Student Management Based on Rough Set and Clustering

Xueli Ren[1, a *] and Yubiao Dai[1, b]

[1]School of Information Engineering Qujing Normal University Qujing,China

[a]oliveleave@126.com, [b]abiaodai@163.com

Keywords: Student management; Rough set; Cluster; K-means

Abstract. The credit system makes teaching students in accordance with their aptitude possible, which provides the basis for the development of students' personality. At the same time, it has brought great challenges for the management of the students. In order to better manage the students, and to provide some useful guidance, K-means algorithm is used to cluster the students to the nearest classifications; rough set is used to reduce attributes in order to improve the efficiency of clustering. The method is applied to cluster students, and the result shows that it is feasible.

Introduction

With the deepening of teaching reform in higher education, all the colleges and universities have actively explored and implemented the management mode of the credit system in order to meet the requirements of the popularization of higher education and the diversification of the stage[1-3]. The implementation of credit system provides a good development space for students that perfect their personality and all-round development. But the students' management under the credit system is different from the existing student management system, which will put forward new requirements and challenges to students' management [4-5]. The original fixed class management mode is broken; course selection, classes and examination of students are uncertainty. Due to the different in courses selected, a class may from different majors and grades of the students; even the students live in same dormitory, curriculum is also different. Therefore, students have the characteristics of dispersion and mobility, which brings great challenges to the daily management.

Under the credit system, degree of freedom increased for each student, but all the students in the whole school will have some of the same or similar circumstances for a long time, the so-called "Like attracts like, if according to the Birds of a feather flock together". If students are clustered based on the performance for a long time, so that students have more in common in the same clustering, and different clustering are different. The different methods of management can be used for different clustering, which can not only improve the pertinence of management, but also provide some useful suggestions for student in course selection, employment and other aspects. K-means algorithm which has the advantages of classification for large data sets is a very influential technology in scientific and industrial applications. Therefore, the method is used to cluster the students in this paper.

K-means Clustering and Rough Set

K-means Clustering. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [6-8].

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq$ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in s_i} \left\| x - \mu_i \right\|^2 \tag{1}$$

Where μ_i is the mean of points in S_i.

Rough Set. Rough set first described by Polish computer scientist Zdzisław Pawlak to deal with imprecise or vague concepts. Here, the basic notation is introduced only from rough set approach used in the paper [9-13].

An information system is denoted as S=(U, A, V ,f) where U={ U1,U2,U3,…,U |u|} denotes the set of all objects in the system, A={a1,a2,a3,…, a|A|} is the set of all attributes. C is the set of conditional attributes and D is the set of decision attributes. C and D are mutually exclusive and C $\cup$ D = A, C $\cap$ D= ϕ, then S is viewed as a decision table. V= $\cup$ Va where a $\in$ A, Va is the range of the attribute a; f: U $\times$ A $\rightarrow$ V is an information function, if q $\in$ A, x $\in$ U, then f(x, q) $\in$ Va is the attribute value of the object in U.

F(x, q) denotes the value of attribute q $\in$ A in object x $\in$ U. f(x, q) defines an equivalence relation over U. With respect to a given q, the function partitions the universe into a set of pairwise disjoint subsets of U:

$$R_q = \{x : x \in U \wedge f(x,q) = f(x_0,q) \qquad \forall x_0 \in U\} \tag{2}$$

IND(P) denotes the indiscernibility relation for all $P \in A$. $U / ind(P)$ denotes the partition of U given IND(P) and which is calculated by formula 3.

$$U / IND(P) = \otimes\{q \in P : U / IND(q)\} \tag{3}$$

Where $A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y = \phi\}$

The lower and upper approximation of a set P $\subseteq$ U (given an equivalence relation IND(P))is defined as:

$$\underline{PY} = \cup\{X : X \in U / IND(P), X \subseteq Y\} \tag{4}$$

$$\overline{PY} = \cup\{X : X \in U / IND(P), X \cap Y = \phi\} \tag{5}$$

The positive region contains all objects in U that can be classified in attributes Q using the information in attributes P. The negative region is the set of objects that cannot be classified this way.

Clustering Based on Rough Set and K- means

Clustering can not only reduce the difficulty of student management, but also make student management have the advantages of pertinence and high efficiency; K means is an effective clustering algorithm, so it is used in student management in this paper.

The height, gender, age of student and other factors have little effect on student management, so they are not considered in the paper. In addition, although character is important, but it is very difficult to collect data, it also isn't considered in the paper. The first fundamental task of students in the school is learning, and a test is used to know the effect of learning after completion of the course. Therefore, these data not only are large, but also can guide the students select courses that is suitable for their own. The grades are used to cluster students in this paper. The specific process is shown in Fig. 1.

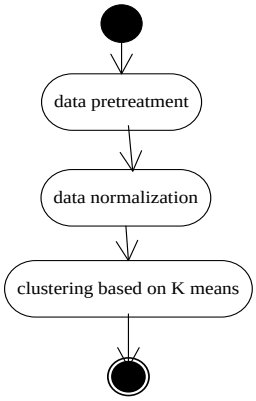

Figure 1. The process of cluster

Data Pretreatment. A table is constructed where courses are columns and students are rows, courses are labeled from C1, students are labeled from S1.Missing value in the table should be processed before reduction. The techniques of missing value imputation are: listwise deletion, mean imputation and some types of hot-deck imputation [14]. The listwise deletion is used to deal with missing value in the paper.

Attribute Reduction Based on Rough Set. Reducing the attribute of datasets is one of the core contents in rough set theory. The two steps are required to reduce attribute. Firstly, data must be discrete to improve its general. These methods are commonly used: equal width algorithm, equal frequency algorithm, Naive Scaler methods and so on. As the scores have the characteristics of the normal distribution, the equal frequency algorithm is used in the paper. Finally, remove the attributes that are not as important or redundant in data set, the process is realized by the following algorithm.

Input: A decision table.
Output: the attribute reduction, Red
Core(A)=Φ
For each a∈ A
Count the Positive Region $POS_{(A-\{a\})}(d)$
If $\left(POS_{(A-\{a\})}\ (\mathrm{d}) \neq POS_A(d)\right)$
Core(A)=Core(A)+{a}
End if
End for
RED=Core(A)
Output RED

Clustering Based on K-means. The data may be clustered after pretreatment; the students will be divided into 3 categories in accordance with the scores. Firstly the three initial clustering center is determined, were normalized scores 0, 1 and 0.5 of C1; then each student will be assigned to the nearest class based on Euclidean distance to three clustering center; finally, the cluster center is recalculated until convergence.

Example

As an example, some grades of students of a class in specialized in computer are taken. This grade table is pretreated by the method in step 1. The scores are discrete based on the equal frequency algorithm. A part of results are shown in Table 1.

The set of {C1, C2, C4, C7, C12} is formed after reduction.

The students are clustered to 3 categories based on k means without reduction and with reduction; the result is shown in the Fig. 2.

Table 1 Results of attribute discrete

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
S1	3	2	2	3	3	2	2	2	2	3	3	3	2	2
S2	2	2	4	3	2	3	3	3	3	2	4	2	4	4
S3	3	3	3	2	3	3	2	3	3	3	3	4	3	4
S4	4	3	2	4	1	1	3	1	2	4	2	2	2	2
S5	3	4	2	3	1	2	4	2	3	2	2	1	2	3
S6	1	3	3	2	2	3	1	3	2	2	2	1	2	1

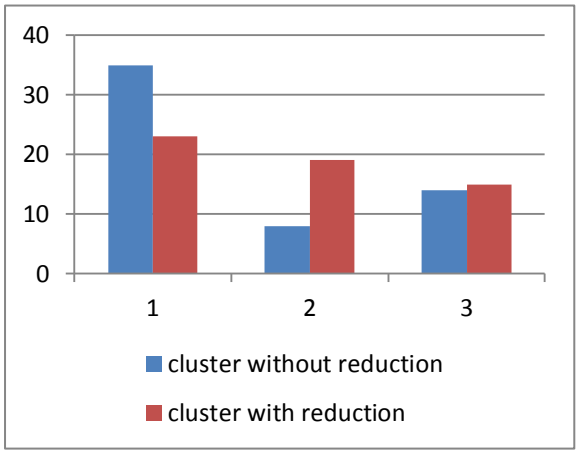

Figure 2. The result of clustering

Conclusion

The student management of the credit system is discussed in this paper, and the students are clustered based on K-means clustering algorithm, the rough set is used to attribute reduction in order to improve the computational efficiency.

References

[1] Zhu Wenyu. Innovation of Student Management in Local Colleges and Universities based on Credit System Information Platform [J]. The Educational technology and equipment of China，2011.11：6-10

[2] Yan Manli. Reflection on strengthening the management of students in Colleges and Universities under the condition of credit system [J]. Journal of Sichuan College of Education, 2006, 22(11):5-7, 13.

[3] Zheng Hua. Explore to discuss Music teaching proficient reform in higher colleges [J]. Journal of Jiamusi Education Institute,2010.6:102

[4] Liu Chang. The innovation of College Students' management idea under the background of the credit system [J].CHINA ADULT EDUCATION,2016:44-46

[5] Yang shuxia. The Problems and Strategies about the Management of College Student under the Credit System[J].JOURNAL OF DALI UNIVERSITY.2011.7:79-82

[6] HU Wei. Improved hierarchical K-means clustering algorithm. Computer Engineering and Applications[J].Computer Engineering and Applications.2013.2:157-158

[7] HELing, WULing-da, CAIYi-chao. Survey of Clustering Algorithmsin Data Mining [J]. Application Research of Computers.2007.1:10-13

[8] K means [EB/OL]. http://baike.baidu.com/link?url=UnBBssdPNTCpndeuvXF0_1scFqvXb9PIVVPJ3Fp0qAuzZpK-Xvi31wEwQhd6bKqO68IU7tNDWb-n9jhM8dUE6q.2016.4

[9] Pawlak Z.Rough set [J]. Internaional Journal of Computer and Information Science,1982,11(5):341-356

[10] WANG Guo-Yin, YAO Yi-Yu, YU Hong.A Survey on Rough Set Theory and Applications[J].chinese journal of computers,2009.7:1229-1232

[11] WANG Xue-en, HAN Chong-zhao, HAN De-qiang. A Survey of Rough sets Theory[J].Control Engineering of China.2013.1:1-4

[12] He Chaobo, Chen Qimai. Rough set analysis model for correlation between courses and its application. Computer Engineering and Applications, 2011, 47(27):233-235.

[13] Rough Set [EB/OL]. http://baike.baidu.com/view/452607.htm.2016.5

[14] Missing Value [EB/OL]. http://baike.baidu.com/view/1578358.htm.2015.10