International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 11 (2018) pp. 10048-10057

Double-Blind Key-Attribute Based Encryption Algorithm for Personalized Privacy-Preserving of Distributed Data Mining

S.Urmela, M.Nandhini

[1,2]*Department of Computer Science, Pondicherry University, Puducherry, India.*

Abstract

Distributed Data Mining (DDM) system contain large amount of private and sensitive data from each local heterogeneous distributed datasites. These private and sensitive data cannot be shared to other local data sites or to global level, so privacy protection of data is required in DDM. To deploy the protection of data in global and local level, Double-Blind Key-attribute based Encryption (DBKE) algorithm is proposed. Proposed DBKE algorithm is evaluated with Electronic Health Records (EHRs) dataset. At local level, encryption of key-attribute followed by hashing function is applied. The encrypted key-attribute is decrypted at global level. The results are analyzed with state-of-art privacy preserving approaches: anonymization and randomization.

Keywords: Distributed Data Mining, distributed data-sites, privacy-preserving, Key-Attribute Based Encryption, Electronic Health Records.

INTRODUCTION

Data Mining (DM) is the process of data extraction using well-known pattern matching techniques. DM technique includes clustering, regression, rule-association, etc. Conventional or traditional DM performs computation on warehouse or database situated at a single geographical location. Current trend of DM computes data at different geographical locations termed DDM/CDM (Distributed Data Mining)/(Collective Data Mining). DDM computes data at different geographical location, hosting computing units at each heterogeneous points[1]. In recent years, DDM has gradually become an important task in decision-making. Increased usage of DDM in both the sectors has led to face many issues. Usually the data is very sensitive to privacy issues. The local data sites of distributed environment wish to work collaboratively but doesn't completely trust neighbouring data sites and global level. Privacy preserving DDM (PPDDM) is a new research area in DDM that builds valid DM models and extracts useful patterns without compromising the confidentiality of sensitive information. Health-care records, financial transactions, personal profile of clients are few examples. The main purpose of PPDDM is to develop new approaches/techniques to modify the original data without compromising the privacy-preserving part of sensitive data. There is a necessity to protect the sensitive data throughout the DM process[2].

Privacy-preserving algorithms designed for centralized DM cannot be used for DDM because in centralized environment, data is stored in a single geographical location whereas in distributed environment, data is distributed among local and global level. PPDDM mainly considers two aspects, guarantee that sensitive information is not revealed in data manipulation process (Confidential information is removed from the original/initial database) and designing more efficient privacy-preserving algorithms[2].

Since the data are huge and voluminous, general-purpose privacy algorithms are ineffective for specific applications. Research in this area seeks more efficient and reliable privacy preserving algorithm for specific functions. There are many variations of applying PPDDM, depending on how the data in datasite is distributed, type of DM algorithm to be applied, and amount of sensitive information to be privacy-preserved[3]. Inspired by the research work in this area, this paper proposes a privacy-preserving algorithm for DDM.

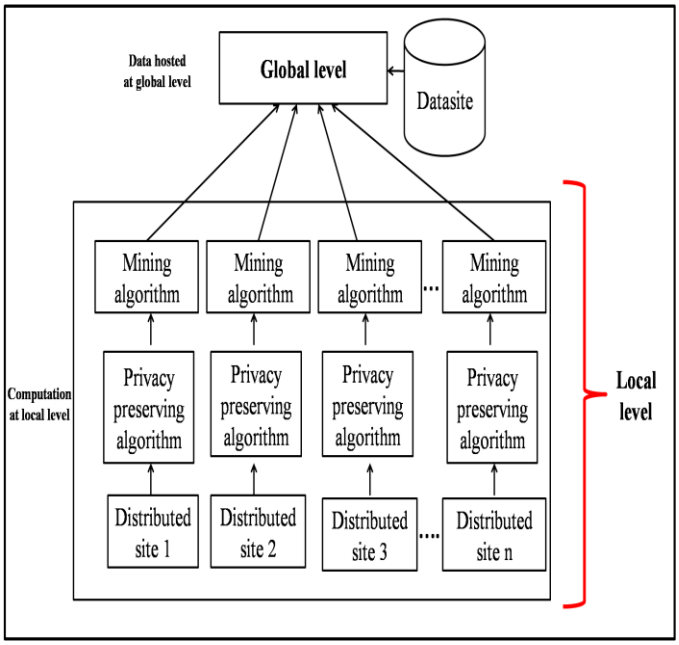

Figure 1. DDM – privacy preserving model

Fig 1. depicts privacy preserving model of DDM. At local level for each distributed data site, on analyzing the sensitive data privacy preserving algorithm is applied followed by DM algorithm. The accumulated privacy-preserved data from local level is merged at global level for final prediction.

PRIVACY-PRESERVING APPROACHES OF DDM

The second level is privacy preserving approach: confidential and sensitive information to be communicated is protected among heterogeneous data sites. Privacy preserving approach for DDM can be classified into two approaches: anonymization and randomization approach is shown in fig. 2 based on level of privacy [4].

Anonymization approach

Anonymization technique prevents the disclosure of critical/sensitive data identity to preserve the privacy. Initial solution of anonymization approach is de-identification, removing all critical data prior to spreading the data in distributed environment. But to identify a particular entity, certain key attributes are stored in external database in global/local level of DDM. Anonymization approach preserves these key attributes by modifying the values of external database[5].

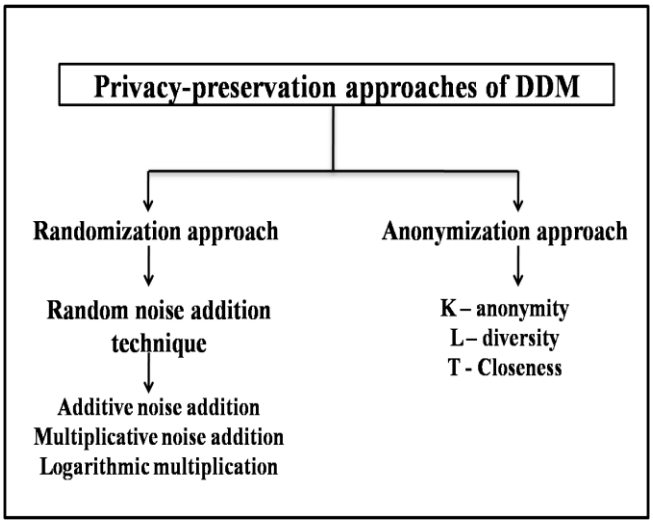

Figure 2. Privacy-preserving approaches of DDM

K-anonymity technique

Each entity must be distinguished by k-1(k – number of attributes) with other entity in a dataset. Data generalization or suppression is used to achieve k-anonymity. This technique consists of some drawbacks. It is very difficult for data owners of each data site to decide on which key attributes is available in external database. This technique suffers from homogeneity attack (same key attributes of each datasite) and external database attack (knowledge of key attributes of neighbouring datasite)[6].

L-diversity technique

This technique was proposed to solve homogeneity attack of k-anonymity technique by storing a variety of sensitive key attributes of each data site instead of single key attribute. But this technique needs to store unnecessary key attributes and it is difficult to achieve. Again this technique suffers from similarity attack/homogeneity attack[7].

T-closeness technique

L-diversity technique treats all attributes value same as sensitive data. In this technique the distance value of sensitive attribute in its data site and distribution in whole environment should be within a threshold range. This technique suffers in calculation of distribution value[8].

On the whole, anonymization technique is scalable and simple to implement. Any data which is added new to the environment can be privacy-preserved without starting from the scratch. Moreover, this technique is prone to attacks like similarity attack, external database attack, etc[8].

Randomization approach

Randomization approach modifies a part or whole of data to preserve data privacy. This approach adds random noise to the original data. This approach is less efficient than anonymization approach because from the spectral values of randomized data, original data can be identified. If privacy is least important and result accuracy is more important, then randomization approach is suitable[9].

Random noise addition technique

This technique works on by adding random noise to sensitive attributes to provide data privacy. It works by adding or multiplying random number to sensitive attributes. This random number is chosen from probability distribution[10].

Additive noise addition technique

The original data is transformed to a privacy-preserved data by adding additive noise computed from probability distribution. Consider D as original data, then transformed/perturbed data T is obtained by adding random variable, β (noise) to original data. Thus,

$$T = D + \beta,$$

by this technique random data is added to original data to conceal the sensitive attributes[10].

Multiplicative noise addition technique

The original data is transformed to a privacy-preserved data by adding multiplicative noise computed from probability distribution. Consider D as original data, the transformed data T is obtained by multiplying random variable β(noise) to

$$T = D\beta,$$

by this technique random data is multiplied to original data to conceal the sensitive attributes of original dataset[10].

Logarithmic multiplication technique

Privacy-preserved data is obtained by adding a logarithmic alteration to original dataset. The random variable (noise) is generated and added to the D, original data. Thus,

$$Y = \log D,$$

The computed random variable, Y is added to the noise generated,

$$T = Y + \beta,$$

by this technique random data is added to original data to conceal the sensitive attributes[10].

On the whole, it is found that there is privacy-leakage on random noise technique. From the spectral values of transformed data, random noise can be identified easily to obtain original data. Establishing a balance between privacy and accuracy is tedious task. Combined effect of anonymization and randomization approaches can balance this trade-off. The proposed work in this paper concentrates on both randomization and anonymization approaches.

RELATED WORKS - PRIVACY-PRESERVING BASED APPROACH FOR DDM

Privacy preserving is a critical issue when different data owners' exclusively gains access to the knowledge from heterogeneous data by pattern evaluation techniques but they inclusively reveal private data to neighbour data sites during execution. PPDDM success relies on building an efficient DM model for finding useful and meaningful patterns by preserving the confidentiality of private data[11]. From the dimensional levels of PPDDM, DM algorithms are built with classification, clustering and rule association. PPDDM can be implemented in two-ways: cryptographic technique and randomization approach. Randomization approach suffers from tradeoff between privacy and accuracy but cryptographic approach provides better accuracy but lacks in data privacy preservation. PPDDM is particularly applicable in almost all mining areas namely ensemble methods, rule association mining, clustering, decision tree and bayesian model[12].

Privacy preserving in Electronic Medical Record (EMR) mining is on demand in distributed system since confidentiality of data from several heterogeneous data sites needs to be preserved. Yan Li et al.[13] (2018) discussed a ensemble distributed technique for extracting patient data. Each heterogeneous site will extract the knowledge pattern from local level data. At global level, final prediction model is generated from accumulated-local models. Merits of this proposed novel technique: less communication cost and computational complexity.

Masooda Modak et al.[14] (2018) discussed horizontal and vertical data privacy preservation with association rule mining. Vertically partitioned data uses distributed Apriori T-tree algorithm on vertical partitioned heterogeneous data. Horizontal partitioned data uses collaborative approach; only association rule values calculated will be shared to neighbouring sites. Each distributed site on generation of association rule will decide whether to reveal its value to neighbouring sites or not. A controller takes responsibility of migrating rule values calculated to distributed data sites. Transfer rate of generated rules is increased and execution time is 204 seconds, half the time of conventional approach.

Yiannis Kokkinos et al.[15] (2018) discussed ensemble method privacy preservation by confidence ratio affinity propagation. Confidence ratio affinity propagation is used for selecting neural network classifiers by privacy computing. At local level, ensemble classifier classifies the homogeneous data. Most suitable confidence ratios are computed between each 2 data site. Pruning is done after calculating confidence ratio affinity propagation among classifiers. Time complexity of O(CE) for C classifiers and E examples is computed for UCI and KEEL data repository.

Hemanta Kumar Bhuyan et al.[16] (2017) discussed confidential data security protection in DDM by fuzzy method. Borel set is used to generate two-fuzzy sets, which helps in determination of sub-features within certain range. Proposed approach shows effective performance compared to conventional methods. Data privacy is maintained. Experimental implementation depicts reduced computation time on hepatitis (157 instances), yeast (1484 instances) and heart disease datasets compared to conventional approaches. Merits of this proposed technique is privacy preservation of original data and efficient sub-feature selection.

Feng Zhang et al.[17] (2016) discussed Distributed Association Rules Mining over homogeneous data along with data privacy. Proposed method uses commutative encryption for sensitive data-preserving distributed association rules mining. A computation protocol by secure division is developed as core protocol for the same. Proposed approach reveals increased computation cost and communication cost compared to conventional approach.

Xinjun Qi et al.[18] (2016) discussed DM privacy preserving classification by five approaches. Data distribution, itemsets distortion, DM algorithms, itemsets hiding and confidential data protection are the five approaches. Other approach discussed is association rule mining by perturbation, hiding association rules generated and sharing rules value to neighbouring distributed homogeneous data sites.

Yanguang Shen et al.[19] (2016) discussed personalized PPDDM which combines K-anonymity technique along with decision tree and multi-party secure computations. Initially less-confidential data are anonymized, used for classifying data but becomes useless for sharing with other distributed homogeneous data sites. Multi-party secure computations are used to share sensitive data to prevent sensitive data leakage. two-divisions: non-sensitive data and sensitive data are computed. Non-sensitive data are shared via K-anonymity technique and sensitive data are shared via multi-party secure computations. Linking rate is reduced; by removing sensitive attribute-value pair. Collaborative calculation of information gain for each attribute is computed assisting in building decision tree and targeted information is chosen of attribute with maximum value. Multi-party secure computations are used to recover attribute of sensitive data. Merits of proposed

approach depict efficiency and minimum overhead in communication and cost.

Rebecca N.Wright et al.[20] (2015) discussed four protocols for DDM. Bayesian network is implemented; dividing a huge problem into smaller sub-problems for privacy-preserving of original data. On Bayesian network a secured two-party computation is applied. Experimental results depict accuracy and efficiency. A fully distributed k-anonymization model, privacy preserving learning classification model is represented.

Xun Yi et al.[21] (2015) discussed confidential data preserving Distributed Association Rule mining model by semi-trusted mixer. Data from each heterogeneous data site is fed to a semi-trusted mixer model which mixes messages and dispatches result to neighbouring data sites. A strong global association rule is obtained from the local level distributed data sites. Two messages per data communication are required to send and receive via mixer and distributed data site. Minimized communication cost and storage cost are the merits of this work.

Though all the above discussed works concentrates on privacy-preserving of distributed data, privacy-preserving of universal data like universal EHRs is not done. Privacy-preserving DDM works with EHRs concentrates on only centralized privacy (local level of each datasite) and not on global level privacy. The proposed PPDDM algorithm concentrates on both local and global level privacy. Table 1 summarizes privacy-preserving approach of DDM along with its pros, cons, performance evaluation results and dataset used for approving the approach.

Table 1. Privacy-preserving approach for DDM

Title	Author	Technique	Pros	Cons	Dataset	Evaluation Result
A distributed ensemble approach for mining health care data under privacy constraints	Yan Li et al.[13] (2018)	**Ensemble Learning** Adaptive distributed privacy-preserving DM by AdaBoost	Communication cost lesser compared to star network (untrusted third party), Complexity is lesser when new participator is added	Memory overhead in learning other participator models	Type 2 – diabetes Real-time EHRs of 9948 patients from top 14 states	F-measure of 0.7 average
Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy	Masooda Modak et al.[14] (2018)	Distributed Association Rule Mining by concept of hierarchy	Privacy-preserved by revealing only association rules calculated	Accuracy	Supermarket database	Time complexity is 204 secs half value of conventional approaches
Confidence ratio affinity propagation in ensemble selection of neural network classifiers for distributed privacy preserving DM	Yiannis Kokkinos et al.[15] (2018)	Confidence ratio affinity propagation is used to select neural network classifiers by privacy computing	Time complexity	Doesn't consider per-class information	Datasets from UCI and KEEL data repository	Cost is O(CE) for C classifiers and E examples reduced than conventional approaches
Privacy preserving sub-feature selection in distributed DM	Hemanta Kumar Bhuyan et al.[16] (2017)	Sub-feature selection involves fuzzy methodology Borel set generates 2 fuzzy set which determines sub-feature selection	Efficient sub-feature selection approach and Privacy of original data	During developing fuzzy membership function outlier values are discarded	Hepatitis (1527instances) yeast (1484instances) heart disease (1568instances) datasets	Computation time reduced than conventional approaches
Privacy-Preserving Two-Party Distributed Association Rules Mining on Horizontally Partitioned Data	Feng Zhang et al.[17] (2016)	Computation protocol by secure division for privacy preserving	Communication and computation cost	Result accuracy by certain loopholes present	– – – – – –	Computation cost reduced than conventional approaches
Classification of Privacy-preserving Distributed Data Mining Protocols	Zhuojia Xu et al.[18] (2016)	Four dimension approach for DDM privacy-preserving	Survey paper	Survey paper	– – – – – –	– – – – – –
Research on the Personalized Privacy Preserving Distributed DM	Yanguang Shen et al.[19] (2016)	Multi-party secure computations and K-anonymity technique with decision tree classification	Efficient and minimum overhead in computation	Accuracy of mined data	Adult dataset of U.S census data	Minimum computation overhead

Title	Author	Technique	Pros	Cons	Dataset	Evaluation Result
Distributed DM Protocols for Privacy: A Review of Some Recent Results	Rebecca N.Wright et al.[20] (2015)	k-anonymity algorithm	Survey paper	Survey paper	_ _ _ _ _ _	_ _ _ _ _ _
Privacy-preserving distributed association rule mining via semi-trusted mixer	Xun Yi et al.[21] (2015)	Semi-trusted mixer, mixes the messages and then dispatches the result to other distributed data sites	Minimized computation cost since for each data communication only 2 messages are sent to and fro mixer and distributed data site	Accuracy of mined data	Medical dataset	Minimized computation cost of nearly 50% compared to Kantarcioglu Clifton protocol

PROPOSED DOUBLE-BLIND KEY-ATTRIBUTE BASED ENCRYPTION (DBKE) ALGORITHM

Privacy preserving of local data at heterogeneous distributed data sites and at global level is crucial for applications like Electronic Health Records, military information, etc. As discussed in dimensional level of PPDDM, the proposed algorithm fits in second dimensional level of PPDDM - privacy preserving technique. Each heterogeneous datasite includes data management engine which depicts 3 phases of data processing.

1. Filling missing values
2. Data selection
3. Privacy preservation algorithm - Double-Blind Key-attribute based Encryption (DBKE) algorithm

Filling missing values

At each distributed data sites, the missing values are filled from the relative data of corresponding entity. For any type application, there are 2 types of filling missing values.

Filling demographic data and filling application-specific data. Demographic data includes filling general information. Example: filling entity age from DOB, etc. Application-specific data includes filling application-oriented data[22]. For medical dataset, clinical results of patient are filled by relative computation of corresponding data. For example, say in hypertension dataset if the value of SBP (Systolic Blood Pressure)/DBP (Diastolic Blood Pressure) for a continuous range is 121/76, 134/74, 152/78 then the next value is computed using relative mean as (136/76)[22].

Data selection

In proposed work, the data structure used for storage of heterogeneous and homogeneous distributed data sites (horizontally and vertically partitioned data) is multi-linked list. Consider EHRs dataset, the data storage at local distributed data site is depicted in fig. 3.

From fig. 3, say, for a patient at each local datasite, if SBP/DBP value is monitored daily, it leads to huge volume of data values. To overcome the problem of maintaining history of records, for every 20 data values of SBP and DBP database is consolidated. Max and min value of SBP and DBP among the old 20 records along with time recorded of max and min values and the time data was consolidated is noted.

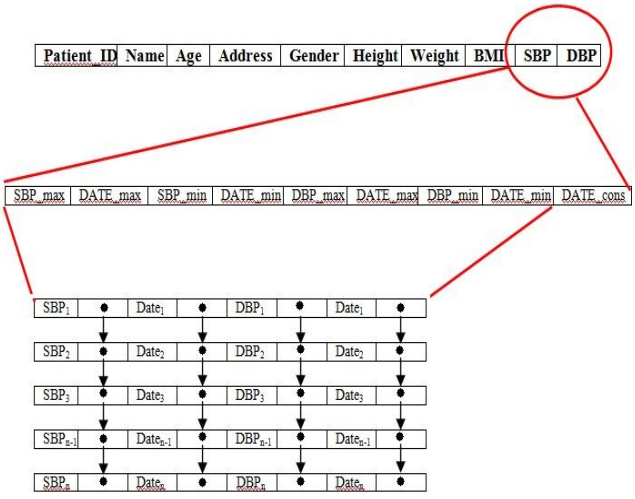

Figure 3. Data Structure representation

Calculation of MAX and MIN of key-attributes

Before explaining DBKE algorithm, consider EHRs dataset, in fig. 3 SBP and DBP values are stored as multi-linked list data. From the clinical data, MAX and MIN values of key-attributes are calculated using which clusters are formulated for mining. For example if SBP value for a single entity reading is 120,134,115,132,134,121,117,128 then MAX = 134 and MIN = 115. Similarly DBP value reading is 80,79,83,82,82,84,82,80 then MAX = 84 and MIN = 79. At each heterogeneous distributed datasite, attributes name is different.

Privacy preservation – Double-Blind Key-attribute based Encryption (DBKE) Algorithm

The calculated (MAX,MIN) pair is highly-sensitive data which needs to be protected and kept confidential at local level of each distributed datasite and during migration of data from local level of each distributed site to global level for resultant query. The value of (MAX,MIN) pair of key attribute is used for mining. The proposed Double-Blind Key-

attribute based Encryption (DBKE) algorithm at local level of each heterogeneous distributed datasite employs an encryption function followed by hashing function. The key used for encryption function and hashing function of each local distributed datasite is stored as a secret share between each heterogeneous distributed local datasite and global level.

Blind function is a service provided by agent to client which computes function for encrypting the data without revealing the original data. In proposed algorithm, Double-Blind technique is followed which computes 2 encryption functions at local level and 2 decryption functions at global level to protect sensitive key-attribute data.

Processing at local level

The (MAX,MIN) pair computed data of key-attribute is twice privacy-preserved with an encryption function and a hashing function. At first level, the (MAX,MIN) value of key-attribute is encrypted using symmetric encryption key.

The encryption function is computed as,

Encrypted data, $E_1 = E(K,I)$

Encrypted data, $E_1 = E(K,(MAX,MIN))$

where I - input data (MAX,MIN) and K - key value.

At second level, hashing function is applied to the encrypted data, which computes hashing function individually for MAX and MIN of key-attribute.

Hash function, $F(E_1) = E_1 \text{ MOD } N$

where E_1 – Encrypted data and N – key value.

The hashed (encrypted) value of (MAX,MIN) pair obtained is depicted in table 2.

Table 2. Hashing Function Explanation

Hash key	Hashing function	Hashing value
E(MAX,MIN)	E_1MOD N	$Value_1$

Processing at global level

The Double-Blind encrypted (MAX,MIN) pair migrated from local level of each distributed datasite is decrypted in reverse. At first level, unhashing function (Reverse decryption) is applied on double-blind encrypted (MAX,MIN) data. On knowing the N, key value used in hashing function, reverse decryption process is possible which computes the MAX and MIN value individually. The output of unhashing function is D_1.

At second level, decryption function is applied on the single-decrypted (MAX,MIN) data to obtain the original (MAX,MIN) pair.

The decryption function,

Input data, $I = D(D_1,K)$

where, K – key value used for encryption at local level,

D_1 – single-level decryption output (Unhashing function).

Double-Blind Key-attribute based Encryption algorithm on EHRs :

The DBKE algorithm working on EHRs at local level and global level is discussed in this section. Consider (MAX,MIN) pair of key-attribute (say, SBP), DBKE algorithm works as follows,

At local level depicted in fig. 4,

Step 1: Encryption function by symmetric encryption key

Step 2: Hashing function by key-value

At global level depicted in fig. 4,

Step 3: Unhashing function by reverse decryption

Step 4: decryption function by symmetric decryption key

At local level above discussed, step 1 and step 2 on (MAX,MIN) pair of SBP takes place. Input data (MAX,MIN) pair is shown in table 3.

Table 3. (MAX,MIN) of SBP

MAX	MIN
120	115
140	121
134	120
145	130
138	131

Step 1: Encryption function by symmetric encryption key

Encrypted data, $E_1 = E(K,(MAX,MIN))$

Table 4. Single-encryption by symmetric key encryption

MAX	MIN
127	122
147	128
141	127
152	137
145	138

Step 2: Hashing function by key-value

Hash function, $F(E_1) = E_1 \text{ MOD } N$

The key value K is distributed as a secret share among each local level and distributed datasites. For each local level different keys are generated at distributed key management center and shared secretly only to corresponding local level and global level.

Table 5. Hashing function on (MAX,MIN) pair of SBP

Hashing of MAX value of SBP			Hashing of MIN value of SBP		
Hash key	**Hashing function**	**Hashing value**	**Hash Key**	**Hashing function**	**Hashing value**
127	127 MOD 7	1	122	122 MOD 7	3
147	147 MOD 7	0	128	128 MOD 7	2
141	141 MOD 7	1	127	127 MOD 7	1
152	152 MOD 7	5	137	137 MOD 7	4
145	145 MOD 7	5	138	138 MOD 7	5

Step 3: Unhashing function by reverse decryption

At global level, reverse decryption, unhashing function on double-blind encrypted (MAX,MIN) pair takes place.

Table 6. Unhashing function on encrypted(MAX,MIN) pair of SBP

UnHashing of MAX value of SBP		UnHashing of MIN value of SBP	
Hashing value	**Hashing key**	**Hashing value**	**Hashing key**
1	127	3	122
0	147	2	128
1	141	1	127
5	152	4	137
5	145	5	138

Step 4: decryption function by symmetric decryption key

Table 7. Decryption function on unhashed (MAX,MIN) pair of SBP

Decryption of MAX value of SBP		Decryption of MIN value of SBP	
Hashing key	**MAX value**	**Hashing key**	**MIN value**
127	120	122	115
147	140	128	121
141	134	127	120
152	145	137	130
145	138	138	131

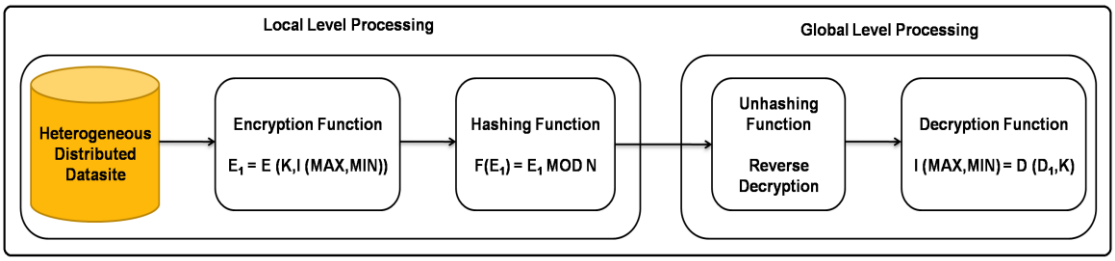

Figure 4. Double-Blind Key-attribute based Encryption algorithm

Double-Blind Key-attribute based Encryption algorithm

Input: **(**MAX,MIN) pair of each entity

Intermediate results: First-encrypted (MAX,MIN) pair

Second-encrypted (MAX,MIN) pair

First-decrypted (MAX,MIN) pair

Output: Second-encrypted (MAX,MIN) pair

Second-decrypted (MAX,MIN) pair

// first blind encryption

Encrypted data, $E_1 = E(K,I)$

$E_1 = E(K,(MAX,MIN))$

// second blind encryption

Hash function, $F(E_1) = E_1 \text{ MOD } N$

$F(E_1) = (E(K,(MAX,MIN)) \text{ MOD } N)$

// first blind decryption

Computing unhashing function (reverse decryption)

// second blind decryption

Input data, $I (MAX,MIN) = D (D_1,K)$

EMPIRICAL ANALYSIS

Proposed personalized privacy preserving algorithm is implemented on EHRs dataset openly available real world dataset from university research websites. The experiments

were implemented with five distributed local data sites. The proposed algorithm is implemented in HDFS distributed environment coded in C# language. Table 8. depicts dataset description of EHRs used for evaluating the proposed DBKE algorithm.

Table 8. EHRs Dataset

Distributed sites	No. of attributes	No. of records initially
Site 1	07 (variable)	2500 (variable)
Site 2	09 (variable)	8000 (variable)
Site 3	07 (variable)	3000 (variable)
Site 4	08 (variable)	7500 (variable)
Site 5	10 (variable)	5500 (variable)

Performance Evaluation Metrics

Performance of proposed algorithm is analyzed on two measures: accuracy and confidence level. Accuracy represents a measure of effective retrieval of EHRs based on user query under privacy-preservation algorithm. The mathematical formulation is expressed as,

$$\text{Accuracy}\ (\%) = \frac{\text{TP} + \text{TN}}{\text{Total Records}}$$

Where TP – True positive, TN – True negative

Confidence level measure how well the original data can be retrieved from the function-encrypted data. If original value is estimated to lie in interval [x1,x2], then the interval (x2 – x1) is the range of privacy at c% confidence[23].

Performance Analysis

Performance analysis of proposed DBKE algorithm is analyzed with accuracy and confidence level. The proposed privacy preserving algorithm is evaluated with state-of-art randomization and anonymization approaches: additive noise addition[23], multiplicative noise addition[24],L-diversity[25] and T-Closeness[26].

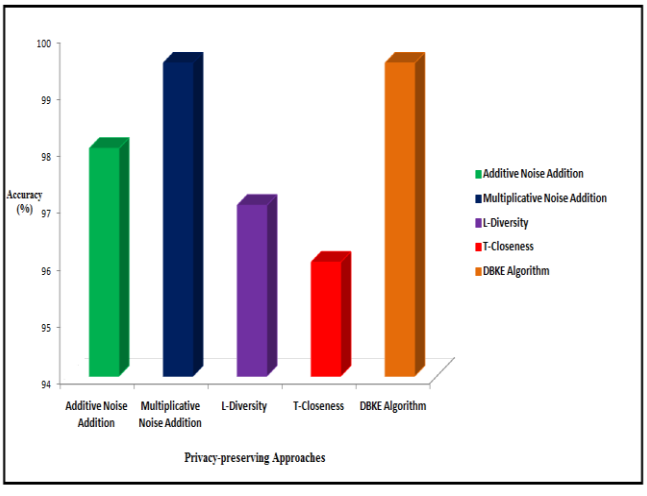

Figure 5. Accuracy Vs privacy-preserving approaches

Fig. 5 depicts accuracy comparison of proposed DBKE algorithm with the state-of-art privacy preserving approaches. Proposed DBKE algorithm exhibits more accuracy (99.38%) compared to other randomization and anonymization approaches because of combining both randomization and anonymization approach. Randomization approaches, additive noise addition (98.23%) and multiplicative noise addition (99.32%) exhibits more accuracy value but privacy/confidence level is less because original data can be retrieved on analyzing the spectral values of randomized data. Anonymization approaches, L-Diversity (97.14%) and T-Closeness (96.73%) exhibits less accuracy compared to randomization approaches but confidence/privacy level is more than randomization approaches. In conventional approaches there is some data loss incurred while converting the original data to protected data.

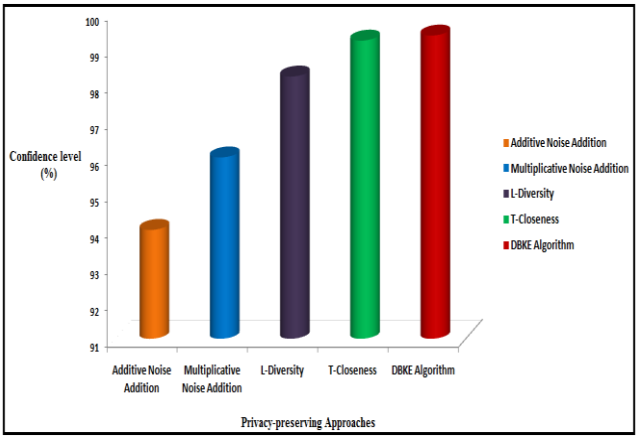

Figure 6. Confidence level Vs privacy-preserving approaches

Fig. 6 depicts confidence level comparison of proposed DBKE algorithm with the state-of-art privacy preserving approaches. Proposed algorithm exhibits more confidence level (99.40%) compared to randomization and anonymization approaches because the distribution of original data is considered in proposed DBKE algorithm. Randomization approaches, additive noise addition (94.37%) and multiplicative noise addition (96.21%) exhibits less confidence level compared to anonymization approaches and proposed DBKE algorithm. Similarly, anonymization approaches, L-Diversity (98.37%) and T-Closeness (99.42%) exhibits higher or equal confidence level compared to proposed DBKE algorithm but result accuracy is less. In conventional approaches data distribution at each local datasite and global level is not considered while converting the original data to protected data.

CONCLUSION

In this paper, a privacy preserving personalization method based on encryption and hashing function is proposed. Experimental implementation on EHRs datasets show that the DBKE algorithm is effective for privacy preserving DDM. Performance is evaluated with confidence level and accuracy

and compared with state-of-art privacy preserving approaches: anonymization and randomization.

REFERENCES

[1] Vinaya Sawant, Ketan Shah, A review of Distributed DM using agents, International Journal of Advanced Technology & Engineering Research (IJATER). 3(5) (2013) 27-33.

[2] S. V. S. Ganga Devi, A Survey on Distributed DM and its Trends, International Journal of Research in Engineering & Technology (IJRET). 2(3) (2014) 107-120.

[3] Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.V.: Efficient Privacy Preserving K-Means Clustering. PAISI 2010. LNCS, vol. 6122, pp. 154–166. Springer, Heidelberg (2010).

[4] Agrawal, R., and Srikant, R. (2000). Privacy Preserving Data Mining. ACM SIGMOD International Conference on Management of Data, SIGMOD‟00, Dallas, USA. 439-450.

[5] Pinkas, B.: Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor. Newslett. 4(2), 12–19 (2002).

[6] Bertino, E., Fovino, I., Provenza, L.: A Framework for Evaluating Privacy Preserving DataMining Algorithms. Data Mining and Knowledge Discovery 11(2), 121–154 (2005).

[7] Kamalika Das, Kanishka Bhaduri, Hillol Kargupta, "A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to peer networks", *Journal of Knowledge and Information Systems*, Volume 24, Issue 3, September 2010, pp. 341-367.

[8] Sung Baik, Jerzy Bala, "A Decision Tree Algorithm for Distributed Data Mining: Towards Network Intrusion Detection", *Computational Science and Its Applications – ICCSA 2004*, Volume 3046 of the series Lecture Notes in Computer Science, pp 206-212.

[9] Josenildo Costa da Silva, Matthias Klusch, " Inferences in Distributed Data Mining", *Engineering Applications of Artificial Intelligence*, Volume 19, 2006, pp. 363–369.

[10] Trilok Nath Pandey, Niranjan Panda, Pravat Kumar Sahu, "Improving performance of distributed data mining (DDM) with multi-agent system", *(IJCSI) International Journal of Computer Science Issues*, Vol. 9, Issue 2, No 3, March 2012, pp. 74-82.

[11] Kawuu W. Lin, Sheng-Hao Chung, " A fast and resource efficient mining algorithm for discovering Frequent patterns in distributed computing environments", *Journal of Future Generation Computer Systems*, Volume 52, November 2015, pp. 49-58.

[12] A.O. Ogunde, O. Folorunso, A.S. Sodiya, "A partition enhanced mining algorithm for distributed association rule mining systems" *Egyptian Informatics Journal*, Volume 16, Issue 3, November 2015, pp. 297-307.

[13] YanLi, ChangxinBai, ChandanK.Reddy, A distributed ensemble approach for mining health care data under privacy constraints, Journal of Information Sciences. 330 (2018) 245-259.

[14] Masooda Modak, Rizwana Shaikh, Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy, 7th International Conference on Communication, Computing and Virtualization. 29 (2018) 993 – 1000.

[15] Yiannis Kokkinos, Konstantinos G.Margaritis, Confidence ratio affinity propagate on in ensemble selection of neural network classifiers for distributed privacy-preserving data mining, Neuro-computing. 150 (2018) 513–528.

[16] Hemanta Kumar Bhuyan, Narendra Kumar Kamila, Privacy preserving sub-feature selection in distributed data mining, Journal of Applied Soft Computing. 36 (2017) 552-569.

[17] Feng Zhang, Chunming Rong, Gansen Zhao, Jinxia Wu, Xiangning Wu, Privacy-Preserving Two-Party Distributed Association Rules Mining on Horizontally Partitioned Data, International Conference on Cloud Computing and Big Data. (2016) 633-640.

[18] Xinjun Qi, Mingkui Zong, An Overview of Privacy Preserving Data Mining, Procedia Environmental Sciences. 12 (2016) 1341 – 1347.

[19] Yanguang Shen, Hui Shao, Yan Li, Research on the Personalized Privacy Preserving Distributed Data Mining, Second International Conference on Future Information Technology and Management Engineering. (2016) 436-439.

[20] Rebecca N. Wright, Zhiqiang Yang, Sheng Zhong, Distributed Data Mining Protocols for Privacy: A Review of Some Recent Results, MADNES 2005. (2015) 67–79.

[21] Xun Yi, Yanchun Zhang, Privacy-preserving distributed association rule mining via semi-trusted mixer, Data & Knowledge Engineering. 63 (2015) 550–567.

[22] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. IEEE Int'l Conf. Data Mining, Nov. 2003.

[23] Wang, J., Zhong, W. J., & Zhang, J. (2006). NNMF-based factorization techniques for high-accuracy privacy protection on nonnegative-valued datasets.

Proceedings of the IEEE Conference on Data Mining 2006, International Workshop on Privacy Aspects of Date Mining (PADM 2006), pp. 513-517, Hong Kong, China.

[24] Stanley R. M. Oliveira, and Osmar R. Zaïane1, "Towards Standardization in Privacy-Preserving Data Mining", In ACM SIGKDD 3rd Workshop on Data Mining Standards, 2004, pp. 7–17.

[25] Inan, A., Kaya, S.V., Saygin, Y., Savas, E., Hintoglu, A.A., Levi, A.: Privacy preserving clustering on horizontally partitioned data. Data Knowledge Eng., 646–666 (2017).

[26] Zhuojia Xu, Xun Yi. Classification of Privacy-preserving Distributed Data Mining Protocols, 978-1-4577-1539-6/11 IEE – 2011.