Extracting Patterns from Educational Traces via Clustering and Associated Quality Metrics

Marian Cristian Mihăescu[1], Alexandru Virgil Tănasie[1], Mihai Dascalu[2(✉)], and Stefan Trausan-Matu[2]

[1] Department of Computer Science, University of Craiova, Craiova, Romania
mihaescu@software.ucv.ro, alexandru_tanasie@yahoo.com
[2] Computer Science Department, University Politehnica of Bucharest, Bucharest, Romania
{mihai.dascalu,stefan.trausan}@cs.pub.ro

Abstract. Clustering algorithms, pattern mining techniques and associated quality metrics emerged as reliable methods for modeling learners' performance, comprehension and interaction in given educational scenarios. The specificity of available data such as missing values, extreme values or outliers, creates a challenge to extract significant user models from an educational perspective. In this paper we introduce a pattern detection mechanism with-in our data analytics tool based on k-means clustering and on SSE, silhouette, Dunn index and Xi-Beni index quality metrics. Experiments performed on a dataset obtained from our online e-learning platform show that the extracted interaction patterns were representative in classifying learners. Furthermore, the performed monitoring activities created a strong basis for generating automatic feedback to learners in terms of their course participation, while relying on their previous performance. In addition, our analysis introduces automatic triggers that highlight learners who will potentially fail the course, enabling tutors to take timely actions.

Keywords: Clustering quality metrics · Pattern extraction · k-means clustering · Learner performance

1 Introduction

Finding educational patterns reflective of learner's performance represents a research question of particular interest that has been addressed from different perspectives in various contexts. As clustering has emerged as a de facto method for finding items that naturally group together, there is a wide range of algorithms that can be used to classify individual, differentiated mainly into three classes: hierarchical, non-hierarchical and spectral.

In terms of available data, items may be represented by actors (e.g., students or tutors) or learning assets (e.g., concepts, quizzes, assignments, etc.). Moreover, due to the variety of e-learning platforms, there is also a wide variety of structured and well defined input data, for example the PLSC DataShop [1] or the UCI Machine Learning Repository Student Performance Data Set [2]. However, designing and running a clustering process

© Springer International Publishing Switzerland 2016
C. Dichev and G. Agre (Eds.): AIMSA 2016, LNAI 9883, pp. 109–118, 2016.
DOI: 10.1007/978-3-319-44748-3_11

on a particular dataset in order to find meaningful educational insights represents a challenge.

In this paper we introduce an automatic feedback mechanism derived from pattern extraction data analytics applied on educational traces. Our approach for finding patterns is based on: (a) a continuous evaluation in terms of clustering quality metrics (CQM) applied on data models generated by the *k*-means algorithm and (b) the automatic highlight and manual removal by means of visual inspection of outliers and extreme values. From an educational perspective, generated patterns enable us to categorize new students based on their continuous performance evaluation, thus providing automatic personalized feedback. In other words, our main contributions cover the following aspects:

1. Define a custom data analytics pipeline that identifies relevant clusters by usage of clustering quality metrics (CQMs) in an educational context. The CQMs are implemented for Weka [3]. The current release does not support any standardized data models as outputs (i.e., cluster assignments based on standard CQMs).
2. Develop a software tool suited for finding educational patterns and providing automatic feedback to both learners and tutors.

The paper continues with a detailed state of the art, followed by methods and results. Afterwards, it discusses key findings derived from our dataset, points out strengths and limitations of our approach, and provides a roadmap for future work centered on increasing the efficacy of the educational tasks at hand.

2 Related Work

The quality of a clustering process is strongly related to the available dataset, the chosen clustering algorithm, as well as the implemented CQMs. The domain specific training dataset and the implemented quality metrics represent the building blocks for interpreting results in correspondence to the proposed learning analytics task. Within our experiments, we have opted to select a dataset from the Tesys e-learning platform [4]. As algorithm, we have used the *k*-means clustering [5] algorithm implemented in Weka [3]. Weka also implements Expectation Maximization [6] and COBWEB [7] for hierarchical clustering, as well as DBSCAN [8], but the selected algorithm was most adequate for our experiments in terms of visual representation.

However, the main drawback of Weka is the lack of implemented CQMs that would enable an easy and reliable analysis of the clustering outcomes for various clustering algorithms. Multiple CQMs are defined in literature and may be classified into various types, depending on the underlying approach: similarity metrics, clustering evaluations, clustering properties, distance measures, clustering distances, as well as cluster validity measures.

First, Jackson, Somers and Harvey [9] define similarity metrics as a way to compute coupling between two clusters. In contrast, Sneath and Sokal [10] cover the following distance, association and correlation coefficients: Taxonomic (distance), Camberra (distance), Jaccard (association), Simple matching (association), Sorensen-Dice (association), and Correlation (correlation).

Second, silhouette statistical analyses [11, 12] and visualization techniques have been successfully used by Jugo, Kovačić and Tijan [13] as main analytics engine for the cluster analysis of student activity in web-based intelligent tutoring systems.

Third, Hompes, Verbeek and van der Aalst [14] define specific clustering properties, i.e., cohesion, coupling, balance and clustering score. Their methodology evaluates weighted scoring functions that grade a clustering activity with a score between 0 (bad clustering) and 1 (good clustering).

Forth, the similarity measures between two clusters as defined by Meila et al. [15, 16] cover Clustering Error (CE), the Rand index (as a well-known representative of the point pair counting-based methods), and the Variation of Information (VI) [15]. Additional measures for comparing clusters were also defined: Wallace indices [17], Fowlkes-Mallows index [18], Jaccard index [19], or Mirkin metric [20]. These later distance measures are necessary to perform external cluster validation, stability assessments based on internal cluster validation, meta-clustering, and consensus clustering [16]. Classic cluster validity measures defined by Stein, Meyer zu Eissen and Wißbrock [21] include precision, recall, f-measure, Dunn index, Davies-Boulden, Λ and ρ measures. These internal measures determine the usefulness of obtained clusters in terms of structural properties.

From the information theory point of view, CQMs should meet the following primary axioms set by Kleinberg: (a) scale invariance, (b) consistency, and (c) richness. In addition, CQMs should follow the principles of relative point margin, representative set and relative margin [22]. However, aside from the wide variety of previously mentioned metrics, *SSE* (Sum of squared errors) is one of the most frequently used functions when inferring clusters in *k*-means algorithms. Minimization of this objective function represents the main goal and SSE may be regarded as a baseline quality threshold for most clustering processes.

From the application development perspective, many educational applications make intensive use of clustering processes. For example, Bogarín, Romero, Cerezo and Sánchez-Santillán [23] improve the educational process by clustering, Li and Yoo [24] perform user modeling via clustering, while Bian [25] clusters educational data to find meaningful patterns.

3 Method

Our learning analytics pipeline is built into two core components, one centered on data processing (Clustering and CQM identification) and the other focused on visual analytics. Data processing is performed at server side based on raw data from the e-learning platform. Visual analytics are performed in the web client using D3js (https://d3js.org/), a JavaScript library for manipulating network graphs. Figure 1 presents the high level workflow of our learning analytics pipeline.

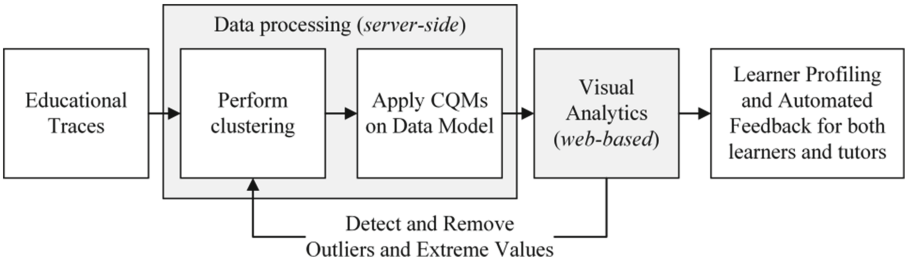


Fig. 1. Data analytics pipeline.

The input dataset has been obtained from logging activity performed during the academic year by students interacting with the Tesys e-learning platform [4]. The platform has specific logging mechanisms for extensively recording raw activity data that is divided into logging, messaging and quiz answering, indices described in detail in Sect. 4.

The data processing module creates the data model based on the input educational traces, the selected algorithm and the associated CQMs, generating in the end specific clusters and the computed quality metrics. Besides the data model itself, the data processing module is also responsible for identifying outliers and extreme values in accordance to the setup and the threshold values enforced by the data analyst.

The visual analytics module displays the generated clusters and marks outliers and extreme values as described by the data model. Its main goal is to properly present the model in a visual manner to the data analyst, facilitating the decision making process of manually removing outliers and extreme values under continuous evaluation of CQMs and centroid values. From this point forward, the design of our data analysis pipeline allows continuous trial and error sessions for the data analyst until relevant results from an educational perspective are obtained. The main characteristic of this module and of the entire pipeline is that it provides two types of relevant data: numerical information (by means of CQMs) and visual representations (by spatially generating presentations of clusters, centroids and outliers/extreme values). This enables the data analyst to properly decide if the current result is satisfactory or may potentially indicate that additional refinements need to be performed for improving the clustering results.

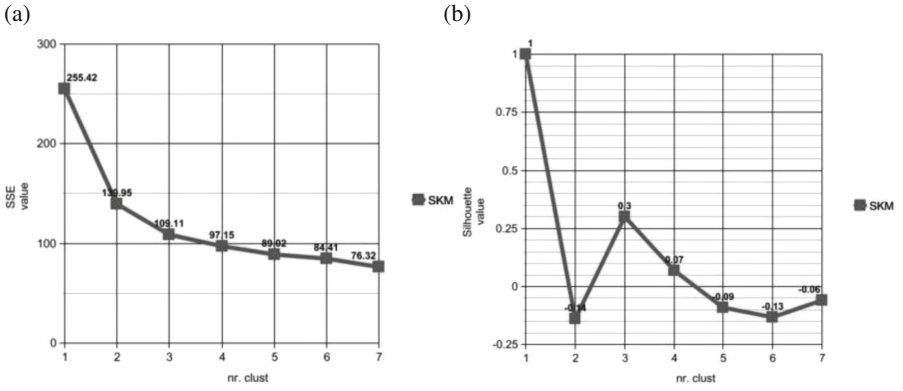
Once the patterns have been identified, the characterization of a new student may be performed. Assuming that three student clusters have been identified and manually labeled as “underperforming”, “average” and “well-performing”, new students may be characterized as belonging to a certain cluster via *k*-Nearest Neighbors algorithm. The labeling is performed by the tutor while taking into account the feature names and their intuitive interpretation. Proper usage of the CQMs in the labeling process is accomplished through the guidance of a data analyst. The overarching goal remains to improve each student’s predicted knowledge level and to reduce as much as possible the number of underperforming students. Furthermore, parameters with a high variance need to be taken into consideration in order to determine the adequate course of action, which becomes in turn the automatic feedback provided by our system.

4 Results

The input dataset consisted of 558 students characterized by 17 attributes described in Table 1. Our initial run of the k-Means clustering algorithm produces the graphs depicted in Figs. 2 and 3 in terms of SSE, Silhouette, Dunn and Xi-Beni indices, as well as the centroids presented in Table 2.

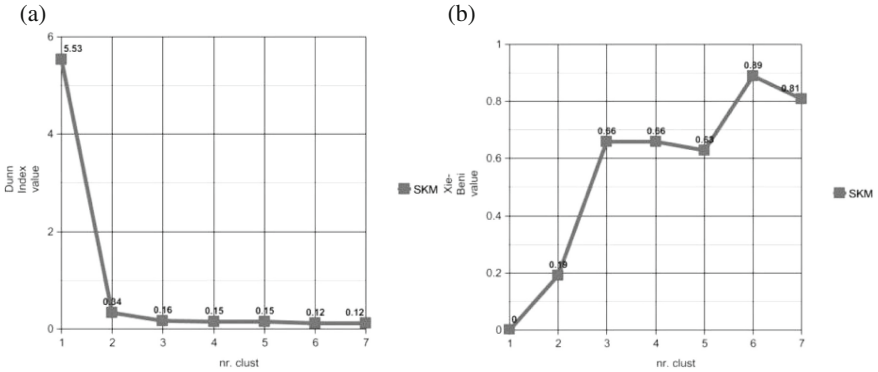
Table 1. Initial classification attributes.

Attribute name	Description
A. Session parameters	
Total	Sessions started by the user
W[n]	Sessions in the n-th week, where n is takes values from 0 to 6
B. Message parameters	
TS	Total sent messages
TR	Total received messages
AS	Average duration between two consecutive sent messages
AR	Average duration between two consecutive received messages
WS	Words from sent messages
WR	Words from received messages
C. Self-assessment parameters	
TQ	Total answered questions
CQ	Correctly answered questions
PQ	Percentage of correctly answered questions



SSE values. Silhouette values.

Fig. 2. (a) SSE values. (b) Silhouette values.



Dunn index values. Xi-Beni values.

Fig. 3. (a) Dunn index values. (b) Xi-Beni values.

Table 2. Feature values for centroids.

	Total	W0	Wl	W2	W3	W4	W5	W6
low	*27.17*	*3.66*	*4.48*	*4.47*	*4.10*	*3.86*	*3.62*	*2.98*
high	**75.21**	**11.98**	**12.31**	**1133**	**10.59**	**11.9**	**9.41**	**7.69**
average	31.33	4.1	4.58	4.84	5.00	5.09	4.10	3.61
	TS	TR	AS	AR	WS	WR	TQ	CQ	PQ
low	22.29	**9.92**	19.34	**319.70**	560.73	**231.28**	*151.51*	*58.42*	*21.10*
high	**50.2**	6.76	**23.93**	144.35	**1022.76**	155.88	**1540.35**	**1042.16**	**68.40**
average	*9.99*	*2.50*	*5.20*	*75.86*	*234.72*	*78.23*	215.76	82.50	32.88

Figure 2a shows an elbow for k equals three clusters. For two clusters, a significant decrease is observed −45.2 % from the SSE maximum value. While moving forward to three clusters, there is also an important decrease of 22 % from the previous SSE value. The subsequent decreases in SSE values are smaller than 10 %, thus leading to the conclusion that the optimal number of clusters in terms of the steepest descent may be two or three. Decision among these values is a matter of domain knowledge, but additional CQMs provide insight in terms of the optimal selection.

The silhouette values from Fig. 2b are a clear indicator of the quality of generated clusters. A value between .25 and .5 reflects a week structure, while a value between .5 and .7 highlights a reliable structure. Due to the nature of analysis, we do not expect strong structures since we are monitoring activities performed by learners. The .3 value obtained for three clusters is by far the highest score and indicates an appropriate number of clusters, with a reasonable structure. All other values are close to zero or negative, meaning that no substantial structure could be identified.

Dunn index values (see Fig. 3a) lays between zero and ∞ and should be maximized as higher values indicate the presence of compact and well separated clusters. The obtained index values show that the clusters are not well defined and this is usually due to the presence of noise in the input dataset. For our particular experiment, this index

yields to the conclusion that a high percentage of extreme values and outliers are found in the dataset.

The Xi and Beni index (see Fig. 3b) is significant when a value is a minimum of its neighbors. Therefore, two and five clusters represent minimum values while disregarding the zero value for one cluster. Still, the index value for three clusters is 4.5 % higher that the index value for five clusters; thus, from an educational perspective, we may consider that three clusters are more representative than five.

Based on all previous CQM analyses, we can conclude that the optimal number of clusters for our analysis is three. Starting from this categorization, Table 2 presents each centroid's feature values. After color coding each minimum, maximum and middle values, we can easily observe that each centroid becomes representative for a corresponding learner level ("under-performing", "average" and "well-performing"), highlighting also that the obtained centroids have educational significance.

In addition, a Principal Component Analysis was applied in order to better represent our data and to reduce the dimensionality of the initial classification attributes. The *Total* attribute was eliminated due to multicollinearity to W[0–6], while the *AS* attribute was disregarded due to low communality. In the end, 4 principal components were identified as having corresponding eigenvalues of 1 or greater and accounting for 90.13 % variance. The identified components represent a refinement of the initial classification, as messages are now split into in-degree and out-degree while relating to each student's activity (Table 3).

Table 3. Rotated component matrix using varimax with Kaiser normalization.

Classification attribute	Component 1	2	3	4
1. Logging activity				
W[0]	.923			
W[1]	.941			
W[2]	.943			
W[3]	.944			
W[4]	.940			
W[5]	.916			
W[6]	.888			
2. Testing activity				
TQ		.967		
CQ		.967		
PQ		.872		
3. In-degree messaging activity				
TR			.971	
AR			.864	
WR			.942	
4. Out-degree messaging activity				
TS				.975
WS				.985

For visualization purposes, three emerging components – *Logging*, *Testing* and *Messaging activities* (aggregation of both in-degree and out-degree) – were used to create dedicated views that facilitate the identification of outliers and extreme values. Figure 4 presents a print screen from our visual analytics application in which the centroids are marked with a large circle and the assigned items have similar colors. Three clusters (orange, green and blue) were annotated as representing "underperforming", "average" and "well-performing" learners. As the coordinate axes represent Testing and Messaging activities, it becomes obvious that students who were most engaged and performed best at their tests were clustered into the well-performing group, while the ones with the lowest participation defined the underperforming cluster.

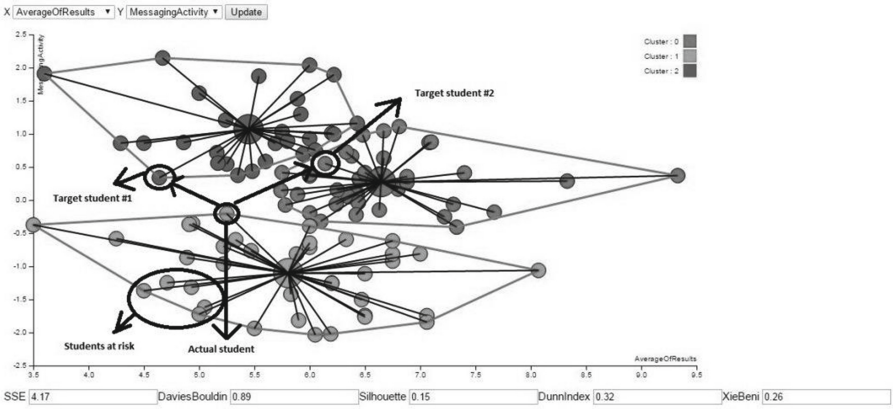


Fig. 4. Clustering visualization with corresponding CQMs.

Once students have been assigned to a cluster, they may select target students with better performance from another cluster or from their own cluster and just-in-time recommendations are offered as a means to improve their activity (e.g., X relevant additional messages should be posted, or Y supplementary tests should be taken). Moreover, the tutor is presented with a list of at-risk students who represent a subset of

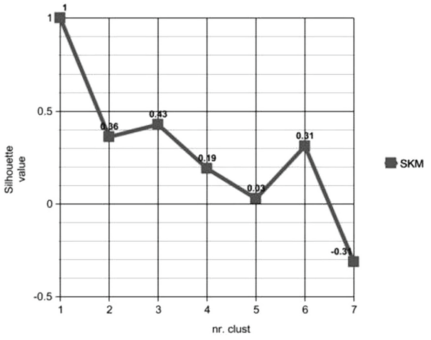


Fig. 5. Silhouette values after the removal of extreme values and outliers.

underperforming students with the lowest activity level. For example, 5 students were marked as "at-risk" in Fig. 4 and will receive personalized feedback from the tutor, encouraging them to become more actively involved.

Starting from the initial distribution, outliers and extreme values are removed using Interquartile Range with an extreme values factor set to 1.5 and an outlier factor set to 1.1. The outlier and extreme value removal led to a Silhouette value of .43 (see Fig. 5). This significant increase from .3 highlights a more solid structure of clusters, thus creating a reliable baseline for providing feedback to learners.

5 Conclusions and Future Work

Our approach consists of a learning analytics solution that integrates clustering algorithms, PCA transformations, visual analytics and clustering quality metrics. In order to obtain a detailed insight on the quality of the clustering process, several CQMs were implemented in order to facilitate the categorization of learners. Outlier and extreme values removal produced more robust structures, effective for running learning analytics and feedback mechanisms. From an educational perspective, reference learners were clustered and used for inferring educational traces corresponding to newcomers in the course.

Future work includes the integration of additional features to represent learners, implementation of other CQMs for a better evaluation of the obtained models, as well as comparative clustering algorithms which might be more appropriate for specific learning analytics tasks. A timeline view focused on displaying the knowledge level of students or centroids for each successive week will be introduced in order to provide a more fine-grained perspective of each learner's engagement.

Acknowledgements. The work presented in this paper was partially funded by the FP7 2008-212578 LTfLL project and by the EC H2020 project RAGE (Realising and Applied Gaming Eco-System) http://www.rageproject.eu/ Grant agreement No. 644187.

References

1. Koedinger, K.R., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: the PSLC DataShop. In: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R. (eds.) Handbook of Educational Data Mining. CRC Press, Boca Raton (2010)
2. Cortez, P., Silva, A.: Using data mining to predict secondary school student performance. In: 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), Porto, Portugal, pp. 5–12 (2008)
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
4. Burdescu, D.D., Mihaescu, M.C.: TESYS: e-learning application built on a web platform. In: International Conference on e-Business (ICE-B 2006), Setúbal, Portugal (2006)

5. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–38 (1977)
7. Dasgupta, S., Long, P.M.: Performance guarantees for hierarchical clustering. J. Comput. Syst. Sci. **70**(4), 555–569 (2005)
8. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231. AAAI Press (1996)
9. Jackson, D.A., Somers, K.M., Harvey, H.H.: Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? Am. Nat. **133**(3), 436–453 (1989)
10. Sneath, P.H.A., Sokal, R.R.: Principles of Numerical Taxonomy. W.H. Freeman, San Francisco (1963)
11. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
12. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. An Introduction to Cluster Analysis. Wiley-Interscience, New York (1990)
13. Jugo, I., Kovačić, B., Tijan, E.: Cluster analysis of student activity in a web-based intelligent tutoring system. Sci. J. Maritime Res. **29**, 75–83 (2015)
14. Hompes, B.F.A., Verbeek, H.M.W., van der Aalst, W.M.P.: Finding suitable activity clusters for decomposed process discovery. In: Ceravolo, P., Russo, B., Accorsi, R. (eds.) SIMPDA 2014. LNBIP, vol. 237, pp. 32–57. Springer, Heidelberg (2015). doi: 10.1007/978-3-319-27243-6_2
15. Meilă, M.: Comparing clusterings by the variation of information. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 173–187. Springer, Heidelberg (2003)
16. Patrikainen, A., Meilă, M.: Comparing subspace clusterings. IEEE Trans. Knowl. Data Eng. **18**(7), 902–916 (2006)
17. Wallace, D.L.: Comment. J. Am. Stat. Assoc. **383**, 569–576 (1983)
18. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **383**, 553–569 (1983)
19. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)
20. Mirkin, B.: Mathematical Classification and Clustering. Kluwer Academic Press, Boston (1996)
21. Stein, B., Meyer zu Eissen, S., Wißbrock, F.: On cluster validity and the information need of users. In: 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 2003), Benalmádena, Spain, pp. 404–413 (2003)
22. Ben-David, S., Ackerman, M.: Measures of clustering quality: a working set of axioms for clustering. In: Neural Information Processing Systems Conference (NIPS 2008), pp. 121–128 (2009)
23. Bogarín, A., Romero, C., Cerezo, R., Sánchez-Santillán, M.: Clustering for improving educational process mining. In: 4th International Conference on Learning Analytics and Knowledge (LAK 2014), pp. 11–15. ACM, New York (2014)
24. Li, C., Yoo, J.: Modeling student online learning using clustering. In: 44th Annual Southeast Regional Conference (ACM-SE 44), pp. 186–191. ACM, New York (2006)
25. Bian, H.: Clustering student learning activity data. In: 3rd International Conference on Educational Data Mining, Pittsburgh, PA, pp. 277–278 (2010)