STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LI**, Number 1, 2006

RELATIONAL ASSOCIATION RULES AND ERROR DETECTION

ALINA CÂMPAN, GABRIELA ŞERBAN, AND ANDRIAN MARCUS

Abstract. In this paper we introduce a new kind of association rules, relational association rules, which are an extension of ordinal association rules ([1]). The relational association rules can express various kinds of relationships between record attributes, not only partial ordering relations. We use the discovery of relational association rules for detecting errors in data sets. We report a case study for a real data set which validates this data cleaning approach and shows the utility of relational rules.

Keywords: Data Mining, Relational Association Rules, Data Cleaning.

1. Introduction

Association rule mining techniques are used to search attribute-value pairs that occur frequently together in a data set ([4], [5]).

Ordinal association rules ([1]) are a particular type of association rules. Given a set of records described by a set of attributes, the ordinal association rules specify ordinal relationships between record attributes that hold for a certain percentage of the records. However, in real world data sets, attributes 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, we define *relational association rules* in order to be able to capture various kinds of relationships between record attributes.

Discovering the ordinal rules that hold in a data set was already used for identifying possible errors in that data set ([1]). We apply relational association rules discovery to the same purpose. We provide an example that illustrates the utility

Received by the editors: March, 1, 2006.

2000 *Mathematics Subject Classification.* 68P15, 68U35.

1998 *CR Categories and Descriptors.* H.2.8[**Computing Methodologies**]: Database Applications – *Data Mining*; H.4.2[**Information Systems**]:Information Systems Applications – *Types of systems*;

of discovering relational rules in data. By using ordinal rules discovery not all types of errors that have been discovered using relational rules can be detected.

2. Relational Association Rules

We extend the definition of ordinal association rules ([1]) towards *relational association rules*.

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

Definition 1. A *relational association rule* is an expression $(a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_\ell}) \Rightarrow (a_{i_1}\ \mu_1\, a_{i_2}\ \mu_2\ a_{i_3} \ldots \mu_{\ell-1}\ a_{i_\ell})$, where $\{a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_\ell}\} \subseteq \mathcal{A} = \{a_1, \ldots, a_m\}$, $a_{i_j} \neq a_{i_k}$, $j, k = 1..\ell$, $j \neq k$ and $\mu_i \in \mathcal{M}$ 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 records, then we call s the *support* of the rule,
and
b) we denote by $R' \subseteq R$ the set of records where $a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_\ell}$ occur together and $\Phi(r_j, a_{i_1})\ \mu_1\ \Phi(r_j, a_{i_2})\ \mu_2\ \Phi(r_j, a_{i_3})\ \ldots\ \mu_{\ell-1}\ \Phi(r_j, a_{i_\ell})$ is true for each record r_j din R'; then we call $c = |R'|/|R|$ the *confidence* of the rule.

We call the length of a relational association rule the number of attributes in the rule. The length of a relational association rule can be at most equal to the number of the attributes describing the data.

The users usually need to uncover interesting relational association rules that hold in a data set; they are interested in relational rules which hold between a minimum number of records, that is rules with support at least s_{min}, and confidence at least c_{min} (s_{min} and c_{min} are user-provided thresholds).

Definition 2. We call a relational association rule in R *interesting* 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}.

We developed in [2] an algorithm, called *DOAR* (Discovery of Ordinal Association Rules), that efficiently finds all interesting ordinal association rules, of any

length, that hold over a data set. This algorithm can be used for finding interesting relational association rules, as well.

3. Data Cleaning

Real-world data tend to be incomplete, noisy and inconsistent. Data cleaning refers to detect and correct or remove corrupt or inaccurate records (inconsistencies) from a record set, to fill in missing values or to smooth out noise while identifying outliers ([4]).

We aim to detect and report (not correct) record values that represent potential error in the analyzed data. We proceed in the same manner as for ordinal association rules discovery ([1]):

- We detect all the interesting binary relational rules (rules between two attributes), with respect to the user-provided support and confidence thresholds). Even if the *DOAR* algorithm can be used to discover all the relational rules, of any length, in a data set, we used it to discover only the binary rules. Binary rules are sufficient in order to detect errors in data sets.
- We detect and mark each record value that brokes any of the discovered binary relational rules.
- We report as potential errors those record values marked as possible errors more times than the average.

4. Case Study

For conducting our case study, we used a programming interface, presented in [3] and designed for the discovery of interesting relational association rules. This interface implements the *DOAR* algorithm. Based upon this interface, we developed an error detection application, following the steps described in section 3.

The data set we used in our case study consists in records containing information about students in a university department. There are 2012 records in the data set. Each record is described by the following attributes: StudentID - number, FirstName - text, LastName - text, CNP (Numerical Personal Code) - 13 digits number, BirthDate - date, RegistrationDate - date.

Between these attributes the following semantic relationships must hold: the CNP value must contain the BirthDate value and the BirthDate value must be earlier than the RegistrationDate value for every student record. We want to discover what are the erroneous records in the data set and which attribute value is most likely to be inconsistent with the rest of the record.

There are such data sets for which the semantic of some of the relationships between attributes describing the data are known. Exceptions from these known rules can be easily detected, but is more difficult, when other external information are not available, to establish which of the conflictual data are real errors. When other unknown regularities also exist in the data set, relational rule discovery, used as described in this paper, can help to estimate where an error resides.

We executed the *DOAR* algorithm with minimum support threshold of 0.95 and minimum confidence threshold of 0.93. The algorithm discovered that two binary interesting relational rules hold in the data set, as we expected:

CNP ‘‘=’’ BirthDate (support=0.970, confidence=0.937)

BirthDate $\leq$ RegistrationDate (support=0.970, confidence=0.967)

The difference between the support and confidence values of these two rules indicate that there are small irregularities in data, which represent potential errors. The average rules broken by the record values, as reported by our application, is 1.014. So, every record value that brokes both rules is reported as a potential error.

As the two binary rules discovered in data have only one common attribute, only this attribute values are reported as possible errors. Usually, when there are more relational rules having more common attributes, it is possible that errors to be detected at record values of different attributes.

We report below the potential errors found by our application.

```
s34 (1790521311822, May 24 1979, Jan 01 1978) :  2 errors at BirthDate
    Cnp(1790521311822) ‘‘=’’ BirthDate(May 24 1979);
    BirthDate(May 24 1979) ≤ RegistrationDate(Jan 01 1978);
s34 (1790521311822, May 24 1979, Jan 01 1978) :  2 errors at BirthDate
    Cnp(1790521311822) ‘‘=’’ BirthDate(May 24 1979);
    BirthDate(May 24 1979) ≤ RegistrationDate(Jan 01 1978);
s2572 (1771103062952, Nov 03 1997, Jan 01 1996) :  2 errors at BirthDate
    Cnp(1771103062952) ‘‘=’’ BirthDate(Nov 03 1997);
    BirthDate(Nov 03 1997) ≤ RegistrationDate(Jan 01 1996);
s2572 (1771103062952, Nov 03 1997, Jan 01 1996) :  2 errors at BirthDate
    Cnp(1771103062952) ‘‘=’’ BirthDate(Nov 03 1997);
    BirthDate(Nov 03 1997) ≤ RegistrationDate(Jan 01 1996);
```

5. Conclusions and Further Work

The concept of relational association rules, introduced in this paper, is a generalization of ordinal association rules. Relational rules discovery has a larger applicability, in different application domains where ordinal rules are not powerful enough to express all existing relationships between data attributes.

Further work can be done in the following directions:

- Defining relational association rules that contain repeating attributes; developing a technique similar to *DOAR* for the discovery of such interesting rules.
- Applying discovery of relational association rules in other application domains, such as medical diagnosis.
- Using relational association rules of arbitrary length together with other data mining techniques such as classification or regression to increase the accuracy of the predictive models ([6]). Binary association rules are currently used in building predictive models in e-banking services ([7]).

References

[1] Marcus, A., Maletic, J. I., Lin, K.-I., "Ordinal Association Rules for Error Identification in Data Sets", CIKM 2001, 2001, pp. 589–591.

[2] Campan, A., Serban, G., Truta, T. M., Marcus, A., "An Algorithm for the Discovery of Arbitrary Length Ordinal Association Rules", submitted to DMIN'06.

[3] Serban, G., Campan, A., Czibula, I.G. , "A Programming Interface For Finding Relational Association Rules", submitted to ICCCC 2006.

[4] Han, J., Kamber, M., "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems, 2001.

[5] Tan, P.-N., Steinbach, M., Kumar, V., "Introduction to Data Mining", Addison Wesley, cap. 8,9, 2005.

[6] Hong, S. and Weiss, S., "Advances in predicitve model generation for data mining", IBM Research Report RC-21570.

[7] Aggellis, V., and Christodoulakis, D., "Association Rules and Predictive Models for e-Banking Services", in Proceedings of 1st Balkan Conference in Informatics, Tessaloniki, Greece, 2003.

Babeş-Bolyai University, Faculty of Mathematics and Computer Science, Cluj-Napoca, Romania
E-mail address: alina@cs.ubbcluj.ro

Babeş-Bolyai University, Faculty of Mathematics and Computer Science, Cluj-Napoca, Romania
E-mail address: gabis@cs.ubbcluj.ro

Department of Computer Science, Wayne State University, USA
E-mail address: amarcus@wayne.edu