ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

Outlier Detection Using Unsupervised and Semi-Supervised Technique on High Dimensional Data

Ms. Gayatri Attarde[1], Prof. Aarti Deshpande[2]

M. E Student, Department of Computer Engineering, GHRCCEM, University of Pune, Pune, India [1]

Professor, Department of Computer Engineering, GHRCCEM, University of Pune, Pune, India[2]

ABSTRACT: Outlier detection is useful for credit card fraud detection. Due to drastic increase in digital frauds, there is a lot of financial losses and therefore various techniques are developed for fraud detection and applied to diverse business fields. In high-dimensional data, outlier detection presents some challenges because of increment of dimensionality. In this paper, the proposed model aims to implement unsupervised outlier detection technique using KNN, AntiHub and AntiHub2 algorithm and semi-supervised outlier detection technique in which first extracting negative instances by KNN and then with fuzzy clustering of both positive and negative example with distributed approach to find credit card fraud. Hence this method will provide more accurate results as compared to other methods.

KEYWORDS: outlier detection, semi-supervised learning, unsupervised learning, high dimensional data

I. INTRODUCTION

Outlier detection is useful for detection credit card fraud. Fraud is defined as the usage of one's asset for personal enrichment through misbehavior. In real world fraudulent activity may arise in many areas such as online banking, telecommunication networks, E-Commerce, mobile communications. Fraud is increasing extremely with globalization and modern technology which results in major loss to the businesses. Fraud detection refers to the act of finding frauds as early as possible. In recent years fraud detection has been implemented using techniques such as neural networks, data mining and outlier detection. Finding outliers in data defined as finding patterns in data that do not fit in to normal behavior. An Outlier is a reflection in data instances which is dissimilar from the others in dataset. Data Labels associated with data instances displays whether that instance belongs to normal data or anomalous. Based on the availability of labels for data instance, the anomaly detection techniques work in one of the three modes are 1)Supervised Anomaly Detection, techniques worked in supervised mode consider that labeled instances availability for normal and anomaly classes in a training dataset. 2) Semi-supervised Anomaly Detection, techniques skilled in supervised mode consider that the availability of labeled instances for normal, do not need labels for the anomaly class. 3) Unsupervised Anomaly Detection, techniques that work in unsupervised mode do not require training data.

Unsupervised methods include distance-based methods that based on measure of distance or similarity to detect outliers. This technique judge a point based on the distance(s) to its neighbors. As distance measures concentrate, due to the increase of dimensionality, means pairwise distances become different as dimensionality increases and distance becomes meaningless. The impact of distance concentration outlier detection on unsupervised was that every point in high dimensional space appears as good outlier.
The distance based methods are combined to implement a
new method with distributed approach to improve performance. In semi-supervised method to detect outliers by using few positive examples and unlabeled data, propose method based on fuzzy clustering methods.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

The rest of the paper is structured as follows. Section 2 describes related work. Section 3 gives programmer's design in which proposed framework for outlier detection is described. Section 4 includes the result and section 5 concludes the paper.

II. LITERATURE SURVEY

As per classification in [2], the scope of investigation is to observe: (1) point anomalies i.e. individual points that can be considered as outliers (2) unsupervised methods, and (3) methods that provide an "outlier score" to each point, creating as output a list of outliers ranked by their scores. The described scope of study is the focus on most research of outlier detection [2].

Among the most commonly applied methods based on nearest neighbors, which assume that outliers appear distant from their closest neighbors. Such methods depend on a distance or similarity measure to catch the neighbors, with Euclidean distance being the most common option. One of the variant of neighbor-based method is k^{th} nearest neighbor [1]. In KNN, data points are separated into several separate classes to predict the classification of new sample point. Runtime performance is poor when training set is large.

In [1], the reverse k nearest neighbour count is defined outlier score of a point where threshold parameter determines whether a point is outlier or not. In unsupervised learning of reverse nearest neighbour technique, if the data has normal instances that do not have enough adjacent neighbors or if the data has anomalies that have enough adjacent neighbors, the technique fails to label them properly, resulting in missed anomalies. The proposed work implements semi-supervised method for outlier detection. These techniques achieve better than unsupervised techniques in terms of lost anomalies.

The angle-based outlier detection (ABOD)[5] technique detects outliers in high-dimensional data by considering the variances of a measure over angles between the difference vectors of data objects.LOF(Local Outlier Factor)[10] technique used for finding anomalous data points by measuring local deviation of given data point with respect to its neighbors. Influenced outlierness (INFLO) [11] is a technique based on a symmetric relationship that considers both neighbors and reverse neighbor of a point.

III. FRAMEWORK

The block diagram of outlier detection is shown in fig.1

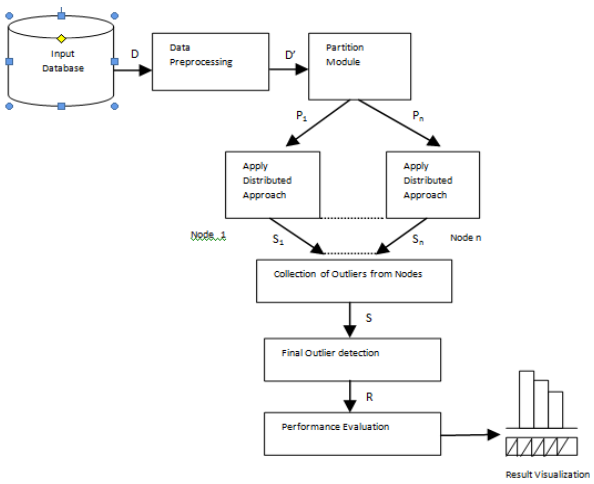

Fig1: Block Diagram

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

A. System Architecture Description

- *Data Collection and Pre-processing :*

Data is preprocessed before forwarding for future steps. Data mining methods such as data integration, cleaning, transformation will be used to preprocess large dataset contents and to generate required clean data.

- *Data Separating:*

In this module, the preprocessed data is distributed into number of datasets as per the data size. This partitioned data then managed by individual clients to balance the data.

- *Proposed Outlier detection methods:*

The technique proposed for detecting outliers will be applied initially at distributed clients and detected outliers would be collected on server machine for final stage computation of outliers. For this, we proposed KNN with AntiHub and AntiHub2 algorithm for unsupervised method and for semi-supervised KNN with fuzzy clustering for outlier detection. The Distributed approach with proposed Method based on anomaly detection techniques based on nearest neighbor.

- *Performance evaluation and Result Visualization:*

Performance is calculated based on the evaluation parameters. The performance evaluation delivers details about calculated system performance, restrictions and metrics for future work. The system execution will be made more clear and explorative for its evaluators with appropriate visualization of results.

IV. APPROACHES FOR OUTLIER DETECTION

A. *Outlier Detection using unsupervised learning:*

The frequently occurring data points in k neighbour sets are mentioned as hubs and very rarely occurring points are mentioned as AntiHub.

Weakness of AntiHub method is hubness and inherent discreteness to add more discrimination one approach is to raise k, possibly up to some value comparable with n. we will explore this option, but the approach raises two concerns with increasing k, the outlierness moves from local to global, if local outliers are there and interesting can be missed and k values comparable with n raise issues with computational complexity. For these reasons we propose method AntiHub2, which refines outlier scores produced by the AntiHub technique by considering the N_k scores of the neighbors of x with $N_k(x)$ itself. For each point x, AntiHub2 adds (1 - α) • $N_k(x)$ to α times the addition of N_k scores of the k nearest neighbors of x, where $\alpha \in [0, 1]$. Summation used to aggregate the neighbors scores, proportion α is determined by maximizing discrimination between outlier scores of the strongest outliers, and controlled by two user provided parameters i.e. the ratio of strongest outliers and step size.

AntiHub (D, k)

Input

• Distance measure dist

• Ordered data set $D = (x_1, x_2, \ldots, x_n)$, where $x_i \in R^d$, for $i \in \{1, 2, \ldots, n\}$

• No. of neighbors $k \in \{1, 2, \ldots\}$

Output

• Vector $s = (s_1, s_2, \ldots, s_n) \in R^n$, where s_i denotes outlier score of x_i, for $i \in \{1, 2, \ldots, n\}$

Temporary variables

• $t \in R$

Steps:

1) For each $I \in (1, 2, \ldots, n)$

2) $t := N_k(x_i)$ computed w.r.t. dist and data set $D \backslash x_i$

3) $s_i := f(t)$, where $f: R \rightarrow R$ is a monotone function

AntiHub2 (x, k, p, step)

Input

• Distance measure dist

• Ordered data set $D = (x_1, x_2, \ldots, x_n)$, where $x_i \in R^d$, for $i \in \{1, 2, \ldots, n\}$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

• No. of neighbors k $\in \{1, 2, \ldots\}$
• Ratio of outliers to maximize discrimination p $\in (0, 1]$
• Search parameter step $\in (0, 1]$
Output
• Vector s = $(s_1, s_2, \ldots, s_n) \in \mathbb{R}^n$, where s_i is the outlier score of x_i, for i $\in \{1, 2, \ldots, n\}$
Temporary variables
• AntiHub scores a $\in \mathbb{R}^n$,
• Sums of nearest neighbors AntiHub scores ann $\in \mathbb{R}^n$
• Proportion $\alpha \in [0, 1]$ $\lceil np \rceil$
• (Current) discrimination score) cdisc, disc $\in \mathbb{R}$
• (Current) raw outlier scores ct, t $\in \mathbb{R}^n$
Local functions
• discScore (y, p) : for y $\in \mathbb{R}^n$ and p $\in (0,1]$ outputs the number of unique items among $\lceil np \rceil$ smallest members of y, divided by $\lceil np \rceil$
Steps
1) a := AntiHub (D, k)
2) For each i $\in (1, 2, \ldots, n)$
3) $ann_i := \sum_{j \in NN_{dist}(k,i)} a_j$, where $NN_{dist}(k, i)$ is the set of indices of k nearest neighbors of x_i
4) disc := 0
5) For each $\alpha \in (0, step, 2.step, \ldots, 1)$
6) For each i $\in (1, 2, \ldots, n)$
7) $ct_i := (1 - \alpha) \cdot a_i + \alpha \cdot ann_i$
8) cdisc := discScore(ct, p)
9) If cdisc > disc
10) t := ct, disc := cdisc
11) For each i $\in (1, 2, \ldots, n)$
12) $s_i := f(t_i)$ where f: R→ R is a monotone function

B. *Outlier Detection using semi-supervised learning:*

- *Extracting Negative Examples by KNN*

In this paper, proposed model aims to implements semi-supervised outlier detection approach for positive and unlabeled data approach to solving the problem where there are few very few positive examples available. For this problem in semi-supervised method, first extracting negative instances by KNN and then with fuzzy clustering of both positive and negative examples for outlier detection. Extracting negative instances using KNN. K-nearest neighbor classification use pattern recognition. KNN is a instance-based learning in which the function is approximated locally and all computation is deferred up to classification. The k-nearest neighbor algorithm is one of the machine learning algorithms. KNN use the unprocessed training set for classification.

- KNN Algorithm with Fuzzy Clustering

Input
P positive examples set,
U unlabeled examples set,
K number of nearest neighbors,
N number of negative examples
Output
NE set of negative instances
Steps
1. For each unlabeled instance u_i
For each positive example v_j
Compute sim (u_i, v_j)
Compute rank (u_i, v_j)

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

End For
2. End For
3. Rank the outcomes for $v_j(j=1, \ldots, p)$
4. Select k nearest neighbors vj (j=1, ..., p) and eliminate them
5. Select best N instances with a most average similarity from vj (j=1, ... , p)

II. RESULT

To examine outlier detection for credit card fraud we use standard German Credit Card Fraud dataset which is available on UCI Machine Learning Repository. This dataset consist of 20 attributes and 1000 instances.
Outlier detection accuracy is calculated, in order to find out the number of outliers detected. Detection rate refers to the ratio between the numbers of correctly detected outliers and to the total number of outliers.

The figure 2 shows the outliers detected by unsupervised technique , figure 3 shows the outliers detected by semi-supervised technique and figure 4 shows the comparative study of outlier detection rate with existing and proposed algorithms for outlier detection.

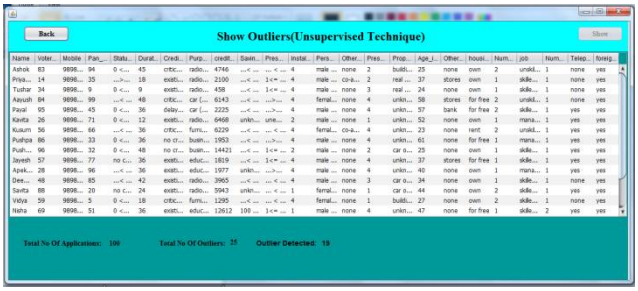

Fig2: Outliers detected by unsupervised technique

Fig3: Outliers detected by semi-supervised technique

Fig4: Outlier Detected

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

III. CONCLUSION

This paper proposed a novel framework for outlier detection by using unsupervised learning in which we are using KNN, AntiHub and AntiHub2 Algorithm and with semi-supervised learning using KNN with fuzzy clustering using distributed approach aims to implement and comparing unsupervised outlier detection methods and semi-supervised outlier detection methods. We propose method to improve them in terms of accuracy. The proposed system can be implemented using HDFS framework to improve execution speed because it will process the data parallelly on high dimensional as well big data.

REFERENCES

[1] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic, "Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection", IEEE Transactions and Data Engineering, VOL. 27, NO. 5, (2015)279-284.
[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM ComputSurv, vol. 41, no. 3, p. 15, 2009.
[3] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Proc 13th Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD), pp. 813–822. 2009.
[4] C. Lijun, L. Xiyin, Z. Tiejun, Z. Zhongping, and L. Aiyong, "A data stream outlier detection algorithm based on reverse k nearest neighbors," in Proc 3rd Int Symposium on Computational Intelligence and Design (ISCID), pp. 236–239, 2010.
[5] H. P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in Proc 14th ACMSIGKDD IntConf on Knowledge Discovery and Data Mining(KDD), pp. 444–452, 2008.
[6] A. Zimek, E. Schubert, and H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statistical Analysis and Data Mining, vol. 5, no. 5, pp.363–387, 2012.
[7] Z. He, X. Xu, and S. Deng. "Discovering cluster-based local outliers", 2003, 24, 9-10, 1641{1650.
[8] D. Zhang and W. S. Lee., " A simple probabilistic approach to learning from positive and unlabeled examples", In UKCI, 2005.
[9] C. Elkan and K. Noto., "Learning classifiers from only positive and unlabeled data", In KDD, 2008.
[10] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," SIGMOD Rec, vol. 29,no. 2, pp. 93–104, 2000.
[11] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in Proc 10th Pacific-Asia Conf on Advances in Knowledge Discovery and Data Mining (PAKDD), 2006, pp. 577–593