Robust Clustering with Distance and Density

Hanning Yuan, School of Software, Beijing Institute of Technology, Beijing, China
Shuliang Wang, School of Software, Beijing Institute of Technology, Beijing, China
Jing Geng, Beijing Institute of Technology, Beijing, China
Yang Yu, Beijing Institute of Technology, Beijing, China
Ming Zhong, Beijing Institute of Technology, Beijing, China

ABSTRACT

Clustering is fundamental for using big data. However, AP (affinity propagation) is not good at non-convex datasets, and the input parameter has a marked impact on DBSCAN (density-based spatial clustering of applications with noise). Moreover, new characteristics such as volume, variety, velocity, veracity make it difficult to group big data. To address the issues, a parameter free AP (PFAP) is proposed to group big data on the basis of both distance and density. Firstly, it obtains a group of normalized density from the AP clustering. The estimated parameters are monotonically. Then, the density is used for density clustering for multiple times. Finally, the multiple-density clustering results undergo a two-stage amalgamation to achieve the final clustering result. Experimental results on several benchmark datasets show that PFAP has been achieved better clustering quality than DBSCAN, AP, and APSCAN. And it also has better performance than APSCAN and FSDP.

KEYWORDS

Clustering, Density, Distance, Images, Parameter Free Affinity Propagation (PFAP)

1. INTRODUCTION

To take advantage of big data, clustering is a fundamental process of dividing dataset into a series of subsets (Rodriguez & Laio, 2014; Seife, 2015). The effective issues of a clustering algorithm may include scalability to large data volume and dimensions, identifying clusters with arbitrarily shapes, and handling data with noises (Aggarwal & Reddy, 2014; Wang et al., 2011).

However, the drawbacks of existent clustering algorithms are distinct. For example, AP (affinity propagation) (Dueck & Frey, 2007) replies on data centers, which makes it difficult to deal with dataset of clusters of arbitrary shapes. Thus, it is bad at the time-complexity, and unable to deal with non-convex clusters. DBSCAN (density-based spatial clustering of applications with noise) (Ester et al., 1996) forces users to specify or adjust the parameters (e.g. the number of clusters, the density of clusters) under their professional knowledge. If there are human misinterpretation of data and errors, the algorithm might be ineffective against the practical scenarios, where it is impossible to choose the “best” set of parameters. In order to overcome the algorithmic drawbacks, APSCAN (affinity propagation spatial clustering of applications with noise) tries to integrate AP and DBSCAN, but it is only suitable for data set with clusters in specific density distribution.

DOI: 10.4018/IJDWM.2017040104
To leverage the performance of APSCAN and further get a more genic algorithm, a parameter free AP is proposed in this paper, to take full advantage of both AP and DBSCAN for grouping data sets under the density. The rest of this paper will be organized as the following. Based on the related work in Section 2, Section 3 gives the principles. Then Section 4 perform the experiments and their analysis. Section 5 finally concludes the work.

2. RELATED WORK

AP is a distance-based algorithm for identifying exemplars in a dataset by imitating the message passing and feedback routine between the data objects (Dueck & Frey, 2007; Frey & Dueck, 2007). It enjoys lower error than traditional methods, which is computationally efficient in many applications (Dueck & Frey, 2007; Dueck et al., 2008). The measurements of the mutual similarity among \( N \) objects are recorded in an input matrix of \( N \times N \). The diagonal of the matrix, \( p(k,k) \), is treated as the reference for the data object \( k \) to become the cluster center. The responsibility \( r(i,k) \) that is sent from data object \( i \) to the candidate clustering center \( k \), indicates how suitable an object \( k \) can be used as a cluster center for the object \( i \). The availability \( a(i,k) \) that is sent from the candidate cluster center \( k \) to the data object \( i \), reflects how likely the object \( i \) chooses \( k \) as its cluster center. The larger the value of \( r(i,k) \) and \( a(i,k) \), the higher the probability that object \( k \) is to become the cluster center. Consequently, increase the chance that an object \( i \) belongs to a cluster with its center at object \( k \). During this iterative process, AP keeps updating \( r(i,k) \) and \( a(i,k) \) between the data objects until the predefined convergence criteria is met. The AP parameters adhere to what are used in its original settings, for example, a maximum of iterations is 1000, the upper limit of steady times is 100, and the damping coefficient is 0.9. The reference of clustering center is chosen to be the median value of similarity matrix.

To improve the performance of AP, some researchers attempted. For example, to address the low-efficiency in information spreading amongst the N×N dimension edges, Fujiwara, Irie, and Kitahara proposed to prune the unnecessary information exchange during the iteration process, and to calculate the convergence information after clusters are created (Yasuhiro et al., 2011). Under the scenarios where there ought to be a required number of clusters, AP may become slow. To improve its efficiency under this circumstance, Wang and Zheng presented a fast algorithm on multi-grid that the calling times of AP was reduced by using multi-grid search, and the searching scope was reduced by improving the upper bound of preference parameter (Wang & Zheng, 2010). It could largely enhance the speed of AP under the fixed clusters numbers. To achieve the improvement over fixed damping factor which has negative impact on efficiency and convergence, Liu and Fu put forward a clustering algorithm F-AP that a constriction factor was given to speed up the convergence by dynamically adjusting the convergence coefficient along with the process of AP algorithm, and simultaneously ensured the same clustering results with original one from traditional AP (Liu & Fu, 2011). Apart from the performance, the quality of clustering results is another fundamental issue. To overcome its inability of obtaining ideal clustering from datasets that are non-convex and uneven distributed, Feng and Yu firstly mapped datasets nonlinearly to a high dimensional space by using Kernel methods (Feng & Yu, 2012). Secondly, according to the similarity magnanimity method shared by the nearest objects, a non-sensitive AP algorithm, DIS-AP was presented to improve the traditional AP in terms of addressing data which is non-convex and in high dimension.

APSCAN (affinity propagation spatial clustering of applications with noise) uses DDBSCAN (Double-Density-Based SCAN), which emphasizes the central density must be higher than the marginal density. The features were grouped by using double-density and non-parametric clustering based on AP algorithm, and then its density parameters of the clusters were normalized (Chen et al., 2011). This dual density-clustering framework is applied to the multiple density clustering before
Related Content

A Hyper-Heuristic for Descriptive Rule Induction
[www.igi-global.com/article/hyper-heuristic-descriptive-rule-induction/1778?camid=4v1a](www.igi-global.com/article/hyper-heuristic-descriptive-rule-induction/1778?camid=4v1a)

Handling Local Patterns in Collaborative Structuring
Ingo Mierswa, Katharina Morik and Michael Wurst (2008). *Successes and New Directions in Data Mining* (pp. 167-186).
[www.igi-global.com/chapter/handling-local-patterns-collaborative-structuring/29959?camid=4v1a](www.igi-global.com/chapter/handling-local-patterns-collaborative-structuring/29959?camid=4v1a)
A Clustering Rule Based Approach for Classification Problems
Philicity K. Williams, Caio V. Soares and Juan E. Gilbert (2012). *International Journal of Data Warehousing and Mining* (pp. 1-23).
[www.igi-global.com/article/clustering-rule-based-approach-classification/61422?camid=4v1a](www.igi-global.com/article/clustering-rule-based-approach-classification/61422?camid=4v1a)

Semantics-Aware Advanced OLAP Visualization of Multidimensional Data Cubes
[www.igi-global.com/article/semantics-aware-advanced-olap-visualization/1791?camid=4v1a](www.igi-global.com/article/semantics-aware-advanced-olap-visualization/1791?camid=4v1a)