Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices

Mirka Saarela
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
mirka.saarela@gmail.com

Tommi Kärkkäinen
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
tommi.karkkainen@jyu.fi

ABSTRACT

The Programme for International Student Assessment, PISA, is a worldwide study to assess knowledge and skills of 15-year-old students. Results of the latest PISA survey conducted in 2012 were published in December 2013. According to the results, Finland is one of the few countries where girls performed better in mathematics than boys. The purpose of this work is to refine the analysis of this observation by using education data mining techniques. More precisely, as part of standard PISA preprocessing phase certain scale indices are constructed based on information gathered from the background questionnaire of each participating student. The indices describe, e.g., students' engagement, drive and self-beliefs, especially related to mathematics, the main assessment area in PISA 2012. However, around 30% of the scale indices are missing so that a nonstructured sparsity pattern must be dealt with. We handle this using a special, robust clustering technique, which is then applied to Finnish subset of PISA data. Already direct interpretation of the created clusters reveals interesting patterns. Clusterwise analysis through relationship mining refines the confidence on our final conclusion that attitudes towards mathematics which are often gender-specific are the most important factors to explain the performance in mathematics.

Keywords

PISA, robust clustering, frequent itemset, association rule

1. INTRODUCTION

PISA (Programme for International Student Assessment) is an international assessment programme by the Organisation for Economic Co-operation and Development (OECD) that studies students' learning outcomes in reading, mathematics, and scientific literacy triennially. It is referred as the "world's premier yardstick for evaluating the quality, equity and efficiency of school systems" [21]. More than seventy countries and economies have already participated in PISA. Finland has consistently been one of the top-performing countries in the assessment [11]. Each time the study is repeated the main learning outcome focus area changes. In the latest assessment (PISA 2012) it was mathematics. A database of the results is publicly available[1].

One general key finding from PISA 2012 was the gender difference in mathematics performance: On average, boys outperform girls in mathematics. Finland, however, is, according to the assessment, one of the eight countries where girls perform better than boys in mathematics: The mean score of girls in mathematics was 520 while boys had the mean score of 517 [23]. Despite the slightly better performance in mathematics women are, also in Finland, underrepresented in mathematics related jobs [28].

The purpose of this work is to apply educational data mining approch and corresponding techniques to study the performance of Finnish student population in mathematics, focusing especially on gender-related findings. As part of standard PISA preprocessing phase, certain *scale indices* are constructed based on information gathered from the background questionnaire for each participating student [21]. These indices describe, e.g., students' engagement, drive and self-beliefs, especially related to mathematics. However, around 30% of the scale indices are missing due to lack of reliable student responses for the background questions. This means that the knowledge discovery process is realized with data having a nonstructured sparsity pattern. We handle this using a special, robust clustering technique as proposed in [4]. Furthermore, the clustering result obtained is further analyzed using itemset mining [1] to foster the generation of novel information and new knowledge.

The contents of the paper is as follows: First, we provide a short summary on PISA data and how students' capabilities and attributes are presented. We then describe a certain set of scale index variables that are associated with the performance in mathematics. Subsequently, we apply methods from two (see [7] for a complete categorization) of the main branches in educational data mining. In Section 3, we utilize a special clustering approach to find groups of students with similar characteristics with respect to scale indices. In order to further refine the characterization of student groups, we then apply frequent itemset mining and association rule learning to selected clusters in Section 4. Finally, we sum-

[1]See `http://www.oecd.org/pisa/pisaproducts/`.

marize and conclude our study in Section 5.

2. ON PISA DATA

We apply educational data mining for the PISA 2012 data subset of Finland. In each country participating PISA, the schools and students selected for the survey reflect the whole population and characteristics of the educational context. In Finland, 311 schools and 10157 students from these schools were sampled for the assessment in 2012. Out of the sampled students 8829 participated in the actual PISA test. Hereby, each student that takes part has to (i) solve a set of cognitive items/tasks and (ii) fill out one background questionnaire[2] with demographic questions.

Finnish PISA data is stored in two different data sets: One data set includes all the students that participated in the test, and the second one includes all sampled schools. The student data set has more than 600 variables. A set of those variables directly encode the students answers given in the background questionnaire. Moreover, since the participating students should reflect all 15-year-old students in Finland, certain weights are assigned to each student to align the sample with the true population. In PISA reports and learning analysis, student abilities are not given as direct responses to task questions but in the form of the so-called *Plausible Values* (PVs).

Since a very broad domain of knowledge and skills should be tested but the testing time for each student is limited, only certain subsample of students respond to each item/task. In order to reliably compare results of different students, even if they have not answered exactly to the same set of items, PISA uses a generalized form of the Rasch Model [19]. Depending on how many students have solved a task correctly, a certain "difficulty value" is assigned to each tasks and depending on how many tasks a student solved, a certain "competence value" is assigned to each student. PVs are estimated based on difficulty and competence scores and then scaled so that the OECD average in each domain (mathematics, reading and science) is 500 and the standard deviation is 100.

Usually, five PVs are drawn from each student's competence distribution for each main assessment area to describe the performance. For instance, in the Finnish data set for 2012 we have have five PVs for each student in reading, science, and mathematics. Moreover, since mathematics was the main assessment area, five PVs for each of the 7 subscales, i.e. subtopics in mathematics (change and relationship, quantity, space and shape, uncertainty and data, formulate, employ, interpret) are enclosed.

2.1 PISA Scale Indices

PISA scale indices (see Table 1) are derived variables based on information gathered from the background questionnaires. The scale indices are constructed in order to better characterise students dispositions, behaviours, and self-beliefs. Indeed, many of the self-reported indicators of engagement in school are strongly associated with the performance in mathematics. Especially, the *index of economic, social and cultural status* (ESCS) explains 46% of the performance variation among OECD countries so that a socio-economically more advantaged student scores 39 points higher in mathematics[3] than a less advantaged student [20]. Furthermore, according to [19], the ESCS is the "strongest single factor associated with performance in PISA".

Table 1 provides an overview of the PISA scale indices used in this study. In the first two columns, we provide the name of the index and it's abbreviation used in the data set. It should be noted that some indices emphasize negative orientation with respect to mathematics. For example, it usually is not beneficial to the performance in mathematics if a student has a high value in the index which measures the anxiety towards mathematics (ANXMAT). Each index in the PISA data is standardized to have mean zero and scaled to have standard deviation one across OECD countries. Hence, a positive score index does not necessarily mean that a student has replied positively to the corresponding questions but that the answers are above the OECD average.

Correlations between the scale indices and the overall performance in mathematics are provided in the third column in Table 1. In the fourth column, ranking of the correlations based on their absolute values is given. We notice that the three indices having highest linear relationship with performance in mathematics are mathematics specific whereas the fourth index in ranking describes readiness for problem solving, and only the fifth one is the already mentioned status indicator ESCS. The correlations are computed using the subset of Finnish students for which a particular index is available. In order to obtain reliable estimates we have, as recommended in [19], analyzed each PV separately. This means that we have first computed five correlation coefficients and then used their mean as the actual result.

As already observed, not every student in the data set has a value for each of the indices. In fact, 33.24% of the index values are missing/invalid. There are different reasons why a specific scale index for a particular student is unusable. First of all, not all background questions were administered to all students. Students, that were not administered the questions included in the index had missing value by design. Second of all, it might be that the student got the questions but did not answer them. Finally, a reason for a missing index value can be that questions were answered but answers were found to be unreliable or invalid in manual scanning.

3. CLUSTER ANALYSIS USING ROBUST PROTOTYPES

Clustering is an unsupervised data analysis technique, where a given set of objects is divided into subsets (clusters) such that objects in the same cluster are similar to each other and dissimilar to objects in other clusters. Even if this appears as a simple rule, there are many approaches for clustering [10]. The classical division of algorithms is the separation into *partitional* and *hierarchical* clustering methods [16, 29]. Hierarchical clustering is usually applied for small data sets since most of the algorithms have quadratic or higher computational complexity [9]. However, the main difference be-

[2]An example of such background questionnaire can be found from `http://nces.ed.gov/surveys/pisa/pdf/MS12_StQ_FormA_ENG_USA_final.pdf`.

[3]39 score points equal nearly one year of schooling.

Table 1: PISA scale indices and correlation to mathematics performance

PISA scale index	abbreviation	corr	rank
economic, social and cultural status	ESCS	0.36	5
sense of belonging	BELONG	0.01	15
attitude towards school: learning outcome	ATSCHL	0.15	11
attitude towards school: learning activities	ATTLNACT	0.08	12
perseverance	PERSEV	0.31	6
openness to problem solving	OPENPS	0.42	4
self-responsibility for failing in mathematics	FAILMAT	-0.20	10
interest in mathematics	INTMAT	0.25	7
instrumental motivation to learn mathematics	INSTMOT	0.23	9
self-efficacy in mathematics	MATHEFF	0.51	2
anxiety towards mathematics	ANXMAT	-0.44	3
self-concept in mathematics	SCMAT	0.52	1
behaviour in mathematics	MATBEH	0.04	13
intentions to use mathematics	MATINTFC	0.23	8
subjective norms in mathematics	SUBNORM	-0.02	14

tween these methods is related to the shape of clusters which is readily amplified in the interpretation of the clustering result. Hierarchical clustering is based on connecting locally similar objects so that the global shape of a cluster can be almost arbitrary. Partitional methods, which rely on creating subsets with respect to global similarities, are quaranteed to produce geometrically closed subsets. Moreover, the special prototype characterizing the properties of all the cluster members provides a well-defined pattern for the interpretation of the clustering result.

Prototype-based partitional clustering methods, such as *k-means*, a popular algorithm utilized also in many EDM studies [30], can be described using an iterative relocation algorithmic skeleton with an explicitly defined score function [12] (see Algorithm 1). Partitional clustering creates a $k-$partion $C = \{C_1, ..., C_k\}$ $(k \leq n)$ of data $\mathbf{X}$, such that

1) $C_i \neq \emptyset$ with $i = 1, ..., k$;
2) $\bigcup_{i=1}^{k} C_i = \mathbf{X}$; and
3) $C_i \bigcap C_j = \emptyset$ with $i, j = 1, ..., k$ and $i \neq j$.

In order to realize a prototype-based partitative clustering algorithm some further issues need to be addressed. First of all, all iterative relocation algorithms search better partitions locally so that the final result depends on the initialization. Although a lot of work has been attributed to this problem, still no universal method for identifying the initial partition exists (actually such an approach would provide an approximate solution to the clustering problem itself). Another main issue is to define the similarity measure that reflects the closedness in the data space. To this end, the amount of clusters must be determined in order to end up with one, final clustering result for the interpretation.

Our data to be clustered is problematic, because there is an arbitrary pattern of missing scale indices to deal with. Such missing values could be considered as extreme outliers because they can have any value from each variable's value range. Hence, second order statistics and least-mean-squares estimates that are sensitive to nonnormal degredations are not suitable, and we use instead the so-called nonparametric, robust statistical techniques and distance measures [15, 27, 14]. Out of the simplest robust location estimates, median and spatial median, we use spatial median due to it's multidimensional nature which allows better utilization of the local/clusterwise available data pattern [17]. Spatial median has many attractive statistical properties and, especially, it's breakdown point is 0.5, i.e. it can handle up to 50% of contaminated data.

In [4], a robust approach utilizing the spatial median to cluster sparse and noisy data was introduced. The *k-spatial-medians* clustering algorithm is based on the algorithmic skeleton as presented in Algorithm 1. As the score function one utilizes

$$\mathcal{J} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \|\boldsymbol{P}_i(\boldsymbol{x}_i - \boldsymbol{c}_j)\|_2, \quad (1)$$

where the last sum is computed over the subset of data attached to cluster j. Here the projections $\boldsymbol{P}_i, i = 1, \ldots, N$, capture the existing variable values of the ith observation, i.e.

$$(\boldsymbol{P}_i)_j = \begin{cases} 1, \text{if } (\boldsymbol{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases}$$

In Algorithm 1, the projected distance as defined in (1) is used in the first step, and recomputation of the prototypes, as spatial median with the available data, is realized using the SOR (Sequential Overrelaxation) algorithm [4] with the overrelaxation parameter $\omega = 1.5$.

3.1 Initialization and Number of Clusters

It is a well-known problem that all iterative clustering algorithms are highly sensitive to the initial placement of the cluster prototypes and, thus, such algorithms do not guarantee unique clustering [18, 9, 6, 16]. Numerous methods have been introduced to address this problem. Random initialization is still often chosen as the general strategy [31]. However, several researchers (e.g., [3, 5]) report that having some other than random strategy for the initialization often improves final clustering results significantly. Having these issues in mind, we developed the following deterministic and context-sensitive approach to find good initial prototypes.

For a subset of 2520 students in the Finnish data, there are

Algorithm 1: Iterative relocation clustering algorithm

Input: Dataset $\mathbf{X}$ with n observations and a given number of clusters k.

Output: A set of k clusters, which minimizes the score function.

Select k points as the initial prototypes;

repeat

1. Assign individual observation to the closest prototype;
2. Recompute the prototypes with the assigned observations;

until *The partition does not change*;

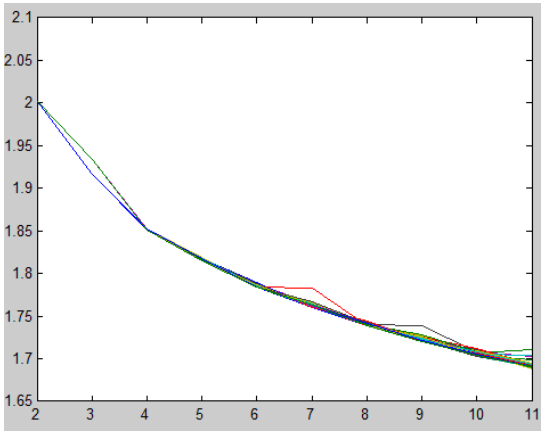

Figure 1: Ray-Turi index for $k = 2, \ldots, 11$

no missing scale index values. For this subset we want to find (i) the most suitable amount of clusters k and (ii) the initial prototypes for the clustering algorithm with the whole data. For this purpose, we utilize a simple search strategy with two nested loops. The first loop iterates through different values of k and the second loop repeats the *k-spatialmedians* algorithm with random initialization ten times. For each clustering result, we then compute the so-called Ray-Turi index, see [25]. This index captures the principal purpose of clustering prototypes, i.e. accurate presentation of separate subset of data, and it is computed by simply dividing the score function (1) with the distance of the two closest prototypes. Figure 1 visualizes the plot of the Ray-Turi index for a set of values for the number of clusters. From the visualization we observe that the clustering result (Ray-Turi index) is decreasing when more clusters are introduced. However, after four clusters the speed of improvement is decreased. Moreover, for four clusters the result is very stable because all the ten random repetitions provide exactly the same clusters and prototypes. To this end, based on these observations, $k = 4$ is used as the number of clusters and the unique result for the full data as initialization for the whole, sparse data set clustering with Algorithm 1. The obtained result, characterized by four prototypes with available value for all scale indices, is to be interpreted next.

3.2 Interpretation of Clustering Result

The four cluster prototypes are depicted in Figure 2. Table 2 provides information about the students in the different clusters. Hereby, *valid indices* shows the percentage of existing index values in each cluster. As can be seen, the available data is quite evenly distributed among the clusters. While *sample size* denotes the actual number of students in the data, *population size of target group* is the same but each student is weighted so that they represent the whole Finnish population of 15-year-old students. *WA math score* is the weighted average of the mathematics scores from the students in the respective cluster.

As can be inferred from Figure 2 in combination with Table 2, we have one clear "high performance" and one clear "low performance" national cluster: The students in *Cluster 1* have mean performance in mathematics of 571.53 and they are on average the most advantaged students with highest beliefs in themselves. In all indices that are associated with highperformance in mathematics, the prototype that represents this cluster has the highest value. Solely in the "intentions" to use mathematics later in their life, the students in *Cluster 1* lack behind the students in *Cluster 3*. *Cluster 4*, on the other hand, represents the most disadvantaged students in Finland, with lowest mean score in mathematics, and also lowest beliefs in themselves.

Cluster 2 and *Cluster 3* are, at the same time, similar and very different. According to the average performance of the students in those two clusters, both belong to PISA score Level 3 (see Table 4). As specified in the proficiency level descriptions in [22] this means that students in both of these clusters are able to, for example, solve tasks with clearly described procedures, but are unlikely to be able to (this proficiency is attributed to students from higher levels) also solve tasks that involve constraints or call for making assumptions. However, the prototypes (see Figure 2) show that students from these clusters can be opposite to each other by means of many scale indices.

While the students in *Cluster 2* generally are slightly more socially and economically advantaged, feel that they belong to school, and commonly have very positive attitude towards school, they definitely have below OECD average intentions to use mathematics, so that they also score worse in mathematics. *Cluster 2* is predominantly populated by girls. *Cluster 3*, on the other hand, has the lowest percentage of girls in it. This cluster consists of mostly boys who do not have the best attitude towards school. They also do not feel like they belong to school and generally are socially and economically less advantaged than the students in *Clusters 1* and *2*. However, they have the highest intentions to use mathematics later in their life, and pursue mathematics-related studies or careers in the future. They also tend to attribute failure in mathematics more to external factors than to themselves, have less anxiety towards mathematics than the OECD average, and are (although they do not seem to be interested in school in general) more interested in mathematics than the OECD average. It seems that they have already decided to have a career in a mathematics related profession, on the contrary to the (mostly female) students in *Cluster 2*.

As for the correlations before, we also created a ranking of indices to clarify the interpretation of the clustering result. The distance that defines the ranking to distinguish *Clusters 2* and *3* is just the absolute difference between the

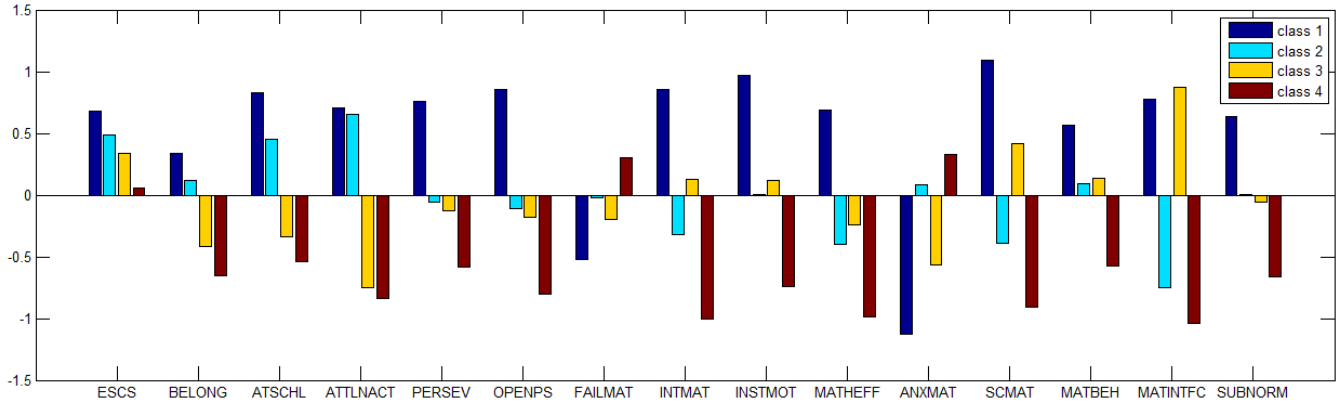

Figure 2: Clustering results

Table 2: Facts of clusters

cluster	valid indices	sample size	population size of target group: all	♀ (in %)	♂	WA math score: ∅	♀	♂
C1	64%	1967	12884	5302 (41%)	7582	571.53	578.66	566.55
C2	69%	2192	14038	8598 (61%)	5440	509.82	516.76	498.85
C3	67%	2450	16751	6434 (38%)	10317	536.02	541.74	532.45
C4	66%	2220	16374	8876 (54%)	7498	467.21	472.96	460.40
C1-C4	67%	8829	60047	29210 (49%)	30837	518.75	520.19	517.39

Table 3: Separation of clusters

index	all clusters: distance	rank	Cluster 2 -3: distance	rank
ESCS	0.62	15	0.15	10
BELONG	0.98	13	0.53	6
ATSCHL	1.38	9	0.78	4
ATTLNACT	1.54	7	1.40	2
PERSEV	1.35	10	0.07	13
OPENPS	1.66	6	0.08	12
FAILMAT	0.83	14	0.17	8
INTMAT	1.86	3	0.44	7
INSTMOT	1.71	4	0.11	11
MATHEFF	1.68	5	0.16	9
ANXMAT	1.46	8	0.65	5
SCMAT	2.00	1	0.81	3
MATBEH	1.14	12	0.04	15
MATINTFC	1.91	2	1.63	1
SUBNORM	1.30	11	0.06	14

index values of the two prototypes. This is generalized as the distance between *all clusters* by simply summing the three absolute differences between individually ordered prototype indices. These two distances and the implied rankings are provided in Table 3. As can be seen from Table 3, the students' self-concept in mathematics, the index which also correlates the most with the performance in mathematics (see Table 1), discriminates all the clusters the most. It seems that students' beliefs in their own mathematics abilities capture their true knowledge and skills fairly well. Additionally, the intentions to use mathematics and the interest in this subject provide a good separation of the four clusters. Those two indices describe the students' drive and interest to learn mathematics because they perceive this subject as profitable and appealing to their future. The two interesting clusters, *Cluster 2* and *Cluster 3*, are separated the most by the intentions to pursue a career in mathematics and by the attitudes towards school concerning learning activities.

4. ASSOCIATION RULE DISCOVERY

The goal of association rule mining, one of the most utilized methods in EDM according to [8, 26], is to automatically find patterns that describe strongly associated attributes in data. The discovered patterns are usually represented in the form of implication rules or attribute subsets [1, 32]. We have two explicit clusters - *Cluster 1* which consists of the highest performing students and *Cluster 4* which consists of the lowest performing students - but for the two remaining clusters with mixed profile, *Cluster 2* and *Cluster 3*, we want to find patterns/rules that further characterize these students. Hence, we form for each student that belongs to one of these two clusters an itemset which contains the gender of the student (first subset in Table 4), all the scale indices (central subset in Table 4), and the categorized proficiency level in mathematics (last subset in this table).

PISA score levels define the performance level of the students. For example, for PISA 2012 the range of difficulty of tasks generates six levels of mathematics proficiency. Students with a performance score within the range of Level 1 are likely to be able to successfully complete Level 1 tasks, but are unlikely to be able to complete tasks at higher levels. Level 6 reflects tasks that are the most difficult in terms of mathematical skills and knowledge [22]. On average, both student clusters of interest belong to performance Level 3 (see Table 2). Therefore, in the corresponding item, we only distinguish three categories: below, within, or above Level 3 (see the last subset in Table 4).

Table 4: Items for Association Rules

id	item
1	girl
2	boy
3 & 4	(+, −) ESCS
5 & 6	(+, −) BELONG
7 & 8	(+, −) ATSCHL
9 & 10	(+, −) ATTLNACT
11 & 12	(+, −) PERSEV
13 & 14	(+, −) OPENPS
15 & 16	(+, −) FAILMAT
17 & 18	(+, −) INTMAT
19 & 20	(+, −) INSTMOT
21 & 22	(+, −) MATHEFF
23 & 24	(+, −) ANXMAT
25 & 26	(+, −) SCMAT
27 & 28	(+, −) MATBEH
29 & 30	(+, −) MATINTFC
31 & 32	(+, −) SUBNORM
33	Level 2 or below: ≤ 482.38
34	Level 3: $482.38 - 544.68$
35	Level 4 or above: ≥ 544.68

In order to separate an individual student from main bulk of students, we fix a threshold value of 0.2 to define whether an item is part of the itemset for that particular student. The threshold 0.2 is chosen because it provides the median (rounded to one decimal place) of the absolute values of scale indices of all cluster prototypes. If a positive index value for a certain student is above the threshold, then the first *id* in the matrix (see Table 4) will be part of the itemset. Similarly, if a negative index value is below the negative threshold, then the second *id* (see Table 4) will belong to the itemset. Again, we utilize only the available indices. This means that in case the student's index value is inside $[-0.2, 0.2]$ or missing/invalid, it is not included in the itemset. For finding frequent itemsets based on the encoding, we used the implementation described in [13], and for generating association rules from the obtained frequent itemsets we utilized the implementation explained in [2].

4.1 Basic Concepts of Frequent Itemsets

Let I be the set of all items. An important property of an itemset is its *support count*, which refers to the number of transactions that contain a particular itemset. Let S_1 be a subset of the set of items ($S_1 \subseteq I$). Logically, a transaction $t_i \in T$, where T denotes the set of all transactions, is said to contain itemset S_1 if S_1 is a subset of t_i. Mathematically, the support count, $\sigma(S_1)$, for an itemset S_1 can be stated as follows:

$$\sigma(S_1) = |\{t_i \mid S_1 \subseteq t_i, t_i \in T\}|,$$

where $|\cdot|$ stands for the number of elements in a set. An *Association Rule* is then an implication expression of the form $S_1 \rightarrow S_2$, where $S_1, S_2 \subseteq I$ and $S_1 \cap S_2 = \emptyset$.

The support, $s(S_1 \rightarrow S_2)$, determines how often a rule is applicable to a given data set. Furthermore, the confidence, $c(S_1 \rightarrow S_2)$, determines how frequently items in S_2 appear in the transactions that contain S_1. Mathematically this can be expressed as follows:

$$s(S_1 \rightarrow S_2) = \frac{\sigma(S_1 \cup S_2)}{|T|} \text{ and } c(S_1 \rightarrow S_2) = \frac{\sigma(S_1 \cup S_2)}{\sigma(S_1)},$$

Support measures how well a rule is covered by the data. Therefore, if a rule has a too low support, it could be that it occurred solely by chance. Confidence is an important measures as it provides the the reliability and accuracy of a rule.

4.2 Obtained Rules and Interpretation

When we use the applied implementation of the famous Apriori Algorithm, we obtain many trivial rules. For example, it is already obvious from the clustering prototypes that those students who have highly positive attitude towards learning activities have also highly positive attitude towards learning outcomes. However, as already discussed, our itemsets can be divided into three subsets: the set that contains the gender, the set which contains the performance in mathematics, and the set which contains the different scale indices. We are interested in the gender differences and the performance in mathematics. Therefore, we search inside the algorithm's output for rules that have items of the gender and/orperformance interval subsets at the right hand side of the rule.

We start with high values for support and confidence and lower then the confidence threshold. Since we are especially interested in rules that contain the gender, the support has to have a relatively small value, so we choose the minimum value 0.1 while trying to keep the confidence value as high as possible. Starting with confidence of 1 and lowering it successively, we obtain the first rule that has gender on the right side with confidence 0.71:

$$\{\text{-ATTLNACT, +SCMAT, +MATINTFC}\} \Rightarrow \{boy\} \quad (2)$$

In words (2) means that those students who have negative attitudes towards school but a high self-concept and high intentions in mathematics are boys.

The first rule that we obtain for girls with confidence 0.69 is of the form:

$$\{\text{ -MATHEFF, - MATINTFC}\} \Rightarrow \{girl\} \quad (3)$$

Rule (3) says that those students who have negative self-efficacy and no intention to use mathematics are girls.

If we lower the minimal acceptable support into 0.095, we obtain the following interesting rule (4): Those students who have positive attitudes towards school but no intention to use mathematics later in life are girls.

$$\{\text{+ATTLNACT, -MATINTFC}\} \Rightarrow \{girl\} \quad (4)$$

Next, with the same minimal support we are searching explicitly for rules that have performance value below or above Level 3 at the left-hand side of the rule and gender at the right-hand side. Here, we first obtain the following rule with a confidence value of 0.6:

$$\{\text{+ATTLNACT, above Level 3 performance}\} \Rightarrow \{girl\} \quad (5)$$

According to (5), those students with a proficiency level above 3 and a clearly above average positive attitude towards learning activities in school are girls.

With confidence 0.52 we obtain the first rule for boys:

$$\{+\text{SCMAT, above Level 3 performance}\} \Rightarrow \{boy\} \quad (6)$$

Rule (6) means that those students with a proficiency level above 3 and a clear above average self-concept in mathematics are boys.

Subsequently, we are searching for rules wich have both gender and below or above Level 3 performance at the left-hand side of the rule. Such rule with the highest confidence (0.65) reads as:

$$\{\text{-ATSCHL -ATTLNACT +OPENPS -FAILMAT +SCMAT}\} \Rightarrow \{boy, \text{above Level 3 performance}\} \quad (7)$$

According to (7), those students with negative attitudes towards school (both, learning outcome as well as learning activities) but with clearly above average openness to problem solving, a high self-concept in mathematics and strictly below average self-responsibility for failing in mathematics, are boys that perform above Level 3.

For girls the rule with the highest confidence (0.63) is given by (8):

$$\{\text{-ESCS +ATTLNACT +ANXMAT -SCMAT}\} \Rightarrow \{girl, \text{below Level 3 performance}\} \quad (8)$$

This means that those students who are socially and economically less advantaged, have high anxiety towards mathematics and a low self-concept in mathematics, but still clearly above average attitude towards school, are girls who perform below Level 3.

If we unite the rules given in (2)-(8), we see that in all the rules that contain boys the item which represents the high self-concept in mathematics is present. In general, high-performing boys are also convinced that they can succeed (see 6). Moreover, even when they fail in mathematics, they are more likely to see other individuals or factors responsible on this than themselves (see 7). In addition, they have the highest intentions to use mathematics later in their life (see 2). However, according to the rules, male students can have negative attitude towards school (see 2 and 8), whereas the most positive attitudes appear only in the rules that include girls. Even the below average performing and socially and economically more disadvantaged girls with low self-concept and high anxiety towards mathematics, perceive the learning activities in their schools as very important (see 8). The same positive attitude towards school is also associated with the highest performing girls (see 5). Moreover, female students are much less confident about their mathematic skills (see 3) and have least intentions to pursue a mathematics related career (see 3 and 4).

To sum up, we conclude that specific characteristics and attitudes in the two middle performing clusters are, indeed, often gender-specific. Since we explicitly searched for rules that have certain items in them, we can not express precisely how typical these situations are. Nevertheless, when we combine all obtained rules with the clustering result two main characterizations appear: On the one hand, we have a specific subgroup of mainly girls who we nominated "to-be-nurses": they seem to be capable of performing well if they want to, having strongly positive attitude towards school. However, these students have low beliefs in themselves to be able to succeed in mathematics, and even a somewhat fear towards mathematics. On the other hand, we have a subgroup of mainly boys which we refer as "to-be-engineers". These students do not seem very interested in school in general. Yet, they trust in their capabilities and are extremely confident about their skills to perform well in mathematics. Even if they fail, they attribute this failure rather to other external factors than to themselves.

5. SUMMARY AND CONCLUSIONS

Although Finland is one of the few countries in which, on average, girls perform slightly better than boys in mathematics, professional careers related to this subject are also in here still dominated by men. We have applied methods from two of the main educational data mining branches on PISA data to obtain more gender-specific knowledge which might explain this observation.

First of all, we utilized a special robust clustering approach to group the students according to those PISA scale indices that are associated with performance in mathematics. The index that represents the students' self-concept in mathematics (SCMAT), and which also was the variable that correlates the most with the students' performance in mathematics (see Table 1), is the most important discriminator for the four clusters that we obtained (see Table 3). Combined with the other attributes we conclude that those students who have a higher self-concept, and tend to be socially and economically more advantaged, perform better than their less advantaged peers. They also have better attitudes to school, trust more in their own capabilities, and have greater expectation for their future careers (see Figure 2).

Two of the clusters we obtained, *Cluster 1* representing the "high performing" and *Cluster 4* representing the "low performing" students, can to a large extend be explained by these differences. However, the two "medium" clusters show the opposite behaviour: Socially and economical more advantaged students with very positive attitudes towards school and learning from *Cluster 2* perform worse in mathematics than the somewhat more disadvantaged students in *Cluster 3*. We found that these clusters are separated the most by the index that measures the student's intentions to pursue a mathematics related career. Since *Cluster 2* is with 61% dominated by girls, while *Cluster 3* consists of a larger percentage (62%) of boys we assumed that this difference could be explained by the gender of the student.

Association rule mining in the data subset of these two remaining medium clusters revised the gender-specific attitudes even more, and confirmed our assumption. Those 15-year-old students from this subset who already seem to have decided to pursue a mathematics related career are mostly boys. On the other hand, the attribute that is the most ascribable to girls is the positive attitude towards school. Altogether, the results of our study suggest that there are distinct groups of high and low performing students. However, the bulk of the girls with average performance seem to have no intentions to pursue a mathematics related profession. This is neither connected to their social status nor to their attitudes towards school. In fact, they often show a

better feeling of belonging to school and have very positive attitudes towards school and learning. While boys often consider mathematics as a great part of their future even when they do not show obvious skills, girls tend to be discouraged much faster and to easier favour other subjects. We feel that this is an important finding that should be studied further, especially concerning when such a gender-specific orientation starts to emerge.

6. REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[2] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[3] R. T. Aldahdooh and W. Ashour. DIMK-means "Distance-based initialization method for K-means clustering algorithm". *International Journal of Intelligent Systems and Applications (IJISA)*, 5(2):41, 2013.

[4] S. Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä, 2006.

[5] L. Bai, J. Liang, and C. Dang. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. *Knowledge-Based Systems*, 24(6):785–795, 2011.

[6] L. Bai, J. Liang, C. Dang, and F. Cao. A cluster centers initialization method for clustering categorical data. *Expert Systems with Applications*, 39(9):8022–8029, 2012.

[7] R. Baker et al. Data mining for education. *International Encyclopedia of Education*, 7:112–118, 2010.

[8] R. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1, 2010.

[9] M. Emre Celebi, H. A. Kingravi, and P. A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 2012.

[10] V. Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75, 2002.

[11] A. L. Goodwin. Perspectives on high performing education systems in Finland, Hong Kong, China, South Korea and Singapore: What lessons for the US? In *Educational Policy Innovations*, pages 185–199. Springer, 2014.

[12] J. Han, M. Kamber, and A. Tung. Spatial clustering methods in data mining: A survey, 2001.

[13] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[14] T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods*. Edward Arnold, London, 1998.

[15] P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., New York, 1981.

[16] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[17] T. Kärkkäinen and E. Heikkola. Robust formulations for training multilayer perceptrons. *Neural Computation*, 16:837–862, 2004.

[18] M. Meilă and D. Heckerman. An experimental comparison of several clustering and initialization methods. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 386–395. Morgan Kaufmann Publishers Inc., 1998.

[19] OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009.

[20] OECD. *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II)*. PISA, OECD Publishing, 2013.

[21] OECD. *PISA 2012 Results: Ready to Learn - Students' Engagement, Drive and Self-Beliefs (Volume III)*. PISA, OECD Publishing, 2013.

[22] OECD. *What Makes Schools Successful? Resources, Policies and Practices (Volume IV)*. PISA, OECD Publishing, 2013.

[23] OECD. *PISA 2012 Results: What Students Know and Can Do. Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. PISA, OECD Publishing, 2014.

[24] N. Raheja and R. Kumar. Optimization of association rule learning in distributed database using clustering technique. *International Journal on Computer Science & Engineering*, 4(12), 2012.

[25] S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143, 1999.

[26] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, 2010.

[27] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons Inc., New York, 1987.

[28] M. Saari. Promoting gender equality without a gender perspective: Problem representations of equal pay in Finland. *Gender, Work & Organization*, 20(1):36–55, 2013.

[29] M. Steinbach, L. Ertöz, and V. Kumar. The challenges of clustering high dimensional data. In *New Directions in Statistical Physics*, pages 273–309. Springer, 2004.

[30] B. Xu, M. Recker, X. Qi, N. Flann, and L. Ye. Clustering educational digital library usage data: A comparison of latent class analysis and k-means algorithms. *Journal of Educational Data Mining*, 5(2):38–68, 2013.

[31] R. Xu and D. C. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[32] Q. Zhao and S. S. Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 2003.