Data & Knowledge Engineering 86 (2013) 19–37

Contents lists available at SciVerse ScienceDirect

Data & Knowledge Engineering

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

Grammar-based multi-objective algorithms for mining association rules

J.M. Luna, J.R. Romero, S. Ventura *

Dept. of Computer Science and Numerical Analysis, University of Cordoba, Rabanales Campus, Albert Einstein Building, 14071 Cordoba, Spain

ARTICLE INFO

Article history:
Received 28 April 2011
Received in revised form 31 December 2012
Accepted 7 January 2013
Available online 16 January 2013

Keywords:
Association rule mining
Genetic programming
Data mining
Mining methods and algorithms

ABSTRACT

In association rule mining, the process of extracting relations from a dataset often requires the application of more than one quality measure and, in many cases, such measures involve conflicting objectives. In such a situation, it is more appropriate to attain the optimal trade-off between measures. This paper deals with the association rule mining problem under a multi-objective perspective by proposing grammar guided genetic programming (G3P) models, that enable the extraction of both numerical and nominal association rules in only one single step. The strength of G3P is its ability to restrict the search space and build rules conforming to a given context-free grammar. Thus, the proposals presented in this paper combine the advantages of G3P models with those of multi-objective approaches. Both approaches follow the philosophy of two well-known multi-objective algorithms: the Non-dominated Sort Genetic Algorithm (NSGA-2) and the Strength Pareto Evolutionary Algorithm (SPEA-2).

In the experimental stage, we compare both multi-objective algorithms to a single-objective G3P proposal for mining association rules and perform an analysis of the mined rules. The results obtained show that multi-objective proposals obtain very frequent (with support values above 95% in most cases) and reliable (with confidence values close to 100%) rules when attaining the optimal trade-off between support and confidence. Furthermore, for the trade-off between support and lift, the multi-objective proposals also produce very interesting and representative rules.

1. Introduction

Given the growing interest in information storage, both the number of available datasets and their sizes are increasing. Nowadays, the extraction of knowledge or high level information hidden in data has become essential for predicting future behavior. A popular technique for discovering knowledge in datasets is association rule mining (ARM) [1–4], an unsupervised learning method that includes approaches having a descriptive nature [5,6]. Let $\mathcal{I} = \{i_1, \ldots, i_n\}$ be a set of items, and let A and C be item-sets, i.e., $A = \{i_1, \ldots, i_j\} \subset \mathcal{I}$ and $C = \{i_1, \ldots, i_k\} \subset \mathcal{I}$. An association rule [7] is an implication of the type $A \rightarrow C$ where $A \subset \mathcal{I}$, $C \subset \mathcal{I}$, and $A \cap C = \emptyset$. The meaning of an association rule is that if antecedent A is satisfied, then it is highly likely that consequent C will also be satisfied. ARM was originally designed for market basket analysis to obtain relations between products like *diapers* → *beer* that describes the high probability of someone buying diapers also buying beer. It would allow shop-keepers to exploit this relationship by moving the products closer together on the shelves.

Originally, the ARM problem was studied under an exhaustive search strategy. The first algorithm in this field was *a priori*, an approach suggested by Agrawal et al. [8,9] that served as the starting point for many algorithms in the ARM field [10–12]. Nevertheless, these sorts of algorithms require a very high computational cost and large amount of memory. Also, more and more companies currently gather useful information and, sometimes, this information is purely numeric, so exhaustive search algorithms require a previous discretization of the numerical attributes. To solve these issues, the study of association rules by

* Corresponding author. Tel.: +34 957212218; fax: +34 957218630.
E-mail addresses: jmluna@uco.es (J.M. Luna), jrromero@uco.es (J.R. Romero), sventura@uco.es (S. Ventura).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.01.002

means of evolutionary algorithms (EA), and especially genetic algorithms (GA), is obtaining promising results [13]. Recently, an initial grammar-guided genetic programming (G3P) [14] proposal was presented in the ARM field. G3P is considered as an extension of genetic programming (GP) [15] that makes use of a grammar to enforce syntactic constraints on GP trees. This new G3P algorithm, called Grammar-Guided Genetic Programming Association Rule Mining (G3PARM) [16], has turned into an area of interest for further exploration.

An important issue in ARM is that, regardless of the methodology used for the extraction of these rules, it is necessary to evaluate them properly. The process of extracting association rules from a dataset often requires the application of more than one quality measure and, in many cases, such measures involve conflicting objectives, so it is necessary to attain the optimal trade-off between them. These problems, called multi-objective optimization problems, need to simultaneously reach more than one objective but do not have a single solution that optimizes them all. Multi-objective algorithms has been used in ARM [37,38] to evaluate rules based on different measures [17,18] but only using nominal attributes. Support and confidence are two of the most commonly used measures. The former states the frequency of occurrence of the rule, while the latter stands for the reliability of the rule. However, as discussed in subsequent sections, these two measures have some limitations, so the lift measure is adequate. Lift calculates how many times the antecedent and the consequent occur together more frequently than expected if they were statistically independent. At this point, we consider dealing with the ARM problem under a multi-objective methodology and for any application domain, not requiring a previous discretization step. Thus, application of multi-objective optimization together with the GP3 methodology could give rise to a promising model especially well suited to optimizing rules in diverse application domains, and using different quality measures.

In this paper, we present two new G3P proposals for mining association rules following a multi-objective strategy. These proposals benefit from the advantages of both G3P [14] and consequently EA [19], and combine them with those of multi-objective models [20]. More specifically, the proposals presented here are based on two well-known multi-objective algorithms: the Non-dominated Sort Genetic Algorithm (NSGA-2) [21] and the Strength Pareto Evolutionary Algorithm (SPEA-2) [22]. Because of the specific grammar definition, these G3P proposals enable the extraction of rules from both numerical and categorical domains. Finally, in order to demonstrate the usefulness of the proposed algorithms, different measures are considered as objectives to obtain a set of optimal solutions. More specifically, the experiments performed combine both the support-confidence and support-lift measures. The results obtained have shown to be very frequent (with support values above 95% in most cases) and reliable (with confidence values close to 100%). Furthermore, for the trade-off between support and lift, the multi-objective proposals also produce very interesting and representative rules.

This paper is structured as follows: Section 2 presents some related work; Section 3 describes the multi-objective G3P proposals; Section 4 describes the datasets used in the experimental stage, the experimental set-up and the results obtained; finally, some concluding remarks are outlined in Section 5.

2. Related work

This section presents the most widely used measures in the ARM field. Next, an introduction to the most relevant multi-objective approaches is outlined, paying special attention to their applicability in the ARM field. We consider that expert readers in Evolutionary Computation could omit this section since it provides basic background in both fields.

2.1. Quality measures

Despite the large number of measures used to evaluate the quality of association rules, most researchers [9,10,23] concur with the application of support and confidence measures because of their simplicity when determining the frequency and reliability of an association rule ($A \rightarrow C$). Given a set of all transactions $\mathcal{T} = \{t_1, t_2, t_3, \ldots, t_n\}$ in a dataset, the support of an item-set A is defined is the number of transactions satisfied by the item-set, which is considered frequent iff $\{support(A) \geq minimum_{support} | support(A) = |\mathcal{S}|, \mathcal{S} \subseteq \mathcal{T}\}$, $|\mathcal{S}|$ being the number of transactions satisfied by the item-set. Each single item of an association rule is known as a condition of the rule,

Table 1
Sample market basket dataset.

Transactions	Diapers	Beer	Milk
T-1	0	1	1
T-2	1	0	0
T-3	1	1	0
T-4	1	0	0
T-5	0	1	0
T-6	0	1	0
T-7	1	1	1
T-8	1	0	0
T-9	0	0	1
T-10	0	1	1