Knowledge-Based Systems 104 (2016) 156–164

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A weight-incorporated similarity-based clustering ensemble method based on swarm intelligence

Qi Kang[a,*], ShiYao Liu[a], MengChu Zhou[b,c,*], SiSi Li[d]

[a] *Department of Control Science and Engineering, Tongji University, Shanghai 201804, China*
[b] *Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA*
[c] *Renewable Energy Research Group, King Abdulaziz University, Jeddah, Saudi Arabia*
[d] *Department of Mathematics and Computer Sciences, Mercy College, Dobbs Ferry, NY 10522, USA*

ARTICLE INFO

Article history:
Received 22 October 2015
Revised 30 March 2016
Accepted 20 April 2016
Available online 22 April 2016

Keywords:
Data clustering
Clustering ensemble
Swarm intelligence
Optimization

ABSTRACT

Clustering methods play an important role in data mining and various other applications. This work investigates them based on swarm intelligence. It proposes a new clustering method by combining K-means clustering method and mussels wandering optimization algorithm. A single cluster method is well recognized to achieve limited performance when it is compared with a clustering ensemble (CE) that integrates several single ones. Hence, this work introduces a new CE method called weight-incorporated similarity-based CE. The commonly-used datasets with varying size are used to test the performance of the proposed methods. The simulation results illustrate the validity and performance advantages of the proposed ones over some of their peers.

1. Introduction

Clustering analysis is fundamentally important to further data analysis and processing. Clustering methods divide a given dataset into certain classes according to the data similarity, thus enabling the element of every class to have the same or similar characteristics and keeping the distance among data items in different classes as large as possible [1]. They have received growing attention due to their role in data mining research and applications [2–6], such as information retrieval, image segmentation, pattern recognition, web analysis, text mining, and visualization.

The traditional clustering algorithms, for example, K-means clustering method [7,8], called K-means for short, have some drawbacks in searching for optimal solutions. To overcome the drawbacks and limitations of a single clustering algorithm, researchers have introduced swarm intelligence to clustering analysis. Swarm intelligence is a promising computing technology especially when used for data analysis, since it inherits effective processing mechanisms and desired characteristics of biological systems. It has borrowed some proven good ideas in the area of biological evolution. The existing researches show that swarm intelligence is a promising method, and can be used to solve global optimization problems effectively. Because of its advantages, it is deeply studied and has achieved many important research results. For instance, the combination of particle swarm optimization (PSO) with traditional clustering algorithms, i.e., K-means and C-means clustering methods, yields better clustering methods [9–11]. Clustering algorithms based on swarm intelligence have rapid convergence and great adaptability. Moreover, in comparison with the traditional clustering methods, they can reveal the correlations among data better, thus greatly improving the clustering quality and efficiency. New clustering algorithms have emerged in pattern recognition, neurocomputing, big data analytic, etc. [12–22].

In this paper, along with the above line, we propose a new algorithm based on K-means and mussels wandering optimization (MWO), called K-MWO for short. MWO is a new effective global optimization algorithm [23]. It aims to reach an optimal solution by mathematically modeling mussels' leisurely locomotion behavior when they form their bed pattern in a habitat. MWO is a simple and easy-to-use algorithm. Its details and effectiveness by simulation experiments are given in [23]. It has relatively better results compared with four well-known algorithms according to the experiments in [23], i.e., genetic algorithm, particle swarm optimization (PSO), biogeography-based optimization, and group search optimization when solving many large-scale optimization problems [24–27]. Considering its effective global optimization ability, we aim to develop MWO for better clustering analysis and clustering algorithms. We combine K-means [7,28,29] with MWO to propose K-MWO to overcome the shortcomings of K-means.

[*] Corresponding authors. Tel.: +19735966282.
E-mail addresses: qkang@tongji.edu.cn (Q. Kang), sherryliuya@126.com (S. Liu), zhou@njit.edu (M. Zhou), sophieandli@gmail.com (S. Li).

http://dx.doi.org/10.1016/j.knosys.2016.04.021
0950-7051/© 2016 Elsevier B.V. All rights reserved.

Although many clustering algorithms are efficient in dealing with specific problems, every one of them has its own limitations, for example, a clustering result is sensitive to parameters and initialization; most of them are unreliable in determining a true number of clusters. They may produce different results on the same dataset. No clustering algorithm is applicable to various structures and different types of datasets. Clustering ensemble (CE) is recognized to be an important way to address the above problems [30–33].

CE uses an ensemble technology to produce a new clustering result by integrating several clustering results obtained from different clustering methods or the same method with different parameters. It is generally called the solution of a consensus problem. Compared with a single clustering algorithm, a CE method has the following advantages: it can improve the quality of clustering results, cluster a dataset with a categorical attribute, and detect and handle isolated points and noises. It can also deal with distributed data sources and process the data in parallel. Therefore, it is significant to study CE in order to achieve better clustering results.

CE methods can be divided into four categories: similarity, graph, relabeling and transformation-based methods [34]. Fred proposed a clustering method based on a co-association matrix. Fred and Jain improved the above algorithm in [35]. They calculated the ratio of the number of times a pair of data points that were clustered in the same cluster to the total number of cluster members. The ratio is considered as the similarity measure between pairs of data points. Strehl and Ghosh proposed three CE methods based on graph partitioning: cluster-based similarity partitioning algorithm, meta-clustering algorithm and hyper-graph partitioning algorithm. Dudoit and Fridlyand unified the class labels produced by every cluster member and conducted CE via voting in [36]. In [35], a novel method called selective spectral CE is proposed, which achieves a better clustering result if only a part of all available clustering results are combined. The results are generated by spectral clustering in which a novel nearest neighbor rule-based selection strategy is introduced to choose and build an ensemble committee from a part of the promising. In [37], a weighted co-clustering based CE is proposed. It assigns a confidence score to each partition in the ensemble and compute weighted co-association for each pair of objects. In [38], a new selective CE algorithm selects the best reference partition based on clustering validity evaluation and uses a new selection strategy and the method of a member's weight. In [39], a coupled CE is proposed. It not only considers but also integrates the coupling relationships between base clustering and objects. In [30], Domeniconi and AL-Razgan propose two CE methods to form the clustering ensemble of weighted cluster results produced by a locally adaptive clustering algorithm [40] with different parameters. In [31], a weighted CE algorithm based on graphs first clusters the datasets to obtain cluster members. It then sets weights for each data object with a proposed ensemble function, and determines the relationship between a data-pair by setting weights to the edges between them, thereby obtaining a weighted nearest neighbor graph. It finally performs clustering based on graph theory. As indicated in [41], more and more CEs appear and can be very useful to solve a problem at hand.

In these CE methods, the weight information has its own updating strategy. Yet none has used the boosting idea to do so. When thinking about the CE progress, we can obtain some information of each data to find if it is suitable to use boosting. This work represents the first try to adopt the boosting idea to decide the weights used in CE. Our proposed method is based on a similarity-based ensemble one. In [42], a Selective Spectral Clsutering (SELSCE) is proposed, which introduce a novel selection strategy to choose the promising committee. In this work, we present a weight updating strategy.

This paper presents a new similarity-based clustering method based on weight information. Different from the existing weighted ensemble methods, the proposed one is a similarity-based one in which each data entry in a dataset is assigned with a weight. Ensemble is complete in only several iterations, which suffices in terms of the clustering performance. When performing ensemble, by means of a similarity matrix form, the data clustered in the same class in different clustering results are found. Then we reduce their weights, cluster again by the updated weight and finally obtain new clustering results. There is no weight information to the clustering methods as input. Weight calculation is human introduction and shows good application. The method absorbs some of the ideas of Adaboost [33]. In an iterative process, data that can be easily classified are weakened; data easy to be misclassified or hard to be classified are focused on, thus leading to significantly better results.

This paper is organized as follows: In Section 2, K-means is introduced first as a basic yet widely-used clustering method. K-PSO is then introduced from which the idea of combining K-means and PSO is borrowed. Finally, we propose K-MWO as a new clustering method. Section 3 gives the description of CE methods with its emphasis on a similarity-based one and our newly proposed weight-incorporated similarity-based clustering ensemble (WSCE). Their input algorithms are selected to be K-MWO, K-means, and K-PSO. Section 4 analyzes the simulation results of the proposed method and others by using datasets with distinct characteristics and size. The paper is concluded in Section 5.

2. Single clustering algorithm

2.1. K-means

The K-means algorithm [7] is a single iterative clustering algorithm that partitions a given dataset into a user-specified number of clusters, K. It is simple to implement and run, relatively fast, easy to adapt, and common in practice. Via a clustering algorithm, data are grouped by some notion of “closeness” or “similarity”. In K-means, the default measure of closeness is Euclidean distance. It attempts to minimize the following nonnegative cost function:

$$\text{Cost} = \sum_{i=1}^{N} \left(\arg\, min_j \|X_i - C_j\|_2^2\right) \tag{1}$$

where N is the population size of a dataset, X_i is the ith data point in the dataset and C_j is the closest cluster center to X_i. Eq. (1) is often referred to as the K-means objective function.

The K-means works as follows:

Step 1: Initialize the cluster centers by picking k points in R^d randomly.
Step 2: Assign each data point to its closest representative $C_i, i \in \mathbf{N}_\mathrm{K} = \{1, 2, \ldots, \mathrm{K}\}$.
Step 3: Update the cluster center such that C_j is the mean of data in the jth cluster.
Step 4: Examine the objective function, finish if a termination condition is met, or go back to Step 2.

It has some drawbacks: (1) it is very sensitive to its initial value as different ones may lead to different solutions; and (2) it is based on an objective function simply and usually solves the extreme value problem by a gradient method [43].

2.2. K-PSO

As an optimization algorithm, initially, a particle swarm optimization (PSO) algorithm is used for optimization of the continuous space system. Mathematically, PSO is described as follows.