International Journal of Engineering Innovation & Research
Volume 1, Issue 6, ISSN: 2277 – 5668

A Novel Approach for Information Hiding Using Data Mining

Sonali Soni
AISECT University Bhopal

Pradeep Chouskey
Deptt. of Computer Engineering
TIT, Bhopal

Gajendra Vaiker
AISECT University Bhopal

***Abstract* – Digital representation of media facilitates access and potentially improves the portability, efficiency, and accuracy of the information presented. Undesirable effects of facile data access include an increased opportunity for violation of copyright and tampering with or medication of content. The motivation for this work includes the provision of protection of intellectual property rights, an indication of b content manipulation, and a means of annotation. Data hiding represents a class of processes used to embed data, such as copyright information, into various forms of media such as image, audio, or text withal minimum amount of perceivable degradation to the "host" the embedded data should be invisible and inaudible to a human observer.**

***Keywords* – Support and confidence, Apriori Algorithm, combined transaction, Evaluation table.**

I. INTRODUCTION

Rule Hiding by Reducing the Confidence

Two approaches for rule hiding using confidence reduction are given, [10]. The first approach is based on replacing 1's by "?" marks, while the second approach replaces 0's with "?" marks. It is important to have these two different approaches for the safety of the rule hiding. If only the first approach is used , then it would be obvious that all the "?" marks are actually 1's. Therefore, the two approaches should be used in an interleaved fashion for rule hiding via confidence reduction. The simplest way of Interleaving could be to hide the first half of sensitive rules by the first approach and the second half using the second approach. The first algorithm shown in hides a sensitive rule r by decreasing the support of the generating itemset of r. The difference between this and the approach presented is that items in the consequent of r only are chosen for hiding. This is due to the fact that by placing a "?" mark for the items in the antecedent of a rule r will cause the minsup(lr) (lr is the left hand side of the rule r) to decrease, leading to an increase in the maxconf(r), and in this the rule hiding process which tries to decrease confidence values of sensitive rules. The hiding process goes on until the minsup(r) or the minconf(r) goes below the MST and MCT thresholds by SM. The algorithm Elmagarmid 2001,. In this method of rule hiding, the transactions that do not fully support are considered rr. Otherwise, by replacing 0 values for the items in lr in the transactions that partially support lr and fully support rr, we will increase the maxsup(r) leading to an increase in the maxconf(r) which is not desirable. The transaction that partially supports lr while supporting the maximum number of items in lr is choosen. In the best case, such a transaction will support |lr|- 1 of the items in lr and in this situation only one of the 0 values will be replaced by a "?" mark, achieving in this way the desired increase in the confidence while making the minimum change on the rest of the rules.

II. PROBLEM DESCRIPTION

Sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy, as we will indicate. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem [6].

III. SOLUTION DOMAIN

The experiments are pursued on both synthetic and real data sets. The synthetic data sets which we used for our experiments were generated using the procedure described in We refer readers to it for more details on the generation of data sets. We report experimental results on two synthetic data sets. In this data set, the average transaction size and average maximal potentially frequent item set size are set to 4 and 5, respectively, while the number of transactions in the dataset is set to 100 K. It is a sparse dataset. The frequent item sets are short and not numerous. The second synthetic data set we used is. The average transaction size and average maximal potentially frequent item set size are set to 20 and 25, respectively. There exist exponentially numerous frequent item sets in this data set when the support threshold goes down. There are also pretty long frequent item sets as well as a large number of short frequent item sets in it. It contains abundant mixtures of short and long frequent item sets[8].

IV. PROPOSED ALGORITHM

Algorithms for Rule Hiding

Given a rule A=> B, the confidence in terms of its support as follows:

Conf (A=>B) = Supp(A∪B)/Supp(A)

According to the relationship between the confidence and the support of a rule, there are two strategies to hide a rule.

1. Decrease the confidence of the rule

(a) By increasing the support of the rule antecedent A, through transactions that partially support it.

(b) By decreasing the support of the rule consequent B, in transaction that support both A and B.

2. Decrease the support of the rule

By decreasing the support of either the rule antecedent A, or the rule consequent

The first one focuses on reducing the minimum confidence of the rules. The second one focuses on hiding the rules by reducing the minimum support of the itemsets that generated these rules (i.e., generating itemsets). Support hiding is adequate against an association rule mining algorithm that uses support pruning to reduce the search space of rules which is usually the case for the currently available commercial products. However, algorithms that can efficiently extract high confidence rules without support pruning have recently been developed.

V. TECHNIQUES FOR DATA HIDING

Note that data hiding, while similar to compression, is distinct from encryption. Its goal is not to restrict or regulate access to the host signal, but rather to ensure that embedded data remain inviolate and recoverable. Two important uses of data hiding in digital media are to provide proof of the copyright, and assurance of content integrity. Therefore, the data should stay hidden in a host signal, even if that signal is subjected to manipulation as degrading as altering, resembling, cropping, or loss data compression. Other applications of data hiding, such as the inclusion of augmentation data, need not be invariant to detection orremoval, since these data are there for the beneath of both the author and the content consumer. Thus, the techniques used for data hiding vary depending on the quantity of data being hidden and the required invariance of those data to manipulation. Since no on method is capable of achieving all these goals, a class of processes is needed to span the range of possible reapplications. The technical challenges of data hiding are formidable. Any “holes” to ill with data in host signal, either statistical or perceptual, are likely targets for removal by loss signal compression. The key to successful data hiding is the ending of holes that are not suitable for exploitation by compression algorithms. A further challenge is to these holes with data in a way that remains invariant to a large class of host signal transformations.

V. CONCLUSION

In this work, the database privacy problems in data mining have been discussed and an algorithm for hiding sensitive data in association rules mining is proposed. Form the experiment it clear that compressed transaction data set methods is much better as compare to the other methods a more efficient approach, called Mining Merged Transactions with the Quantification Table is proposed, which can compress the original database into a smaller one and perform the data mining process efficiently.

REFERENCES

[1] S. A. Raut, S. R. Sathe, and A. Raut, “Bioinformatics: Trends in GeneExpression Analysis,” proceedings of 2010 International ConferenceOn Bioinformatics and Biomedical Technology, 16-18 April 2010, Chengdu, China.

[2] S. A. Raut, S. R. Sathe, and A. P. Raut, “Gene Expression Analysis-AReview for large datasets,” Journal of Computer Science andEngineering, vol.4, Issue 1, November 2010.

[3] S.Prakash, R.M.S.Parvathi., An Enhanced Scaling Apriori for Association Rule Mining Efficiency. European Journal of Scientific Research,ISSN 1450-216X Vol.39 No.2 (2010), pp.257-264

[4] Praveen Ranjan Srivastava and Tai-hoon-Kim “Application of Genetic Algorithm in Software Testing” in IJSE,2009

[5] “ Sensitive Items In Privacy Preserving Association Rule Mining” K Duraiswamy and H.Maheswari , Journal of Information & Knowledge 2008.

[6] Akhilesh Tiwari, R. K. Gupta, D.P. Agrawal, Mining Frequent Itemsets Using Prime Number Based Approach. In Proc. 3rd International Conference on Advanced Computing and Communication Technologies (ICACCT), India, November 08-09, 2008, pp: 138-141.

[7] Y-H. Wu, C.M. Chiang, A.L.P. Chen, “Hiding Sensitive Association Rules with Limited Side Effects,”IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 29–42, Jan. 2007

[8] Hillol Kargupta, Kun Liu, Souptik Datta, and Jessica Ryan KrishnamoorthySiva Kumar:” Homeland Security and Privacy Sensitive Data Mining from Multi-Party distributed Resources” The IEEE International Conference on FuzzySystems pages 727-764, 2003