International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume 2, Issue 12 (December '2016)

OUTLIER DETECTION ON FINANCIAL CARD OR ONLINE TRANSACTION DATA USING MANHATTAN DISTANCE BASED ALGORITHM

V.Kathiresan,
Assistant Professor,
Department of Computer Science and Engineering,
Coimbatore Institute of Engineering and Technology,
Coimbatore, Tamilnadu.

Dr.N.A.Vasanthi,
Professor and Head,
Department of Information Technology,
Dr.N.G.P Institute of Technology,
Coimbatore, TamilNadu.

Abstract: Outlier detection is a technique in statistics to find anomaly in a dataset or data distribution. Anomaly or outlier is a data object which is deviated from the existing group of data objects. Finding outlier in financial card or online transaction leads fraud detection or fraud suspicion.With the increase in the number of credit and debit card transactions, there has been a substantial increase in the number of fraudulent card transactions too. As the data of RBI, Indian banks are reported close to 27614 and 3835 cases of credit and debit-card related frauds between April 2011 and September 2014. There were additional 2000 cases of internet-banking fraud in that same period. In transaction details two dimensions are taken into account to detect outlier those are transaction number and amount. The outlier is detected based on the MDBA(Manhattan Distance Based Algorithm) Manhattan Distance is mostly used to find distance between two data objects.

Keywords: Outlier detection, anomaly, Data objects, Financial card,MDBA

I.INTRODUCTION

Outlier detection can be identified by various techniques in datamining, Outlier is the data object which is lies outside from the most of the data objects are lies. Outlier objects are also called anomaly in dataset. Outlier can be identified both positive and negative purpose, outlier objects may be considered as noise in data distributions to smoothening the dataset more. In sameway outliers are focused and seen further more in financial card or online transactions becausethose transactions may be a fraudulentone, further investigation on those fraudulent or suspected fraud transactions are very essential.

Here the outlier is identified for the purpose of fraud detection or fraud suspicion, In financial card or online transaction. Only two dimensions of the customer or organizations transaction is taken into account those are transaction number and amount. When the new transaction is coming that will be checked with the existing transaction if the new one will be an outlier, the transaction is suspect to be fraud, further more investigation is essential to approve transaction. Manhattan Distance based Algorithm is used to find when the incoming transaction is outlier or not.

II.RELATED WORK

Alejandro Correa Bahnsen et al (2014) proposed a method to improve credit card fraud detection with calibrated probabilities [3] .this approach requires good probability estimates that not only separate well between positive and negative examples, but also assess the real probability of the event. Unfortunately, not all classification algorithms satisfy this restriction. In this paper, two different methods for calibrating probabilities are evaluated and analyzed in the context of credit card fraud detection, with the objective of finding the model that minimizes the real losses due to fraud.

AbhinavSrivastava et al (2008) stated that Hidden Markov Model (HMM)[1] provides a pragmatic method of detecting the anomaly by analysing the spending pattern of the customer. Basically HMM consists of sequence of states that works on Markov chain property. The input to the model is the observed sequence and the output is the probability for that sequence. For each cardholder HMM is trained and maintained. In HMM based approach there is an extreme decrease in the false positive rate. The objective of the system is to detect the anomaly during the transaction and then the fraud is confirmed with the cardholder by asking some secret questions.

The fraud detection in credit card transaction is an application of classification. Dipti Thakur and Shalini Bhatia (2009) provides a technique to perform classification using decision tree methodology[2] in data mining and also the rules are shared among different credit card companies without sharing the data using agent based classification.

Trend offset analysis (TOA) is a local outlier-based supervised learning technique implemented for credit card fraud detection. It focuses on identifying pattern changes at an individual account level. TOA is used credit card fraud detection in such a way that a signature is assigned to each account based on the most recent transaction. Any significant

deviation in current behavior from the assigned signature was used for outlier detection.

In genetic approach a authentication mechanism is used while transaction is done, to secure cash card by asking secret question to user for verification in case of credit card & SMS feedback system for ATM transactions. It secures cash card from being cloned via skimmed device& providing more security during the transaction. This work shows AI, image Processing & data mining techniques are used for fraud prevention there by implementing as/which ask secret questions i.e. ATM feedback SMS system with reply and by thumb impressions instead of detecting a fraud, a fraud can also be prevented.

Manhattan Distance based Algorithm (MDBA)

Sample Data: (Spending behaviors of the vendee)

Transaction Number(x)	1	2	3	4	5	6	7	8	9	10
Transaction Amount(y)	2000	2500	1500	4200	6300	3500	5600	5800	4900	6000

Table 1: Sequence of Transaction Number and Transaction Amount

Step 1:

Find the Centroid men of (N, A) where

N –Mean of Transaction Number

A –Mean of Transaction Amount

For the above sample data

N –5.5

A –4230

Step 2:

Find the distance among each data point and centroid using Manhattan Distance Formula

$MD = | Xi - N | + | Yi – A |$

Where i= 1,2,3,4…..n

MD from centroid for each data point:

Data Objects	(1,2000)	(2,2500)	(3,1500)	(4,4200)	(5,6500)	(6,3500)	(7,5600)	(8,5800)	(9,4900)	(10,6000)
MD from Centroid	2234.5	1733.5	2732.5	31.5	2070.5	730.5	1371.5	1572.5	673.5	1774.5

Table 2: Manhattan distance between centroid and each transaction or data point

Step 3: Fix the Maximum Distance as threshold

Threshold Distance ($\mathbf{T_d}$) =2732.5

Step 4: When new Transaction ($\mathbf{T_{new}}$) takes place find MD (Manhattan Distance) between Centroid and new data object. If the distance greater than Td the Transaction suspect to be a fraud otherwise the data object will be added into existing collection of data objects and do step 1 to 3 again for another upcoming transaction.

New Transaction:

Transaction Number = 11

Transaction Amount = 35000

MD between Centroid and New Transaction:

MD (T_{new} , Centroid) =30775.5

MD (T_{new} , Centroid) >T_d

So T_{new} suspect to be a fraud and proceed with **Step 5 and 6**

Condition to suspect fraud transaction

If (MD (T_{new} , Centroid))>T_d
T_{new} suspect to be a fraud

Step 5: Once the fraud is suspected the transaction is undergone further investigations by asking secret questions.

MDBA condition

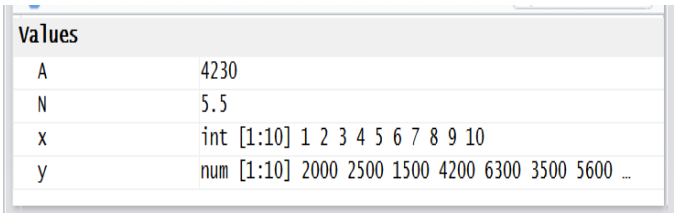
If (MD (T_{new}, Centroid)) > Td
T_{new} suspect to be a fraud and ask secret questions;
If (properly answered)
False positive – added into database
True positive - Transaction denied
Else
Allow transaction

Step 6: If fraud confirmed the transaction denied and consider the suspicion by MDBA is True positive otherwise allow the transaction and consider the suspicionto MDBAas False positive.

III.SIMULATION RESULT

x and y are the two dimensions of the dataset which is taken from the spending behavior of the card holder or account holder. xand y are respectively called Transaction number and Transaction amount.A and N are respectively, Mean of transaction number and amount. In order to find the centroid of the dataset mean value is essential.

Figure 1: Value of x,y,A and N



N – Mean of Transaction Number

A – Mean of Transaction Amount

x- List of transaction number

y- List of transaction amount

The distribution of transaction number is shown in the boxplot(Fig2) .Transaction number is distributed sequentially so the minimum is 1 ,maximum is 10,1st quartile is 3,3rd quartile is 8 and the IQR (Inter Quartile Range) is 5.So the data is distributed evenly.

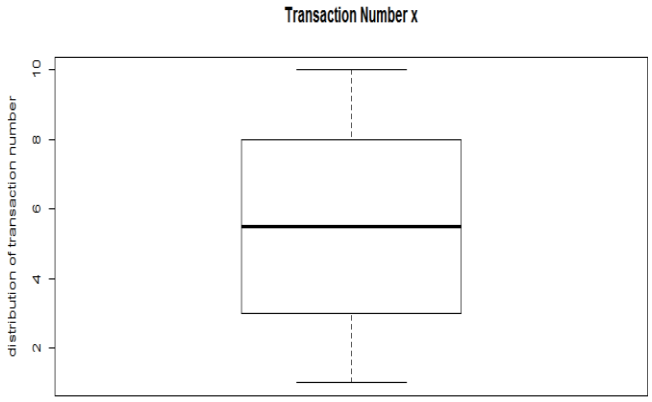


Figure 2:boxplot of the distribution of transaction number

The distribution of transaction amount is shown in the boxplot (Fig3). Transaction amounts are mostly distributed in the range 2500 to 5500 that is IQR(Inter Quartile Range). 1st quartile is 2750,3rdquartile is 5750,minimum value is 1500 and maximum value is 6300.

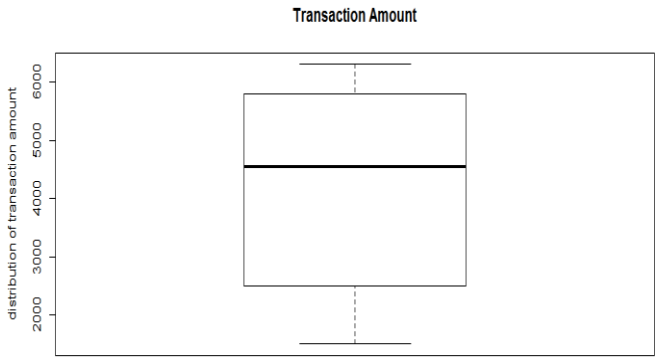


Figure 3:Boxplot of distribution of transaction amount

After the identification of centroid N and A the MD(Manhattan Distance) have calculated for the each data point from the centroid. Manhattan distance of the each transaction have shown in the given barplot(Fig 5).

```
> MD
 [1] 2234.5 1733.5 2732.5   31.5 2070.5  730.5 1371.5 1572.5  673.5 1774.
```

Figure 4:Manhattan Distance from centroid to each point simulation tool result

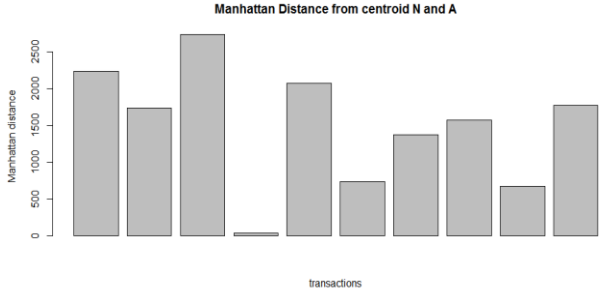


Figure 5: barplot of Manhattan Distance from centroid N and A

The maximum distance among all distances which have calculated between centroid and the dataobject considered as threshold valueT_d .When the new transaction T_{new}arrived the MD(T_{new},Centroid) will be calculated and the value is checked with the threshold T_d .

```
[1] 30775.5
>
```

Figure 6:Manhattan distance from centroid to Tnew Simulation Tool Result(MD(T_{new}, Centroid))

Mnhattan distance MD from T_{new}Centroid have plotted in the scatter plot.

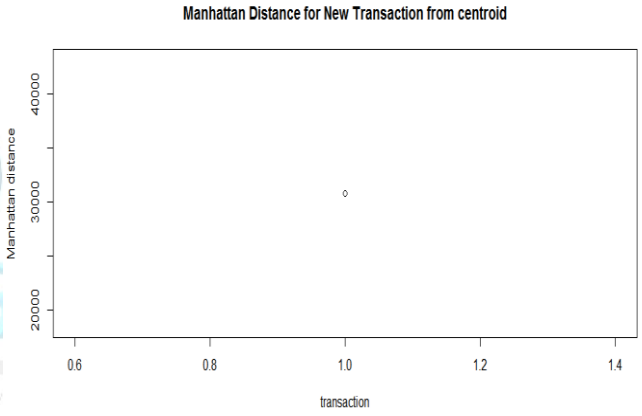


Figure 7: scattorplot of Manhattan distance from centroid to Tnew Simulation Tool Result

The T_{new}Transaction is considered as an outlier because the MD of the transaction $T_{new}$is greater than the threshold limit T_d . So the transaction is suspect to be a fraud. Outlier detection is shown in the scatterplot(Fig7) and the outlier object is marked as red.

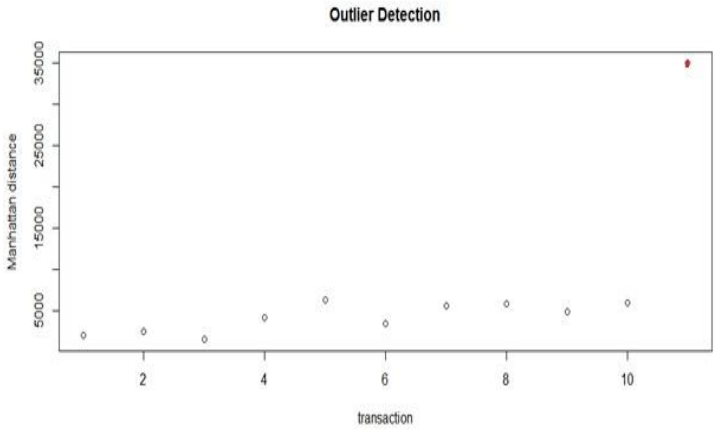


Figure 8:scattorplot of Manhattan distance with outlier detection Simulation Tool Result

If the transaction is considered as a fraud in outlier

IV.CONCLUSION

MDBA is simple and effective technique to handle or detect fraud in financial card or online finance transactions. Implementation of the MDBA algorithm is simple and less time and resource consuming process. In this paper MDBA is used to find outlier using transaction number and transaction amount. In future in order to improve the efficiency in fraud detection and to reduce the false positive location of the amount withdrawal or purchase, ip address of the machine which is used to do online transaction are taken into account and again the outlier is identified for these parameters leads effective fraud detection.

V.REFERENCES

[1]. AbhinavSrivastava, AmlanKundu, ShamikSural and ArunK. Majumdar(2008) 'Credit Card Fraud Detection UsingHidden Markov Model' IEEE Transactions onDependable and Secure Computing vol. 5 No. 1.
[2]. Dipti Thakur and Shalini Bhatia (2009) 'Distributed Data Mining Approach to Credit Card Fraud Detection' Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India Vol. 4, 48.
[3]. Von Matt, Urs, and D. Dacunha-Castelle. "Improving credit card fraud detection with calibrated probabilities." (2014).
[4]. https://xlinux.nist.gov/dads/HTML/manhattanDistance.html
[5]. Lawrence R. Rabiner, 'A tutorial on Hidden Markov Models and Selected applications in Speech Recognition', Proceedings of the IEEE, VOL.77, No 2, February 1989.
[6]. Leila Seyedhossein and Mahmoud Reza Hashemi (2010) 'A Timelier Credit card Fraud Detection by Mining Transaction Time series' International Journal of Information & Communication Technology vol 2 No 3.
[7]. ParulBhanarkar and Pratiksha L. Meshram (2012) 'Credit and ATM card Detection using Genetic Approach' International Journal of Research & Technology Vol 1 Issue 10 ISSN:2278-0181.
[8]. Sherwood, Timothy, et al. "Automatically characterizing large scale program behavior." ACM SIGARCH Computer Architecture News. Vol. 30. No. 5. ACM, 2002.