Australasian Journal of Educational Technology

2011, 27(3), 481-498

Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation

Chih-Fong Tsai
National Central University, Taiwan
Ching-Tzu Tsai
National Chung Cheng University, Taiwan
Chia-Sheng Hung
Nanhua University, Taiwan
Po-Sen Hwang
National Chung Cheng University, Taiwan

Enabling undergraduate students to develop basic computing skills is an important issue in higher education. As a result, some universities have developed computer proficiency tests, which aim to assess students' computer literacy. Generally, students are required to pass such tests in order to prove that they have a certain level of computer literacy for successful graduation. This paper applies data mining techniques to make predictions about students who are going to take the computer proficiency test and fail. A national university in Taiwan is considered as the case study. Three different clustering techniques are used individually to cluster students into different groups, which are *k*-means, self-organising maps (SOM), and two-step clustering (i.e. BIRCH). After the best clustering result is found, the decision tree algorithm is used to extract useful rules from each of the identified clusters. These rules can be used to warn or counsel students who have higher probability of failing the test. The results can help the university identify a number of student groups who need to pay much more attention to preparing for the test, which is likely to help conserve resources. Furthermore, this study can be regarded as a guideline for future developments in assessing students' English literacy, as this is also an important graduation requirement in many universities.

1. Introduction

In higher education, computer literacy is the most important and basic skill in the knowledge-based economy, and a country's competitiveness may be directly affected by undergraduate literacy (Hannon, 2001; O'Hanlon, 2002; Tesch, Murphy & Crable, 2006). Therefore, governments and educational institutions in many countries have paid much attention to enhancing the computer literacy of undergraduate students.

Nowadays, in order to promote basic computer skills of undergraduate students, some universities have developed specific certification programs to examine students' computer proficiency. In particular, students may be required to pass this kind of test as a requirement for graduation, as occurs for example with some national universities in Taiwan.

However, not all students will pass such tests at the first attempt, and they need to retake it until they fulfill this graduation requirement. To take the case university considered in this paper (3.1 The case university), only about 60% of the students from the 2002 to 2005 academic years passed this test. As a result, the university needed additional resources to take care of these failed students, and they also needed to spend extra time to prepare for retaking the test.

These considerations lead to the purpose reported in this paper, to develop an early warning mechanism for students, by analysing the characteristics of passed and failed students, by applying data mining techniques. In general, data mining focuses on the discovery and extraction of latent knowledge in a database (Romero & Ventura, 2007; Shih, Chiang, Lai & Hu, 2009). Many studies have applied data mining techniques to understand learners' behavioural patterns and usage rules, and to improve interactive educational systems (Guo & Zhang, 2009; Guruler et al., 2010; Hamdi, 2007; Lee, Chen et al., 2009; Romero & Ventura, 2006, 2007; Romero et al., 2008; Romero et al., 2009; Shih et al., 2009; Wang et al., 2009).

Specifically, clustering analysis, one type of data mining technique aimed at searching for hidden patterns, has been widely used in many fields (Jain, 2009; Paterlini & Krink, 2006; Romero et al., 2008; Yang et al., 2009; Tang & McCalla, 2005). Clustering analysis is suitable for the research aim of this paper because it can find a number of student groups (i.e. clusters) with similar characteristics, e.g. the place of a student's senior high school. Then, based on analysing the relationship between the characteristics and passed/failed student groups, some decision rules can be extracted from these groups to 'predict' whether the students who are going to take the test at the first time will pass or fail.

This paper is organised as follows. Section 2 describes the research background about previous works related to the computer proficiency test, and the clustering techniques used in this paper. Section 3 presents the research methodology, including the collected dataset, the procedure for developing different clustering models for comparisons, etc. Section 4 shows the experimental results and the conclusion is provided in Section 5.

2. Literature review

2.1 Computer proficiency test

Since computer literacy has become an important and basic skill for various jobs today, this leads to the necessity of providing computer related curriculums for undergraduate students (Verhey, 1999). Universities must help students to become computer literate to take advantage of the best ways to use computers. For example, students are required to learn several computer applications, such as Microsoft *Word, PowerPoint, Excel*, etc. O'Hanlon (2002) argued that students cannot learn research skills if they are unable to handle these kinds of tools effectively. Therefore, students who are seeking to graduate should be certified for some level of computer skills or capabilities. On the other hand, Hannon (2001) recommended that colleges need to develop some related voluntary certification programs. Hence, in order to promote students' computer skills for the international competitive advantage, many universities have subsequently introduced the computer proficiency test as one requirement for successful graduation, especially in Taiwan.

The computer proficiency test is based on assessing students' ability (usually the entry-level computer skill) to perform some specific tasks using computer applications. Although different universities executing computer proficiency tests for the requirement of undergraduate graduation have slightly different regulations, the goal is the same, which is to promote students' computer skills. Currently, there are 12 universities in Taiwan which require students to pass their computer proficiency tests before graduation.

2.2 Clustering analysis

Clustering analysis is a common unsupervised learning technique. Its aim is to group objects into different categories. That is, a collection of data objects that are similar to one another are grouped into the same cluster and the objects that are dissimilar are grouped into other clusters (Han & Kamber, 2006; Hosseini et al., 2010; Kao et al., 2008; Qiu, 2010; Yang et al., 2009). It is an important technique in data mining to analyse high-dimensional data and large scale databases.

Clustering algorithms can be classified into hierarchical and non-hierarchical algorithms (Han & Kamber, 2006). The hierarchical procedure produces a tree-like structure, which is able to see the relationship among entities. The hierarchical clustering procedure can be agglomerative or divisive. On the other hand, non-hierarchical methods do not possess tree-like structures but assign some cluster seeds to central places, also called *k*-means clustering. There are three methods to assign an object to a group, namely the sequential threshold, parallel threshold and optimisation partitioning procedures.

K-means
The *k*-means algorithm is one of the best known and simplest clustering algorithms. It was proposed over 50 years ago and still widely used (Hosseini et al., 2010; Jain, 2009; Yang et al., 2009). This is due to its ease of implementation, simplicity, and superior feasibility and efficiency in dealing with a large amount of data. However, it is sensitive to initialisation and is easily trapped in local optima (Hosseini et al., 2010; Kanungo et al., 2002; Mingoti & Lima, 2006; San et al., 2004; Yang et al., 2009).

In addition, the main shortcoming of the *k*-means algorithm is that it depends heavily on the initial choice of the cluster centres, which reduces its convergence reliability and efficiency (Kao et al., 2008; Mingoti & Lima, 2006; Qiu, 2010; Yang et al., 2009).

The *k*-means algorithm is a non-parametric approach that aims to partition objects into *k* different clusters by minimising the distances between objects and cluster centers (Qiu, 2010). The *k*-means algorithm contains the following steps:

1. Select initial centers of the *k* clusters,
2. Assign each object to the group that is closest to the centroid,
3. Compute new cluster centers as the centroids of the clusters,
4. Repeat Steps 2 and 3 until the centroids no longer move.

Self-organising maps
The self-organising map (SOM) or self-organising feature map network (SOFM) was proposed by Kohonen (1982, 2001). SOM is an unsupervised neural network consisting of an input layer and the Kohonen layer. It is usually designed as a two-dimensional

arrangement of neurons that maps an n-dimensional input to a two-dimensional map (Budayan et al., 2009; Mingoti & Lima, 2006). Particularly, SOM provides a topological structure imposed on the nodes in the network, and preserves neighborhood relations from the input space to the clusters (Kohonen, 1989; 2002). The learning algorithm of SOM is described as follows (David & Yong, 2007):

1. Initialise the map: this stage aims to initialise reference vectors, set up the parameters of the algorithm, such as the distance of neighborhoods and the learning rate;
2. Determine the winning node: select the best matching node that minimises the distance between each input vectors by the Euclidean distance;
3. Update reference vectors: updating reference vectors and its neighborhood nodes based on the learning criterion;
4. Iteration: iterate Steps 2 and 3 until the solution can be regarded as steady.

Two-step clustering: BIRCH
The BIRCH (balanced iterative reducing and clustering using hierarchies) algorithm contains two main steps and hence is known as a *two-step clustering* (Markov & Larose, 2007). BIRCH is an integrated hierarchical clustering method (Han & Kamber, 2006). It introduces the concepts of clustering feature (CF) and clustering feature tree (CF tree), and these structures help achieve good speed and scalability for very large datasets. A CF is a triple that stores the information about sub-clusters of objects; a CF tree is a height balanced tree used to store the clustering features (Han & Kamber, 2006).

Unlike k-means and SOM, the BIRCH clustering algorithm represents a desirable exploratory tool, for which the number of clusters does not need to be specified at the beginning (Markov & Larose, 2007). BIRCH performs the following steps:

1. Load data into memory by building a CF tree;
2. Condense the initial CF tree into a desirable range by building a smaller CF tree (optional);
3. Perform global clustering;
4. Perform cluster refining (optional).

2.3 Related work

In recent years, some researchers have used various data mining techniques to help instructors and administrators to improve e-learning systems (Guo & Zhang, 2009; Guruler et al., 2010; Hamdi, 2007; Lee et al., 2009; Romero & Ventura, 2006, 2007; Romero et al., 2008; Romero et al., 2009; Shih et al., 2009; Wang et al., 2009). For example, Romero and Ventura (2007) stated that educational data mining has been an emerging discipline, and they surveyed the application of data mining to some different types of educational systems. In addition, they illustrated how data mining techniques could be applied to some educational problems.

Guruler et al. (2010) employed data mining techniques to explore some factors having an impact on the success of university students. Similarly, Lee et al. (2009) analysed some important factors which can influence the preferences of learners from diverse backgrounds. For web based systems, Hamdi (2007) presented a method for extracting and inferring useful knowledge for student learning by web mining techniques. Romero et al. (2008) focused on mining e-learning data for online instructors and e-

learning administrators. Further, a hybrid data mining technique is proposed by Shih et al. (2009) to evaluate the important characteristics of study strategy scales and their inter-relationships for freshmen students in a web-based self-assessment system.

Romero et al. (2009) proposed the architecture of a recommender system that utilises web usage mining to recommend the links to visit next in an adaptive, web-based educational system in order to help the instructor to carry out the web mining process. For teaching and learning content, Wang et al. (2009) employed a decision tree algorithm to discover the most adaptive learning sequences based on students' profiles for a particular teaching content. Guo and Zhang (2009) presented a method for representing and extracting a dynamic learning process and learning patterns to support students' deep learning, efficient tutoring and collaboration in a web-based learning environment.

In summary, numerous examples of data mining techniques have been applied or developed in order to help various educational problems, such as understanding factors affecting students' learning outcomes, teaching contents, etc. However, very few consider predicting students' performances upon taking some required test, such as the computer proficiency test considered in this paper. Therefore, we introduce a new problem in educational data mining, developing a decision support system to warn those students who have high probability of failing a graduation requirement test.

3. Research methodology

3.1 The case university

The case university considered in this paper is the National Chung Cheng University (CCU) in Taiwan. It was established in 1989, and consists of 7 colleges, 35 departments and 47 research institutes. In total, there are approximately 12,000 students (including undergraduate, masters and doctoral programs). In order to enhance the computer literacy of undergraduate students, since 2002 CCU has required undergraduate students to pass the computer proficiency test before graduation.

Students may take this test anytime, with an e-learning platform provided by the computer center enabling them to take it online. The test contains 'discipline based' and 'skill based' questions and the minimum score for passing the test is 70 out of 100. The discipline based test includes five types of questions, such as introduction to computer science, official editing, electronic spreadsheet, presentation software, and introduction to Internet. The skill based test focuses on the understanding of using computer applications, which are Microsoft *Word, Excel*, and *PowerPoint*. There is not a required order, but most students took the discipline based test first.

3.2 Data collection and pre-processing

The data collected from the computer center of CCU are from 2003 to 2005 inclusive. In total, there are 4513 samples in this dataset, i.e 4513 undergraduate students who have taken the computer proficiency test. Table 1 shows the College distribution information for these students and their pass rates.

Table 1: Summary statistics of students

College	Number	Pass rate for the discipline based test	Pass rate for the skill based test
College of Humanities	520 (11.5%)	54.62%	61.54%
College of Sciences	575 (12.7%)	54.78%	56.17%
College of Social Sciences	683 (15.1%)	56.08%	61.49%
College of Engineering	965 (21.4%)	68.19%	62.49%
College of Management	1165 (25.8%)	61.29%	67.12%
College of Law	392 (8.7%)	53.32%	60.20%
College of Education	213 (4.7%)	61.50%	55.40%
Total/Average	4513	58.54%	60.63%

After integrating different database tables containing students' information for a single dataset, data transformation is performed. Table 2 shows all of the input variables in the processed dataset.

Table 2: Description of the input variables

Items	Type	Description
Student number	Category	Student identification
Graduated department	Category	By department code, for example: physics = 2204, business administration = 5204, electrical engineering = 4154, etc.
Graduated class	Category	1 = class A; 2 = class B (note that for some departments, there is only one class.)
Gender	Category	1 = Male; 2 = Female
Blood type	Category	1 = A type; 2 = B type; 3 = AB type; 4 = O type
Date of birth	Category	(dd/mm/yy)
Place of birth	Category	By postcode, for example: 100 = Taipei city, 200 = Keelung city, 207 = Taipei county, etc.
Academic year of admission	Category	2003/2004/2005
Dept. name of admission	Category	By department code
Nationality	Category	Definition by country code, for example: 1 = United States of America, 27 = South Africa, 54 = Argentina, 60 = Malaysia, 62 = Indonesia, 66 = Thailand, etc.
College name	Category	By college code, for example: 1000 = humanities, 2000 = sciences, 3000 = social sciences, 4000 = engineering, 5000 = management, 6000 = law, and 7000 = education
Name of senior high school	Category	By school code in Taiwan (including overseas), for example: National Lo-Tung senior high school = 040004.
Place of senior high school	Category	By place code, for example: Taipei city = 10; Chung-hua county = 22, etc.
Admission status	Category	1 = domestic student; 2 = overseas student; 3 = aborigine student; 4 = foreign student; 5 = China student
Admission channel	Category	1 = general student; 2 = dispense student; 3 = admission by application; 4 = admission by recommendation and screening; 5 = foreign student; 6 = test of recommendations; 7 = athletic scholarship; 8 = transfer student; 9 = handicapped student
Other items	Category	1 = none; 2 = in-service student (by status of admission exam); 3 = demobilised soldier; 4 = student superior in athletic accomplishments; 5 = autism; 6 = mild limb handicapped student; 7 = moderate limb handicapped student; 8 = severe limb handicapped student; 9 = visual impairment student; 10 = hearing impaired student

Status in school	Category	1 = graduate; 2 = drop out of school
Times discipline based test taken	Numeric	Total frequency of taking the discipline based test before graduation
Times skill based test taken	Numeric	Total frequency of taking the skills based test before graduation
Score for discip-line based test	Numeric	The score of the discipline based test taken for the first time
Score for the skill based test	Numeric	The score of the skill-based test taken for the first time

3.3 Experimental procedure

After the dataset is processed for the experiment, it is divided into the training and testing datasets, which are based on the data from 2003 to 2004 and the data of 2005 respectively. Then, the training set is used to develop three clustering models based on two-step, *k*-means, and SOM clustering algorithms respectively and the testing set is for evaluating their clustering performances. Next, the best clustering result of a specific clustering model will be further analysed. In particular, the relation between students' characteristics and the computer proficiency test in each cluster will be examined. The analysis result is able to assist students who have similar characteristics as the former failed students to pay increased attention to preparing for the test.

The experimental procedure contains three parts. The first one is to develop the clustering models. Then, the clustering results from each of the three clustering models are analysed in order to find out a number of different clusters with different features or characteristics. Finally, a C5.0 decision tree algorithm is applied to extract important decision rules from each of the clusters.

Development of clustering models
Given a training set, the two-step, *k*-means, and SOM clustering models can be constructed respectively. The cluster number for two-step and *k*-means is set from 3 to 6 groups. For SOM, 2×2, 3×1, 3×2, 3×3, 4×2, 4×3, 4×4, 5×1, 5×2, 5×3, 5×4, and 5×5 SOMs are constructed. After these clustering algorithms are constructed, the testing set is used to test these clustering results. Figure 1 shows the procedure of training and testing these clustering algorithms.

Analysis of clustering results
Once the clustering results are obtained, this stage compares the significant characteristics of each cluster produced by each of the three clustering algorithms, and aggregates all of the significant characteristics. Then, the best clustering result of each clustering algorithm can be identified by cross-comparison analyses. The purpose of this stage is to discover the most optimal clustering results (i.e. groups) to explain different students having different characteristics. Figure 2 shows the procedure for analysing the clustering results.

Decision rules
When the best clustering result is identified, the final stage is to use all of the data samples composed of the training and testing sets to feed into the best clustering algorithm, in order to obtain the 'best' student groups. This is because if we only take the best clustering result over two years of the training data or one year of testing data for extracting useful decision rules, it might be insufficient.

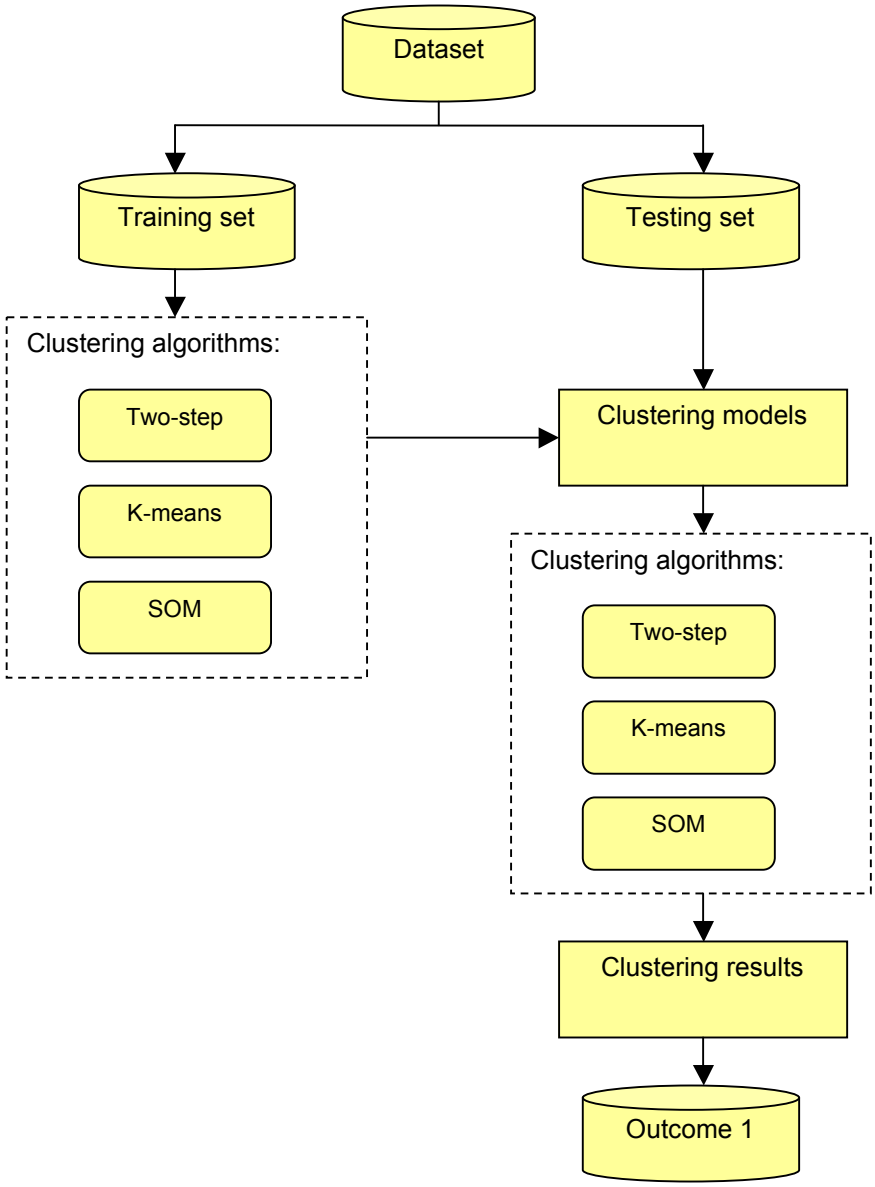

Figure 1: The procedure of training and testing the clustering algorithms

After the clustering results by the whole dataset are obtained, i.e. each cluster contains a number of students, the C5.0 decision tree algorithm is used to find out the decision rules for each student groups. The primary stages for constructing a decision tree are :

1. Divide the original data into the training set and testing set.
2. Open the training data and input the data into the decision tree algorithm (e.g. the root node).

3. Build the decision tree by using the training set, and choose what character to be the classification base through the information theory on each leaf node.
4. Prune the decision tree through applying the testing data until only one node remains in each category.
5. Repeat these stages until all internal nodes become leaf nodes.

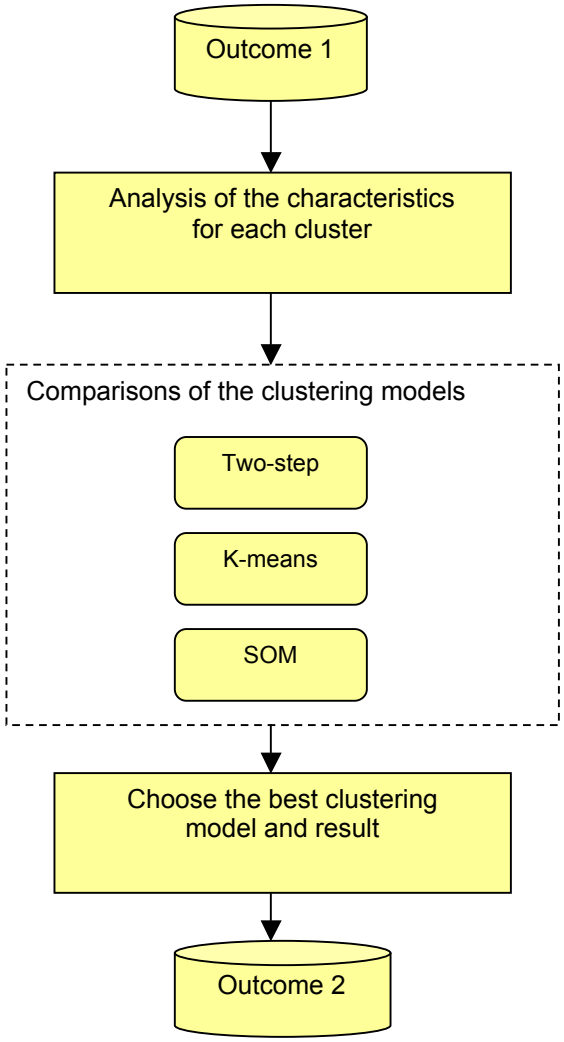

Figure 2: The procedure for analysing the clustering results

This strategy by using decision trees to identify useful rules from some clustering results has been considered in literature, such as Tsai et al. (2009). Figure 3 shows the procedure for extracting decision rules from the clustering results.

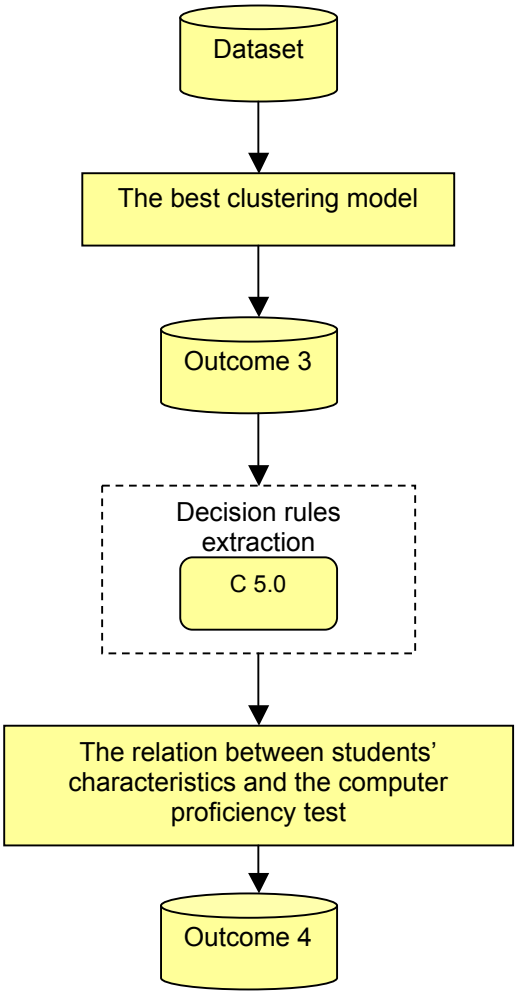

Figure 3: The procedure for identifying decision rules from the clustering results

4. Results

4.1 Clustering results

K-means
Table 3 shows the clustering results of k-means (k = 3 to 6). Particularly, the number and percentage of observations in each cluster is listed. In order to find out the representative attributes (i.e. the input variables in Table 2) for analysing the clustering results, the attribute which has more than 80% observations in a specific cluster is defined as a representative attribute. As a result, Table 4 shows five significant attributes of the clustering results, which are Graduated class, Gender, College name, Status in school, and Score for the skill based test.

Table 3: Clustering results of *k*-means

k-means		Training data		Testing data	
		Observations	Percentage(%)	Observations	Percentage(%)
$k = 3$	C1	1199	41.5	532	39.88
	C2	1189	41.2	580	43.48
	C3	501	17.3	222	16.64
$k = 4$	C1	1199	41.5	532	39.88
	C2	628	21.74	302	22.64
	C3	739	25.58	362	27.14
	C4	323	11.18	138	10.34
$k = 5$	C1	1199	41.5	532	39.88
	C2	628	21.74	302	22.64
	C3	178	6.16	84	6.30
	C4	323	11.18	138	10.34
	C5	561	19.42	278	20.84
$k = 6$	C1	693	23.99	296	22.19
	C2	79	2.73	37	2.77
	C3	174	6.02	84	6.30
	C4	318	11.01	135	10.12
	C5	828	28.66	419	31.41
	C6	797	27.59	363	27.21

Table 4: The five significant attributes

Attribute	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Graduated class	V	V	V	V
Gender	V	V	V	V
College name	V	V	V	V
Status in school				V
Score for the skill based test				V

Self organising maps

Table 5 shows the clustering results of SOM. Note that only the results of 4×2 and 5×2 SOM are present here since significant attributes can only be identified from them. In addition, Table 6 shows seven significant attributes of the clustering results, which are Graduated class, Gender, Place of birth, Department name of admission, College name, Place of senior high school, and Times skill based test taken.

Table 5: Clustering results of SOM

SOM		Training data		Testing data	
		Observations	Percentage(%)	Observations	Percentage(%)
4×2	C1	1199	41.5	532	39.88
	C2	323	11.19	138	10.34
	C3	178	6.16	84	6.3
	C4	1189	41.16	580	43.48
5×2	C1	1199	41.5	532	39.88
	C2	501	17.34	222	16.64
	C3	1189	41.16	580	43.48

Table 6: The seven significant attributes

Attribute	4×2 SOM	5×2 SOM
Graduated class	V	V
Gender	V	V
Place of birth	V	V
Department name of admission	V	V
College name	V	V
Place of senior high school		V
Times skill based test taken	V	V

BIRCH

For the BIRCH clustering algorithm, Table 7 shows its clustering results and Table 8 lists the eleven significant attributes identified from the clusters respectively.

Table 7: Clustering results of BIRCH

k-means		Training data		Testing data	
		Observations	Percentage(%)	Observations	Percentage(%)
$k = 3$	C1	314	10.87	206	15.44
	C2	1513	52.37	630	47.23
	C3	1062	36.76	498	37.33
$k = 4$	C1	188	6.51	96	7.2
	C2	145	5.02	142	10.64
	C3	1508	52.2	614	46.03
	C4	1048	36.28	482	36.13
$k = 5$	C1	188	6.51	96	7.2
	C2	142	4.92	140	10.49
	C3	1063	36.79	415	31.11
	C4	1045	36.17	492	36.88
	C5	451	15.61	191	14.32
$k = 6$	C1	188	6.51	96	7.2
	C2	139	4.81	123	9.22
	C3	843	29.18	365	27.36
	C4	493	17.06	149	11.17
	C5	775	26.83	410	30.73
	C6	451	15.61	191	14.32

Table 8: The eleven significant attributes

Attribute	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Graduated class	V	V	V	V
Academic year of admission	V	V	V	V
Nationality	V	V	V	V
College name	V	V	V	V
Name of senior high school	V	V	V	V
Place of senior high school	V	V	V	V
Admission status	V	V	V	V
Admission channel	V	V	V	V
Other items	V	V	V	V
Status in school	V			
Times skill based test taken				V

Clusters and their attribute characteristics

Regarding the testing results of k-means, SOM, and BIRCH, BIRCH with $k = 5$ provides the best clustering result. That is, there is the smallest difference between the training and testing results in terms of the data distribution in each cluster. Then, the whole dataset is fed into the BIRCH clustering algorithm with five clusters (c.f. Figure 3). Table 9 shows the clustering information of the five clusters.

Table 9: Clustering information of BIRCH ($k = 5$)

Cluster ID	Attributes	Distributions
C1	• Graduated class	• 84.4% for class A
	• Academic year of admission	• 97.9% for 2004
	• Nationality	• 100% for non-Taiwanese
	• Name of senior high school	• 100% for 060198 (School ID)
	• Place of senior high school	• 100% for non-Taiwan
	• Admission status	• 100% for non-domestic students
	• Admission channel	• 100% for general students
	• Status in school	• 5.1% for drop out of school
C2	• Graduated class	• 88.6% for class A
	• Academic year of admission	• 97.7% for 2005
	• Nationality	• 100% for Taiwanese
	• Name of senior high school	• 82.4% for non-ID school
	• Place of senior high school	• 100% for non-place code
	• Admission status	• 100% for domestic students
	• Admission channel	• 97.7% for transfer students
	• Other items	• 100% for special students
	• Status in school	• 98.6% for graduate
C3	• Graduated class	• 100% for class A
	• Academic year of admission	• 100% for 2004
	• Nationality	• 100% for Taiwanese
	• College name	• 84% for the College of Law, 88.3% for the College of Education
	• Admission status	• 100% for domestic students
	• Admission channel	• 87.5% for general students
	• Status in school	• 100% for graduate
C4	• Graduated class	• 100% for class A
	• Academic year of admission	• 100% for 2004
	• Nationality	• 100% for Taiwanese
	• College name	• 83.1% for the College of Sciences, 81% for the College of Social Sciences
	• Admission status	• 100% for domestic students
	• Admission channel	• 80.5% for general students
	• Status in school	• 100% for graduate
C5	• Graduated class	• 100% for class B
	• Academic year of admission	• 100% for 2004
	• Nationality	• 100% for Taiwanese
	• Admission status	• 100% for domestic students
	• Admission channel	• 100% for general students
	• Status in school	• 100% for graduate

Regarding these clustering results, we can assign a group name for each of the five clusters. For clusters 1 to 5, they are the non-Taiwanese group, transfer student group, Colleges of Law and Education group, Colleges of Sciences and Social Sciences group, and class B group.

4.2 Extraction of decision rules

After identifying the best clustering results and their characteristics, the next step is to extract decision rules from each of these five clusters. That is, the extracted rules can be used to predict students who are going to take the computer proficiency test and have higher probability of failing the test. Therefore, the decision tree algorithm can be applied for this purpose.

The training and testing sets to construct and test the decision tree model are based on BIRCH (k = 5) (c.f. Table 7). Prediction accuracy of the decision tree model is examined. For the example of a two-class prediction problem, given the testing set, prediction accuracy can be obtained by the confusion matrix shown in Table 10.

Table 10: Confusion matrix

Actual	Predicted	
	Class 1	Class 2
Class 1	a	b
Class 2	c	d

$$\text{Prediction accuracy} = \frac{a+d}{a+b+c+d}$$

Since the computer proficiency test contains discipline and skill based tests and students are required to pass both, the five clusters with these two tests are examined individually in terms of prediction accuracy and decision rules. Table 11 shows prediction accuracy of the decision rules extracted from each of the five clusters.

Table 11: Prediction accuracy of the extracted rules in the five clusters

	Cluster ID				
	C1	C2	C3	C4	C5
Discipline based test	78.62%	82.42%	80.96%	78.78%	82.95%
Skill based test	86.91%	85.28%	79.95%	81.33%	83.18%

The prediction results indicate that the extracted rules from the identified five groups are highly reliable for predicting students who cannot pass the discipline and skill based tests. More specifically, by using these decision rules we can correctly predict about 80% of the students who will fail the computer proficiency test over the testing set. Table 12 lists the extracted 19 rules of the five clusters for CCU to decide which students will not pass the tests.

These decision rules are very simple to use in practice. The following steps are performed to forecast whether a new student S will fail this test.

1. Collect the 21 required variables for S shown in Table 2;
2. Input these variables to BIRCH with five clusters;
3. Examine the clustering result of S, i.e. the cluster ID that S belongs to;
4. Use the rules of the cluster ID of S shown in Table 12 to produce the prediction.

For example, there is a student, Mike, who is going to take the computer proficiency test. We then input the 21 required variables of Mike to BIRCH with five clusters. If Mike was grouped into the third cluster (C3), i.e. the Colleges of Law and Education

Table 12: Decision rules of the five identified groups

Test	Cluster ID	Decision rules
Discipline based test	C1	• If Admission channel is general student, College name is {Management, Law, or Education}, and Gender is Female, then there is a 78.62% chance of failing the test.
	C2	• If Other items are {2 (in-service overseas student), 4, 5, 6, 7, or 8}, then there is an 82.42% chance of failing the test.
	C3	• If College name is {Law or Education}, and Admission channel is non-general student, then there is an 80.96% chance of failing the test. • If Place of birth is {Chiayi city, Chiayi county, or Yunlin county}, College name is Education, and Gender is Female, then there is an 80.96% chance of failing the test. • If Place of birth is {Kaohsiung, Penghu, Kinmen, Pingtung, Taitung, or Hualien county}, and College name is Education, then there is an 80.96% chance of failing the test. • If Place of birth is {Tainan city or county}, Blood type is O, and College name is {Engineering, Management, Law, or Education}, then there is an 80.96% chance of failing the test. • If College name is {Law or Education}, Place of birth is {Tainan city or county, Kaohsiung city or county, Penghu county, Kinmen county, Pingtung county, Taitung county, or Hualien county}, Gender is male, and Blood type is {B, AB, or O}, then there is an 80.96% chance of failing the test.
	C4	• If Blood type is A, the score of the Skill based test $\leq$ 55, and Gender is Female, then there is a 78.78% chance of failing the test. • If Place of senior high school is Hsinchu city, the score of the Skill based test $\leq$ 55, and College name is {Humanities, Science, and Social Science}, then there is a 78.78% chance of failing the test. Note that these rules are only suitable for warning students who take the skill based test first.
	C5	• If Place of senior high school is {Miaoli county, Taichung county, Nantou county, Chunghua county, Hsinchu city, Yunlin county, Chiayi county, Tainan county, Kaohsiung county, Penghu county, or Hualien county}, College name is {Management, Law, or Education}, Gender is male, and Admission channel is {general student, dispense student, or admission by application}, then there is an 82.95% chance of failing the test.
Skill based test	C1	• If Place of birth is {Taipei county, Keelung city, or Taipei city}, Gender is female, and Admission channel is {general student, dispense student, or admission by application}, then there is an 86.91% chance of failing the test. • If College name is {Management, Law, or Education}, Gender is male, and Place of senior high school is {Miaoli county, Hsinch county, Taoyuan county, Yilan county, Taipei county, Kaohsiung city, Tainan city, Keelung city, Taichung city, Taipei city, or foreign countries}, then there is an 86.91% chance of failing the test. • If Other items are {2 (in-service overseas student), 4, 5, 6, 7, or 8}, then there is an 86.91% chance of failing the test. • If Graduated class is A, Gender is female, and Nationality is one of foreign countries, then there is an 86.91% chance of failing the test.
	C2	• If Gender is Female, College name is {Engineering, Social Sciences, Sciences, or Humanities}, Other items is none, and Place of senior high school is outside of Kinmen, Penghu, and Matsu counties, then there is an 85.28% chance of failing the test.

	C3	• If Admission channel is {dispense student or admission by application}, Blood type is A, and Place of senior high school is outside of Taipei and Taichung cities and from foreign countries, then there is a 79.95% chance of failing the test.
	C4	• If Place of senior high school is {Chiayi, Tainan, or Kaohsiung counties}, College name is Humanities, Admission channel is general student, and Gender is Female, then there is an 81.33% chance of failing the test. • If Department name of admission is Chinese, Other items are {2 (in-service overseas student), 4, 5, 6, 7, or 8}, and Gender is Female, then there is an 81.33% chance of failing the test.
	C5	• If Department name of admission is {Business Administration, Accounting and Information Technology, Information Management, Law, Financial & Economic Law, Adult & Continuing Education, or Criminology}, Place of birth is outsider of Taipei city, and Place of senior high school is {Yilan county, Taipei county, Kaohsiung city, Tainan city, Keelung city, or Taichung city}, then there is an 83.18% chance of failing the test.

group, the variables of the decision rules in C3 shown in Table 12 are compared with the 21 variables of Mike. In C3, there are five rules for the discipline based test and one for the skill based test. Therefore, if the variables of Mike exactly match the first rule of C3 in the discipline based test, then we can say that Mike has an 80.96% chance of failing the test. In this case, the university will send him a 'warning' message to make him understand this predicted result and suggest ways to prepare better for this test.

5. Conclusion

Educational institutions in many countries require fresh graduates to be equipped with IT skills. That is, they have paid much attention to enhancing computer literacy amongst their undergraduate students. Hence, many universities invest lots of money and human resources to promote student's computer proficiency.

This paper focuses on examining students' computer proficiency test at one national university in Taiwan by taking advantage of data mining. Due to the complexity of students' backgrounds and learning situations, the experiments are first of all based on clustering the data samples into groups, and then extracting some useful rules from the identified groups. Regarding these rules, we can discover students who have higher probability of failing the computer proficiency test. Besides showing the applicability of data mining in this domain problem, our findings indicate the strong relationship between some representative attributes (or factors) and the failure of the computer proficiency test, such as the Place of senior high school. Moreover, based on these rules, we can remind students whose failure rates are potentially high and help them from failing the computer proficiency at the first attempt. As a result, students do not need to spend extra time to prepare for a second test and this helps the university to conserve resources.

In summary, the contribution of this paper is two-fold. Our experimental results can help the university (CCU) to identify a number of groups who need reinforcement training and promote their computer proficiency more efficiently. As English proficiency is another major focus for many universities in Taiwan, they also deploy lots of resources to promote students' English capability. This research can be applied also to the English language assessment for helping students pass this kind of exam.

6. References

Budayan, C., Dikmen, I. & Birgonul, M. T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. *Expert Systems with Applications,* 36(9), 11772-11781.

David, O. & Yong, S. (2007). *Introduction to business data mining.* Boston: McGraw-Hill/Irwin.

Guo, Q. & Zhang, M. (2009). Implement web learning environment based on data mining. *Knowledge-Based Systems,* 22(6), 439-442.

Guruler, H., Istanbullu, A. & Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education,* 55(1), 247-254. http://dx.doi.org/10.1016/j.compedu.2010.01.010

Hamdi, M. S. (2007). MASACAD: A multi-agent approach to information customization for the purpose of academic advising of students. *Applied Soft Computing,* 7(3), 746-771. http://hdl.handle.net/10576/10576

Han, J. & Kamber, M. (2006). *Data mining: Concepts and techniques.* New York: Morgan Kaufman.

Hannon, C. (2001). Information literacy in the undergraduate curriculum. *Educause Quarterly,* 24(4), 41-42. http://www.educause.edu/ir/library/pdf/EQM0146.pdf

Hosseini, S. M. S., Maleki, A. & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications,* 37(7), 5259-5264. http://dx.doi.org/10.1016/j.eswa.2009.12.070

Jain, A. K. (2009). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters,* 31(8), 651-666. http://dx.doi.org/10.1016/j.patrec.2009.09.011

Kanungo, T., Mount, D.M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A.Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24(7), 881-892. http://dx.doi.org/10.1109/TPAMI.2002.1017616

Kao, Y.-T., Zahara, E. & Kao, I-W. (2008). A hybridized approach to data clustering. *Expert Systems with Applications,* 34(3), 1754-1762. http://www.sciencedirect.com/science/journal/09574174

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics,* 43(1), 59-69. http://dx.doi.org/10.1007/BF00337288

Kohonen, T. (1989). *Self-organization and associative memory,* 3rd edition. New York: Springer-Verlag.

Kohonen, T. (2001). *Self-organizing maps*, 3rd edition. New York: Springer-Verlag.

Lee, M. W., Chen, S. Y., Chrysostomou, K. & Liu, X. (2009). Mining students' behavior in web-based learning programs. *Expert Systems with Applications,* 36(2), 3459-3464. http://dx.doi.org/10.1016/j.eswa.2008.02.054

Markov, Z. & Larose, D. T. (2007). *Data mining the web: Uncovering patterns in web content, structure, and usage.* New York: John Wiley & Sons.

Mingoti, S. A. & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research,* 174(3), 1742-1759.

O'Hanlon, N. (2002). Net knowledge: Performance of new college students on an Internet skills proficiency test. *The Internet and Higher Education,* 5(1), 55-66. http://dx.doi.org/10.1016/S1096-7516(02)00066-0

Paterlini, S. & Krink, T. (2006). Differential evolution and particle swarm optimisation in partitional clustering. *Computational Statistics & Data Analysis,* 50(5), 1220-1247. http://dx.doi.org/10.1016/j.csda.2004.12.004

Qiu, D. (2010). A comparative study of the K-means algorithm and the normal mixture model for clustering: Bivariate homoscedastic case. *Journal of Statistical Planning and Inference,* 140(7), 1701-1711. http://dx.doi.org/10.1016/j.jspi.2009.12.025

Romero, C. & Ventura, S. (2006). *Data mining in e-learning*. Southampton. UK: WIT Press.

Romero, C. & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications,* 33(1), 135-146. http://dx.doi.org/10.1016/j.eswa.2006.04.005

Romero, C., Ventura, S. & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers and Education,* 51(1), 368-384. http://dx.doi.org/10.1016/j.compedu.2007.05.016

Romero, C., Ventura, S., Zafra, A. & de Bra, P. (2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers & Education,* 53(3), 828-840. http://dx.doi.org/10.1016/j.compedu.2009.05.003

San, O. M., Huynh, V.-N. & Nakamori, Y. (2004). An alternative extension of the *k*-means algorithm for clustering categorical data. *International Journal of Applied Mathematics and Computer Science,* 14(2), 241-247.

Shih, C.-C., Chiang, D.-A., Lai, S.-W. & Hu, Y.-W. (2009). Applying hybrid data mining techniques to web-based self-assessment system of Study and Learning Strategies Inventory. *Expert Systems with Applications,* 36(3), 5523-5532. http://www.amcs.uz.zgora.pl/?action=paper&paper=198

Tang, T. & McCalla, G. (2005). Smart recommendation for an evolving e-learning system. *International Journal on E-Learning,* 4(1), 105-129.

Tesch, D., Murphy, M. & Crable, E. (2006). Implementation of a basic computer skills assessment mechanism for incoming freshmen. *Information Systems Education Journal,* 4(13), 1-11. http://isedj.org/4/13/ISEDJ.4(13).Tesch.pdf

Tsai, C.-F., Lin, Y.-C. & Wang, Y.-T. (2009). Discovering stock trading preferences by self-organizing maps and decision trees. *International Journal on Artificial Intelligence Tools*, 18(4), pp. 603-611. http://dx.doi.org/10.1142/S0218213009000299

Verhey, M. P. (1999). Information literacy in an undergraduate nursing curriculum: Development, implementation, and evaluation. *Journal of Nursing Education,* 38(6), 252-259.

Wang, Y.-H., Tseng, M.-H. & Liao, H.-C. (2009). Data mining for adaptive learning sequence in English language instruction. *Expert Systems with Applications,* 36(4), 7681-7686. http://dx.doi.org/10.1016/j.eswa.2008.09.008

Yang, F., Sun, T. & Zhang, C. (2009). An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization. *Expert Systems with Applications,* 36(6), 9847-9852. http://dx.doi.org/10.1016/j.eswa.2009.02.003

Authors: Chih-Fong Tsai (corresponding author), Department of Information Management, National Central University, Taiwan. Email: cftsai@mgt.ncu.edu.tw

Ching-Tzu Tsai, Department of Business Administration
National Chung Cheng University, Taiwan

Chia-Sheng Hung, Department of Accounting and Information Science
Nanhua University, Taiwan

Po-Sen Hwang, Computer Center, National Chung Cheng University, Taiwan

Please cite as: Tsai, C.-F., Tsai, C.-T., Hung, C.-S. & Hwang, P.-S. (2011). Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. *Australasian Journal of Educational Technology*, 27(3), 481-498. http://www.ascilite.org.au/ajet/ajet27/tsai.html