The Application of DM in University Admissions Decision Making

Zaixun Guo, Liying Fang, Lei Yu, Hang Su, Zhifeng Liu
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
E-mail:guozaixun@sina.com

***Abstract* — With the purpose of extracting potential knowledge from the vast amounts of data, as well as providing auxiliary information for decision making for college admissions departments, in this paper, based on SPSS 12.0 software clementine, it mainly uses C5.0 decision tree algorithm and association rules Apriori algorithm to study the college entrance examination enrollment data, students' performance in college and employment data. It mainly focuses on the data preparation process, and shows the mining process aimed at two different goals and the valuable information dug out during these processes. The results of study shows that this kind of information can provide decision making for the department in charge of admissions.**

Keywords-data mining; admissions; clementine; decision tree; association rules

I INTRODUCTION

With the development of universities admissions informatization, more and more college enrollment data are accumulated, and how to utilize the existing information resource for admissions decision service has already become an urgent subject that we need to face. According to the application of DM in University Admissions, it is easy to find a variety of potentially valuable rules in the historical data, to scientifically guide the admission, and to efficiently conduct the propaganda; it will be helpful to improve the quality of students. There is also an important practical significance to the development of the whole university and the improvement of the education quality. This article will discuss how to use the DM to do a confluence analysis on the enrollment data of college students, student status data and student employment data, and then provide decision support to the leaders of universities.

II THE OVERVIEW OF DM

The Data Mining (DM), which is also known as the Knowledge Discovery from Database (KDD),is a complex process that extracts and digs out the unknown and valuable knowledge from large amounts of data[1]. DM is used to mine information and discover knowledge without explicit assumptions. At the same time, there are three characteristics of DM,i.e.The information obtained by DM should be previously unknown, effective and practical. The previously unknown information refers to the information which is unforeseen in advance, and in other words, DM isused to discover the knowledge or information which cannot be found by intuition, even if it is counterintuitive. The more unexpected the information dug out is, the more valuable it might be.

This article intends to use DM in the admissions of Beijing University of Technology, and dig out the valuable knowledge from vast and historical data to provide strong support for university. In this paper, the research targets are: 1.To try to research the potential relationship among the weighted scores of students and cast archives voluntary enrollment data, college entrance examination, the examinee category, the choice of major, origin of student and the source of middle school,and 2. To try to research the potential relationship among students' employment after graduation, admissions and student status.

KDD is a complex process[2], and the steps are shown in figure 1.

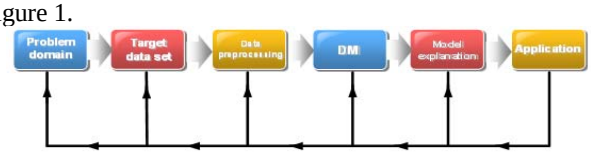

Figure 1 Steps of KDD

Based on this article, the actual operation of each steps in figure 1 is shown as follows

- Problemdomain search: including to understand university admissions and employment related business process,to learn relative background knowledge,and to determine the data mining tasks.
- Target data set selection: according to the requirement of step 1,to dig out the admissions data, student status data and employment data of the undergraduate students at Beijing University of Technology from the year of 2005 to 2008.
- Data preprocessing: tointegrate, cleanse and transform the data in step 2, and to make these data into high-quality ones which can be directly applied into DM tools
- DM: to choose the appropriate data mining tools according to the DM tasks and the nature of data. In this paper, SPSS clementine 12.0 will be used as the data mining tool.
- Model explanation and evaluation:to remove the useless or redundant mode, and to store or submit the valuable mode to the decision maker in an easy way.

• Application: use the valuable mode or knowledge by above steps to guide the next phase of the admissions work.

III DATA PREPARATION

An accurate and reliable data source is the prerequisite of decision analysis, and the data preparation is crucial for the whole process of mining. The preparation work will directly influent the efficiency and quality of data mining. The data preparation process is shown in figure 2.

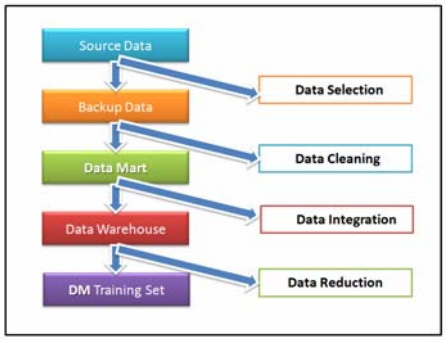

Figure 2 Steps of Data preparation

In this paper, the source data consists of three parts: The undergraduate admissions data, student status data and employment data of Beijing University of Technology from the year of 2005 to 2008. The main data types are numerical data and text data.

A. *Data Selection*

For the Beijing University of Technology, 70 percent of the students come from Beijing, and most of them are enrolled in science majors. These students constitute the main body of the school students, which is enough to provide support for the two objectives in this paper. And the students who come from other provinces or enrolled in literature and history majors only take a small percentage, and meanwhile this group of students become special because of the different policies in annual admissions plan. And the next phase of study will be designed for these students as a special case study. So this article chooses the data of the Beijing students who are enrolled in headquarters and science and engineering majors as the source data, and all the rest of data is deleted.

B. *Data Cleaning*

The data cleaning generally includes: missing value processing, noise data processing, disposition of abnormal indexes, duplicate data check, and the validity of the data validation. This article adopts the methods of ignoring tuple, filling and speculating to deal with missing value. For example, for a small number of high-level athletes, for whom the different admissions way lead to the information missing or incomplete, these data can be deleted directly; for individual students, for whom there is a lack of employment records, these data can be manually filled in according to the tripartite agreement which is retained in the university; however, when the data quantity is large, a maximum probability dispatch can be used to fill in these data automatically, or the possible employment is determined based on the student class and the general contract rate in the major.

C. *Data Integration*

In the process of data integration, the main problems are data redundancy, physical inconsistencies and data values conflict. For example, admission, employment, and school roll will contain the same basic information of students, such as the origin of student and major, and after the standard unified induction, the redundant data can be deleted directly; For the political affiliations of the students, the status shown in the admissions data is the one upon the entrance, while the employment data is updated to the graduation. Therefore, the employment data will become the standard for correction; when the place of birth data are inconsistent in admission data and employment data, these data need to be integrated.

D. *Data Reduction*

Data reduction is the main point of data processing in this paper, and data generalization and the continuous data discretization are used as the main ways to realize the data reduction. Here, are three kinds of methods of data processing.

1) *Generalization sub-file of source middle school:* The middle schools are abstractly generalized into four classes of A, B, C and D, D represents for the non demonstration senior high schools in Beijing, and on the other hand, the demonstration senior high schools are divided into three levels of A, B, and C.The following points need to be considered for the ranking: 1. the total number of students in the most recent 8 years, and the class of A is assigned to the schools with a large number; 2. whether it is a key high school in Beijing, and the class of A is assigned to the key high school; 3. the senior high school entrance examination grades, and the class of A is assigned to the ones with a high score; 4. whether it is a base school of Beijing University of Technology.

2) *The college entrance examination score:* Because the difficulty of the annual examination paper, admission scale and provincial plan number vary each year, the "gold content" of the same score of the college entrance examination varies in different years. Therefore, it is necessary to standardize the college entrance examination scores. In this paper, Z-Scores[3] is used as the standardized method, and its computation formula is:

$$Z_i = \frac{X_i - \overline{X}}{S}$$ Where: $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ and $$S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$. X_i is the raw score of student i, $\overline{X}$ is the average original points of all admitted candidates in the same year, S is the standard deviation of the annual admissions examinee distribution, and Z_i is the standard score of student i. Table 1 shows the average college entrance examination score and annual standard deviation of

science and engineering students in Beijing University of Technology from the year of 2005 to 2008.

TABLE 1 THE TABLE OF $\overline{X}$ AND S IN DIFFERENT YEARS

Years	Average value $\overline{X}$	Standard deviation S	Number of people
2005	521	30	1622
2006	578	26	1583
2007	576	23	1544
2008	554	25	1344

College entrance examination scores can be achieved by functions AVERAGE and STDEVP in Mircosoft Office Excel 2010. The actual result of Z_i is distributed into the interval [-4, 4], and standardized scores classifying is shown in figure 3, which means that $Z\,i \geqslant 1$ is excellent, $Z\,i \in [0,1)$ is good, $Z\,i \in [-1,0)$ is medium and $Z\,i < -1$ is fall.

Figure 3 Steps of Standardized scores

3) *The discretization of the weighted scores:* Since the scores of students of different grades and majors are also independent, theoretically, it is also necessary to use Z-scores for the standardization. But there are more than 40 majors in Beijing University of Technology, and the score of the selective courses of each student are different, the workload for the standardization will be beyond imagination, and its effects might not be proportional to the workload.

For the discretization of the weighted scores, most researchers divide the different quality directly according to the score interval. Usually, the boundary of scores is taken as: excellent（>80）, good（70-80）, medium（60-70） and fall（<60）. This method, which is based on the experience, is not suitable for the division of course for the university in consideration of the fact that the difficulty is also different in different examination. This paper uses Z-Scores to determine the suitable interval step method through the experiment based on the actual situation of Beijing university of Technology. Firstly, the scores of five senior students, for whom the admitted number of their major is largest, are extracted as the sample, then Z-Scores is used for standardization, those students are divided into different levels according to figure. 3, and the standardized results are obtained. Then three common interval dividing methodsare used as shown in table 2 to distribute the sample students into different levels respectively. The number of students in each level is shown in table 2. Finally, the sums of the students' number with standardized results are compared with three dividing methods to get coupling factor, i.e. the higher the coupling factor is, the closer the standardized results are.

TABLE 2 CONTRASTS OF SCORES STEPS

	Standardized	Method 1		Method 2		Method 3	
		Way	number of people	Way	number of people	Way	number of people
Excellent	318	80↑	525	85↑	134	82↑	325
Good	593	70↑	1138	75↑	912	75↑	721
Medium	715	60↑	259	65↑	855	70↑	617
Fail	299	60↓	3	65↓	24	70↓	262
Coupling	--	39.32%		69.14%		84.05%	

As is shown in table 2, the coupling factor in method one is too low, which means the number of people distribution is not reasonable, and the coupling factor in method three is higher than the others.

Although there are other methods in which the coupling factor is over 90 percent, it should be pointed out that even though in mathematical sense the higher coupling factor is better, in order to make the mining results with more practical value, this paper still needs to consider other objective factors, such as that the number of distribution should be in accordance with normal distribution, that the score interval should be integer, and generally it should not be less than 5 points, and other factors. So this article chooses method three as the main standard, in which >82 is excellent, 75-82 is good, 70-75 is medium, and <70 is fall. The results fulfill all the requirements of high coupling factor, normally distribution, and the interval difference which is not less than 5 points.

IV DATA MINING BASED ON CLEMENTINE

A. *Common mining methods*

DM can set up six models: classification, regression, time series, clustering, association rules and sequential rules. Classification and regression are mainly used to make predictions, and the association rules and sequential rules are mainly used to describe the behavior. Clustering can be used for both purposes [4-5].

The mining methods and algorithms used in this paper are C5.0 decision tree algorithm and Apriori association rules algorithm [6].

B. *Software implementation of data preparation*

SPSS Clementine itself has a powerful data processing function, and it is helpful to complete the data preparation work [7-10]. This article illustrates how to use Clementine function of "grouping derived new variables" to realize the discretization. Figure 4 shows the data building flow.

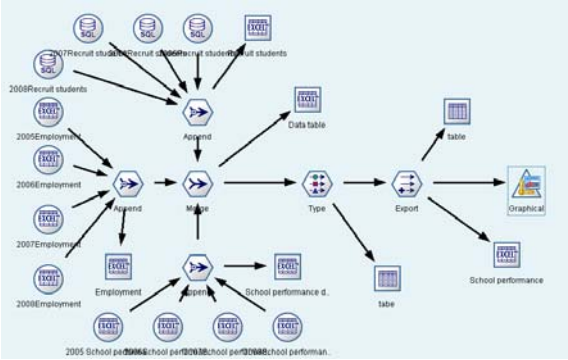

Figure 4 Data flow diagram

The data processing in figure 4 can be disassembled into four steps:

- To use “addition” function to integrate the data of admission, employment and student status in different years into their summary tables respectively.
- To use “merge” function to combine three summary data into one total table.
- To set and filter the data types and fields in the total table.
- Finally, to complete the discretization of student achievement. The settings for the rules of function are shown in figure5, and the discretization results are shown in a form of column diagram in figure 6.

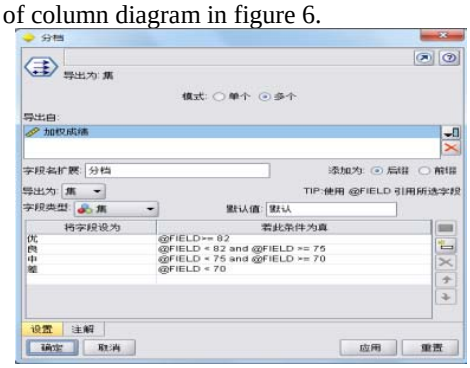

Figure 5 The rules of function setting

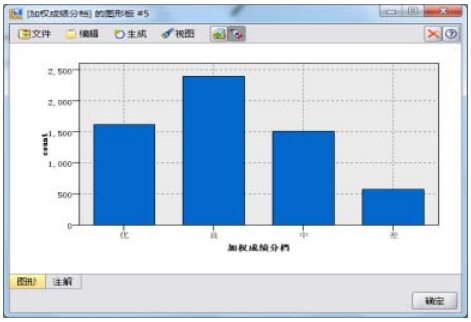

Figure 6 Results in the form of column diagram

C. *Data mining by using Clementine*

Based on the two goals mentioned above, this paper adopts C5.0 decision tree algorithm for student grades in university and Apriori association rules algorithm for graduate students to run the data mining with Clementine[11-13]. The data mining framework is shown in figure 7.

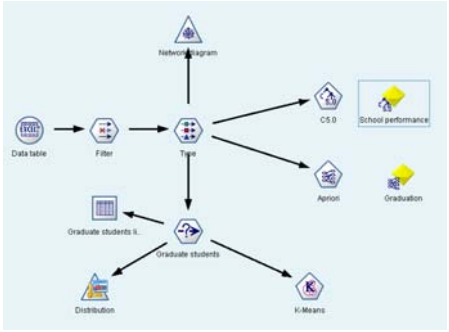

Figure 7 Data flow diagram of DM

With students grades as the goal, the gender, the examinee category, middle school level, cast archives volunteer, major volunteer, the university entrance exam score, the origin of student and major adjustment permission or rejection as the inputs, the decision-making tree is got as shown in figure 8, and the rule set is got as shown in figure 9.

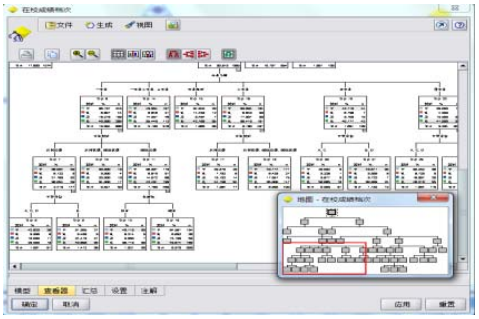

Figure 8 Shown of the decision-making tree

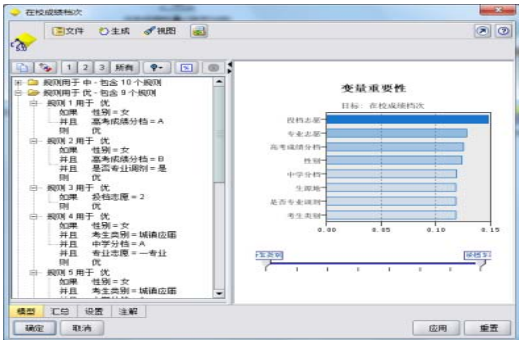

Figure 9 Shown of the rule set

The decision-making tree shows that the excellent scores are usually from female students, and the medium are fail are mainly from male students. Therefore, it is necessary to carry out specific work in the schools, in which the male students take a high percentage, in the future management to improve the scores of the male students. Besides, the students who are enrolled in their first three major volunteer

admission have higher scores, which means that the students who have been admitted into their ideal major have stronger interest and motivation. So we might need to increasestudents' propaganda for professional cognitive efforts to improve the first professional volunteer acceptance rate in the future admission work. Additionally, most of the students with an excellent or good score come from a level A high school, and it is obvious that the students from level A schools might have higher learning literacy, study capacity and general proficiency. Thus, we need to increase the propaganda in such high schools to get more high-quality students. And on the other hand, the male students from level D high schools, and the students with a D score in the university entrance exam in the past usually get unsatisfactory scores, and we also need to pay more attention to them. And those agricultural registered permanent residence students, whose university entrance exam score is not high either, will have higher scores in college due to the economic pressures. Meanwhile, there is no obvious correlation between the scores and whether the students come from the suburbs.

With graduate destination as the consequent, and the examinee category, middle school level, cast archives volunteer, major volunteer, the university entrance exam score, the origin of student, major adjustment permission or rejection, the marks at school, and the fail courses as the antecedent, the minimum support is set to be 10, the minimum confidence is set to be 30, and then the mining of association rules is made as the figure 10 shows:

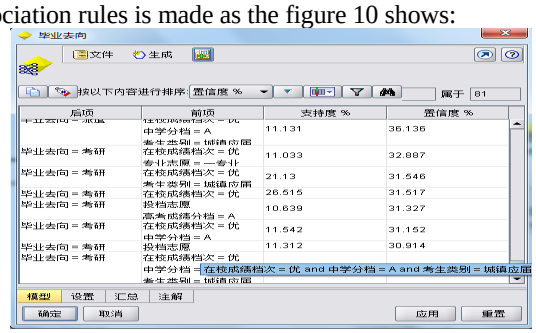

Figure 10 Results shown of association rules

The results of association rules shows that 30% of the graduating students who has excellent scores, come from level A high schools, or come from towns successfully, and 33% of the students who are enrolled in their first major volunteer admission and have excellent scores in university, pass the master admission exams. So we can tell the influence of good scores in college, good hight schools and ideal major admission on the success rate of passing the master admission exams are significant. For the students who come from suburbs and whose scores in the university entrance exam are C, 81% of them choose to work. For the graduating students who come from town and whose scores are medium or fail, 76% of them choose to work. And there is no specific trend for the students who choose to go abroad, and this might be due to the facts that whether the student chooses to go abroad or not is associated with family economic conditions, or the low proportion of oversea students leads to the result that the trend is submerged. Moreover, there is little effect of students' gender on the employment.

Based on the above mining, valuable information for admission work can be concluded: 1. In order to achieve the goal to improve the scores and the success rate of passing the master admission exams, the university should carry out the admission propaganda work in depth aimed at level A high school, strengthen the construction of base school, and therefore get more high-quality students. 2. In the admission propaganda work, the university should emphasize and strengthen the major cognition or adjust the admission policy to improve the ideal major volunteer acceptance rate. And also the university can be more flexible with the major transfer and try to meet the demands of students to study in their ideal majors as much as possible, so that positive impacts will be on school scores and the success rate of passing master admission exams. 3. in the student management work, for the majors in which male students' percentage is high, the management should be strengthened and specific aimed work should be carried out to improve their grades in college and make them more competitive after graduation.

V CONCLUSION

This paper will apply the DM technology to the admission, student status and employment data of Beijing University of Technology, and use decision tree algorithm and association rules algorithm to dig up large quantities of knowledge which is beneficial to the admissions decisions, in order that the huge and uninteresting "data tomb" is converted into the "data gold", and strong scientific basis is provided to make student management work and admission students' propaganda more in-depth and niche targeting in the future. So the research of this paper has a value of practical importance. During the research, some experiences can be summarized from this paper: 1. the data preparation work is huge, the quality of data processing affects the mining results directly, and therefore we need to continue to explore the scientific methods and verify the results through a lot of experiments. In this paper, the mining results are proof from the side for the reasonability of using discretization processing method to deal with the university entrance exam scores and marks in school. 2. Data processing also needs to be considered based on the objective reality of the environment in research subject and the experiences of the staff, in order to improve the practicability of mining results. 3. For the huge amounts of source data, it is also necessary to create data warehouse to improve the efficiency of data mining, which proposes higher requirement for the future research in this article. 4. C5.0 and Apriori algorithm are capable of digging up effective information for the research subject in this article, but some problems also exists during the experimental process. This paper will do a more in-depth study on the improvement of mining algorithm based on the practical situation of the subject.

REFERENCES

[1]P.XIONG. Data mining algorithms and Clementine Practice. BeiJing: Tsinghua University press.2011.2-3

[2]L.WANG,L.ZHOU,H.CHEN,Q.XIAO. The principle and Application of Data warehouse and Data mining. BeiJing: Science Press.

[3]F.ZHOU Etc. Application of mathematical statistics. Renmin University of China press.1989

[4]B.C.XIE.Data mining Clementine application. Machinery Industry Press.

[5]J.W.HAN, M.Kamber. Data Mining: Concept and Techniques. Beijing: China Machine Press, 2001

[6] X.XU. Data mining algorithm and application. Beijing: Peking University Press.2007：194.

[7]Clementine 12.0 system help document.

[8]W.XUE,H.CHEN. Data mining based on Clementine. Renmin University of China press.

[9]C.G.YUAN. The principle of data mining and SPSS Clementine Application collection. BeiJing: Electronic Industry Press.2009.

[10]C.L. YU,B.X.LIU. Application of Association Rules and Cluster Analysis in Supermarket. Development and application of computer.2012.8.

[11]Y.C. SONG, Y.F.FANG. Application Research of Association Analysis with Clementine. The 2nd International Conference on Software Engineering and Data Mining.2010.

[12]X.Z.FENG,X.H.HE,B.L.FENG,A Study on Association Rules Mining of Adverse Drug Reactions. 2010 International Forum on Computer Science-Technology and Applications.

[13]Y.C.SONG, Grady M J O, Hare G M P O. Applications of Attributes Weighting in Data Mining. In proceedings of IEEE Cybernetic Systems Conference, IEEE Publishers, 2007: 41-45.