

Title	大規模カテゴリーカル混合データセットのためのクラスタリングアルゴリズムへの局所性鋭敏ハッシュの組み込み手法
Author(s)	Nguyen, Mau Toan
Citation	
Issue Date	2021-12
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/17603
Rights	
Description	Supervisor:Huyhn Nam Van, 先端科学技術研究科, 博士

氏名	NGUYEN, Mau Toan		
学位の種類	博士 (情報科学)		
学位記番号	博情第 464 号		
学位授与年月日	令和 3 年 12 月 24 日		
論文題目	Incorporating the Locality Sensitive Hashing Technique into Clustering Algorithms for Massive Categorical and Mixed Datasets		
論文審査委員	主査	HUYNH Van Nam	北陸先端科学技術大学院大学 教授
		金子 峰雄	同 教授
		田中 清史	同 教授
		NGUYEN Le Minh	同 教授
		本多 克宏	大阪府立大学 教授

論文の内容の要旨

Cluster analysis has been an object of research since the 1980s for finding the natural groups of objects in the data so that similar objects stay within the same clusters while different objects stay in different clusters. It is undoubted that cluster analysis is important for a wide range of scientific and industrial processes such as data mining, computer vision, signal processing, and census research. k -means-like algorithms are the most used unsupervised machine learning techniques for handling such problems of cluster analysis. In the context of a k -means-like algorithm, a proper structure for representing the clusters and an appropriate distance measure for measuring the distance between objects and clusters must be accordingly defined. In general, a k -means-like algorithm seeks for the optimal clusters that can minimize the total distance from all objects to their nearest clusters, which makes the cluster representation and distance measure become very important to achieve such clustering target. With different types of data, the formulas of cluster representations must differ accordingly. For instance, *centroid* is specified for numerical data while *mode* and *representative* are specified for categorical data. For the data with both numerical and categorical attributes, the *prototype* structure can be effectively applied by hybridizing the *centroid* and *representative*.

On one hand, a basis k -means algorithm is a local optimization technique that can easily return a locally optimal solution. To achieve a better or the global solution, the clustering algorithm must seek the solutions several times with different initial states. For this reason, a "good" initial state is very important to achieve the global optimum. In this research, we first aim to propose a new scheme to use a dimension reduction technique so-called Locality-Sensitivity Hashing (LSH) to predict the "good" initial state of the cluster so that the global optimal can be potentially obtained. The empirical experiment using real and synthetic datasets showed that our proposed method LSH- k -representatives and LSH- k -prototypes not only can outperform other related works in terms of clustering accuracy but also have the best consistency for clustering categorical and mixed data, respectively. However, the proposed LSH-based cluster prediction requires extra processes in order to create the LSH hash table, which makes the proposed method not ready for handling big data yet.

On the other hand, the complexity of a k -means-like algorithm varies linearly with the

volume of data while the volume of data is exploded day by day. Thus, it is essential to reduce the complexity of k -means-like algorithms so that they become capable of processing big and real data. Dimension reduction and data sample are the most used techniques that can approximate the clustering procedures. However, these approaches change the nature of the data instead of changing the algorithm to make it more appropriate. This dissertation also fills such shortcoming by proposing a new heuristic approach for approximately reducing the complexity of a typical k -means-like algorithm. In detail, the proposed method can avoid the potential unnecessary distance computations from objects to cluster representations in each iteration. Consequently, after applying our proposed method into LSH k -representatives, the incorporated algorithm can process up to 2 to 32 times faster than its own original version with comparable clustering accuracy.

Moreover, we also extend a kernel-based representation of so-called LSH k -prototypes to make it capable of fuzzy clustering of categorical data. The LSH-based cluster prediction technique is then extended to estimate the fuzzy clusters of categorical data. Eventually, the proposed fuzzy clustering algorithm so-called LSHF k -centers can outrun other state-of-the-art fuzzy clustering approaches.

Keywords: Cluster analysis, fuzzy cluster analysis, categorical data, mixed data, LSH, cluster initialization, big data.

論文審査の結果の要旨

Unsupervised learning is getting more and more attention because labeling data is both costly and time consuming, and sometimes impossible especially in the context of big data. As an important branch in unsupervised learning, clustering has re-emerged as an active research topic recently. Despite the efforts made in the literature, clustering of massive data sets with categorical and mixed-type data is still a significant challenge in many applications of big data mining, due to the lack of inherently meaningful measure of similarity between categorical objects and the high computational complexity of existing clustering techniques. The main objective of this dissertation was to develop new scalable and effective clustering methods for massive categorical and mixed datasets. Extensive experiments were conducted to evaluate the performance of the proposed clustering algorithms against other previously developed clustering methods. The main results of this research are summarized as follows.

Firstly, a novel approach that incorporates the Locality-Sensitive Hashing (LSH) technique for cluster initialization into the k -means-like clustering was proposed for categorical data. Essentially, in the proposed approach the LSH technique is applied for predicting quality clusters at the initialization stage and a nearest neighbor search is used at each iteration for cluster reassignment of data objects to improve the clustering complexity. These solutions were eventually incorporated into the k -representatives algorithm resulting in the so-called LSH- k -representatives algorithm. Secondly, the proposed approach was extended to make it applicable for clustering mixed numeric and categorical data that resulted in the so-called LSH- k -prototypes algorithm. Finally, a new

method for fuzzy clustering based on the new LSH-based initialization method and the kernel-based representation of cluster centers was also developed. All newly developed clustering algorithms were experimentally tested on multiple real-world and synthesis datasets. Experimental results have shown that the new algorithms yield comparable or better clustering results in comparison to the previously developed methods in terms of clustering quality indexes in most cases and are scalable as well.

This dissertation has made significant contributions to methodological and experimental developments within the area of cluster analysis. The research work presented in this dissertation has resulted in 2 journal papers (1 published and 1 under review), and two refereed conference papers.

In summary, Mr. NGUYEN Mau Toan has completed all the requirements in the doctoral program of the School of Information Science, JAIST and finished the examination on November 5, 2021, and all committee members approved awarding him a doctoral degree.

Date: *09 November 2021*