iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8298

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

Perspective Challenges & Proposed Solution for Ppdm Using Data Mining Techniques

Charu Sharma[1], Kanwal Garg[2]
[1,2]*Department of Computer Science & Applications, Kurukshetra University*

Abstract: ***Data mining is usually used to retrieve unseen information, sensitive insensitive data, and knowledge from the high volume of databases. In general raw data consist of confidential and secret information which can only be accessed via authorized analyst and cannot leak. Therefore, to secure data from being accessed by some unauthorized users and organizations privacy preservation is requisite. The task of mining data by preserving its privacy is called Privacy Preserving Data Mining (PPDM). Privacy preserving is done to acquire precise paradigms without letting accession to secure information in the data records, hence finessing the difference between privacy and data mining. Recently many events have increased the attention of researchers for fact excavation toward security-related challenges thus ruling to fascinate the proficient challenges for PPDM. So, the purpose of this paper is to exhibit the modern arena of PPDM framework and methods. Various methods have been established and implemented for PPDM. This article provides a review of different PPDM schemes and approaches that can be used to hide secret information from an unauthorized user by using association rule mining. In the end guidelines for the future scope are laid out.***

Keyword: Association rule, Data Mining, Distributed database, Cryptography, Secret sharing.

I. INTRODUCTION

Information Space / Databank protection next to online network phishing became a necessity. Online web phishing called high safety productive interest on customer and companies globally such as mobile commerce, virtual banking, investing online impose upon an equal person and program susceptibility got distressed from terrible economic failure. Therefore, strengthen PPDM methods that are continuously demand protected and authenticated data transaction over the Internet [1].

The impressive growth of storing customer's private information led to an upgrade involvement of data mining algorithms and substantial influence on information sharing. Amongst various current algorithm, PPDM relies on excellent results. Indeed, privacy preservation can come into existence by three mining perspective consisting of the association rule, classification, and clustering. The difficulties faced in data mining have schemed in many states such as database statistics and cryptography.

Each of them is connected to aggregate efficiency to store user's data together in such a manner that it leads in growing intricacy of data mining algorithms that influence data interchanged over the network [1].

II. PRIVACY PRESERVING DATA MINING

Data mining is an approach of selecting rules and paradigms from data. It is acknowledged as knowledge discovery from data (KDD).The conventional data mining operates on a data warehouse for aggregating a variety of data on a centralized location and running a set of algorithms for the particular data repository [2]. This model works fine when the whole data owns a specific superintendent who primarily induced the data and use data mining model for creating results without broadcasting the results to any of the third party. However, if we consider all the real time applications of data mining, then privacy is the primary concern.

Privacy is the principal concern for various approaches. The first problem is the fact that certain attributes of data i.e. Service Security Number (SSN), or a combination of attributes that might result in leaking of personal identity information. The second problem is that the data gets horizontally split across multiple keepers and without allowing them to transfer data to another site. The data goes vertically partitioned in that case, different superintendents own different dimensions of evidence, and they have same allocation constraints [2]. Finally, we can say that data mining model has some limitations i.e. some rules can be limited, and some control may lead to a precise formulation which restricts to law. Cryptographic techniques are for performing modifications to ensure required privacy level. In most of the cases, the constraints for privacy preserving are accurate data, generated models and performance of mining and maintaining privacy.

The several approaches used with PPDM are

A. Data renovate before it gets transmitted to data mine.

B. Data gets distributed over two or more sites and is combined using semi-honest protocol for learning global results by without revealing any information/data at the individual site.

C. The model only provides classification results to the designated/authorized party, who cannot learn anything but can check the existence of rules without revealing rules [3].

III. KEY ISSUES

The various key issues in the field of PPDM are

A. Privacy Protection Data Publishing

These methods tend to study unique conversion methods collaborated among privacy. These different methods are randomization, k-anonymity, l-diversity. Therefore, some other related issue with this is that how it gets associated with various data mining methods such as association rule mining. Other connected issues include to keep data valuable and to study different solutions to privacy and comparing them regarding effectiveness.

B. Modifying the outcome of Data Mining applications to secure privacy

In most of the cases, the results of data mining applications such as association and classification rule mining leads to an adjustment with the privacy of the information. Thus we have generated a field of privacy in which outcomes are modified in such a manner to preserve the privacy of data. So, the methods used are association rule hiding methods, which involve suppression of association rule to sustain privacy.

C. Problem Auditing

In this methods are comparable to previous methods for modifying outcomes. Here, it either modify or limit the result of problems.

D. Crypto-graphical Methods for Scattered Privacy

The Data may get distributed across many sites, but the owner of data at different locations wants to compute same/similar function. Therefore, various cryptographic techniques are used to communicate in several places, so that secure method computation is probably without acknowledging precise information ([4] - [5]).

IV. METHODS

At present, there are different privacy preserving methods available for data mining. PPDM techniques safeguard the statistics by altering the data to mask and removes the unique, sensitive data. Typically, the methods based on the concepts of privacy loss, the extent to determine authentic user data from altered one, the failure of loss of data and an evaluation of data efficiency loss. The other approaches that apply cryptographic methods to prevent the loss of knowledge are estimated to be very expensive. Contrarily, PPDM's use data partitioning and it takes place in the form of horizontal and vertical partitioning [1

Table 1: Various Privacy Preserving Methods in Data Mining [1]

PPDM Techniques	Description
Data Distribution	Data is partitioned in vertical and horizontal manner.
Data Distortion	This technique involves data blocking, data perturbation and data swapping etc
Association rule mining	Association rule mining is the improvised version of data distortion technique that is used for mining frequent item sets.
Rule hiding	It hides the sensitive data so that the privacy remain preserved.
K-anonymity	It uses quasi identifier to identify the public and private attributes.
L-diversity	It keeps the group size to k and maintains the diversity of sensitive attributes.
Data Randomization	I It hides the individual data by performing masking of data or modification of data.
Privacy Protection	Protects the data by maintaining the quality of data.
Cryptographic Technique	This technique is used to encrypt the data such that the actual data is not revealed to third party.

V. LITERATURE REVIEW

"A Literature Review is a content of a scholar paper, which includes the present knowledge including the substantive conclusion, as well abstract and methodological improvements on a particular topic." Literature reviews are secondary sources and do not report original or experimental work.

Agrawal, Rakesh et al. (2000) [8] in this paper the primary task is to build up accurate models without admittance to private information. Kantarcıoglu, Murat et al. (2003) [9] in this data is split among multiple organizations and such organizations utilize data in such a manner that neither its training data/database nor instances can be classified. Friedman, Arik et al. (2008) [10] proposed a new process for achieving k-anonymity by suppression. The values are quenched only on fixed records depending on attribute values of other documents without the prerequisite of the standard hierarchy tree. Belwal, Ramesh Chandra, et al. (2008) [11], Yogendra Kumar et al. (2011) [17] & Dehkordi Mohammad Naderi et al. (2009) [13] recommended an algorithm that is used to hide information. It is done by increasing and decreasing the confidence and support of sensitive global rules by using association rule hiding. Harnsamut, Nattapon et al. (2008) [12] &Kamakshi, P., et al. (2010) [16] identified a heuristic algorithm by applying data transformation takes place by perturbing the original data, and then customized information is submitted as an applicant's query through cryptography. Gkoulalas D., Aris, et al. (2009) [14] they provide an exact solution to hide sensitive frequent item sets by applying border based approach. Zhang, Xiaolin, et al. (2010) [15] concluded a privacy perpetuating process by using perturbation approach. Jain, He, Yeye et al. (2011) [18] proposed a graph based anonymization algorithm suggests the solutions for "min-cut/max-flow" problem and deals with equivalence attacks. Dev, Himel, et al. (2012) [19] &Ibrahim, Ayad et al. (2012) [21] introduced an efficient cryptographic approach for mining cloud data. In this, a client commits a data to a single cloud, gives provider and outside muggers an illegal admittance to data present on the cloud. Karim R, et al. (2012) [20]stated an algorithm that is used to develop the entire set of maximum frequent item sets and remove null transactions by using 'ComMapReduce' framework. Keng, Joseph Chan Joo et al. (2013) [22] delivered a distributed architecture algorithms that rely on Row-level data and field partitioning. Giannotti, Fosca et al. (2013) [23] & Arunadevi, M., et al. (2014) [24] concentrated on the issue of outsourcing the successive thing set inside corporate protection saving system. Xu, Lei et al. (2014) [25] Sharma, Manish et al. (2014) [26] & Rathna, S. Selva, et al. (2015) [27] observed that Information mining has turned out to be progressively famous. It permits sharing of touchy security information for investigation purposes. Therefore, proposed a proficient approach for the safety safeguarding in data mining. Modak, Masooda, et al. (2016) [28] proposed a code based approach which permits the information to be partitioned into various shares and handled independently at multiple servers.

Privacy persevering discovers different functions in a police investigation that are imagined to be "secrecy infringement" applications. Most ways surveyed in our literature for privacy estimation use some style of transmutation on the information to perform data protection. Often, there are such methods that scale back the roughness of illustration to cut back the privacy. The decrease in roughness ends up with fall in efficiency of knowledge mining finding. It can be the fundamental trade-off between information loss and confidentiality. There are various other issues with privacy approaches, depicted in next section [4].

VI. RESEARCH ISSUES & CHALLENGES

Nowadays, a large amount of data contributes to several sectors such as business, military, health, government sectors, etc. Thus, conservation of secrecy against leaks and intrusion are of demanding concernment. Respective large industries or organizations and ministries worldwide are relying on knowledge or information through online network discloses substantial interest on top of concealment issues. Therefore, the expeditious improvement of informatics is facing new challenges regarding privacy of data.

Data processing acquires the proficiency to draw out the splash of useful, attractive patterns and recognition from a large amount of database that requires précised preservation. The basic interpretation of privacy preserving is to consolidate acknowledged data excavation methods in modifying the data to conceal sensitive data. The primary demand is to intensively modify data and retrieve its mining result from the changed one and provide filtered data to the existing users and clients. The security threats are cut back by precise content concealing by the refinement of a sensitive rule. The data that contains private information result in seclusion risk as the data gets corrupted. The heuristic methods wrap data cleaning. To hide raw data some secondary effects of missing and artificial costs takes the design ([1] – [4]). Following process is to be followed by filtering of rules.

A. No pattern is considered to be touchy according to the proprietor's point of view, and that can mine from the initial database at pre-processing edges of support and certainty that remain additionally uncovered from the sterilized database.

B. The entire non-touchy standards that show up when mining the master database at pre-defined boundaries of comfort and confidence can extract from the cleaned information at similar tips or higher.

C. Association lead concealing procedure relies on upon support or assurance of the control, there are two approaches to shroud govern, one by diminishing bolster up to secure edge or lessening certainly up to the particular edge, so the calculation in mining deals with support not ready to excavate delicate guidelines [7].

VII. PROPOSED SOLUTION

Association Rule Hiding can achieve the proposed solution for issues mentioned above. In recent years a terrific advancement has been done in expertise to perform an efficient association rule mining. Alike patterns usually conceal the primary objective as a commercial information. Earlier, some work has done by providing database security on various challenges of association rule mining. The two broader approaches used for control hiding are:

Distortion: In deformation technique, the particular entry for a given transaction is altered by a distinctive value. It deals with binary transactional data sets where the access value point is flip.

Blocking: In blocking technique, the entry point is not altered and left inadequate Thus, the unexplored values are used to prevent sensitive rules.

Both the blocking and deformation process has some after effects on the non-sensitive rules in data. Creation of new ghost rules because of the distortion or blocking process as some non-sensitive rules are missing with sensitive rules. Therefore, the after effects are inconvenient as they lessen the utility of the data. These techniques modify the data by manipulating the values by replacing 1values with 0-values. The support of the similar sensitive rule gets lower by applying this method. According to the above approach, the utility of the data can be defined as the support of non-sensitive rules also get lowered. The appealing nature of blocking approach is that it holds the integrity of the underlying primitive data, as it replaces an attribute value with some unknown value i.e. '?' rather than by replacing it with a false value ([6] – [7]).

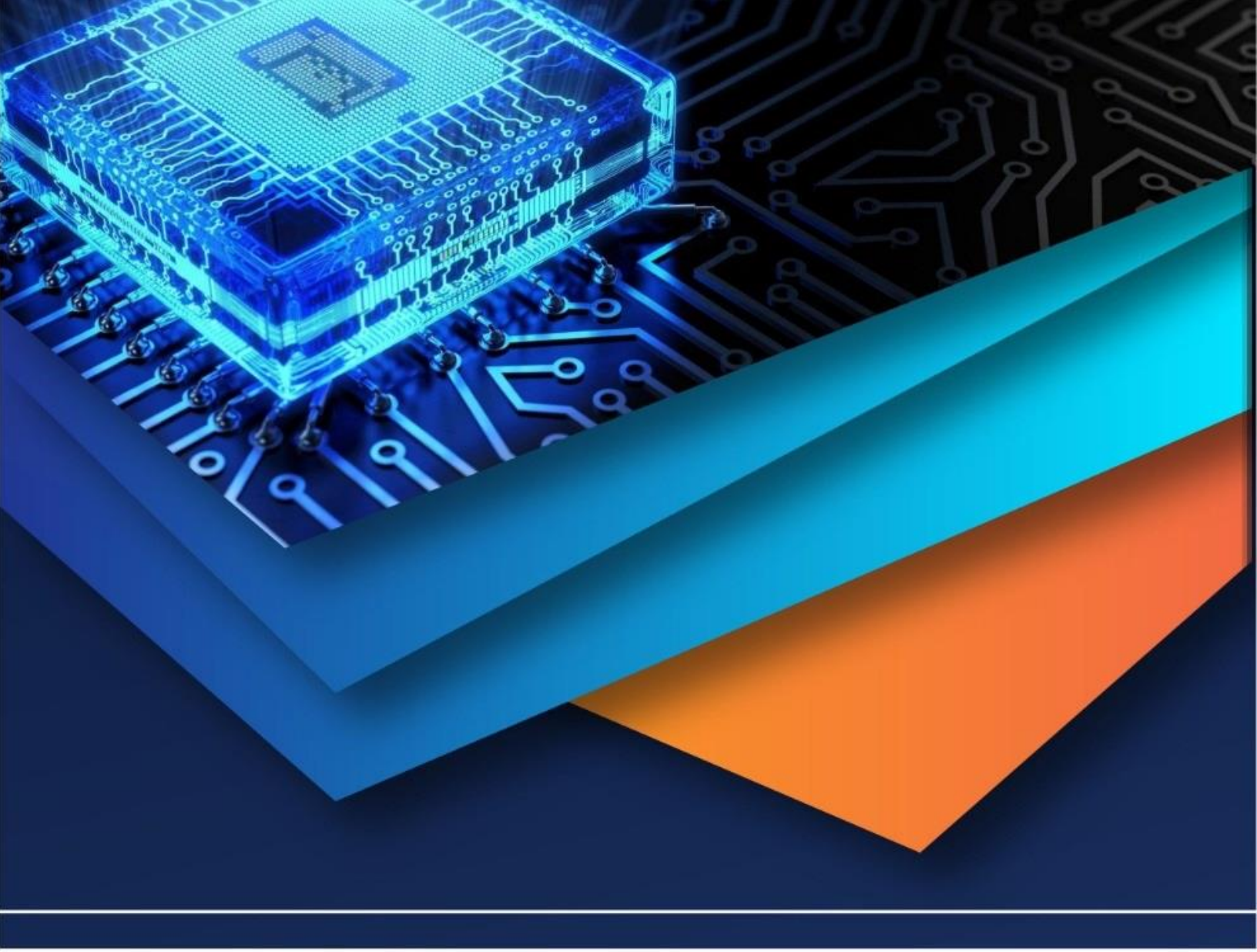
VIII. CONCLUSION

From above-reviewed papers, it concludes that cryptographic techniques provide top results, but their communication and computation cost is high there exist some loss of information. Cryptography based Secure Multiparty Computation (SMC) has the highest accuracy concerning data mining capability. Therefore, study work can be carried out via performing an analysis between various privacy preserving data mining techniques so that the performance can be better and filtering of data can also perform at different levels in such a manner that only filtered and sanitized data can only be in the reach of the users or clients.

REFERENCES

[1] Aldeen, Y.A.A.S., Salleh, M. and Razzaque, M.A., 2015. A comprehensive review of privacy preserving data mining. SpringerPlus, 4(1), (p.694).
[2] Patel, D.S., Tiwari, S. 2013. Privacy Preserving Data Mining. International Journal of Computer Science and Information Technologies, 4(1), pp.139-141.
[3] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M.Y., 2002. Tools for privacy preserving distributed data mining. ACM Sigkdd Explorations Newsletter, 4(2), pp.28-34.
[4] Aggarwal, C.C., and Philip, S.Y., 2008. A general survey of privacy-preserving data mining models and algorithms. In Privacy-preserving data mining. Springer US. (pp. 11-52)
[5] Midhun, B. and Rana, A., Privacy Preserving Data mining with Reduced Communication overhead. International Journal of Emerging Technology and Advanced Engineering, ISSN, pp.2250-2459.
[6] Gkoulalas-Divanis, A. and Verykios, V.S., 2010. Association Rule Hiding for Data Mining.
[7] Mistry Nikita, J. and Kharwar Ankit, R., A Review on Association Rule Hiding.
[8] Agrawal, R. and Srikant, R., 2000, May. Privacy-preserving data mining. ACM Sigmod Record, (pp. 439-450).
[9] Kantarcıoglu, M., Vaidya, J. and Clifton, C., 2003. Privacy preserving naive Bayes classifier for horizontally partitioned data. IEEE ICDM workshop on privacy preserving data mining (pp. 3-9).
[10] Friedman, A., Wolff, R., and Schuster, A., 2008. Providing k-anonymity in data mining. The VLDB Journal, 17(4), (pp.789-804).
[11] Belwal, R.C., Varshney, J., Khan, S.A., Sharma, A. and Bhattacharya, M., 2008, Hiding sensitive association rules efficiently by introducing new variable hiding counter. Service Operations and Logistics, and Informatics. IEEE/SOLI. IEEE, (pp. 130-134)
[12] Harnsamut, N. and Natwichai, J., 2008. A novel heuristic algorithm for privacy preserving of associative classification. PRICAI 2008: Trends in Artificial Intelligence, (pp.273-283).
[13] Dehkordi, M.N., Badie, K. and Zadeh, A.K., 2009. A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms. JSW, 4(6), (pp.555-562).
[14] Gkoulalas D., A. and Verykios, V.S., 2009. Exact knowledge is hiding through database extension. IEEE Transactions on Knowledge and Data Engineering, 21(5), (pp.699-713).
[15] Zhang, X. and Bi, H., 2010, Research on the classification of privacy preserving data mining based on random perturbation. Information Networking and Automation (ICINA), IEEE, (pp. 51-173).
[16] Kamakshi, P. and Babu, A.V., 2010. Preserving privacy and sharing the data in distributed environment using the cryptographic technique on perturbed data. ar xiv preprint ar xiv:1004.4477.
[17] Jain, Y.K., Yadav, V.K. and Panday, G.S., 2011. An efficient association rule hiding algorithm for privacy preserving data mining. International Journal of Computer Science and Engineering, 3(7), (pp.2792-2798).

[18] He, Y., Barman, S. and Naughton, J.F., 2011, Preventing equivalence attacks in updated, anonymized data. International Conference on Data Engineering (ICDE), (pp. 529-540). IEEE.
[19] Dev, H., Sen, T., Basak, M., and Ali, M.E., 2012, an approach to protecting the privacy of cloud data from data mining based attacks. High-Performance Computing, Networking, Storage, and Analysis (SCC), 2012 SC Companion: (pp. 1106-1115).
[20] Karim, M., Rashid, M., Jeong, B.S. and Choi, H.J., 2012. Privacy preserving mining maximal frequent patterns in transactional databases. Database Systems for Advanced Applications, Springer Berlin/Heidelberg (pp. 303-319).
[21] Ibrahim, Ayad, Hai Jin, Ali A. Yassin, and Deqing Zou. 2012, "Towards privacy-preserving mining over distributed cloud databases." Cloud and Green Computing (CGC), (pp. 130-136).
[22] Chan, J., and Keng, J., 2013. Privacy protection in outsourced association rule mining using distributed servers and its privacy notions.
[23] Giannotti, F., Lakshmanan, L.V., Monreale, A., Pedreschi, D. and Wang, H., 2013. Privacy-preserving mining of association rules from out-sourced transaction databases. IEEE Systems Journal, 7(3), (pp.385-395).
[24] Arunadevi, M. and Anuradha, R., 2014. Privacy preserving outsourcing for frequent itemset mining. International Journal of Innovative Research in Computer and Communication Engineering, 2(1), (pp.3867-3873).
[25] Xu, L., Jiang, C., Wang, J., Yuan, J. and Ren, Y., 2014. Information security in big data: privacy and data mining. IEEE Access, 2, (pp.1149-1176).
[26] Sharma, M., Chaudhary, A., Mathuria, M., Chaudhary, S. and Kumar, S., 2014, an efficient approach for privacy preserving in data mining. International Conference on Signal Propagation and Computer Technology (ICSPCT), (pp. 244-249).
[27] Rathna, S.S., and Karthikeyan, T., 2015. Survey on novel algorithms for privacy-preserving data mining. International Journal of Computer Science and Information Technologies, 6(2), (pp.1835-40).
[28] Modak, M. and Shaikh, R., 2016. Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy. Procedia Computer Science, 79, (pp.993-1000).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY