Ordinal Association Rules for Error Identification in Data Sets

Andrian Marcus
Department of Computer Science
Kent State University
Kent, OH 44242-0001
01 (330) 672- 9039
amarcus@cs.kent.edu

Jonathan I. Maletic
Department of Computer Science
Kent State University
Kent, OH 44242-0001
01 (330) 672-9039
jmaletic@cs.kent.edu

King-Ip Lin
Division of Computer Science
The University of Memphis
Memphis, TN 38152
01 (901) 678- 3135
linki@msci.memphis.edu

ABSTRACT

A new extension of the Boolean association rules, ordinal association rules, that incorporates ordinal relationships among data items, is introduced. One use for ordinal rules is to identify possible errors in data. A method that finds these rules and identifies potential errors in data is proposed.

Keywords

Data Mining, Data Cleansing, Association Rules, Ordinal Rules.

1. INTRODUCTION

The work presented here extends data mining techniques and applies these extensions to the problem of detecting errors in data sets. The types of errors we are trying to detect lie outside the standard integrity constraints. The method presented here aims to uncover relationships (e.g., numerical ordering or equality) between attributes that commonly occur over the data set and then use this information to identify attributes that do not conform to these uncovered (partial) orderings. Few methods exist that directly tackle this problem and our method represent a useful and practical approach. The proposed method is intended to be use in conjunction with existing ones to solve part of the data cleansing problem (i.e., identification of potential errors in data).

2. RELATED WORK

Our work is directly related to association rule mining and therefore, we first outline the work in this area. Next, we discuss previous work on using database/data mining techniques for data cleansing (cleaning).

The term, *association rule* was first introduced by Agrawal et al. [1] in the context of market basket analysis. Association rules of this type are also referred to in the literature as *classical* or *Boolean* association rules. For the purposes of this paper Boolean association rules between single items are considered as basis for subsequent definitions and the rules between item sets are considered generalizations.

The problem of mining association rules from large databases has been subject of numerous studies. Some of them focus on developing faster algorithms for the classical method and/or adapting the algorithms to various situations, like parallel mining and incremental mining. Hipp et al. [6] provides an excellent survey on this topic. Another direction is to define rules that modify some conditions of the classical rules to adapt to new applications (like imposing constraints on item sets, or adapting rules to time-series data). The work of Ng et al. [10] contains a comprehensive list of references related to the above-mentioned studies. Association rules as defined above apply to Boolean or categorical data. Srikant et al. extended the categorical definition to include quantitative data. Such rules are called *quantitative association rules* [12]. A stronger set of rules is defined in [7] as *ratio-rules*. This time the strength of the rule allows multiple applications, including data cleansing and outlier detection. However, the paper does not exploit this idea. In this paper, we focus on rules that describe ordinal relationships among attributes. The work of Guillaume et al. [4], which independently uses the term ordinal rules, is not related, and focuses on the development of ordinal objective measures. Likewise, in [2], the authors produce association rules on ordinal data and their goal is more akin to the quantitative rules mentioned above.

Various techniques have been developed to tackle the problem of data cleansing. For example, [11] proposed the use of an interactive spreadsheet to allow users to perform transformation, while [3] allow users to specify rules and conditions on an SQL-like interface. Apart from general approaches, in many cases there are specific data cleansing problems that need to be solved. The merge/purge problem [5] aims at removing duplicates in data. All the approaches above typically require the user to specify the rules beforehand. While it is reasonable in many cases, it is also important that the data cleansing system be able to automatically discover rules and detect errors. This is the approach of this paper. One must note that the error identification part of the data cleansing problem is difficult and no single method can solve it entirely or completely automatically.

3. ORDINAL ASSOCIATION RULES

The objective here is to find ordinal relationships between record attributes that tend to hold over a large percentage of records. If attribute A is less than B most of the time then a record that contains a B that is less than A may be in error. One flag on B may not mean much, but if a number of such rules that deal with B are broken, the likelihood of error goes up. These considerations lead to a new extension of the association rules – *ordinal association rules* or simply *ordinal rules*. The following more formally defines this concept.

Definition: Let $R = \{r_1, r_2, \ldots, r_n\}$ a set of records, where each record is a set of k attributes $(a_1, \ldots, a_k)$. Each attribute a_i in a particular record r_j has a value $f(r_j, a_i)$ from a domain D. The value of the attribute may also be empty and is therefore included in D. The following relations (partial orderings) are defined over D, namely less or equal (≤), equal (=) and, greater or equal (≥) all having the standard meaning.

Then $(a_1, a_2, a_3, \ldots, a_m)$ Þ $(a_1\, m_1\, a_2\, m_2\, a_3\, \ldots\, m_{m-1}\, a_m)$, where each $m_i \in \{\leq, =, \geq\}$, is a an *ordinal association rule* if:

1. $a_1 \ldots a_m$ occur together (are non-empty) in at least *s%* of the n records, where *s* is the *support* of the rule;
2. and, in a subset of the records $R' \acute{I} R$ where $a_1 \ldots a_m$ occur together and $f(r_j, a_1)\, m_1 \ldots m_{m-1}\, f(r_j, a_m)$ is true for each $r_j \hat{I} R'$. Thus $|R'|$ is the number of records that the rule holds for and the *confidence*, c, of the rule is the percentage of records that hold for the rule $c = |R'|/|R|$.

The work here currently focuses on ordinal rules that use only two attributes. The process to identify potential errors in data sets using ordinal association rules is composed of the following main steps:

a) Find ordinal rules with a minimum confidence *c*;

b) Identify data attributes that broke the rules and which can be considered potential errors.

Here, the manner in which support of a rule is important differs from the typical data-mining problem. We assume all the discovered rules that hold for more than two records represent valid possible partial orderings. Future work will investigate user-specified minimum support and rules involving multiple attributes.

The method first normalizes the data if necessary and then computes comparisons between each pair of attributes for every record. Only one scan of the data set is required. An array with the results of the comparisons is maintained in the memory. Figure 1 contains the algorithm for this step. The complexity of this step is only $O(N*M^2)$ where N is the number of records in the data set, and M is the number of fields/attributes. Usually M is much smaller than N. The results of this algorithm are written to a temporary file for use in the next step of processing.

```
Algorithm compare items.
for each record in the data base (1...N)
    normalize or convert data
    for each attribute x in (1 .. M-1)
        for each attribute y in (x+1 ... M-1)
            compare the values in x and y
            update the comparisons array
        end for.
    end for.
    output the record with the normalized data
end for.
output the comparisons array
end algorithm.
```

Figure 1. The algorithm for the first step.

In the second step, the ordinal rules are identified based on the chosen minimum confidence. There are several researched methods to determine the strength including interestingness and statistical significance of a rule (minimum support and minimum confidence, chi-square test, etc.). Using confidence intervals to determine the minimum confidence is currently under investigation. However, previous work on the data set [8] used in our experiment showed that the distribution of the data was not normal. Therefore, the minimum confidence was chosen empirically, several values were considered and the algorithm was executed. The results indicated that a minimum confidence between 98.8 and 99.7 provide best results (less number of false negative and false positives).

The second component extracts from the temporary file and stores in memory the data associated with the rules. This is done with a single scan of the comparisons file (complexity $O(C(M,2))$. Then for each record in the data set, each pair of attributes that correspond to a pattern it is check to see if the values in those fields within the relationship indicated by the pattern. If they are not, each field is marked as possible error. Of course, in most cases only one of the two values will actually be an error. Once every pair of fields that correspond to a rule is analyzed, the average number of possible error marks for each marked field is computed. Only those fields that are marked as possible errors more times than the average are finally marked as containing high probability errors. Again, the average value was empirically chosen as threshold to prune the possible errors set. Other methods to find such a threshold, without using domain knowledge or multiple experiments, are under investigation. The time complexity of this step is $O(N*C(M,2))$, and the analyzes of each record is done entirely in the main memory. Figure 2 shows the algorithm used in the implementation of the second component. The results are stored so that for each record and field where high probability errors were identified, the number of marks is shown.

```
Algorithm analyze records.
for each record in the data base (1…N)
    for each rule in the pattern array
        determine rule type and pairs
        compare item pairs
        if pattern NOT holds
            then mark each item as possible error
    end for.
    compute average number of marks
    select the high probability marked errors
end for.
end algorithm.
```

Figure 2. Algorithm for the second step.

4. EXPERIMENTS AND RESULTS

Two sets of experiments were executed to date. For the first set of experiments, we used synthetically generated data to validate the algorithms. A set of data with 100 attributes and 10,000 records was randomly generated. Each attribute had a known distribution and range. Then a number of errors were introduced. A number of these errors broke the existing ordering in data and additionally, a number were statistical outliers. Using statistical measures (e.g., means, standard deviation, etc.) some of these

errors were not identifiable. Using the ordinal rules in the manner described above, all of the errors that broke the orderings were identified. By combining the two methods (i.e., identification of statistical outliers and ordinal rules) all the induced errors were detected. The number of false positives and false negatives was in direct correlation with the chosen confidence for ordinal rules. The best value for the confidence is data dependent and we are currently investigating methods to identify this value automatically.

The second set of experiments was performed on real world data supplied by the Naval Personnel Research, Studies, and Technology (NPRST). The data set is part of the officer personnel information system including midshipmen and officer candidates. Many of the attributes represent dates of particular events (e.g., first enlistment, promotion dates, etc.). For the experiment, a subset of this data set was chosen representing an important class of Navy personnel and contained 32,721 records with 226 attributes.

The attributes of type date are the only ones examined. Given that these attributes are all of the same type, the comparison operators make perfect sense and the generated ordinal rules map directly into the problem domain. For instance, all dates in an individual record should be greater then their date of birth. While this may seem an obvious relationship, the use of ordinal rules should automatically uncover such relationships (obvious or not). At this time, the results are under investigation by the domain experts at NPRST.

The results were compared with results of standard statistical outlier detection methods obtained in previous work [8]. These possible errors not only matched many of the previously discovered ones, but also yielded (as expected) a number of errors unidentified by the other methods. The distribution of the data dramatically influenced the error identification process in the previous utilized methods. Ordinal rules are not influenced as much by the distribution of the data and is proving to be more robust.

5. CONCLUSIONS

Association rule mining proves to be useful in identifying not only interesting patterns for fields such as market basket analysis or census data, but also, by extension to ordinal association rules, patterns that uncover errors in other kind of data sets. The classical notion of association rules has been extended to include ordinal relationships between pairs of numerical attributes, thus defining ordinal association rules. This extension allows the uncovering of stronger rules that yielded potential errors in the data set, while keeping the computation simple and efficient. Ordinal association rules bear some similarity with the above-mentioned extensions of Boolean association rules. However, they are better suited to the problem of identifying possible errors in the type of data sets being analyzed for the following reasons:

- They are easier and faster to compute than quantitative association rules or ratio-rules.
- Although they are weaker than quantitative association rules or ratio-rules, they give very good results in the case of finding (partial) ordering trends.
- Distance-based association rules (over interval data) [9] could be also used in this for this problem, but it is inherently hard to find the right intervals in the absence of specific domain knowledge, and the methods tend to be rather expensive.

The results of the current experiments are promising and new ones are in progress to extend the use of the ordinal rules to cope with attributes of different types and to find relationships between rules that involve more than two attributes.

6. ACKNOWLEDGMENT

This research was supported in part by grants from the Office of Naval Research (N00014-99-1-0730) and the National Science Foundation (C-CR 98-18323).

7. REFERENCES

[1] Agrawal, R., Imielinski, T. and Swami, A., Mining Association rules between Sets of Items in Large Databases. In ACM SIGMOD International Conference on Management of Data, (Washington D.C., 1993), 207-216.

[2] Buchter, O. and Wirth, R. Exploration of Ordinal Data Using Association Rules. Knowledge and Information Systems, 1 (4).

[3] Galhardas, H., Florescu, D., Shasha, D. and Simon, E. An Extensible Framework for Data Cleaning, Institute National de Recherche en Informatique ét en Automatique, 1999.

[4] Guillaume, S., Khenchaf, A. and Briand, H., Generalizing Association Rules to Ordinal Rules. In The Conference on Information Quality (IQ2000), (MIT, Boston, MA, 2000), 268-282.

[5] Hernandez, M. and Stolfo, S. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Data Mining and Knowledge Discovery, 2 (1). 9-37.

[6] Hipp, J., Guntzer, U. and Nakhaeizadeh, G. Algorithms for Association Rule Mining - A General Survey and Comparison. SIGKDD Explorations, 2 (1). 58-64.

[7] Korn, F., Labrinidis, A., Yannis, K. and Faloustsos, C., Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. In 24th VLDB Conference, (New York, 1998), 582--593.

[8] Maletic, J.I. and Marcus, A., Data Cleansing: Beyond Integrity Checking. In The Conference on Information Quality (IQ2000), (MIT, Boston, MA, 2000), 200-209.

[9] Miller, R.J. and Yang, Y. Association Rules over Interval Data. ACM SIGMOD, 26 (2). 452-461.

[10] Ng, R.T., Lakshmanan, S., L.V., Han, J. and Pang, A., Exploratory Mining and Pruning Optimizations of Constrained Association Rules. In ACM SIGMOD, (Seattle, Washington, 1998), 13-24.

[11] Raman, V. and Hellerstein, J.M., Potter's wheel: an interactive data cleaning system. In 27th International Conference on Very Large Databases, (Rome, 2001), To appear.

[12] Srikant, R., Vu, Q. and Agrawal, R., Mining Association Rules with Item Constraints. In SIGMOD International Conference on Management of Data, (Montreal, Canada, 1996), 1-12.