure between categorical values. In such a way, the resulting algorithm called k-representative algorithm is then formulated in a similar fashion to the k-means algorithm. In fact, it has been shown that the k-representative algorithm is very effective in clustering categorical data [22].

More recently, Chen and Wang [6] have proposed a new kernel density based method for defining cluster centers in central clustering of categorical data. Then the so-called k-centers algorithm that incorporates the new formulation of cluster centers and the weight attributes calculation scheme has been also developed. The experimental results have shown that the k-centers algorithm has good performance especially for the task of recognizing biological concepts in DNA sequences.

3 k-Means Algorithm and Its Extensions for Categorical Data

Assume that DB is a data set consisting of N objects, each of which is characterized by a set of D attributes with finite domains $O_1, \ldots, O_D$, respectively. That is, each object in DB is represented by a tuple $t \in O_1 \times \ldots \times O_D$, and the d^{th} attribute takes $|O_d|(> 1)$ discrete values. In addition, the categories in O_d will be denoted by o_{dl}, for $l = 1, \ldots, |O_d|$, and each data object in DB will be denoted by X, with subscript if necessary, which is represented as a tuple $X = (x_1, ..., x_D) \in O_1 \times ... \times O_D$. Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the set of k clusters of DB, i.e. we have

$$C_j \cap C_{j'} = \emptyset \text{ if } j \neq j' \text{ and } DB = \bigcup_{j=1}^{k} C_j$$

Regarding the clustering problem discussed in this paper, we consider two types of data: *numeric* and *categorical.* The domain of numerical attributes consists of continuous real values. Thus, the distance measure based on geometric concept such as the Euclid distance or Manhattan distance can be used.

A domain O_d is defined as categorical if it is finite and unordered, so that only a comparison operation is allowed in O_d. It means, for any $x, y \in O_d$, we have either $x = y$ or $x \neq y$.

3.1 *k*-Means Algorithm

The k-means algorithm [21] is one of the most popular algorithm in *partitional* or *non-hierarchical* clustering methods. Given a set DB of N numerical data objects, a natural number $k \leq N$, and a distance measure $\text{dis}(\cdot, \cdot)$, the k-means algorithm searches for a partition of DB into k non-empty disjoint clusters that minimizes the overall sum of the squared distances between data objects and their cluster centers. Mathematically, the problem can be formulated in terms of an optimization problem as follows:

Minimize

$$P(U, \mathcal{V}) = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{i,j} \text{dis}(X_i, V_j) \tag{1}$$

subject to

$$\sum_{j=1}^{k} u_{i,j} = 1, \quad 1 \leq i \leq N,$$
$$u_{i,j} \in \{0, 1\}, \quad 1 \leq i \leq N, 1 \leq j \leq k, \tag{2}$$

where $U = [u_{i,j}]_{N \times k}$ is a partition matrix ($u_{i,j}$ take value 1 if object X_i is in cluster C_j, and 0 otherwise), $\mathcal{V} = \{V_1, \ldots, V_k\}$ is a set of cluster centers, and $\text{dis}(\cdot, \cdot)$ is the squared Euclidean distance between two objects.

The problem P can be solved by iteratively solving two problems:

- Fix $\mathcal{V} = \hat{\mathcal{V}}$ then solve the reduced problem $P(U, \hat{\mathcal{V}})$ to find $\hat{U}$.
- Fix $U = \hat{U}$ then solve the reduced problem $P(\hat{U}, \mathcal{V})$.

Basically, the k-means algorithm iterates through a three-step process until $P(U, \mathcal{V})$ converges to some local minimum:

1. Select an initial $\mathcal{V}^{(0)} = V_1^{(0)}, \ldots, V_k^{(0)}$, and set $t = 0$.
2. Keep $\mathcal{V}^{(t)}$ fixed and solve $P(U, \mathcal{V}^{(t)})$ to obtain $U^{(t)}$. That is, having the cluster centers, we then assign each object to the cluster of its nearest cluster center.
3. Keep $U^{(t)}$ fixed and generate $\mathcal{V}^{(t+1)}$ such that $P(U^{(t)}, \mathcal{V}^{(t+1)})$ is minimized. That is, construct new cluster centers according to the current partition.
4. In the case of convergence or if a given stopping criterion is fulfilled, output the result and stop. Otherwise, set $t = t + 1$ and go to step 2.

In numerical clustering problem, the Euclidean norm is often chosen as a natural distance measure in the k-means algorithm. With this distance measure, we calculate the partition matrix in step 2 as below, and the cluster center is computed by the mean of cluster's objects.

$$\begin{aligned} &\text{if } \text{dis}(X_i, V_j) \leq \text{dis}(X_i, V_p) \text{ then} \\ &u_{i,j} = 1, \text{and } u_{i,p} = 0, \quad \text{for } 1 \leq p \leq k, p \neq j \end{aligned} \tag{3}$$

3.2 Extensions of *k*-Means for Categorical Data

k-Modes Algorithm. It was also shown in [13] that the k-means method can be extended to categorical data by using a simple matching distance measure for categorical objects and the most frequent values to define the "cluster centers" called modes. Let X_1, X_2 are two categorical objects in DB, with $X_1 = (x_{11}, \ldots, x_{1D})$ and $X_2 = (X_{21}, \ldots, X_{2D})$. The dissimilarity between X_1 and X_2 can be computed by the total matching of the corresponding attribute values of the two objects. Formally,

$$\mathrm{dis}(X_1, X_2) = \sum_{d=1}^{D} \delta(x_{1d}, x_{2d}) \tag{4}$$

where

$$\delta(x_{1d}, x_{2d}) = \begin{cases} 0 & \text{if } x_{1d} = x_{2d}, \\ 1 & \text{if } x_{1d} \neq x_{2d}. \end{cases}$$

Given a cluster $\{X_1, \ldots, X_p\}$ of categorical objects, with $X_i = (x_{i1}, \ldots, x_{iD})$, $1 \leq i \leq p$, its mode $V = (o_1, \ldots, o_D)$ is defined by assigning o_d, $1 \leq d \leq D$, the value most frequently appeared in $\{x_{1d}, \ldots, x_{pd}\}$. With these modifications, Huang [14] developed the k-modes algorithm that mimics the k-means method to cluster categorical data. However, as mentioned previously, by definition the mode of a cluster is not in general unique. This makes the algorithm unstable depending on the selection of modes during the clustering process.

k-Representative Algorithm. In stead of using modes for cluster centers as in [13], San et al. [24] proposed the notion of *representatives* for clusters defined as follows.

Again, let $C = \{X_1, \ldots, X_p\}$ be a cluster of categorical objects and

$$X_i = (x_{i1}, \ldots, x_{iD}), 1 \leq i \leq p.$$

For each $d = 1, \ldots, D$, let us denote O_d^C the set forming from categorical values $x_{1d}, \ldots, x_{pd}$. Then the representative of C is defined by $V_C = (v_1^C, \ldots, v_D^C)$, with

$$v_d^C = \{(o_{dl}, f_C(o_{dl})) \,|\, o_{dl} \in O_d^C\}, \tag{5}$$

where $f_C(o_{dl})$ is the relative frequency of category o_{dl} within C, i.e.

$$f_C(o_{dl}) = \frac{\#_C(o_{dl})}{p} \tag{6}$$

where $\#_C(o_{dl})$ is the number of objects in C having the category o_{dl} at d^{th} attribute. More formally, each v_d^C is a distribution on O_d^C defined by relative frequencies of categorical values appearing within the cluster.

Then, the dissimilarity between object $X = (x_1, \ldots, x_D)$ and representative V_C is defined based on the simple matching measure δ by

$$\mathrm{dis}(X, V_C) = \sum_{d=1}^{D} \sum_{o_{dl} \in O_d^C} f_C(o_{dl}) \cdot \delta(x_d, o_{dl}) \tag{7}$$

As such, the dissimilarity $\mathrm{dis}(X, V_C)$ is mainly dependent on the relative frequencies of categorical values within the cluster and simple matching between categorical values.

***k*-Centers Algorithm.** More generally, Chen and Wang [6] have recently proposed a generalized definition for centers of categorical clusters as follows. The center of a cluster C_j is defined as

$$V_j = [\boldsymbol{\nu}_{j1}, \ldots, \boldsymbol{\nu}_{jD}] \tag{8}$$

in which the d^{th} element $\boldsymbol{\nu}_{jd}$ is a probability distribution on O_d estimated by a kernel density estimation method [1]. More particularly, let denote X_d a random variable associated with observations x_{id}, for $i = 1, \ldots, |C_j|$, appearing in C_j at d^{th} attribute, and $p(X_d)$ its probability density. Let O_{jd} be the set forming from categorical values $\{x_{id}\}_{i=1}^{|C_j|}$. Then the kernel density based estimate of $p(X_d)$, denoted by $\hat{p}(X_d, \lambda_j | C_j)$, is of the following form (see, e.g., [27]):

$$\hat{p}(X_d, \lambda_j | C_j) = \sum_{o_{dl} \in O_{jd}} f_j(o_{dl}) K(X_d, o_{dl} | \lambda_j) \tag{9}$$

where $K(\cdot, o_{dl} | \lambda_j)$ is a so-called kernel function, $\lambda_j \in [0,\, 1]$ is a smoothing parameter called the bandwidth, and f_j is the frequency estimator for C_j, i.e.

$$f_j(o_{dl}) = \frac{\#_j(o_{dl})}{|C_j|} \tag{10}$$

with $\#_j(o_{dl})$ being the number of o_{dl} appearing in C_j. Note that another equivalent form of (9) was used in [6] for defining a kernel density estimate of $p(X_d)$.

Also, Chen and Wang [6] used a variation of Aitchison and Aitken's kernel function [1] defined by

$$K(X_d, o_{dl} | \lambda_j) = \begin{cases} 1 - \frac{|O_d|-1}{|O_d|} \lambda_j & \text{if } X_d = o_{dl} \\ \frac{1}{|O_d|} \lambda_j & \text{if } X_d \neq o_{dl} \end{cases} \tag{11}$$

to derive the estimate $\hat{p}(X_d, \lambda_j | C_j)$, which is then used to define $\boldsymbol{\nu}_{jd}$.

It is worth noting here that the kernel function $K(X_d, o_{dl} | \lambda_j)$ is defined in terms of the cardinality of the whole domain O_d but not in terms of the cardinality of the subdomain O_{jd} of the given cluster C_j.

From (9)–(11), it easily follows that $\boldsymbol{\nu}_{jd}$ can be represented as

$$\boldsymbol{\nu}_{jd} = \left[P_{jd}(o_{d1}), \ldots, P_{jd}(o_{dl}), \ldots, P_{jd}(o_{d|O_d|})\right]$$

where

$$P_{jd}(o_{dl}) = \lambda_j \frac{1}{|O_d|} + (1-\lambda_j) f_j(o_{dl}) \qquad (12)$$

and $\lambda_j \in [0,\, 1]$ is the bandwidth for C_j.

When $\lambda_j = 0$, the center degenerates to the pure frequency estimator, which is originally used in the k-representative algorithm to define the center of a categorical cluster.

To measure the dissimilarity between a data object and its center, each data object X_i is represented by a set of vectors $\{y_{id}\}_{d=1}^{D}$, with

$$y_{id} = \big[I(x_{id} = o_{d1}), \ldots, I(x_{id} = o_{dl}), \ldots, I(x_{id} = o_{d|O_d|})\big]$$

Here $I(\cdot)$ is an indicator function whose value is either 1 or 0, indicating whether x_{id} is the same as $o_{dl} \in O_d$ or not. The dissimilarity on the d^{th} dimension is then measured by

$$\text{dis}_d(X_i, V_j) = ||y_{id} - \boldsymbol{\nu}_{jd}||_2 \qquad (13)$$

We can see that, k-centers uses the different way to calculate the dissimilarities between objects and cluster centers, but the idea of comparing two categorical values is still based on the simple matching method (represented by indicator function $I(\cdot)$). The remains of the k-center mimics the idea of k-means algorithm.

4 The Proposed Algorithm

In this section we will introduce a new extension of the k-means clustering algorithm for categorical data by combining a slightly modified concept of cluster centers based on Chen and Wang's kernel-based estimation method and an information theoretic based dissimilarity measure.

4.1 Representation of Cluster Centers

Similar as in k-centers algorithm [6], for each cluster C_j, let us define the center of C_j as

$$V_j = [\boldsymbol{\nu}_{j1}, \ldots, \boldsymbol{\nu}_{jD}]$$

where $\boldsymbol{\nu}_{jd}$ is a probability distribution on O_d estimated by a kernel density estimation method.

As our aim is to derive a kernel density based estimate $\hat{p}(X_d, \lambda_j | C_j)$ for the d^{th} attribute of cluster C_j, instead of directly using Chen and Wang's kernel function defined in terms of the cardinality of the domain O_d as above, we use a slightly modified version as follows.

For any $o_{dl} \in O_d$, if $o_{dl} \in O_{jd}$ then we define

$$K(X_d, o_{dl}|\lambda_j) = \begin{cases} 1 - \frac{|O_{jd}|-1}{|O_{jd}|}\lambda_j & \text{if } X_d = o_{dl} \\ \frac{1}{|O_{jd}|}\lambda_j & \text{if } X_d \neq o_{dl} \end{cases} \tag{14}$$

otherwise, i.e. $o_{dl} \notin O_{jd}$, we let $K(X_d, o_{dl}|\lambda_j) = 0$. Then, from (9), (10) and (14) it easily follows that $\boldsymbol{\nu}_{jd}$ can be obtained as

$$\boldsymbol{\nu}_{jd} = \left[P_{jd}(o_{d1}), \ldots, P_{jd}(o_{dl}), \ldots, P_{jd}(o_{d|O_d|})\right] \tag{15}$$

where

$$P_{jd}(o_{dl}) = \begin{cases} \lambda_j \frac{1}{|O_{jd}|} + (1-\lambda_j) f_j(o_{dl}) & \text{if } o_{dl} \in O_{jd} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

and $\lambda_j \in [0,\, 1]$ is the smoothing parameter for C_j.

The parameter λ_j is selected using the least squares cross validation (LSCV) as done in [6], which is based on the principle of selecting a bandwidth that minimizes the total error of the resulting estimation over all the data objects. Specifically, the optimal λ_j^* is determined by the following equation:

$$\lambda_j^* = \frac{1}{|C_j| - 1} \frac{\sum_{d=1}^{D} (1 - \sum_{o_{dl} \in O_{jd}} [f_j(o_{dl})]^2)}{\sum_{d=1}^{D} (\sum_{o_{dl} \in O_{jd}} [f_j(o_{dl})]^2 - \frac{1}{|O_{jd}|})} \tag{17}$$

4.2 Dissimilarity Measure

Instead of using the simple matching measure as in [13,24] or the Euclidean norm as in [6], we first introduce a dissimilarity measure for categorical values of each attribute domain based on an information-theoretic definition of similarity proposed by Lin [20], and then propose a new method for computing the distance between categorical objects and cluster centers, making use of the kernel density based definition of centers and the information-theoretic based dissimilarity measure for categorical data.

In [20], Lin developed an information-theoretic framework for similarity within which a formal definition of similarity can be derived from a set of underlying assumptions. Basically, Lin's definition of similarity is stated in information theoretic terms, as quoted "the similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are." Formally, the similarity between A and B is generally defined as

$$\text{sim}(A, B) = \frac{\log P(\text{common}(A, B))}{\log P(\text{description}(A, B))} \tag{18}$$

where $P(s)$ is the probability of a statement s. To show the universality of the information-theoretic definition of similarity, Lin [20] also discussed it in different settings, including ordinal domain, string similarity, word similarity and semantic similarity.

In 2008, Boriah et al. [5] applied Lin's framework to the categorical setting and proposed a similarity measure for categorical data as follows. Let DB be a data set consisting of objects defined over a set of D categorical attributes with finite domains denoted by $O_1, \ldots, O_D$, respectively. For each $d = 1, \ldots, D$, the similarity between two categorical values $o_{dl}, o_{dl'} \in O_d$ is defined by

$$\mathrm{sim}_d(o_{dl}, o_{dl'}) = \begin{cases} 2\log f_d(o_{dl}) & \text{if } o_{dl} = o_{dl'} \\ 2\log(f_d(o_{dl}) + f_d(o_{dl'})) & \text{otherwise} \end{cases} \tag{19}$$

where

$$f_d(x) = \frac{\#(x)}{|DB|}$$

with $\#(x)$ being the number of objects in DB having the category x at d^{th} attribute. In fact, Boriah et al. [5] also proposed another similarity measure derived from Lin's framework and conducted an experimental evaluation of many different similarity measures for categorical data in the context of outlier detection.

It should be emphasized here that the similarity measure $\mathrm{sim}_d(\cdot, \cdot)$ does not satisfy the Assumption 4 assumed in Lin's framework [20], which states that the similarity between a pair of identical object is 1. Particularly, the range of $\mathrm{sim}_d(o_{dl}, o_{dl'})$ for $o_{dl} = o_{dl'}$ is $[-2\log|DB|, 0]$, with the minimum being attained when o_{dl} occurs only once and the maximum being attained when $O_d = \{o_{dl}\}$. Similarly, the range of $\mathrm{sim}_d(o_{dl}, o_{dl'})$ for $o_{dl} \neq o_{dl'}$ is $\left[-2\log\frac{|DB|}{2}, 0\right]$, with the minimum being attained when o_{dl} and $o_{dl'}$ each occur only once, and the maximum value is attained when o_{dl} and $o_{dl'}$ each occur $\frac{|DB|}{2}$ times, as pointed out in [5].

Based on the general definition of similarity given in (18) and its application to similarity between ordinal values briefly discussed in [20], we introduce another similarity measure for categorical values as follows.

For any two categorical values $o_{dl}, o_{dl'} \in O_d$, their similarity, denoted by $\mathrm{sim}^*_d(o_{dl}, o_{dl'})$, is defined by

$$\mathrm{sim}^*_d(o_{dl}, o_{dl'}) = \frac{2\log f_d(\{o_{dl}, o_{dl'}\})}{\log f_d(o_{dl}) + \log f_d(o_{dl'})} \tag{20}$$

where

$$f_d(\{o_{dl}, o_{dl'}\}) = \frac{\#(\{o_{dl}, o_{dl'}\})}{|DB|}$$

with $\#(\{o_{dl}, o_{dl'}\})$ being the number of categorical objects in DB that receive the value belonging to $\{o_{dl}, o_{dl'}\}$ at the d^{th} attribute. Clearly, we have $\mathrm{sim}^*_d(o_{dl}, o_{dl'}) = 1$ if o_{dl} and $o_{dl'}$ are identical, which satisfies the Assumption 4 stated as above.

Then, the dissimilarity measure between two categorical values $o_{dl}, o_{dl'} \in O_d$ can be defined by

$$\mathrm{dis}^*_d(o_{dl}, o_{dl'}) = 1 - \mathrm{sim}^*_d(o_{dl}, o_{dl'}) = 1 - \frac{2\log f_d(\{o_{dl}, o_{dl'}\})}{\log f_d(o_{dl}) + \log f_d(o_{dl'})} \tag{21}$$

Let $X_i = [x_{i1}, x_{i2}, \ldots, x_{iD}] \in DB$ and $V_j = [\boldsymbol{\nu}_{j1}, \ldots, \boldsymbol{\nu}_{jD}]$ be the center of cluster C_j. We are now able to extend the dissimilarity between categorical values of O_d to the dissimilarity on the d^{th} attribute between X_i and V_j, i.e. the dissimilarity between the d^{th} component $x_{id} \in O_d$ of X_i and the d^{th} component $\boldsymbol{\nu}_{jd}$ of the center V_j, as follows. Without danger of confusion, we shall also use dis^*_d to denote this dissimilarity and

$$\mathrm{dis}^*_d(X_i, V_j) = \sum_{o_{dl} \in O_{jd}} P_{jd}(o_{dl})\mathrm{dis}^*_d(x_{id}, o_{dl}) \tag{22}$$

4.3 Algorithm

With the modifications just made above, we are now ready to formulate the problem of clustering categorical data in a similar way as k-means clustering. Adapted from Huang's W-k-means algorithm [16], we also use a weighting vector $W = [w_1, w_2, \ldots, w_D]$ for D attributes and β being a parameter for attribute weight, where $0 \leq w_d \leq 1$ and $\sum_d w_d = 1$. The principal for attribute weighting is to assign a larger weight to an attribute that has a smaller sum of the within cluster distances and a smaller one to an attribute that has a larger sum of the within cluster distances. More details of this weighting scheme can be found in [16]. Then, the weighted dissimilarity between data object X_i and cluster center V_j, denoted by $\mathrm{dis}^*(X_i, V_j)$, is defined by

$$\mathrm{dis}^*(X_i, V_j) = \sum_{d=1}^{D} w_d^{\beta}\mathrm{dis}^*_d(X_i, V_j) = \sum_{d=1}^{D} w_d^{\beta} \sum_{o_{dl} \in O_{jd}} P_{jd}(o_{dl})\mathrm{dis}^*_d(x_{id}, o_{dl}) \tag{23}$$

Based on these definitions, the clustering algorithm now aims to minimize the following objective function:

$$J(U, \mathcal{V}, W) = \sum_{j=1}^{k}\sum_{i=1}^{N}\sum_{d=1}^{D} u_{i,j} w_d^{\beta}\mathrm{dis}^*_d(X_i, V_j) \tag{24}$$

subject to

$$\begin{array}{lr} \sum_{j=1}^{k} u_{i,j} = 1, & 1 \leq i \leq N \\ u_{i,j} \in \{0, 1\}, & 1 \leq i \leq N, 1 \leq j \leq k \\ \sum_{d=1}^{D} w_d = 1, & 0 \leq w_d \leq 1 \end{array}$$

where $U = [u_{i,j}]_{N \times k}$ is a partition matrix.

The proposed algorithm is formulated as below.

Algorithm 1.1. The Proposed Algorithm

```
Select an initial $\mathcal{V}^{(0)} = \{V_1^{(0)}, \ldots, V_k^{(0)}\}$, and set $t = 0$, $\lambda_j = 0$ for $j = 0, \ldots, k$, set $W^{(0)} = [1/D, \ldots, 1/D]$.
repeat
    Keep $\mathcal{V}^{(t)}$ and $W^{(t)}$ fixed, generate $U^{(t)}$ to minimize the distances between objects and cluster mode (using Eq. (23)).
    Keep $U^{(t)}$ fixed, update $\mathcal{V}^{(t+1)}$ using Eq. (16) and Eq. (17).
    Generate $W^{(t+1)}$ using formulas from [16].
    $t = t + 1$.
until The partition does not changed.
```

5 Experiments Results

In this section, we will provide experiments conducted to compare the clustering performances of k-modes, k-representatives and three modified versions of k-representatives briefly described as below.

- In the first modified version of k-representatives (namely, Modified 1), we replace the simple matching dissimilarity measure with the information theoretic-based dissimilarity measure defined by Eq. (21).
- In the second modified version of k-representatives (namely, Modified 2), we combine the new dissimilarity measure with the concept of cluster centers proposed by Chen and Wang [6], i.e. the Algorithm 1.1 uses Eq. (12) instead of Eq. (16) to update the cluster centers).
- The third modified version of k-representatives (namely, Modified 3) is exactly Algorithm 1.1, which incorporates the new dissimilarity measure with our modified representation of cluster centers.

5.1 Datasets

For the evaluation, we used real world data sets downloaded from the UCI Machine Learning Repository [4]. The main characteristics of the datasets are summarized in Table 1. These datasets are chosen to test our algorithm because of their public availability and since all attributes can be treated as categorical ones.

5.2 Clustering Quality Evaluation

Evaluating the clustering quality is often a hard and subjective task [18]. Generally, objective functions in clustering are purposely designed so as to achieve high intra-cluster similarity and low inter-cluster similarity. This can be viewed as an internal criterion for the quality of a clustering. However, as observed in the literature, good scores on an internal criterion do not necessarily translate into good effectiveness in an application. Here, by the same way as in [19], we use three external criteria to evaluate the results: Purity, Normalized Mutual

Table 1. Categorical datasets

Datasets	Number of instances	Number of attributes	Number of classes
Car	1728	6	4
Soybean (Small)	47	35	4
Soybean (Large)	683	35	19
Breast cancer	286	9	2
Nursery	12960	8	5
Mushroom	8124	22	2

Information (NMI) and Adjusted Rand Index (ARI). These methods make use of the original class information of each object and the cluster to which the same objects have been assigned to evaluate how well the clustering result matches the original classes.

We denote by $C = \{C_1, \ldots, C_J\}$ the partition of the dataset built by the clustering algorithm, and by $P = \{P_1, \ldots, P_I\}$ the partition inferred by the original classification. J and I are respectively the number of clusters $|C|$ and the number of classes $|P|$. We denote by N the total number of objects.

Purity Metric. Purity is a simple and transparent evaluation measure. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned objects and dividing by the number of objects in the dataset. High purity is easy to achieve when the number of clusters is large. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters.

$$Purity(C, P) = \frac{1}{N} \sum_{j} \max_{i} |C_j \cap P_i| \tag{25}$$

NMI Metric. The second metric (NMI) provides an information that is independent from the number of clusters [26]. This measure takes its maximum value when the clustering partition matches completely the original partition. NMI is computed as the average mutual information between any pair of clusters and classes

$$NMI(C, P) = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} |C_j \cap P_i| \log \frac{N|C_j \cap P_i|}{|C_j||P_i|}}{\sqrt{\sum_{j=1}^{J} |C_j| \log \frac{|C_j|}{N} \sum_{i=1}^{I} |P_i| \log \frac{|P_i|}{N}}} \tag{26}$$

ARI Metric. The third metric is the adjusted Rand index [17]. Let a be the number of object pairs belonging to the same cluster in C and to the same class in P. This metric captures the deviation of a from its expected value corresponding to the hypothetical value of a obtained when C and P are two random, independent partitions.

The expected value of a denoted by $E[a]$ is computed as follows:

$$E[a] = \frac{\pi(C)\pi(P)}{N(N-1)/2} \tag{27}$$

where $\pi(C)$ and $\pi(P)$ denote respectively the number of object pairs from the same clusters in C and from the same class in P. The maximum value for a is defined as:

$$\max(a) = \frac{1}{2}(\pi(C) + \pi(P)) \tag{28}$$

The agreement between C and P can be estimated by the adjusted rand index as follows:

$$ARI(C, P) = \frac{a - E[a]}{\max(a) - E[a]} \tag{29}$$

when $ARI(C, P) = 1$, we have identical partitions.

In many previous studies, only purity metric has been used to analyze the performance of clustering algorithm. However, purity is easy to achieve whens the number of cluster is large. In particular, purity is 1 if each object data gets its own cluster. Beside, many partitions have the same purity but they are different from each other e.g., the number of object data in each clusters, and which objects constitute the clusters. Therefore, we need the other two metrics to have the overall of how our clustering results matches the original classes.

5.3 Results

The experiments were run on a Mac with a 3.66 GHz Intel QuadCore processor, 8 GB of RAM running Mac OSX 10.10. For each categorical dataset, we run 300 times per algorithm. We provide the parameter k equals to the number of classes in each dataset. The performance of three evaluation metrics are calculated by the average after 300 times of running. The weighting exponent β was set to 8 as experimentally recommended in [16].

As we can see from Tables 2, 3 and 4, the modified versions 2 and 3 produce the best results in five out of six datasets. The results are remarkably good in the soybean (small) dataset, mushroom dataset, car dataset, soybean (large) dataset (when modified version 3 outperformed in all three metric) and breast cancer dataset (when the purity is slightly lower than the best one but the other two criteria are significantly higher). Comparing the performance of modified versions 2 and 3, we can see that the proposed approach yields better results in many cases, especially in NMI values and ARI values. In conclusion, the new approach has been proved to enhance the performance of previously developed k-means like algorithms for clustering categorical data.

Table 2. Purity results of categorical dataset algorithms

Datasets	K-mode	K-representative	Modified 1	Modified 2	Modified 3
Car	0.7	0.7	0.703	**0.705**	**0.705**
Soybean (Small)	0.873	0.961	0.967	0.981	**0.986**
Soybean (Large)	0.538	0.595	0.6	0.69	**0.71**
Breast-cancer	0.702	0.707	**0.713**	0.71	0.709
Nursery	0.409	0.425	0.435	0.451	**0.468**
Mushroom	0.518	0.83	0.864	**0.87**	**0.87**

Table 3. NMI results categorical dataset algorithms

Datasets	K-mode	K-representative	Modified 1	Modified 2	Modified 3
Car	0.051	0.077	0.101	0.119	**0.125**
Soybean (Small)	0.823	0.96	0.965	0.98	**0.981**
Soybean (Large)	0.533	0.708	0.71	0.72	**0.732**
Breast-cancer	0.0015	0.036	0.039	0.051	**0.057**
Nursery	0.044	0.047	0.055	**0.071**	**0.071**
Mushroom	9.26E-05	0.448	0.493	0.51	**0.521**

Table 4. Ajusted rand index results categorical dataset algorithms

Datasets	K-mode	K-representative	Modified 1	Modified 2	Modified 3
Car	0.028	0.024	0.043	0.049	**0.052**
Soybean (Small)	0.729	0.929	0.938	0.963	**0.975**
Soybean (Large)	0.306	0.379	0.38	0.44	**0.46**
Breast-cancer	-0.002	0.082	0.085	0.128	**0.135**
Nursery	0.034	0.028	0.034	0.046	**0.049**
Mushroom	4.62E-06	0.5	0.554	**0.589**	0.586

6 Conclusions

In this paper, we have proposed a new k-means like algorithm for clustering categorical data based on an information theoretic based dissimilarity measure and a kernel density estimate-based concept of cluster centers for categorical objects. Several variations of the proposed algorithm have been also discussed. The experimental results on real datasets from UCI Machine Learning Repository have shown that the proposed algorithm outperformed the k-means like algorithms previously developed for clustering categorical data. For the future work, we are planning to extend the proposed approach to the problem of clustering mixed numeric and categorical datasets as well as fuzzy clustering.

References

1. Aitchison, J., Aitken, C.G.G.: Multivariate binary discrimination by the kernel method. Biometrika **63**(3), 413–420 (1976)
2. Andritsos, P., Tsaparas, P., Miller, R.J., Sevcik, K.C.: LIMBO: A Scalable Algorithm to Cluster Categorical Data (2003)
3. Barbara, D., Couto, J., Li, Y.: COOLCAT: an entropy-based algorithm for categorical clustering. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 582–589 (2002)
4. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Databases. Dept. of Information and Computer Science, University of California at Irvine (1998). http://www.ics.uci.edu/mlearn/MLRepository.html
5. Boriah, S., Chandola, V., Kumar V.: Similarity measures for categorical data: a comparative evaluation. In: Proceedings of the SIAM International Conference on Data Mining, SDM 2008, pp. 243–254 (2008)
6. Chen, L., Wang, S.: Central clustering of categorical data with automated feature weighting. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 1260–1266 (2013)
7. Ganti, V., Gehrke, J., Ramakrishnan, R.: CATUS–clustering categorical data using summaries. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, San Diego, CA, pp. 73–83 (1999)
8. Gibson, D., Kleinberg, J., Raghavan, P.: Clustering categorical data: an approach based on dynamic systems. In: Proceedings of the 24th International Conference on Very Large Databases, New York, pp. 311–323 (1998)
9. Guha, S., Rastogi, R., Shim, K.: CURE: an efficient clustering algorithm for large databases. In: Proceedings of ACM SIGMOD International Conference on Management of Data, New York, pp. 73–84 (1998)
10. Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes. Inf. Syst. **25**(5), 345–366 (2000)
11. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco (2001)
12. Hathaway, R.J., Bezdek, J.C.: Local convergence of the c-means algorithms. Pattern Recogn. **19**, 477–480 (1986)
13. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Lu, H., Motoda, H., Luu, H. (eds.) KDD: Techniques and Applications, pp. 21–34. World Scientific, Singapore (1997)
14. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. **2**, 283–304 (1998)
15. Huang, Z., Ng, M.K.: A fuzzy k-modes algorithm for clustering categorical data. IEEE Trans. Fuzzy Syst. **7**, 446–452 (1999)
16. Huang, Z., Ng, M.K., Rong, H., Li, Z.: Automated variable weighting in k-means type clustering. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 657–668 (2005)
17. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1995)
18. Ienco, D., Pensa, R.G., Meo, R.: Context-based distance learning for categorical data clustering. In: Adams, N.M., Robardet, C., Siebes, A., Boulicaut, J.-F. (eds.) IDA 2009. LNCS, vol. 5772, pp. 83–94. Springer, Heidelberg (2009)
19. Ienco, D., Pensa, R.G., Meo, R.: From context to distance: learning dissimilarity for categorical data clustering. ACM Trans. Knowl. Discov. Data **6**(1), 1–25 (2012)
20. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, pp. 296–304 (1998)

21. MacQueen, J.B.: Some methods for classification, analysis of multivariate observations. In: Proceedings of the Fifth Symposium on Mathematical Statistics and Probability, Berkelely, CA, vol. 1(AD 669871), pp. 281–297 (1967)
22. Ng, M.K., Li, M.J., Huang, J.Z., He, Z.: On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 503–507 (2007)
23. Ralambondrainy, H.: A conceptual version of the k-means algorithm. Pattern Recog. Lett. **16**, 1147–1157 (1995)
24. San, O.M., Huynh, V.N., Nakamori, Y.: An alternative extension of the k-means algorithm for clustering categorical data. Int. J. Appl. Math. Comput. Sci. **14**, 241–247 (2004)
25. Selim, S.Z., Ismail, M.A.: k-Means-type algorithms: a generalized convergence theorem and characterization of local optimality. IEEE Trans. Pattern Anal. Mach. Intell. **6**(1), 81–87 (1984)
26. Strehl, A., Ghosh, J.: Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2003)
27. Titterington, D.M.: A comparative study of kernel-based density estimates for categorical data. Technometrics **22**(2), 259–268 (1980)