ISSN 2347-4289

An Efficient Method For Discrimination Prevention Using Differentiated Virtual Passwords And Secret Little Functions

Swathi.S

Dept. of Computer Science and Engineering, P.A College of Engineering and Technology, Coimbatore, Tamilnadu;
Email: laxmiswathiks@gmail.com

ABSTRACT: In classification, discrimination is a type of treatment that includes denying the membership in one group opportunities that are available in another group. Discrimination based on age, religion, gender, caste, disability, employment, language, race and nationality. In this technique, direct and indirect discrimination is prevented using rule protection and rule generalization methods and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is used to perform hierarchical clustering for huge data-sets. To enhance privacy, we propose a virtual password concept to secure user's passwords in banking process. We applied user-determined randomized linear generation functions to secure user's passwords based on the fact that a server has more information than any adversary does. We are evaluating metrics for proposed methods that impact on information loss and data quality in data mining.

Keywords : Classification, Direct discrimination, indirect discrimination, data transformation, differentiated virtual passwords, Code books, Secret little functions.

I. INTRODUCTION

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. In the beginning, automating decisions may give a sense of fairness but the decision rule does not learn itself by personal preferences. The classification rules are actually learned by the system model based on historical data. If the original data are inherently biased against a particular community, the learned model also shows the negative impact on it. Users with important accounts on the Internet face many kinds of attacks, e.g., a user ID and password can be stolen and misused. 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. Meanwhile, even though a password can be passed through protected channel, this approach is still susceptible to attacks as follows. Phishers attempt to illegally obtain sensitive data, such as passwords and debit card details, by concealed as a reliable person or business in an electronic communication. Password Stealing Trojan is a program that contains or installs malicious code. There are many such Trojan codes that have been found online today, so here we just briefly introduce two types of them. Key loggers confine keystrokes and accumulate them in the machine, or drive them back to the opponent. Once a key logger program is activated, it provides the opponent with any strings that a person might enter online, consequently placing individual data and online account details at risk. Trojan Redirector was designed to redirect end-users network traffic to a location to where it was not intended. This includes crime ware that changes hosts files and other DNS specific information, crime ware browser-helper objects that redirect users to counterfeit sites, and crime ware that may install a network level driver or filter to redirect users to fraudulent locations. Shoulder surfing is a method of stealing other's passwords and other sensitive personal information by looking over victims' shoulders while they are sitting in front of terminals. Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. Most traditional data mining techniques analyze and model the data set in aggregation, while privacy preservation is primarily concerned with protecting against disclosure individual data records. This area partition points to the technical feasibility of PPDM. The cryptographic approach to PPDM assumes that the data are stored at several private parties who agree to disclose the result of a certain data mining computation performed jointly over their data.

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 AprioriHybrid. 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]. Preferential sampling introduced the idea of Classification with No Discrimination (CND) [5] and proposed a solution based on "massaging" the data to remove the discrimination from it with the least possible changes [6]. For a survey of work in statistical databases, see Adam and Wortmann (1989) and Willenborg and de Waal (2001).The term privacy-preserving data mining was introduced in the papers Agrawal and Srikant (2000) and Lindell and Pinkas (2000). These papers considered two fundamental problems of PPDM: privacy-preserving data collection and mining a data set partitioned across several private enterprises. Agrawal and Srikant devised a randomization algorithm that allows a large number of users to contribute their private records for efficient centralized data mining while limiting the disclosure of their values; Lindell and Pinkas invented a cryptographic protocol for decision tree construction over a data

set horizontally partitioned between two parties. Phishing attacks are relatively new but very effective. There are two typical types of phishing. To prevent phishing emails [8], [9], [10], a statistical machine learning technology is used to filter phishing emails but the content filter does not always work correctly. Blacklists of spamming mail servers are built in [11] and [12]; 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 [13] and [14] developed some web browser toolbars to inform a user of the reputation and origin of the websites which they are currently visiting. In [15], 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. It only makes it slightly more difficult because it is very easy to record all the keys, mouse events, and applications of the focus. They used a dynamic pad for the login system, which allows a user to click the dynamic keyboard on the screen instead of typing on the physical keyboard, but such a dynamic keyboard faces some of the same problems as previous methods. Alphanumeric password systems are simply attacked by this approach, in which an opponent can record the user motions by a hidden camera when the user types in the password. In [16], the authors adopted a game-like graphical method of authentication to combat shoulder-surfing; it requires the user to pick out the passwords from hundreds of pictures, and then complete rounds of mouse-clicking in the Convex Hull.

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'. It is important that each record in the dataset used to represent the classifier already have a value for the attribute used to describe classes. Because each record has the attribute value used to define the classes. 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

$$Elift(A, B-> C) = \frac{Conf(A, B-> C)}{Conf(B-> C)}$$

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 [4].

B. BIRCH ALGORITHM

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a data mining algorithm used to perform clustering in discrimination environment. It can be used efficiently in multi-dimensional datasets and it has minimized I/O cost than Apriori algorithm (1 or 2 scans).

1. First, it scans the data set and construct clustering feature tree in its memory as shown in Figure 1.
2. Then it condenses large clustering feature tree into smaller one and performs global clustering by using its centric points.
3. Finally it performs cluster refining for removing outliers.

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 4.

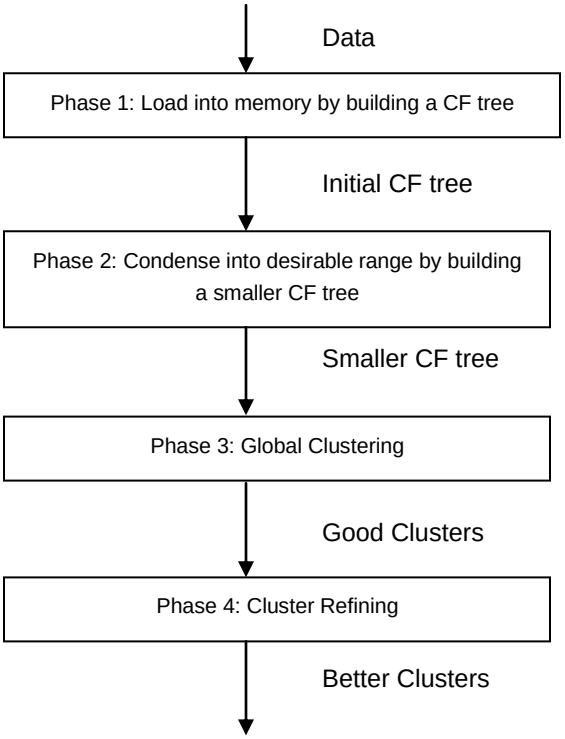

Figure 1: Birch algorithm

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 [19]. The functions are P= (F, B) and B (F, R) = Pd, R is a random number and Pd is a dynamic password where B is a virtual function. The user input consists of (ID, Pd), where ID is user ID. In server side, the server cal also evaluates Pd and compares it with password. First the server will find user information based

on ID and compute Pd. The bijective function allows the system to use reverse function to remove password.

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 are kept very secret and cannot know by others.

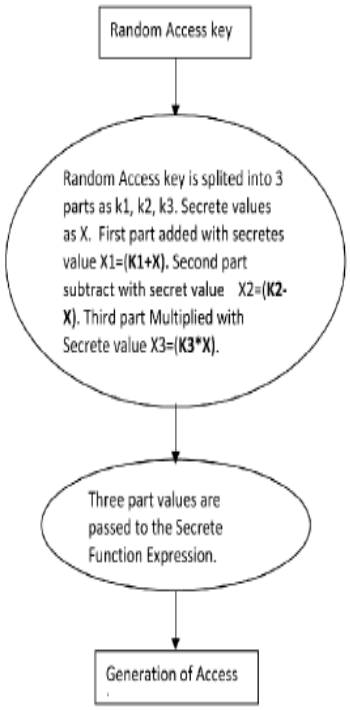

Figure 2: Flow chart for user defined program

User specified functions can be unbounded and the attackers do not know the function forms that are much protected as shown in Figure 2. Otherwise, it would be easy to attack these functions. These simple and protected functions are named as secret little functions. It is a problem that extra effort is needed to program the function into server for creating an account. 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.

Figure 3: Helper applications 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 [20] 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.

IV. RESULTS

NetBeans 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. During registration, human computing is needed.

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.

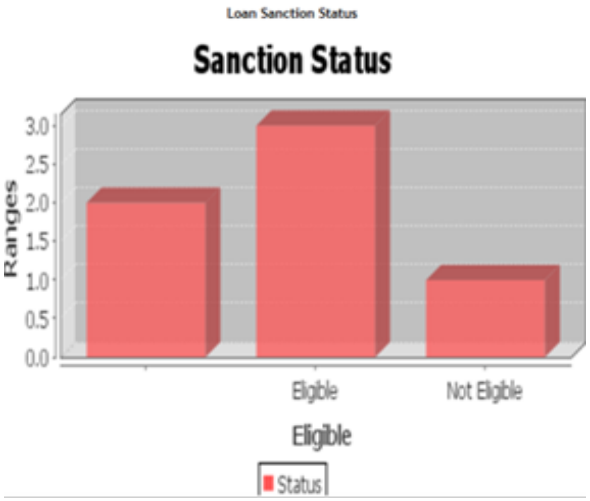

Figure 4: Loan sanction status based on eligible and not eligible criteria

V. CONCLUSION

The purpose of this paper was to develop a new preprocessing discrimination prevention technique that consists of various data transformation methods includes rule protection and rule generalization used to prevent direct discrimination and indirect discrimination and provide privacy. We anticipated differentiated security mechanisms that the user can choose virtual password method. It performs tradeoff between security and complexity and less human computing. In user specified programs, secret little functions can be used to improve protection by hiding secret functions. 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] S. Hajian, J. Domingo-Ferrer ,“Rule Protection for Indirect Discrimination Prevention in Data Mining”, Proc. Eighth Int’l Conf. Modeling Decisions for Artificial Intelligence (MDAI ’11), pp. 211-222, 2011.

[5] F. Kamiran and T. Calders, “Classification without Discrimination”, Proc. IEEE Second Int’l Conf. Computer, Control and Comm.(IC4 ’09), 2009.

[6] F. Kamiran and T. Calders, “Classification with no Discrimination by Preferential Sampling”, Proc. 19th Machine Learning Conf. Belgium and The Netherlands, 2010.

[7] F. Kamiran, T. Calders, “Discrimination Aware Decision Tree Learning”, Proc. IEEE Int’l Conf. Data Mining (ICDM ’10), pp. 869-874, 2010.

[8] J. Mason, “Filtering spam with SpamAssassin,” in Proc.HEANetAnnu. Conf., 2002.

[9] 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.

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

[11] 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

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

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

[14] Net craft. Anti PhishingToolbar,http://www.mail-abuse.com/services/mds−rbl.html

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

[16] S. Wiedenbeck, J. Waters, L. Sobrado, and J. Birget, “Design and evaluation of a shoulder- surfing resistant graphical password scheme,” in Proc. Working Conf. Adv. Vis. Interfaces.

[17] R. Kohavi and B. Becker, “UCI Repository of Machine Learning Data bases”, 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] 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.

[20] A. Perrig, R. Szewczyk, J. D. Tygar, V. Wen, and D. E. Culler, “SPINS: Security protocols for sensor networks,” Wirel. Netw. vol. 8, no. 5, pp. 521–534, 2002.