Effective Clustering Algorithms for Categorical and Mixed Data

DINH DUY TAI

S1720019

Supervisor: Prof. HUYNH VAN NAM

Second Supervisor: Prof. TAKASHI HASHIMOTO

A dissertation presented for the degree of

Doctor of Philosophy

Graduate School of Advanced Science and Technology

Japan Advanced Institute of Science and Technology

Knowledge Science

March, 2020

Abstract

This dissertation focuses on several topics for categorical and mixed data clustering. It provides a thorough background, theoretical models and empirical studies of the proposed frameworks. Key concepts and terminologies are also introduced. First, we design 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 mea-

sure 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