Game learning analytics for understanding reading skills in transparent writing system

Marko Niemelä, Tommi Kärkkäinen, Sami Äyrämö, Miia Ronimus, Ulla Richardson and Heikki Lyytinen

Marko Niemelä is a PhD student at the Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland. His research interests include machine learning, data analytics, and optimization. Tommi Kärkkäinen is a professor at the Faculty of Information Technology, University of Jyväskylä. His main research fields include computational sciences and learning analytics. Sami Äyrämö is an adjunct professor of data analytics at the Faculty of Information Technology, University of Jyväskylä. His research interests include machine learning and predictive modelling with applications in sport, health and medicine. Miia Ronimus is a postdoctoral researcher at the Niilo Mäki Institute, Jyväskylä, Finland. Her research interests include digital game -based learning, student motivation, and dyslexia. Ulla Richardson is a professor at the Centre for Applied Language Studies, University of Jyväskylä. Her research interests include technology enhanced language learning, reading development, dyslexia, and reading skill assessment. Heikki Lyytinen is a professor at the Department of Psychology, University of Jyväskylä. He has UNESCO Chair on Inclusive Literacy Learning for All. His areas of recent research include dyslexia, reading acquisition, and digital learning environments. Address for correspondence: Mr. Marko Niemelä, Faculty of Information Technology, University of Jyväskylä, PO Box 35, FI-40014 Jyväskylä, Finland. Email: marko.p.niemela@jyu.fi

Abstract

Serious games are designed to improve learning instead of providing only entertainment. Serious games analytics can be used for understanding and enhancing the quality of learning with serious games. One challenge in developing computerized support for learning is that learning of skills varies between players. Appropriate algorithms are needed for analyzing the performance of individual players. This paper presents a novel clustering-based profiling method for analyzing serious games learners. *GraphoLearn*, a game for training connections between speech sounds and letters, serves as the game-based learning environment. The proposed clustering method was designed to group the learners into profiles based on game log data.

The obtained profiles were statistically analyzed. For instance, the results revealed one profile consisting of 136 players who had difficulties with connecting most of the target sounds and letters, whereas learners in the other profiles typically had difficulties with specific sound-letter pairs. The results suggest that this profiling method can be useful for identifying children with a risk of reading disability and the proposed approach is a promising new method for analyzing serious game log data.

Keywords: learning analytics, serious game, letter knowledge, reading difficulties

Introduction

Background and motivation

Differentiated instruction is a framework supporting diverse needs and ability levels of students in classrooms by flexible use of time, space, materials, and strategies (Regan et al., 2014). Computer-assisted instruction, including intelligent tutoring systems and serious games is one way to differentiate traditional teaching (Boone & Higgins, 2007). Intelligent tutoring systems usually focus on embodying learning principles and adapting for differences between students, where as serious games emphasize student's motivation and engagement (Yanjin & Vincent, 2017). Serious games provide a considerable alternative for improving learning experience in comparison to traditional teaching methods such as classroom lessons (Wendel et al., 2012). Many serious games share the features of intelligent tutoring systems by including individually adaptive learning content, and by logging game events and contextual information during the gameplay (Wendel et al., 2012). Adaptation usually includes automatic content creation and adaptation of difficulty level for individual users as well as adaptation rules for gameplay (Wendel et al., 2012). Therefore, serious games provide an excellent platform for collecting data about individual differences in learning, which can be analyzed and utilized in the development of differentiated instruction.

Practitioner Notes

What is already known about this topic

- Serious games are used to improve learning and to tailor learning environments for people with various difficulties in learning.
- Learning analytics and serious games analytics are growing research fields, applying and developing data analysis methods to analyze, profile, and understand learning using serious games.
- *GraphoLearn* is a learning game for training reading skills. The game provides preventive support for learners with varying skill levels including individuals who are struggling with reading.

What this paper adds

- The paper develops and presents a novel approach for serious games analytics to analyze *GraphoLearn* players.
- The proposed data analysis approach produces an interpretable set of error profiles, which characterize the learning difficulties in a unique way.
- The profiling method can be used for longitudinal studies and applied to analyzing logs of other serious games.

Implications for practice and/or policy

- It is possible to reveal and understand profiles of serious game players.

- The proposed data analysis method can be used to identify players who have a potential risk for reading difficulties or disabilities.
- Even though the proposed method provides only limited information about players' future skills, it offers a good starting point for other studies in which players' development can be monitored more accurately.

Learning analytics focuses on the development and utilization of analysis methods for data from educational settings. The main ambition of learning analytics is to measure, collect, analyze, and report data about learners, for purposes of understanding and optimizing learning, teaching, and the environment in which it all occurs (Mor et al., 2015). It aims for the discovery of meaningful patterns about learners in their learning environment by using methods originated from statistics, information visualization, data mining, and social network analysis (Chatti et al., 2012; Peña-Ayala, 2017). Learning analytics can respond to a wide range of different needs, including visualization of learning activities, assessing learning behavior, predicting student performance, learning personalization, profiling, evaluation of social learning, and improving learning materials and tools (Nguyen et al., 2017).

In the present study serious games analytics is applied to the Finnish version of *GraphoLearn*. The game was originally developed during the Jyväskylä Longitudinal Study of Dyslexia (JLD) (Lyytinen et al., 2009). The aim was to support the basic decoding skills of Finnish children at risk for reading difficulties by helping the learner to connect spoken items (e.g. speech sounds) to their written counterparts (e.g. letters). Nowadays, the game has been adapted accordingly to a high variety of languages around the world.

We combine the methods of clustering, missing values handling, and cluster validation to offer an approach for profiling *GraphoLearn* players. The proposed model categorizes learners into distinct profiles based on players' game log data informing about the choices the players have made in the game. The number of profiles is selected by using cluster validation indices. We excluded other more complex clustering methods because we do not aim at discovering clusters with any specific or anomalous shapes, but rather partition the data into subsets of similar observations using a clustering model that is straightforward to interpret both with respect to input variables and players (Steinbach et al., 2004). Further, the study presents statistics of the different profiles, which can be used for analyzing learners' risk for a reading difficulty. The purpose of the research is to identify a distinct set of learner profiles, which are interpretable and applicable to practice.

On serious games analytics

Serious games analytics can be used to improve learning and to tailor learning environments for people with various difficulties in learning. Lameras et al. (2017) investigated how learning attributes (e.g., learning activities, learning outcomes,

assessment, and feedback) and game properties can be planned, designed, and implemented by university teachers interested in using games for teaching and learning in higher education. The study identified 165 papers providing empirical evidence and conceptual assumptions concerning specific learning activities that could be linked with game elements (e.g., leaderboard, virtual currencies, and in-game hints), feedback and progress indicators, and teacher's roles designing and facilitating game play. Nguyen et al. (2018) provided a framework and a design tool for people with intellectual disabilities to address each learner's individual needs. The proposed framework is valuable for the design, implementation, evaluation, and adaptation of serious games for more enhanced learning and teaching at the group or individual level

Serious games analytics can also be successfully applied for analyzing the individual differences and behavioral patterns of serious game learners. For instance, Hicks, Eagle, et al. (2016) analyzed gameplay patterns of the *Quantum Spectre* physics game to understand player dropout in the game. By using survival analysis, interaction network analysis, and the results from player surveys they were able to identify particular problem spots where players dropped out of the game due to its complexity. Hicks, Liu, Eagle, and Barnes (2016) also compared three different level creation editors, which are helpful for players learning about the *BOTS* game's core mechanic. Based on the results of a zero-inflation model, programming editor and building editor were more effective than drag-and-drop editor in the case of encouraging the creation of levels, which contained more game play affordances for players. Horn et al. (2016) explored player strategies in *GrAZE*, an educational puzzle-based game that is designed to support algorithmical thinking for middle school students. The aim was to understand by using hierarchical clustering how players learn and progress in the game. The study identified problem areas in the game design for further development of the game. Harpstead and Aleven (2015) utilized learning curve analysis from serious games analytics in *BeanStalk* physics game designed to teach the concept of balance beam system for young children. The aim was to find implications for the level design to better accomplish its educational goals. The results show that analytical methods can yield actionable design recommendations.

Research questions

The study uses learning analytics for analyzing the playing patterns of *GraphoLearn* players based on a group-level information extracted from cluster profiles. The variables of interest are error rates, contexts of the errors, progression information, total playing times, and interval times between playing sessions. The provided clustering method is a novel alternative for analyzing partially incomplete learning data and it is modifiable for a high volume of data. The developed method is aimed to help characterizing and monitoring players and their learning process. In addition, the method can help researchers identify groups of individuals who have a risk of reading difficulty. A diverse set of profiles is expected to be found because of a relatively large sample of different learners. The research questions are following:

RQ1: Is it possible to identify a set of distinct and interpretable cluster profiles by using the proposed clustering method?

RQ1.1: Can internal cluster validation indices be used for finding the number of clusters in the models that are well-separated, interpretable, and useful for practice.

RQ2: What are the typical bottlenecks compromising the learning of letter-sound correspondences?

We set the following hypotheses to the research questions:

H1: Because of the variability in the starting skills of the learners, we expect to identify several distinct profiles which can be interpreted for further application.

H1.1: We expect that the use of cluster validation indices lead to a number of clusters that are well-separated, interpretable and thereby useful as well (see e.g. Hämäläinen et al., 2018).

H2: We anticipate children to confuse especially letters that either look or sound similar (see Lyytinen et al., 2009).

Context of the study

Reading skill development

The basic reading skill is based on connection building between spoken and written language. Thus, learning the skill requires storing of those connections. *GraphoLearn* is designed as a training environment for this purpose (Richardson & Lyytinen, 2014). In alphabetic writing systems, such connection building is based on the smallest imaginable units, phonemes, and their written equivalents, that is, letters (or graphemes when more than one letters is used to represent one sound). Phonemes and graphemes are consistently connected in transparent orthographies. Thus, one has to learn only the sounds of letters and invent that assembling such sounds in the order of letters means reading. In less transparent writings systems such as English the same principle works but only by using larger units such as rimes (e.g. *ing* in English) to make the connections more "learnable", that is, true in all contexts of writing. Learning to differentiate phonetically similar sounds such as */g/* and */d/* and visually similar letters such as *n* and *h* can be considered as the most challenging part of storing the connections. The method described here helps to understand reading difficulties and disorders, which result from, for example, biological factors or inadequate education. This is made by showing how the difficulties appear during the learning (i.e., connection building) process.

GraphoLearn

GraphoLearn is a game proven to provide preventive support for learning to read (Saine et al., 2011). The game was originally developed as a way to observe how the difficulties in learning appear and later to supplement for reading instruction provided by schools. There are dozens of different *GraphoLearn* versions built for helping the learner to master the connection building in different linguistic and orthographic

contexts. The present study used the version designed for Finnish students.

In transparent orthographies the game starts by introducing speech sounds and corresponding letters. First, phonetically and visually distinct and easy to perceive letters (e.g., *a*, *e*, and *i*) are presented and then one moves on to present correspondences that are more similar and thus less distinguishable (e.g., *b*, *d*, and *p*). In the game, the player first hears speech sound and then identifies and selects the corresponding letter from the several alternatives shown on the screen. The player receives immediate visual and auditory (corrective) feedback after each response. When the player has learned to connect most of the sounds and letters flawlessly, the game proceeds to training larger units such as spoken and written syllables and then words, starting from two letter syllables and eventually moving on to long words consisting of several letters. The player is expected to grasp the idea that reading occurs by assembling the speech units represented by the letters of a word.

An important feature of *GraphoLearn* is that the progression of the game adapts to the learner's current level of performance. This is done, for example, by using the Bayesian principle to present new learning tasks (Kujala et al., 2010). The adaptation techniques aim for a mean success rate of at least 80%, offering both challenge and success, which together makes playing more rewarding. Important features are also a personalizable avatar and rewards. Such rewards and graphically different game levels are efficient ways to sustain the learners' motivation in playing and to expose them repeatedly to strengthen the correct connections. The game also involves the static assessment levels of learners' development in the tasks during playing.

Figure 1 shows the user interface of an assessment task included in specific versions of *GraphoLearn*, and chosen for a closer inspection in the present study. The assessment task evaluates the player's skill in identifying the letters corresponding to the 23 speech sounds of the Finnish language. In the assessment, the player hears each of the sounds, one by one, and selects the corresponding letter from the alternatives shown on the screen. The sound is repeated if player does not response within 5 seconds. If the player does not answer within 15 seconds, an option for skipping the trial becomes available. The assessment is first presented when the game is started and is then repeated at intervals of 1 hour.

Methods

Participants

Learners were recruited by sending an information letter about *GraphoLearn* and upcoming study to an email list of teachers registered as *GraphoLearn* users. The information letter was sent in September, about six weeks after the start of the school year. Teachers were asked to consider if they had a first grade student with risk factors for dyslexia (difficulties at learning to read, poor letter knowledge, family members with dyslexia) and who spoke Finnish as first language. *GraphoLearn* was recommended for such students. Teachers needed a written consent from the child's guardian before registration. Before the game could be used, parents and teachers also

Figure 1: Appearance of the sound-letter assessment task included in GraphoLearn

needed to accept the terms and conditions stating that the game log data would be saved in the secure *GraphoLearn* server and could be used for research purposes

Although we were unable to control the type of children who started the use the game, we expect that the suggestions given in the information letter had an effect on the characteristics of the sample, and we expect it to consist mostly of Finnish-speaking first graders who had not yet acquired the level of letter-sound correspondence skills needed for learning to read and who may also have a risk of dyslexia.

Eventually, data was gathered from 1632 players who were 6.5–8.75 years old (*M=7.39, SD=0.46*). The data from the sound-letter assessment task indicates that children could correctly identify 13.68 letters (*SD*=5.09) out of 23, suggesting they had not yet learnt to master all the associations between sounds and letters and would likely benefit from training with *GraphoLearn.*

Majority of the players were boys (61.1%), which is probably because reading difficulties are more common among boys (e.g. Rutter et al., 2004). The players came from more than 200 municipalities with all regions of Finland being represented. Largest numbers of players came from the cities of Helsinki and Jyväskylä. The number of adults, who had registered the children as *GrahoLearn* players, was 669. These adults, 88.6% being teachers and 11.4% parents, were in charge of supervising the player. The number of registered players per adult ranged from 1 to 46, but only 3% were in charge of more than 10 players. The median number of registered players per adult was 1.

Data collection

The players can learn to use *GraphoLearn* within 1–2 minutes. They were advised to use headphones and play short (about 10 minutes) sessions at time, and several sessions per day in consecutive days. Teachers and parents were responsible of supervising the

playing and ensuring that children used the game in a quiet place to avoid distractions. Teachers and parents were advised not to help children with game tasks, so that the difficulty level determined by the adaptation would not increase too much relative to child's skill level.

The player's actions during the game were logged into a database. The personal log files include, for example, starting times, ending times, number of playing sessions, target items, durations, correct/incorrect selections, and skipped tasks. For research purposes, the most important information to be logged were player's inputs and time spent with each task (from perceiving the stimuli to the selection of the corresponding written unit), which is commonly referred as response time. The sample was divided into two groups based on the type of letters used in practice (lowercase or uppercase) that was chosen by user. In total 1275 players played with lowercase letters and 357 players used uppercase letters. The lowercase letters are used in the initial stages of formal reading instruction at schools, which is the likely reason for them being chosen more often. The main limitation of the data was that 4.66% of the responses were missing because some players stopped playing before all 23 targets had been presented. This was taken in account when algorithms were developed for the present analysis (see the next section).

The realized profiling approach

Clustering is an unsupervised technique for organizing empirical observations into different groups called clusters so that observations in the same cluster are more similar to each other than observations in the other clusters. K-means is probably the most common prototype-based partitional clustering approach, which has a long history (Jain, 2010). The algorithm is broadly used due to its ability to solve general purpose problems. K-means finds a partition such that the squared Euclidean error between cluster prototype, and the observations in the cluster is minimized (see more details in Supplement S2).

Many clustering algorithms require the number of clusters as an input parameter. However, this information is not often available and it can be a challenging task to determine the number, especially in the cases of multidimensional data. Even though there exist different tricks to illustrate multidimensional data, for example, using different multidimensional visualization techniques or dimension reduction techniques, perceiving the data structure may not be obvious. Cluster validity measures provide a way of validating the quality of results of clustering methods to find a partition that best fits the nature of data. Because of multidimensional data structures, cluster validation measures, for example, cluster validation indices, are very suitable, even essential methods, for determining the number of clusters (Arbelaitz et al., 2013). The internal cluster validity index is one of categories of cluster validity, which utilizes the results of a clustering algorithm in terms of quantities of the data set itself (see more details in Supplement S3).

This study consisted of implementing K-means clustering, K-means++ initialization, and cluster validation indices algorithms. Since some observations in *GraphoLearn* data

included missing values, distance calculations in all of the implemented algorithms was needed to replace with the general similarity measure (Gower, 1971).

The players were divided into two groups based on whether they used lowercase or uppercase letters. The game log data of both groups were transformed into two binary matrices. The number of rows corresponded to the number of players and the number of columns to the number of distinct target-letter pairs. Non-zeros in the matrices indicated selected erroneous selections. Matrix dimensionalities were reduced, because of the computational cost of clustering. This was applied by filtering out columns, which did not consisted noticeable number of erroneous selections.

After the pre-processing step, the clustering was performed by gradually increasing the number of clusters, *K*, from 2 to 10. The maximum number was selected as 10, because a high number of clusters makes the interpretation and analysis of the results more challenging. In addition, a small number of *K* generalizes data the most. For instance, Saarela and Kärkkäinen (2015) used 11 as the maximum number of clusters in their study of the Finnish student population in PISA 2012. For each value of *K*, clustering was repeated 200 times and the best prototypes with the lowest clustering error were saved. These were also used as initial points for the next value of *K*, where the additional initial point was generated using K-means++ initialization algorithm.

The quality of distinct data partitions and obtained cluster profiles were evaluated using internal cluster validation indices (CVIs). Eight CVIs were selected from our previous study (Niemelä et al., 2018) for calculating clustering index values. Multiple indices were selected to the current study since the previous studies revealed that there does not exist one superior index which overcomes others (see e.g. Hämäläinen et al., 2017). Each CVI produced one quality measure of clustering for each value of *K*. These values were used when deciding the final number of clusters for lowercase and uppercase data sets. Index values from different indices were scaled to the same range of [0, 1] to easy up their comparison.

The number of clusters was decided by analyzing the index curves of validation indices. First, the index values were grouped together and the speed of improvement (i.e., strength of decreasing trend based on group distributions) was analyzed using statistical testing. The aim was to reject weak candidates, that is, to eliminate regions where improvements were not statistically significant. The Wilcoxon statistical ranksum test was performed for each two-pair of successive groups. In the final stage, the number of clusters were decided benefiting the statistical measures and analyzing figures obtained from the CVIs. Regarding the source codes of algorithms, they are available online[1].

Results

Interpretation of the learner profiles

Figures S1.1, S1.2, and S1.3 in Supplement S1 show the learner profiles in a confusion

[1] http://users.jyu.fi/~mapeniem/BJET/Kmeans/

matrix format for the lowercase letter data set. The profiles were calculated also by using the uppercase letter data set but because of similar confusion patterns and low number of cases in certain profiles they are omitted here. Nevertheless, these results are available in Supplement S3.

In Figures S1.1, S1.2, and S1.3 darker colors indicate higher average confusion percentages for the target-distractor pairs over the players in the profiles. The confusions in the matrix diagonals are zero because they indicate correct selections. Most of the observed confusion can be explained by phonetic and visual similarity of the sounds and letters. These two main categories of confusion are marked with "circle" and "square" symbols in the matrices. It is also possible that the errors are associated to both or neither categories, which are marked with "star" and "rectangle" symbols.

Main errors in the profiles

Confusion symbols are summarized in Table 1. Only confusion percentages exceeding 10% are illustrated to clarify presentation. Further, noticeable confusions exceeding 15% are underlined. Table 1 shows that many profiles have something in common, for example, the letter *n* is often mixed to letters *h*, *m*, and the letter *f* is mixed to letters *s* and *v*. Especially, *n* is strongly confused with *m* and this can be concluded to be the most challenging sound-letter pair for the players possibly, because both acoustic and visual similarity compromises building the connection. An interesting finding is that the confusion between commonly mixed letters *f* and *v* cannot be explained by concrete phonetic nor visual similarity of the letters. This may be related to *f* being a foreign letter in the Finnish language, and being pronounced as */v/* in certain dialects.

Table 1 shows that all the profiles have some unique errors regarding target letters. *Profile 1* players have difficulties in connection building due to the difficulties in separating both visually and phonetically similar items represented by the *b* and *d* letters, which is not as often appearing in other profiles. *Profile 2* players mix sound */g/* to sounds */d/* and */k/*, whereas *profile 3* players mix sound */t/* to sound */s/*. *Profile 4* players have difficulties with both of the two main confusion categories, that is, they often do not differentiate visually and phonetically similar letters. The main problems of *profile 5* and *profile 6* players are related to the visual similarity of the letters.

Calculated statistics

Table 2 provides information about the performances in the assessment tasks and playing patterns of the players in the different profiles. The error rate refers to the mean percentage of incorrect selections of players within a profile. The players' development in connecting speech sounds to letters from the first assessment to the second assessment (after about 60 minutes of playing) was calculated by subtracting the error rate in the second assessment from the error rate in the first assessment. Only players who completed both assessments were included and clustering was not repeated in the second assessment. The total playing time refers to the time the game was used within the first five months of usage. The interval time refers to the median time gap between play sessions during the first month of playing.

Table 1: Symbol table of similarities for different profiles

	target | distractor																								
	b|d	b|p	b|v	d|b	d|g	f|h	f|s	f|v	g|b	g|d	g|k	i|l	j|l	m|n	n|h	n|m	p|b	p|d	t|f	t|s	u|o	y|ö	ö|o	ö|ä
profile																									
p1	☆			☆		○		▢	○	○		□		☆		☆		☆				○			
p2		☆			○		○	▢		○	○	□			□	☆	☆		□			○		□	
p3							○	▢		○					□	☆	☆			○					
p4			○				○	▢		○		□			□	☆	☆				○	○	□		
p5		☆		☆				▢	○	○		□			□	☆	☆		□						
p6												□	▢	☆	□				□						
total	1	2	1	2	1	1	3	5	2	5	1	5	1	2	5	5	4	1	3	1	1	3	1	1

phonetic similarity=○, visual similarity=□, phonetic and visual similarity=☆, unknown category=▢

According to Table 2, majority of players (34.8%) were grouped in the *profile 3*. In this profile, all the statistical values were near the average values of all profiles. The players in the *profile 4* have average error rate of 55.3% and median total playing time of 130.1 minutes. These values are much higher than the values in the other profiles. The *profile 4* players seem to have had difficulties with almost all target letters. The players have approximately 71% higher total playing time compared to the median value of all players, suggesting that they have needed more training than others. The average error rate in the *profile 6* is also high but this is caused by players who skipped most of the target tasks. The high percentage of the skipped tasks may imply that the players of this profile (5.4% of all players) were not motivated to complete the assessment in the beginning of the training. Although the players in the *profile 4* and *profile 6* had the highest error rates in the beginning, they also showed more progress than the players in other profiles according to the calculated differences in error rates, 25.1% and 23.9%, respectively. The players in *profile 1* and *profile 4* had the shortest time intervals between the playing sessions, suggesting more frequent playing.

Determination of the number of profiles

Using validation index curves obtained from different CVIs, minimums correspond to the best clustering structures. However, instead of the minimums, the speed of improvements of the index values was the main interest. Thus, because if the value of an individual CVI does not change much, it usually means that increasing the number of clusters does not notably improve the final solution. The results of CVIs are given in Figure 2. Numbers of clusters are in x-axes and y-axes show index values which were scaled to the range of [0, 1]. All indices except *Pakhira-Bandyopadhyay-Maulik (PBM)* and *Silhouette* obtain the minimum at the highest *K* value. Especially, *Calinski-Harabasz*, *kCE*, *PBM*, and *WB* indices provided the high speed of improvement of the cluster validation measures.

Table 2: Findings of profiles

	p1	p2	p3	p4	p5	p6	p
statistics							
size (in %)	14.3%	19.0%	34.8%	10.7%	15.8%	5.4%	100.0%
error rate	35.7%	39.1%	40.1%	55.3%	30.7%	54.2%	40.2%
progression*	14.3%	17.3%	17.6%	25.1%	11.5%	23.9%	17.8%
playing time	67.2 min	89.7 min	75.4 min	130.1 min	59.8 min	75.7 min	76.0 min
interval time	3.0 days	6.5 days	5.0 days	4.0 days	5.5 days	6.5 days	5.1 days

*Only players who completed both assessments are included.

Figure 3 shows a box plot presentation of all index values combined in the groups based on values of *K*. On each box the central mark indicates median of eight indices, and the bottom and the top edges of the box indicate 25th and 75th percentiles, respectively. The whiskers show to the most extreme data points and outliers are plotted using a '+' symbol. Table 3 presents statistical differences between each two pairs of groups, which were measured by Wilcoxon ranksum test so that only the successive groups which showed a decreasing trend in index values were compared. The bolded numbers indicate statistically significant differences between groups ($p<0.05$). Using the measured index values for lowercase letter data, statistically significant differences were obtained in two comparisons of the distributions. The measured difference between the median values of groups 5 and 6 was the highest (0.464) and therefore K=6 was the selected number of cluster profiles. Analogously, using the measured values for uppercase letter data, in total two comparisons were statistically different. The calculated difference between the median values of groups 2 and 5 was the highest (0.361) and therefore K=5 was the selected number. Nevertheless, our experiments showed that the fifth uppercase letter cluster profile included only few players and therefore four profiles were considered in the future analysis.

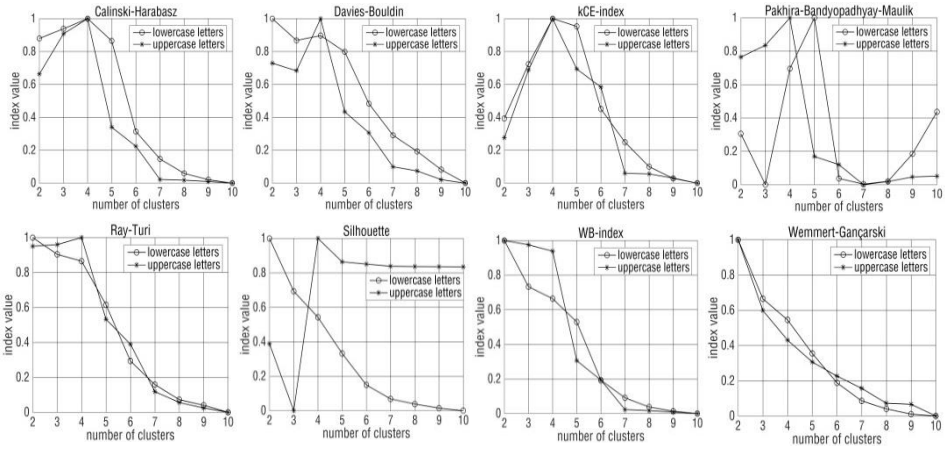


Figure 2: Values of cluster validation indices for K=2,…,10

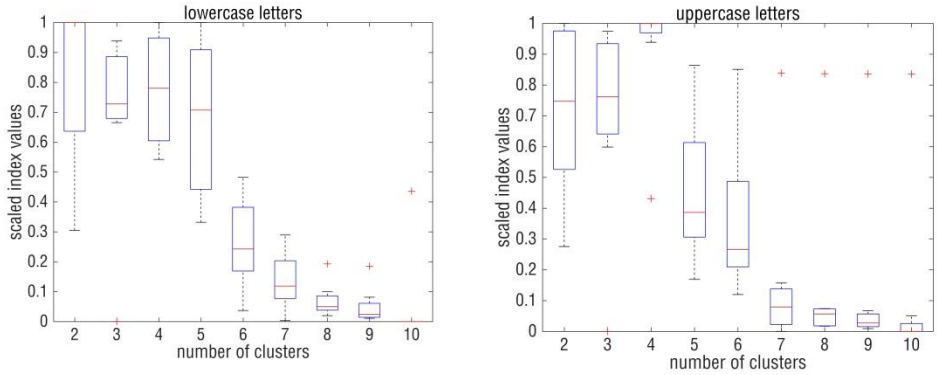


Figure 3: Box plot presentation of scaled index values

Table 3: Statistical p values obtained by Wilcoxon ranksum test

	compared pairs of distributions							
	g2, g3	g2, g5	g3, g5	g5, g6	g6, g7	g7, g8	g8, g9	g9, g10
lowercase letters	0.088	–	0.850	**0.002**	0.054	0.140	0.162	**0.003**
uppercase letters	–	**0.038**	–	0.104	**0.004**	0.326	0.521	0.238

Discussion

This paper presents a new clustering based approach for identifying different profiles of serious game players. We applied this method to *GraphoLearn* game log data. Based on the results, a set of profiles with different error types and rates were found. Even though there were errors common to all profiles, there were also many specific errors, which differentiated the profiles. According to Table 2, there were one "high" performing, three "medium" performing, and two "low" performing profiles with the different sound-letter pair errors. The players in the two weakest profiles showed the best progression while playing the game, which suggest that the combination of *GraphoLearn* and school-provided reading instruction helps children who have difficulties in reading acquisition. These findings are applicable to the practice and, therefore, the first hypothesis H1 is supported.

We found support to the hypothesis H2, because most of the errors were related to confusing phonetically and visually similar letters (see Table 1 for more details). Taking into account the confusions exceeding 10%, we realized that only 6 cases out of 57 confusions were not explainable by phonetic or visual similarity of letters.

Lyytinen et al. (2009) believed that children with familiar risk of dyslexia and/or low letter knowledge during the few months before school entry benefit from preventive playing in terms of avoiding unwanted failure experiences during the first months of school instructions. The study shows that the most challenging game tasks are related to visually and phonetically similar letters. In addition, uncommon letters in the Finnish language (e.g. *d* and *b*) showed to be challenging for the beginners.

The hypothesis H1.1 was supported. We used the Wilcoxon's ranksum test and the real differences of combined groups of CVIs to identify the number of clusters for lowercase and uppercase letter data. The results revealed 6 profiles for lowercase data and 5 profiles for uppercase data and we consider them as the most appropriate number for the clustering models.

Clustering methods are very commonly used in learning analytics. Saarela and Kärkkäinen (2017) have made a small survey of educational clustering methods. Three main approaches were hierarchical clustering, K-means clustering, and expectation maximization. These methods were used student modelling which included behavior and performance based models. The set of papers was identified scanning through relevant publication forums including the *Journal of Learning Analytics* and the *Conference on Learning Analytics & Knowledge*.

The used K-means clustering method and provided data analysis differentiates the current study from the related work as described in the section *On serious games analytics*. Horn et al. (2016) used the hierarchical clustering method to analyze game progression of learners. The main difference between clustering approaches is that the K-means clustering produce a single-layer clustering structure whereas the hierarchical method generates a tree-type clustering structure. The computational complicity of the hierarchical method is much higher and, therefore, it is not recommended for large-sized data sets. Further, the hierarchical method produces arbitrary shaped clusters whereas K-means produces easily interpreted geometrically closed subsets (Jain, 2010).

In the present study, the game data is limited only to the assessment tasks. To obtain more accurate and reliable clustering results, a larger sample size should be used. Further, other interesting variables could also be clustered, for example, larger units such as syllables or words, to achieve player profiles revealing differences in the types of errors children make in the actual reading. Further, more efficient clustering algorithms are required for a larger pool of samples. More specifically, a parallel implementation of algorithms into multiple machines with shared memory resources could be realized (Hämäläinen et al., 2018). Since *GraphoLearn* data contain missing values and outliers it is important to consider use of a robust clustering method in future algorithm design. For instance, spatial median is a statistically robust location estimate in clustering which can handle up to 50% of missing values or outliers (Hämäläinen et al., 2017)

A possible direction for future research could be repeating the clustering at regular time intervals to see how players divide into profiles in the follow-up cluster models. The approach offers a way to monitor players' progression in the game by detecting their connections to varied skill profiles. This new framework can be beneficial for validating the design of the original game, for example, it might be advantageous to improve the adaptation mechanism of the *GraphoLearn* for learners from different profiles (Kujala et al., 2010). For instance, Cano et al. (2018) have previously used learning analytics for validating the design of a learning game for adults with intellectual disabilities. In the study, the data tracker sent out relevant information about the behavior of the users and their learning patterns while playing the game. Further,

statistical learning models, for example, neural networks, can be used for predicting players' game progression. Interesting variables to be predicted are, for example, player's inputs to different tasks and a particular time when the player will stop playing.

Conclusions

The growth of learning games and e-learning platforms imply that volumes of data on learning and learners are increasing rapidly. This means that special techniques are needed for analyzing learners with varying skills and their needs to enhance their learning process. We applied the clustering method from a branch of learning analytics to analyze performance of *GraphoLearn* players. The results indicated that it is possible to identify different types of learners using the given clustering method. The calculated statistics offered valuable information about the cluster profiles. This information can be used, for example, as a support for tracking children with a risk of reading disability due to certain types of bottlenecks compromising learning. Clustering was performed for data obtained at a very early stage in the game. Therefore, the used approach gives limited evidence about players' future skills. However, the future research direction is to extend the developed algorithms so that many other interesting learner patterns can be extracted from the data, for example, players' development in the game is one main interest. The present study offered the method, which is a considerable alternative for analyzing learners of alphabetical learning games and it is a good starting point for developing more effective analytical tools in different contexts of learning.

Statements on open data, ethics and conflict of interest

The data that support the findings of this study are available on request from the first author. The data are not publicly available due to them containing information that could compromise privacy of the participants.

Data collection, analysis and publishing followed the modes of action endorsed by the research community: integrity, meticulousness and accuracy in conducting research. The research ethics guidelines of the Finnish Advisory Board on Research Integrity (2019) were followed throughout this study.

There is no conflict of interest.

Acknowledgements

This study was supported by a grant from the Academy of Finland for the Unesco professor Heikki Lyytinen (decision numbers 292493 and 311737), the professor Tommi Kärkkäinen (decision numbers 311877 and 315550), and the professor Ulla Richardson (decision number 274050). The study was also supported by a grant from the Foundation Botnar for the professor Ulla Richardson (decision number 6066). All the authors gratefully acknowledge the postdoctoral researcher Harri Ketamo for the valuable discussions related to serious games and data analytics.

References

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256.

Boone, R. & Higgins, K. (2007). The role of instructional design in assistive technology research and development. *Reading Research Quarterly*, 42:135–140.

Cano, A. R., Fernández-Manjón, B., & García-Tejedor, A. J. (2018). Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities. *British Journal of Educational Technology*, 49:659–672.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Joural of Technology Enhanced Learning*, 4(5- 6):318–331.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* (pp. 857-871). Washington, United States: International Biometric Society.

Hämäläinen, J., Jauhiainen, S., & Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms*, 10(3):105.

Hämäläinen, J., Kärkkäinen, T., & Rossi, T. (2018). Scalable robust clustering method for large and sparse data. In *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 449–454). Bruges, Belgium: ESANN.

Harpstead, E., & Aleven, V. (2015). Using empirical learning curve analysis to inform design in an educational game. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (pp. 197–207). New York, United States: Association for Computing Machinery.

Hicks, A., Liu, Z., Eagle, M., & Barnes, T. (2016). Measuring gameplay affordances of user-generated content in an educational game. In T. Barnes, M. Chi, & M. Feng (Eds.), *Educational Data Mining* (pp. 78–85). North Carolina, United States: International Educational Data Mining Society (IEDMS).

Hicks, D., Eagle, M., Rowe, E., Asbell-Clarke, J., Edwards, T., & Barnes, T. (2016). Using game analytics to evaluate puzzle design and level progression in a serious game. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 440–448). Edinburgh, United Kingdom: Association for Computing Machinery.

Horn, B., Hoover, A. K., Barnes, J., Folajimi, Y., Smith, G., & Harteveld, C. (2016). Opening the black box of play: Strategy analysis of an educational game. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (pp. 142–153). Austin, Texas, United States: Association for Computing Machinery.

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

Kujala, J., Richardson, U., & Lyytinen, H. (2010). A bayesian-optimal principle for child-friendly adaptation in learning games. *Journal of Mathematical Psychology*, 54:247–255.

Lameras, P., Arnab, S., Dunwell, I., Stewart, C., Clarke, S., & Petridis, P. (2017). Essential

features of serious games design in higher education: linking learning attributes to game mechanics. *British Journal of Educational Technology*, 48(4):972–994.

Lyytinen, H., Erskine, J., Kujala, J., Ojanen, E., & Richardson, U. (2009). In search of a science-based application: a learning tool for reading acquisition. *Scandinavian Journal of Psychology*, 50:668–675.

Mor, Y., Ferguson, R., & Wasson, B. (2015). Editorial: Learning design, teacher inquiry into student learning and learning analytics: A call for action. *British Journal of Educational Technology*, 46:221–229.

Nguyen, A., Gardner, L. A., & Sheridan, D. (2017). A multi-layered taxonomy of learning analytics applications. In R. A. Alias, P. S. Ling, S. Bahri, P. Finnegan, & C. L. Sia (Eds.), *21st Pacific Asia Conference on Information Systems.* Langkawi, Malaysia: PACIS.

Nguyen, A., Gardner, L. A., & Sheridan, D. (2018). A framework for applying learning analytics in serious games for people with intellectual disabilities. *British Journal of Educational Technology*, 49(4):673–689.

Niemelä, M., Äyrämö, S., & Kärkkäinen, T. (2018). Comparison of cluster validation indices with missing data. In *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 461–466). Bruges, Belgium: ESANN.

Peña-Ayala, A. (2017). *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning*. Switzerland: Springer International Publishing.

Regan, K., Berkeley, S., Hughes, M., & Kirby, S. (2014). Effects of computer-assisted instruction for struggling elementary readers with disabilities. *The Journal of Special Education*, 48(2):106–119.

Richardson, U. & Lyytinen, H. (2014). The Graphogame method: The theoretical and methodological background of the technology-enhanced learning environment for learning to read. *Human Technology*,10:39–60.

Rutter, M., Caspi, A., Fergusson, D., Horwood, L., Goodman, R., Maughan, B., Moffitt, T., Meltzer, H., & Carroll, J. (2004). Sex differences in developmental reading disability. *Journal of the American Medical Association*, 291:2007–2012.

Saarela, M., & Kärkkäinen, T. (2015). Weighted clustering of sparse educational data. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 337–342). Bruges, Belgium: ESANN.

Saarela, M., & Kärkkäinen, T. (2017). Knowledge discovery from the programme for international student assessment. In A. Peña-Ayala (Ed.), *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-Learning. Studies in systems, decision and control, 94* (pp. 229–267). Switzerland: Springer International Publishing.

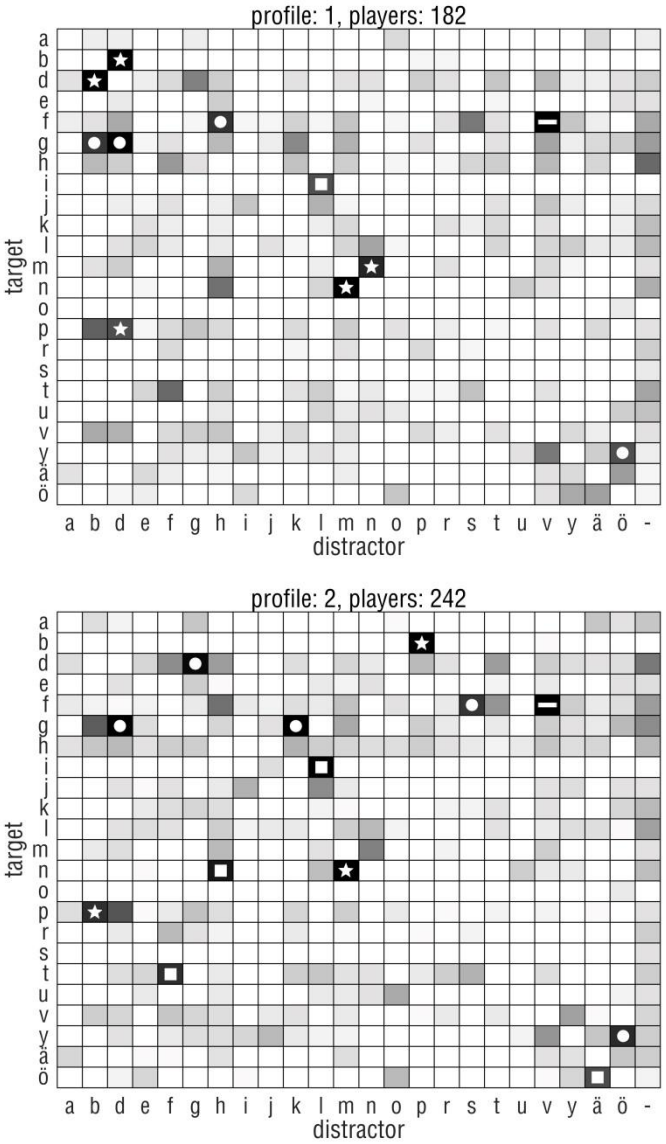
Saine, N., Lerkkanen, M.-K., Ahonen, T., Tolvanen, A., & Lyytinen, H. (2011). Computer-assisted remedial reading intervention for school beginners at risk for reading disability. *Child development*, 82:1013–1028.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In L. T. Wille (Ed.), *New Directions in Statistical Physics*, (pp. 273–309). Berlin, Heidelberg: Springer.

Wendel, V., Göbel, S., & Steinmetz, R. (2012). Game mastering in collaborative multiplayer serious games. In S. Göbel, W. Müller, B. Urban, & J. Wiemeyer (Eds.), *E-Learning and Games for Training, Education, Health and Sports* (pp. 23–34). Berlin, Heidelberg: Springer.

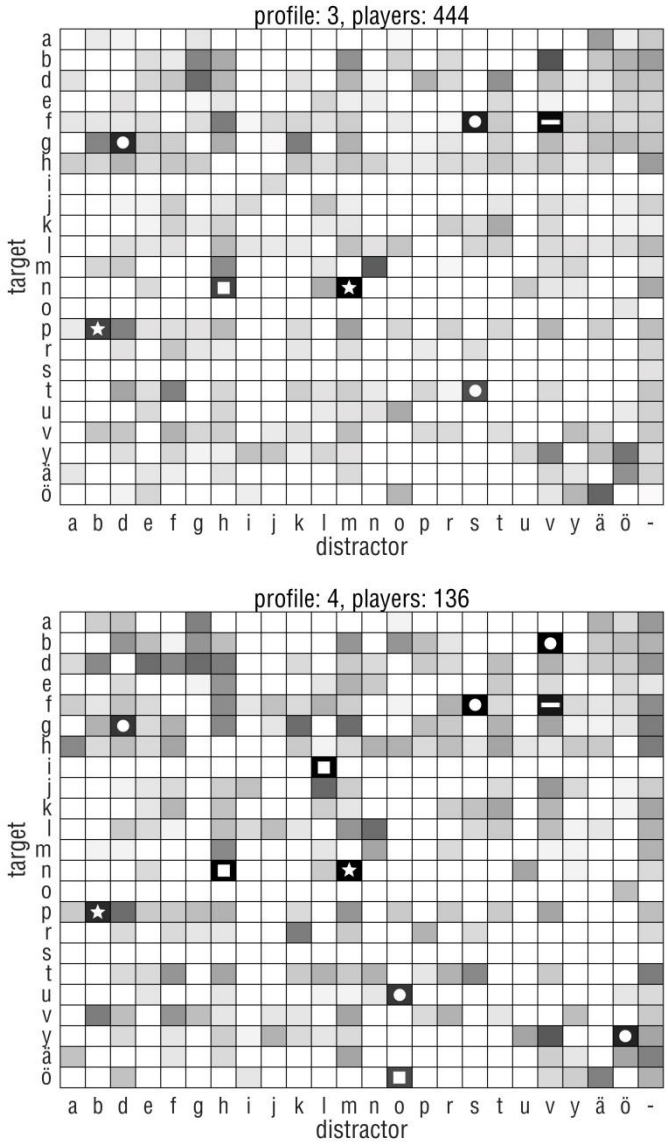
Yanjin, L. & Vincent, A. (2017). Educational game and intelligent tutoring system: A classroom study and comparative design analysis. *ACM Transactions on Computer-Human Interaction,* 24(3):1-27.

Supplement S1: Confusion matrices for lowercase letter data



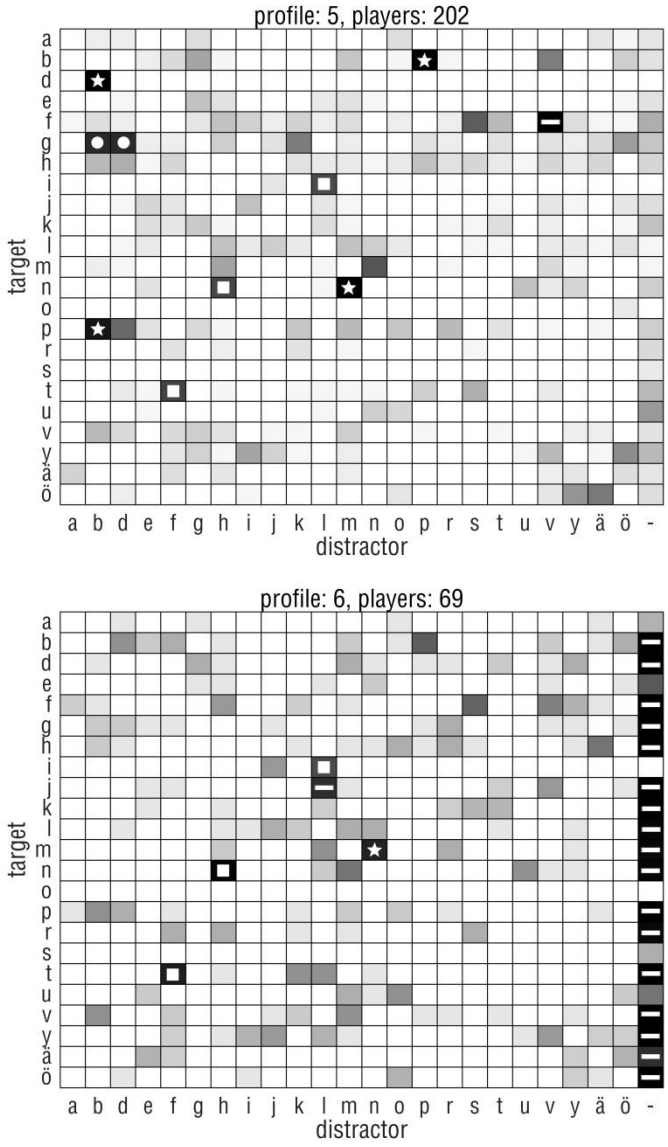
Darker colors indicate more confusion. Confusions exceeding 10 % are marked with symbols. There are two main categories of letters' similarity: phonetic similarity (marked with a "circle") and visual similarity (marked with a "square"). It is also possible that both or neither categories are occurring (marked with a "star" and a "rectangle", respectively).

Figure S1.1: Profiles 1 and 2 for lowercase letter data



Darker colors indicate more confusion. Confusions exceeding 10 % are marked with symbols. There are two main categories of letters' similarity: phonetic similarity (marked with a "circle") and visual similarity (marked with a "square"). It is also possible that both or neither categories are occurring (marked with a "star" and a "rectangle", respectively).

Figure S1.2: Profiles 3 and 4 for lowercase letter data



Darker colors indicate more confusion. Confusions exceeding 10 % are marked with symbols. There are two main categories of letters' similarity: phonetic similarity (marked with a "circle") and visual similarity (marked with a "square"). It is also possible that both or neither categories are occurring (marked with a "star" and a "rectangle", respectively).

Figure S1.3: Profiles 5 and 6 for lowercase letter data

Supplement S2: K-means clustering and validation indices

K-means clustering with missing data

The objective function for K-means clustering can be defined as:

$$\arg\min_{\mathbf{C}} \sum_{\mathbf{x}\in\mathbf{X}} d^2(\mathbf{x}, \mathbf{c}_k), \tag{1}$$

where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}, \mathbf{x}_i \in \mathbb{R}^n$, refers to a set of N observations, and $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K}$ are obtained cluster profiles. $d()$ denotes modified version of the l_2-norm. Since partially incomplete data, the modified norm is needed for clustering. The main idea of the modified approach is to use pairwise available components and scale the result to the missing components (Gower, 1971) .

K-means clustering method consists of two main steps: an initialization and local refinement steps (see Algorithm 1). These steps are usually performed using multiple restarts and the result with the smallest clustering error will be selected. In an initialization step a local partition of data is decided. The quality of clustering depends on the initialization step since clustering acts locally. A local refinement step perform local search which improve quality of initial partition. The aim of this step is to minimize clustering error, that is, summed distance of observations to the nearest prototypes. The step is performed in an iterative way assigning observations to the nearest prototypes and updating prototype locations. An advantage of K-means with K-means++ type of initializations is that it has only a linear time complexity and comparable fast convergence since K-means++ favors distinct prototypes in a data space (Arthur and Vassilvitskii, 2007).

Algorithm 1: Prototype-based clustering with K-means++ initialization

Input: Data set $\mathbf{X}$ and given number of profiles K

Output: Obtained cluster profiles, which minimize the objective function (1)

Select the first profile, $\mathbf{c}_1$, as an average value of observations in $\mathbf{X}$

for $j = 2, j = j + 1, j \leq K$ **do**

Select $\mathbf{c}_j$ randomly from $\mathbf{X}$ with probability:

$\min d^2(\mathbf{x}, \{\mathbf{c}_k\}_{k=1}^{j-1}) / \sum_{\mathbf{x}\in\mathbf{X}} \min d^2(\mathbf{x}, \{\mathbf{c}_k\}_{k=1}^{j-1})$

repeat

1. Assign each observation to the closest profile using $\min d^2(\mathbf{x}, \{\mathbf{c}_k\}_{k=1}^{j})$
2. Recompute the profiles as average values of the assigned observations.

until *The partition does not change*

end

Internal cluster validation indices

In K-means setting the number of clusters is essential to be determined. Internal cluster validation indices (CVIs) identify the number of clusters such that any external/prior information is not needed in the calculations. The most of the CVIs are defined by compactness and separability of the clustering result. The validity index provides a measure for each number of clusters. Depending on the used index formula, the lowest

or the highest measure is usually selected as the final number of clusters. Further, the number of clusters can be also selected using the speed of improvement of the cluster validation measures, for example, using a classical knee-point method (Thorndike, 1953).

Our previous study (Niemelä et al., 2018) presented the most commonly used validation indices. The reduced formulas were used since constant terms and monotone functions offered in the original formulas do not affect to the final solutions. In addition, the used formulas were extended for the general similarity measure. In the study, compactness was defined by Intra and separability by Inter. Compactness is usually defined by using summed variances of observations around prototypes in different clusters. Separability indicates how well distinct clusters are for each other. Minimum or maximum values of distances of all prototypes or variance of prototypes are popularly used variables. The study proposed formulas in the form where Intra was divided by Inter and thus they were attempted to be minimized.

In general, the decision of the number of clusters by using CVIs involves the following procedure:

1) Repeat clustering iteratively ranging K from $K_{\min}$ to $K_{\max}$. Obtain calculated cluster profiles and data partitions for each value of K based on Algorithm 1.
2) Calculate index measures using CVIs for each value of K. Form index curves based on the measured values.
3) Select the optimal number of clusters according to some decision criteria, for example, minimum/maximum values of cluster validation index curves or using speed of improvements of index measures.

Regarding to the described methods, the source codes are available online: http://users.jyu.fi/~mapeniem/BJET/Kmeans/

References

Arthur, D. & Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027–1035). New Orleans, Louisiana, United States: Society for Industrial and Applied Mathematics.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* (pp. 857-871). Washington, United States: International Biometric Society.

Niemelä, M., Äyrämö, S., & Kärkkäinen, T. (2018). Comparison of cluster validation indices with missing data. In *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 461–466). Bruges, Belgium: ESANN.

Thorndike, R. L. (1953). Who belongs in the family. *Psychometrika*, *18*(4): 267–276.

Supplement S3: Results for uppercase letter data

Uppercase data

The analyses of *GraphoLearn* game play data, which was originally performed for the lowercase data set were repeated by using uppercase letter data set. These results are given in Tables S3.1 – S3.2 and Figures S3.1 – S3.2, which can be shortly summarized.

Table S3.1 shows symbols for confusions exceeding 10 % and confusions exceeding 15 % are illustrated with underlined symbols. The most frequently mixed letters were *G*, *D*, *N*, and *M* similarly to the players who used lowercase letter data. Table S3.2 shows error rates from four profiles which were in the range of 34.5 % – 42.7 %. The results are mostly better than the calculated error rates from six profiles of lowercase letter data (30.7 % – 55.3 %). This may be related to fact that uppercase letters are visually less similar than lowercase letters. The progression information was calculated based on only few players because many of players played less than one hour and did not complete the second assessment. Therefore, these numbers give only limited information about the players' progression. The players of this data set have not actively played the game because the total playing times were remarkably smaller and the interval times were higher compared to the times gained from the players who used the lowercase letter data set.

Table S3.1: Symbol table of similarities for different uppercase data profiles

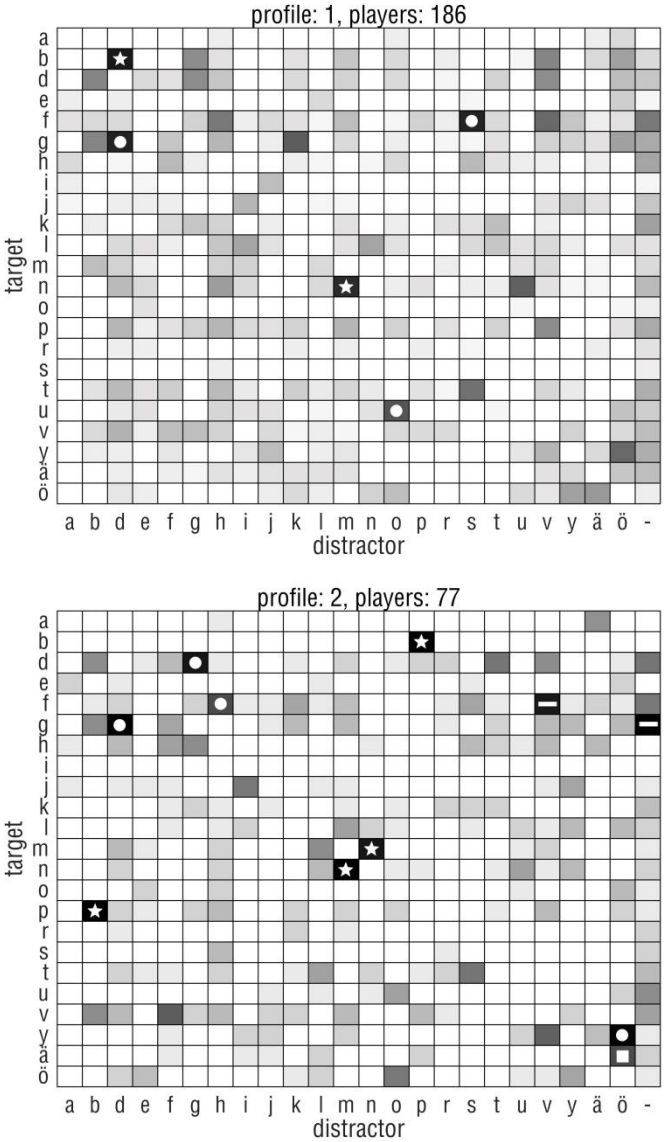
	target | distractor																	
	B|D	B|P	D|B	D|G	F|H	F|S	F|V	G|B	G|D	K|F	M|N	N|M	P|B	U|O	V|F	Y|Ö	Ä|H	Ä|Ö
profile																		
P1	☆					○			○			☆		○				
P2		☆		○	○		▢		○		☆	☆	☆			○		□
P3			☆	○				○	○		☆	☆	☆					
P4	☆				○	○			○	▢	☆	☆			▢	○	▢	
total	2	1	1	2	2	2	1	1	4	1	3	4	2	1	1	2	1	1

phonetic similarity=○, visual similarity=□, phonetic and visual similarity=☆, unknown category=▢

Table S3.2: Findings of uppercase data profiles

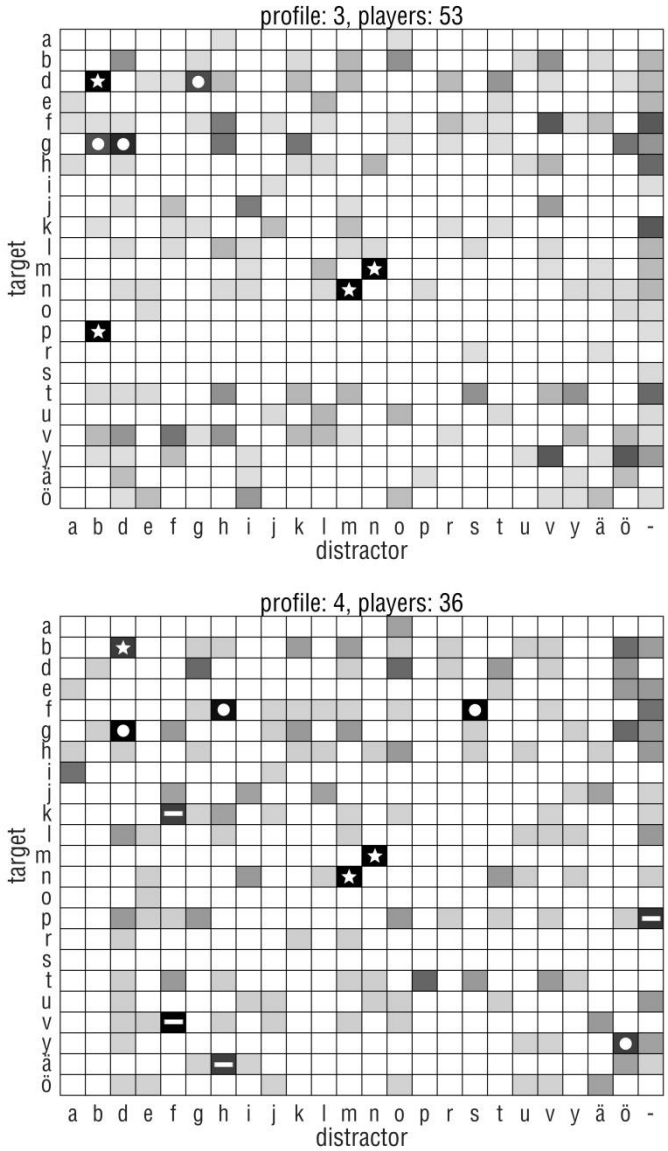
	profile				**all**
	P1	P2	P3	P4	P
statistics					
size (in %)	52.8%	21.9%	15.1%	10.2%	100.0%
error rate	34.5%	38.7%	36.2%	42.7%	36.5%
progression*	10.2%	23.0%	13.1%	11.9%	13.8%
playing time	36.8 min	46.5 min	38.2 min	29.7 min	39.5 min
interval time	7.0 days	8.1 days	7.0 days	9.5 days	7.5 days

*Only players who completed both assessments are included.



Darker colors indicate more confusion. Confusions exceeding 10 % are marked with symbols. There are two main categories of letters' similarity: phonetic similarity (marked with a "circle") and visual similarity (marked with a "square"). It is also possible that both or neither categories are occurring (marked with a "star" and a "rectangle", respectively).

Figure S3.1: Profiles 1 and 2 for uppercase letter data



Darker colors indicate more confusion. Confusions exceeding 10 % are marked with symbols. There are two main categories of letters' similarity: phonetic similarity (marked with a "circle") and visual similarity (marked with a "square"). It is also possible that both or neither categories are occurring (marked with a "star" and a "rectangle", respectively).

Figure S3.2: Profiles 3 and 4 for uppercase letter data