Cloud Framework for Association Rule Mining

Manoj Kumar, Peeyush Varshney
Department of Computer Engineering, Delhi Technological University
Delhi, India

***Abstract*—**In this paper, we propose model which has statedtheunaccustomed competency to figure out associations in large data sets. There is the need to justify the privacy of the disclosed data with the permissible needs of the data users. One methoddefined as sensitive if its acknowledgment risk is above a certain privacy threshold. Sensitive rules should not communicate to the public, since among other things, they may be used for driving sensitive data, or they may yield business competitors with a leverage.

Keywords*—Data Privacy, Association Rules, Hiding Techniques.*

I. INTRODUCTION

Data mining is the use of analyzing database and compute of sorting patterns in massive data sets requiring methods at the intersection of machine learning, artificial intelligence,database systemsandstatistics. It is the practice of discovering useful information or knowledge from large databases. Data mining has evolved as important techniquesforvast databases. Data mining applications are marketing, business,medical analysis, quality improvement,products control and scientific research etc.

APPLICATIONS

A. Crime and Anti-terrosim Agencies[1]

Total Information Awareness (TIA) is the name of a great U.S. data-mining project focused on scrutinizing financial,travel,communicational and other data from public and private sources with the target of observing and impeding transnational warning to national security. TIA has also called Terrorism Information Awareness. The program has developed for the Homeland Security Act and, has managed by the Defense Advanced Research Projects Agency (DARPA). The basic idea was to gather as much information as possible about everyone, refine through it with massive computers, and examine patterns that might indicate terrorist plots.

B. Service Providers

Service providers is the correct sample of Data Mining techniques and Business Intelligence to consumer based enterprises. Such enterprise implement these techniques and make out of better result. They uses customer's information, collect data of users then organize accordingly as it needs. Given user parameter makes one index that can utilize directly further.Aim is not to drop weak nodes, which found lower index during of this exercise, to overcome provide recommended services and other rewards to customer. Enterprise's objective is to develop higher sell using mined rules, proposed method will help in doing the same, and it gives opportunity to choose better one.

C. E-Commerce

There are many use-cases of Data Mining, which used in commercial web sites. Many enterprisesdevelop Data Mining techniques and Business Intelligence to persuadebetter deal to consumer and grow business through their websites. One of the most famous of these enterprise is, of course, Amazon, that uses experienced mining techniques to control their 'Costumer who visited that product on websites, also interested this' functionality.

Privacy-preserving data mining has appeared to targetabove-mentioned problem. One method is to modify the data before transferring it to the data miner. The second method is to distribute data between two or more places, and those members collaborate to assimilate the global data-miningoutcome without disclosing the data at their individual member. Members togetherconstruct a decision tree without either memberknowing anything about the other member's data, except what might release through the ultimate decision tree. Objective is to provide privacy in data mining. We devise and implement a model for data hiding to preserve this privacy by storing data on a cloud.

II. THEORETICAL BACKGROUND AND LITERATURE SURVEY

We have identified two broad implementation areas of Privacy preservation in data mining namely,

1. Secure Multiparty Computation (SMC)
2. Data Hiding

Association rules are constructed using if-then blocks that assist bare relationships between seemingly not associated data in a relational database or other information warehouse. Association rule is repercussion of form A⇒ B, where the left side statement isknown as premise and it depict a condition, which must be true, for the right side statement(conclusion) to carry. A rule A⇒B can interpret as "If A happens, then B happens." There is no way to tell which rule is more effective; it is impractical to differentiate them. To get past this restriction, we can add various classifiers to the rule, which will describe the robustness of the rule. The two classical measures are:

1. Support is anindication, which explain how often the rule requested. It is a proportion of all transaction, where the items in the rule detected.

2. Confidence is a proportion of all transactions, which contain items on the left and on the right side of the rule.

A. *Security Multiparty Compulatation*

Secure Multiparty Computation is an ability for Privacy Preserving Data Mining in which several parties perform a joint computation and each party only gets the results of computation without knowing the inputs from other parties.Security must be maintain in the face of adversarial behavior by some of the contributor, or by an external member. Each organization knows nothing except the final computation results. The fundamental approach of Secure Multiparty Computation is that a result is secure if at the end of the assessment, no participants knows anything except its own input and the results. One way to view this is supposed to be a trusted third party. Secure Multiparty Computation (SMC) suggested in the literature has proven communication overheads. We use the paradigm of data hiding, which permits the data to divide into variouspart and processed separately at different servers. Using the protocol of data hiding, enables me to design a provably secure, cloud computing based solution, which has insignificant disclosure overhead compared to SMC.

B. *Data hiding*

Data mining approachextensively apply in various applications. Thewrong way of using these methods may result to the leakage of confidentialinformation. However, unwanted side effects, e.g., non-sensitive rules incorrectly masked and bogus rules inappropriately generated in the rule hiding process. Iproposed a novel technique that deliberatelyrefines a few transactions in the transaction database to decrease the supports or confidences of sensitive rules without generating the side effects. Since the relationshipamong rules make it impractical to reach this target, in this paper, I present heuristic methods for increasing the number of hidden sensitive rules and reducing the number of modified entries. The experimental results show the effectiveness of this approach, i.e., unwanted side effects have eluded in the rule hiding process. The good adaptability of this technique in terms of database size and the effectiveness of the relationship among rules on rule hiding noticed.

C. *Hiding Techniques*

Consider a database D, a set R of applicable rules that mined from D and a subset RH (sensitive rules) of R. We have to convert D into a modified database D' so that mining can still be possible on the rules in R. Two main approaches for implementing the above are:

1. We can either block the rules in RH to generate by hiding the frequent item sets from which they have generated.

2. We can decrease the confidence of the sensitive rules by minimizing it below a user-specified limit (min_confi).

On changing the unanalyzed database, there will be negative impacts that can classify into two parts:

1. Desired rules have been vanished.

2. New (unwanted) rules have constructed artificially.

Accuracy of the hiding technique will depend on how it hides all sensitive rules in less time complexity along with reducing these negative impacts.

We have shown a comparative study of the two rules hiding algorithms namely ISLF and MDSRRC

Disadvantages of ISLF

- In this algorithm, I am adding items that are present in antecedents of sensitive rules, in the transactions that does not support these sensitive rules.
- Thus, support of antecedent increases and in turn confidence of the rule decreases.
- This will lead to false rule generation.
- There will be chances that same antecedents are also present in some useful rules. Thus, useful rules will also be lost

Advantages of MDSRRC

- Sensitive rules are hidden more efficiently.
- No false rule generation
- Sensitive rules are decided by the user/database owner instead of deciding it on the basis of support and confidence.
- Less time complexity.
- Less modification done in database as deletion perform after analyzing all sensitive rules.

III. PROPOSED MODEL AND WORK

We have discussed Apriori Algorithm for Association Rule Mining and three Data Hiding techniques. Apriori Algorithm [3]used to generate association rules for a given data set efficiently by pruning the irrelevant sets. The Data Hiding techniques used to modify the database so that the sensitive rules cannot mine and hence privacy is preserved

A. *Approach to Association Rule Mining*

1. Frequent Item set Generation – Generate all item sets whose support ≥ minsup (Threshold support)

2. Rule Generation – Generate high confidence rules from each frequent item set, where each rule is a binary partitioning of a frequent item set.

Suppose there are 'd' number of items. Generate all possible subsets of an item sets, excluding the empty set (2d - 1) and use them as rule consequents (the remaining items form the antecedents).Select rules with high confidence (using a threshold). So for given d unique items:

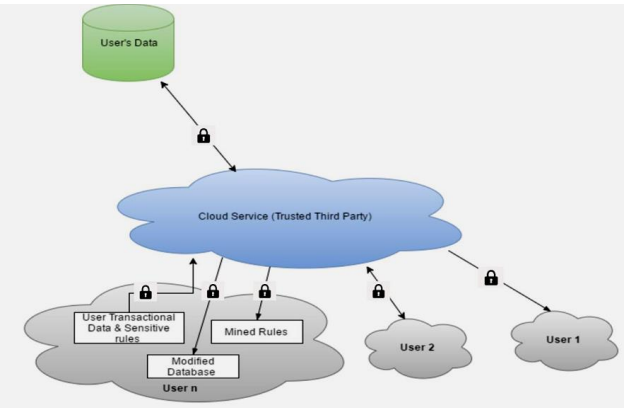
Total number of possible association rules [3]:

$$R = \sum_{k=0}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$

$$= 3^d - 2^{d-1} + 1$$

Apriori algorithm is used as pruning technique to reduce total number of item sets while dealing with large data items.

We have designed a software as a Service (SaaS) cloud model. It is a web service, which provides a user friendly GUI with interactive and easy to use interface features. The cloud model's web service has an HTML front end for UI and Java back end, which runs the various algorithms.



B. Strength of Model

a) Easy accessibility of the cloud model from anywhere and everywhere

b) Hassle-free computation of data to obtain desired results.

c) Increased availability of resources like storage capacity and computing power.

d) Large database can be easily stored on the cloud server.

e) Data integrity is provided by this model.

C. Performance Evaluation Parameter

We have compared and evaluated the two algorithms based on following parameters.

1. Hiding Failure [10]: This measure quantifies the percentage of the sensitive patterns that remain disclosed in the sanitized dataset.

$$\text{Hiding Failure} = \frac{|D_R(S')|}{|D_R(S)|}$$

2. Artificial Pattern [10]: This measure quantifies the Percentage of the discovered patterns that are artificial facts.

$$\text{AP} = \frac{|S'| - |S \cap S'|}{|S'|}$$

3. Dissimilarity (DISS) [10]: This measure quantifies the amount by which the database is modified while hiding sensitive association rule.

$$\text{Diss}(S, S') = \frac{1}{\sum_{i=1}^{n} f_S(i)} \times \sum_{i=1}^{n} [f_S(i) - f_S(i)]$$

4. Misses Cost (MC) [10]: It is a measure of the number of useful rules that are preserved after modification of database

$$\text{MC} = \frac{|D'_R(S)| - |D'_R(S')|}{|D'_R(S)|}$$

D. User Interface

This model provides following features:

a) Provide user a separate space

b) File storage and opreations like upload, download, delete

c) Mining Association Rule by Apriori Alogrithm

d) Modify dataset by MDSRRC and ISLF for sensitive rule hiding

e) Evaluating modifications and Integrity check

E. Data Integrity

This cloud model provides integrity, which secures the data from adversaries. Data integrity has implemented using Java's inbuilt function of MD5. The MD5 algorithm hashes the entire file content. The user can store this hash value so that the next time he opens his file he can compare the current generated hash value with the previous value which is stored with him. If the hash values do not match it means that some outsider modified the file. Along with hashing the file data, the algorithm also displays the last accessed time, which lets user detect any unwanted access or attack on his data. The user can, therefore, check the integrity of his data this way.

IV. SIMULATION AND RESULTS

The hiding algorithms implemented ondifferent dataset, which covers all the aspects i.e. having different support and confidence variations so that rule derivation can be easily studied. The dataset consists of number of transaction, which are composed of different product ids that are bought. We have applied Apriori on the given dataset and generated mined rules.

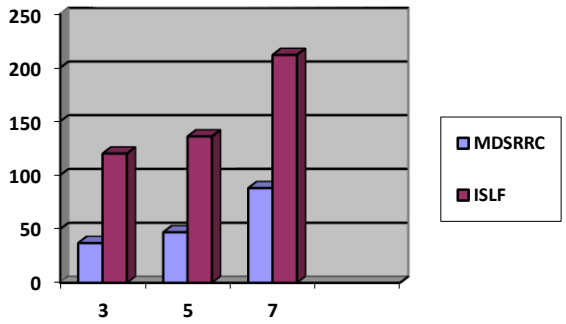


Fig. 1 Sensitive Rule Count Vs Dissimilarity

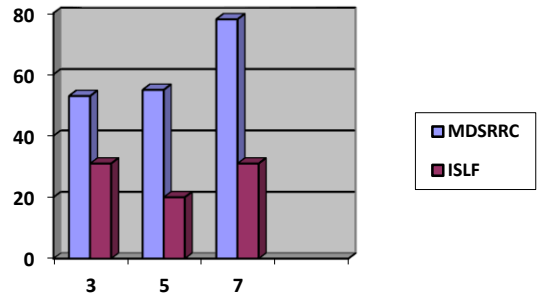


Fig. 2 Sensitive Rule Count Vs Misses Cost

Hiding techniques have applied on user specifies the sensitive rules by decreasing the support and confidence of these rules. Have varied the number of sensitive rules and compared the two algorithms based on various parameters. I have used synthetic dataset generated by TARtools[11].

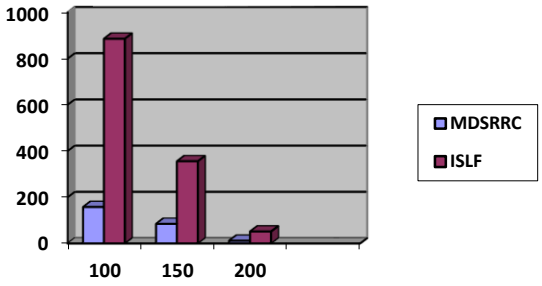


Fig. 3 Dissimilarity Vs Support

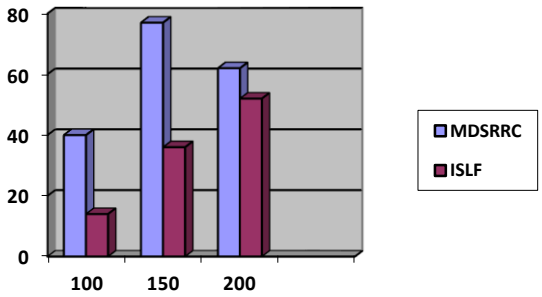


Fig. 4 Support Vs Misses Cost

V. CONCLUSION

We proposed a model which provides a cloud software service that is the whole package of modifying database, publishing it and efficiently mining association rules from it with data integrity.Simulation results proves that MDSRRC can be more efficiently used in hiding knowledge in database as compared to ISLF algorithm in terms of dissimilarity.MDSRRC algorithm make minimum modification in database to hide sensitive rules, time complexity and number of false rules generated are also less.

REFERENCES

[1] Techtarget. [Online]. Available: http://searchsecurity.techtarget.com/definition/Total-Information-Awareness.

[2] Maneesh Upmanyu, Anoop M. Namboodiri,Kannan Srinathan, and C.V. Jawahar,"Efficient Privacy Privacy Preserving K-Means Clustering," presented at International Institute of Information Technology Hyderabad, 2010.

[3] Linköping University, [Online]. Available: http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec7.pdf .

[4] Nidhi Porwal, Mahavir singh, Sunil Kumar, (2012, Nov.) "An Algorithm for Hiding Association Rules on Data Mining," International Journal of Computer Applications. [Online]. Available: http://www.ijcaonline.org/proceedings/ctngc/number3/9061-1022 File: ctngc1022.pdf

[5] Udai Pratap Rao, Nikunj H. Domadiya, "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database," in *Proceedings of IEEE International AdvanceComputing Conference IACC*, 2013.

[6] Prof. Hitesh Gupta, Dharmendra Thakur, (2014, Feb.) "Privacy Preservation by Hiding Highly Sensitive Rules with Fewer Side Effects," International Journal of Advanced Research in Computer Science and Software Engineering, pp. 546-551.

[7] Chen mu yehi, P. C. Yi-Hung Wu, "Hiding Sensitive Association Rules with Limited Side Effects, " in Proceeding of IEEE Transaction on Knowledge and data engineering, 2007. vol. 19, pp. 29-42.

[8] M. Gupta. & R. C. Joshi, (2009, Oct) "Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data," International Journal of Computer Theory and Engineering, vol. 1, no. 4, pp. 382-388.

[9] "Association rule mining," [Online]. Available: http://tkramar.blogspot.in/2008/09/introduction-to-association-rules-minig.html.

[10] Atefe, Naderi Dehkordi, Faramarz Safi Esfahani, Ramezani, (2014, June) "Hiding Sensitive Association Rules by Elimination Selective Item among RHS Items for each Selective Transaction," Indian Journal of Science and Technology, vol. 7, no. (6), pp. 831-837.

[11] Omari, Asem, Regina Langer, and Stefen Conrad "Tartool: A temporal dataset generator for market basket analysis." in Advanced Data Mining and Applications, Springer Berlin Heidelberg, 2008, pp. 400-410.