International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Real Time Feature Convergence Measure for Efficient Discrimination for Transactional Data Set

M.A.Jamal Mohamed Yaseen Zubeir, A.R. Mohamed Shanavas

Abstract: ***The problem of discrimination in transactional data set has been well studied. Numerous techniques has been recommended by various researchers but suffer to achieve higher performance. To handle this issue, a real time feature convergence measure based discrimination prevention algorithm is presented in this paper. The method first eliminates the noisy records by preprocessing the transactional data set. Second, the transactional data set has been grouped into number of clusters according to the pattern relevancy measure (PRM). Using the clusters generated, the the feature convergence measure (FCM) is computed for each item towards each cluster. The value of FCM is used to select a subset of items as sensitive one. Based on identified sensitive items, the method performs sanitization using probabilistic mapping scheme. The FCM algorithm supports the performance development of sanitization and discrimination prevention.***

Index Terms: ***Transactional Data, FCM, PRM, Discrimination, PMS..***

I. INTRODUCTION

The human society performs their daily transaction in purchasing various things through the web. Whatever they purchase has been maintained by the organizational transactional data set. The logs present the data set has been used for several purposes. The product manufacturer has the responsibility in meeting the needs of the customers. The interest and expectation of the customers gets changing every day and the product manufacturer has to be more sensible in producing the goods according to the needs of the users. Not only that, any manufacturer has to monitor the movement of other products which also impact on the sale of their own products.

The consumers are more confident on the product manufacturers and the organization which maintains the logs related to their purchase. The organization maintains several information related to the consumer purchase, their personal details and so on. The transactional data set would contain different sensitive information which can be obtained from the logs. For example, a user A, who purchased different medicine and sexual products would contained in the data set. The user would not like such information being exposed to the external world. The organization is the sole authority to provide data security to the users.

In most situations, the data present in the transactional data set has been used to generate business intelligence to measure the growth of the organization. Sometimes, the data set has been shared with the other organizations who are partnered with the organization. In this case, if the original data set is given, then the third party would get information related to the user and his privacy will be breached. This reduces the trustworthy of the organization among the customers. To handle this issue, sanitization is performed by the data owner. The sanitization is the process of hiding sensitive information without the loss of data in the original and published data set. The published data set contains all the necessary information but the sensitive information will be hidden and modified. To perform this, several algorithms like Dot matrix, averaging scheme are available. But each suffer with the accuracy of data publishing.

To identify the sensitive items and to perform discrimination prevention, an efficient real time feature convergence measure based discrimination algorithm is presented and explained.

II. RELATED WORKS

Various algorithms has been presented for the discrimination prevention in literature. This section discusses few of them related to the problem.

The data hiding in transactional data set has been perform using cryptography techniques to restrict the user access in [1]. The restriction has been enforced with role based approach.

The data hiding in consumption data has been handled with legacy and distributed approach which partition a data in different locations. The selection of data is performed by measuring the similarity between the data [2].

Fuzzy based data hiding is presented in [3], which transform the values of data set in any dimension into fuzzy values. This restricts the identification of original values and preserves the privacy of user data.

Manuscript published on November 30, 2019.
* Correspondence Author
M.A.Jamal Mohamed Yaseen Zubeir*, Corresponding Author, Research Scholar, Department of Computer Science, Jamal Mohamed College (Autonomous) (Affiliated to Bharathidasan University), Tiruchirappalli, Tamilnadu, India. (Email: zubi2208@gmail.com)
Dr. A.R. Mohamed Shanavas, Associate Professor, Department of Computer Science, Jamal Mohamed College (Autonomous) (Affiliated to Bharathidasan University), Tiruchirappalli, Tamilnadu, India.

A Heuristic approach for data hiding is presented in [4], which users the association Rule the method maintains different rules to perform clustering and update, which uses the support values. The process of data hiding any data set has impact on the originality of data being published. In [5], a hiding missing artificial utility (HMAU) algorithm is proposed. The method identifies a sensitive and non-sensitive items are identified and deleted. This improves the performance of data hiding.

In different methods the support value of sensitive items has been altered before publishing. To perform data hiding without altering the support values, an association rule based approach is presented [6]. The original data set has been changed before publishing but the support value of sensitive items has not been changed.

In [7], privacy preservation is approached with pattern based technique. The frequent items and sequential patterns are identified using an index based approach. The SSAPP approach support the reduction of frequency of cryptography operations as the method maintains the patterns in form of trees.

The problem of community detection in social network has been approached using learning automata theory. The theory splits the patterns and indexes them and form of graphs. The method groups the communities at each alteration to identify the community [8].

The privacy preservation in health care systems has been enforced using the user identity [9]. The method uses cryptographic techniques which are performed according to the user identity. The key agreement is performed using bilinear pairing.

K-Anonymity in Multidimensional dataset has been presented in [10]. The anonymity has been approached with greedy approximation algorithm.

The importance of data preprocessing has been discussed in [11]. The convex optimization has been used to transform the data, which has been used group detection. A rule based anti- discrimination is used which works according to classification rules. Similarly a taxonomy based approach is presented. In [12], the problem of direct and indirect discrimination prevention is handled with data mining techniques. The method uses different rules to support the above mentioned problem. Similarly, in [13], a set of taxonomy has been used for the prevention of discrimination in both ways.

The problem of Discrimination prevention towards intrusion detection is presented in [14]. The data mining algorithm has been used to generate classification rules and has been used to identify the sensitive items and performs sanitization.

III. REAL TIME FEATURE CONVERGENCE MEASURE BASED DISCRIMINATION PREVENTION

The input transactional data set has been read and identifies the list of features and values. According to the feature list, the preprocessing is performed to eliminate the noisy records from the data set. With the noise removed data set, a set of transactional pattern has been generated and according to the pattern set generated, the method clusters the data set into number of groups. With he clusters, the method estimates feature convergence measure (FCM) to identify the sensitive items. Based on the value of FCM, the method performs sanitization using probabilistic mapping scheme. The detailed approach is discussed in this section.

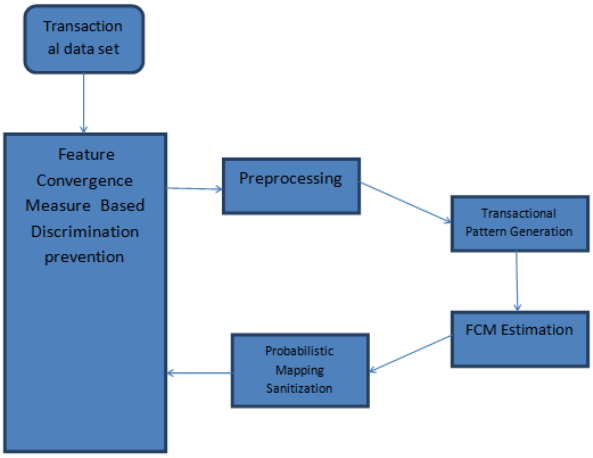

Figure 1: FCM based Probabilistic Mapping Sanitization Architecture

The general block diagram of FCM based Probabilistic Mapping Sanitization algorithm is presented in Figure 1.

IV. PREPROCESSING

The preprocessing is performed to eliminate the noisy records from the input transactional set. The transactional log would contain much information but there will be records which would have missing values. Such missing records are removed to work with the remaining process. First, the lists of dimension are identified and for each dimension there should be values. If there is no value for any of the feature, then it is considered as noisy record and it has been removed from the data set. The remaining will be used towards pattern generation.

Preprocessing Algorithm:

Given: Data set Trds
Obtain: Preprocessed Data Set Prds
Begin
Fetch Trds.
Find the list of all attributes.
lA = $\int_{i=1}^{size(Trds)} lA \cup (\sum Trds(i).Attr \ni lA)$
For each transaction Ti
If $\int_{i=1}^{size(lA)} if\ LA(i) \in Ti \&\& Ti(LA(i)) == Null$ then
Eliminate the log
Else
Add the log to Prds.
Prds = $\sum DataPoints(Prds) \cup Ti$
End
End
Stop

The above discussed algorithm represents the way preprocessing is performed to do data cleaning preprocess which remove the noisy data points from the data set.

V. TRANSACTIONAL PATTERN SET GENERATION

The noise removed data set is the key for generating the transactional data set. To generate that, for each transaction items, number of possible combinations is generated. For example, if it contains N number of items in the item set, then the patterns are generated from 1 to N. It will be created as one item set, 2 item set, …, N item set. Generated pattern set is used to perform clustering.

VI. PATTERN RELEVANCY MEASURE (PRM) CLUSTERING

Clustering the transactional patterns is performed according to the pattern relevancy measure. The relevancy of pattern has been measured not only based on the similarity among the items of pattern between the input and the patterns present in any cluster. First, random patterns are added to each cluster according to the number of clusters. Then for each pattern from the pattern set, the method measures the pattern relevancy according to the size as well as features. The method estimates the PRM for each cluster for any pattern from the set. According to the value of PRM, a single cluster is identified and indexed.

PRM Clustering Algorithm:

Given: Transactional Pattern Set Trps.
Obtain: Clusters Cs
Begin
Fetch Trps.
Initialize Cluster set Cs.
For each cluster c
Select random pattern p from Trps.
C(i) = $\int_{i=1}^{size(Trps)} Random(1, size(Trps))$
End
For each pattern p from Trps
For each cluster C
Compute the dimensional similarity Dsim = $\int_{i=1}^{size(C)} \frac{\sum C(i).Dim==Size(p)}{size(C)}$
Compute feature similarity Fs = $\int_{i=1}^{size(C)} \frac{\sum_{i=1}^{size(C(i))} C(i)(j)==p(j)}{size(C(i))}$
Compute Pattern Relevancy Measure PRM.
PRM = Dsim×Fs
End
Choose cluster with maximum PRM.
Index the pattern to the selected cludter.
End
While true
For each cluster
For each pattern
For each cluster
Estimate PRM
End
End
Choose the cluster with maximum PRM.
Index the pattern to the selected cluster.
End
Continue if there is movement
end
Stop

The patterns relevancy measure in clustering algorithm presented above shows how the clustering is performed according to the pattern and the data points of the data set.

VII. FEATURE CONVERGENCE MEASURE ESTIMATION

The feature convergence measure represents the attraction of the feature towards the data points of the cluster. It is measured according to the frequency of item in the patterns of the cluster. It is measured according to the number of patterns contains the specific item and total number of patterns. Similarly, the frequency of the item in other cluster also measured. Using both of them, the FCM measure is estimated for each cluster. Finally, if the item has higher FCM for the specific cluster, then the item has been selected for sanitization.

FCM Algorithm:
Given: Cluster set Cs
Obtain: FS, FCMs
Begin
Fetch CS.
For each cluster c
Identify list of all features Fl = $\int_{i=1}^{size(C)} \sum Feature(C(i)) \ni Fl$
For each feature f
Compute number of occurrence Noc = $\int_{i=1}^{size(C)} \sum C(i) \in f$
Compute Frequency measure Fm = Noc/Total Patterns
For each other cluster
Compute frequency measure OFm.
End
Estimate FCM = Fm×OFm
C = choose the cluster with maximum FCM.
If C==C then
Add to feature fs.
end
end
end
Stop

The FCM algorithm presented above shows how the feature selection is performed for any cluster. Identified feature set has been used to perform sanitization.

VIII. PROBABILISTIC MAPPING SANITIZATION

The sanitization on the transactional data set has been performed based on the Feature convergence measure. For each feature, the method compute the FCM value and if the FCM value is higher than specific threshold then it has been selected for sanitization. With the features identified, the method estimates the probability value of the feature and places at the published data set.

Algorithm:

Input: Cluster Set Cs, Feature Set Fs, FCMs
Output: Publication Set Ps
Start

```
Read Cs, Fs, FCMs.
Initialize publication set ps.
Identify list of features Fl.
For each feature f
   If Fcms(f)>Th then
      Add f to sanitization feature list Sfl
   End
End
For each feature from sfl
   Compute probability value Pv.
   Replace publication set with pv.
end
```

Stop

The Probabilistic Mapping sanitization algorithm presented above shows how the sanitization is performed and publication set has been prepared.

IX. EXPERIMENTAL RESULTS

The proposed time orient feature convergence measure orient probabilistic mapping scheme is hard coded in Advanced Java where its performance is computed and compared with other techniques.

Parameter	Value
Data Set	Transactional Amazon
Total Items	3000
Total Logs	3 million
Platform	Advanced Java

Table 1: Simulation details

The details of simulation being used for the evaluation of proposed algorithm is presented in Table 1. The performance of the method has been measured in various parameters.

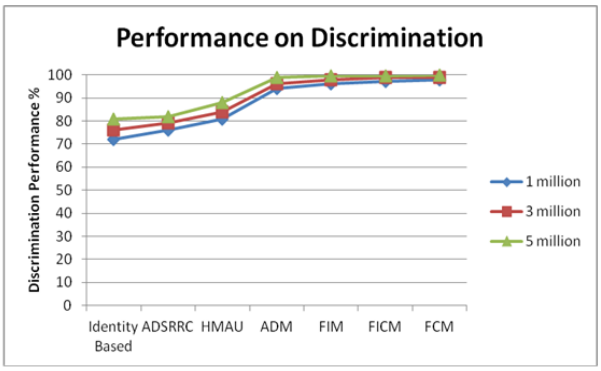

Fig 2: Performance on discrimination

The performance on discrimination produced by different methods has been presented in Fig 2. The proposed FCM based approach improves the discrimination performance than other methods.

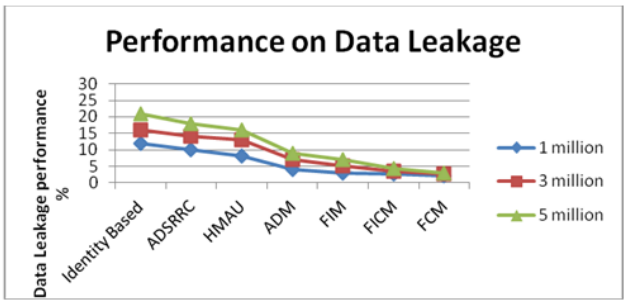

Fig 3: Performance on Data Leakage

The amount of data leakage produced by the FCM algorithm is measured. It has produced less leakage compare to other methods.

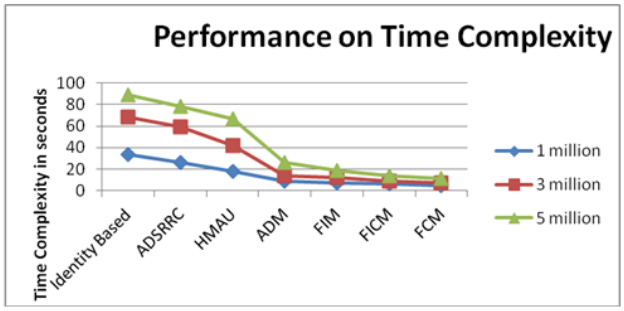

Fig 4: Performance on time complexity

The performance in time complexity is measured for the FCM algorithm where it achieved less time complexity than other techniques.

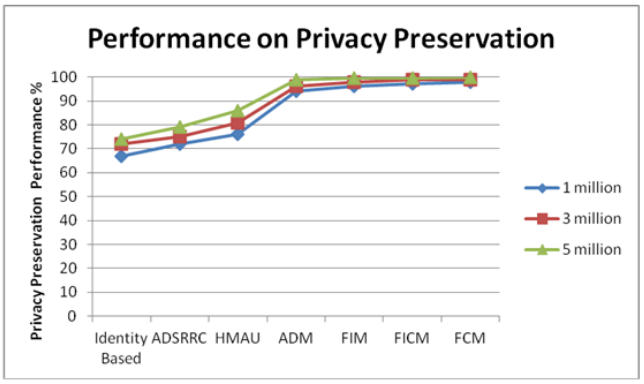

Figure 5: Performance on privacy preservation

The performance on privacy preservation produced by different methods have been measured and compared with the result of proposed methods. The proposed FCm algorithm have achieved higher performance on privacy preservation than other methods.

X. CONCLUSION

In this paper, an efficient real time feature convergence measure (FCM) based discrimination prevention approach is presented. The method uses the feature convergence measure to perform sanitization. The method first preprocesses the data set to eliminate the noise records and use them to

generates transactional pattern set. The pattern set are clustered into number of groups according to the pattern relevancy measure. From the clusters generated, the method estimates the feature convergence measure for each attribute, based on which a subset of features has been selected. According to the probabilistic mapping scheme, the method performs sanitization. The proposed method improves the performance of discrimination prevention and reduces the time complexity.

REFERENCES

1. Murugeswari.s, "An Efficient Method for Knowledge Hiding Through Database Extension", IEEE conference on ITC, 2010.
2. V. Thavavel, "A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment", IJCSI , Issues, Vol. 9, Issue 1, No 2, 2012
3. M Sridhar, "A Fuzzy Approach for Privacy Preserving in Data Mining", IJCA , 57(18):1-5, 2012.
4. Komal Shah, Amit Thakkar ,"Association Rule Hiding by Heuristic Approach to Reduce Side Effects and Hide Multiple R.H.S. Items",IJCA, 45(1),pp:1-7, 2012.
5. Chun Wei Lin, "Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", The Scientific World Journal Volume 2014 (2014).
6. Dhyanendra Jain , " Hiding Sensitive Association Rules without Altering the Support of Sensitive Item", IJAIA, Vol.3, No.2, 2012.
7. Marcin Gorawski," 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, 2014.
8. Fatemeh Amiri, "A Novel Community Detection Algorithm for Privacy Preservation in Social Networks", Intelligent Informatics Advances in Intelligent Systems and Computing, Volume 182, 2013, pp 443-450,2013.
9. Kambombo Mtonga, "Identity-Based Privacy Preservation Framework over u-Healthcare System", Multimedia and Ubiquitous Engineering Lecture Notes in Electrical Engineering, Volume 240, 2013, pp 203-210.
10. K. LeFevre, D.J. DeWitt, "Mondrian Multidimensional k-Anonymity," IEEE Int'l Conf. Data Eng.(ICDE), 2006.
11. Flavio du Pin Calmon, "Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis", IEEE Journal of Selected Topics in Signal Processing, Volume: 12 , Issue: 5 , 2018.
12. Sara Hajian, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining", IEEE Transaction on Knowledge and Data Engineering, 2013.
13. Sara Hajian, "Direct and Indirect Discrimination Prevention Methods", Springer, Discrimination and Privacy in the Information Society, pp 241-254, 2013.
14. Sara Hajian , "Discrimination prevention in data mining for intrusion and crime detection", IEEE Symposium on Computational Intelligence in Cyber Security (CICS), 2011.