Data & Knowledge Engineering 69 (2010) 881–900

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

New fuzzy *c*-means clustering model based on the data weighted approach

Chenglong Tang *, Shigang Wang, Wei Xu

School of Mechanical and Dynamical Engineering of Shanghai Jiao Tong University, No.800 Dong Chuan Road, Minhang District, Shanghai 200240, PR China

ARTICLE INFO

Article history:
Received 5 July 2009
Received in revised form 5 May 2010
Accepted 24 May 2010
Available online 4 June 2010

Keywords:
Fuzzy clustering
Data weighted approach
Exponent impact factor
Influence exponent
Outliers mining

ABSTRACT

This paper proposes a new kind of data weighted fuzzy *c*-means clustering approach. Different from most existing fuzzy clustering approaches, the data weighted clustering approach considers the internal connectivity of all data points. An exponent impact factors vector and an influence exponent are introduced to the new model. Together they influence the clustering process. The data weighted clustering can simultaneously produce three categories of parameters: fuzzy membership degrees, exponent impact factors and the cluster prototypes. A new fuzzy algorithm, DWG-K, is developed by combining the data weighted approach and the G-K. Two groups of numerical experiments were executed. Group 1 demonstrates the clustering performance of the DWG-K. The counterpart is the G-K. The results show the DWG-K can obtain better clustering quality and meanwhile it holds the same level of computational efficiency as the G-K holds. Group 2 checks the ability of the DWG-K in mining the outliers. The counterpart is the well-known LOF. The results show the DWG-K has considerable advantage over the LOF in computational efficiency. And the outliers mined by the DWG-K are global. It was pointed out that the data weighted clustering approach has its unique advantages when mining the outliers of the large scale data sets, when clustering the data set for better clustering results, and especially when these two tasks are done simultaneously.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Artificial intelligence research and application involve a number of sub-areas. This paper discusses two of them: cluster-based pattern recognition and outlier mining. For a given data set, cluster-based pattern recognition refers to dividing the data set into several patterns using the cluster methods. Outlier mining refers to finding the abnormal data points in the data set and mining the information that they contain [1]. On one hand, cluster-based pattern recognition and outliers mining hold close connectivity; the outliers in the data set must be detected and appropriately processed, for example, replaced by the normal points or directly eliminated when necessary. The goal is to reduce their negative influences on the results of the cluster-based pattern recognition [2]. On the other hand, the treatment of the outliers is obviously different for cluster-based pattern recognition and outlier mining. For the cluster-based pattern recognition, outliers are usually regarded as "harmful", the main measures taken here are to minimize or eliminate this harm and the outliers are detected as the by-products of the clustering. While for the outlier mining, the outlier itself becomes the focus. The main task is to mine, not only to detect the outliers. In addition, the methods used in these two areas are different under most circumstances. For the outlier mining, the main methods are based on statistics, density, distance, and feature deviation [1]. Cluster-based methods have been reported but they are not very popular. In reality, there is a common demand that can be summarized as follows: for a given data set which contains a certain number of outliers, the analysis of the data set simultaneously involves three tasks. The first is to cluster the data set. The data set is clustered into several groups, and then the belonging of each data point to the prototypes is subsequently determined. The second is to establish the classifiers. The classifiers are established by the cluster prototypes. Lastly, to mine the outliers. This includes detecting the outliers and

* Corresponding author.
E-mail addresses: tang_chenglong@hotmail.com, tang_chenglong@sjtu.edu.cn (C.L. Tang).

0169-023X/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2010.05.001

Symbols and abbreviations

c	The number of clusters
$\boldsymbol{d_{ij}}$	Distance between $\boldsymbol{x_j}$ and $\boldsymbol{v_i}$
k	The number of outliers in the data set
m	Fuzzy exponent for $\boldsymbol{u}_{ij}$
n	The number of data points
q	The number of features
s	Influence exponent
$\boldsymbol{t}_j$	Weight of $\boldsymbol{x_j}$ under data weighted approach
$\boldsymbol{t_{ij}}$	Typicality value
$\boldsymbol{u}_{ij}$	Fuzzy membership degree
$\boldsymbol{v_i}$	The ith prototype
$\boldsymbol{w}$	Weights
$\boldsymbol{x_j}$	The jth data
$\boldsymbol{E}$	Exponent impact factor vector
$\boldsymbol{J}$	Objective function
$\boldsymbol{T}$	Typicality matrix
$\boldsymbol{U}$	Membership degree matrix
$\boldsymbol{V}$	Prototype matrix
$\boldsymbol{X}$	Data set
CMP	Compactness of fuzzy partition
SPT	Separation of fuzzy partition
EVA	***CMP/SPT***
$\boldsymbol{EIF(j)}$	Exponent impact factor of $\boldsymbol{x}_j$
$\boldsymbol{LOF}(\boldsymbol{x}_j)$	Local outlying degree under the LOF
$\boldsymbol{O}_{\boldsymbol{DWF}}(\boldsymbol{x}_j)$	Outlying degree under DWG-K
ε	Convergence values
φ	Lagrange multiplier operator
η	Fuzzy exponent for $\boldsymbol{t_{ij}}$

discovering the information that they contain. Currently, these tasks are solved in different areas. In real applications, these three tasks are often expected to be solved simultaneously. Little research about solving these three tasks simultaneously has been reported.

In order to solve the above problems, this paper proposes a new fuzzy clustering approach, which is called data weighted fuzzy clustering approach. The core idea of this novel approach is the nature of each data point in the data set is "different" from one another, and internal connectivity exists among all data points. A constraint which describes the internal connectivity is given in the data weighted fuzzy clustering approach. A set of exponent impact factors and an influence exponent are introduced to the novel objective function. Together they influence the clustering process and realize the goal of treating each data point differently. Because the data weighted clustering approach considers the internal connectivity of all data points, it holds a strong ability to handle the outliers. When the data weighted approach is used to cluster the data set, not only can it get much better clustering qualities than the existing clustering models do, it can also effectively detect the outliers and easily mine information related to the outliers. In contrast, most existing fuzzy clustering models neglect the internal connectivity of all data points. They treat them equally in the process of clustering. This paper gives the theoretical model of the data weighted fuzzy clustering approach, and numerical experiments are given to verify the performance of the data weighted fuzzy clustering approach in clustering and mining outliers, particularly when they are done simultaneously.

This paper is presented as follows: Section 2 reviews related work carried out on the existing fuzzy clustering and outlier mining. Section 3 describes the data weighted fuzzy clustering approach. First, the mathematical model and some update equations are reasoned. Second, a conventional fuzzy clustering algorithm, Gustafson–Kessel (hereinafter, in short, G-K), is introduced and is combined with the proposed data weighted approach, as a result, the data weighted G-K algorithm, DWG-K for short, is developed. Section 4 tests and verifies the DWG-K by the numerical experiments on two real data sets. The experiments have been divided into two groups, group one verifies the clustering performance of the DWG-K, compared with the G-K. Group two verifies the ability of the DWG-K in mining the outliers, compared with the LOF. Section 5 discusses the roles of two new parameters: the exponent impact factors and the influence exponent. Section 6 concludes the research results.

2. Related work

Cluster analysis is one of several important tools in modern data analysis. The general philosophy of cluster analysis is to divide the data set into several homogeneous groups. Division is based on similarity or dissimilarity. The objects in the same group tend

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات