Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2019, 6(11):151-154**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Application of Clustering Method Based on Python in the Evaluation of Local University Science and Technology Innovation

Han Zhang

School of Mathematics, Nanjing Normal University, Taizhou College, Taizhou, China

Abstract This paper mainly uses Python as the programming language and machine learning library Scikit-Lear as a tool to classify the regions of local colleges and universities for scientific and technological innovation evaluation, comparing the results of K-means cluster analysis with the results of factor analysis, it is found that the degree of fit is very high.

Keywords Python, K-means cluster, factor analysis

Introduction

Cluster analysis is an important topic in data mining. This paper uses Scikit-Lear, a machine learning library in Python, to effectively analyze the scientific and technological innovation data of local universities. In 2017, the author used the factor analysis in SPSS to study the evaluation system of science and technology innovation in local universities. In order to compare the results of cluster analysis with the results of factor analysis, the data in this paper is the same as the data used in factor analysis at that time. It isfrom the Science and Technology Department of the Ministry of Education of the People's Republic of China, "Collection of Science and Technology Statistics of Colleges and Universities in 2015".

First, the K-Means clustering algorithm idea

Clustering centered on k points in space, classifying the objects closest to them, and successively updating the values of each cluster center by iterative method until the best clustering result is obtained.

Algorithm flow summary:

1. Appropriately select the initial center of k classes
2. For any sample, find the distance to each center and classify the sample to the class with the shortest center.
3. Update the center value of each cluster by some means such as mean.
4. Repeat the iterations of Process 2 and Process 3 above until the k center point values remain unchanged, then the iteration ends, otherwise continue the iteration

The K-means clustering algorithm (K-average/K-means algorithm) is the most classical and widely used distance-based clustering algorithm. The distance-based clustering algorithm refers to the use of distance as the evaluation index of similarity measure. That is to say, when the two objects are close together, the distance between the two is relatively small, then the similarity between them is relatively large.

Common similarity/distance evaluation criteria are:

- Euclidean distance

The meaning is the set distance of two elements in Euclidean space, because it is intuitive and understandable, and is widely used to identify the dissimilarity of two scalar elements.

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}$$

- Manhattan distance

$$d(X,Y)=|x_1-y_1|+|x_2-y_2|+\cdots+|x_n-y_n|$$

- Minkowski distance

$$d(X,Y)=\sqrt[p]{|x_1-y_1|^p+|x_2-y_2|^p+\cdots+|x_n-y_n|^p}$$

Clustering performance evaluation criteria

The squared error sum criterion function formula is: $E=\sum_{i=1}^{k}\sum_{p\in X_i}\|p-m_i\|^2$.

The K-means clustering algorithm usually uses the error squared criterion function to evaluate the clustering performance. Given data set X , which contains only description attributes, not category attributes. Hypothesis X contain K Clustering subset $X_1, X_2, \cdots, X_k$ The number of samples in each cluster subset is $n_1, n_2, \cdots, n_k$.

The mean representative points (also called cluster centers) of each cluster subset are $m_1, m_2, \cdots, m_k$. The main idea of the algorithm is to divide the data set into different categories through an iterative process, so that the criterion function for evaluating the clustering performance is optimal, so that each cluster (also called cluster) generated is compact and independent.

Second, K-Means clustering code example implemented in Python

```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.metrics import homogeneity_score, completeness_score, v_measure_score
Data = pd.read_excel('indicator.xls', header=None,
Names=['C1', 'C2', 'C3', 'C4', 'C5', ' C6', 'C7',' C8',' C9', 'C10',
                'C11', ' C12', ' C13', 'C14', 'C15', 'C16', 'C17', 'category'])
x = data[['C1', 'C2', 'C3', 'C4', 'C5', ' C6', 'C7',' C8',' C9', 'C10',
                'C11', ' C12', ' C13', 'C14', 'C15', 'C16', 'C17']]
model = KMeans(n_clusters=3, init='k-means++')
model.fit(x)
y_pred = model.predict(x)
Print('homogeneity_score = ', homogeneity_score(data['category'], y_pred))
Print('completeness_score = ', completeness_score(data['category'], y_pred))
Print('v_measure_score = ', v_measure_score(data['category'], y_pred))
data['Predict'] = y_pred
print(data)
data.to_csv('result.csv', sep=',', encoding='gbk', index=False)
print('Data Save OK....')
```

Cluster Analysis of the Evaluation of Science and Technology Innovation in Local Universities Based on Python

In the sample code, n_clusters=3, you can also change it to other values, such as n_clusters=4, use the code in the above example to process and analyze the data in the statistical yearbook to get cluster analysis of k=3 and k=4. The results are as follows:

area	Python based k=3 clustering results	area	Python based k=4 clustering results	area	SPSS-based factor analysis comprehensive score	SPSS-based factor analysis ranking
Jiangsu Province	2	Jiangsu Province	2	Jiangsu Province	2.079288426	1
Guangdong Province	1	Guangdong Province	0	Guangdong Province	0.717647683	2
Shandong Province	1	Shandong Province	0	Shandong Province	0.528215622	3
Liaoning Province	1	Liaoning Province	0	Liaoning Province	0.499791967	4
Zhejiang Province	1	Zhejiang Province	0	Zhejiang Province	0.447180585	5
Hunan Province	1	Henan Province	0	Hunan Province	0.360866261	6
Henan Province	1	Beijing	0	Henan Province	0.358512771	7
Anhui Province	1	Shanghai	0	Anhui Province	0.329671127	8
Sichuan Province	1	Hunan Province	1	Sichuan Province	0.286882354	9
Beijing	1	Anhui Province	1	Beijing	0.158032257	10
Shanghai	1	Sichuan Province	1	Shanghai	0.14202706	11
Hebei Province	1	Heilongjiang Province	1	Heilongjiang Province	0.095733392	12
Jiangxi Province	1	Hebei Province	1	Hebei Province	0.080092718	13
Shaanxi Province	1	Jiangxi Province	1	Jiangxi Province	0.004277264	14
Hubei Province	1	Shaanxi Province	1	Shaanxi Province	0.003176886	15
Heilongjiang Province	0	Hubei Province	1	Hubei Province	-0.008761172	16
Fujian Province	0	Fujian Province	1	Fujian Province	-0.020217878	17
Guangxi Zhuang Autonomous Region	0	Guangxi Zhuang Autonomous Region	1	Guangxi Zhuang Autonomous Region	-0.106591674	18
Jilin Province	0	Jilin Province	1	Jilin Province	-0.140910607	19
Tianjin	0	Tianjin	1	Tianjin	-0.157757647	20
Chongqing	0	Chongqing	1	Chongqing	-0.208191983	21
Yunnan Province	0	Yunnan Province	1	Yunnan Province	-0.239145978	22
Shanxi Province	0	Shanxi Province	1	Shanxi Province	-0.300624346	23
Guizhou Province	0	Guizhou Province	3	Guizhou Province	-0.427639131	24
Inner Mongolia	0	Inner Mongolia	3	Inner Mongolia	-0.487620539	25

Autonomous Region		Autonomous Region		Autonomous Region		
Xinjiang Uygur Autonomous Region	0	Xinjiang Uygur Autonomous Region	3	Xinjiang Uygur Autonomous Region	-0.552342528	26
Gansu province	0	Gansu province	3	Gansu province	-0.581540486	27
Ningxia Hui Autonomous Region	0	Ningxia Hui Autonomous Region	3	Ningxia Hui Autonomous Region	-0.665963802	28
Hainan	0	Hainan	3	Hainan	-0.710686354	29
Qinghai Province	0	Qinghai Province	3	Qinghai Province	-0.726689215	30
Tibet Autonomous Region	0	Tibet Autonomous Region	3	Tibet Autonomous Region	-0.756725096	31

From the results of the cluster analysis, it can be seen that Jiangsu Province has a separate category. Jiangsu Province is China's economic province. The economic aggregate has been ranked second in the country for many years. At the same time, education is also very developed. In history, Jiangsu is a place where people come out. As of 2019, there were more than 170 colleges and universities in Jiangsu Province, including three 985 universities and eight 211 universities. The economy of Beijing and Shanghai is developed and the quality of education is high, but the results are not classified as Jiangsu. This may be because there are more universities directly under the central government (unit) of the People's Republic of China in Beijing and Shanghai. There are 25 in Beijing, followed by Shanghai and 8 in Shanghai.

The two columns in the table are the comprehensive scores and rankings of the provinces using SPSS for factor analysis. Compare it with the cluster analysis results in the table and find that:

The cluster analysis results of K=3 showed that except for Heilongjiang, the results of clustering are in good agreement with the results of SPSS factor analysis.

The cluster results of K=4 show that except for the SPSS factor analysis of Henan, Anhui and Sichuan provinces before Beijing and Shanghai, the Python clustering is not in the same category as the two cities. The class results are in good agreement with the SPSS factor analysis results.

In summary, the cluster analysis model is relatively straightforward, and the conclusion form is concise. In the study of practical problems, multiple methods can be combined to process the data, making the research results more scientific and reasonable.

References

[1]. Anil K J. Data clustering : 50 years beyond K-Means [J]. Pattern Recognition Letters, 2010, 31 (8): 651-666.

[2]. YANG Jinhua, LIU Xianwei. Initial Center Selecting Using K-means Clustering Algorithm [J]. Henan Science, 2016, 34(3): 348-351.

[3]. CHEN Baolou. The Research and Application in Text Clustering of K-Means Algorithm [D]. Hefei: Anhui University, 2013.