Expert Systems With Applications 76 (2017) 59–70

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A novel clustering algorithm based on data transformation approaches

Rasool Azimi[a], Mohadeseh Ghayekhloo[a], Mahmoud Ghofrani[b,*], Hedieh Sajedi[c]

[a] Young Researchers and Elite Club, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[b] School of Science, Technology, Engineering and Mathematics (STEM), University of Washington, Bothell, USA
[c] Department of Computer Science, College of Science, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 28 November 2015
Revised 29 October 2016
Accepted 24 January 2017
Available online 25 January 2017

Keywords:
Data mining
Clustering
K-means
Data transformation
Silhouette
Transformed K-means

ABSTRACT

Clustering provides a knowledge acquisition method for intelligent systems. This paper proposes a novel data-clustering algorithm, by combining a new initialization technique, K-means algorithm and a new gradual data transformation approach to provide more accurate clustering results than the K-means algorithm and its variants by increasing the clusters' coherence. The proposed data transformation approach solves the problem of generating empty clusters, which frequently occurs for other clustering algorithms. An efficient method based on the principal component transformation and a modified silhouette algorithm is also proposed in this paper to determine the number of clusters. Several different data sets are used to evaluate the efficacy of the proposed method to deal with the empty cluster generation problem and its accuracy and computational performance in comparison with other K-means based initialization techniques and clustering methods. The developed estimation method for determining the number of clusters is also evaluated and compared with other estimation algorithms. Significances of the proposed method include addressing the limitations of the K-means based clustering and improving the accuracy of clustering as an important method in the field of data mining and expert systems. Application of the proposed method for the knowledge acquisition in time series data such as wind, solar, electric load and stock market provides a pre-processing tool to select the most appropriate data to feed in neural networks or other estimators in use for forecasting such time series. In addition, utilization of the knowledge discovered by the proposed K-means clustering to develop rule based expert systems is one of the main impacts of the proposed method.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

The expert systems are computer applications that contain stored knowledge and are developed to solve problems in a specific field in almost the same way in which a human expert would (Shuliang, Barry, Edwards, Kinman, & Duan, 2002). Acquisition of the expert knowledge is a challenge for developing such expert systems (Yang, Li, & Qian, 2012). One of the major problems and most difficult tasks in developing rule based expert systems is representing the knowledge discovered by data clustering (Markic & Tomic, 2010). The K-means algorithm is one of the most commonly used clustering techniques, which uses the data reassignment method to repeatedly optimize clustering (Lloyd, 1982). The main goal of clustering is to generate compact groups of objects or data that share similar patterns within the same cluster, and isolate these groups from those which contain elements with different characteristics.

Although the K-means algorithm has features such as simplicity and high convergence speed, it is totally dependent on the initial centroids which are randomly selected in the first phase of the algorithm. Due to this random selection, the algorithm may converge to locally optimal solutions (Celebi, Kingravi, & Vela, 2013). Different variants of K-means algorithm have been proposed to address this limitation. The K-Medoids algorithm was proposed in Kaufman and Rousseeuw (1987) to define each cluster by the most central medoid in which it is located. First, K data are considered as initial centroids (medoid) and then each data is assigned to the closest Medoid, and the initial clusters are formed. In an iteration-based process, the most central data in each cluster is considered as the new centroid and each data is assigned to the nearest centroid. The remaining steps of this algorithm match the K-means procedure. Fuzzy C-means (FCM) clustering introduced the partial membership concept (Dunn, 1973; Bezdek, Ehrlich, & Full, 1984). In fact, in the FCM algorithm, each data belongs to all clusters. The degree of belonging is represented by a partial

* Corresponding author at: UWBB room 227, 18807 Beardslee Blvd, Bothell, WA 98011, USA
E-mail addresses: r.azimi@qiau.ac.ir (R. Azimi), m.ghayekhloo@qiau.ac.ir (M. Ghayekhloo), mghofrani@uwb.edu (M. Ghofrani), hhsajedi@ut.ac.ir (H. Sajedi).

http://dx.doi.org/10.1016/j.eswa.2017.01.024
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

membership determined by a fuzzy clustering matrix. A genetic algorithm-based K-means (GA-K-means) was proposed in Krishna and Murty (1999) to provide global optimum for the clustering. In this method, the K-means algorithm was used as a search operator instead of crossover. A biased mutation operator was also proposed for clustering to help the K-means algorithm to avoid local minima. Global K-means algorithm was developed in Likas, Vlassis, and Verbeek (2003) to provide experimentally optimal solution for clustering problems. However, it is not appropriate for clustering medium-sized and large-scale datasets due to its heavy computational burden. K-means++ initialization algorithm was proposed in Arthur and Vassilvitskii (2007) for obtaining an initial set of centroids that is near-optimal. The main drawback of the K-means++ is its inherent sequential nature, which limits the effectiveness of the method for high-volume data. An artificial bee colony K-means (ABC-K-means) clustering approach was proposed in Zhang, Ouyang, and Ning (2010) for optimal partitioning of data objects into a fixed number of clusters. A hybrid of differential evolution and K-means algorithms named DE-K-means was introduced in Kwedlo (2011). The differential evolution algorithm was used as a global optimization method and the resultant clustering solutions were fine-tuned and corrected using the K-means algorithm. Dogan, Birant, and Kut (2013) proposed a hybrid of K-means++ and self-organizing map (SOM) (Kohonen, 1990) algorithm to improve the clustering accuracy. It first uses K-Means++ initialization method to determine the initial weight values and the starting points, and then uses SOM to find an appropriate final clustering solution. However, the aforementioned limitation of the K-means++ was not addressed. A new clustering technique using a combination of the global K-means algorithm and the topology neighborhood based on Axiomatic Fuzzy Sets (AFS) theory was developed in Wang, Liu, and Mu (2013) to determine initial centroids. A new clustering algorithm, named K-means*, was presented in Malinen, Mariescu-Istodor, and Fränti (2014) that generates an artificial dataset X^* as the input data. The input data are then mapped one-by-one to the generated artificial data ($X \rightarrow X^*$). Next, the inverse transformation of the artificial data to the original data is performed by a series of gradual transformations. To do so, the K-means algorithm updates the clustering model after each transformation and moves the data vectors slowly to their original positions. The K-means* algorithm uses a random data swapping strategy to deal with the problem of generating empty clusters. However, the random selection of the data vectors as the cluster centroids may reduce the other clusters' coherence and decrease the efficiency of the K-means* algorithm. Moreover, the convergence rate of the K-means* algorithm significantly reduces as the number of clusters increases, especially with increasing data volumes. Density based clustering methods were proposed in Mahesh Kumar and Rama Mohan Reddy (2016) to speed up the neighbor search for clustering spatial databases with noise. Density Based Spatial Clustering of Applications with Noise (DBSCAN) provides a graph based index structure for high dimensional data with large amount of noise. It was shown that running time of the proposed method is faster than DBSCAN with exactly same clustering results. The proposed method solved the inefficacy of the DBSCAN method to work with clusters with large differences in densities. A novel clustering algorithm named CLUB (CLUstering based on Backbone) was developed in Chen et al. (2016) to determine optimal clusters. First, the algorithm detects the initial clusters and finds their density backbones. Then, the algorithm finds out the outliers in each cluster based on K Nearest Neighbor (KNN) method. Finally by assigning each unlabeled point to the cluster with the nearest higher density neighbor, the algorithm yields the final clusters.

CLUB has several drawbacks: The KNN method lacks an efficient algorithm to determine the value of parameter K (number of nearest neighbors); Computational cost of this method is too high because it requires calculating distance of each instance query with respect to all training samples. Two particle swarm optimization (PSO) based fuzzy clustering methods were proposed in Silva Filho, Pimentel, Souza, and Oliveira (2015) to deal with the shortcomings of the PSO algorithms used for fuzzy clustering. The proposed methods adjust parameters of PSO dynamically to achieve a balance between exploration and exploitation to avoid trapping in local optimum. The proposed methods lack precision for high-dimensional applications. In addition, the iterative process of the proposed methods significantly decreases the convergence rate. Generally, the speed at which a convergent sequence approaches its limit is defined as the rate of convergence. Three clustering algorithms named Near Neighbor Influence (CNNI), an improved version of time cost of Near Neighbor Influence (ICNNI), and a variation of Near Neighbor Influence (VCNNI) were presented in Chen (2015). The clustering results showed that ICCNNI is faster than CNNI and also, CNNI requires less space than VCNNI. These methods suffer from large scale computing and storage requirements. A growing incremental self-organizing neural network (GISONN) was developed in Liu and Ban (2015) to select appropriate clusters by learning data distribution of each cluster. The proposed method is however not applicable for large-volume or high-dimensional datasets due to its computational complexity (Abdul Nazeer and Sebastian, 2009). In addition, the neighborhood preserving feature of the algorithm is violated when the output space topology does not match with the structure of the data in the input space.

In spite of the improved performance of the K-means variants for synthetic datasets with Gaussian distribution, their performance on real datasets is neither very promising nor different from the original K-means algorithm. In addition, all K-means based algorithms lack an efficient method to determine the optimal number of clusters. This requires the user to determine the number of clusters either arbitrarily or based on practical and experimental estimates, which might not be optimal.

In this paper, we propose a novel clustering approach called transformed K-means to provide more accurate clustering results compared to the K-means algorithm and its improved versions. The proposed clustering method combines a new initialization technique, K-means algorithm and a new gradual data transformation approach to appropriately select the initial cluster centroids and move the real data into the locations of the initial cluster centroids that are closer to the actual positions of the associated data. By doing this, the data are placed in an artificial structure to properly initiate the K-means clustering. The inverse transformation is then performed to gradually move back the artificial data to their original places. During this process, K-means updates the clustering centroids after any changes in the data structure. This provides more optimal clustering results for both synthetic and real datasets. In addition, the proposed data transformation solves the empty cluster problem of K-means algorithm and its improved versions. An efficient method based on the principal component transformation and a modified silhouette algorithm is also proposed in this paper to determine the optimal number of clusters for the K-means algorithms.

The proposed clustering method develops a rule-based expert system by means of knowledge acquisition through data transformation. Significances of the proposed method include addressing the limitations of the K-means based clustering and improving the accuracy of clustering as an important method in the field of data mining and expert systems. The proposed method can be used for intelligent system applications such as forecasting time-series including solar, wind, load and stock market series.

Contributions of the paper are outlined as follows:

1. A new initialization technique is proposed to select initial centroids which are closer to the optimum centroids' locations.

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