International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Correlation Based Transformation For Privacy Preserving Medical Data Publication

K.Saranya[1], K. Premalatha[2]

[1,2]*Department of Computer Science and Engineering Bannari Amman Institute of Technology, Erode, Tamil Nadu*

Abstract

Privacy protection is a high level of encouragement to provide confidential information by posting the details of the customers. Data mining tasks are incorporated in the database for data intelligence and the knowledge recovery. Initially different isolated transformation techniques are proposed for masking the sensitive information, which combines correlation analysis and data transformation provides intended level of privacy preservation. In this research work correlation based transformation techniques are considered to preserve the sensitive information and provides the complete privacy of the original information. The classifiers Decision Tree (DT), Random Forest(RF), Linear model, Ada Boost, Support Vector Machine and Neural Network are used to identify the performance of the proposed works.

Keywords: Correlated matrix, Data sanitization, Data transformation and Anonymization.

1. INTRODUCTION

The `modern society performs most of their purchase activity through the web. Such purchases made by different people are logged in the database, from that data set you can learn and analyze various information which supports the product manufactures in targeting the product pioneer[1]. For example, the product organization or manufacturer generates intelligence in enhancing their business to the next level[2]. From the market database, the product manufacturers involve in acquiring the information like where the product is moving in higher level and as well as in the average level. Based on these levels, they can also identify the reason for the low mobility of products and they can find the competing product is. Similarly, the financial organization acquires the information like, the purchase habit of people living in specific region and their buying capacity[3][4].

On the other side, the medical data set is shared among the medical organization, in order to analyze the disease rate that exists in various regions and the medicine which is highly preferred on a specific disease[5]. Also, from the medical data set, It can be identified that the people who have purchased different medical products. Such information is used to mobilize the medical products among the users[6]. However, the medical data set is available with the information like patient details, personal information and the causes of disease, the prescribed medicine, the present status of the patient and so on. Such information contains huge personal and professional information which are more sensitive and cannot be exposed to the third party and other organization or external world.

The customer is fully relying on the organization and the organization has its own responsibility in maintaining the secrecy of the patient[7]. If the data set is shared between different organizations, their personal information may be exposed to the external world which affects the customer morale over the organization[8]. In order to safeguard such personal information, the information hiding or sanitization is performed. The sanitization is the process of hiding sensitive information from original data set. It is performed in several ways, in simple manner; the data hiding is performed by undertaking the dot matrix operation which places the star in the place of sensitive items[9]. But in this case, the originality of the data may get affected and the end user or third party cannot identify any useful information from the published one. Similarly, there are many techniques like probabilistic approach which places all the values with a set of probability values[10]. This in turn the user can identify only the probability of the products and leads less possibility in attaining any useful information. Similarly there are number of techniques available to perform sanitization but suffers with poor performance in sanitization and originality.

2. LITERATURE REVIEW

There are number of methods are declared for the problem of privacy preservation of transactional data sets using data mining techniques[11][13]. We perform a detailed review on the earlier methods.

Ali et.al, (2009), presented a new method for uncertain data using building classifiers for modeling anonymized data. During data release, it replaces all the statistics which are collected and implicit probability distribution towards the anonymized data.

Zhengli Huang et.al, (2010), proposed data reconstruction methods based on data correlations. Firstly Principal Component Analysis (PCA) techniques are used for modifying sensitive information and the other method the Bayes Estimate (BE) technique is used for estimating the private information.

Maurizio Atzori et.al, (2011) , addresses an intention of privacy compliance of patterns which lacks the reference to a predetermined knowledge of sensitive and non-sensitive information, on the basis of rather intuitive and practical restriction that the anonymity of individuals should be guaranteed. In particular, the problem addressed here arises from the possibility of assuming from the output of frequent itemset, the existence of patterns with very low support than an anonymity threshold.

Thomas G. Dietterich, (2011), outlines the statistical report for determining whether one of the machine learning algorithms that performs another task on particular on learning assessments[12]. These tests are compared practically and it verify the incorrectness of the output if exists (type I error).

V. Thavavel, (2012), proposed a method that manages the unstructured data into a structured data using legacy system and it performs distributed data for multi-text document using partitioned method.

Dharmendra Thakur and Prof. Hitesh Gupta, (2013), presented a classification of Association Rule mining and sensitive rule hiding algorithm, which is one of the hottest research areas in PPDM. The process of modifying the original dataset refers to the certain sensitive association rule sometimes it depart without affecting the data and the non-sensitive association rules.

C.V.Nithya and A.Jeyasree, (2013), developed a new pre-processing discrimination prevention methodology, which has different methods for data transformation. It prevents the data which includes direct discrimination, indirect discrimination or both of them at the same time. To accomplish this objective, firstly it measures the discrimination based on that it identify categories of the dataset and it groups the individuals by the decision making process. Secondly, the data transformation will be performed in appropriate way to eliminate all the discrimination biases. Finally, discrimination free data models can be formed from the transformed data set without seriously damaging data quality.
X. Zhang, (2017), proposed a novel method for preventing upper bound privacy leakage, constraint-based approach is identified. From this method, it classifies which intermediate datasets are needed to be encrypted and which do not. So, the privacy preservation cost also saves the privacy requirements of the data holder.

Jinyan Wang, (2018), presented a sliding window concept to ρ-uncertainty for structuring the defected sensitive rules. By using generalization and suppression, it outperforms the data transaction based on addition and deletion of transaction to satisfy ρ-uncertainty.

ChaobinLiu, (2019) , addressed a new privacy preserving method for data publication using conditional probability distribution and machine learning techniques. To attain different prior idea for different transactions, cross sampling algorithm and a complete cross sampling algorithm are modelled respectively for the settings of single and multiple sensitive attributes. In order to improve data utility, complete algorithm is developed by using Gibbs sampling id used, when data are not sufficient.

3. PROBLEM DEFINITION

The task of sanitizing original data set is to protect the originality of the data and to reform the original data from the sanitized one. A new sanitization process is introduced, which uses probability values of sensitive items to generate the publishing data. The user can be able to interpret the required information from the published data. At the same time it is impossible to find out the person whom does the information is belongs to.

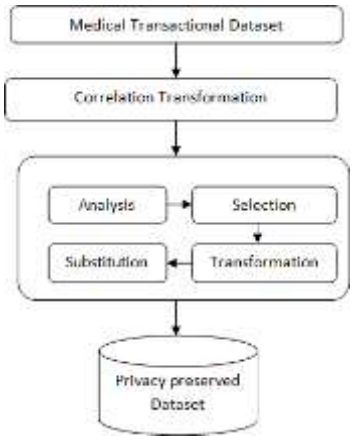

Figure 1 Architecture of proposed Correlation based Transformation

The Figure 1 shows the architecture of proposed Correlated based transformation. Each functional component has been discussed in detail in this section.
A new sanitization process is introduced with the help of correlation based transformation technique. Given Medical data containing both sensitive and non-sensitive information, Correlation based transformation determines the relations between the sensitive and non sensitive items. Correlation matrix is calculated using Pearson's correlation coefficient. The transformation techniques are tending to connect the observation similarity between the attributes and generating new sanitization approach. If the non sensitive items are highly correlated with sensitive item, then replace it by the low sensitive item value. Our work focuses on perturbation techniques on the correlated data of sensitive information.

It has four steps namely: Analysis, Selection, Transformation and Substitution.

1) **Analysis**: By using Pearson, Correlation Matrix is computed. It is the primary step in identifying the relationship between items, in particular with the sensitive items. To perform this, the sensitive items in the medical data set must be given to the system and which finds the correlations between each item X to each other item Y using Pearson correlation.
The correlation matrix is a X ∗ Y values where X- Number of non-sensitive items
Y- Number of sensitive items

$$\rho_{x,y}=\frac{con(X,Y)}{\sigma x \sigma y} \quad (1)$$

2) **Selection:** In this technique, if non-sensitive item X is highly correlated with a sensitive item Y, then Y can be replaced by X, if it satisfies statistical property. To find the accurate level for correlation, first it proceed from low level value to higher level value for correlating the non-sensitive and sensitive item. Correlated value should satisfy the threshold values which are formed for analyzing the Correlation Matrix for the entire sensitive item. There are 3 conditions that arise while processing the dataset.

1) If the values are highly correlated with any one of the non sensitive value, then substitute the related non-sensitive value to the sensitive value.

2) If the values are correlated within the boundary, then the transformation can continue.

3) If no values are found for the limit in such case threshold lower value are formed incrementally.

By doing this process sensitive values are removed, which is replaced by non-sensitive values and now waiting for transformation which is done in next steps.

3) **Transformation:** From the remaining sensitive item Y identify which are not correlated with any of the non sensitive item *X*, the subsets of sensitive data are formed .The result of this step is processed for substitution.

4) **Substitution:** Sensitive values are substituted with the related attributes derived from the transformation and the sensitive value will be replaced in the original data.

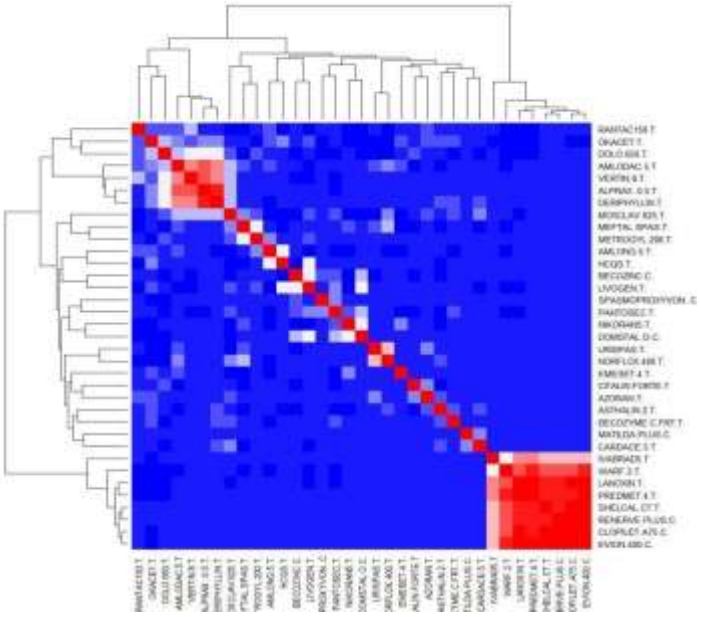

Figure 2 Sample Correlations for Medical Dataset

The figure 2 represented the correlation matrix for medical dataset with the sample of 30 records. The dendrogram of samples was divided into two parts based on the correlation between sensitive and non-sensitive attributes respectively. In particular the method finds the similarity between the sensitive and non-sensitive attributes.

4. RESULTS AND DISCUSSION

The proposed attribute probability matrix based privacy preservation algorithm is implemented and evaluated for its performance under different conditions. The result produced by the algorithm is given in this section.

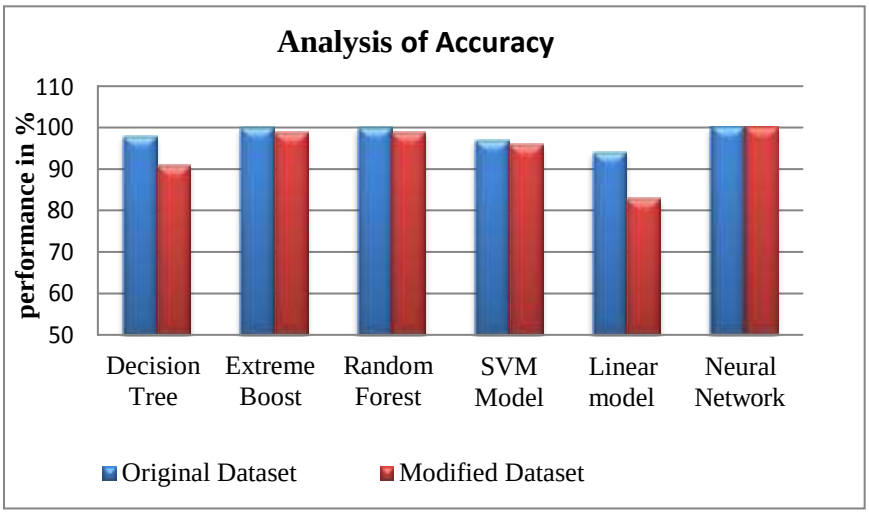

Figure 3 Comparison of Accuracy

The performance on privacy preservation has been measured for various methods. The proposed APM algorithm has produced higher performance than other methods. It Figure 3 shows the comparison of accuracy produced by various classifiers are Extreme Boost, Randam Forest and Neural Network produce 100% ,whereas SVM model and Decision tree produce more than 90% shows that the proposed approach has produced higher performance.

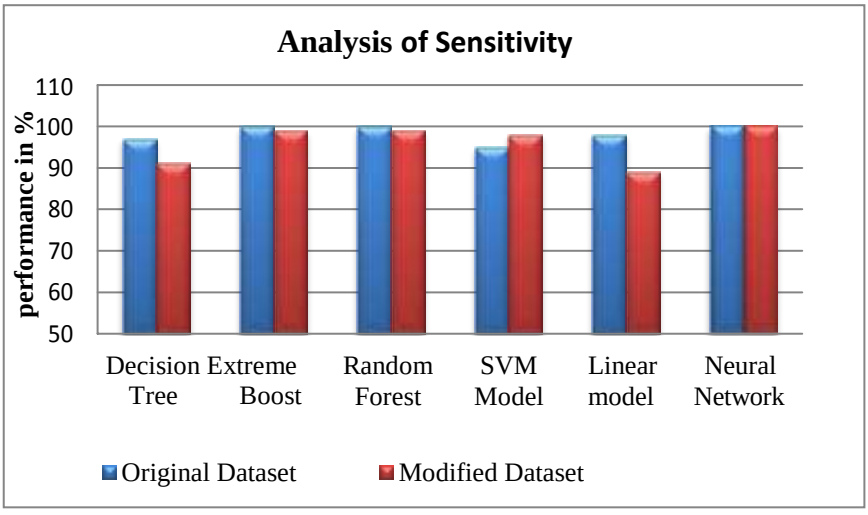

Figure 4 Comparison of Sensitivity

Above Figure 4, Shows the comparison of Sensitivity ratio produced for Extreme Boost, Random Forest and Neural network it has 100%, whereas for Decision tree, SVM model has more than 90% and the proposed approach produces higher performance ratio.

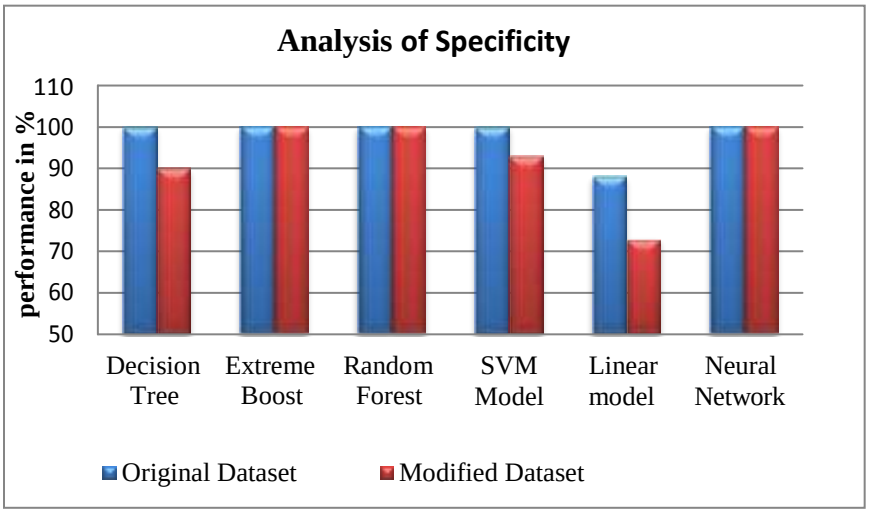

Figure 5 Comparison of Specificity

Above Figure 5, shows the comparison of specificity ratio produced by different classifiers and it produce 100% result for Decision tree, Extreme Boost, Randam Forest, SVM model, Neural Network, Whereas for linear model it shows 83%.

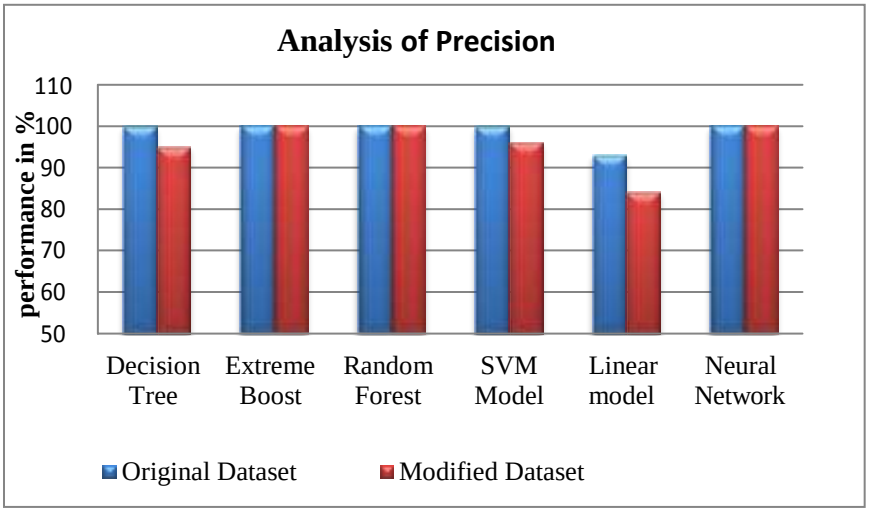

Figure 6 Comparison of Precision

Above Figure 6, shows the comparison of specificity ratio produced by different classifiers and it produce 100% result for Decision tree, Extreme Boost, Randam Forest, SVM model, Neural Network, Whereas for linear model it shows 93%.

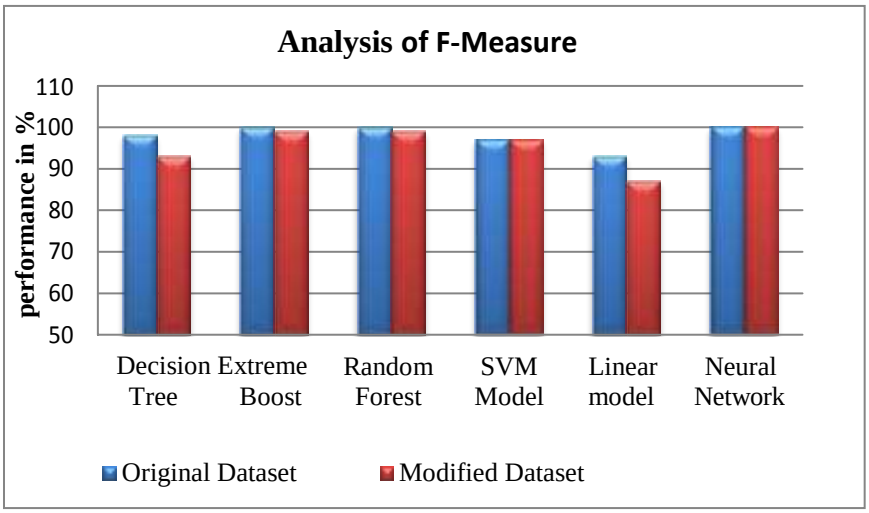

Figure 7 Comparison of F-Measure

Above Figure 7, shows the comparison of F-Measure produced by different classifiers and it produce 100% result for Extreme Boost, Randam Forest, SVM model, Neural Network, Whereas for linear model and Decision tree it shows above 95%.

5. CONCLUSION

The proposed graph based Anonymization process for privacy preservation using Correction Sanitization outperforms better performance than the early methods. The proposed Correlation based transformation method produces good results. The published dataset holds only the data transformation value, still the user can gather sufficient information but they cannot bring back the original data set. The proposed method has produced higher efficiency on privacy preservation with 96%.

6. REFERENCES

[1]. Dharmendra Thakur and Prof. Hitesh Gupta," An Exemplary Study of Privacy Preserving Association Rule Mining Techniques", P.C.S.T., BHOPAL C.S Dept, P.C.S.T., BHOPAL India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 11,2013.

[2]. C.V.Nithya and A.Jeyasree,"Privacy Preserving Using Direct and Indirect Discrimination Rule Method", Vivekanandha College of Technology for WomenNamakkal India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 12,2013.

[3]. T.J. Trambadiya,and P. bhanodia , "A Heuristic Approach to Preserve Privacy in Stream Data with Classification", International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 1, pp.1096-1103, Jan -Feb 2013.

[4]. Dhyanendra Jain , Hiding Sensitive Association Rules without Altering the Support of

Sensitive Item, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[5]. MarcinGorawski, An Efficient Algorithm for Sequential Pattern Mining with Privacy Preservation, Advances in Systems Science Advances in Intelligent Systems and Computing Volume 240, 2014, pp 151-161.

[6]. FatemehAmiri, A Novel Community Detection Algorithm for Privacy Preservation in Social Networks, Intelligent Informatics Advances in Intelligent Systems and Computing Volume 182, 2013, pp 443-450.

[7]. Jinyan Wang, Two Privacy-Preserving Approaches for Publishing Transactional Data Streams, IEEE access, Volume 4, 2018.

[8]. ChaobinLiu, A novel privacy preserving method for data publication, ELSEVIER (IS), Volume 501, 2019, PP 421-435.

[9]. ChongjingSun, Personalized Privacy-Preserving Frequent Itemset Mining Using Randomized Response, HINDAWI (SWJ), 2014.

[10]. Maurizio Atzori, Francesco Bonchi, FoscaGiannotti, and Dino Pedreschi. Blocking anonymity threats raised by frequent itemset mining. In ICDM, pages 561–564, 2010.

[11]. Sreeja mole, Sujatha Krishnamoorthy* (2019) An efficient Gait Dynamics classification method for Neurodegenerative Diseases using Brain signals, Published in Journal of Medical System, Springer

[12]. Venkatachalam K, Karthikeyan NK (2018) A framework for constraint based web service discovery with natural language user queries. J Adv Res Dyn Control Syst, Elsevier Publication 05-Special Issue, 1310–1316

[13]. S. Ramamoorthy, G. Ravikumar, B. Saravana Balaji, S. Balakrishnan, and K. Venkatachalam, "MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-8, 2020.