Data Analysis and Application of Retail Enterprises Based on Knime

Xiaochuan Pu¹,², Ning Qi¹,³, Junli Huang¹,²
¹Graduate School of Management of Technology Pukyong National University, Busan, South Korea, 48513, China
²Zunyi Normal University, Zunyi, 563002, China
³Hexi university Zhangye, 734000, China

Abstract. With the rapid development of information technology, the application of cross-discipline has shown explosive growth. Data mining, big data and other technologies are rapidly entering all walks of life. Research on data mining technology and its application in data analysis of related enterprises have certain theoretical significance and practical value. This paper starts with a cross-border sales data set and uses Knime visual data analysis tool to illustrate that the widespread application of visual data mining technology in small and medium-sized enterprises is a feasible way to improve the decision-making ability of enterprises.

1. Introduction
With the continuous growth of the Internet economy, China's consumption means are constantly changing, and more and more urban and rural residents are shopping online. Therefore, in recent years, the competition among online retail enterprises in China has become more and more fierce, and business operators pay more attention to valuable information such as sales performance. At present, all online sales enterprises use computers to manage their own internal affairs, but most of them only record and inquire about their own sales transaction data, without paying attention to the potential value of their long-term accumulated transaction data. Even some businesses delete the past transaction data because of limited storage space. They wake up only after discovering the potential value of these data.

Online retail enterprises gather a large amount of commodity and customer data in their daily operations, which are also the core "assets" of enterprises. How to further explore the information transmitted by these data and make in-depth analysis to form valuable "intelligence" for the operation of enterprises has always been a problem for most retail enterprises. How to make data serve enterprises and make enterprises understand consumers better requires data mining and business intelligence technology, which provides new thinking and technical solutions to these problems.

2. Introduction to Knime tool and K-means algorithm

2.1. Introduction to Knime Tools
The first version of KNIME was released in 2006. many pharmaceutical companies began to use KNIME and many life science software vendors began to integrate their tools into KNIME. Later that
year, after an article was published in a german magazine, users from many other fields joined KNIME. As of 2012, KNIME has been used by more than 15,000 actual users and has been applied to banks, distributors, automobile manufacturers, telecommunications companies, consulting companies and various other industries, with research teams all over the world. The latest updates to the KNIME server and KNIME big data extensions provide support for ApacheSpark2.3, Parquet, and HDFS type storage.

KNIME is a workflow platform based on GUI. The analysis platform provides an easy-to-use graphical interface and can simply drag and drop various pre-built data processing/machine learning modules without writing any code. Many pre-built functions and modules. KNIME tools provide you with basic modules, such as I/O, data processing, data conversion, data visualization, and machine learning models (from linear regression to advanced depth learning). You can simply link these modules together in the workflow and provide specific instructions to fine-tune the model.

KNIME supports Python and r through wrapper scripts. This means that you can merge your own code into the workflow or customize some procedures in the workflow when needed.

KNIME has been named the leader of data science and machine learning platform for the sixth consecutive year. Now KNIME is divided into three systems:

1. KNIME Analysis Platform
   KNIME Analytics Platform is an open source software for creating data science. KNIME is intuitive, open and integrates new development results continuously, so that everyone can understand data and design data scientific workflow and reusable components.

2. KNIME Server
   KNIMEServer is enterprise software for team-based collaboration, automation, management, and deployment of data science workflows as analytical applications and services. Non-experts can access data science through KNIMEWebPortal or use RESTAPI.

3. Extension of KNIME Analytics platform
   Use KNIMEAnalyticsPlatform extension to process more data. Some were developed and maintained by us in KNIME, while others were developed and maintained by the community and our trusted partners. We have also integrated with some very cool open source projects.

This paper chooses to use KNIME's analysis platform to help small and medium-sized enterprises complete data mining analysis independently.

2.2. Basic Idea and Algorithm Flow

K-means algorithm is a clustering analysis method based on partition. In data mining technology, it is one of the most classical algorithms used in intrusion detection. The algorithm itself is simple in thought, easy to implement in intrusion detection and low in complexity. The basic idea of k-means algorithm is to divide N data into K classes, and the final clustering should satisfy that each data is divided into an appropriate class. For these classifications, the data divided into the same classification should be as similar as possible, while the data divided into different classifications should be as different as possible. The k-means algorithm divides the data into categories. Its algorithm flow is as follows: At the beginning, the algorithm will randomly select K points from the data set D as the initial clustering center, where the K value needs to be input in advance by the user. Then, the algorithm will calculate the similarity between the data according to the previously adopted method of calculating similarity. Generally, the distance is used to judge the similarity, and each data is divided into the classification with the largest similarity. After the first clustering of the data is completed, the center of each cluster is adjusted, and then the classification steps are repeated until the clustering process ends when the clustering centers before and after clustering are consistent. At this time, the criterion function of the algorithm has converged and the algorithm is finished [9].
2.3. Clustering Step

Suppose there is currently a data set D to be clustered, in which there are N data objects, and each data object contains Q characteristic attributes. That is \( D_i = (D_{i1}, D_{i2}, D_{i3}, ..., D_{iq}) \), \( D_i \in D, 1 \leq i \leq N \leq n \).

The number of clusters required is k, and k \(<\ n\).

(1) At the beginning of clustering, K initial clustering centers need to be selected first. The first K data objects can be directly selected, or determined by a random function, and finally K data objects are selected as initial clustering centers;

(2) For each data object \( D_i \), calculate its Euclidean distance to each cluster center, and then compare the sizes of these distance values to divide \( D_i \) into clusters with the smallest distance. At this time, k clusters \( C_1, C_2, C_3, ..., C_k \) are obtained;

(3) adjusting the clustering center of the newly generated classification, namely calculating the centroid of each class and taking them as initial clustering centers for re-clustering;

(4) comparing the clustering center with the previous clustering center, and executing step (5) when the clustering center is the same as the previous clustering center, otherwise executing step (2).

(5) outputting and saving the clustering result to a database [9].

3. Application of Visual Data Mining in Retail Data Analysis

3.1. Data Acquisition

The data source of this experiment is UCI Machine Learning Repository. The data set contains 250,000 transaction order data from an online retailer in the UK from December 1, 2012 to December 1, 2013. The company mainly sells daily necessities and gifts. Many customers are wholesalers. They read in the node and data through KNIME data. The expanded structure is shown in Fig. 1.

![Fig. 1 Data set](image)

3.2. Optimal Clustering Number and Customer Value Modeling

We will consider from this perspective: the customer's purpose and requirements—which aspects are demonstrated—which data to reflect. Only when you know what data you need, can you choose the right data source, can you guide the next step of data preprocessing, and can you get the potential value. When we observe the data set, it is the customer's consumption data, then customer segmentation and customer value are the direction of the data set mining.

Customer segmentation refers to the method of selecting certain segmentation variables and classifying customer values according to certain classification standards. The most representative customer segmentation models based on customer consumption behavior are RFM model and customer value matrix model [6].
Table 1. Customer value model

| Name                        | Type                      |          |          |          |
|-----------------------------|---------------------------|----------|----------|----------|
| Key value customers         | Consumption interval ↑    | Consumption frequency ↑ | Total consumption ↑ |          |
| Key development customers   | Consumption interval ↑    | Consumption frequency ↓ | Total consumption ↑ |          |
| General value customers     | Consumption interval ↑    | Consumption frequency ↑ | Total consumption ↑ |          |
| General development         | Consumption interval ↑    | Consumption frequency ↓ | Total consumption ↓ |          |

RFM model: the consumption time interval r: r value refers to the difference between the latest purchase time of the customer and the current time, the larger the difference, the lower the customer value. Consumption Frequency F: refers to the total number of times retail customers consume in a limited period of time. Total Consumption M: refers to the total consumption of customers in a certain period of time. Since we only want to obtain the valuable customer class of the enterprise, we also use these indexes to do clustering research. According to the suggestions of the enterprise's sales experience and many iterations, we get the optimal clustering K value of 4. Considering the comprehensive practical factors, each clustering customer has its special performance characteristics. According to the characteristics, we can temporarily divide them into key value customers, key development customers, general value customers and general development customers, as shown in Fig. 2. Here we use Python tools to calculate R, F and M values according to the customer value model.

3.3. Data Preprocessing
1) data cleaning: according to the data requirements, the Missing Value node in the knime tool is used to clean the values, and missing values are found in the data. If the customer number is "?" Generally speaking, they are non-member customers. They can be replaced (for example, using specific symbols, etc.) or columns exceeding a certain missing value can be deleted.
2) Attribute Specification: There are many attributes in the original data, and attributes that are not related or weakly related to them will be deleted.
3) Data exploration and analysis: statistics of empty records and some simple statistical values: maximum and minimum. According to the customer value model, the values of r, f and m are calculated.

Table 2. Partial results of RFM values

| CustomerID | R    | F   | M    |
|------------|------|-----|------|
| 13705      | 207.0208 | 103 | 318.14 |
| 13706      | 136.0681 | 189 | 330.79 |
| 13707      | 100.7056 | 351 | 806.41 |
| 13708      | 104.884  | 306 | 296.88 |
| 13709      | 2.065972 | 930 | 1539.08 |

3.4. Parsing Process
The customer value analysis process is mainly composed of two parts:
1) The first part is clustering and grouping according to customer value model.
2) The second part is to analyze the characteristics of each customer group in combination with business and formulate corresponding marketing strategies.
Import the above data into KNIME software, use the K-means clustering function, set the maximum iteration number to 99 and the classification number to 4, and run the program. The specific process is shown in Fig. 3. Import the data in the accident data statistical table (Table 1) and run the program to obtain the final clustering result, as shown in Fig. 4.
2) Customer Value Analysis

Individual consumption of users has certain regularity, with most users spending less than 2000 pounds. Consumer spending reflects the 2/8 rule, with the top 30% of consumers contributing 80% of their spending. Therefore, it is an unchangeable principle to pay close attention to high-quality customers. These high-quality customers are all "members" and need to optimize the shopping experience for members, such as special offers, etc. The clustering results are listed in Table 2, and the clustering results are visually represented as shown in Fig. 6. The classification results show that among the four types of users, there are 617 Cluster 2 users. In a year, their average consumption is the highest, and the number of purchases is relatively frequent. The purchase interval is about 15 days. They belong to high-value users. There are 4310 users in Cluster 3, 9721 in Cluster 1 and 12092 in Cluster 0, which are general development customers.

4. Case Inspiration and Conclusion

On the basis of in-depth analysis of RFM model, this paper is based on data obtained from model indexes. Through Knime tool and K-means algorithm, a customer segmentation method based on Knime is proposed, which effectively screens out high-quality customers and is successfully applied to an online sales enterprise. The experimental results show that the method can fully reflect the current value and value-added potential of customers and provide scientific basis for enterprises to treat customers differently.

The application of data mining and big data technology has always been a research hotspot. However, the current application technology has a high threshold, and the potential commercial value of the data can be extracted in detail and accurately only if there is a corresponding information technology background and the working background of the industry in which the data are analyzed. It is hoped that with the maturity of technology application, more data mining business cases can be developed, and more profits can be won for enterprises to form a benign business model.

References

[1] Yan Bin, Guo Wenchuan. Detection method of 'Hayward' kiwi swollen fruit based on K-means clustering and fruit capsule shape [J / OL]. Journal of Northwest A & F University (Natural Science Edition), 2020 (05) : 1-8 [2019-11-30].
[2] Wang Luyao, Gaoshan, Li Jun, Fan Hong. Analysis of retail sales behavior based on K-means clustering and spatial correlation [J]. Bulletin of Surveying and Mapping, 2019 (09): 51-54.
[3] Tong Qiang. Value analysis of large department store members based on RFM model [J]. Journal of Lanzhou Petrochemical College of Technology, 2019, 19 (02): 16-18.
[4] Zhang Kuan. Application Research of Customer Churn Early Warning Technology Based on TFM Model for Automobile 4S Stores [D]. Lanzhou University of Technology, 2019.

[5] Bao Zhiqiang, Zhao Yuanyuan, Zhao Yan, Hu Xiaotian, Gao Fan. Baidu takeaway customer segmentation based on RFA model and cluster analysis [J]. Computer Science, 2018, 45 (S2): 436-438.

[6] Liu Zhiyi, Chen Gong. Research on RFAT customer segmentation based on improved K-means algorithm [J]. Journal of Nanjing University of Science and Technology, 2014, 38 (04): 531-536.

[7] Wang Jikui, Li Hong. Examples of social network analysis based on KINME platform [J]. Journal of Jilin Institute of Technology and Technology, 2013, 29 (12): 91-93.

[8] Peng Hu, Tian Junfeng, Yu Mali. Application of Data Mining Technology in Cost Analysis of Retail Enterprises [J]. Science and Technology Information (Science Education and Research), 2007 (32): 83-84.

[9] Zhao Sen, Wei Mingjun. Research on intrusion detection based on k-means algorithm [J]. Journal of Hebei Energy Vocational and Technical College, 2019, 19 (02): 66-69.