Mining Association Rules for Estimation and Prediction

Takashi Washio, Hiroki Matsuura⋆ and Hiroshi Motoda

Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567, Japan
washio@isir1.sanken.osaka-u.ac.jp

1 Introduction

The standard Basket Analysis derives all frequent itemsets and all association rules having support and confidence levels greater than their thresholds, and filters out trivial rules in statistical sense (1994, 1991). This framework gives comprehensive "*descriptions*" of regularities contained in the data. Another major purpose of data mining is to derive important knowledge for "*estimation and prediction*" on the underlying system which has generated the data (1996). We propose a novel principle to derive association rules for the latter purpose, where the rules provide maximal guesses from minimal facts about the system while maintaining their support and confidence levels as uniform as possible. The principle and its evaluation through real world data are described in the later sections.

2 Criteria for Estimation and Prediction

Given a set of transactions each of which consists of some items, the Basket Analysis derives "*association rules*" having the following form where "*Body*": B stands for an itemset and "*Head*": H another itemset (a superset of B). [2]

$$B \Rightarrow H, \text{ where } B \subset H.$$

The "*support*" values of B and H, i.e., $sup(B)$ and $sup(H)$, are ratios of the number of transactions including each set to the total number of transactions respectively. The "*confidence*" value, $conf(B \Rightarrow H)$, stands for the credibility of the rule, and is defined as a ratio of the number of transactions including H to the number of transactions including B. The itemset having its support value greater than a threshold $l - sup$ is called a "*frequent itemset*"(1997).

The standard Basket Analysis generates all association rules, where its head is a frequent itemset, and its confidence value is greater than another threshold value $s - conf$. Furthermore, some statistical "*rule-filters*" are adopted to the generated rules to extract only the rules "*describing*" interesting features embedded in the objective data (1991).

⋆ Current affiliation is Fujitsu Kansai Communication Systems Ltd.

[2] This representation of association rules is different from the standard notion $B \Rightarrow R$ where $R = H - B$. We use H instead of R for ease of our explanation.

On the other hand, our interest is to establish a novel principle to generate and extract association rules for "*estimation and prediction*" on the system underlying the data. First, we propose the following criteria for the basis.

Support threshold: The head of every association rule must have the support greater than a threshold "*lowest support*": $l - sup$.

Uniform confidence: Every association rule must have a confidence close to but not less than a level "*specified confidence*": $s - conf$.

Maximal estimation: Every association rule must estimate a maximally specific consequence from a minimal fact.

The following definitions are introduced to implement these criteria.

Minimal bodyset : For a specified confidence $s - conf$, if a rule $B \Rightarrow H$ satisfies the following condition, B is said to be a "*minimal bodyset*" of $Head : H$ under $s - conf$.

$$conf(B \Rightarrow H) \geq s - conf \text{ and } conf(B' \Rightarrow H) < s - conf \quad \forall B' \subset B$$

Maximal headset : For a specified confidence $s - conf$, if a rule $B \Rightarrow H$ satisfies the following condition, H is said to be a "*maximal headset*" of $Body : B$ under $s - conf$.

$$conf(B \Rightarrow H) \geq s - conf \text{ and } conf(B \Rightarrow H') < s - conf \quad \forall H' \supset H$$

Maximal estimation rule : For a specified confidence $s - conf$, if $Body : B$ and $Head : H$ of a rule $B \Rightarrow H$ are the minimal bodyset and the maximal headset respectively, $B \Rightarrow H$ is said to be a "*maximal estimation rule*".

The maximal estimation rule satisfies the aforementioned criteria.

The maximal estimation rules contain some redundancy in terms of the estimation and prediction. We apply a logical "*rule-filter*" where the rule $AB \Rightarrow ABR$ is removed, when two maximal estimation rules

$$AB \Rightarrow ABR \text{ and } B \Rightarrow BCR$$

are obtained. Here, every intersection among A, B, C, R is empty, and $AB = A + B, ABR = A + B + R$ and $BCR = B + C + R$. This rule-filtering does not violate the criteria of support threshold and uniform confidence.

3 Evaluation through call tracking data

The practical performance of our proposing framework has been evaluated in comparison with the standard approach (1994, 1991). The data used are a set of real call tacking data acquired in a telecommunication company in which each call of a person is recorded as a transaction (1996). Total number of the transactions involved is 65,525, where each item stands for calling date, calling type (oral, cellular, data, and busy), names of cities of calling and called persons, marital status and sex of those persons. Totally 221 types of items appear.

Table 1 shows the results of the evaluation. The numbers of generated rules before the rule-filtering indicated in the second column show the efficient reduction of the rules in our method because it enumerates the maximal estimation rules only. The numbers of filtered rules in the third column are also smaller in our method. The forth column shows the ratio of the average cardinalities of the bodyset and the headset, i.e., [*ave. card. of headsets*]/[*ave. card. of bodysets*], in each method. We have also checked the union of bodysets and the union of headsets of all rules for each method in every condition, and found the complete agreement between our method and the standard. These results indicate the higher ability of each rule derived in our method for estimation and prediction. The fifth column indicates the average and the standard deviation of the rule confidence in each rule set. The values are lower and closer to $l - conf$ in our method, and this shows an advantage that the consequences of all rules can be accepted under uniform statistical belief levels.

Table 1. Performance of rule derivation

$s - conf$	Upper:Conventional method, Lower:Our method			
	generated	filtered	card. ratio	ave. & std. of conf.
90.0%	3,695	578	2.90	100% ± 0%
	2,658	104	3.59	100% ± 1%
70.0%	4,954	593	2.75	96% ± 9%
	2,046	141	2.86	83% ± 9%
50.0%	7,470	455	2.76	89% ± 17%
	3,211	161	3.01	61% ± 10%
30.0%	11,691	160	2.93	77% ± 26%
	2,915	136	3.08	41% ± 14%

$l - sup$ is fixed to be 1.0%

4 Conclusion

A novel framework to mine association rules dedicated to estimation and prediction on the system underlying the data has been proposed in this paper. Its evaluation indicates 1) efficient reduction of redundant rules, 2) high ability and 3) uniform confidence for estimation and prediction. Our framework is expected to provide a new and effective tool for data mining in practical fields.

References

G. Piatestsky-Shapiro.: Discovery, analysis, and presentation of strong rules. Knowledge Discovery in Databases. AAAI/MIT Press (1991)

R.Agrwal and R.Srikant.: Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference,* (1994) 487–499

R.Kimball.: The Data Warehouse Toolkit. Wiley Computer Publishing (1996)

U.Fayyad , G.Piatesky-Shapiro and P.Smyth.: From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE FALL, (1996) 37–54

R.Srikant, Q. Vu and R.Agrwal.: Mining Association Rules with Item Constraints. In *Proc. of 3rd Conference on Knowledge Discovery and Data Mining,* (1997) 67–73

This article was processed using the LaTeX macro package with LLNCS style