ABSTRACT: Identity Crime is well known, established, and costly. Identity Crime is the term used to refer to all types of crime in which someone wrongfully obtains and uses another person’s personal data in some way that involves fraud or deception, typically for economic gain. Forgery and use of fraudulent identity documents are major enablers of Identity Fraud. It has affected the e-commerce. It is increasing significantly with the development of modern technology and the global superhighways of communication, resulting in the loss of lots of money worldwide each year. Also along with transaction the application domain such as credit application is hit by this crime. These are growing concerns for not only governmental bodies but business organizations also all over the world. This paper gives a brief summary of the identity fraud. Also it discusses various data mining techniques used to overcome it.

Keywords: Credit Application, Data Mining, Identity Crime, Online Transaction.

I. INTRODUCTION

Identity Crime may occur when someone steals victim’s personal information, to open credit card accounts using the victim’s name without permission, and charges products to those accounts. Identity Crime is a substitute of unlawful identity change. It points to unlawful activities that use the identity of another person or of a non-existing person as a principal tool for products procurement. Identity Crime can be committed by forging the related documents in two ways first one is synthetic identity fraud which refers to the use of plausible but fictitious identities. These are easy to create but more difficult to apply successfully. The second one is, real identity theft refers to illegal use of innocent people’s complete identity details. These can be harder to obtain but easier to successfully apply. In reality, identity crime can be done with a mix of both synthetic and real identity details [1]. The manufacturing and use of forged identity documents is a financial burden for governments, social welfare institutions, and financial institutions.

Prominently there are two domains where the identity crime is a big apprehension in financial field: Transaction and Credit Application.

Transaction domain is concerned with the identity crime in the online financial transaction which is done through credit card, online banking. Here fraudster does the transaction online through the victim’s credit card or bank details disguised as the victim. In Credit Application domain the identity fraud is when someone applies for a credit card, mortgage loan, home loan with false information.

Fraud that involves mobile phones, insurance claims, tax return claims, credit card transactions etc. denote significant problems for governments and businesses, however detecting and preventing fraud is not an easy task. Fraud is an adaptive crime, so it needs special methods of intelligent data analysis to detect and prevent it. These methods exist in the fields of Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning and Statistics. They propose applicable and successful solutions in different areas of fraud crimes.
Section II gives an overview on data mining in identity crime detection, section III discusses the various data mining methods used to detect identity crime.

II. DATA MINING: OVERVIEW

Data mining is about finding insights which are statistically dependable, unknown previously, and actionable from data [4]. This data must be available, applicable, sufficient, and clean. Also, the data mining problem must be precisely defined, cannot be solved by query and reporting tools, and directed by a data mining process model [5]. Data mining is used to detect to classify, cluster, segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud. So, if data mining results in discovering meaningful patterns, data turns into information and this information is used in detecting anomalies which results in fraud.

The purpose of Data Mining tools is to have a fine-grained, adaptable understanding of the personal data of the people, and of the activity logs of the document issuance system. Data Mining enables an all-inclusive view on the data related to one citizen, from the enrolment step to each transaction made with the identity documents.

Data mining tools take data and construct a depiction of reality in the form of a model. The resultant model describes patterns and relationships existing in the data. From a process orientation, data mining activities fall into three general categories:

Discovery-the process of observing a database to find hidden patterns without a predetermined idea or hypothesis about what the patterns may be.

Predictive Modelling-the process of taking patterns discovered from the database and using them to predict the future.

Forensic Analysis-the process of applying the extracted patterns to find anomalous or unusual data elements.

Data mining is used to construct six types of models aimed at solving business problems: classification, regression, time series, clustering, association analysis, and sequence discovery [6]. The first two, classification and regression, are used to make forecasts, while association and sequence discovery are used to describe behaviour. Clustering can be used for either forecasting or report. Companies in various industries can get a competitive edge by mining their expanding databases for valuable, thorough transaction information.

The use of Data Mining technology helps reduce the workload of analysts and enables them to focus on investigating activities or Individuals that have been tagged as suspicious.

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. The transaction domain concerns with the online transactions which are done mostly through credit card. For the transaction domain following are the methods:

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 behaviour or unusual transactions. These methods model a baseline distribution that represents normal behaviour and then detect observations that show greatest departure from this norm. In supervised methods, models are trained to discriminate between fraudulent and non-fraudulent behaviour 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. Bolton and Hand proposed unsupervised credit card fraud detection, using behavioural outlier detection techniques [5].

Neural Networks- 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 [2] 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 the neuro-adaptive approach [7]. Falcon developed by HNC uses feed-forward Artificial Neural Networks trained on a variant of a backpropagation training...
algorithm[11]. 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 modelling and reduce the number of “false” rejections.

Hidden Markov Model-It is a double embedded stochastic process with used to model much more complicated stochastic processes as compared to a traditional Markov model. If an incoming credit card transaction is not accepted by the trained Hidden Markov Model with sufficiently high probability, it is considered to be fraudulent transactions. In HMM, Baum Welch algorithm is used for training purpose and K-means algorithm for clustering. HMM stores data in the form of clusters depending on three price value ranges low, medium and high[15].

Genetic Algorithm-For predictive purposes, 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, one algorithm that has been suggested by Bentley et al. (2000) that is based on genetic programming. However, this method follows the scoring process. In the experiment as described in their study, the database was made of 4,000 transactions along with 62 fields. As for the similarity, tree, training and testing samples were employed. 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. Chan et al. (1999) has developed an algorithm for prediction of suspect behavior. Origin of their research is that cost model evaluated and rated b whereas other studies use evaluation based on their prediction rate the True Positive Rate (TPR) and the error rate the False Negative Rate (FNR). Wheeler & Aitken (2000) formed the idea of combining different algorithms to maximize the power of prediction. 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. The inspiration for GANN, by combining genetic algorithms with neural networks comes from nature. In GANN, the genetic algorithm is used to find some parameters. Main query is how exactly Genetic Algorithm and Neural Network can be combined. Neural Network has been encoded in the genome of the Genetic Algorithm. In GANN the procedure involves generation of number of random individuals. Designing of neural network is according to the genome information which helps in evaluation of parameter strings. Performance can be easily determined after back-propagation training. To find an optimal network, few GANN strategies rely only on the GA. In this case no training set takes place which are further evaluated and ranked according to parameter performance. Genetic Algorithm (GA) is a search heuristic that copies the process of natural evolution and is used to generate useful and appropriate solutions for optimization problems and search problems. Genetic algorithms (GAs) belong to the larger class of Evolutionary Algorithms (EAs), generate solutions to optimization problems using some techniques such as mutation, inheritance, selection, and crossover.

Logistic Regression-Two advanced data mining approaches, support vector machines and random forests, together with the well known logistic regression [16], as part of an attempt to better detect (and thus control and prosecute) credit card fraud. The study depends on real-life data of transactions from an international credit card operation. It is well-understood, easy to use, and remains one of the most commonly used for data-mining in practice. It thus provides a useful baseline for comparing performance of newer methods. Supervised learning methods for fraud detection face two challenges. The first is of unbalanced class sizes of legitimate and fraudulent transactions, with legitimate transactions far outnumbering fraudulent ones. For model development, 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 imuse; others include directed sampling. The second problem in developing supervised models for fraud can arise from potentially undetected fraud transactions, leading to mislabelled cases in the data to be used for building the model. For the purpose of this study, fraudulent transactions are those specifically identified by the institutional auditors as those that caused an unlawful transfer of funds from the bank sponsoring the credit cards. These transactions were observed to be
fraudulent ex post. The transaction data is aggregated to create various derived attributes.

Decision Trees – It is statistical data mining technique that expresses 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 [9]. Decision tree usually separates the complex problem into many simple ones and resolves the sub problems through repeatedly using [9][10]. Decision trees are predictive decision support tools that create mapping from observations to likely consequences. There are number of popular classifiers construct decision trees to generate classmodels. Decision tree methods C5.0, C&RT and CHAID. The work demonstrates the advantages of applying the data mining techniques including decision trees and SVMs to the credit card fraud detection problem for the purpose of reducing the bank’s risk. The results show that the proposed classifiers of C&RT and other decision tree approaches outperform SVM approaches in solving the problem under investigation.

For the credit application domain, Logistic regression, neural networks, or Support Vector Machines (SVM), cannot achieve scalability or handle the extreme imbalanced class [17] in credit application data streams. As fraud and legal behaviour 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:

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 thresholded nearest neighbour 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 analyse the recovered cases. CBR has 20 percent higher true positive and true negative rates than common algorithms on credit applications [1][8].

CD(Communal Detection) and SD(Spike Detection) – These are the two layers which are used to spot the identity fraud in credit application where first layer, CD algorithm is the whitelist-oriented approach on a fixed set of attributes. It finds real social relationships to reduce the suspicion score and is tamper resistant to synthetic social relationships. For strengthening CD, the second new layer is SD which is the attribute-oriented approach on a variable-size set of attributes. SD discovers spikes in duplicates to increase the suspicion score and is probe-resistant for attributes. Both CD and SD can detect more types of attacks, better reason for changing legal behaviour and eliminate the redundant attributes [1].

IV. CONCLUSION

This paper gives a comprehensive summary on the identity crime and the data mining techniques which are used to detect it in both the transaction and credit application domain in financial field.

REFERENCES


◆◆◆