Mining Weak Rules

Huan Liu
School of Computing
National University of Singapore
Singapore, 117599
liuh@comp.nus.edu.sg

Hongjun Lu
Department of Computer Science
Hong Kong University of
Science and Technology*
luhj@cs.ust.hk

Abstract

Finding patterns from data sets is a fundamental task of data mining. If we categorize all patterns into strong, weak, and random, conventional data mining techniques are designed only to find strong patterns, which hold for numerous objects and are usually consistent with the expectations of experts. In this paper, we address the problem of finding weak patterns (i.e., reliable exceptions) from databases. They are valid for a small number of objects. A simple approach is proposed which uses deviation analysis to identify interesting exceptions and explore reliable ones. Besides, it is flexible in handling both subjective and objective exceptions. We demonstrate the effectiveness of the proposed approach through a benchmark data set.

1. Introduction

Data mining has attracted much attention from practitioners and researchers in recent years. Combining techniques from the fields of machine learning, statistics and database, data mining works towards finding patterns from huge databases and using them for improved decision making. We categorize patterns into three types: (a) strong patterns - regularities for numerous objects; (b) weak patterns - reliable exceptions representing a relatively small number of objects; and (c) random patterns - random and unreliable exceptions. We argue that weak patterns are an important part for data mining of discovery nature.

Conventional data mining techniques are only designed to find strong patterns which have high predictive accuracy or correlation. This is because we normally want to find such kinds of patterns that can help the prediction task. However, in certain tasks of data mining, we may seek more than predicting: as strong patterns are usually consistent with the expectations of experts, we want to know what is in the data that we do not know yet. Therefore, in some cases, we are more interested in finding out those weak patterns with respect to the strong ones. Usually, such patterns (reliable exceptions) are unknown, unexpected, or even contradictory to what the user believes. Hence, they are novel and potentially more interesting than strong patterns to the user who can then act upon the weak patterns. For example, if we are told that "some kind of jobless applicants are granted credit", that will be more novel and interesting, as compared to "jobless applicants are not granted credit". Moreover, an exception rule is often beneficial for tasks with discovery nature since it differs from a common sense rule which is often a basis for people's daily activity.

There is no sufficient support from current data mining techniques. Most current data mining techniques such as association rules, decision trees cannot effectively support weak pattern mining. Intuitive extensions cannot help either. We cannot simply use the existing data mining techniques to mine weak rules, or simply resort to generate-and-test: generate rules R on data D; remove D' from D that are covered correctly by R; then generate rules R' on $D - D'$, repeat the process until no data is left. Generating all possible rules R_{all} will only produce too many rules: much more than the number of instances of data.

2. Our Approach

A simple yet flexible approach is proposed which uses deviation analysis to find reliable exceptions. Different from previous work [3], we shun searching for strong (or common sense) patterns, directly identify those exceptional instances and mine weak patterns from them. Besides the promised efficiency for weak pattern mining, the proposed method can also handle both subjective and objective exceptions. To demonstrate the effectiveness of this novel method, we apply it to some benchmark data sets [2] and report one case here, and indeed find some interesting exception patterns.

*on leave from School of Computing, National University of Singapore.

0-7695-0368-3/99 $10.00 © 1999 IEEE

Our approach is based on the following observations: (O1) Any exception would have a low support [1] found in the data, otherwise it might be a strong pattern. (O2) A reasonable induction algorithm can summarize data and learn rules. (O3) Attributes in the rules are salient features. Observation (O1) suggests that exceptions cannot be discovered from the data by applying standard machine learning techniques. Observations (O2) and (O3) allow us to focus on the important features so that we are more focused and an efficient method is possible for finding reliable exceptions. Our approach consists of the four phases: P1. Rule induction and focusing for subjective or objective interests (a window of the data is created); P2. Building contingency tables to find negative deviation; P3. Searching for reliable exception candidates; and P4. Obtaining exception rules with common sense and reference rules.

Before we verify the proposed approach, we would like to say a few words about the exceptions found by our approach and those defined in [3]. Briefly, Suzuki suggested that for an exception rule, we should also be able to find its corresponding common sense rule and reference rule. A reference rule should hold low support and low confidence. Due to the nature of negative deviations, it is obvious that every exception rule found by our approach has a corresponding common sense rule (with positive deviation). A reference rule can also be found in our effort. In summary, our work suggests that common sense and reference rules can be obtained based on exception rules.

3. Case Study

We apply our approach to the mushroom data set [2]. It has 22 attributes, each has 2 to 12 possible values, and 8124 instances with binary class. The two types (classes) of mushrooms are *poisonous* or *edible*. We focus on attribute *Stalk-root* and the class attribute here. With the contingency table, we obtain significant deviations as

Deviation analysis of the "Stalk-root = ?" classification.

Stalk-root	Type	x_{ij}	n_{ij}	δ
?	p	1760	1195	+.47
?	e	720	1285	-.44
...				

Using the negative deviation, we create a window to search for exception candidates and find the following.

Attribute-Value	Overall Conf	Chosen
stalk-root = ?	.29	Y
veil-type = p	.52	Y
stalk-shape = e	.46	Y
gill-size = b	.70	Y
bruises = f	.31	Y

We find one weak rule below:

#1 CS: bruises=f, g-size=b, stalk-shape=e $\rightarrow$ class=p
RE: C, stalk-root=? $\rightarrow$ class=e
RR: stalk-root=? $\rightarrow$ class=e

where CS is a common sense rule, RE is a reliable exception, C in RE is the conditional part of CS, and RR is a reference rule. Detailed results on more weak rules will be provided upon request.

4. Conclusion

We propose here a simple approach that enables us to study reliable exceptions with respect to a rule of our interest or attributes specified by a user. The major techniques are deviation analysis, windowing, and conventional mining tools (e.g., Apriori for association rule mining). For the concerned attributes, we first find their negative deviations which determine the window, and then search reliable patterns from the window using any data mining tools we want. Reliable exceptions are those patterns that are only valid or strong in the window.

This approach is efficient because (1) it can work around focused attributes, thus avoiding search all attributes in the database; (2) we are only concerned about negative deviations; (3) it only scans the data once to create the window; and (4) the window size (i.e., number of instances) is usually much smaller than the number of instances in the data. Besides, such an approach is also flexible to handle both subjective and objective prior knowledge. Mining weak pattern is an important area for effective, actionable, and focused data mining.

Acknowledgments

We would like thank Gurmit Singh, Danny Oh, and Soh Seng Guan for implementing some of the programs used for this project. Special thanks go to Kian Sing Ng and Farhad Hussain for their excellent help in this project.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th Conference on Very Large Data Bases*, pages 478–499, Santiago, Chile, September 1994.

[2] C. Merz and P. Murphy. UCI repository of machine learning databases. Technical Report http://www.ics.uci.edu/ mlearn/MLRepository.html.Irvine, CA: University of California, Department of Information and Computer Science, 1996.

[3] E. Suzuki. Autonomous discovery of reliable exception rules. In *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 259–263, Newport Beach, CA, USA., August 1997.