# A detection algorithm of customer outlier data based on data mining technology

*Jia Ren*
Xi’an Fanyi University, Xi’an, Shaanxi,China
Corresponding author: Jia Ren, Master Degree,495032517@qq.com

**Abstract**

For the outlier data detection problem of customer transactional retail data in a large-scale chain supermarket, customer transaction data are detected by data mining technology and database technology, the sample data of abnormal customer behavior have been chosen in the customer transaction database, the abnormal customer behavior will be found out for outlier samples data fusion by the Dempster/Shafer evidence theory. The experimental result shows that the algorithm is more accurate and efficient than other algorithms to detect abnormal customer transactional retail behavior by the Dempster/Shafer evidence theory.

**Keywords:** *outlier membership data; dempster/shafer evidence theory; algorithm; data fusion*

## 1. Introduction

At present, most supermarkets in China have collected a large amount of customer transaction data by POS system, the transaction data contain the transaction time, commodity name, transaction amount. The manager analyzes the transaction data by database technology before implementing of the corresponding marketing strategy for customer behavior change. Abnormal customers behavior data are an integral part of the enterprise customers. The enterprise not only pays attention to transactional retail data changes of the normal customer but also pays attention to the behavior of abnormal customer because abnormal customer behavior analysis is beneficial for enterprises to prevent the loss of customers. High level of employee retention has become the most important factor in business success. Reichheld and Sasser found out the industry average profit rate will increase between 25% and 95% if customer retention rate increases 5% by survey data of nine industries in the United States in 1990. High level of customer retention has a great impact on enterprise profits because cost of returning visitors is much lower than the cost of acquiring new customers. Therefore, the manager of enterprises should pay attention to the existing customers and take corresponding measures to avoid the loss of existing customers. Statistics show that most of the company's sales come from the 12 percent of key customers, while 88 percent of the common customer is a small profit for the business[1,2]. The cost of developing a new customer is 5 times than the cost of retaining an old customer. Therefore, the enterprise manager need to pay attention to the factors of affecting customer behavior and study the characteristics of abnormal customer behavior in order to provide guidance for enterprise's marketing decision.

In twenty-first Century, Customer needs are diversified and complex, and consumer behavior appear diversification. The enterprise manager should study various characteristics of abnormal customer behavior according to changes of customer behavior. This shows that, research on abnormal customer plays a very important role in customer relationship management. Abnormal customer behavior is difficult to be detected because of its characteristics of uncertainty. Dempster/Shafer theory expresses the important concepts of uncertainty and has simple reasoning form without necessary prior probability[3]. Unfortunately, there are some shortcomings about Dempster/Shafer evidence theory, such as the large amount of calculation and the tedious process[4]. Therefore, this paper uses the distance based outlier data detection method to select sample data from the massive transaction data.

### 1.1 Customer behavior analysis.

The abnormal sample data were extracted to fusion of the outlier sample data by Dempster/Shafer evidence theory. We adopt the isolated datum method based on distance between one transactional retail data and other transactional retail data. The original data have an impact on calculation of two transactional retail data distance. To calculate distance of two transactional retail data with processed data, Assuming $x_j$ represents jth element, $R_j$ represents absolute deviation, $S_j$ represents standard deviation.

$$x_{ij}' = \frac{x_{ij} - \overline{x_j}}{R_j} \quad \text{or} \quad x_{ij}' = \frac{x_{ij} - \overline{x_j}}{S_j} \tag{1}$$

### 1.2 Extraction of detection samples based on distance and abnormal customer

The $R_j$ and Sj will have an effect on the isolated datum among the sample data, the effect of the isolated datum is reduced because the deviation of the attribute value and the mean value cannot been squared in the calculation. Literature review points out that $R_j$ is better robustness than $S_j$ in outlier customer behavior detection, we hope to highlight the isolated point as much as possible after standardization of data points. The common distance is absolute distance and

Euclidean distance based on distance between two customer transactional retail data and the automatic detection method of isolated points. The distance of the absolute datum is adopted in this paper[5,6].

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right| \qquad (2)$$

The advantage of outlier customer behavior based on distance detection is that we do not need to know the distribution model of transactional retail data, which can be applied to any feature space by some distance formulas. The customer behavior characteristics are measurement spaces based on distance formulas. Therefore, distance based outlier data detection method is used to extract the isolated point data of customer behavior. But it is necessary to determine the parameters p and D Based on distance outlier data detection. The formulas are difficult to calculate P and D. Therefore, outlier data detection method based on distance is improved in this paper. In the outlier data detection, the parameters P and D cannot be determined.

### *1.3. Outlier detection overview*

At present, the algorithms of isolated point datum detection include method based on statistical, method based on distance, method based on deviations and method based on density. The concept of isolated points datum based on distance among some transactional retail data was first proposed by E.M.K norr and R.T.N g, S.D.Bay Ramaswamyetal and improved by S.D.B. ay. the nearest datum point is generally judged to be an isolated point by this method[7].

## 2. Outlier customer behavior detection based on distance

### *2.1 Isolated points datum detection*

Large-scale chain supermarket has accumulated a large amount of transactional retail data, such as customer transaction record, goods in and out data, consumption records by POS and CRM. The manager can find the valuable information through rules mining and data analysis. The abnormal behavior customer was detected by the outlier detection method based on the distance and located according to the information characteristics of customer transaction records in large-scale chain supermarket.

First, the raw data usually have a specific unit. The different unit measures affect the results of calculation the distance. The raw data should be pretreatment before calculating distances. $x_j$ represents means of jth element, $R_j$ represents the absolute deviation and $S_j$ represents the standard deviation[8].

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad R_j = \frac{1}{n}\sum_{i=1}^{n}\left|x_{ij} - \overline{x_j}\right| \qquad S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2} \tag{3}$$

$$X_{ij}' = \frac{X_{ij} - \overline{X_j}}{S_j} \qquad \text{Or} \qquad X_{ij}' = \frac{X_{ij} - \overline{X_j}}{R_j} \tag{4}$$

Among above formulas, j represents the order of customer transaction retail data in each month, and n represents the customer transaction times, and $X_{ij}$ presents the transaction money amount of the ith month and the jth time. $S_j$ has an effect on the isolated point of customer transactional retail data because the property value and the average value are not squared before calculating $R_j$, the influence of calculating the isolation point datum is reduced. $R_j$ has better robustness than $S_j$. We hope to highlight the isolation point datum as much as possible after standardization of data points, therefore, $S_j$ is used.

Second, the most common distance of customer transactional retail data is absolute distance and Euclidean distance based on distance and isolated point automatic detection method, absolute distance is also called Manhattan distance.

Manhattan distance is defined $d_{ij} = \sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|$ as:

Euclidean distance：

$$d_{ij} = \sqrt{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2} \tag{5}$$

Above formulas, m is the dimension of the customer transactional retail data, $X_{ij}$ represents the value of the jth month of the ith customer. The Minkowski distance can be defined as:

$$d_{ij} = \left[\sum_{k=1}^{m}\left|X_{ik} - X_{jk}\right|^q\right]^{\frac{1}{q}} \tag{6}$$

Third, Establishment distance matrix based on distance and isolated point detection, the method is defined as:

After standardizing the original data, calculate the distance between two transactional retail data, accumulating the distance. All $d_{ij}$ calculated based on the absolute distance formula produce the distance matrix R.

$$R = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n,1} & d_{n,2} & \cdots & d_{n,n} \end{bmatrix} \tag{7}$$

Fourth, The largest Pi is the farthest away the customer I, and the customer I may be an isolated point of the abnormal customer behavior.

Fifth, Finding the average of $p_i$, then compare $p_i$ with p, if $p_i > p$, then this point i is the isolation data point of the abnormal customer behavior.

### *2.2 The algorithm for outlier data of the abnormal customer behavior*

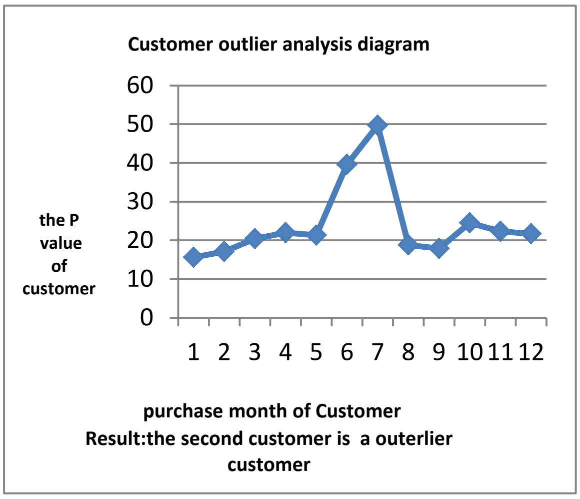
$$\overline{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

In order to determine the isolation point datum, the customer transactional retail data need to be standardized before data analysis before the algorithm of isolation point detection is carried out. The following is a description of the isolation point algorithm based on the distance:

```
Private function GetMaxOutlier()
  StandardizeDataProcess ()    //standardized processing of the original data
  Get_nextData(user,result)    //Get the result data for the next customer
        For j=1 to month_Num    //Get the value of Pi
              For k=1 to Times_Num
                    P(j)=GetResult.p1(j,k)
              Next k
        Next j
    GetMaxoutlierPjm(p(j))    // Gets the largest object P (J, m) value in P (J)
End function
…….
```

### *2.3 Experimental data analysis*

In order to verify the validity of the algorithm, we choose customer transaction retail data in the experiment from a large-scale chain supermarket for 12 months, the experimental data from June 2015 to June 2016, we choose the variables of the customer transaction frequency and the transaction amount, the images in the form of chart express tendency of an abnormal customer behavior data because the outlier datum is difficult to understand, the algorithm in this paper takes two typical customer behavior data to test results with one abnormal customer data and another abnormal customer data, result shown in the Fig.1 and Fig.2 as below.



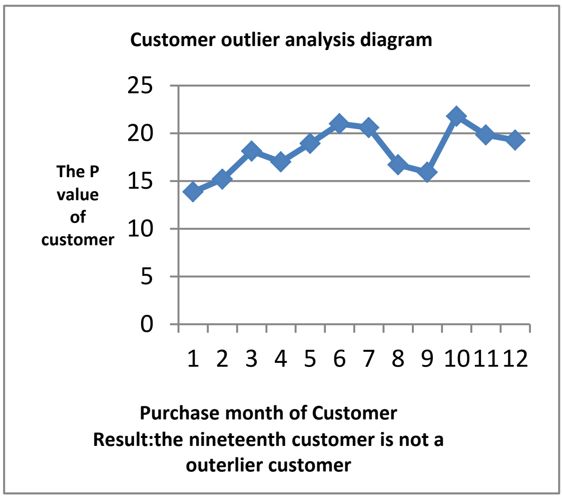


*Fig.1*- deviation with the larger waveform

*Fig.2*- deviation with the small waveform

## 3. Conclusions

The algorithm of outlier customer behavior detection based on distance was discussed in a large-scale chain supermarket, the algorithm has been verified effectiveness of outlier customer behavior detection method in a large-scale chain supermarket, the results proves that the algorithm can not only be used to find out the customer purchasing behavior volatility of those processes in the customer transactional retail database but also make outlier data detection of the consumer transactional behavior. The algorithm provides effective marketing basis for customer management of a large-scale chain supermarket and lays a theoretical foundation for customer forecasting research.

## Acknowledgement

In this paper, the research was sponsored by the science research projects of education department in Shaanxi provincial government (Project No. 16JK2078).

## References

1. *J.Dezert, D. Q.Han, Z. G. Liu, et al.* Hierarchical DSmP transformation for decision-making under uncertainty[C].The 15th Int Conf on Information Fusion. Singapore, 2012:294-301.
2. *S.Yao, Y. J. Guo, W. Q. Huang,* An improved method of aggregation in DS/AHP for multi-criteria group decisionmaking based on distance measure[J]. Control and Decision, 2010, 25(6): 894-898.)

3. *A. O.Boudraa, A.Bentabet, F.Salzenstein, et al.* Dempster-Shafer’s basic probability assignment based on fuzzy membership functions[J]. Electronic Letters on Computer Vision and Image Analysis, 2004, 4(1): 1-9.
4. *H. Y.Liu, Z. G. Zhao, X. Liu,* Combination of conflict evidences in D-S theory[J]. J of University of Electronic Science and Technology of China, 2008, 37(5): 701-704.
5. *B.Dierynck, W. R.Landsman, A.Renders,* Do Managerial Incentives Drive Cost Behavior? Evidence about the Role of the Zero Earnings Benchmark for Labor Cost Behavior in Private Belgian Firms. The Accounting Review, 2012, 87(4): 1219-1246.
6. *P. N.Patatoukas,* Customer-base Concentration: Implications for Firm Performance and Capital Markets. The Accounting Review,2012, 87(2): 363-392
7. *X.Yang, Y. S. Dong, et al.* Online correlation analysis for multiple dimensions data streams [ J] .Journal of Computer Research and Development , 2006 , 43(10):1744-1750.
8. *J.Podesta, P.Pritzker, E. J.Moniz, et al.* Big Date: Seizing Opportunities, Preserving Values[R]. American: Executive Office of the President, 2014.