早稲田大学大学院情報生産システム研究科

博　士　論　文　概　要

論　文　題　目

Study on Student Classification Using Logistic Learning Curve and GNP-based Class Association Rule Mining

申　　請　　者

Rong ZHANG

情報生産システム工学専攻
ニューロコンピューティング研究

2013 年 12 月

In order to solve the problem of seriously unbalanced distribution of students' scholastic competence, remedial education programs have been established in many Japanese universities to help students to compensate for their weaknesses in the study of some common subject areas. But, it has become a very challenging task for teachers to select qualified candidate students for an established remedial course, because there have been very few convincing studies on the issue of student classification using the analytical research. Teaching effectiveness can be improved when students are allocated to the courses which suit their individual ability level and proper learning contents are offered.

The commonly-accepted method for student classification is to divide the students according to their placement test scores, i.e., their original scholastic competence before the course training. But, this method is ineffective for classification, because it neglects the very basic criterion for the education evaluation that students' benefits from a remedial course should always have the maximum priority.

Therefore, the study on student classification in this dissertation focuses on the following: ① Student classification should be conducted based on students' learning progress in addition to their original scholastic competence; ② Students' learning progress can be predicted by differentiating the logistic learning curve based on their pre-course test scores. This is the method proposed for for student classification when students' pre-course test scores are based; ③ GNP-based class association rule mining is proposed for the prediction of students' class labels based on the extracted class association rules. This method suggests that student classification can be conducted through class association rule mining using inquiries; ④ Advanced GNP-based class association rule mining is proposed to improve the classification accuracy using the attribute selection process through Genetic Algorithm (GA) and with consideration of the negative attributes.

Based on the above, this dissertation manages a rigorous control over student classification and realizes the integration of mathematical and engineering skills into pedagogical instructions in the field of engineering education. The essence of this dissertation is as follows:

Chapter 1 introduces the background and objective of the analytical study by highlighting the critical significance of conducting student classification based on their learning progress in remedial education programs. Previous studies on student classification suggest that an intensified study is in urgent need, because the commonly-accepted method doesn't pay enough attention to the issue of how to help students to achieve their best from the course study. This study attempts to realize student classification through predictions on students' learning progress and their class labels before the course study by using the differential curve of the logistic learning curve and GNP-based class association rule mining in order to make sure that the contents of a remedial course fit the individual level of every candidate student.

Chapter 2 conducts fundamental statistical analysis for the 240 sample data collected from the webbased English remedial education program adopted by Nishinippon Institute of Technology (Japan). Pre-course testing and post-course testing were held. The changes in students' test scores, called post-course score changes (PSC) in this study, are used as the index for the measurement of their progress in learning. The optimized prediction model for post-course score changes is obtained through multiple regression analysis and pre-course test score (PRTS) is found to be the only decisive attribute for post-course score changes (PSC). Based on the advice of experienced teachers of English, 10≦PSC is considered as a desirable criterion for student classification. Therefore, 48 point of PRTS is proposed to be the classification point for student classification. That is, students with 20≦PRTS≦48 are qualified students because their PSC are predicted to be higher than 10 points according to the optimized prediction model; On the contrary, students with 48<PRTS<100 are unqualified students because their PSC are predicted to be less than 10 points. The prediction accuracy is 49.2% for the training data and 46.7% for the testing data, which is found to be not as impressive as expected.

Chapter 3 proposes the differential curve of the logistic learning curve to improve the relationship between students' pre-course test scores and their post-course score changes. In this study, the logistic learning curve and its differential curve share the same horizontal axis - pre-course test scores. The vertical axis of the logistic learning curve represents total amount of knowledge students have mastered during the course learning, while the vertical axis of its differential curve stands for their post-course score changes, i.e., students' learning progress. After fitting the testing data to the various forms of learning curves, the differential curve of the logistic learning curve is chosen to express the relationship between pre-course test scores and post-course score changes. Based on the same criterion used in Chapter 2, 61 point of PRTS is proposed to be the classification point for student classification. That is, students with 20≦PRTS≦61 are qualified for the remedial course and students with 61<PRTS<100 are unqualified candidates. The prediction accuracy is around 70% for the training data and around 65% for the testing data.

Chapter 4 discusses the possibility of conducting student classification through data mining using inquiries. GNP-based class association rule mining is used for examining the impacts of attributes related to students' motivation and attitudes in learning on their post-course score changes. GNP (Genetic Network Programming) is an effective technology in class association rule mining and has a superior ability in handling a large amount of information. It generates a sufficient number of important class association rules for classification by using a compact directed-graph structure and never causes the bloat. GNP-based class association rule mining is used in this research, because important attributes and their combinations, which have significant influences on students' progress in the remedial learning, can be discovered through the extracted class association rules. Important attributes and their combinations are included in those extracted rules with a strong explanation ability about the features of both qualified students and unqualified students in the remedial course study. The simulation result shows that the student classification accuracy is 56.5% for the training data and 51.2% for the testing data when the conventional GNP method is applied.

Therefore, in order to increase the classification accuracy, Optimal Attribute Subset through Genetic Algorithm (OASGA) is proposed by integrating the attribute selection process through Genetic Algorithm (GA) into the conventional GNP-based class association rule mining. Attribute selection is a technique to prune less relevant information and discover high-quality knowledge so that the predictor attributes given to the mining algorithm become more relevant to the class attribute. When the attribute selection process through GA is added to the conventional system, class association rule mining is conducted with a small data subset rather than the original large number of attributes, the function of GNP has been enhanced by the attribute selection process through GA. Thus simple but important rules are obtained and higher classification accuracy is available. The simulation results show that the average classification accuracy has been largely improved to around 71.9% for the training data and 64.7% for the testing data.

Chapter 5 reports another attempt to promote the classification accuracy by taking into account of the negative class association rules. The original database is amplified by adding the negative aspect of all the predictor attributes. When class association rule mining through OASGA is conducted using the amplified database, important class association rules are available with the negative rules being included. Thus, the extracted class association rules give a more accurate description about the features of students both in the qualified class and the unqualified class. The simulation results show that the prediction accuracy for student classification is improved to around 74.5% for the tranining data and 70.2% for the testing data. Detailed analysis of the frequently appearing attributes and their combinations are also carried out in this chapter.

Chapter 6 summarizes the whole dissertation. This study focuses on the issue of study classification, i.e., how to select qualified students for an established English remedial program. Consequently, two methods are proposed for the selection process: (1) When the pre-course test scores are basically used, student classification can be conducted by using the differential curve of the logistic learning curve. (2) When inquireies are available, student classification can be conducted based on the extracted class association rules obtained through GNP-based class association rule mining.

Compared with the commonly-accepted method for student classification, this dissertation serves as the first study to examine the issue through scientific streamlining and structural optimization rather than focusing on their original scholastic achievement before the course learning. Although there are still limitations in this study and future work is needed to refine the technical skills, the results have highlighted the application of engineering technologies to practical teaching activities and will throw light on the future research on student classification.