Information Sciences 258 (2014) 1–28

Contents lists available at ScienceDirect

Information Sciences

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

# QAR-CIP-NSGA-II: A new multi-objective evolutionary algorithm to mine quantitative association rules

D. Martín [a], A. Rosete [a], J. Alcalá-Fdez [b,*], F. Herrera [b]

[a] Department of Artificial Intelligence and Infrastructure of Informatic Systems, Higher Polytechnic Institute José Antonio Echeverría, Cujae, 19390 La Habana, Cuba
[b] Department of Computer Science and Artificial Intelligence, University of Granada, CITIC-UGR, 18071 Granada, Spain

ARTICLE INFO

*Article history:*
Received 24 May 2012
Received in revised form 11 June 2013
Accepted 2 September 2013
Available online 14 September 2013

*Keywords:*
Data mining
Quantitative association rule
Multi-objective evolutionary algorithm
NSGA-II

ABSTRACT

Some researchers have framed the extraction of association rules as a multi-objective problem, jointly optimizing several measures to obtain a set with more interesting and accurate rules. In this paper, we propose a new multi-objective evolutionary model which maximizes the comprehensibility, interestingness and performance of the objectives in order to mine a set of quantitative association rules with a good trade-off between interpretability and accuracy. To accomplish this, the model extends the well-known Multi-objective Evolutionary Algorithm Non-dominated Sorting Genetic Algorithm II to perform an evolutionary learning of the intervals of the attributes and a condition selection for each rule. Moreover, this proposal introduces an external population and a restarting process to the evolutionary model in order to store all the nondominated rules found and improve the diversity of the rule set obtained. The results obtained over real-world datasets demonstrate the effectiveness of the proposed approach.



## 1. Introduction

Association discovery is one of the most common Data Mining (DM) techniques used to extract interesting knowledge from large datasets [34]. Association rules identify dependencies between items in a dataset [65] and are defined as an expression of the type $X \rightarrow Y$, where $X$ and $Y$ are sets of items and $X \cap Y = \emptyset$ [1,2]. Many previous studies for mining association rules focused on datasets with binary or discrete values, however the data in real-world applications usually consists of quantitative values. Thus, designing DM algorithms able to deal with various types of data is a challenge in this field [6,13,36,56,61]. A commonly used method to handle continuous domains in the extraction of association rules is to partition the domains of the attributes into intervals. For instance, an association rule could be $Income \in [1200, 2000] \rightarrow MortgageExpenses \in [360, 600]$. These kinds of rules are known as quantitative association rules (QARs) [56].

In recent years, Evolutionary Algorithms (EAs), particularly Genetic Algorithms (GAs) [23], have been used by many researchers to mine QARs from datasets with quantitative values [4,8]. The main motivation for applying GAs to knowledge extraction tasks is that they are robust and adaptive search algorithms that perform a global search in place of candidate solutions (for instance, rules or other forms of knowledge representation).

Recently, some researchers have presented the extraction of association rules as a multi-objective problem (instead of single objective), removing some of the limitations of the current approaches. Several objectives are considered in the process of extracting association rules, obtaining a set with more interesting and accurate rules [5,33]. In this way, we

* Corresponding author. Tel.: +34 958 241000x46080.
*E-mail addresses:* dmartin@ceis.cujae.edu.cu (D. Martín), rosete@ceis.cujae.edu.cu (A. Rosete), jalcala@decsai.ugr.es (J. Alcalá-Fdez), herrera@decsai.ugr.es (F. Herrera).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.09.009

can jointly optimize measures such as support, confidence, and so on, which can present different degrees of trade-off depending on the dataset used and the type of information that can be extracted from it. Since this approach has a multi-objective nature, the use of Multi-Objective Evolutionary Algorithms (MOEAs) [14,18] to obtain a set of solutions with different degrees of trade-off between the different measures could represent an interesting way of working (by considering these measures as objectives).

In this paper, we propose a new multi-objective evolutionary model to mine a set of QARs with a good trade-off between interpretability and accuracy which maximizes three objectives: comprehensibility, interestingness and performance, understanding by performance the product of Certainty Factor (CF) [54] and support. To accomplish this, the model (called QAR-CIP) extends the well-known MOEA Non-dominated Sorting Genetic Algorithm II (NSGA-II) [19] to perform an evolutionary learning of the intervals of the attributes and a condition selection for each rule, therefore the algorithm is called QAR-CIP-NSGA-II. Moreover, this proposal introduces an external population (EP) and a restarting process to the evolutionary model in order to store all the nondominated rules found and promote diversity in the population. Notice that this proposal follows a dataset-independent approach which does not rely on the minimum support (minSup) and the minimum confidence (minConf) thresholds, which are hard to determine for each dataset.

In order to assess the performance of the proposed approach, we have presented an experimental study using 9 real-world datasets. We have developed the following studies. First, we have compared our approach with the original evolutionary model of NSGA-II in order to analyze the performance of the new components introduced. Second, we have compared the performance of our approach with four mono-objective approaches and three MOEAs to mine QARs. Third, we have shown the results obtained from the comparison with two other classical approaches for mining association rules. Furthermore, in these studies, we have made use of some nonparametric statistical tests for the pairwise and multiple comparison [21,30,29,31] of the performance of these approaches over 22 real-world datasets. Finally, we have analyzed the scalability of the proposed approach.

This paper is arranged as follows. Section 2 introduces a brief study of the existing MOEAs for general purposes [67], some basic definitions of QARs and some quality measures. Section 3 details the evolutionary learning components proposed to mine a set of high quality QARs. Section 4 shows and discusses the results that are obtained over 9 real-world datasets. Section 5 presents some concluding remarks. Finally, Appendix A shows the results obtained by the analyzed algorithms in the 22 real-world datasets considered for the statistical analysis.

## 2. Preliminaries

In this section, we first introduce the basic definitions of QARs and some quality measures. Then, we present a brief study of MOEAs.

### 2.1. Quantitative association rules

Association rules are used to represent and identify dependencies between items in a dataset [34,65]. As we mentioned above, they are expressions of the type $X \rightarrow Y$, where $X$ and $Y$ are sets of items, and $X \cap Y = \emptyset$. There are many previous studies of mining association rules that are focused on datasets with binary or discrete values; however, data in real-world applications usually consist of quantitative values. When the domain is continuous, the association rules are known as QARs, in which each item is a pair attribute-interval. For instance, a QAR could be $Age \in [30,52]$ and $Salary \in [3,4] \rightarrow NumCars \in [3,4]$.

Support and Confidence are the most common measures to assess association rules. These measures for a rule $X \rightarrow Y$ are defined as:

$$Support(X \rightarrow Y) = \frac{SUP(XY)}{|D|} \tag{1}$$

$$Confidence(X \rightarrow Y) = \frac{SUP(XY)}{SUP(X)} \tag{2}$$

where $SUP(XY)$ is the number of patterns of the dataset covered by the antecedent and consequent of the rule, $SUP(X)$ is the number of patterns of the dataset covered by the antecedent of the rule and $|D|$ is the number of patterns in the dataset.

The classic techniques for mining association rules attempt to discover rules whose support and confidence are greater than the user-defined thresholds minSup and minConf. However, several authors have pointed out some drawbacks of this framework that lead it to find many more rules than it should [10,12,55]. For instance, confidence is unable to detect statistical independence or negative dependence between items because it does not take into account the support of the consequent. Moreover, itemsets with very high support are a source of misleading rules because they appear in most of the transactions, and hence any itemset (despite its meaning) seems to be a good predictor of the presence of the high-support itemset.

In recent years, several researchers have proposed other measures to select and rank patterns according to their potential interest to the user [3,12,32,48,50,54]. We briefly describe some of them below.