

DOCTORAL THESIS

Advances in categorical data clustering

Zhang, Yiqun

Date of Award:
2019

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

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

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Tables	ix
List of Figures	xi
Chapter 1 Introduction	1
1.1 Ordinal and Nominal Attributes	3
1.2 Categorical Data Clustering	5
1.3 Categorical Data Metrics	6
1.4 Fast and Incremental Hierarchical Clustering	7
1.5 Main Contributions	8
1.6 Organisation	11
Chapter 2 Literature Review of Related Works	12
2.1 Categorical Data Clustering	12
2.1.1 Partitional Categorical Data Clustering Algorithms	12
2.1.2 Fast Hierarchical Clustering Approaches	16
2.2 Distance Measurement	22
2.2.1 Hamming Distance Metric	23

2.2.2	Association-based Distance Metric	23
2.2.3	Ahmad’s Distance Metric	24
2.2.4	Context-based Distance Metric	25
2.2.5	Jia’s Distance Metric	26
2.3	Inter-Attribute Dependence Measurement	27
2.3.1	Nominal Measures	27
2.3.2	Ordinal Measures	28
2.4	Validity Indices for Clustering Performance Assessment	30
2.4.1	Clustering Accuracy	30
2.4.2	Adjusted Rand Index	31
2.4.3	Normalised Mutual Information	31
2.4.4	Fowlkes Mallows Index	32
2.5	Summary	32

Chapter 3 Distance Metric for Ordinal Data Clustering 34

3.1	Introduction	34
3.2	Preliminaries	37
3.3	The Proposed Metric	39
3.3.1	Basic Idea	39
3.3.2	EBDM: Entropy-Based Distance Metric	40
3.3.3	Scale Normalisation	41
3.4	Discussions	41
3.4.1	Mathematical Properties	42
3.4.2	Distance Measurement	43
3.4.3	Time Complexity Analysis	43
3.5	Experiments	45
3.5.1	Experimental Settings	45
3.5.2	Comparative Studies	46
3.5.3	Distance Matrices Demonstration	48
3.5.4	Evaluation of the Order Information Exploiting	49
3.6	Summary	52

Chapter 4	Unified Distance Metric for Categorical Data Clustering	54
4.1	Introduction	54
4.2	Preliminaries	59
4.3	The Unified Distance Metric	60
4.3.1	Limitations of EBDM	60
4.3.2	Attribute Weighting	61
4.3.3	UEBDM: Unified Entropy-Based Distance Metric	62
4.4	Discussions	63
4.4.1	Mathematical Properties	63
4.4.2	Distance Measurement	64
4.4.3	Time Complexity Analysis	65
4.5	Experiments	68
4.5.1	Experimental Settings	68
4.5.2	Clustering Performance on Ordinal and Mixed Categorical Data	70
4.5.3	Clustering Performance on Nominal Data	76
4.5.4	Evaluation of UEBDM and UEBDM ^{nom}	79
4.5.5	Weighting Scheme Evaluation	80
4.5.6	Distance Matrices Demonstration	82
4.6	Summary	85
Chapter 5	Fast Hierarchical Clustering of Categorical Data	86
5.1	Introduction	86
5.2	Preliminaries	90
5.2.1	Hierarchy	90
5.2.2	Topology	91
5.3	The Proposed Method	91
5.3.1	GMTT: Growing Multi-layer Topology Training	91
5.3.2	Fast Hierarchical Clustering Based on GMTT	98
5.3.3	Incremental Hierarchical Clustering Based on GMTT	104
5.4	Time Complexity Analysis	106
5.4.1	Time Complexity Analysis for GMTT-UEBDM	106

5.4.2	Time Complexity Analysis for IGM-TT-UEBDM	109
5.4.3	Time Complexity Analysis for MST-Hierarchy Transformation	110
5.4.4	Discussions	110
5.5	Experiments	111
5.5.1	Experimental Settings	111
5.5.2	Performance Evaluation of G-M-TT-UEBDM	113
5.5.3	Performance Evaluation of IGM-TT-UEBDM	120
5.5.4	Study of the Branching Factor	123
5.6	Summary	125
Chapter 6 Conclusions and Future Work		127
6.1	Conclusions	127
6.2	Future Work	129
List of Publications		147
Curriculum Vitae		149