Credit card fraud detection using anti-k nearest neighbor algorithm

VENKATA RATNAM GANJI

Dept of Computer Science and Engineering,
VKR, VNB and AGK College of Engineering, Gudivada
A.P,India.
gvr.jntuk@gmail.com

SIVA NAGA PRASAD MANNEM

Dept of Computer Science and Engineering,
VKR, VNB and AGK College of Engineering, Gudivada
A.P,India.
csesiva547@gmail.com

Abstract:

Banks have used early fraud warning systems for some years. Improved fraud detection thus has become essential to maintain the viability of the payment system. Outlier mining in data mining is an important functionality of the existing algorithms which can be divided into methods based on statistical, distance based methods, density based methods and deviation based methods. In this article I propose the concept of credit card fraud detection by using a data stream outlier detection algorithm which is based on reverse k-nearest neighbors (SODRNN). The distinct quality of SODRNN algorithm is it needs only one pass of scan. Whereas traditional methods need to scan the database many times, it is not suitable for data stream environment.

Keywords: *Outlier; Reverse k nearest neighbors; Sliding Window; Data mining, SODERNN.*

1. Introduction:

Credit Card Fraud is one of the biggest threats to business establishments today. However, to combat the fraud effectively, it is important to first understand the mechanisms of executing a fraud. Credit card fraudsters employ a large number of modus operandi to commit fraud. In simple terms, Credit Card Fraud is defined as: When an individual uses another individuals' credit card for personal reasons while the owner of the card and the card issuer are not aware of the fact that the card is being used. Further, the individual using the card has no connection with the cardholder or issuer, and has no intention of either Contacting the owner of the card or making repayments for the purchases made.

Credit card frauds are committed in the following ways:

- An act of criminal deception (mislead with intent) by use of unauthorized account and/or personal information
- Illegal or unauthorized use of account for personal gain
- Misrepresentation of account information to obtain goods and/or services.

Contrary to popular belief, merchants are far more at risk from credit card fraud than the cardholders. While consumers may face trouble trying to get a fraudulent charge reversed, merchants lose the cost of the product sold, pay chargeback fees, and fear from the risk of having their merchant account closed.

Increasingly, the card not present scenario, such as shopping on the internet poses a greater threat as the merchant (the web site) is no longer protected with advantages of physical verification such as signature check, photo identification, etc. In fact, it is almost impossible to perform any of the 'physical world' checks necessary to detect who is at the other end of the transaction. This makes the internet extremely attractive to fraud perpetrators. According to a recent survey, the rate at which internet fraud occurs is 12 to 15 times higher than

'physical world' fraud. However, recent technical developments are showing some promise to check fraud in the card not present scenario.

There are many ways to commit credit card fraud. Fraudsters are very inventive, fast moving people. Mainly there are two common forms of credit card fraud:

1) Traditional Techniques

2) Modern Techniques

The first type of credit card fraud to be identified by this paper is *Application Fraud*, where an individual will falsify an application to acquire a credit card. Application fraud can be split into assumed identity, where an individual pretends to be someone else; and financial fraud, where an individual gives false information about his or her financial status to acquire credit. This investigation then goes on to look at *Intercept Fraud*, where a card is applied for legitimately, but is stolen from the post service before it reaches its final destination. There is also the illegal use of *Lost and Stolen Cards*, which makes up a significant area of credit card fraud.

There are more sophisticated credit card fraudsters, starting with those who produce *Fake and Doctored Cards*, there are also those who use *Skimming* to commit fraud. This is where the information held on either the magnetic strip on the back of the credit card, or the data stored on the smart chip is copied from one card to another. *Site Cloning and False Merchant Sites* on the Internet are becoming a popular method of fraud for many criminals with a competent ability for hacking. Such sites are designed to get people to hand over their credit card details without realizing they have been scammed. *Triangulation* is also a new phenomenon. Triangulation is when a merchant offers a product at a very cheap price through a web-site. When a customer seeks to buy the product the merchant tells to customer to pay via email once the item is delivered. The merchant uses a fraudulent card number to purchase the product from a Web site and sends the product to the consumer, who then sends the merchant his or her credit card details via email. The merchant goes on operating in this way using the credit card numbers that have been sent from the consumers to purchase products, appearing for a short time to be a legitimate merchant before he or she closes the Web site and starts a new one. There is also the more sophisticated fraudsters, who use *Credit Card Generators*; computer emulation software that creates valid credit card numbers and expiry dates. These generators are highly reliable at creating valid credit card details and are available for free download off the internet. Making them available to many individuals who run fraudulent operations

Credit card fraud is divided into two types: offline fraud and online fraud. Offline fraud is committed by using a stolen physical card at storefront or call center. In most cases, the institution issuing the card can lock it before it is used in a fraudulent manner. Online fraud is committed via web, phone shopping or cardholder not present. Only the card's details are needed, and a manual signature and card imprint are not required at the time of purchase.

2. Credit Card Fraud Detection

Credit card fraud detection is quite confidential and is not much disclosed in public. Some available techniques are discussed as follows.

2.1 Outlier Detection.

An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Unsupervised learning approach is employed to this model. Usually, the result of unsupervised learning is a new explanation or representation of the observation data, which will then lead to improved future responses or decisions. Unsupervised methods do not need the prior knowledge of fraudulent and non fraudulent transactions in historical database, but instead detect changes in behavior or unusual transactions. These methods model a baseline distribution that represents normal behavior and then detect observations that show greatest departure from this norm. Outliers are a basic form of non-standard observation that can be used for fraud detection. In supervised methods, models are trained to discriminate between fraudulent and non-fraudulent behavior so that new observations can be assigned to classes. Supervised methods require accurate identification of fraudulent transactions in historical databases and can only be used to detect frauds of a type that have previously occurred. An advantage of using unsupervised methods over supervised methods is that previously undiscovered types of fraud may be detected. Supervised methods are only trained to discriminate between legitimate transactions and previously known fraud. In order to explain our

point a little bit better, let us consider the example illustrated in Fig. 1. In the example, we have shown two two-dimensional cross sections of a very high-dimensional data set.

Bolton and Hand proposed unsupervised credit card fraud detection, using behavioral outlier detection techniques. Abnormal spending behavior and frequency of transactions will be identified as outliers, which are possible fraud cases.

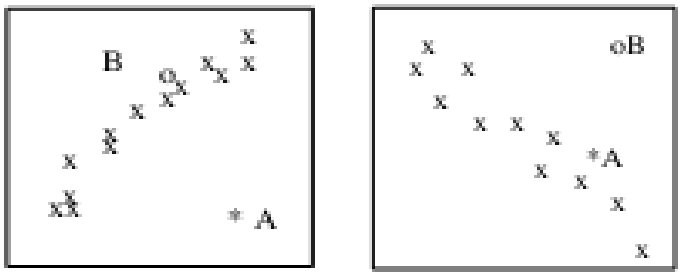

Fig. 1. Illustrations of outliers in various views of the data.

3. Algorithm

3.1 **Algorithm method**

In this section the algorithm SODRNN, standing for Stream Outlier Detection based on Reverse k Nearest Neighbors, is described. This algorithm consists of two procedures: the Stream Manager and the Query Manager. And the entire window should be allocated in memory. The former procedure receives the incoming data stream objects and efficiently updates the current window. When new stream object comes, in order to maintain current window perfectly, it needs only update the knnlist and rknnlist of the influenced objects in the current window rather than that of all the data stream objects in the current window. When new coming object is inserted, it needs only one pass of scan to the current window to find all objects whose k nearest neighbors are influenced. The update of the knnlists of the influenced objects in the current window can update their rknnlists at the same time. The deletion of the expired object needs only update the rknnlists of the influenced objects in the current window according to its knnlist, and then update the knnlists of the influenced objects in the current window according to its rknnlist. When user demands a query of the top m outliers, the latter procedure will make a scan of the current window and return m objects whose RNNk(p) is small as outliers of this query.

3.2 **Algorithm description**

ALGRITHM: SODRNN
INPUT: DS, current window size N, integer k,
query time Uquery, number of outlier m
OUTPUT: m outliers

METHOD:

BEGIN

SM(DS,N,k);

when (Uquery) QM(m);

END

1) Stream Manager Procedure

PROCEDURE SM(DS,N,k)

BEGIN

(1) FOR each data stream object obj with arrival time tDO

(2) IF the oldest object q of current window

```
    expIres
(3) FOR all objects 0 III q.knnlist DO
    o .rknnlistdelete( q);
(4) FOR all objects 0 in q.rknnlist DO
    o .knnlistdelete( q);
(5) ENDIF
(6) remove object q from current window
(7) object p(obj,t,cD,cD);
(8) FOR all objects 0 in current window DO
(9) dist=o.distance(p);
(10) p.knnlistinsert( 0); lithe k nearest neighbors
    ofp
(11) o.rknnlistinsert(p);
(12) IF dist<=o.k_distanceO
(13) o.knnlistinsert(p);
(14) p.rknnlistinsert( 0);
(15) ENDIF
(16) END FOR
(17) Insert object p into current window.
(18) ENDFOR
  END
```

2). Outlier Query Management Procedure

```
PROCEDURE QM(m)
BEGIN
(1) perform a single scan of current window;
(2) return m objects with minimal I RNNk(p) I as
    outliers.
END
```

4. Experimental verification

Our experiments were performed on several synthetic data and real data sets. In all our experiments, we have assumed that we have information about the outliers in the data set, so we could evaluate the detection performance. In order to improve the knn query performance, we also make a change to the X-tree index structure. We omit its Split History information field of the node structure and use a simple clustering algorithm for the split node algorithm which supports knn query better. Then we implement the program SODRNN based on this index structure with VC++6.0. The experiments are conducted on an Intel Pentium D 3.1GHz PC with l GB main memory under Windows XP.

5. Experimental memory space requirements

In the experiment, we choose a certain data set size, the random number generator produced in high dimensional space uniformly distributed data, including the 10,000 multi-dimensional space point data, on the different dimensions of uniformly distributed data were tested the X * X-tree index structure tree and actual take up memory space. As can be seen from Figure 2, with the increase of dimension, X * X-tree index structure tree and the actual memory space occupied by a corresponding increase, because each node within the array of data items with the MBR with the increase of dimension up more space, and in any peacekeeping number, X Shu

than X * tree takes up more memory space, because the X directory tree needs all the nodes in all the data items additional storage node split in the history record information.

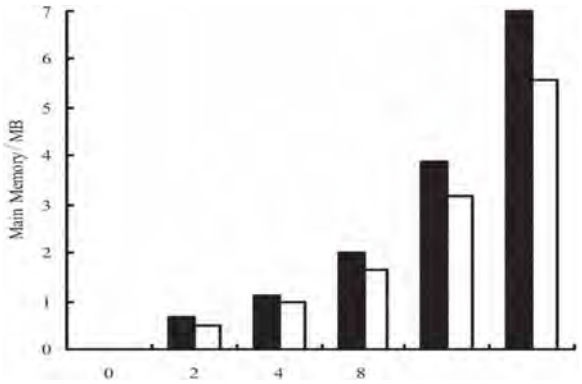

Fig. 2. Main memory requisition of the two indexes structure for different dimensions.

6. Advantages of proposed model

1. Lost and stolen card feature makes it easier to stop fraudulent transactions
2. Credit card validation checks detects errors in a sequence of numbers hence detects valid and invalid numbers easily.

Conclusion

In this paper, a novel data stream outlier detection algorithm SODRNN is presented. This algorithm reduces the number of scans to only one. Experiments conducted on both synthetic and real data sets show that the proposed method is efficient and effective.

References

[1] Lijun Cao, Xiyin Liu, Tiejun Zhou , Zhongping Zhang Aiyong Liu; Based on the flow of anti-k nearest neighbors algorithm for data mining outliers; In Proceedings of IC-BNMT2010

[2] Yufeng Kou, Chang-Tien Lu, Sirirat Sinvongwattana , Yo-Ping Huang; Survey of Fraud Detection Techniques; in Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control Taipei, Taiwan, March 21-23, 2004

[3] Y. Dora Cai, David Clutter, Greg Pape, Jiawei Han. MAIDS; Mining alarming incidents from data streams; In SIGMOD, Paris, 2004:919-920

[4] Charu C. Aggarwal, Philip S. Yu; An effective and efficient algorithm for high-dimensional outlier detection; In The VLDB Journal (2005) 14: 211–221

[5] Aleksandar Lazarevic, Vipin Kumar; Feature Bagging for Outlier Detection; In Proceedings of KDD'05, August 21–24, 2005, Chi cago, Illinois, USA.

[6] Jiaheng Lu,Ying Lu, Gao Cong; Reverse Spatial and Textual k Nearest Neighbor Search; in proceedings of SIGMOD'11,June 12–16, 2011, Athens, Greece.

[7] Wen Jin, Anthony K. H. Tung, Jiawei Han and We Wang; Ranking Outliers Using Symmetric Neighborhood Relationship; PAKDD 2006 LNAI 3918 pg 577-593,2006.

[8] Yufeng Kou, Chang-Tien Lu, Sirirat Sinvongwattana Yo-Ping Huang; Survey of Fraud Detection Techniques; Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control Taipei, Taiwan, March 21-23, 2004

[9] Philip k.chan, WeiFan,Andreas Prodromidis,andSalvatore J.Stolfo; Distributed Data Mining in Credit Card Fraud Detection; Submitted to IEEE Intelligent Systems Special Issue onDataMining,1999.

[10] Pokrajac D, Lazarevic A ; Incremental local outlier detection for data streams; IEEE CIDM,2007:S04-S1S