Expert Systems with Applications 42 (2015) 2582–2592

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Association rule mining with mostly associated sequential patterns

Ömer M. Soysal *

Highway Safety Research Group, Louisiana State University, 3535 Nicholson Ext., Baton Rouge, LA, USA

ARTICLE INFO

Article history:
Available online 10 November 2014

Keywords:
Association rule mining
Interesting rules
Pattern recognition
Big data
Knowledge discovery
Data mining

ABSTRACT

In this paper, we address the problem of mining structured data to find potentially *useful* patterns by association rule mining. Different than the traditional *find-all-then-prune* approach, a heuristic method is proposed to extract mostly associated patterns (MASPs). This approach utilizes a maximally-association constraint to generate patterns without searching the entire lattice of item combinations. This approach does not require a pruning process. The proposed approach requires less computational resources in terms of time and memory requirements while generating a long sequence of patterns that have the highest co-occurrence. Furthermore, *k*-item patterns can be obtained thanks to the sub-lattice property of the MASPs. In addition, the algorithm produces a tree of the detected patterns; this tree can assist decision makers for visual analysis of data. The outcome of the algorithm implemented is illustrated using traffic accident data. The proposed approach has a potential to be utilized in big data analytics.

1. Introduction

In the last century, data-driven decision making is becoming more challenging due to production and processing of extremely huge amount of data from a variety of sensors. The decision makers are often required to understand relations within the multi-dimensional space before taking an action, making a law, producing a product, setting up regulations, etc. In this paper, we attempt to reveal *useful relations* by means of extracting mostly associated patterns from a structured data.

Many approaches have been proposed to discover *useful* information from structured data. Among these approaches, association rule mining (ARM) plays an important role. The ARM algorithms aim to discover hidden rules among enormous pattern combinations based on their individual and conditional frequencies. The traditional ARM algorithms first generate all of the possible patterns from the data while pruning out non-frequent ones and then produce rules from these frequent patterns. Once the rules are generated, some interesting measures (IMs) are applied to obtain interesting rules that can be used in decision making. A general process flow of an ARM framework is shown in Fig. 1. A brief description of the modules in this general framework is as follows: (a) the Preprocess module is used to localize data by filtering, to summarize data by sampling, or to transform data to speed up rule detection, (b) the C-Generator finds candidate patterns, (c) a pruning is applied before rule generation, (d) the R-Generator is used to generate *k*-items rules, (e) interesting rules are obtained by the R-Filter. The constraints define the rule search strategy. The ARM algorithms differ mainly from each other based on utilization of these constraints. Among the many, both thresholds, the minimum support and minimum confidence, would be considered as default constraints.

As we summarized in the section 'Related Works' of this paper, interesting rules are extracted through an exhaustive search if no constraint other than the default ones is used on patterns. In this paper, we propose an approach that imposes an interestingness constraint on patterns to detect the highest co-occurring ones without searching all possible pattern combinations (entire lattice) and filtering them out later. This approach offers an advantage of consuming significantly less computational resources for finding long rule sequences. During the search process, a most associated sequential pattern (MASP) tree is formed. After generating the MASP tree, the rules are generated in significantly less computation time. Besides obtaining MASPs, a traditional rule mining can be conducted within a relatively small data set of each MASP; the outcome of both MASPs' rules and traditionally obtained rules can be combined to find interesting rules as explained in the method section of this paper; this combination is named as 'MASP+'. Readers should refer to Lemma 4. Furthermore, the MASP tree has a sub-lattice rule generation property that reveals *k*-items rules from MASPs as stated in Theorem 1.

In general, real data to be mined has 'attribute = value' imbalances; that is, some distinct values of an attribute are

* Tel.: +1 225 578 6297; fax: +1 225 578 0240.
E-mail address: omsoysal@lsu.edu

http://dx.doi.org/10.1016/j.eswa.2014.10.049
0957-4174/

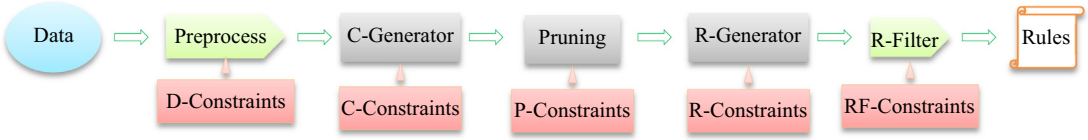

Fig. 1. The general process flow of an ARM framework for detection of interesting rules.

over-represented than other values of the same attribute. As an example, events of property-damage-only cases are extensively more than the injury-only cases, and injury-only cases are relatively higher than fatal-only cases in traffic accident data. When applied to such a data, a traditional ARM algorithm will favor the over-representing frequent items; consequently, these over-representing items will show up in most of the rules. The proposed approach can find these most *favorable* rules without spanning the entire search lattice. In addition, the proposed approach is capable of discovering long patterns while utilizing less resources compared to the exhaustive search approaches that consume a significant amount of resources.

The rest of the paper is as follows: The literature review about searching lattice and mining interesting rules are presented in the section *Related Works*. The proposed approach is introduced in the *Method* section. Data, the experiments conducted, and their results are summarized in the *Experiments and Results*. The paper is finalized with conclusion and future works.

2. Related Works

In general, the problem of mining association rules is solved in two steps (Das, Ng, & Woon, 2001): (1) first, all frequent itemsets are found, (2) then, association rules are generated from the frequent itemsets. Once the rules are obtained, the rules are ranked by their interestingness measure. In this section, we provide a brief review of approaches that aim to search combinatorial pattern space for finding frequent itemsets and to filter the interesting rules among the rule set.

Association rule mining algorithms can be classified based on the search strategy used to find frequent itemsets and on the scope of the search. The scope can be the entire lattice or a sub-lattice determined by constraints. The search strategies show varieties based on how to traverse data space; some algorithms find frequent itemsets directly from the transactional data while others form an intermediate data structure. In the former group, the Apriori algorithm (Agrawal & Srikant, 1994) utilizes the bread-first approach, and the Eclat algorithm (Klbsgen, 1996) uses the depth-first approach. In the latter group, the FP-growth algorithm (Han, Pei, Yin, & MAO, 2004) transforms transactional data into the form of a tree; the A-Close proposed in Pasquier, Bastide, and Taouil (1998) finds frequent closed itemsets from which all frequent itemsets are derived or rules are directly generated from the closed set; the MAFIA proposed in Burdick et al. (2001) obtains the maximal itemsets before rule generation.

The constraint-based algorithms perform a filtering operation on the data itself, on the patterns while being generated, or on the patterns after being generated (Kotsiantis & Kanellopoulos, 2006). A constraint can belong to a data set, to a measure (such as a statistic) for discovering patterns, or to the type of patterns to be discovered (Wojciechowski & Zakrzewicz, 2002); note that temporal and spatial constraints would be considered under the 'type of patterns'. Among the constraint-based mining, the RARM (Das et al., 2001) finds frequent 2-itemsets and utilizes an Apriori-based strategy to find frequent k-itemsets where $k \geqslant 3$. In Das et al. (2001), a schema constraint, which defines the structure of the patterns, and the opportunistic confidence constraint, which aims to discriminate significant and redundant rules, are introduced. The category-based (or concept hierarchy-based) approaches, e.g. in Do et al. (2003) at each pass, check whether the transaction has items belonging to the "categories" (or concepts) specified by the user.

Multiple-minimum supports proposed in Wojciechowski and Zakrzewicz (2002) discover sequential patterns by means of a tree structure. An improved version of the predictive (n,p) approach proposed in Denwattana and Getta (2001) is introduced in Hong, Horng, Wu, and Wang (2009), where the frequent itemsets are discovered through promising and non-promising candidate itemsets using two threshold parameters of minimum itemsets' length and minimum frequency. A review of association rule mining algorithms from the subgroup discovery perspective is provided in Herrera, Carmona, González, and Jesus (2011).

Among the most recent research on finding frequent patterns, as an emerging topic, mining top-k frequent patterns that does not require to set a minimum support value is studied by Pyun and Yun (2014), Deng (2014). The closed itemsets can be extracted from these top-k frequent patterns. The former researchers developed a new algorithm based on the FP-growth structure and the later proposed a new data structure named Node-list. Tseng (2013) addresses the problem of mining large databases. The author proposed a hierarchical partitioning approach on both the database and solution space. In discovery of patterns from large database, Király, Laiho, Abonyi, and Gyenesei (2014) reduced two well-known problems of frequent closed itemset mining and biclustering into a single problem for binary data. In Chen, Lan, Hong, and Lin (2013), propositional logic is utilized to find coherent rules that take into consideration of negations; this approach addresses to find an appropriate minimum support as well. Jin, Wang, Huang, and Hu (2014) employed causality between antecedent and consequent to discover interesting rules; they used causality as an objective measure. The frequent itemsets and useful rules are explored by similarity instead of attribute–value equivalence in Rodríguez-González, Martínez-Trinidad, and Carrasco-Ochoa (2013). They adapted the algorithm proposed in Agrawal and Srikant (1994) to generate interesting rules. In Vo, Coenen, and Le (2013), significance of items are considered while finding frequent itemsets and interesting patterns. They proposed the WIT-tree (Weighted Itemset-Tidset tree) as a data structure to mine high utility itemsets.

2.1. Rule interestingness

In ARM, the second main step after discovering frequent patterns is to generate the rules. As in the most cases, the ARM- based information discovery suffers from producing many trivial or uninteresting patterns when all possible rules are produced first and then redundant ones are eliminated (Ashrafi, Taniar, & Smith, 2004, 2005; Omiecinski, 2003). Sahar (2010) classifies IMs in three main categories as objective, subjective, and semantics-based measures. Many criteria have been proposed for elimination of redundant rules (or for revealing interesting ones) (Heravi & Zaïane, 2010; Sahar, 2010). Discovery of non-redundant rules based on