氏　　名	DINH, Duy Tai			
学位の種類	博士(知識科学)			
学位記番号	博知第 267 号			
学位授与年月日	令和 2 年 3 月 25 日			
論文題目	Effective Clustering Algorithms for Categorical and Mixed Data			
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
		藤波　努	同	教授
		DAM　Hieu Chi	同	准教授
		由井薗　隆也	同	准教授
		YAN Hongbin	East China Univ. of Sci. and Tech.	教授

論文の内容の要旨

This dissertation focuses on several topics for categorical and mixed data clustering. It provides a thorough background, theoretical models and empirical studies of the pro- posed frameworks. Key concepts and terminologies are also introduced. First, we de- sign a novel clustering algorithm for categorical data. The algorithm uses a kernel-based method for the formation of cluster centers. This approach provides an interpretation of cluster centers being consistent with the statistical interpretation of the cluster means in numeric data clustering. In addition, taking the underlying distribution of categorical attributes into consideration, we define an information-theoretic based measure of dissimilarity for categorical data. This dissimilarity measure is used for computing the distance between categorical objects and cluster centers. The kernel-based method and information-theoretic based measure will be further used for clustering steps of all proposed frameworks in the dissertation. Second, we design an integrated framework for clustering categorical data with missing values. The proposed model can impute missing values occurring in data objects and assign them into appropriate clusters. For the imputation, we use a decision tree-based method to fill in missing values within data. This method has shown to be suitable for categorical data since it can find the set of complete objects that are highly correlated with the data object having missing values. From that, appropriate values are selected for missing positions. The kernel-based method and information-theoretic based measure are used for clustering steps. Third, we extend the second model to solve the problem of clustering mixed numeric and categorical data with missing values. For the imputation, the model splits an input data set into two sub-datasets based on their data types. The decision-tree based method is also used for imputing missing values inside objects constituted by categorical attributes. The missing values in numeric attributes are imputed by using the mean of corresponding attributes from the correlated set. For the clustering, we use the mean and the kernel-based method to define cluster centers for numeric and categorical attributes, respectively. The squared Euclidean and information-theoretic based dissimilarity measure is used to calculate distances for numeric and categorical attributes, respectively. Fourth, we design a framework to address the limitation

of random initialization in categorical data clustering. Specifically, a maximal frequent itemset mining approach is used to find the sets of correlated itemsets (patterns). Each pattern describes the largest set of categories occurring in the corresponding categorical object. The group of data objects containing each pattern is considered as an initial cluster. The kernel-based method and information-theoretic based measure are used for clustering steps. Fifth, we design a framework to estimate the optimal number of clusters (k) in categorical data clustering. The silhouette analysis-based approach is used to evaluate different clustering results so as to choose the best k for each data set. The kernel-based method and information-theoretic based measure are used for clustering steps. All proposed frameworks are tested on real benchmark data sets from open access data repositories. We compare them with previous clustering algorithms in terms of clustering quality and computational complexity by using several internal and external validation metrics. In general, the proposed frameworks can enhance clustering results and can be used to perform clustering tasks for any real categorical and mixed data sets as long as their formats match the input requirement of algorithms.

Keywords: clustering, partitional clustering, categorical data, mixed data, missing values, kernel-based method, information-theoretic based dissimilarity, cluster center initialization, optimal number of clusters

論文審査の結果の要旨

Clustering is one of fundamental operations in data mining and machine learning, and has been widely applied in a variety of fields: ranging from medical sciences, economics, computer sciences, engineering to social sciences and earth sciences. Despite recent efforts, 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 a k-means like clustering framework capable of handling missing data for categorical and mixed datasets. A series of experiments tested on real datasets from UCI Machine Learning Repository was also conducted to evaluate the performance of the proposed clustering framework against other previously developed clustering methods. The main results of this research are summarized as follows.

Firstly, based on a newly developed similarity measure for categorical data, a novel k-means like clustering framework making use of kernel-based methods for representation of cluster centers was developed. Essentially in the proposed clustering framework it allows us to formulate the problem of clustering categorical data in the fashion similar to k-means clustering, while kernel-based representation of centers also provides an interpretation of cluster means being consistent with the statistical interpretation of the cluster means for numerical data. Within this framework, a class of k-means like algorithms for clustering categorical data have been experimentally implemented and tested. Secondly, a so-called k-CCM algorithm for clustering categorical data with missing values was proposed by

incorporating a decision tree-based imputation method into clustering process. An extensive experimental evaluation was conducted on benchmark categorical datasets to evaluate the performance of the k-CCM algorithm. Finally, the proposed k-means like clustering framework was extended so as to make it applicable for clustering mixed numeric and categorical datasets with missing data. In particular, a so-called k-CMM method for clustering mixed numeric and categorical datasets with missing values by integrating the imputation step into the proposed clustering framework was developed and experimentally tested. Experimental results have shown that k-CMM is more efficient than the previously developed methods in terms of clustering quality indexes (namely, Purity, Normalized Mutual Information (NMI) and Adjusted Rand Index) in most cases and 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 4 journal papers (2 published and 2 under review), and several refereed conference papers.

In summary, Mr. DINH Duy Tai has completed all the requirements in the doctoral program of the School of Knowledge Science, JAIST and finished the examination on February 7, 2020, all committee members approved awarding him a doctoral degree in Knowledge Science.