IDEA GROUP PUBLISHING

701 E Chocolate Avenue, Hershey PA17033, USA
Tel: 717/533-8845; Fax: 717/533-8661; URL: www.idea-group.com

ITB7255

Chapter VI

Clustering Mixed Incomplete Data

José Ruiz-Shulcloper
University of Tennessee-Knoxville, USA
Institute of Cybernetics, Mathematics and Physics, Cuba

Guillermo Sánchez-Díaz
Autonomous University of the Hidalgo State, Mexico

Mongi A. Abidi
University of Tennessee-Knoxville, USA

In this chapter, we expose the possibilities of the Logical Combinatorial Pattern Recognition (LCPR) tools for Clustering Large and Very Large Mixed Incomplete Data (MID) Sets. We start from the real existence of a number of complex structures of large or very large data sets. Our research is directed towards the application of methods, techniques and in general, the philosophy of the LCPR to the solution of supervised and unsupervised classification problems. In this chapter, we introduce the GLC and DGLC clustering algorithms and the GLC+ clustering method in order to process large and very large mixed incomplete data sets.

Idea Group Inc
Idea Group Inc
Idea Group Inc
Idea Group Inc

This chapter appears in the book, *Heuristic and Optimization for Knowledge Discovery*, edited by Ruhul Sarker, Hussein Abbass and Charles Newton.
 Copyright 2002, Idea Group Inc.

CLUSTERING MIXED INCOMPLETE DATA

In the process of Knowledge Discovery from Data (KDD), one of the most important tasks is to classify data. It is well known that one of the most powerful tools to process data in order to extract knowledge is the class of clustering algorithms, whose purpose is (in the KDD context) to solve the following problem. Given a similarity measure Γ (not necessarily a distance function) between pairs of object descriptions in some representation space and a collection of object descriptions in that space, find a structuralization of this collection. These sets could form hard or fuzzy cover or partition (Martínez-Trinidad, Ruiz-Shulcloper, & Lazo-Cortés, 2000a; Ruiz-Shulcloper, & Montellano-Ballesteros, 1995) of the data set. In other words, finding the similarity relationship between any pair of objects under a certain clustering criterion without utilizing a priori knowledge about the data and with the following additional constraints: i) the use of computing resources must be minimized and ii) the data set could be *large* or *very large*.

Also, it is well known today that in some areas such as finance, banking, marketing, retail, virtual libraries, healthcare, engineering and in diagnostic problems in several environments like geosciences and medicine among many others, the amount of stored data has had an explosive increase (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). In these areas, there are many instances where the *description of the objects is nonclassical*, that is, the features are not exclusively numerical or categorical. Both kinds of values can appear simultaneously, and sometimes, even a special symbol is necessary to denote the absence of values (missing values). A *mixed and incomplete* description of objects should be used in this case. Mixed in the sense that there are simultaneously categorical and numerical features; incomplete because there are missing values in the object descriptions.

The study of the similarity relationships with mixed incomplete descriptions of objects is the principal aim of LCPR (Martínez-Trinidad et al., 2000a; Ruiz-Shulcloper, & Montellano-Ballesteros, 1995; Dmitriev, Zhuravlev, & Krendelev, 1966).

In order to gain clarity and understanding, we will establish conventional differences between *Data Set* (DS), *Large Data Set* (LDS) and *Very Large Data Set* (VLDS). In a mining clustering (also in a supervised classification) process DS will be understood to mean a collection of object descriptions where the size of the set of descriptions together with the size of the result of the comparison of all pair wise object descriptions, that is, the *similarity matrix*, does not exceed the available memory size. LDS will mean the case where only the size of the set of descriptions does not exceed the available memory size, and VLDS will mean the case where both sizes exceed the available memory size.

In addition, we propose conventional differences between the *Very Large Data Set Clustering Algorithm* (VLDSCA), the *Large Data Set Clustering Algorithm* (LDSCA), and the *Data Set Clustering Algorithm* (DSCA). If we denote OT_A as the run-time complexity, and OE_A as the space complexity of a clustering algorithm CA, then we have a VLDSCA iff $OT_A < O(m^2)$ and $OE_A < O(m)$. We have a LDSCA iff

Idea Group Inc

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/clustering-mixed-incomplete-data/22151

Related Content

Examining University Retention Efforts of Non-Traditional Students
Valerie McGaha-Garnett (2012). *Cases on Institutional Research Systems (pp. 228-237).*
www.irma-international.org/chapter/examining-university-retention-efforts-non/60850

Recognizing Threats From Unknown Real-Time Big Data System Faults
William H. Moneyand Stephen J. Cohen (2020). *Current Issues and Trends in Knowledge Management, Discovery, and Transfer (pp. 331-366).*
www.irma-international.org/chapter/recognizing-threats-from-unknown-real-time-big-data-system-faults/244891

Using Grids for Distributed Knowledge Discovery
Antonio Congiusta, Domenico Taliaand Paolo Trunfio (2008). *Mathematical Methods for Knowledge Discovery and Data Mining (pp. 284-298).*
www.irma-international.org/chapter/using-grids-distributed-knowledge-discovery/26146

Opinion Mining with SentiWordNet
Bruno Ohanaand Brendan Tierney (2011). *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains (pp. 266-286).*
www.irma-international.org/chapter/opinion-mining-sentiwordnet/46900

A Composite Risk Model for Optimizing Information System Security
Yahel Giatand Michael Dreyfuss (2020). *Optimizing Data and New Methods for Efficient Knowledge Discovery and Information Resources Management: Emerging Research and Opportunities (pp. 74-97).*
www.irma-international.org/chapter/a-composite-risk-model-for-optimizing-information-system-security/255752