Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Identify Crime Detection Using Data Mining Techniques

K.S.Arthisree, M.E, A.Jaganraj, M.E.
CSE DEPT.,
Arulmigu Meenakshi Amman College Of Engineering,
Thiruvannamalai district, Near Kanchipuram, India

Abstract— Credit-card-based purchases can be categorized into two types: 1) physical card and 2) virtual card. In a physical-card based purchase, the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In the second kind of purchase, only some important information about a card (card number, expiration date, secure code) is required to make the payment. Such purchases are normally done on the Internet or over the telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information.The only way to detect this kind of fraud is to analyze the spending patterns on every card and to figure out any inconsistency with respect to the "usual" spending patterns. Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. The existing nondata mining detection system of business rules and scorecards, and known fraud matching have limitations. To address these limitations and combat identity crime in real time, this paper proposes a new multilayered detection system complemented withtwo additional layers: communal detection (CD) and spike detection (SD).CD finds realsocial relationships to reduce the suspicion score, and is tamper resistant to synthetic social relationships. It is the whitelist-oriented approach on a fixed set of attributes. SD finds spikes in duplicates to increase the suspicion score, and is probe-resistant for attributes.

Key words— communal detection, spike detection, fraud detection, support vector machine

1. INTRODUCTION

Identity crime is defined as broadly as possible in this paper. At one extreme, synthetic identity fraud refers to the use of plausible but fictitious identities. These are effortless to create but more difficult to apply successfully. At the other extreme, real identity theft refers to 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) but easier to successfully apply. In reality, identity crime can be committed with a mix of both synthetic and real identity details.

Identity crime has become prominent because there is so much real identity data available on the Web, and confidential data accessible through unsecured mailboxes. It has also become easy for perpetrators to hide their true identities. This can happen in a myriad of insurance, credit, and telecommunications fraud, as well as other more serious crimes. In addition to this, identity crime is prevalent and costly in developed countries that do not have nationally registered identity numbers.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.

The US law requires offending organizations to notify consumers, so that consumers can mitigate the harm. As a result, these organizations incur economic damage, such as notification costs, fines, and lost business . Credit applications are Internet or paper-based forms with written requests by potential customers for credit cards, mortgage loans, and personal loans. Credit application fraud is a specific case of identity crime, involving synthetic identity fraud and real identity theft. As in identity crime, credit application fraud has reached a critical mass of fraudsters who are highly experienced, organized, and sophisticated. Their visible patterns can be different to each other and constantly change. They are persistent, due to the high financial rewards, and the risk and effort involved are minimal. Based on anecdotal observations of experienced credit application investigators, fraudsters can use software automation to manipulate particular values within an application and increase frequency of successful values.

Duplicates (or matches) refer 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 argues that each successful credit application fraud pattern is represented by a sudden and sharp spike in duplicates within a short time, relativeto the established baseline level. Duplicates are hard to avoid from fraudsters' point-ofview 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. It will be shown later in this paper that many

fraudsters operate this way with these applications and that their characteristic pattern of behavior can be detected by the methods reported. In short, the new methods are based on white-listing and detecting spikes of similar applications. White-listing uses real social relationships on a fixed set of attributes. This reduces false positives by lowering some suspicion scores. Detecting spikes in duplicates, on a variable set of attributes, increases true positives by adjusting suspicion scores appropriately. 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

1.1 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 . The detection system needs "defence-in-depth" with multiple, sequential, and independent layers of defence to cover different types of attacks. These layers are needed to reduce false negatives. In other words, any successful attack has to pass every layer of defence without being detected. The two greatest challenges for the data mining-based layers of defence are adaptivity and use of quality data. These challenges need to be addressed in order to reduce false positives.

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 (such as rising/falling numbers of siblings) 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. The detection system needs to exercise caution with applications which reflect communal relationships. It also needs to make allowance for certain external events. 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 reentered due to human error or for other reasons. It also needs to ignore redundant attributes which have many missing values, and other issues.

1.2 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. They must add up to at least 100 points, where a passport is worth 70 points. Another business rule is to contact (or investigate) the applicant over the telephone or Internet. The above two business rules are highly effective, but human resource intensive. To rely less on human resources, a common business rule is to match an application's identity number, address, or phone number against external databases. This is convenient, but the public telephone and address directories, semipublic voters' register, and credit history data can have data quality issues of accuracy, completeness, and timeliness. In addition, scorecards for credit scoring can catch a small percentage of fraud which does not look creditworthy; but it also removes outlier applications which have a higher probability of being fraudulent. The second existing defence is known fraud matching. Here, known frauds are complete applications which were confirmed to have the intent to defraud and usually periodically recorded into a blacklist. Subsequently, the current applications are matched against the blacklist. This has the benefit and clarity of hindsight because patterns often repeat themselves. However, there are two main problems in using known frauds.

First, they are untimely due to long time delays, in days or months, for fraud to reveal itself, and be reported and recorded. This provides a window of opportunity for fraudsters. Second, recording of frauds is highly manual. This means known frauds can be incorrect , expensive, difficult to obtain, [3], andhave the potential of breaching privacy.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 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 (converted from the original, sparse string attributes) and too few known frauds.

1.3 New Data Mining-Based Layers of Defence

The main objective of this research is to achieve resilience by adding two new, real time, data mining-based layers to complement the two existing nondata mining layers discussed in the section. These new layers will improve detection of fraudulent applications because the detection system can detect more types of attacks, better account for changing legal behavior, and remove the redundant attributes. These new layers are not human resource intensive. They represent patterns in a score where the higher the score for an application, the higher the suspicion of fraud (or anomaly). In this way, only the highest scores require human intervention. These two new layers, communal and spike detection, do not use external databases, but only the credit application database per se. And crucially, these two layers are unsupervised algorithms which are not completely dependent on known frauds but use them only for evaluation. 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.

2. BACKGROUND

Many individual data mining algorithms have been designed, implemented, and evaluated in fraud detection. Yet until now, to the best of the researchers' knowledge, resilience of data mining algorithms in a complete detection system has not been explicitly addressed. Much work in credit application fraud detection remains proprietary and exact performance figures unpublished, therefore there is no way to compare the CD andSD algorithms against their leading industry methods and techniques. For example, has ID Score-Risk which gives a combined view of each credit application's characteristics and their similarity to other industry-provided or Web identity's characteristics. In another example, 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. Case-based reasoning (CBR) is the only known prior publication in the screening of credit applications . CBR analyzes the hardest cases which have been misclassified by existing methods and techniques.

Retrieval uses thresholded nearest neighbor 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 analyze the retrieved cases. CBR has 20 percent higher true positive and true negative rates than common algorithms on credit applications. The CD and SD algorithms, which monitor the significant increase or decrease in amount of something important (Section 3), are similar in concept to credit transactional fraud detection and bioterrorism detection. Peer group analysis [2] monitors interaccount behavior over time. It compares the cumulative mean weekly amount between a target account and other similar accounts (peer group) at subsequent time points. The suspicion score is a t-statistic which determines the standardized distance from the centroid of the peer group. On credit card accounts, the time window to calculate a peer group is 13 weeks, and the future time window is 4 weeks. Break point analysis [2] monitors intraaccount behavior over time. It detects rapid spending or sharp increases in weekly spending within a single account. Accounts are ranked by the t-test. The fixed-length moving transaction window contains 24 transactions: the first 20 for training and the next four for evaluation on credit card accounts. Bayesian networks uncover simulated anthrax attacks from real emergency department data.Wong surveys algorithms for finding suspicious activity in time for disease outbreaks. Goldenberg et al. 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). Control-chart-based statistics, exponential weighted moving averages, and generalized linear models were tested on the same bioterrorism detection data and alert rate.

The SD algorithm, which specifies how much the current prediction is influenced by past observations (Section 3.3), is related to Exponentially Weighted Moving Average (EWMA) in statistical process control research. In particular, like EWMA, the SD algorithm performs linear forecasting on the smoothed time series, and their advantages include low implementation and computational complexity. In addition, the SD algorithm is similar to change point detection in biosurveillance research, which maintains a cumulative sum (CUSUM) of positive deviations from the mean . Like CUSUM, the SD algorithm raises an alert when the score/CUSUM exceeds a threshold, and both detects change points faster as they are sensitive to small shifts from the mean. Unlike CUSUM, the SD algorithm weighs and chooses string attributes, not numerical ones.

3. THE METHODS

This section is divided into four sections to systematically explain the CD algorithm (first two sections) and the SD algorithm (last two sections). Each section commenceswith a clearer discussion about its purposes.

3.1 Communal Detection

This section motivates the need for CD and its adaptive approach. Suppose there were two credit card applications that provided the same postal address, home phone number, and date of birth, but one stated the applicant's name to be John Smith, and the other stated the applicant's name to be Joan Smith. These applications could be interpreted in three ways: 1. Either it is a fraudster attempting to obtain multiple credit cards using near duplicated data. 2. Possibly there are twins living in the same house who both are applying for a credit card. 3. Or it can be the same person applying twice, and there is a typographical error of one character in the first name. With the CD layer, any two similar applications could be easily interpreted as (1) because this paper's detection methods use the similarity of the current application to all prior applications (not just known frauds) as the suspicion score. However, for this particular scenario, CD would also recognize these two applications as either (2) or (3) by lowering the suspicion score due to the higher possibility that they are legitimate. To account for legal behavior 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 behaviors and false positives. Communal relationships are near duplicates which reflect the social relationships from tight familial bonds to casual acquaintances: familymembers, housemates, colleagues, neighbors, or friends .The family member relationship can be further broken down into more detailed relationships such as husbandwife, parent-child, brother-sister, male-female cousin (or both male, or both female), as well as uncle-niece (or

unclenephew, auntie-niece, auntie-nephew). Self-relationships highlight the same applicant as a result of legitimate behavior (for simplicity, self-relationships are regarded as communal relationships). Broadly speaking, the whitelist is constructed by ranking link-types between applicants by volume. The larger the volume for a link-type, the higher the probability of a communal relationship. On when and how the whitelist is constructed, please refer to Section 3.2, Step 6 of the CD algorithm. However, there are two problems with the whitelist.

First, there can be focused attacks on the whitelist by fraudsters when they submit applications with synthetic communal relationships. Although it is difficult to make definitive statements that fraudsters will attempt this, it is also wrong to assume that this will not happen. The solution proposed in this paper is to make the contents of the whitelist become less predictable. The values of some parameters (different from an application's identity value) are automatically changed such that it also changes the whitelist's link types. In general, tampering is not limited to hardware, but can also refer to manipulating software such as code. For our domain, tamper resistance refers to making it more difficult for fraudsters to manipulate or circumvent data mining by providing false data.

Second, the volume and ranks of the whitelist's real communal relationships change over time. To make the whitelist exercise caution with (or more adaptive to) changing legal behavior, the whitelist is continually being reconstructed. 3.2 CD Algorithm Design This section explains how the CD algorithm works in real time by giving scores when there are exact or similar matches between categorical data; and in terms of its nine inputs, three outputs, and six steps. This research focuses on one rapid and continuous data stream of applications. For clarity, let G represent the overall stream which contains multiple and consecutive f.gx_2; gx_1; gx; gxþ1; gxþ2;g Minidiscrete streams gx: current Minidiscrete stream which contains multiple and consecutive fux;1; ux;2; ux;pg microdiscrete streams.x: fixed interval of the current month, fortnight, or week in the year. . p: variable number of microdiscrete streams in a Minidiscrete stream. Also, let ux;y represent the current microdiscrete stream which contains multiple and consecutive fvx;y;1; vx;y;2;vx;y;qg applications. The current application's links are restricted to previous applications within a moving window,and this window can be larger than the number of pplications within the current microdiscrete stream.y: fixed interval of the current day, hour, minute, or second. q: variable number of applications in a microdiscrete stream. Here, it is necessary to describe a single and continuous stream of applications as being made up of separate chunks: a Minidiscrete stream is long-term (for example, a month of applications); while a microdiscrete stream is short-term (for example, a day of applications). They help to specify precisely when and how the detection system will automatically change its configurations. For example, the CD algorithm reconstructs its whitelist at the end of the month and resets its parameter values at the end of the day; the SD algorithm does attribute selection and updates CD attribute weights at the end of the month. Also, for example, long-term previous average score, longterm previous average links, and average density of each attribute are calculated from data in a Minidiscrete stream; short-term current average score and short-term current average links are calculated from data in a microdiscrete stream. With this data stream perspective in mind, the CD algorithm matches the current application against a moving window of previous applications. It accounts for attribute weights which reflect the degree of importance in attributes. The CD algorithm matches all links against the whitelist to find communal relationships and reduce their link score. It then calculates the current application's score using every link score and previous application score.

Inputs
v_i (current application)
W number of v_j (moving window)
$\Re_{x,link-type}$ (link-types in current whitelist)
$T_{similarity}$ (string similarity threshold)
$T_{attribute}$ (attribute threshold)
η (exact duplicate filter)
α (exponential smoothing factor)
T_{input} (input size threshold)
SoA (State-of-Alert)

Outputs
$S(v_i)$ (suspicion score)
Same or new parameter value
New whitelist

CD algorithm

Step 1: Multi-attribute link [match v_i against W number of v_j to determine if a single attribute exceeds $T_{similarity}$; 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 [calculate single-link score by matching Step 1's multi-attribute links against $\Re_{x,link-type}$]

Step 3: Single-link average previous score [calculate average previous scores from Step 1's linked previous applications]

Step 4: Multiple-links score [calculate $S(v_i)$ based on weighted average (using α) of Step 2's link scores and Step 3's average previous scores]

Step 5: Parameter's value change [determine same or new parameter value through SoA (for example, by comparing input size against T_{input}) at end of $u_{x,y}$]

Step 6: Whitelist change [determine new whitelist at end of g_x]

3.2 Spike Detection

This section contrasts SD with CD; and presents the need for SD, in order to improve resilience and adaptivity.Before proceeding with a description of SD, it isnecessary to reinforce that CD finds real social relationships to reduce the suspicion score, and is tamper resistant to synthetic social relationships. It is the whitelist-oriented approach on a fixed set of attributes. In contrast, SD finds spikes to increase the suspicion score, and is probe resistant for attributes. Probe resistance reduces the chances a fraudster will discover attributes used in the SD score calculation. It is the attribute-oriented approach on a variable-size set of attributes. A side note: SD cannot use a whitelist-oriented approach because it was not designed to create multiattribute links on a fixed-size set of attributes. CD has a fundamental weakness in its attribute threshold. Specifically, CD must match at least three values for our data set. With less than three matched values, our whitelist does not contain real social relationships because some values, such as given name and unit number, are notunique identifiers. The fraudster can duplicate one or two important values which CD cannot detect.

SD complements CD. The redundant attributes are either too sparse where no patterns can be detected, or too dense where no denser values can be found. The redundant attributes are continually filtered, only selected attributes in the form of not-too-sparse and not-too-dense attributes are used for the SD suspicion score. In this way, the exposure of the detection system to probing of attributes is reduced because only one or two attributes are adaptively selected. Suppose there was a bank's marketing campaign to give attractive benefits for it's new ladies' platinum credit card.

This will cause a spike in the number of legitimate credit card applications by women, which can be erroneously interpreted by the system as a fraudster attack. To account for the changing legal behavior caused by external events, SD strengthens CD by providing attribute weights which reflect the degree of importance in attributes. The attributes are adaptive for CD in the sense that its attribute weights are continually determined. This addresses external events such as the entry of new organizations and exit of existing ones, and marketing campaigns of organizations which do not contain any patterns and are likely to cause three natural changes in attribute weights. These changes are volume drift where the overall volume fluctuates, population drift where the volume of both fraud and legal classes fluctuates independent of each other, and concept drift which involves changing legal characteristics that can become similar to fraud characteristics. By tuning attribute weights, the detection system makes allowance for these external events.In general, SD trades off effectiveness (degrades security because it has more false positives without filtering out communal relationships and some data errors) for efficiency (improves computation speed because it does not match against the whitelist, and can compute each attribute in parallel on multiple workstations). In contrast, CD trades off efficiency (degrades computation speed) for effectiveness (improves security by accounting for communal relationships and more data errors).

Inputs
v_i (current application)
W number of v_j (moving window)
t (current step)
$T_{similarity}$ (string similarity threshold)
θ (time difference filter)
α (exponential smoothing factor)

Outputs
$S(v_i)$ (suspicion score)
w_k (attribute weight)

SD algorithm

Step 1: Single-step scaled counts [match v_i against W number of v_j to determine if a single value exceeds $T_{similarity}$ and its time difference exceeds θ]

Step 2: Single-value spike 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 $S(v_i)$ from Step 2's value scores and Step 4's w_k]

Step 4: SD attributes selection [determine w_k for SD at end of g_x]

Step 5: CD attribute weights change [determine w_k for CD at end of g_x]

Fig. 1. Real Application Data Set (RADS).

4. EXPERIMENTAL RESULTS

4.1 *Identity Data—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 governmentregulated 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. The following publications support this argument: Jonas ranks SSN as most important, followed by personal name, DoB, and address. Jost assigns highest weights 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). Sweeney states that DoB, gender, and postcode can uniquely identify more than 80 percent of the United States (US) population. Head and Kursun et al. [20] regard name, gender, DoB, andaddress as the most important attributes. The most important identity attributes differ from database to database. They are least likely to be manipulated, and are easiest to collect and investigate. They also have the least missing values, least spelling, and transcription errors, and have no encrypted values.

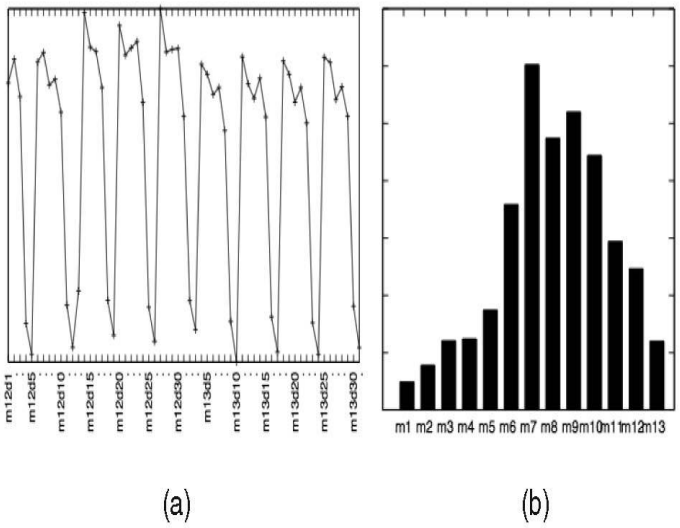

Fig (a) Daily application volume for 2 months. (b) Fraud percentage across months.

This real data set was chosen because, at experimentation time, it had the most recent fraud behavior. Although this real data set cannot be made available, there is a synthetic data set of 50,000 credit applications which is available at The specific summaries and basic statistics of the real credit application data are discussed below. For purposes of confidentiality, the application volume and fraud percentage in Fig. 1 have been deliberately removed. Also, the average fraud percentage (known fraud percentage in all applications) and specific attributes for application fraud detection cannot be revealed. The impact of fewer known frauds means algorithms will produce poorer results and lead to incorrect evaluation.

To reduce this negative impact and improve scalability, the data have been rebalanced by retaining all known frauds but randomly under sampling unknown applications by 90 percent. 4.4 CD and SD's Results and Discussion The CD F-measure curves skew to the left. The CD-related F-measure curves start from 0.04 to 0.06 at threshold 0, and peak from 0.08 to 0.25 at thresholds 0.2 or 0.3. On the other hand, the SD F-measure curves skew to the right.Without the whitelist, the results are inferior. From Fig. 2 at threshold 0.2, the no-whitelist experiment (F-measure below 0.09) performs poorer than the CD-baseline experiment (F-measure above 0.1). From Fig. 3, the no-whitelist experiment has about 10 percent more false positives than the CD-baseline experiment. This verifies the hypothesis that the whitelist is crucial because it reduces the scores of these legal behavior and false positives; also, the larger the volume for a link type, the higher the probability of a communal relationship.

5. CONCLUSION

The main focus of this paper is Resilient Identity Crime Detection; in other words, the real-time search for patterns in a multilayered and principled fashion, to safeguardcredit 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 onstraints 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).

References

[1] A. Bifet and R. Kirkby Massive Online Analysis, Technical Manual, 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, "Fraud Classification 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 Discoveryand 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.

Author's Biographies

Arthisree K.S , received the BE degree from Pallavan College of Engineering Which affiliated to Anna University, Chennai, India, in 2011. He also pursued Master Degree in Arulmigu Meenakshi Amman college of engineering which is Affiliated to Anna University, Chennai. His research interests include Cyber Network, cloud computing, Internet security.
E-mail: artiisri@gmail.com.

Jaganraj *A* received BE degree from Arunai Engineering College Which affiliated to Anna University, Chennai, India, in 2010. He pursued the Master Degree in Arulmigu Meenakshi Amman college of engineering which is Affiliated to Anna University, Chennai. His research interests include malicious node analysis, web and internet security, and wireless network security.
E-mail: jagan_math88@yahoo.co.in.

