IDENTITY CRIME DETECTION

Vaibhav Kapse
Department of Computer Science
State University of New York at Albany
Albany, New York 12222, USA
vkapse@albany.edu

Shashwat Parashar
Department of Computer Science
State University of New York at Albany
Albany, New York 12222, USA
sparashar@albany.edu

ABSTRACT

Due to a rapid advancement in the electronic commerce technology, Identity fraud has dramatically increased. As it is largest growing fraud in developed countries, cases of fraud associated with it are also rising. To avoid and predict the fraudulent activities, in this paper a method of detecting the fraud over available records with the help of anomalous pattern. This provides effective way to detect fraudulent data from the set of input, the proposed system is an efficient way of matching the data provided by the applicants to predict the fraudsters. The existing process of fraud detection is computationally intensive and difficult in scalability for various data, In order to predict the fraudulent applicants with an appropriate time constraints, we used link scores then matched with the anomalous pattern.

Keywords

Identity fraud, link score, anomalous pattern

1. INTRODUCTION

Identity Theft is a crime in which an imposer tries to steal someone's identity and pretends to be him. The identity theft is classified as a criminal offence globally. Identity theft occurs when someone uses another's personally identifying information, like their name, identifying number, or credit card number, without their permission, to commit fraud or other crimes. The major fields where the fraud occurs are in financial domain (Banking), health sector, telecommunication and other serious crime[2]. There are various ways to commit the identity fraud,one of the most difficult to identify among them is the synthetic identity fraud.The synthetic identity fraud either creates the completely synthetic identity or tries to get some information of an individual and tweak a few values and create a synthetic identity. The synthetic identity creation is complemented with the fact that so much of data is available in the public domain. [3]

The fraudsters make best use of the synthetic identity by utilising the information available to them by either phishing, hacking or unsecured mailboxes. This is further exacerbated by the fact that there are number of countries which do not have nationally registered identification numbers, for e.g. (Social Security Numbers).[4] Many countries need physical presence of the applicant with biometric scanning in order for a person to obtain the credit card, voter id as a rule of law. But these things are slow, time consuming and in future nearly a bottleneck with the fast growing industry,

although the safest and most secure way to minimize the identity theft.

2. PROBLEM DEFINITION

The existing system usually needs every application to be verified by a separate authority which checks for the exact record match by checking in the verification agency. Or else it can be checked against a list of fraudulent records to find if a current application is one of fraudulent nature or not. There is a time gap in which a new user becomes fraudulent and be updated as fraud but this time window may be utilised to obtain other card by either tweaking a few parameters or by changing a few values of the attributes. The nature of the fraud keeps of changing and not necessarily be exactly duplicate of the previous record in the database but changes to change different parameters to keep us guessing.

Generating a completely new synthetic identity does not yield great results for the fraudsters, as it can be caught at many levels of filtering in existing systems. But near duplicates increases their rate of success .Hence it is unavoidable by the fraudster's point of view and almost intricate as a part of their success.

By defining a new algorithm on has to ensure not only the frauds are caught but also it keeps accommodating the changing pattern of the fraud. By this we are able to not only hold the frauds but also able to determine the fraudster pattern and in future might be able to nab him if the pattern is monitored.

3. RELATED WORK

The previous work against which our baseline was done was Resilient Identity Crime Detection[1].This paper was relying on finding certain patterns within the application which form a special relation with each other called a communal relationship. This was based on an assumption that the more the number of repeating pattern in the application, higher the chances of that pattern happens to be a communal relationship.

The example as below.

ID	First Name	Last Name	Address	Apt no	DOB	Phone
1	John	Smith	Collins circle	1	1/1/1982	5187504567
2	Joan	Smith	Collins circle	1	1/1/1982	5187504589
3	Jack	Jones	Manning Blvd	2	1/1/1957	5187894567
4	Ella	Jones	Manning Blvd	2	1/11/1959	5187789567

TABLE 1

The Application ID 1 and 2 if found with same address and same Date of birth and last names are found this application could be that of 2 twins applying .Application ID 3 and 4 if found with same address and same last names are found this application could be that of spouses applying and if these patterns are taken into account thus can be reduced in suspicion. If these reoccurring patterns are encountered ,they may be assumed to be genuine.

The above algorithm on a high level clusters of the genuine applicants on conformity with the patterns that are closely related. The patterns with the communal relationship are mined from an existing database of application. Then any new incoming application is checked for any patterns from the existing applications if they follow any of the communal patterns (relationships).[7] If any pattern is found ,then the suspicion score for the current application is lowered and hence

the communal relationship is obtained for the application.

4. ALGORITHM AND IMPLEMENTATION

The proposed algorithm which we are trying to develop is about finding the anomalous patterns in the already identified fraudulent records and checking if a new application is following the same anomalous pattern or not. If the current application is found to have the same anomalous pattern, it is declared fraud if not then true. For example.

ID	First Name	Last Name	Address	Apt no	DOB	Phone
1	John	Smith	Collins circle	1	1/1/1982	5187504567
2	Joan	Smith	Collins circle	1	1/1/1982	5187504589
3	Jack	Jones	Manning Blvd	2	1/1/1957	5187894567
4	Ella	Jones	Manning Blvd	2	1/11/1959	5187789567

TABLE2

Here we can observe that the fraudster has tried to obtain multiple cards by utilising a fraud pattern by interchanging the content of address 1 with address 2.We observed that the various fraudsters try to synthesize the existing users identity in this manner.

In order to represent these patterns in a meaningful way we have used the multi attribute link.

First Name	Last Name	City	Phone Number
John	Smith	Albany	252536
Jack	Smith	Albany	252536
0	1	1	1

TABLE 3

The pattern as for example "**0111"** is recorded for all the anomalous patterns already existing and hence a list of initial anomalous patterns is made.

Algorithm:

Step 1: Filtering Unique Identifiers like SSN, Passport number.[4]

Using this filter one is able to identify the repeating applications and remove them. This gives an edge of over finding applications which are non-repeating and is better suited for finding anomalous patterns.

Step 2: String Similarity Check:Use of Jaro-Winkler similarity[5]

$$e_k = \begin{cases} 1, & \text{if } Jaro - Winkler(a_{i,k}, a_{j,k}) \geq T_{similarity}, \\ 0, & \text{otherwise,} \end{cases}$$

Tsimilarity =similarity threshold of strings, Tsimilarity=0.8 in our approach. This similarity is utilised to check if any application is how much similar since the fraudsters need to tweak a few parameters like 'first name' and 'last name' by changing the name as 'William' to 'Wiliam' to make this application be treated as a new application .This can be pacified by utilising this.

Step 3.Finding MultiAttribute links: Every application is considered as a record entry and its multi attribute links are found by comparing the records against the entire existing records.[1]

First Name	Last Name	City	Phone Number
John	Smith	Albany	252536
Jack	Smith	Albany	252536
0	1	1	1

TABLE 4

Step 4: Assigning Scores to every link. As we compare every application to 'N' number of application ,there is a possibility of obtaining N multiattribute links .We only obtain only the highest attribute matching patterns.

Score = total no attribute matching in a multiattribute link.

For Example"100111000111" score =7.

Step 5: Matching the multiattribute links identified with the list of existing anomalous patterns.

The anomalous patterns as discussed have been obtained by comparing the known frauds.

If the pattern matches from the list then it is declared Fraud.

AnomalousPattern List.

ID	PatternType	Frequency
1	100111111111111101	10
2	111110111111011111	8
3	100100111111111111	9

TABLE 5

The frequency of the list is the number of times the pattern is observed in the series of current applications.

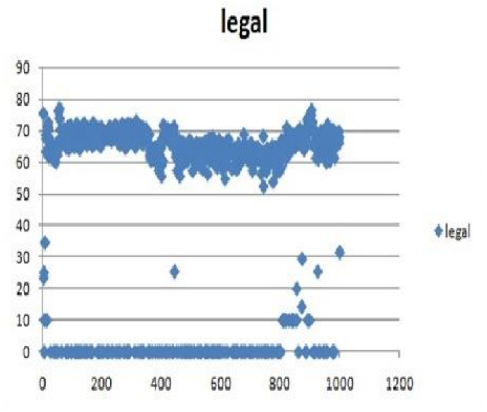
Step 6: Updating the Anomalous Pattern List:

The various highest scoring multiattribute links obtained for every current application are stored in the Anomalous Pattern List and the least frequent Anomalous Pattern are removed from the List.

Hence the changing behaviour of the Anomalous pattern are thus catered.

5. EVALUATION RESLUTS

For the evaluation we implemented our algorithm and the algorithm mentioned in the baseline[1]. The data set had a set of 19 attributes. This real data set chosen because, at experimentation time, it had the most recent fraud behaviour. Although this real data set cannot be made available, there is a synthetic data set of 50,000 credit applications. The graph in fig 2 illustrates the pattern generated over dataset after applying baseline communal detection algorithm. Fig 2(a) shows the legal pattern we obtained ,the X-axis represents the records in which 1000 records are to be considered for the legal pattern generation and the Y-axis denotes the suspicious score achieved through the communal detection algorithm on the dataset. It shows that suspicious score of legal record is below 80 percent means if the new record obtained suspicious score below 80 percent then that record is



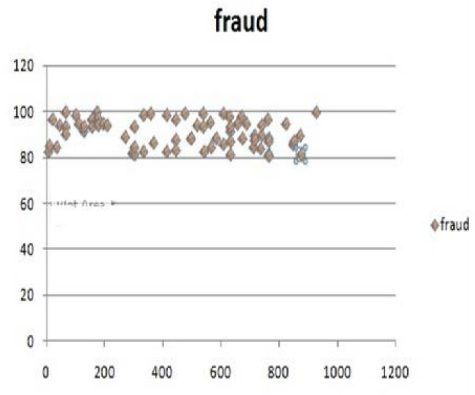


Fig 2(a) : legal pattern obtained from data set after applying baseline algorithm

Fig 2(b) : fraud pattern obtained from data set after applying baseline algorithm

Fig 2: Legal and fraud pattern obtained from data set after applying baseline algorithm over 1000 records (where X-axis =records , Y-axis = suspicious score)

considered to be legal record. Suspicious score 0 indicates that there is no communal relationship with the existing records means it is legal new application. Fig 2(b) shows that the fraud pattern we obtained ,the X-axis represents the records in which 1000 records are to be considered for the fraud pattern generation and the Y-axis denotes the suspicious score achieved through the communal detection algorithm on the dataset It shows that the suspicious score of fraud record is above 80 percent means if the new record obtained suspicious score above 80 percent then that record is considered to be fraud record.

Putting the unique identifier filter on the dataset it was observed that around 35% applications are filtered hence the initial step is able to achieve its objective of removing duplicate entries for a better analysis of the dataset.[6] Out of the dataset of the first month, the known frauds are filtered out and the Anomalous Pattern List is made. The frequency of every pattern is set to one. The one month data is kept as a window of the records to be compared with the next. for example the data for the month of February acts as the database for the applications to be compared for the month of March .Every months applications are checked for their respective multiattribute links and then the step 4 and 5 are applied on them respectively.

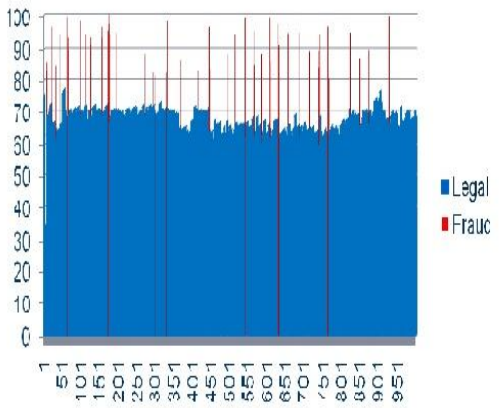


Fig 3(a) The graph presents the Number of legal records and fraud records obtained from dataset over 1000records using baseline algorithm

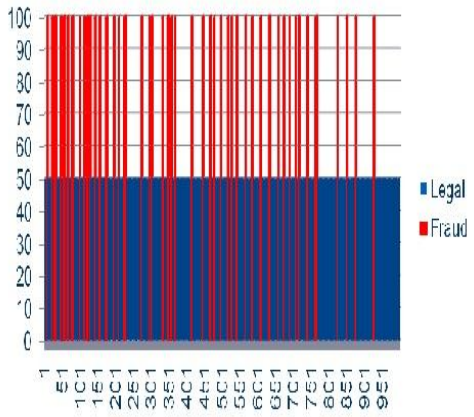


Fig 3(a) The graph presents the Number of legal records and fraud records obtained from dataset over 1000records using proposed algorithm

Fig 3: Number of Legal and fraud records obtained from data set after applying both baseline algorithm and proposed algorithm over 1000 records (where X-axis =records , Y-axis = suspicious score)

It was observed that the algorithm was able to identify the frauds for the entire dataset. To establish the ground truth, a dataset of 1000 records was given an input to both the baseline algorithm and also to the new algorithm. The behaviour of the previously known records from dataset was provided for the training. Hence legality or fraudness of the applications in the dataset was predetermined and only the mapping of the both algorithm had to be observed. The graphs for the same are provided here. The graph in fig 3 illustrates the Number of legal and fraud records found over the dataset after applying both baseline communal detection algorithm as well as proposed algorithm. Fig 3(a) shows the number of legal and fraud records we obtained by using baseline communal detection algorithm ,the X-axis represents the records in which 1000 records are to be considered for finding number of legal and fraud records from dataset and the Y-axis denotes the suspicious score achieved through the communal detection algorithm on the dataset. It shows that the suspicious score of legal record is below 80 percent which indicated by blue colour means the legal record must possess suspicious score below 80 percent[1]. Suspicious score above 80 percent indicated by red colour which are nothing but fraud records means fraud records must possess the suspicious score above 80 percent[1]. Over 1000 records , we found 74 records as a fraud record and 926 records as a legal records.Fig 3(b) shows the number of legal and fraud records we obtained by using proposed algorithm ,the X-axis represents the records in which 1000 records are to be considered for finding number of legal and fraud records from dataset and the Y-axis

denotes the suspicious score achieved through the communal detection algorithm on the dataset. Proposed method does not involve any computational suspicious score as it involves only anomalous pattern matching. Here we considered suspicious score is to be 100 when the new record is matched with anomalous patterns and then that record is considered to be fraud record. Similarly, we considered suspicious score to be 50 when the new record does not match with anomalous patterns and then that record is considered to be legal record Over 1000 records , we found 112 records as a fraud record and 888 records as a legal records.

The baseline was able to evaluate the exact behaviour of the frauds. Our algorithm was also able to identify the same number of frauds but it gave certain false positives because of a couple of patterns were found in the legal records which invoked the fraud check for the same

6. CONCLUSION AND FUTURE WORK

The algorithm is easy to implement, as the computation involved is less compared to the baseline against which this is being compared. The changing pattern of the frauds can be made by looking at the anomalous pattern list and be utilised by someone to identify if the applications is being sent from the same fraudster .However we were unable to draw out any conclusions from the same as we lacked the information of the applications is being submitted from where. We have assumed this to be a case of separate analysis and future work. Scalability of the algorithm was derived by applying on other dataset of Voter Id Fraud.

7. REFERENCES

[1] Clifton Phua, Kate Smith-Miles, Vincent Cheng-Siong Lee, Ross Gayler, "Resilient Identity Crime Detection", *IEEE Transaction on knowledge & Data Engineering* ,vol.24,no.3, pp. 533-546, March2012, doi :10.1109/TKDE.2010.262

[2] B.Schneier, Beyond Fear : Thinking sensibly about security in an Uncertain World. Copernicus, 2003.

[3]G. Gordon,D.Rebovich, K.Choo, and J. Gordon, "Identify Fraud Trends and Patterns:Building a Data Based Foundation for proactive Enforcement," center for Identity Management and Information Protection, Utica College,2007.

[4]J.Jones, "Non-Obvious Relationship Awareness (NORA),"Proc. Identity Mashup,2006.

[5] W, Winkler, "Overview of Record Linkage and Current Research Directions, "Technical report RR 2006-2,US Census Bureau,2006.

[6]R. Wheeler and S. Aitken, "Multiple Algorithms for Fraud Detection, "Knowledge-Based Systems, vol.13,no.3,pp.93-99,2000,doi,10.1016/S0950-7051(000502)

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