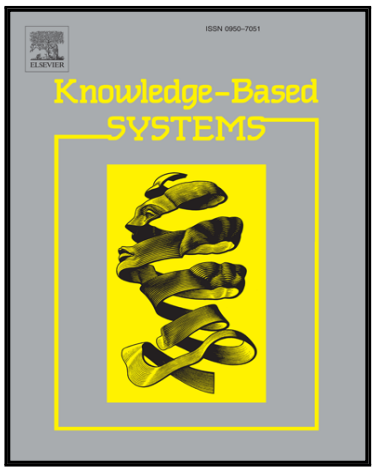

An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood

Shifei Ding , Mingjing Du , Tongfeng Sun , Xiao Xu , Yu Xue

PII: S0950-7051(17)30349-0
DOI: 10.1016/j.knosys.2017.07.027
Reference: KNOSYS 3991

To appear in: *Knowledge-Based Systems*

Received date: 27 October 2016
Revised date: 4 June 2017
Accepted date: 20 July 2017

Please cite this article as: Shifei Ding , Mingjing Du , Tongfeng Sun , Xiao Xu , Yu Xue , An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.07.027

An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood

Shifei Ding[1], Mingjing Du[1], Tongfeng Sun[1], Xiao Xu[1], Yu Xue[2]
[1](School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)
[2](School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Most clustering algorithms rely on the assumption that data simply contains numerical values. In fact, however, data sets containing both numerical and categorical attributes are ubiquitous in real-world tasks, and effective grouping of such data is an important yet challenging problem. Currently most algorithms are sensitive to initialization and are generally unsuitable for non-spherical distribution data. For this, we propose an entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood (DP-MD-FN). Firstly, we propose a new similarity measure for either categorical or numerical attributes which has a uniform criterion. The similarity measure is proposed to avoid feature transformation and parameter adjustment between categorical and numerical values. We integrate this entropy-based strategy with the density peaks clustering method. In addition, to improve the robustness of the original algorithm, we use fuzzy neighborhood relation to redefine the local density. Besides, in order to select the cluster centers automatically, a simple determination strategy is developed through introducing the γ-graph. This method can deal with three types of data: numerical, categorical, and mixed type data. We compare the performance of our algorithm with traditional clustering algorithms, such as K-Modes, K-Prototypes, KL-FCM-GM, EKP and OCIL. Experiments on different benchmark data sets demonstrate the effectiveness and robustness of the proposed algorithm.
Keywords: Entropy; Density peaks clustering; Mixed type data; Fuzzy neighborhood.

1. Introduction

Clustering analysis is aimed at finding correlations within subsets of the dataset and assessing similarity among elements within these subsets [1-2]. Clustering has many applications in various domains including biology, economics and medicine. Its applications include data mining, document retrieval, image segmentation, and pattern classification [3-4]. Traditional clustering methods, e.g., K-Means [5], can only handle numerical values. Nevertheless, in some real world applications, one has to deal with features, such as gender, color, and type of disease that are categorical attributes. In other words, data sets containing both numerical and categorical attributes are ubiquitous in real-world tasks. Designing an effective clustering algorithm for this type of data is a challenging problem. For convenience, we use the "mixed type" data to denote this type of data with numerical and categorical attributes in this paper. But the mixed type data may contain ordinal attribute or some other attributes in other literatures.

A straightforward way to deal with mixed type data has a pre-processing that is able to convert categorical attributes to new forms, e.g. the binary strings, and then apply the aforementioned numerical value based clustering methods. However, binary encoding has three drawbacks. First and foremost, this method destructs the original structure of categorical attributes. In other words, transformed binary attributes are meaningless and their values are hard to interpret [6]. Second, if the domain of a categorical attribute is large, then transformed binary attributes will have a much larger dimensionality. The last disadvantage is the difficult of maintenance. If an attribute value is added into a categorical attribute, all of the objects will be changed. In order to better solve the problem, numerous researchers study on clustering based on similarity metrics dealing with categorical values directly, during the last decade. Based on a similarity (or dissimilarity) metric that takes into account both numeric and categorical attributes, some methods, e.g., K-prototypes (KP) [7] and its variations which are applicable to numerical and categorical data are presented. In order to circumvent parameter adjustment between categorical and numerical values, some works, e.g., Similarity-Based Agglomerative Clustering (SBAC) algorithm [6], based on a new similarity metric for mixed type data, are presented. However, SBAC is high computational complexity. Some methods based on a parameter-free similarity metric, e.g. OCIL [8], are proposed. But this metric only can measure the similarity between an object and a cluster. And like KP and its variations, OCIL uses the K-Means paradigm to cluster mixed type data and is an iterative clustering algorithm. Hence, such kind of algorithms is sensitive to initialization and is more suitable for spherical distribution data.