Association Rule Hiding Based on Evolutionary Multi-Objective Optimization by Removing Items

Peng CHENG, Jeng-Shyang PAN
Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong, 518055, China
E-mail: {chengp.mail, jengshyangpan}@gmail.com

Abstract

Today, people benefit from utilizing data mining technologies, such as association rule mining methods, to find valuable knowledge residing in a large amount of data. However, they also face the risk of exposing sensitive or confidential information, when data is shared among different organizations. Thus, a question arises: how can we prevent that sensitive knowledge is discovered, while ensuring that ordinary non-sensitive knowledge can be mined to the maximum extent possible. In this paper, we address the problem of privacy preserving in association rule mining from the perspective of multi-objective optimization. A new hiding method based on evolutionary multi-objective optimization (EMO) is proposed and the side effects generated by the hiding process are formulated as optimization goals. EMO is used to find candidate transactions to modify so that side effects are minimized. Comparative experiments with exact methods on real datasets demonstrated that the proposed method can hide sensitive rules with fewer side effects.

Introduction

Data often need to be shared among different organizations during business collaboration in order to gain more reciprocal interests. People can utilize data mining techniques to extract useful knowledge from the shared large data collection. However, despite its benefits to business decision making, data mining technology could also pose the threat of disclosing sensitive knowledge to other parties. To address this issue, a feasible solution is to modify the original database in some way so that the sensitive knowledge can not be mined out. In this paper, we focus on privacy preserving in association rule mining. Modification could lead to non-sensitive rules also to be concealed. The challenge is how to hide the sensitive rules while the non-sensitive ones still can be mined out in the modified database to the largest extent possible.

Atallah et al. (Atallah et al. 1999) first proposed the protection algorithm for data sanitization and proved the optimal solution to this problem is NP-hard. Dasseni (Dasseni et al. 2001) and Verikios (Verikios et al. 2004) extended the itemset hiding to association rules and proposed three heuristic hiding approaches, i.e., algorithm *1.a*, *1.b*, *2.a* and *2.b*. These approaches hide sensitive rules by deleting or inserting items to decrease the supports or confidences of sensitive rules below the specified thresholds. Amiri (Amiri 2007) proposed heuristic algorithms to hide itemset (not rules) by removing transactions or items, in terms of the number of sensitive and non-sensitive itemsets related. Although the rule hiding problem well beholds the characteristic of multi-objective optimization, as far as we know, there is no related work to solve this problem from a multi-objective optimization point view.

In view of this, we adopted the evolutionary multi-objective optimization (EMO) algorithm to solve this problem. The side effects were formulated as optimization goals to be minimized. The model we adopted to modify database and hide rules was to remove selected items in identified transactions which support sensitive rules, so that sensitive rules could escape the mining in the modified database at some predefined thresholds.

The main contribution of this paper is as follows. First we took the rule hiding problem as a multi-objective optimization process and adopted the EMO method to solve it for the first time. Secondly, compared with deterministic methods, the proposed hiding approach based on EMO can hide all sensitive rules with fewer side effects in most cases at the cost of more running time.

Problem Formulation

The rule hiding problem can be formulated as follows.

- D: The original transactional database.
- *MST*: the minimum relative support threshold.
- *MCT*: the minimum confidence threshold.
- R: the set of rules mined from D with given *MST*, *MCT*.
- R_S: a set of sensitive rules to be hidden, and $R_S \subset R$.

The hiding problem is to transform D into a sanitized database D' such that only the rules belong to $R \backslash R_S$ can be mined from D'. Let R' denote the strong rules mined from sanitized database D'.

There are three possible side effects after transforming D into D'. The sensitive rules subset which is not hidden in the modified database D' is called as S-N-H (Sensitive rules Not Hidden). S-N-H $= \{r \in R_S \mid r \in R'\}$. Some of non-

Dataset	*MCT*	|*R*|	Side effects: (|S-N-H|, |N-S-L|, |S-F-G|)					
			NSGA-II -RH	SMS-EMO -RH	1.a	1.b	2.a	2.b
Mushroom *MST*=5%	0.6	849	(0,7,1)	(0,7,1)	(0,7,0)	(0,16,1)	(0,9,1)	(0,9,1)
	0.7	678	(0,10,0)	(0,9,0)	(0,10,0)	(0,13,0)	(0,12,0)	(0,12,0)
	0.8	560	(0,4,0)	(0,5,0)	(0,10,0)	(0,5,0)	(0,5,0)	(0,6,0)
	0.9	461	(0,2,0)	(0,3,0)	(0,12,0)	(0,4,0)	(0,4,0)	(0,10,0)
BMS-WebView-1 *MST*=0.1%	0.3	325	(0,3,0)	(0,4,0)	(0,7,0)	(0,4,0)	(0,4,0)	(0,11,0)
	0.4	131	(0,1,0)	(0,1,0)	(0,4,0)	(0,1,0)	(0,1,0)	(0,5,0)
	0.5	34	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,5,0)
	0.6	11	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,1,0)
BMS-WebView-2 *MST*=0.3%	0.3	482	(0,7,0) (0,6,1)	(0,12,0) (0,6,1)	(0,9,0)	(0,18,0)	(0,12,0)	(0,13,0)
	0.4	283	(0,5,0) (0,4,1)	(0,8,0) (0,4,1)	(0,8,0)	(0,13,0)	(0,9,0)	(0,10,0)
	0.5	112	(0,6,0)	(0,6,0) (0,3,2)	(0,8,0)	(0,6,1)	(0,6,1)	(0,7,1)
	0.6	29	(0,0,0)	(0,0,0)	(0,2,0)	(0,2,0)	(0,2,0)	(0,3,0)

Table 1. Comparison of side effects with different *MCT*s to hide 5 sensitive rules

sensitive rules are falsely hidden and lost/missing in the modified database *D*', which is denoted as N-S-L (Non-Sensitive rules Lost). N-S-L= $\{r \in R_N \mid r \notin R'\}$.In addition, some rules falsely generated in sanitized database *D*' is marked as S-F-G (Spurious rules Falsely Generated). S-F-G= $\{r \in R' \mid r \notin R\}$. Thus, the sensitive rules hiding task can be formulated as a multi-objective optimization problem:

Minimize f = [| S-N-H |, | N-S-L |, | S-F-G |]

The EMO-based Hiding Method

The sensitive rules are hidden by removing items to decrease their support or confidence below the thresholds *MCT* or *MST*. We need to solve two problems before database modification.

1) Find suitable transactions to be modified in the database.
2) Determine which items to be removed in each identified transaction.

We adopted the selection mechanism of the NSGA II algorithm (Deb et al. 2002) and SMS-EMO (Beume et al. 2007) respectively to solve the first problem. We named the two EMO-based methods as NSGA-II-RH and SMS-EMO-RH. For the second problem, the selected item to remove is the one which corresponds to the consequent part of the sensitive rule in the identified transaction and beholds the highest support/frequency.

In order to improve the quality of population, two specific devised solutions in the first generation are created by the heuristic way to accelerate the convergence speed.

Performance Evaluations

We tested the proposed algorithm on three representative real databases. The population size is 40 and the maximum generation is 100. 5 sensitive rules were selected randomly for each dataset to perform hiding task. Table 1 shows the experiment result on the above three datasets with various *MCT*s. We compared the two EMO-based hiding approaches with four heuristic methods proposed by Dasseni (Dasseni et al. 2001) and Verikios (Verikios et al. 2004). As indicated in Table 1, the EMO-based method could achieve better results in most cases. Note that *1.a* used the strategy of adding items to hide rules. All other methods adopted the strategy of removing items.

Conclusion

In this paper, we propose an EMO-based method to solve the association rule hiding problem. NSGA-II and SMS-EMO were utilized respectively to drive the evolution process forward. Comparative experiments demonstrated that EMO-based methods can effectively hide all sensitive rules with fewer side effects.

References

E. Dasseni, V.S. Verykios, and et al. 2001. Hiding association rules by using confidence and support. In: Proceedings of the 4th International Workshop on Information Hiding, pp. 369–383.

V.S. Verykios, A.K. Elmagarmid, and et al. 2004. Association rule hiding. IEEE Transactions Knowledge and Data Engineering 16(4): 434-447.

M. Atallah, E. Bertino, and et al. 1999. Disclosure limitation of sensitive rules. In: Proceedings of IEEE Workshop on Knowledge and Data Engineering Exchange. Chicago, IL, pp. 45–52.

K. Deb, A. Pratap, S. Agarwal, T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation 6 (2): 182–197.

N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA multiobjective selection based on dominated hypervolume. European Journal of Operational Research, 181(3):1653-1669, 2007.

Ali Amiri. 2007. Dare to share: Protecting sensitive knowledge with data sanitization. Decision Support Systems 43(1): 181-191.