International Core Journal of Engineering Volume 8 Issue 8, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0047

Abnormal Commodity Identification based on Optimized K-means Algorithm

Jinjin Wang[a], Yang Li[b], Boyang Ji[c], Jie Chen[d], Tao Han

School of Information Engineering, Yancheng Teachers University, Yancheng 224000, China

[a]wangjj@yctu.edu.cn, [b]1642479722@qq.com, [c]1614649552@qq.com, [d]chensj0330@126.com

Abstract

With the popularity of online shopping, while the scale of various online shopping platforms continues to expand and the number of products continues to increase, some improper business practices, such as false price quotations and swiping behaviors, also appear. Therefore, commodity data needs to be identified to ensure the interests of consumers. Regarding the issue above, this paper proposes an Abnormal Commodity Identification Algorithm based on the Optimized K-means Algorithm(ACI-OKA). In this paper, Abnormal commodity identification is divided into three parts: (i) Classification of commodity data; (ii)Determination of K value; (iii) Identify the commodity data under each classification. Experiments show that it takes 24 hours for manual detection of the same classification of commodity, while ACI-OKA only takes 25 seconds. Compared with the KNN algorithm, ACI-OKA is more convenient. Compared with the Isolation Forest algorithm, ACI-OKA reduces the running time by 30% and improves the accuracy by 27%.

Keywords

Abnormal Commodity Identification; Optimized K-means Algorithm; Classification.

1. Introduction and Related Work

In recent years, Internet technology has developed rapidly, and the e-commerce industry has also embarked on the fast track of development. "Online shopping" is more and more popular because of its convenience, time saving and door-to-door delivery. As the scale of each platform continues to expand and the number of commodities continues to increase, some improper business practices also appear, which seriously violates the e-commerce law, and it is necessary to accurately identify commodity data.

Therefore, the identification of abnormal commodities and the detection of outliers have become a hot topic. Dhiren Ghosh et al. [1] made a methodological evaluation on outliers and defined outliers as correct data. By comparing other data, it was found that the data information of outliers was contrary to other data information, indicating that this data was outliers. The accuracy of outliers detected by this method was higher than 95%, and outliers were detected effectively. Steven Walfish et al. [2] studied outlier detection and found that with the increase of data, the value of outliers was skewed. First, the author wanted to find outliers by deleting data points, and then used the weighted least square regression method. The authors believed that there are many methods for outlier detection, which depend on the data set. Therefore, the author had studied Box plot, Trimmed means, Extreme studied deviate and other methods to detect outliers. Irad Ben-Gal et al. [3] thought that outlier detection in data mining was usually based on distance, measurement, clustering and spatial methods and classified outlier methods into univariate methods proposed in earlier works in this field and multivariable methods usually formed. The former divided the data into two non overlapping sets: outliers and non outliers. The latter provided ranking by assigning an outlier category to each data.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0047

Hongzhi Wang et al. [4] thought that outlier detection could be transformed into important operational information in various applications, such as fraud detection, intrusion detection and health diagnosis in network security and introduced many outlier detection methods, including statistics based methods (similar to [1]), distance based methods (mentioned in [3]), density based methods, clustering based methods, graph based methods, integration based methods, and learning based methods. Different methods could be used to detect outliers in different scenarios. Erich Schubert et al. [5] conductd outlier detection based on ranking and score. Similar to [3], the author calculated the distance between scores, used statistical knowledge to judge whether the scores were abnormal, and then compared the scores with the rankings to obtain the abnormal ranking.

2. System Model

2.1 Application of K-means Algorithm

We assume that the data sample is X and contains n objects $X = \{x_1, x_2, x_3, \dots x_n\}$, where each object x_i has two attributes, price x_{ip} and sales volume x_{isv}. The goal of the K-means algorithm is to cluster n objects into K clusters $C = \{C_1, C_2, C_3, \dots C_k\}, 1 < K \leq n$ according to the similarity between objects, and each object belongs to one and only one cluster whose distance from the center of the cluster is the smallest.That is to minimize the squared error.The formula for squared error is [6]:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - u_i|^2 \tag{1}$$

u_i in the formula (1) is the mean vector, which has two attributes, price u_{ip} and sales volume $u_{isv}$and the formula is as follows[6]:

$$u_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

Minimizing formula (1), that is, finding its optimal solution, needs to consider all possible cluster divisions of sample X, which is an NP-hard problem. Therefore, the K-means algorithm approximately solves (1) by iterative optimization. The algorithm steps are as follows. First, K-means algorithm selects K samples from the data samples as the initial mean vector $U = \{u_1, u_2, u_3, \dots u_k\}$, and then calculates the Euclidean distance from each object to each mean vector. The Euclidean distance formula is as follows:

$$dis(x_i, u_j) = \sqrt{(x_{ip} - u_{jp})^2 + (x_{isv} - u_{jsv})^2} \tag{3}$$

Compare the $dis(x_i, u_j)$ from each object x_i to each cluster center C_j in turn, assign the object to the cluster with the closest cluster center, and get K new clusters $U' = \{u'_1, u'_2, u'_3, \dots u'_k\}, 1 < K \leq n$. Repeat the above steps until U' no longer changes.

2.2 Optimization of K-means Algorithm

In the K-means algorithm, the value of K, that is, the number of clusters, is critical. Different K values get different results. The K value of the K-means algorithm is generally determined by the user. The

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0047

value of k is generally determined by the elbow rule. The formula for the cost function of the elbow rule is as follows:

$$cost = \frac{1}{K}\sum_{i=1}^{k}\sum_{x\in C_i}|x-u_i|^2 \quad (4)$$

When the selected K value is smaller than the real K, every time K increases by 1, the cost value will be greatly reduced; When the chosen value of K is larger than the true K, the change in $cost$ value will not be so obvious for each increase of 1 in K. Thus, the correct value of K will be at this turning point.

Sometimes the elbow rule cannot accurately determine the value of K in all cases, because the turning point may not occur. Therefore, we use a threshold φ to constrain the clusters. When $dis(x_i, u_j) > \varphi$, x_i does not cluster.

3. ACI-OKA

In this section, we consider applying optimized K-means algorithm to abnormal commodity identification. The process of identifying abnormal commodity in this paper is as follows. First, classify the commodity data, then determine the K value, perform algorithm clustering, and finally identify abnormal commodity data.The purpose is to find abnormal commodity data in commodity data. This paper proposes ACI-OKA to obtain abnormal commodity data.

3.1 Classification of Commodity Data

There are a wide variety of commodities on the platform, and there is also a lack of comparability between various commodities. The commodity data in this paper is Taobao background data, provided by related companies. The attributes of product data include commodity price, sales volume, brand, category and store information, etc. There are at most 5 levels in the category of commodity data. There are many kinds of commodities on the commodity data platform, and there is also a lack of comparability among various commodities. Therefore, it is necessary to first classify all commodities according to the fifth level category. The classification method of commodity data in this paper is as follows. Traverse all commodity data to get all fifth-level categories $Class = \{class_1, class_2, class_3 \ldots class_m\}$,m is the number of classification. Then, according to the fifth-level classification attribute of the commodity data, every commodity x_i is divided into corresponding classification $class_z^i$. $class_1^i$ indicates that x_i is in the class $class_z$. Each classification $class_z$ generates K clusters $C_z = \{C_z^1, C_z^2, C_z^3, \ldots C_z^k\}$,$1 < K \leq |class_z|$. φ_z Represents the threshold under classification.

3.2 Determination of K Value

Prices and sales of normal commodity are generally positively correlated. Normal data points are concentrated, and abnormal commodities are generally located at the end edge of the X-axis or Y-axis. That is, the optimized K-means algorithm generates an area with relatively concentrated data points, and the data in this area are all normal data. On the contrary, abnormal data is abnormal commodity. Figure 1, Figure 2, Figure 3 and Figure 4 are the clustering results of the optimized K-means algorithm when K=1, K=2, K=3, K=6, respectively. The X-axis of the axes represents price and the Y-axis represents sales. 0 represents normal data, 1 represents abnormal data. The experimental results show that when K=1, the effect is the best.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0047

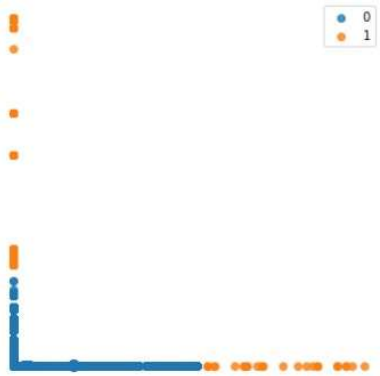

Figure 1. The clustering result of K=1

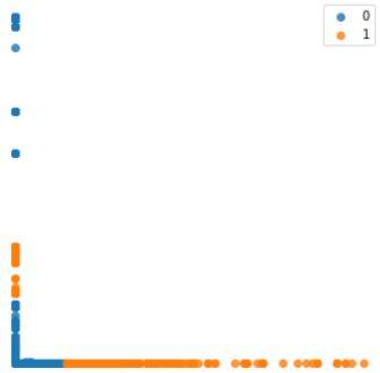

Figure 2. The clustering result of K=2

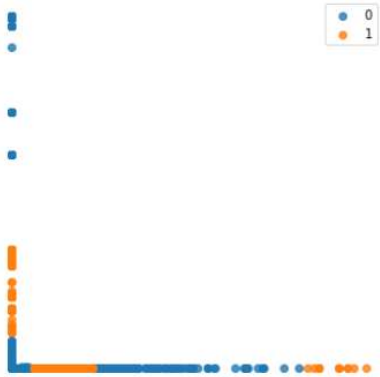

Figure 3. The clustering result of K=3

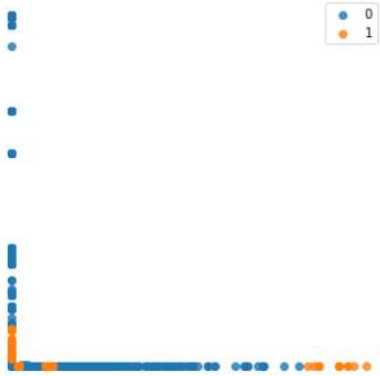

Figure 4. The clustering result of K=6

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0047

3.3 ACI-OKA Identification Algorithm

Algorithm 1 is the execution flow of the ACI-OKA algorithm. The input to this algorithm is set of commodity data $X = \{x_1, x_2, x_3, \dots x_n\}$, the K value of the K-means algorithm, and classification of all commodities $C_z = \{C_z^1, C_z^2, C_z^3, \dots C_z^k\}$. The output is a set of abnormal commodity data. This algorithm is described in detail below.

First, The result of abnormal commodity data is represented by the set $result$ (line 1). Iterate over each classification (line 2). The original cluster C_z^j is set to empty (line 4) . Randomly select x_i from $class_z^j$ as the initial mean vector $\{u_j\}$ (line 5). Traverse the sample x_i under each classification, calculate $dis(x_i, u_j)$, if $dis(x_i, u_j) < \varphi$, then add x_i to the cluster C_z^j, until the cluster with the largest value $|C_z^j|$ is found (line 6-11). Finally, the data that does not belong to the cluster C_z^j is added to $result$ to obtain abnormal commodity data (line 12).

Algorithm 1: ACI-OKA Identification algorithm

Input: $X = \{x_1, x_2, x_3, \dots x_n\}$:Set of commodity data

K: K value of K-means algorithm.

$C_z = \{C_z^1, C_z^2, C_z^3, \dots C_z^k\}$:Classification of all commodities.

Output: $result$: Set of abnormal commodity data

1, $result \leftarrow \emptyset$
2, for $class_z \in Class$ do
3, repeat
4, $C_z^j = \emptyset$
5, Randomly select x_j from $class_z^j$ as the initial mean vector $\{u_j\}$
6, for $x_i \in class_z$ do
7, Calculate $dis(x_i, u_j)$
8, if $dis(x_i, u_j) < \varphi_z$
9, $C_z^j \cup \{x_i\}$
10, end
11, until $|C_z^j|$ is the largest
12, $result \cup \{\{x_i | x_i \in class_z\} \setminus C_z^j\}$
13, end
14, return $result$

The following is the time complexity analysis of algorithm. The time complexity of classification of commodity data is $O(n)$, where n represents the number of data. The time complexity of identifying the abnormal data model is $O(|Class| * \max_{x_j \epsilon class_j} |class_j|)$, $|Class|$ represents the number of all classification, and $\max_{x_j \epsilon class_j} |class_j|$ represents the number of commodities under the classification with the largest number of commodities. Therefore, the time complexity of the ACI-OKA algorithm is $O(\max\{n, |Class| * \max_{x_j \epsilon class_j} |class_j|\})$.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0047

4. Experiment and Result Analysis

In this section, we verify the algorithm ACI-OKA by experiments, mainly considering two performance indicators: time and accuracy, mainly including the following: (1) Compare the convenience of KNN algorithm and ACI-OKA algorithms; (2) Comparing the time and accuracy of identifying abnormal commodities by manual detection and ACI-OKA algorithms. (3) Comparing the time and accuracy of identifying abnormal commodities by the Isolation Forest algorithm and the ACI-OKA algorithm

4.1 Experimental Design

For an introduction to sample data, see Section 3.1. The sample data is about 16 million. The experimental code is written in Python and runs on the local computer. The configuration of the local computer is as follows: CPU: Intel Core i7, memory: 8G, external memory: 512G.

4.2 Analysis of experimental Results

4.2.1 Comparison with KNN

When KNN identifies abnormal commodity data, it needs to be trained on the test set and then supervised learning. The samples need to be labeled, which takes a lot of manpower and time. ACI-OKA can directly identify abnormal products, which is more convenient.

4.2.2 Comparison with Manual Detection

In this experiment, the commodity data under the same classification is used as the experimental data. The number of people is 5. This paper assumes that the accuracy of manual detection is 100%. Table 1 shows the experimental results of manual detection and ACI-OKA Algorithms. According to the results of manual detection, the accuracy of ACI-OKA algorithm is 95%. However, it takes more than 24 hours to manually detect commodities under the same classification, and the ACI-OKA algorithm only takes 25 seconds. As the amount of data increases, the time required for manual detection also increases exponentially. And with the increase of time, the time required for manual detection also increases exponentially at the same time.

Table 1. Experimental results of manual detection and ACI-OKA Algorithms

Algorithm	Time spent identifying abnormal commodities	The accuracy of identifying abnormal commodity	The relationship between data volume and time consumption
Manual detection	More than 24 hours	100%	When the amount of data increases, the time spent increases exponentially
ACI-OKA	25 seconds	95%	positive correlation

4.2.3 Comparison with Isolation Forest algorithm

Table 2 shows the experimental results of Isolation Forest algorithm and ACI-OKA Algorithms. In this experiment, the experimental data is 100 classifications of commodity data. In terms of the time it takes to identify abnormal commodities, the Isolation Forest algorithm takes 55 minutes, while the ACI-OKA algorithm takes only 40 minutes. Because the Isolation Forest algorithm needs to give the abnormal data ratio of commodities in advance, it cannot accurately obtain abnormal commodities. In terms of the accuracy of identifying abnormal commodities, the accuracy of the isolated forest algorithm is only 68%, while the accuracy of the ACI-OKA algorithm is as high as 95%.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0047

Table 2. Experimental results of Isolation Forest algorithm and ACI-OKA Algorithms

Algorithm	Time spent identifying abnormal commodities	The accuracy of identifying abnormal commodities	The relationship between data volume and time consumption
Isolation Forest	55 minutes	68%	positive correlation
ACI-OKA	38 minutes	95%	positive correlation

5. Conclusion

This paper studied the problem of abnormal commodity identification from the aspects of commodity price and sales. The K-means algorithm is applied to the problem of abnormal commodity identification, and an optimization-based K-means algorithm ACI-OKA is proposed. ACI-OKA classifies the commodities, determines the K value using the elbow algorithm and threshold, and finally identifies the commodities. Three experiments were performed in this paper. Experiments show that the algorithm ACI-OKA proposed in this paper has high efficiency and accuracy.

References

[1] Ghosh, Dhiren, and Andrew Vogt. "Outliers: An evaluation of methodologies." Joint statistical meetings. Vol. 2012. 2012.

[2] Walfish S. A review of statistical outlier methods[J]. Pharmaceutical technology, 2006, 30(11): 82.

[3] Ben-Gal, I. (2005). Outlier detection. In Data mining and knowledge discovery handbook (pp. 131-146). Springer, Boston, MA.

[4] Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. Ieee Access, 7, 107964-108000.

[5] Schubert E, Wojdanowski R, Zimek A, et al. On Evaluation of Outlier Rankings and Outlier Scores[C]// 2012.

[6] Kanungo T , Mount D M , Netanyahu N S , et al. An efficient k-means clustering algorithm: analysis and implementation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(7):881-892.