A Classification Algorithm: Data Mining and Mathematical Model

R Cheng*, X Kong, M Yu and N Wang
Beijing Polytechnic, Beijing, 100028

*Corresponding author email: tianxiaocheng0925@126.com

Abstract. In this paper, we propose a classification algorithm based on Recency-Frequency-Monetary (RFM) model and K-means data mining method. In addition, the designed algorithm is verified by the experiments on the member data in a large shopping mall. The experiments results show that the proposed algorithm can provide an accurate classification of the members. Finally, some marketing strategies for different classes of members are given according to the classification results.

Keywords: Classification algorithm, RFM model, K-means data mining method.

1. Introduction
Data classification refers to the process of distinguishing and classifying data using certain principles and methods according to its attributes or characteristics, and establishing a certain classification system and order. As a common data mining method, data classification is generally used to improve management ([1],[2]). Therefore, the classification problem has been a popular research topic in recent years, and many studies have proposed various methodologies for the problem ([3]-[5]). Because of the speed and accuracy in classification, k-means is considered as a popular classification method ([6],[7]). For example, based on a real-world customer data of an enterprise, Wu et al. [8], designed a classification strategy according to K-means method; Zhang et al. [9], cluster highly-correlated neighboring samples into multi-dimensional vectors using K-means method; The weighted k-means method is applied to forecast rainfall by identifying related meteorological data in Zhang et al. [10].

In this paper, we proposed a classification algorithm based on k-means method and RFM model. Next, the algorithm is used to make a scientific classification of the mall members as a real-data based case study. In the case study, we first find the proper number of classes of the members. And then, the classification results are compared between the algorithms using different normalization methods. Finally, we find the proper classification algorithm for the mall members. It must be mentioned that the classification method proposed in this paper is not limited to the classification of members, but can also be used for graduate employment analysis and student mental health analysis.

The remainder of this paper is organized as follows. In Section 2, we provide an introduction of k-means algorithm and the relevant normalization methods. In Section 3, a real-data based case study is conducted by using the propose algorithm. In Section 4, the experiment results are provided.

2. Classification Algorithm
K-means is one of the classical clustering algorithms in partition methods [11]. Because of its high efficiency, this algorithm is widely used in clustering large-scale data. This paper designs a classification algorithm for mall members by using k-means clustering method.
The basic idea of K-means clustering method is to find \( K \) clusters by iteration with the goal of minimizing the loss function corresponding to the clustering result. Here, the loss function can be defined as the sum of distance between each sample and the centre node of the cluster, which can be formulated as

\[
F(c, \mu) = \sum_{j=1}^{N} \| x_j - \mu_j \|_2^2
\]

(1)

where \( x_j \) is the \( j \)th sample of the member consumption data, \( c_j \) is the cluster to which \( x_j \) belongs, \( \mu_j \) refers to the centre node of \( c_j \), and \( N \) is the number of samples.

The detailed steps of K-means classification algorithm are listed in Table 2.

Table 1. K-means classification algorithm.

| Step | Description |
|------|-------------|
| 1:   | Selection \( K \) samples from the \( N \) samples as the centre nodes \( \mu_1, \mu_2, \ldots, \mu_K \), set \( M \), the maximum number of iterations, \( \varepsilon \), the accuracy requirements, and \( D' = \text{Inf} \); |
| 2:   | While \( m < M \) & \( D' < \varepsilon \) //Iterate until the maximum number of iterations is reached or the accuracy requirements are met; |
| 3:   | set \( C_i = \emptyset \) \((1 \leq i \leq K) \) //Initially, all clusters are empty; |
| 4:   | for \( j = 1: N \) // \( N \) is the number of samples; |
| 5:   | \( d_{ji} = \| x_j - \mu_i \|_2 \) //Calculate the distance between \( x_j \) and \( \mu_i \), i.e., \( d_{ji} \); |
| 6:   | \( \pi_j = \arg\min_{i \in \{1,2,\ldots,K\}} d_{ji} \) //Find the cluster closest to \( x_j \), i.e., \( C_{\pi_j} \); |
| 7:   | \( C_{\pi_j} = C_{\pi_j} \cup \{x_j\} \) //Classify \( x_j \) into the closest cluster \( C_{\pi_j} \); |
| 8:   | end for |
| 9:   | for \( i = 1: M \) //Update centre node of the cluster; |
| 10:  | \( \mu'_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \) //generate new centre node of the cluster; |
| 11:  | if \( \mu'_i \neq \mu_i \), then |
| 12:  | \( D = D + \| \mu'_i - \mu_i \|_2 \), \( \mu_i = \mu'_i \) //Accumulation the distance between \( \mu_i \) and \( \mu'_i \), and update centre node |
| 13:  | end if |
| 14:  | end for |
| 15:  | \( D' = D \) |

K-means is essentially a data classification method based on Euclidean distance measurement. The dimensions with big difference on mean and variance will have a decisive impact on the clustering results of data. Therefore, it is very important to normalization the data (specifically the characteristics of each dimension) before clustering. There are two types of common normalization methods, which are for data with big difference on mean and variance respectively [12].

**Min-Max normalization**

This method generally maps the data to a specified range to remove the influence of dimension and unit, which is usually for data with big difference on mean and specifically is formulated as,

\[
x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}
\]

(2)

According to this method, the data of each dimension is mapped to interval \([0,1]\), where \( x'_i = 0 \) when \( x_i = x_{\min} \) and \( x'_i = 1 \) when \( x_i = x_{\max} \).

**Z-score normalization**
The normalization method prefers to the case with data having big difference on variance, which is modelled as

\[ x'_i = (x_i - \mu) / \sigma \]  

where \( \mu \) is the mean value and \( \sigma \) is the standard deviation value of the data.

### 3. Case study: Classification of Members in A Shopping Mall

Here we use the proposed classification algorithm to figure out the members classes in a shopping mall based on classical RFM model. RFM model is commonly used for member clustering, the principle of which is to classify elements with last consumption time (Recency), consumption frequency (Frequency), and consumption amount (Monetary) as indicators. Therefore, we take Recency, Frequency and Monetary as the classification index by applying RFM model in this paper.

Firstly, we pre-process the data to remove the default data. Through statistical analysis, there are great differences in the mean and variance of the data corresponding to the three indicators, so the data need to be normalized before classification. After that, the number of classifications needs to be determined, therefore, the value of loss function \( F(c,u) \) with different classification numbers (elbow figure) is generated (see Figure 2).

![Figure 1. Value of loss function \( F(c,u) \) with different classification numbers.](image)

When the cluster number \( K \) is less than the real number of classifications, the increase of \( K \) will greatly increase the aggregation degree of each cluster, so the decrease of \( F(c,u) \) will be large. When \( K \) reaches the real number of clusters, the return of aggregation degree obtained by increasing \( K \) will decrease rapidly, so the decrease of \( F(c,u) \) will decrease sharply, and then tend to be flat with the continuous increase of \( K \), In other words, the figure of \( F(c,u) \) and \( K \) is the shape of an elbow, and the \( K \) value corresponding to this elbow is the real cluster number of the data. As shown in Figure 2, \( K \) shall be taken as 5. Therefore, next, the shopping mall members will be divided into five classes.

- Classification results with Min-Max normalization

Firstly, we processing the data by using the Min-Max normalization method, and then obtain the classification results as shown in Figure 3 and the tables (Table 3 and Table 4). Figure 3 shows the class to which each member belongs. And Table 3 lists the center nodes of each class (the values of the nodes are transformed by Min-Max normalization). From the classification results shown in Figure 3 and table 3, it can be seen that variable Recency plays a major role in the classification process, while the other two variables Monetary and Frequency have little impact on the classification results, which can be verified by the analysis of variance (see Table 4). In Table 4, \( F \) is the ratio of the sum of squares of deviations between and within groups to the degree of freedom. The greater its value, the greater the proportion of variance between groups, that is, the more significant the corresponding variables have on the classification results. As is shown in the table, variable Frequency has a much
larger $F$ compared with the other two. Consequently, the classification with Min-Max normalization is not appropriate for the member data of the shopping mall.

![Figure 3. Classification results with Min-Max normalization.](image)

Table 3. Center nodes of Classification results with Min-Max normalization.

| Variable | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|----------|-----------|-----------|-----------|-----------|-----------|
| Recency  | 973.5     | 309.7     | 573.6     | 80.8      | 35.6      |
| Frequency| 2.6       | 3.2       | 2.7       | 4.0       | 23.7      |
| Monetary | 6650      | 9360      | 6871      | 11685     | 104856    |

Table 4. Analysis of variance for the classification results with Min-Max normalization.

| Variable | $F$     |
|----------|---------|
| Recency  | 10286   |
| Frequency| 605386  |
| Monetary | 23279   |

• Classification results with Z-score normalization

Then, a classification with Z-score normalization method is carried out. And the results of the classification are shown in Figure 4, Table 5 and 6. Though the $F$ value of Frequency is still the largest, the gap with the other two is no longer obvious, which means that the classification with Z-score normalization method is effective (see Table 6). From Figure 4, we can clearly see to which class each member belongs. Furthermore, Table 5 gives the center nodes of each class. As shown in Figure 4 and Table 5, there are five classes of the members: 1) Class 1 are the members with the highest Monetary and Frequency but lowest value of Recency, which is named premium member. This kind of members do not need special marketing means to stimulate consumption, but only need to recommend some products according to their needs. 2) Members in class 4 with the second higher Monetary, Frequency and the second lower Recency are called key member. 3)Class 5 are formed by members with a medium level of the three variables, which here we call normal member. In order to stimulate the consumption of these two types of members, shopping malls can take some small preferential activities and publicity such as point exchange to improve their consumption level. 4) because of the lower Monetary, Frequency and the higher Recency, class 2 is identified as inactive member. 5) class 3 is regarded as dead member due to the lowest Monetary, Frequency and the highest Recency. As shown in the table, such kind of members have not consumed in the mall for more
than two years. Because the cost of activating such members is high and the return is small, activating such members is of little significance.

Figure 4. Classification results with Z-score normalization.

Table 5. Center nodes of Classification results with Z-score normalization.

| Cluster | 1   | 2   | 3   | 4   | 5   |
|---------|-----|-----|-----|-----|-----|
| Recency | 41  | 226 | 885 | 112 | 153 |
| Frequency | 54.4 | 2.5 | 2.2 | 28.3 | 12.4 |
| Monetary | 523351 | 6157 | 5276 | 142499 | 39437 |

4. Conclusion
With the purpose of improving the membership management of the shopping malls, this paper designed a classification algorithm based on the RFM model and k-means method. By applying the classification algorithm to the case of a shopping mall, the validity of the proposed method is verified. The results show that the members of the mall are divided into 5 classes according to three variable indicators, which are called premium member, key member, normal member, inactive member and dead member respectively. Membership classification can provide theoretical support for managers to formulate targeted membership management strategies. For example, the members most in need of promotion are the normal members who have the moderate consumption level in the mall. In contrast, premium members with the highest level of consumption do not need to spend too much effort on management. For inactive members, a series of promotional offers are required to activate them. In addition, removing “dead members” while developing new members.

Furthermore, the application of the algorithm proposed in this paper is not limited to the classification of shopping mall members, and can also be used in other cases. For example, this classification method can be used for mental health assessment of college students. Moreover, analysis of graduate employment can also be carried out by using this classification algorithm.
Acknowledgments
This research is supported by Foundation Sciences of Beijing Polytechnic under grant 2020Z109-KXZ.

References
[1] Guan J, Tang C, Ou J 2019 The portrait depiction of the market members based on data mining International Journal of Pattern Recognition and Artificial Intelligence 34 p 2059024
[2] Deng M 2019 Analysis of "member portraying for a large department store" International Mathematical Modeling and Its Application 8 pp 66-70
[3] Adeniyi DA, Wei Z, Yang Y 2016 Automated web usage data mining and recommendation system using k-nearest neighbor (KNN) classification method Applied Computing and Informatics 12 pp 90-108
[4] Dong X, Qian L, Guan Y, Huang L, Yu Q, Yang J 2016 A multiclass classification method based on deep learning for named entity recognition in electronic medical records New York Scientific Data Summit (NYSDS) pp 1-10
[5] Wang A, Xue D, Wu H, Iwahori Y 2020 LiDAR data classification based on improved conditional generative adversarial networks IEEE Access 8 pp 209674-209686
[6] Shahri A, Hosseini M, Phalp K, Taylor J, Ali R 2019 How to engineer gamification Journal of Organizational and End User Computing 31 pp 39 – 60
[7] Namasudra S, Roy P 2018 PpBAC: popularity based access control model for cloud computing Journal of Organizational and End User Computing 30 pp 14 – 31
[8] Wu J, Shi L, Lin W, Tsai SB, Xu G 2020 An empirical study on customer segmentation by purchase behaviors using a RFM model and k-means algorithm Mathematical Problems in Engineering 6 pp 1-7
[9] Lu Z, Pang X, Ozolins O, Udalcovs A, Chen J 2018 K-means clustering based multi-dimensional quantization scheme for digital mobile fronthaul Optical Fiber Communication Conference pp 1 – 3
[10] Zhang P, Cao W, Li W 2020 Surface and high-altitude combined rainfall forecasting using convolutional neural network Peer-to-Peer Networking and Applications pp 1-13
[11] Li H and Wang S 2019 Modeling and verification of mall member portraits based on RFM model Journal of Zhengzhou Railway Vocational and technical College 3 pp 14-16
[12] Henderi H 2021 Comparison of min-max normalization and z-score normalization in the K-nearest neighbor (KNN) algorithm to test the accuracy of types of breast cancer International Journal of Informatics and Information Systems 4 pp 13-20