Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 4, August 2021, pp. 2201~2211
ISSN: 2302-9285, DOI: 10.11591/eei.v10i4.2502

Determining patterns of student graduation using a bi-level learning framework

Lalida Nanglae[1], Natthakan Iam-On[2], Tossapon Boongoen[3], Komkrit Kaewchay[4], James Mullaney[5]
[1]School of Science, Mae Fah Luang University, Chiang Rai, Thailand
[2,3]Center of Excellence in Artificial Intelligence and Emerging Technologies, School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand
[4]Department of Aeronautical Engineering, Navaminda Kasatriyadhiraj Royal Air Force Academy, Thailand
[5]Department of Physics and Astronomy, University of Sheffield, Sheffield, United Kingdom

Article Info

Article history:

Received Apr 8, 2020
Revised May 20, 2021
Accepted Jun 15, 2021

Keywords:

Classification
Clustering
Data science
Higher education
Student performance

ABSTRACT

The practice of data science, artificial intelligence (AI) in general, has expanded greatly in terms of both theoretical and application domains. Many existing and new problems have been tackled using different reasoning and learning methods. These include the research subject, generally referred to as education data mining (or EDM). Among many issues that have been studied in this EMD community, student performance and achievement provide an interesting, yet useful result to shaping effective learning style and academic consultation. Specific to this work at Mae Fah Luang University, the pattern of students' graduation is determined based on their profile of performance in different categories of courses. This course-group approach is picked up to generalize the framework for various undergraduation programmes. In that, a bi-level learning method is proposed in order to predict the length of study before graduation. At the first tier, clustering is applied to derive major types of performance profiles, for which classification models can be developed to refine the prediction further. With the experiments on a real data collection, this framework usually provides accurate predictive outcomes, using several conventional classification techniques.

Corresponding Author:

Natthakan Iam-On
Center of Excellence in AI and Emerging Technologies
School of Information Technology
Mae Fah Luang University, 333 Moo.1, Ta-sud, Muang District, Chiang Rai 57100, Thailand
Email: natthakan@mfu.ac.th

1. INTRODUCTION

Catching up the changing world, regarding technology advancement and life style, almost all organizations have embraced tools and techniques to derive useful knowledge from a pool of transactional data. This also applies to the context of higher education, where conventional and new sources of such a data have an important role to play [1], [2]. These include a simple student grading profile that is normally obtained from a university registration system, history of course enrollment, and student logs with online learning sessions [3], [4]. To a university and alike education institutes, this has proven critical to maintain competitive and meet expectations of young generation and the government. In particular to the study of [5], the trend of applying data mining that is recently renewed to a general concept of data science, to various educational data and problems keeps increasing over the years. With this methodology of educational data mining (EDM) [6], [7], effective planning and decision making can well be improved by transferring a goldmine of data specific to each university to working knowledge about student behavior, preferences of

learning methods and materials, communication channels and other factors to their achievement. Examples of past development include the prediction of student performance, recommendation systems for courses or a personalized learning plan, determination of atypical learning patterns and causes [1], [8].

Drilling down to the topic of student performance or achievement, a number of previous studies exploit newly customized and existing data mining models to commonly demonstrate the benefits of identifying students at risks. Given this, a university may be able to act quickly or even prevent undesirable events to take place, hence reducing the damage to both student and university. The work of [9] focuses on inventing a predictive model that accurately categorize new students to different programmes of student retention on campus. In addition, others [10]-[12] also propose models that determine groups of students with distinct preferences. Such a division leads to appropriate policy and treatment being implemented to ensure student retention. Similar to these, there are other investigations that make use of a range of data mining methods to modeling student performance and dropout. These include supervised learning models like Naive Bayes classifier [13] and decision tree [14], [15], with an unsupervised learning approach like k-means [16] being an efficient alternative for a big set of data.

For Mae Fah Luang University (MFU) and other universities in Thailand, the problem of student retention has gained a great deal of attention. It is due to the country moves closer to the aging society, with the ratio between young and old population groups is geting smaller and smaller, hence less students will pursue higher education. This is also motivated by initial attempts [17]-[21] that make use of basic classification algorithms, and another set of studies by [8], [22] that explores both existing methods and their extensions. According to [8], a new data transformation is introduced prior the usual classification process. For that, the concept of consensus clustering [23]-[25] is adopted to transform an original data to the corresponding matrix with sample-cluster-relation embedding. Instead of modeling student performance solely as a classification problem, it might be feasible to include an unsupervised model like data clustering to determine the obvious cases, before forwarding the rest to a more complex classifier. Of course, this makes the training procedure more efficient with less samples. Besides, it might help to solve another difficulty of class imbalance, which is rather common as the amount of at-risk students is often much smaller than that of the other group. As such, this paper introduces a bi-level learning framework that first relates a new case to one of the pre-defined clusters. Then, for a particular cluster that sees almost all of its members belonging to one class, a pattern of student graduation can be justified right away. On the other hand, for a cluster with low purity, the prediction is produced by the cluster-specific classifier.

The proposed framework is exploited to determine the graduation patterns, or whether a student finishes the enrolled programme within a regular period of 4 years or else. This knowledge provides an opportunity for students together with advisors to adjust the plan of courses, which may help the student to perform better or graduate on time. This model is designed in such a way that it is applicable for different programmes across schools at MFU. To be precise, courses are groups to categories that are common to all students, thus generalizing the target learning model. For the current research, the framework is evaluated with a real data collection, which covers students graduating in 2016. The rest of this paper is organized as follows. Section 2 presents the research methodology of this study, including details of the data mining process, investigate data collection, and the proposed framework of bi-level learning. After that, experiment design, the corresponding results and discussion are provided in section 3. The paper is then concluded in section 4 with a perspective of future research.

2. METHOD

This research follows those data mining or data science studies, especially those focusing on EDM [8], [9], [20]. In particular, the target data is firstly identified, followed by the preparation stage that ensures the readiness and quality of final data set. Having completed this, the bi-level learning framework can be described, with respect to characteristics of the data under investigation. These issues are discussed in the following sections.

2.1. Data acquisition and preparation

In order to obtain an effective framework, it is designed based on transactional data maintained in the MFU registration system. Due to the concern of data privacy, the current project is to initially exploit only academic records of those undergraduate students who graduated in 2016 (or 2559 in B.E.). This population consists of 1,162 cases from 2 schools of management and information technology. The retrieval of these is subjected to conditions that a selected sample has to complete the number of required courses for three subject categories. These include general education course, specific required course, and free elective course, respectively. Moreover, those belonging to students with a record of programme transfer or exchange are excluded.

Within the registration database, two important tables from which the target data is retrieved are shown in Figure 1: 'Student personal information' and 'Student enrollment information'. In the former, each student is represented with personal identification number (ID), year of entry that specified in B.E., name of school that administrates the enrolled programme, and graduation GPAX. The latter describes a number of enrolled courses, course categories and the grades achieved. Given these, the target data can be obtained by joining the aforementioned two tables by student IDs. Following that, the 'Student data for analysis' table in Figure 1 can be generated by collapsing multiple rows of a single student (each representing one course) to one record. For such a purpose, course names are ignored, whilst frequencies of different grades (i.e., A, B+, B, C+, C, D+, D, F, P, S, U, V, and W) are accumulated. Note that three sets of grade frequencies are formed, one for each course category. Table 1 represents details of these sets of grade frequencies that are considered attributes or features of the intermediate data.

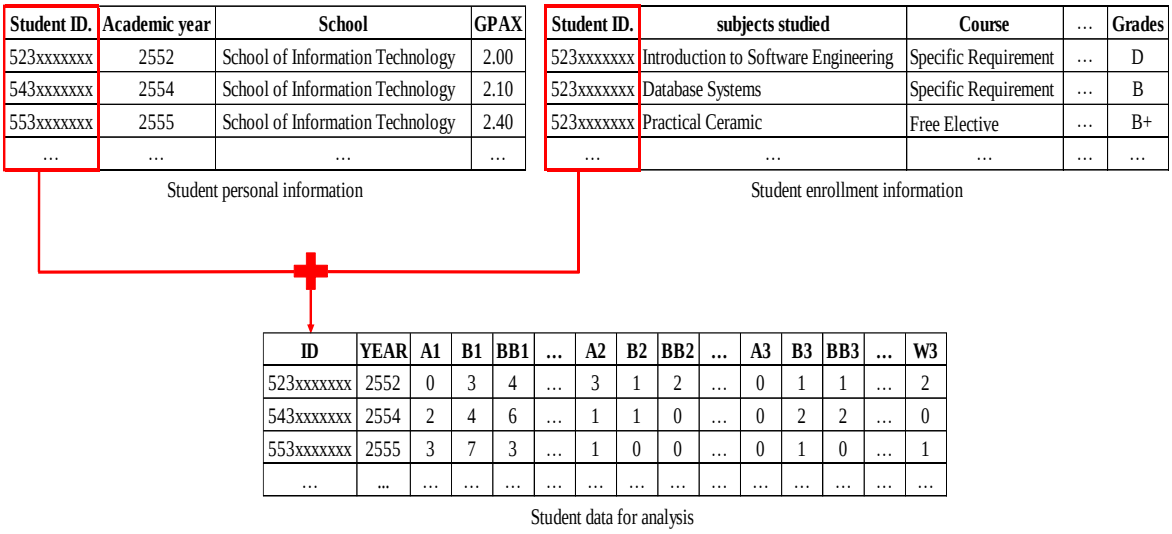

Figure 1. Initial target data ('Student personal information' and 'Student enrollment information' tables) and the initial data preparation procedure that produces the final data set (i.e., 'Student data for analysis' table)

Having obtained this intermediate data set, the following pre-processing steps are needed to create the final data set, which will be analyzed using the proposed framework.

(i) Each grade frequency such as A1, A2 and A3 in Table 1 is normalized such that its value domain is transformed to be within the range of [0, 1]. This is to ensure the absence of biases among different attributes in the analyzing process (i.e., these data attributes are equally important). Furthermore, it helps to overcome the problem that different programmes may consist of different number of courses in those three categories. As a result, the normalization of each grade frequency f_{xi} in the category x is defined as f_{xi^*}, which can be estimated by the following.

$$f_{xi^*} = \frac{f_{xi}}{\sum_{\forall j \in x} f_{xj}} \quad (1)$$

(ii) Then, the attribute ID is removed in order to protect the privacy of personal information.

(iii) At last, the attribute YEAR that represents the entry year in B.E., is transformed to a number of year each student has spent in the programme before graduation. Note that those students that graduate in year y actually started the programme in year y - *3* or before that. Given this knowledge, the new value y_{k^*} of YEAR attribute for a student k (where k=*1*, ..., N; N=1,162) can be calculated from the entry year y_k as follows, where y_{grad} is the year of graduation and set to 2559 in B.E. (or 2016 as mentioned earlier) for the present study. Note that other data batches may be available for future studies, where can be y_{grad} can be 2559, 2560, 2561 and so on.

$$y_{k^*} = y_{grad} - y_k + 1 \quad (2)$$

After these steps of data preparation, the final data set is composed of N=1,162 samples, and D=40 features or attributes. These can be summarized as follows. In that, 911 samples belong to School of management, and the other 251 cases represent students from School of information technology.

- 13 normalize grade frequencies in the category of specific required courses; $d_1, ..., d_{13}$ *in [0, 1].*
- 13 normalize grade frequencies in the category of free elective courses; $d_{14}, ..., d_{26}$ *in [0, 1].*
- 13 normalize grade frequencies in the category of general education courses; $d_{27}, ..., d_{39}$ *in [0, 1].*
- YEAR that is now the number of years before graduation; d_{40} *in {4, 5, 6, 7}*. It is noteworthy that the minimum numebr of years anyone at MFU has to be in a programme is 4 years. Also, it is possible for a student to spend up to 7 years in a specific programme before graduation.

Table 1. Description of student information and different grading frequencies (i.e., the intermediate data)

No	Attribute Name	Description
1	ID	Student identification number
2	YEAR	Year of entry (in B.E.)
3	A1	Number of grade A obtained from specific required courses
4	BB1	Number of grade B+ obtained from specific required courses
5	B1	Number of grade B obtained from specific required courses
6	CC1	Number of grade C+ obtained from specific required courses
7	C1	Number of grade C obtained from specific required courses
8	DD1	Number of grade D+ obtained from specific required courses
9	D1	Number of grade D obtained from specific required courses
10	F1	Number of grade F obtained from specific required courses
11	P1	Number of grade P obtained from specific required courses
12	S1	Number of grade S obtained from specific required courses
13	U1	Number of grade U obtained from specific required courses
14	V1	Number of grade V obtained from specific required courses
15	W1	Number of grade W obtained from specific required courses
16	A2	Number of grade A obtained from free elective courses
17	BB2	Number of grade B+ obtained from free elective courses
18	B2	Number of grade B obtained from free elective courses
19	CC2	Number of grade C+ obtained from free elective courses
20	C2	Number of grade C obtained from free elective courses
21	DD2	Number of grade D+ obtained from free elective courses
22	D2	Number of grade D obtained from free elective courses
23	F2	Number of grade F obtained from free elective courses
24	P2	Number of grade P obtained from free elective courses
25	S2	Number of grade S obtained from free elective courses
26	U2	Number of grade U obtained from free elective courses
27	V2	Number of grade V obtained from free elective courses
28	W2	Number of grade W obtained from free elective courses
29	A3	Number of grade A obtained from general education courses
30	BB3	Number of grade B+ obtained from general education courses
31	B3	Number of grade B obtained from general education courses
32	CC3	Number of grade C+ obtained from general education courses
33	C3	Number of grade C obtained from general education courses
34	DD3	Number of grade D+ obtained from general education courses
35	D3	Number of grade D obtained from general education courses
36	F3	Number of grade F obtained from general education courses
37	P3	Number of grade P obtained from general education courses
38	S3	Number of grade S obtained from general education courses
39	U3	Number of grade U obtained from general education courses
40	V3	Number of grade V obtained from general education courses
41	W3	Number of grade W obtained from general education courses

2.2. Model development

This section presents the process of model development, including cluster analysis that is conducted initially to observe the grouping structure within the final data set, and details of the proposed bi-level model with its evaluation being reported in section 3.

2.2.1. Initial cluster analysis

At first, it is trivial to observe the structure of data whether it is appropriate to develop the desired bi-level learning framework. In other words, after applying a clustering algorithm to the data set, there should be a cluster that is pure or almost pure (i.e., almost all samples in a cluster belong to the same class). Besides, there also are other clusters of the same clustering result that are nor pure, and needed additional classifiers to justify an appropriate class of their members. The final data set X is further divided into two subsets of school

specific samples: X_1 for School of management and X_2 for School of information technology. To accomplish this, k-means clustering algorithm is applied to the final data set X_q for M times, for a particular number of clusters k. These multiple trials are required to draw a reliable conclusion from a non-deterministic model like k-means. For each run $p=1 \ldots M$, the result C^k_p is assessed with two well-known validity indices of DB and Dunn (see [5] and [6] for more details). For each C^k_p, there will be two measurements of DB^k_p and $Dunn^k_p$. Then, the averages across M runs can be estimated and presented as DB^k_{p*} and $Dunn^k_{p*}$, respectively.

$$DB^k_{p^*} = \frac{\sum_{p=1,\ldots,M} DB^k_p}{M} \quad (3)$$

$$Dunn^k_{p^*} = \frac{\sum_{p=1,\ldots,M} Dunn^k_p}{M} \quad (4)$$

The aforementioned procedure is repeated for a range of different k values, i.e., k *in {2, 3, ..., k_{max}}*. As such, the optimal k is selected from this range as the value that provides the best values of DB^k_{p*} and $Dunn^k_{p*}$. To accomplish this, a rank-based approach is exploited such that the parameter k with the minimum overall ranking score (RS^k) is preferred. As a low DB measure indicates a good clustering, $DB^k_{p^*}$ for diffent k values are ranked from minimum to maxmum. Given this ranked list, the k-specific ranking score RS^k_{DB} can be determined, where the first in this list is assigned with 1 and the last with k_{max}-1. In case of a tie, the average of ranking score is given to related parties. Likewise, the k-specific ranking score RS^k_{Dunn} can also be estimated from the ranked list, in which high $Dunn^k_{p^*}$ measures appear in the front as they represent better clustering than those with lower Dunn values. Provided these, the overall ranking score specific to k can be simply calculated as follows. After that, the optimal k value is identified with the minimum $RS^k, k \in \{2, \ldots, k_{max}\}$.

$$RS^k = RS^k_{DB} + RS^k_{Dunn} \quad (5)$$

With k_{max} being 10, clustering results with two clusters (or $k=2$) proves to be better than those using other k values. Figures 2 and 3, for School of management and School of information technology, illustrate the two clusters that are obtained from the trial with the best quality measures. According to Figure 2, Cluster 1 is almost pure with 444 out of 447 samples (i.e., 99%) having the entry year of 2556 (in B.E.) or YEAR is 4, while only 1% spends 5 years before graduation. However, with Cluster 0, it is less pure with the majority of 85% finishes on time, or YEAR=4. The other 15% is a mixture between samples with YEAR values of 5 (13%), 6 (1%), and 7 (1%). Similar observations of the two clusters are also obtained with samples of School of information technology, see Figure 3 for more details. Henceforth, a clustering process may well be used to provide an accurate prediction model for specific clusters, such as those Cluster 1 in both cases. Nonetheless, a classifier is also required in addition to the initial clustering for some other clusters, for instance Cluster 0 in Figures 2 and 3. This finding leads to the proposed framework that will be explained next.

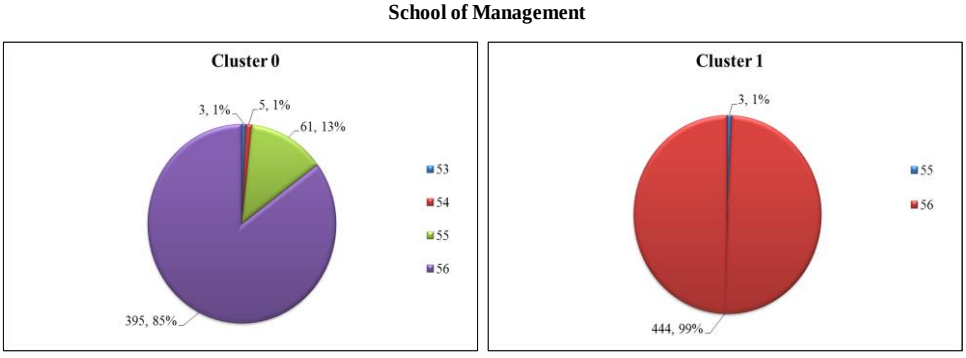

Figure 2. The best clustering result with k=2, for samples belonging to school of management

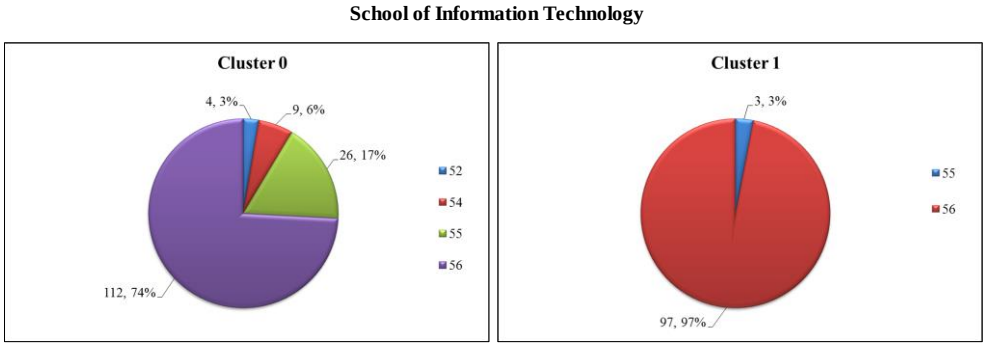

Figure 3. The best clustering result with k=2, for samples belonging to school of information technology

2.2.2. Proposed model

This section provides details of the proposed framework of bi-level learning, in which both types of unsupervised and supervised learning approaches are systematically combined to produce an accurate, yet efficient learning and prediction processes. The steps taken to generate or train a model are given as:

Step 1: For a given specific case q (e.g., school), suppose that $X_{q,train}$ and $X_{q,test}$ are training and test data, respectively. The process of model generation will make use of only the former, while the latter is used to assess the quality of the resulting model. With a clustering Φ, the procedure explained in section 2.2.1 is conducted on $X_{q,train}$ to find the optimal number of clusters. Then, select among M alternative of clustering results with that best k, to represent the knowledge model in the first level. Note that for this stage, the YEAR feature is left out such that groups of students can be formulated based solely on grade achievement. This problem is designed as a binary classification, with two classes of A (YEAR=4) and B (YEAR > 4).

Step 2: For each cluster c^k_t in the clustering C^k from Step 1 (where $t=1 \ldots k$), its centroids z^k_t is used as a reference for a new sample in the test or prediction phase. Please refer to [20] for details of estimating a centroid from cluster members.

Step 3: Again, for each cluster, find the percentage of majority class among samples in that cluster. The analysis process stops only at this clustering level, 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 centroid. Otherwise, a classifier is to be built using samples of this specific cluster (see Step 4).

Step 4: When one cluster is not pure up to the expected level of α, samples in that cluster will be used to train a classifier using the classification algorithm β. Please note that a conventional feature-based classification like a Naïve Bayes model can be used here. Please see section 3.1 for all methods that are employed in the present investigation.

After going through those steps explained above, the resulting bi-level model can be exploited to predict a class of a new instance in $X_{q,test}$ as follows.

Level 1: For a sample g in the test data $X_{q,test}$, find the optimal centroid z^k_t amongst k alternatives that provides the minimum distance to the sample g. This is defined by the following equation. Note that $d(.)$ is a distance function, with Euclidean being used in the current research.

$$\min_{z^k_t, t=1\ldots k} d(g, z^k_t) \tag{6}$$

If the optimal centroid z^k_t represents a cluster with the final class prediction (i.e., without additional classifier), the predicted class is simply provided. Otherwise, classify the sample g using the cluster-specific classifier in Level 2.

Level 2: Given the sample g, produce a class prediction using the classifier specifically developed for the cluster c^k_t (whose centroid is z^k_t that is identified earlier in Level 1).

3. RESULTS AND DISCUSSION

In this section, the design of empirical study is explained, which includes the investigated data and evaluation approach, settings of algorithm parameters, and compared methods. Furthermore, results and important findings are discussed in such a way to amend useful information and guideline.

3.1. Experimental design

This experiment makes use of the final data set of 1,162 samples, which is described in section 2.1. Two cases are formed regarding two schools where these samples belong to: i) School of management with 911 samples, and School of information technology with the other 251. Other settings are listed as:

a. k-means is used as the clustering algorithm Φ in bi-level learning framework, with $M=10$ for the number of trials to be investigated for a particular number of cluster k. Also, note that k is selected from a range of 2 to k_{max}, where k_{max}=10.
b. The minimum level of cluster purity is determined by the proportion of majority class, which is specified by the variable α=90%.
c. Four algorithms are examined as the choice to create the classifier β in Level 2 of the proposed model. These include: Naive Bayes (using Gaussian distribution for numerical features), K-nearest neighbors or KNN (using $K \in \{1, 3\}$ as to generalize the findings), Decision Tree (with the maximum depth=10), and Random Forest (with the size of forest=20).
d. 10-fold cross validation is exploited as the evaluation approach here, such that each sample is a member of test data once. As such, a confusion matrix is produced for this binary classification problem.
e. In addition, there are two compared methods that are considered as baseline counterparts of the bi-level learning framework.
f. Clustering-only prediction, i.e., only Level1 in the proposed model is implemented.
g. Classification-only prediction, where cluster analysis is not included and a classifier is generated from the entire training data set. The same collection of four classification algorithms specified above is also examined in this specific use case.

3.2. Experimental results and discussion

Based on the design described in the previous section, Table 2 shows the evaluation results of 6 different models with the case of School of management. Both overall as well as class-specific accuracies $\in [0, 100]$ are exploited here to compare predictive performance of different methods. For instance, the accuracy of Class A is estimated as: the number of Class A samples that are predicted correctly devided by the total number of Class A samples. In this table, all variants of the bi-level model have higher overall accuracies than that of the clustering-only counterpart. In addition, Random Forest (RF) obtains the highest overall accuracy of 93.96%. With respect to the accuracy of Class A, all the models are able to generate exceptional performance, with RF is the best again. However, for Class B, Naive Bayes (NB) achieves the highest accuracy of 79.71%, with RF obtains only at 42.03%. Unfortunately, the clustering-only or Level1 model is not able to identify any sample of Class B, with resulting in an accuracy of 0%. Another observation is with the KNN model performing better with K=1 than a bigger neighbor set of K=3.

Table 2. Evaluation results with different models, for the case of School of management

Model	Confusion Matrix			Class specific accuracy	Overall acuuracy
	A	B	Classified as		
Level1 (Clustering only)	839	72	A	92.10%	92.10%
	0	0	B	0.00%	
Bi-level (Naive Bayes)	792	50	A	94.06%	92.97%
	14	55	B	79.71%	
Bi-level (KNN, K=1)	815	27	A	96.79%	92.86%
	38	31	B	44.93%	
Bi-level (KNN, K=3)	810	32	A	96.19%	91.99%
	41	28	B	40.58%	
Bi-level (Decision Tree)	819	23	A	97.27%	93.19%
	39	30	B	43.48%	
Bi-level (Random Forest)	827	15	A	98.22%	93.96%
	40	29	B	42.03%	

In addition to the results reported in Table 2, Figure 4 depicts the comparison of accuracies specific to Class A, which are achieved by different variations of the bi-level framework (shown in Table 2) and four simple classifiers (NB, KNN, DT, and RF are trained with the whole training set). Note that for KNN, results with only K=1 are reported since they demonstrate the best performance among different K values. According to this, all of the four bi-level variations perform better than their corresponding baselines. For instance, the bi-level model implementing RF acquires the accuracy of 98.22%, almost 2% higher than the score achieved by a simple RF classifier. The largest improvement is witnessed with the case of NB, with the bi-level version reaches 94.06% and a simple NB is only at 88.10%. Likewise, Figure 5 shows a similar set of results for the Class-B prediction. This figure suggests that the bi-level framework usually outperforms the

corresponding simple classification models. In particular, NB obtains the highest accuracy of 79.71%, while the lowest of 42.03% is seen with RF. However, this is still a significant improvement from using a simple RF that is accurate at only 27.55%.

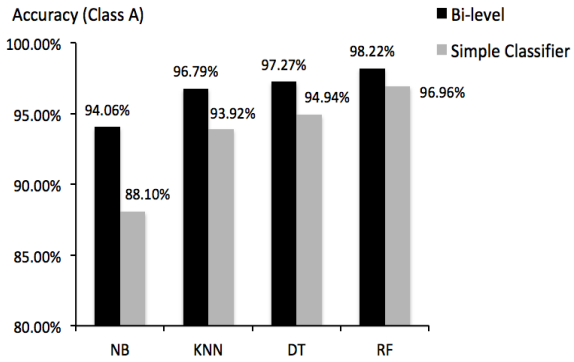

Figure 4. Class-A accuracies obtained by bi-level and basic classifiers, categorized by classification algorithm exploited for the training process. Note that the results with KNN are obtained using K=1

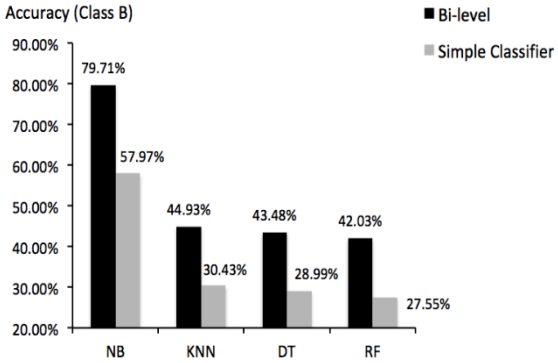

Figure 5. Class-B accuracies obtained by bi-level and basic classifiers, categorized by classification algorithm exploited for the training process. Note that the results with KNN are obtained using K=1

Similar to Table 2, Table 3 shows details of the evaluation results with the data belonging to School of information technology. For the overall accuracy, the bi-level (NB) and the clustering-only model obtain the highest and the lowest scores, respectively. The bi-level (RF) is the most effective for Class-A classification at 97.17%, while the bi-level (NB) proves to be exceptional for Class B. It reaches a high value of 92.31%. These results lead to a conclusion that the proposed framework is more accurate than using only the clustering results to guide prediction. Again, the KNN model with K=1 performs better than the other using K=3. Besides these, Figures 6 and 7 compare the accuracies obtained by bi-level variations and basic classifiers for Class A and Class B, respectively. Like the previous case, trends found with School of management also appear here with students from School of information technology. So, the findings that the proposed framework is better than simple classifiers and a clustering-only prediction are confirmed by these two set of results. In fact, it is generalized and applicable across different schools.

Table 3. Evaluation results with different models, for the case of School of information technology

Model	Confusion Matrix			Class specific accuracy	Overall acuuracy
	A	B	Classified as		
Level1 (Clustering only)	209	42	A	83.27%	83.27%
	0	0	B	0.00%	
Bi-level (Naive Bayes)	199	13	A	93.87%	93.63%
	3	36	B	92.31%	
Bi-level (KNN, K=1)	197	15	A	92.92%	86.06%
	20	19	B	48.72%	
Bi-level (KNN, K=3)	195	17	A	91.98%	84.06%
	23	16	B	41.03%	
Bi-level (Decision Tree)	197	15	A	92.92%	89.24%
	12	27	B	69.23%	
Bi-level (Random Forest)	206	6	A	97.17%	92.43%
	13	26	B	66.67%	

In order to digest those results further, Figure 8 reveals an important finding regarding the problem of class imbalance. According to Tables 2 and 3, the accuracies reported for Class A are usually better those of Class B. This is pretty much due to the uneven cardinality of samples belonging to these binary classes. In fact, based on the original class distribution for School of management shown in Figure 7, the proportion of instances of Class A is 92.10% and only 7.90% of the other. It is slightly better for School of information technology, with the ratios being 83.27% and 16.73%. It can be summarized from Figures 6 and 7 that most models included in this empirical study exhibit better performance with Class B in the case of School of

information technology, compared to the other case. The level of imbalance between classes in the former is less than the latter, which may well explain that observation. Another point worth noted here is that the bi-level framework can ease the imbalance problem with higher proportions of Class-B samples are included in the stage of classification modeling, see Figure 8 for details. Hence, bi-level variants are more accurate than their corresponding baseline counterparts, i.e., simple classifiers.

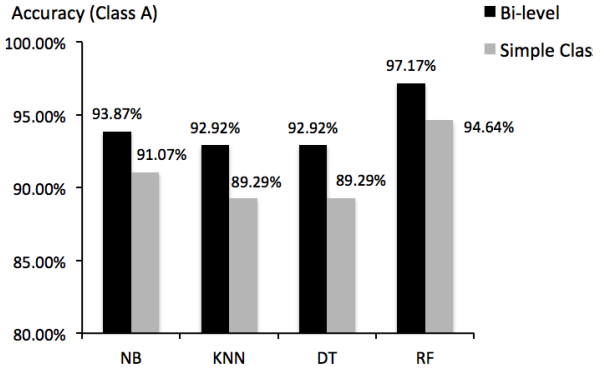

Figure 6. Class-A accuracies obtained by bi-level and basic classifiers, categorized by classification algorithm exploited for the training process. Note that the results with KNN are obtained using K=1

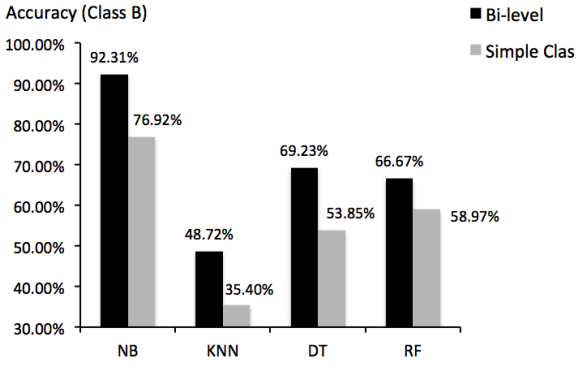

Figure 7. Class-B accuracies obtained by bi-level and basic classifiers, categorized by classification algorithm exploited for the training process. Note that the results with KNN are obtained using K=1

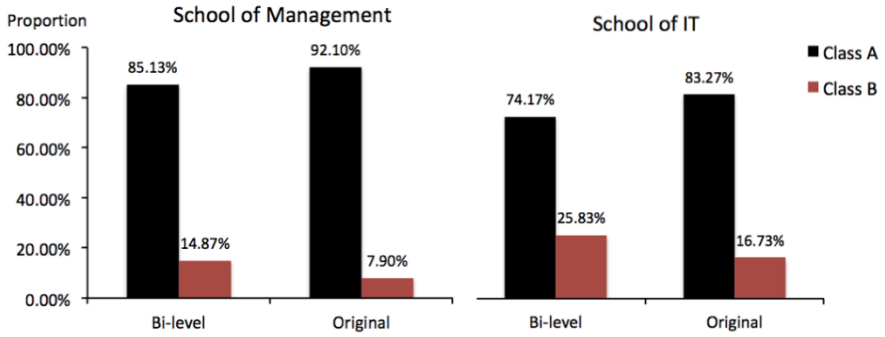

Figure 8. Comparison of class distributions between the entire original data (without clustering process) and those samples belonging to the cluster going through the second level of bi-level framework

4. CONCLUSION

This paper has presented an original work on the application of bi-level learning framework to determine patterns of student graduation. It is designed around a real collection of student enrollment and personal information. The proposed framework is divided into two tiers, with the initial applying a clustering technique to obtain clusters of student samples. A cluster of high quality is used as a reference for prediction, whereas those with the purity below a user-defined threshold are further analyzed using a choice of classifier. Evaluated on a data set specific to Mae Fah Luang University, the bi-level variations usually perform better than adopting simple classifiers to the whole data, or relying on the clustering result alone. This is due to the ability to solve the class imbalance to a certain extent. In fact, the application of Naive Bayes (NB) and Random Forest (RF) in the bi-level learning framework has proven more effective than other alternatives in this empirical study. While the former is the most accurate for Class B, the latter is exceptional for Claass A.

Despite such a positive finding, there are a few issues that might lead to future works. In addition to the methodology of bi-level learning model, an oversampling or undersampling technique may well be exploited to resolve the problem of class imbalance further. Also, the concept of classifier ensemble may be useful to aggregate predictions made by different classifiers, which are deployed at the second level of proposed framework. Another direction is with the use of consensus clustering and recent variants to provide an accurate clustering in the intial layer of proposed model.

ACKNOWLEDGEMENTS

This research is part of an MSc dissertation and also partly funded by MFU. It is also partly supported by STFC-GCRF2018: From Stars to Baht (collaboration between MFU & University of Sheffield).

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. e1355, 2020, doi: 10.1002/widm.1355.
[2] S. A. H. Basha, G. N. R. Prasad, M. Rao K and M. G. Vardhan, "A Review of Predictive and Descriptive Techniques in Higher Education Domain," *International Journal of Computer Engineering and Applications*, vol. XIII, no. VI, pp. 1-7, 2021.
[3] P. Baepler and C. J. Murdoch, "Academic analytics and data mining in higher education," *International Journal of Scholarship of Teaching and Learning*, vol. 4, no. 2, pp. 1-9, 2010, DOI:10.20429/ijsotl.2010.040217.
[4] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12-27, 2013.
[5] S. Z. Erdogan and M. Timor, "A data mining application in a student database," *Journal of Aeronautic and Space Technologies*, vol. 2, no. 2, pp. 53-57, 2005.
[6] R. Bakerand and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3-17, 2009, doi: 10.5281/zenodo.3554657.
[7] M. Bala and D. B. Ojha, "Study of applications of data mining techniques in education*," International Journal of Research in Science and Technology*, vol. 1, pp. 1-10, 2012.
[8] N. Iam-On and T. Boongoen, "Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 2, pp. 497-510, 2017, 10.1007/s13042-015-0341-x.
[9] S. H. Lin, "Data mining for student retention management," *Journal of Computing Sciences in Colleges*, vol. 27, no. 4, pp. 92-99, 2012, doi: 10.5555/2167431.2167450.
[10] K. Koedinger, K. Cunningham, A. Skogsholm, and B. Leber, "An open repository and analysis tools for fine-grained, longitudinal learner data," in *Proceedings of First International Conference on Educational Data Mining*, 2008, pp. 157-166.
[11] J. Mostow and J. Beck,"Some useful tactics to modify, map and mine data from intelligent tutors," *Natural Language Engineering*, vol. 12, pp. 195-208, 2006, doi:10.1017/S1351324906004153.
[12] M. Ramaswami and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining," *International Journal of Computer Science*, vol. 7, no. 1, pp. 10-18, 2010.
[13] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411-426, 2004, doi: 10.1080/08839510490442058.
[14] R. R. Kabra and R. S. Bichkar, "Performance prediction of engineering students using decision trees," *International Journal of Computer Applications*, vol. 36, no. 11, pp. 8-12, 2011, doi: 10.5120/4532-6414.
[15] C. Sung-Hyuk and C. Tappert, "Constructing binary decision trees using genetic algorithms," GEM, pp. 49-54, 2008.
[16] C. Yu, S. D. Gangi, A. Jannasch-Pennell, and C. Kaprolet, "A data mining approach for identifying predictors of student retention from sophomore to junior year," *Journal of Data Science*, vol. 8, pp. 307- 325, 2010, doi: 10.6339/JDS.201004_08(2).0007.
[17] K. Kongsakun and C. C. Fung, "Neural network modeling for an intelligent recommendation system supporting srm for universities in Thailand," *WSEAS Transactions on Computers*, vol. 11, no. 2, pp. 34- 44, 2012.
[18] R. Sittichai, "Why are there dropouts among university students? experiences in a thai university," *International Journal of Educational Development*, vol. 32, no. 2, pp. 283-289, 2012, doi: 10.1016/j.ijedudev.2011.04.010.
[19] S. Subyam, "Causes of dropout and program incompletion among undergraduate students from the faculty of engineering, King Mongkut University of Technology North Bangkok," in *Proceedings of 8th National Conference on Engineering Education*, 2009.
[20] R. Trakunphutthirak, Y. Cheung and V. Lee, "A Study of Educational Data Mining: Evidence from a Thai University," In *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 734-741, 2019, doi: 10.1609/aaai.v33i01.3301734.
[21] W. S. Nuankaew, P. Nuankaew, D. Teeraputon, K. Phanniphong and S. Bussaman, "Perception and Attitude Toward Self-Regulated Learning of Thailand's Students in Educational Data Mining Perspective," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 9, pp. 34-49, 2019, doi: 10.3991/ijet.v14i09.10048.
[22] N. Iam-On and T. Boongoen, "Generating descriptive model for student dropout: a review of clustering approach," *Human-centric Computing and Information Sciences*, vol. 7, no. 1, 2017, doi: 10.1186/s13673-016-0083-0.
[23] M. Pattanodom, N. Iam-On and T. Boongoen, "Clustering data with the presence of missing values by ensemble approach," *2016 Second Asian Conference on Defence Technology (ACDT)*, 2016, pp. 151-156, doi: 10.1109/ACDT.2016.7437660.
[24] N. Iam-On and T.Boongoen, "Diversity-driven generation of link-based cluster ensemble and application to data classification," *Expert Systems with Applications*, vol. 42, no. 21, pp. 8259-8273, 2015, doi: 10.1016/j.eswa.2015.06.051.

[25] P. Panwong, T. Boongoen and N. Iam-On, "Improving Consensus Clustering with Noise-Induced Ensemble Generation," *Expert System with Applications*, vol. 146, pp. 113-138, 2020, doi: 10.1016/j.eswa.2019.113138.

BIOGRAPHIES OF AUTHORS

Lalida Nanglae is an MSc student in Computational Science, School of Science, Mae Fah Luang University, Chiang Rai, Thailand. She received Bachelor Degree in Statistics from the Faculty of Science, Maejo University, Chiang Mai, Thailand in 2012. Her research interests are data science and education data mining.

Natthakan Iam-On is an Assistant Professor at School of Information Technology, Mae Fah Luang University. She obtained PhD in Computer Science from Aberystwyth University, UK in 2011. Also she received both MSc and BSc in Computer Science from Chiang Mai University, Thailand. She has published research and review articles in high-impact international journals like Machine Learning, IEEE Transactions of Pattern Analysis and Machine Intelligence, IEEE Transactions of Knowledge and Data Engineering, Bioinformatics, for instance. Her research interests include data science, biomedical data analysis, cluster analysis and ensemble.

Tossapon Boongoen is an Associate Professor at School of Information Technology, Mae Fah Luang University. He received his PhD in Computer Science from Cranfield University, UK in 2003. During 2007-2011, he has been a PDRA and visiting research fellow at Department of Computer Science, Aberystwyth University, UK. He has published articles in well-known venues like IEEE Transactions of Knowledge and Data Engineering, IEEE Transactions of Cybernetics, Machine Learning, Expert Systems with Applications, AI and Law, for instance. In additional, he serves as associate editors of IEEE Access and PeerJ Computer Science. His research interests are AI, machine learning, data science, uncertainty and fuzzy systems.

Komkrit Kaewchay is an Assistant Professor at Department of Aeronautical Engineering, Navaminda Kasatriyadhiraj Royal Air Force Academy. He received MSc in Aerospace Engineering from University of Texas at Arlington in 2004. He also received BS in Aeronautical Engineering from Republic of Korea Air Force Academy. His research interests include dynamic model, control, navigation and guidance of aerospace vehicle. He has also published article in Journal of Guidance, Control, and dynamics.

James Mullaney is a Senior Lecturer at Department of Physics and Astronomy, University of Sheffield, UK. He obtained his PhD in Astrophysics from Durham University, UK in 2008. Before that, he also received MSc in Physics and Astronomy The University of Nottingham, UK in 2004. He has published over 60 journal articles, six of which have more than100 citations. His interest is to develop new ways to extract useful information from large amounts of astronomical data.