Neurocomputing 176 (2016) 36–47

Contents lists available at ScienceDirect

Neurocomputing

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

Obtaining optimal quality measures for quantitative association rules

M. Martínez-Ballesteros [a,*], A. Troncoso [b], F. Martínez-Álvarez [b], J.C. Riquelme [a]

[a] Department of Computer Science, University of Seville, Spain
[b] Department of Computer Science, Pablo de Olavide University of Seville, Spain

ARTICLE INFO

Article history:
Received 10 March 2014
Received in revised form
9 October 2014
Accepted 21 October 2014
Available online 8 May 2015

Keywords:
Quality measures
Quantitative association rules
Fitness function
Evolutionary algorithms

ABSTRACT

There exist several works in the literature in which fitness functions based on a combination of weighted measures for the discovery of association rules have been proposed. Nevertheless, some differences in the measures used to assess the quality of association rules could be obtained according to the values of the weights of the measures included in the fitness function. Therefore, user's decision is very important in order to specify the weights of the measures involved in the optimization process. This paper presents a study of well-known quality measures with regard to the weights of the measures that appear in a fitness function. In particular, the fitness function of an existing evolutionary algorithm called QARGA has been considered with the purpose of suggesting the values that should be assigned to the weights, depending on the set of measures to be optimized. As initial step, several experiments have been carried out from 35 public datasets in order to show how the weights for confidence, support, amplitude and number of attributes measures included in the fitness function have an influence on different quality measures according to several minimum support thresholds. Second, statistical tests have been conducted for evaluating when the differences in measures of the rules obtained by QARGA are significative, and thus, to provide the best weights to be considered depending on the group of measures to be optimized. Finally, the results obtained when using the recommended weights for two real-world applications related to ozone and earthquakes are reported.

1. Introduction

The discovery of association rules is an effective computational technique focused on the extraction and representation of meaningful relationships among different variables. It was first defined by Agrawal et al. in [1] but the authors only considered the use of discrete variables. Nonetheless, when the domain of the variables involved in the rule extraction process is continuous, the rules obtained are called quantitative. In this case, they are called quantitative association rules and will hereinafter be referred to QAR. One of the first and most used algorithms is Apriori, which was also proposed by Agrawal et al. [2] one year later.

The mining of QAR is typically associated with the non-supervised learning, in which datasets lack a priori information about the internal structure of data or about how attributes are interrelated. Given this kind of input, the challenging task faced by QAR is to find groups of attributes that exhibit similar behavior. These groups must be formulated as comprehensive rules so that the relationships existing among the attributes can be easily interpreted.

There exist several measures to assess the quality of the QAR. All these measures are conceived to separately evaluate different properties of the rules. In this sense, it is quite common to model the rule extraction process by means of a multi-objective (MO) fitness function. This function aims at jointly maximizing a set of measures and its election depends on the type of rules searched for.

Different strategies can be found in the literature to solve MO problems. On one hand, there exist Pareto-based MO algorithms, which attempt at discovering the best trade-off among conflicting objectives. On the other hand, many fitness functions seeking for the optimization of a single objective can also be found in the literature. In them, some input parameters are required to weight, and therefore favor or penalize, the relevance of the measures. The improvement of certain measures is easily achieved by using these functions, insofar as such measures do not assess conflicting properties.

A vast majority of works focused on QAR mining only include the measures support and confidence in the fitness function. Nevertheless, there are many other measures, as gathered and reported in [17], that are ignored in this configuration. The negative effect that might be caused in other properties is simply not taken into account. All the measures considered in this work as well as their mathematical formulation are summarized in Table 1, where $n(X)$ is the number of occurrences of the itemset X in the dataset and N is the total number of instances in the dataset. *ND*

[*] Corresponding author. Tel.: +34 954556949.
E-mail addresses: mariamartinez@us.es (M. Martínez-Ballesteros), ali@upo.es (A. Troncoso), fmaralv@upo.es (F. Martínez-Álvarez), riquelme@us.es (J.C. Riquelme).

http://dx.doi.org/10.1016/j.neucom.2014.10.100
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Table 1
Measures to assess the quality of QAR.

Measures	Equation	Description	Range
$Sup(X)$	$n(X)/N$	Coverage of X	[0, 1]
$Sup(X \Longrightarrow Y)$	$n(X \cap Y)/N$	Generality of the rule	[0, 1]
$Conf(X \Longrightarrow Y)$	$sup(X \Longrightarrow Y)/sup(X)$	Reliability of the rule	[0, 1]
$Lift(X \Longrightarrow Y)$	$sup(X \Longrightarrow Y)/(sup(X) \cdot sup(Y))$	Interest of the rule • Value <1 : X and Y (ND) • Value$=1$: X and Y (I) • Value >1 : X and Y (PD)	$[0, +\infty)$
$Gain(X \Longrightarrow Y)$	$conf(X \Longrightarrow Y) - sup(Y)$	Implication of the rule	$[-0.5, 1]$
$Certainty\ Factor(X \Longrightarrow Y)$	• If $conf(X \Longrightarrow Y) > sup(Y)$: $(conf(X \Longrightarrow Y) - sup(Y))/(1 - sup(Y))$ • If $conf(X \Longrightarrow Y) < = sup(Y)$: $(conf(X \Longrightarrow Y) - sup(Y))/sup(Y)$	Gain normalized • Value <0 : X and Y (ND) • Value$=0$: X and Y (I) • Value >0 : X and Y (PD)	$[-1, 1]$
$Leverage(X \Longrightarrow Y)$	$sup(X \Longrightarrow Y) - sup(X)sup(Y)$	Strength of the rule • Value <0 : X and Y (ND) • Value$=0$: X and Y (I) • Value >0 : X and Y (PD)	$[-0.25, 0.25]$
$Accuracy(X \Longrightarrow Y)$	$sup(X \Longrightarrow Y) + sup(\neg X \Longrightarrow \neg Y)$	Veracity of the rule	[0, 1]

stands for negatively dependent, *PD* for positively dependent and *I* for independent.

A preliminary study about how the weights, in a fitness function based on a sum of weighted measures, have an influence on the quality measures was provided in [20]. However, this study was reduced as it was applied only for the weights associated to the support and confidence measures.

The main goal of this work is to conduct an extensive study to evaluate the effect of varying different weights for different ranges in a fitness function. Another significant goal is to provide the researcher with several guidelines to set the weights of any fitness function, according to the preset objectives that are wanted to be maximized. Additionally, multiple relationships between the weight variations and the quality measures are inferred in this work. The algorithm selected to perform such tasks, to which different experimental setups have been applied, is QARGA [16].

The remainder of the paper is as follows. Section 2 overviews the most relevant works recently published with techniques using a weighted sum-based fitness function. The QARGA algorithm as well as the design of the experimental setup is described in Section 3. Section 4 provides a description of the datasets. It also includes the setup of the parameters involved in the process. Moreover the analysis itself is reported as well as some statistical tests. Finally, Section 5 summarizes the conclusions drawn from the analysis conducted.

2. Related work

This section is devoted to examine the latest and most relevant works recently published. In particular, those focused on the extraction of QAR by means of fitness functions constructed as a combination of weighted objectives.

Although the optimization of only support and confidence is a usual strategy for defining fitness functions, some works use fitness functions even simpler. Such is the case of EARMGA, an evolutionary algorithm introduced in [28] that only considered the confidence as quality measure to be maximized. It is worth mentioning that the authors did not compute the actual minimum support.

An approach called QuantMiner was introduced in [26]. It used confidence and support as metrics to evaluate its performance. This genetic algorithm was defined in order to provide satisfactory intervals for numeric attributes. The algorithm was compared to GAR [22,23], another algorithm with similar features, to assess the accuracy.

An approach based on an evolutionary algorithm providing an antecedent with a variable number of attributes was published in 2001 [21]. This algorithm, called GENAR, included support, confidence and number of recovered instances in the fitness function. A similar fitness function, but adding the comprehensibility and the amplitude of the intervals to the aforementioned set of attributes, was used in Pachón et al. [24] naming the algorithm GAR-Plus. A comparative analysis of the quality of EARMGA, GAR and EARMGA was presented in [5], where all the three algorithms were applied to two different datasets than those of the original papers. The coverage and their efficiency were reported.

Another genetic algorithm was proposed in 2011 by Soto and Olaya-Benavides [27]. This algorithm included four weighted measures in the fitness function: support, confidence, comprehensibility and interest. A weighted support based on the individual weight of the items according to their importance in the dataset was calculated in [25].

The extraction of QAR can also be found in bioinformatics. For instance, the analysis of microarray gene-expression data by means of QAR based on half-spaces was introduced in [10]. The measures selected to form the fitness function were interestingness, support, confidence and coverage.

All the algorithms mentioned so far only explored positive dependencies. However, it is well-known that negative dependencies can also provide meaningful information. In this sense, a genetic algorithm to extract QAR with positive and negative dependencies was proposed in [3]. To achieve such task, the authors formed a fitness function composed of the five quality measures: support, confidence, number of attributes, recovered records and amplitude. A multiobjective version was introduced two years later by the same authors using, this time, a Pareto-based evolutionary algorithm [4].

Recently, the study of the positive and negative dependencies by means of a multi-objective evolutionary algorithm has been proposed [19]. MOPNAR's fitness function consists of three measures to be optimized: the interestingness, the comprehensibility and what they named performance, a modification of the coverage. The performance and efficiency of MOPNAR was evaluated by comparing the algorithm to up to nine different approaches. One of the main benefits of this algorithm claimed by the authors is its low computational cost. Note that this algorithm extends MOEA/D [14] by introducing two new components into the evolutionary model: an external population and a restarting process. MOEA/D was designed to deal with especially complicated Pareto sets.

The extraction of fuzzy QAR has also been addressed. In particular, the use of Pareto-optimal fuzzy rules as candidate rules