Identifying Influence Factors of Students Success by Subgroup Discovery

F. LEMMERICH AND M. IFLAND AND F. PUPPE

University of Würzburg, Germany

The identification of influence factors on students' success rate is of high interest. Especially knowledge about factors, which can be determined before the start of a student's degree program, is important, since most drop-outs occur in the very first semesters. In this paper we show, how the data mining technique of subgroup discovery can be utilized to identify such influence factors on the overall success of students. Additionally, we also discuss several interesting measures for this purpose..

Key Words and Phrases: Subgroup Discovery, Pattern Mining

1. INTRODUCTION AND METHOD

In order to avoid drop-outs in degree programs, early knowledge about influence factors on students' success is a key issue. In this case study, we show how the technique of *subgroup discovery* can be applied to identify important combinations of factors, which are known before students start their degree program, e.g., age, sex, regional origin or previous activities.

Subgroup discovery [Klössgen 1996] aims at finding *descriptions* (conjunctions of attribute-value pairs) of subsets in the population that show an interesting behavior with respect to a certain target concept, e.g., the drop-out rate. The results can be interpreted as a set of association rules, in which the consequent is always the target concept and the antecedent is the description of the respective subgroup. For example, the subgroup with description *age = '20-21' ∧ previous_courses = false* can be reformulated for a search with the target concept drop-out as: *age = '20-21' ∧ previous courses = false* → *drop_out = true*.

In a subgroup discovery task the top rules with respect to a predefined measure of interestingness, see [Geng 2006], are returned. The most important quality measures q can be formulated as: $q_a\ (sg) = n^a\ (p\text{-}p_0)$, where n denotes the size of the subgroup sg, that is, the number of individuals, for which the description of sg applies. p is the share of the target concept in the subgroup and p_0 the share of the target in the total population. The parameter a weights the sizes of subgroups with respect to the deviation in the target share. For example, for $a=1$ the resulting quality function $q_1\ (sg) = n\ (p\text{-}p_0)$ equals the *Weighted Relative Accuracy*, for $a=0$ we get the *Added Value* $q_0\ (sg) = (p\text{-}p_0)$, which has been applied to evaluate association rules in the educational domain [Merceron 2008]. In addition, we will use adapted *Relative Measures*, cf. [Grosskreutz 2010], which replace the target share in the total population p_0 in the above formula by the maximum target share in any generalized subgroup, i.e., subgroup descriptions, that contain only a subset of describing attribute-value pairs: $r_a\ (sg) = n^a\ (p - max_i\ (s\_i))$. Thus, a specialization is only considered as interesting, if it significantly increases the target share with respect to all its generalizations.

2. CASE STUDY AND DISCUSSION OF RESULTS

For our experiments we first extracted the data of the students, from the multi-relational warehouse into a tabular form, resulting in a total row count of 9400. In additional preprocessing steps attributes were discretized and categorized. The resulting attributes included: drop-out, sex, age at the start of the degree program, school final exam grade,

Authors' addresses:
Universität Würzburg, Fakultät für Mathematik und Informatik
Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik (Informatik VI)
Am Hubland, D-97074 Würzburg

nationality, a flag for previous courses, which were passed before its start, a flag for previous degree programs at this university, and the categorized place of final school exams. The overall drop-out rate in the population was 26.3%. Important single factors, that increase the likelihood of drop-outs are for example *age > 30*, *foreigner=true* (both 40% drop-outs.) or *age < 20* (39% drop-outs).

To find interesting combinations of influence factors, we performed subgroup discovery using the interesting measures q_0, $q_{0.5}$, q_1 and their relative counterparts r_0, $r_{0.5}$, and r_1. As an example, the top resulting subgroups for descriptions with two influence factors are presented below with their basic statistics and rankings according to the different interesting measures. Statistics for the respective generalizations are given below each subgroup. For example, the last three rows indicate, that there were 4394 students not originating in the universities state. 28% of these were drop-outs. The drop-out rate for the 4668 male students was equally at 28%. However, the 2066 students, for which both these influence factors apply, had a drop-out rate of 32%. This subgroup was ranked at 14 for the interesting measure q_0, but was ranked first for the interesting measure r_1.

SG	Description	size	target share	q_0	$q_{0.5}$	q_1	r_0	$r_{0.5}$	r_1
S_1	grade < 1.5∧ age = [26;30]	23	61%	1	-	-	1	1	-
	grade < 1.5	716	35%						
	age = [26;30]	619	21%						
S_2	Prev._course = false ∧ grade = ‘?’	861	44%	4	1	1	14	7	8
	Prev._course = false	9027	27%						
	grade = ‘?’	986	42%						
S_3	Origin_in_state = false ∧ sex = male	2066	32%	-	14	5	13	2	1
	Origin_in_state = false	4394	28%						
	sex = male	4668	28%						

The results show significant differences depending on the used interesting measure. As expected, the measures with a higher parameter *a*, i.e., q_1 and r_1 prefer larger subgroups, while those with a lower parameter *a* prefer smaller subgroups with an higher deviation of the target share. The deviation of the target share for combinations of influence factors that result from the absolute measures can sometimes be almost completely be explained by one factor alone. For example, the high deviation of the target share in subgroup S_2 can be almost completely explained by its generalization *grade = ‘?’*. Therefore, we consider this combination of influence factors as less interesting. Such subgroups are ranked significantly lower by utilizing relative measures.

For the result presentation to the head of degree programs we chose subgroups resulting from the interesting measure $r_{0.5}$, as it provided a nice balance between large subgroups and high deviation of the target share. The project received positive feedback and continues to regularly report results each semester.

REFERENCES

GENG, L., Hamilton, H.J. 2006. Interestingness measures for data mining: A survey. ACM Cmp. Surv. 38, 3, 9.

GROSSKREUTZ, H., BOLEY, M., AND KRAUSE-TRAUDES, M. 2010. Subgroup Discovery for Election Analysis: A Case Study in Descriptive Data Mining. In: *Discovery Science*, B. Pfahringer, G. Holmes, and A. Hoffmann, Eds. Lecture Notes in Computer Science Series, vol. 6332. Springer, 57-71.

KLÖSSGEN, W. 1996. A Multipattern and Multistrategy Discovery Assistant. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press, 249–271.

MERCERON, A. AND YACEF, K.. 2008. Interestingness Measures for Association Rules in Educational Data. In: *EDM 2008: 1st International Conference on Educational Data Mining, Proceedings*. 57–66.

ROMERO, C., GONZALEZ, P., VENTURA, S., DELJESUS, M., AND HERRERA, F. 2009. Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. Expert Systems with Applications 36, 2, 1632–1644.