International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Feature Correlation Measure Based Real Time Discrimination Prevention with Transactional Data Sets using Social Networks

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

Abstract: ***The effect of social network in anti-discrimination has been analyzed in detail.Various approaches towards anti-discrimination on transactional data have been identified, but does not produced expected performance level. To improve the performance, a feature correlation measure based approach has been presented in this article. The method reads the transactional data set and generates number of patterns based on the purchase details. For each pattern generated, the method estimate pattern impact measure towards the data set. Based on the value of PIM (Pattern Impact Measure), a subset of patterns is selected. With the selected pattern set, the method reads the social network user data and identifies the list of items being discussed. According to identified items, the method estimates the feature correlation measure (FCM) for each items identified. According to the value of FCM, a subset of items with higher value has been selected. The selected items are identified as more sensitive and being sanitized with the publishing data set. The sanitization is performed using probabilistic sanitization algorithm. The FCM algorithm has leverage the discrimination prevention and sanitization performance.***

Keywords: Discrimination, Transactional Data, Social Networks, FCM, PIM, Probabilistic Sanitization.

I. INTRODUCTION

The growth of information technology allowed the organizations in maintaining various information of different customer in their database. The customers purchase various products through E-commerce and other gates. Such purchase details are stored in the transactional data set. Similarly, the organizations maintain information related to their purchase as well as other personal information. Both, transactional and personal information would contain many sensitive items belong to the users. The organization has the responsibility in maintaining the secrecy of user data. In reality the data belongs to an organization has been shared with others for business analysis. What happens in this is, the other organization would misuse the data and perform criminal activities. This must be avoided and eliminated. The discrimination prevention is the process of avoiding any criminal activities and safeguarding the user data efficiently. Privacy preservation is the way or sanitization is the way to secure the user data.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
M. A Jamal Mohamed Yaseen Zubeir.*, Research Scholar Jamal Mohamed College,Department of computer science,Trichy.
A.R. Mohamed Shanavas, Associate Professor, Department of computer science,Jamal Mohamed College,Trichy-20.E-mail: jamalcsemohamed47@yahoo.com

For example, when a organization manufactures many products, their sale would not be at the expected level. To improve the sale, they would like the collaborate with the other product manufactures in promoting the product.

But sharing the original data would lead to leakage of customer information and lead to privacy breach. To secure the user data, there are number of sanitization techniques available. In simple manner, a frequency based approaches identifies the most or least frequent items as more sensitive and hides such information. Dot matrix approaches places dots in the publishing data set which leads to incomplete analysis which would not support efficient business intelligence generation. Transactional data set would contain many information which has been used to identify most moving items. By identifying the most selling product, the organization would look to increase the manufacturing and would be used to develop the process. If the entire data set has been published as it is, the other organization would find the customers who purchased the data and may overlook the partner. This really affect the business of the home organization which must be avoided. The sanitization of the transactional data set can be performed by identifying the most impacting product. By identifying the impacting product, such data can be sanitized without modifying the meaning of the data set. The pattern mining techniques has been used in different problems. From the transactional data set, you can generated different patterns of purchase, which can be used to identify set of items which are more sensitive. Each pattern of purchase would present in number of times in the transactional data set. By computing some frequency measures and impact measures, a subset of patterns can be identified. The sanitization can be performed by identify a subset of patterns from the pattern set.

The social networks has more impact on different problems. In modern social networks, the users shares many information. By monitoring the logs of social network, the item being spoke by many users can be identified. Such products can be considered for the sanitization. By identifying the sensitive items through the support of social network data set, the most impacting products can be identified to perform sanitization. To achieve this, a feature correlation measure has been discussed in this paper. The FCM has been estimated according to the impact and correlation with the items of patterns generated. The detailed approach is discussed in the next section.

II. RELATED WORKS

There are number of algorithms available for the sanitization and discrimination of transactional data set. This section discusses set of methods towards discrimination in detail. A database extension based algorithm towards knowledge hiding is presented in [1].

The method focused to reduce the information loss and loss in privacy. Crypto techniques are used restrict the data access using profiles.

In [2], a privacy preservation algorithm is presented towards unstructured data. The legacy systems are used to move unstructured data into structured one using the legacy systems. In [3], a fuzzy based approach is presented which transform the data attributes in to set of fuzzy values. By presenting the range values, the original value of any feature cannot be identified. In [4], present a rule based approach which computes advanced decrease support using RHS rule and remove/reinsert with LHS rule. The remove reinsert with LHS rule improves the performance of privacy preservation. Similarly, the side effect on privacy preservation has been reduced with HMAU (hiding missing artificial utility) approach which sanitizes the sensitive items by deleting set of transaction with higher ratio. The transactions with higher ratio and sensitive items are removed. In [6], a privacy preservation algorithm is presented which sanitize the rules with higher sensitivity. The method never changes the support of any item but alters the items support by using distortion technique. In [6], a sequential pattern mining algorithm for privacy preservation is presented which uses indexing technique. The method uses equivalent form in reducing crypto operations. The privacy preservation in social network is presented in [8], which uses learning automata. The method detects the community using learning automata theory. The method cluster the data by dividing the nodes under different hierarchy like graph where min cut algorithm is used. In [9], a privacy preservation algorithm is presented which uses identity of users. The method has tested over E-health care. The method uses identity-Based and non-interactive key management the privacy preservation is performed using bilinear paring In [10], the author presents a multi dimension model which provides higher flexibility. The method uses a greedy algorithm to approximate with simple way. In [11], the author presents a discrimination prevention approach towards mitigating discrimination attack performed by a group of people. The method mitigates group threat and reduces the distortion In [12], a discrimination prevention technique is presented which handle the direct and indirect discrimination. The method uses outsourced data in generating nondiscriminatory rule to mitigate in direct and indirect rules. Similarly, [13] discuss the approach for privacy preservation. The method uses taxonomy in the preservation of user data. In [14], the author presents a discrimination prevention technique towards crime detection. The author discuss different cleaning algorithm to perform cleaning and uses outsourced datasets in such a way that legitimate classification rules can still be extracted but discriminating rules based on sensitive attributes cannot. All the above discussed methods has produces poor results in sanitization and discrimination.

Feature Correlation and Impact Measure Based III.

III. PREPARE YOUR PAPER BEFORE STYLING

The proposed feature correlation and impact measure based discrimination approach reads the input transactional data set and remove noise based on the presence and absence of features and values. The noise removed data has been used for the generation of pattern. For each pattern from the set, the method estimates the Feature impact measure and select set of patterns with higher feature impact measure. Similarly, the method computes the feature covariance measure using social data and the pattern. Finally, discrimination is performed using probabilistic approach. The detailed approach is discussed in this section.

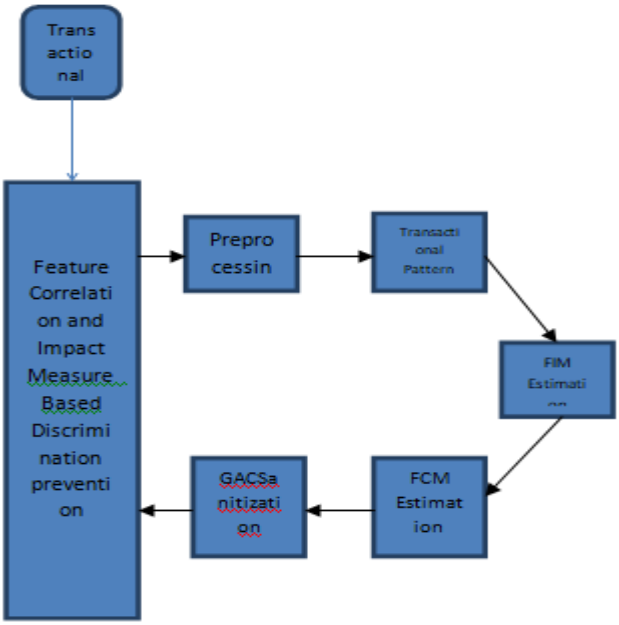

Fig 1: Architecture of Proposed Discrimination Prevention Technique

The Fig 1, present the architecture of proposed discrimination prevention technique and shows various stages involved.

A. Preprocessing

The input transactional data set has been taken and the list of features present in the data set has been identified. Based on identified features, the method identifies the presence of all the features for each data record. The records found incomplete has been eliminated from the data set and has been used to generate transactional patterns. Generated patterns has been used to perform discrimination prevention.

Preprocessing Algorithm:

Input: Transactional Data set Tds
Output: Preprocessed Data Set Pds
Start
Read Tds.
Initialize Attribute List Atl.

Find the list of attributes present in data set Tds

Atl = $\int_{i=1}^{size(Tds)} Atl \cup (\sum Tds(i).Attr \ni Atl)$

For each record Ri

If $\int_{i=1}^{size(Atl)} if\ Atl(i) \in Ri \&\& RiAtli == Null$ then

Eliminate the data point from data set.

Else

Add to preprocessed data set.

Pds = $\sum Rk(Prds) \cup Ri$

End

End

Stop

The working of preprocessing algorithm has been presented and it removes the noise records from data set to generate transactional patterns.

B. Transactional Pattern Generation

The pattern generation is performed with the use of transactional data set being preprocessed. At each number of item sets, the method generates pattern with varying items. The combinatory has been used to generate possible patterns. It has been performed at 1 to N number of pattern where N decides the total number of items. Generated patterns have been used for discrimination prevention.

C. FIM Estimation

The feature impact measure representthe impact of feature in the pattern set. The impact measure has been measured according to the presence of feature in different patterns of transaction. It has been measured for all the patterns transaction. Based on the impact measure, the method selects a sub set of features which are having higher impact in various classes. Identified features are used for sanitization process.

Algorithm:

Input: Preprocessed Data set Pds, Pattern Set Ps

Output: Feature ListFl

Start

Read Preprocessed data set Pds and pattern set Ps

For each pattern p

Identify list of all features PFl = $\int_{i=1}^{size(p)} \sum Feature(P(i)) \ni PFl$

Identify list of transactions from Pds.

Ptl = $\int_{i=1}^{size(Pds)} \sum Pds(i).Pattern == P$

For each feature f

Compute number of occurrence Noc = $\frac{\int_{i=1}^{size(Ptl)} \sum Ptl(i) \in f}{size(Ptl)}$

Compute Number of Occurrence in othersNooc = $\frac{\int_{i=1}^{size(Pds)} \sum Pds(i) != P \&\& Pds(i) \in f}{size(Pds)}$

Compute Feature Impact Measure FIM = NoC×Nooc

If FIM>Th then

Add to feature list Fl = $\sum$(Features ∈ Fl) ∪ F

end

end

end

Stop

The above discussed algorithm how the feature impact measure has been estimated and how the feature selection is performed.

D. FCM Estimation

The feature covariance measure of any feature represents how it correlates with the social network data. It has been measured by analyzing the social network data and the features selected in the previous stage. For each feature selected, the method estimates the feature covariance measure, according to their presence in the conversations of social media belongs to different users. For each feature with the social data, the method estimates the FCM measure which has been used for feature selection in sanitization.

Algorithm:

Input: Social Data Sd, Feature f

Output:FCM

Start

Read social data Sd and Feature f

Compute Feature Covariance measure FCM.

FCM = $\int_{i=1}^{size(SD)} \frac{\sum SD(i) \in F}{size(SD)}$

Stop

The above discussed algorithm shows how the feature covariance measure has been estimated using the social network data available. Estimated FCM measure has been used to perform sanitization.

E. Probabilistic Sanitization

The probabilistic model has been used for sanitizing the data set. The method preprocesses the transactional set and generates set of patterns. Then the estimates Feature impact measure for each attribute. Similarly, for the attributes the method estimates the feature covariance measure for all the attributes. Finally, using these two measures, the method selects a subset of features. For the selected features, the method estimates the probability of appearance. Based on the probability value, the method performs sanitization.

Algorithm:

Input: Transactional Set Ts, Social Data Sd

Output: Publication Set Ps

Start

Read Ts, Sd.

Initialize publication set ps.

Pds = Preprocessing(Ts)

Ps = Pattern Generation(Pds)

For each pattern p

Compute frequency measure Fm = $\frac{\sum_{i=1}^{size(ps)} \sum Ts(k) \in p}{size(Ts)}$

If Fm>Th then

Leave

Else

Eliminate from Ps.

End

end

Feature List Fl.
For each feature f
FIM = Estimate Feature Impact Measure.
FCM = Estimate Feature Covariance Measure.
Compute sanitization support measure SSM.
SSM = FIM×FCM.
If SSM>Th then
Add to feature list Fl.
End
End
For each feature f from feature list

Estimate probability value pv = $\frac{\sum_{i=1}^{size(Ts)} Ts(i) \in F}{size(Ts)}$

End
Ps=Generate publishing data with Pv.
Stop

The above discussed algorithm shows how the probability based sanitization is performed and shows the different stages involved in data publishing.

IV. RESULTS AND DISCUSSION

The proposed feature impact and feature covariance measure based discrimination prevention algorithm has been implemented and measured for its performance in various parameters. The method has produced higher efficient results in all the factors considered. The results produced has been presented in this section.

Table 1: Details of Simulation

Parameter	Value
Data Set	Transactional Amazon
Number of Items	2000
No of rows	2 million
Tool Used	Advanced Java

The Table 1, shows the details of evaluation being used for the performance measure of different algorithms.

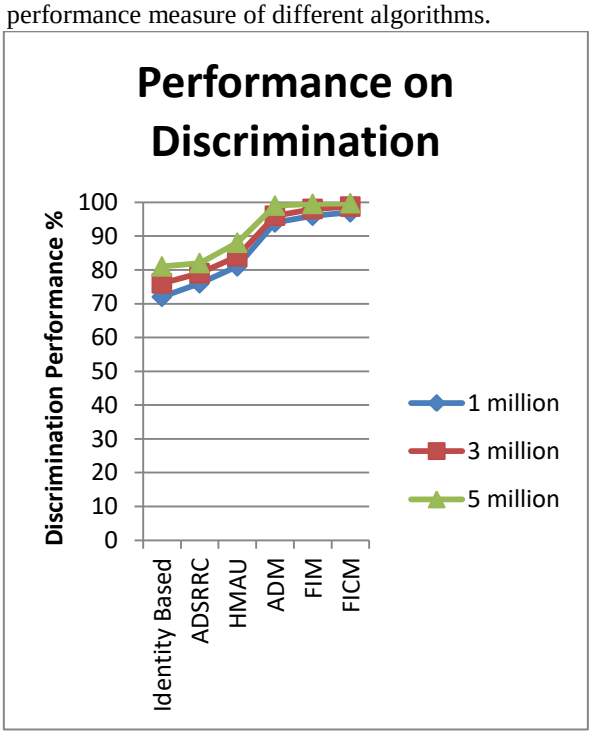

Fig 2: Performance on discrimination

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

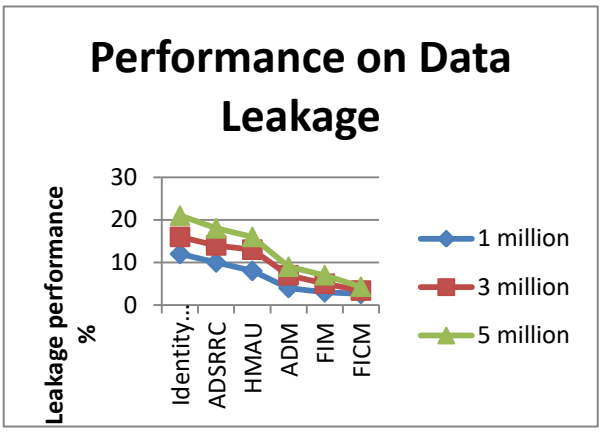

Fig 3: Performance on Data Leakage

The performance in data leakage has been measured and compared with the results of other methods. The result obtained has been presented in Figure 3, which shows the proposed method has reduced the leakage ratio than any other methods.

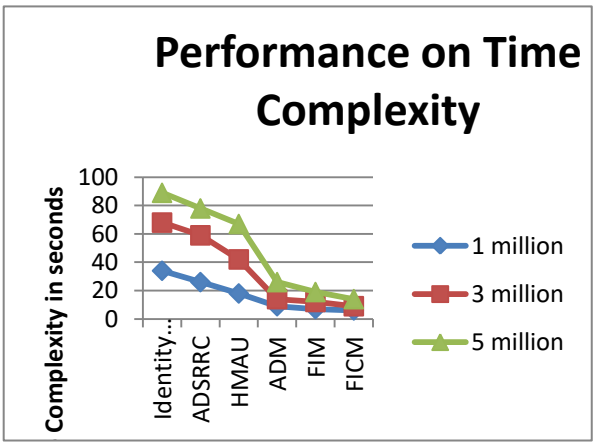

Fig 4: Performance on time complexity

The time complexity introduced by various methods has been measured and compared with the results of other methods. The proposed FICM algorithm has produced less time complexity than other methods.

V. CONCLUSION

The problem of discrimination prevention and sanitization has been well studied and different approaches available have been analyzed. To improve the performance, an efficient feature impact measure and feature covariance measure based probabilistic model has been presented. The method preprocess the transactional data to remove the noise features and records.

The noise removed data has been used to generate the transactional patterns. The patterns with higher frequency has been selected. Further, for the items present in the patterns, the method estimates feature impact and feature covariance measures. Estimated measures are used to estimate sanitization support measure which is used to select the item for sanitization.

At the publication, the method estimates the probability values for the items identified. 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 Recent Trends in Information, Telecommunication and Computing (ITC), Page(s): 342 – 344, 2010.
2. V. Thavavel, A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012
3. M Sridhar and Raveendra B Babu. A Fuzzy Approach for Privacy Preserving in Data Mining. International Journal of Computer Applications 57(18):1-5, November 2012.
4. Komal Shah, AmitThakkar and AmitGanatra. Article: Association Rule Hiding by Heuristic Approach to Reduce Side Effects and Hide Multiple R.H.S. Items. International Journal of Computer Applications 45(1):1-7, May 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, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.
7. 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.
8. 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.
9. KambomboMtonga, 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, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. 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 , Oct. 2018)
12. SaraHajian, A Methodology for Direct and Indirect Discrimination Prevention in Data Mining, IEEE Transaction on Knowledge and Data Engineering, 2013.
13. SaraHajian, 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.