Binarization and Validation in Formal Concept Analysis

Mostafa A. Salama, Department of Computer Science, British University in Egypt, El Sherouk City, Cairo, Egypt
Aboul Ella Hassanien, Faculty of Computers and Information, Cairo University, Giza, Cairo, Egypt

ABSTRACT

Representation and visualization of continuous data using the Formal Concept Analysis (FCA) became an important requirement in real-life fields. Application of formal concept analysis (FCA) model on numerical data, a scaling or Discretization / binarization procedures should be applied as preprocessing stage. The Scaling procedure increases the complexity of computation of the FCA, while the binarization process leads to a distortion in the internal structure of the input data set. The proposed approach uses a binarization procedure prior to applying FCA model, and then applies a validation process to the generated lattice to measure or ensure its degree of accuracy. The introduced approach is based on the evaluation of each attribute according to the objects of its extent set. To prove the validity of the introduced approach, the technique is applied on two data sets in the medical field which are the Indian Diabetes and the Breast Cancer data sets. Both data sets show the generation of a valid lattice.

Keywords: Binarization, ChiMerge, Discretization, Feature Selection, Formal Concept Analysis (FCA)

1. INTRODUCTION

Formal Concept Analysis (FCA) is a data mining model that introduces the relation among attributes in a visual form. It was introduced in the early 80s by Wille (1982) to study how objects can be hierarchically grouped together according to their common attributes. It serves in different fields such as biology (Bertaux et al., 2007), medicine (Motameny et al., 2008) and system analysis (Düwel, 2009). This tool is of a great interest for mining association rules in real life data, specially the numerical ones. The basic structure of FCA is the formal context which is a binary-relation between a set of objects and a set of attributes. The formal context is based on the ordinary set, whose elements has one of two values, 0 or 1 (Kaytoue et al., 2009; Cole et al., 1998). A context materializes a set of individuals called objects, a set of properties called attributes, and a binary relation usually represented by a binary table relating objects to attributes.

These mappings are called Galois connections or concepts. Such concepts are ordered in FCA within a lattice structure called concept lattice within the FCA. Concept lattices can be represented by diagrams giving clear visualiza-
tion of classes of objects in each domain. At the same time, the edges of these diagrams give essential knowledge about objects, by introducing association rules between attributes which describe the objects (Sergei, 2002). Mostly, the real-world data are not available as binary data. Such data could be either numerical or categorical. To represent a numerical or a categorical data in the form of a formal context, such data should be transformed using conceptual scaling. In this approach, the attribute of numerical values are discredited. Then each interval of entry values, have to be considered as binary attributes (Motameny et al., 2008). The transformation of such data, i.e., conceptual scaling, allows one to apply FCA tools. Such procedure may dramatically increase the complexity of computation and representation. Hence, it worsens the visualization of results. This scaling may produce large and dense binary data (Snsel et al., 2008), which are hard to process with the existing FCA algorithms. As it is based on arbitrary choices, the data may be scaled in a lot of different ways that lead to different results. Its interpretations could lead also to classification problems. The work done by Fu and Nguifo (2003) proposed a scalable lattice-based algorithm ScalingNextClosure to decompose the search space for finding formal concepts in large datasets into partitions and then generate independently concepts (or closed item sets) in each partition. This paper replaces the scaling technique by a method that uses of the chiMerge algorithm into binarization of the numerical data attributes and into the validation of the generated formal concept lattice. The binarization technique is applied using the chiMerge algorithm through discretizing the continuous attribute values into only of two values, 0 or 1. Then the resulted binary table is used in generation of the formal concept lattice. Then the chiMerge method here is used to validate this generated lattice. For continuous data sets, the ChiMerge method is used to automatically select proper Chi-square $\chi^2$ values to evaluate the worth of each attribute with respect to the corresponding classes (Thabtah et al., 2002; Kumar & Rao, 2009). These $\chi^2$ values are used to compare value of each attribute. Such values were calculated according to a novel formula based on the generated formal concept lattice. If both evaluations are matched, then the generated lattice is considered to be representing the actual structure of the data. Hence, the binarization method does not corrupt the generated lattice. Finally, it led to a valid lattice. The conceptual computation and the lattice visualization are performed using a tool for formal concept lattice generation named (OpenFCA, http://code.google.com/p/openfca). The introduced techniques, the binarization, visualization and validation methods, is applied on two data sets in the medical field from the UCI database; the Indian Diabetes data set and the Breast Cancer data set.

The rest of this paper is organized as follows: Section 2 gives an overview about related work in FCA. Section 3 shows the proposed approach. The experimental results and conclusions are presented in Section 4 and Section 5 respectively.

2. RELATED WORK

Formal concept analysis (FCA) explores attributes of objects and their corresponding contexts, and classifies systems based on relationships among objects and attributes (Stumme et al., 2008). It visualizes the inherent properties in data sets, allowing meaningful and comprehensible interpretation. In Formal Concept Analysis, the elements of one type are called formal objects or extents, the elements of the other type are called “formal attributes” or intents. The sets of formal objects and formal attributes together with their relation to each other form a formal context. And a pair of a set of formal objects and a set of formal attributes that is closed in this manner is called a formal concept (Sergei, 2004). A concept lattice consists of the set of concepts of a formal context and the subconcept-supercconcept relation between the concepts. If a formal concept has a formal attribute then its attributes are inherited by all its subconcepts. In general, formal context is
Exerting Cost-Sensitive and Feature Creation Algorithms for Coronary Artery Disease Diagnosis
www.igi-global.com/article/exerting-cost-sensitive-feature-creation/74695?camid=4v1a

Biomedical Librarianship in the Post-Genomic Era
Shubhada Prashant Nagarkar (2017). Library and Information Services for Bioinformatics Education and Research (pp. 1-17).
www.igi-global.com/chapter/biomedical-librarianship-in-the-post-genomic-era/176134?camid=4v1a