Outlier Detection for Business Intelligence using Data Mining Techniques

Mohiuddin Ali Khan
Computer Science Department
Utkal University
Bhubaneshwar, India

Sateesh Kumar Pradhan
Computer Science Department
Utkal University
Bhubaneshwar, India

M. A. Khaleel
Computer Science Department
Sambalpur University
Sambalpur, India

ABSTRACT
In this paper we have made a review of various outlier detection techniques from data mining perspective. Existing studies in data mining focus generally on finding patterns from large datasets and using it for organizational decision making. However, finding exceptions and outliers did not receive much attention in the data mining field as other topics received. Finally, this paper concludes some advances in outlier detection recently.

Keywords: Data mining, Outlier, Business Intelligence, Architecture.

1. INTRODUCTION
Data mining is a process of extracting previously unknown information from large datasets, using it for organizational decision making [1]. Whereas an outlier is a data object that deviates significantly from the normal objects as if it were generated by a different mechanism. The identification ofoutliers can lead to the discovery of useful and meaningful knowledge. In recent years, credit card fraud, corporate fraud and financial fraud, has made a great deal of concern and attention. Outlier detection has been extensively studied in the recent years. However, most existing analysis and research focuses on the algorithm, based on special background, compared with outlier detection approach remains rare. Outliers can be categorized into two categories as classic outlier approach and spatial outlier approach. The classic outlier approach analyzes the outliers based on the transaction dataset, which can be grouped into statistical-basedapproach. However, there are several problems existing in mining data for large datasets such as data redundancy, incomplete data the value of attributes is not specific[2].Whenever there is an outlier there arouses a suspicious that it was generated by a different mechanism from the other observations [3]. The identification of outliers can lead tothe discovery of useful and meaningful knowledge andhas a number of practical applications in areas such as Business Intelligence, public safety, public health, credit card transactions, and location based services. The outlier detection is implemented in order to measure the distance between the data objects to detect those objects that are quite different from or inconsistent with the remaining data set. Recently, few studies have been conducted on outlier detection for large dataset [4]. This paper mainly discusses about outlier detection approaches from datamining perspective.

In a typical process mining scenario, a group of traces registering the sequence of tasks performed on many enactments of a transactional system, such as a Workflow Management (WFM), a Enterprise Resource Planning (ERP), a Customer Relationship Management (CRM), a Business to Business (B2B), or a Supply Chain Management (SCM) system – is given to hand, and the goal is to automatically derive a model explaining the scenario recorded in it. Eventually, the "mined" model can be used to design a detailed process schema capable of supporting a forthcoming validation, or to explore on its actual behavior.

2. PREVIOUS WORK
The classic definition of an outlier is due to Hawkins [3] who defines "an outlier is an observation that deviates so much from other observations as to arise suspicions that it was absolutely created by a different mechanism". Outliers are defined based on the probability distribution. Knorr et al. proposed a definition based on the concept of distance, which regard a point p in data set as an outlier with respect to the parameters K and L, if no more than k points in the data set are at a distance Lor lessthan p [6].Arninget a1. Proposed a deviation-based method, which identifies the outliers by inspecting the main characteristics of objects in a dataset and those objects that deviate from these features arc considered outliers [7].

Breunig et al. introduced the concept of local outlier, a kind of density-based outlier, which assigns each data a local outlier factor LOF of being an outlier depending on their neighborhood. The outlier factors can be computed very efficiently only if some multi-dimensional index structures such as R-tree and X-tree [8] are employed. A top-n based local outlier mining algorithm which uses distance bound micro-cluster to estimate the density was presented in [9]. Lazarevic and Kumar proposed a local outlier detection algorithm with a technique called feature bagging. Shekhar et al. proposed the definition of spatial outlier: "A spatial outlier is a spatially referred object whose non spatial values are significantly different from those of other spatially referenced objects in its spatial region".

Kou et al. developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weight when comparing non-spatial attributes. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbors. Liu and Jezek proposed a method for detecting outliers in an irregularly distributed spatial data set.

3. OUTLIER DETECTION APPROACH
Outlier detection has been studied in the past and numerous approaches have been made. Here we have created the cluster using K-means algorithm. In Figure 1 we have divided the datasets into three major clusters and there are some values which are outliers as they are far from other neighbors, hence these objects should be studied and examined carefully in order to minimize the web crime.

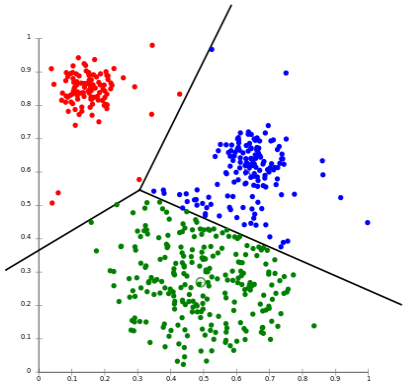

Figure 1: Cluster analysis

The outlier detection approach can be classified mainly in two categories

1: classic outlier approach and 2: spatial outlier approach. The classic outlier approach analyzes outliers based on the transaction dataset, which can be further grouped into statistical based approach, distance, deviation, density based approaches. The spatial outlier approaches analyzes the outliers based on spatial dataset, which can be grouped into space based approach and graph based approach as well.

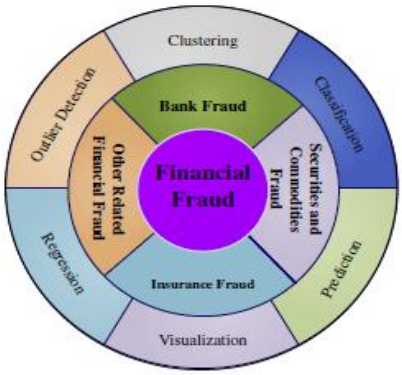

Figure 2: Classification of the applications of data mining in financial fraud

3.1. Classic Outlier

Classic outlier approach analyzes the outliers based on the transaction dataset, which consists of collection of items. A typical example is market basket analysis example, where each transaction is the collection of items purchased by a customer in a each transaction. Such data can also be augmented by additional items, describing the customer or the context of the transaction. Commonly, transaction data is relative to other data to be simple for the outlier detection. Thus, most outlier approaches are researched on transaction data.

(1) Statistical Approach Statistical approaches were the earliest algorithms used for outlier detection, which assumes a distribution or probability model for the given data set and then identifies outliers with respect to the model using a discordancy test. In fact, many of the techniques described in both Barnettand Lewis [5] and Rousseeuw and Leroy are single dimensional. However, with the dimensions increasing, it becomes more difficult and inaccurate to make a model for dataset.

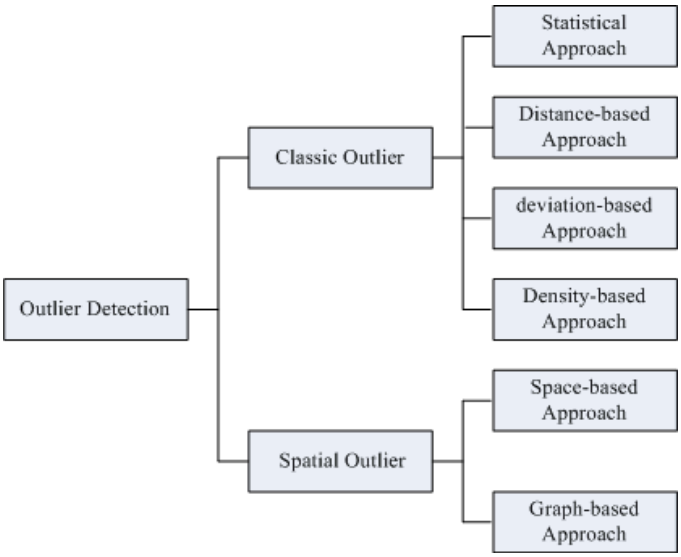

Figure 3: Outlier Detection Approach

(2) Distance-based Approach

The concept of distance-based outlier relies on the notion of their neighborhood of a point, the k nearest neighbors, and has been first introduced by Knorr and Ng [10]. Distance-based outliers are those points for which there are less than k points within the distance in the input data set. This definition does not provide a ranking of outliers and needs to determine an appropriate value of the parameter.

Ramaswamy et al. [9] modified the definition of outlier introduced by Knorr and Ng and consider as outliers the top n point's p whose distance to their k nearest neighbor is greatest. To detect outliers, a partition based algorithm is presented that, initially it partitions the input points using a clustering algorithm and, then, prunes those partitions that cannot contain outliers.

The distanced based approach is effective in small dimensions, because of the scarcity of large dimensional points, the approach is sensitive to the parameter L and it is hard to figure out a-priori. As the dimensions increase, the method's effectiveness and accuracy quickly decline.

(3) Deviation-based Approach Arninget a1. Proposed a deviation-based method, which identify outliers by inspecting the main characteristics of objects in a dataset and objects that deviate from these features arc considered outliers.

(4) Density-based Approach The density-based approach estimates the density, local outlier factor (*LOF*) to each point based on the local density of its neighbor, which is determined by a user-given minimum number of points (*MinPts*).

Papadimitriou et al. [11] present *LOCI* (Local Correlation Integral) which uses statistical values based on the data itself to tackle the issue of choosing values for *MinPts*.

Density-based techniques have the advantage that they can detect outliers that would be missed by techniques with a single, global criterion. However, data is usually sparse in high-dimensional spaces rendering density-based methods problematic. Distribution of the data and identifies outliers as those lying in low-density regions.

3.2. Spatial Outlier

For spatial data, classic approaches should be modified because of the qualitative difference between spatial and non-spatial attributes. Spatial datasets could be defined as a collection of spatially referenced objects, such as buildings and cities. Attributes of spatial objects fall majorly into two categories: spatial attributes and non-spatial attributes. The spatial attributes include location, shape and other geometric properties. Non-spatial attributes include building age, name, length, height the details of owner,

A spatial neighborhood of a spatially referenced object is a subset of the spatial data based on the spatial dimension using spatial relationships, Example distance and adjacency. Comparisons between spatially referenced objects are based on non-spatial attributes.

Spatial outliers are spatially referenced objects whose non-spatial attribute values are significantly different from those of other spatially referenced objects in their spatial neighborhoods. Informally, a spatial outlier is a local instability, or an extreme observation with respect to its neighboring values, even though it may not be significantly different from the entire population. Detecting spatial outliers is useful in many applications of geographic information systems and spatial dataset.

The identification of spatial outliers can reveal hidden but valuable information in many applications, For example, it can help locate severe meteorological events, discover highway congestion segments, pinpoint military targets in satellite images, determine potential locations of oil reservoirs, and detect water pollution incidents. (1) Space-based Approach

Space-based outliers use Euclidean distances to define spatial neighborhoods. Kou et al. developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weight when comparing non-spatial attributes. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbors[3]. Liu et al. proposed a method for detecting outliers in an irregularly-distributed spatial data set Graph-based Approach Graph-based Approach uses graph connectivity to define spatial neighborhoods. Yufeng Kou et al. proposed a set of graph-based algorithms to identify spatial outliers, which first constructs a graph based on k-nearest neighbor relationship in spatial domain, assigns the non-spatial attribute differences as edge weights, and continuously cuts high- weight edges to identify isolated points or regions that are much dissimilar to their neighboring objects. The algorithms have two major advantages compared with the existing spatial outlier detection methods: accurate in detecting point outliers and capable of identifying region outliers.

4. RECENT ADVANCES IN OUTLIER DETECTION

Along with the fast development of data mining technique, identification of outliers in large dataset has received more and more attention. Traditional outlier detection methods may not be efficiently applicable to large dataset. So some new methods are specially designed for special background.

(1) High Dimension-based Approach. High dimension space is a difficult problem for outlier detection. According to the criterion of the technique designed for high dimension proposed in the literature [4], a new method ODHDP based on the concept of projection is proposed in this paper, it can well deal with the sparcity of high dimensional points. The basic idea of the approach is to find the outliers by clustering the projections of data set. So, firstly, clustering the projections of data set in each dimension, and putting different weight to each dimension; secondly, selecting the dimension which has the maximum weight in the rest of dimensions for

Descartes combination clustering in turn, then pruning the candidate clusters in which the number of the points is less than threshold, until all dimensions are scanned; thirdly, computing the similarity of the points in the remains based on their relationship with the clusters in full dimension, by which the outliers is distinguished from the remains [10].

(2) SVM-based Approach

A SVM-based outlier detection approach was proposed [11]. The method uses several models of varying complexity to detect outliers based on the characteristics of the support vectors obtained from SVM-models. This has the advantage that the decision does not depend on the quality of a single model, which adds to the robustness of the approach. Furthermore, since it is an iterative approach, the most severe outliers are removed first. This allows the models in the next iteration to learn from cleaner data and thus reveal outliers that were masked in the initial model. Other outlier detection efforts include Support Vector approach[8], using Replicator Neural Networks (RNNs)[7], and using a relative degree of density with respect only to a few fixed reference points [12].

5. CONCLUSIONS

This paper mainly discusses about the different outlier detection approaches from data mining perspective. Firstly, we have reviewed the related work in outlier detection. Next, we discuss and compare various algorithms of outlier detection, which can be categorized into two categories classic outlier approach and spatial outlier approach. The classic outlier approach analyzes the outliers based on the transaction datasets, which can also be grouped into statistical based approach, deviation, distance, density based approaches. The spatial outlier approach analyzes outliers based on spatial dataset, which can be grouped into spacebased approach, graph-based approach. Also we conclude some advances in outlier detection recently.

6. ACKNOWLEDGMENT

With immense pleasure we would like to thank Dr. Sateesh Kumar Pradhan for his endless support and guidance in order to do the research on this research paper.

7. REFERENCES

[1] Data Mining:Concepts and Techniques, Third Edition, Jiawei Han and Micheline Kamber, ISBN-13, 978-0123814791.

[2] Data mining techniques by Arun K. Pujari, ISBN 9788173713804.

[3] Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, no. 1, 2010.

[4] Data Mining Techniques for E-Business Intelligence, International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October 2013, ISSN 2229-5518 by Mohiuddin Ali Khan and Sateesh K Pradhan.

[5] Mining students behavior in web-based learning programs Man Wai Lee, Sherry Y. Chen, Kyriacos Chrysostomou, Xiaohui Liu Expert Syst. Appl. 36(2): 3459-3464 (2009).

[6] Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, (IJCSIS) International

Journal of Computer Science and Information Security, Vol. 7, no. 1, 2010.

[7] Data mining in course management systems: Moodle case study and tutorialCristo´bal Romero , Sebastia´n Ventura, Enrique Garcı´a, Volume 51 Issue 1, August, 2008.

[8] A. Merceron and K. Yacef, " Educational Data Mining: A Case study," in 'Proc. Int. Conf. Artif Intell Educ., Amsterdam, The Netherlands, 2005 pp 1-8.

[9] A. F. D. Costa and J. T Lopes. Os Estudantes e os seus Trajectos no Ensino Superior: Sucesso e Insucesso, Factores e processos, Promocao de Boas Praticas. 2008 Retrieved July 2009 from http:// etes.cies.iscte.pt

[10] Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, 2005, vol. 14, pp. 211-221.

[11] Yufeng Kou, Chang-Tien Lu, Dos Santos, R.F." Spatial Outlier Detection: A Graph-Based Approach", ICTAI 2007,Volume 1, 2007,pp.281 – 288.

[12] Y. Kou, C.-T. Lu, and D. Chen. "Spatial weighted outlier detection". In Proceedings of the Sixth SIAM International Conference on Data Mining, pp. 614–618, Bethesda,Maryland, USA, 2006.