DOCTORAL THESIS

Advances in categorical data clustering

Zhang, Yiqun

Date of Award:
2019

Abstract

Categorical data are common in various research areas, and clustering is a prevalent technique used for analyse them. However, two challenging problems are encountered in categorical data clustering analysis. The first is that most categorical data distance metrics were actually proposed for nominal data (i.e., a categorical data set that comprises only nominal attributes), ignoring the fact that ordinal attributes are also common in various categorical data sets. As a result, these nominal data distance metrics cannot account for the order information of ordinal attributes and may thus inappropriately measure the distances for ordinal data (i.e., a categorical data set that comprises only ordinal attributes) and mixed categorical data (i.e., a categorical data set that comprises both ordinal and nominal attributes). The second problem is that most hierarchical clustering approaches were actually designed for numerical data and have very high computation costs; that is, with time complexity $O(N^2)$ for a data set with N data objects. These issues have presented huge obstacles to the clustering analysis of categorical data.

To address the ordinal data distance measurement problem, we studied the characteristics of ordered possible values (also called 'categories' interchangeably in this thesis) of ordinal attributes and propose a novel ordinal data distance metric, which we call the Entropy-Based Distance Metric (EBDM), to quantify the distances between ordinal categories. The EBDM adopts cumulative entropy as a measure to indicate the amount of information in the ordinal categories and simulates the thinking process of changing one's mind between two ordered choices to quantify the distances according to the amount of information in the ordinal categories. The order relationship and the statistical information of the ordinal categories are both con-

sidered by the EBDM for more appropriate distance measurement. Experimental results illustrate the superiority of the proposed EBDM in ordinal data clustering.

In addition to designing an ordinal data distance metric, we further propose a unified categorical data distance metric that is suitable for distance measurement of all three types of categorical data (i.e., ordinal data, nominal data, and mixed categorical data). The extended version uniformly defines distances and attribute weights for both ordinal and nominal attributes, by which the distances measured for the two types of attributes of a mixed categorical data can be directly combined to obtain the overall distances between data objects with no information loss. Extensive experiments on all three types of categorical data sets demonstrate the effectiveness of the unified distance metric in clustering analysis of categorical data.

To address the hierarchical clustering problem of large-scale categorical data, we propose a fast hierarchical clustering framework called the Growing Multi-layer Topology Training (GMTT). The most significant merit of this framework is its ability to reduce the time complexity of most existing hierarchical clustering frameworks (i.e., $O(N^2)$) to $O(N^{1.5})$ without sacrificing the quality (i.e., clustering accuracy and hierarchical details) of the constructed hierarchy. According to our design, the GMTT framework is applicable to categorical data clustering simply by adopting a categorical data distance metric. To make the GMTT framework suitable for the processing of streaming categorical data, we also provide an incremental version of GMTT that can dynamically adopt new inputs into the hierarchy via local updating. Theoretical analysis proves that the GMTT frameworks have time complexity $O(N^{1.5})$. Extensive experiments show the efficacy of the GMTT frameworks and demonstrate that they achieve more competitive categorical data clustering performance by adopting the proposed unified distance metric.

Keywords: Categorical data, clustering analysis, distance metric, information theory, large-scale data, nominal attribute, ordinal attribute, streaming data.

Table of Contents