Volume: 11 Issue: 05 Pages: 4393-4398(2020) ISSN: 0975-0290

Intelligent Recommendation System Based on K-means Clustering Algorithm

School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China

Tang Zhi-hang

Email: zhtang@hnie.edu.cn

Guo Tao

Email: 2206159759@qq.com

Li Jun

Email:642615662@qq.com

Wu Shi-qi

Email:1305056857@qq.com

----------**ABSTRACT**----------

Use python web crawler to collect data from Trade website. The collected data is down jacket information. The fields are shell material, structure type, filling material, process information and style information. This information can be used for data mining, using clustering algorithms, correlation algorithms, etc. to identify potential value, providing decision-making reference for the management of textile and garment enterprises, with strong practical value. This paper provides a new idea for the development of textile and garment enterprises. The employees of the company screen, deal with the missing data and standardize the data, and then conduct data mining. The management of the enterprise makes decisions based on the results of data mining to improve decision-making basis and correctness.

Keywords: **K-means clustering algorithm, decision-making, Intelligent Recommendation System**

Date of Submission: Apr 27, 2020 Date of Acceptance: May 08, 2020

I. INTRODUCTION

With the advent of the era of big data and the maturity of data mining technology, more and more Dead data is being fully utilized. This is especially true in the field of textiles and clothing. This article uses clothing style data as the research target for analysis. Clothing style data is an important reference indicator for the production of clothing by textile and apparel companies. With the increasing diversity of clothing styles, the amount of pattern data is constantly increasing, and traditional data analysis methods can no longer meet the current needs of clothing version style analysis. In response to this problem, this article will use big data analysis methods to analyze the relevant data of clothing styles. The overall analysis process is divided into three steps: the acquisition of clothing source data, the use of K-means clustering algorithm.

The first is the acquisition of clothing source data. The author will use Python to write a web crawler program to crawl clothing style data and related review information from Jingdong Mall. The crawled attributes mainly include commodity number, style, pattern, collar type, sleeve type, number of favorable comments and total number of comments, which will be stored in the MongoDB.

Second is the application of K-means clustering algorithm. The author uses Python to write programs to pre-process the data. The pre-processing operations include missing value processing, outlier processing, standardization processing, and quantization processing. Through the preparation of K-means clustering algorithm,

Volume: 11 Issue: 05 Pages: 4393-4398(2020) ISSN: 0975-0290

the data after preprocessing is analyzed, and different K-values are selected for many times. The best one is selected from different clustering results, and the clothing pattern characteristics of each category are analyzed. By this, the author finds out the relationship between the combinations of garment pattern and style, analyzes the reasons and summarizes them.

The K-means clustering analysis for clothing pattern data collected by web crawler can help the management of textile and clothing enterprises and sellers to make business decisions, reduce the decision-making error rate of decision-makers, and enable decision-makers to grasp the market information in time and make reasonable adjustments.

II. K-MEANS CLUSTERING ALGORITHM

Cluster analysis is an exploratory and unsupervised method of data analysis. Cluster analysis can classify data so that similar data is classified into the same category, and data with low similarity is classified into different categories [1]. However, the method of judging the similarity is to calculate the distance between two data elements, the similarity of the distance is high, and the similarity of the distance is low. The distance between data elements mainly includes Euclidean distance, Mahala Nobis distance, etc. The K-means cluster used in this paper is to calculate the similarity by calculating the Euclidean distance. Now every data item is seen. Make a coordinate point, With $x = (x_1,...,x_n)$ and $y = (y_1,...,y_n)$, the Euclidean distance formula is:

$$d(x,y) := \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

(1)

The clustering analysis algorithm is mainly divided into density-based clustering algorithm, grid-based algorithm, model-based algorithm, hierarchical clustering algorithm, and partitioning clustering algorithm [2].

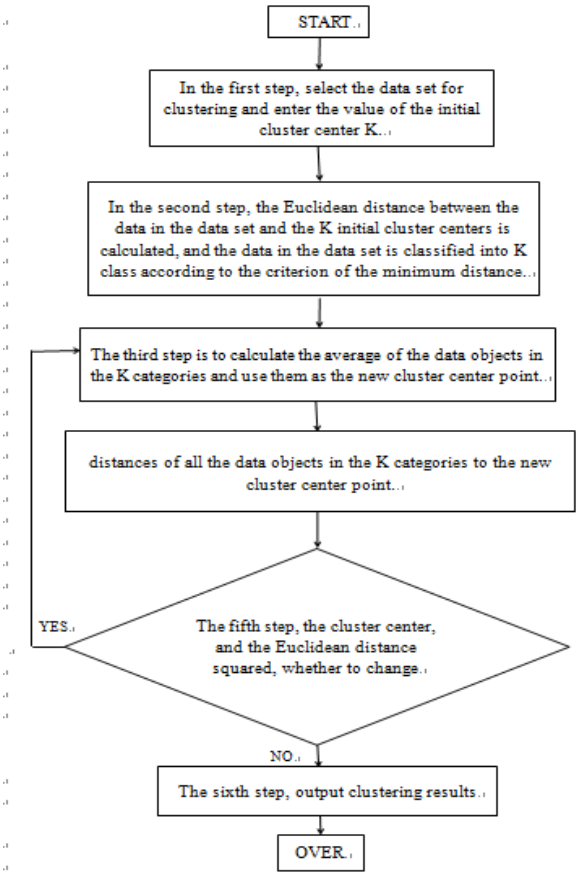
K-means clustering algorithm belongs to the clustering algorithm in the clustering algorithm. It is a classical clustering algorithm and has applications in many fields such as biology, medical, marketing, and image [3].

The K-means clustering algorithm was first proposed by J.B.MacQueen in 1967. The specific idea of the algorithm is: in a preprocessed data set, select k data points as the cluster center point of the k-means algorithm, and then calculate this The Euclidean distance of all the data points in the data set to the previously selected k data points, and the similarity between the two data points is judged by the magnitude of the Euclidean distance, and the similarity of the large distance is low, and vice versa. The distance is small and the similarity is high. Here, the distance from each data point to the center point of the k clusters is compared. Which point is the closest to the cluster point of the data point, which is in which cluster, so after returning to the class the new cluster generated will also be calculated, updating the "cluster center". The K-means clustering algorithm has three key points. The first is the selection of the k-value of the cluster center point, and the second is the calculation of the Euclidean distance of all the data points to the previously selected k data points. It is a newly-divided cluster, and the "cluster center" of each cluster is recalculated [4].

The K-means clustering algorithm is applied to the field of textiles and garments [5]. Firstly, the pre-processing operations of the collected garment style data are carried out. The specific data pre-processing operations have missing values, quantization, and standardization. Then use Python to write the K-means clustering algorithm [6], select the attribute column for clustering, and set the K value of the cluster. After modifying the K value multiple times, select the best clustering effect as the final K value of the cluster. Finally, the results of the classification are expanded in the form of a visual chart, and the data analysis operation is performed according to the result of the classification [7].

Volume: 11 Issue: 05 Pages: 4393-4398(2020) ISSN: 0975-0290

Fig1. The flow chart of the K-means clustering algorithm



III. APPLICATION OF K-MEANS CLUSTERING ALGORITHM APPLICATION OF K-MEANS CLUSTERING ALGORITHM

3.1 Source of clothing version style data

This paper uses Python language to program and imports K-Means clustering from sklearn.cluster machine learning clustering package in Python to implement the K-means clustering algorithm [8].

The clothing version style data used in this article is crawled in JingDong Mall by using a Python web crawler. The clothing fields crawled mainly includes product number, style, material, sleeve length, type, collar, and sleeve type [9]. The amount of data is 20,000.

Table1. The part of collected data

product number	style	version	collar	sleeve type	comments	praise
35100099252	2	9	15	5	200	200
37219234639	2	7	15	5	2	2
36497353859	2	9	9	5	100	100
37209168083	1	7	15	3	400	400
32954113259	2	7	15	5	2100	2100
35736556950	2	7	15	5	900	900
17256439942	2	7	15	5	2500	2500
37064854882	1	8	15	1	100	100
34106321994	2	7	15	5	900	900
34050609608	2	7	14	5	1700	1700
34103811389	2	7	15	5	1000	900
21138549700	2	7	14	5	600	600
33121227356	2	7	9	5	700	700
34700032175	1	8	15	1	600	600
33484899010	2	7	14	5	600	500
36084683597	2	7	3	5	300	300

Volume: 11 Issue: 05 Pages: 4393-4398(2020) ISSN: 0975-0290

37075586383	2	7	14	5	400	400
35068406425	2	7	14	5	700	700
35470388912	6	9	4	1	900	900
33267078531	2	9	9	5	20	20
25020816291	2	7	14	5	200	200
32965482396	6	7	4	10	1700	1700
37471613932	2	9	9	5	200	200
37204538391	6	7	3	1	900	900
35487748509	6	9	19	1	900	900

3.2 Data preprocessing

We can see that the data from Figure. 1 needs further pre-processing to be used in K-means cluster analysis. First, the missing value processing is performed, the missing data items are deleted, and then the processing is repeated to remove the duplicates in the data. Delete the duplicated items, remove the two materials that affect the size and size of the clothing version, and then quantify the data, and replace the subordinates of the style, type, collar and sleeve with numbers. Achieve quantitative results. After the quantization is completed, the results are shown in Table2.

Table2.Quantitative result

	style	version	leader	sleeve type
0	2	9	15	5
1	2	7	15	5
2	2	9	9	5
3	1	7	15	3
4	2	7	15	5
5	2	7	15	5
6	2	7	15	5
7	1	8	15	1
8	2	7	15	5
9	2	7	14	5
10	2	7	15	5
11	2	7	14	5
12	2	7	9	5
13	1	8	15	1
14	2	7	14	5
15	2	7	3	5
16	2	7	14	5
17	2	7	14	5
18	6	9	4	1
19	2	9	9	5
20	2	7	14	5
21	6	7	4	10
22	2	9	9	5

Volume: 11 Issue: 05 Pages: 4393-4398(2020) ISSN: 0975-0290

3.3 Program Implementation of K-means Clustering

Open the pycharm editor, create the k-means.py file, import numpy, pandas, matplotlib and other packages, use pd.read_excel to import the pre-processed data, select 'style' through iloc[:,1:5], Type ', 'collar', 'sleeve' four columns, then import KMeans cluster from sklearn.cluster machine learning clustering package, set K-means clustering K value by kms = KMeans(n_clusters = 5) 5, the clothing version of the style data is divided into 5 categories, and then clustered by kms.fit_predict (), after the completion of the cluster, a two-dimensional table is formed by pd.DataFrame and output to excel, and finally, the product is the abscissa, the category is the ordinate, draw a k-means cluster scatter plot. As shown in Figure 2.

In the scatter plot, you can see that category 2 is the most intensive, and other categories are relatively loose. Use python's matplotlib package to visualize the data after clustering, and display it in the form of a pie chart. As shown in Figure 3. In the pie chart, the maximum proportion of the two types of distribution is 60.36%, the type 0 and the class 1 are 11.79%, class 3 is 5.65%, and class 4 is 10.41%.

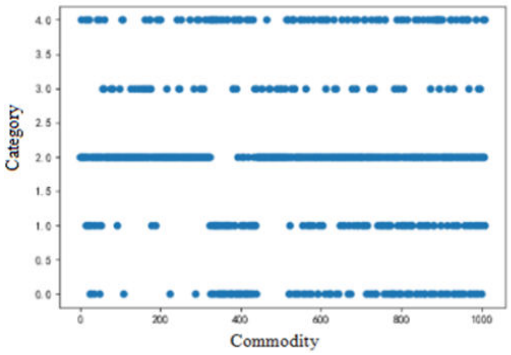


Fig2.K-means clustering map

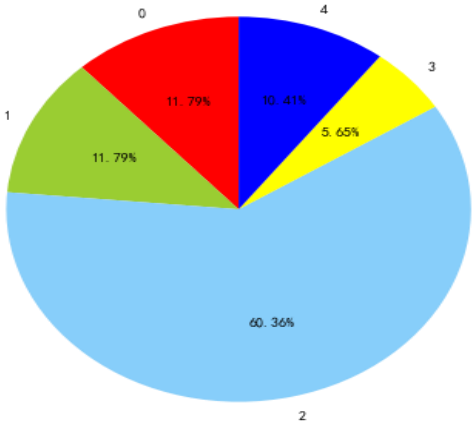


Fig3. Cluster grouping map

3.4 Analysis of K-means clustering results

According to the results of K-means clustering, it can be observed that in the five groups of grouping graphs, the second group has the highest data volume ratio of 60.36%, and the clustering results are output to the table file by using Python programming. In the observation, the second group of clothing data can be found. The styles are mostly regular and cardigan models. The models are mostly loose, slim and straight. The collar is mainly hooded, and the sleeves are mostly sleeves and Regular sleeves, which also shows that the combination of this type of style is the most used, in the process of the company's production of clothing can increase the combined production of these types of styles, the third group of data volume accounted for a minimum of 5.65 %, the styles are mainly regular models, the models are mainly loose and slim, the collar is mainly hooded, the sleeves are mainly closed sleeves and other types, which shows that the combination of this type of style is used. At the very least, the combined production of these types of styles can be reduced during the company's production of garments. The grouping results for the zero groups, the first group and the fourth group can also be analyzed according to the above analysis method, and the combination of the type of the pattern with a large amount of grouped data should be incrementally produced, and the version type with less packet data is used. The combination should be reduced in production. However, after using the clustering algorithm for classification, it still needs to be improved. In the

Volume: 11 Issue: 05 Pages: 4393-4398(2020) ISSN: 0975-0290

selection of attributes, several attributes can be added as appropriate, such as scoring, price, etc. Analysis of the results can improve the quality of the analysis the results.

IV. CONCLUSION

This paper classifies clothing data by k-means clustering algorithm, and then analyzes which type of clothing type combination is the most valuable, predicting Which styles are combined will be more popular with customers, providing decision-making reference for the management of textile and apparel companies to help predict the direction of the apparel market. However, this paper is not perfect enough. It can also be optimized on the k-means clustering algorithm. Overall, the data analysis method of this paper provides a new idea for textile and apparel enterprises. With the rapid development of big data, more big data analysis methods will be used in the future to solve some difficult problems.

ACKNOWLEDGEMENTS

Project supported by Provincial Natural Science Foundation of Hunan（2018JJ4047）

REFERENCES

[1]. He Xiao. BIRCH algorithm and data management in financial enterprises based on dynamic panel GMM test [J]. Cluster Computing, 2019, 22(2): 4231-4237.

[2]. Mohamed Saied, Mona Nasr. Blended Learning Model Supported By Recommender System And Up to-Date Technologies [J]. International Journal of Advanced Networking and Applications,2019,10(5):3829-3832

[3]. Shen Hong, Liu Shun. Application Research of Data Analysis Model Based on K-means Clustering Algorithm [J]. Software Guide, 2017, 16(3):103-107.

[4]. Yu Qilin.Optimization of initial clustering center selection based on K-means algorithm [J]. Computer System Application, 2017, 26(5):170-174.

[5]. Mattias Thuvander,Frederick Meisenkothen,Gang Sha. The Application of the OPTICS Algorithm to Cluster Analysis in Atom Probe Tomography Data [J].Microscopy and Microanalysis, 2019, 25(2):1-11.

[6]. Reddy D,Jana P K. Initialization for k-means Clustering using Voronoi Diagram[J].Procedia Technology, 2012(4):395-400.

[7]. Soroor Sarafrazi,Hossein Nezamabadi-pour,Saeid R. Seydnejad. A novel hybrid algorithm of GSA with Kepler algorithm for numerical optimization [J]. Journal of King Saud University - Computer and Information Sciences, 2015, 27(3):288-296.

[8]. Lv Jia, Cheng Dongsheng.Clothing Sensitive Data Mining Method Based on Clustering Algorithm [J]. Journal of Textile Research, 2014, 35(5):108-112

[9]. LIN Peng,WANG Yinghui,QI Hongsheng,HONG Yiguang.Distributed Consensus-Based K-Means Algorithm in Switching Multi-Agent Networks[J].Journal of Systems Science & Complexity,2018,31(05):1128-1145