Fraud Detection using Dual Distance Method

Dinker G. Mattam
Analytics Data Labs
DXC Technology
Bangalore, India
dinker.g.mattam@hpe.com

***Abstract*—Fraud is a significant problem faced by companies across the world. In securities trading, fraud can lead to major losses and regulatory sanctions. Detecting fraud from millions of transactions through manual effort is next to impossible. Machine learning provides a scalable alternative through automation or partial automation of the fraud detection process.**

This paper discusses a novel unsupervised anomaly detection technique to identify fraud transactions. We present a distance-based method, which we believe has advantages over other related conventional techniques, and is faster than density based methods. It is a semi-automated process, which leverages on domain knowledge to choose the input attributes. We also add an ensemble transformation to capture various types of outliers. We implement the system on a real world dataset and identify suspicious trades. We think the technique is suitable for fraud detection in a wide range of applications not limited to financial services domain.

Keywords—fraud; anomaly; outlier; detection; machine learning; unsupervised; distance; clustering; automation

I. INTRODUCTION

Fraud is a major problem faced by companies across the world. This paper deals with detecting fraud in the commodity trading domain. Trading fraud leads to significant losses to businesses and affects efficient price discovery.

In regulated markets, it is the responsibility of the corporates to ensure that all transactions on their behalf are fair. Corporates usually have compliance sections which make use of domain knowledge to track and monitor transactions for fraud. Since transactions number in millions, it would be next to impossible to scrutinize all of them manually. So the compliance departments review samples of transactions for various known modes of fraud. This domain knowledge based fraud detection is tedious, usually not comprehensive and not scalable. The clustering method expounded in this paper can be used independently or in alliance with domain knowledge-based expert systems. It scales, saves time and is agnostic to the type of fraud.

The major challenge that machine learning based fraud detection systems face is that businesses seldom have clear or enough examples of fraudulence in the past. Unavailability of previously known instances makes it difficult to apply supervised techniques. Frauds are also diverse and dynamic in nature rendering the application of supervised algorithms rigid and sometimes antiquated. In this paper, we use an unsupervised method for fraud detection, which does not require labeled examples of fraud. The method outlined here is generic and can be applied to various cases and industries.

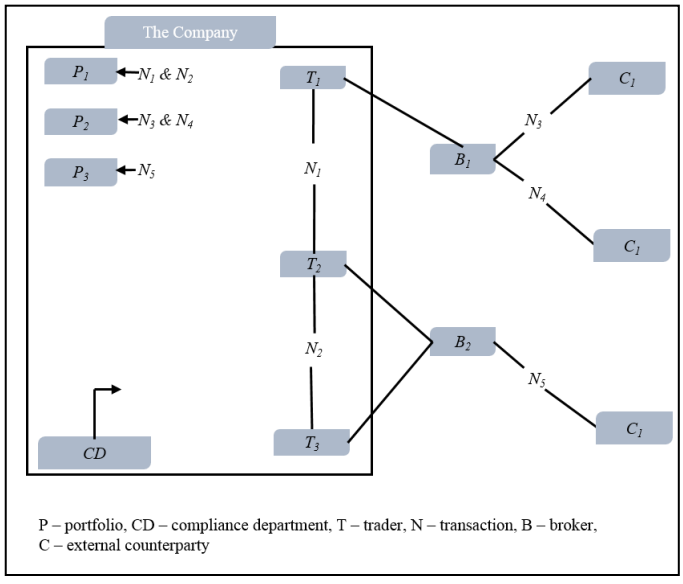
A. Background and Objectives

For the purpose of this work, we use the real use case of a commodity trading company. The company has a group of employees who trade on its behalf. Trading in this context has the twin objectives of hedging and generating profits. Fig. 1 exhibits the sample structure of trading outfit within the company.

Fig. 1 shows example traders, T_1, T_2, and T_3. The traders execute transactions between themselves (transaction labeled N_1 and N_2), and also with external traders C_1 and C_2 (transactions labeled as N_3, N_4, and N_5). Brokers B_1 and B_2 mediate the transactions. The trader books proceedings from transactions into various portfolios P_1, P_2, and P_3 which are also known as trading books. A portfolio is akin to a ledger that holds financial instruments or assets with some common features that bind them together. The value of the portfolio is updated periodically based on the market price of the assets that it holds.

Fig. 1. Structure of a trading outfit

The traders place most of the transactions in forward markets. Forward market is an informal financial market



through which traders enter into contracts for future delivery. Fig. 2 Shows the example of a forward contract. The traders transact using forward contracts for a variety of underlying assets such as physical commodity, energy, and LNG.

The two parties agree on a forward contract (which we call a deal) at time point t_1 as shown in Fig. 2. As per the

agreement, delivery of the underlying assets in lieu of payments are carried out at time points, t_2, t_3, and t_4. Many such transactions usually make up a deal. Here, t_1 is the initiation date of the deal, t_2 is the delivery start date, and t_4 is the delivery end date. We define time horizon of the deal as $t_2 - t_1$, and delivery length of the deal as $t_4 - t_2$. The amount received at the time of delivery is based on a specific price curve or an index, which in turn is mutually decided and specified on t_1. Sometimes counterparties make amendments to the deal features between t_1 and t_4.

We start out with a dataset that provides the date, location, delivery start and end dates, traded volume, currency, underlying asset class, identifiers for the deal, trader, portfolio, broker and counterparty and the basis price curve among other information for a large number of historical transactions. It contains information for about a million deals and does not have fraud identifier labels for any of them. The fraud detection system that we propose uses this data to identify:

1. Fraud deals and a degree of confidence with which we categorize them as fraudulent
2. Traders who are most likely to have been involved in fraudulent deals

Fraud comes in different forms; the popular ones from the perspective of a commodity trader include:

- Insider trading – using confidential information to gain advantage through trading
- Fake deals to boost action in certain counters
- Increase the quoted price by a significant margin and retract to help an accomplice
- Non-delivery of goods in forward contracts
- Internal deals between traders to boost profits for one of the parties and thus increase bonus.

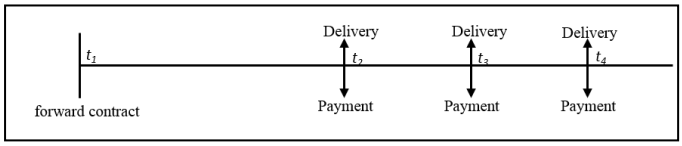


Fig. 2. Example of a forward contract

B. Literature Review

Research in machine learning for anomaly detection is rich with numerous techniques. For the purpose of this paper, we restrict the review of existing research to unsupervised methods used to find anomalies in the so-called metric datasets [1], which are datasets without pre-existing linkages across records (for example, time series does not qualify). We review the methods suitable to deal with such datasets for which the concepts of "distance" and "density" have direct relevance.

Clustering is one of the most popular unsupervised techniques used in anomaly detection. Research has shown that k-means can be used as a unified approach for clustering and outlier detection [2]. Unlike the standard k-means clustering, the unified approach recalculates centroids by removing the outliers in each iteration. In this way, the process ensures that outliers do not influence the centroids or the cluster shapes. Post the clustering exercise, the outliers are identified based on the distance to the nearest cluster centroid. This modified k-means technique forms the basis of our fraud detection system. We use a procedure that is improvised from the standard distance based anomaly detection method.

Another popular technique for outlier detection is the local outlier factor (LOF) [3], a density based method. The method estimates the density of the neighborhood of each data instance and declares those that lie in low-density regions as anomalous. Although LOF is an effective method for outlier detection, it has constraints around the choice of parameter values and computational memory requirements. Also, density-based methods may not identify anomalous but dense clusters. The LOF method falls into the broad category of the nearest neighbor based anomaly detection methods. At the core, nearest neighbor based techniques model the k nearest neighbor distances as a mixture distribution and identify the outliers [4].

Close on the heels of the nearest neighbor category is another technique known as peer group analysis, which detects individual accounts that begin to behave in a way distinct from accounts to which they had previously been similar [5]. This technique ideally uses time series or longitudinal data. From the perspective of a trading application, the method can be useful for a trader-wise analysis. However, data in each group should have a reasonable scale for the peer group analysis to work well. In the case of our current example of trading fraud where we have only a few trades for many of the traders, the technique may not be comprehensive.

Another popular method for outlier detection when past anomaly examples are unavailable is one-class classification. This category of methods assumes that all training instances have only one class label. Such techniques learn a discriminative boundary around the typical cases using a one-class classification algorithm, for example, one-class support vector machines [6, 7]. Any test case that does not fall within the learned boundary is declared as anomalous. The method is more suited to identifying global outliers than local ones.

Isolation-based anomaly detection is a relatively new unsupervised technique, fast gaining in popularity [8]. It is a tree-based method and works on the principle that because of the susceptibility to isolation, anomalies are more likely to be isolated closer to the root of a tree, whereas regular instances are more likely to be isolated at the deeper end. The proposed method, called Isolation Forest builds an ensemble of trees for a given data set. Instances with short average path lengths on the trees are classified as anomalies.

A lot of practical anomaly detection focusses on the statistical techniques, based on the assumption that regular data instances occur in high probability regions of a stochastic model, while abnormalities take place in the low probability regions [7]. The statistical techniques include both parametric and non-parametric ones. Non-parametric methods (for example, histograms) are more relevant to scenarios such as the trading fraud case considered in this paper. Nonetheless, many of these techniques, including the histogram-based one, lack the ability to account for multi-variate linkages.

Another interesting statistical technique for outlier detection uses Chebyshev inequality, which does not require any assumption regarding the distribution of the underlying data [9]. We leverage this technique as a part of the outlier detection system presented in this paper.

II. Our Approach

We make an assumption that fraudulent deal is equivalent to an anomaly. The assumption has its basis in domain knowledge, which hypothesizes that deals that are very different from the others in one or more dimensions are suspicious. Fig. 3 shows the process flow of our approach.

The primary challenge in outlier detection is the high level of heterogeneity of the deals. Broadly, we go about the outlier detection process in two phases:

1. Partition the deals into groups with reasonable homogeneity
2. Identify outliers in each group and obtain a standardized outlier score

As shown in fig. 3, in the beginning, we aggregate the transactions into deals with all the relevant attributes. We do not categorize the deals data as sequential because of the absence of any periodicity. We start out with about 1 million independent records from which we aim to flag a few as likely to be fraudulent.

A. Partitioning into Homogeneous Groups

Before outlier detection, we segregate of the deal universe into similar groups. This segregation is a critical preprocessing step as it has a direct influence on the probability of deals getting identified as outliers. We can perform the partitioning in two ways – using analytical techniques or using domain knowledge. In the trading fraud use case, we leverage domain knowledge as there is an available clear basis for partitioning into homogeneous groups. We use the portfolio as the core unit for grouping. Deals in a portfolio align on many dimensions and have shared characteristics. Since we have too many portfolios, we further combine them into homogeneous groups, again leveraging the available domain knowledge.

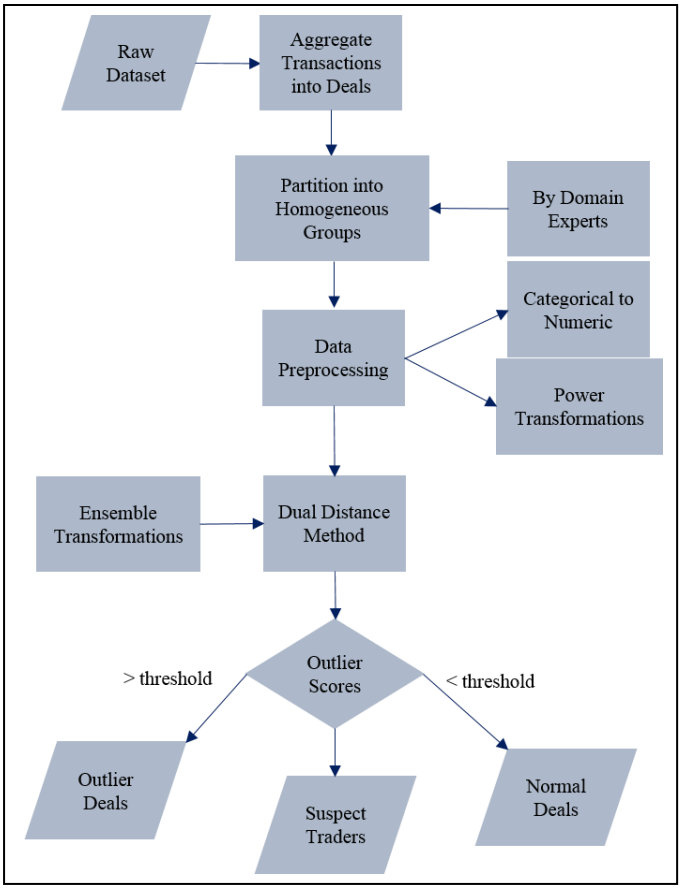


Fig. 3. Process flow

Alternatively, one can use analytical methods to segregate the instances into homogeneous groups. Clustering is an obvious candidate as a relevant analytical method for this purpose. However, one has to be careful on the choice of clustering dimensions here. A wrong choice of dimensions could lead to errors in the following outlier detection step. Also, the attributes used in clustering cannot be used again in the outlier detection process.

From the next step onwards, we explain the process as applicable to a single homogeneous group obtained from the partitioning process. We repeat the process across groups, one by one.

B. Data Preprocessing

We choose the dimensions for outlier detection based on discussions with domain experts on the most likely attributes that fraudulent deals outlie on. We use the following attributes for the trading use case:

- Time horizon of the deal
- Delivery length of the deal
- Number of amendments for the deal
- Location
- Currency

- Underlying asset class
- Basis price curve

Among the features mentioned above, the last four variables are of the nominal type. For the purpose of clustering, we convert these into numeric values based on rarity. Each categorical label, L_{ij} within the nominal variable j is mapped to a numeric value V_{ij} as follows:

$$V_{ij} = 1 - freq(L_{ij}) / n$$

where $freq(L_{ij})$ is the number of records for which the nominal variable j takes the label L_{ij}, and n is the total number of records. So the corresponding numeric value represents how rare the label is within the particular column.

A major challenge with the kmeans clustering based approach is the likely skewed distributions of the attribute dimensions. The presence of skewed distributions breaks the assumption of k-means clustering and gives more importance to some of the attributes than others. In other words, if applied without any transformation, most of the outliers generated using the approach tend to be the deals that outlie along the highly skewed dimensions. We can remedy this by applying power transformations that bring the distributions as close to normal as possible. Power transformation is a family of functions used to create a monotonic transformation of data using power functions [10]. In the trading fraud example, we use Box-cox transformation, a traditional method of power transformation.

C. Outlier Detection using the Dual Distance Method

The outlier detection process starts with the application of the unified approach to clustering and outlier detection using the k-means algorithm (henceforth referred to as the unified kmeans) [2]. As mentioned in Section I.B, unlike the standard k-means clustering, the unified kmeans approach recalculates centroids by removing the outliers in each iteration. The process outputs k deal clusters, d_{c1}, d_{c2}, …, d_{ck}. We decide the number of deal clusters k, based on the variance explained, popularly known as the elbow method [11].

In the unified method, outliers are identified based on the distance from the nearest cluster centroid. We modify this in the novel dual distance method by considering both the distances from the nearest cluster centroid and that from the overall center of the homogeneous group D_{g1}. All the distances are evaluated in the space defined by the dimensions chosen for clustering. Formally, for a deal d_i in the deal cluster d_{ci}, we look at the composite distance defined as:

$$dist_{composite} = (dist_{centroid}^{\ 2} + dist_{center}^{\ 2})^{1/2}$$

where $dist_{centroid}$ is the Euclidian distance from d_i to the centroid of the nearest cluster d_{cj}, and $dist_{center}$ is the Euclidean distance from d_i to the center of the homogeneous group D_{g1}. The degree of abnormality of each deal is a function of the composite distance $dist_{composite}$.

We believe the dual distance method provides a few advantages over unified kmeans:

- The approach considers both the abnormality of each deal d_i in the context of the other deals and at how different d_i is with respect to the overall characteristics of the homogeneous group D_{g1}.
- The approach can capture deals with characteristics distinct from that of the most likely deals in the group D_{g1}, as outliers. In the unified kmeans method, these deals may not be identified as outliers if there are many of them, as they would form a cluster and the $dist_{centroid}$ would be small.
- On the other hand, if the deal d_i is far away from the nearest cluster, but has characteristics close the most likely ones of D_{g1}, the unified kmeans approach will classify it as an outlier while the dual distance method may not. We believe it is not a bad outcome, as the method assigns the most likely characteristics of D_{g1} its due importance in the outlier detection process.

When compared to the density based outlier detection approach, the dual distance method is much faster and lighter.

Though the dual distance method makes only a minor change to the unified kmeans approach, we believe it leads to a non-trivial shift in the outlier detection process. We have not seen this approach used in the past. We can extend the dual distance approach to many other situations with an appropriate choice of features as dimensions for clustering.

D. Ensembling Transformations to capture more Outliers

A challenge with the technique is from the perspective of the nature of outliers. It is quite unlikely that the fraudulent deals outlie in all dimensions. Fraudsters try their best to make the deals look as typical as possible, which means, many of the deals may outlie on just one or two dimensions. The composite distance measure may not identify such outliers.

To remedy this challenge, we devised a novel approach of using the Box-cox exponents as a set of tunable parameters. Let us say we have an optimized vector of Box-cox exponents, **b** as shown below:

$$\mathbf{b} = \begin{matrix} b_1 & b_2 & b_3 & b_4 & b_5 & b_6 & b_7 \end{matrix}^{\mathrm{T}}$$

where b_i is the optimized box-cox exponent corresponding to the i-th dimension used in clustering.

In the typical process, we use vector **b** to transform the attribute values and then identify a set of outliers through clustering. This process most likely yields outliers that are abnormal across all or many dimensions. To broaden our search for outliers, we propose scaling the vector b in different ways and repeating the outlier detection process. For instance,

we can use matrix **A** shown below, to scale the vector **b** in various ways.

$$\mathbf{A} = \begin{matrix} 1.3 & 1.7 & 0.5 & 0.4 & 1.7 & 1.7 & 1.6 \\ 0.6 & 1.5 & 1.3 & 0.1 & 0.9 & 1.2 & 1.4 \\ 0.8 & 0.2 & 1.2 & 1.5 & 1.6 & 1.4 & 0.2 \\ 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{matrix}$$

Broadcasting **b** and operating **A** on **b** yields a matrix **C** of dimension 5 × 7. Each row of matrix **C** gives a new set of exponents with which we can transform the attribute values. Clustering after each of the five transformations generates five sets of outliers O1, O2, O3, O4 and O5. We define the final set of outliers O as,

O = O1 U O2 U O3 U O4 U O5

The first three rows of A has scaling factors generated from a uniform distribution between 0 and 2. The fourth row, a set of 1s, keep the original exponent vector b as it is, while the last row of 0s leads to direct clustering without transformation of attribute values. To further broaden the set of outliers we can increase the number of rows in the matrix A, with more randomly generated scaling factors. This approach of expanding the set of outliers is particularly suitable for applications like the trading fraud where the cost of false positive is much lesser than that of false negatives. Also, the use of scaling factors to the exponents does not lead to a linear increase in the number of outliers, as these sets usually have a high degree of intersection.

An alternative to applying scaling factors on the Box-cox exponents is to assign random sets of weights to scale the attribute values. This approach would have an equivalent effect on the outlier detection process. Nonetheless, we believe that our method of using the Box-cox exponents as tunable parameters provides a much more intuitive way to think about repeated clustering to expand the outlier list. The optimized exponents' vector forms a basis that gives almost equal importance to various attributes, and scaling this vector provides more control on the transformations than using random weights.

E. Obtaining Standardized Scores

The composite distance values obtained from the dual distance method provides a comparable score for the degree of abnormality of deals within a homogeneous group. But a user would prefer scores on a universal scale allowing comparisons of deals across the various groups. We can use z-scores to standardize the composite distances. However, the application of z-scores is limited to normal distributions. So we propose using Chebyshev inequality to arrive at standardized outlier scores.

Chebyshev inequality guarantees that, for a wide class of probability distributions, no more than $1/z^2$ of the distribution's values can be more than z standard deviations away from the mean [12]. We calculate a Chebyshev confidence score, *CS* defined as below:

$$CS_i = sign(z_i) \times \min(1 / z_i^2, 1)$$

$$z_i = (dist_{composite_i} - x_g) / s_g$$

where CS_i is the Chebyshev confidence score for the deal d_i, $dist_{composite_i}$ is the composite distance for the deal d_i (defined in section II.C), x_g and s_g are the mean and standard deviation respectively of the composite distance for all the deals in the homogeneous group D_{g1}.

We can think of the unsigned part of CS_i, the Chebyshev confidence score, similar to the p-value, ranging from 0 to 1. One can interpret this value as the probability that the deal belongs to the homogeneous group D_{g1}. A value close to 0 implies very low confidence that the deal d_i belongs to D_{g1}. However, we are only interested in the CS_i values that represent high composite distances, as the ones with small distances are typical deals in the group with proximity to the centroid of the nearest cluster and the center of the homogeneous group D_{g1}. To incorporate this aspect, we define CS_i in such a way that it carries a sign. From the outlier detection perspective, we are interested in deals that have positive signed CS_i with value zero or very close to zero. With the CS_i available, the user has the option to set thresholds based on the desired confidence interval, to classify deals into typical and atypical.

F. Evaluating New Deals for Fraud

In the previous sections, we examined the dual distance method used to flag outliers or possible fraud instances among a large historical population. We can easily extend the technique to evaluate new deals on a live basis, which is more relevant from a practical point of view. Ideally, we conduct this validation on a new deal, once all the transactions pertaining to the deal are concluded. We follow the steps described below:

1. Create the feature vector for the new deal
2. Transform the feature vector using the sets of exponents (as in matrix **C** presented in section 2.3). We obtain as many transformed feature vectors as the number of rows in **C**.
3. For each transformed feature vector, find the composite distance ($dist_{composite}$ as explained in section II.C) on the clustered space of historical deals, with a similar transformation.

Obtain the Chebyshev confidence scores for the deal, and flag as an outlier or a fraud suspect if any of the scores are higher than the chosen threshold.

G. Identifying Suspect Traders

We use an indirect method of rolling up the obtained deal outlier scores to identify the suspect traders. A roll up of deals at the trader level helps us get the trader outlier score *TS*, defined as below:

$$TS_i = x_i + ws_i$$

$$w = x / s$$

where TS_i is the trader outlier score for the i-th trader, x_i and s_i are the mean and the standard deviation of outlier scores of deals executed by the i-th trader, and x and s are the mean and standard deviation of the outlier scores of the entire deal population. We use w as a weight to give equal importance to both mean and standard deviation of outlier scores.

Higher the TS_i, more suspect the trader. Along with the mean deal outlier score, we believe standard deviation is also an important determinant of trader fraud, as deals executed by a trader normally should not have very high variability. We can get further insights by creating a scatterplot of traders across x_i and s_i.

III. Results and Discussion

Application of the dual distance method and its extension to identify suspect traders result in the following broad outputs:

- Deal outlier scores, and a list of outlier deals based on the chosen score threshold
- Trader outlier scores, and a list of suspect traders based on a threshold

We had about 0.5% of the deal population or about 5000 deals flagged as outliers. A reduction in the sample from a million deals to 5000 deals represents a huge saving of the manual effort for fraud checking.

A. Does the Outlier Set Capture Fraud?

The significant test for the effectiveness of the outlier detection system is whether the outliers we obtained include fraudulent deals. We could not perform this, because of the unavailability of such deal examples from the past. But since the test is vital to the conclusions from the exercise, we got domain experts to create artificial deals with fraudulent attributes so that we can test the model on those deals. The domain experts created the artificial deals based on the various known modes of fraud.

We applied the dual distance method for outlier detection on the enhanced set of deals including 100 artificial deals with fraudulent characteristics. The test yielded an outlier list that contained about 5000 deals, out of which 60 were the newly added artificial deals. Assuming all the other deals excluding the newly added ones were genuine, we had a true positive rate of 60% and a false positive rate of about 0.5%. Though the result is not a conclusive proof for the accuracy of the model considering the unavailability of the nature of the deals other than the newly created ones, the test returns positive signals.

In addition to the above, we tested the technique for two lower level objectives:

1. Examine the effectiveness of the dual distance-based outlier detection method through visual inspection.
2. Check whether the use of Box-cox exponents as tunable parameters help in capturing outliers of diverse characteristics.

B. Visual Representation of the Dual Distance Method

To verify the effectiveness of outlier detection using the dual distance technique, we create a biplot, which gives an alternate visual perspective. The biplot is a scatterplot across the two principal components of the dimensions used in outlier detection. Dimensionality reduction to the two principal components leads to loss of some information. But since the two principal components explain the maximum possible variance in the data, they represent the core underlying characteristics of the features.

Fig. 4 demonstrates the dual distance method used to identify outliers from a homogeneous group of deals. Fig. 5 shows the outliers on a newly transformed space. The group shown in the figure has a total of 179,753 deals, out of which we identify 77 as outliers.

In both Fig. 4 and Fig. 5, we mark the outliers in red. In Fig. 5, we have a visual check whether the detected outliers look abnormal when viewed from a different perspective involving fewer dimensions. Based on visual inspection, we conclude that outliers in fig. 5 more or less match those that we would visually pick if all the deals were plotted on such a biplot.

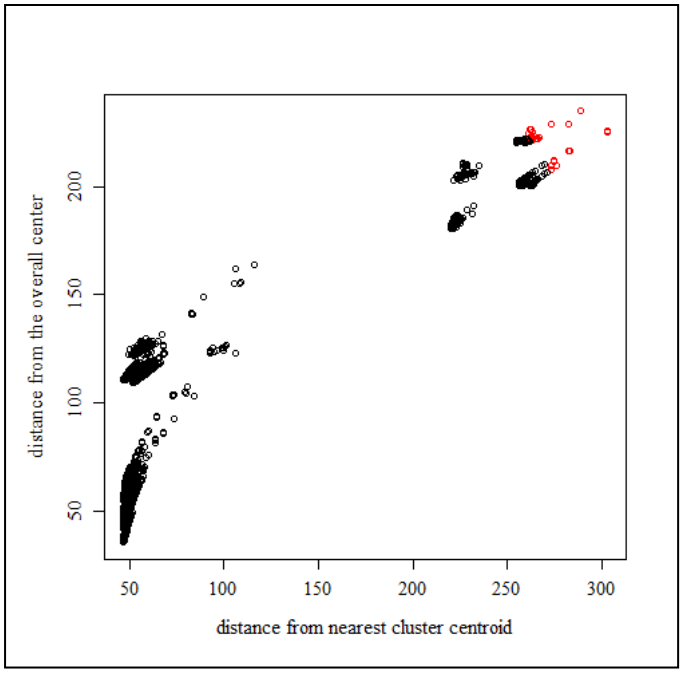


Fig. 4. Representation of the dual distance method of outlier detection

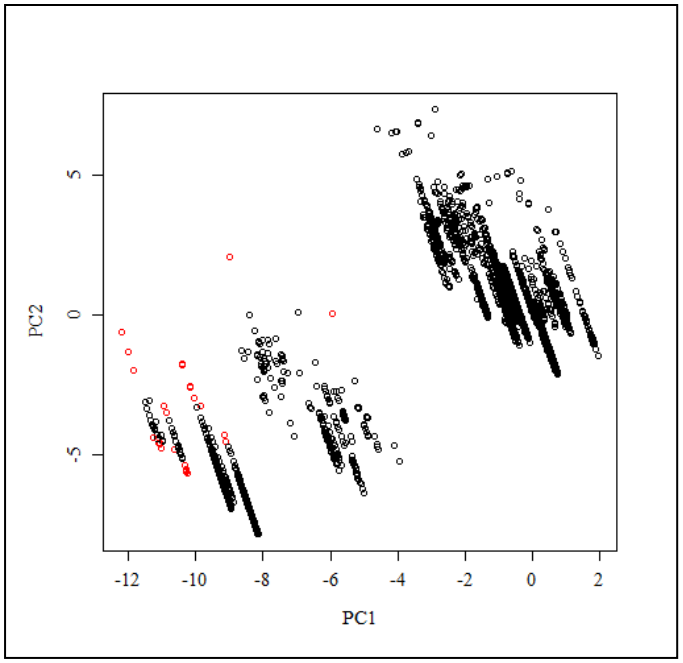


Fig. 5. Visual inspection of outliers on a representation using principal components

C. Validating the Impact of Tunable Parameters

We use Box-cox components as tunable parameters to enhance the list of outliers. We believe the use of these parameters reduces the chance of missing some of the atypical deals that do not outlie in multiple dimensions (for details, see section II.D). In other words, fraudulent deals are of various kinds, and we do not want our system to detect just one or two types of them. To examine the impact of the exponents used as tunable parameters, we look at the nature of the outliers obtained from two different sets of exponent values. We compute the average coordinates of outliers on various dimensions and convert the values into percentiles within the corresponding distributions of all the deals in the homogeneous group. The gray region in the radar charts in fig. 6 and fig. 7 shows the percentiles of outlier values across dimensions.

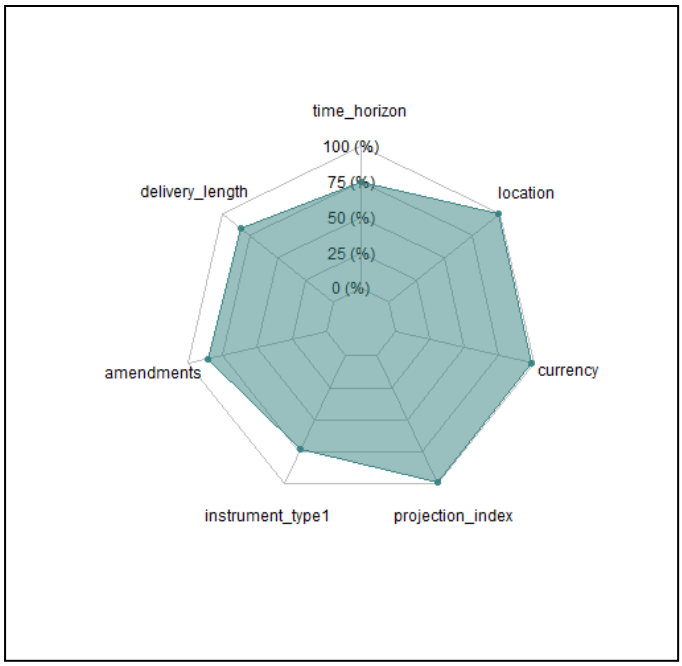


Fig. 6. Percentiles of average coordinates of the outliers, with Box-cox transformation using optimized exponents

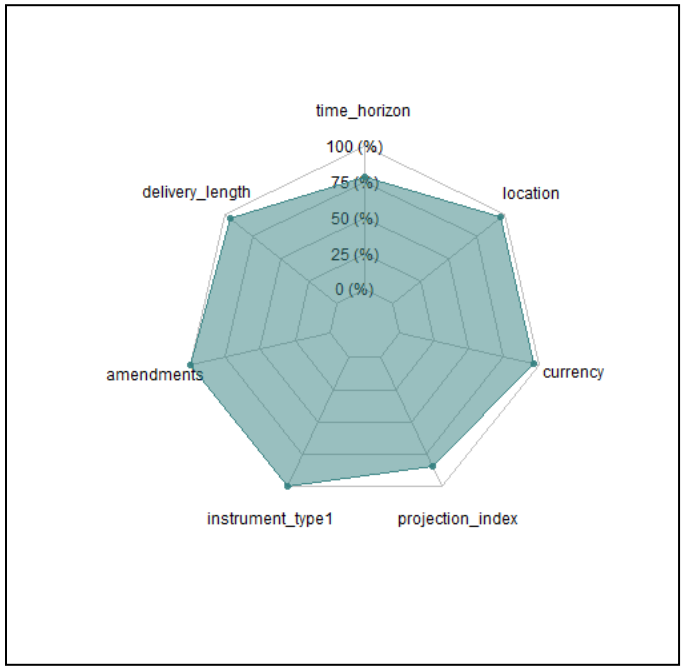


Fig. 7. Percentiles of average coordinates of the outliers, without transformation

The two approaches – one with transformation of features and the other without – leads to two different sets of outliers though partially overlapping. Based on fig. 6, identified outliers are the ones with the rarest possible labels for currency, location and projection index, and reasonably high values for the number of amendments, delivery length, and time horizon. The outliers obtained without transformation as shown in fig. 7 have the highest possible value for the number of amendments, relatively large value for delivery length and the rarest possible label for instrument type.

We do not delve into the results from identification of traders here as the process follows from the detection of outlier trades, as explained in section II.G.

D. Limitations

This paper focuses on the application of a modified outlier detection technique and its utility from a practical point of view. We do not explore the theoretical underpinnings of the method or compare its effectiveness with that of the other existing ones.

From the utility perspective, we believe it can lead to a semi-automated outlier detection system. The choice of features is a major determining factor in the effectiveness of the outlier detection method. The model relies on domain knowledge for the selection of features and so does not provide a fully automated solution to the problem. The method requires preprocessing, regarding partitioning the population into homogeneous groups, also done using domain knowledge.

The paper does not answer some of the pertinent questions – Does the identified outlier set encompass all the possible fraudulent deals? Does the set also erroneously include a few genuine deals? As mentioned in section III.A, we do not have examples of fraudulent deals to answer these questions. Nonetheless, the test conducted with the addition of artificial deals gives us significant pointers to the effectiveness of the model. A comprehensive test of the model may not be hard if we have another dataset with identified fraudulent cases, even in a different domain.

Lastly, the presented technique has its basis on distance-based outlier detection methods. Arguably, density based methods work as well or better with the kind of outlier detection problem encountered here. We think the dual distance method indirectly takes the density characteristics of the space into account by including the distance from the overall center. For instance, the composite measure identifies cases in a tiny cluster far away from the overall center as outliers. More often than not, a conventional distance-based method misses out on such situations.

IV. Conclusion

The paper takes up the problem of identifying outliers from a group of heterogeneous instances. It presents a dual distance method, flexible enough to be used for fraud detection in wide-ranging domains including those that are different from the one presented here. The framework allows the user decide the dimensions for outlier detection based on the defined problem. The flexibility in choosing the dimensions provided by the outlier detection method also allows it to address new applications unrelated to fraud. For instance, we can apply the method with certain modifications to a problem of identifying high risk investments/loans. Outlier detection provides an alternate to the conventional absolute measure based methods for such applications.

The dual distance method can scale to larger datasets with minimal pressure on computational requirements. The dual distance method, Chebyshev inequality based scoring method and parameter tuning to vary weights across dimensions use algorithms with linear time complexity.

References

[1] Methods for Anomaly Detection: A Survey – Kalinichenko, Shanin, and Taraban

[2] k-means --: A Unified Approach to Clustering and Outlier Detection - Chawla and Gionis

[3] LOF: Identifying Density Based Local Outliers –Breunig, Kriegel, Ng and Sander

[4] Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes –Byers and Raftery

[5] Unsupervised Profiling Methods for Fraud Detection –Bolton and Hand

[6] Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection – Amer, Goldstein, and Abdennadher

[7] Anomaly Detection: A Survey – Chandola, Banerjee, and Kumar

[8] Isolation-based Anomaly Detection – Liu, Ting, and Zhou

[9] Data Outlier Detection using Chebyshev Theorem – Amidan, Ferryman, and Cooley

[10] [https://en.wikipedia.org/wiki/Power_transform

[11] https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

[12] https://en.wikipedia.org/wiki/Chebyshev%27s_inequality