International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 2, February 2017

ISSN: 2394-4404

Implementation Of Privacy Preservation Using Anonymization Methods For Discrimination Prevention

S. Swathi

PG Scholar, Department of Computer Science and Engineering,
P. A. College of Engineering and Technology, Pollachi

Abstract: ***Privacy preserving data mining (PPDM) refers to the part of data mining used to safeguard sensitive information illegal disclosure.*** **Discrimination** ***is the detrimental process of people based on their association with a certain classes or groups. Direct discrimination restricts a certain group of working class based on sensitive reasons. Indirect discrimination restricts a certain group of working class based on non sensitive ones. Both direct and indirect discrimination can be prevented using data transformation methods such as rule protection and rule generalization. Balanced iterative reducing and*** **clustering** ***using hierarchies (BIRCH) algorithm is used for analyzing discrimination datasets based on eligible criteria. In this paper, privacy can be enhanced using differentiated virtual password schemes and anonymization techniques. We provide a differentiated virtual password that applies user-specified randomized linear generation functions to protect user passwords. We provide an anonymization algorithm that processes inferring approach to prevent attacks in discrimination environment. We are evaluating these methods on Adult dataset and provide metrics for proposed methods that impact on information loss and data quality in data mining.***

Keywords: Classification, Rule protection, Rule generalization, Birch algorithm, differentiated virtual passwords, Secret little functions, Inferring algorithm, Privacy, Codebooks

I. INTRODUCTION

Nowadays, the process of hardware technology paved the way of storing large amount of transactional information. Many commercial websites ask their user's password for security purpose. Although these processes authenticated through a secured channel, it remains unprotected one. The secure protocol SSL/TLS [1] for transmitting private data over the web is well-known in academic research, but most current commercial websites still rely on the relatively weak protection mechanism of user validation through plaintext password and user ID. The attackers such as phishing, shoulder-surfing and Trojan viruses attempt to illegally obtain sensitive data, such as passwords and debit card details, by concealed as a reliable person in an electronic communication. An opponent can intrude privacy of personal details using social networks and background knowledge. A huge database may contain information about specific individuals becoming public for open government laws. Such datasets includes health histories of patients, transaction details and sensitive data paved the way of privacy risks. Even if the data such as names and security numbers are removed, the attackers can use background knowledge and combine with other datasets to identify individual information.

For example, the Centers for Disease Control want to use data mining approaches to predict patterns in disease information of people. Bank insurance process contains data that would be useful but they are not interested to reveal due to people privacy concerns. An alternative approach for these problems is to provide some statistics on data that cannot ne reveal to individual, but can be used to predict patterns to CDC. An opponent can intrude privacy of personal details using social networks and background knowledge. Discrimination involves the group's initial reaction that influencing the individual's actual behavior towards the group, restricting members of one group from privileges that are available to another group, leading to the rejection of the individual or entities based on logical decision making.

Discrimination based on age, religion, gender, caste, disability, employment, language, race and nationality. Most conventional data mining approaches are mainly focuses on protecting against disclosure individual data records. Privacy preserving data mining includes *k*-anonymity, *l*-diversity, and *t*-closeness. Border-based process that modifies the original borders in lattice of patterns based on rule hiding methods. Condensation based privacy preserving data mining that generates fake document from constrained clusters.

II. RELATED WORK

Numerous direct and indirect discrimination schemes have been proposed previously. Those schemes either eliminate direct or indirect discrimination. Fast algorithms for mining association rules that defines the issues of discovering association rules between items in a large database of sales transactions[2]. The proposed algorithms can be combined into a hybrid algorithm named as ApriorіHybrid. Toon Calders investigated that to modify the naive Bayes classifier in order to perform classification that is restricted to be independent with respect to a given sensitive attribute [3].

Phishing is the process of attempt to obtain information such as passwords and credit card details indirectly. There are two typical types of phishing. First, to prevent phishing emails [4], [5], [6], a statistical machine learning technology is used to filter the likely phishing emails; however, such a content filter does not always work correctly. Blacklists of spamming mail servers are built in [7] and [8]; however, these servers are not useful when an attacker hijacks a virus-infected PC and key distribution architecture and a particular identity-based digital signature scheme were proposed to make email trustworthy. Second, to defend against phishing websites, the authors in [9] and [10] developed some web browser toolbars to inform a user of the reputation and origin of the websites which they are currently visiting. In [11], the author presented a tricky method which can confuse a key logger, which works as follows. Instead of typing your whole password into the login field, the user changes focus outside the login form and types some random characters between any two successive password characters. However, this trick does not shield the user from key logger attacks. It only makes it slightly more difficult because it is very easy to record all the keys, mouse events, and applications of the focus.

In a survey of work, the process of providing security to statistical databases using suppression methods [12] can be explained. Agrawal[13] and Lindell[14] introduced privacy preserving data mining in their papers. They defines two concepts: Privacy preserving information gathering and mining a data set partitioned across various sources. Agrawal and Srikant define a randomization scheme that allows a huge amount of users to contribute their records for central data mining approach that limits the confession of values. Linkell and Pinkas describe a cryptographic technique for a decision tree construction on a dataset between two parties. Privacy preserving data mining includes database security, database query auditing for disclosure detection, database privacy and secure multiparty computation. Neenu Mary Kuruvila and V.Vennila[15] proposed rule protection and rule generalization methods. Differential privacy and rule privacy can provide high privacy ratio that integrated with previous privacy methods to find synergies between privacy preserving and rule hiding methods [16]. Heuristic approaches involve well-organized and fast algorithms to hide the sensitive information.

III. CLASSIFICATION BASED ON DISCRIMINATION PREVENTION USING DATA TRANSFORMATION TECHNIQUES

Classification is the task of generalizing known structures applies to new data. Classification is supervised learning. For example, classes are used to represent that a customer defaults on a loan decisions like ‘Yes’ or’ No’. Classification is a machine learning technique used to predict group membership for data instances. It assigns items in a collection to target categories. The aim of classification is to accurately determine target class for each and every case in data.

A. DISCRIMINATION MEASUREMENT AND DATA TRANSFORMATION

The purpose of Discrimination measurement is to identify discriminatory rules and redlining rules using Potentially Discriminatory (PD) and potentially non-discriminatory (PND) rules [18]. Direct discrimination is measured by identifying α-discriminatory rules among the PD rules using a direct discrimination measure (elift) and a discriminatory threshold (α).The extended lift can be calculated as

$$\mathrm{elift}(A, B \rightarrow C) = \frac{\mathrm{Conf}(A, B \rightarrow C)}{\mathrm{Conf}(B \rightarrow C)}$$

The indirect discrimination is measured by identifying redlining rules among the PND rules that correlated with background knowledge based on an indirect discriminatory measure (elb), and a discriminatory threshold (α).Transform the original data DB in such a way to remove direct and indirect discriminatory biases, with minimum impact on the datasets.

B. BIRCH ALGORITHM

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an unsupervised data mining algorithm. BIRCH can be used to perform clustering in discrimination environment. It can be used in multi-dimensional datasets and it has minimized I/O cost than Apriori algorithm (1 or 2 scans). First, it scans the data set and construct clustering feature tree in its memory. Then it condenses large clustering feature tree into smaller one and performs global clustering by using its centroid points [20]. Finally it does cluster refining one more time for removing outliers

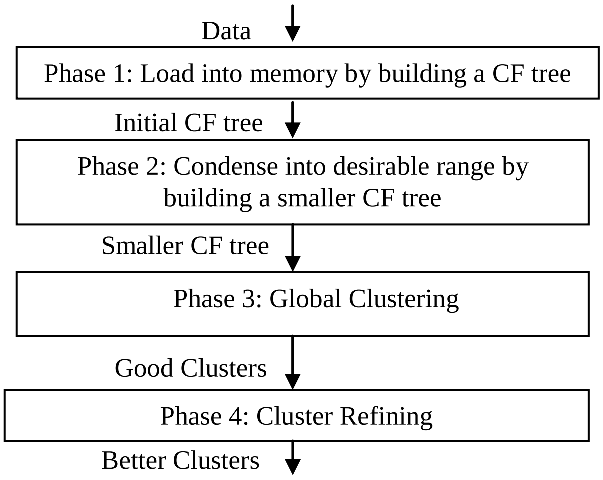
Data
Phase 1: Load into memory by building a CF tree
Initial CF tree
Phase 2: Condense into desirable range by building a smaller CF tree
Smaller CF tree
Phase 3: Global Clustering
Good Clusters
Phase 4: Cluster Refining
Better Clusters

Figure 1: Birch Algorithm

BIRCH algorithm can be divided into two phases: It scans the transformed data set in memory and generate model based on eligible and not eligible criteria as shown in Figure 7.The process of BIRCH algorithm is explained in Figure 1.

C. DIFFERENTIATED VIRTUAL PASSWORDS

A virtual password is a scheme that cannot applied directly, it produces dynamic password that submitted to server for validation. A dynamic password consists of two parts such as function B and fixed alphanumeric F form the domain ψ to ψ, where ψ is letter space used for passwords In the Registration Module, and the users have to make registration here [19]. As per the registration a jar will be downloaded as per the random value.

User has to install the jar in the java supporting mobile. In the jar there will be expression calculation. Expression varies for each jar. Expression will be stored in the database. In the login form the user will give the user name and password first. If the username and password is same, the random key will be sent to the access page. User has to install the jar and enter the random key contain in access page. As per the user expression calculation will be done and viewed in the access code text field.

D. SECRET LITTLE FUNCTIONS

In modern ciphers, encryption algorithms are open to the public but keys of these algorithms are kept secret. One reason that modern ciphers seldom choose secret encryption algorithms is that secret encryption algorithms prevent communication among parties such as commercial products, networking protocols, and so on. Therefore, the approach in which only keys are kept as secrets and algorithms are open to the public for implementation is very popular in modern ciphers. The reason behind using user specified programs is that information is kept very secret and cannot know by others.

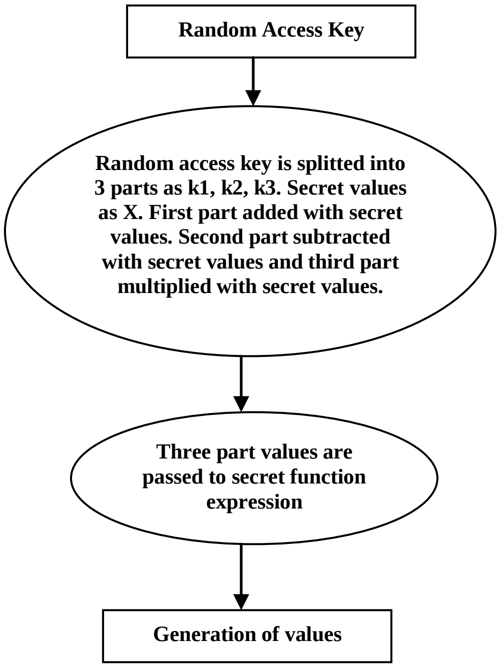
Random Access Key
Random access key is splitted into 3 parts as k1, k2, k3. Secret values as X. First part added with secret values. Second part subtracted with secret values and third part multiplied with secret values.
Three part values are passed to secret function expression
Generation of values

Figure 2: Flow chart for user defined program

User specified functions can be unbounded and the attackers do not know the function forms that are much protected as shown in Figure 2. The condition is that secret little functions should use the random number provided by the server; otherwise, it may be subject to Key logger attacks since the attackers do not need to know the function but can simply input the same capture inputs again to gain access.

Figure 3 defines helper application for a mobile device. There will be 11 jars the secret value and secret Function will vary for each jar. Calculation Part in the Secret Little Function module is as Follows: The access code values will be split into 3 parts and split the value in 3 parts and assign to the 3 variables such as a, b, c. Then a will be added with X variable b will be subtract with x variable and c will be multiplied with x. Here x value will vary for each jar. Assign the value as a1, b1, c1. Secret Function will vary for each user. The expression calculation will be in a1 b1 c1 format only.

The values will be passed to the expression and generated code will be generated.

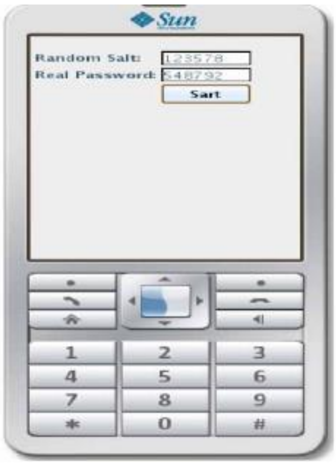
Random Salt: 123578
Real Password: 548792

Figure 3: Helper application for cell phone

E. CODEBOOKS

A codebook should be small that can accumulate in a PDA and easy to carry. It is not possible that the use remembers the whole codebook the server should have effective computing power to run Random Number Generator (RNG), so if user loses their codebook they can use new one without changing parameters. Linear Congruential Generators are not possible in this environment.

A codebook is a type of text used for collecting and accumulates codes in it. In the setup part, the user assigns the length of the password, n and then server generates n 64-digit random numbers. The server generates four random numbers; R0, R1, R2, and R3 with each have 64-digits. Let r(i,0), r(i,1), r(i,2).. r(i,63) denote the 64- digits of Ri. ACT is a code protection scheme used for sensors to validate a transmit message sender in networks based on hash function. ACT generates chain-key that stored in codebook. The user's codebook consists of hidden password, constant value and the user specified function, authentication decrypting key and ACT keys.

F. ANONYMIZATION APPROACH

In anonymization approach, when identifiers are linked with public available data, individual patterns can be identified with higher probability is also known as linking attacks.

Using generalization and suppression, this technique conceals sensitive information about record owners.

Appln Id	Age	Income	Nationality	Loan applied for	Loan amount
10000	A1	I1	N1	Business loan	2500000
10001	A1	I1	N1	Business loan	3000000
10002	A1	I1	N1	Business loan	2500000
10003	A1	I1	N2	Car loan	6000000
10004	A1	I1	N1	Personal loan	2500000
10005	A1	I1	N1	Business loan	2400000
10006	A1	I1	N1	Business loan	1600000

Figure 4: Direct Discrimination

Appln Id	Education	Property	Savings	Loan applied for	Loan amount
10000	Q4	A1	S1	Business loan	2500000
10001	Q4	A1	S1	Business loan	3000000
10002	Q0	A1	S1	Business loan	2500000
10003	Q4	A1	S2	Car loan	6000000
10004	Q4	A1	S1	Personal loan	2500000
10005	Q4	A1	S1	Business loan	2400000
10006	Q4	A1	S1	Business loan	1600000

Figure 5: Indirect Discrimination

Appln Id	Direct	Indirect	Total	Inference	Loan applied for
10000	3	3	6	1	Business loan
10001	3	3	6	1	Business loan
10002	3	3	6	1	Business loan
10003	2	1	3	0	Car loan
10004	3	3	6	1	Personal loan
10005	3	3	6	1	Business loan
10006	3	3	6	1	Business loan

Figure 6: Inferring approach

Such data when released for mining that reduces risk of identification even if the data linked with public, but it reduces transformation accuracy. Figure 4 describes the direct discrimination that represents age as A1, income as I1 and Nationality as N1. Figure 5 describes the indirect discrimination that represents education as Q1, property as A1 and Savings as S1. Customer will registered their personal details regarding loan form. The manager will verify direct and indirect discrimination based on inferring approach for preprocessing. It classifies loan applicants based on eligible and not eligible criteria.

IV. RESULTS

Net Beans is an integrated development environment (IDE) for developing primarily with Java. Using secret little functions, phishing and other types of attacks can be defeated. The dynamic password approach is processed in mobile-application using ACT and after entering random number in sign-on screen, virtual password can be calculated. In this approach, the bank will send access code to user after entering dynamic password.

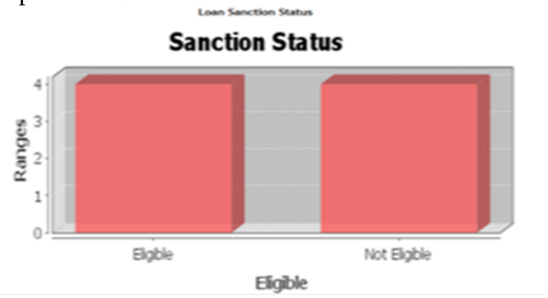

Figure 7: Loan sanction status

ADULT DATA SET: We used the Adult data set [17], also known as Census Income, in our experiments. Adult data set consists of 48,842 records, split into a "train" part with 32,561 records and a "test" part with 16,281 records. The data set has 14 attributes. The prediction task associated with the Adult data set is to determine whether a person makes more than 50K$ a year based on census and demographic information about people.

V. CONCLUSION

The purpose of this paper is to extend discrimination prevention methodologies and to enhance privacy using anonymization and differentiated virtual passwords to safeguard sensitive information. We anticipated a differentiated security mechanism that allows the user to choose dynamic password that submitted to server for authentication. In user specified programs, secret little functions can be used to improve protection by hiding secret functions. Anonymization methods such as suppression and generalization can be used to protect data against linking attacks. The experimental results reported demonstrate that the proposed techniques are efficient in both goals of removing discrimination and preserving data quality.

REFERENCES

[1] T. Dierks and C. Allen. The TLS Protocol— Version 1.0. IETF RFC 2246, January 1999.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.

[3] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification", Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.

[4] J. Mason, "Filtering spam with SpamAssassin," in Proc.HEANetAnnu. Conf., 2002.

[5] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering june e-mail. In learning for text categorization," in Proc. Workshop, May 1998.

[6] T. A. Meyer and B. Whateley, "SpamBayes: Effective open-source, Bayesian based, e-mail classification system" in Proc. CEAS, 2004.

[7] MAPS. (1996). RBL—Realtime Blackhole List [Online]. Available and Phishing Attacks, Cryptology ePrint Archive, Rep.2004/155 [Online]. Available: http://eprint.iacr.org/2004/155

[8] The Spamhaus Project. The Spamhaus Block List [Online]. Available http://www.spamhaus.org/sbl

[9] Herzberg and A. Gbara. (2004). Trustbar: Protecting (Even Naive) Web Users from Spoofing.

[10] Net craft. AntiPhishing Toolbar, http://www.mail-abuse.com/services/mds−rbl.html

[11] C. Herley and D. Florencio, "How to login from an Internet cafe without worrying about keyloggers," in Proc. SOUPS, 2006.

[12] Adam, N. R. & Wortmann, J. C, "Security-Control Methods for Statistical Databases: A Comparative Study", ACM Computing Surveys, Vol. 21, N. 4, pp.515–556,1989.

[13] Agrawal, R. & Srikant, R," Privacy Preserving Data Mining", In Proc. of ACM SIGMOD Conference on Management of Data (SIGMOD'00), 2000.

[14] [14] Lindell, Y. & Pinkas, B, "Privacy Preserving Data Mining". In Proc. of Advances in Cryptology – Crypto'00, LNCS 1880, Springer-Verlag, pp. 20–24, 2000.

[15] Neenu Mary Kuruvila, V. Vennila, "Privacy Preservation Using Discrimination Prevention Methods in Data Mining", In Proc of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, 2013.

[16] S. Subbulakshmi, B. Arulkumar, Syed Farmhan.S, "Survey on Discrimination Techniques for Privacy Preserving Data Mining", In Proc of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 11, 2013.

[17] R. Kohavi and B. Becker, "UCI Repository of Machine LearningDatabases",http://archive.ics.uci.edu/ml/datasets/Adult,1996.

[18] Sara Hajian and Josep Domingo-Ferrer , "A Methodology for direct and indirect discrimination in data mining", Knowledge and Data Engineering, vol. 25, no. 7, pp. 1445-1459, 2013.

[19] Yang Xiao, Chung-Chih Li, Ming Lei, and Susan V. Vrbsky,"Differentiated Virtual Passwords, Secret Little Functions, and Codebooks for Protecting Users From Password Theft", Systems Journal, IEEE (Volume: PP, Issue: 99),2012.

[20] Rui Liu, Xiao-long Qian and Shu Mao," Research on Anti-Money Laundering Based on Core Decision Tree Algorithm", Control and Decision Conference (CCDC), pp .4322 – 4325, 2011.