# Research and Analysis of Learning Factors Based on the Foundation of Computer Big Data

Yiwei Ma(✉)

University of Glasgow, Glasgow, UK
mayiweiuk@163.com

**Abstract.** In the context of big data from computers, this paper explores Chinese students' low engagement behaviors in the UK classroom, supported by a foundation of data mining. In detail, it examines some of the current pedagogical reasons taught by Chinese graduate students about silence at the University of Glasgow through the theory of planned behavior. This exploratory study aims to gain a clearer understanding of what causes Chinese students to be silent in UK classrooms. The study also aims to provide innovative strategies for postgraduate lecturers at the University of Glasgow to deal with the silence of Chinese students, even as others teaching Chinese students in the UK encounter the same situation. In this case, it could give more pointed support to better meet the needs of Chinese students. This issue is important because Chinese students' silence in the classroom is a common phenomenon that has caused some problems for UK educators. This paper applies computerized big data to research that explores factors of student learning, also lays some groundwork for new areas of research with the development of new ways of research.

**Keywords:** Computer models · Data mining · Higher education research · Data dimensional features

## 1 Introduction

With the development and evolution of computer network technology in modern society, the data analysis capability of computer models has been applied to various fields. This study investigates and analyzes computer-integrated data as a factor of learning in higher education, which is not only an innovative point in this research field, but also a fundamental contribution to subsequent studies [1]. The main stakeholders of the study include lecturers in UK higher education and Chinese students in UK universities, including the study participants and the institutions they belong [2]. A potential outcome of the study is an overall idea for UK educators that a theoretical framework could help UK university teachers and Chinese students to better manage classroom time in the UK, especially in universities with large numbers of Chinese students.


C. F. Peng et al. (Eds.): EIMT 2023, AHSSEH 8, pp. 645–650, 2023.
https://doi.org/10.2991/978-94-6463-192-0_84

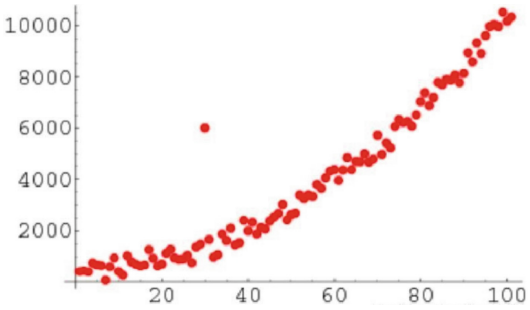


**Fig. 1.** Outlier

## 2 Anomaly Detection and Solutions for Data Mining

### 2.1 Definition of Anomalies

As shown in Fig. 1, the outlier point is the point that is not part of the vast majority of the data growth trend and has a different observation in the figure.

## 3 Basic theory of Anomaly Detection Based on Data Mining

### 3.1 Key Issues in Anomaly Detection

Anomalies in an abstract sense, as patterns of abnormal behavior that do not conform to expectations [5]. Therefore, anomaly detection methods need to first define rules or regions that represent normal behavior, and data that do not conform to the rules and regions are considered as anomalous data, as shown in Table 1.

### 3.2 Distance-based KNN Algorithm

The k-Nearest Neighbor calculates the mean, median, and maximum of the distance between each sample point and its nearest K samples, and the KNN algorithm is described in more detail by J. Zou (2017) [6], and either one is chosen as the anomaly score, and since the outliers are different from the vast majority of data points, that will have a greater distance than similar data points in terms of distance, and the distance is compared to the threshold is compared, and if it is greater than the threshold, it is considered an outlier [7]. Alternatively, the K-average distance of all samples is taken and the first n largest ones are considered as outliers. The Euclidean distance is generally used to calculate the distance, and the angular distance can also be used.

**(1) Algorithm example**

As shown in the figure below, the selected GDP data of each province in 2018 can be seen to have an obvious growth trend [8]. When using the KNN distance algorithm, the k-neighbor distance should increase when the data is larger, and it is easier to be judged as an outlier; while when the data points are small, even if the data points change significantly from the usual data, the data is relatively tight and the k-neighbor distance is small, which makes it not easy to be judged as an outlier.

**Table 1.** Advantages and disadvantages of the algorithm

| Methods | Advantages | Disadvantages |
|---|---|---|
| Based on distance | In the short training time and simple thinking | Only for spherical clusters of samples, the calculation is complicated |
| Based on density | Does not judge the situation due to data density dispersion | There is a density difference in the detected data |
| Tree-based | There is no need to calculate metrics about distance, density | A higher proportion of abnormal samples is less effective |
| Based on clustering | Very fast detection efficiency | Classification results depend on the initialization of the classification center |
| Statistics | The assumptions of the data distribution are valid for easy operation | Less effective with higher data dimensionality |

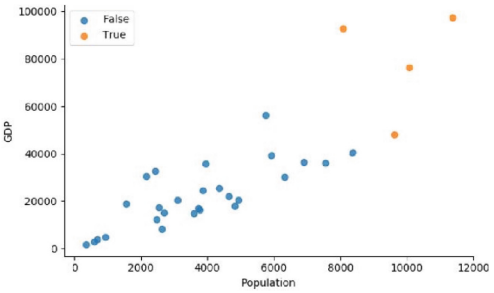


**Fig. 2.** KNN anomaly detection example

### 3.3 Local Density LOF-based Algorithm

**1) Abnormal score**

In the LOF algorithm, the outlier factor is the key to detect whether the sample data points are anomalous or not. Starting from the basic principle of the algorithm, this paper involves the calculation of four data, the distance between sample points $d(p, o)$, the reachable distance of local sample points, the third is the local reachable density $\rho_k(P)$, and finally the local outlier factor $LOF_k(P)$, so their detailed calculation is as follows.

(1) Calculate the kth distance: $d_k(P)$. The kth distance of the point O satisfies the following conditions.

a. There are at least $k$ points $o^{'} \in C\{x \neq p\}$ in the set that do not include p and satisfy $d(p, o) \leq d(p, o^{'})$.

b. There are at most $k-1$ points $o^{'} \in C\{x \neq p\}$ in the set that do not include p and satisfy $d(p, o) < d(p, o^{'})$, in short of the point $p$ is the $k$ th nearest point to O.

(2) Calculate the $k$ th distance neighborhood: Calculate the number of $k$ th distance neighborhoods.

(3) Calculation of reachable distance

$$reach - distancek(p, o) = max\{k - distance(o), d(p, o)\} \tag{1}$$

If it is less than the $k$th distance, then the reachable distance is the $k$ th distance, and if it is greater than the $k$ th distance, then the reachable distance is the true distance.

(4) Calculation of local achievable density

$$\rho_k(P) = \frac{1}{\sum_{0 \in N_k(P)} d(p, o)/|N_k(P)|} \tag{2}$$

The inverse of the average reachable distance from the $k$ th neighborhood point to the point P.

(5) Calculation of local outlier factors

$$LOF_k(P) = \frac{\sum_{0 \in N_k(P)} \frac{\rho_k(0)}{\rho_k(P)}}{|N_k(P)|} \tag{3}$$

At this point there are two scenarios.

a. If p and the surrounding neighborhood points are close, the reachable distance may be the smaller $k$ th distance $d_k(o)$, $LOF_k(P)$ tends to be close to 1, indicating that the point is about the same density as the neighborhood points, and the higher the density value, the more likely it is the same cluster.

b. If p and the surrounding neighborhood points are far away, then the reachable distance will take a larger actual distance $d(p, o)$, the local reachable density $\rho_k(P)$ value becomes smaller, $LOF_k(P)$ greater than 1, indicating that the point is less than the density with the neighborhood points, the density value is lower and may belong to the anomaly.

**2) Algorithm example**

LOF algorithm is different from the perspective of measuring the similarity of data points to determine the anomaly, and the second is built on the data point density comparison an outlier detection algorithm with high accuracy.

As shown above, in density-based outlier detection $C_1$ and $C_2$ are two types of clusters with different densities. If based on the distance algorithm, the distribution of sample points within the $C_2$ cluster is sparse and the outlier score will be greater than that of sample points within the $C_1$ cluster, however, the actual situation may be that the sample points within the $C_2$ cluster have the same density and are normal data points; similarly $O_2$ has a lower outlier score, however, the actual situation is that $O_2$ is relatively sparse for other points within the $C_2$ cluster and has a greater possibility of being an outlier point.

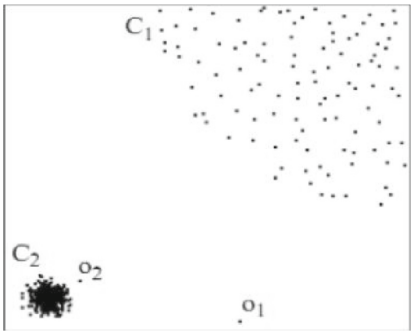


**Fig. 3.** LOF anomaly detection example

## 4 Conclusion

This paper is divided into two main parts. The first part introduces the concept, sources, and causes of outliers. Also based on the computer processing method of big data statistics, the data model is applied to the report of investigating the current status of student learning, which is a good pavement and foundation for further data acquisition as well as data processing and analysis. It makes a certain basic reference and basic experiment for future statistical research on data. The research results obtained in this paper after doing certain data statistics are as follows: (1) We can first define the distribution of the data and the outliers, and select the model according to the assumptions if we have a general understanding of the data distribution, and select the algorithm that targets and screens the outliers. (2) After combining example analysis for anomaly detection, further digging important information behind the anomaly points, and analyzing in the category of data anomaly causes. If it conforms to the actual information then it is considered as a good data point generated by new mechanism and new things, if it does not conform to the actual information behind the data, then it is considered as a data impurity point, and the data is removed and re-corrected.

## References

1. Qi Zhijiang. Computer intrusion detection data mining model design and system experimental validation[J]. Science and Technology Innovation,2022(07):69-72.
2. Jin Xianhao. Application of MapReduce model in parallel computer data mining [J]. Journal of Jingdezhen College,2021,36(06):114-116.
3. Sun, B. Han. Analysis of an empirical research model of computer teaching in colleges and universities based on data mining learning analysis[J]. Modern Electronics Technology,2019,42(19):127–131.DOI:https://doi.org/10.16652/j.issn.1004-373x.2019.19.030.
4. Xie Jingwei,Cheng Huaan. Construction of computer audit model based on data warehouse and data mining technology[J]. Journal of Hunan Mass Media Vocational Technology College,2016,16(02):82-84.DOI:https://doi.org/10.16261/j.cnki.cn43-1370/z.2016.02.023.
5. Wang, Yuan. Research on computer audit model and its application based on data warehouse and data mining technology[D]. Northeast University of Finance and Economics,2011.

6. Chi Wei, Lai Shihong, Du Guifang, Li Mingzhong. Hotspot Analysis and Future Prospects of Higher Education Research in China: Based on the Data of Higher Education in NPC Reprographic Journal of 2021[J]. Journal of Hebei Science and Technology Teacher's College (Social Science Edition),2022,21(03):109-116.
7. Liu Hui. Beyond the "unfruitful tree": The realistic possibility of higher education research in China: A review of "The history of contemporary Chinese educational scholarship: higher education research"[J]. Journal of Educational Science Exploration,2022,40(05):97.
8. Shen Wenqin. Definition and types of higher education theories [J]. Higher Education Research,2022,43(07):69-88.
9. Yang Pei. Research on the mechanism of visual analysis of high-dimensional data based on information entropy[D]. Lanzhou University of Technology, 2021. DOI:https://doi.org/10.27206/d.cnki.ggsgu.2021.000918.