ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 3, March 2014*

# A comprehensive and relative study of detecting deformed identity crime with different classifier algorithms and multilayer mining algorithm

**Sohini Bhattacharya Chakraborty**[1], **M. Z. Shaikh**[2]

ME Student, Department of Computer Science, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India[1]

Principal, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India[2]

**Abstract:** Now a day's identity crime is well known, prevalent and prominent in our society. There are some algorithms which have been implemented to detect or resolve resilience identity. The existing data mining and non data mining algorithm and known fraud matching have some limitation. To achieve a complete and transparent view of different resilience identity crime detection system a comparative study and approach is a prime focus of this paper. Here different classifiers algorithm are compared with data mining based algorithm with multilayer mining stage of defense (territory detection and suspicion score detection algorithm) to probe the synthetic identity crime. Through this comprehensive approach we can find out the relative measurement of different classifier algorithm using proposed algorithm territory detection and suspicion score detection algorithm as baseline. Although multilayer mining algorithm is specific to credit application fraud detection, but the concept of resilience or deformation with this comparative study discussed in this paper are general to design, implement and evaluate of all detection system.

**Keywords:** Data mining based fraud detection, security, anomaly detection, data stream mining.

## I. INTRODUCTION

Identity crime is well known important and very costly in our society. 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. At one 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 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 loans. Credit application fraud is a specific case of identity crime, involving Synthetic identity fraud and real identity theft. 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 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. To resolve and combat against synthetic identity theft this paper proposes a multi layer data mining stage of defenses, territory detection and suspicion score detection algorithm and compare it with four classifier algorithm( supervised algorithm) like logistic regression, SVM, decision tree and neural network and also with CBR analysis. Territory detection finds real social relationships to shrink the suspicion score, and is corrupt opposed to synthetic social relationships. It is the white list-oriented approach on a fixed set of attributes. Suspicion score detection finds spikes in duplicates to enhance the suspicion score, and is probe-resistant for attributes. It is the attribute-oriented approach on a variable-size set of attributes.

## II. RELATED WORKS

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.

## III. METHODS

This section discusses the various data mining methods which are used in the identity crime detection in both the transaction domain also the application domain. These are the following method for the transaction and application domain.

Outlier Detection- An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Unsupervised learning is a new explanation or representation of the observation data, which will then lead to better future responses or decisions. Unsupervised methods do not need the preceding knowledge of fraudulent and non-fraudulent transactions in old database, but instead detect changes in behavior or unusual transactions. These methods model a baseline distribution that represents normal behavior and then detect observations that show greatest departure from this norm. In supervised methods, models are trained to discriminate between fraudulent and non-fraudulent behavior so that new observations can be assigned to classes. Supervised methods need accurate identification of fraudulent transactions in old databases and can only be used to detect frauds of a type that have previously occurred. A benefit of using unsupervised methods over supervised methods is that previously undiscovered types of fraud may be detected.

For detection of the credit transactional and application fraud analysis the following classifier methods will be discussed.

- Neural network and Bayesian Network
- Logistic regression
- Decision tree
- Support vector Machine
- CBR analysis

These comprehensive methods are studied and later on a relative discussion will be there on this approach with currently proposed territory detection and suspicion score detection algorithm.

**Neural Networks and Bayesian Network:** A neural network is a set of interconnected nodes designed to imitate the functioning of the human brain. Each node has a weighted connection to several other nodes in adjacent layers. Individual nodes take the input received from connected nodes and use the weights together with a simple function to compute output values, Neural networks can be constructed for supervised or unsupervised learning. CARDWATCH [16] features neural networks trained with the past data of a particular customer. It makes the network process the current spending patterns to detect possible anomalies. Brause and Langsdorf proposed the rule- based association system combined with then euro-adaptive approach [17]. Falcon developed by HNC uses feed-forward Artificial Neural Networks trained on a variant of a back propagation training algorithm [18]. A neural MLP-based classifier is another example using neural networks. It acts only on the information of the operation itself and of its immediate previous history, but not on historic databases of past cardholder activities. A parallel Granular Neural Network (GNN) method uses fuzzy neural network and rule- based approach. The neural system is trained in parallel using training data sets, and then the trained parallel fuzzy neural network discovers fuzzy rules for future prediction. Cyber Source introduces a hybrid model, combining an expert system with a neural network to increase its statistic modeling and reduce the number of "false" rejections.

Bayesian and Neural network approach is automatic credit card fraud detection system and type of artificial intelligence programming which is based on variety of methods including machine learning approach, supervised and data mining for reasoning under uncertainty. The advantage of neural network is that it learns and does not need to be reprogrammed. Its processing speed is higher than Bayesian neural networks but it needs high processing time for large neural networks. Whereas Bayesian neural networks provide good accuracy but needs training of data to operate and requires high processing speed.

**Logistic regression:** Logistic Regression-Two advanced data mining approaches, support vector machines and random forests, together with the well known logistic regression [19], as part of an attempt to better detect (and thus control and prosecute) credit card fraud. Logistic regression (LR) is useful for situations in which we want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model and it is applicable to a broader range of research situations than feature analysis. (**Ohlson,** 1980; **Martin,** 1997) estimating the odds of a firm's failure with probability. For model development in logistic regression some form of sampling among the two classes is typically used to obtain training data with reasonable class distributions. Various sampling approaches have been proposed in the literature, with random oversampling of minority class cases and random under sampling of majority class cases being the simplest and most common in use; others include directed sampling The second problem in developing supervised models for fraud can arise from potentially undetected fraud transactions, leading to mislabeled cases in the data to be used for building the model.

**Decision Tree:** Decision trees are statistical data mining technique that express independent attributes and a dependent attributes logically AND in a tree shaped structure. Classification rules, extracted from decision trees, are IF-THEN expressions and all the tests have to succeed if each rule is to be generated [20]. Decision tree usually separates the complex problem into many simple ones and resolves the sub problems through repeatedly using [20]. Decision trees are predictive decision support tools that create mapping from observations to possible consequences. There are number of popular classifiers construct decision trees to generate class models. From the well-known decision tree algorithms, ID3, C5.0 and C&RT methods use impurity measures to choose the splitting attribute and the split value/s. ID3 [21] uses information gain while the successor, C5.0 uses gain ratio, and C&RT [26] uses Gini coefficient for impurity measurements. Unlikely, CHAID uses chi-square or F statistics to choose the splitting variable [22]. As the tree is grown, the resultant tree may over fit the training data containing possible errors or noise or some of the branches of the resultant tree may contain anomalies. So, the resultant tree should be checked whether removal of some nodes, starting from the leaf ones, make a significant effect on the tree's classification performance. This operation is called as pruning. After the tree is grown, a new observation or record is classified by tracing the route on the tree up to a leaf node according to the values of the attributes of the record. This is done by recursively checking the values of the splitting attribute of the record at each node and following the required branch of the tree until a leaf node is reached. The label of the leaf node reached gives the class which the new observation or record is classified in.

**Support vector machine:** Unlike the decision tree methods, SVM tries to find a hyper plane to separate the two classes while minimizing the classification error. SVM is a new and promising classification and regression technique proposed by Vapnikand his group at AT&T Bell Laboratories [23]. SVM learns a separating hyper plane which maximizes the margin and produces good generalization ability [24]. In prior literature, SVM has been successfully applied to many areas such as telecommunication fraud detection , pattern recognition , system intrusion detection . SVM's basic idea is to transform the attributes to a higher dimensional feature space and find the optimal hyper plane in that space that maximizes the margin between the classes. Briefly, SVM does this by using a polynomial, sigmoid, radial basis or a linear kernel function which satisfies the Mercer condition [25]. Two key elements in the implementation of SVM are the techniques of mathematical programming and kernel functions. The parameters are found by solving a quadratic programming problem with linear equality and inequality constraints; rather than by solving a non-convex, unconstrained optimization problem. The flexibility of kernel functions allows the SVM to search a wide variety of hypothesis spaces.

**CBR:** Case-based reasoning (CBR)–It is detection technique for the credit applications. CBR analyses the hardest cases which have been misclassified by existing methods and techniques. Retrieval process uses threshold nearest neighbor matching. Diagnosis utilizes multiple selection criteria which are probabilistic curve, best match, negative selection, density selection, and default and resolution strategies which are sequential resolution-default, best guess, and combined confidence to analyze the recovered cases. CBR has 20 percent higher true positive and true negative rates than common algorithms on credit applications [1].

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.

## IV. RELATIVE SYUDY WITH PROPOSED METHOD

The new methods are based on white-listing and detecting suspicion score of similar applications. White-listing or Territory detection 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 [10].

## V. METHODS

This section is divided into four subsections to systematically explain the Territory detection algorithm (first two subsections) and the suspicion score detection algorithm (last two subsections). Each subsection commences with a clearer discussion about its purposes.

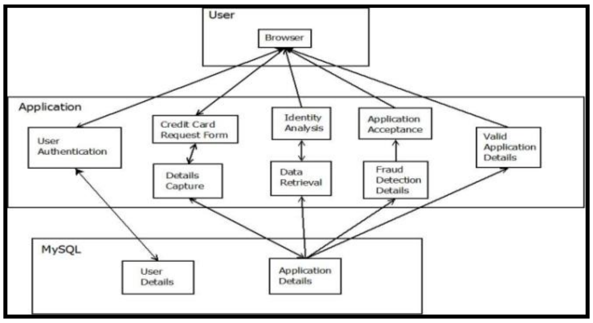

Fig1: System architecture of proposed system

### 5.1 Territory 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 Multi attributes link count by Jaro-Winkler formula

The Jaro–Winkler[15] distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match. The Jaro distance of two given strings and is

The Jaro distance $d_j$ of two given strings $s_1$and $s_2$is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

Where:

· $m$is the number of matching characters

Two characters from $s_1$and $s_2$respectively, are considered matching only if they are the same and not farther than

$$\left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor - 1 .$$

### 5.3 Suspicion score 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 gx]

Step 5: Territory attributes weights change [determine wk for Territory at end of gx]

## VI. CONCLUSIONS

This comprehensive and relative study of different classifier algorithm and the proposed two algorithm have been discussed through this paper highlights the prime functionality of these previous algorithms for detecting credit application and transactional fraud detection. . This paper also 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, the proposed approach produces three concepts which increase the detection system's effectiveness (at the expense of some efficiency). These concepts are resilience (multi-layer defense), 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 Territory 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. http://www.experian.com/ products/pdf/experian detect.pdf.

[7] Experian. 2008. Experian Detect: Application Fraud Prevention System .Whitepaper.

[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.

[16] E. Aleskerov, B. Freisleben, and B. Rao. Cardwatch: a neural network based database mining system for credit card fraud detection. In Proceedings of Computational Intelligence forFinancial Engineering, pages 173-200, 1997.

[17] R. Brause, T. Langsdorf, and M. Hepp. Credit card fraud detection by adaptive neural data mining. In Proceedings of the11th IEEE International Conference on Tools with Artificial Intelligence, pages 103-106, 1999.

[18] Experian.ExperianDetect:ApplicationFraudPreventionSystem,Whitepaper,http://www.experian.com/products/pdf/experian_detect.pdf, 2008.

[19] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, J. Christopher Westland, “Data mining for credit card fraud: A comparative study”, Decision Support Systems 50 pp. 602–613, 2011.

[20] S. Benson Edwin Raj, A. Annie Portia, “Analysis on Credit Card Fraud Detection Methods”, International Conference on Computer, Communication and Electrical Technology – ICCCET2011, 18th & 19th March, 2011

[21] Han, J., & Camber, M. (2000). Data mining concepts and techniques. San Diego, USA: Morgan Kaufman

[22] Koh, H. C., & Low, C. K. (2004). Going concern prediction using data mining techniques. Managerial Auditing Journal, 19(3), 462– 476

[23] Cortes, C., Vapnik, V. 1995. Support vector network. Machine Learning, vol:20 pg.273–297

[24] Burges, C.J.C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, Vol. 2, No.2, pp. 955-974.

[25] Mercer, J. Function of positive and negative type and their connection with theory of integral equations, Philosophical Transactions of theRoyal Society, London A 209 (1909) 415–446.

[26] Clifton Phua, Kate Smith-Miles, Vincent Cheng- Siong Lee and Ross Gayler, “Resilient Identity Crime Detection”, IEEE Transactions on Knowledge and Data Engineering, vol.2, no. 3,pp.533-546, 2012.