Volume 5, Issue 7, July 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Resilient Identity Crime Detection Using Malay Phonetic Coding

Swapna Kashid[*]
Pursuing ME (CE) From RMD Sinhgad School of Engg.
Warje, Pune, Maharashtra, India

Prof. Manisha Desai
[M-Tech] Faculty at RMD Sinhgad School of Engg.
Warje, Pune, Maharashtra, India

Abstract—Identity crime is well well-known, prevalent, and expensive; and credit computer application fraud is a particular case of identity crime. Theexisting nondata mining detection system computer of business rules and scorecards, and well-known fraud similar have boundaries. To addressthese boundaries and contest identity crime in real time system, inthis proposes a new multifaceted computerdetection system complemented withtwo additional layers: communal detection (CD) and spike detection (SD). CD finds real social relationships to decrease the suspicionscore of that, and is tamper resistant to synthetic social relationships. It is the whitelist-oriented system on a set of records (attribute). SD searchspikes in duplicates to boost the suspicion score, and is probe-resistant for attributes. It is the attribute-oriented approach on avariable-size set of attributes. Together, CD and SD can sense more types of attacks, better account for varying legal behavior, anddelete the redundant records attributes. Experiments were accepted out on CD and SD with more than a few million real credit applications.

Apart from a all technical data quality problem, in diverse culture, personal identity attributes such as user names and user postal address contain language syntax. Hence, enhanced identity matching for specific culture could improve algorithmperformance. In this paper, we aim to evaluate effectiveness and efficiency of identity matching algorithm using Malay phonetic coding algorithm in Malaysian identity records.

Index Terms- whitelist-oriented approach,synthetic social relationships.

I. INTRODUCTION

DENTITYcrime is set as broadly as feasible in thispaper. At one extremecondition, synthetic identity fraud refers tothe use of possible but fictitious identities. These areeasy to create but more difficult to successfully used.At the other extreme, real identity robbery refers to black-hat useof innocent people's complete identity information. These can beharder to search (although large volumes of some identitydata are widely available) but easier to used.In real life, identity crime can be committed with a mix ofboth synthetic and real identity information.

Identity crime has become well-known because there is somuch real identity information available on the Net, andsecret data accessible through unsecured mailboxes.

It has also become effortless for perpetrators to hide their correctidentities. This can ensue in a myriad of insurance, credit and telecommunications fraud, as well as other newserious crimes. In addition to this, identity crime isblocked and costly in developed countries that do nothave nationally registered identity numbers.

Data sub breaches which involve stolen consumers'identity user data can lead to any other frauds such as payment card fraud and taxreturns, home equity. Consumers

Can incur thousands of dollars in out-of-pocket expenses.The US law requires criminal organizations to warnconsumers, so that consumers can moderate the harm. As a

final result, these organizations incur economic damage, such aswarning costs, fines, and lost business [24].Credit applications are Internet or paper-based formswith written information requirements by potential customers for creditcards, mortgage loans, and personal loans.

Credit application fraud is a particular case of identity crime, involvingsynthetic identity fraud and real identity robbery.As in identity crime, credit computer application fraud has reacheda critical mass of fraudsters who are extremely experienced,sophisticatedand organized [10]. Their observable patterns canbe different to each other and constantly change. They areunrelenting, due to the towering financial rewards, and the riskand effort involved are minimal. Based on anecdotalobservations of knowledgeable credit application investigators,fraudsters can use software automation to manipulateparticular values within an application and boostfrequency of successful record values.

Duplicates (or matches) refer to computer applications whichshare general values. There are types of duplicates:near (or approximate) duplicates have some same values(or characters),exact (or identical) duplicates have the all same values, some similar values with slightly alteredspellings, or both. Inthis paper argues that each successfulcredit application fraud pattern is represented by a suddenand sharp spike in duplicates within a short time, relativeto the established baseline level.

Identity matching algorithm used in retrieval domain anddifferent identity management, monetary computerapplication; and especially related to record security i.e. law enforcement [8, 14], casino security [6], insurance application fraud [11], telecommunication subscription fraud [1], etc. It is a challenging and crucial work because of unfinished identity information and errors can caused by many real-world factors, like privacy, fraudulent activity, data quality problem and legitimate identity changes. Mistaken information entry at help-desk is of human errors factor in information quality problem and it could also being done intentionally by an individual to gain illegal access on waged services. One of the political bloggers in Malaysia has been said in [17], to have more than one identity and fake passports to keep away from him being monitored by law enforcement; this is a special case of privacy in which it also involved fraudulent activity.

Main Challenges for Detection Systems 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). In fact, any form of security involves tradeoffs [26]

II. LITERATURE SURVEY

A. Bifet and R. Kirby, [1]. Massive Online Analysis (MOA) is a software environment for implementing algorithms and running experiments for online learning from evolving data streams. MOA includes a collection of offline and online methods as well as tools for evaluation. In particular, it implements boosting, bagging, and Hefting Trees, all with and without Naıve Bayes classifiers at the leaves. MOA supports bi-directional interaction with WEKA, the Waikato Environment for Knowledge Analysis, and is released under the GNU GPL license.

R. Bolton and D. Hand, [2].Credit card fraud falls broadly into two categories: behavioralfraud and application fraud. Application fraud occurs when individuals obtain new credit cards from issuingcompanies using false personal information and then spend as much as possible in a short space of time. However, most credit card fraud is behavioral and occurs when details of legitimate cards have been obtained fraudulently and sales are made on a 'Cardholder Not Present' basis. These sales include telephone sale s and e-commerce transactions where only the card details are required. In this paper, we are concerned with detecting behavioral fraud through the analysis of longitudinal data. These data usually consist of credit card transactions over time, but can include other variables, both static and longitudinal. Statistical methods for fraud detection are often classification (supervised) methods that discriminate between known fraudulent and non-fraudulent transactions; however, these methods rely on accurate identification of fraudulent transactions in historical databases – information that is often in short supply or non-existent. We are particularly interested in unsupervised methods that do not use this information but instead detect changes in behavior or unusual transactions. We discuss two methods for unsupervised frauddetection in credit data in this paper and apply them to some real data sets.

Hernandez, M.A., Stolfo, S.J. 1995. The merge/purge

Problem for large databases. [23], Many commercial organizations routinely gather large numb- beers of databases for various marketing and business analysis functions. The task is to correlate information from different databases by identifying distinct individuals that appear in a number of different databases typically in an inconsistent and oft env incorrect fashion. The problem we study here is the task of merging data from multiple sources in as efficient manner as possible, while maximizing the accuracy of the re- salt. We call this the merge/purge problem. In this paper we detail the sorted neighborhood method that is used by some to solve merge/purge and present experimental results that demonstrates this approach may work well in practice but at great expense. An alternative method based upon clustering is also presented with a comparative evaluation to the sorted neighborhood method. We show a means of improving the accuracy of the results based upon a multi-pass approach that succeeds by computing the Transitive Closure over the results of independent runs considering alternative primary key attributes in each pass.

Jonas, J. 2006. [24], This paper as Vegas, Nevada, is possibly the most interesting real-world setting for a high-stakes game of data surveillance. Most of the 38 million people who visit the city annually are attracted by the gam- bling, entertainment, shopping, architecture, dining, and shows. 1 However, among them are a few thousand “op- opportunists” who converge on Las Vegas solely to exploit its vulnerabilities. Some have become so infamous that gaming regulators have banned them from ever again stepping foot in a Nevada casino. In fact, if a casino gets caught doing business with such a person, it can be heavy- iy fined or, worse, lose its gaming license. If you’re a casino operator, knowing with whom you’re doing business isn’t just good business in terms of protecting corporate assets—it’s a matter of legal res- possibility. Finding a few bad actors, while minimizing the disruption, inconvenience, and privacy invasions to tens of millions of innocent tourists, has by necessity grown from an art mastered by a few practitioners into a teachable discipline. Elements of that discipline include regulatory policy, industry best practice procedures, staff development, and information technology.

Levenshtein, V.I. 1966,[25],Identity matching algorithm used in various identity management and retrieval domain, in which approximate string coding and similarity measure algorithm employed for attribute matching and indexing. Apart from a data quality problem, in diverse culture, personal identity attributes such as names and postal address contain specific language syntax. Hence, enhanced identity matching for specific culture could improve algorithm performance. In this paper, we aim to evaluate effectiveness and efficiency of identity matching algorithm using Malay phonetic coding algorithm in Malaysian identity records. Our experimental result shows promising accuracy, in which average F-measure increased 77%, as compared to identity matching without phonetic algorithm.

III. IMPLEMENTATION DETAILS

3.1 Main Challenges for Detection Systems:

Data sets: Resilience is the ability to degrade kindly when under most real attacks. The basic things asked by all detection systems are whether they can reach resilience. To do so, the detection system trades off a little degree of efficiency for a much bigger degree of effectiveness (improves security by detecting most real attacks). In fact, any form of security involves tradeoffs [26].

Adaptivity accounts for morphing fraud behavior, as the attempt to observe fraud changes its behavior. But what is not obvious, yet equally important, is the need to also account for changing legal (or legitimate) behavior within a changing environment. In the credit application domain, changing legal behavior is exhibited by communal relationships and can be caused by external events (such as introduction of organizational marketing campaigns). This means legal behavior can be hard to distinguish from fraud behavior, but it will be shown later in this paper that they are indeed distinguishable from each other.

Quality dataare 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 reentered due to human error or for other reasons. It also needs to ignore redundant attributes which have many missing values, and other issues.

Existing Identity Crime Detection System: There are nondata mining layers of defence to protect against credit application fraud, each with its unique strengths and weaknesses.

The first existing defence is made up of business rules and scorecards. In Australia, one business rule is the hundred-point physical identity check test which requires the applicant to provide sufficient point-weighted identity documents face-to-face.

The main contribution of this paper is the demonstration of resilience, with adaptivity and quality data in real-time data mining-based detection algorithms. The first new layer is Communal Detection (CD): the whitelist-oriented approach on a fixed set of attributes. To complement and strengthen CD, the second new layer is Spike Detection (SD): the attribute-oriented approach on a variable-size set of attributes. The second contribution is the significant extension of knowledge in credit application fraud detection because publications in this area are rare. In addition, this research uses the key ideas from other related domains to design the credit application fraud detection algorithms. Finally, the last contribution is the recommendation of credit application fraud detection as one of the many solutions to identity crime. Being at the first stage of the credit life cycle, credit application fraud detection also prevents some credittransactional fraud. Section 2 gives an overview of related work in credit application fraud detection and other domains. Section 3 presents the justifications and anatomy of the CD algorithm, followed by the SD algorithm. Before the analysis and interpretation of CD and SD results, Section 4 considers the legal and ethical responsibility of handling application data, and describes the data, evaluation measures, and experimental design.

3.2 Proposed Work

We employed Malay phonetic coding [10] for indexing and similarity measure in Wang et al. [16] identity matching framework, depicted in Figure 1. Our algorithm uses personal identity features, which are name, address and Malaysian identification card (IC) number. These three personal identity features are commonly available in identity management database and has been proven as effective in identifying matching identities.

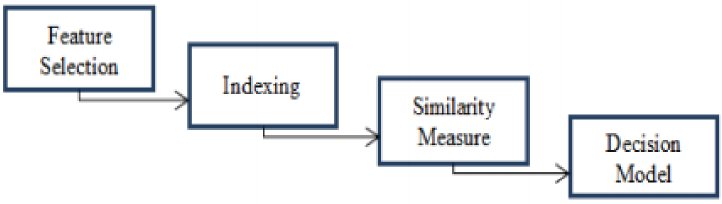

Fig. 1 Identity matching framework.

Identity matching is being used in diffents identity management and retrieval domain and cultures. Thus, enhanced identity matching for specific language and culture is possible by using phonetic coding algorithm, in which it can increase identity matching accuracy. Our experiment of identity matching using Malay phonetic coding algorithm is focused on the indexing atribute value and similarity measure algorithm for attribute matching. It is being tested in synthetic identity data set, in which contain various identity data problems including phonetic errors.

3.3 Algorithm:

CD Algorithm:

Step 1: Every application value is compared against a list of previous application values to find the links.

Step 2: Every application's link is matched against the white list to find communal relationships among applications and reduce their Link score.

Step 3: Every previous application's score is to be included into the current application's score. Previous score acts as a baseline level.

Step 4: Calculate every current application score using link and previous application's score.

Step 5: The algorithm updates one random parameter's value such that there is a tradeoff between effectiveness with efficiency, or vice versa.

Step 6: A new white list is constructed on the current Mini discrete stream links.

CD with Multi attributes:

-Select Application: Select month then show that month's application.
-Link Type and Weight: Attribute Name is e (legal/fraud, legal/fraud).
Link type calculates and shows in binary format (i.e. 0 or 1) and weight is calculated.

CD with Single Link:

In this show that records we select moth in previous (CD with Multi attribute).
-Single Link Score calculated.
-Suspicious score calculated.
-Parameter Performance is check.

SD Algorithm:

Step 1: Every application value is compared against a list of previous application values step by step.
Step2: Calculate application's current value score by integrating the steps to find the spikes.
Step 3: Calculate application's score using attribute weights.
Step4: Identify the key attributes to calculate the SD suspicion score.
Step 5: The final step updates the weights of the attributes.

SD with Multi attributes:

-Select Application: Select month then show that month's application
-Link Type and Weight: Attribute Name is e (legal/fraud, legal/fraud
Link type calculates and shows in binary format (i.e. 0 or 1) and weight is calculated
- Single scale is calculated.

SD with Single Link:

In this show that records we select moth in previous (CD with Multi attribute).
-Single value calculated
-Weight of attribute calculated
-Multiple value score calculated

CD Attribute Weight change:

-Here we update CD white List

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.

3.4Mathematical Model:

Photonic Coding:

-Here we load only non-duplicate data for processing
-Then we calculate Levenshtein distance by below formula
The Levenshtein distance between two strings $a, b$is given
By $\operatorname{lev}_{a,b}(|a|,|b|)$where

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j)=0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j)+1 \\ \operatorname{lev}_{a,b}(i,j-1)+1 \\ \operatorname{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Where $1_{(a_i \neq b_j)}$is the indicator function equal to 0 when $a_i = b_j$and equal to 1 otherwise.Note that the first element in the minimum corresponds to deletion (from a to b), the second to insertion and the third to match or mismatch, depending on whether the respective symbols are the same.

3.5 Performance Study:

Show results in graph F- Measure of CD and SD F-measure of SD and Photonic.

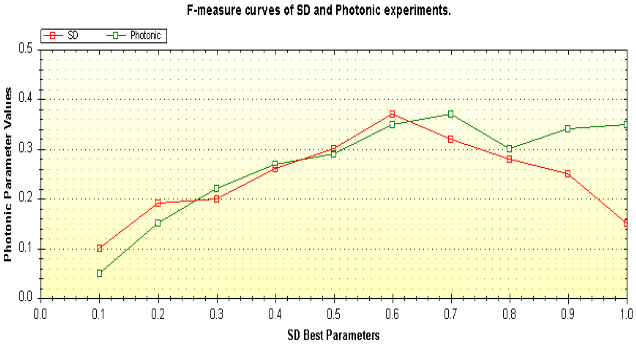

Fig .2 F-measure curves of SD and Photonic experiments.

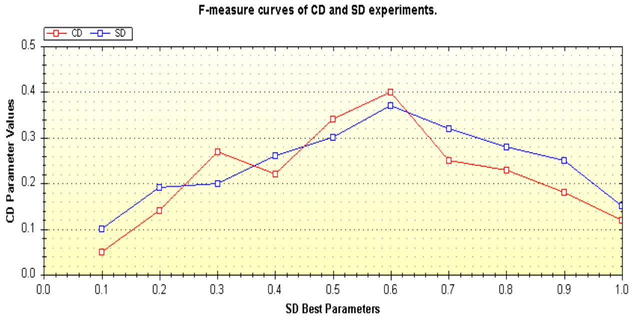

Fig .3 F-measure curves of CD and SD experiments.

IV. CONCLUSION

This paper describes an important field that has lots of problems relevant to other information mining research. It has documented the improvement and valuation in the information mining layers of cover for a real-time credit computer application fraud detection system. In doing so, this examine produced force multiplierswhich dramatically boost the detection system's usefulness (at the expense of some efficiency). These concepts are resilience (multilayer defence), accounts for changing fraud and legal behaviour, andreal-time removal of data errors. These concepts are fundamental to the evaluation, implementation, and designof total fraud detection, identity crime-related detection systems, and adversarial-related detection.

The working of CD and SD algorithms is practical because these algorithms are calculated for real use to complement the existing detection system. Nevertheless, there are limitations. The first limitation is extreme imbalanced class, as scalability issues,effectiveness, and time constraints dictated the use of rebalanced information in this paper. The main heart of this paper is Resilient Identity Crime Detection; in previous words,Identity matching is being used in various identity management and retrieval domain and cultures. Thus, improved identity matching for definite language and culture is likely by using phonetic coding algorithm, in which it can boost identity matching accuracy. Our experiment of identity matching using Malay phonetic coding algorithm is focused on the indexing attribute records value and similarity measure algorithm for attribute records matching.

REFERENCES

[1] A. Bifet and R. Kirkby Massive Online Analysis, TechnicalManual, Univ. of Waikato, 2009.

[2] R. Bolton and D. Hand, "Unsupervised Profiling Methods for Fraud Detection,"Statistical Science,vol. 17, no. 3, pp. 235-255,2001.

[3] P. Brockett, R. Derrig, L. Golden, A. Levine, and M. Alpert, "FraudClassification Using Principal Component Analysis of RIDITs," The J. Risk and Insurance, vol. 69, no. 3, pp. 341-371, 2002, doi: 10.1111/1539-6975.00027.

[4] 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, doi: 10.1145/1014052.1014063.

[5] 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, doi: 10.1007/978-3-540-44918-8.

[6] C. Cortes, D. Pregibon, and C. Volinsky, "Computational Methods for Dynamic Graphs," J. Computational and Graphical Statistics, vol. 12, no. 4, pp. 950-970, 2003, doi: 10.1198/1061860032742.

[7] Experian. Experian Detect: Application Fraud Prevention System, Whitepaper, http://www.experian.com/products/pdf/experian_detect.pdf, 2008.

[8] T. Fawcett, "An Introduction to ROC Analysis,"Pattern Recognition Letters,vol. 27, pp. 861-874, 2006, doi: 10.1016/j.patrec. 2005.10.010.

[9] A. Goldenberg, G. Shmueli, R. Caruana, and S. Fienberg, "Early Statistical Detection of Anthrax Outbreaks by Tracking Over-theCounter Medication Sales,"Proc. Nat'l Academy of Sciences USA (PNAS '02), vol. 99, no. 8, pp. 5237-5240, 2002.

[10] G. Gordon, D. Rebovich, K. Choo, and J. Gordon, "Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement," Center for Identity Management and Information Protection, Utica College, 2007.

[11] D. Hand, "Classifier Technology and the Illusion of Progress," Statistical Science,vol. 21, no. 1, pp. 1-15, 2006, doi: 10.1214/088342306000000060.

[12] B. Head, "Biometrics Gets in the Picture," Information Age,pp. 10-11, Aug.-Sept. 2006.

[13] L. Hutwagner, W. Thompson, G. Seeman, and T. Treadwell, "The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS),"J. Urban Health,vol. 80, pp. 89-96, 2006.

[14] IDAnalytics, "ID Score-Risk: Gain Greater Visibility into Individual Identity Risk," Unpublished, 2008.

[15] M. Jackson, A. Baer, I. Painter, and J. Duchin, "A Simulation Study Comparing Aberration Detection Algorithms for Syndromic Surveillance,"BMC Medical Informatics and Decision Making, vol. 7, no. 6, 2007, doi: 10.1186/1472-6947-7-6.

[16] J. Jonas, “Non-Obvious Relationship Awareness (NORA),” Proc. Identity Mashup,2006.

[17] M. Kantarcioglu, W. Jiang, and B. Malin, “A Privacy-PreservingFramework for Integrating Person-Specific Databases,” Proc. UNESCO Chair in Data Privacy Int’l Conf. Privacy in Statistical Databases (PSD ’08),pp. 298 314, 2008, doi: 10.1007/978-3-540-87471-3_25.

[18] J. Kleinberg, “Temporal Dynamics of On-Line Information Streams,” Data Stream Management: Processing High-Speed DataStreams,M. Garofalakis, J. Gehrke, and R. Rastogi, eds., Springer, 2005

[19] Becker, R.A., Volinsky, C., Wilks, A.R. 2010. Fraud detection in telecommunication: history and lessons learned. J.Technometrics52,1(Feb.2010),20-33. DOI=http://amstat.tandfonline.com/doi/abs/10.1198/TECH 2009.08136

[20] Bose, R. 2006. Intelligent technologies for managing fraud and identity theft. In Proceedings of the Third International Conference on Information Technology: New Generations. ITNG '06. IEEE, Washington, DC, 446-451. DOI=http://dx.doi.org/10.1109/ITNG.2006.78

[21] Christen, P. 2012. A Survey of indexing techniques for scalable record Linkage and deduplication. J. IEEE Transactions on Knowledge and Data Engineering24, 9(Sep. 2012), 1537-1555. DOI=http://doi.ieeecomputersociety.org/10.1109/TKDE.2011.127

[22] Ghosh, M. 2010. Telecoms fraud. J. Computer Fraud & Security2010, 7 (July 2010), 14-17. DOI=http://dx.doi.org/10.1016/S1361-3723(10)70082-8

[23] Hernandez, M.A., Stolfo, S.J. 1995. The merge/purge problem for large databases. In Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. SIGMOD '95. ACM, New York, NY, 127-138. DOI=http://dx.doi.org/10.1145/568271.223807.

[24] Jonas, J. 2006. Threat and fraud intelligence, Las Vegas style. J. IEEE Security & Privacy4, 6 (Nov. 2006), 28-34. DOI=http://dx.doi.org/10.1109/MSP.2006.169.

[25] Mutalib, N.S.A., Noah, S.A. 2011. Phonetic coding methods for Malay names retrieval. In Proceedings of the Semantic Technology and Information Retrieval (STAIR), 125-129. DOI=http://dx.doi.org/10.1109/STAIR.2011.5995776