ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research (ASSEHR), volume 300

2018 International Workshop on Education Reform and Social Sciences (ERSS 2018)

Analysis of Postgraduates' Behavior and Learning Achievements based on Clustering Method

Yongchao Shen[1, a], Jiawen Li[2, b] and Menghua Huo[3, c]

[1]Graduate School, Beihang University, Beijing 100191, China;

[2] School of Mathematics and Systems Science, Beihang University, Beijing 100191, China;

[3] School of Economics and Management, Beihang University, Beijing 100191, China.

[a]xxgl@buaa.edu.cn, [b]jiawenli@buaa.edu.cn, [c]1021364586@qq.com

Abstract. With the rapid development of information technology, the application of big data in the education management has attracted more and more scholars' attention. The widespread use of information recognition methods, especially the Ecards' swiping technology provides an important support for the collection of students' data. In this paper, the data of dormitory access, library access, breakfast consumption, published paper and course grades are combined to describe the characteristics of graduate students. Then academic graduate students are clustered into seven categories, from which data portraits for "straight A student" and "top researcher" are obtained. The colleges are divided into three categories according to the nature of their students' paper, thus we can explore the differences of students' behavior in different colleges. The research shows the prospect of machine learning in education management, and provides some inspiration to managers in this field.

Keywords: Hierarchical Clustering, Big Data, Ecards, Students' Behavior.

1. Introduction

With the rapid development of the information construction in universities, artificial intelligence and data mining technology have made it possible to discover new knowledge from massive data. The application of technology related to big data in education management has attracted more attention than any before [1]. Analyzing the behavioral rules behind the mined data provides managers with decision-making support and enlightenment, so as to advance the innovation of school governance.

From having meal at canteen to entering the library, swiping Ecards is ubiquitous in students' daily routine [2]. These records generate massive data in the database day after day. In addition, the course grades of postgraduates and the publication of their papers are also available, which makes it possible to analyze the implicit association between students' behavior at school and their accomplishment. In this paper, we use the hierarchical clustering method to acquire the rules.

The following will briefly describe hierarchical clustering. It is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other, and the objects within each cluster are broadly similar. Hierarchical clustering is a special kind of clustering method. Given a set of N items to be clustered, and a N×N distance (or similarity) matrix, the basic process of hierarchical clustering [3] is this:

Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances between the items they contain.

Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

Compute distances between the new cluster and each of the old clusters.

Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

2. Data and Modeling

For the sake of data integrity, we selected the 1061 academic master students who have graduated as the research subjects, and analyzed their school behavior and learning outcomes. For the description of their school behavior, records of dormitory access at unusual times, library access and breakfast consumption are utilized. For the qualification of learning outcomes, we selected two indicators of the student's course scores and publication of papers.

2.1 Data Preprocessing.

As for dormitory access data, the problem of missing data is quite noticeable. On the one hand, the access control system is off-duty by day. On the other hand, a batch of students will only need one card to enter the building. In order to avert these two types error, the paper decides to only count the access records during the abnormal time period. The abnormal time period is set as from 0 a.m. to 4 a.m. In this way, the total number of dormitory access at unusual times from August 2016 to April 2017 is counted. The more times, the more the student is used to returning late.

Since students can leave the library without swiping the card, we can't get the records of the student's specific departure time, nor can we calculate the further data such as the learning time span. So, for the data of library access, we can only count the number of their entrance in this semester.

Students' consumption during breakfast is an important indicator of a student's early habits. For breakfast consumption data, this article selects the canteen consumption records before 10 am, and does not accumulate the multiple purchases of breakfast at one day.

A very important indicator of student excellence is his learning grades. Taking into account the differences in the selection of different graduates in different majors, this article uses the average score of each student during this semester as a measure.

For graduate students, scientific research is as important as their course learning. The achievement of their research is generally measured by the level of published papers, in which the number of papers and the author order are really critical. Consider the following formula to evaluate the student's research ability.

$$Ability\ of\ scientific\ research = \sum_{\substack{The\ number \\ of\ papers}} \frac{1}{Author\ order} \tag{1}$$

2.2 Concept Hierarchy

Concept hierarchy is a data discretization technique. It is usually used to divide continuous data for some convenience reasons. For example, people's age can be divided into juvenile, youth, middle-aged and elders. Students' learning grades can be divided into excellent, good, medium and poor. Although the details of the data are lost, the discrete data is more interpretable and more meaningful in the analysis[4] .

In our data, the differences between the samples are very large. For example, in terms of the number of library accesses, a large number of students may not be used to studying in the library. They may have only visited several times during a semester, but a small number of students may have gone more than a hundred times. Such samples are already out of the group in the data set, which may introduce great interference to the model. Therefore, we transformed our data into five grades with values 1 to 5, from low to high. So, we get our raw data set

$$\left\{ x^i \in V^5 : x^i = \left(x_1^i, x_2^i, x_3^i, x_4^i, x_5^i \right), i = 1, 2, \cdots, N \right\} \tag{2}$$

Among which $V = \{1,2,3,4,5\}$ is the value range of each feature, $N = 1061$ is the total number of graduate students we selected. Each student is represented as a x^i .

3. Experiment and Analysis

The original data set is imported into SPSS, and the function of hierarchically clustering is applied. The rules are quite detectable when clustered into 7 categories. For each type, the average of the five characteristics is obtained, shown in Fig 1 and Table 1.

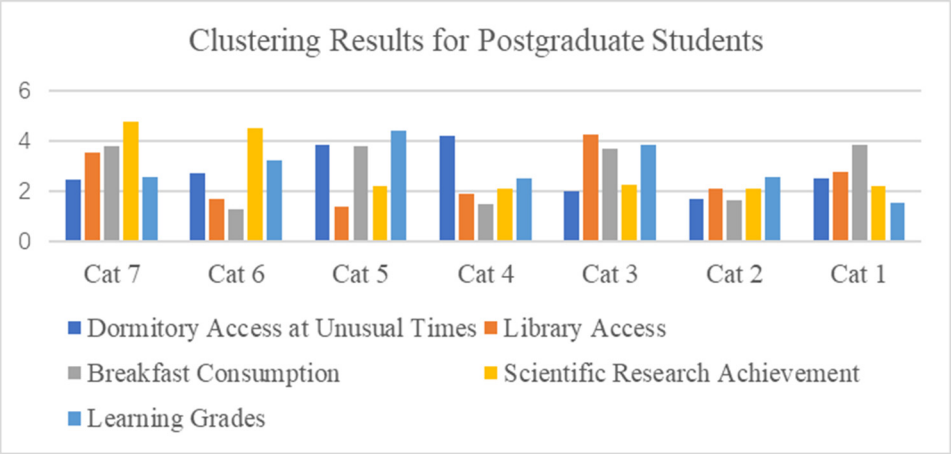

Fig. 1 Clustering Results

Table 1. Population of Each Category

Category	7	6	5	4	3	2	1
Number of Students	24	41	24	189	117	386	280

In Fig1 we can see Cat7 and Cat6 can be regarded as two types of "top researcher". Students in Cat7 are used to going to library and eating breakfast at canteen, while those in Cat6 are on the opposite. However, to some extent, they both have some late return issues.

Those "straight A student" are centered on Cat5 and Cat3. Students in Cat5 go out early and come back really late, which epitomize the habitual figure of top students. As for Cat3, they show incomparable zeal for library.

We are also interested at those who have preponderance on other three features. Students in Cat7 and Cat3 pay a lot of visit to library, which gives them positive feedback on courses learning or scientific research. Cat7, 5, 3, 1 are used to get up early and have breakfast at canteen, and we can see the promotion on three categories of them (except for poor Cat1). Cat5 and Cat4 are quite similar on the number of their abnormal dormitory access, while Cat5 are extolled as top students and Cat4 does not have notable advantages on their study or research. As the population of Cat4 is 7 times bigger than that of Cat5, it is reasonable to consider late return as a bad habit.

In order to explore the connections and differences between colleges, we divide all the colleges into three types according to the dominant type of their papers. If more than 60% of this college's science papers are published on journals, then we will label this college as J-type. Similarly, it will be labeled as C-type if more than 60% are published on conferences. Rests of colleges are regarded as B-type for having traits on both journal papers and conference papers. Several colleges are abandoned for their insufficient population. Finally we get 378 J-type students, 414 C-type students and 258 B-type students, their distribution in seven categories are shown in Table 2.

Table 2. Distribution of Three Types of Colleges

	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
J-type	96	137	79	48	7	8	3
B-type	87	80	16	51	6	8	10
C-type	94	163	20	90	11	25	11

Sort the relative populations of each cluster, we get Table 3

Table 3. Relative Population of Three Types of Colleges

J-type	Cat3	Cat2	Cat1	Cat5	Cat4	Cat6	Cat7
Relative Population	0.686957	0.360526	0.34657	0.291667	0.253968	0.195122	0.125
B-type	Cat7	Cat1	Cat4	Cat5	Cat2	Cat6	Cat3
Relative Population	0.416667	0.314079	0.269841	0.25	0.210526	0.195122	0.13913
C-type	Cat6	Cat4	Cat5	Cat7	Cat2	Cat1	Cat3
Relative Population	0.609756	0.476191	0.458333	0.458333	0.428947	0.33935	0.173913

It can be seen that these three types of colleges contribute the most to Cat 3, Cat 7 and Cat 6, respectively. As Fig1 shows, during the transition from the J-type to the C-type, the number of dormitory access control in the abnormal period is increasing, and the number of times to library and the number of eating breakfasts are gradually decreasing, students' advantages have progressively shifted from course achievement to scientific research outcomes.

4. Summary

As we can see from the clustering results, there are two kinds of "top researcher", one goes to library a lot and often has breakfast at canteen while another kind is on the opposite. Both kinds will return late sometimes. "straight A student" is basically an image of getting out at dawn and returning at dusk, or the students who get immersed in the library can also achieve excellent grades. Spending much time in library or getting up early to eat breakfast is of great help to increase course scores or scientific research, while late return should be lessened in most cases.

Colleges are separated into three types according to the differences of published papers. It is found that the students who are more likely in J-type colleges generally work and rest at regular hours, and the advantage lies in their course studies. C-type students may lead to stay up late to work on their latest papers, and the scientific research achievements are generally more. That is maybe because journals usually spend more time on peer review than conferences.

In the future we may add other features to the model to portray the students' image more accurately. Due to the accommodation arrangements for postgraduates, it is inconvenient for students to go to the canteen for breakfast. Therefore, some students may choose to buy breakfast at the shops on roadside, which leads to lower credibility of the breakfast consumption data. In addition, most postgraduates have their own assigned offices on campus, so they may not choose to use library for studying. All of these factors will have some impact on our analysis. In the future, we may need to try various methods to minimize these deviations, and take part in other features like internet using data to strengthen our model. It is promising to increase the support of the data to the conclusions and discover the unknown rules.

Acknowledgments

This study was financially supported by the fundamental research funds for the central universities (YWF-18- XGB-004). Yuxi Zhang's work on collecting data is sincerely appreciated.

References

[1]. Xiaohao Ding. The Educational Research under the Context of Big Data[J]. Tsinghua Journal of Education. Vol. 38 (2017) No. 05, p. 8-14.

[2]. Hua Wang, Ling Li, Fan Yang. Research on campus card data analysis and application in big data era[J]. Modern Electronics Technique. Vol. 41 (2018) No. 04, p. 56-59.

[3]. S. C. Johnson. Hierarchical Clustering Schemes. Psychometrika, 1967, p. 241-254.

[4]. B. C. Chien, C. H. Hu, M. Y. Ju. Learning fuzzy concept hierarchy and measurement with node labelling[J]. Information Systems Frontiers. Vol. 11 (2009) No. 05, p. 551-559.