World Applied Sciences Journal 31 (12): 2087-2092, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.31.12.745

Improved Genetic Algorithm Approach for Sensitive Association Rules Hiding

[1]Azam Khan, [2]Muhammad Shuaib Qureshi, [1]Ayyaz Hussain

[1]Department of Computer Science,
International Islamic University, Islamabad, Pakistan
[2]Ministry of Interior, Islamabad, Pakistan

Abstract: Association rule mining is interesting area of data mining research which discovers correlations between different item sets in a transaction database. Efforts have been made for efficient hiding of sensitive association rules, but these techniques do not consider the consequences such as loss of information, lost rules and increase in ghost rules production. In this paper, we propose improved genetic algorithm architecture with a new fitness function for hiding sensitive rules by reducing loss of information, lost rules and generation of ghost rules. Different datasets have been used for experimental analysis. The results show the superiority of our work over the existing techniques.

Key words: Association rules · Genetic algorithm · Rules hiding · Ghost rules · Privacy preservation

INTRODUCTION

The procedure of retrieving the secrete information from large data is called as data mining. Organizations such as customer relationship management (CRM), telecommunication industry, financial sector investment trends, web technologies, demand & supply analysis, direct marketing, health industry, e-commerce, stocks & real estates, understanding consumer research marketing, e-commerce and product analysis generate huge amount of data.This massive quantity of data holds useful unseen knowledge or confidential information. Using data mining approaches, we can discover the useful information. Agarwal *et al*. [1] introduced association rules which is a common technique of data mining for the purpose of revealing useful hidden information from dataset. This technique is popular in discovering behavior from large datasets. Market basket data analysis is a popular example of this kind. Association rule mining is a dual step process. In step 1, Algorithm is used to identify frequent k-itemsets In step 2, association rules are derived from the frequent k-itemsets [1]. Furthermore, a rule is called sensitive if its discloser threat is greater than a user specified threshold. In addition, sensitive rules contain confidential data that we do not want to release to community. Privacy preserving data mining PPDM techniques are used to preserve such confidential information or restrictive patterns from illegal entrance [2-7].

The genetic algorithm (GA) is a probabilistic searching mechanism using the Darwinian principle which transforms an initial population into new population called as offsprings using crossover and mutation.

GA is initiated with a set of solutions called population which is represented by a chromosome. New population is generated by using the solutions of the old population and supposed that the new population will be superior to the old population. The new offsprings for reproduction are selected the same process is repeated until some condition is contented [8].

Related Work: Wang *et al*., [10] introduced pattern inverse tree (PI tree). This technique is use to hide informative association rules, the least association rule set that carry out the similar guess as the entire association rule rest by confidence precedence. The side effects in term of hiding failure, lost rules and ghost rules are high.W. K. Wong, David W. Cheung [11] discussed the security and integrity of Association Rule Mining.Cryptography approach in enforcing data security is adopted. Two common approaches K-anonymity & Data perturbation are used to ensure security and

Corresponding Author: Muhammad Shuaib Qureshi, Ministry of Interior, Islamabad, Pakistan.

integrity of association rules mining [11]. Clifton *et al*. [12] discussed the security and privacy implication of data mining in a broad scale in order to achieve privacy preserving in data mining. They presented the idea of limiting access to the database, eliminate unnecessary grouping, augmenting data, audit and fuzzy data. In this research they did not propose any specific algorithm. Yuhong *et al*. [9] proposed a reconstruction base technique in the domain of privacy preserving association rules. In this research, a new method, called FP- tree, was introduced for inverse frequent set mining. This technique fails to hide each sensitive association rules and also fall short to control the ghost rules and lost rules side effects. Chieh *et al*. [13] discussed greedy approach for privacy preservation of association rules [2]. In this approach two different methodologies are used for hiding sensitive rules and transactional retrieval engine based on FCET index tree are combined and the frame work is proposed. The author claimed that there are drawbacks in the proposed system. These drawbacks are removed through the rule hiding procedure all the sensitive rules are hidden and generated no false rules,and this procedure is performed without any limit in term of scalability of database size and it generate no extra ordinary rules and thus causes no hiding failure. E.Poovammal & M.Ponnavaikko [14] discussed task independent technique for privacy preserving association rules. A task based privacy preserving algorithm is developed, which secure the information, confidentiality and effectiveness of the data.This technique comparing with any privacy preservation technique, no information loss and several numbers of sensitive elements can be hold.

A new algorithm for carry out helpful PPDM actions while preserving data of the underlying data base is developed by Igor *et al*. [15]. This mathod is efficient against the information fraudulence due to the PPDM sanitization. This technique yields helpful information without neglecting the confidentiality of data holders. Chih-Chia *et al*. [16] proposed a novel algorithm called FHSAR, for fast hiding sensitive association rules. The schem can hide at all known SAR by scanning the database at once time.This will minimize the processing time.The goal of the technique is to convert the original database into release database D',in which none of the SAR is derived and the side effects are minimized.Remesh *et al*. [17] discusses the problem of sensitive association rule hiding. However, this approach did not mention about the modified database D', from which the sensitive association rules may not derived. These side effects will limit the scope of this approach. Vessilios *et al*. [18] discussed the issues regarding privacy preserving association rules. In this research, the author introduce five algorithm techniques namely algorithm *1.a*, *1.b*, *2.a*, *2.b*, *2.c*. Duraiswamy *et al*. [19] proposed an method called SRH (Sensitive rule Hiding) in the domain PPDM. According to this approach a rule is called sensitive rule that contain sensitive item in the RHS (Right Hand Side) of the rule. This approach adds together sensitive rules in to a cluster. The hiding failure of this approach is high because it is unable to hide sensitive rules that contain sensitive item in both sides.This method provides more side effects in the form of ghost rules and lost rules. Wang *et al*. [20] introduced two methods,increase support of the LHS (ISL) and decrease support of RHS (DSR). Here blocking technique (replace a value with unknown?) is used to hide sensitive predictive association rules. Similarly, a sensitive predictive association rule is defined as a rule in which the predictive set consist sensitive items on the left hand side of the rule.Generally, the proposed technique based on support and confidence. The performance of the algorithms evaluate with Saygin *et al*. [21] method. The proposed algorithms need little number of databases scanning. Moreover the approach proves more side effects in term of lost rules. Similarly,Gupta *et al*. [22] discusses the problem of fuzzy association rule hiding derived from computed data.A lot of research is performed to hide boolean association rules. This technique based on support and confidence framework. The performance of this approach is better in term of hiding failure and transaction modification. Stanley *et al*. [23] proposed a frame work for privacy in mining frequent itemset. In this research, taxonomy of algorithm: Naïve algorithm, Minimum Frequent Item Algorithm (MinFIA), Maximum Frequent Item Algorithm (MaxFIA) and Item Grouping Algorithm (IGA) were introduced. Zhang *et al*. [24] proposed a TAR algorithm (transaction adding and removing) for hiding sensitive association rules. This algorithm performs two procedures, adding weak association transaction (WAT) and removing strong association transaction (SAT). The main limitation of this approach is single rule hiding and high side effect in term of lost rules and ghost rules.The method proposed by Modi *et al*. [25] addressed privacy preservation in association. In this approach a new heuristic, called decrease support of right hand side item or rule clusters (DSRRC) were introduced in the domain of PPDM. This approach uses distortion; replacing 1s by 0s and

vice versa, as a modification technique. The proposed technique generates lost and ghost rule side effects. A novel algorithm named ADSSI (Advanced Decrease Support of Sensitive items) discussed the problem PPDM [26]. Introduced to preserve privacy for sensitive association rule in database.This approach is the advance version of DSSI (Decrease Support of Sensitive Items) proposed by Chang *et al*. [27]. The DSSI algorithm can completely hide SAR with the side effects of few non sensitive association rules wrongly hidden. The goal of the ADSSI algorithm is to change the original dataset D into sanitized dataset D', in which none of the SAR is derived and the side effects in term of lost rules and ghost rules are minimized. Besides the support and confidence of association rules, Malik *et al*. [28] have proposed other measure in the domain of PPDM. In this approach they define five measures namely correlation, Coefficient, Laplace, kappa and J-Measure. They presented that these measures are better in result as compare to conventional support and confidence frame work. Naeem *et al*. [29] have proposed a novel architecture in the domain of PPDM. The technique is only applicable on dataset whose attributes not more then 26. The authors claim that this technique does not create ghost rules side effects. The side effects in term of lost rules are still available. M. Naderi Dehkordi *et al*. [9] proposed a novel method for hiding SAR using genetic algorithm. This technique based on the traditional support and confidence framework. In this approach four fitness strategies are used namely Confidence based fitness strategy, Support based fitness strategy, Hybrid fitness strategy and Min-Max fitness strategy for the specification of fitness function. All these fitness strategies based on weighted sum function. In this paper authors claim that we minimize the number of lost rules and ghost rules with minimum modification in the original dataset, but number of lost rules and ghost rules are not mentioned.

Problem Description: Lot of research has been carried out on hiding sensitive rules using variety of techniques as given in literature. Most of the techniques focused on hiding sensitive rules but doesn't consider minimization of loss of information, lost rule and ghost rules. PPDM using genetic algorithm [9] gives good results for sensitive rules hiding, but does not consider minimization of penalties in response of the applied technique. Therefore an improved fitness function is required to be developed which can guide genetic algorithm to hide sensitive rules while minimizing the above mentioned three penalties.

Proposed Scenario: Privacy preserving data mining techniques are based on sensitive rules hiding. Most of the techniques are suffering from penalties such as hiding failure (rule hiding distance), lost rules, ghost rules and loss of information.These consequences played an important role in the motivation of the development of proposed architecture. In the proposed architecture we are trying to minimize the aforementioned issues. We are using genetic algorithm approach to preserve privacy in association rules. The proposed frame work is shown in Fig 1.

Fitness Function: Fitness is a measure of suitability or success of chromosomes. It measures the suitability of survival and reproduction of a genome. The fitness of an organism is measured by success of the organism in its life [9].The fitness function is defined over the genetic demonstration and determine the excellence of the represented solution. The fitness function is always probem depende.Parameters of proposed fitness function are Max of Hiding sensitive rules; Min lost rules, Min loss of information and Min ghost rules. The fitness functions of the existing and proposed frame works are calculated.

Hiding Failure (HF): Hiding Failure (HF) quantify the fraction of the sensitive patterns that remain exposed in the released dataset. It can be defined as the portion of the limited association rules to become visible in the released database divided by the all of those appearing in the original dataset. Formally, it can be computed by the equation 1.

$$HF = \frac{|A|}{|B|} \quad (1)$$

where |A| denotes total number of sensitive association rules exposed in the released (sanitized) dataset *D*. |*B*| corresponds to sensitive association rules presented in the novel dataset *D*.

Loss of Information (LI): Some rules are to be modified during rules hiding process due to this modification some information is to loss. Therefore it is called loss of information. So we can write as;

Loss of Information = Number of Data Items Modified

D. Lost Rules (LR)

This measure quantifies the percentage of the non-restrictive association rules which are hidden as a side-effect of the sanitization process. It can be calculated by the equation 2.

Table 1: Datasets for experimentation

Datasets	Total No of Instances	Total No of Attributes	Missing Instances
Synthetic	10000	8	None
Zoo	101	18	None

Table 2: Experimentational results

|D|	SAR	MST (%)	MCT (%)	LRs	GRs
10000	4→7	0.35	0.75	1	0
	6→4	0.35	0.65	0	1
	1→6	0.35	0.75	1	0
	7→4,6	0.30	0.65	2	1
101	9→13	0.65	0.85	1	2
	9→8	0.58	0.65	0	1
	13,9→10	0.58	0.58	2	0
	9→10,13	0.58	0.65	1	1

Table 3: Performance measurements

	Dataset			
	Synthetic		Zoo	
Factors	Dehkordi *et al.*	Proposed	Dehkordi *et al.*	Proposed
Rule Hiding Distance	4	2	4	1
Lost Rules	8	1	3	0
Ghost Rules	5	0	6	0
Loss of Information	6	1	7	1

$$Lost\ Rules = abs\ (|NSAR' - NSAR|) \quad (2)$$

where |NSAR| is the number of all non-sensitive association rules in the novel dataset D and |NSAR'| is the number of those non-sensitive association rules revealed in released (sanitized) dataset *D'*. It is noticeable that there presence a compromise of the ignore cost and the hiding failure, as the abundant sensitive association rules are necessary to keep, the rules secrete, the most valid association rules are likely to be ignore.

Ghost Rules (GR): This measure which is also known as artifactual rules is a quantifier of the fraction of the discovered association rules that are non genuine and artifact. This measure is computed as shown in equation 3.

$$GR = \frac{|A| - |B \cap A|}{|A|} \quad (3)$$

where |*B*| is the number of association rules exposed in the novel database *D* and |*A*| is the number of association rules expoaed in released (sanitized) dataset *D'*.

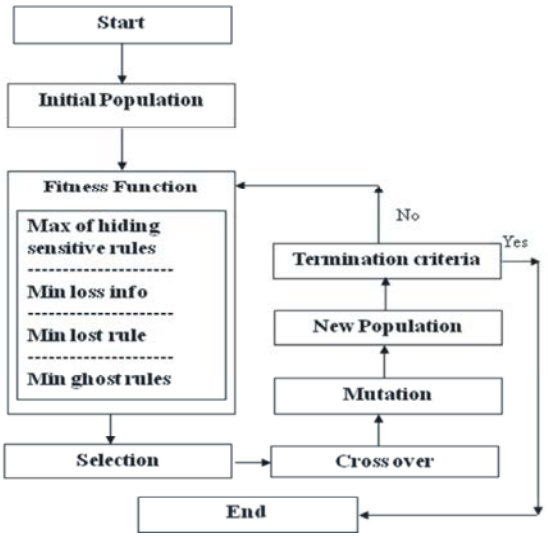

Fig. 1: Proposed framework for sensitive rules hiding

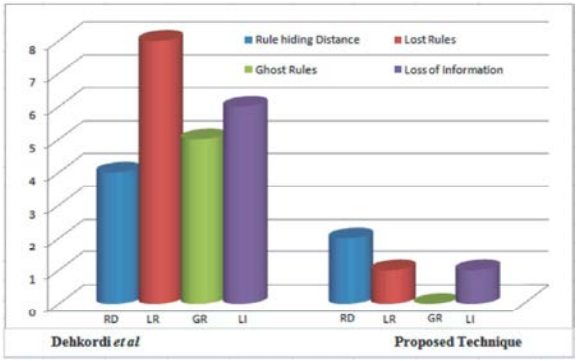

Fig. 2: Synthetic dataset

Proposed Fitness Function:

Fitness Function =
W_1* Rule Hiding Distance* + W_2**Number of Lost Rules* + W_3** Number of Ghost Rules* + W_4** Number of Modifications*

where $W_1+W_2+W_3+W_4 = 1$

For the validation of our proposed model, we have implemented it by using two datasets i.e. Synthetic dataset and Zoo dataset. [http://mlearn.ics.uci.edu/databases]

Table 2.describes the association rules generated from frequent k-itemsets before and after sanitization. During this hiding process no ghost rules are generated Initially, Sensitive Association Rule SAR, Minimum Supporting Threshold MST, Minimum Confidence Threshold MCT and original dataset pass to proposed framework.

Table 4: Fitness values of the proposed technique

Number of Generations	Synthetic Dataset Fitness Value	Zoo Dataset Fitness Value	Number of Generations	Synthetic Dataset Fitness Value	Zoo Dataset Fitness Value
10	3.5	4.1	110	1.4	2.2
20	3.3	3.9	120	1.3	2.0
30	3.1	3.6	130	1.2	1.8
40	2.9	3.5	140	1.1	1.7
50	2.6	3.3	150	1	1.6
60	2.5	3.1	160	1	1.5
70	2.2	2.8	170	1	1.5
80	2.0	2.7	180	1	1.5
90	1.7	2.5	190	1	1.5
100	1.6	2.2	200	1	1.5

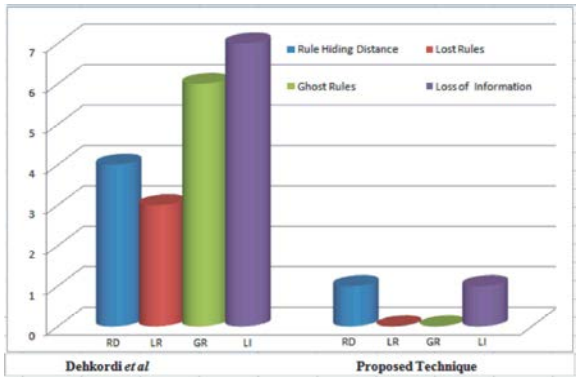

Fig. 3: Zoo dataset

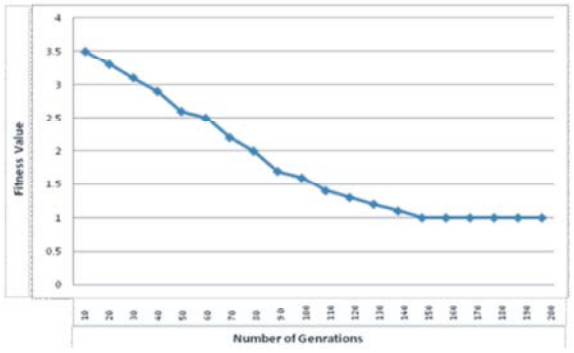

Fig. 4: Fitness graph of synthetic dataset

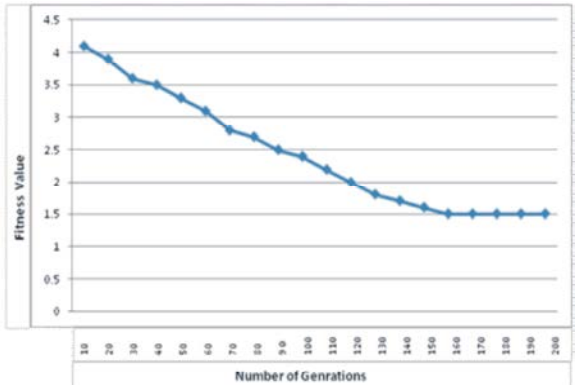

Fig. 5: Fitness graph of zoo dataset

CONCLUSION

Association rules mining and hiding without causing any consequences like loss of information, rules and generation of ghost rules, is a real challenge for the research community in the field of data mining. In this work, we have presented novel approach for hiding sensitive association rules with very minimum loss of information, lost rules and reduced ghost rules, by improving traditional genetic algorithm. The proposed fitness function calculates fitness value of each transaction that modifies some transactions in original dataset. Experimental results reveal remarkable reduction in the aforementioned consequences which outperforms the existing counterpart. In future work we would like to extend the approach by applying advanced privacy preserving techniques and revolutionary approaches to the sensitive rules.

REFERENCES

1. Agarwal, R., T. Imielinski and A. Swami, 1993. Mining ssociations Between Sets of Items in Large Databases, SIGMOD93, pp: 207-216, Washington, D.C, USA.
2. Agrawal, R. and R. Srikant, 2000. Privacy preserving data mining", In ACM SIGMOD Conference on Management of Data, pp.439-450, Dallas, Texas.
3. Ljiljana Brankovic and Vladimir Estivill-Castro, 1999. Privacy Issues in Knowledge Discovery and Data Mining, A ustralian Institute of Computer Ethics Conference Lilydale.
4. Clifton, C. and D. Marks, 1996. Security and Privacy Implications of Data Mining, in SIGMOD Workshop on Research Issues on Data Mining and knowledge Discovery.
5. Lindell, Y. and B. Pinkas, 2000. Privacy preserving data mining", In CRYPTO, pp: 36-54.

6. Leary, D.E.O., 1991. Knowledge Discovery as a Threat to Database Security, In G. Piatetsky-Shapiro and W. J. Frawley,editors, Knowledge Discovery in Databases, pp: 507-516, AAIPress/ MIT Press, Menlo Park, CA.
7. Verykios, V., E. Bertino, I.G. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis, 2004. State-of-the-art in Privacy Preserving Data Mining, SIGMOD Record, 33(1): 50-57.
8. Holland, 1992. Genetic Algorithm," Scientific American.
9. Young, G., 2007. Reconstruction Based Association Rule Hiding. In Proc. SIGMOD2007 ph.d Workshop on Innovative Database Research 2007(IDAR 2007) Beijing China, june 10, 2007.
10. Wang, S.L. and A. Jafari, 2005. Using Unknowns for Hiding Sensitive Predictive Association Rrules, In Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration (IRI 2005), pp: 223-228.
11. Wong, W.K. and David W. Cheung, 2008. Security and Integrity of Association Rule Mining.
12. Clifton, C. and D. Marks, 1996. Security and Privacy Implications of Data Mining, in Proc. ACM Workshop Research Issues in Data Mining and Knowledge Discovery.
13. Chieh-Ming Wu, Yin-Fu Huang and Jian-Ying Chen, 2009. Privacy Preserving Association Rules by Using Greedy Approach.
14. Poovammal, E. and M. Ponnavaikko, 2009. Task Independent Privacy Preserving Data Mining on Medical Dataset.
15. Igor Nai Fovino and Alberto Trombetta, 2008. Information Driven Association Rule Hiding Algorithms.
16. Chih-Chia, W., C. Shan-Tai and L. Hung-Che, 2008. A Novel Algorithm for Completely Hiding Sensitive Association Rules, in Proc. Eighth International Conference on Intelligent Systems Design and Applications Taiwan.
17. Ramesh, C.B., V. Jitendra, A.K. Sohel and S. Anand, 2008. Hiding Sensitive Association Rules Efficiently By Introducing New Variables Hiding Counter, IEEE International Conference on Service Operations and Logistics and Informatics, Beijing.
18. Verykios, V., A. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, 2004. Association Rules Hiding, IEEE Transactions on Knowledge and Data Engineering, 16(4): 434-447.
19. Duraiswamy, K., D. Manjula and N. Maheswari, 2008. A New approach to Sensitive Rule Hiding, Journal of Computer and Information Science.
20. Wang, S.L. and A. Jafari, 2005. Hiding Sensitive Predictive Association Rules, Systems, Man and Cybernetics, 2005 IEEE International Conference on, 1: 164-169 Vol. 1, 10-12 Oct. 2005
21. Saygin, Y., V. Verykios and C. Clifton, 2001. Using Unknowns to Prevent Discovery of Association Rules", SIGMOD Record 30(4): 45-54.
22. Gupta, M., *et al*., 2009. Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data, International Journal of Computer Theory & Engineering, pp: 4.
23. Stanley R.M. Oliveira and Osmar R. Zaïane, 2002. Privacy Preserving Frequent Itemset Mining, Proceedings of the IEEE international conference on Privacy, security and data mining - Volume 14, Austrlia, 2002.
24. Xiaoming Zhang and Xi Qiao, 2008. New Approach for Sensitive Association Rule Hiding, ettandgrs, vol. 2: 710-714, International Workshop on Education Technology and Training & International Workshop on Geoscience and Remote Sensing, 2008
25. Modi, C.N., U.P. Rao and D.R. Patel, 2010. Maintaining privacy and data quality in privacy preserving association rule mining, Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on, pp: 1-6, 29-31 July 2010.
26. Shan-Tai Chen, Shih-Min Lin, Chi-Yii Tang and Guei-Yu Lin, 2009. An Improved Algorithm for Completely Hiding Sensitive Association Rule Sets, Computer Science and its Applications, CSA '09. 2nd International Conference on, pp: 1-6, 10-12 Dec. 2009.
27. Chang, Y.C. and S.T. Chen, 2008. Fast Algorithm for Completely Hiding Sensitive Association Rule Sets, Proceedings of the Cryptology and Information Security Conference (CISC2008), Hualien, Taiwan, R.O.C., pp: 547-560.
28. Malik, H.H. and J.R. Kender, 2006. Clustering Web Images using association Rules, Interestingness Measures and Hypergraph Partitions, ICWE 06', july 11-14, 2006.
29. Naeem, M. and S. Asghar, 2010. A Novel Architecture for Hiding Sensitive Association Rules, Proceedings of the International Conference on Data Mining, Las Vegas, Nevada, USA, pp: 12-15 July 2010