Outlier Detection and Processing Technology and Its Application

Yong Wang, Jiahe Cui,Zebao Zhang, Zhigang Li
College of Computer Science and Technology
Harbin Engineering University,Harbin, China
wangyongcs@hrbeu.edu.cn

Keywords: Outlier detection,Data cleaning,Attribute anomalies,Duplicatedrecords.

Abstract. Faced with the development of information technology and the arrival of the era of big data, data has become the advanced productive forces and strategic resources. However, the existence of outliers often has a negative impact on the storageand analysis of the data, andeven causes disastrous consequences. This paper starts with the definition, concept and classification of outliers, and then compares and analyzes various methods of outlier detection and cleaning technology based on attribute anomaly and the duplicate or similarrecords.Finally, the research and application of outliers are prospected.

Introduction

In recent years, with the rapid development of society and economy, the informatization degree of various fields and industries has been greatly improved, and accumulated a lot of data. Data quality issues have become increasingly prominent with the rapid expansion of data. In situations where big data analysis and application have had a substantial impact on decision-making, the presence of outliers can affect the results of the analysis and may even have serious consequences.

On the other hand, outliers are not only regarded as noise, isolated point or novelty, but also have important semantics. Through analyzing the anomaly data, it can find hidden information and potential value.

The detection and cleaning of outliers has become one of the research hotspots in data-warehouse, data mining and comprehensive data quality management. It has important practical significance for public health, weather forecasting, finance, network security and government decision-making.

Definition and Classification of Outliers

Definition. High quality data should have the following characteristics: accuracy, completeness, consisitency and minimality[1]， the existence of outliers is contrary to such expectations of data quality.

Hawkins [2] put forward the essential definition of outliers in 1980: Outliers is the data set out of the unusual data, which makes it suspect that these data are not random errors, but are generated in a completely different mechanism.

Outlier[3] is also called noise data, isolated points, novel points, departure points, exception points, rare class, abnormal data, minorityclass and so on.

Classification. (a) Classification method based on causation**.** Outliers may be caused by errors in the process of data collection or mining, or due to loss of timeliness (inherent variability of data [4]). Common outlier types and their causes[5,6] are shown in Table 1 below.(b) Classification method based on the linkage of data points. According to the definition of data set, a data set is a set of data points, with each data point represented by a series of features. Generally speaking, each data point is connected with each other. According to the linkage of data points, such outliers can be divided into three categories, as shown in Table 2.

Table1 Classification Method Based On Causation

Classification	Outlier Type	Causation
Abnormal attribute	Missing value	The real value for the field has not been determined;Missing data item when typing;The integrity constraint is missing
	Misspelling	Manual input errors;Lack of data validation mechanism
	Multiple information exits in single field	Design defects of schema
	The value does not match the field name	Heterogeneous data integration process lacks control
Duplicate recording[7]	Similar or duplicate records	Multiple records correspond to the same real entity
	Contradictory records	An attribute of the same real entity has several different values in different records

Table2 Classification Method Based On Linkage of Data Points

Outlier Type	Description
Point anomaly	Compared with other points,anisolatedpoint is anomalous in the overall data set
Context anomaly	In a specific context, the data point is abnormal
Setanomaly	An interrelated set of data is abnormal relative to the entire data set

Detecting and Cleaning Technology of Outliers

Abnormal attribute value detection. The attribute errors caused by accidental reasons such as human error are usually noise data, while the attribute anomalies caused by the system evolution or the inherent timeliness of the data are usually the meaningful outliers. The detection of attribute anomalies is the basis of outlier mining and analysis.

The method of manually detecting attribute anomalies is usually done by domain experts. Although this method can guarantee the accuracy of abnormal data detection, it will consume a great deal of manpower and material resources, and it is very inefficient. Therefore, efficient algorithm for automatic attribute detection is an important approach to attribute anomaly detection, including statistical methods, neural network based methods, association rules based methods, etc. Table 3 shows some typical anomaly attributes Value detection method.

Table 3 Data Detection Method Based on Attribute Anomaly

Classification	Method	Basic Idea	Privilege	Disadvantage
MethodsBased on Statistical Principle	Algorithm based on simple statistical principle	Suppose that the sample obey Gaussian distribution [10] or regressive model[11], samples deviated from the mean or regressive directionare abnormal	If the data set is known to obey a certain distribution, its detection accuracy is high	Easily affected by the "dimension disaster" with high data dimension; the data often will not obey a certain distribution

Table 3 Data Detection Method Based on Attribute Anomaly (Continued)

Classification	Method	Basic Idea	Privilege	Disadvantage
MethodsBased on Statistical Principle	Histogram-Based Algorithm [12]	Establish feature histogram through the training set.Those do not fall into the corresponding histogram areoutliers	The algorithm is intuitive and easy to understand	Not suitable for high dimensional data
	k-NearestNeighbor Algorithm (KNN[13])	Conventional data are always close to each other in the feature space, while oueliers are always far away from the neighbor	Do not make any assumptions about the data set, applicable to a variety of data sets	Does not suitable for the case of unbalanced sample [8]; high algorithm complexity
	Kernel DensityEstimation[14]	Construct kernel density estimation function ;the outliers are determined based on the density threshold	The generalization ability is strong, and it can be guaranteed to converge to the complex unknown density.	High Algorithm complexity; which is not conducive to the construction of efficient clustering algorithm
	Clustering Algorithm [9]	Each point in the dataset has a membership degree for different clusters, the pointswith membership degree lower than the threshold are outliers	The algorithm is simple and efficient	A large number of unrelated attributes in high dimensional data reduces the possibility of clustering; and the data distribution is sparse
Methods Based on Neural Network	Algorithm based on MLP (Multilayer Perceptrons) [15]	A backward-propagated supervised learning method is used to train the network, through the input layer, node hidden layer, and the output layer to detect outliers	Can deal with the problem of nonlinear separability	Learning speed is slow, easy to fall into local extremes, learning is not sufficient
	Algorithm based on SOM(Self-Organizing Maps) [16]	High-latitude data distribution is visualized in a lower-dimensional topology to detect outliers	The network has self-stability, and no external evaluation is needed	The convergence time is too long; number of clusters needs specify in advance
Rule-based Method	Association rules method	Through a relationship between data entities, a series of rules are defined to define the exception pattern	The algorithm is simple and easy to implement	Experience of specific domain experts is required

Duplicate or similar records detection. Duplicate or similar records mean that two or more records with the same or similar attribute values in the data set represent the same entity in the real world. Such anomaly data has a serious impact on the integrity, consistency, and analysis of the data set.

For the detection of duplicate records, the basic method is to merge - sorting algorithm, that is, first sort the data set to make similar records cluster, then compare the similarity of recordsint the same cluster to detect duplicate data, as shown in Fig. 1.

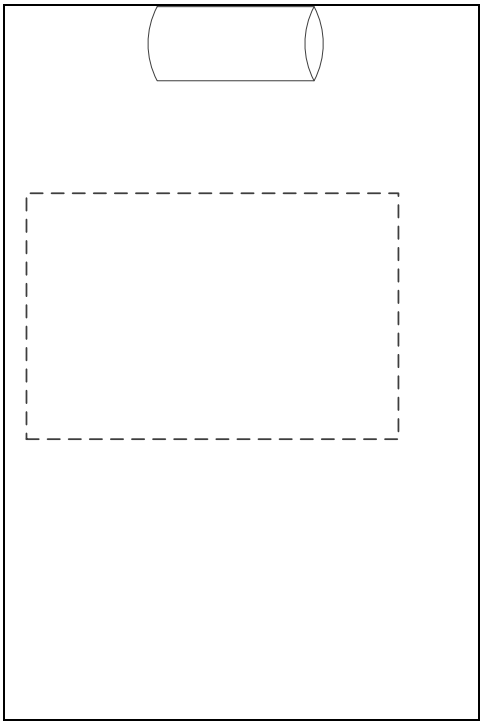

Fig. 1. Duplicate Similarity Records Cleaning Process

It should be notedin Fig. 1that①the selection of the sorting key and ②the matching of the field similarity are the key to the duplicate similarity records detection. The selection of keywords determines whether the sorting of data sets can accurately cluster duplicate records to adjacent positions, and the matching of field similarity determines whether the neighboring data can be accurately judged as similar duplicated data. Field similarity matching algorithm is basic to similar or duplicate record detection.The algorithm is mainly based on the string matching algorithm, etc, as shown in Table 4.

Table 4 Field Matching Degree Algorithm

Method	Basic Idea	Privilege	Disadvantage
Basic String Matching Algorithm	String matching algorithm based on KMP or LCS	The algorithm is simple, intuitive and efficient	Very limited, can not handle the case of sub string shift or abbreviation
Algorithm Based on Edit Distance	Comparing the minimum operation cost of two-string conversion [19]	Good at detecting short substring miss or repetition and spelling errors	Not suitable for cases where the substring position is reversed and absence or duplication of a longer substring
Smith-Waterman distance algorithm [18]	Calculating the similarity matrix	Do not rely on domain knowledge;Have good adaptability in fields with missing information or spelling errors	Can not handle the case of the sub string reversal

The time complexity of cleaning algorithm based on merging - sorting duplicate record detection is high, so researchers propose Basic SortedNeighborhoodMethod andMulti-PassSortedNeighborhood algorithm [17,20], as shown in the table

Table 5 Comparison of Similar Duplicate Data Detection Methods

Method	Basic Idea	Privilege	Disadvantage
Merge - Sort method	First sort and clusterthe similar records, then compare the field matching degree	The algorithm is simple and intuitive	The time complexity of the algorithm is high and the accuracy is low, which makes it easy to miss similar records
Basic SortedNeighborhoodMethod(SNM)	When the selected keywords are sorted, only the field match degrees within the fixed window are compared	More efficient	The accuracy of the algorithm depends on the selection of the keys and the size of the windows
Multi - PassNeighbor Sorting Algorithm (MPN)	Multiple sorting, using different keywords each time, and using a smaller sliding window, duplicate data are determined through mutiple calculating	The omission degree of the similarity record is less	Transitivity of similar records may cause detection errors

Application of Outlier Detection

Outlier detection technology has important applications in various fields. In the ETL (extract-transform-load) part of the data warehouse construction process, the detection and processing of outliers is a very important step [21], most ETL frameworks provide the API for outlier cleaning [22, 23], and the accuracy of the detection has a direct impact on the confidence level of decision making based on data warehouse; Nowadays, as a kind of self-descriptive, semi-structured mainstream data transmission standard, XML has its complex hierarchical structure and data source[24]. The outlier detection technology based on XML is more complex than it of conventional data set. The research of XML has important significance with a rapid increase of WEB data [25,26]; In the financial field, outlier detection technology can be applied to departments such as banks, which have high requirements for data accuracy and security, etc., such as analyzing unusual account information, abnormal bank card business, and detecting duplicate customer data.

In addition to detecting anomalous noise data generated by human or accidental factors, anomaly data detection can also be performed by analyzing the semantics implied by outlier, since anomaly data often reflect some evolution and variation of the physical world. As in the field of seismology, anomalous data mining and precursor data observation can be used to detect outliers in earthquake precursor data and provide the basis for earthquake prediction [27]. In the field of Internet security, real-time anomaly data detection and analysis of network packets can be used for intrusion detection [28]. Outlier detection also provides a theoretical and methodological basis for outlier mining[29].

Conclusion and Prospect

Good data quality is the guarantee of correct data analyzing and decision making. Under the current situation where massive data has been accumulated in various fields, the importance of outlier detection, analysis and processing technology for improving data quality, ensuring data availability and abnormal data mining is self-evident.

Although many researchers have done a lot of researches on outlier detection and analysis, there are still some problems to be explored:

(1) Outlier detection in dynamic data (or data flow). At present most of the outlier detection algorithm is based on static data sets. Facing the characteristics of rapidly (Velocity in 4Vs) data generation, it has important practical significance to study how to construct high efficient, real-time and accurate outlier detection algorithm in dynamic data set at the time of rapidly change;

(2)Outlier detection in the era of big data.In big data environment, the amount of data is soaring and complicated. Traditional outlier detection technology can not guarantee its correctnessand efficiency. How to find outliers for exponential growth of heterogeneous data is worthy of further study;

(3) Outlierdetection of domain data. Traditional anomaly detection methods, which lack domain knowledge and ignore the potential semantics, can not fully reflect the real causes of outliers and business logic. How to improve the outlier detection effect of the domain data and the practical application value of the analysis results need further research.

Acknowledgment

The research work was supported by The Fundamental Research Funds for the Central Universities under Grant No.HEUCF100610.The National Key Technology R&D Program of the Ministry of Science and Technology under Grant No. 2012BAH81F02, and The Youth Foundation of Heilongjiang Province of China under Grant No. QC2016083.

References

[1]Davoudi S, Dooling J A, Glondys B, et al. Data Quality Management Model (Updated).[J]. Journal of Ahima, 2015, 83(7):62-71.

[2] HawkinsDM. IdentificationofOutliers[J]. MonographsonAppliedProbability&Statistics, 1980, 80(2):21-8.

[3] Wang T, Li Z. Outlier detection in high-dimensional regression model[J]. Communications in Statistics-Theory and Methods, 2016.

[4] Han Han, Xu Lizhen, Dong Yisheng. Review of data quality research [J]. computer science, 2008,02:1-5+12.

[5] Rahm E, Hong H D. Data Cleaning: Problems and Current Approaches[J]. IEEE Data Engineering Bulletin, 2000, 23(23):3-13.

[6]Song X, Wu M, Jermaine C, et al. Conditional anomaly detection[J]. IEEE Transactions on Knowledge & Data Engineering, 2007, 19(5):631-645.

[7] Tamilselvi J J, Saravanan V. Detection and Elimination of Duplicate Data Using Token-Based Method for a Data Warehouse: A Clustering Based Approach[J]. International Journal of Computational Intelligence Research, 2009, 2:145-164.

[8]Bijalwan V, Kumar V, Kumari P, et al. KNN based machine learning approach for text and document mining[J]. International Journal of Database Theory and Application, 2014, 7(1): 61-70.

[9]Krishnan S, Wang J, Wu E, et al. Activeclean: Interactive data cleaning while learning convex loss models[J]. arXiv preprint arXiv:1601.03797, 2016.

[10]DudarO,HartpE.Pattern classification[M].NewYork:JohnWilry&Sons,2001:87-92

[11]Chen C, Liu L M. Joint Estimation of Model Parameters and Outlier Effects in Time Series[J]. Journal of the American Statistical Association, 2012, 88(421):págs. 284-297.

[12]Rahm E, Hong H D. Data Cleaning: Problems and Current Approaches[J]. IEEE Data Engineering Bulletin, 2000, 23(23):3-13.

[13]Östermark R. A fuzzy vector valued KNN-algorithm for automatic outlier detection.[J]. Applied Soft Computing, 2009, 9(4):1263-1272.

[14]Steury T D, Mccarthy J E, Roth T C, et al. Evaluation of Root‐n Bandwidth Selectors for Kernel Density Estimation[J]. Journal of Wildlife Management, 2014, 74(3):539-548.

[15]Markou M, Singh S. Novelty detection: a review—part 2: : neural network based approaches[J]. Signal Processing, 2003, 83(12):2499-2521.

[16] WANG Jiawei, HAN Bingqing, CHEN Dafeng, et al.SOM-based Unifying framework for mining outliers [J] .Application Research of Computers, 2007, 24 (10): 44-47.

[17]Burrows M. Technique for deleting duplicate records referenced in an index of a database: US, US6745194[P]. 2004.

[18]Smith T F, Waterman M S. Identification of common molecular subsequences[J]. Journal of molecular biology, 1981, 147(1): 195-197.

[19]RistadES, YianilosPN. LearningString-EditDistance[J]. IEEETransactionsonPatternAnalysis&MachineIntelligence, 1998, 20(5):522-52.

[20]Xiao-Sheng Y U, Sun-Zhi H U. Research on Eliminating Duplicate Records Based on SNM Improved Algorithm[J]. Journal of Chongqing University of Technology, 2016:91-95.

[21] Ghosh S, Goswami S, Chakrabarti A. Outlier detection from ETL execution trace[C]// Electronics Computer Technology. 2012:343 - 347.

[22] Hanine M, Boutkhoum O, Tikniouine A, et al. Application of an integrated multi-criteria decision making AHP-TOPSIS methodology for ETL software selection[J]. Springerplus, 2016, 5(1):1-17.

[23] Milman I M, Oberhofer M, Saillet Y. CODE ANALYSIS FOR PROVIDING DATA PRIVACY IN ETL SYSTEMS:, US20160246986[P]. 2016.

[24] Ciancarini P, Tolksdorf R, Zambonelli F. Coordination middleware for XML-centric applications[J]. 2016:336-343.

[25] Cuzzocrea A, Manco G, Masciari E. Effective Detection of XML Outliers[J]. 2012:1221-1232.

[26] Milano D, Scannapieco M, Catarci T. Using Ontologies for XML Data Cleaning[J]. Lecture Notes in Computer Science, 2005, 3762:562-571.

[27] Wu Y M, Chen D Y, Lin T L, et al. A High-Density Seismic Network for Earthquake Early Warning in Taiwan Based on Low Cost Sensors[J]. Seismological Research Letters, 2013, 84(6):1048-1054.

[28] Zhang J, Zulkernine M. Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection[J]. 2006, 5:2388-2393.

[29] Xiang X U, Liu J W, Luo X L. Research on outlier mining[J]. Application Research of Computers, 2009, 26(1):34-40.