Applied Soft Computing 36 (2015) 519–533

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel approach to adaptive relational association rule mining

Gabriela Czibula, Istvan Gergely Czibula*, Adela-Maria Sîrbu, Ioan-Gabriel Mircea

Department of Computer Science, Babeş-Bolyai University, 1, M. Kogalniceanu Street, 400084 Cluj-Napoca, Romania[1]

ARTICLE INFO

Article history:
Received 21 July 2014
Received in revised form 28 April 2015
Accepted 30 June 2015
Available online 29 July 2015

Keywords:
Data mining
Relational association rule
Adaptive algorithm

ABSTRACT

The paper focuses on the adaptive relational association rule mining problem. Relational association rules represent a particular type of association rules which describe frequent relations that occur between the features characterizing the instances within a data set. We aim at re-mining an object set, previously mined, when the feature set characterizing the objects increases. An adaptive relational association rule method, based on the discovery of interesting relational association rules, is proposed. This method, called *ARARM* (*Adaptive Relational Association Rule Mining*) adapts the set of rules that was established by mining the data before the feature set changed, preserving the completeness. We aim to reach the result more efficiently than running the mining algorithm again from scratch on the feature-extended object set. Experiments testing the method's performance on several case studies are also reported. The obtained results highlight the efficiency of the *ARARM* method and confirm the potential of our proposal.

1. Introduction

It is well known that mining different kinds of data is of great interest in various domains such as medicine, bioinformatics, bioarchaeology, as it can lead to the discovery of useful patterns and meaningful knowledge.

Association rule mining [6] means searching attribute-value conditions that occur frequently together in a data set [25,42,44]. Ordinal association rules [8] are a particular type of association rules. Given a set of records described by a set of characteristics (features or attributes), the ordinal association rules specify ordinal relationships between record features that hold for a certain percentage of the records. However, in real world data sets, features with different domains and relationships between them, other than ordinal, exist. In such situations, ordinal association rules are not powerful enough to describe data regularities. Consequently, *relational association rules* were introduced in [40] in order to be able to capture various kinds of relationships between record features. The *DRAR* method (*Discovery of Relational Association Rules*) was introduced for mining interesting relational association rules within data sets [40].

Relational association rule mining can be used in solving problems from a variety of domains, such as: data cleaning, natural language processing, databases, healthcare, bioinformatics, bioarchaeology, etc. We have previously applied, so far, relational association rule mining in different data mining tasks such as: medical diagnosis prediction [41], predicting if a DNA sequence contains a promoter region or not [15], software defect prediction [16], software design defect detection [17], data cleaning [8].

The method *DRAR* for relational association rule mining starts with a known set of objects, measured against a known set of features and discovers interesting relational association rules within the data set. But there are various applications where the object set is dynamic, or the feature set characterizing the objects evolves. Obviously, for obtaining the interesting relational association rules within the object set in these conditions, the mining algorithm can be applied over and over again, beginning from scratch, every time when the objects or the features change. But this can be inefficient.

In this paper, we propose an adaptive relational association rule method, named *Adaptive Relational Association Rule Mining (ARARM)*, that is capable to efficiently mine relational association rules within the object set, when the feature set increases with one or more features. The *ARARM* method starts from the set of interesting rules that was established by applying *DRAR* before the feature set changed and adapts it considering the newly added features. The result is reached faster than running *DRAR* again from scratch on the feature-extended object set.

We have to mention that the adaptive relational association rule mining method, proposed in this paper, is a novel approach. There exist in the data mining literature approaches which consider the

* Corresponding author.
E-mail addresses: gabis@cs.ubbcluj.ro (G. Czibula), istvanc@cs.ubbcluj.ro (I.G. Czibula), adela@cs.ubbcluj.ro (A.-M. Sîrbu), mircea@cs.ubbcluj.ro (I.-G. Mircea).
URLs: http://www.cs.ubbcluj.ro/ gabis (G. Czibula), http://www.cs.ubbcluj.ro/ istvanc (I.G. Czibula).
[1] Tel.: +40 264 405 327; fax: +40 264 591 906.

http://dx.doi.org/10.1016/j.asoc.2015.06.059
1568-4946/© 2015 Elsevier B.V. All rights reserved.

adaptive association rule mining process for particular problems, but none of them deal with *relational* association rules as in our proposal.

The remaining of the paper is organized as follows. A background on relational association rule mining is given in Section 2. The *Adaptive Relational Association Rule Mining* (*ARARM*) method is described in Section 3. Section 4 presents the experimental evaluation of our approach and shows the efficiency of the proposed method on several case studies. An analysis of the adaptive approach introduced in this paper, as well as a discussion on the obtained results and comparison to related work are given in Section 5. Section 6 outlines the conclusions of the paper as well as directions for further improvements.

2. Background on relational association rule mining

There is a continuous interest in applying association rule mining [39] in order to discover relevant patterns and rules in large volumes of data. Data mining methods [46,2] are applied in various domains such as medicine, bioinformatics, bioarchaeology, software engineering.

In order to be able to capture various kinds of relationships between record attributes, the definition of ordinal association rules from [8,7] was extended in [40] toward *relational association rules*.

In the following we will briefly review the concept of *relational association rules*, as well as the mechanism for identifying the relevant relational association rules that hold within a data set.

Let $R = \{r_1, r_2, \ldots, r_n\}$ be a set of *instances* (entities or records in the relational model), where each instance is characterized by a list of m attributes, $(a_1, \ldots, a_m)$. We denote by $\Phi(r_j, a_i)$ the value of attribute a_i for the instance r_j. Each attribute a_i takes values from a domain D_i, which contains the empty value denoted by ε. Between two domains D_i and D_j relations can be defined, such as: less (<), equal (=), greater or equal ($\geq$), etc. We denote by M the set of all possible relations that can be defined on D_i x D_j and by $\mathcal{A} = \{a_1, \ldots, a_m\}$ the attribute set.

Definition 1. *[40]* A *relational association rule* is an expression $(a_{i_1}\ \ \mu_1\, a_{i_2}\ \ \mu_2\ \ a_{i_3} \ldots \mu_{\ell-1}\ \ a_{i_\ell})$, where $\{a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_\ell}\} \subseteq \mathcal{A}$, $a_{i_j} \neq a_{i_k}, j, k \in \{1 \ldots \ell\}, j \neq k$ and $\mu_i \in M$ is a relation over $D_{i_j} \times D_{i_{j+1}}$, D_{i_j} is the domain of the attribute a_{i_j}. If:

a) $a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_\ell}$ occur together (are non-empty) in s% of the n instances, then we call s the **support** of the rule,
   and
b) we denote by $R' \subseteq R$ the set of instances where $a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_\ell}$ occur together and $\Phi(r\prime, a_{i_j})\ \ \mu_1\ \Phi(r\prime, a_{i_{j+1}})$ is true $\forall\, 1 \leq j \leq \ell - 1$ and for each instance r' from R'; then we call $c = |R'|/|R|$ the **confidence** of the rule.

The *length* of a relational association rule is given by the number of attributes in the rule. Users usually need to uncover interesting relational association rules that hold within a data set; they are interested in relational rules which hold in a minimum number of instances, that are rules with support at least $s_{\min}$, and confidence at least $c_{\min}$ ($s_{\min}$ and $c_{\min}$ are user-provided thresholds).

A relational association rule in R is called *interesting* [40] if its support s is greater than or equal to a user-specified minimum support, $s_{\min}$, and its confidence c is greater than or equal to a user-specified minimum confidence, $c_{\min}$.

In [7] an A-Priori [1] like algorithm, called *DOAR* (Discovery of Ordinal Association Rules), was introduced in order to efficiently find all ordinal association rules (i.e. relational association rules in which the relations are ordinal) of any length, that hold over a data set. The *DOAR* algorithm was proven to be correct and complete and it efficiently explores the search space of the possible rules [7].

The *DOAR* algorithm was further extended in [40,15] toward the *DRAR* algorithm (*Discovery of Relational Association Rules*) for finding interesting relational association rules, i.e. association rules which are able to capture various kinds of relationships between record attributes. The *DRAR* algorithm provides two functionalities: (a) it finds all interesting relational association rules of any length; (b) it finds all maximal interesting relational association rules of any length, i.e. if an interesting rule r of a certain length l can be extended with one attribute and it remains interesting (its confidence is greater than the threshold), only the extended rule is kept.

So far, *relational association rules* were successful in different data mining tasks in domains like: *medicine* (for diagnosis prediction [41]), *bioinformatics* (for predicting if a DNA sequence contains a promoter region or not [15], *software engineering* [16,17], as well as for *data cleaning* tasks [8].

2.1. Example

It is well known that mining medical data (like the example considered in the following) is of great interest in modern medicine, as it can lead to the discovery of useful patterns and knowledge that can be important for the diagnosis and treatment of different diseases. That is why researchers are still focusing on applying data mining techniques to medical data in order to discover interesting patterns [13].

In order to better explain the concept of relational association rules and the *DRAR* algorithm [40] that is used for discovering interesting relational association rules, we give an example that illustrates how it can be applied on a medical data set sample. The data set considered in our experiment is a subset of the breast cancer database obtained from the University of Wisconsin Hospitals, Madison, Dr. William H. Wolberg. The file for this experiment was obtained from [35].

The instances (entities) in this experiment are patients: each patient is identified by 9 attributes [47]. The attributes represent measurements from malignant and benign tumors and have integer values between 1 and 10. Each instance has one of 2 possible classes: *benign* or *malignant*.

In this example we have considered only the first 25 records representing "malignant" instances. The data set considered in our example is given in Table 1.

As all attributes in the experiment have integer values, we have defined three possible binary relations between integer valued attributes: =, <, >.

We executed the *DRAR* algorithm with a minimum support threshold of 1 and a minimum confidence threshold of 0.65. The discovered interesting relational rules are shown in Table 2 and the maximal interesting association rules are given in Table 3. For each discovered rule, its confidence is also provided. The attributes characterizing the instances are denoted by $a_1, a_2, \ldots, a_9$.

Each line from Table 2 expresses a relational association rule of a certain length, which was discovered in the data set indicated in Table 1 with a specified confidence. For example, the first line in Table 2 refers to the relational association rule $a_1 > a_9$ of length **2** (i.e the rule contains two attributes) having a confidence of **0.88**. That is, the value of the attribute a_1 is greater than the value of the attribute a_9 in 88% of instances within the data set (i.e in **22** instances).

As it can be seen in the results above, interesting relational association rules can be discovered within the set of malignant patients. Further analysis of these relational association rules may give relevant information regarding the diagnosis process.