Journal of Computer Science 9 (9): 1086-1091, 2013
ISSN: 1549-3636

doi:10.3844/jcssp.2013.1086.1091 Published Online 9 (9) 2013 (http://www.thescipub.com/jcs.toc)

PRESERVATION OF THE PRIVACY FOR MULTIPLE CUSTODIAN SYSTEMS WITH RULE SHARING

[1]B. Murugeshwari, [2]C. Jaya Kumar and [3]K. Sarukesi

[1]Department of Computer Science and Engineering, SA Engineering College, India
[2]Department of Computer Science and Engineering, RMK Engineering College, India
[3]Department of Computer Science and Engineering, Hindustan University, India

Received 2012-07-07; Revised 2013-07-02; Accepted 2013-07-10

ABSTRACT

The aim of the research work is to mine the data set available with each custodian in a semi honest model, securely without disclosure of any data amongst various custodians involved. No custodian discloses any information. In the proposed scheme in order to reduce the computational complexity, the data partitioning has been done in the horizontal way. The proposed research work consists of a well skilled and planned architecture implementation for achieving the proposed privacy preservation in the data mining filed and used a new hybrid data mining model which is developed for combining commutative RSA and a C5.0 algorithm to generate classification rules. This study utilized real world data collected from an UCI repository and experiments are conducted based on the parameters like time complexity, accuracy and error rate. The proposed model preserve expected level of privacy without any information loss, take less time for computation, lower error rate and improves accuracy.

Keywords: Privacy Preserving Data Mining, University of California Irvine Data Repository (UCI), Semi Honest Model, Secure Multiparty Computation (SMC)

1. INTRODUCTION

In order to achieve the secured data mining a number of algorithms have been implemented which played the significant role in the contemporary, but the increase in the data complexity and the higher security requirements made effect insufficient. This scenario ignited the scientific society to optimize the data mining technique especially considering the privacy preservation. At first, the Privacy preservation Data Mining (PPDM) was introduced in 2000, in which many dominant issues related to this area were discussed. Researchers and the scientific society from then on have proposed numerous issues related to the PPDM.

One must know inputs from all the participants to conduct the constraint based security computations. However, if nobody can be trusted enough to know all the inputs, privacy will become a primary concern. One of the solutions to this is Secure Multiparty Computation (SMC). This SMC is based on cryptographic functionality which plays a major role in the context of privacy preserving data between different participants in sharing authorized data. SMC is a computational system in which the value based on individually held secret bits of information that compute multiple parties wish to join. Das *et al*. (2009) proposed a scalable, local privacy-preserving algorithm for distributed Peer-to-Peer (P2P) data aggregation useful for many advanced data mining/analysis tasks such as average/sum computation, decision tree induction, feature selection and more. Unlike most multi-party privacy-preserving data mining algorithms, this approach works in an asynchronous manner through local interactions and therefore, is highly scalable. Karthik *et al*. (2011) proposed rule extraction technique for liver disease. Ukil and Sen (2010) developed a scheme for secure multiparty data aggregation with the help of modular arithmetic concept. Shaneck *et al*. (2006) addressed the issue of secure multi

Corresponding Author: B. Murugeshwari, Department of Computer Science and Engineering, SA Engineering College, India

party computation which formed the kernel of many data mining applications. Harnsamut *et al*. (2008) focused on maintaining the data quality in the scenarios in which the transformed data was used to build associative classification models. Lin and Chen (2008) decided which instances of training dataset were support vectors, i.e., the necessarily informative instances to form the classifier. The support vectors are intact tuples taken from the training dataset. Fung *et al*. (2005) proposed a k-anonymization solution for classification. The goal was to find a k-anonymization, not necessarily optimal in the sense of minimizing data distortion, which preserves the classification structure. Lin and Chen (2008) proposed an approach to the SVM classifier process to transform it to a privacy-preserving classifier which does not disclose the private content of support vectors. Chen *et al*. (2011), a new discriminate diagnosis model constructed by attribute selection, decision tree C5.0 algorithm and discrimination analysis was proposed, which consists of two phases. The critical attributes were filtered out from the original attributes. Wang *et al*. (2009) focused on comparing the classification performance and accuracy for short-term urban traffic flow condition using decision tree algorithms (CHAID, CART, QUEST and C5.0).

The classification rules are derived from the decision tree in the form of if-then- else. These rules are used to classify the records with unknown value for class label. Sethi *et al*. (2012) proposed second order decision table using rule generation. Hou and Su (2007) proposed to develop a fuzzy rule-based reasoning system to set a nano- particle milling process. The characteristics of the proposed system were to use data-driven to do fuzzification and rule extraction instead of directly using domain experts. Qiong and Xiao-Hui (2009) proposed the privacy preservation and the mining efficiency, an effective privacy preserving distributed mining algorithm of association rules. Combining the advantages of both RSA public key cryptosystem and homomorphism encryption scheme, a model of hierarchical management on the cryptogram was put forward in the algorithm. Karthikeswaran *et al*. (2012) presented a novel based approach that strategically modified a few transactions in the database. It modifies support or confidence values for hiding sensitive rules without producing many side effects Jiri *et al*. (2009) presented one of many possibilities of decision theory that could be used in the modeling of the quality of life in a given city. Real data sets were analyzed, pre-processed and used in the classification models. Naeem *et al*. (2010) proposed a novel architecture which acquired other standard statistical measures instead of conventional framework of Support and Confidence to generate association rules. The new architecture generated no ghost rules with complete avoidance of failure in hiding sensitive association rules. Pourebrahimi *et al*. (2010) presented an opportunity to increase significantly the rate at which the volumes of data generated through the maintenance process could be turned into useful information. This could be done using classification algorithms to discover patterns and correlations within a large volume of data. Parmar *et al*. (2011) proposed a blocking based approach for sensitive classification rule hiding.

2. MATERIALS AND METHODS

Let us consider a training database D_m to be considered for mining. Let a set $c = \{c_1, c_2, c_3.... ....c_p\}$ represent the set of data custodians amongst which the database D_m is portioned horizontally. Also p represents the total number of data custodians. The research work presented here considers the hybrid model which combines both C5.0 algorithm for classification and commutative RSA Algorithm for encryption and decryption.

It is assumed that each data custodian embodies two datasets one training dataset d_{cx} which is pre classified and another dataset dt_{cx} which represents a test dataset or an unclassified dataset. The goal of the research work proposed here is to mine the test dataset dt_{cx} available with each data custodian in a semi-honest model, securely without disclosure of any data dt_{cx} or d_{cx} amongst the varied custodians involved.

The aim is to mine the dataset available with each data custodian in a semi-honest model, without the disclosure of any data amongst the varied custodian involved. The proposed scheme does not consider an central server for group establishment. No database exchange between the parties instead only exchange the rules.

From the **Fig. 1**, the overall framework has been accomplished in the following phases: Local classification rules ($R_{r_1}, R_{r_2}, R_{r_3}....R_{r_p}$) are generated in step 1. In step 2, encryption keys are generated using commutative RSA. Every custodian sends solely encrypted rules to the opposite custodians who in flip encrypt the encrypted rules received by them using their encryption keys in step 3 ($C_E R_{r_1}, C_E R_{r_2}, ...C_E R_{r_p}$). The combined secure rule set is a collection of all the rules available with each data custodian and discovered in step 4. The use of commutative RSA decryption method at every knowledge custodian to get the combined rule set in step 5.

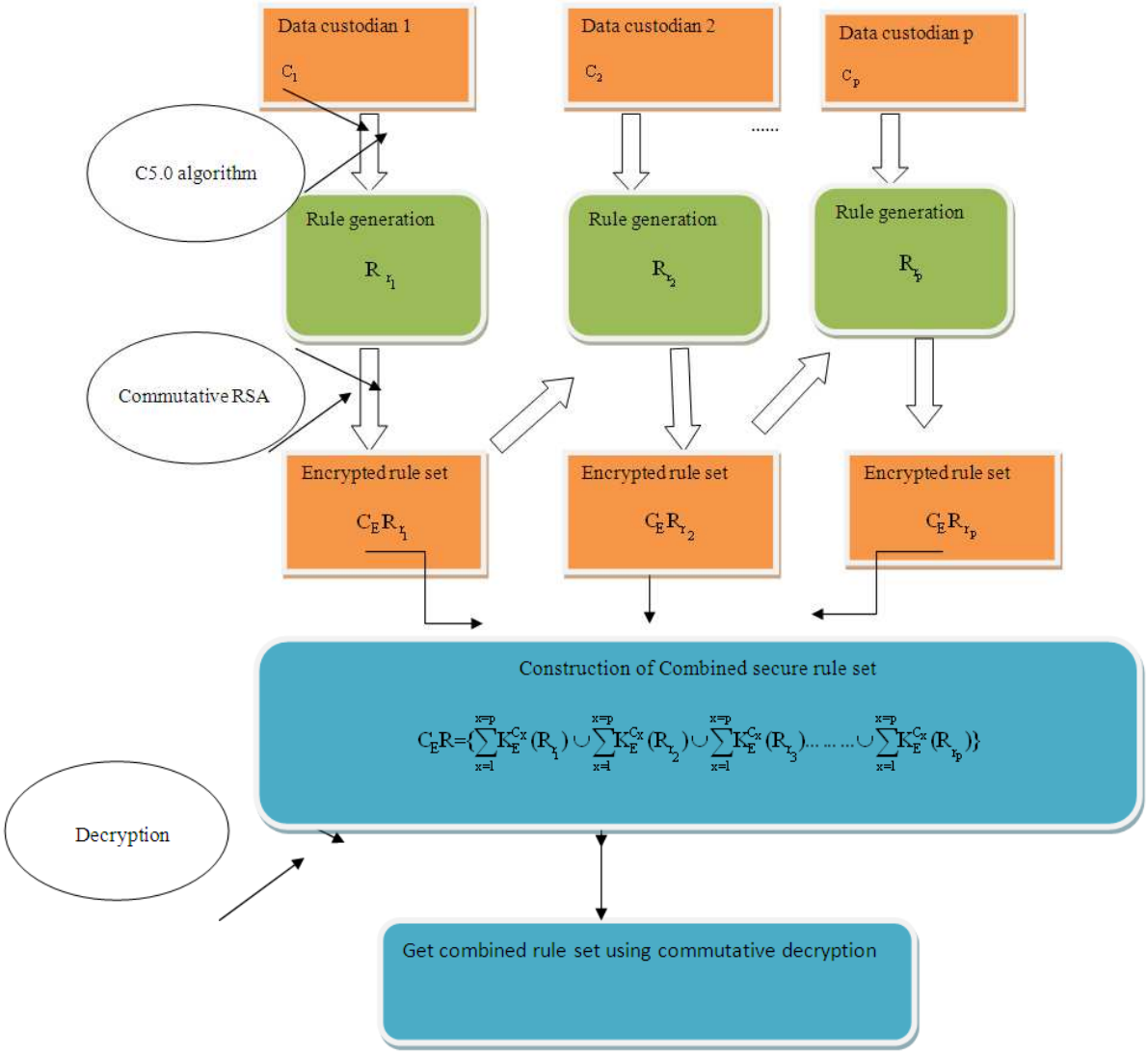

Fig. 1. Framework for the proposed privacy preserving data mining system

3. RESULTS AND DISCUSSION

In order to perform the implementation successfully the researcher has utilized many standard data sets from UCI for analyzing execution time, accuracy and error rate. The results have been obtained for individual data sets for difference performance parameters. In this research work the researcher has implemented his developed technique for assuring the most robust privacy preservation feature in data mining and on the other hand in order to make the system more efficient and accurate, the researcher has implemented C5.0 classification algorithm. The results obtained have illustrated comparatively better results among other existing techniques like IDE3, CART and C4.5.

3.1. Performance Metrics of the Proposed System

3.1.1. Accuracy

The accuracy of the model can be evaluated based on the following measures:

True positive = TCT/tot
True negative = TIT/tot
False positive = TCF/tot
False negative = TIF/tot

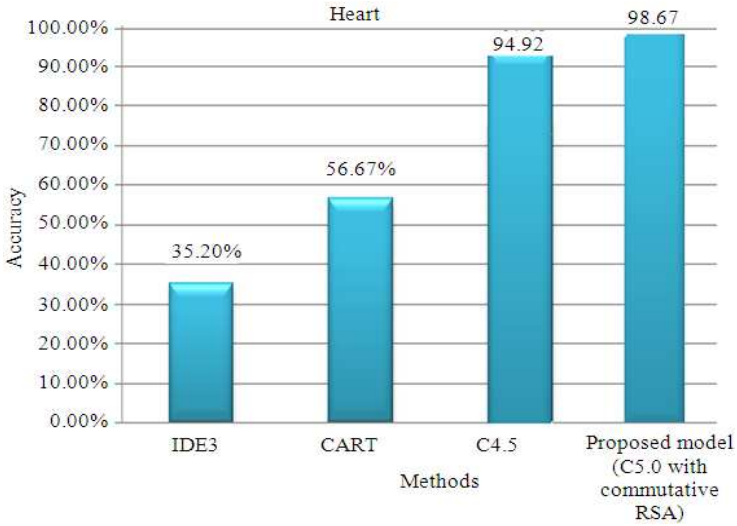

Fig. 2. Accuracy of heart data set

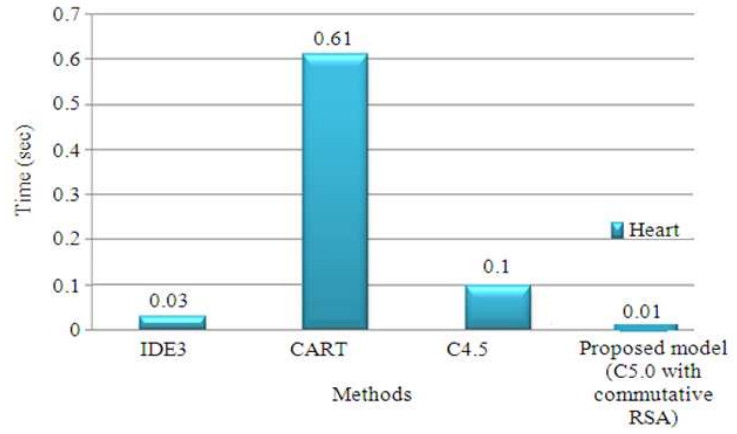

Fig. 3. Time complexity of heart dataset

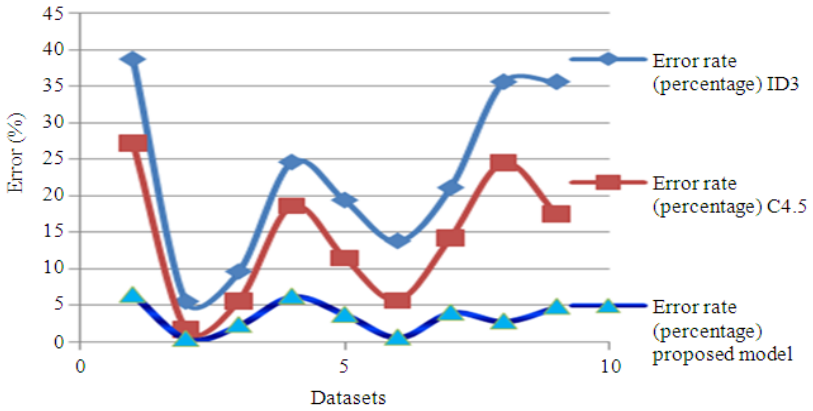

Fig. 4. Error rate of heart data set

TCT = Total number of items that are correctly classified as true
TCF = Total number of items that are correctly classified as false
TIT = Total number of items that are incorrectly classified as true
TIF = Total number of items that are incorrectly classified as false.
Tot = Total number of items

In the research implementation, the researcher has considered three dominant parameters for comparative study. These are:

- Accuracy
- Time complexity
- Error rate

The above mentioned **Fig. 2** represents the result obtained for the study of accuracy for heart data sets. It contains 22 attributes and 267 instances. From the result obtained it can be found that the proposed method is more accurate than other methods. In this result, the proposed method has exhibited the accuracy of 98.67 while the other three (IDE3, CART, C4.5) are having 35.20%, 56.67, 94.92 respectively.

3.2. Time Complexity

Time required to build and execute the model.

Figure 3 shows comparative results of proposed approach with existing approaches in terms of time complexity using heart dataset. For the above mentioned heart data set, CART is found to be more time consuming and the proposed technique is fastest among all other techniques.

3.3. Error Rate

Error rate is the difference between actual result or outcome and expected result or outcome. From **Fig. 4**, the proposed model has lower error rate than other techniques.

4. CONCLUSION

The presented research work for Privacy preservation in Data mining gives a new privacy preserving data mining system providing prominence to the privacy and security of the horizontally portioned data available with the multiple parties. The results obtained here, have demonstrated that the proposed system can accomplish the goal of most efficient, lower error rate and optimum accuracy in data mining applications. On the other hand the presented technique can play a vital role in reducing time complexity in mining operation. Even the results obtained has illustrated and established itself as best solution for reducing time complexity.

5. REFERENCES

Chen, X., L. Ma, N. Chu and Y. Hu, 2011. Diagnosis based on decision tree and discrimination analysis for chronic hepatitis b in TCM. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops, Nov. 12-15, IEEE Xplore Press, Atlanta, GA., pp: 817-822. DOI: 10.1109/BIBMW.2011.6112478

Das, K., H. Kargupta and K. Bhaduri, 2009. A local distributed peer-to-peer algorithm using multi-party optimization based privacy preservation for data mining primitive computation. Proceedings of the IEEE 9th International Conference Peer-to-Peer Computing, Sep. 9-11, IEEE Xplore Press, Seattle, WA., pp: 212-221. DOI: 10.1109/P2P.2009.5284514

Fung, B.C.M., K. Wang and P.S. Yu, 2005. Top-down specialization for information and privacy preservation. Proceedings of the 21st International Conference on Data Engineering, Apr. 5-8, IEEE Xplore Press, pp: 205-216. DOI: 10.1109/ICDE.2005.143

Harnsamut, N., J. Natwichai and B. Seisungsittisunti, 2008. Privacy preserving of associative classification and heuristic approach. Proceedings of the 9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Aug. 6-8, IEEE Xplore Press, Phuket, pp: 434-439. DOI: 10.1109/SNPD.2008.155

Hou, T.H. and C.H. Su, 2007. A fuzzy data-driven and rule-based reasoning system for setting the Nano-particle milling process parameters. Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, Dec. 2-4, IEEE Xplore Press, Singapore, pp: 582-586. DOI: 10.1109/IEEM.2007.4419256

Jiri, K., J. Pavel, K. Miloslava and M. Jan, 2009. Quality of life investigation case study in the czech republic. Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery, Aug. 14-16, IEEE Xplore Press, Tianjin, pp: 264-268. DOI: 10.1109/FSKD.2009.44

Karthik, S., A. Priyadarishini, J. Anuradha and B.K. Tripathy, 2011. Classification and rule extraction using rough set for diagnosis of liver disease and its types. J. Adv. Applied Sci. Res., 2: 334-345.

Karthikeswaran, D., V.M. Sudha, V.M. Suresh and A.J. Sultan, 2012. A pattern based framework for privacy preservation through association rule mining. Proceedings of the International Conference on Advances in Engineering, Science and Management, Mar. 30-31, IEEE Xplore Press, Nagapattinam, Tamil Nadu, pp: 816-821.

Lin, K.P. and M.S. Chen, 2008. Releasing the SVM classifier with privacy-preservation. Proceedings of the 8th IEEE International Conference on Data Mining, Dec. 15-19, IEEE Xplore Press, pp: 899-904. DOI: 10.1109/ICDM.2008.19

Naeem, M., S. Asghar and S. Fong, 2010. Hiding sensitive association rules using central tendency. Proceedings Of the 6th International Conference on Advanced Information Management and Service, Nov. 30-Dec. 2, IEEE Xplore Press, Seoul, pp: 478-484.

Parmar, A.A., U.P. Rao and D.R. Patel, 2011. Blocking based approach for classification rule hiding to preserve the privacy in database. Proceedings of the International Symposium on Computer Science and Society, Jul. 16-17, IEEE Xplore Press, Kota Kinabalu, pp: 323-326. DOI: 10.1109/ISCCS.2011.103

Pourebrahimi, A., S. Mokhtar, S. Sahami and M. Mahmoodi, 2010. Status detection and fault diagnosing of rotatory machinery by vibration analysis using data mining. Proceedings of the 2nd International Conference on Computer Technology and Development, Nov. 2-4, IEEE Xplore Press, Cairo, pp: 131-135. DOI: 10.1109/ICCTD.2010.5646134

Qiong, G. and C. Xiao-Hui, 2009. A privacy-preserving distributed method for mining association rules. Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence, Nov. 7-8, IEEE Xplore Press, Shanghai, pp: 294-297. DOI: 10.1109/AICI.2009.486

Sethi, K.K., D.K. Mishra and B. Mishra, 2012. KDSODTEX: A novel technique to extract second order decision table using KDRuleEx. Int. J. Scient. Eng. Res., 3: 645-649.

Shaneck, M., Y. Kim and V. Kumar, 2006. Privacy preserving nearest neighbor search. Proceedings of the 6th IEEE International Conference on Data Mining Workshops, Dec. 18-22, IEEE Xplore Press, Hong Kong, pp: 541-545. DOI: 10.1109/ICDMW.2006.133

Ukil, A. and J. Sen, 2010. Secure multiparty privacy preserving data aggregation by modular arithmetic. Proceedings of the 1st International Conference on Parallel Distributed and Grid Computing, Oct. 28-30, IEEE Xplore Press, Solan, pp: 344-349. DOI: 10.1109/PDGC.2010.5679976

Wang, J.J., J.F. Wang, F. Lu, Z.D. Cao and Y.L. Liao *et al*., 2009. Comparison study on classification performance for short-term urban traffic flow condition using decision tree algorithms. Proceedings of the WRI World Congress on Software Engineering, May 19-21, IEEE Xplore Press, Xiamen, pp: 434-438. DOI: 10.1109/WCSE.2009.255