An Improved Data Utility Model For Privacy Preserving Data Mining Using Decision Tree Algorithm

Md Ilyas, Dhanraj Verma, Manoj kumar Deshpande

Abstract: The paper demonstrates the PPDM model improvements. The proposed work focused on recovering privacy-preserving "IF-THEN-ELSE" rule mining. Thus a previously proposed technique is improved with some modifications. First, the over-fitting cases of data are explored and inappropriate attributes are reduced. Further in place of encryption-based data sanitization process is replaced with the noise-based data model. Therefore a noise mixing algorithm is proposed in this paper. In order to reduce the data dimension, the client-side process involves the utility measurement technique using the correlation coefficient based technique. Additionally for recovering the contributed attributes are recovered at the client thus a rule and decision recovery algorithm is also involved in the proposed modified PPDR model. Using these modifications the data utility, and computational resource consumption is improved. Both the aspects are also justified using the implementation and experimental study with previously introduced PPDM model and proposed a modified model. According to the obtained results, we found the proposed enhanced model is efficient and accurate as compared to the previously proposed data model. Finally, some essential future extension of the work is also reported in this paper.

Index Terms: privacy preserving data mining, decision rule mining, noise based model, data utility improvement, resource conservation, performance evaluation.

1 INTRODUCTION

Data mining is a technique for analyzing data, in various applications where classification, prediction, categorization, and recognition required data mining techniques are used [1]. The industries related to banking, finance, engineering, medical, research, security, and much more domain usage these methodologies [2]. The different kinds of algorithms can be used for patterns recovering and solving real-world problems [3]. The privacy-preserving data mining [4] is a technique that mine data with focusing on security, privacy and the sensitivity of data [5]. Multiple data suppliers are involved in this environment, they don't want to disclose their data. But they are agreed to contribute data for research and decision-making, due to the privacy of data owners [6]. The next component of PPDM is the authority under the supervision data is mined. The authority usages data mining technique, and security algorithm for mining confidential data securely and privately without damaging data utility (application) [7] thus before data discloser to any party it is sanitized [8]. This work is aimed to find an efficient privacy-preserving decision rule (PPDR) mining technique [9]. Therefore, in this condition, the following key features in the existing system are involved.

1. A cryptographic algorithm at the end of data supplier
2. Managing data at the server end for secure mining and data discloser
3. Only authentic parties can recover the disclosed decisions.

In order to fulfill the objectives an article [10] were published. In this study association rule mining algorithm (ARM) and decision tree (DT) involved, and the following conclusion were made:

1. ARM technique are costly with respect to decision tree computationally
2. Quantity of rules for ARM is higher with respect to the DT
3. The quantity of Item-set and transactions decides the number of rules.

And according to the conclusion made. The following future work is proposed.

1. The aggregated data, from different parties, increases dimensions of data, thus a dimensionality reduction techniques is required
2. The ARM techniques generate a significant amount of rules with respect to DT. Thus, the DT is used for further experiments

Based on the future extension the following objectives are prepared.

1. Implement a privacy preserving dimensionality reduction technique
2. Enhancing the previously proposed [10] PPDR technique by implementing the dimensionality reduction technique with existing system

This section includes the preface of the work, the next section offers recent contributions in PPDM as a literature survey. After that, a new privacy-preserving dimensionality reduction technique is proposed for mining decision rules. In the next performance is measured and conclusions are discussed.

2 LITERATURE SURVEY

This section provides the study about the privacy preserving model and their dimensionality reduction problem. Thus PPDM techniques and their contributions are discussed.

2.1 Privacy Preserving Data Models

R. Mendes et al [11] surveys the PPDM techniques and evaluate them with applications of methods. Additionally, current challenges and issues are discussed. Y. A. A. S. Aldeen et al [12] reviews a list of published literature using categories and subcategories. Additionally, PPDM methods, advantages, and limitations are discussed. The categories are distortion, association rule, hide association rule, taxonomy, clustering, associative classification, k-anonymity and distributed, outsourced data mining, with pros and corns. K. Xu et al [13] propose a framework for PPDM where dataset is distributed and huge. The data locality property of Apache Hadoop is used to reduce cryptographic operations. I. San et al [14] investigate low performance of Paillier cryptosystem.

First, utilize parallelism between operations and interleaving among independent operations. Then, develop hardware using field-programmable gate arrays. Z. Gheid et al [15] propose a privacy-preserving k-means algorithm based on the multiparty additive scheme for horizontally partitioned data. R. Lu et al [16] present a Lightweight Privacy-preserving Data Aggregation (LPDA). It is defined by Chinese Remainder Theorem, Paillier encryption, and one-way hash chain to aggregate IoT devices' into one and filter injected false data. It is secure and enhanced with differential privacy. Y. Kokkinos et al [17] distributed PPDM in decentralized data locations using several neural networks and select best of them using confidence ratio affinity propagation and privacy-preserving (PP) computing. Classifiers validate each other. The training set of one becomes the validation set for next. S. Sharma et al [18] present challenges for designing real world privacy-preserving system for healthcare. Thus a personalized healthcare system is developed for disease supervision. The necessities for data suppliers, to recover resources, analyze existing techniques, and discuss tradeoff among efficiency, privacy, and quality. The EHRs can resolve many problems with disease diagnosis, with patient's data privacy and sensitivity. The sharing of records between different facilities is a major concern. Y. Li et al [19] develop two distributed privacy-preserving algorithms using ensemble learning. The idea is to build a model for each facility to accurately learn data distribution and can transfer useful information using decision models without sharing sensitive data. H. Hammami et al [20] suggest an approach to combine the extraction of frequent closed patterns with the privacy of sites using cloud-based homomorphic encryption. The mechanism requires less communication and computation overheads. K. Birman et al [21] propose smart metering that will allow utilities to use data effectively consumers' privacy and work is focused on mining association rules. N. Domadiya et al [22] is proposed PPDARM for combine mining of association rules. This also analyzed the heart disease dataset. Privacy-Preserving Utility Mining (PPUM) discussed by W. Gan et al [23]. First, present utility mining, and then introduce related preliminaries and problems, with evaluation criteria. Additionally, advantages and deficiencies are highlighted, finally challenges and future directions discussed. J. C. W. Lin et al [24] focus on issues of HUIM and PPUM, and present two algorithms to mine and hide sensitive high-utility itemsets. G. Kalyani et al [25] has been planned and idea to protect association rules. To select transactions to update it using binary TLBO optimization is used. The disclosure is lack of protection. B. Abidi et al [26] introduce a new micro-aggregation HM-PFSOM, based on fuzzy possibilistic clustering. The anonymization process is used to decrease information loss and risk of confidentiality. C. Y. Lin et al [27] focuses on data stream and sliding window design with reversible privacy-preserving for real-time data, termed as continuous reversible privacy-preserving (CRP). Data accurately recovered additionally, using a watermark, data integrity verified.

2.2 Dimensionality Reduction

Nan Zhang et al [28] data mining is successful in many applications, but the concerns are handling private data. Architecture for a systemic view of issues, are established for collection, inference control, and data sharing. Bhupendra Kumar Pandya et al [29] study a data perturbation technique for PPDM using randomized multiplication. Theoretical results were provided for projections. Additionally explores the possibility of multiplicative random projection. A hiding-missing-artificial utility (HMAU) algorithm is proposed by Chun-Wei Lin et al [30] to hide perceptive item-sets. The transaction with the maximal sensitive to non-sensitive is selected for removal. Three cases of hiding failures are considered. Sheryl Parmar et al [31] characterize the consequence of PPDM and its techniques to keep the data safe and prevent it from abuses. Samir Patel et al [32] propose a simple PCA based transformation for datasets to preserve privacy and maintain accuracy on clustering. Thanveer Jahan et al [33] discuss different data perturbation methods such as matrix decomposition methods, Fuzzy logic and Hybrid methods. And propose a Multiplicative data perturbation method and experiments have proved its efficiency with k-means clustering. Hessam Zakerzadeh et al [34] demonstrate few common properties of real data to upgrade the negative effects of dimensionality. In real data sets, dimensions contain high levels of inter-attribute relationships. They introduced an approach to k-anonymity, ℓ-diversity, and t-closeness. In the presence of inter-attribute relations, that approach offers more robust with large dimensional data, without compromising accuracy. The purpose of Reinhard Heckel et al [35] is to measure the impact of dimensionality-reduction in random projection on performance of the sparse subspace clustering (SSC) and the thresholding based subspace clustering (TSC). Both algorithms reduce down the subspace dimensions without performance humiliation. Aristos Aristodimou et al [36], proposed a new anonymisation algorithm for PPDP, based on k-anonymity using pattern-based multidimensional suppression (kPBMS). The feature selection technique is used for reducing dimensions. Yining Wang et al [37] propose a framework to analyze a optimization-based algorithm for SSC, when data dimension is compressed. The SSC succeeds if random projection is a subspace embedding.

2.3 C4.5 Decision Tree

The decision rules (DR) can also be mined using C4.5 or J48 algorithms. The J48 is an extension of ID3. To prepare tree Entropy and Information Gain (IG) required. The higher IG decides the position of node in tree. Using multiple iterations the branches of tree developed. First the entropy is measured for computing IG. The entropy is basically measured for the entire data. For a dataset which contains two classes, P and N. Thus, for binary scenario entropy E:

$$E(D) = -P\ log_2 P - N\ log_2 N$$

Where P is the ratio of Positive samples and N is negative

The minimum entropy attributes are used first. Thus IG is drop in entropy. The IG, Gain (E, A) for attribute A is:

$$Gain(E, A) = Entropy(s) - \sum_{n=1}^{v} \frac{E_v}{E} XEntropy(E_v)$$

The gain is deciding positions of node. The J48 algorithm returns a DT as learning [38]. The following steps are used:

Input: Dataset (D).

Output: A decision tree T.

A. Create an initial node
B. If object in the given class.
    a. Create a leaf node and assign label C;
C. If attribute list is null,
    a. Create a leaf node and most frequent class is assigned;
    b. Select attribute with highest IG, and mark it test-attribute;

D. If X is test-attribute;
   a. Add new branch in tree relevant to test-attribute X;
E. $If\ (B_i == Null)$
   a. Create leaf node, of frequent class;
F. Else
   a. Create leaf node
G. Returned.

3 PROPOSED WORK

This section provides the details about the proposed privacy preserving dimensionality reduction technique. In this context the previously proposed PPDM method is optimized for recovering the effective attributes for improving time and memory resource consumption.

3.1 Background

The previously proposed methodology of PPDM technique [10] offers the three layers for privacy-preserving decision rule mining as given in figure 3.1.

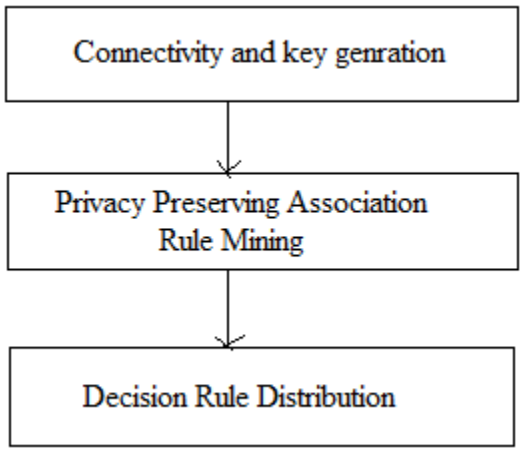

Figure 3.1 *proposed system*

The process is initiated with data supplier for combining and mining rules. Thus, when a connection requested, server accepts it and generate a key for secure communicates. The encryption results ciphered data to communicate. The received key is used with MD5 algorithm for hash key generation. The hash key and dataset produced to AES for cipher-text generation. At server the rules prepared using Apriori and C4.5 DT. At server, data from all clients are used to prepare a common dataset. Further the dataset is encoded. The encoded data were used for IF-THEN-ELSE rule preparation. The rules encrypted again for prevention of distribution data leakage risk. Thus, rules were recovered at client end using same key.

3.2 Proposed system

In this work the future work of the article [10] is carried out. In this context the decision tree algorithm is used for computing the decision rules. Second aim is to work on data dimension which are increases during the aggregation of data. Therefore, a new algorithm is introduced to handle data dimensions and their utility. Over-fitting based attribute removalIn this context first we involve the technique to counter the cases of over fitting of the data. Due to over fitting cases of data the decision tree is not learning with the entire data. Additionally we cannot recover the required fruitful outcomes. Therefore first the algorithm 3.1 which is described in table 3.1 is developed. That algorithm explores the dataset attributes and concludes which attributes can be used for learning.

Input: Dataset D
Output: Reduced data R

Process:

A. $[row, col] = readDataset(D)$
B. $for(i = 1; i \leq col; i++)$
   a. $for(j = 1; j \leq row; j++)$
      i. $if(j = 1)$
         1. $R[i,j].insert(Val(i,j))$
      ii. $else\ if\big(Val(i,j) == R(i,j-1)\big)$
         1. $count++$
         2. $R[i,j].insert(Val(i,j))$
      iii. *end if*
   b. *End for*
   c. $U = R.getUniqueCount()$
   d. $If(count == row)$
      i. $R.Eliminate(i)$
   e. $if(U == row)$
      i. $R.Eliminate(i)$
   f. *End if*
C. *End for*
D. *Return R*

Table 3.1 *Reducing dimensions for over fitting*

Using the above given algorithm the over fitted data attributes. Adding Noise on dataAfter that in place of using the encryption algorithm, we create a range based noisy attribute creation algorithm. According to the available nature of data there are two kinds of data attributes present in dataset i.e. quantitative and qualitative. In order to process the quantitative attributes we usages the concept of data normalization. The min-max normalization can be defined using the following equation.

$$newVal = \frac{current\ value - min}{max - min}$$

In addition of that we need a noise range varies between 00-10. That random noise is introduced by the party who are contributing for data. That input is given by NR. Thus new value can be given by:

$$newVal = newVal * NR$$

On the other hand for dealing with the qualitative attributes we mutate and place new values for the existing values. Both the processes of the handling data is incorporated to the algorithm steps given in table 3.2.

Input: input dataset from previous phase R, Noise Range NR, circular queue $Q = \{a, b, c, \ldots, z\}$
Output: Noise added dataset ND

Process:

1. $[row, col] = R.GetDimension$
2. $for(i = 1; i \leq col; i++)$
   a. $for(j = 1; j \leq col; j++)$
      1. $if(R[j,i] == Neumeric)$
         1. $newVal = \frac{R[j,i] - R.min}{R.max - R.min}$
         2. $R[j,i] = newVal * NR$
      2. *Else*
         1. $Temp = R[j,i]$
         2. $C[x] = Temp.GetChar()$
         3. $for(k = 1; k \leq C.length; k++)$
            a. $C[k] = Q[k + NR]$
         4. *end for*
         5. $R[j,i] = C$
      3. *End if*
   b. $ND.Add(R[j,i])$
   c. *End for*

3. End for
4. Return ND

Table 3.2 *Noise Adding algorithm*

After adding noise over the dataset the data is transformed into the noisy attributes but the noise is distributed uniformly therefore the utility of dataset is not being changed.

Measuring Data Utility

In order to measure data utility in PPDM surrounding the correlation coefficient is suggested to be use. The correlation coefficient is denoted by r that told us relation between two vectors to define closeness of data. The closest value of r is 1. If r =1 or r = -1 then relationship perfectly defined. Data sets with values of r = zero show no relationship among them. Using this property of correlation coefficient we use this for first electing the optional columns. Therefore first at client end correlation coefficient is calculated to find effective attributes. To calculate the r we use the following eq.

$$r_{x,y} = \frac{\sum_{i=1}^{n} d_x * d_y}{n\sqrt{\sum_{i=1}^{n} {d_x}^2 * \sum_{i=1}^{n} {d_y}^2}}$$

Where,

Deviation of x variable $d_x = x_i - \bar{x}$

Deviation of y variable $d_x = x_i - \bar{x}$

N = number of instances

Now the following algorithm is used for selection of optional attributes.

Input: noise added data ND
Output: Optional Data attributes O

Process:

A. $[row, col] = ND.GetDimension$
B. $C = ND[col]$
C. $for(i = 1; i \leq col; i++)$
   a. $V = R[i]$
   b. $CR_{V,C} = Calculate(r_{V,C})$
   c. $O.Add(CR_{V,C})$
D. End for
E. $M = CalculateMean(O)$
F. $for(j = 1; j \leq col; j++)$
   a. $if(O_i \leq \frac{M}{2})$
      1. $O_i.Mark(Optional)$
   b. End if
G. End for
H. Return O

Table 3.3 *Selecting optional attributes*

After that the encryption algorithm is avoided to secure the dataset. The transformed noisy data is transmitted to server. Here the computation of optional attributes demonstrate less likely hood among attributes and class labels. But it may possible the optional attributes will work better then locally involved data and also can enhance the classification accuracy. Therefore entire data is combined and used with correlation coefficient for finding relevancy among class labels and attributes. We found no change in their r value. Thus, if the attribute not has the strong relationship then client can reduce the attributes to reduce the computational overhead by reducing them. Server accepts the data from all the concerned clients and organized using the class labels. After organization of data the decision tree over the data is applied. The decision tree initially produces the tree structure which is further converts into the IF-THEN-ELSE rules. These rules are distributed to all the clients.

Data recovery

To recover the data actual values at the end of concerned client the similar technique is used as described in table 2. Therefore to implement the required process we need to reverse the values which are generated. Therefore in order to recover the categorical values the sequences of characters are replaced with the existing values of character. Additionally to recover the numerical values, the following two steps are used.

$$newVal = \frac{Ot}{NR}$$

Where Ot is the value which is received by the server and NR is the noise value included before, and the $newVal$ is min-max normalized data, we can recover the actual value using this $newVal$. Initially we have,

$$newVal = \frac{current\ value - min}{max - min}$$

Thus,

$$newVal(max - min) = current\ value - min$$

And finally,

$$current\ value = newVal(max - min) + min$$

Thus if we have

$$\delta = (max - min)$$
$$current\ value = newVal * \delta + min$$

Using the above given formulation the following process is taken place as given in table 3.4.

Input: Obtained rules R, Noise Range NR, circular queue $Q = \{a, b, c, \dots, z\}$, $\delta = (max - min)$
Output: Noise free data D

Process:

1. $[row, col] = R.GetDimension$
2. $for(i = 1; i \leq col; i++)$
   a. $for(j = 1; j \leq col; j++)$
      1. $if(R[j,i] == Neumeric)$
         1. $newVal = \frac{R[j,i]}{NR}$
         2. $R[j,i] = newVal * \delta + min$
      2. Else
         1. $Temp = R[j,i]$
         2. $C[x] = Temp.GetChar()$
         3. $for(k = 1; k \leq C.length; k++)$
            a. $C[k] = Q[k - NR]$
         4. $end\ for$
         5. $R[j,i] = C$
      3. End if
   b. $D.Add(R[j,i])$
   c. End for
3. End for
4. Return D

Table 3.4 *data recovery algorithm*

In this section the modification on existing model is provided which is experimented in next section.

4 RESULTS ANALYSIS

This section includes the comparison among proposed modified privacy persevering rule mining techniques and previously proposed technique. The similar parameters are used for comparison. We start with the objectives of experiments, then the evaluation of parameters described.

4.1 Aim of experiments

The proposed work is motivated to enhanced previously proposed PPDR model. That model includes encryption technique for sanitization of confidential data. That is replaced here with the noise based technique for hiding the sensitive and private information. In addition of that to reduce the time and memory usages of the system the utility of data is measured and unreliable attributes were reduced. That may help to improve the computational cost of rules. Therefore with the following aim the experiments are conduced.

1. To compare the current modified technique with previously proposed model for finding the performance improvements due to involved modifications
2. To validate the utility of data after sanitization process adopted

This section provides the aim of experiments next section offers the evaluation of required parameters similar as the previous article [10].

4.2 Parameters

The overview of performance analysis of the proposed PPDR technique is provided here.

4.2.1 Accuracy

Accuracy of a data mining algorithm is correctly reorganization rate of a trained algorithm. Thus, total correctly identified and total samples are used for measuring it. The following Eq. can be use.

$$accuracy(\%) = \frac{total\ correct\ decisions * 100}{total\ samples\ for\ decision\ making}$$

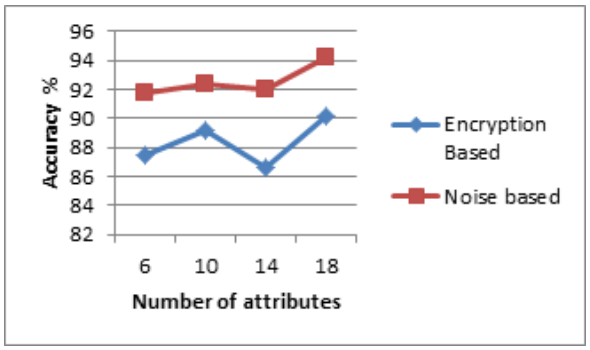

Figure 4.1 *Accuracy*

Figure 4.1 shows accuracy of developed techniques. The Y-axis contains accuracy in percentage and, X-axis describes attribute size as input. Accuracy for both schemes noise based and cryptographic algorithms is similar. However, the mean accuracy of modified noise based PPDR algorithm is higher with respect to previously technique [10].

4.2.2 Space complexity

The space complexity or memory usages of an algorithm are basically computed on the basis of process, when a process execution is initiated the system assigns a fixed size of memory. The free size of memory over total is reported as memory usage. In JAVA it is computed as.

$$memory\ usages = total\ assigned - free\ memory$$

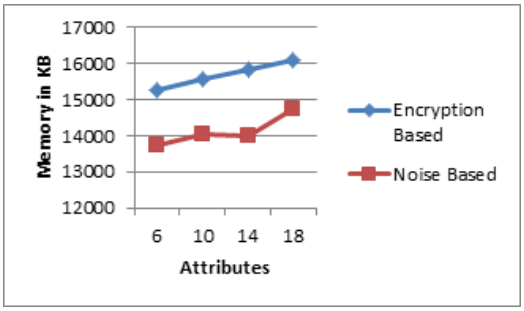

Figure 4.2 *memory usages*

According to observation of memory usages for both techniques (i.e. noise based and cryptography based) the figure 4.2 prepared. That is a line graph where in Y-axis the expended memory is reported and in X-axis the number of attributes is listed. The results describe memory usages for a PPDR technique increases with the attribute size. Moreover, the encryption based technique is expensive as with respect to noise based technique.

4.2.3 Time complexity

The time complexity or usages is the quantity of time used for data processing. The following Eq. can be used:

$$time\ consumed = Algorithm\ end\ time - start\ time$$

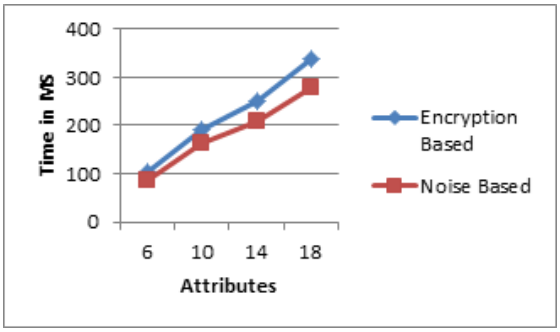

Figure 4.3 *time expenses*

The time complexity of encryption based and noise based technique is reported in figure 4.3. The time consumption is noticed here in milliseconds (MS). The Y-axis of line graph shows the time consumed and the number of attributes used is shown in X-axis. The demonstrated result says the time consumption of noise based technique is low as compared to encryption based technique. The line graph clearly demonstrates the gap between lines is increases with the size of attribute set. Therefore the proposed technique contributes significantly for reducing the cost of algorithm execution.

4.2.4 Number of Rules

The performance of both the PPDR model i.e. noise based technique and encryption based technique is described using number of rule generation. Figure 4.4 describes the volume of rules generated. The generated amount of rules is demonstrated in Y-axis and X-axis reports number of attributes in experiment.

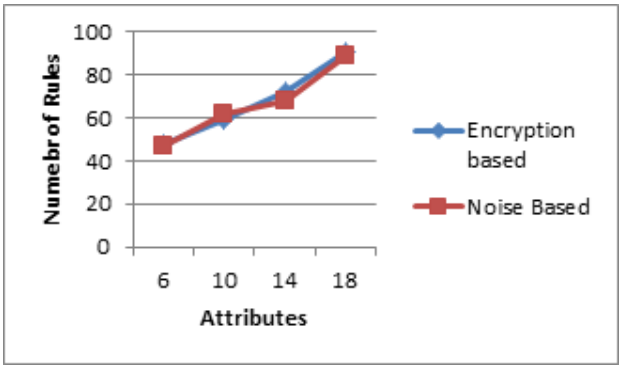

Figure 4.4 *Number of rules*

The results clearly show the noise based technique generates the similar rules as the encryption based technique. But the noise based technique respectively produces less rules as compared to cryptographic technique.

5 CONCLUSION

The findings of the efforts placed in order to improve the existing PPDR model is reported in this section. Additionally future extension of the work is also reported.

5.1 Conclusion

The number of applications exists where additional or delegated attributes are required for making precise decisions. In this context agreed parties are outsource their data but they are worried on discloser of data and decisions. Thus applications needed to sanitize and prevent the privacy of the end data owners. Therefore, the paper is indented to improve the previous data model for reducing the data dimensions and improving the data utility and performance in terms of resources consumption. Therefore first the decision tree algorithm is targeted to study and implementation. Additionally in order to handle the over fitting cases the algorithm is proposed at the client. Using this over-fitting based conditions user can reduce the attributes their own end. Further the noise mixture algorithm is proposed and implemented with the system. That helps to sanitize the sensitive information. After noise mixing the data can be used for mining the privacy-preserving IF-THEN-ELSE rules. But the data utility is main concern for the transformed data. Therefore the proposed enhancement also includes the optional attribute computation. These optional attributes having less bounding with the target class labels. Therefore the data can be reduced at the client end to reduce the communication overhead as well as computation overhead. After that process data attributes are transmitted to the server where the data is processed using decision tree algorithm. During this rules are recovered and distributed to all the parties. The rules are recovered at the client's machine using a recovery algorithm which is also reported. The proposed models (i.e. encryption based and noise based model) are developed using the JAVA. Moreover to maintain and preserve the performance the MySQL Database is used. The system is evaluated using different parameters as described in table 5.1.

S. No.	*Parameters*	*Encryption based*	*Noise based*
1	*Accuracy*	*Low*	*High*
2	*No. of rules*	*Higher*	*Low*
3	*Memory usages*	*Higher*	*Low*
4	*Time expenses*	*Higher*	*Low*

. Table 5.1 *performance summary*

According to the obtained performance the proposed work enhances the existing system. The improvement is noticed for data utility, time consumption and the memory requirements. Therefore the proposed extension is helpful for performance improvement of rule based model development with preserving privacy.

5.2 Future Work

The enhanced model for rule mining is efficient and accurate for measuring the content fitness and utility for effective rule mining. In near future the following work is proposed for enhancement.

1. Most of models are developed for either the rule based mining (association rules and decision rules). Additionally the supervised learning approaches are used. In near future the work is focused on unsupervised learning process
2. The current work is focused on the intentionally data discloser cases in near future the model is developed which works for human error based data discloser

6 REFERENCES

[1] Toch, B. Lerner, E. B. Zion, I. B. Gal, "Analyzing large-scale human mobility data: a survey of machine learning methods and applications", Knowl Inf Syst, https://doi.org/10.1007/s10115-018-1186-x

[2] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal 15 (2017) 104–116

[3] Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng, "A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing (2016) 2016:67, DOI 10.1186/s13634-016-0355-x

[4] C. W. Lin, T. P. Hong, H. C. Hsu, "Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", Hindawi Publishing Corporation Scientific World Journal Volume 2014, Article ID 235837, 12 pages

[5] P. S. Rao, S. Satyanarayana, "Privacy preserving data publishing based on sensitivity in context of Big Data using Hive", J Big Data (2018) 5:20, https://doi.org/10.1186/s40537-018-0130-y

[6] O. Tene, J. Polonetsky, "Big Data for All: Privacy and User Control in the Age of Analytics", 11 Nw. J. Tech. & Intell. Prop., 239 (2013).

[7] Urquhart, N. Sailaja, D. McAuley, "Realising the right to data portability for the domestic Internet of things", Pers Ubiquit Comput (2018) 22:317–332, DOI 10.1007/s00779-017-1069-2

[8] Ray, T. C. Ong, I. Ray, M. G. Kahn, "Applying Attribute Based Access Control for Privacy Preserving Health Data Disclosure", 978-1-5090-2455-1/16/$31.00 ©2016 IEEE

[9] T. Pawar, Prof. S. Kamalapur, "A Survey on Privacy Preserving Decision Tree Classifier", International Journal of Engineering Research and Applications, Vol. 2, Issue 6, November- December 2012, pp.843-847

[10] First paper reference
[11] R. Mendes, J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications", Vol. 5, 2017, IEEE
[12] Y. A. A. S. Aldeen, M. Salleh, M. A. Razzaque, "A comprehensive review on privacy preserving data mining", SpringerPlus (2015) 4:694, DOI 10.1186/s40064-015-1481-x
[13] Xu, H. Yue, L. Guo, Y. Guo, Y. Fang, "Privacy-preserving Machine Learning Algorithms for Big Data Systems", 2015 IEEE 35th International Conference on Distributed Computing Systems, 1063-6927/15© 2015 European Union DOI 10.1109/ICDCS.2015.40
[14] San, N. At, I. Yakut, H. Polat, "Efficient paillier cryptoprocessor for privacy-preserving data mining", Security and Communication Networks 2016; 9:1535–1546, Wiley Online Library, DOI: 10.1002/sec.1442
[15] Z. Gheid, Y. Challal, "Efficient and Privacy-Preserving k-Means Clustering for Big Data Mining", IEEE TristCom, Aug 2016, Tianjin, China pp.791 - 798, ff10.1109/TrustCom.2016.0140ff ffhal-01466904f
[16] R. Lu, K. Heung, A. H. Lashkari, A. A. Ghorbani, "A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT", Special Section on Security and Privacy in Applications and Services for Future Internet of Things, Vol. 5, 2017, IEEE
[17] Y. Kokkinos, K. G. Margaritis, "Confidence ratio Affinity Propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining", Neurocomputing, (2015), vol. 150, pp. 513−528
[18] S. Sharma, K. Chen, A. Sheth, "Towards Practical Privacy-Preserving Analytics for IoT and Cloud Based Healthcare Systems", IEEE Internet Computing, March-April 2018
[19] Y. Li, C. Bai, C. K. Reddy, "A Distributed Ensemble Approach for Mining Healthcare Data under Privacy Constraints", Inf Sci (Ny). 2016 February 10; 330: 245–259. doi:10.1016/j.ins.2015.10.011
[20] H. Hammami, H. Brahmi, I. Brahmi, S. B. Yahia, "Using Homomorphic Encryption to Compute Privacy Preserving Data Mining in a Cloud Computing Environment", EMCIS 2017, LNBIP 299, pp. 397–413, DOI: 10.1007/978-3-319-65930-5_32, Springer International
[21] Birman, M. Jelasity, R. Kleinberg, E. Tremel, "Building a Secure and Privacy-Preserving Smart Grid", ACM SIGOPS OSR 49(1) pp131–136, http://dx.doi.org/10.1145/ 2723872.2723891
[22] Domadiya, U. P. Rao, "Privacy-preserving association rule mining for horizontally partitioned healthcare data: a case study on the heart diseases", Sådhanå (2018) 43:127 Indian Academy of Sciences, https://doi.org/10.1007/s12046-018-0916-9
[23] W. Gan, J. C. W. Lin, H. C. Chao, S. L. Wang, P. S. Yu, "Privacy Preserving Utility Mining: A Survey", arXiv:1811.07389v1 [cs.DB] 18 Nov 2018
[24] C. W. Lin, W. Gan, P. F. Viger, L. Yang, Q. Liu, J. Frnda, L. Sevcik, M. Voznak, "High utility-itemset mining and privacy-preserving utility mining", Perspectives in Science (2016) 7, 74—80
[25] Kalyani, M. V. P. Chandra Sekhara Rao and B. Janakiramaiah, "Privacy-Preserving Association Rule Mining Using Binary TLBO for Data Sharing in Retail Business Collaboration", Advances in Intelligent Systems and Computing 515, © Springer Nature Singapore Pte Ltd. 2017
[26] B. Abidi, S. B. Yahia, C. Perera, "Hybrid Microaggregation for Privacy-Preserving Data Mining", arXiv:1812.01790v1 [cs.CR] 4 Dec 2018
[27] C. Y. Lin, Y. H. Kao, W. B. Lee and R. C. Chen, "An efficient reversible privacy‑preserving data mining technology over data streams", SpringerPlus (2016) 5:1407, DOI 10.1186/s40064-016-3095-3
[28] Zhang, W. Zhao, "Privacy-Preserving Data Mining Systems", Published by the IEEE Computer Society 0018-9162/07/$25.00 © 2007 IEEE
[29] B. K. Pandya, U. K. Singh, K. Dixit, "A Study of Projection based Multiplicative Data Perturbation for Privacy Preserving Data Mining", International Journal of Application or Innovation in Engineering & Management, Vol. 3, Issue 11, Nov. 2014
[30] C. W. Lin, T. P. Hong, H. C. Hsu, "Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", Hindawi Publishing Corporation Scientific World Journal Volume 2014, Article ID 235837, 12 pages, http://dx.doi.org/10.1155/2014/235837
[31] S. Parmar, Mrs. P. Gupta, Ms. P. Sharma, "A Comparative Study and Literature Survey on Privacy Preserving Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 4, April 2015, pg.480 – 486
[32] S. Patel, K. R. Amin, "Privacy Preserving Based on PCA Transformation Using Data Perturbation Technique", International Journal of Computer Science & Engineering Technology, Vol. 4 No. 05 May 2013
[33] T. Jahan, G. Narasimha, V. G. Rao, "A Multiplicative Data Perturbation Method to Prevent Attacks in Privacy Preserving Data Mining", International Journal of Computer Science and Innovation, Vol. 2016, no. 1, pp. 45-51, ISSN: 2458-6528
[34] Zakerzadeh, C. C. Aggrawal, K. Barker, "Towards Breaking the Curse of Dimensionality for High-Dimensional Privacy: An Extended Version", arXiv:1401.1174v1 [cs.DB] 6 Jan 2014
[35] R. Heckel, M. Tschannen, H. B¨olcskei, "Subspace clustering of dimensionality-reduced data", arXiv:1404.6818v1 [cs.IT] 27 Apr 2014
A. Aristodimou, A. Antoniades, C. S. Pattichis, "Privacy preserving data publishing of categorical data through k-anonymity and feature selection", Healthcare Technology Letters, 2016, Vol. 3, Iss. 1, pp. 16–21, doi: 10.1049/htl.2015.0050
[36] Y. Wang, Y. X. Wang, A. Singh, "A Deterministic Analysis of Noisy Sparse Subspace Clustering for Dimensionality-reduced Data", Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).
[37] B. Hssina, A. Merbouha, H. Ezzikouri, M. Erritali, "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications, Issue No10, page 13