Piloting an automatic clustering algorithm to supplement placement test results for large groups of students

Marie-Pierre Jouannaud
Université Grenoble Alpes, France

Sylvain Coulange
Université Grenoble Alpes, France

Anne-Cécile Perret
Université Grenoble Alpes, France

Abstract

We report on the initial piloting of an online application using a co-clustering algorithm to supplement the results of a placement test (SELF) developed at Université Grenoble Alpes in six languages, and used at a number of partner universities in France. Automatic clustering models aim to group together similar objects (in our case, test-takers) according to certain variables (test items), and do this without supervision. Our co-clustering algorithm groups test-takers and items simultaneously by identifying groups of test-takers who answered similarly to groups of items (a first step toward learner profiles). The application, developed in R, provides a graphic interface enabling users to visually compare SELF and co-clustering results. It is possible to set the number of groups desired, which might be useful when many students receive the same test placement results but institutions want to make finer-grained groupings.

Introduction

The SELF placement test is a semi-adaptive multi-stage test developed at Université Grenoble Alpes in six languages (English, French as a Foreign Language, Italian, Japanese, Mandarin Chinese and Spanish) and used at a number of partner universities in France. The first stage of the test (the initial testlet) is common to all test-takers, but the items in the second stage depend on test-takers' results in the first. Results in the second stage are used to refine the estimation of learners' level and arrive at placement results expressed in Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) levels that are as reliable as possible. SELF was designed and developed following Association of Language Testers in Europe (ALTE) guidelines (2011): needs analysis, construct definition, choice of reference level descriptors, test and item specifications, item writing, reviewing and piloting, pretesting, standard setting, final test assembly, and post-administration analyses. Currently, SELF only provides test users with a CEFR 'aggregate level' (corresponding to proposed course level enrollment), and a CEFR level in three macro skills: listening comprehension (L), reading comprehension (R), and 'limited' writing (W), but does not provide further diagnostic information about test-takers. Our goal is to explore the use of automatic clustering to identify subgroups of learners, or to discover learner profiles, which could then be used to enrich the feedback given to learners and help them (and their teachers) decide what skills or areas they need to work on.

Co-clustering models

Automatic clustering models aim to group together similar objects (in our case, test-takers) according to certain variables (test items), and do this without supervision. The co-clustering algorithm we are using, derived from Latent Block Modeling or LBM (Brault & Mariadassou, 2015), groups test-takers and items simultaneously by identifying groups of test-takers who answered similarly to groups of items (right or wrong, since our items are dichotomous). LBM is especially suited to our needs because, being derived from mixture models, it creates homogeneous, non-overlapping groups (Brault & Lomet, 2015). Our application, developed in R, provides a graphic interface enabling users to visually compare SELF and co-clustering results. It is possible to set the number of groups desired, which might be useful when many students receive the same test placement results

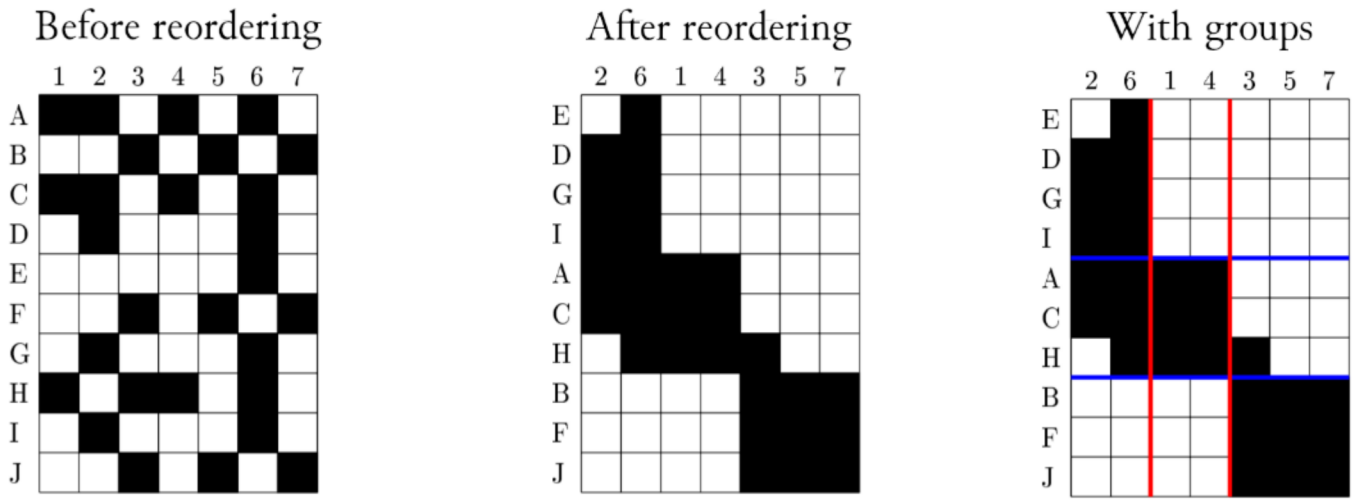

Figure 1 Three matrixes divided into 7 columns and 10 rows to group students according to item response patterns

but institutions want to make finer-grained groupings. Lastly, the algorithm identifies items whose results do not contribute significantly to the classification of test-takers (a so-called 'noise cluster').

In SELF, item levels were determined by a panel of experts in a standard-setting session convened before the test was originally assembled, which means that the results are interpretable in terms of CEFR levels and are thus directly useful to stakeholders (students, teachers, as well as administrative staff, for managing groups, scheduling, etc.). However, the result is ultimately based on total score, and does not distinguish between two students who received the same score but had different patterns of responses (i.e. did not answer the same items correctly). The co-clustering algorithm works in the opposite way: it only looks at patterns of responses and tries to identify groupings based on these patterns (Figure 1). Since it is unsupervised, no meaning is attached to items beforehand. Our objective is to see whether it is possible to make sense of these automatic groupings and whether they can be meaningfully interpreted in terms of learner profiles for diagnostic or placement purposes. In our example (the far right matrix in Figure 1), students A, C and H (lines) did well on items 1 and 4 (columns), but not on 3, 5 and 7, and students B, J and F responded in the opposite way. The question is whether we can identify what items 1 and 4 (or 3, 5, and 7) have in common, and whether we can use them to characterize the difference between students A, C, H on the one hand, and B, J, F on the other.

Initial results

Comparison of placement test and co-clustering results for student groupings

Because the algorithm has not been used with language test results before, the initial step in our analysis is a simple comparison of SELF placement results and the co-clustering algorithm in two groups of students (studying English or Japanese). We only used SELF results in the initial testlet, which separates test-takers into three groups (for the English test, A1/A2, B1 and B2/C1, B1 being the most common level observed in incoming students; for the Japanese test, A1/A2, A2/B1 and B1/B2), and we set the number of groups desired to three in the application.

Figure 2 shows (part of) the output of the application, with SELF results (here, for Japanese students) on the left-hand side, and co-clustering reordering on the right. The white squares correspond to right answers, and the dark ones to incorrect answers (all of the students answered all of the questions). The red/orange/green color coding corresponds to student groups according to SELF results in the initial testlet. We observe that, according to the co-clustering algorithm (on the right), some of the 'intermediate' (orange) students in SELF are more similar to 'advanced' (green) students than to other intermediate students, and are thus placed in the same group. The group of 'beginner' (red) students is essentially the same in SELF and with co-clustering.

For the Japanese as a foreign language cohort (n=101), the correlation between the two grouping methods (SELF and co-clustering) is 0,67 (Kendall's tau rank correlation coefficient), and in English (n=228), $\tau = 0{,}82$ (in both cases, $p \leftarrow .000$). We find high correlations for both languages, which is not surprising, given that the application uses patterns of successful responses and thus indirectly takes the total score into account. We feel that this validates the use of the co-clustering algorithm, which is capable of arriving at similar results as the placement test when the number of groups is the same, i.e. it can classify students into groups that are interpretable in terms of language level. We find similar results when we increase the number of groups desired.

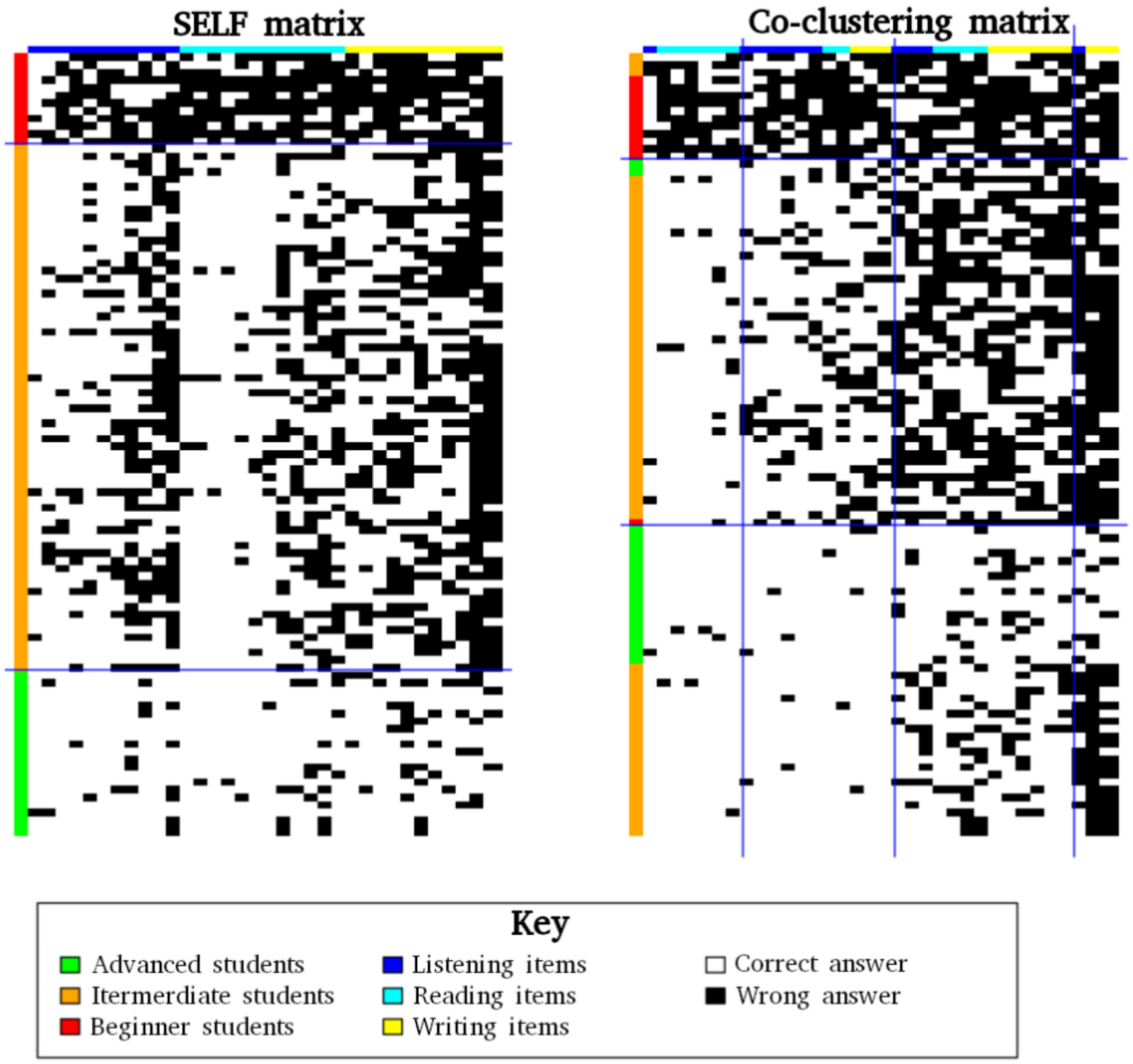

Figure 2 Screenshot of output of the application, with SELF results on the left, and co-clustering reordering on the right

Interpretability of item subgroups

Our original goal is not simply to use the co-clustering algorithm to group students according to levels (since we can already do that with our placement test), but to create more subgroups and/or to gain deeper insight into the characteristics of students placed in the same subgroup. In order to do this, we can analyze the item subgroups with the algorithm used to create the student groupings.

In the example above (Figure 2, right hand side), it is difficult to identify commonalities between items in each item subgroup. The first subgroup (first group of columns) is mostly composed of items targeting reading (light blue), with one listening item (dark blue). This first group might be said to contain 'comprehension' items, and could be used to characterize students according to their receptive skills (regardless of their results in productive skills). The next two item subgroups, however, are composed of items targeting all three language skills included in the test (listening, reading, and writing, in dark blue, light blue and yellow, respectively). The main difference between these two subgroups of items seems to be item difficulty, with the third group containing more difficult items than the second (as can be seen by the greater number of dark squares in the columns of the third item subgroup). The last group of items contains listening and writing items (two skills that do not necessarily have much in common) which are all difficult, as can be seen by the large number of dark squares in the last column, indicating that most students failed to answer correctly.

Thus, targeted language skill does not seem to be a relevant variable (over and above item difficulty) to interpret item groupings and define learner profiles according to these groupings. We are exploring the role of other item characteristics such as language focus (the critical information that item writers believe test-takers need to understand in order to answer the question correctly, which can be lexical, morphosyntactic or pragmatic), discourse type (the prevalent genre of the text the item bears on: narrative, informative, argumentative, etc.), and other characteristics laid out in item specifications, to see if these play a larger role in determining item groupings and students' response to them. Although the method is very different, the goal is similar to what cognitive diagnosis assessment (CDA) approaches have tried to accomplish (Liu, 2015): using results of large-scale tests, and characteristics of the items included in these tests, to provide diagnostic information to learners beyond general language results.

Conclusion and further study

SELF is currently used by more than 25 French universities and language centers, and more than 150,000 students have taken the test in one (or more) of its six foreign languages since it became operational in 2016. Data from administration to two groups of students (tested in Japanese and English) were used to explore the use of unsupervised co-clustering models to automatically create student groups based on their patterns of responses to groups of items, in an effort to automatically uncover learner profiles. We have shown that test-taker subgroupings by the co-clustering algorithm are interpretable in terms of language level and are very similar to SELF results based on CEFR levels. Item groupings, on the other hand, are interpretable in terms of item difficulty, but cannot at present be easily used to give finer-grained information about learner profiles.

These results are only preliminary, and we are exploring avenues for further study. One is the use of more item characteristics to try to interpret item subgroups created by the algorithm (test-taker characteristics could also be used to enrich the interpretation of test-taker subgroups). Another avenue is the analysis of items identified as 'noise' by the co-clustering algorithm (items that do not help in the definition of learner groups): are these items also characterized by lower discrimination in more traditional analyses (classical test theory)? Lastly, the algorithm in its present form does not respond well to missing data, which is why we have only used results to the first stage of our multistage test (the initial testlet), completed by all test-takers. We are working on integrating results to the second stage to enrich the data the co-clustering algorithm has access to.

References

Association of Language Testers in Europe. (2011). *Manual for Language Test Development and Examining*. Strasbourg: Language Policy Division, Council of Europe.

Brault, V., & Lomet, A. (2015). Revue des méthodes pour la classification jointe des lignes et des colonnes d'un tableau. *Journal de la Société Française de Statistique, 156*(3), 27–51.

Brault, V., & Mariadassou, M. (2015). Co-clustering through Latent Bloc Model: A review. *Journal de la Société Française de Statistique*, 156(3), 95–119.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Liu, H. H-T. (2015). The conceptualization and operationalization of diagnostic testing in second and foreign language assessment. *Working Papers in TESOL and Applied Linguistics*, *14*(1), 1–12.