Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7778426

MEASURING INFLUENCE OF NOISE IN SUPERVISED LEARNING PERFORMANCE UNDER PRIVACY PRESERVING ENVIRONMENT

Mayur Rathi[1], Anand Rajavat[2]
Department of Computer Science & Engineering
Shri Vaishnav Vidyapeeth Vishwavidyala
Indore, India
[1]mayurrathi.31dec@gmail.com, [2]directorsviit@svvv.edu.in

Abstract—Data mining combined with security and privacy is known as Privacy-Preserving Data Mining (PPDM). In this setting, multiple data owners are aggregating their data with unknown parties for utilizing the combined knowledge based on data intelligence techniques. The PPDM outcomes are sensitive against different factors like dimensions and sanitization techniques. In this paper, we aimed to investigate the performance influence of PPDM classifiers due to data dimensions and sanitization techniques. In this context, first, a public dataset KDD CUP is obtained. Additionally, the dimensionality reduction techniques such as PCA (Principle component analysis), KPCA (kernel principle analysis), and CRC (correlation coefficient) are applied for the observing the impact of dimensions in PPDM systems.
Additionally, noise based data sanitization technique is investigated for investigating the impact of noise on PPDM systems. Further random noise, is used to sanitize the data. But, the categorical data can not be utilized with random noise. Therefore, we extended random noise algorithm as controlled noise algorithm. The controlled random noise algorithm is producing a new sanitized dataset without disturbing the data utility. The newly generated datasets are trained with two supervised learning algorithms, i.e. C4.5 and CART.
The experiments on five public UCI datasets are performed. The results prove that the accuracy of classifier is highly influencing with classical random noise. Beside that, the proposed controlled noise-based technique has low impact on classifier's accuracy, because less statistical difference between original and controlled noise-based sanitized data. In addition, the controlled noise may expensive in terms of time and memory utilization due to moditification of data.

Keywords—PPDM, Effect of Data Dimension, Effect of Data Sanitization, Random Noise, Noise Inclusion Algorithm.

Introduction

With the aim of analyzing data for recovering usful patterns to understand and distinguish data is known as data mining. This technology is used to classify a large amount of data or predicting the relevant values. Data mining generally used on centralized databases [1]. Some of the time, for making next level of decisions, the DM requires deligated information, but without security and privacy it is not feasible. In such scenarios we need a technique known as Privacy-Preserving Data Mining (PPDM) [2]. Essentially, PPDM is used where multi-parties are sharing information. However, nobody consented to disclose information to another [3]. However, PPDM stages are influencing with different issues and challenges [4]. Among them two factors significantly influence the PPDM methodologies.

Data dimension [5]
Data sanitization [6]
This paper study the PPDM and discusses the effect of data sanitization and dimension.Therefore, the following work is highlighted in this paper:

Perform an experimental comparative study among three popular dimensionality reduction techniques principal component analysis (PCA), kernel principal component analysis (KPCA), and correlation coefficient (CRC). The aim of this study is simulate how the dimensions of data can influence the PPDM methods performance in terms of effiency and classification accuracy.

Performing the experimental study for classifying the data based on original dataset, traditional random noise based, and modified random noise based data sanitization techniques. The aim is to study the influence of data sanitization process in PPDM modeling. Additionally, introduce a new noise inclusion technique, which make balance between the security and data utility.

The paper will be useful for identifying the best approaches to select during the PPDM system design. The paper expected to deliver a highlight to deal with the issues of PPDM in terms of performance i.e. efficiency and learning correctness. This section is a general idea of the work involved in this paper. Next, section includes the investigations conducted in this paper. The section II discuss the performance impact based on dimensionality of the data and section III provides the details about how the data sanitization technique will effect on the PPDM modeling. Additionally a modified random noise inclusion technique is also described. Finally the results are measured and the conclusion of the study has been discussed.

Effect of dimensions

In Data mining the dimensionality reduction techniques are used to select most appropriate features for classification task [12]. Figure 1 shows a model to demonstrate the effect of dimensionality reduction on data mining performance. In this model we need a dataset for simulation, here we considered two popular and publically available high-dimension experimental datasets i.e. VAN [14] and KDD CUP 99's [13] dataset. Additionally three dimensionality reduction alsogorithms i.e. PCA, CRC and K-PCA is used to measure influence. The dataset is divided into training sample (70%) and test samples(30%) [15]. Further for performing training and testing we have considered the K-nearest neighbor (k-NN) classifier. The Euclidean distance is used to differentiate between two classes [16]. That can be given as:

$$D(M,N) = \sqrt{\sum_{i=1}^{n}(M_i - N_i)^2} \qquad (8)$$

Where, distance D(M,N) between vectors M and N, and n is sample size.

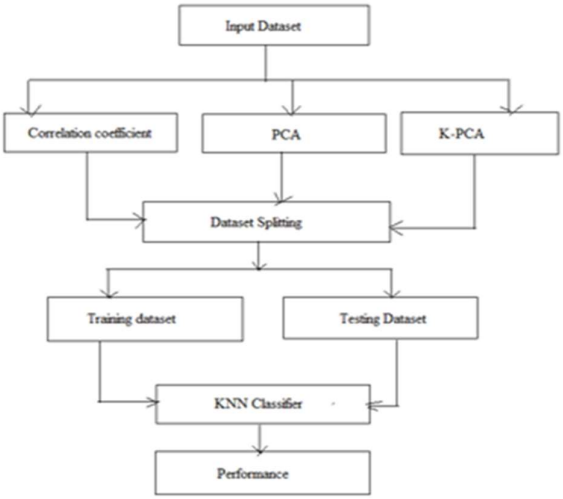

Figure 1 Model for Measering Effect of Dimensions

The classification of test samples are performed and for performing training and testing of KNN table 1 contains the steps:

Table 1 Process to Perform Training and Test

Input: Dataset D, Algorithm List $L_2 = \{L_0, L_1, L_2\}$, Number of features U **Output:** C predicted classes
Steps: 1. $D_n = readDataset(D)$ 2. $SA = L_2.SelectAlgorithm(U)$ 3. $F_m = SA.ReduceDimension(D_n)$ 4. $[Train, Test] = F_m.Split(70,30)$ 5. $for(j = 1; j \leq Test.Lenght; j++)$ a. $C = KNN.Classify(Test_j)$ 6. $end\ for$ 7. Return C

The influence on efficiency is measured based on memory uses and training time. The time taken to train the model using the given examples [17] is known as training time. The training time after transforming data using different dimensionality reductioon technques is shown figure 2(a) in millisecond (MS). According to the outcomes, PCA is costly compared to KPCA and CRC. Moreover, KPCA and CRRC is less expensive in execution, yet the CRC shows superior performance.

The amount of assigned main memory utilized for executing the process is known as memory usages or space complexity. It is given in figure 2(b) and measured in KB (kilobytes). The results shows PCA and CRC are expensive as compare to KPCA. Next matrix is accuracy, which is here measured to validate the data utility. The accuracy of the classifiers after applying dimensionality reduction algorithms is given in figure 2(c). According to results, CRC is high accurate as compared to KPCA and PCA. The utility is also measured in error rate, which is given in figure 2(d). According to this comparision PCA has low accuracy.

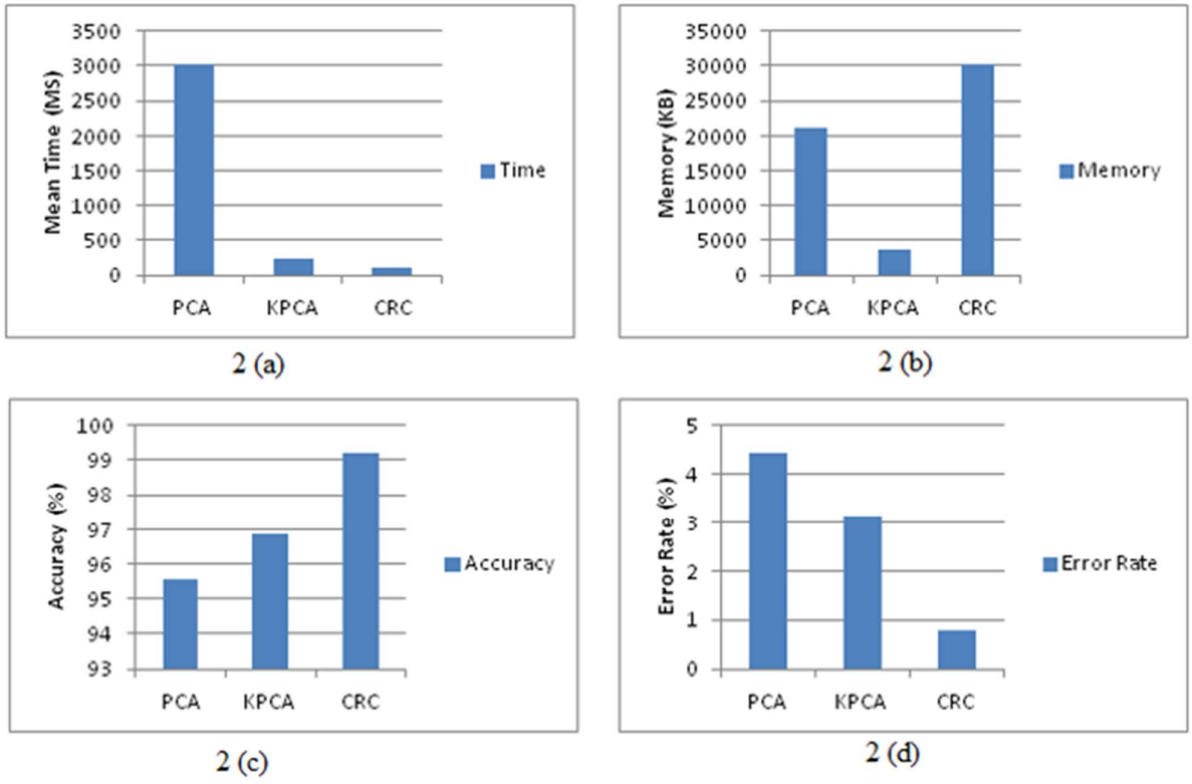

Figure 2 Effect of Dimensions on Classifier Performance in Terms of (a) Training Time (b) Memory Usage (c) Accuracy and (d) Error Rate

This part shows the performance influence of dimensionality reduction in classification performance. Thus, three popular algorithms are compared. Based on examinations, the CRC based method is proficient and exact. Next the problem of higher data utility is the main point of interest [18]. The required investigation for influence of data utility is described in next section.

EFFECT OF Noise

The preprocessing is one of the essential step of the ML based technologies. Using preprocessing we reduce the unwated data and noise. The aim is to improve the data quality and make clean by which the learning algorithms can effectively train on the data [19]. But in PPDM applications we utilize the noise to sanitize sensitive and private information. In this context, the unregulated noise level can harm the data utility [20]. Therefore, it is required to investigate the effects of data sanitization on the data utility. This paper aimed to explore the data sanitization approaches and their influence on classifier's performance under PPDM settings.

Data sanitization teechniques

Some essential and frequently adopted data sanitization techniques are discussed in table 3 [22]. Table also discuss ttthe advantages and limitations of the sanitization approaches.

Table 3 Techniques of data sanitization

Technique	Advantages	Disadvantages
Perturbation	Efficient, Simple, Accurate, preserves statistical feature, Treated Attributes Individually	Data loss in multiple dimensions, Modified techniques Required
Condensation or Aggregation	Save statistical features, support modified data, improve security	Data loss due to larger number of grouped records, data mining results affected
Suppression	Private data is completely hidden, secure against intruders	Data loss, Results of data mining is affected
Anonymization	Save individuals identity	High data loss, Less secure against linking attacks
Cryptography	Multiple parties support, offers tools for cryptographic algorithms	Large number of parties are not supported, less secure output, Less security
Swapping	Difficult to recover data, originality improved	Time taking, low strength against diversity attack
Randomization	Efficient, Simple, Not needed prior data	Treats all data equally, Data is not recoverable

The table, first, include Perturbation-based PPDM, which is appropriate for both cenralizeed and distributed databases. In this method, new dataset is prepared by original data modification using two methodologies disfigurement and likelihood. Next is Condensation or Aggregation, which is appropriate for a centralized databases. The data is grouped based on application requirements. Then grouped data is used to produce sanitized data by using the patterns of groups. Suppression-based sanitization utilized for statistical calculation, which is appropriate for centralized databases. Here, sensitive attributes are removed before the information discloser. Next is Anonymization based sanitization, which is used with centralized databases. This technique is useful to preserve the individual's data, it utilizes generalization and suppression. Cryptography based sanitization is a popular method and frequently used with centralized databases. It is used when different parties are working together to figure out conclusions based on their own part of data, without considering third party data. The swapping is a popular technique of data sanitization and appropriate for distributed as well as centralized databases. This method keeps the original values in the database. Additionally, few new values are used to replace old values. Randomization is also a frequently used sanitization method for distributed and centralized databases. The information is randomized to include noise during sanitization.

Applying data sanitization

Among different data sanitization technique the randomization is a frequently used and powerful approach. Nevertheless, the randomization is only successful for the neumerical data. Hence, an improved version of random noise is set up to sanitize data. Moreover, a data-mining model is used to discover the effect on learning performance. The required model for this task is shown in figure 3.

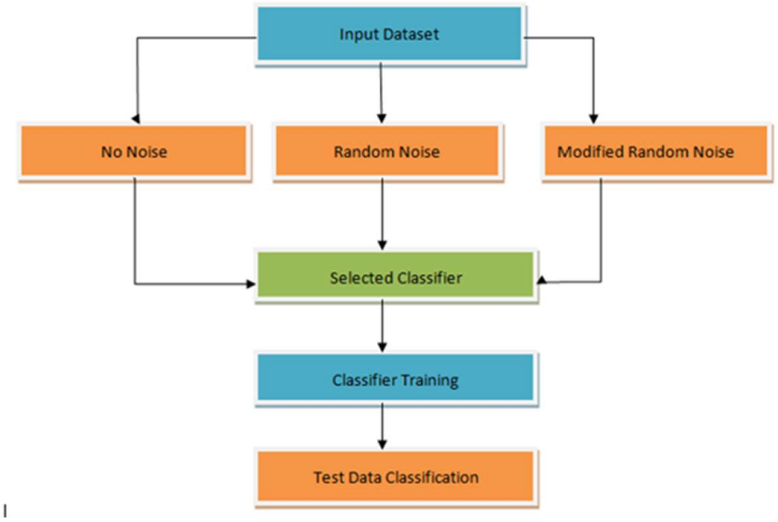

Figure 3 Noisy data Classifier

The dataset is obtained from UCI archive [23]. The used datasets are Iris (Numeric), Forest (Numeric), Auto-mpg11 (Numeric), Heart (Numeric) and Forest Fires (Mix). The mixed dataset contains Numeric and categorical attributes. Next, we have to prepare sanitized dataset by including noise on original dataset. Three type of datasets are used as given in figure 3. In this diagram no noise indicate the original dataset, random noise indicate traditional randomization based sanitization. In this method random values generated between a range of values, which is used to replace original values. But it cannot modify categorical values. Therefore, a modified version of random noise is introduced to modify the categorical data also. The noise inclusion process of the proposed randomization algorithm is shown in table 4 and table 5. The estimation of required noise level for the dataset is given in table 4, and table 5 provides the noise inclusion.

Table 4 Noise Factor Computation

Input: Dataset D

Output: noisy to be add ξ

Process:

1. $[row, col] = readDataset(D)$
2. $Var = 0$
3. $for(i = 1; i \leq col; i++)$
   a. $if\ (D_i.isNeumeric == true)$
      i. $\mu = \frac{1}{row}\sum_{j=1}^{row} D(j,i)$
      ii. $\sigma^2 = \frac{1}{row}\sum_{j=1}^{row}(D(j,i) - \mu)^2$
   b. $else$
      i. $\mu = \frac{1}{row}\sum_{j=1}^{row} D(j,i).length$
      ii. $\sigma^2 = \frac{1}{row}\sum_{j=1}^{row}(D(j,i).Length - \mu)^2$
   c. End if
   d. $var = var + \sigma^2$
4. $End\ for$
5. $\xi = \frac{var}{col}$
6. $return\ \xi$

Table 5 shows the process to compute the noise level which is suitable to use. First, the number of instances (rows) and number of attributes (columns) are calculated. Then, for all the

attributes, we check whether a column is numeric or categorical. The mean value μ is calculated for numerical attributes. Using μ, we calculate the variance σ^2 of the attribute. Next, for the case of categorical attributes, the value's length is used for mean μ and variance σ^2 computing. After that, the error factor is computed:

$$\xi = \frac{\sigma}{A_i}$$

Where σ is the combination of all attribute's varience, number of attributes A_i, and ξ noise to add.

The original dataset values are altered based on ξ. The steps to include noise is given in table 5. All the dataset values are treated individually, and a new dataset is prepared by manipulation of the original dataset. This method make use of noise factor ξ and dataset D.

Table 5 Noise Add-on Algorithm

Input: noise component ξ, Dataset D

Output: Sanitized dataset N

Steps:

1. $[row, col] = readDataset(D)$
2. $for(i = 1; i \leq col; i++)$
   a. $if\ (D_i.isNeumeric == true)$
      i. $min = getMin(D_i)$
      ii. $max = getMax(D_i)$
      iii. $for(j = 1; j \leq row; j++)$
         1. $Norm = \frac{D(j,i)-min}{max-min}$
         2. $NewD(j,i) = Norm * \xi$
      iv. $end\ for$
   b. Else
      i. $for(j = 1; j \leq row; j++)$
         1. $NewD(j.i) = Randomize(D(j,i), \xi)$
      ii. $end\ for$
   c. End if
   d. $N.Add(NewD_i)$
3. End for
4. Return N

Then min-max normalization is used [24] as:

$$A_N = \frac{A - A_{min}}{A_{max} - A_{min}}$$

The normalized data is used with the noise factor ξ to compute the multiplicative noisy value. The original dataset value are replaced with noisy values. For manipulating the categorical values table 6 provide an algorithm.

Table 6 Algorithm for Manipulating Categorical Attribute

Input: String S, Noise factor ξ, Character Array $C = \{a, \dots z, \& 0, \dots, 9\}$

Output: Randomize String R

Process:

1. $SC_n = String2CharArray(S)$
2. $for(i = 0; i \leq n; i++)$
   a. $NewIndex = i + \xi$

b. $|Diff| = (NewIndex) mod (36)$
c. $R.Add(SC_i.replace(C_{Diff}))$
3. End for
4. Return R

The generated noisy dataset is further used with two rule-based classification methods, namely CART [25] and C4.5 [26]. These models are used to identify the performance influence or data utility. Next section discusses the experimental performance of the proposed random noise inclusion technique.

Results analysis

The main aim is to analyze the effect of dataset sanitization method in the classifier's performance. In this context, first the accuracy of the classifier's has been measured to identify the utility of sanitized data. Figure 4 contains accuracy and figure 5 shows the error rate. Correctness of classifier is measureable in error rate and accuracy. The random noise, controlled noise and original datasets are used for experimentatioons. The classifiers CART and C4.5 is used for data classification. The bar graph as a performance indicator is given for the experimental results. Five datasets are used among four datasets has numerical values, and one of them has mix values. The mixed dataset is not suitable to use with traditional random noise. Finally, based on the results in terms of accuracy and error rate, we found the controlled noise-based data has preserve the data utility more effectively as compared to traditional randomization technique.

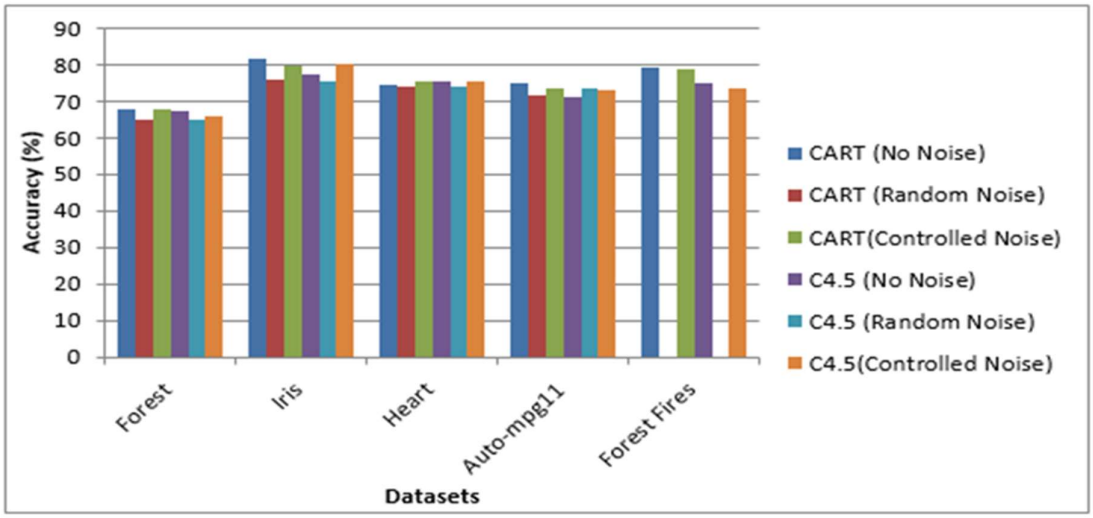

Figure 4 Accuracy

In this experiment we found the traditional random noise based sanitization can degrad the accuracy and error rate, and sometimes it can improve performance. Therefore the classical random nosie can majorly influence the classifier's performance.

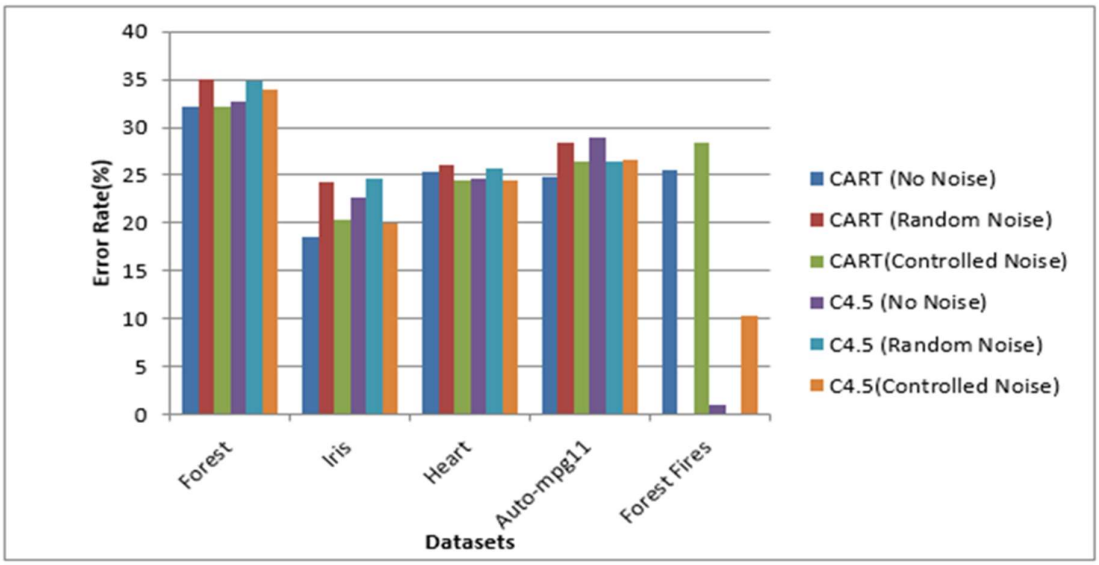

Figure 5 Error Rate

Next, for measuring the effieciency training time and space is measured. The space complexity are given in Figures 6 and 7. The time is deliberate in milliseconds(MS) and memory in MB, The results show that time and space is not fluctuating highly by data sanitization. But with mixed dataset, the effiency is degraded.

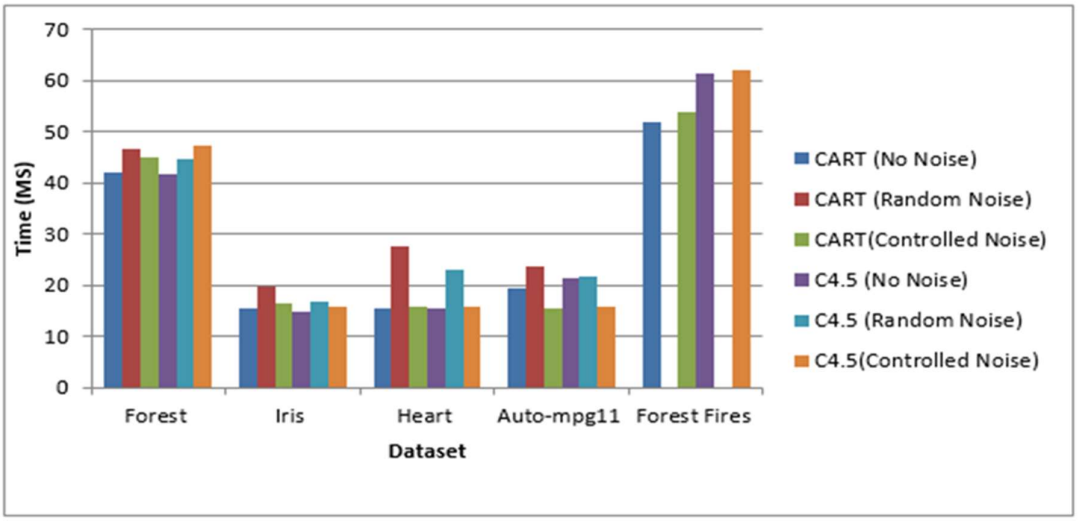

Figure 6 Time consumption

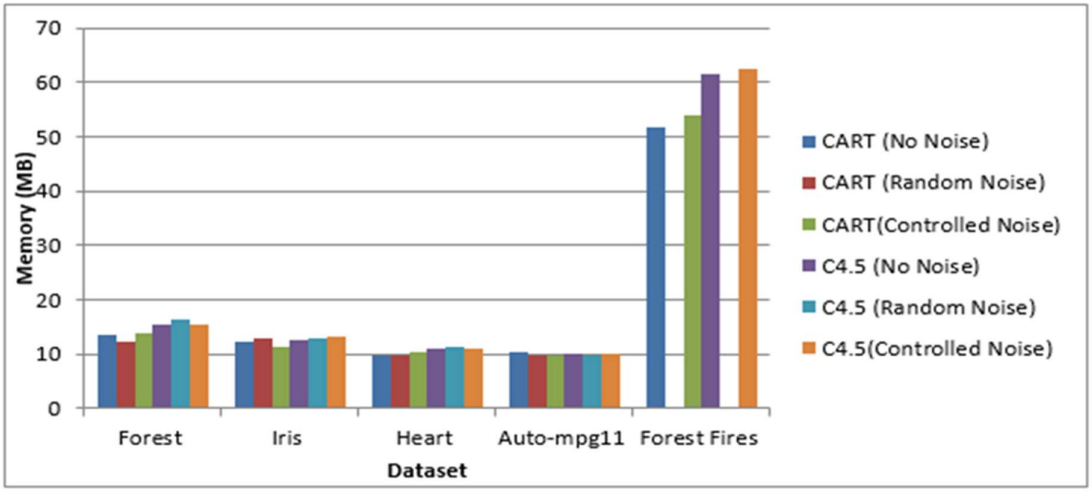

Figure 7 Memory Usage

conclusion

The PPDM enviournment is sensitive against sanitization technique and dataset dimensions. Therefore, the knowledge of impact on PPDM method is needed to know. In order to identify the influence of these factors on PPDM models we have discussed two experimental scenarios.

All parties are merge their data vertically, thus increasing number of parties are also increases data dimensions. By results, it is found the dat dimensions has no affect on learning, but can influence on computational cost.

The experiments with noisy and non noisy datasets has been carried out using two classifiers. Then the difference among them is measured. Results show the noise not impacts the computational performance, but can degrade the data utility.

Based on the experiments, the controlled random noise model include limited noise thus, has less influence on classification performance. Shortly the following extension has been planned.

Combine the experience gained for preparing future PPDM model

Aim to design a secure and less resource expensive PPDM model

REFERENCES

A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, C. A. Coello, "A survey of Multi-objective Evolutionary Algorithms for Data Mining: Part I", IEEE Trans. on Evolutionary Computation. 18(1), pp. 20-35, 2014.

Y.A.A.S. Aldeen, M. Salleh, M.A. Razzaque, "A comprehensive review on Privacy Preserving Data Mining" Springer Plus 4:694, pp. 1-36, 2015.

J. Danasana, R. Kumar, D. Dey, "Mining Association Rules for Horizontally Partitioned Databases using CK Secure Sum Technique", Int. J. of Dist. and Parallel Sys. 3(6), pp. 149-157,2012.

S. Hariraman, Dr. S. Velmurugan, "An Enhanced Privacy Preserving Techniques for Asynchronous Streaming Data Mining in Distributed Environment", International Journal of Engineering Research & Technology,Special Issue - 2020 Volume 8, Issue 07

Y. Dong, B. Du, L. Zhang, and L. Zhang, "Dimensionality Reduction and Classification of Hyperspectral Images Using Ensemble Discriminative Local Metric Learning", IEEE Transactions on Geoscience and Remote Sensing, 0196-2892 © 2017 IEEE

J. C. W. Lin, J. M. T. Wu, P. F. Viger, Y. Djenouri, C. H. Chen, Y. Zhang, "A Sanitization Approach to Secure Shared Data in an IoT Environment", Volume 7, 2019 2169-3536 2019 IEEE

F. Artoni, A. Delorme, S. Makeig, "Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition", Neuroimage. 2018 July 15; 175: 176–187. doi:10.1016/j.neuroimage.2018.03.016

H. Binol, "Ensemble Learning Based Multiple Kernel Principal Component Analysis for Dimensionality Reduction and Classification of Hyperspectral Imagery", Hindawi Mathematical Problems in Engineering Volume 2018, Article ID 9632569, 14 pages

T. J. Abrahamsen, L. K. Hansen, "A Cure for Variance Inflation in High Dimensional Kernel Principal Component Analysis", Journal of Machine Learning Research 12 (2011) 2027-2044 Submitted 1/11; Published 6/11

A. Ly, M. Marsman, E. J. Wagenmakers, "Analytic posteriors for Pearson's correlation coefficient", Statistica Neerlandica (2018) Vol. 72, nr. 1, pp. 4–13

S. Kumar, I. Chong, "Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States", Int. J. Environ. Res. Public Health 2018, 15, 2907

Z. M. Hira, D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", Advances in Bioinformatics, Volume 2015 |Article ID 198363 | https://doi.org/10.1155/2015/198363

K. K. Vasan, B. Surendiran, "Dimensionality reduction using Principal Component Analysis for network intrusion detection", Perspectives in Science (2016) 8, 510—512

https://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29

M. I. Sameen, B. Pradhan, S. Lee, "Self-Learning Random Forests Model for Mapping Groundwater Yield in Data-Scarce Areas", Natural Resources Research (2018) https://doi.org/10.1007/s11053-018-9416-1

N. Ali, D. Neagu, P. Trundle, "Evaluation of k nearest neighbour classifer performance for heterogeneous data sets", SN Applied Sciences (2019) 1:1559 | https://doi.org/10.1007/s42452-019-1356-9

Y. Cao, P. Li, Y. Zhang, "Parallel processing algorithm for railway signal fault diagnosis data based on cloud computing", Future Generation Computer Systems 88 (2018) 279–283

W. Fang, X. Z. Wen, Y. Zheng, M. Zhou, "A Survey of Big Data Security and Privacy Preserving", IETE Technical Review, 2016

S. R. Gallego, B. Krawczyk, S. García, M. Wozniak, F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions", Neurocomputing 239 (2017) 39–57

J. S. Pyo, L. J. Seong, L. Juyeon, "Method of improving the performance of public-private innovation networks by linking heterogeneous DBs: Prediction using ensemble and PPDM models", Technological Forecasting & Social Change 161 (2020) 120258

H. Vaghashia, A. Ganatra, "A Survey: Privacy Preservation Techniques in Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015

https://archive.ics.uci.edu/ml/index.php

L. Munkhdalai, T. Munkhdalai, K. H. Park, H. Gyu Lee, M. Li, K. H. Ryu, "Mixture of Activation Functions With Extended Min-Max Normalization for Forex Market Prediction", VOLUME 7, 2019, IEEE

M. Hasanipanah, R. S. Faradonbeh, H. B. Amnieh, D. J. Armaghani, M. Monjezi, "Forecasting blast induced ground vibration developing a CART model", Engineering with Computers DOI 10.1007/s00366-016-0475-9, © Springer-Verlag London 2016

A. Lohani, J. Singh, A. Lohani, "Comparative Analysis Of Classification Methods Using Privacy Preserving Data Mining", International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 02, Issue 04; April – 2016