The Future of Education
International Conference
The Future of Education
4th Edition

Data Mining Techniques for Detecting Behavioural Patterns of Gifted Students in Online Learning Environment (Case Study)

Zdena Lustigova
Faculty of Mathematics and Physics, Charles University in Prague (Czech Republic)
lustigo@plk.mff.cuni.cz

Abstract

Presented article deals with selected results of data mining analysis of psychological and educational data, collected within 5 years (2006-2011) on the group of 91 children, who participated in a specific online program for gifted children, organised by faculty of Mathematics and Physics, Charles University in Prague, Czech Republic. The information record for each case (student) is represented by 151 variables, of both categorical and metric nature, describing 1) personal characteristics, including motivation and intelligence 2) behavioural and action records, including individual decision making records, and 3) particular educational results and learning path.

The study is based on comparison of students with very similar personal characteristics like motivation and intelligence and their study results. The success in a selected online course seems to be related to the nature of particular student talent. Student, who had chosen the right courses, that matched his/her nature of talent, succeed. An individual, who for some reason chooses the course that does not match his /her nature of talent, is more likely to fail.

1. Introduction

1.1. Data mining among gifted children

The education domain offers a fertile ground for many interesting and challenging data mining applications. These applications can help both educators and students to improve the quality of education. In [2] Ma, Liu, Wong and others deal with Gifted Education Programme (GEP) of the Ministry of Education (MOE) in Singapore. They focus mainly on better selection of students for remedial classes, since traditional methods choose too many participants, which increase the teaching load of the instructors and slows down the real GEP students.

Identification of gifted children using neural networks within the online environment is also the topic of the paper [3], where Bae, Ha and Park offer a special questionnaire and use it to measure the implicit capabilities of giftedness and to cluster the students with similar characteristics. The neural network and data mining techniques are applied to extract a type of giftedness, their characteristics, and their learning path. Kamath and Srimani in [4] deals with gifted children performance analysis using generic data mining model, that validates the accuracy and efficiency of the learning model and leads, according the authors, to more reliable and authentic predictions. Cluster analysis as a specific technique was used few times by Parker [5] and [6] to identify perfectionism. A nationally gathered sample of 820 academically talented sixth graders took the Multidimensional Perfectionism Scale, and scores were cluster analysed using both hierarchical and non-hierarchical cluster analysis with cross-validation. A three-cluster solution was indicated. Parent perceptions of the children were consistent with the students' self-perceptions. The construct of perfectionism was primarily associated with conscientiousness and secondarily with agreeableness and neurosis. Both studies did not identify statistically significant differences between gifted students and the general cohort.

2. Research methodology

2.1. Data collection methods

The data were collected within 5 years (2006-2011) on the group of 91 children, participating in the specific online program for gifted children in the Czech Republic.

The collected information record for each case (child) is represented by 151 variables, both categorical and metric nature, describing on one side 1) personal characteristics, including motivation and intelligence (set of specialized psychological tests, e.g. LMI The Achievement Motivation Inventory, etc.,); on the other side 2) behavioural and action records, including individual decision making records (e.g. course set selection, preferences in teacher selection, etc.) and 3) their learning path based mainly on particular educational results.

The whole data set included more than 12000 data cells both measured and recorded. Approximately one eighth (12%) of expected data set (13740 values) was unavailable (1740 missing data cells). The data unavailability reasons were mainly external, caused by parents' awareness of their child's IQ or personality tests, or by lack of teachers' online records, concerning particular student's educational results.

2.2. Data processing methods

Cluster (cluster) analysis was used to identify homogeneous subgroups of variables or cases in a given population of all 151 variables and 91 gifted students.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields and areas of research.

Cluster analysis is used wherever the researcher does not know the number of groups in advance, but wants to identify and then to analyse group membership.

Within the study we used just two from existing three general approaches to cluster analysis, namely hierarchical clustering and two stage cluster analysis.

2.3. Selected research results

Case analysis

Approximately 12 research questions were set up, focused mostly on the existence of a connection between 1) Motivational characteristics of students and their

a) course selection,
b) study results or intelligence (as a whole) or particular components of intelligence,
c) gender.

Other research questions were focused on 2) the difference in motivation for a) course selection, b) achieved results between boys and girls.

When adding additional variables of courses selected in different years (categorical variables Nos. 6-14) the following interesting connections were revealed:

Basically, there seems to be strong relationship between course topic selection and Dominance.

Biology (and biology connected subjects e.g. environmental studies) were connected with less dominant individuals, while Physics (and physics connected subjects) proved much higher dominance.

In connection with the higher dominance also both programming and math appeared (in particular years), but the relationship was not as strong as in the case of physics.

The variables like “Engagement”, “Confidence in success” and “Flexibility”, did not split the population onto two clusters. These variables probably do not influence the course selection.

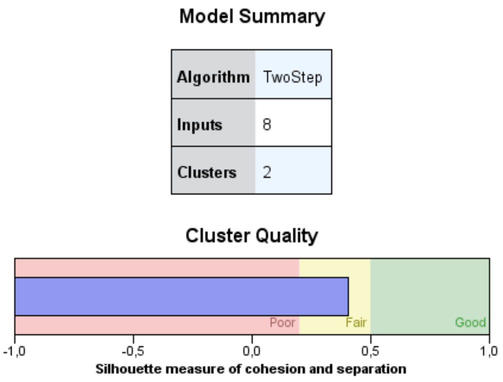
The same zero or weak influence seems to be observed at other variables like Fearlessness, Internality, Compensation effort, Pride in productivity, Eagerness to learn and Preference for difficult tasks.

On the other side, the influence of the variable Independence seems to be strong.

Independence has caused the formation of two clusters in every particular year and during the whole program as well. Its influence and predictive value is similar to that of Dominance.

Again, the first cluster is dominated by biology and biology connected topics, the second cluster is dominated by different items in different years, but mostly connected with physics, followed by programming, and also by math (appeared once).

Effect of variables Autonomy, Status orientation, and Competitiveness was not approved, quite surprisingly. These variables did not cause the appearance of the two clusters.



Model Summary

Algorithm	TwoStep
Inputs	8
Clusters	2

Fig. 1. The quality of the two clusters formatted on the base of Course Selection and variable Dominance

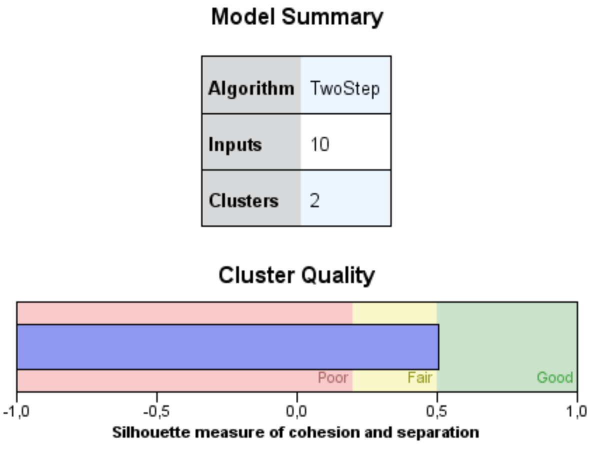
The Goal setting variable proved a relatively high predictive value. It caused the splitting of the Talent population onto the same two clusters as Dominance and Independence did.

Note: Notice all these hypothesis and results should not be overestimated or misinterpreted, as well as the previous. It is necessary to verify them on a larger sample, particularly a sample with a lower percentage of missing values.

2.4. The influence of Gender

The general method used was a two-step cluster analysis, which included variables 6-14 as categorical variables and variable gender as well as categorical variable.

Under these conditions the two clusters of a relatively good quality appeared. Variable gender was identified as the most important predictor.



Model Summary

Algorithm	TwoStep
Inputs	10
Clusters	2

Fig. 2. The quality of the two clusters formatted on the base of Course Selection variables 6-14 and variable Gender (No 15).

Cluster 1 includes only girls, while in Cluster 2 we can find all boys and some girls, however.

Of these variables is the first jurisdiction in the most significant predictor of a particular cluster of courses is therefore gender. The following interpretation of both clusters is not entirely accurate, but it is understandable for laics.

Students in Cluster 1 selected within the large offer of topics those, connected with biology, chemistry, and ecology. They could also choose no course for a particular year. Learners included into the 2nd cluster primarily chose physics, mathematics and programming.

Gender significantly influences the selection of courses. However, it is not a single factor. In the following we will try to find the other factor (or factors) that caused the appearance of a significant percentage of girls in the second cluster. The likely hypothesis is that this factor may be either the total IQ, or any of its ingredients. It may also be one of the "components of motivation” or some combination of motivational and intellectual components.

2.5. The influence of the intelligence

The general method used was two steps cluster analyses, Included variables 6-14 as categorical variables and variable 21 (IQ) as a continual variable, further other components of Intelligence (variables 22-28).
All components of intelligence, as well as the IQ score, with the exception of crystalline, caused the formation of two clusters. Their cluster quality was not high. The most significant effect had verbal intelligence and numerical. The weakest influence on the choice of courses was revealed with IQ knowledge and absolutely no with crystalline IQ.

2.6. The influence of motivational factors on Course topics selection

The effect of variable Persistence
Variable persistence did not cause the two clusters formation, but not unnoticed should remain the fact that while omitting the variables 13 and 14, which relate to courses in the academic year 2010-2011 (problematic), the effect of persistence is relatively high.
The effect of variable Dominance
Dominance has caused the formation of two clusters in all studied years and in Talnet as a whole. Into cluster 1 fall mostly students who chose either no course or biology connected course, the second cluster includes all others (see figure below). This does not mean that all those who have chosen biology are submissive individuals. In 2009, for example, the first cluster included almost everyone, with the exception of physicists.
The effect of variables Independence and Goal setting
Independence and Goal setting both have caused the formation of similar two clusters in all the years particularly and within the experiment as a whole. Their influence and predictive values are similar to that of dominance.
Again, the first cluster is dominated by "no course" (or biology), in the second cluster is the dominant item physics, but also programming (and sometimes math) occurs in certain years.

3. Discussion

Dominance and Independence have lower predictive value than Goal setting. It is probably due to the large number of participants who did not choose any course. Since the column of those who did not participate in any run, in each histogram always dominant (highest column), the Goal setting has expectedly high predictive value.
Based on the above presented analysis, we can make the hypothesis that dominance and independence are rather related to the selection of a particular type of course, while Goal setting motivational parameter is probably more connected with " no course " than to "some (any) course” selection.
To verify the hypothesis we would need a sample with a smaller number of missing values, and, in particular, information on whether a student really did not choose any course, or whether he finished his studies due to lack of interest or because of exceeding the age limit (e.g.to replace missing data values with non-zero meaningful values).
Note: In the current quality of data, a high percentage of missing values (almost 12%) could influence or even distort some of the results of this challenging analysis.

4. Conclusions

The course selection (and its prediction) is affected mostly by:

a) Intelligence and all of its components except of crystalline. The most significant impact was observed in case of verbal and numerical component of intelligence.

b) Motivation, but only by the following components: Dominance, Independence, Persistence and greatly by Goal setting, which affects especially the fact that a student chooses a course at all or not. The course selection is not influenced by TKM and its various components.

It can be assumed that the analysis of the variables split the Talnet population into two clusters, where the dominant, but not sole role gender plays. The remaining factors were especially Dominance, Independence, Goal setting (or Persistence) and verbal and numerical intelligence.

The success in a selected course is also related to the nature of particular student's talents. If the students chose the appropriate courses (courses that matched their nature of talent), they succeed. An individual who, for some reason, chooses the course that does not match his/her nature of talent, is more likely to fail.

The nature of talent seems to be determined by:

a) all components of intelligence, with the exception of crystalline,

b) selected components of motivation: Dominance, Independence and Persistence,

c) selected components of creativity (Elaboration, Fluency, Flexibility).

References

[1] Mooi,E., Sarstedt,M.: A concise guide to market research. The process, data and Methods Using IBM SPSS statistics. 2011, XX,307 p, ISBN: 978-3-642-12540-9

[2] Ma,Y.,Liu,B., Wang, Ch. At al: Targeting the Right Students Using Data. Available online ftp://ftp.cse.buffalo.edu/users/azhang/disc/disc01/cd1/out/papers/kdd/p458-ma.pdf

[3] Bae, S., Ha, S.H., Park, S.C.: Identifying gifted students and their learning paths using data mining techniques. Available online http://library.witpress.com/pages/PaperInfo.asp?PaperID=18312

[4] Srimani, P.K., Kamath, A.: Data Mining Techniques for the Performance Analysis of a Learning Model – A Case Study. International Journal of Computer Applications (0975 – 8887) Volume 53– No.5, September 2012

[5] Parker, W.D.: An Empirical Typology of Perfectionism in Academically Talented Children. Am Educ Res J September 21, 1997vol. 34 no. 3 545-562

[6] Parker, W.D.: The Incidence of Perfectionism in Gifted Students. Gifted Child Quarterly Fall 1996 vol. 40 no. 4 194-199