Abstract

Today, Money Laundering (ML) poses a serious threat and unique challenge to financial institutions. Most of the financial institutions internationally have been implementing Anti-Money Laundering solutions (AML) to fight against money laundering activities. The volume of transaction data in banking is huge and contains a lot of useful information. Detecting money laundering is one of the most valuable information which we can discover from transaction data. Various data mining techniques have been applied in Money Laundering detection system used by financial institutions today. Outlier detection is a data mining technique to detect rare events, deviant objects, and exceptions from client or customer transaction data. This paper, discusses on different outlier techniques, comparison between them and framework for selection of right mining algorithm for implementing the same.

Keywords:
AML, Data Mining, Outlier detection, Money Laundering, LOF.

1. INTRODUCTION

Money Laundering is the process of hiding the identity, source and destination of illicitly-obtained “black” money so as to legitimize it. Banks are commonly used as channels to launder money. The money laundering process can be broken down into three stages as shown in Figure 1.

First, the illegal activity that garners place the money in the launderer’s hands. Second, the launderer passes the money through a complex scheme of transactions to obscure who initially receives the money from the criminal enterprise. Third, the scheme returns the money to the launderer in an obscure and indirect way[27].

Failure to detect and prevent such illegal transactions can invite hefty fines both monetarily and operationally which can prove very costly for the bank and even can make its survival difficult. Conventional rule-based transaction analysis based on reports and tools will not be sufficient to detect more complicated transaction patterns. Here data mining techniques can be applied to dig out transaction patterns that can lead to money laundering.

Outlier detection is a key element for financial surveillance systems which intend to identify fraud and money laundering by discovering unusual customer behavior pattern as shown in Figure 2. Hawkins (Hawkins, 1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it is generated by a different mechanism. An Outlier is a financial transaction activity that significantly deviates from normal transaction pattern or behavior by the client [3].

Fig.1. Money Laundering is process

Fig.2. Outlier Detection
In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations [4].

A key challenge in outlier detection is that it involves exploring the unseen space. At an abstract level, an outlier can be defined as a pattern that does not conform to expected normal behavior. A straightforward approach will define a region representing normal behavior and declare any observation in the data which does not belong to this normal region as an outlier. But several factors make this apparently simple approach very challenging such as defining a normal region which encompasses every possible normal behavior is very difficult.

2. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

Knowledge Discovery in Databases is the process of searching for hidden knowledge in the massive amounts of data that we are technically capable of generating and storing[12]. Data, in its raw form, is simply a collection of elements, from which little knowledge can be gleaned. With the development of data discovery techniques the value of the data is significantly improved. A variety of methods are available to assist in extracting patterns that when interpreted, provides valuable, possibly previously unknown, insight into the stored data. This information can be predictive or descriptive in nature[11]. Data mining, the pattern extraction phase of KDD, can take on many forms, the choice dependent on the desired results. KDD is a multi-step process that facilitates the conversion of data to useful information. Steps in the KDD process are depicted in figure 3.

2.1 DATA CLEANING

This step concerns data quality in the database and the data warehouse. Data must be checked and cleaned prior to moving it forward in the KDD process[5]. Many quality problems are handled at this stage including: outlier or noisy data, missing fields and inaccurate data.

2.2 DATA SELECTION

This phase is very useful for reducing the dimensionalities of the dataset. In the data selection stage, users need to select useful features to represent the data. The selection of such features varies and depends on the goal of the data mining task[10].

2.3 DATA TRANSFORMATION

In this stage, the data is transformed and consolidated based on the specified data mining tasks. Transformation methods include: normalization, aggregation, generational and attribute redesign, which can be used in transforming data[10].

2.4 DATA MINING

This stage refers to the data mining tasks that users tend to adopt in a nominated KDD project. There involve the number of data mining tasks: pattern summarization, classification, clustering and association rule mining. Based on the data mining tasks, there are number of techniques and algorithms that can be used to identify the patterns from the data. This usually results in huge and meaningless numbers of patterns[5].

2.5 PATTERN EVALUATION - INTERPRETATION

Data mining tasks often produce an overwhelming number of meaningless patterns. Users need to evaluate and interpret these patterns to identify those interesting patterns that are relevant to the targeted application[11].

2.6 KNOWLEDGE REPRESENTATION

After locating interesting patterns, users need to encapsulate these patterns in knowledge. This knowledge can be incorporated and represented by users or the system in order to apply this knowledge to unseen data[10]. The final step is the interpretation and documentation of the results from the previous steps. Actions at this stage could consist of returning to a previous step in the KDD process to further refine the acquired knowledge, or translating the knowledge into a form understandable to the user. A commonly used interpretive technique is visualization of the extracted patterns. The result should be critically reviewed and conflicts with previously believed or extracted knowledge should be resolved.[12]

2.7 DATA MINING VS OLAP

Online Analytical Processing (OLAP) and data mining are used to solve different kinds of analytical problems. OLAP is a design paradigm, a way to seek information out of the physical data store. OLAP is all about summation[23]. It aggregates information from multiple systems, and stores it in a multi-dimensional format. For example, it answers operational questions like “What are the average sales of cars, by region and by year?” These could be a star schema, snowflake schema or a hybrid kind of a schema. On the other hand data mines leverage information within and without the organization to aid in answering business questions. For instance, in a telecom industry where customer churn is a key factor, Data
mining would answer questions like, “Who is likely to shift service providers and what are the reasons for that?” They involve ratios and algorithms like decision trees, nearest neighbor classification and neural networks, along with clustering of data[23].

OLAP and data mining can complement each other. Framework of OLAP is shown in figure 4. For instance, while OLAP pinpoints problems with the sales of a product in a certain region, data mining could be used to gain insight about the behavior of the individual customers. Similarly, after data mining predicts something like a 10% increase in sales, OLAP could be used to track the net income[2].

3. DATA MINING TASKS AND TECHNIQUES

The goal of any data mining effort can be divided in one of the following two types:

i. Using data mining to generate descriptive models to solve problems.

ii. Using data mining to generate predictive models to solve problems.

The goals of descriptive and predictive data mining are achieved by using specific data mining techniques that fall within certain primary data mining tasks. The goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patterns and relationships one they have been found.

Clustering is an unsupervised data mining technique. In clustering, instances are divided and grouped into a number of clusters based on the resemblance between instances. Those instances belonging to the same cluster share many characteristics[1]. The objects are grouped together based on self-similarity as shown in figure 6.

Fig.4. Online Analytical Processing

Fig.5. Data Mining Tasks and Techniques

The descriptive data mining tasks characterize the general properties of the data in the database, while predictive data mining tasks perform inference of the current data in order to make prediction steps included as shown in figure 5. Descriptive data mining focus on finding patterns describing the data that can be interpreted by humans, and produces new, nontrivial information based on the available data set. Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation, while the goal of descriptive data mining is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets[2]. The goal of a descriptive data mining model therefore is to discover patterns in the data and to understand the relationships between attributes represented by the data, while the goal of a predictive data mining model is to predict the future outcomes based on passed records models into the following two approaches:

i. Supervised or directed data mining modeling.

ii. Unsupervised or undirected data mining modeling.

The goal in supervised or directed data mining is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The task is to explain the values of some particular field. The user selects the target field and directs the computer to determine how to estimate, classify or predict its value. In unsupervised or undirected data mining however variable is singled out as the target. The goals of predictive and descriptive data mining are achieved by using specific data mining techniques that fall within certain primary data mining tasks. The goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patterns and relationships one they have been found.

Clustering is an unsupervised data mining technique. In clustering, instances are divided and grouped into a number of clusters based on the resemblance between instances. Those instances belonging to the same cluster share many characteristics[1]. The objects are grouped together based on self-similarity as shown in figure 6.
target marketing purposes? A classic clustering technique, which is based on K-means, involves the user initially specifying the number of desirable clusters, as K. Then, based on the ordinary Euclidean distance metric, instances are assigned to the closest clusters.

Summarization is a key data mining concept which involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied for data analysis, data visualization and automated report generation[5]. Clustering is another data mining technique that is often used to summarize large datasets. Summarization can be viewed as compressing a given set of transactions into a smaller set of patterns while retaining the maximum possible information[6]. A trivial summary for a set of transactions would be itself. The information loss here is zero but there is no compaction. Another trivial summary would be the empty set, which represents all the transactions. In this case the gain in compaction is maximum but the summary has no information content[7]. A good summary is one which is small but still retains enough information about the data as a whole and also for each transaction.

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together[3]. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

Sequential patterns analysis is one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data. Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data. Sequential pattern mining is trying to find the relationships between occurrences of sequential events by looking for any specific order of occurrences. In other words, sequential pattern mining is aiming at finding the frequently occurred sequences to describe the data or predict future data or mining periodical patterns. Sequential pattern is a sequence of item sets that frequently occurred in a specific order, all items in the same item sets are supposed to have the same transaction-time value or within a time gap.

Classification is a supervised data mining technique. It aims to correctly classify a set of features related to set classes. The function or the model that emerges between set features and classes in the training data can then be used to predict the classes for new data in the testing set. The accuracy of the model depends on accuracy when assigning a set of features or objects as belonging to classes.

A Decision Tree is a tree-shaped graphical predictive algorithm that represents alternative sequential decisions and the possible outcomes for each decision. This algorithm provides alternative actions that are available to the decision maker, the probabilistic events that follow from and affect these actions, and the outcomes that are associated with each possible scenario of actions and consequences[3]. Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flow-chart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules[3]. When decision trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data. In decision-tree based algorithms, the description of a subset of examples in a leaf node of a tree is uniquely described as a series of feature tests from the root to the bottom of a tree. This approach does not have the flexibility of describing a target concept in different ways.

Rule-based algorithms have the ability to generate multiple descriptions of a concept[3]. Rule induction is one of the most important techniques data mining. Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools of data mining at the same time. Usually rules are expressions of the form:

\[
\text{if} \ (\text{attribute } i, \text{value } i) \ \text{and} \ (\text{attribute } j, \text{value } j) \ \text{and} \ \cdots \ \text{and} \ (\text{attribute } n, \text{value } n) \ \text{then} \ (\text{decision}, \text{value})
\]

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain[19]. Data from which rules are induced are usually presented in a form similar to a table in which cases (or examples) are labels (or names) for rows and variables are labeled as attributes and a decision[4].

Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data[9]. Neural networks essentially comprise three pieces: the architecture or model; the learning algorithm; and the activation functions. Neural networks are programmed or trained to store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill-defined problems. In summary, to estimate sampled functions when we do not know the form of the functions[15]. It is precisely these two abilities (pattern recognition and function estimation) which make Artificial Neural Networks (ANN) so prevalent a utility in data mining. As data sets grow to massive sizes, the need for automated processing becomes clear. With their
model-free estimators and their dual nature, neural networks serve data mining in a myriad of ways [14]. Nearest Neighbor is quite similar to clustering, but it will only look at others records in the dataset that are “nearest” to a chosen unclassified record based on a “similarity” measure. Records that are “near” to each other tend to have similar predictive values as well. Thus, if we know the prediction value of one of the records, we can predict its nearest neighbor. This algorithm works similar to the way that people think – by detecting closely matching examples. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions)[13]. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee.

Regression is a powerful and commonly used algorithm that evaluates the relationship of one variable, the dependent variable, with one or more other variables, called independent variables. By measuring exactly how large and significant each independent variable has historically been in its relation to the dependent variable, the future value of the dependent variable can be estimated[25]. Regression models are widely used in applications, such as seasonal forecasting, quality assurance and credit risk analysis.

A time series represents a collection of values obtained from sequential measurements over time. Time-series data mining stems from the desire to reify our natural ability to visualize the shape of data. Humans rely on complex schemes in order to perform such tasks[26]. We can actually avoid focusing on small fluctuations in order to derive a notion of shape and identify almost instantly similarities between patterns on various time scales. A time series is often the result of the observation of an underlying process in the course of which values are collected from measurements made at uniformly spaced time instants and according to a given sampling rate. A time series can thus be defined as a set of contiguous time instants[21]. Query by content is the most active area of research in time-series analysis. It is based on retrieving a set of solutions that are most similar to a query provided by the user.

Predictive modeling is a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or dependent variable and various predictor or independent variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable.

4. OUTLIER

An Outlier is a data object that significantly deviates from normal objects as if it were generated by different mechanisms[7][8]. Outlier is different from noise as noise is a random error or measured variance and it should be removed before outlier detection. Outlier detection aims to find patterns in data that do not conform to expected behavior.

Outlier detection is extensively used in a wide variety of applications such as fraud detection for credit cards and Money Laundering crime detection[27]. In outlier detection the normal behavior is to characterized by the model and deviation comes from the model is an outlier. The normal behavior model represents the positive patterns that are allowed and also negative patterns that are detecting as an outlier as shown in Figure 7.

A key challenge in outlier detection is that it involves exploring the unseen space. As mentioned earlier, at an abstract level, an outlier can be defined as a pattern that does not conform to expected normal behavior[11]. A straightforward approach will be to define a region representing normal behavior and declare any observation in the data which does not belong to this normal region as an outlier. But several factors make this apparently simple approach very challenging.

i. Defining a normal region which encompasses every possible normal behavior is very difficult.

ii. Often times normal behavior keeps evolving and an existing notion of normal behavior might not be sufficiently representative in the future.

iii. The boundary between normal and outlying behavior is often fuzzy. Thus an outlying observation which lies close to the boundary can be actually normal and vice versa[1].

iv. The exact notion of an outlier is different for different application domains.

v. Every application domain imposes a set of requirements and constraints giving rise to a specific problem formulation for outlier detection.

vi. Availability of labeled data for training/validation is often a major issue while developing an outlier detection technique.

vii. In several cases in which outliers are the result of malicious actions, the malicious adversaries adapt themselves to make the outlying observations appear like normal, thereby making the task of defining normal behavior more difficult[18].

viii. Often the data contains noise which is similar to the actual outliers and hence is difficult to distinguish and remove.

In the presence of above listed challenges, a generalized formulation of the outlier detection problem based on the abstract definition of outliers is not easy to solve. There are number of outlier’s detection methods, categorized into four approaches: statistical approach, distance-based approach, density-based local outliers approach and frequency-based approach [20].
5. TYPE DEGREE AND DIMENSION OF OUTLIERS

An important input to an outlier detection technique is the definition of the desired outlier which needs to be detected by the technique.

5.1 TYPE OF DETECTED OUTLIERS

Outliers can be identified as either global or local outliers. A global outlier is an anomalous data point with respect to all other points in the whole data set, but may not with respect to points in its local neighborhood[4]. A local outlier is a data point that is significantly different with respect to other points in its local neighborhood, but may not be an outlier in a global view of the data set.

5.2 DEGREE OF BEING AN OUTLIER

A data point can be considered as an outlier in two manners, scalar (binary) or outlierness. The scalar fashion is that the point is either an outlier or not. On the other hand, the outlierness fashion provides the degree of which the point is an outlier when compared to other points in a data set[25]. This outlierness is also known as anomaly score or outlier score, which usually can be calculated by using specific measure methods.

5.3 DIMENSION OF DETECTED OUTLIERS

Whether a data point is an outlier is determined by the values of its attributes. A univariate data that has a single attribute can be detected as an outlier only based on the fact that a single attribute is anomalous with respect to that of other data[26]. On the other hand, a multivariate data that has multiple attributes may be identified as an outlier since some of its attributes together have anomalous values, even if none of its attributes individually has an anomalous value.

5.4 NUMBER OF DETECTED OUTLIERS AT ONCE

Outlier detection techniques can be designed to identify different number of outliers at a time. In some techniques, one outlier is identified and removed at a time, then the procedure will be repeated until no outliers are detected. These techniques may be subject to the problem of missing some real outliers during the iteration. On the other hand, for other techniques[14], they can identify a collection of outliers at once. However, these techniques may cause some normal data to be declared as outliers in operation.

6. CHARACTERISTICS OF OUTLIER DETECTION

6.1 USE OF PRE-LABELLED DATA

Outlier detection approaches can generally be classified into three basic categories, i.e., supervised, unsupervised and semi-supervised learning approaches[7]. This categorization is based on the degree of using pre-defined labels to classify normal or abnormal data.

Supervised learning approaches initially require the learning of normality and an abnormality models by using pre-labelled data, and then classify a new data point as normal or abnormal depending on which model the data point fits into. These supervised learning approaches usually are applied for many fraud detection and intrusion detection applications. However, they have two major drawbacks, i.e., pre-labelled data is not easy to obtain in many real-life applications, and also new types of rare events may not be included in pre-labelled data. Unsupervised learning approaches can identify outliers without the need of pre-labelled data [7]. For example, distributed-based methods identify outliers based on a standard statistical distribution model. Similarly, distance-based methods identify outliers based on the measure of full dimensional distance between a point and its nearest neighbors. Compared to supervised learning approaches, these unsupervised learning approaches are more general because they do not need pre-labelled data that are not available in many practical applications. Semi-supervised approaches only require training on pre-labelled normal data to learn a boundary of normality, and then classify a new data point as normal or abnormal depending on how well the data point fits into the normality model. These approaches require no pre-labelled abnormal data, but suffer from the same problem as supervised learning approaches, i.e., a set of representative normal data difficult to obtain in many real-life applications.

6.2 USE OF PARAMETERS OF DATA DISTRIBUTION

Unsupervised learning approaches can be further grouped into three categories, i.e., parametric, non-parametric and semi-parametric methods, on the basis of the degree of using the parameters of the underlying data distribution.

Parametric methods assume that the whole data can be modelled to one standard statistical distribution (e.g., the normal distribution), and then directly calculate the parameters of this distribution based on means and covariance of the original data[24]. A point that deviates significantly from the data model is declared as an outlier. These methods are suitable for situations in which the data distribution model is a priori known and parameter settings have been previously determined. However, in many practical situations, a priori knowledge of the underlying data distribution is not always available and also it may not be a simple task to compute the parameters of the data distribution. Non-parametric methods make no assumption on the statistic properties of data and instead identify outliers based on the full dimensional distance measure between points. Outliers are considered as those points that are distant from their own neighbors in the data set. These methods also use some user-defined parameters ranging from the size of local neighborhood to the threshold of distance measure. Compared to parametric methods, these non-parametric methods are more flexible and autonomous due to the fact that they require no data distribution knowledge. However, they may have expensive time
complexity, especially for high dimensional data sets. Also, the choice of appropriate values for user-defined parameters is not really easy[21]. Semi-parametric methods do not assume a standard data distribution for data, but instead map the data into a trained network model or a feature space to further identify if these points deviate from the trained network model or are distant from other points in the feature space, on the basis of some classification techniques such as neural network and support vector machine.

7. OUTLIER DETECTION APPROACH

Outlier detection approach can be categorized into five approaches which are:

- Statistical approach,
- The distance-based approach,
- Frequency based approach,
- Density based approach
- Deviation-based approach.

7.1 STATISTICAL APPROACHES

Statistical approaches were the earliest algorithms used for outlier detection. Some of the earliest are applicable only for single dimensional data sets. In fact, many of the techniques are single dimensional or at best univariate. One such single dimensional method is Grubbs’ method which calculates a Z value as the difference between the mean value for the attribute and the query value divided by the standard deviation for the attribute where the mean and standard deviation are calculated from all attribute values including the query value[12]. The Z value for the query is compared with a 1% or 5% significance level. The technique requires no user parameters as all parameters are derived directly from data. However, the technique is susceptible to the number of exemplars in the data set. The higher the number of records the more statistically representative the sample is likely to be. Statistical models are generally suited to quantitative real-valued data sets or at the very least quantitative ordinal data distributions where the ordinal data can be transformed to suitable numerical values for statistical (numerical) processing[5]. This limits their applicability and increases the processing time if complex data transformations are necessary before processing.

7.2 DISTANCE BASED APPROACH

In Distance based approach, similarity between two objects is measured with the help of distance between the two objects in data space, if this distance exceeds a particular threshold, then the data object will be called as the outlier. There are many algorithms under this category. One of the most popular and simple to implement is K neighbor technique. This technique operates under the assumption that normal points have several closely located neighbors, while outliers are located far from other points. In the first step a neighborhood for each point is computed, using a distance or similarity measure defined between two data instances[20]. Here Euclidean distance can be used to measure the distances between the point under consideration and every data point in the dataset. In the second step the neighborhood is analyzed to determine if the point an outlier or not. Proximity-based techniques are simple to implement and make no prior assumptions about the data distribution model. In a distance-based approach, the outlier is an object O in dataset T is a DB(p, D) outlier if at least fraction p of the objects in T lies greater than distance D from O. They are suitable for both type 1(Determine the outliers with no prior knowledge of the data) and type 2(Model both normality and abnormality). This approach is analogous to supervised classification and requires pre-labelled data, tagged as normal or abnormal) outlier detection. However, they suffer exponential computational growth as they are founded on the calculation of the distances between all records[18]. The computational complexity is directly proportional to both the dimensionality of the data m and the number of records n. Hence, methods such as k-nearest neighbor (also known as instance-based learning) with O(n2m) runtime are not feasible for high dimensionality data sets unless the running time can be improved[26]. There are various flavors of k-Nearest Neighbor (k-NN) algorithm for outlier detection but all calculate the nearest neighbors of a record using a suitable distance calculation metric such as Euclidean distance or Mahalanobis distance.

7.3 PARAMETRIC METHODS

Parametric methods allow the model to be evaluated very rapidly for new instances and are suitable for large data sets; the model grows only with model complexity not data size. However, they limit their applicability by enforcing a pre-selected distribution model to fit the data. If the user knows their data fits such a distribution model then these approaches are highly accurate but many data sets do not fit one particular model[19]. One such approach is Minimum Volume Ellipsoid estimation (MVE) which fits the smallest permissible ellipsoid volume around the majority of the data distribution model (generally covering 50% of the data points). A similar approach, Convex Peeling peels away the records on the boundaries of the data distribution’s convex hull and thus peels away the outliers. In contrast MVE maintains all points and defines a boundary around the majority of points[24]. In convex peeling, each point is assigned a depth. The outliers will have the lowest depth thus placing them on the boundary of the convex hull and are shed from the distribution model. Both MVE and Convex Peeling are robust classifiers that fit boundaries around specific percentages of the data irrespective of the sparseness of the outlying regions and hence outlying data points do not skew the boundary. Both however, rely on a good spread of the data. Figure 2 has few outliers so an ellipsoid circumscribing 50% of the data would omit many normal points from the boundary of normality.

7.4 DEVIATIONS-BASED

Deviation-Based outlier detection does not use statistical tests or distance-based measures to identify exceptional objects. Instead, it identifies outliers by examining the main characteristics of objects in a group. Objects that “deviate” from this description are considered...
outliers[6]. Hence, in this approach the term deviations are typically used to refer to outliers.

7.5 DENSITY-BASED APPROACHES

Density-based Approaches compares the density around a point with the density around its local neighbors. The relative density of a point compared to its neighbors is computed as an outlier score. Local Outlier Factor (LOF) is the first concept of an object which also quantifies how outlying an object is and the LOF value of an object is based on the average of the ratios of the local reachability density of the area around the object and the local reachability densities of its neighbors[10]. The size of the neighborhood of the object is determined by the area containing a user-supplied minimum number of points (MinPts). LOF’ algorithm proposed a better formulation compared with LOF. Unlike the method of Connectivity based Outlier factor (COF) in which the focus is on outlier detections for low density patterns, this enhancement scheme improves the efficiency and effectiveness of LOF for general datasets. It can be seen that the notion of LOF is quite complex. Three components including MinPts-dist, reachability distance and local reachability density are to be understood before the understanding of the LOF formulation. Local reachability density is an indication of the density of the region around a data point. LOF’ argue that MinPtsdist already captures this notion: a large MinPts-dist corresponds to a sparse region; a small MinPts-dist corresponds to a dense region. LOF” introduced a slight variation in LOF’ and hence it is named as LOF”[10]. Sometimes outlying objects may be quite close to each other in the data space, forming small groups of outlying objects. Since MinPts reveals the minimum number of points to be considered as a cluster, if the MinPts is set too low, the groups of outlying objects will be wrongly identified as clusters. On the other hand, MinPts is also used to compute the density of each point, so if MinPts is set too high, some outliers near dense clusters may be misidentified as clustering points. LOF” uses two different neighbourhoods: Neighbours in computing the density and Neighbours in comparing the densities. In LOF, these two neighbourhoods are identical. In DSNOF algorithm, each object in dataset is assigned a Density-Similarity-Neighbor based Outlier Factor (DSNOF) to indicate the degree (intensity) of outlier possessed by an object. This algorithm calculates the densities of an object and its neighbors and constructs the Similar Density Series (SDS) in the neighborhood of the object. Based on the SDS, the algorithm computes the Average Series Cost (ASC) of the objects. Finally, it calculates the DSNOF of the object based on the ASC of the object and its neighbors.

7.6 FREQUENCY BASED OUTLIER DETECTION

Frequency based Outlier Detection helps us to identify outliers in non-numerical data set too. Statistical, distance-based and density-based approaches work well only for numerical data[7]. When we have data with categorical attributes it is assumed that the categorical attributes could be easily mapped into numerical values. However, there are cases of categorical attributes, where mapping to numerical attributes is not a straightforward process, and the results greatly depend on the mapping that is used, e.g., the mapping of a marital status attribute (married or single) or a person’s profession (engineer, financial analyst, etc.) to a numerical attribute. Frequency-based approaches have been defined to detect outliers in categorical data.

8. MONEY LAUNDERING

Money laundering consists in granting legitimacy to operations on properties, which are criminal offenses. Most industries deal with funds in some ways whether it is cash, cheque, credit card or electronic transfers. In a banking and finance environment all mediums are used, this is why building an AML solution is not an easy task because ML instances are not self-revealing[27]. Instances of ML reporting are likely to be rare. Today, ML activities are more and more sophisticated because of this reason. ML crimes are well hidden within a normal distribution as it mimics normal behavior. Hence, they exist in the large majority of legal transactions. Therefore, data volumes and the nature of ML are challenges to the first generation of AML solutions that are rule-based mechanisms based on predefined sets of fixed thresholds for example, the using mean and standard deviation rules for volume and quantity of transactions in a period of time. Data mining techniques (DM) have been proven to be well suited for identifying trends and patterns in large datasets. Therefore, DM techniques are expected to be applied successfully in AML[26].

9. CHALLENGES OF USING DATA MINING

9.1 DATA QUALITY

In banking and finance, datasets has a different set of quality problems at the instance level. Some of them can be listed as: missing values, dummy values or null. This would happen in most of data fields in all databases except the CID, the customer type (corporate, individual and joint) and the fund name; misspellings, usually typos and phonetic errors. Besides, banking datasets are normally managed in distributed way for the flexibility and security reasons. The independence and heterogeneity of each data source can also be data quality issues when an integrating task is required, as all conflicts must be solved. Fundamentally, data preprocessing step is applied to deal with data quality issues.

9.2 DATA VOLUME AND HETEROGENEITY DATA

Large and growing volume of datasets in financial institutions with regard to the relatively small number of suspicious ML cases in them become a challenge because the analysis of such large volumes is a time-consume tasks for AML experts. Moreover, these large data needed for analyzing is normally not available at a single place[26]. The distribution of datasets requires integration process in
data preprocessing and that can lead to a consequence to data quality issues as mentioned in the section above. Furthermore, financial datasets for investigating ML are usually high dimension and heterogeneous. For instance, a massive dimensional support vector that consists of “customers x accounts x products x geography x time”. Data type of account value is continuous; meanwhile the geography obtains discrete value.

9.3 THE NATURE OF ML

Most industries deal with funds in some ways whether it is cash, cheque, credit card or electronic transfers. In a banking and finance environment all mediums are used, this is why building an AML solution is not an easy task because ML instances are not self-revealing[27]. Instances of ML reporting are likely to be rare. Today, ML activities are more and more sophisticated because of this reason. ML crimes are well hidden within a normal distribution as it mimics normal behavior. Hence, they exist in the large majority of legal transactions. Therefore, data volumes and the nature of ML are challenges to the first generation of AML solutions that are rule-based mechanisms based on predefined sets of fixed thresholds for example, the using of mean and standard deviation rules for volume and quantity of transactions in a period of time.

10. CURRENT DATA MINING APPROACHES IN AML

Clustering is the process of grouping the data into classes so that objects within the same cluster have high similarity and objects within different clusters are very dissimilar. There are different clustering methods in the literature and they have been successfully exploited for scientific datasets, spatial datasets, business datasets, etc. In AML, clustering is normally used for grouping transactions/accounts into clusters based on their similarities. This technique helps in building patterns of suspicious sequence of transactions and detecting risk patterns of customers/account[20]. One of the most challenges in clustering financial datasets is the size, this technique, for instance, should deal with millions of transactions during hundreds/thousands of time instances applied a discretization process on their datasets to build clusters. They map their feature space “customer x time x transaction” to n+2 dimensional Euclidean space: n customer dimensions, 1 time dimension and 1 transaction dimension. They firstly discretize the whole timeline into difference time instances. Hence, each transaction is viewed as a node in one-dimensional timeline space[4]. They project all transactions of customers to the timeline axis by accumulating transactions and transaction frequency to form a histogram as shown in figure 8.

They create clusters based on segments in the histogram. This approach improves firstly the complexity by reducing the clustering problem to a segmentation problem. Next, it avoids the iterative search existing in other clustering algorithms such as K-means[6]. Furthermore, it is more or less appropriate for analyzing individual behaviors or group behaviors by their transactions to detect suspicious behaviors related to “abnormal” hills in their histogram. However, as we have to analyze many customers with many transactions of variety amounts for a long period, it is difficult to detect suspicious cases, as there are very few or no “peak hills” in the histogram. Another global analysis is firstly needed and we can then apply this method for further analysis in this case.

Support Vector Machine (SVM) is a kind of statistical learning that is widely used for classification and regression. AML task involves the detection of unusual behavior of all dimensions (transactions, accounts, product types, etc.). Hence, the AML problem becomes a pattern classification and divides datasets in two parts: normal and abnormal sets. Besides, results of classification depend strongly on the training datasets. Therefore, the training set should be large enough in order to get stable results with high accuracy. Meanwhile, in money laundering, finding a popular training dataset is a challenge. In some financial institutions, for instance, there is only one or two suspicious transactions per month compared to thousands of “clean” transactions per day[17]. This is why SVM, which is a classification method based on small training datasets is suitable for classifying normal and abnormal patterns in AML. Moreover, SVM is also not sensitive to the dimensionality disorder feature that is popular in financial datasets. Traditional SVM is a supervised learning method that requires labeled training datasets to create classification rules. One-class SVM is an unsupervised learning approach used to detect outliers based on unlabeled training datasets which is highly suitable for ML training sets. One-class SVM can be summarized as follow: Given a set of unlabeled training set x1, x2..., xn is in X, and X is chosen in such a way that its most data have a common feature while a small number of elements are outlier. This approach attempts to find a kernel function f where f(x) takes the value +1 with most of the data x in X and it takes the value -1 on outlier. The advantage of one-class SVM as mentioned above is that it requires a small set of unlabeled training data. However, finding efficient parameters for the optimization cost function to avoid the over fitting with a
given datasets is still an open question. Finally yet importantly, the financial datasets is normally heterogeneous with continuous and discrete data. Therefore, additional techniques are needed to extend this SVM-based approach for analyzing heterogeneous datasets. Author exaggerates support vectors to generate an enormous adaptive probabilistic matrix (customer x account x products x geography x time) to compute the likelihood of each customer’s behaviors based on simple weighted aggregations[15]. The association rules can be used to detect hidden relationships between financial transactions based on their co-occurrence. Frequent sequence mining finds patterns of transactions that occur frequently. Regression analysis is used to predict the possibility of an account being used as a conduit for ML based on demographic and behavioral variables.

11. PROBLEM AND OBJECTIVES

The objective is to study different outlier detection techniques and then do comparison by analyzing their different aspects, essentially, time complexity against data size/availability with respect to Money Laundering suspicious activity detection. Identify the various possible permutation and combination of detection techniques through experiment. Various mining techniques will be identified and applied across multiple levels to yield insights[9]. Identify the right technique based on associated data constraints. Look for possible enhancements and improvements to existing outlier detection methods and algorithm based on pattern discovered. Create an intelligent based outliers detection technique, framework, decision table and metrics for implementation. Finding will be examined against the known Money Laundering scenarios. Pattern which are unidentified but had a well-known impact will be further investigated. Extend the research and define a technique to identify those hidden patterns and relationships in data in order to effectively predict future results.

12. METHODOLOGY AND EVALUATION PROCESS

The figure.9 shows structured model that will be applied to evaluate different outlier algorithm based on various data set to derive the target model.

Based on the experimental study over various data attributes and outlier algorithms, final model and framework will be defined. One of the major attribute considered for this experiment is data type and data size.

Based on the assessment the characteristics of choosing the right outlier algorithm will be defined. The figure.10 depicts the interactive flow of defining the outlier algorithm detecting framework for any given data set.

Fig.9. The experimental study over various data attributes and outlier algorithms

Fig.10. characteristics of choosing the right outlier algorithm

13. INITIAL EXPERIMENT – SAMPLE PERFORMED

Initial experiment performed by combining distance-based unsupervised clustering and local outlier detection. Clustering is used for the purpose of pre-processing data for the consequent anomaly identification. As far as the nature of ML is concerned, the chosen clustering algorithm should be able to generate the number of clusters automatically (with no need for pre-establishment) and all the clusters are to be ranked according to the number of the components in each.
Given a point o in the dataset, its LOF value is:

\[
\text{LOF}(o) = |c_i| \times \min \{\text{distance}(o,c_j)\}
\]

Where oC ci, ci C SC, cj C LC.

The higher the LOF value is, the farther the point o deviates from the normal transactional behavioral patterns. Once the LOF value is fixed for each object, we can get to know how suspicious the transactional behavioral patterns are in the given account. Rank the data points as per LOF values, we can get a feature-oriented ordering of Suspicious Money Laundering Transactional Behavioral Patterns (SMLTBPs) to help FIs choose n number of objects as they like for a detailed exploration.

We are more interested in the transactional behavioral attributes like amount and frequency than in the account owner’s subjective characters, thus transaction amount, transaction amount deviation coefficient, and transaction frequency (i.e., withdrawal frequency and deposit frequency) are chosen to be research variables with the following definitions:

Transaction amount (Tai) is the total amount of all the transactional segments or subsequences, that is,

\[
T_{ai} = \sum_{j=1}^{n} t_{aij}
\]  

Where, Ta is the transaction amount of the ith transaction segment and Ta_{ij} is the amount of \(j_{th}\) transaction in the \(i_{th}\) segment. Transaction amount is a critical criterion for us to determine whether a transaction is suspicious or not since large cash transaction is viewed as a special kind of suspicious transaction[16].

Transaction amount deviation co-efficiency (Tadi) is the ratio between Transaction amount variations (Ts) and average transaction amount (Tai), that is

\[
\text{Tadi} = \frac{T_{di}^2}{T_{ai}}
\]  

Tadi is used to measure the degree of equalization of transaction amount (i.e., Structuring) which means a large cash transactions, either deposit or withdrawal, is purposefully divided into several transactions of a nearly equal amount in order to exempt from filing Currency Transaction Report (CTRs) as required by authority[18]. The less the Tadi value is the more equalized the transaction amount is, and the more suspicious the transaction is as far as CTR regulations are concerned.

Withdrawal/deposit frequency is the ratio of the number of withdrawal/deposit transfers to the aggregated frequency of transactions. Analyzing withdrawal frequency and deposit frequency can identify two novel capital flows within a short time frame: one is centralized capital in-transfers followed by decentralized capital out-transfers and the other is decentralized capital in-transfers followed by centralized out-transfers.
Objective is to identify the most interesting data patterns that deviate from the normal operational data. So historical transaction records are to be transformed into several segments or subsequences of neighboring single transactions, with one segment (subsequence) representing one behavioral pattern, and the transactional data embedded in SMLTBPs (suspicious money laundering transactional behavioral patterns) are just the suspicious objects we hope to find out. For each feature as above mentioned, calculate its feature value for each segment and take the feature vectors composed of feature values as research samples.

14. DESIGN OF EXPERIMENTS

In this experiment, 100 sample data have been collated from test environment as experimental samples. After pre-processing the experimental data, obtained 80 experimental samples.

Experiments were performed with the CBLOF algorithm on the sample set. As global outlier detection cannot mine all the outliers, given LOF value to each sample, and then identified n number of samples with the highest LOF values for further investigation and final reporting. Let clustering threshold $\varepsilon = 0.15$ and categorization parameters $\alpha = 75\%$ and $\beta = 4$, categorize LC and SC, and compute LOF values of transaction segments as shown in Table 1.

<table>
<thead>
<tr>
<th>Account, Sample No.</th>
<th>$T_{Ca}$</th>
<th>$T_{Cd}$</th>
<th>$T_{Fw}$</th>
<th>$T_{Fc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 17), (3, 5)</td>
<td>0.411964</td>
<td>0.027473</td>
<td>0.133</td>
<td>0.867</td>
</tr>
<tr>
<td>(1, 20), (10, 5)</td>
<td>0.720696</td>
<td>0.769771</td>
<td>0.973</td>
<td>0.007</td>
</tr>
<tr>
<td>(2, 23), (19, 19)</td>
<td>0.512147</td>
<td>0.017695</td>
<td>0.97</td>
<td>0.13</td>
</tr>
<tr>
<td>(20, 10), (21, 17)</td>
<td>0.357604</td>
<td>0.010843</td>
<td>0.67</td>
<td>0.33</td>
</tr>
<tr>
<td>(20, 14), (21, 6)</td>
<td>1.302777</td>
<td>0.851729</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>(2, 3), (4, 17)</td>
<td>3.200964</td>
<td>0.884096</td>
<td>0.97</td>
<td>0.12</td>
</tr>
<tr>
<td>(6, 7), (7, 4)</td>
<td>2.956707</td>
<td>0.832994</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>(6, 13), (7, 5)</td>
<td>2.890437</td>
<td>0.826036</td>
<td>0.93</td>
<td>0.07</td>
</tr>
<tr>
<td>(12, 10), (13, 20)</td>
<td>2.963899</td>
<td>0.006847</td>
<td>0.73</td>
<td>0.27</td>
</tr>
<tr>
<td>(12, 11), (13, 22)</td>
<td>2.980796</td>
<td>0.002417</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>(5, 2), (9, 9)</td>
<td>3.022736</td>
<td>1.049567</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>(5, 20), (9, 12)</td>
<td>3.522792</td>
<td>1.004896</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>(9, 5), (14, 2)</td>
<td>3.557539</td>
<td>1.015748</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(9, 29), (14, 20)</td>
<td>3.529355</td>
<td>1.049067</td>
<td>0.53</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Five of the samples with top LOF values are listed for each account. They are the five transactions with the highest degree of suspiciousness, as shown in Table 2.

<table>
<thead>
<tr>
<th>Account, Sample No.</th>
<th>LOF Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 17), (3, 5)</td>
<td>0.668371</td>
</tr>
<tr>
<td>(1, 20), (10, 5)</td>
<td>0.439310</td>
</tr>
<tr>
<td>(2, 23), (19, 19)</td>
<td>0.300100</td>
</tr>
<tr>
<td>(20, 10), (21, 17)</td>
<td>0.260535</td>
</tr>
<tr>
<td>(20, 14), (21, 6)</td>
<td>0.984055</td>
</tr>
<tr>
<td>(2, 3), (4, 17)</td>
<td>0.672394</td>
</tr>
<tr>
<td>(6, 7), (7, 4)</td>
<td>0.712111</td>
</tr>
<tr>
<td>(6, 13), (7, 5)</td>
<td>0.379222</td>
</tr>
<tr>
<td>(12, 10), (13, 20)</td>
<td>0.223848</td>
</tr>
<tr>
<td>(12, 11), (13, 22)</td>
<td>1.005104</td>
</tr>
<tr>
<td>(5, 2), (9, 9)</td>
<td>0.1015748</td>
</tr>
<tr>
<td>(5, 20), (9, 12)</td>
<td>0.53</td>
</tr>
<tr>
<td>(9, 5), (14, 2)</td>
<td>0.47</td>
</tr>
</tbody>
</table>

15. CONCLUSIONS

This paper, presents a comprehensive taxonomy framework for contemporary data mining techniques and outlier detection techniques based on simple finance domain of Money Laundering together with specific semantics of data. Also, it discusses about detail performance evaluation of each class of outlier detection techniques together with an simple initial experiment.

Based on initial experiment result, making a good use of the advantages of both distance-based unsupervised clustering and local outlier detecting, the CBLOF (Cluster-Based Local Outlier Factor) algorithm can effectively identify the synthetic data suspicious of ML transactions with a high processing speed and a satisfactory accuracy. Needing neither prior samples to serve as training data nor the number of clusters to be designated in advance can solve the problem that AML research is always in short of case data. In particular, the algorithm is self-adaptable to the evolution of ML methods and can recognize SMLTBPs(suspicious money laundering transactional behavioral patterns) that haven’t been detected before, which is quite beneficial in saving limited investigation resources and preventing FIs from filing defensive SARs. However, only a few transactional behavioral features of amount and frequency are studied in this first experimental study, so relative subjective efficiency in similar situation is needed to be evaluated and compared.

16. REFERENCES


AUTHOR

Sathasivam Kannan has about twenty two years of extensive professional experience in Information Technology industry in managing and leading wide range of program/Project management function across industry, globally involving technology application systems development, IT transformation, enterprise integration, ERP system implementation, B2B/B2C implementation and Data warehousing mainly for finance and supply chain domain.

Dr. K. Somasundaram, Research Guide (Karpagam University), is a Professor with the Department of Computer Science and Engineering, Jaya Engineering College, Thiruvallur, Tamilnadu.