INVESTIGATION OF PROCESSING TEST RESULTS BASED ON KNOWLEDGE SIMILARITY

Bence Sipos[1]
Budapest University of Technology and Economics
Budapest, Hungary
0000-0002-6546-6966

Brigitta Szilágyi
Budapest University of Technology and Economics, Covinus University of Budapest
Budapest, Hungary
0000-0002-2566-0465

Conference Key Areas: *Mathematics at the heart of Engineering, Assessment*
Keywords: *similarity, knowledge-based, classification*

ABSTRACT

The traditional scoring is based on the difficulty of the task. However, the same total score can be earned with different knowledge, hence it is difficult to create homogenous groups by only relying on the total score. In our work, we aim to present a new scoring method where such knowledge-based groups can be created, as opposed to the previous point-based method.

For comparison of the test result done by the students, we utilized different distance measures. The main challenge with finding the similarities between the results is the high dimensionality of the data compared to the total number of observations.

First, we used the traditional Minkowski distance with different p values, then we used local similarity hashes and high dimensional embedding techniques designed originally for natural language processing. With these never before used techniques, we were able to identify students with similar skillsets and knowledge. Furthermore, we utilized dimension reduction methods (t-SNE and UMAP) to make a lower dimension representation where we can cluster the data easier. These clusters and pairwise similarities were assessed by oral exam subjective scores.

The teacher's subjective scores correlated more with this new metric-based method of scoring than the total test score itself. The presented procedures can not only be used effectively in research but can also help to get a more complete picture of the students' knowledge we teach in everyday practice.

[1] *Corresponding Author*

Bence Sipos

bence.sipos.sb@gmail.com

doi:10.5821/conference-9788412322262.1383

1 INTRODUCTION

Testing has several functions in the learning process. However, in many educational contexts, it is still used only to check knowledge: the test is the way in which the student reports on the extent to which he or she has mastered what has been learned. However, recent brain research suggests that tests may have a greater role: testing can be seen as part of the learning process, because rather than repeatedly introducing the knowledge to be learned, recall is a more effective tool for learning. With tests of the right difficulty and taken at the right time, we can acquire long-term knowledge.

The analysis of the results obtained is also an essential element. We can now visualise data in a number of ways, some of which are almost artistic, but they are not all about beauty . A well-done visualisation can also lead to the discovery of new correlations. Assessing the results obtained in different ways can enrich our analysis.

Evaluation as feedback is an important part of the learning process. There is evidence that timely feedback during learning increases the effectiveness of the process [1], [2]. Correct assessment and appropriate feedback also inform the learner whether his/her preparation has been adequate. As well as being fair, good assessment should also be motivating. It should encourage the student to progress to the next level by doing more work.

Good assessment of knowledge and ability is also important for the person doing the assessment, for example, if you want to create smaller, more homogeneous groups in a large, heterogeneous group of students to increase the effectiveness of teaching, it is important to define the groups well.

It is not uncommon for students to come to a course with very different levels of knowledge. Then, when evaluating their tests, it may be important to be able to find errors that result from a lack of knowledge (for example, a student cannot solve a problem because he or she does not know how to calculate scalar multiplication), in order to distinguish them from those that result from, for example, an incorrect algebraic transformation. With this knowledge, we can form groups of students much more effectively.

As the drop-out rate in STEM fields is also high in Europe, many higher education institutions are measuring first-year students to identify those at risk and start catching them up as soon as possible. It is also important to find first-year students with exceptional potential, for whom talent programmes offer the opportunity to develop their skills in a way that is appropriate to their abilities.

For many first-year university courses, it is a prerequisite that the student has passed a test from the secondary school curriculum at the appropriate level. Such tests, which can cover several subjects, are designed to check the knowledge required for entry to studies. It should be borne in mind, however, that if an incoming student fails a few tests in the first week, this may discourage him from starting his studies. It is also not ideal if, after the admission procedure, a lengthy series of tests is used to determine who needs catching up. In addition, there is often insufficient time to complete tests in class, and extra-curricular activities are difficult to organise for large classes. It is important to use the same or very similar assessment tests, but in this case, it is not possible to write at different times.

In the case of voluntary participation, we have found that the very people who do not attend the sessions advertised in this way are those who later have difficulties in progressing according to the curriculum. We therefore consider this solution to be of limited effectiveness.

It can therefore be seen that, while measurement is certainly desirable, it is not so easy to implement. And once we have measurement results, it is worth making the most of them. We now have increasingly efficient algorithms to help us with the evaluation.

In the following, we would like to show examples of what we can learn about a population by evaluating the results of a test using different methods. Often, we are looking for answers to the question of who wrote similar tests, how far apart the students are according to a metric of our choice.

One of the major problems to be solved in online education was to reduce the amount of cheating. The EduBase online education platform that we use has several features that try to reduce the use of illegal devices [3]. For example, the system alerts when the screen is being recorded or when the person completing the test leaves the page. The order in which the test taker progressed through the tasks and the time spent on each task can also be accurately tracked. But there are still plenty of opportunities for cheating. With the analysis we present, we have been able to filter out students who may have been communicating on another device.

Furthermore, we demonstrate a data visualisation procedure using graphs that displays the results of a test in an easy-to-understand format. In all cases, we will try to present our computational results in a visual way.

2 DATA

For our analysis, we use the results of a test we developed at the Budapest University of Technology and Economics, which consists of mathematical and linguistic elements and is used to identify those at risk of dropping out and those who deserve talent support. However, the procedures presented here can be used to analyze the results of any test. Only a handful of parameters are required for the analysis, the most important is the results of each question. This can be a continuous variable, an integer, or a binary value. Other parameters also can be used, for example, the time it takes to answer each question or the background of the student (ie: high school, study group). At our university, we have study groups and in online settings, they tend to work together, even if they should work independently. Previously, we used a single-component test with only mathematics problems to measure the proficiency of first-year students, but we found that the picture is more nuanced when language problems are included [4], [5]. Indeed, the results of the mathematics tests are distorted by the fact that students focus on practising different types of tasks already in preparation for the exam. This proved to be a winning strategy in that situation, the lack of real understanding did not significantly affect the outcome - the grade obtained. We often find that students who come to us are calculating derivatives or doing integration, but they do not have the correct understanding of the meaning of derivation or infinitesimal summation. The greater variability of language tests makes it much more difficult to learn without a deeper understanding, so the use of a complex test has proved more effective.

The maths part of the test consisted of 14 tasks, while the language part consisted of 30 tasks, with 90 minutes to complete. (The larger number of language tasks is due to the more detailed breakdown of the tasks for electronic evaluation, while in mathematics it was not necessary to break down the tasks into smaller parts.) Language questions are similarly varied. The tasks are weighted by difficulty. This provides us with a data set that measures a wide range of knowledge.

The other test on which we present a possible form of evaluation and data visualisation is the results of a test measuring geometric reasoning, compiled by Usiskin based on van Hiele's theory.

The subjects of both our tests were first-year students of mechatronics and energy engineering admitted to the Faculty of Mechanical Engineering of BME in 2020. A total of 153 participants took the test. They are students with high admission scores, high aptitude and the most homogeneous knowledge levels. So, when we define distances, we are actually looking at a population of very similar elements, according to their secondary school results, their matura exam, and trying to find the subtle differences that can be taken into account to create effective groups of students and to identify the gaps that need to be filled for them to succeed in the obstacles.

3 METHODS

Determining the distance between two data points is not always straightforward. Mathematicians have developed several methods to characterise this distance. There are two main directions: one is designed to measure the distance of binary vectors (vectors containing ones and zeros), and the other is for real-valued vectors. An example of the former is the Jaccard distance, which we use, and the latter is the Minkowski distance. The Minkowski distance reproduces several well-known distances with the appropriate choice of parameter, such as the Manhattan distance ($p = 1$) or the Euclidean distance ($p = 2$). There are also much more complex metrics with area-specific properties, but these will not be discussed due to space limitations.

These metrics can also serve as a basis for the construction of a graph. One way to create such a graph is to construct a matrix of pairwise distances, which becomes the adjacency matrix of the graph. On this basis, the weight of the edge between two data points will be proportional to the distance between them. The resulting matrices can now be well studied using graph theory methods but can also be used as a basis for manual studies using, for example, a Force Atlas visualization. In this case, the edge weights act as virtual springs, with samples of similar properties pulled together and outlier points pushed apart.

The best known and frequently used dimensional reduction method is principal component analysis. This allows us to identify the directions with the largest variance, which is useful information in the reduction process because it tries to keep the differences in those directions that allow the data to show distinct groups. This procedure gives a good picture of the distribution of the data globally but is not as detailed locally.

The t-SNE method tries to preserve this local relationship by keeping points that are nearby in the higher dimensions as well as in the lower dimensions. This closeness is mostly controlled by the adjustment of the number of nearest neighbours. The UMAP method similarly focuses on maintaining local similarities but is also more effective at maintaining global similarities. This method is also based on nearest neighbours and builds a graph based on their distances, optimizing the lower dimensional similarity using a stochastic gradient descent method. More modern versions of these latter two methods have appeared (TriMAP, PaCMAP), which are now able to produce good dimensionality reductions without complex parameterization, considering global and local features, but they do not always provide better results. These were also tested but did not give significantly better results.

We used Python for our analysis because it provides easy access to huge amounts of statistical and computational methods. The first step is data ingestion and cleaning. The data comes in XLSX format for the online learning platform. After removing the unnecessary variables and missing values we can compute the distances between the students. Based on these distances we can find the most similar students. A graph can be created based on this distance matrix, the distance matrix is as assigned as the graph’s adjacency matrix.

This similarity can be due to knowledge similarity or due to cooperation. Based on the timing of the test and the order of the questions we can sort out the cheaters. Overlapping time and the same order of completion can be telling in this situation. In the last part, we reduce the dimensions of the data with the aforementioned methods. This step is for visualization, manual similarity checking, and clustering.

4 RESULTS

The most common way of evaluating a test is to weight the tasks according to difficulty and then add up the scores for each task to get a final score, which can be used to give the student a grade. Students with the same score will then receive the same mark, but their knowledge may be different. Consider an essay where you want to check a wide range of knowledge. If you have hundreds of students taking a test, online testing is preferable as it is quicker to correct. Another advantage is that the results of the tasks corrected by the machine can be easily retrieved and used immediately for analysis.

Figure 1 illustrates the results obtained using UMAP on the study population. Each dot represents a student, the same colour indicates that part of the maths-language test that was designed to test how the knowledge acquired in high school was applied in situations that they had encountered in high school. (The first third of the math test assessed procedural skills, here were algebraic transformations, and calculations with identities. The second third was the part mentioned above, while the last third was to solve problems for which the students had acquired the necessary knowledge in high school, but the task was unusual, most likely they had not or had only rarely encountered the application of their knowledge in such a setting.) Although the colouring was done according to the scores on the middle block of the mathematics test, the separation, finding students with close knowledge of each other, was done based on the overall test scores. We can see that there are dense clusters of the same colour in the point cloud, marking students who are close to each other in terms of knowledge, but it is also clear that points of the same colour can be very far apart. If we want to create groups with homogeneous knowledge, then students represented by the distinct point clusters in the figure can be placed in a group.

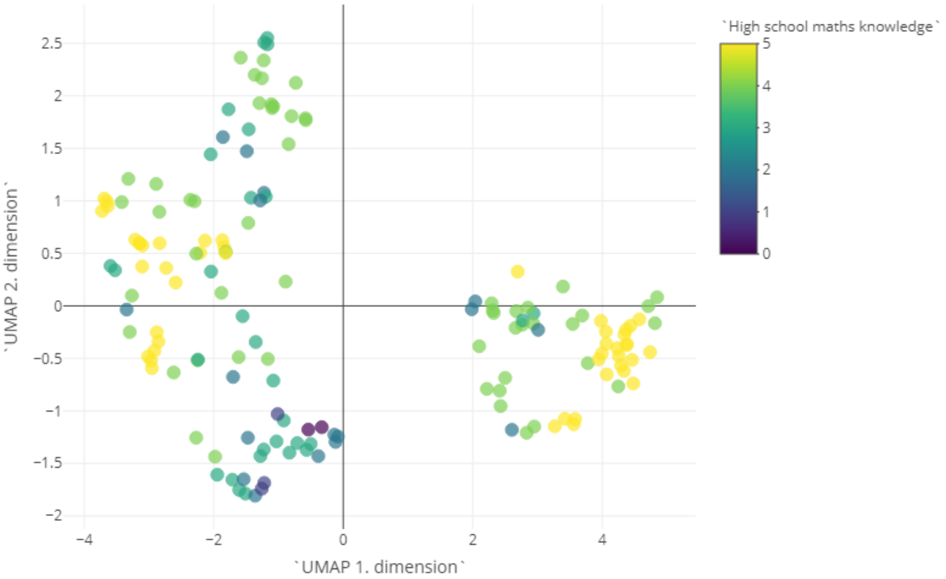

Fig. 1. UMAP results – the colour represents the high school math knowledge

The result of the clustering obtained by t-SNE is shown in Figure 2. Here again, each point represents one student, and the colours are defined according to the maths scores in the maths-language test. Red represents a score of 0, while purple represents the maximum score. The figure shows several small clusters, with those making the same type of error closer together. No larger, distinct clusters have emerged. It is also striking that students with a medium score are almost evenly distributed in the point cloud. This follows from the fact that they may have made the most type of mistakes.

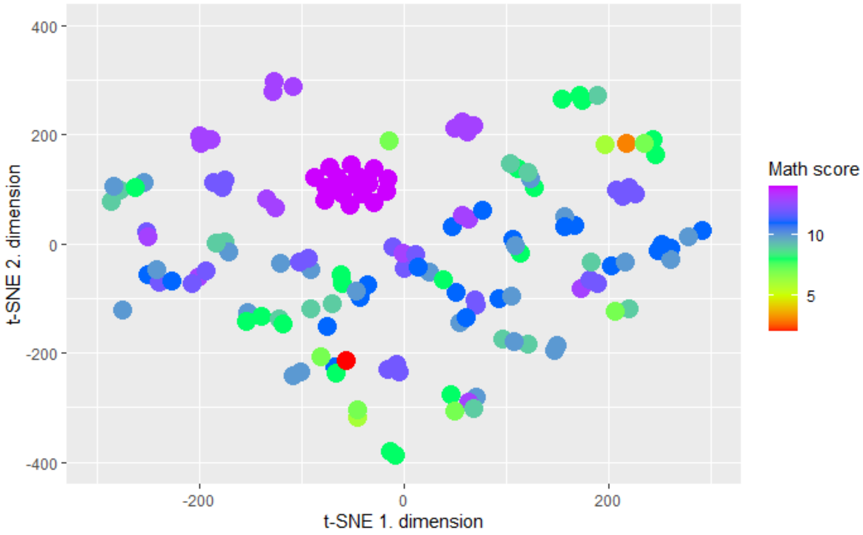

Fig. 2. Result of t-SNE based on the test – the color represents the total math score

In Figure 3, the dots are coloured according to the score on the mathematics test, representing one student. The figure illustrates the distance of each completer from a correct completion.

In the figure on the left, we have used the so-called L1 norm (Manhattan distance), while in the figure on the right, we have used L2 (Eucledian distance). The left graph shows that students with identical scores in mathematics are close to each other in this respect but can be very far apart in terms of their scores on the language test. Note, however, that those who scored poorly in maths were also far from perfect in the language test.

The figure on the right shows the distance according to the L2 norm. Here the reason for the clustering is due to the metric calculation, not to a real separation.

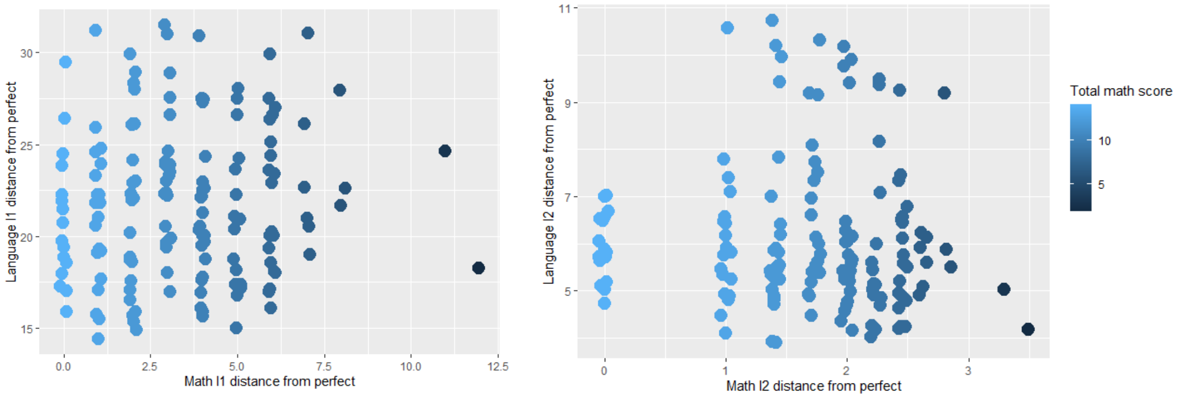

Fig. 3. L1 (left) and L2 (right) distance from a perfect test both in the math and language part

4.1 Investigating cooperations and cheating

The use of distance metrics and the overlapping of time periods allows cheating to be detected. If the match is high and the temporal relationship is given, cooperation may be at work in many cases. Proving this is difficult, but the closeness suggests that the students were working together. Cheaters are detected by examining the distance matrix and analysing it. There does not have to be a perfect match, just look for deviations much smaller than the average distances, for example, based on the histogram. The next step in the analysis is to examine the overlap in time. This procedure can play a key role in making online examinations more secure. A visualisation of our calculations is shown in Figure 4. The group circled in blue is the group of near-perfect candidates, where the reason for near-perfection is not cheating but few errors, while the group of a few candidates circled in green has the same number of errors and a similar temporal pattern. There is a suspicion that they wrote their test jointly.

4.2 Finding redundant questions

The 25 vertex graph in Figure 5 illustrates the results of the 25-question van Hiele test in the study population. The test reveals levels of geometric thinking. The van Hiele pair defined 5 levels of thinking, each level is measured by 5 questions. The first level is assessed by the first five questions, the second level by the second five questions, and so on. The test, questions and evaluations are described in detail in Usiskin's work [6]. The vertices of the graph are the tasks, which are connected by weighted edges. The weight of the edges is given by the probability of solving the problem of the two vertices of the edge. In the figure, the weight of the edges is represented by their thickness, length and colour. A thicker edge indicates a higher probability of solving. These are the short green edges in the graph. For example, many people solved problem 13 and 18 correctly. While the problems represented by the thin, long pink vertices (e.g., 2 and 9) were solved by few.

When assessing performance, we would like to assess the knowledge of those completing the tasks in depth. We would like to set a wide range of tasks, but the completion time should not be too long. The question arises as to which tasks can be discarded. The results of a test written in the framework of a pilot study can be visualised using the above procedure and it quickly becomes apparent that the middle tasks can be omitted from those that are linked with a heavy weighting edge. In this way we can optimise the completion time and reduce redundant items. This procedure can also help us in the teaching process: we can easily identify the types of tasks that need further practice and those that our students are already solving with good results.

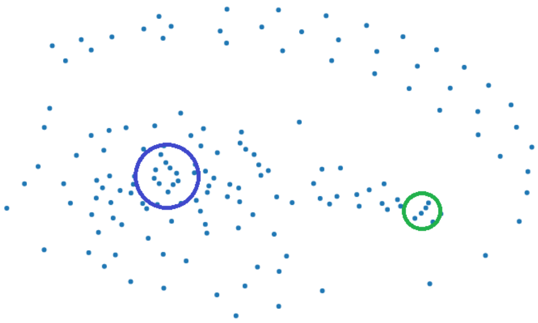

Fig. 4. Force-directed graph drawing of the knowledge similarity graph – blue circle shows the perfect tests, green circle for the suspicious students

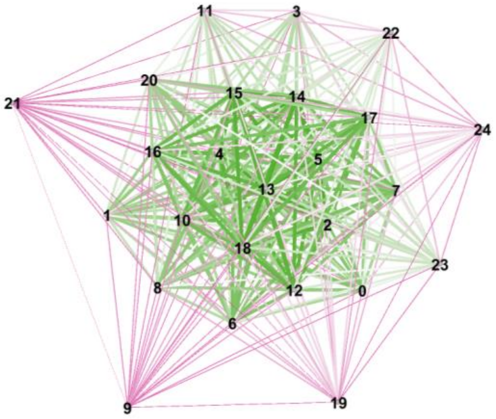

Fig. 5. Question similarity graph – green edge connects the similarly completed questions, purple for the large differences

5 CONCLUSION

Cheating became a huge problem in online exams during the pandemic and its detection requires manual reviews which is not feasible for large student groups. We encountered problems with knowledge homogeneity, some students lack a basic understanding of math due to inefficient online learning during their last year in high school. These large disparities make catch-up lessons essential.

Above, we have presented methods for evaluating test results, which not only visualize the results obtained in the exams but can also be useful for detecting cheating or for group creation based on similar knowledge, which can be crucial for efficient teaching.

ACKNOWLEDGEMENT

This research has been supported by the EFOP-3.4.4-16-2017-00025 project of the Hungarian Government and the European Social Fund of the European Union.

REFERENCES

[1] Kornell, N., Castel, A., Eich, T. S., Bjork, R. A. (2010): Spacing as the friend of both memory and induction in young and older adults. Psychology and Aging, 25, 2, pp. 498–503.

[2] Wheeler, M. A., Ewers, M., Buonanno, J. F. (2003): Different rates of forgetting following study versus test trials. Memory, 11, 6, pp. 571–580.

[3] www.edubase.net (last downloaded: 07. 04. 2022.)

[4] Szilágyi, B., Hornyánszky, G., Berezvai, Sz., Hives, Á., Horváth D., (2019), Novel prediction test for freshmen at BME, Faculty of Chemical Technology and Biotechnology, *Varietas delectat... Complexity is the new normality : SEFI 47th Annual Conference Proceedings,* pp. 1937-1947.

[5] Sipos, B., Oláh, T., Széles, K., Balogh, J., Szilágyi, B., (2021) How to utilize test results effectively? *Blended Learning in Engineering Education: challenging, enlightening – and lasting? SEFI 2021 49th Annual Conference Proceedings*, pp. 500-508.

[6] Usiskin, Z. (1982) Van Hiele levels and achievement in secondary school geometry, Final Report of the Cognitive Development and Achievement in Secondary School Geometry Project Department of Education, University of Chicago, US.