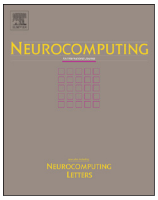
Neurocomputing 126 (2014) 3–14

Contents lists available at ScienceDirect

Neurocomputing

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



Selecting the best measures to discover quantitative association rules

M. Martínez-Ballesteros [a,*], F. Martínez-Álvarez [b], A. Troncoso [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 30 March 2012
Received in revised form
29 November 2012
Accepted 10 January 2013
Available online 27 July 2013

Keywords:
Quantitative association rules
Quality measures
Optimal fitness function
Evolutionary algorithms

ABSTRACT

The majority of the existing techniques to mine association rules typically use the support and the confidence to evaluate the quality of the rules obtained. However, these two measures may not be sufficient to properly assess their quality due to some inherent drawbacks they present. A review of the literature reveals that there exist many measures to evaluate the quality of the rules, but that the simultaneous optimization of all measures is complex and might lead to poor results. In this work, a principal components analysis is applied to a set of measures that evaluate quantitative association rules' quality. From this analysis, a reduced subset of measures has been selected to be included in the fitness function in order to obtain better values for the whole set of quality measures, and not only for those included in the fitness function. This is a general-purpose methodology and can, therefore, be applied to the fitness function of any algorithm. To validate if better results are obtained when using the function fitness composed of the subset of measures proposed here, the existing QARGA algorithm has been applied to a wide variety of datasets. Finally, a comparative analysis of the results obtained by means of the application of QARGA with the original fitness function is provided, showing a remarkable improvement when the new one is used.

1. Introduction

Hybrid artificial intelligent systems are rapidly gaining relevance in the scientific community due to the ability shown to deal with real-life problems [1,13,14]. These systems combine the use of both extracted knowledge and raw data to solve problems.

High volume of data can be stored nowadays; therefore, the use of efficient computational techniques has become a task of the utmost importance. In this context, the discovery of association rules (AR) – and particularly of quantitative association rules (QAR) in this work – is a popular methodology that allows the discovery of significant and apparently hidden relations among variables that form databases [3,4,27,28].

The AR extraction process consists in using a non-supervised strategy to explore data properties. The main goal pursuit is, then, to find groups of attributes appearing frequently together in a dataset, so to provide comprehensive rules able to explain the existing relations among them.

A review of the literature reveals that there exist many algorithms to find AR. Most of them are based on the methods proposed by Agrawal et al. such as AIS [2], Apriori [3] or SETM [26].

Nonetheless, there is another big group of techniques to extract AR that are based on evolutionary algorithms (EA). EA are search algorithms that generate solutions for optimization problems using techniques inspired in natural evolution [18,23], in which a population of abstract representations (chromosomes) of candidate solutions (individuals) evolves toward better solutions. EA can be used to discover AR, since they offer several advantages for knowledge extraction and for rule induction processes [7].

The algorithms that discover AR are normally assessed by means of certain interestingness measures that are able to evaluate the quality of a rule. From all of them, support and confidence highlight although lift, gain, certainty factor or leverage are also indicators that provide useful information about the extracted rules.

A review on AR learning based on the use of EA applied to boolean, categorical, quantitative and fuzzy variables has been described in [16]. However, as this work is focused on quantitative variables only the works using this kind of data are reviewed in this section. Table 1 summarizes the measures used for both evaluation and optimization in several works recently published. From the observation of this table, one conclusion can be easily drawn: There is no uniformity on the selection of measures to assess the algorithms' performance.

For instance, an EA called EARMGA was used in [45] to obtain QAR. The confidence was the only objective to be optimized in the fitness function. To achieve this goal, the authors avoided the

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

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.01.056

Table 1
Quality measures used in the literature.

Algorithm	Quality measures considered					
	Support	Confidence	Re-covered	Comprehensibility or # Attributes	Amplitude	Interest
GENAR [33]	√	√	√			
GAR [34]	√	√	√	√	√	
Tong et al. [43]	√	√				
QuantMiner [40]	√	√				
Kaya and Alhajj [27]	√	√				√
Kaya and Alhajj [28]	√	√				√
Dehuri et al. [15]		√		√		√
Alatas and Akin [5]	√	√	√	√	√	
MODENAR [6]	√	√		√	√	
Orriols-Puig et al. [36]	√	√				
Ayubi et al. [9]	√	√				
EARMGA [45]		√				
Quodmanan et al. [39]	√	√				
NSGA-II-QAR [29]	√	√		√	√	√
GAR-plus [37]	√	√	√	√	√	

specification of the actual minimum support, which can be considered the main contribution of the work.

The combination of confidence and support as only quality measures can be found in several works. Hence, the work introduced in [43] proposes an approach to discover QAR by clustering items of a dataset and projecting the clusters into the domains of the quantitative attributes to form meaningful intervals. Also, the algorithm called QuantMiner [40] proposed a genetic algorithm to mine QAR and optimize support and confidence, by using a fitness function based on the gain measure proposed in [19].

The extraction of QAR has also been applied to the data streams field. A classifier, whose main novelty lied on its adaptability to online gathered data was presented in [36]. By contrast, a multi-objective approach was proposed in [39]. The algorithm did not consider the minimum support and confidence and applied the FP-tree algorithm [25]. The fitness function maximized both support and confidence of the rule. Finally, some works such as [9] have proposed the use of an extended set of operators to mine general association rules and have evaluated the proposal in terms of confidence and support.

Additionally to support and confidence, the authors of the work introduced in [33] used the number of recovered instances to evaluate their approach, called GENAR. GENAR is an EA-based approach capable of obtaining an undetermined number of quantitative attributes in the antecedent of the rule. The same quality measures plus the comprehensibility and the amplitude of the intervals forming the rule were used to evaluate the GAR algorithm [34] (and in its extension [37]). The comprehensibility measure [22] is defined as the logarithm of the number of attributes in the consequent divided by the logarithm of the number of attributes in the rule. The amplitude measure is defined as the addition of the amplitudes for each interval of the attributes which belong to the rule divided by the number of attributes. The authors proposed another EA but, this time, it was necessary to select which attributes formed the antecedent and which one the consequent. Recently, a comparative analysis of the effectiveness in QAR extraction has been presented [7], in which the algorithms GENAR [33], GAR [34] and EARMGA [45] were applied to two different datasets showing their efficiency in terms of coverage and confidence. These five features (support, confidence, recovered, comprehensibility and amplitude) were also evaluated on a multi-objective Pareto-based EA called MODENAR [6]. The same authors proposed an optimization metaheuristic based on rough particle swarm techniques to mine QAR [5]. The fitness function was composed of four different objectives in both works: Support, confidence, comprehensibility of the rule (to be maximized) and the amplitude of the intervals that forms the rule (to be minimized).

Alternatively, the support and confidence have been combined with the interest to form fitness functions in some works [27,28]. Their main particularity lies on the use of genetic algorithms to mine fuzzy association rules. The authors in [29] went one step further and used, in addition to the three measures aforementioned, the amplitude of the intervals as well as the comprehensibility of the rule to form the fitness function.

Finally, the authors in [15] proposed a fast and scalable multi-objective GA for mining AR from large datasets using parallel processing and a homogeneous dedicated network of workstations. The confidence, comprehensibility and interest were the measures maximized.

There is no unanimity in choosing the set of quality measures to be optimized, thus it becomes essential to propose a methodology to automatically select a subset of them whose optimization leads to the optimization of the entire set. Therefore, this work is focused on finding relations among different quality measures in order to determine which measures must be optimized in the fitness function. This way, it is expected that better rules can be extracted, regarding the whole set of measures and not only those included in the fitness function. To fulfill this task, this subset is generated according to a principal component analysis (PCA). The QARGA algorithm [31] has been used to check the new fitness function composed of the selected measures versus the original fitness function based on a weighting scheme that involved several evaluation measures such as support, confidence, number of attributes and amplitude of intervals of the attributes belonging to the rules. In particular, datasets from the public Bilkent University Function Approximation (BUFA) repository [24] have been used. Likewise, four different real-world datasets have been analyzed, specifically, datasets from biological, meteorological and seismological nature.

The remainder of the paper is as follows. Section 2 introduces the foundations underlying QAR. It also explores the most used measures found in the literature as well as some of their inherent