ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

Privacy Preserving in Data Mining using Bi-party Data Releasing Method

K. Noel Binny[1], Nisha Varghese[2]

Assistant Professor, Kovai Kalaimagal College of Arts and Science, Coimbatore, India[1]

Student, Kovai Kalaimagal College of Arts and Science, Coimbatore, India[2]

Abstract: Nowadays, tremendous amount of data is created and distributed to different repositories. With reduction in cost of storing information and existing infrastructures such as cloud computing and grid computing, there is the opportunity to extract knowledge or confidential data from that resources. In this topic the privacy preserving in Data mining works with Bi-Party Data Release Method. The exponential mechanism chooses a candidate that is close to optimum with respect to a utility function while preserving differential privacy. In the distributed setting, the candidates are owned by two parties and, therefore, a secure mechanism is required to compute the same output while ensuring that no extra information is leaked to any party. The proposed distributed exponential mechanism takes (candidate, score) pairs as inputs. The score is calculated using a utility function. The proposed distributed exponential mechanism is therefore independent of the choice of the utility function. In the case of vertically-partitioned data, we can use two types of utility functions: First, utility functions such as information gain, maximum function, and the widest (normalized) range of values that can be calculated locally by each party or Second, utility functions that cannot be computed locally. In the latter case, secure function evaluation techniques can be used by the parties to compute these utility functions. Once the scores of the candidates are computed using the utility functions in either case, they are ready to be used as inputs to execute the distributed exponential mechanism. The third party can request the data on the basis of anonymity and can view the data with the digital signatures provided by the both parties.

Keywords: Privacy Preserving, Anonymity, Utility Function

I. INTRODUCTION

Data mining is the process of analyzing large data sets from different types of repositories and uncovers the pattern and correlation to summarize data according to the problem. It is a multistep process, basically divided in to Data gathering and Preparation and Model building and Evaluation.

A. Data mining Implementation Process: Data mining implementation process lies in various sectors, they are

- Business understanding is the establishment of business and data-mining objectives. Initially, needs to understand about the business and client goals. It is required to specify what is demanded by the client (many a times this is not known to the clients themselves) and have a control over the present data mining environment.
- Data Understanding stage focus on the data pre-processing to assure the quality of data, with minimum noises in dataset.
- Data preparation contains cleaning and integration. Data cleaning removes the noisy and inconsistent data and the integration merges the multiple data sources.
- Data transformation makes the data in a transformed or consolidated form, which appropriate for mining. Some strategies for transformation are Smoothing, Aggregation, Generalization, Attribute construction or feature construction and Normalization.

B. Modeling: Modeling is the demonstration of the sequential pattern as a real world construct, on other hand it is the representation of knowledge. Depending on the business goals, appropriate modeling approaches have to be chosen for the dataset prepared, like hypercube.

C. Privacy Concerns: Nowadays our daily life is blended with Internet Of Things, so our daily activities are stored in databases, activities like a phone call or interview, browsing, smart or credit card details, any types of ticket booking, online registration and application, online consultation, E-commerce etc., in all cases a lot of information would be known about the concerned person of that site. This is a possibility that is real. This type of information i.e., the client side information (cookies) is already saved in any data source. In this situation privacy preserving takes advantage.

D. Privacy Preserving in Data Mining (PPDM): Privacy-preserving data mining focuses on the issue of executing data mining algorithms over sensitive or confidential data, which is not supposed to be exposed even to the party that

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

runs the algorithm. There are two important aspects to be taken into consideration for PPDM. First, confidential raw data such as identifiers, account numbers, credit card details, names, addresses etc., need be changed or cut out from the actual database, so that the recipient of the information is not capable of compromising the privacy of another individual. Secondly, confidential information that can be mined from a database by making use of data mining algorithms, also has to be left out, as such kind of knowledge can be an equal compromise over data privacy. Therefore, privacy preservation happens in two important dimensions, which include private information of users and information related to their collated activity.

- **Individual privacy preservation**

The foremost objective of data privacy is the preservation of individually recognizable information. Generally, information is regarded to be personally recognizable when it can be linked, either directly or indirectly, to an individual. Therefore, if private data are put to mining, the attribute values related to individuals are personal and need to be safeguarded from exposure. Then the miners are capable of learning from global models instead of learning the features of a specific person.

- **Collective privacy preservation**

Protection of confidential data might not be sufficient. At times, protection against learning of confidential knowledge that represents the activities of a certain group is required. The protection of confidential knowledge is known as collective privacy preservation.

E. Models of PPDM: In the research work involving privacy-preserving data mining, the important models are

- **Trust Third Party Model**

The standard objective for security is the supposition that there is a trustworthy third party to whom all data can be given. The third party carries out the computation and provides just the results – with an exception being the third party, it is evident that nobody has any information about anything, except those which are inferable from its individual input and the results. The aim of secure protocols is to attain this same degree of privacy preservation, without the issue of getting a third party that is trusted by everyone.

- **Semi-honest Model**

In the semi-honest model, each party adheres to the rules set by the protocol utilizing its right input, but then the protocol can use anything whatever is visible to it during the protocol execution to allow security compromise.

- **Malicious Model**

In the malicious model, no constraints are enforced on any of the persons taking part. Therefore any party has the entire freedom to be involved in whatever actions it desires. Generally, it is very hard to design effective protocols, which still hold their validity under the malicious model. But, the semi-honest model does not yield adequate protection for several applications.

F. Evaluation of privacy preserving algorithms: One significant aspect in the construction and evaluation of algorithms and tools, to be used for privacy preserving data mining is the discovery of appropriate evaluation criteria and the design of associated standards. A preliminary list consisting of evaluation parameters to be utilized for the assessment of the quality of privacy preserving data mining algorithms is provided as follows:

Privacy level provided by a privacy preserving approach that show how much approximate estimation on the confidential information, which has been concealed, can be made, in other words the make secure the transactional data between parties. Concealing failure, which is, the part of confidential information or sensitive data not concealed by the application of a privacy preservation approach;

Data quality after using a privacy preserving technique, regarded to be both as the quality of data themselves and the quality of the data mining outcomes once the hiding mechanism is used;

Complexity, that is, the ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm, i.e., the efficiency of the algorithm and techniques used much better preserve the confidential data.

II. REVIEW OF LITERATURE

A. Data-intensive applications, technologies and data analytics: Data Intensive Scientific Discovery (DISD) also termed as Data Mining problems are becoming highly popular. These Data Mining problems have been seen in varied applications of science, engineering and technology involving small scale industries as well as large scale and corporate organizations. Thus, it has become a social problem that has to be handled immediately in the present scenario. According to Jiawei Han and Micheline Kamber "Datamining refers to extracting or “mining” knowledge from large amounts of data”. Data Mining has already given its depth significance in various applications of scientific disciplines, which in turn resulted in the economic growth as well as advancements in technological innovations. Data analytics associated with database searching, mining, and analysis can be seen as an innovative IT capability that can improve firm performance. In the context of data analytics, capitalizing such external information may turn

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

out to be highly valuable for corporate decision making or accumulating business knowledge (Chen et al., 2012). Thus, processing external data for sense making becomes an integral part of big data analytics. A data source that particularly interests big data adopters is customer generated information from social media or social networking services, stored in multiple locations across several geographies.

B. Mining constrained frequent itemsets from distributed uncertain data: Nowadays, high volumes of massive data can be generated from various sources (e.g., sensor data from environmental surveillance). Many existing distributed frequent itemset mining algorithms do not allow users to express the itemsets to be mined according to their intention via the use of constraints. Consequently, these unconstrained mining algorithms can yield numerous itemsets that are not interesting to users. Moreover, due to inherited measurement inaccuracies and/or network latencies, the data are often riddled with uncertainty. These call for both constrained mining and uncertain data mining.

In the study of this topic choose many different datasets like transactional data, real-life databases as well as those from the Frequent Itemset Mining Implementation (FIMI) Dataset Repository. Regardless whether they are Apriori-based or tree-based; many frequent itemset mining algorithms provide little or no support for user focus when mining precise or uncertain data. In many real-life applications, the user may have to focus in mining for a small knowledge from this tremendous amount of data. In the execution of the algorithms for extraction, our system took the shortest amount of time to mine frequent itemsets because it pushes user-defined constraints into the mining process. The higher the selectivity of the constraints or retrieval queries, the longer was the runtime for our system. Both U-Apriori and UF-growth were not designed to handle constraints for uncertain data, let alone pushing the constraints into the mining. To handle constraints, U-Apriori and UF-growth first ignored the constraints and found all frequent itemsets. Then, they applied constraint checking as a post-processing step to prune those infrequent itemsets.

C. Enabling Multilevel Trust and Anonymity in PPDM: Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. A widely studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data are published. Previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. In this work, we relax this assumption and expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). Anonymity technologies enable Internet users to maintain a level of privacy that prevents the collection of identifying information such as the IP address. Understanding the deployment of anonymity technologies on the Internet is important to analyze the current and future trends.

Anonymity systems can be categorized by their latency, trust level, network type, anonymity properties, or adversary capabilities. From a usability point of view, anonymity systems are classified into two categories: high latency systems mostly used by non-interactive applications to provide strong anonymity, and low latency systems, mostly used by anonymous web browsing to have better performance. Anonymization-based privacy protection ensures that published data cannot be linked back to an individual. The most common approach in this domain is to apply generalizations on the private data in order to maintain a privacy standard such as k-anonymity while generalization-based techniques preserve truthfulness, relatively small output space of such techniques often results in unacceptable utility loss especially when privacy requirements are strict.

III. RESEARCH METHODOLOGY

A. Existing System: Previous, privacy preserving data mining has been studied widely. Association rule mining can cause potential threat toward privacy of data. So, association rule hiding techniques are employed to avoid the risk of sensitive knowledge leakage. Many researchers have been done on association rule hiding, but most of them focus on proposing algorithms with least side effect for static databases (with no new data entrance), while now the authors confront with streaming data which are continuous data. In this concept, new big data association rule hiding technique is presented, which uses fuzzy logic approach. The Fuzzy logic is an approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0) and the Boolean logic on which the modern computer is based. Fuzzy logic works similar to the working of brain, the fuzzy logic tries to decrease undesired side effect of sensitive rule hiding on non-sensitive rules in data streams. As mentioned, association rules should not be disclosed since they may be used to infer sensitive information. Although, many researches have been done in association rule hiding, there are significant drawbacks in most of them:

• Boolean logic versus fuzzy logic approach to check whether the association rule is sensitive or not in a frequent item set.
• Undesired side effect of hiding sensitive data using association rules on non-sensitive rules.

To solve mentioned problems, membership degree in fuzzy logic and Boolean logic is used to specify appropriate hiding level of each association rule apply to a particular dataset. Furthermore, anonymity techniques would be used for

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

rule hiding as an alternative for deleting some of the most repeated items and which makes secure confidential data items.

Fuzzy logic approach can be considered as an important part in large data set association rule hiding:

- Rules with confidence value near defined threshold are not as non-sensitive as rules with low confidence value.
- Regarding velocity feature of big data, probability of being sensitive rules for rules with confidence value near the defined confidence threshold is high.

Disadvantages

- Association rule mining is one of the most important data mining techniques. However, misuse of this technique may lead to the disclosure of sensitive or confidential information about users.
- Many researches have been done in association rule hiding and most of them present algorithms that delete items from transactions for hiding sensitive rules.
- Unfortunately, undesired side effect is obvious in these algorithms.
- To solve this problem, researchers try to use greedy-based approaches.

However, these approaches cannot guarantee finding an optimal solution and only try to increase their efficiency.

B. Proposed System: The proposed method for association rule data hiding is Bi-Party Data Release Method. The exponential mechanism chooses a candidate that is close to optimum with respect to a utility function while preserving differential privacy. In the distributed setting, the candidates are owned by two parties and, therefore, a secure mechanism is required to compute the same output while ensuring that no extra information is leaked to any party. The proposed distributed exponential mechanism takes (candidate, score) pairs as inputs. The score is calculated using a utility function (Frequent Pattern Mining).

The proposed distributed exponential mechanism is therefore independent of the choice of the utility function. In the case of vertically-partitioned data, we can use two types of utility functions: (1) utility functions such as information gain, maximum function, and the widest (normalized) range of values that can be calculated locally by each party or (2) utility functions that cannot be computed locally. In the latter case, secure function evaluation techniques can be used by the parties to compute these utility functions. Once the scores of the candidates are computed using the utility functions in either case, they are ready to be used as inputs to execute the distributed exponential mechanism. Rabin hash algorithm is used to generate the key while sharing of privacy data to the third parties.

Advantages

- In this research paper, new Bi-Party Data Release technique is presented, which uses Rabin hash algorithm, tries to decrease undesired side effect of sensitive rule hiding on non-sensitive rules in data streams.
- The proposed distributed exponential mechanism takes (candidate, score) pairs as inputs. The score is calculated using a utility function (Frequent Pattern Mining).
- Once the scores of the candidates are computed using the utility functions in either case, they are ready to be used as inputs to execute the distributed exponential mechanism.

IV. SYSTEM DESIGN AND DEVELOPMENT

The following are the important modules used in the proposed system.

- Party data splitting
- Candidate score Calculation
- Utility based Candidate prioritization
- Hash key value Generation
- Bi-party Data release Function

Party data splitting: This module is for collecting the data records with splitting of data for two parties. Both parties have a common primary key for identification and further integration. The data consists of sensitive information's and non-sensitive information's. The sensitive information should be kept privacy during maintaining and sharing of data to others. Candidate score Calculation: Encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events. Candidate score Calculation is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element.

Utility based Candidate prioritization: The proposed distributed exponential mechanism takes (candidate, score) pairs as inputs. The score is calculated using a utility function -Frequent Pattern Mining. The Frequent Pattern Mining Algorithm, proposed is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

patterns named frequent-pattern tree (FP-tree). In his study, Han proved that his method outperforms other popular methods for mining frequent patterns.

Hash key value Generation: A hash key value Generation is any function that can be used to map data of arbitrary size to data of a fixed size. The values returned by a hash function are called hash values, hash codes, digests, or simply hashes. Hash functions are often used in combination with a hash table, a common data structure used in computer software for rapid data lookup. Hash functions accelerate table or database lookup by detecting duplicated records in a large file. One such application is finding similar stretches in DNA sequences. They are also useful in cryptography. A cryptographic hash function allows one to easily verify that some input data maps to a given hash value, but if the input data is unknown; it is deliberately difficult to reconstruct it (or any equivalent alternatives) by knowing the stored hash value. This is used for assuring integrity of transmitted data, and is the building block for HMACs, which provide message authentication.

- **Rabin hash algorithm**

In this project Rabin Karp algorithm makes secure keys to access the data for a third party in an asymmetric or a public key encryption method, which provides a base level security for authentication.
A brute-force substring search algorithm checks all possible positions:
1 function NaiveSearch(string s[1..n], string pattern[1..m])
2 for i from 1 to n-m+1
3 for j from 1 to m
4 if s[i+j-1] ≠ pattern[j]
5 jump to next iteration of outer loop
6 return i
7 return not found

Bi-party Data release Function: This module takes the output of last module which discussed. This module is to find out the candidates from the two party splatted data. The sensitive information's are taken into account and used for further splitting, it is called candidate selection. The both party will have the candidate selection and the original data are grouped with the candidates. The Bi-party algorithm for differentially private data release for vertically partitioned data. We present our Distributed Differentially-private anonymizations algorithm based on Generalization for two parties. The algorithm first generalizes the raw data and then adds noise to achieve differential privacy.

V. EXPERIMENTAL RESULT AND ANALYSIS

We shall explain the proposed results in detail, and then summarize the relationship between these results and the data distributions. We evaluate their execution and performance based on three important criteria: Accuracy, Precision, and Recall. The most often metric used to determine the performance of classifier is accuracy. Since the accuracy is not desirable when data is imbalanced, we used another metrics to compare the performance. It shows constant accuracy even though the data has been randomized 30 times. Bi-Party method can classify the result better than other classifiers. Recall measures how often a positive class instance in the dataset was predicted as a positive class instance by the classifier. Precision measure how often an instance that was predicted as positive that is actually positive.
Here we compare proposed and existing algorithms are
1. Bi-Party Release (Bi-PR),
2. Association rule mining (ARM),
3. Heuristic for Confidence and Support Reduction Based on Intersection Lattive (HCSRIL).

Accuracy (%): We can observe that our proposed outperforms the others in almost all of the cases. Our proposed linear structure to its trees instead of the previous tree form in order to minimize access times to search nodes. As a result, its advantages have a positive effect on reducing runtime in whole experiments. Especially as the minimum support threshold becomes lower, the difference of runtime between our algorithm and the others are bigger.

Table1 Accuracy Results

No of Data	Bi-PR	ARM	HCSRIL
100	71.0	73.8	84.6
200	71.3	76.6	87.8
300	71.5	78.9	89.5
400	71.8	79.5	90.3

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

Precision (%): Proposed algorithm shows the best Precision while the others have relatively poor performance, which indicates that our scheme can store these increasing attributes more efficiently than the other structures of the competitor algorithms. Through the above experimental results, we know that the proposed algorithm outperforms the others with respect to increasing transactions and items in terms of scalability as well as runtime and memory usage for the real datasets.

Table2 Precision Results

No of Data	Bi-PR	ARM	HCSRIL
100	72.5	76.7	87.5
200	73.5	77.5	89.4
300	74.5	78.7	90.5
400	75.0	79.5	91.0

Recall (%): Through the above experimental results, we know that the proposed algorithm outperforms the others with respect to increasing transactions and items in terms of scalability as well as runtime and memory usage for the real datasets.

Table 3 Recall Results

No of Data	Bi-PR	ARM	HCSRIL
100	77.0	81.5	91.0
200	78.5	82.5	92.4
300	79.5	83.5	93.9
400	80.5	84.8	95.5

VI. GRAPH RESULTS

Accuracy

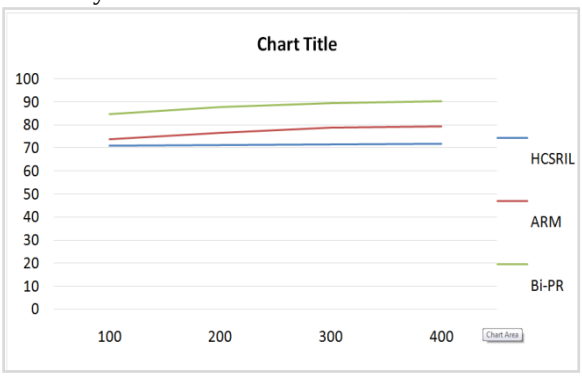

Fig 1 Accuracy Results

Precision

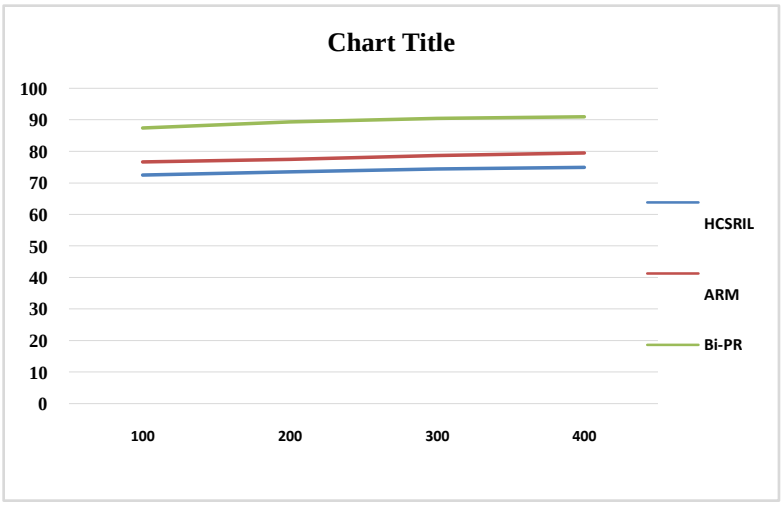

Fig 2 Precision Results

DOI 10.17148/IJARCCE.2018.71017

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

Recall

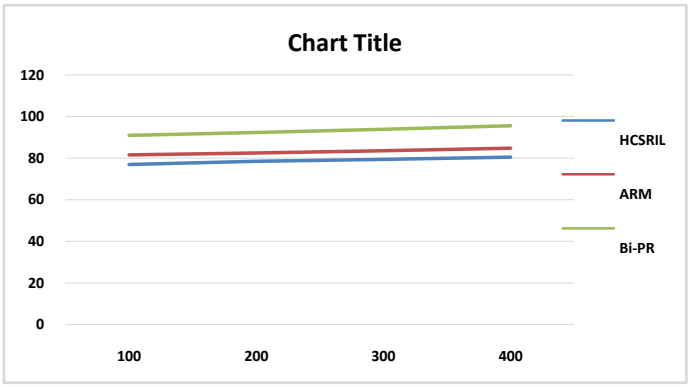

Fig 3 Recall Results

Our experimental results showed that Bi-Party Release scheme outstanding performance in terms of accuracy, precision, recall, memory usage, and scalability. Our proposed work shows significant improvement in classification accuracy, precision, recall and in survival probability. This significant effect indicates that use of a Bi-Party Release in decision support systems has multiple benefits, both in terms of system accuracy and in terms of system transparency.

CONCLUSION

In this framework, we have obtainable the first Bi-Party Release Algorithm differentially-private data release procedure for vertically-partitioned data. We have shown that the suggested procedure is differentially-private and protected in the safety classification of the adversary model. Moreover, we have experimentally assessed the data utility for classification analysis. The recommended process can excellently hold crucial evidence for taxonomy investigation. It delivers alike data utility associated to the freshly projected single party algorithm and improved records utility than the dispersed Frequent Pattern Mining algorithm for utility classification analysis. The taxonomy perfections of this process, in concurrence with a sureness portion articulating the per-sample prospect of sorting failure is defined and unrushed. Our outcome expressions that the Bi-Party Release classifier is translucent, constant, forthright, unpretentious to understand, great tendency to hold necessary assets and informal to contrivance than supreme other contrivance erudition methods definitely when there is little or no prior knowledge about data distribution.

SCOPE OF FUTURE WORK

In the future work, we can improve the Rabin-Karp algorithm and utility function; instead of Rabin-Karp algorithm we can also use any cryptographic algorithms (like IDEA, DES, BLOWFISH, etc.) and digital signatures to provide more internal security. The improvements of anonymity methods also provide. In the future enhancement, we can increase the efficiency of the item set by using Dynamic Item set Counting algorithm and the bi-directional method like Pincer Search algorithm are suitable to big sized data item set or the most recent association mining algorithms provide best and more reliable results and the improve the performance of the system.

REFERENCES

[1]. Philip, C.L.C., Zh, C.-Y.: 'Data-intensive applications, challenges, techniques and technologies: a survey on big data', Inf. Sci., 2014, 275, pp. 314–347
[2]. Ohbyung, K., Namyeon, L., Bongsik, S.: 'Data quality management, data usage experience and acquisition intention of big data analytics', Int. J. Inf. Manage., 2014, 34, (3), pp. 387–394
[3]. Alfredo, C., Carson, K.S.L., Richard, K.M.: 'Mining constrained frequent item-sets from distributed uncertain data', Future Gener. Comput. Syst., 2014, 37, pp. 117–126
[4]. Xuyun, Z., Chang, L., Surya, N.S., et al.: 'A hybrid approach for scalable subtree anonymization over big data using MapReduce on cloud', J. Comput. Syst. Sci., 2014, 80, (5), pp. 1008–1020
[5]. Yaping, L., Minghua, C., Qiwei, L., et al.: 'Enabling multilevel trust in privacy preserving data mining', IEEE Trans. Knowl. Data Eng., 2012, 24, (9), pp. 1589–1612
[6]. Yi-Huang, W., Chia-Ming, C., Arbee, L.P.C.: 'Hiding sensitive association rules with limited side effects', IEEE Trans. Knowl. Data Eng., 2007, 19, (1), pp. 29–42
[7]. Aris, G.D., Vassilios, S.V.: 'Exact knowledge hiding through database extension', IEEE Trans. Knowl. Data Eng., 2009, 21, (5), pp. 699–713
[8]. Hai, Q.C., Somjit, A.I., Huy, X.N., et al.: 'Association rule hiding in risk management for retail supply chain collaboration', Comput. Ind., 2013, 64, (4), pp. 776–784
[9]. Yu-Chiang, L., Jieh-Shan, Y., Chin-Chen, C.: 'MCIF: an effective sanitization algorithm for hiding sensitive patterns on data mining', Adv. Eng. Inf., 2007, 21, (3), pp. 269–280
[10]. Anna, M., Gennady, A., Natalia, A., et al.: 'Movement data anonymity through generalization', Trans. Data Priv., 2010, 3, (2), pp. 1–31
[11]. Gaofen, Z., Yun, Y., Xiao, L., et al.: 'A time-series pattern based noise generation strategy for privacy protection in cloud computing'. Int. Symp. Cluster, Cloud and Grid Computing (CCGrid), Ottawa, Canada, May 2012, pp. 458–465
[12]. Hui, W.: 'Quality measurement for association rule hiding', AASRI Procedia, 2013, 5, pp. 228–234