Journal of Engineering Sciences
Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

Incremental And Adaptive Fuzzy Clustering For Virtual Learning Environments Data Analysis

Apoorva A[1]
Assistant professor
BCA Department
New Horizon College-Marathalli

Dr. Bharathi S[2]
Associate Profesor
MCA Department
Dr. Ambedkar Institute of Technology

Abstract—Virtual Learning Environments (VLE) offer a wide range of courses and learning supports for students. Such innovative learning platforms generate daily a huge quantity of data, regarding the interactions among the students and the VLE. To analyze these big educational data a new research branch called educational data mining (EDM) has emerged, that puts together computer scientists and pedagogues researchers' expertise. So far, educational data have been studied as stationary data by traditional machine learning methods. Rather, educational data are non-stationary in nature and can be better analyzed as data streams. In this paper we investigate the use of an adaptive fuzzy clustering algorithm called DISSFCM (Dynamic Incremental Semi-Supervised FCM) to process educational data as data streams and predict the students' outcomes to one exam module. Numerical experiments on the Open University Learning Analytics Dataset (OULAD) show the reliability of DISSFCM in creating good classification models of educational data. Index Terms—Educational Data Mining, Virtual Learning Environments, Data stream classification, Fuzzy clustering.

I. INTRODUCTION

Educational tools have drastically changed in the last decades, thanks to the advent of digitization and Internet. The use of Virtual Learning Environments (VLEs) has exponentially increased, because, on the one hand a great reduction in management costs is achieved, if compared to physical universities, and on the other hand the students' enrolling is facilitated, by eliminating the physical distance between them and the university. Moreover, VLEs allow personalized student support measures that take into account their needs, their weaknesses and strengths. The daily interaction of students with VLE platforms produces a large amount of data describing the student himself. It is a digital footprint of how each student is engaging with the learning materials and activities.

Thanks to the increased availability of this kind of data, a new research branch called Educational Data Mining (EDM) has recently attracted lot of interest. EDM uses Machine Learning techniques to analyze educational data in order to extract students' behavior models useful to predict their future performances. The possibility to detect, during the learning process, any risks of failure for the enrolled students, represents a very powerful tool for all the stakeholders that are involved in VLEs, such as teachers, tutors, students, and managers. Indeed, all of them could take advantage from information embedded in students models by different point of views. Particularly adaptive feedback, customized assessment, more personalized attention to prevent student failures [27] and to improve student retention [32], could be implemented by considering the suggestions coming from a data analysis process.

II. EXISTING SYSTEM

Several studies proved that machine learning techniques can be successfully used in the educational field [9]–[11], [16]. In [3], [6] the hidden learners skills, that are necessary to pass a test, are automatically extracted in form of Q-matrix from the questionnaires results. In [28], [29], [31], [34] learning analytics methods to support the big quantity of data coming from student-VLE interactions have been proposed. Several methods have been proposed to predict students' performances [2], [12], [21] or to measure students' satisfaction [33]. In [8], [20] learners are grouped in categories that are extracted from empirical data. Finally, visualization techniques have been used to directly observe the interactions between students and VLEs [15], [18], [23].

However, to the best of our knowledge, none of the proposed solutions takes into account the intrinsic streaming nature of educational data. They are big data that are continuously produced and that may evolve during the time. To analyze such kind of stream data we need incremental algorithms that are able to process the data sequentially and maintain a summary of the data using less space than the size of the data. Furthermore, although data are possibly unlimited, algorithms should use limited computational and storage resources, and have limited direct access to the data but need to provide answers in nearly real time [1], [13], [14]. Incremental and adaptive algorithms fit naturally to this scheme, since they can continuously incorporate new information into the constructed model, and traditionally aim for minimal processing time and space. Due to their ability of continuous large-scale and realtime processing, adaptive learning algorithms have gained more attention particularly in the context of Big Data [17]

III. PROPOSED SYSTEM

In [5] we proposed an incremental and adaptive clustering algorithm called DISSFCM (Dynamic Incremental Semi-Supervised FCM) which is specifically designed for data stream classification. DISSFCM is an incremental and semi-supervised version of the well known Fuzzy C-Means (FCM) clustering algorithm that is applied to subsequent,

non-overlapping chunks of data assumed to be continuously available during time. The clusters are formed from a chunk via a Semi-Supervised FCM clustering and when the next chunk becomes available the clustering is run again starting from cluster prototypes inherited from the previous chunk, and a new model is created, by adapting the previous one to the new data [7]. The Semi-supervised nature of DISSFCM enables the construction of classification models leveraging unlabeled samples together with a few labeled ones, thus overcoming the limitation of most existing data stream classification methods requiring the availability of completely labeled data. The DISSFCM method has been successfully applied in the classification of data streams coming from different sources [4].

In this paper we investigate the use of DISSFCM as a tool to analyze educational data and derive useful models to predict the students' behavior. In particular, we study the effectiveness of DISSFCM on the Open University Learning Analytics Dataset (OULAD) [19]. Preliminary experimental results show that DISSFCM can be an effective method to perform educational data stream mining.

IV. METHODOLOGY

According to [27], the analysis of educational data involves four main iterative steps (fig. 1):

1) Data collection: different kind of data, coming from the VLE interactions and contents, are collected and represented in a structured form, as a database.

2) Data pre-processing: depending on the data analysis goal, data are filtered, aggregated or manipulated, in order to obtain a significant dataset.

3) Data analysis: one or more machine learning methods are applied to the filtered data to derive a model that could be used, for example, to predict the students' performances.

4) Data interpretation: results coming from automatic methods need to be interpreted by domain experts to support their decisions [30].

In order to have significant results, the analysis should be conducted with a clear idea of the final actors: students, teachers and/or managers [22]. Students may be interested in discovering strengths and weaknesses in their learning method so as to remedy to any deficiencies. Teachers may be interested in grouping the students according with their study level, to better assess their teaching methodologies. Managers could be interested in statistics about abandonment rate and its causes. In this work we focus on educational data mining to predict students' outcomes, by analyzing general and behavioral factors that influence the results. This kind of analysis turns

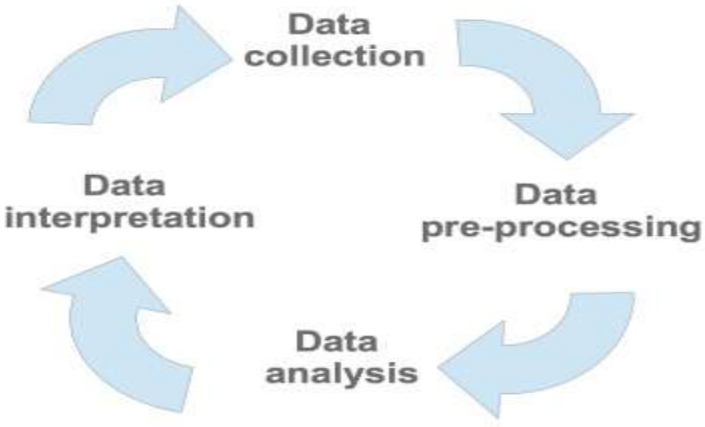

Fig. 1. Main steps of educational data mining process.

out to be helpful for managers and teachers, who can use the analysis feedback to limit the failure causes and to improve their courses. For example, if the interaction with a particular resource is crucial for the students' success, teachers will be strongly invited to use this resource in their courses. On the contrary, if a particular condition, such as coming from a risk area or having a low instruction level, is shown to be correlated to the students' failures, support activities should be undertaken to help the learning process of these categories of students.

In this work the data collection step has been skipped, since a freely available dataset has been used. Specifically, we have used the Open University Learning Analytics Dataset (OULAD)1 for the analysis. It provides data containing on-line courses information, such as students' general information, students' interactions with the VLE, courses' information.

A. Data pre-processing

A selection of the OULAD dataset including data coming from a single course (code 'DDDD') has been considered. It includes information about 100 students, including general information (such as gender, age, to be grown up in risk areas, etc..), behavioral information coming from the interaction with the VLE, and assessment results. Each student is described by the following attributes:

A1) Gender: the students' gender (M/F)

A2) Highest education: the highest UK student education level when enrolled (A Level or Equivalent, Lower Than A Level, HE Qualification or No Formal quals)

A3) Imd band: the Index of Multiple Deprivation2 band value of the place where the student lived when enrolled. It is an UK government measure to evaluate deprived areas in English local councils. It could assume percentage values in 0−100 representing the risk levels (0 for low, 100 for high).

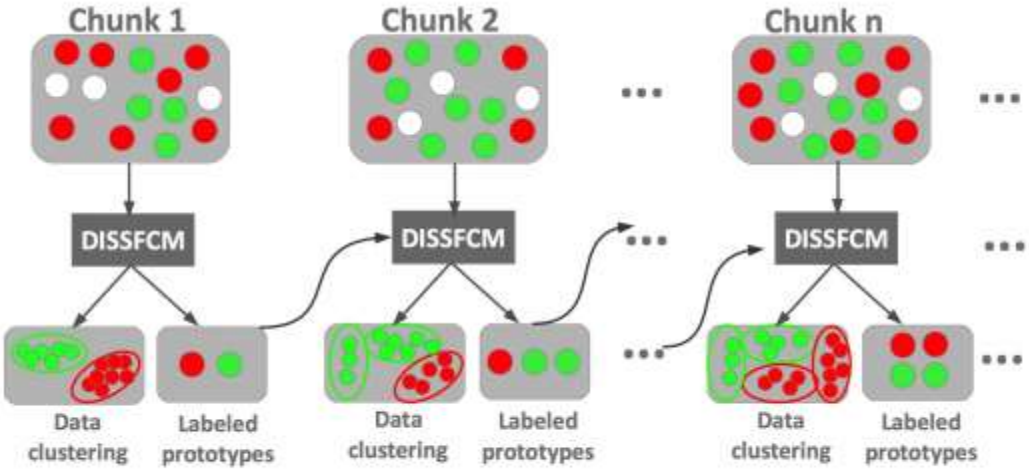

Fig. 2. Scheme of DISSFCM.

A4) Age band: Three students' age bands have been considered: [0-35], [35-55], greater than 55

A5) Number of previous attempts: the number of times the student has attempted the current module

A6) Studied credits: the total number of credits that the student is interacting with, simultaneously with the module

A7) Disability: student's disability conditions (Y/N)

A8) Number of assessments: number of the student's submitted assessments for the module

A9) Average assessments score: the average student's assessments score for the current module

Additional attributes describe the number of interactions (clicks/visualizations) that the student has performed on the VLE using ten different educational supports:

A10) quiz: questionnaires regarding the module contents

A11) forum: interactive platform to connect students together

A12) glossary: hyper-link dictionary that explains particular words in the module

A13) homepage: module homepage

A14) out collaboration: collaborations among students

A15) out content: extra platform material suggested by the professor

A16) resource: extra material given by the professor

A17) wiki: wiki pages, students and professor can interact with

A18) subpage: course subpages that focus on a particular topic

A19) url: external resources, liked by the professor.

The class attribute C indicates the Final result i.e. whether the student has passed or not the current module (fail/pass). Ordinal attributes have been converted into numerical values for subsequent processing through the DISSFCM algorithm.

B. Data analysis

To analyze the educational data and derive a classification model capable to predict the final result of a student, we used our data stream classification method based on an incremental semi-supervised fuzzy clustering algorithm called DISSFCM [5]. The method assumes that data belonging to different classes (fail/pass) are continuously available during time in form of chunks, i.e. sub-sets of data with their

own class distribution. The clusters (and their prototypes) are formed starting from a single chunk via the SSFCM (Semi-Supervised FCM) clustering [24] and when the next chunk becomes available the clustering is run again starting from cluster prototypes inherited from the previous chunk in order to preserve memory of the previously discovered relations.

In real-world contexts, such as in the case of educational data, the underlying distribution of data may change over the time. The DISSFCM algorithm can cope with these changes because it is able to evolve the classification model by dynamically adapting the produced clusters to the new incoming data. Specifically, the model adaptation is based on a splitting mechanism that is applied from time to time depending on the quality of current clusters. When the cluster quality deteriorates from one data chunk to another, the cluster with lowest quality is split in two clusters via a contextual fuzzy clustering where the context is determined by the cluster to be split [25]. The cluster quality is evaluated in terms of the reconstruction error [26], which measures the difference between the original data and their reconstruction operated by averaging the prototypes weighted by the membership degrees of data to clusters.

The algorithm requires the data as a sequence of chunks and an initial collection of labeled prototypes such that each class label is represented by at least one prototype. For each chunk, the SSFCM (Semi-Supervised FCM) algorithm is applied in order to group data into a number of clusters. Each cluster is represented by a prototype that is labeled automatically due to the semi-supervised nature of SSFCM. Therefore, after application of SSFCM on a chunk, each data sample is matched against all the labeled prototypes and assigned to the class of the best-matching prototype. The matching mechanism is based on the standard Euclidean distance. At the end, the algorithm returns the most recent collection of the prototypes, reflecting the data structure of the last data chunk. The algorithm is incremental since the collection of prototypes returned from one chunk can be used as starting point for a new run of the algorithm as long as new chunks are available from the data stream, as it is shown in fig. 2

C. Data interpretation

The output of DISSFCM is a collection of cluster prototypes that synthesize the data in a chunk. Specifically, each cluster prototype is a medoid that summarizes the attributes of all the student items belonging to that cluster. In this sense, each prototype can be interpreted as a typical student profile arising from data. The analysis of these profiles can provide more insight into common characteristics among student items. For example, teachers may be able to discover the causes of the students' failures by analyzing in detail each cluster prototype. Hence the resulting clusters can provide additional information useful to take proper actions oriented to help students avoiding failures.

III. Conclusions & Future Work

In this paper we have presented a case study of educational data mining process involving the application of DISSFCM, an incremental semi-supervised fuzzy clustering algorithm for data stream classification. The Open University Learning Analytics Dataset (OULAD) has been processed as a data stream via DISSFCM to extract a classification model capable to predict the students' outcomes. The DISSFCM algorithm has shown to be able to adapt and evolve the classification model to new incoming data. Preliminary numerical results have shown the effectiveness of the proposed method in correctly classifying students' outcomes by processing educational data as a stream.

Further work is in progress to carry out a deeper analysis of the clusters resulting from DISSFCM in order to investigate the main factors that influence the students' failures or successes. This kind of analysis may help all the stakeholders involved in the VLEs to better design courses on the basis of different student profiles. In addition, future work will be addressed to better investigate the effectiveness of DISSFCM when dealing with huge collections of more complex and heterogeneous educational data.

V. REFERENCE

[1] A. Abdullatif, F. Masulli, and S. Rovetta. Clustering of nonstationary data streams: A survey of fuzzy partitional methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1258, 2018.

[2] A. F. Agudo-Peregrina, ´ A. Hern ´ andez-Garc ´ ´ıa, and S. Iglesias-Pradas. Predicting academic performance with learning analytics in virtual learning environments: A comparative study of three interaction classifications. In 2012 International Symposium on Computers in Education (SIIE), pages 1–6. IEEE, 2012.

[3] T. Barnes. The q-matrix method: Mining student response data for knowledge. In American Association for Artificial Intelligence 2005 Educational Data Mining Workshop, pages 1–8, 2005.

[4] G. Casalino, G. Castellano, A. Fanelli, and C. Mencar. Enhancing the dissfcm algorithm for data stream classification. In M. F. Fuller R., ´ Giove S., editor, Fuzzy Logic and Applications. WILF 2018., volume 11291 of Lecture Notes in Computer Science. LNAI 10614., pages 109–122. Springer, Cham, Genova, Italy, September 6-7, 2018. 2019. DOI:10.1007/978-3-030-12544-8 9.

[5] G. Casalino, G. Castellano, and C. Mencar. Incremental adaptive semisupervised fuzzy clustering for data stream classification. In Proc. of the 2018 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS 2018), pages 1–7, Rhodes, Greece, 5 2018.

[6] G. Casalino, C. Castiello, N. Del Buono, F. Esposito, and C. Mencar. Q-matrix extraction from real response data using nonnegative matrix factorizations. In International Conference on Computational Science and Its Applications, pages 203–216. Springer, 2017.

[7] G. Castellano and A. Fanelli. Classification of data streams by incremental semi-supervised fuzzy clustering. In Int. Workshop on Fuzzy Logic and Applications, volume 10147 of Lecture Notes in Computer Science, pages 185–194, 2016.

[8] G. Castellano, A. Fanelli, and T. Roselli. Mining categories of learners by a competitive neural network. In IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222), volume 2, pages 945–950. IEEE, 2001.

[9] P. Donaldson, N. Ntarmos, and K. Portelli. A systematic review of the potential of machine learning and data science in primary and secondary education. 2017.

[10] I. Duru, G. Dogan, and B. Diri. An overview of studies about students' performance analysis and learning analytics in moocs. In 2016 IEEE International Conference on Big Data (Big Data), pages 1719–1723. IEEE, 2016.

[11] A. Dutt, M. A. Ismail, and T. Herawan. A systematic review on educational data mining. IEEE Access, 5:15991–16005, 2017.

[12] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala. Predicting student performance using personalized analytics. Computer, 49(4):61–69, 2016.

[13] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy. Mining Data Streams: A Review. SIGMOD Rec., 34(2):18–26, 6 2005.

[14] J. Gama. Knowledge Discovery from Data Streams. Chapman and Hall/CRC, 1st edition, 2010.

[15] A. F. D. Gonc¸alves, A. M. A. Maciel, and R. L. Rodrigues. Development of a data mining education framework for visualization of data in distance learning environments. In The 29th International Conference on Software Engineering and Knowledge Engineering, Wyndham Pittsburgh University Center, Pittsburgh, PA, USA, July 5-7, 2017., pages 547–550, 2017.

[16] K. Govindasamy and T. Velmurugan. A survey on the result based analysis of student performance using data mining techniques. International Journal of Data Mining Techniques and Applications, 5(1):91–93, 2016.

[17] H. He, S. Chen, K. Li, and X. Xu. Incremental Learning From Stream Data. IEEE Transactions on Neural Networks, 22:1901–1914, 2011.

[18] A. Hern ´ andez-Garc ´ ´ıa, I. Gonzalez-Gonz ´ alez, A. I. Jim ´ enez-Zarco, and ´ J. Chaparro-Pelaez. Visualizations of online course interactions for ´ social network learning analytics. International Journal of Emerging Technologies in Learning (iJET), 11(07):6–15, 2016.

[19] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. Scientific data, 4:170171, 2017.

[20] H. Nen-Fu, I. Hsu, L. Chia-An, C. Hsiang-Chun, T. Jian-Wei, F. TungTe, et al. The clustering analysis system based on students' motivation and learning behavior. In 2018 Learning With MOOCS (LWMOOCS), pages 117–119. IEEE, 2018.

[21] Y. Nieto, V. Garc´ıa-D´ıaz, C. Montenegro, and R. G. Crespo. Supporting academic decision making at higher educational institutions using machine learning-based algorithms. Soft Computing, pages 1–9, 2019.

[22] E. Osmanbegovic and M. Suljic. Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business, 10(1):3–12, 2012.

[23] R. Paiva, I. I. Bittencourt, W. Lemos, A. Vinicius, and D. Dermeval. Visualizing learning analytics and educational data mining outputs. In International Conference on Artificial Intelligence in Education, pages 251–256. Springer, 2018.