An Approach, For Shielding Sensitive Information Using Data Mining Technique

Suman, Ravinder Shahrawat, Sarita

Abstract*: large volume of data is regularly collected and shared, for the purpose of various data mining application. The data includes a lot of personal and provisional information. Some of the examples like shopping pattern, criminal record, Diagnosis history, bank details, address, personal emails etc. Now, there is very urgency, as to preserve the privacy of these kind of sensitive data. These data are sent to data analyst for further knowledge discovery. In order to share and keeping the goal in mind as to preserve the privacy of such data, this approach is designed. Data Sharing is very important phase in data transformation stage. Need to make sure that ,if any technique is implemented as to preserve the data security ,that must preserve data behavior and characteristic .Here, data is divided in two parts as character and numerical. Character data is shielded using data masking technique. Numerical data is shielded using clustering algorithm

Index Terms: Large, collected, shopping, Character

I. Introduction

Considering the huge amount of data processing done at various level of Information system, maintaining the data privacy is a very vital part of any organization. Here DSDMT (Data Shielding Using Data Mining Technique) Technique is proposed as to shield the sensitive data or information before sending that to any third party. This technique is proposed for centralized database (data source) environment. In distributed database environment ,there are various level of privacy setting are faced like different vendors and terms and conditions for sharing the data source ,which is not so in centralized data source. DSDMT maintains/preserve the original data integrity and consistency. Data is modified to some extend while keeping the data characteristic and behavior untouched to large extend. Therefore, the data lost is very minimal (negligible) in this approach. DSDMT technique deals with both character and numerical dataset. The original dataset containing the sensitive information is assumed and classified in character and numerical dataset. Then this technique deals individually with each of these dataset as to implement the proposed method and finally join them, to form the full dataset. The data flow of DSDMT can be shown as:

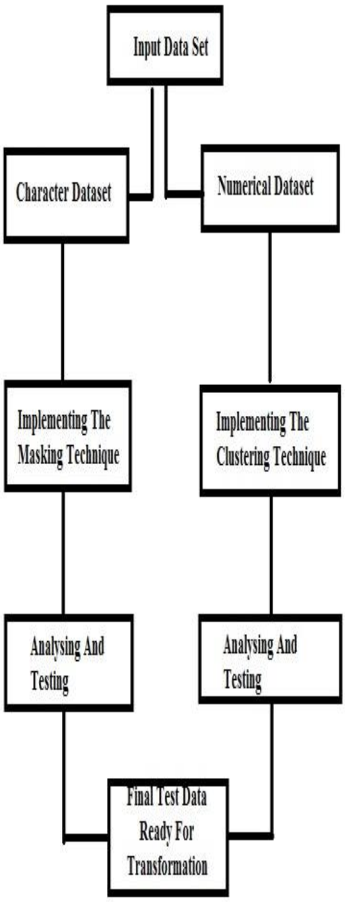
Revised Manuscript Received on 30 September 2014.
* Correspondence Author

Suman*, Department of Information technology, World College of Technology & Management, Gurgaon, India.

Ravinder Shahrawat, Department of Information Technology, GGSIPU, India

Sarita, Department of ECE, World college of Technology and management, Gurgaon, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/


Input Data Set
Character Dataset
Numerical Dataset
Implementing The Masking Technique
Implementing The Clustering Technique
Analysing And Testing
Analysing And Testing
Final Test Data Ready For Transformation

Pulling Data Source: Centralized data source is considered for Input step for DSDMT technique.

Data Modification :Here the Original dataset is divided in 2 parts as Character dataset and Numerical dataset. Character dataset is shielded using Data masking technique[2] . Numerical dataset is shielded using data clustering technique of data mining .The dataset obtained from the above shielding methods are further analyzed and tested using the F Measure and distortion factors.

Data transformation: Here the transformed dataset obtained from above step are combined together as to form the final test dataset. Both character and numerical dataset are combined to form the original dataset, which is consistent and integrated as per the original dataset. This transformed dataset is the output for this approach ,which is shielded and its privacy is maintained .

II. Conceptual Theory Of DSDMT Technique

A. *Character Solution Definition (Dc):*

Data masking technique is used to shuffle the data so randomly, that it's challenging for reverse engineering to re-generate the original dataset from the transformed one[2]. There are 2 ways of data shuffling in a dataset :

1. By Row wise
2. By Column wise

Data is shuffled by column wise .Data can't be shuffled by row wise as by doing so, the data may lost its consistency in term of data types, formatting, values and behaviour for that column. An address , SSN, name, city, country can't be shuffled row wise as by doing so , it will lost its originality.

The degree of shuffling depends on the probability of presence of the data value at its original position, after shuffling is done on it .The more efficient shuffling is where the probability of finding is very minimal at its original position. Mathematical expression for calculating the chances for a data value to be present at its original position is derived as:

Pn = 1/n

Where:

Pn= Probability of occurrence of a data value at its original position after data shuffling is done.

n= Number of rows in a dataset.

The probability parameter (Pn) is inversely proportional to the number of rows in a dataset. That means, more rows ,less is the probability and better is the shuffling.

B. *Numerical Solution Definition (Dn):*

Steps Involved In Numerical Data Solution Process:

1. *Input Data*

Input is dataset which is stored in file which contains sensitive information ,which is to be shielded, such that there is less information loss and hence good clustering result.Each row of data is sequence of real value X = x1 x2 x3…….xm . Dataset contain "n" rowof data.The row values don't contain any character value .There is no missing or special character value in the dataset.The source of dataset for this is centralized source.

In our case we use Computer Hardware dataset available at UCI Repository[11].

2. *Dividing in K Clusters*

The input data is divided in K cluster using K Mean clustering algorithm[10]. The K mean algorithm can be explained as :

k-mean, is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. This algorithm follows a simple, easy and straight way to cluster a given data set through a certain number of clusters (assume k clusters) as per requirement . The main emphasis is to define k centers, one for each cluster/Group. These centers must be placed in a cunning way as different location causes different result. So, the best choice is to configure/select them as much as possible far away from each other as that can give better clustering. The k-means algorithm(clustering) steps can be listed as:

Select a random value of K(number of clusters) and then follow the below steps:

1)Pick randomly 'k' cluster centers from the original dataset.

2) Generate the distance between each data point and cluster centers (Centroid).

3) Pair the data point to the cluster center(Centroid), whose distance from the cluster center is minimum of all the cluster centers(Centroids).

4) Regenerate the new cluster center(Centroid) using Euclidean distance:

5) Recalculate the distance between each data point and note the new centroid for these clusters.

6) If no data point(centroid) moves then stop, otherwise repeat from step 3.

Note: As to calculate the distance between centroid and each object ,Euclidean distance is used.

The Euclidean distance between points p and q can be evaluated as[14] :

if p = (p1, p2,...,pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space:

$$d(p,q)=d(q,p)=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

The Original dataset (Dn) is divided in K clustered. The centroid is calculated for each cluster using Euclidean distance. Since, the centroid represent the closest value for the values(rows) in a cluster .The value of each row in a cluster is replaced by its centroid of that cluster.

Also, the degree or dimension of a centroid is same as of rows of that cluster.

Mathematically ,it can be represented as :

Let D be the original dataset then it can be represented in the clustered form as :

D_n= D_1 D_2 D_3 D_4…………………………D_{K-1} D_K

Now since, here the centroid is of same dimension as of the cluster, so the centroid can be represented as:

C_1 C_2 C_3 C_4……………………………….C_{K-1} C_K

3. *Transformed dataset*

After rows are replaced by their corresponding centroids ,the new transformed data set can be represented as:

D_n'= C_1 C_2 C_3 C_4…………………………C_{K-1} C_K

The function which is used in transformation is:

T(Dj) = {Centroid(Cj): only if d(Dj , centroid(Cj)) < d(Dj , centroid(Cp)) for ∀p }

D is migrated to D', with having similar approximated value but not exact what it had in original dataset.

4. *Data Aggregation:*

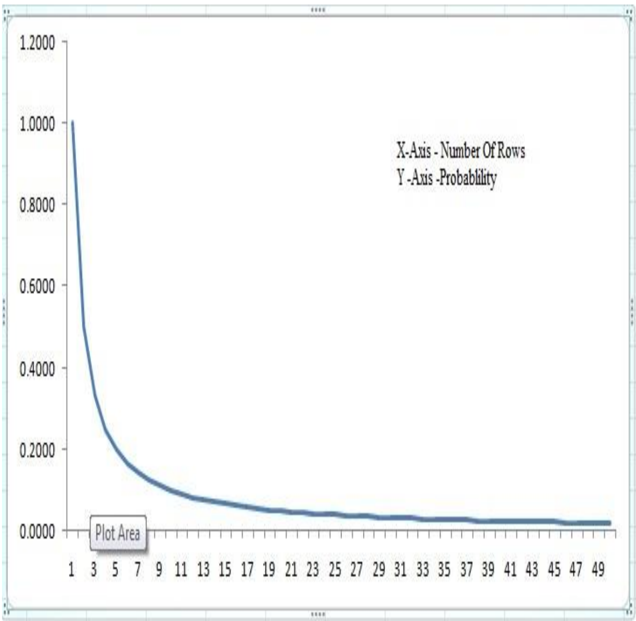
The final dataset is obtained by aggregating the dataset obtained from Data masking on character datset(Dc) and K mean clustering implementation on numerical dataset(Dn).

Let D was the original dataset as:
D = Dc +Dn
Where:
Dc =Dataset with character variables
Dn =Dataset with Numeric variables
Then final transformed dataset from point 2 and 3 can be stated as:
D'= Dc' +Dn'
Where:
Dc' = Dataset with character variable obtained after transformation done using data masking technique.
Dn'=Dataset with numerical variables obtained after transformation done using K mean Clustering technique

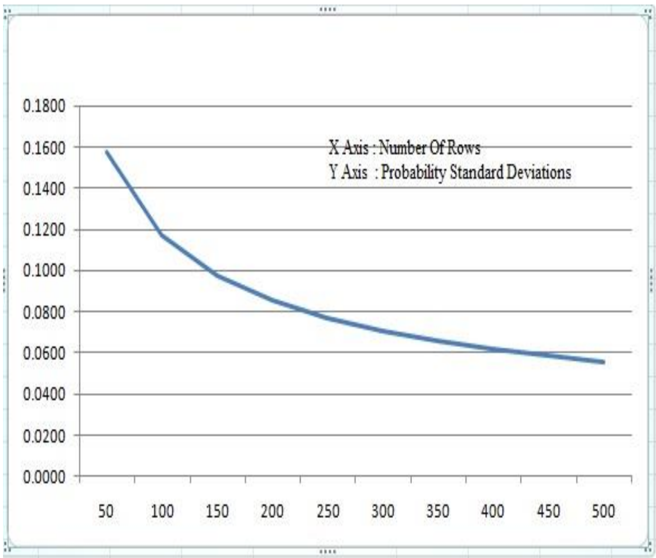
III. Results And Verifications

***Data source*:Credit** Approval Data Set from UCI Machine Learning Repository[11].

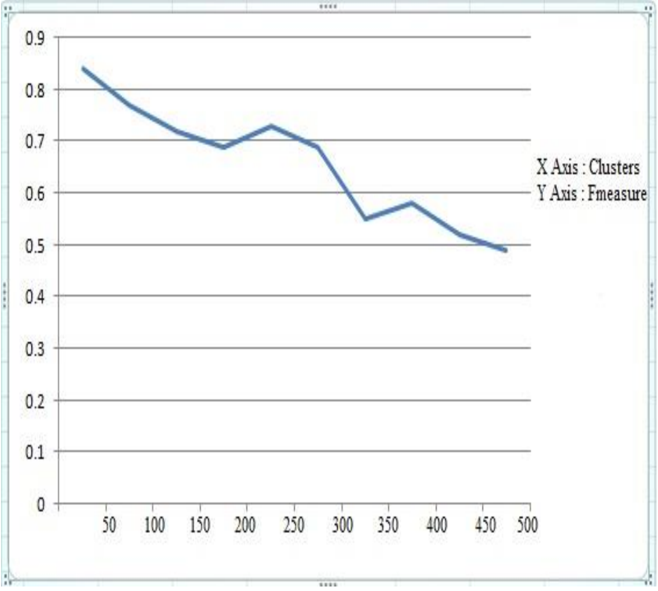
For Character Dataset :Shuffling using data masking technique Behaviour of Probability with respect to the number of rows in dataset



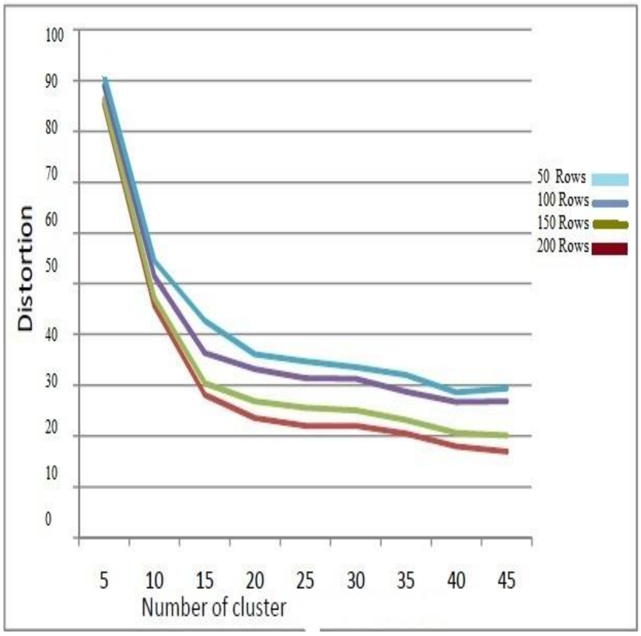
Behaviour of standard deviation of probability with respect of number of rows :



For Numerical Dataset :K mean clustering in data masking technique Behaviour of FMeasure with respect of number of cluster:



Behaviour of Distortion factor with respect to number of cluster:



IV. FUTURE SCOPE

Considering the current scenarios in the technology and market trends, the privacy of the data is the major concern .The sensitive data is required to be shielded, before sending it to any third party. Also, precautions need to be taken ,while shielding the data as to keep its integrity and consistency untouched. Also, the there must be minimal data lost. Keeping all these things the future scope for DSDMT Technique can be listed as:

i) DSDMT Technique can be viewed as a potential method to shield the privacy of the sensitive data.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved

ii) Here, for character data shielding, data shuffling is used .Data shuffling is the most common and easy to implement technique. As observed in experiment and results ,it works very well for large dataset .So, there can be an approach which can handle both the small and large dataset in same manner and their privacy can be maintained ,which is independent of the size of dataset. Additional complex rules can also be added into any masking solution regardless of how the masking methods are implemented. Row Internal Synchronisation Rules, Table Internal Synchronisation Rules and Table to Table Synchronisation Rules are some from the list.

iii) The numerical dataset is shielded using the K mean clustering technique. K means centroid are the major data transformation parameters used . Other techniques like ,K nearest neighbour approach is one of the approach which can give better result.

REFERENCES

1. Micheline Kamber, Data Mining concepts and techiniques, Second Edition, Jiawei Han University of Illinois at urbana-Champaign.
2. Data Masking overview: http://en.wikipedia.org/wiki/Data_masking .
3. Data Mining overview : http: http://en.wikipedia.org/wiki/Data_mining updated on 25th june 2014.
4. Anomaly detection: http://en.wikipedia.org/wiki/Anomaly_detection , latest updated on 24th june2014.
5. Association rule learning: http://en.wikipedia.org/wiki/Association_rule_learning, latest updated on 4th june2014.
6. Clustering : http://en.wikipedia.org/wiki/Cluster_analysis , latest updated on 18th may 2014.
7. Classification : http://en.wikipedia.org/wiki/Statistical_classification latest updated on 18th may 2014.
8. Regression: http://en.wikipedia.org/wiki/Regression_analysis, latest updated on 18th may 2014.
9. Summarization: http://en.wikipedia.org/wiki/Automatic_summarization, latest u pdated on 10th june 2014.
10. Z. Huang. : “Extensions to the K-mean algorithm for clustering large data sets with categorical values”, Data mining and knowlwdge discovery, 2:283-304,1998.
11. Source dataset: UCI Repository of machine learning databases, University of California, Irvine, http://archive.ics.uci.edu/ml/
12. Berkhin Pavel, A survey of Clustering data mining Techniques, Springer Berlin Heidelberg,2006.
13. Wu Xiaodan,Chu Chao-Hsien, Wang Yunfeng, Liu Fengli, Yue Dianmin, Privacy Preserving data Mining Research: Current status and key issues,Computational Science-ICCS 2007,4489(2007), 762-772.
14. Agarwal Charu C., Yu Philip S., Privacy Preserving Data Mining: Models and Algorithms, New York Springer, 2008.
15. Kurt Thearling, Information about data mining ans analytic technologies, http://www.thearling.com/

AUTHOR PROFILE

Suman, M.Tech in Software Engineering from world college of technology and management, Gurgaon. Area of interest are Data Mining, Data clustering and Statstical Analysis.

Ravinder Shahrawat, B.Tech in IT from GGSIPU, India. Presently working in Aon Hewitt. Area of Interest are Data management, Data programming and Statstical Analysis.

Sarita, M.tech in ECE from world college of technology and management, Gurgaon.Her research interest are in digital designing.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

© Copyright: All rights reserved