Volume 4, Issue 9, September 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Deformed Identity Crime Detection of Online Credit Application as A First Stage of Credit Life Cycle

[1]Sohini Bhattacharya Chakraborty, [2]M. Z. Shaikh
[1]ME Student (Dept of Comp Sc), Bharati Vidyapeeth College of Engineering, Navi Mumbai, India
[2]Principal, Bharati Vidyapeeth College of Enggineering, Navi Mumbai, India

Abstract- Now a day's identity crime is well known, prevalent and prominent in our society. The fraud of credit application is a specific case of identity crime. There are some algorithms which have been implemented to detect or resolve resilience identity. To address the limitation of existing algorithm and to detect and combat identity crime in real time this paper proposes a new multilayer detection system complemented with two additional layer click detection and deformation of spike detection. Click finds real social relationship to reduce the suspicion score with fixed set of attribute. Deformation of spike detects spikes in duplicates to increase the suspicion score and finally rejects the invalid entries. Although multilayer mining algorithm is specific to credit application fraud detection, but the concept of resilience or deformation are general to design, implement and evaluate of all detection system to detect and remove fraud identities.

Keywords - Data mining based fraud detection, security, anomaly detection, synthetic data, data stream mining.

I. INTRODUCTION

Identity crime is well known prominent and very important in our society. To some extent, synthetic identity fraud Refers to the use of plausible but fictitious identities. These are easy to create but more difficult to apply. At one extreme, real identity theft refers to the illegal use of innocent people's complete identity Details. These can be harder to obtain (although large volumes of some identity data are widely available on web) but easier to efficiently apply. In reality, identity crime can be committed with a mix of both synthetic and real identity details. Identity crime has developed into a further numerous approaches as there is so much real identity data available on the net and private data available through unsecured mailboxes. It has furthermore developed into straightforward for fraudster to conceal their true identities. This can happen in credit cards, and telecommunications fraud with other more serious crimes. Credit applications are Internet or paper-based forms with written requests by potential customers for credit cards, mortgage loans, and personal loan. In case of duplication of data, similarity (or matches) refers to applications which share common values. There are two types of duplicates: exact (or identical) duplicates have the all same values; near (or approximate) duplicates have some same values (or characters), some similar values with slightly altered spellings, or both. This paper has studied that each successful credit application fraud pattern is represented by a sudden and sharp spike in duplicates within a short time, relative to the established baseline level. Duplicates are hard to avoid from fraudsters' point of-view because duplicates increase their success rate. The synthetic identity fraudster has low success rate, and is likely to reuse fictitious identities which have been successful before. The identity thief has limited time because innocent people can discover the fraud early and take action, and will quickly use the same real identities at different places. In this method click detection finds real social relationships to shrink the suspicion score, It is the white list-oriented approach on a fixed set of attributes. Deformation of spikes finds spikes in duplicates to enhance the suspicion score to detect and reject duplicate identities. It is the attribute-oriented approach on a variable-size set of attributes.

II. LITERATURE SURVEY

Many individual data mining algorithm has been designed, implemented and evaluated in fraud detection analysis. There is some pattern in identity crime which can be highly indicative in early symptom in identity fraud especially in synthetic identity crime [3]. In this scheme [14] has ID score risk which gives a combined view of each credit application's characteristics and their similarity to other industry. In another example, it can be detected the application of fraud prevention system [7]. But case based reasoning (CBR) is the only known prior publication in the screening of credit application [8]. My proposed approach which monitors the significant increase or decrease in amount of something important is similar in concept to credit transactional fraud detection and bio terrorism detection. In case of fraud detection peer group analysis [2] monitors inter account behavior over time. It compares the cumulative mean weekly amount between a target account and other similar accounts (peer group) at subsequent time points. Bayesian Network [4] uncovers simulated anthrax attack from real emergency department data. Surveys algorithms [5] are used for finding

suspicious activity in time for disease outbreaks. [9] Uses time series analysis to track early symptoms of synthetic anthrax outbreaks from daily sales of retails medication .Control chart based statistics, exponential weighted moving averages and generalized linear models were tested on the same bio terrorism detection of data and alert rate [15]. In addition my proposed algorithm suspicion score detection is similar to change point detection in bio surveillance research, which maintains the cumulative sum (CUSUM) of positive derivation from the mean [13].In the real-time credit application fraud detection domain, this paper argues against the use of classification (or supervised) algorithms which use class labels. In addition to the problems of using known frauds, these algorithms, such as logistic regression, neural networks, or Support Vector Machines (SVM), cannot achieve scalability or handle the extreme imbalanced class [11] in credit application data streams. As fraud and legal behavior changes frequently, the classifiers will deteriorate rapidly and the supervised classification algorithms will need to be trained on the new data. For detection of the credit transactional and application fraud analysis the following classifier methods are

□ □ □ □ □ □ □ □ □ □ Neural network and Bayesian Network

- Logistic regression
- Decision tree
- Support vector Machine
- CBR analysis

For the credit application domain Logistic regression, neural networks, or Support Vector Machines (SVM), cannot achieve scalability or handle the extreme imbalanced class [11] in credit application data streams. As fraud and legal behavior changes frequently, the classifiers will deteriorate rapidly and the supervised classification algorithms will need to be trained on the new data. But the training time is too long for real-time credit application fraud detection because the new training data have too many derived numerical attributes and too few known frauds. Many individual data mining algorithms have been designed, implemented, and evaluated in fraud detection. Following are the data mining techniques.

III. PROPOSED SYSTEM

The proposed system detects new multilayer mining stage of defense complemented two additional layer clique detection and deformation of spike detection. Click finds real social relationship to reduce suspicion score. This reduces false positives by lowering some suspicion scores with fixed set of attribute. Deformation of spike finds duplicate set of entries to increase the suspicion score with variable attribute. Throughout this paper, data mining is defined as the real-time search for patterns in a principled (or systematic) fashion. These patterns can be highly indicative of early symptoms in identity crime, especially synthetic identity fraud.

IV. OBJECTIVE OF PROPOSED METHOD

The problem statement of this proposed method is synthetic, resilient, duplicated deformed identities will be detected through proposed algorithm and will be rejected also through proposed method. There are three types of objective are achieved through this method. Firstly resilience means reinforcement of data from deformed position to original position. Secondly adaptivity of data, means the account of morphing the fraud changes of legal behavior where the synthetic legal behavior are found in CD and the score value of spike in duplicates are found in SD algorithm. Thirdly the quality of data means the reliability and application efficiency of data can be improved through filtering of noisy and error or duplicated data. So for better account of changing fraud changes of legal behavior this multilayer mining stage of defense is very helpful and prominent.

V. METHODOLOGY

This section is divided into four subsections to systematically explain the clique detection algorithm (first two subsections) and the suspicion score detection algorithm (last two subsections). Each subsection commences with a clearer discussion about its purposes. There are two types of algorithm implemented click detection with some input parameter to reduce suspicion score by applying a threshold value with set of fixed attribute and deformation of spike value by increasing the suspicion score to detect and reject the duplicated spike value.

The following steps maintains a system architecture to get the transparent view of overall system with two module, user and Admin module. The user authentication, credit card request form, identity analysis, application acceptance are coved in user module and acceptance or rejection of application will be covered in admin module. In my sql database the user details and application details will be captured. So accordingly we will discuss the algorithm details of clique detection and deformation of spike detection through the successive path.

According to the basic objective of the system, there are two module user and admin module, complemented with one module that is active verification of the user. Means if the user login the system, then after proper valid authentication the user will be able to allow entering the credit application request, After proper valid authentication the legitimate user will be allowed to complete the process. But in case of fraudster having duplicate entries, then database will go to the admin and the algorithm will detect the valid user and invalid user having suspicion score value. In case of active verification module the algorithm will detect the duplicate entries or invalid entries in user module also and the applicants can also see the alert messege.

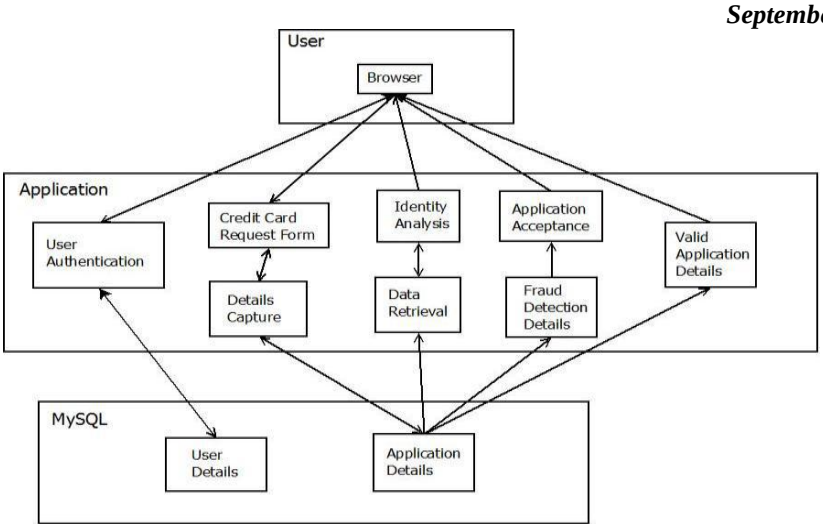

Fig1: System architecture of proposed system

Inputs
vi (current application)
W number of vj (moving window)
Y_link-type (link-types in current whitelist)
S_ similarity (string similarity threshold).
S attribute (attribute threshold)
η (exact duplicate filter)
S input (input size threshold)

Outputs
Sc(vi) (suspicion score)
Same or new parameter value
New whitlist.

5.1 Clique detection basic algorithm layout
Step 1: Multi-attribute link establishment [match vi against W number of vj to determine if a single attribute exceeds S_similarity; and create multi-attribute links if near duplicates' similarity exceeds T attribute or an exact duplicates' time difference exceeds η]
Step 2: Single-link score value [calculate single-link score by matching Step 1's multi-attribute links against _a,link-type]
Step 3: Single-link average previous score formation [calculate average previous scores from Step 1's linked previous applications]
Step 4: Multiple-links score formation [calculate Sc(vi) based on weighted average (using α) of Step 2's link scores and Step 3's average previous scores]
Step 5: Whitelist change [determine new whitelist at end of the result].

5.2 Suspicion score detection method
Inputs
vi (current application)
W number of vj (moving window)
t (current step)
S similarity (string similarity threshold)
θ (time difference filter)
α (exponential smoothing factor)

Table 1.1 C Deetction analysis

Name	Surname	DoB	Mobile	Email	Address	Pin code	Suspicion score
User2	User2	19/09/1981	9433214566	User2@yahoo.com	india	700891	1
User2	User2	12/07/1985	8991234390	ruma@gmail.com	India	790560	1
Aruna	Yadav	13/06/1980	9012045534	aruna@gmail.com	India	710234	0

Saheli	Ray	11/05/1985	9833417 34	saheli@yahoo.com	india	700045	3
Saheli	Ray	11/05/1985	9545537 89	Sims@gmail.com	india	720344	3

Table 2.2 S Detection Analysis

Name	Surname	Pancard	D.licence	Voter id	Time	Suspicion score
Salendra	Yadav	SDRTG6787C	FGTYT87566	BHG8977854	19.07.42	6
Raja	Sahani	SDRTG6787C	BGRTD34912	VGF7866721	11.09.19	6

Outputs
Black list

5.3 Deformation of spike detection algorithm
Step 1: Single-step scaled counts measurement [match vi against W number of vj to determine if a single value exceeds S_similarity and its time difference exceeds θ]
Step 2: Single-value spike or deformation detection [calculate current value's score based on weighted average (using α) of t Step 1's scaled matches]
Step 3: Multiple-values score [calculate S(vi) from Step 2's value scores and Step 4's wk]
Step 4: Suspicion score attributes selection [determine wk for Spike at end of the result]
Step 5: Territory attributes weights change [determine wk for Territory at end of the result]

5.4 Analysis and discussion of this method
Here Experiment is carried out on Tomcat application server5.0.Here server side script is Java server page and scripts are java script. All data are stored in MySql database. The front end application is HTML, Java, Jsp, XML and database connectivity is JDBC. There are ten no of attributes are taken and Pan card, voter id , D.licence are set as highest priority accordingly. Email id, mobile are set as high priority in C detection and name, surname ,DoB, address are set as les priority accordingly. Six attribute are set as C detection method and four attribute such as Pan card, voter id, D license are set as S detection algorithm. The priority of attributes depends upon database to database. Identity data - Real Application Dataset (RADS)Substantial identity crime can be found in private and commercial databases containing information collected about customers, employees, suppliers, and rule violators. If there are three no of attribute same having less priority, that will remain in C detection, otherwise if more than three no of attribute become same that will go for S detection, including Pan card, Voter id and D.licenceFor C detection there are sis no of fixed set of attribute, it is positively valid set of entries, within this fixed set the suspicion score will remain within the threshold value, the value is 4, if it becomes greeter than 4 then the record will enter into S detection or in blacklist. If the value of Pan card or voter id or D.license become duplicate, or more than four values become duplicate the record will enter into the S detection.
For C detection there are sis no of fixed set of attribute, it is positively valid set of entries, within this fixed set the suspicion score will remain within the threshold value, the value is 4, if it becomes greeter than 4 then the record will enter into S detection or in blacklist. If the value of Pan card or voter id or D.license become duplicate, or more than four values become duplicate the record will enter into the S detection(by table1.1 and 2.2)

VI. ADVANTAGE OF PROPOSED SYSTEM

The proposed method can detect synthetic identities in user level as well as admin level with using score value with the help of threshold violation basis. Much work in credit application fraud detection remains proprietary and exact performance figures unpublished, therefore there is no way to compare the clique and SD algorithms against their leading industry methods and techniques. So in that case this is one beneficiary purpose of this proposed system.

VII. CONCLUSION

The main focus of this paper is deformation and spike of deformation find out Identity Crime Detection; in other words, the real-time search for patterns in a multi-layered and principled fashion, to safeguard credit applications at the first stage of the credit life cycle. This paper describes an important domain that has many problems relevant to other data mining research. It has documented the development and evaluation in the data mining layers of defense for a real-time credit application fraud detection system. In doing so, this research produced three concepts which increase the detection system's effectiveness (at the expense of some efficiency). But in this method has some limitation, means the scalability is a factor when huge set of data continuously increased. These concepts are fundamental to the design, implementation,

and evaluation of all fraud detection, adversarial-related detection,and identity crime-related detection systems. The implementation of CLIQUE and suspicion score algorithms is practical because these algorithms are designed for actual use to complement the existing detection system.

REFERENCES

[1] Bifet, A. and Kirkby, R. 2009. Massive Online Analysis, Technical Manual, University of Waikato.
[2] Bolton, R. and Hand, D. 2001. Unsupervised Profiling Methods for Fraud Detection, Proc. of CSCC01.
[3] Oscherwitz, T. 2005. Synthetic Identity Fraud: Unseen Identity Challenge, Bank Security News 3: p.7.
[4] Wong, W., Moore, A., Cooper, G. and Wagner, M. 2003. Bayesian Network Anomaly Pattern Detection for Detecting Disease Outbreaks, Proc. of ICML03. ISBN: 1-57735-189-4.
[5] Wong, W. 2004. Data Mining for Early Disease Outbreak Detection, PhD thesis, Carnegie Mellon University.
[6] Cortes, C., Pregibon, D. and Volinsky, C. 2003. Computational methods for dynamic graphs, Journal of Computational and Graphical Statistics 12(4): pp. 950-970. DOI: 10.1198/ 1061860032742.
[7] Experian. 2008. Experian Detect: Application Fraud Prevention System. Whitepaper, http://www.experian.com/products/pdf/experian detect.pdf.
[8] Wheeler, R. and Aitken, S. 2000. Multiple Algorithms for Fraud Detection, Knowledge-Based Systems 13(3): pp. 93-99. DOI: 10.1016/S0950-7051(00)00050-2.
[9] Goldenberg, A., Shmueli, G. and Caruana, R. 2002. Using Grocery Sales Data for the Detection of Bio-Terrorist Attacks, Statistical Medicine.
[10] Gordon, G., Rebovich, D., Choo, K. and Gordon, J. 2007. Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement, Center for Identity Management and Information Protection, Utica College.
[11] Hand, D. 2006. Classifier Technology and the Illusion of Progress, Statistical Science 21(1): pp. 1-15. DOI: 10.1214/088342306000000060.
[12] Head, B. 2006. Biometrics Gets in the Picture, Information Age August-September: pp. 10-11.
[13] Hutwagner, L., Thompson, W.,Seeman, G., Treadwell, T. 2006. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS), Journal of Urban Health 80: pp. 89-96.PMID: 12791783.
[14] ID Analytics. 2008. ID Score-Risk: Gain Greater Visibility into Individual Identity Risk. Unpublished.
[15] Jackson, M., Baer, A., Painter, I. and Duchin, J. 2007. A Simulation Study Comparing Aberration Detection Algorithms for Syndrome Surveillance, BMC Medical Informatics and Decision Making 7(6). DOI: 10.1186/1472-6947-7-6.

