Pattern Recognition Letters 97 (2017) 46–53

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# A novel density peaks clustering algorithm for mixed data

Mingjing Du[a], Shifei Ding[a,b,*], Yu Xue[c]

[a] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China
[b] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[c] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China

ARTICLE INFO

*Article history:*
Received 31 May 2016
Available online 3 July 2017

*Keywords:*
Data clustering
Density peaks
Entropy
Mixed data

ABSTRACT

The density peaks clustering (DPC) algorithm is well known for its power on non-spherical distribution data sets. However, it works only on numerical values. This prohibits it from being used to cluster real world data containing categorical values and numerical values. Traditional clustering algorithms for mixed data use a pre-processing based on binary encoding. But such methods destruct the original structure of categorical attributes. Other solutions based on simple matching, such as K-Prototypes, need a user-defined parameter to avoid favoring either type of attribute. In order to overcome these problems, we present a novel clustering algorithm for mixed data, called DPC-MD. We improve DPC by using a new similarity criterion to deal with the three types of data: numerical, categorical, or mixed data. Compared to other methods for mixed data, DPC absolutely has more advantages to deal with non-spherical distribution data. In addition, the core of the proposed method is based on a new similarity measure for mixed data. This similarity measure is proposed to avoid feature transformation and parameter adjustment. The performance of our method is demonstrated by experiments on some real-world datasets in comparison with that of traditional clustering algorithms, such as K-Modes, K-Prototypes EKP and SBAC.



## 1. Introduction

Clustering analysis has attracted a lot of research attention due to its usefulness in many applications, including community detection, image processing, document processing, and so forth [1-7]. Clustering analysis has attracted a lot of research attention due to its usefulness in many applications, including most clustering algorithms rely on the assumption that data simply contain numerical values, but what should be dealt with is categorical values or mixed data containing both numerical and categorical values on data sets in the real world. For clustering algorithms dealing with mixed data, the core of these methods is how to measure the similarity for categorical attributes. Roughly, the existing clustering algorithms for mixed data can fall into two categories according to dealing with categorical attribute values. The first category of the methods is based on the pre-processing methods. The original attributes are transformed to new forms. Then, traditional distance functions are used to measure the transformed data in the new relation. The second category of approaches is based on similarity metrics dealing with categorical values directly.

Traditional clustering algorithms for mixed data have a pre-processing that is able to convert categorical attributes to new forms and facilitates processing. Binary encoding is the most common pre-processing method. This method transforms each categorical attribute to a set of binary attributes. For example, Ralambondrainy's algorithm [8] transforms categorical attributes into a set of binary attributes. Then, new forms are treated as numeric in the K-Means algorithm. Hence, we can directly adopt most traditional distances which are often used in numerical clustering, such as Euclidean distance, to define similarity between transformed objects. However, this method destructs the original structure of categorical attributes. In other words, transformed binary attributes are meaningless and their values are hard to interpret [9]. Apart from binary encoding, there are also other pre-processing methods. For example, in order to handle categorical data, Hsu [10] presents a new mechanism, distance hierarchy, which encodes a data set into a weighted tree structure. But it has a serious drawback that both the assignment of weights and the construction of distance hierarchies rely on domain knowledge.

In the respect of similarity metrics for categorical values, the K-Prototypes algorithm [11] is one of the most famous clustering algorithms for mixed data. Nevertheless, the choice of the weight $\gamma$ has a significant effect on clustering results. As a variation of K-Prototypes algorithm, evolutionary K-Prototypes algorithm (EKP) [12], an unsupervised evolutionary clustering algo-

* Corresponding author at: School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China.
*E-mail address:* dingsf@cumt.edu.cn (S. Ding).

http://dx.doi.org/10.1016/j.patrec.2017.07.001
0167-8655/

rithm for mixed type data, which integrates evolutionary computation framework with KP, also has a weight $\gamma$. Also, these algorithms [13,14] take into account the significance of different attributes towards the clustering process. However, a new parameter, the degree of fuzziness $\alpha$, is introduced into these clustering algorithms. Hence, it will come out that choosing the parameter is a delicate and difficult task for users that may be a roadblock for using K-Prototypes and its variations efficiently. In addition, some algorithms [15,16] use entropy-type measures to group objects. However, these methods only deal with categorical data instead of mixed data and these entropy-type criteria can only measure the similarity between an object and a cluster. Besides, OCIL [17] gives a unified similarity metric which can be applied to mixed data using the entropy-based criterion. This similarity metric is also based on the concept of object-cluster similarity. In other words, it only can measure the similarity between an object and a cluster. In addition, OCIL is an iterative clustering algorithm. This means that this method requires a random initialization and may trap into local optimum. Similar to OCIL, Lim, et al. [18] propose a clustering framework for mixed attribute type dataset based on the entropy concept. It also needs to adjust the parameter which is used to balance attribute type between categorical attribute and numerical one. Besides, Li and Biswas [9] propose a Similarity-Based Agglomerative Clustering (SBAC) algorithm based on a new similarity metric that deals with the mixed data. But this method is high computational complexity and only suitable for some small data sets.

From the above discussion, most of clustering algorithms use the K-Means paradigm to cluster data having values. It means that those methods have an iterative process and probably trap into local optimum. A new algorithm, density peaks clustering (DPC) [19], proposed by Rodriguez and Laio is published in the US journal Science. This algorithm is able to detect non-spherical clusters without specifying the number of clusters. And more important, DPC does not need to iterate. Some studies [20–24] have been going on around this method. However, there are still some shortcomings. For example, DPC algorithm cannot find the correct number of clusters automatically. In order to overcome this difficulty, Liang and Chen [25] propose the 3DC clustering based on the divide-and-conquer strategy and the density-reachable concept. Du et al. [26] propose a density peaks clustering based on k nearest neighbors (DPC-KNN) which introduces the idea of k nearest neighbors (KNN) into DPC and has another option for the local density computation.

This paper presents a novel clustering algorithm, DPC-MD, based on a new similarity measure for mixed data. Actually, the proposed algorithm is the generalization of the original DPC algorithm. In order to assess the performance of the proposed algorithm, we compare the proposed algorithm with other algorithms on some UCI data sets. As a result, our algorithms have achieved satisfactory results in most data sets.

## 2. Related works

### 2.1. Notations

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ denote a dataset of $n$ mixed data objects, where for each $i$, $1 \le i \le n$, $\mathbf{x}_i$ with $m$ features consists of $m_r$ numerical features and $m_o$ categorical features. Therefore, for each $i$, $1 \le i \le n$, and for $k$, $1 \le k \le m_r$, let $x_{i,k}^{(r)}$ be the $k$th feature of $\mathbf{x}_i^{(r)}$, where $\mathbf{x}_i^{(r)}$ is the numerical part. Similarly, for each $i$, and for $k$, $1 \le k \le m_o$, $x_{i,k}^{(o)}$ denotes the $k$th feature of $\mathbf{x}_i^{(o)}$, where $\mathbf{x}_i^{(o)}$ is the categorical part. The domain of numerical feature $F_k^{(r)}$ is represented by continuous values. And categorical feature $F_k^{(o)}$ has $t_k$ categories, i.e., $\mathrm{DOM}(F_k^{(o)}) = \{f_{k,1}, f_{k,2}, \cdots, f_{k,t_k}\}$, where $\mathrm{DOM}(F_k^{(o)})$ contains all possible values that can be chosen by attribute $F_k^{(o)}$. Therefore, $\mathbf{x}_i$ can be represented as $[\mathbf{x}_i^{(r)}, \mathbf{x}_i^{(o)}] = [x_{i,1}^{(r)}, x_{i,2}^{(r)}, \cdots, x_{i,m_r}^{(r)}, x_{i,m_r+1}^{(o)}, \cdots, x_{i,m}^{(o)}]$.

Distance functions such as Euclidean distance are used as similarity measure for numerical attribute. The Euclidean distance $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$ between the object $\mathbf{x}_i$ and the object $\mathbf{x}_j$ is defined as:

$$\mathrm{dist}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2. \tag{1}$$

The definition of the information entropy $H(x)$ is given, as follow:

$$H(x) = -\sum_{x \in V} p(x) log(p(x)). \tag{2}$$

where $p(x)$ is the probability mass function of the random variable $x$. $V$ is the finite set of possible outcomes of $x$.

### 2.2. Density peaks clustering

Its idea is that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. This method utilizes two important quantities: One is the local density $\rho_i$ of each point $\mathbf{x}_i$, and the other is its distance $\delta_i$ from points of higher density. The two quantities correspond to two assumptions with respect to the cluster centers. One is that the cluster centers are surrounded by neighbors with a lower local density. The other is that they have relatively larger distance to the points of higher density. In the following, we will describe the computation of $\rho_i$ and $\delta_i$ in much more detail.

DPC represents data objects as points in a space and adopts a distance metric, such as (1), as a similarity between objects.

The local density of a point $\mathbf{x}_i$, denoted by $\rho_i$, is defined as

$$\rho_i = \sum_j exp\left(-\frac{\mathrm{dist}\left(\mathbf{x}_i, \mathbf{x}_j\right)^2}{{d_c}^2}\right), \tag{3}$$

where $d_c$ is an adjustable parameter, controlling the weight degradation rate.

$d_c$ is the only variable in (3). The choice of $d_c$ is actually the choice of the average number of neighbors of all points in data set. Let $v = n_d \times (p/100)$, where $n_d = \binom{n}{2}$ and $p$ is a percentage. And $n$ denotes the number of points in data set.

In the code presented by Rodriguez and Laio, $d_c$ is define as

$$d_c = d_{\lceil \tau \rceil}, \tag{4}$$

where $d_{\lceil \tau \rceil} \in D = [d_1, d_1, \cdots d_{n_d}]$. $D$ is a set of all the distances between every two points in data set, which are sorted in ascending order. $\lceil \tau \rceil$ is the subscript of $d_{\lceil \tau \rceil}$, where $\lceil \cdot \rceil$ is the ceiling function.

The computation of $\delta_i$ is quite simple. The minimum distance between the point of $\mathbf{x}_i$ and any other points with higher density, denoted by $\delta_i$,

$$\delta_i = \begin{cases} \min\limits_{j:\rho_i < \rho_j} \left(\mathrm{dist}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right), if\ \exists j\ s.t. \rho_i < \rho_j \\ \max\limits_{j} \left(\mathrm{dist}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right), otherwise \end{cases} \tag{5}$$

When the local density and delta values for each point have been calculated, this method identifies the cluster centers by searching anomalously large parameters $\rho_i$ and $\delta_i$. On the basis of this idea, cluster centers always appear on the upper-right corner of the decision graph.

After cluster centers have been found, DPC assigns each remaining points to the same cluster as its nearest neighbors with higher density. A representation named as decision graph is introduced to help one to make a decision. This representation is the plot of $\delta_i$ as a function of $\rho_i$ for each point.