Application of Correlation Analysis and Cluster Analysis in Teaching Reform for Big Data Course

Jianhua Zhang(✉)

School of Information, Guangdong NanFang Institute of Technology, Jiangmen, China
2727611989@qq.com

Abstract. According to the needs of big data course teaching reform, Using IBM SPSS as the tool, BNUZ has carried out the correlation analysis and the cluster analysis on the data samples of the big data courses in recent three years. The results of qualitative and quantitative analysis are conducive to correction and implementation of specific teaching contents. Cluster analysis of data samples is conducive to improvement and planning of the overall teaching reform scheme. This paper emphasizes that applying advanced mathematical statistical analysis methods to teaching reform is a scientific process that must implement in teaching research, which is quite necessary. This paper also explains these algorithms used and application, such as the correlation analysis, the cluster analysis, K-means, k-medoids and so on, conducive to other disciplines' teaching research and convenient to learn from this example.

Keywords: Correlation analysis · cluster analysis · application of IBM SPSS · precision teaching · teaching reform of big data course

1 Introduction

In July 2014, this explosive news caused a huge shock: IBM invested US $100 million to support the cultivation of Chinese big data and data analysis talents. According to the active application of colleges and universities, the Ministry of education allocated this resource at the end of 2014. Therefore, more and more colleges and universities began to apply to set up for big data majors in the own university [4], and their enthusiasm was constantly stimulated. China's Ministry of Education has paid more attention to the demand for talents in national construction, guided colleges and universities to aim at the forefront of world science and technology, continuously improved their scientific and technological innovation ability, and provided strategic support for the development of the new economy [2].

It is generally believed that the teaching and teaching research of big data technology in Chinese universities began in 2016. In the past six years, teaching and teaching research in Colleges and universities can be divided into three stages. With the rapid development, the world has been changing. For the technical field, a huge change is that it opens the door to big data. With the promotion and implementation of the national big data strategy and the implementation of supporting policies, the development environment of the big

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 155–164, 2023.
https://doi.org/10.2991/978-94-6463-034-3_17

Table 1. This is a statistical table from 2016 to 2020.

Time	B.E.	coll.	sum
2016	3	0	3
2017	32	64	96
2018	253	212	465
2019	256	460	710
2020	231	619	840

data industry has been further optimized, and the demand for big data services in all social and economic fields has been further enhanced. With the approval of colleges and universities to set up the major of "data science and big data technology", compound talents will graduate from colleges and universities one after another.

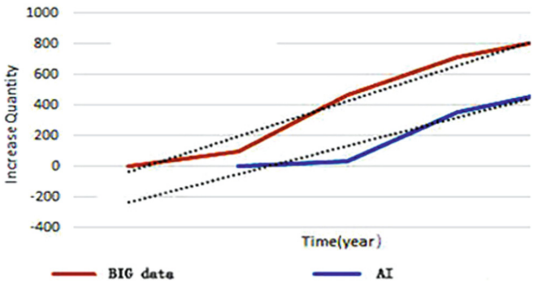
1.1 Curriculum System of Big Data Undergraduate Major

In August 2014, Professor Yihua Huang of Nanjing University wrote and published the book "Deep Understanding of Big Data: Big Data Processing and Programming Practice" [5]. The reason why the book has attracted extensive attention is not only that the content of the book is a "model of the perfect combination of theory and practice", but also because he put forward the training goal of cultivating big data technical talents and the curriculum system of undergraduate teaching for the first time in China. This pointed out the direction of teaching research for educators in other colleges and universities at that time, and soon they began big data teaching. Since then, hundreds of colleges and universities have successively applied and been approved by Ministry of Education to set up the major of data science and big data. The teaching system of big data course had established. Table 1 is a statistical table of the number of big data majors applied by colleges and universities approved by Ministry of Education in recent years.

(The statistical data in this table are collected from the public information on the Internet in China in recent years and compiled by the author of this article.) Here, we note that in the process of approving colleges and universities to set up undergraduate courses of big data, the Ministry of Education also reviewed and agreed to set up specialized courses of big data in Higher Vocational Colleges at almost the same time. Because the development of national construction is in urgent need of professionals at different levels and all kinds, this is a wise measure of the Ministry of Education.

1.2 Curriculum System of Big Data Higher Vocational Colleges Major

According to statistics, in 2017, 64 vocational colleges were approved for the major of "big data technology and application", 212 Vocational Colleges in 2018, 460 Vocational Colleges in 2019 and 619 Vocational Colleges in 2020. By 2020, a total of, 1355 Vocational Colleges have successfully applied for the major of big data technology and application, laying a foundation for the cultivation of big data application-oriented and practical talents.



Graph 1. Linear rate

By the end of 2017, a total of 30 higher vocational colleges had been successfully approved to set up the specialty of "big data management and application". This is a new major established by Higher Vocational Colleges under the background of the rapid development of big data technology. At that time, there were not many colleges offering this major, and its courses were basically engineering courses. Previously, some colleges and universities opened economic management disciplines with similar names (Graph 1).

The data of Table 1 comes from the Internet, and its information may not be very accurate, but we can draw the following conclusion: the data growth of undergraduate courses shows an obvious growth trend. What we need to pay attention to is that in more than 1000 colleges and universities and higher vocational colleges, how to cultivate professional talents and how to carry out teaching work in each college or higher vocational college, which is the problem of teaching research and teaching reforms, we should pay attention to this.

2 Implementation Scheme of BNUZ

2.1 Implementation Scheme

In the first half of 2016, Beijing University of Aeronautics and Astronautics opened a series of lecture courses with big data technology as the content, and all students of this university can choose whether to study or not. This may be a precedent for domestic colleges and universities to set up science popularization courses of big data technology. After the summer vacation of 2016, BNUZ opened a general course with the content is "big data technology system and data mining". This may be the first case of a general course with big data technology in domestic colleges and universities.

BNUZ provided full-time study for students, focusing on undergraduate education. The university has 14 branches, including School of education, School of Chinese, School of Information Technology, International Business Faculty, School of Management, College of Real Estate, School of Government and Law, College of Design, School of Art and Communication, School of Engineering Technology, School of Foreign Languages, School of Applied Mathematics, Logistics College, School of Sports Leisure, etc. Therefore, taking the data in the data table as the statistical sample data, these data are widely representative. For the convenience of statistical analysis and operation, the above 14 branches are represented by symbols, which are A, B, C, D, E, F, G, H, I, J, K, l, M and N.

Table 2. This table data are from 14 colleges or schools.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
0	0	37	28	5	4	2	4	2	7	2	6	2	0
2	0	88	65	25	10	4	8	32	6	4	32	38	6
8	0	91	75	32	15	8	10	34	16	4	61	32	8
10	0	216	168	62	29	14	22	68	29	10	99	72	14

Table 2 is the number of students who choose to study this course in one of the three years from 2016 to 2019. These students are distributed in 14 colleges or schools in BNUZ.

It should be explained here: BNUZ does not implement the full credit system. Students' own curriculum is basically carried out in accordance with the established teaching plan and scheme of the school. In each semester, students have less opportunity to choose courses freely according to their hobbies, interests and life plans. Even so, the information recorded in the data sheet can still reflect the teaching situation at that time. We have also carried out this teaching and research work on this basis.

2.2 Research Contents

Foremost, set up universal courses of big data technology serving the whole schools, and hope that this teaching can be developed year by year. This is conducive to students in the stage of higher education, learning and access to the latest cutting-edge information of science and technology, and understanding more advanced scientific research methods and work skills. This is very beneficial to students' employment after graduation. Of course, cultivating high-quality talents is more beneficial to national construction.

Second, by counting the number of students participating in the course and the professional category to which the students belong, we can analyse the students' cognition of this new technology field, and even predict this subject and major development prospects in the next few years. Because big data technology is being integrated into various disciplines and plays an indispensable role in the development of various disciplines and the progress of undertakings in their respective fields. The progress of education is the driving force, the development of national scientific and technological productivity.

Third, universal teaching and specialized training of students with personal characteristics, is our work policy. We are particularly concerned about the ways or methods to promote the teaching of big data popularization courses and improve the quality of students and the overall teaching quality of the school.

3 Calculation Process of Correlation Analysis Algorithm

3.1 Purpose of Work

At the end of each semester, we conduct correlation analysis on students' academic performance and the change of the number of students participating in the course, which

is one of the methods to evaluate the effect of teaching work and the basis to improve teaching methods, which is to improve teaching quality and do more detailed work [3].

3.2 Algorithm Description

To determine the closeness of the correlation between phenomena, the correlation coefficient r is usually calculated, and the absolute value above 0.8 indicates a high correlation. If necessary, the significance test of R shall be carried out.

3.3 Analysis and Conclusion

First, we remove the unqualified data in the original table to form a new data table as shown below. Then calculate the correlation between the data in Table 3 to obtain the results of the analysis data we need.

Figures 1, 2 and 3 are SPSS operation step 1.

Please pay attention to its operation.

Therefore, we can draw the following conclusions: the teaching situations of the three academic years are closely related to each other. This shows that the continuity and cohesion of teaching work are very good. On the other hand, it also shows that the teaching situation of the second school year and the third school year is better connected. As shown in the table, the exact value of correlation degree is $0.961 > 0.868$.

Table 3. This table data is sorted from the Table 2.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
0	0	37	28	5	4	2	4	2	7	2	6	2	0
2	0	88	65	25	10	4	8	32	6	4	32	38	6
8	0	91	75	32	15	8	10	34	16	4	61	32	8
10	0	216	168	62	29	14	22	68	29	10	99	72	14

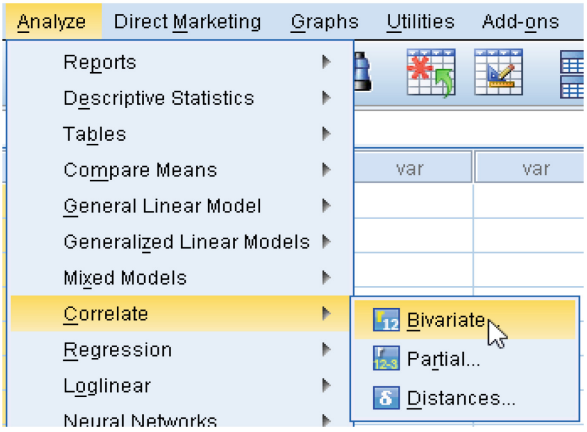


Fig. 1. This item is the choice of correlation analysis.

Descriptive Statistics

	Mean	Std. Deviation	N
M1	7.07	11.132	14
M2	23.00	26.282	14
M3	28.14	28.463	14

Fig. 2. Statistical table of the calculated data.

Correlations

		M1	M2	M3
M1	Pearson Correlation	1	.894**	.868**
	Sig. (1-tailed)		.000	.000
	Sum of Squares and Cross-products	1610.929	3400.000	3574.857
	Covariance	123.918	261.538	274.989
	N	14	14	14
M2	Pearson Correlation	.894**	1	.961**
	Sig. (1-tailed)	.000		.000
	Sum of Squares and Cross-products	3400.000	8980.000	9347.000
	Covariance	261.538	690.769	719.000
	N	14	14	14
M3	Pearson Correlation	.868**	.961**	1
	Sig. (1-tailed)	.000	.000	
	Sum of Squares and Cross-products	3574.857	9347.000	10531.714
	Covariance	274.989	719.000	810.132
	N	14	14	14

**. Correlation is significant at the 0.01 level (1-tailed).

Fig. 3. Calculated variable relationship table.

4 Calculation Process of Cluster Analysis Algorithm

4.1 Purpose of Work

At the end of each semester, we used the cluster analysis is carried out for the student's information who participated in the course. We use the calculation results obtained by the algorithm of mathematical analysis to evaluate whether the teaching course content is conducive to students' learning or not. In this way, we can prepare for the next step of layered teaching and arranging different course contents.

4.2 Algorithm Description

Cluster analysis is an exploratory analysis. In the process, people do not need to give a classification standard in advance. Cluster analysis can automatically classify from the

sample data. From the perspective of practical application, cluster analysis is one of the main tasks of data mining. Moreover, cluster analysis can be used as an independent tool to obtain the distribution of data and observe the characteristics of each cluster of data.

Clustering analysis has many algorithms, such as k-means, k-medoids and so on [6], which have their own characteristics and different applications. We use k-means algorithm.

4.3 Analysis and Conclusion

The original data set is shown in Table 3.

Scheme 1: We pre-set two aggregation points and 10 iterations. The operation is as shown in Fig. 4.

Next: step (Figs. 5 and 6).

Please note that we have selected two aggregation points this time. Then, next time, we will select three or more aggregation points and analyse their calculation results respectively.

The calculation results are as shown in Figs. 7 and 8.

Scheme 2: We pre-set three aggregation points and 10 iterations. The operation is as shown in Figs. 9, 10 and 11.

The calculation results are as shown in Fig. 12.

Our focus is this: This is the calculation result of selection of three aggregation points. The calculation results are shown in Fig. 13.

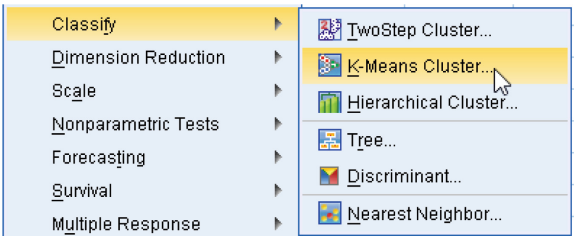


Fig. 4. The following is SPSS operation step 1.

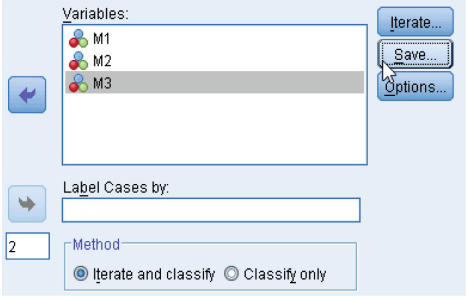


Fig. 5. The following is SPSS operation step 2.

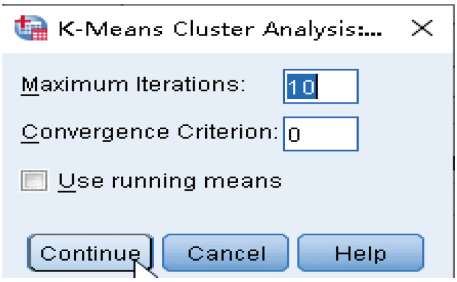


Fig. 6. The following is SPSS operation step 3.

Number of Cases in each Cluster

Cluster	1	2.000
	2	12.000
Valid		14.000
Missing		.000

Fig. 7. Statistical table of the calculated data.

Final Cluster Centers

	Cluster	
	1	2
M1	33	3
M2	77	14
M3	83	19

Fig. 8. The calculation results are shown below.

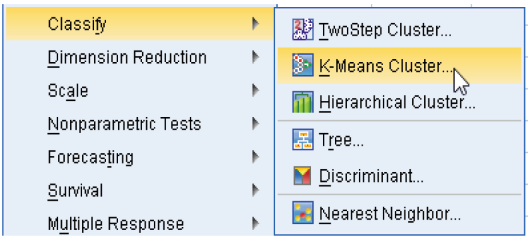


Fig. 9. The following is SPSS operation step 1

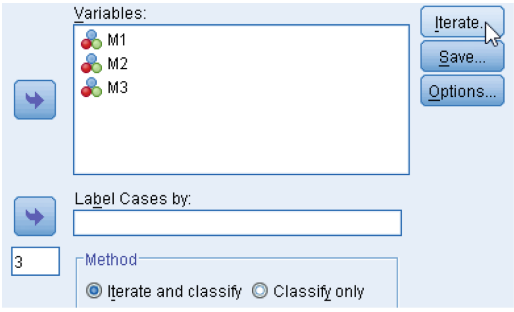


Fig. 10. The following is SPSS operation step 2.

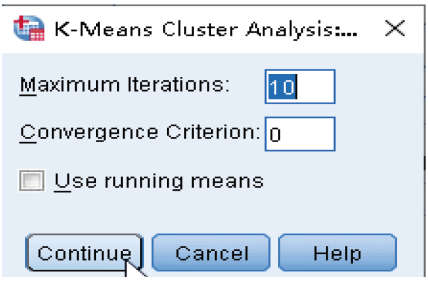


Fig. 11. The following is SPSS operation step 3.

Number of Cases in each Cluster

Cluster	1	4.000
	2	8.000
	3	2.000
Valid		14.000
Missing		.000

Fig. 12. Cases in each Cluster

Final Cluster Centers

	Cluster		
	1	2	3
M1	4	2	33
M2	32	5	77
M3	40	9	83

Fig. 13. This is the final calculation result.

From the results of the above analysis, it is obvious that if we adopt the second scheme, it will be more conducive to the future teaching work. If this teaching information through modern data analysis method is applied to every academic year or even every semester, it may be more effective.

5 Conclusions

The content described in this paper is only one of the applications of IBM SPSS in teaching reform. In fact, IBM SPSS can play a more important role in the field of educational research. However, in teaching and research, we should avoid those troublesome things involving personal privacy, information security, trade secrets, morality and law, and so on. Harmonious development and common progress are our original intention.

Acknowledgements. This paper is one of the achievements of the research project of "Big Data Technology Professional Group" of Guangdong Provincial Education Department. (No.: GSPZYQ2021052).

I would like to thank all the teachers and staffs. I would also like to thank the Department of Education of Guangdong Province for its strong support to our Guangdong NanFang Institute of Technology in carrying out this research work.

References

1. *HaiBin Wang*, 2021. *BASE AND APPLICATION OF ARTIFICIAL INTERLLIGENCE*, China Electronic Industry Press, ISBN: 987-7-121-41296-7
2. JunJun Cen 2021. *CHINA BIG DATA APPLICATION DEVELOPMENT REPORT NO. 5 (2021)*, China Social Sciences Literature Press. ISBN: 9787520191531
3. Li Lu, 2021. *MATHEMATICAL FOUNDATIONS OF DATA SCIENCE,* China People's Posts and Telecommunications Press, ISBN: 978-7-115-55288-1
4. Phil Simon, 2014. *BIG DATA APPLICATION*, China People's Posts and Telecommunications Press, ISBN: 9787115365262
5. YiHua Huang, 2014. *UNDERSTANDING BIG DATA: BIG DATA PROCESSING AND PROGRAMMING*, China Machine Press, ISBN: 9787111473251
6. ZhiHua Zhou, 2016. *MACHINE LEARNING*, China QingHua University Press, ISBN: 9787302423287