Iam-On and Boongoen *Hum. Cent. Comput. Inf. Sci.* (2017) 7:1
DOI 10.1186/s13673-016-0083-0

Human-centric Computing and Information Sciences

REVIEW Open Access

Generating descriptive model for student dropout: a review of clustering approach

Natthakan Iam-On[1*] and Tossapon Boongoen[2]

*Correspondence:
natthakan@mfu.ac.th
[1] School of information Technology, Mae Fah Luang University, 333 Moo 1, Ta-sud, Muang, Chiang Rai 57100, Thailand
Full list of author information is available at the end of the article

Abstract

The implementation of data mining is widely considered as a powerful instrument for acquiring new knowledge from a pile of historical data, which is normally left unstudied. This data driven methodology has proven effective to improve the quality of decision-making in several domains such as business, medical and complex engineering problems. Recently, educational data mining (EDM) has obtained a great deal of attention among educational researchers and computer scientists. In general, publications in the field of EDM focus on understanding student types and targeted marketing, using both descriptive and predictive models to maximize student retention. Inspired by previous attempts, this paper aims to establish the clustering approach as a practical guideline to explore student categories and characteristics, with the working example on a real dataset to illustrate analytical procedures and results.

Keywords: Educational data mining, Clustering, Student performance, Retention, Dropout

Background

Given an increasing number of higher educational institutes and learning assisted technology, many universities have to adapt to changes in business environment and student expectation. This leads to a critical revision of their strategies and effectiveness [1], since they are held accountable for learning outcome and stakeholders' satisfaction. It appears that one response to this challenge is the application of decision-support tools, including analytical and data mining (DM) techniques [2]. Such an approach is in line with the need of most universities to handle and make the best use of large repositories of data, which normally cover enrollment and registration, learning materials and resources, course and student details [3]. With respect to [4], this data collection can be regarded as a goldmine, from which knowledge about students' behavior, preference and performance can be discovered.

Having recognized its potential, educational data mining (EDM), has been a fast-growing interdisciplinary research field [5]. It concerns with developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze [1]. As such, disclosed knowledge is highly useful to better understand how students learn and effects of

different settings to their achievement. This can help to improve educational outcomes and to gain insights into various educational phenomena, with several applications of EDM being put forward in the recent years [6]. Examples of these include evaluation of student performance, course recommendation and personalized learning plan, and identification of atypical learning pattern [7].

Specific to the understanding of student performance, several researches have shown that EDM can help to disclose at-risk students. This allows universities to become more proactive in identifying and assisting those students. For instance, with the aim to yield student retention, Lin [8] has studied a variety of machine learning algorithms to develop predictive models based on incoming students' data. As such, the models are able to provide short-term accuracy for predicting which types of students would benefit from student retention programs on campus. To achieve alike models of performance and dropout, different techniques have been explored. These include Naive Bayes [9], *k*-means [10], decision tree [11, 12]. In addition, recent researches have also focused on understanding student groups and corresponding policy [13]: predictive models to maximize student retention [14, 15], an enrollment prediction system, for instance.

Student retention has become a common problem encountered by any university around the world, including Mae Fah Luang University (MFU) and others in Thailand. However, this does not draw much attention among researchers in Thai agencies and neighboring countries, with a handful of investigations being pursued in the past few years. Examples are the work conducted at King Mongkut University of Technology North Bangkok [16], and another for Prince of Songkla University [17]. Note that the model is limited to predictive purpose with the use of conventional classification algorithms [18]. Failing to improve or even sustain the retention rate would negatively impact students, parents, university and the society as a whole. On the other hand, the success will bring about several benefits such as better graduates' career, higher university's ranking, and more funding from both government agencies and private sector. As suggested by [19], universities with high attrition or dropout rates may face the significant loss of tuition fees and potential alumni contributions. Note that a significant portion of student attrition occurs in the first year of university. According to the research of [20], more than 50% of the student attrition can be attributed to the freshmen. Therefore, it is essential to identify vulnerable students who are prone to dropout as early as possible. This allows institutions to better and faster progress towards achieving their retention management goals.

Note that almost all the aforementioned attempts are constrained to the use of analytical algorithms only to generate a predictive model of students' success and failure. As such, the prediction result is narrowed to the likely achievement of any student under examination, typically as either graduate or dropout. Unfortunately, a predictive method often fails to provide insights to understand factors and characteristics of those two student categories. In response, this review paper aims to boost the quality of analytical results by exploring the development of a descriptive model that can largely complement the predictive side, or even provide a unique and useful viewpoint hardly obtained before. One of the major approaches to deliver a desired descriptive model is data clustering, which is an unsupervised learning process for the exploration of data structural setting and properties. It is capable of revealing natural groups of objects of

interest, especially for a new domain with minimal prior knowledge. As a result, clustering has been coupled with many real problems, including bioinformatics [21], medical and health informatics [22], psychological study [23], marketing research [24], customer relationship [25], and recommender systems [26]. Furthermore, the development of clustering for microarray gene expression data motivates a large number of contributions regarding both theoretical advancement and applications [27–29].

The rest of this paper is organized as follows. As for the development of a descriptive model, one of the recent developments in subspace clustering model is employed. Therefore, its basic assumption and process are presented in the second section, including its baseline technique. This also provides details of the model evaluation, in which different quality metrics are made available to ensure the reliability of clustering result. Following that, the third section illustrates a working example of descriptive model generation and interpretation, based on the case study of MFU. This discovery will allow a more in-depth analysis where significant factors to a particular group-wise character can be revealed. The review is concluded in the forth section, with a discussion of future research directions.

Basis of cluster analysis

Principally, the core of cluster analysis is the clustering process, which divides data objects into groups or clusters such that objects in the same cluster are more similar to each other than to those belonging to different clusters [30]. Objects under examination are normally described in terms of object-specific (e.g., attribute/feature values) or relative measurements (e.g., pairwise dissimilarity). Unlike supervised learning to which classification is categorized, clustering is 'unsupervised' and does not require class information. This is typically achieved through a manual tagging of category labels on data objects, by a domain expert (or through the consensus of multiple experts). Given its potential, a large number of research studies focus on several aspects of cluster analysis; for instance, clustering algorithms and extensions for particular data type [31], dissimilarity (or distance) metric [32], optimal cluster number [33], relevance of data attributes per cluster [34], evaluation of clustering results [35], and cluster ensembles [36]. This section aims to set the scene for the following section by emphasizing the clustering technique used for generating a descriptive model of student performance. In addition, a section of model evaluation is also included to shed the light on measuring goodness of the obtained model.

Model generation

Clustering is branded an unsupervised learning approach as the measurement of similarity is conducted without knowledge of class assignment. This knowledge-free scenario brings about a series of difficult decisions, with respect to selecting appropriate algorithm, similarity measure, criterion function, and initial parameter condition [37, 38]. For a given data $X \in R^{n \times d}$, each $x_i, i = 1 \ldots n$, corresponds to a sample or data point, which can be represented by a profile of d features, i.e., $x_i = (x_{i1}, \ldots, x_{id})$. A clustering algorithm searches for the partition $\pi = \{C_1, \ldots, C_k\}$ of samples $(x_1, \ldots, x_n)$ into k clusters, such that samples in the same cluster are more similar to each other than to those in the other clusters.

There are a large number of clustering algorithms developed in the literature. Among these, *k*-means is perhaps, the best known clustering technique. Its name comes from the mechnism of representing each of *k* clusters by the mean of its members or so-called 'centroid'. *k*-means is an iterative algorithm that exploits a square-error as a criterion function (i.e., the total distance between each data point and its cluster center [39]). It begins with initializing centroids randomly and then allocates data points to clusters such that the square-error is minimized. This criterion function tends to work well with separated and compact clusters. Given a dataset X, the square-error e^2 of a clustering π is defined as the following.

$$e^2(X,\pi) = \sum_{p=1}^{k} \sum_{\forall x \in C_p} \|x - \overline{c_p}\|^2, \tag{1}$$

where $\|.\|$ denotes the Euclidean norm and $\overline{c_p}$ is the center of the pth cluster. A general description of the *k*-means algorithm is summarized below.

1. k data points are first randomly selected as initial cluster centers.
2. Repeat:
   a. Assign each data point to its closest cluster center. The Euclidean metric is commonly used to compute the distance between data points and centroids.
   b. The centroid of each cluster is updated as the mean of all current data points in that cluster.
3. Until the termination criteria are met.

Examples of termination criteria are: (i) no change is made to the cluster centers (i.e., no reassignment of any data point from one cluster to another), (ii) the maximum number of iterations is exceeded, and (iii) there is no improvement in the objective function such as a decrease in the square-error. The *k*-means algorithm is popular largely due to its efficiency, with time complexity of $O(Nkr)$, where N is the number of data points, k is the number of clusters and r is the number of iterations. However, it is sensitive to the choice of initial cluster centers. In other words, different initial states can lead to different output partitions. Another drawback relates to the fact that all features equally contribute to the distance measure, which is hardly the case for many problem domains. To tackle this, a number of soft subspace clustering methods are proposed to determine significance degrees of different features in the clustering process. One of these called R-KM [40] has been recently introduced with superior performance than existing counterparts. Specific to this research, it is employed to generate a descriptive model for the case study discussed in the next section.

This method consists of two major stages. The first involves exploiting the data reliability measure [34] to construct a sample-feature association matrix. It represents the locally relevance degree of each feature per sample. Let $\alpha \in \{1 \ldots (n-1)\}$ be the number of nearest neighbors of any sample under examination. The sample-feature association matrix $\Theta^{\alpha} \in R^{n \times d}$ is a collection of information entries $\Theta_{ij}^{\alpha} \in [0,1]$ representing the

strengths that a sample $x_i,\ i = 1 \ldots n$, is similar to (or associates with) a set $N_{ij}^{\alpha} \subset X$ of its α nearest-neighboring samples in a given feature dimension $g_j,\ j = 1 \ldots d$. Formally, the underlying strength measure is defined as follows.

$$\Theta_{ij}^{\alpha} = 1 - \left(\frac{\Pi_{ij}^{\alpha}}{\Pi_{*}^{\alpha}} \right), \tag{2}$$

where $\Pi_{*}^{\alpha} = \max_{\forall i,j}\ \Pi_{ij}^{\alpha}$, with Π_{ij}^{α} being estimated by

$$\Pi_{ij}^{\alpha} = \frac{1}{\alpha} \sum_{\forall q \in N_{ij}^{\alpha}} \sqrt{(x_{ij} - q_j)^2} \tag{3}$$

Note that the estimation of data reliability measure relies on the search for α nearest neighbors of any sample in question. See the study of [34] for the algorithm that is employed to efficiently find $N_{ij}^{\alpha}, \forall i = 1 \ldots n, j = 1 \ldots d$. The measure Θ_{ij}^{α} has an intuitive interpretation towards the problem of subspace clustering. As it approaches 1, feature g_j is highly relevant to the local cluster in which sample x_i is an element. If however, the underlying measure is close to 0, the feature becomes irrelevant to the clustering of x_i.

The next stage accounts for the clustering model of R-KM. It extends the conventional k-means such that the association values in matrix Θ^{α} are automatically employed in the formulation of sample clusters. Unlike the existing approaches to soft subspace clustering, dimensional weights are updated using sample-specific reliability measures, which represent true characteristics of locally relevance and remain unchanged over time. The R-KM algorithm aims to minimize the following objective function:

$$J_R(U, Z, W) = \sum_{l=1}^{\beta} \sum_{i=1}^{n} \sum_{j=1}^{d} u_{il}\ w_{lj} (x_{ij} - z_{lj})^2 \tag{4}$$

At the initial stage where cluster centroids $Z = \{z_1, \ldots, z_k\}$ correspond to a set of randomly selected samples in X, the weight w_{lj} of the j-th feature in cluster $C_l \in \pi$ is estimated as

$$w_{lj} = \frac{\Theta_{ij}^{\alpha}}{\sum_{t=1}^{d} \Theta_{it}^{\alpha}}, \quad j = 1 \ldots d, \tag{5}$$

given that $x_i \in X$ is selected as z_l. In the following iterations, feature-specific weight w_{lj} of each cluster $C_l \in \pi$ is updated by

$$w_{lj} = \frac{\Phi_{lj}^{\alpha}}{\sum_{t=1}^{d} \Phi_{lt}^{\alpha}}, \quad j = 1 \ldots d, \tag{6}$$

where Φ_{lj}^{α} is the association measure to the j-th feature. This is minimally shared by all members of cluster C_l:

$$\Phi_{lj}^{\alpha} = \min_{\forall x_i \in C_l} \Theta_{ij}^{\alpha} \tag{7}$$

With both Z and W being fixed, the crisp cluster membership $u_{il} \in U,\ i = 1 \ldots n,\ l = 1 \ldots k$ can be specified as

$$u_{il} = \begin{cases} 1 & \text{if } l = \arg\min\limits_{s=1\ldots k} \sum\limits_{j=1}^{d} w'_{sj}(x_{ij} - z_{sj})^2, \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $w'_{sj}, s = 1 \ldots k,\ j = 1 \ldots d$, is defined by

$$w'_{sj} = \frac{\min\left(\frac{\Theta^{\alpha}_{ij}}{\sum_{t=1}^{d}\Theta^{\alpha}_{it}}, w_{sj}\right)}{\sum_{t'=1}^{d}\min\left(\frac{\Theta^{\alpha}_{it'}}{\sum_{t=1}^{d}\Theta^{\alpha}_{it}}, w_{st'}\right)} \tag{9}$$

Similar to the classical k-means method, the set of centroids Z is updated using the following.

$$z_{lj} = \frac{\sum_{i=1}^{n} u_{il}\, x_{ij}}{\sum_{i=1}^{n} u_{il}} \tag{10}$$

Given these definitions, the R-KM algorithm can be summarized as follows.

ALGORITHM: $R - KM(\beta, \Theta^{\alpha})$

(1) Randomly initialize Z

(2) Calculate initial weights by Eq. 5

(3) **Repeat**

(4) Update U by Eq. 8

(5) Update Z by Eq. 10

(6) Update W by Eq. 6

(7) **Until** the objective function obtains its local minimum

Model evaluation

Having achieved a clustering result, with respect to the preferred number of clusters or k, the quality of this data partition can be assessed using one of many quality indices published in the literature. These can be largely categorized into internal and external measures. The former makes use of information regarding data attributes and cluster assignment only, while the other also includes class information that may not always exist for comparison. Hence, the internal family is sought to be appropriate for justifying the goodness of clustering in an unsupervised fashion. Examples of these metrics are explained below.

- *Davies-Bouldin (DB)* makes use of similarity measure R_{ij} between the clusters C_i and C_j, which is defined upon a measure of dispersion (s_i) of a cluster C_i and a dissimilarity measure between two clusters (d_{ij}). According to [41], R_{ij} is formulated as

  $$R_{ij} = \frac{s_i + s_j}{d_{ij}}, \tag{11}$$

  where d_{ij} and s_i can be estimated by the following equations. Note that v_x denotes the center of cluster C_x and $|C_x|$ is the number of data points in cluster C_x.

  $$d_{ij} = d(v_i, v_j), \tag{12}$$

$$s_i = \frac{1}{|C_i|} \sum_{\forall x \in C_i} d(x, v_i) \tag{13}$$

Following that, the DB index of a clustering π with k clusters, is defined as

$$DB(\pi) = \frac{1}{k} \sum_{i=1}^{k} R_i, \tag{14}$$

where $R_i = \max_{j=1...k, i \neq j} R_{ij}$. The DB index measures the average of similarity between each cluster and its most similar one. As the clusters have to be compact and separated, the lower DB index indicates better goodness of a data partition.

- *Dunn* is introduced by [42]. Its purpose is to identify compact and well-separated clusters. For a given number of clusters k, the definition of the Dunn index is given by the following equation.

$$Dunn(\pi) = \min_{i=1...k} \left(\min_{j=i+1...k} \left(\frac{d(C_i, C_j)}{\max_{z=1...k}(diam(C_z))} \right) \right), \tag{15}$$

where $d(C_i, C_j)$ is the distance between two clusters C_i and C_j, which can be defined as

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{16}$$

In addition, $diam(C_i)$ is the diameter of a cluster C_i, which is defined as follows:

$$diam(C_i) = \max_{x,y \in C_i} d(x, y) \tag{17}$$

In a dataset containing compact and well-separated clusters, the distances between the clusters are expected to be large and the diameters of the clusters are expected to be small. Therefore, a large value of the Dunn index signifies compact and well-separated clusters.

Clustering approach to generate a descriptive model of student dropout

To accomplish objectives of the review set out earlier, this section presents the application of clustering approach to a real case study. It illustrates the research methodology used to deliver the working data-mining model, with the capability to provide descriptive insight towards student academic performance. Following a number of EDM studies found in the literature, the conventional DM framework is employed. This consists of a sequence of different stages regarding problem definition, data acquisition and preparation, model development and evaluation, as well as model interpretation, respectively. Figure 1 presents summarizes these stages that will be discussed in the following sections.

Problem definition

Specific to MFU, the retention problem has not been investigated nor properly treated since the establishment in 1998. As the number of enrollment exceeding 10,000 in the academic year of 2015, the loss due to student attrition becomes more crucial. This would project a serious issue when other universities in the region have proven to be more attractive, in terms of student's achievement, tuition fee and personalized

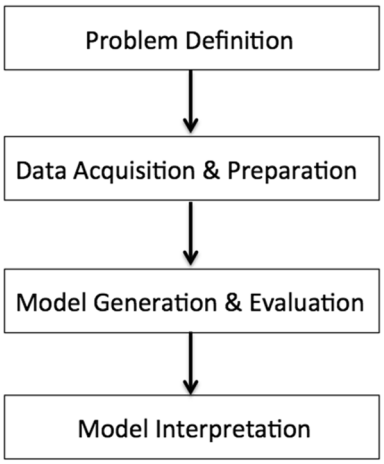

Fig. 1 Framework. The analysis framework for generating a descriptive model of student dropout

assistance. An internally funded research is set to initialize the application of EDM to MFU problems, such that the primary findings may reveal significant patterns, trends and relations useful for future management. It kicks off with the definition of problem and context, which are in line with the requirements of Admission and Registrar Divisions. One of several preferences is to obtain major groups of student behavior, in which factors to academic performance can be further determined. In order to achieve this, the research problem is defined with respect to different perspectives. Two contexts of the underlying problem can be examined, regarding scope of investigated data and the application of analysis results:

- Problem Context1—this focuses on the student cases before starting the first-year study at MFU. The data under examination covers prior-university academic capability, demographic and degree enrollment details. The outcome will be beneficial for university admission, with students being helped to choose an appropriate degree. This is based on the degree outcome suggested by the data mining model and student's preference. Of course, the level of student attrition might be reduced given this guidance. Furthermore, factors related to both desired and undesired performance can be exploited for an effective enrollment strategy.
- Problem Context2—the focus has been shifted to the scenario observed after the first-year study. In addition to the aforementioned collection of data, first-year academic performance is also covered for this purpose. The expected results reflect student stereotypes as taking on university courses. They can be used to identify at-risk freshmen, and possible early assistive measures. Also, performance associated factors can lead to an effective degree/course planning.

Data acquisition and preparation

Having problem and application contexts defined, the next phase deals with data acquisition and preparation prior to model development. Initial data for the following generation of data mining models is retrieved from MFU MIS. Among 484 relational data

tables, only a few that contribute to students' information are targeted. As with the organizational policy regarding data privacy and security matter, those staffs in Admission and Registrar divisions have authorized this project the use of a limited snapshot of the working database. In particular, the following four views are provided for research purpose only.

View1 (VW_STUDENT), students' general data, for instance:

- STUDENTID (student's identification number)
- LEVELID (student's degree level, with '3' specifying Bachelor degree)
- DEPARTMENTID (identification number of academic department)
- SEX (student's gender)
- HOMEPROVINCEID (identification number of student's home province)
- ADMITACADYEAR (year of student's admission)
- ENTRYTYPE (code denoting type of admission)
  - RQ (Regional Quota),
  - DA (Direct Admission),
  - ADA (Additional Direct Admission),
  - CA (Conditional Admission, with school GPAX above 2.0)
- STUDENTSTATUS (code representing student's status)
  - status '40' means student graduated,
  - status '50' means student resigned,
  - status '61' means student dropout with GPAX less than 1.5,
  - status '62' means student dropout with GPAX less than 1.8,
  - status '63' means student was dropout with GPAX less than 2.0
- FINISHDATE (date of termination - graduate or dropout)

View2 (VW_APPLICANT), students' personal information and pre-university grading data:

- APPLICANTID (identification number of an university-entry applicant)
- GPAX (student's school overall grade)
- GPA1 (student's school grade, with respect to English subjects)
- GPA2 (student's school grade, with respect to Mathematical subjects)
- GPA3 (student's school grade, with respect to Science subjects)
- GPA4 (student's school grade, with respect to General subjects)

View3 (VW_TRANSCRIPT), students' academic performance:

- ACADYEAR (academic year in which student takes a specific course)
- SEMESTER (semester in which student takes a specific course)
- COURSECODE (code denoting academic course)
- GRADE (course grade, i.e., A, B+, B, C+, C, D+, D, F, S, U, or W)

View4 (VW_ENTRYTYPE), description of entry type:

- ENTRYTYPE (code denoting type of admission)
- ENTRYTYPEDES (description of entry type)

The project initially aims to make use of student data covering the admission years of 2007–2009. As such, SQL Query1 is employed to extract a set of 'STUDENTID' that specifies the target group. Note that year 2007 A.D. is equivalent to year 2550 B.E., where STUDENTID takes the format '50xxxxxxx'. Unfortunately, school-performance details of those students admitted in 2007–2008 were not recorded. This makes the target group of students smaller, with the focus on the admission year of 2009, i.e., 2552 B.E. The final data collection consists of 811 records each belonging to a specific undergraduate student (LEVELID $= 3$) who either graduates (STUDENTSTATUS $= 40$) or dropouts (STUDENTSTATUS $\in \{50, 61, 62, 63\}$). The initial fact observed with this dataset is that 271 students (33.42%) dropout right after the first year or later.

SQL Query1:

$$
\begin{aligned}
&\textbf{select } S.STUDENTID \\
&\textbf{from } VW\_STUDENT\ S \\
&\textbf{where } S.LEVELID = 3 \\
&\quad \textbf{and } \textbf{year}(S.FINISHDATE) > S.ADMITACADYEAR \\
&\quad \textbf{and } S.STUDENTSTATUS\ in\ (40, 50, 61, 62, 63) \\
&\quad \textbf{and } S.STUDENTID > 5000000000 \\
&\quad \textbf{and } S.STUDENTID < 5300000000
\end{aligned}
$$

SQL Query2:

$$
\begin{aligned}
&\textbf{select } S.STUDENTID,\ S.SEX,\ S.HOMEPROVINCEID, \\
&\quad S.ENTRYTYPE,\ S.DEPARTMENTID \\
&\textbf{from } VW\_STUDENT\ S \\
&\textbf{where } S.LEVELID = 3 \\
&\quad \textbf{and } \textbf{year}(S.FINISHDATE) > S.ADMITACADYEAR \\
&\quad \textbf{and } S.STUDENTSTATUS\ in\ (40, 50, 61, 62, 63) \\
&\quad \textbf{and } S.STUDENTID > 5200000000 \\
&\quad \textbf{and } S.STUDENTID < 5300000000
\end{aligned}
$$

Table 1 summarizes 21 features used in this empirical study, with respect to data type and involvement in the aforementioned application contexts. In particular to student's sex, the dataset comprises 253 male and 558 female, who are originally from 72 different home provinces. Their university entries can be categorized into four types of RQ (Regional Quota), DA (Direct Admission), ADA (Additional Direct Admission), and CA (Conditional Admission, with school GPAX above 2.0). The numbers of students belonging to these types are 222 of RQ, 393 of DA, 178 of ADA, and 18 of CA. This dataset cover students from 26 academic departments of MFU, e.g., Law, Business Administration, Information Technology, Nursing Science, and Public Health. These nominal variables can be retrieved using SQL Query2.

In addition to the aforementioned four nominal features, there are 17 numerical variables included in this examination. Five of these represent student's prior-university

Table 1 Description of investigated dataset, with Context1 and Context2 denoting two problem contexts of before- and after-first-year prediction

Feature	Data type	Context1	Context2	Description
Sex	Nominal	Applicable	Applicable	Student's sex
Province	Nominal	Applicable	Applicable	Student's home province
Type	Nominal	Applicable	Applicable	Type of university entry
Department	Nominal	Applicable	Applicable	Academic department
S-GPAX	Numerical	Applicable	Applicable	School grade (Overall)
S-GPA1	Numerical	Applicable	Applicable	School grade (English)
S-GPA2	Numerical	Applicable	Applicable	School grade (Mathematics)
S-GPA3	Numerical	Applicable	Applicable	School grade (Science)
S-GPA4	Numerical	Applicable	Applicable	School grade (General)
GPAX	Numerical	n/a	Applicable	Student's university grade
A ratio	Numerical	n/a	Applicable	Ratio of subject with grade A
B+ ratio	Numerical	n/a	Applicable	Ratio of subject with grade B+
B ratio	Numerical	n/a	Applicable	Ratio of subject with grade B
C+ ratio	Numerical	n/a	Applicable	Ratio of subject with grade C+
C ratio	Numerical	n/a	Applicable	Ratio of subject with grade C
D+ ratio	Numerical	n/a	Applicable	Ratio of subject with grade D+
D ratio	Numerical	n/a	Applicable	Ratio of subject with grade D
F ratio	Numerical	n/a	Applicable	Ratio of subject with grade F
S ratio	Numerical	n/a	Applicable	Ratio of subject with grade S
U ratio	Numerical	n/a	Applicable	Ratio of subject with grade U
W ratio	Numerical	n/a	Applicable	Ratio of withdrawn subject

Note that 'n/a' is the abbreviation of 'not applicable'

academic capability: the overall grade (S-GPAX) and average grades across four subject groups (S-GPA1, S-GPA2, S-GPA3 and S-GPA4). Note that these are extracted from VW_APPLICANT, with the original attribute names of GPAX, GPA1, GPA2, GPA3 and GPA4, respectively. Figures 2, 3, 4, 5, 6 show the distribution of these variables. The other 12 features that are applicable only to the second problem context, regard student's

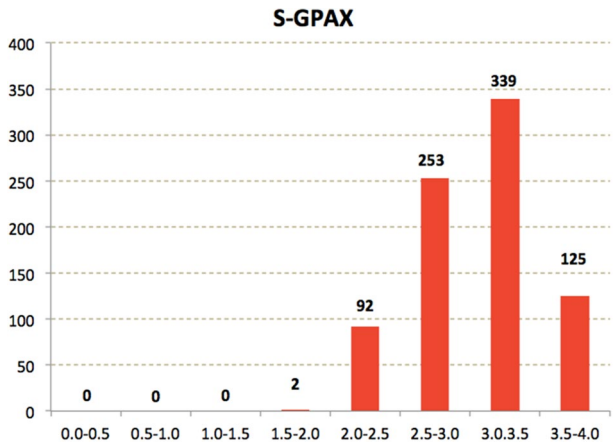

Fig. 2 S-GPAX. Distribution of S-GPAX values, where *x-axis* and *y-axis* correspond to value intervals and number of data instances (or students), respectively

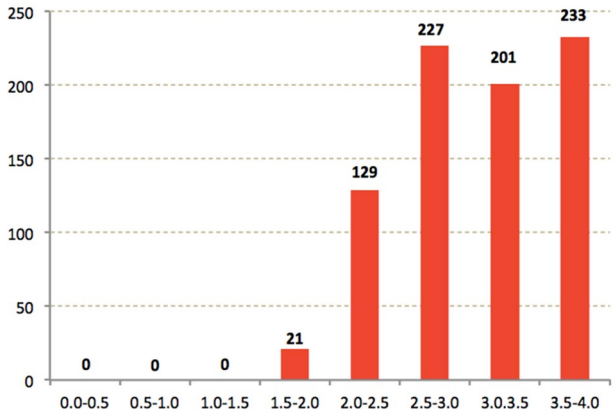

Fig. 3 S-GPA1. Distribution of S-GPA1 values, where *x-axis* and *y-axis* correspond to value intervals and number of data instances (or students), respectively

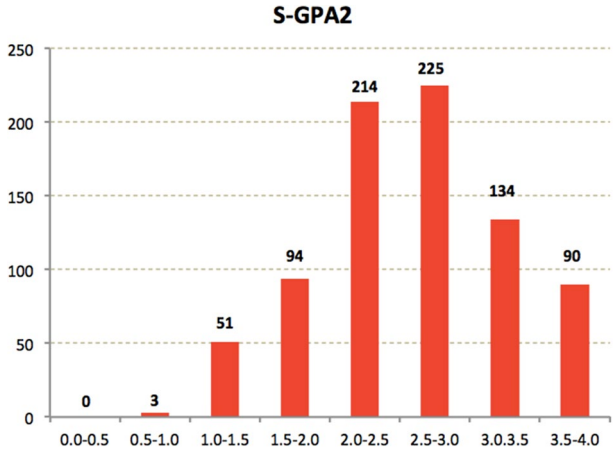

Fig. 4 S-GPA2. Distribution of S-GPA2 values, where *x-axis* and *y-axis* correspond to value intervals and number of data instances (or students), respectively

performance in the first year. These are encoded as ratios of different grades achieved from a collection of registered subjects. Note that the academic assessment is subjected to two grading systems: (a) 8-level setting, i.e., A, B+ and so on, and (b) 2-level setting, i.e., S (Satisfied) and U (Unsatisfied).

Let S_i be the total number of subjects taken by the i-th student in the first year, $S_i = S_i^a + S_i^b$, given that S_i^a and S_i^b denote the number of subjects with 8-level and 2-level assessment systems, respectively. The student's ratio of subject with grade $r \in \{A, B+, B, C+, C, D+, D, F\}$, RT_i^r, is estimated by

$$RT_i^r = \frac{r_i}{S_i^a}, \tag{18}$$

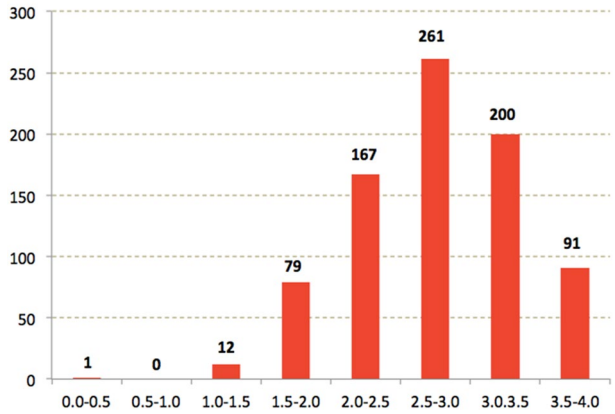

Fig. 5 S-GPA3. Distribution of S-GPA3 values, where *x-axis* and *y-axis* correspond to value intervals and number of data instances (or students), respectively

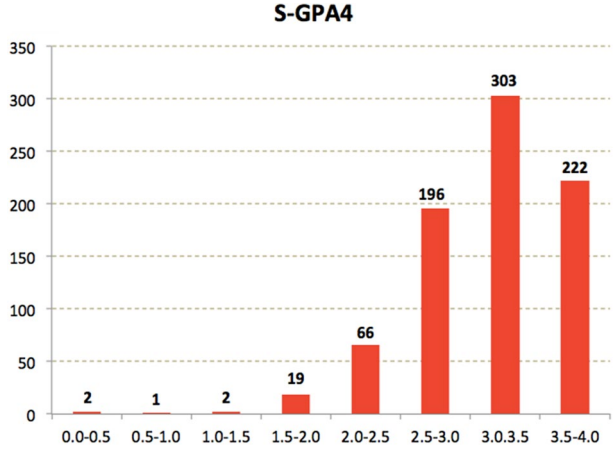

Fig. 6 S-GPA4. Distribution of S-GPA4 values, where *x-axis* and *y-axis* correspond to value intervals and number of data instances (or students), respectively

where $r_i \in \{0, \ldots, S_i^a\}$ is the number of subjects with grade r obtained by student i. Likewise, the student's ratio of subject with grade $t \in \{S, U\}$, RT_i^t, is defined as

$$RT_i^t = \frac{t_i}{S_i^b}, \tag{19}$$

here $t_i \in \{0, \ldots, S_i^b\}$ is the number of subjects with grade t obtained by student i. In addition, the ratio of withdrawn subject is

$$RT_i^W = \frac{W_i}{S_i}, \tag{20}$$

given that $W_i \in \{0, \ldots, S_i\}$ is the number of subjects withdrawn by student i. Table 2 illustrates statistical details for each of numerical features.

Table 2 Statistical details of numerical features

Feature	Range	Max	Min	Mean
S-GPAX	[0.00–4.00]	3.96	1.98	3.06
S-GPA1	[0.00–4.00]	4.00	1.60	3.09
S-GPA2	[0.00–4.00]	4.00	0.75	2.62
S-GPA3	[0.00–4.00]	4.00	0.00	2.78
S-GPA4	[0.00–4.00]	4.00	0.00	3.17
GPAX	[0.00–4.00]	4.00	0.00	2.36
A ratio	[0.00–1.00]	1.00	0.00	0.12
B+ ratio	[0.00–1.00]	0.64	0.00	0.15
B ratio	[0.00–1.00]	0.67	0.00	0.17
C+ ratio	[0.00–1.00]	0.67	0.00	0.16
C ratio	[0.00–1.00]	1.00	0.00	0.14
D+ ratio	[0.00–1.00]	1.00	0.00	0.09
D ratio	[0.00–1.00]	1.00	0.00	0.08
F ratio	[0.00–1.00]	1.00	0.00	0.08
S ratio	[0.00–1.00]	1.00	0.00	0.63
U ratio	[0.00–1.00]	1.00	0.00	0.19
W ratio	[0.00–1.00]	0.67	0.00	0.02

Model generation and evaluation

Data clustering approach is employed to develop a specific descriptive model for each of the two problem contexts under examination. Having obtained natural clusters of student academic profiles, it is possible to discover main characteristics and relations amongst variables of interest. To generate an accurate set of clusters, the aforementioned soft subspace clustering algorithm (i.e., R-KM) is exploited with respect to the following settings.

- The set of data used in this clustering phase includes only academic-performance variables:
  - Problem Context1: S-GPAX, S-GPA1, S-GPA2, S-GPA3 and S-GPA4
  - Problem Context2: S-GPAX, S-GPA1, S-GPA2, S-GPA3, S-GPA4, GPAX, A Ratio, B+ Ratio, B Ratio, C+ Ratio, C Ratio, D+ Ratio, D Ratio, F Ratio, S Ratio, U Ratio and W Ratio
- The number of clusters (k) is determined by the consensus of quality indices, such as Davies-Bouldin (DB) and Dunn. First, clusterings of the investigated data set are created using different values of $k \in \{2, 3, \ldots, 10\}$. Then, the optimal k is justified as the corresponding clustering having the best quality measures, which are summarized from 20 trials of each k-specific study.
- Having achieved the student clusters, their stereotypes (in terms of academic performance profiles) can be derived for future references. In addition, the value distribution of other features such as 'Entry Type' can be examined. This can reveal relations and trends specific to each of the disclosed student clusters, hence the strategies to tackle dropout or underachievement.

Based on the values of DB and Dunn, Figs. 7, 8, 9, 10 similarly illustrate that the optimal number of clusters (k) is 2 for both problem contexts. Note that these measures are concluded from repeated experiments, with low DB and high Dunn values being preferred. Specific to Problem Context1, two academic-profile clusters are generated, with the sizes of namely Cluster1 and Cluster2 being 434 and 377, respectively. As for the other context, a similar set of clusters is created with the sizes of 529 and 282 student profiles. See Figs. 11 and 12 for the corresponding illustrations.

Model interpretation

Having obtained two clusters specific to the study of Context1, it is possible to extract the representative profile for each of these two clusters, i.e., cluster centroids or centers. According to Fig. 13, the difference between profiles of Cluster1 and Cluster2 is obvious,

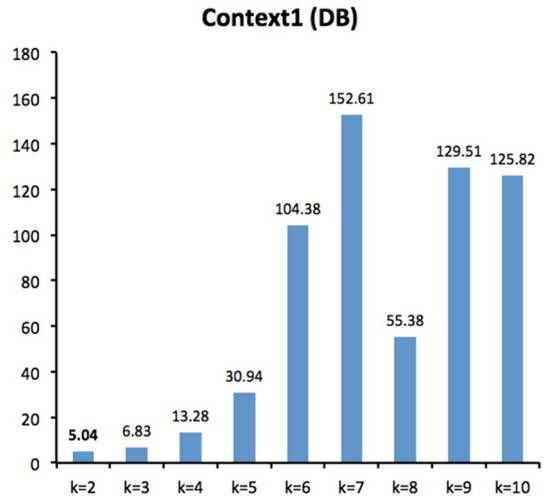

Fig. 7 DB Context1. DB measures with different number of clusters (k), with respect to Problem Context1

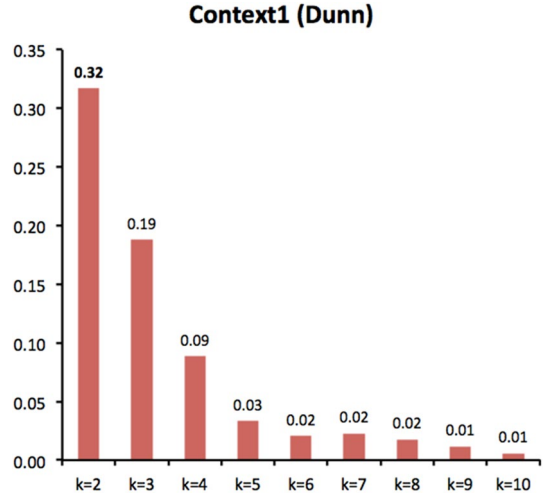

Fig. 8 Dunn Context1. Dunn measures with different number of clusters (k), with respect to Problem Context1

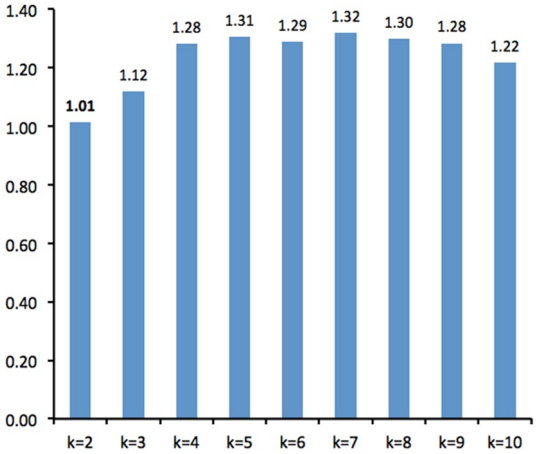

Fig. 9 DB Context2. DB measures with different number of clusters (k), with respect to Problem Context2

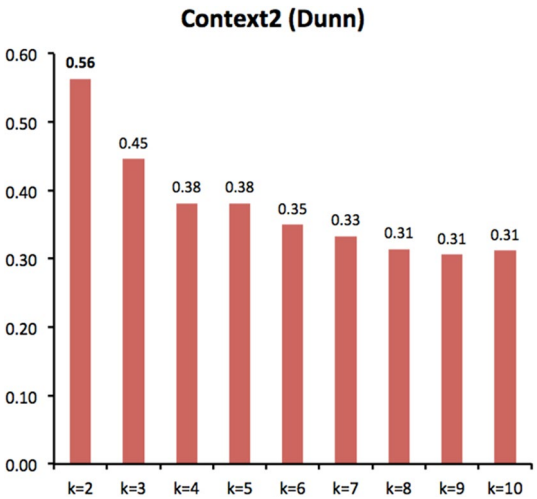

Fig. 10 Dunn Context2. Dunn measures with different number of clusters (k), with respect to Problem Context2

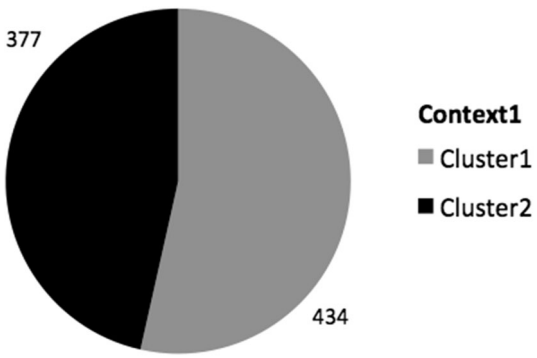

Fig. 11 Cluster Context1. Sizes of discovered clusters, with respect to Problem Context1

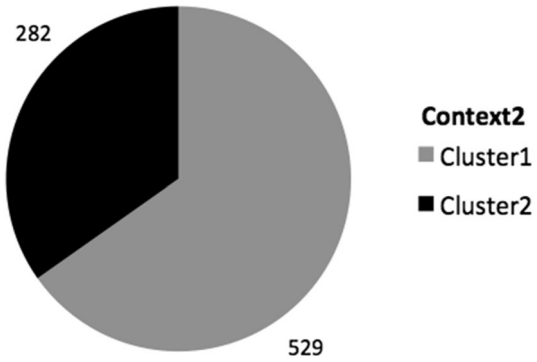

Fig. 12 Cluster Context2. Sizes of discovered clusters, with respect to Problem Context2

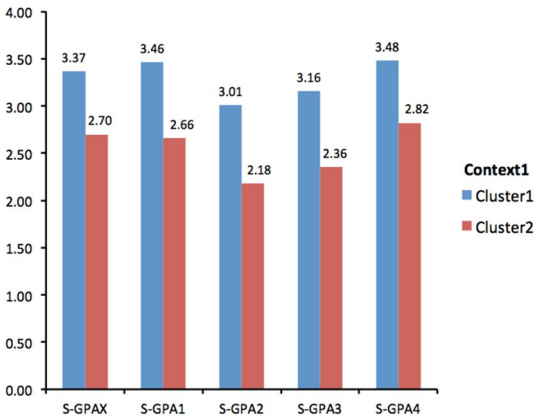

Fig. 13 Profile Context1. Profile representatives of two clusters (with 'high' and 'low' school-level academic capability), with respect to Problem Context1

where S-GPAX, S-GPA1, S-GPA2, S-GPA3 and S-GPA4 values of students belonging to the former are consistently higher than those of the other. Given these statistics, the clusters under examination may be interpreted as students with 'high' and 'low' prior-university performance, respectively.

Similarly, Fig. 14 shows the two obvious trends with additional first-year assessment results. Note that GPAX denotes the overall grade obtained at the end of first-year study. This Context2-specific illustration implies that applicants with good school-level grades are likely to continue delivering high performance at the university. In contrary, those with poor school grades often fall into the same group in Context2, with a small portion being able to improve and become a member of Cluster1. In addition, Fig. 15 presents the aforementioned categories in accordance with other university-performance variables. Students in the so-called 'high' performance cluster or Cluster1 usually possess high ratios of grades A, B+, B and C+ than those in the other group. Another significant observation is that members of Cluster2 have a higher withdraw ratio (i.e., W Ratio), as compared to others belonging to the other cluster.

Besides previous findings that are related to cluster properties, the followings point out plausible associations between the acquired cluster models and the features that have not been used in the clustering process. In particular to Context1, Fig. 16 indicates

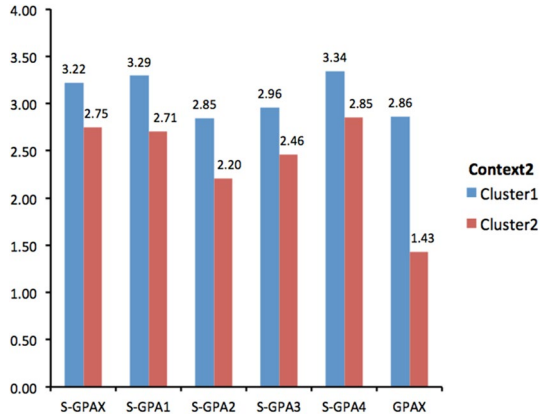

Fig. 14 Profile1 Context2. Profile representatives of two clusters (with 'high' and 'low' academic capability), with respect to Problem Context2

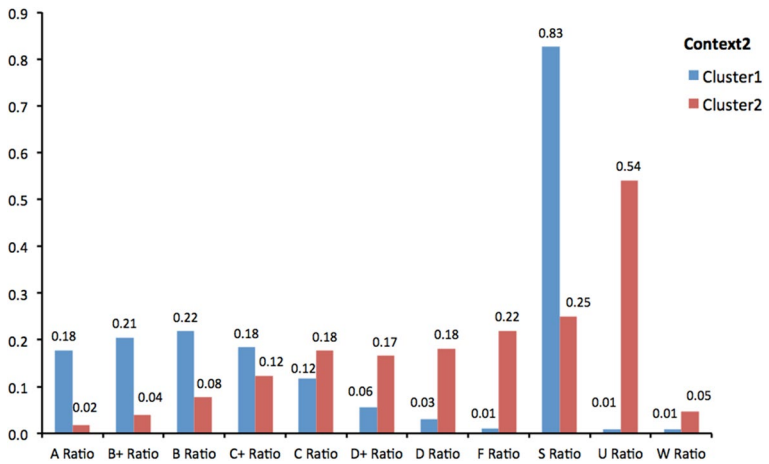

Fig. 15 Profile2 Context2. Profile representatives of two clusters (with 'high' and 'low' academic capability), with respect to Problem Context2

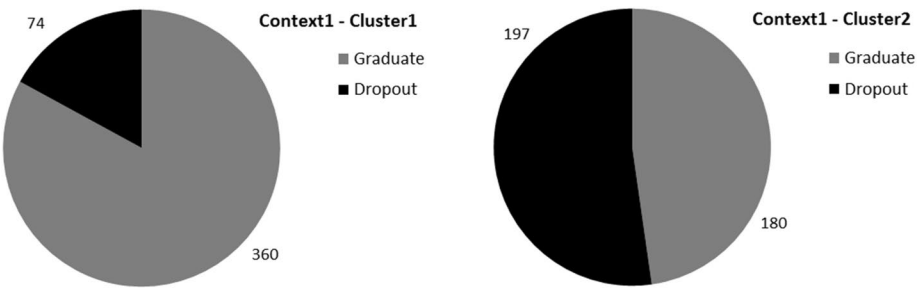

Fig. 16 Dropout Context1. Comparison of graduate and dropout cases, with respect to Problem Context1

that the probability of dropout is around 17.05% (i.e., $\frac{74}{74+360}$) for members of Cluster1 with high school-level performance. On the other hand, this has risen to 52.25% (i.e., $\frac{197}{197+180}$) for those in Cluster2. In other words, for a new applicant with moderate to

low school grade, his or her chance to dropout is three times greater than the case with good school-level profile. Given these figures, Admission division may form a working strategy such that the size of Cluster1 is much larger than the other, thus the amount of dropout can be reduced. This can be achieved by matching a new comer's profile against those of the two clusters, where a close mapping to Cluster1 representative is desired.

As for the analysis of second problem context, Fig. 17 shows that the probability of dropout is around 5.48% (i.e., $\frac{29}{29+500}$) for students in Cluster1 with high school-level and first-year performance. On the other hand, this has risen to 85.82% (i.e., $\frac{242}{242+40}$) for those in Cluster2. These statistical measures bring about an interesting relation between students' academic achievement and dropout possibility. It is rational to suggest that Cluster1 and Cluster2 represent graduate and dropout stereotypes. Therefore, by the end of the first year, Registrar division should identify at-risk students, whose profiles belonging to the second group. Then, an appropriate assistive measure such as course planning or degree transfer may be urgently executed to ensure student retention.

Following the preceding discussion, Figs. 18 and 19 illustrate cluster-wise categorization of male and female students, respectively. At the time of application, 60.08% (i.e., $\frac{152}{152+101}$) of male candidates may encounter dropout in the future, given their school grades. This is 40.32% (i.e., $\frac{225}{225+333}$) for female. With these observations, Admission division may pay a little more attention to male than female cases when it comes to

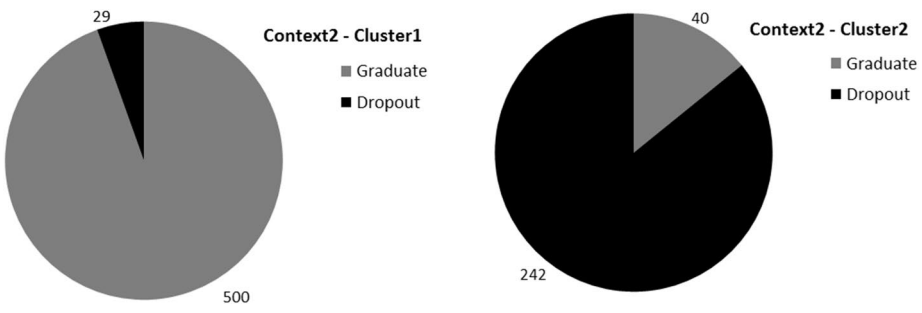

Fig. 17 Dropout Context2. Comparison of graduate and dropout cases, with respect to Problem Context1

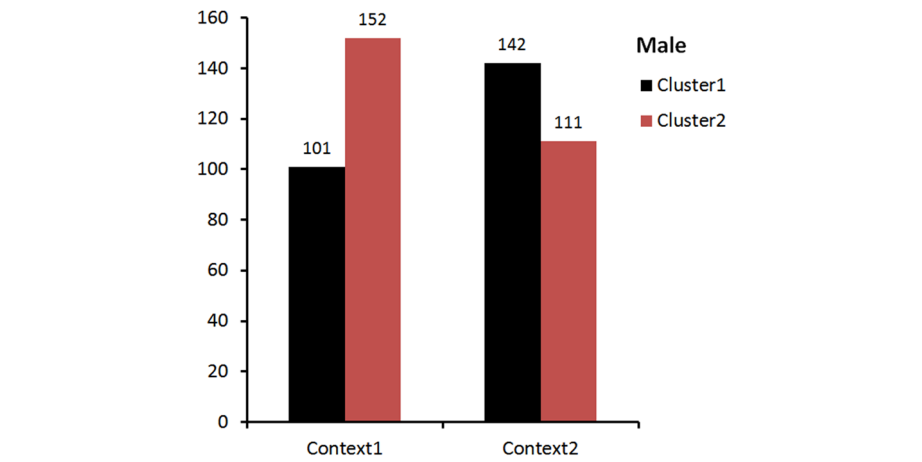

Fig. 18 Male. Categorization of male students

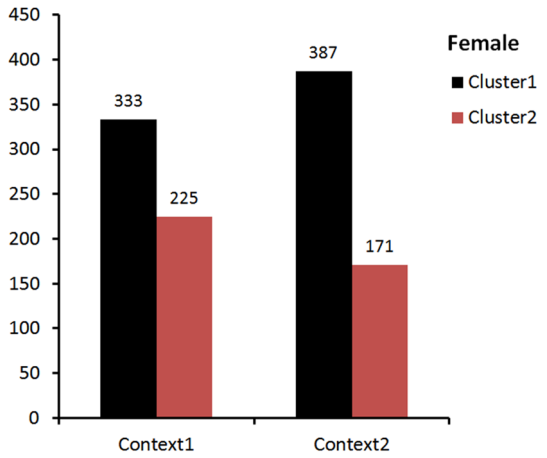

Fig. 19 Female. Categorization of female students

applicants with poor school-assessment records. Handling such a problem is rather difficult at this initial stage, as the forthcoming dropout event is still uncertain with some students in Clusters2 are able to produce an acceptable result in the first year. As shown in the aforementioned figures, the sizes of Cluster1 in the second context are larger than the correspondings in the first, for both male and female viewpoints. Nonetheless, Registrar division and academic advisor should work together to provide an effective consultation for those with poor first-year grading profile, especially male students.

For the association between student achievement and entry type, Fig. 20 suggests that RQ is an effective admission strategy with 19.37% (i.e., $\frac{43}{43+179}$) and 17.57% (i.e., $\frac{39}{39+183}$) of at risk of dropout at the time of application and after the first year at university. Therefore, Admission division should encourage the use of RQ, while minimize the number of applicants with low school grades (i.e., similar to those belonging to Cluster2). Specific to entry type of DA, Fig. 21 illustrates that 51.65% (i.e., $\frac{203}{190+203}$) of applicants may face a dropout problem later on, whilst it declines to to 32.32% (i.e., $\frac{127}{127+266}$) by the end of the freshman year. Of course, it is not as effective as RQ, but it contributes for a large

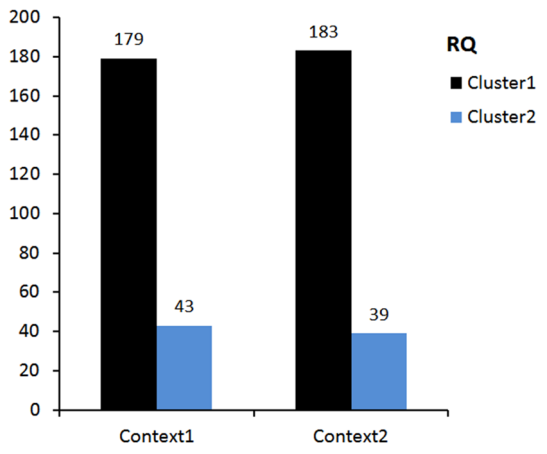

Fig. 20 RQ. Categorization of students with entry type of RQ

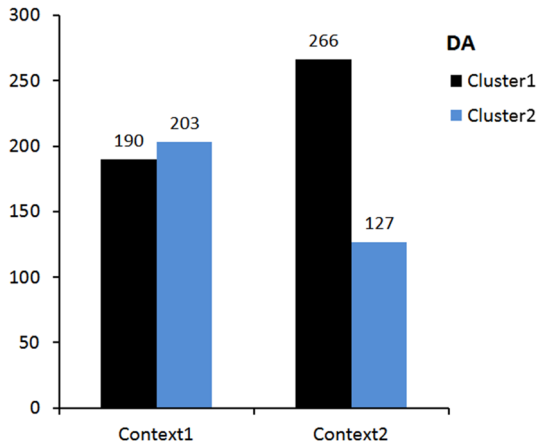

Fig. 21 DA. Categorization of students with entry type of DA

proportion of MFU students, around 48.46% (i.e., $\frac{393}{811}$). Hence, Admission division still needs to execute DA strategy to ensure the university income, with definite precautions. These includes (i) promoting the selection of students with good school grades, and (ii) providing academic help to freshmen with moderate-to-poor school results as soon as possible, since there is a tendency for improvement (sizes of Cluster1 increase from 190 in Context1 to 266 in the other).

In addition, Figs. 22 and 23 provide statistics with respect to the entry types of ADA and CA, respectively. Specific to these admission categories, most applicants are with moderate or low school grades, with little room for improvement during the first year study. A possible solution is to guide those students through degree selection and year planning. All the aforementioned relations may well be useful to increase retention and students' achievement, which will turn out to be advantageous for several parties, including students, families and the university.

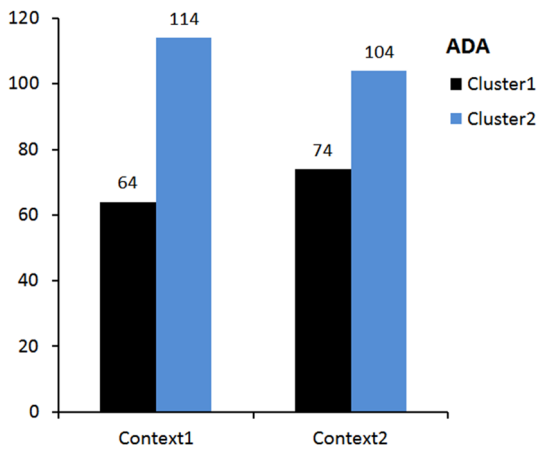

Fig. 22 ADA. Categorization of students with entry type of ADA

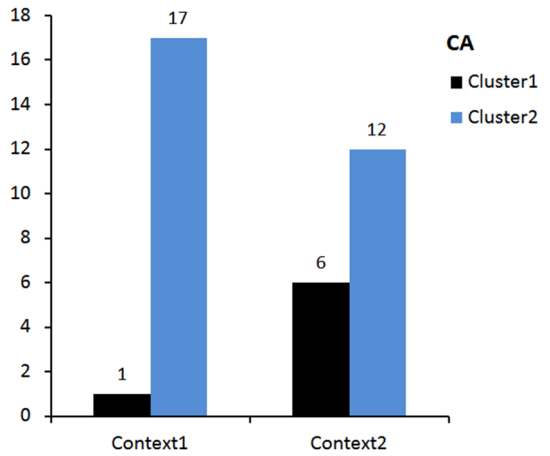

Fig. 23 CA. Categorization of students with entry type of CA

Conclusion

This review has presented the clustering approach to generate a descriptive model from educational data. The underlying methodology aims to discover knowledge, interesting patterns and relations that contribute to student dropout. This so-called retention problem has been recognized as one of the major difficulties commonly encountered by any university. Leaving it unsolved may negatively affect several parties, such as students, parents as well as the university, in terms of financial support and reputation. The paper kicks off with a brief revision of the clustering technique that is used to analyze the desired data collection, and examples of quality measures for the assessment of analytical results. These form ontology of technical tools for those who are interested in EDM and data clustering in general. To consolidate the process of creating and interpreting a descriptive model using the aforementioned methods, an illustrative case study of MFU has been explored and discussed.

The working example demonstrates a sequence of processes matching those of the conventional data mining framework. This begins with problem definition and data acquisition, where the former is conducted in conjunction with staffs from Admission and Registrar divisions, while the latter is achieved by extracting relevant data from MFU MIS. Having obtained the initial collection of student data, it is pre-processed such that errors and missing values are resolved. In particular, it is arranged for the following investigation using descriptive type of data mining model. This is studied with respect to two problem contexts of (i) before starting the first year where only demographic and school-academic details are available, and (ii) after the first year where initial university performance is known in addition to those in (i).

Context-specific data is analyzed using a clustering procedure to reveal natural student groups. A filter approach to soft subspace clustering, namely R-KM, has been exploited for this task. It reveals two clusters at the time of university admission; one corresponds to applicants with good school grades, while the other represents those with moderate to poor grading profiles. Likewise, two student clusters have been disclosed when applying the aforementioned technique to the set of data prepared for the second problem context. In a nutshell, those students with good school-academic background continue to do

well in the first year. The majority of applicants with low school performance may face dropout after the first year, with some being able to adapt to university academic system and survive. Also, entry types implemented by MFU are not equally effective, where RQ appears to be the most successful while others should be used with constraints. The developed models and research findings may be highly useful as a working guideline to formulate an effective admission and consultant strategies. As a result, this can yield the level of student retention, hence optimizing tuition fees and government funding, student achievement, university reputation, and satisfaction of all the parties involved.

To strengthen this line of research, a number of important directions for future work can be highlighted. As suggested by many research works on the subject of student dropout [1], family background, financial support and university-event participation may provide complementary interpretation of student achievement. To some students, academic capability is a major barrier to success, while social and financial factors can be crucial to others. Unfortunately, these attributes may not been properly recorded, thus prohibiting the corresponding investigation. However, through the cooperation with responsible divisions, a better understanding of non-academic motives towards student performance can be acquired with the aforementioned variables being included.

Authors' contributions

NIO collected and pre-processed the data. NIO and TB designed the research as well as analyzed the result. NIO was the lead writer of the paper. Both authors read and approved the final manuscript.

Author details

[1] School of information Technology, Mae Fah Luang University, 333 Moo 1, Ta-sud, Muang, Chiang Rai 57100, Thailand. [2] Department of Mathematics and Computer Science, Navaminda Kasatriyadhiraj Royal Air Force Academy, 171/1 Saimai, Bangkok 10220, Thailand.

Acknowledgements

Natthakan Iam-On—main contributor.

Competing interests

The authors declare that they have no competing interests.

Received: 5 September 2016 Accepted: 8 December 2016
Published online: 02 January 2017

References

1. Romero C, Ventura S (2010) Educational data mining: a review of the state-of-the-art. IEEE Trans Syst Man Cybern Part C 40:601–618
2. Bala M, Ojha DB (2012) Study of applications of data mining techniques in education. Int J Res Sci Technol 1:1–10
3. Koedinger K, Cunningham K, Skogsholm A, Leber B (2008) An open repository and analysis tools for fine-grained, longitudinal learner data. In: Proceedings of first international conference on educational data mining, pp. 157–166
4. Mostow J, Beck J (2006) Some useful tactics to modify, map and mine data from intelligent tutors. Nat Lang Eng 12:195–208
5. Baepler P, Murdoch CJ (2010) Academic analytics and data mining in higher education. Int J Schol Teach Learn 4(2):1–9
6. Romero C, Ventura S (2013) Data mining in education. Wiley Interdiscip Rev Data Min Knowl Discov 3(1):12–27
7. Baker R, Yacef K (2009) The state of educational data mining in 2009: a review and future visions. J Educ Data Min 1(1):3–17
8. Lin SH (2012) Data mining for student retention management. J Comput Sci Coll 27(4):92–99
9. Kotsiantis S, Pierrakeas C, Pintelas P (2004) Prediction of student's performance in distance learning using machine learning techniques. Appl Artif Intell 18(5):411–426
10. Erdogan SZ, Timor M (2005) A data mining application in a student database. J Aeronaut Space Technol 2(2):53–57
11. Sung-Hyuk C, Tappert C (2009) Constructing binary decision trees using genetic algorithms. J Pattern Recognition Res 1:1–13
12. Kabra RR, Bichkar RS (2011) Performance prediction of engineering students using decision trees. Int J Comput Appl 36(11):8–12
13. Antons C, Maltz E (2006) Expanding the role of institutional research at small private universities: a case study in enrollment management using data mining. New Dir Inst Res 131:69–81

14. Ramaswami M, Bhaskaran R (2010) A CHAID based performance prediction model in educational data mining. Int J Comput Sci 7(1):10–18
15. Yu C, Gangi SD, Jannasch-Pennell A, Kaprolet C (2010) A data mining approach for identifying predictors of student retention from sophomore to junior year. J Data Sci 8:307–325
16. Subyam S (2009) 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
17. Sittichai R (2012) Why are there dropouts among university students? Experiences in a thai university. Int J Educ Dev 32:283–289
18. Kongsakun K, Fung CC (2012) Neural network modeling for an intelligent recommendation system supporting SRM for Universities in Thailand. WSEAS Trans Comput 11(2):34–44
19. Scott DM, Spielmans GI, Julka DC (2004) Predictors of academic achievement and retention among college freshmen: a longitudinal study. Coll Stud J 38(1):66–80
20. Delen D (2011) Predicting student attrition with data mining methods. J Coll Stud Retent 13(1):17–35
21. Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng 16(11):1370–1386
22. He Q, Wang J, Zhang Y, Tang Y, Zhang Y (2009) Cluster analysis on symptoms and signs of traditional Chinese medicine in 815 patients with unstable angina. In: Proceedings of international conference on fuzzy systems and knowledge discovery, pp 435–439
23. Henry DB, Tolan PH, Gorman-Smith D (2005) Cluster analysis in family psychology research. J Fam Psychol 19(1):121–132
24. Sheppard AG (1996) The sequence of factor analysis and cluster analysis: differences in segmentation and dimensionality through the use of raw and factor scores. Tour Anal 1:49–57
25. Wu RC, Chen RS, Chang CC, Chen JY (2005) Data mining application in customer relationship management of credit card business. In: Proceedings of international conference on computer software and applications, pp 39–40
26. Kim K, Ahn H (2008) A recommender system using GA K-means clustering in an online shopping market. Expert Syst Appl 34:1200–1209
27. Bredel M, Bredel C, Juric D, Harsh G, Vogel H, Recht L, Sikic B (2005) Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. Cancer Res 65(19):8679–8689
28. Kim E, Kim S, Ashlock D, Nam D (2009) MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. BMC Bioinform 10:260
29. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron J, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100(14):8418–8423
30. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
31. Ahmad A, Dey L (2007) A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl Eng 63(2):503–527
32. Huang Z (1997) Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the first Pacific Asia knowledge discovery and data mining conference, pp 21–34
33. Dudoit S, Fridyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 3(7):0036
34. Boongoen T, Shen Q (2010) Nearest-neighbour guided evaluation of data reliability and its applications. IEEE Trans Syst Man Cybern Part B 40(6):1622–1633
35. Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66:846–850
36. Iam-On N, Boongoen T, Garrett S (2010) LCE: a link-based cluster ensemble method for improved gene expression data analysis. Bioinformatics 26(12):1513–1519
37. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley-Interscience, New York, p 153
38. Xue H, Chen S, Yang Q (2009) Discriminatively regularized least-squares classification. Pattern Recognit 42(1):93–104
39. McQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, pp 281–297
40. Boongoen T, Shang C, Iam-On N, Shen Q (2011) Extending data reliability measure to a filter approach for soft subspace clustering. IEEE Trans Syst Man Cybern Part B 41(6):1705–1714
41. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(2):224–227
42. Dunn JC (1974) Well separated clusters and optimal fuzzy partitions. J Cybern 4:95–104