Retraction

Retraction: Business Data Analysis Based on Hierarchical Clustering Algorithm in the Context of Big Data (J. Phys.: Conf. Ser. 1744 042135)

Published 16 September 2022

This article has been retracted by IOP Publishing following an allegation that raises concerns this article may have been created, manipulated, and/or sold by a commercial entity. In addition, IOP Publishing has seen no evidence that reliable peer review was conducted on this article, despite the clear standards expected of and communicated to conference organisers.

The authors of the article have been given opportunity to present evidence that they were the original and genuine creators of the work, however at the time of publication of this notice, IOP Publishing has not received any response. IOP Publishing has analysed the article and agrees there are enough indicators to cause serious doubts over the legitimacy of the work and agree this article should be retracted. The authors are encouraged to contact IOP Publishing Limited if they have any comments on this retraction.

Retraction published: 16 September 2022
Business Data Analysis Based on Hierarchical Clustering Algorithm in the Context of Big Data

Hao Sun1,*

1Lianyungang Technical College, Educational administration department, Jiangsu, China, 222006

*Corresponding Author Email: shaottx@lygtc.edu.cn

Abstract: With the rapid development of computer technology, computer technology has been applied to all aspects of the daily field, we are now in an era of data explosion. With the passage of time, the data on the network will become more and more, consumers will not be able to meet their consumption needs in the face of so much data. How to deal with these data is an urgent problem to be solved. Based on this, this paper uses a similar analytic hierarchy process algorithm, hierarchical clustering algorithm, to process the data. According to the browsing and consumption records of consumers, the algorithm can recommend the goods that consumers want to buy and the price of goods they want to buy. In the research, this paper analyzes the business data of consumers by hierarchical clustering analysis, and classifies people with different consumption levels into a large group, analyzes the purchasing tendency of consumers and the price of goods they want to buy, and then recommends products to consumers. The experimental results show that the analysis results of this paper using hierarchical clustering algorithm are consistent with the results of selling goods on the online platform. Therefore, according to the experimental results of this paper, businesses can promote more suitable products for consumers.

Keywords: Big data, Hierarchical clustering, Analytic hierarchy process, Cluster analysis

1. Introduction

Big data is massive data and data information [1-3]. Big data turns the real world into a virtual world, making the data of the virtual world react with the real world, and making the real world develop better. It can also be said that big data is the foundation of artificial intelligence [4]. The characteristics of big data are large amount of data, fast reversal, data diversity, low value density and authenticity. How to obtain the information we need from massive amounts of data requires data mining. Data mining is to rely on computer-related technology through data preparation. Look for rules and representation rules to analyze and obtain the data we are interested in. Data mining tasks include...
association analysis, cluster analysis, classification analysis, etc. [5].

The so-called clustering [6, 7] is to put similar things together and divide the dissimilar things into different categories. Clustering is very important in the method of data analysis. In data analysis, clustering and classification are two techniques [8]. Classification refers to things we already know, we only need to learn the classification rules of the sample, and then classify things. Classification is a kind of heuristic learning, and clustering is a kind of unsupervised learning. It can be said that the two processes are opposite. There is shortest distance method, longest distance method, middle distance method, class average method and so on for calculating the distance between classes. Hierarchical clustering is a very intuitive algorithm, it needs to gather layer by layer [9]. Small clusters can be divided from bottom to top for consolidation and clustering, and large clusters can be divided from top to bottom for consolidation and clustering [10].

Based on the above background, in order to mine the data that consumers are interested in from massive data centers, we use a method of classification rules of a hierarchical clustering algorithm. First, the analytic hierarchy process is used to calculate the distance between samples, and each sample is merged into the closest point of the same class. Then calculate the distance between the classes, the nearest class is merged into a large category, and continue to merge until a new category is synthesized. In this way, it is possible to analyze which products consumers are more inclined to buy among various products, analyze consumers' desire for consumption and consumption amount, and then make recommendations on the merchant's product composition and selling prices.

2. Research on Analytic Hierarchy Process and Hierarchical Clustering Algorithm

2.1 Research on Analytic Hierarchy Process

The analytic hierarchy process (AHP) is a kind of decision-making related elements are decomposed into objectives, rules and plans, and on this basis, the decision-making is analyzed qualitatively and quantitatively. The basic principle of analytic hierarchy process (AHP) is to decompose the actual problem into multiple factors, and construct a hierarchical structure according to the selected multiple factors. A weight is determined for each factor, which is used to determine the relative importance of the factor. Then, the weight vector is calculated according to the determined weight. Finally, the comprehensive decision-making is judged and compared to analyze which is the most appropriate answer. AHP is not only suitable for uncertain and subjective information, but also allows experience, insight and intuitive logic application. Its greatest advantage is that it shows the hierarchy itself and allows consumers to carefully consider and measure the relative importance of indicators.

2.2 Research on Hierarchical Clustering Algorithm

Hierarchical clustering is clustering layer by layer. Small clusters can be merged and gathered from bottom to top, or large clusters can be divided from top to bottom. This article uses bottom-up aggregation, that is, first find the two classes with the shortest distance. The Minkowski distance formula is shown in formula (1):

\[ \text{dist}_{mk}(x_i, x_j) = \left( \sum_{\mu=1}^{n} |x_{i\mu} - x_{j\mu}|^p \right)^{\frac{1}{p}} \]  

(1)

Where \( X_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \), \( X_j = (x_{j1}, x_{j2}, \ldots, x_{jn}) \).

When \( p = 2 \), the Minkowski distance is the Euclidean distance:
\[ \text{dist}_{cd}(x_i, x_j) = \left\| x_i - x_j \right\|_2 = \sqrt{\sum_{\mu=1}^{n} (x_{i\mu} - x_{j\mu})^2} \]  

(2)

When \( p = 1 \), Minkowski distance is Manhattan distance:

\[ \text{dist}_{man}(x_i, x_j) = \left\| x_i - x_j \right\|_1 = \sum_{\mu=1}^{n} |x_{i\mu} - x_{j\mu}| \]  

(3)

The hierarchical clustering algorithm is easy to define the distance between classes and the similarity of rules, requires fewer restrictions, does not need to pre-set the number of clusters, and can discover the hierarchical relationship of the classes.

2.3 Hierarchical Clustering Algorithm Process

The hierarchical clustering strategy is to treat each object as a cluster, and then merge the atomic clusters into larger and larger clusters until all the objects are in the same cluster or meet a termination condition, most of which belong to the hierarchical cluster. For hierarchical clustering of classes, the similarity between cluster definitions is different. The algorithm of analytic hierarchy process and hierarchical clustering fusion algorithm is as follows:

1. Weighted evaluation of each object;
2. Take each object as a class and calculate the minimum distance between two objects;
3. Combine the two classes with the smallest distance into a new class;
4. Recalculate the distance between the new class and all classes;
5. Repeat (3) and (4) until all classes are finally merged into one class.

Hierarchical clustering does not have a global objective function similar to the basic K-means, there is no local minimum problem or the initial point is difficult to choose. The merge operation is usually the final operation and will not be undone after the two clusters are merged.

3. Experimental Background and Parameter Settings

The arrival of big data affects all aspects of people’s lives. People continue to browse web pages and videos, and the amount of big data is constantly increasing. With the development of big data, big data has great value in military, politics and entertainment. At the same time, the development of big data also poses huge challenges in e-commerce. How to obtain the consumer's propensity of customers from the rules of customers' browsing products is a problem that Internet companies urgently need to solve. The combination of the emergence of big data and the development of e-commerce is an inevitable result of social development, and it will lead us into a new innovative development. With the passage of time, how to provide personalized services and recommendations for consumers' consumption and preferences will become a research hotspot.

This paper adopts a new simple classification rule method to classify consumer groups by hierarchical clustering algorithm. In the second part, the theory of hierarchical clustering algorithm and analytic hierarchy process have been introduced respectively. In the experiment, the parameter settings used in this article are shown in Table 1.

| Parameter name | Value |
|----------------|-------|
| Experimental tools | PyCharm |
| X               | \([[1,2],[2,3],[4,4],[1,2],[1,3]]\) |
4. Discussion

4.1 Data Mining Concept
Data mining is to extract or "mine" the knowledge we want to extract from a large number of data. To be exact, data mining refers to the process of extracting hidden, unknown, but possibly useful information and knowledge from incomplete, noisy, fuzzy and random practical application data. The above definition of data mining includes the following meanings: data source must be real; a large number of data may contain noise; consumers are interested in knowledge; knowledge discovery should be able to accept, understand and apply knowledge; it is not necessary to find accurate data anywhere. Data analysis, decision support and data mining are similar, which can show that data mining has been applied to various fields, especially in the intelligent processing of big data.

In this chapter, firstly, the business data at three levels are counted. Combined with the actual business data, the data points are calculated based on the hierarchical clustering analysis algorithm, and the experimental data are combined according to the data. The experimental results are compared with the actual business applications. The experimental results show that it has a certain reference value in the actual business data analysis to reflect the role of data in the business field. Data comes from life. In this era of big data explosion, data is an indispensable part of our development. We can correctly analyze and use data to promote the development of all walks of life.

4.2 Data Sources Based on Hierarchical Clustering Algorithm
The data of this experiment mainly come from three parts: according to the number of customers, according to the types of goods and according to the platform.

1. Divide statistics based on the number of customers
   In the data preparation, the customer information in this experiment includes the customer browsing goods, adding goods to the shopping cart, the number of various kinds of goods, customer information (including customer ID, user gender, age, and the total number of customers who have purchased the product). Each customer information is input into each data point. According to the actual amount of data, the customer types are classified according to the actual statistical data, Customers are divided into three categories. The first category is the customers who have browsed the goods. These customers are collectively referred to as customer resources. The second category is the customers who add goods to the shopping cart. In this experiment, these customers are collectively referred to as potential customers. The third category is the total number of people who purchase products. These customers are collectively referred to as actual customers.

2. Statistics based on the number of customers
   The commodity information compiled in this paper mainly includes commodity type, commodity name, commodity price, commodity scale, etc. The customers who have browsed the product are divided into one category, the customers who have purchased the product are divided into one category, and the customers who have added the product into the shopping cart are divided into one category. Then, the average price of goods purchased by customers in each category is divided into new categories. The Euclidean distance between each class is calculated, and then the distance between each data point is calculated. Finally, according to the algorithm mentioned in this paper, each similar class is merged to calculate the satisfaction.

3. Statistics based on platform Division
   In the division of the platform, the number of customers who buy goods on different platforms is mainly selected. In the experiment of this paper, the actual data of resource customers, potential customers, actual customers, jackets and skirts are selected to analyze the fusion algorithm of hierarchical clustering and analytic hierarchy process.

4.3 Data Analysis Based on Hierarchical Clustering Algorithm
This paper uses a hierarchical clustering algorithm to analyze consumer browsing product data, shopping cart data and purchase product data. In this test session, this article selects the consumer data of the item that has been purchased on the platform and the item added to the shopping cart, and selects the top seven products on the platform for display. According to the data of consumers' purchases on the website, the experimental results are shown in the figure below.

![Figure 1. Summary of customer purchase data](image)

From the data on Figure 1, we can see that after the results of the hierarchical clustering algorithm used in this article are applied to online platforms, consumers who buy short sleeves are the most, with 59 people, followed by consumers who buy coats with 47 people. The number of people buying suits and fruits is the least, 21 and 19 people respectively. From the data, we can see that consumers who buy clothes on online platforms account for the majority, followed by consumers who buy snacks, and then sports products. The fruit with the least purchasing power is mainly because the fruit has the shortest shelf life, and the probability of damage during transportation is higher. If you buy dresses for women and basketball for men, we can see that female consumers are higher than male consumers in the number of people who use the website. Therefore, when providing merchandise recommendations to merchants, based on the results of this article, it is recommended that merchants choose clothes first, then snacks, and finally fruit.

Since there are the most people who buy short sleeves, we will analyze the price of browsing records and buying habits of those who buy short sleeves.
Figure 2. Short sleeve price data analysis

From the data in Figure 2, we can see that when 30% of people buy short sleeves, they hope that the price of short sleeves is between 50-60 yuan, and the proportion of consumers who want the price below 40 yuan is 16%, the percentage of consumers who hope the price is above 80 yuan is 5%. 70% of people hope that the price can be less than 60 yuan. After comparing with the actual situation, the results of this experiment are consistent with the results of the short-sleeved shops sold in the actual situation. Reference value. In the future data analysis of the platform, this method can grasp the buying habits of different types of customers and the popularity of different types of goods at any time. After mastering the purchase rules of these customers, we can promote the corresponding products.

5. Conclusions

With the continuous development of information technology, information data has exploded. Data mining is to extract the data we are interested in from the massive data, and promote the development of the industry based on the data prediction results. This article first introduces the analytic hierarchy process and the hierarchical clustering analysis algorithm, and then classifies the data needed for this experiment into three categories according to consumers. Then use hierarchical clustering algorithm to conduct data statistics on the browsing records, consumption records and the records added to the shopping cart of consumers. The experimental results show that the results based on the analytic hierarchy process are consistent with the results of the actual items sold on the platform, indicating that the results of the hierarchical clustering analysis of commercial data in this article have a certain reference value for the platform. This experiment also confirmed that data mining can be applied to all areas of business.

Acknowledgements:

This research has been financed by Major project of philosophy and social science research in Jiangsu Universities (2019SJZDA033) and the research project of higher education reform in Jiangsu Province(2019JSJG499).

References
[1] Yang Yanying. Research on clustering algorithms under the background of big data[J]. Computer Programming Skills and Maintenance, 2019, 403(01): 68-71.

[2] Chen Xiao, Zhao Jingling. Research and Implementation of Hybrid Clustering Algorithm in Big Data Processing[J]. Information Network Security, 2015(04): 45-49.

[3] Yang Jing. Discussion on fingerprint collection and resolution technology based on big data[J]. Journal of Heilongjiang Institute of Technology: Comprehensive Edition, 2019, 019(011): 54-57.

[4] Liu Sihong[1], Yu Fei[1]. Application research of online learning behavior path based on K-means clustering algorithm under big data[J]. Journal of Lanzhou University of Arts and Science (Natural Science Edition), 2019, 33(01): 75 -79.

[5] Jiang Yanwen. Simulation of efficient clustering method for multi-dimensional discrete data under big data analysis[J]. Computer Simulation, 2019, 36(02): 205-208.

[6] Liu Xiaojin. Research on wireless network optimization based on big data application[J]. Journal of Anyang Normal University, 2019, 000(002): 44-47.

[7] Yu Hong, Lin Jihong, Hou Chao, et al. Mining hotspot areas of library business in the world under the background of big data——An empirical analysis based on CNKI bibliographic data[J]. Library Work and Research, 2018, No.264(02): 60-68.

[8] Luo E T, Wang G J. A hierarchical clustering method based on the threshold of semantic feature in big data[J]. Journal of Electronics & Information Technology, 2015, 43(31): 15 -19.

[9] Xia X. Fast search of art culture resources based on big data and cuckoo algorithm[J]. Personal and Ubiquitous Computing, 2020, 24(1): 127-138.

[10] Zhu Q, Wu J, Song M. Efficiency evaluation based on data envelopment analysis in the big data context[J]. Computers & Operations Research, 2017, 98(OCT.): 291-300.