Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Improved Cluster Analysis for Graduation Prediction using Ensemble Approach

Patcharaporn PANWONG [a], Natthakan IAM-ON [ab*] & James MULLANEY [c]
[a]*School of Information Technology, Mae Fah Luang University, Thailand*
[b]*Center of Excellence in AI & Emerging Technologies, Mae Fah Luang University Research & Innovation Institute, Thailand*
[c]*Department of Physics & Astronomy, University of Sheffield, UK*
*natthakan@mfu.ac.th

Abstract: Predicting student performance has been one of major subjects in the educational data mining, for which a bucket of analytical methods has been proposed. Among these, a recent framework of bi-level learning is recently introduced with improved classification performance from a basic supervised paradigm. However, only k-means is exploited to derive data clusters, which are employed as references for context-specific classification modeling. As such, this paper presents an original work that applies ensemble clustering to deliver more accurate data partition, thus lifting the predictive accuracy. Based on data collected from Mae Fah Luang University databases, the new approach are usually more effective, especially to the minority class that is the core of imbalance problem. Besides, a parameter analysis is briefly addressed herein to specify recommended settings for future exploitation.

Keywords: Guidelines, formatting instructions, author's kit, conference publications

1. Introduction

A recent series of emerging applications of artificial intelligence technology and automated learning modules has demonstrated the potential of knowledge discovery and exploitation at large. Of course, insights gained from analyzing operational data may help to foresee outputs and possible problems, for which a timely and effective measure can be developed. It is also the case for education domain at all levels of primary to higher education (Romero and Ventura, 2020; Basha et al, 2020). Different data sources have been a gold mine for educational data science projects, including students' enrollment and grading profiles, activity logs for the utilization of online resources and class participation (Baepler and Murdoch, 2010; Romero and Ventura, 2013). The tendency of this data-driven practice has been on the rise in the past decade (Erdogan and Timor, 2005), perhaps under the concept of Educational Data Mining or EDM (Bakerand and Yacef, 2009; Bala and Ojha, 2012). Examples of research work in this area are the predictive modeling of student academic performance, recommendation of courses and personalized curriculum planning, as well as identification of atypical learning patterns (Romero and Ventura, 2020; Iam-On and Boongoen, 2017a).

Specific to the subject of student performance or achievement, various studies investigate both new and conventional data analytical techniques to disclose students at risk and their major properties. Along this line of research, Lin (2012) develops a predictive model for categorizing new students to appropriate retention programs. With a similar goal of preventing undesirable incidents, hence damages to both students and universities, other studies have proposed different approaches to determine a set of student groups with unique preferences (Koedinger et al, 2008; Ramaswami and Bhaskaran, 2010). In addition to these, several projects explore many specific methods to modeling student performance and dropout. They employ supervised learning algorithms such as Naive Bayes classifier (Kotsiantis, 2004) and decision tree (Kabra and Bichkar, 2011), and unsupervised learning counterparts like k-means (Yu et al, 2010) that is highly efficient for big data analysis.

For higher education institutes in Thailand, the issue of student retention and personalized treatment have grasped an attention of executives as well as data science researchers. Specific to Mae Fah Luang University or MFU, published case studies (Iam-On and Boongoen, 2017a; Iam-On and

Boongoen, 2017b) explore existing techniques and their extensions to discover normal and dropout patterns. For the former, the methodology of ensemble clustering (Iam-On and Boongoen, 2015a; Iam-On and Boongoen, 2015b) has been reused to develop a data transformation procedure to generate informative features for the down-streaming classification process. This meta-learning approach is also exploited in the latter to obtain more accurate clustering results compared to those achieved with basic methods. Recently, Nanglae et al (2021) has put forward a novel framework for improved classification performance, namely a bi-level learning model. This combines cluster analysis and data classification within a unified learning method, with data clustering is conducted initially to derive different contexts for classification strategies.

Despite the success reported for the ability of bi-level framework to overcome the imbalance class problem, there is a gap for further development, especially in the phase of cluster analysis. Instead of using a simple k-means method to deliver a target collection of clusters, the current work presents an application of ensemble clustering such that those clusters are of higher quality, thus raising up the accuracy of this predictive framework. The rest of this paper is organized as follows. At first, details of investigated data, basic stages of developing a bi-level model and the modification of cluster analysis step are provided in Section 2. Then, Section 3 presents experimental design and results in which the proposed approach is compared with its baseline. The paper is concluded in Section 4 with possible future works.

2. Material and Method

2.1 Investigated Data

Based on the previous study of Nanglae et al (2021), this work makes use of academic records of those undergraduate students who graduated in 2016, i.e., 2559 in B.E. The data collection contains 1,162 samples from School of management and School of Information Technology. Each of these instances corresponds to a student that completes a number of required courses for three subject categories of general education, specific required and free elective courses, respectively. Note that students who are categorized as program transfer or exchange are excluded from the present examination. Table 1 shows description of different features extracted from the operational database. Provided these, the following data pre-processing steps are implemented prior the phase of model development.

(i) Each grade frequency, e.g., A1, A2 and A3 in Table 1, is normalized to the standard range of [0, 1]. This helps to overcome the problem that different programs may consist of different number of courses per category. Formally, the normalization of a grade frequency f_{xi} in the category x is defined as f_{xi*}. This can be estimated by $f_{xi*} = f_{xi} / f_X$, where f_X is the summation across all courses in x.

(ii) Then, the attribute YEAR representing the entry year in B.E., is transformed to the number of years each student has spent in the program before graduation. Note that a student that graduates in year y started the program in year $y - 3$ or before that.

After these steps of data preparation, the final data set is composed of N = 1,162 samples (911 of these belonging to School of Management, and the other 251 cases representing the other), and D = 40 features or attributes. These can be summarized as follows.

- 13 normalized grade frequencies for specific required courses; d1, ..., d13 in [0, 1].
- 13 normalized grade frequencies for free elective courses; d14, ..., d26 in [0, 1].
- 13 normalized grade frequencies for general education courses; d27, ..., d39 in [0, 1].
- YEAR that is now the number of years before graduation; d40 in {4, 5, 6, 7}. It is noteworthy that the minimum number of years anyone at MFU has to be in a program is 4 years. Also, it is possible for a student to spend up to 7 years in a specific program before graduation.

2.2 Cluster Analysis in Bi-Level Learning Model

This section provides details of the proposed application of ensemble clustering to the framework of bi-level learning. The steps taken to generate or train a model are given as follows.

Step 1: For a given specific case q (e.g., school), suppose that $Xq,train$ and $Xq,test$ are training and test data, respectively. Note that the process of model generation exploits only the former, while

the latter is used to evaluate the resulting model. To start with, the optimal number of clusters or *k* for training data is determined using the assessment method introduced by Nanglae et al (2021). This is based on the selection of a partition among multiple clusterings with different *k* values, each of which is generated using a classical k-means technique. The best quality is specified by internal validity indices of DB and Dunn.

Table 1. *Description of Student Information and Different Grading Frequencies.*

Attribute Name	Description
YEAR	Year of entry (in B.E.)
A1	Number of grade A obtained from specific required courses
BB1	Number of grade B+ obtained from specific required courses
B1	Number of grade B obtained from specific required courses
CC1	Number of grade C+ obtained from specific required courses
C1	Number of grade C obtained from specific required courses
DD1	Number of grade D+ obtained from specific required courses
D1	Number of grade D obtained from specific required courses
F1	Number of grade F obtained from specific required courses
P1	Number of grade P obtained from specific required courses
S1	Number of grade S obtained from specific required courses
U1	Number of grade U obtained from specific required courses
V1	Number of grade V obtained from specific required courses
W1	Number of grade W obtained from specific required courses
A2	Number of grade A obtained from free elective courses
BB2	Number of grade B+ obtained from free elective courses
B2	Number of grade B obtained from free elective courses
CC2	Number of grade C+ obtained from free elective courses
C2	Number of grade C obtained from free elective courses
DD2	Number of grade D+ obtained from free elective courses
D2	Number of grade D obtained from free elective courses
F2	Number of grade F obtained from free elective courses
P2	Number of grade P obtained from free elective courses
S2	Number of grade S obtained from free elective courses
U2	Number of grade U obtained from free elective courses
V2	Number of grade V obtained from free elective courses
W2	Number of grade W obtained from free elective courses
A3	Number of grade A obtained from general education courses
BB3	Number of grade B+ obtained from general education courses
B3	Number of grade B obtained from general education courses
CC3	Number of grade C+ obtained from general education courses
C3	Number of grade C obtained from general education courses
DD3	Number of grade D+ obtained from general education courses
D3	Number of grade D obtained from general education courses
F3	Number of grade F obtained from general education courses
P3	Number of grade P obtained from general education courses
S3	Number of grade S obtained from general education courses
U3	Number of grade U obtained from general education courses
V3	Number of grade V obtained from general education courses
W3	Number of grade W obtained from general education courses

Step 2: Having obtained the desired number of clusters, the target data partition is created using a clustering algorithm Φ, which is k-means again in the original research. Note that for this stage, the YEAR feature is excluded such that clusters of students are formulated based solely on information of grade achievement. As such, this problem is designed as a binary classification, with two classes of A

(YEAR = 4) and B (YEAR > 4). In spite of a good classification accuracy, performance can be unstable as k-means is a weak method, with random initialization of cluster centers often leading to sup-optimal results. Henceforth, a more reliable alternative of ensemble clustering is then used in place of a single k-means. Specific to this research, a pairwise similarity or CO matrix is built as a summarization of *M* base clustering results, each of which is an output of applying k-means to training data. In addition, the strategy of random-k is included in this creation procedure as to boost diversity within an ensemble. Lastly, the final partition is achieved by applying k-means with the preferred *k* value to the CO matrix. See Panwong et al (2020) for further details of concept and terminology.

Step 3: For each cluster c_t *i*n the clustering *C* from Step 2 (where $t = 1 \ldots k$), its centroids z_t is employed as a reference for a new sample in the test or prediction phase.

Step 4: Next, for each of final clusters, find the percentage of majority class among samples belonging to that cluster. The process terminates only at this level of cluster analysis, if that percentage is greater than or equal to α (i.e., a predefined value of minimum percentage for a pure cluster). As a result, this cluster represents that majority class, which is a prediction of a new instance that is similar to the corresponding cluster center. Otherwise, a classifier is to be built using samples of this specific cluster (see Step 5).

Step 5: When one cluster is not pure up to the expected level of α, samples in that cluster will be exploited to train a classifier using a classification algorithm β.

3. Results

This empirical study is based on the data set of 1,162 samples, which is described earlier in Section 2.1. Other major settings are summarized here. At first, the ensemble clustering is employed here with M = 10, where the minimum level of cluster purity or the variable α is configured as 90%. Based on the results reported in the original work, the two best algorithms are examined as the choice to create the classifier β in Level 2 of the proposed model. These include Naive Bayes or NB (using Gaussian distribution for numerical features), and Random Forest or RF (with the size of forest = 50). As for an assessment framework, 10-fold cross validation is chosen to allow each sample to be a member of test data once. After that, a confusion matrix is produced for this binary classification problem. With the goal of this work as to initially explore the potential of ensemble clustering for bi-level learning modeling, a compared method is the original approach where k-means is used to deliver a reference data partition. Note that baseline and proposed methods will be specified as Original and Proposed hereafter.

Table 2. *Confusion Matrices and Corresponding Accuracies (School of Management).*

Model	Confusion matrix			Class-specific accuracy	Overall accuracy
	A	B	Ground truth		
Original (NB)	792	50	A	94.06%	92.97%
	14	55	B	79.71%	
Original (RF)	827	15	A	98.22%	93.96%
	40	29	B	42.03%	
Proposed (NB)	799	43	A	94.89%	94.07%
	11	58	B	84.06%	
Proposed (RF)	833	9	A	98.93%	95.72%
	30	39	B	56.52%	

Table 2 presents the results with the first collection specific to School of management, with 911 samples (842 of these belong to Class A and 69 to Class B). With this first data collection, the optimal number of clusters $k = 2$ provides the highest measures across internal validity indices. It can be obviously observed that the proposed approach is able to improve predictive performance in both cases of NB and RF classifiers. To be precise, the overall accuracy of NB is lifted from 92.97% to 94.07%, while the similar improvement is made from 93.96% to 95.72% for the other. Given these, RF appears to be a more effective choice than NB, however with higher complexity. Another point worth mentioning here is that the ensemble clustering approach can resolve the problem of imbalance class

significantly better than the original counterpart. This can be similarly witnessed for both classification algorithms examined in this work.

Table 3. *Confusion Matrices and Corresponding Accuracies (School of Information Technology).*

Model	Confusion matrix			Class-specific accuracy	Overall accuracy
	A	B	Ground truth		
Original (NB)	199	13	A	93.87%	93.63%
	3	36	B	92.31%	
Original (RF)	206	6	A	97.17%	92.43%
	13	26	B	66.67%	
Proposed (NB)	201	11	A	94.81%	94.42%
	3	36	B	92.31%	
Proposed (RF)	208	4	A	98.11%	94.82%
	9	30	B	76.92%	

Similarly, Table 3 summarizes predictive performance for 251 cases (212 of these belong to Class A and 39 to Class B) belonging to School of Information Technology. With the optimal number of cluster k being 2 as well, RF still delivers the highest overall accuracy of 94.82% that is improved from the original score of 92.43%. This is pretty much due to a significant increase of the accuracy level specific to the minority class, i.e., Class B. This is initially at 66.67% with Original(RF) and 76.92% with Proposed(RF). Given these findings that imply an ability to partly overcome the imbalance class issue, it is interesting to investigate the association between a hyper parameter of the proposed approach such as ensemble size (M) and classification performance seen with the RF classifier. To this end, Figure 1 illustrates corresponding observations from the case of School of Management, where M is within the range of {10, 20, 30, 40, 50}. It is seen that the accuracy specific to Class B tends to incline with M increasing from 10 to 20 and 30, i.e., from 56.52% to 59.42% and 65.21%, respectively. However, bigger ensembles may not further yield the underlying predictive capability, thus $M = 30$ is generally recommended for future applications. It is noteworthy that a similar finding is also obtained with the case of School of Information Technology, but not included here due to space limitation.

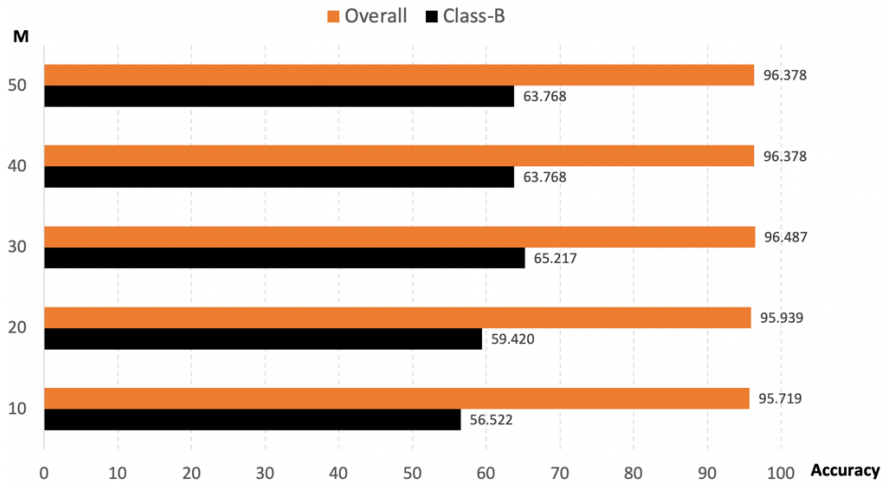

Figure 1. Accuracies obtained with different ensemble sizes (School of Management).

4. Conclusion

This paper presents an original investigation that applies ensemble clustering to the bi-level learning framework that has been recently introduced in the literature. The proposed approach helps to overcome the problem of imbalance class observed in data collected from MFU operational databases. Besides, it delivers a more stable result as compared to the use of a single clustering technique seen in the original work. Despite this positive finding, a few issues are worth further studies. Firstly, an oversampling or

undersampling technique may well be explored to resolve the problem of class imbalance (Tabacolde et al, 2018). In addition, the methodology of classifier ensemble may be useful to combine predictions made by different classifiers. Lastly, other ensemble clustering models and recent variants (Panwong et al, 2018; Pattanodom et al, 2016) can be another subject area for possible improvement.

Acknowledgements

This work is partly supported by Mae Fah Luang University (MFU) and STFC-GCRF2018: From Stars to Baht (collaboration between MFU & University of Sheffield).

References

Baepler, P. & Murdoch, C. (2010). Academic analytics and data mining in higher education. *International Journal of Scholarship of Teaching and Learning*, 4(2), 1–9.

Bakerand, R. & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.

Bala, M. & Ojha, D. (2012). Study of applications of data mining techniques in education. International Journal of Research in Science and Technology, 1, 1–10.

Basha, S., Prasad, G., Rao, M. & Vardhan, M. (2021). A Review of Predictive and Descriptive Techniques in Higher Education Domain. *International Journal of Computer Engineering and Applications*, XIII(VI), 1-7.

Erdogan, S. & Timor, M. (2005). A data mining application in a student database. *Journal of Aeronautic and Space Technologies*, 2(2), 53–57.

Iam-On, N. & Boongoen, T. (2015). Comparative study of matrix refinement approaches for ensemble clustering. *Machine Learning*, 98(1-2), 269–300.

Iam-On, N. & Boongoen, T. (2015). Diversity-driven generation of link-based cluster ensemble and application to data classification. *Expert Systems with Applications*, 42(21), 8259–8273.

Iam-On, N. & Boongoen, T. (2017). Improved student dropout prediction in Thai university using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, 8(2), 497–510.

Iam-On, N. & Boongoen, T. (2017). Generating descriptive model for student dropout: a review of clustering approach. *Human-centric Computing and Information Sciences*, 7(1).

Kabra, R. & Bichkar, R. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications,* 36(11), 8–12.

Koedinger, K., Cunningham, K., Skogsholm, A. & Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. *Proceedings of International Conference on EDM*, 157–166.

Kotsiantis, S., Pierrakeas, C. & Pintelas, P. (2004). Prediction of students performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5), 411–426.

Lin, S. (2012). Data mining for student retention management. *Journal of Computing Sciences in Colleges*, 27(4), 92–99.

Nanglae, L. et al. (2021). Determining patterns of student graduation using a bi-level learning framework. Bulletin of Electrical Engineering and Informatics, 10(4), 2201-2211.

Panwong, P., Boongoen, T. & Iam-On, N. (2018). Improving Consensus Clustering with Noise-Induced Ensemble Generation: A Study of Uniform Random Noise. *Proceedings of International Conference on Machine Learning and Computing*, 390-395.

Pattanodom, M., Iam-On, N. & Boongoen, T. (2016). Clustering Data with the Presence of Missing Values by Ensemble Approach. *Proceedings of Asian Conference on Defence Technology*, 114-119.

Ramaswami, M. & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *International Journal of Computer Science*, 7(1), 10–18.

Romero, C. & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27.

Romero, C. & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1355.

Tabacolde, A., et al. (2018). Transient Detection Modeling as Imbalance Data Classification. *Proceedings of IEEE International Conference on Knowledge Innovation and Invention*, 180-183.

Yu, C., Gangi, S., Jannasch-Pennell, A. & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8, 307– 325.