Evolving fuzzy clustering for data analysis in Virtual Learning Environments

Gabriella Casalino
gabriella.casalino@uniba.it
University of Bari Aldo Moro, Italy

Giovanna Castellano
giovanna.castellano@uniba.it
University of Bari Aldo Moro, Italy

Corrado Mencar
corrado.mencar@uniba.it
University of Bari Aldo Moro, Italy

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.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. 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) have shown the reliability of DISSFCM in creating good classification models of educational data.

Keywords

Educational Data Mining, Virtual Learning Environments, Data stream, Fuzzy clustering.

1 Introduction

The use of Virtual Learning Environments (VLEs) has exponentially increased, because, a great reduction in management costs is achieved, and 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.

Educational Data Mining (EDM) uses Machine Learning techniques to analyze educational data in order to extract students' behavior models useful to predict their future performances. This 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 and to improve student retention could be implemented by considering the suggestions coming from a data analysis process.

Several studies proved that machine learning techniques can be successfully used in the educational field [2–4].

2 Our contribution

However, 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.

In [1] we proposed an 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. 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.

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) [5]. Preliminary experimental results show that DISSFCM can be an effective method to perform educational data stream mining.

3 Conclusions

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. Future works will apply DISSFCM in real scenarios.

References

[1] G. Casalino, G. Castellano, and C. Mencar. Incremental adaptive semi-supervised 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.

[2] 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.

[3] 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.

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

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