ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Fraud Detection Using A New Multilayered Detection System

Kaminee Gurav, Manisha Gurabe, Priyanka Suryawanshi, Prof.Sinu Mathew

BE Student, Dept of CSE, University of Mumbai, Atharva college of Engineering, Mumbai, India

BE Student, Dept of CSE, University of Mumbai, Atharva college of Engineering, Mumbai, India

BE Student, Dept of CSE, University of Mumbai, Atharva college of Engineering, Mumbai, India

Assistant Professor, Dept of CSE, University of Mumbai, Atharva college of Engineering, Mumbai, India

ABSTRACT: Identity theft is a form of stealing someone's identity in which someone pretends to be someone else, usually as a method to gain access to resources or obtain benefits in that person's name. Identity crime is prevalent, and costly; and credit application fraud is a specific case of identity crime or identity theft. The existing non-data mining detection systems that uses business rules and scorecards, and known fraud matching have limitations.

To overcome these limitations and combat identity crime in real-time, we propose a new multi-layered detection system consisting of communal detection (CD) and spike detection (SD) layers that are resilient. Resilience is the long-term capacity of a system to deal with change and continue to develop communal detection (CD) finds real social relationships to decrease the suspicion score, and is tamper-resistant to the synthetic social relationships. It is the whitelist oriented approach on a fixed set of attributes [1]. The CD algorithm matches all links against the whitelist to find communal relationships and reduce their link score. CD can detect more types of attacks; better account for changing legal behavior and spike detection (SD) complements CD.

KEYWORDS: Identity crimes, communal detection, spike detection.

I. INTRODUCTION

In identity crime at one extreme there is synthetic identity fraud that refers to the use of believable but fictitious identities. These are easy to create but more difficult to apply successfully. At the other extreme we have real identity theft that refers to illegal use of innocent people's complete identity details. These can be harder to obtain but effortless to apply. In reality, identity crime can be carried out with a mix of both synthetic and real identity details. Identity crime has become prominent because there is lots of real identity data available on the Web, and also confidential data is accessible through unsecured mails.

Duplicates are of two types: exact and near duplicates .Exact (or identical) duplicates have the all same values whereas near (or approximate) duplicates have some same values (or characters), some similar values with slightly altered spellings, or both. 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 [1].

II. RELATED WORK

Fraud detection has been around for many years. Fraud behaviour has been increased as the financial institutions are providing electronic payment options by issuing credit and debit cards. Banks and other such outfits are worried about possible fraud. Fraud detection has been a challenging job in credit applications. Resilience of data mining algorithms in a complete detection system has not been explicitly addressed even though many data mining algorithms have been designed and implemented.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Much work in credit application fraud detection remains proprietary and exact performance figures unpublished and hence the CD and SD algorithms cannot be compared against their leading industry methods and techniques.

In example, [2] has Detect which provides four categories of policy rules to signal fraud, one of which is checking a new credit application against historical application data for consistency, one which checks validity, one checks whether the application triggers any fraud indicators, last policy checks whether the application matches any record in national fraud database.

Case-based reasoning (CBR) is the only known prior publication in the screening of credit applications [3]. CBR analyses the hardest cases which have been misclassified by existing methods and techniques. Retrieval uses thresholded nearest neighbour matching. Diagnosis utilizes multiple selection criteria (probabilistic curve, best match, negative selection, density selection, and default) and resolution strategies (sequential resolution-default, best guess, and combined confidence) to analyse the retrieved cases. Bayesian networks [4] uncover simulated anthrax attacks from real emergency department data. Wong [5] surveys algorithms for finding suspicious activity in time for disease outbreaks. Goldenberg et al. [6] use time series analysis to track early symptoms of synthetic anthrax outbreaks from daily sales of retail medication (throat, cough, and nasal) and some grocery items (facial tissues, orange juice, and soup). The system includes several layers; each customized to grocery data and tuned to finding footprints of an epidemic. Control-chart-based statistics, exponential weighted moving averages, and generalized linear models were tested on the same bioterrorism detection data and alert rate.

III. AIM AND OBJECTIVE

A. Aim

Data breaches which involve lost or stolen consumers' identity information can lead to other frauds such as tax returns, home equity, and payment card fraud. Consumers can incur thousands of dollars in out-of-pocket expenses. To prevent this we are designing a Resilient Fraud Detection System that detects fraud credit applications using a new multilayered approach consisting of two algorithms CD and SD.

B. Objective

Resilience is the ability to degrade gracefully when under most real attacks. The basic question asked by all detection systems is whether they can achieve resilience. To do so, the detection system trades off a small degree of efficiency (degrades processing speed) for a much larger degree of effectiveness (improves security by detecting most real attacks). The detection system needs "defence-in-depth" with multiple, sequential, and independent layers of defence [7] to cover different types of attacks. These layers are needed to reduce false negatives.

i. The two greatest challenges for the data mining-based layers of defence are adaptivity and use of quality data.
ii. Adaptivity accounts for morphing fraud behaviour, as the attempt to observe fraud changes its behaviour. But what is not obvious, yet equally important, is the need to also account for changing legal (or legitimate) behaviour within a changing environment.
iii. In the credit application domain, changing legal behaviour is exhibited by communal relationships (such as rising/falling numbers of siblings) and can be caused by external events (such as introduction of organizational marketing campaigns). This means legal behaviour can be hard to distinguish from fraud behaviour, but it will be shown later in this paper that they are indeed distinguishable from each other. The detection system needs to exercise caution with applications which reflect communal relationships. It also needs to make allowance for certain external events.
iv. Quality data are highly desirable for data mining and data quality can be improved through the real time removal of data errors (or noise). The detection system has to filter duplicates which have been re-entered due to human error or for other reasons. It also needs to ignore redundant attributes which have many missing values, and other issues.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

III. SYSTEM ARCHITECTURE

The System Architecture is described as follows -
Initially display the GUI for entering credit card application details, then the client details are accepted by the system. The client also submits a new application. Comparison of new applications is done with the existing ones. Then CD and SD algorithms are performed. After identifying whether the application is a genuine one the application is accepted.
The figure 1 illustrates the system architecture-

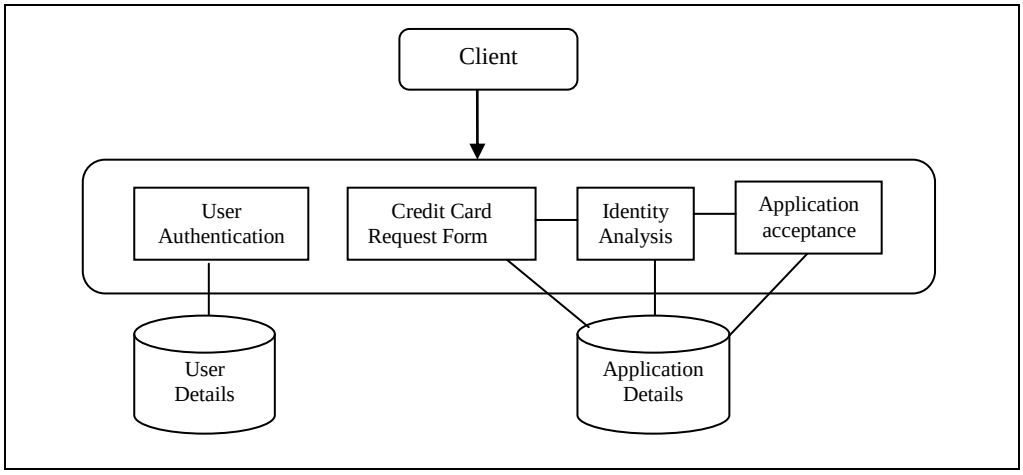


FIG1. SYSTEM ARCHITECTURE

IV. METHODOLOGY

We propose a new multi-layered detection system to combat identity crime in real-time, complemented with two additional layers: Communal Detection (CD) and Spike Detection (SD).

A. Communal Detection

Communal relationships are social or family relationships such as parent-child, brothers, sisters etc.
The usability of communal detection can be explained with an example. Suppose there are two credit card applications with same address ,landline phone number ,date of birth; one with name as John Smith and other Joan Smith.; this application can be interpreted in three ways:-

1. It is a fraudster attempting to obtain multiple credit cards using near duplicated data.
2. Possibly they are twins living in the same house and both are applying for a credit card.
3. Or it can be the same person applying twice, and has done a typing mistake.

The communal detection algorithm layer detects frauds from communal relationships. To account for legal behaviour and data errors, CD is the whitelist-oriented approach on a fixed set of attributes. The whitelist, a list of communal and self-relationships between applications, is crucial because it reduces the scores of these legal behaviours and false positives [1]. A false positive is an error in some evaluation process in which a condition tested for is mistakenly found to have been detected.

B. Spike Detection

SD layer complements CD layer. It strengthens CD by providing attribute weights which reflects the degree of importance of attribute (like name, phone number).

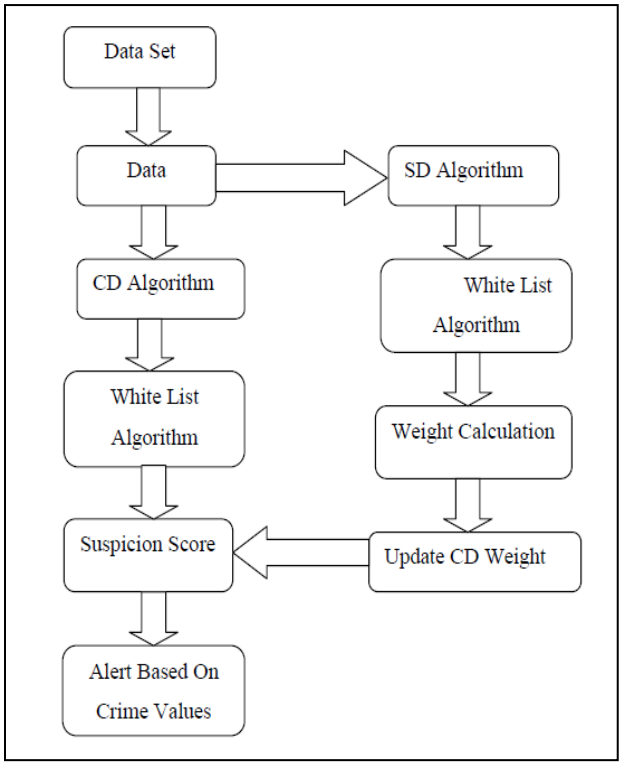
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

C. Basic Flow of System [8]



Step1**:** Initially we take input from user of a new application.
Step2: Now compare the new application with each other application in database and calculate link type. The link type is nothing but a binary string (e.g. 00101011) in which "1" represents matched fields and "0" represents unmatched fields.
Step 3: The applications with unmatched fields make up for the initial white list. The White list is a list that has verified applications, link type, number of applications corresponding to a particular link type and weight.
Next two steps are communal detection and spike detection.
Step 4: CD layer

- The new application is matched with application in the white list to find communal relationships between applications.
- If four or more fields are matched then CD assigns less suspicion score.
- Else application is added to white list.

Fig3. Sample of Six Credit Applications with Six Attributes [1]

i or j	**Given name**	**Family name**	**Unit name**	**Street name**	**Home Phone No.**	**Date of birth**
1	John	Smith	1	Circular road	91234567	1/1/1982
2	Joan	Smith	1	Circular road	91234567	1/1/1982
3	Jack	Jones	3	Square drive	93535353	3/2/1955
4	Ella	Jones	3	Square	93535353	6/8/1957

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

				drive		
5	Riley	Lee	2	Circular road	91215676	5/3/1963
6	Liam	Smyth	3	Circular road	91215676	1/1/1982

Step 5: SD layer

- Spike detection then verifies matched fields for their priority. The unique ID fields are given higher priority.
- If unique IDs are matched then the suspicious score gets increased and the application form is declared as fraud and hence finally rejected.
- If no unique IDs are matched then the application form is added into the white list as a genuine one.

VI. APPLICATION MODULE

i. Registration and Login
This module facilitates authentication of various users and thereby providing access to the selected users within the system.

ii. Apply for Credit Card
This feature will allow various users to apply for credit cards using various details required in the application to perform invalid application scenario.

iii. Track Details
The details of all applications are tracked and utilized in detection mechanism.

iv. Change Password
This module will facilitate changing the password details of the user.

v. Application Validation
The applications will be analysed using Fraud Detection Techniques to identify the identity conflict scenarios with the system and displaying it to the admin.

vi. Application Acceptance
This will be an admin module displaying the conflicts in the system and finally allowing admin to reject or accept these applications.

VII. RESULTS

A. Data set for real application

Data Set Substantial identity crime can be found in private and commercial databases containing information collected about customers, employees, suppliers, and rule violators. The same situation occurs in public and government regulated databases such as birth, death, patient and disease registries; taxpayers, residents' address, bankruptcy, and criminals lists. To reduce identity crime, the most important textual identity attributes such as personal name, Social Security Number (SSN), Date-of-Birth (DOB), and address must be used.

Therefore highest weights must be given to permanent attributes (such as SSN and DOB), followed by stable attributes (such as last name and state), and transient (or ever changing) attributes (such as mobile phone number and email address). The most important identity attributes differ from database to database. In our system we have given pan card no. (pan), license no (lno) and voter id (vid) highest weights; as shown in fig below:-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

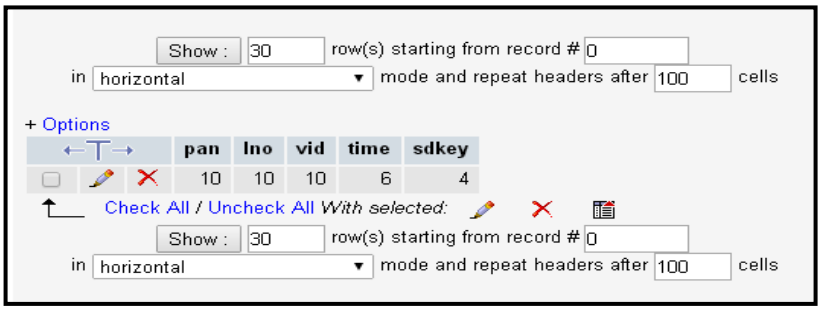


Fig4.

And attributes like mobile and email have been given comparatively lower weights as shown in fig below:-

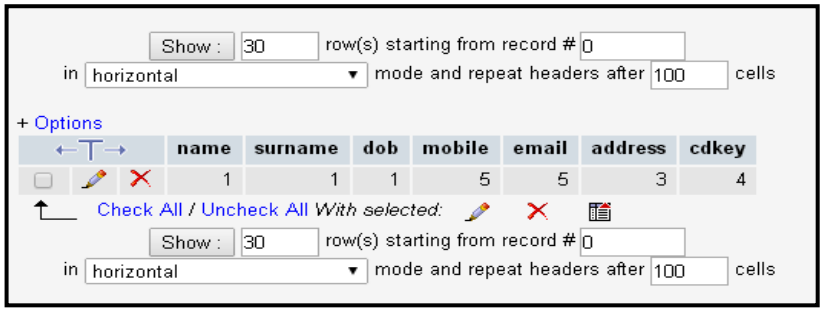


Fig5.

B. Result set

The white-list is constructed form the input data set and a CD suspicious score is assigned to each application as a result of communal detection algorithm. The Table shows the sample white-list constructed from credit applications in fig 3.

Fig6. Sample white list [1]

z	Link-type	No.	Weight
1	010101	2	0.25
2	011111	1	0.5
3	011110	1	0.75
4	001110	1	1

The Spike detection algorithm outputs the SD suspicious score. CD and SD scores are combined together to give a single score. SD updates the CD attribute weights. At last after calculating link type, CD Suspicious score, Multi-attribute score, the system gives the results as application is accept or reject after applying CD & SD algorithms on record.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Sr.No	Figures	References
1.	Basic Flow of System [8]	Aniruddha Kshirsagar, Lalit Dole,Recognizing the theft of identity using data mining ,International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 4, April 2014)
2.	Sample of Six Credit Applications with Six [1]	Clifton Phua, Member, IEEE, Kate Smith-Miles, Senior Member, IEEE, Vincent Lee, and Ross Gayler,Resilient Identity Crime Detection,IEEE transactions on knowledge and data engineering vol.24 no.3 year 2012.
3.	Sample white list [1]	Clifton Phua, Member, IEEE, Kate Smith-Miles, Senior Member, IEEE, Vincent Lee, and Ross Gayler,Resilient Identity Crime Detection,IEEE transactions on knowledge and data engineering vol.24 no.3 year 2012.

Fig7.

VIII. **CONCLUSION**

Resilient Identity Crime Detection; in other words, the real-time search for patterns in a multilayered 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 defence for a real-time credit application fraud detection system. In doing so, this research produced three concepts (or "force multipliers") which dramatically increase the detection system's effectiveness (at the expense of some efficiency). These concepts are resilience (multilayer defence), adaptivity (accounts for changing fraud and legal behavior), and quality data (real-time removal of data errors).

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 CD and SD algorithms is practical because these algorithms are designed for actual use to complement the existing detection system. Nevertheless, there are limitations. The first limitation is effectiveness, as scalability issues, extreme imbalanced class, and time constraints dictated the use of rebalanced data in this paper. The counter-argument is that, in practice, the algorithms can search with a significantly larger moving window, number of link types in the whitelist, and number of attributes. The second limitation is in demonstrating the notion of adaptivity. While in the experiments, CD and SD are updated after every period, it is not a true evaluation as the fraudsters do not get a chance to react and change their strategy in response to CD and SD as would occur if they were deployed in real life (experiments were performed on historical data).

IX. **ACKNOWLEDGEMENT**

We are thankful to our Principal Dr. Shrikant Kallurkar, Project Coordinator Prof. Deepali Maste and other senior faculties of Computer Department for technical assistance and feedback through discussions .Our thanks to some of our colleagues who contributed towards the success of this project.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

X. FUTURE WORK

For future work the security of the system can be improved by making a secure transaction based on a threshold value. Based on this the transaction application may be accepted or denied.

REFERENCES

1. Clifton Phua, Member, IEEE, Kate Smith-Miles, Senior Member, IEEE, Vincent Lee, and Ross Gayler, "Resilient Identity Crime Detection", IEEE transactions on knowledge and data engineering ,vol.24 no.3 year 2012.
2. Experian.ExperianDetect: Application Fraud Prevention System, Whitepaper, http://www.experian.com/products/pdf/experian_detect.pdf, 2008.
3. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java. Morgan Kauffman, 2000.
4. W. Wong, A. Moore, G. Cooper, and M. Wagner, "Bayesian Network Anomaly Pattern Detection for Detecting Disease Outbreaks,"Proc. 20th Int'l Conf. Machine Learning (ICML '03), pp. 808-815, 2003
5. W. Wong, "Data Mining for Early Disease Outbreak Detection," PhD thesis, Carnegie Mellon Univ., 2004.
6. Goldenberg, G. Shmueli, R. Caruana, and S. Fienberg, "Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales," Proc. Nat'l Academy of Sciences USA(PNAS '02), vol. 99, no. 8, pp. 5237-5240, 2002.
7. R. Bolton and D. Hand, "Unsupervised Profiling Methods for Fraud Detection," Statistical Science, vol. 17, no. 3, pp. 235-255, 2001.
8. Aniruddha Kshirsagar, Lalit Dole, "Recognizing the theft of identity using data mining" ,International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 4, April 2014)
9. P. Brockett, R. Derrig, L. Golden, A. Levine, and M. Alpert, "Fraud Classification Using Principal Component Analysis of RIDITs, "The J. Risk and Insurance, vol. 69, no. 3, pp. 341-371, 2002
10. Mr.Shakadwipi Amol J,Prof. P.N. Kalavadekar ,"Real-Time Credit Application Fraud Detection System based on Data Mining," Third Post Graduate Symposium on Computer Engineering Organized by department of Computer Engineering,MCERC Nasik cPGCON2014.
11. Bifet and R. Kirkby Massive Online Analysis, Technical Manual, Univ. of Waikato, 2009.
12. R. Caruana and A. Niculescu-Mizil, "Data Mining in Metric Space:An Empirical Analysis of Supervised Learning Performance Criteria," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '04), 2004
13. P. Christen and K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication," Quality Measures in Data Mining, F. Guillet and H. Hamilton, eds., vol. 43, Springer, 2007