ISSN 2581 – 3781

AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

Available Online at www.ajcse.info
Asian Journal of Computer Science Engineering 2022;7(1):1-10

RESEARCH ARTICLE

A Comprehensive Study on Outlier Detection in Data Mining

Deepti Mishra[1], Devpriya Soni[2]

[1]Department of CSE, Noida International University, Noida, Uttar Pradesh, India, [2]Department of CSE, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

Received on: 25-11-2021; Revised on: 30-12-2021; Accepted on: 10-01-2022

ABSTRACT

The paper presents a survey on the literature of outliers and data mining. The prime focus of the paper is to deliver an outline of outlier with various approaches for its detection. Outliers are the data points which are partially or totally diverge from the residue data set. They can be considered as those data objects which cannot be fitted in any cluster. Outliers can be different from its neighboring data points only or from complete data set. It is a necessary task to identify and detect outliers from the dataset as their presence effect the preciseness of the outcome. Outliers can exist in any kind of data varies from low-dimensional to high-dimensional data set. The detection of an outlier requires some precise mathematical calculations, appropriate domain knowledge, and statistical calculations which are presented in the paper. The paper presents, the significant characteristics of the outliers.

Key words: Clustering, Data mining, Knowledge discovery, Outlier detection, Outliers

INTRODUCTION

It is exceedingly difficult to manage large scale collected data, as nowadays, there is abundance of data in the data base or data warehouses. At present, increment in data is going on very rapidly. It might be petabytes, terabytes, or exabytes of data consisting of billion to trillions of records of million entities gathered from different sources. For such massive data, the manual data analysis and data retrieval system are very time consuming.[1] This huge amount of data generates necessity of data analysis tool since data may be located at different locations. A competent and robust tool can be built using data mining with its techniques. Data mining is an exciting field of computer science. Scientists and researchers find it as a new field for their research.

Data mining is the combination of techniques which are further based on concept of learning. Acquiring knowledge is a big aspect of data mining. Further, it can be stated as the data mining is the part of knowledge discovery. Data mining is the extraction of information from a large data storage which includes many extraction techniques. It can handle a large amount of information. Data mining, that extracts unrevealed predictive knowledge from huge datasets, is a strong innovative technology with influential strength to help organizations for managing the most crucial information in their data bases. The data mining tools help to find out the future trends and patterns, empowering businesses to conclude knowledge generated decisions.[2] The automated potential analysis provided by data mining is far ahead of the analyses of data provided by retroactive tools mainly of decision support systems. The tools used for data mining techniques can resolve conventionally and extremely time-consuming issues related to business in efficient way. The tools examine databases for hidden patterns, detecting predictive information that one may miss easily otherwise.

It is the process of semi-automatically analyzing huge datasets to detect patterns which are: true, useful, fresh, feasible, and easily assimilated.

Everyday life of a person goes through an innumerable kinds of pattern recognition (PR) problems such as images, faces, smells, voices, situations, and many more.[3] Initially, most of these issues are corrected at a sensory level or instinctively, without any help of an explicit technique or algorithm. The moment we get an algorithm the problem or issue becomes trivial and is then handed over to the computer. Now,

Address for correspondence:
Deepti Mishra
E-mail: itsdeepti.s@gmail.com

machines have regularly replaced humans in many previously difficult or impossible way, except only in few unvarying PR tasks, namely, medical test reading, fingerprint recognition, mail sorting, military target recognition, signature verification, DNA matching, meteorological forecasting, and others. PR is the scientific discipline whose aim is to classify data, objects, and patterns into different categories or classes. It utilizes abilities of – unsupervised learning – supervised learning.

The basic ideology is to use data mining techniques to find pertinent and useful patterns of system virtues that elucidate program and user nature, with utilization of set of relevant system features to summarize (learning by induction) classifiers which detect outliers. Nowadays, outlier and its detection are an emerging area of data mining. There is no accurate method to identify outlier.

At present in data mining, outlier detection is the widely research field which entails excessive attention. Uncovering outliers carried out in generated patterns is a pertinent hindrance in the data mining. Outlier mining is the technique of spotting rare events, deviant data points, and exceptional objects.[4] Outlier analysis is a major data mining topic in knowledge discovery. Data mining and knowledge discovery can be used interchangeably. Data mining in database systems points to self-extraction of useful predictive information that is not conspicuous otherwise.[5,6]

The focus of the paper goes around the concept of data mining, outlier detection, and its different approaches.

The work presents the literature survey of outliers in various kinds of data set. The work is designed to demonstrate the study on data mining and various methodologies for outlier detection. We categorized and presented a comprehensive overview of approaches of outlier detection in various taxonomy.

Rest of the paper presents literature survey of outliers with approaches applied for detection additionally with comparison and analysis.

CUMULATIVE KNOWLEDGE FACTS FOR OUTLIERS

This section provides preliminary knowledge required for outlier identification and detection. Before proceeding to outlier, let have a look on key areas for awareness of outliers.

Data warehouse

Data warehouse is the collection of enormous heterogenous data gathered from multiple sources for further prediction and analysis. Key features of data warehouse are that it is subject-oriented, integrated, time variant, and durability.

Data warehouse facilitates the process named extract, transform, and load, that is, ETL which is a key for analyzing and discovering business perceptions. Extraction is the procedure to read and gather data from different data sources in a one database using a tool. The same tool transforms the data into usable and required form which is followed by writing the data into target database called load. The data are pre-processed and cleaned during the procedure. The transferring and loading process is applied with the help of data marts. It can be stated as data mart as subset of data warehouse which concentrates on specific domain of business.[7] OLAP operations are applied for data processing.

It processes the data, incorporate in business point of view such as customers' information, sale details, products information, and review results.

Data warehouse covers numerous benefits such as easy retrieval of information, consistency in database, adaptable to change, modifications in timely manner, and many more.[8] It provides a foundation for good decision-making system.[9]

Data mining

Data mining includes various techniques to facilitates the process of generating beneficial information for prediction and analysis from huge pre-existing database. Data mining comprises different steps such as data pre-processing, learning, and analysis. Data collected from different sources are cleaned, transformed, and reduced to construct a product which is better for the use. As during collection of enormous data, there are chances of loss of information such as geological data, numerical data, etc.[10]

It is composed of three techniques classification, clustering, and association rule mining. The classification followed the concept of supervised learning, while clustering is applied on the concept of unsupervised learning. Decision trees, regression, Bayesian classification, and neural network are few methodologies and techniques of

classification. K-means, DBSCAN, OPTICS, and Chameleon are few techniques of clustering.

In the current times, when data generation is occurring at an exponential rate, it is equally important to treat the data collected to harvest out the useful data from the not so useful data.

Such mined out, data are extremely useful for innumerable purposes, namely, fraud detection, weather forecasting, market trend predictions, customer purchasing trend identification, and many more.

Knowledge discovery

Knowledge discovery is acquiring and achieving the results of prediction from huge data by applying data mining techniques. It is the extraction of knowledge from structured or unstructured data. Data mining and knowledge discovery are directly or indirectly related to each other.

Fundamentally, the resulted patterns generated after applying data mining techniques to find some conclusions are termed as knowledge discovery.

Initially, the objective of KDD procedure is to be identified according to users' point of view. Afterward, the selection of data is accomplished to achieve the goal, while selecting the data set user needs to ponder in right direction with exact methodology of data mining.

OUTLIERS

Outliers are the data points which can be notably differentiated from the residue of the statistics of data points. It is assumed practically that the number of "normal" observations are considerably more than "abnormal" observations (outliers/ anomalies) in the given data.

There are numerous definitions of outliers suggested by many authors.

Large-scaled gathered data needs to be processed and managed properly for efficient analysis. The analysis process requires software tools based on techniques of data mining. During the process of examining of data, it is mandatory to remove unwanted data for precise outcome. That unwanted data may be considered as outliers.

Outliers can be flowed into the data set may be by malicious activity, due to outcomes of faulty experiments and merely by erroneous entries in the data.[22]

There are many reasons to handle outliers, because they have significant impact on the result. Some may hide substantial and useful information for further analysis.

Detecting outliers play a key role in retrieving precise information and have various applications such as intrusion detection, credit card fraud detection, proofing medical diagnosis data, and analysis of satellite images. It has[23] observed as outliers often share the same mathematical features, so it can be difficult to distinguish between them. If there is a lack of relevant domain information that leads to the idea of deciding why they are outside and different and what the data model is doing below it means that it is difficult to identify them as outsiders. To find out, if an outlier is revealing or important other information is needed, for example, an accurate mathematical calculation and appropriate background.[24] External detection algorithms are intended to automatically detect those objects or disturbing visions in large amounts of data. Since there is no standard definition of an outlier, so the whole algorithm is based on a model based on a certain guess of what qualifies as a trader.

Outliers carry few characteristics so that they can be distinguished. Outliers can be identified and separated based on their characteristics such as existence in high-density or low-density region, identified separately or with other data points. Some key feature of outliers to distinguish them are mentioned below.[25]

1. Based on neighborhood – They can be categorized as global and local outliers. They can be stated as global having different characteristics from whole data set, and local if they can be identified only from their neighborhood.
2. Approach of degree for knowing outlier – A SCALAR (binary) value is used to detect whether the data object can be considered as an outlier or not. The degree for considering data point as outliers implies that to generate the score of the data point for declaring it as an outlier. In contrast, OUTLIERNESS signifies the visualization of characteristics and degree, on which the point can be considered as an outlier in respect with other data.
3. Approach of dimensions for detecting outliers – UNIVARATE data can be categorized as, with one feature that can be categorized

externally only on the basis that one feature is unique in relation to that other data. In contrast, MULTIVARIATE data contains multiple attributes and can be considered as for identifying outliers as the combination of few factors have unusual data values.

4. Total count of outliers – The number of outliers can be obtained by different techniques in different ways per second. A few tricks identify one outlier and remove it and repeat the process until more exits are found. Some strategies find a collection of outliers in a single effort. Both have their advantages and disadvantages.

The necessity and requirement to distinguish the outliers according to their characteristics because they can have significant impact on the result. Sometimes, they may have valuable knowledge and vital information. They may have valuable knowledge and vital information.

OUTLIER DETECTION APPROACHES AND TECHNIQUES

Outlier detection techniques and approaches have been comprehensively analyzed in the previous decades and many approaches have been developed to identify outliers [Table 1]. There are few basic approaches for outlier detection – statistical approach, depth-based approach, distance-based approach, density-based approach, and deviation-based approach. Other approaches are – for spatial data, graph-based approach. Approaches are also developed for high-dimensional data.[14,26] It can be defined that approaches to identify outliers are based on supervised learning and unsupervised learning.[27] Some techniques can be applied only on univariate data or some technique are beneficial for multivariate data for outlier detection. Before proceeding further, it is better to remove outliers.

Statistical approach

Statistical algorithms are the foundation and basic algorithms applied for outlier detection, which were based on probability or distribution model for the given data set. The standard deviation, interquartile range, Chi-square test, ANNOVA, and Z-score, these few are methods of statistics to identify outliers. Some of the techniques of statistics assume Gaussian distribution of data. These kinds of statistical methods can be applied to notify rare and unexpected data in bivariate or multivariate dataset. The traditional way is to apply discordancy method by assuming the hypothesis which verifies that whether the data point as outlier or not. A traditional way is to plot the data to identify the odd data point from residue data set. The data points that does not follow the model stated as the outliers. Barnett and Lewis[12] and Rousseuw and Leroy[28] described some techniques which are single dimensional. As the dimension increases, it is very tough to manage the model for data set.

Depth-based approach

In depth-based approach, the data set is organized into different layers. The outer layer is more prone to outliers in comparison to inner layers. Minimum volume ellipsoid is based on depth-based technique. The probability of observation is computed and visualized to notify the outlier.

Distance-based approach

The distance-based approach as name suggested requires distance calculations for outlier detection such as Euclidean distance. Knorr and Ng proposed the outlier detection technique using k nearest neighbors.[29,30]

Some authors also use a distance-based approach to detect external objects such as nested loop, cell-based, index-based methods that reduce unnecessary computations. A new hybrid approach was introduced using a distance-based method and a k-way to identify outliers.[15] Initially, this method uses a different integration algorithm which means K-means dividing the data set into several clusters and then using distance-based methods to find outliers. The scope of the approach necessitates alterations so that it can be made applicable to the textual mining. It can be applied on more complicated data set. Also, can be applied on varying data sets.

The paper presents the new algorithm PLDOF and tried to proof that it is better than LDOF. K-means clustering algorithm is applied to divide the data set into clusters. It is based on the concept that first calculates the distance of each point from the centroid of the cluster.[13]

Table 1: Various definitions of outliers

S. No.	Authors	Definitions
1.	Hawkins[11]	Outlier is a monitored data point, which differ so much from rest data set as to create a suspicious environment that it was generated by different mechanism.
2.	Barnett and Lewis[12]	A data object (or subgroup of observation) which emerges and acts to be irregular or volatile with the remainder of that data set.
3.	Pamula *et al*.[13]	Outlier is the object or point which does not adapt the same characteristics alike to the usual data depicting the data set.
4.	Guo *et al*.[14]	Outlier can act abnormally but may encompass valued information.
5.	Pachgade and Dhande[15]	Outlier can act as a type of pattern such as behaving differently in aspect of residue data in the data set.
6.	Agarwal and Yu[16]	The data point is considered as outlier if it is not identical with residue data on some characteristics.
7.	Li *et al*.[17]	Outlier is a small quantity of data objects with abnormal behavior in data set.
8.	Pham and Phag[18]	Outliers are the items that noticeably differ and cannot fit in the general distribution of the data.
9.	Behera *et al*.[19]	The irregular data points pointed as outliers from residue data if it appears that they are engendered by any faulty circumstances from experiments.
10.	Li and Kitagawa[20]	An outlier is an abnormal data point which is greatly different from the rest of the data.
11.	Breuning *et al*.[21]	Outliers are those data points which can be categorized by density as they lie in the less density region in compared to its neighboring points.

Density-based approach

The density-based approach evaluates the density in the data region to detect outliers as outliers generally lie in low-density region.

Certain methodologies are proposed and implemented concerning with density-based approach. One of the algorithms named outlier finding technique is proposed, which further uses K-means clustering algorithm to cluster the data set and to find out outliers. The functioning of methodology is based on the findings of density and distance values.[19]

Deviation-based approach

Arning *et al*. defined a deviation-based tactic, functions as observing key qualities to detect the outliers. Those object that departs from the qualities are declared as outliers.[31]

Another approach proposed in the paper a definition of outlier that is class outlier and proposed a class outlier factor (COF). It generates as a ranking score named COF to measure a degree of being a class outlier for a data point. The main factors of computing COF are the probability of the instance classes, the deviation of instance from an instance of a same class, and the distance between the instance and its K neighbors.[4]

Outlier approaches for Spatial data set

Spatial data set includes geographic data or in other words geographic information. It can be defined as the data which can be mapped and related to space, for example, building, lake, mountains, and township. Cai *et al*. proposed an iterative self-organizing map with robust distance estimation for spatial outlier detection.[32] The neighboring clusters Kou *et al*. proposed spatial-weighted outlier detection algorithms for spatial data set.[33]

Outliers can be present in the meteorological data and effect the valuable information.[34] Author presented a new approach based on wavelet analysis. Wavelet analysis is applied to study images and signal processing. The signals and images are observed at various focuses to generate conclusion and distinct patterns.

Graph-based approach

The paper introduced a new algorithm SWHOT which is based on weighted hypergraph model. It is the union of BSWH algorithm and CURE algorithm. The algorithm provides the concept of the feature vector and attribute similarity.[17]

High-dimensional data-based approach

Nowadays, high-dimensional data are the big challenge for outlier analysis. It is difficult to decoct out the irrelevant data from high-dimensional data and concurrent segregation of outliers.

Tests are performed on high-quality data set by an outlier detection algorithm. The new angle-based outlier discovery (ABOD) algorithm is proposed and compared to other algorithms. ABOD computes that the scalar product between the vectors of the data points, furthermore, computes

the variance over the angles which are calculated over the data points. There is not any requirement for the selection of parameters for generating outcomes.[35]

Another approach for outlier detection is introduced by Pham and Phag which functions in high-dimensional dataset. The approach is efficient and scalable to extremely high-dimensional data set.[18]

There is another approach named outlier detection in high-dimension based on projection. It discovers the outliers by applying the concept of projection, from the data set. In the methodology, clusters are designed by applying the concept of projection and weight. Smaller clusters are pruned further detecting outliers.[14]

The author presents a paper which describes the concept outliers in high-dimensional data set.[16] Outliers and its techniques are discussed in detail in the paper.

Amit Banerjee suggested a new algorithm for outlier detection which is based on evolutionary search. The genetic algorithm is used to find external objects that are outliers. The method used is to integrate the Euclidean-based approach to distance, partly based on congestion based on the converted value of Lancaster.[36] There is an advantage of the given approach which can also be applied on continues attributes. The scope of algorithm, to generate a procedure which can detect absolute count of outliers by applying fitness function.

For high-dimensional data, the analysis becomes more critical. A new method COMPREX is introduced which apply pattern-based compression. The approach is parameter free, general, scalable, and efficient. The categorical database is used.[37]

Approaches based on classification

Classification is based on concept of supervised learning further part of data mining and machine learning. In classification technique, there are predefined trained models. In supervised learning, the class description of every training sample is provided. It is used to classify the dataset into predefined groups of classes or models. The classification includes many techniques such as decision trees, neural network, and statistics. In classification, there are predefined groups of models or classes to classify the new data. It is extremely useful process for knowledge discovery by identifying the classes and models for new data. Further continuing, the decision tree approach of classification generally uses methodologies of statistical techniques of or in other words concept of clustering to identify the outliers or classification of data precisely. Due to presence of outliers', branches of tree reflect variances in the classes.

Approaches based on clustering

Clustering is based on concept of unsupervised learning to identify the data in clusters or in groups. The data with similar traits lie in same cluster, but they have must have different traits with data objects which lie in another cluster. The clusters identified by any clustering algorithm can be succumbed to outliers. Due to presence of outliers, there may be variation in outcome and result may be inappropriate. Techniques of clustering can identify the outliers such as distance-based methodologies, for example, K-means and density-based methodologies, for example, CURE and DBSCAN. The identification of outlier depends on the distance value between data point and centroid data point which may in turn affect the accuracy of clusters. In that aspect, the density-based approach identifies outliers on the concept of high- and low-density regions. Hence, it can be stated as density-based approaches are better approaches than distance-based approaches to identify outliers.

A new algorithm Balanced Iterative Reducing and Clustering using Hierarchies is applied to generate clusters.[38] Pattern identification is an important task from the large data bases. This methodology works on large data sets to identify patterns. It can also deal with noise. The paper defines the two types of attribute – metric and non-metric. It works in four stages – load database in memory after building CF tree, reduce in smaller CF tree, apply global clustering, and refine the clusters.[38]

The authors presents new approach for clustering applying in huge databases.[17,39] The key objective of the work is to apply clustering in two stages: A basic and fast step that divide data into overlapping divisions named as "canopies," and then distance measure is done only for those data objects that present in common canopy.

The canopies are the subset of the of the data points. The generated clustering which is based on canopies can be suitable to many existing clustering algorithms, for example, K-means, greedy agglomerative clustering, and expectation maximization. In the paper, the canopies are applied with agglomerative clustering.[39]

Authors proposed a new algorithm for outlier detection.[40] The approach is defined in different stages. In initial stage, a modified K-means approach is applied followed by constructing minimum spanning tree. Further, the longest edge is removed connecting with farthest data points considering it as outliers.

OLAP data cube techniques

OLAP data cube techniques are applied to identify unexpected data points in multidimensional data model. It is consisting of numerical facts and data which are termed as dimensions. It is used for analysis and computations on data. Data are gathered from various data collections in data cube. Basic operations on OLAP include drill down, roll up, slice, dice, and pivot. Each data cube is comprised data cells. Each data cell contains numeric data which are further applied for statistical, mathematical, aggregate functions, and many more. If data in any cell deviate too much from data in another cells, then that cell value is spotted as exception or outliers. Data cube is better scope to store data as it offers worthy visualization such as data in cell can be viewed in distinct colors to signify outliers and exceptions.

COMPARATIVE STUDY AND ANALYSIS OF VARIOUS TECHNIQUES OF OUTLIER DETECTION

From the comparison shown in the Table 2, it can be analyzed that different techniques are required to manage different kind of databases to detect outliers. Depth-based technique can identify multiple outliers at a time, since it organizes the data in different layers. It is assumed that maximum number of outliers lie in the outer layer. Usually, statistical- and distance-based techniques work more efficiently on low-dimensional data set, while techniques such as graph based, density based, clustering based can be applied on high-dimensional data set.

Table 2: Comparison of different approaches of outlier detection

Comparison Technique	Approach			Types of data set		Types of outliers		Number of outliers	
	Supervised	Unsupervised	Semi supervised	Single	High-Dimensional	Hybrid	Global	Local	Single	Multiple
Statistical		✓		✓				✓	✓	
Depth Based		✓			✓		✓			✓
Distance Based		✓		✓				✓		✓
Density Based		✓			✓		✓			✓
Deviation Based	✓		✓	✓			✓			✓
Graph		✓			✓		✓			✓
Classification based	✓				✓	✓		✓	✓	
Clustering Based		✓		✓	✓	✓	✓	✓		✓

Table 3: Comparison of different tools on various platforms

	WEKA	R Studio	RapidMiner	Orange
Written in	JAVA	C++	JAVA XML	Python, C++, Cython
Rapidity	FAST	FAST	FAST	FAST
Open source	YES	YES	YES	YES
Large data	Efficiently on small data sets less than R	Efficiently on large and small data sets	Efficient less than R	Efficiently on large and small data sets
Visualization	Less options	Many options	Many options	Less options
Outliers	YES	YES	YES	YES

Maximum number of techniques are based on unsupervised learning such as depth based and density based, while some techniques such as classification and deviation based are lie on the concept of supervised learning. Since these later techniques needs predefined characteristics to identify odd one data points.

Classification-based techniques are based on supervised learning as they apply operations on predefined model. On the contrary, clustering-based approaches are based on unsupervised learning, because there is no trained exemplar to apply mathematical operations.

In general, graph-based approaches can function efficiently on high-dimensional data set. It is the part of unsupervised learning. It generates graph using the distance function and finding neighboring points.

OUTLIER DETECTION TOOLS

At present, numerous software in existence to cope with the identifications and complications generated by outliers such as WEKA, R Studio, RapidMiner, and Orange.

This free software provides strong statistical environment for analyzing data mining techniques which are further capable of identifying outliers.[41] In general, the software is comprising various machine learning techniques, offer vivid visualization and excellent platform for data science. These tools are foundation for data analyzing, computation, and its visualization. They offer an integrated environment for data preparation, data mining, and generating knowledge.

Outliers can be detected using pre-processing techniques in WEKA tool. As it is a user-friendly software, clicking on some buttons such as filter can spot outliers effortlessly. In R studio, user must install packages which initially take few seconds to install to execute the functions to detect outliers. One of the package outliers apply the statistical techniques such as mean values and standard deviation to spot outliers. The function detect outliers are capable enough to identify outliers easily in RapidMiner. Below a comparative study is mentioned in the Table 3. Four different software are compared on various platforms.

CONCLUSION AND FUTURE WORK

Considering the never-ending rise in data collection and concomitant need of processed data for scientific and commercial usage, it is needless to reemphasize about the need of development of data mining tools for the same. The development of new algorithms which possess the capability of detecting outliers by more precision is deeply needed. Furthermore, processed data need to be presented in such a diligent pattern to manifest the secret patterns for public usage.

Ongoing research efforts in artificial intelligence needs to be either combined or utilized at specific points to make out knowledge discoveries from data mining more productive, scientific, and sensitive.

The identification of outliers is a subtopic of data mining. Outlier analysis is a highly research field for scientists. Outliers are the data points that those cannot be fitted in any type of clusters. These objects are somehow objectionably unusual from residue data set. They may differ from entire data sets or may be difficult to locate. The presence of outliers makes the results confusing. Patterns generated after data from data are incorrect and inaccurate for outliers. Every kind of data includes a new problem of outliers and requires the different techniques to handle it. A single straight approach cannot deal with all kind of outliers in various fields and subject areas. Hence, we

need different approaches to identify and deduct outliers. The paper presents the review of outlier and outliers detection techniques. It is valuable for the researchers. Outliers may cover any vital information or may influence the accuracy of result. The idea is, how to notify the outliers, as to retain them for further information or remove them to gain accurate outcome.

REFERENCES

1. Houari ME, Rhanoui M, Asri BE. Hybrid Big Data Warehouse for on-demand Decision Needs. Rabat, Morocco: International Conference on Electrical and Information Technologies (ICEIT); 2017.
2. Ming J, Zhang L, Sun J, Zhang Y. Analysis Models of Technical and Economic Data of Mining Enterprises Based on Big Data Analysis. Chengdu, China: 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA); 2018.
3. Heling J, Yang A, Yan F, Miao H. Research on pattern analysis and data classification methodology for data mining and knowledge discovery. Int J Hybrid Inf Technol 2016;9:179-88.
4. Hewahi NM, Saad MK. Class outliers mining: Distance-based approach. Int J Comput Infor Eng 2007;1:2805-18.
5. Koteeswaran S, Visu P, Janet J. A review on clustering and outlier analysis techniques in data mining. Am J Appl Sci 2012;9:254-8.
6. Velchamy I, Subramanian R, Vasudevan V. A five step procedure for outlier analysis in data mining. Eur J Sci Res 2012;75:327-39.
7. Prakash D, Prakash N. A Requirements Driven Approach to Data Warehouse Consolidation. Brighton, UK: 11th International Conference on Research Challenges in Information Science (RCIS); 2017.
8. Bagambiki E. Enterprise Data Warehouse and Business Intelligence Solution. In: ICEGOV'18 Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance; 2018.
9. Farooqui NA, Mehra R. Design of a Data Warehouse for Medical Information System Using Data Mining Techniques. Solan Himachal Pradesh, India: 2018 5th International Conference on Parallel, Distributed and Grid Computing (PDGC); 2019.
10. Lgkklw EA, Lingling Z. A BP neural network based method for geological missing data processing. Gold Sci Technol 2015;23:53-8.
11. Hawkins DM. Identification of Outliers. Berlin, Germany: Spinger; 1980.
12. Barnett V, Lewis T. Outliers in Statistical Data. Hoboken, New Jersey: Wiley; 1994.
13. Pamula R, Deka JK, Nandi S. An Outlier Detection Method Based on Clustering. Kolkata, India: Second International Conference on Emerging Applications of Information Technology; 2011.
14. Guo P, Dai JY, Wang YX. Outlier Detection in High Dimension Based on Projection. Dalian, China: International Conference on Machine Learning and Cybernetics; 2006.
15. Pachgade SD, Dhande SS. Outlier detection over data set using cluster-based and distance-based approach. Int J Adv Res Comput Sci Softw Eng 2012;2:12-6.
16. Agarwal CC, Yu PS. Outlier Detection for High Dimensional Data. Santa Barbara, California, USA: ACM SIGMOD International Conference on Management of Data; 2001.
17. Li Y, Wu D, Ren J, Hu C. An Improved Outlier Detection Method in High-dimension Based on Weighted Hypergraph. In: Second International Symposium on Electronic Commerce and Security; 2009.
18. Pham N, Phag R. A near-linear Time Approximation Algorithm for Angle Based Outlier Detection in High Dimensional Data. Beijing, China: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2012.
19. Behera HS, Ghosh A, Mishra SK. A new hybridized k-means clustering based outlier detection technique for effective data mining. Int J Adv Res Comput Sci Softw Eng 2012;2:287-92.
20. Li Y, Kitagawa H. DB-outlier Detection by Example in High Dimensional Datasets. Istanbul, Turkey: IEEE International Workshop on Databases for Next Generation Researchers; 2007.
21. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF-identifying Density Based Local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data: Dallas, Texas, USA; 2000.
22. Hodge VJ, Austin J. A survey of outlier detection methodologies. Artif Intell Rev 2004;22:85-126.
23. Cheng JG. Outlier Management in Intelligent Data Analysis. London: University of London; 2000.
24. Zimek A, Campello R, Sander J. Ensembles for Unsupervised Outlier Detection: Challenges and Research Questions. SIGKDD Explorations Newsletter; 2013. p. 11-23.
25. Zhang Y, Meratnia N, Havinga P. A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets. CTIT Technical Report Series; No. Paper P-NS/TR-CTIT-07-79). Enschede: Centre for Telematics and Information Technology (CTIT), 2007.
26. Xi J. Outlier Detection Algorithms in Data Mining. In: 2nd International Symposium on Intelligent Information Technology Application: Shanghai, China; 2008.
27. Zhu C, Kitagawa H, Papadimitriou S, Faloutsos C. Outlier detection by example. J Intell Inf Syst 2011;36:217-47.
28. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. In: Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley; 1987.
29. Knor EM, Ng RT. Finding Intensional Knowledge of Distance Based Outliers. Edinburgh, Scotland: Proceedings of the 25th VLDB Conference; 1999.
30. Knorr EM, Ng RT, Tucakov V. Distance based outlier: Algorithm and applications. VLDB J Int J 2000;8:237-53.
31. Arning A, Agrawal R, Raghavan P. A linear Method

for Deviation Detection in Large Databases. Portland, Oregon: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining; 1996.
32. Cai Q, He H, Man H. Spatial outlier detection based on iterative self organizing learning model. Neurocomputing 2013;117:161-72.
33. Kou Y, Lu CT, Chen D. Spatial Weighted Outlier Detection. Bethesda: Proceedings of SIAM Conference on Data Mining April 20-22; 2006.
34. Zhao J, Lu CT, Kou Y. Detecting Region Outliers in Meteorological Data. In: GIS 03 Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems: New Orleans, Louisiana, USA; 2003.
35. Kriegel HP, Schubert M, Zimek A. Angle Based Outlier Detection in High Dimensional Data. Las Vegas, Nevada, USA: KDD '08 Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2008.
36. Banerjee A. Density-based Evolutionary Outlier Detection. Philadelphia, PA, USA: GECCO'12; 2012.
37. Akoglu L, Tong H, Vreeken J, Faloutsos C. Fast and Reliable Anomaly Detection in Categorical Data. Maui, HI, USA: CIKM'12; 2012.
38. Zhang T, Ramakrishnan R, Livny M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. Montreal, Quebec, Canada: SIGMOD '96 Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data; 1996.
39. McCallum A, Nigam K, Ungar LH. Efficient Clustering of High-dimensional Data Sets with Application to Reference Matching. Boston, Massachusetts, USA: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2000.
40. Jiang MF, Tseng SS, Su CM. Two-phase Clustering Process for Outliers Detection. Amsterdam, Netherlands: Pattern Recognition Letters, Elsevier; 2001. p. 691-700.
41. Al-Khoder A, Harmouch H. Evaluating four of the most popular open source and free data mining tools. Int J Acad Sci Res 2015;3:13-23