Proficient Approaches for Determining Identity Frauds Using Dataset

[1]Vaibhav Sirasat,[2]Sweta Kale
[1]Student,[2]Professor
[1]Department of Information Technology,
[1]RMD Sinhgad School of Engineering, Warje, Pune - 411058, India

Abstract **- Identity crime is well known, popular, and expensive; And a special case of identity fraud crime loan applications. Current non data score cards and business rules mining detection system and match the boundaries of known fraud. These limitations and to combat identity crime in real time this system complemented with multilayered detection. Two additional layers: communal detection (CD) and Spike detection (SD). CD represents actual social relationships to reduce suspicion score. To increase the suspicion score duplicate spikes, and is resistant to examining features characterize this attribute-oriented approach to a variable-size set. Together, CD and SD can detect more types of attacks; better legal practices account for change and remove redundant features, Experimentation on several hundred thousand real CD and SD credit applications. Results are successful credit application fraud data patterns suddenly and in sharp spikes exhibit hypothesis duplicates, Although this research credit application fraud detection, resilience, adaptively and is specific to the concept of quality. To address these limitations and combat identity crime in real time, during this project data mining based method is investigated and extended further. In this project first investigation of existing method is done and further methods are presented to overcome the limitation of effectiveness which will be improve by overcoming the issues related scalability, extreme imbalanced class, time constraints etc. Therefore aiming to present and extend the existing method with goal of improving the effectiveness..**

IndexTerms **- Crime detection, Credit card, Data mining, Identity crime, Scorecards, Fraud detection, Security, Anomaly detection.**

I. INTRODUCTION

Identity crime has become famous because there is so much actual identity data present on the Internet, and confidential data accessible through unsecured mailboxes. It is also easy for hacker to hide their true identities. These frauds can happen in different domains like insurance, credit, and telecommunications as well as other many serious crimes. In addition to this, identity crime is costly in developed countries that do not have nationally registered identity numbers like Social Security Number in US. While fight against fraud, actions fall under two broad categories: fraud prevention and fraud detection. Here Fraud prevention describes measures to stop fraud occurring in the first place. These include different PINs for bankcards, Internet security systems while making transactions using credit card and passwords on telephone bank accounts. On the other side, fraud detection involves identifying fraud as early as possible once it has been occurred. Apply fraud detection once fraud prevention has failed, using detection methods continuously, as this will usually be unaware that fraud prevention has failed. This study is concerned solely with fraud detection. [1], the statistical and insurance domain, a mathematical approach for an a priori classification of objects when group membership is unknown and gives an example of the empirical application of the methodology. Using this technique, Principal Component Iterative Discriminate Technique or the analysis of Principal component of RIDIT scores (PRIDIT), the insurance fraud detector can minimizes uncertainty and increase the possibilities of targeting the appropriate claims so the organization will be more likely to assign investigative resources efficiently to insurance fraud [2].

Data breaches which involve lost or stolen customers identity information can lead to other frauds such as tax returns, home equity, and payment card fraud. The US law requires criminal organizations to notify consumers, so that consumers can ease the harm. As a result, these organizations deserve economic damage, such as notification costs, fines in term of money, and lost business. As in identity crime, credit application fraud has reached a critical group of fraudsters who are highly experienced, organized, and sophisticated. In most of the cases their visible patterns are different to each other and constantly change or not fixed. They are persistent, due to the high financial rewards, and the risk and effort involved are less. Based on unreliable observations of experienced credit application investigators, fraudsters are using software automation to modify particular values within an application and increase rate of successful values.

Here argument is made that each successful credit application cheat pattern is represented by a sudden and high sharp spike in duplicates within the less time, relative to the established baseline level. Duplicates values are hard to avoid from fraudsters' thinking view because duplicates increase their' success rate. Matches (or Duplicates) refer to applications which use common values. There are two types of duplicates: exact (or identical) duplicates have the all same values; approximate (or near) duplicates have some same values (or characters), some similar values with slightly altered spellings, or both. The synthetic identity fraudster has poor success rate, and is likely to reuse fabricated identities which have been successful before. The identity picker has limited time because innocent people can discover the fraud as early as possible and take action, and will quickly use the same real identities at different places.

In next section II the related work is represented over the various methods security at data sharing systems. In section III, the proposed system and its block diagram is depicted. In section IV newly proposed method is represented. Finally conclusion and future work is predicted in section V.

II. RELATED WORK

In the literature survey existing work done related to identity crime detection is discussed. Below some of them are described.

O. Kursun, A. Koufakou, B. Chen, M. Georgiopoulos, K. Reynolds, and R. Eaglin, [1] in this, when the data in hand is not a quality data, a search for specific information by a standard query (e.g., search for a name that is misspelled or mistyped) does not return all needed information. This is an issue of grave importance in homeland security, criminology, medical applications so on. There is a pressing need for name matching approaches that provide high levels of accuracy, while at the same time maintaining the computational complexity of achieving this goal reasonably low. This present ANSWER is a name matching approach that utilizes a prefix-tree of available names in the database. Creating and searching the name dictionary tree is fast and accurate and, thus, ANSWER is superior to other techniques of retrieving name matches in large databases.

L. Sweeney, [2] in this, Consider a data holder, such as a hospital or a bank that has a privately held collection of person-specific, field structured data. Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful? The solution provided by this study includes a formal protection model named k-anonymity and a set of accompanying policies for deployment. A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Datively, Argus and k-Similar provide guarantees of privacy protection.

V. Priyadharshini, G.AdilineMacriga,[3] in this, Today there are millions of credit card transactions are being processed and mining techniques are highly applied to amount transaction and processing then the data's are highly skewed Mining such massive amounts of data requires highly efficient techniques that scaled that can be extend transactions are legitimate than fraudulent fraud detection systems were widely used but this document gives the detection techniques. These contain multilayered techniques for providing the security for the credit card frauds.

In [4], after introducing the concept of learning system, decision tree method has been developed that can deals with continuous data. The decision tree is a table of tree shape with connecting lines to available nodes. Each node is either a branch node followed with more nodes or only one leaf node assigned by classification. With this strategic approach of separating and resolving, decision tree usually detach the complex problem into many simple ones and resolves the sub-problems through repeatedly using, data mining method to discover training various kinds of classifying knowledge by constructing decision tree. The basis of decision tree model is how to construct a decision tree with high precision and small scale. A similarity tree refers to edges are labelled with values of attributes and pertaining nodes that are labelled with attribute names, that satisfy some condition and "leaves", an intensity factor which implies as the ratio of the number of transactions that satisfy these condition(s) over the total number of legitimate transaction in the particular behavior.

In [5], for predictive purposes, genetic algorithms are often acclaimed as a means of detecting fraud. In order to establish logic rules which is capable of classifying credit card transactions into suspicious and non-suspicious classes, genetic algorithm has been suggested that is based on genetic programming. However, this method follows the scoring process. For this purpose, different types of rules were tested with the different fields. The best rule among these is with the highest predictability. Their method has proven results for real home insurance data and could be one best method against credit card fraud. Article by, Wheeler & Aitken, presents different algorithms: diagnostic algorithms, diagnostic resolution strategies, best match algorithms, density selection algorithms, probabilistic curve algorithms and negative selection algorithms. As a conclusion from their investigation that probabilistic algorithms and neighborhood-based algorithms have been taken to be appropriate techniques for classification, and further it may be improved using additional diagnostic algorithms for decision-making in borderlines cases as well as for calculation of confidence measures and relative risk measures.

III. PROPOSED SYSTEM

Problem Definition

To design the application which search for different patterns from real time data in principled fashion to safeguard credit applications at the first stage of credit life cycle.

Communal Detection

- Multi-attribute link: More applications are compared using the link types. Multi-attribute link score to focus on a single link between two applications, not on identical of attributes between the values.
- Single-link score value with average score: Average score is being produced based on the user input value.
- Parameter value change: Determine same or new parameter value by comparing inputs.
- White list creation: Valid user details are stored in the white list and others are discarded and new white list is created.

In CD, any two same applications could be easily interpreted as Multi-attribute link because; detection methods use the similarity of the current application to all former applications as the suspicion score. However, for this particular scenario, CD would also identify these two applications as either Single-link score value with average score or Parameter value alteration by lowering the suspicion score due to the possibility that they are valid. To account for legal behavior and data errors, CD is the white list-based approach between the applications, is crucial because it minimizes the scores of these legal behaviors and false positives. Communal relationship is near duplicates which reflect the social relationships from familial bonds to casual connections like housemates, family members, colleagues, neighbors, or friends etc. Broadly speaking, the white list is created by ranking link-types between applicants by data volume. If the larger the volume for a link-type, then there is higher the possibility of a communal relationship. There are two problems with the white list. One; there can be alert attacks on the white list by fraudsters when they submit applications with synthetic communal relationships. Although it is complicated to make definitive statements that fraudsters will attempt this, it is also not correct to assume that this will not take place. The resolution proposed here is to make the contents of the white list become less predictable. there were two different credit card applications that provided the same postal address, landline phone number, and date of birth, but one stated the applicant's name to be Reena, and the other stated the applicant's name to be Reeta. CD algorithm works in real time by giving scores when there are exact or similar matches between categorical data. CD algorithm is the calculation of each and every linked previous application's score for addition into the current application's score. Therefore, a high score is the result of strong links between two names.

Table 1 Communal Detection Example

Name	ID No	Street	DOB
Reena	10	S.K street	10/12/1980
Reeta	20	S.K street	10/12/1980

Table 2 Communal Below is an example for twins in a home having same address and date of birth and for this a sample white list created by using the table 1 values

No	Link	Score
1	00010	1

Spike Detection

After processing the information the data are then passed for the SD value computation then the weights are calculated by comparing inputs.

- Single-step scaled counts: Elaborates the value exceeds the time difference between each process.
- Single value spike detection: Compute current value score which is based on weighted scaled match values.
- SD attribute selection: At last each attribute weight is automatically updated when processing data stream ends. SD algorithm is the computation of each and every current applications score using all values score and attribute weights.
- CD attribute weights change: At the end of every current discrete data stream process, SD algorithm computes and updates the attribute weight for CD. CD is provided by attribute weights by SD layer .SO both these layers collectively provides the single score for detection any illegal user and their scores are listed to determine the identity crime identification (similar values of the user) is done.

The SD algorithm matches the current application's value against changing window of previous applications' values. It computes the current value's score by combining all steps to find spikes. Then, it computes the Current application's score using all values' scores and attribute weights. When it reaches at the end of the current data stream, the SD algorithm choose the attributes for the SD suspicion score, and change the attribute weights for CD. The capability to perform global search, the study of solutions in parallel, the healthiness to cope with noisy and missing data, and the ability to assess the goodness of the solutions as they are generated. The SD algorithm is the computation of every current value's score by integrating all the steps to detect spikes. Two layers, communal and spike detection do not use external databases, but only the credit application database. Importantly, these two layers are unverified algorithms which are not completely dependent on known frauds but use them only for evaluation.

IV. SYSTEM ARCHITECTURE

The two main challenges for the data mining-based layers are adaptively and use of quality data. These problems need to be addressed in order to reduce false positives. Adaptively is responsible for morphing fraud behavior, as the attempt to observe fraud changes its behavior. But what is not explicit, yet equally important, is the need to also account for altering lawful (or legitimate) behavior within a varying environment. In the credit application field, altering legal behavior is exhibited by communal relationships (such as increasing or decreasing numbers of siblings) and can be caused by outside events (organizational marketing campaigns). Unique data are highly advantageous for data quality and data mining can be enhanced through the real time removal of data errors or noise. The detection system has to clean duplicates which have been re-entered due to human error while entering data or for other reasons. It also needs to remove redundant attributes which have many omitted

values, and other issues. The main contribution of this project is the demonstration of resilience, with adaptively and quality data in real-time data mining-based identity detection algorithms. The first new layer is Communal Detection (CD): the white list oriented or based approach on a fixed set of attributes. To complement and strengthen CD (Communal Detection), the second new layer is Spike Detection (SD): It is the attribute-oriented approach based on a variable-size set of attributes.

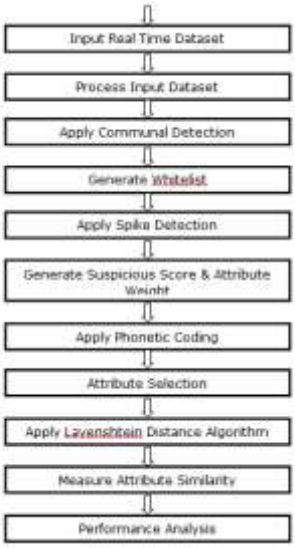

Fig. 1 Proposed System Architecture

In Proposed system, Phonetic code algorithm is applied to calculate the Levenshtein Distance. The Levenshtein distance is a string metric for measuring difference between two sequences. In which it increase identity matching accuracy. The F- measure is greater than CD and SD algorithm. Here only non duplicate data is used for processing. Then calculate the Levenshtein distance by below formula.

The Levenshtein distance between two strings is given by,

(a!=b) = 1 and (a=b) = 0

$$W\text{-}sum(I_1, I_2) = \sqrt{\frac{Lv(Name_{I1}, Name_{I2})^2 + Lv(Addr_{I1}, Addr_{I2})^2 + Lv(IC_{I1}, IC_{I2})^2}{a}}$$

Fig. 2 Levenshtein distance formula

Where, LV(FirstName1, FirstName2) will return the Levenshtein distance for 2 different values of firstname attribute.
Where, LV(Address1, Address2) will return the Levenshtein distance for 2 different values of Address attribute.
Where, LV(IC1, IC2) will return the Levenshtein distance for 2 different values of IC attribute.
Where, W-sum(I1, I2) is a modified weighted-sum function for matching pair of identities I1 and I2.

Also 'a' is a number of available attribute which are non-missing values in both pair of identities.

V. RESULT

During this study, we mainly focused on Proficient Approaches for Determining Identity Frauds Using Data Mining; here we have implemented CD and SD algorithms for actual use to complement the existing detection system. In Phonetic Coding we used Levenshtien distance. This distance is a string metric for measuring the difference between two sequences. Also in result we study F-measure of SD and Phonetic Coding. F-measure of phonetic coding is best than SD F-measure. We can conclude that, by using phonetic code as the attribute value for indexing, the identity matching algorithm F-measure result is increased as compared to non-phonetic indexing value approach. Levenshtein distance based similarity measure however, is performed well in differentiating a matching and non-matching pair of identities as compared to phonetic coding based similarity measure algorithm in many conditions; it also performed faster in a term of identity matching time completion.

.

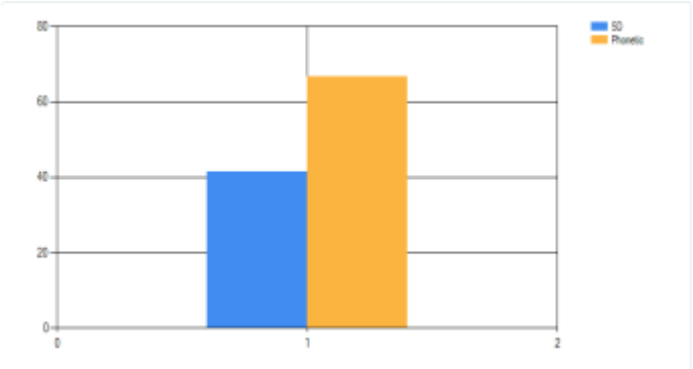

Fig. 1 Study F-measure of (SD & Phonetic Coding)

VI. CONCLUSION

The main focus of this work to study Proficient Approaches for Determining Identity Frauds Using Data Mining; The implementation of CD and SD algorithms is practical because these algorithms are designed for actual use to complement the existing detection system. Levenshtein distance based similarity measure however, is performed well in differentiating a matching and non-matching pair of identities as compared to phonetic coding based similarity measure algorithm in many conditions; it also performed faster in a term of identity matching time completion. The number of identity matching comparison is dependent on window size setting in Array based Sorted Neighborhood indexing approach, in which it went higher when window size is increased accordingly. Main goal is to evaluate effectiveness and efficiency of identity matching algorithm using phonetic coding algorithm in Malaysian identity records.

REFERENCES

[1] O. Kursun, A. Koufakou, B. Chen, M. Georgiopoulos, K. Reynolds, and R. Eaglin, "A Dictionary-Based Approach to Fast and Accurate Name Matching in Large Law Enforcement Databases,"Proc. IEEE Int'l Conf. Intelligence and Security Informatics (ISI '06), pp. 72-82, 2006, doi: 10.1007/11760146.
[2] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, vol. 10, no. 5, pp. 557-570, 2002.
[3] V. Priyadharshini, G.Adiline Macriga,[, An Efficient Data Mining for Credit Card Fraud Detection using Finger Print Recognition
[4] Y. Sahin and E. Duman , "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines " International MultiConference of Engineering and Computer Scientists 2011 Vol I,IMECS2011,Hong Kong.
[5] K.RamaKalyani, D.UmaDevi , "Fraud Detection of Credit Card Payment System by Genetic Algorithm," International Journal of Scientific & Engineering Research Volume 3, Issue 7 , July-2012.