International Journal of Engineering &Technology, 7 (4) (2018) 6309-6312

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.29057
Research paper

A survey on outlier detection methods in data mining

Roy Thomas [1] *, J. E. Judith [2]

[1] *Research Scholar, Department of CSE, Noorul Islam Centre for Higher Education, Kumaracoil, India*
[2] *Associate Professor, Department of CSE, Noorul Islam Centre for Higher Education, Kumaracoil,India*
**Corresponding author E-mail: roygptc@gmail.com*

Abstract

Outliers are data objects whose characteristics differ from the mainstream characteristics of the data objects in a data set. Outlier detection plays a vital role in statistics as well as in data mining. Outlier detection effects to find out hidden and important information from large data sets. It has been a research field with diverse application areas for the past few decades. Outlier detection has been a topic of research in many fields like detecting malicious activity in cyber security, finding fake transactions in banking, detecting abnormality in medical data, identifying defects in industrial products etc. and various methods have been developed for detecting outliers. Most of the methods are developed specifically for certain applications while others are generic methods. Outlier detection methods are grouped into supervised, unsupervised and semi-supervised methods depending on the availability of class labels. Outlier detection methods can also be classified into statistical, proximity-based, clustering-based and classification-based depending on the type of data. We, in this paper, present the relative advantages and limitations of various methods used for detecting outliers.

Keywords: *Classification; Clustering; Data mining; Outliers; Proximity.*

1. Introduction

The volume of data in the world is increasing tremendously at every moment. Data mining is a field in computer science that is intended to find important information from large collections of data sets most of which are unstructured. It is the discovery of the frequent patterns from the data sets.

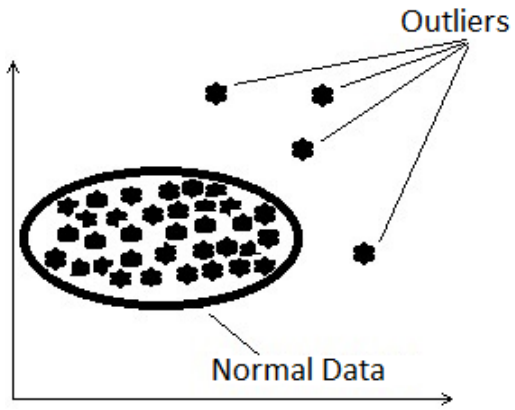

Fig. 1: Outliers.

Outlier detection is aimed at finding infrequent data objects that are not conformable with the majority of data objects in the data set. As shown in Fig.1, an outlier is a data object that is distant from most of the remaining data objects in the data set which arouses suspicion that it does not belong to the data set itself [1]. Outlier detection is defined as the discovery of data objects that are dissimilar or uneven with respect to the majority of data objects in the data set.

Outlier detection has become an essential research field as it provides critical information in a variety of domains. Outlier detection methods can be applied to detect unauthorized access to resources in a computer network, to find fake credit card transactions in banking, to detect abnormality in medical data analysis, to identify the defects in industrial products, to detect imperfections in image processing, to detect anomalies in web applications, to detect irregularities in robot behavior, to discover variances in astronomical data, to find discrepancies in census data etc.

Outliers can be classified into different types such as point outliers, collective outliers and contextual outliers [2]. A data object which differs considerably from other data objects in its set is called a point outlier. For example, normal credit card transactions follow a regular pattern. In the case of a fraud transaction, the pattern may change significantly. The features of the credit card transactions like the type of items purchased, their quantity, etc. by the fraud shall be very much different from the normal purchase features of the credit card transaction by the authenticated card owner. Contextual outlier is a data object that differs considerably from the data set based on a selected situation and is a normal data object with reference to some other situation. For example, rain in summer is a contextual outlier. Collective outlier is a subgroup of data objects which collectively differs from the entire data set. The distinct data objects in a collective outlier are not outliers individually. For example, heavy traffic for a limited time is a normal occurrence in a city. However, if it continues for a long period of time, it becomes a collective outlier.

2. Related work

Outlier detection methods have been a research subject in different research fields and various application domains. Hodge and Austin [3] observed outlier detection approaches as extracted mainly from three fields of computing – statistical, neural networks and

machine learning. Hodge et al. classified the outlier detection approaches into three categories Type-I, Type-II, and Type-III which are analogous to unsupervised, supervised and semi-supervised approaches respectively. Patch and Park [4] accumulated the methods into four groups known as statistical based, classification based, clustering based and nearest neighbor based. Chandola et al. [5] added two more groups of outlier detection methods called spectral methods and information theoretic methods. Han et al. [2] presented methods for detecting collective and contextual outliers as well as methods for finding outliers in subspaces and high dimensional data. Gogoi et al. [6] classified the outlier detection approaches into three categories and classified the outliers into six cases to identify network anomalies. Zimek et al. [7] presented a survey on unsupervised methods for detecting outliers in high dimensional numerical data. Lu et al. [8] proposed a density based structured framework for sequential outlier detection that uses autoencoder models to catch the dissimilarity between outliers and normal data objects. Wu et al. [9] formulated outlier detection as an optimization problem for detecting outliers in extensive categorical data. Papadopoulos et al. [10] proposed graph based descriptors to capture the network billing related outliers in cellular mobile networks.

3. Outlier detection methods

Outlier detection methods are grouped into supervised, semi-supervised, and unsupervised methods depending on the availability of class labels [2].

3.1. Supervised outlier detection

Supervised approaches are used to separate data objects into outliers and inliers in a labeled data set, in which data objects can be labeled as outlier or inlier. These labels are then used to make the model of the normal class or outlier class. However in most situations, the occurrences of outliers in the data sets are very insignificant compared to the extent of normal objects in the data sets. Peculiar classification techniques are needed to detect outliers from such unbalanced data sets.

3.2. Unsupervised outlier detection

Unsupervised methods are used to find outliers in an unlabeled data set by giving each object an outlier score which indicates its degree of abnormality. These scores are usually calculated by comparing the characteristics of each data object with the characteristics of its neighborhood. Unsupervised methods do not need a training data set and assume that the occurrences of outliers are quite small compared to the occurrences of normal data objects in the data set.

3.3. Semi-supervised outlier detection

Semi-supervised methods are used when labels are available for only a small part of normal objects or outliers. The labeled normal objects along with objects which are close to the labeled normal objects are used in the training phase to obtain a model of the normal objects. This information is used to separate the data objects into two sets in which one set contains all the normal objects and the other set contains the remaining objects or the outliers.

Based on the assumptions made on outliers versus normal data, outlier detection methods can be grouped into different categories like statistical methods, proximity based methods, classification based methods and clustering based methods [5].

3.4. Statistical methods

Statistical outlier detection methods assume that normal data objects follow a statistical model, and outliers do not follow this model [3]. Data objects are classified into normal data objects and outliers through a statistical test in which the normal objects occur in high probability region of the model and outliers occur in low probability regions. Statistical models are mostly suitable for quantitative numerical data sets and are not appropriate for categorical data sets. In order to apply statistical model for ordinal data, those data have to be converted into numerical values by suitable transformation. This limits the applicability of statistical techniques and increases the preprocessing complexity.

Parametric methods assume the knowledge of a parametric distribution model to fit the normal data. A small value for the probability density function of the parametric distribution indicates that the data object is an outlier.

Nonparametric methods are used when a prior knowledge of the distribution model is not available. In nonparametric methods, a statistical model is determined based on the nature of the data objects in the data set. Parametric methods are faster than nonparametric methods as the model structure is already determined. The advantage of nonparametric methods is the flexibility in determining the distribution model structure.

Semi-parametric methods combine the advantages of both the parametric and the nonparametric methods. Semi-parametric methods are not limited by a single distribution model, but they use multiple local distribution models.

3.5. Proximity based methods

Proximity based methods are developed on the assumption that the distance from a normal data object to its nearest neighbors is very small compared to the distance from an outlier to its nearest neighbors. Normal data instances are assumed to form dense regions and outliers form sparse regions. Proximity based methods need a similarity or distance measure defined between two data objects to compute the outlier score of each data object. The effectiveness of the methods relies on this measure. Proximity based methods have limited use where a similarity or distance measure cannot be obtained and they are not suitable to detect collective outliers. Proximity based outlier detection methods can be divided into different categories like distance based methods, density based methods etc.

Distance based methods compute the distances among data objects using a suitable distance measure or similarity measure. A data object is considered as an outlier if its distances to most of the data objects are comparatively greater than its distance to other data objects. Distance based methods are not suitable for determining outliers in high dimensional data because of its complexity in calculating distances among all data objects. The outlier score assigned with each data object is determined from the distance to its nearest neighboring data objects.

Density based methods consider the data objects lying in low density regions as outliers and objects in high density regions as normal data objects. These methods are developed on the assumption that there is not any significant difference for the density around the normal objects and the density around their neighbors whereas there is substantial difference for the density of outliers and the density of their neighbors.

3.6. Classification based methods

Classification based approaches use learned classifiers from a data set containing labeled data objects to classify the data objects into classes [5]. These methods build a classifier in the training phase using the labeled data objects and classify the data objects during testing phase into normal objects or outliers.

Based on the nature of class labels, the classification based methods are categorized into one-class and multi-class classification. One-class classification describes normal data objects only and all data objects rejected by the testing phase belong to the outliers. In multi-class classification, boundaries among multiple classes are learnt and test data objects belong to one of the classes which can be either normal classes or outliers.

3.7. Clustering based methods

Clustering based methods divide the data set into different groups called clusters, where each cluster contains similar objects. Data objects in large and dense clusters are normal data objects and other data objects are outliers. Hence outliers can be data objects in small and sparse clusters, or can be data objects that are not part of any cluster, or can be data objects within a cluster which is far from all other clusters [2].

3.8. Comparison of outlier detection methods

The advantages and limitations of various outlier detection methods are given in Table 1. This table shows that the suitability of a particular method depends on the data set domain.

Table 1: Comparison of Outlier Detection Methods

Method	References	Advantages	Limitations
Parametric	Hodge et al [3] Neagu et al [11]	Once the model is learned, simple to implement. Computational complexity is less.	Prior knowledge of data distribution is needed. For sophisticated models, cost for approximating the best parameter values is high.
Non-parametric	Wang et al [12] Hodge et al [3]	Prior knowledge of model structure is not needed Supervised setting is not needed.	Difficult to construct statistical test for high dimensional data sets. Difficult to capture contextual anomalies.
Neural Networks-Based	Lu et al [8] Ferdosi et al [13] Wang et al [14]	Applicable for multi-class settings. Can learn complex class boundaries.	Susceptible to the curse of dimensionality. Suffer slow training.
Support Vector Machines-Based	Liu et al [15] Rajasegarar et al [16]	Faster for one-class setting. Applicable to poorly balanced data sets.	Computationally complex. Not suitable for multi-class settings.
Rule-Based	Ekizoglu et al.[17] Kao et al [18]	Applicable for multi-class settings. More flexible and incremental.	Needed accurate data labels, which may not be available. Computational cost is high for complex domains like text and pattern matching.
Partitioning	Jia et al.[19] Anguilli et al.[20]	Effective for small- to medium-size data sets. Less computational complexity.	Difficult to find clusters of arbitrary shape. Count of clusters to be defined initially.
Hierarchical	Zhang et al. [21] Xu et al. [22]	No need to define the count of clusters initially. Less computational time.	Cannot correct erroneous merges or splits. Not suitable for streaming data.
Density based	Mandhare et al [23] Gagoi et al [6]	Can be used with arbitrarily shaped clusters. Need not specify the count of clusters initially.	Difficult for border points. Threshold setting is difficult for large and dynamic data sets.
Distance based	Anguilli et al.[20] Papadopoulos et al [10]	Algorithms are simple and easy to implement. Applicable to multiple domains.	Difficulty in detecting a group of outliers Computational cost is high for large data sets.
Grid based	Gu et al [24] Xiang et al [25]	Complexity is linear to data size. Does not make any assumption on the underlying distribution.	Grid resolution is difficult for sparse and large data sets. It is a compromise between performance and accuracy.

4. Conclusion

In this fast growing world, outlier detection plays a vital role in a number of application domains. This paper presents the importance of outlier detection in data mining and provides the features of various outlier detection methods in the literature. We have presented different methodologies for outlier detection in data mining area and discussed the merits of each method and their application domain. There is not any single generic outlier detection method, but specific methods are to be used based on the assumptions on class labels, type of data, scalability, distribution model, dimensionality etc. A comprehensive survey of the various categories of outlier detection methods such as statistical methods, proximity based methods, classification based methods and clustering based methods is presented in this paper. The classification of outlier detection methods depending on class labels into supervised, unsupervised and semi-supervised has also been presented.

References

[1] D. Hawkins, Identification of Outliers.Chapman and Hall, London and New York, 1980. https://doi.org/10.1007/978-94-015-3994-4.

[2] J. Han, M. Kamber, and J Pei. Data Mining: Concepts and Techniques, Massachusetts (US): Morgan Kaufmann, 2012.

[3] V. J. Hodge and J. Austin, A survey of outlier detection methodologies, Artificial Intelligence Review, vol. 22 (2), pp. 85-126, 2004. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

[4] A. Patcha, and J-M. Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends, Computer Networks, 2007. https://doi.org/10.1016/j.comnet.2007.02.001.

[5] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: A survey,ACM Comput. Surveys, vol. 41, no. 3, pp. 1–58, 2009. https://doi.org/10.1145/1541880.1541882.

[6] P. Gogoi, D.K.Bhattacharyya, B. Borah, J.K.Kalita, A Survey of Outlier Detection Methods in Network Anomaly Identification,OUP, The Computer Journal,2011. https://doi.org/10.1093/comjnl/bxr026.

[7] A. Zimek, E. Schubert, and H-P Kriegel, A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data, Wiley Periodicals, Inc.,2012. https://doi.org/10.1002/sam.11161.

[8] W. Lu, Y. Cheng, C.Xiao, S.Chang, S.Huang, B.Liang, and T.Huang, Unsupervised Sequential Outlier Detection With Deep Architectures, IEEE Transactions on Image Processing, 2017, Volume: 26, Issue: 9. https://doi.org/10.1109/TIP.2017.2713048.

[9] S. Wu, and S. Wang Information-Theoretic Outlier Detection for Large-Scale Categorical Data, IEEE Transactions on Knowledge and Data Engineering, Volume: 25, No:3, March 2013. https://doi.org/10.1109/TKDE.2011.261.

[10] S. Papadopoulos, A.Drosou, and D.Tzovaras A Novel Graph-Based Descriptor for the Detection of Billing-Related Anomalies in Cellular Mobile Networks, IEEE Transactions on Mobile Computing, 2015. https://doi.org/10.1109/TMC.2016.2518668.

[11] B. C. Neagu, G. Grigoraș, F. Scarlatache, Outliers Discovery from Smart Meters Data Using a Statistical Based Data Mining Approach, IEEE 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE).2017. https://doi.org/10.1109/ATEE.2017.7905046.

[12] W. Wang, Y.Liang, H. V. Poor, Nonparametric composite outlier detection , 2016 50th Asilomar Conference on Signals, Systems and Computers, Year: 2016. https://doi.org/10.1109/ACSSC.2016.7869574.

[13] H. Ferdowsi,S.Jagannathan, and M. Zawodniok, An Online Outlier Identification and Removal Scheme for Improving Fault Detection Performance, IEEE Transactions on Neural Networks and Learning Systems, Volume:25, No:5, May 2014. https://doi.org/10.1109/TNNLS.2013.2283456.

[14] C. Wang, J.Lai , D.Huang , and W.Zheng, SVStream: A Support Vector-Based Algorithm for Clustering Data Streams, IEEE Transactions on Knowledge and Data Engineering, Year: 2013 , Volume: 25 , Issue: 6. https://doi.org/10.1109/TKDE.2011.263.

[15] B. Liu, Y. Xiao , P.S. Yu , Z. Hao,and L. Cao, An Efficient Approach for Outlier Detection with Imperfect Data Labels, IEEE Transactions on Knowledge and Data Engineering, Year: 2014 , Volume: 26 , Issue: 7. https://doi.org/10.1109/TKDE.2013.108.

[16] S. Rajasegarar, C. Leckie , J.C. Bezdek , M. Palaniswami, Centered Hyperspherical and Hyperellipsoidal One-Class Support Vector Machines for Anomaly Detection in Sensor Networks, IEEE Transactions on Information Forensics and Security, Year: 2010 , Volume: 5 , Issue: 3. https://doi.org/10.1109/TIFS.2010.2051543.

[17] B. Ekizoglu , A.Demiriz, Fuzzy rule-based analysis of spatio-temporal ATM usage data for fraud detection and prevention , 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Year: 2015. https://doi.org/10.1109/FSKD.2015.7382081.

[18] L. Kao, Y. Huang, Association rules based algorithm for identifying outlier transactions in data stream, 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Year: 2012. https://doi.org/10.1109/ICSMC.2012.6378285.

[19] H. Jia, Y. Cheung, and J.Liu , A New Distance Metric for Unsupervised Learning of Categorical Data, IEEE Transactions on Neural Networks and Learning Systems, Volume:27, No:5, May 2016. https://doi.org/10.1109/TNNLS.2015.2436432.

[20] F. Angiulli, S.Basta, and C.Pizzuti, Distance-Based Detection and Prediction of Outliers,IEEE Transactions on Knowledge and data engineering, year:2006,Vol:18, Issue: 2. https://doi.org/10.1109/TKDE.2006.29.

[21] Q. Zhang,M. Qiao, R. R. Routray, and W. Shi ,H2O: A Hybrid and Hierarchical Outlier Detection Method for Large Scale Data Protection, IEEE International Conference on Big Data (Big Data), 2016. https://doi.org/10.1109/BigData.2016.7840715.

[22] T-S. Xu, H-D. Chiang,G-Y. Liu, and C-W Tan, Hierarchical K-means Method for Clustering Large - Scale Advanced Metering Infrastructure Data, IEEE Transactions on Power Delivery,Volume: 32 , Issue: 2 , April 2017. https://doi.org/10.1109/TPWRD.2015.2479941.

[23] H. C. Mandhare, and S. R. Idate, A Comparative Study of Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques, IEEE International Conference on Intelligent Computing and Control Systems, 2017. https://doi.org/10.1109/ICCONS.2017.8250601.

[24] Y. Gu, R. K. Ganesan, B. Bischke, A. Bernardi, A. Maier, H. Warkentin, T. Steckel, and A. Dengel,. Grid-based outlier detection in large data sets for combine harvesters. IEEE 15th International Conference on Industrial Informatics (INDIN), 2017. https://doi.org/10.1109/INDIN.2017.8104877.

[25] Y. Xiang, L. Guohua, X. Xiandong, and L. Liandong, A data stream outlier detection algorithm based on grid. IEEE The 27th Chinese Control and Decision Conference, 2015. https://doi.org/10.1109/CCDC.2015.7162657.