Electrical engineering student learning preferences modelled using k-means clustering

Citra Kurniawan†‡, Punaji Setyosari‡, Waras Kamdi‡ & Saida Ulfa‡

Sekolah Tinggi Teknik Malang, Malang, East Java, Indonesia†
State University of Malang, Malang, East Java, Indonesia‡

ABSTRACT: In this research, electrical engineering students' visual-verbal preferences were modelled using the k-means clustering method. The data collected included the level of initial ability of students, student demographic information and student learning preferences. Data were processed using the k-means clustering method, which divides data into several groups or clusters based on the similarity of data attributes. The study identified five clusters, viz. 1) Cluster 0 - informatics, intermediate, male, no, verbal; 2) Cluster 1 - broadcasting, intermediate, male, no, verbal; 3) Cluster 2 - informatics, master, male, no, visual; 4) Cluster 3 - informatics, teachers, men, yes, visual; 5) Cluster 4 - broadcasting, intermediate, male, no, visual. The k-means clustering method is iterative and required six iterations to converge onto a stable solution. Findings indicate that k-means analysis can be used to model student data. Student modelling is essential for learning strategies that when appropriate to the student model can help students to get better outcomes.

Keywords: Student modelling, k-means cluster, learning preferences

INTRODUCTION

Students have various characteristics that affect learning. Students who have similar characteristics form a model called the student model. The student model consists of student demographic information, learning style preferences, initial skill level and other information. The model of students is a reference model of the characteristics of students who are grouped based on similar attributes. Student modelling becomes essential in learning because learning strategies that are appropriate to the student model can help students to get better learning outcomes. Therefore, the *one size fits all* learning approach cannot be used as a learning strategy, because students have different learning style preferences. Characteristics of students, as modelling attributes of students, were the focus in this study. The student attributes consist of student identity, gender, study programme, student preference and level of knowledge.

The process of determining the level of knowledge involves the extraction of evidence, the combination of evidence and the initialisation of the student model. The extraction of evidence uses pre-test questions to identify a student's understanding, and these results are combined with other information. The next step is initialising the student model. Forming a student model profile consists of two stages: initialisation of the student model and student model update. The initialisation of the student model follows the pre-test to determine initial ability. Initial ability is classified into three classifications of *beginner*, *intermediate* and *expert*. Next is to update the student model; when students have completed learning activities, the results are recorded as information that can be traced back. Learning preference refer to how students learn and personalise their choices in the learning process. [2]. The model uses the following nomenclature [1], see Equation (1) below:

$$\theta = \{L, I, M\} \text{ (L: low level; I: intermediate level; M: master level)} \quad (1)$$

Learning preference is how students learn and personalise their choices in the learning process [2].

The development of students during learning updates the student model. The student model represents individual or group characteristics, and can be formed from model profiles, cognitive overlays, predictive models or only overlays. The model profile contains all information related to a student, such as name, learning style, age and gender. Cognitive overlays record skill levels, such as beginner, intermediate or expert. Predictive models present the student's learning resource preferences. Overlay contains information relating to student interaction with learning resources.

The student model can affect the learning process, because each student has different initial abilities. According to Nakic et al, the development of a student involves a set of variables including cognitive ability, meta-cognitive ability, psychomotor skills, cognitive style, learning style, experience, prior knowledge and preference [3]. The age of the student is related to the experience and background knowledge possessed by the student. Meta-cognitive ability is related to the ability of a student to adjust to a change in the way of learning in the learning process. Cognitive style is related to the pattern of information processing in learning. Learning styles relate to the learning environment and describe attitudes and behaviours that determine how a student learns [3].

In this study is presented the modelling of students based on their attributes, in the form of a data mining analysis. Attributes of students were processed using the k-means clustering method. The k-means clustering method is used to organise the available data into groups, where a group contains similar characteristic data, while another group contains data with different characteristics, though the characteristics are similar within that group. K-means clustering is a cluster analysis method that aims to separate the observed n quantity of data into k clusters, where the observed data are grouped into the cluster with the nearest mean [4]. A data mining analysis was performed to determine clusters from the attribute data collected. Data attributes were obtained from student data, assessment of student ability levels, and the measurement of learning preference between verbal and visual. The aim of the research was to model students based on a data mining analysis approach using k-means clustering on the attributes of students.

STUDENT MODELLING

Student modelling is a process of model formation based on grouping similarity attributes that have been collected for each student. According to Jia et al, the student model consists of the level of knowledge, cognitive abilities, and preferences of students [5]. The level of knowledge refers to the knowledge of a student to achieve goals.

The level of knowledge (*k*) refers to Bloom's taxonomy (*h*) on cognitive roles, such as knowledge, comprehension, application, analysis, synthesis and evaluation. Level 1 refers to knowledge (*h1*), level 2 refers to comprehension (*h2*), level 3 refers to application (*h3*), level 4 refers to analysis (*h4*), level 5 refers to synthesis (*h5*) and level 6 refers to evaluation (*h6*). Students with basic skills are at level 0 (*h0*). So, the following relationships in Equation (2) are obtained for knowledge:

$$(k, h) = \{(k_1 , h_1), (k_2 , h_2), ..., (k_n , h_n)\} \quad (2)$$

Cognitive ability is related to the ability of students to complete tasks ranging from simple to the most complex. Cognitive ability can be represented by the ability level. The relationship between ability and level, (a , l), can be seen by the following Algorithm (3):

$$(a , l) = \{(a_1 , l_1), \{(a_2 , l_2), ……., (a_n , l_n)\} \quad (3)$$

Where: a_i is the cognitive ability,
l_i is the level of cognitive ability.

Students' preferences are defined as interests, hobbies and other information of interest in learning. In this case, the student preferences consist of the student's background (*b1*), the learning strategy (*b2*) and the learning time (*b3*), which is represented by Equation (4) below [6]:

$$P (c, \sigma) = \{<b1, \sigma1>, <b2, \sigma2>, <b3, \sigma3>\} \quad (4)$$

Where: c is the preference concept,
b_i is the student preferences,
σ_i is the preference level.

According to Benton, student modelling is influenced by student interactions with the system and system characteristics [7]. The following Algorithm (5) illustrates the interaction of students (*y*):

$$\sum_{i=0}^{n} y_i = |(c_i + d_i x_i + r_i)| \quad (5)$$

Where: c_i is the indirect interaction,
d_i is the impact of the technology,
r_i is the residual factor of the system usage.

System characteristics (s) are described by Equation (6) below:

$$\sum_{j=0}^{n} s_j = a_j + r_j \quad (6)$$

Where: a_j is the j^{th} system object.

So overall, the modelling of students can be described by the algorithm $\Sigma(|s| + |y|)$ [4].

In the modelling process the attributes for a student consists of the student's initial ability level, student identity, the course of study, gender, interaction with the learning and the preferences of the student.

K-MEANS CLUSTERING

Clustering is a process of partitioning a data set into multiple groups or subsets, such that the members of each group have similar attributes, i.e. intra-group attributes are similar, while inter-group attributes are not. The analysis used in this research is k-means clustering. K-means clustering is a method used to divide objects into partitions based on categories by looking at the distance of the object from the closest mean. The k-means clustering algorithm aims to find the k groups iteratively by assigning each data point to one of the k groups [8]. The result of the k-means clustering algorithm is the centroid for each of the k clusters. The iterative process consists of [9]:

- initialising the k-partition M = [m_1, ..., m_k] based on students' previous knowledge;
- assigning each object x_j in the data set to the nearest cluster C_i;

  where: $x_j \in C_l$ if $\|x_j - m_l\| < \|x_j - m_i\|$ for j=1, …, N, $i \neq$ l, and i = l, …, k;
- recalculating the cluster prototype matrix based on the partition $m_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j$.

The process begins by determining the number of clusters. Then, the data are allocated to the clusters and the centroid calculated for each cluster. Data are allocated to the nearest centroid. The iteration process continues until no cluster changes. Cluster determination is based on minimising the mean distance of objects in clusters from the centroids of the clusters.

RESEARCH METHOD

Subjects for this research were 100 electrical engineering students who take Informatics Engineering and Broadcasting Engineering at a technical high school in Malang, Indonesia. Demographic information was collected from students. It consisted of student identity, majors, gender and interest in image. Data analysis was performed based on the k-means clustering method. Visual and verbal preferences were measured using the visualiser-verbaliser questionnaire (VVQ) of Richardson [10] adapted to the visual-verbal preferences approach of Kirby [11]. The level of student ability was based on a pre-test value. Data processing using k-means clustering analysis was performed using the data mining software, Weka. The k-means process divides the dataset into several groups based on the data patterns. The division is done by an iterative process to produce a stable set of groups of the student model. The student model was obtained through a recurring process by determining the number of groups (k) and the initial centroids at random. The measurement of the distance between the dataset and the centroid allocates members of the dataset to the nearest centroid. Workflow of student modelling is shown in Figure 1.

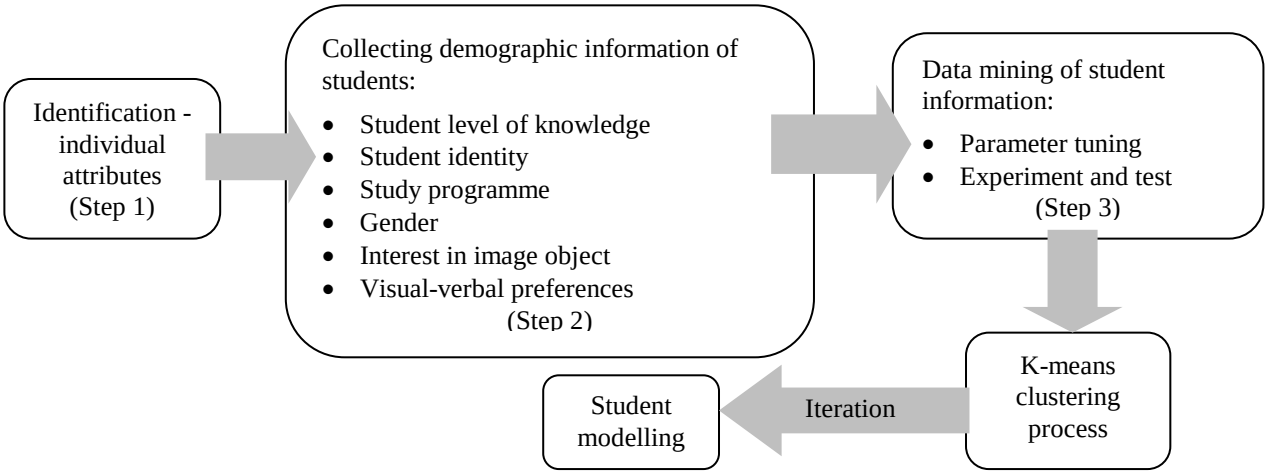

Figure 1: Workflow of student modelling.

The k-means clustering analysis was performed using the data mining tool, Weka 3.8. The collected data were processed into Attribute-Relation File Format (ARFF). The ARFF used in this study consisted of an @RELATION naming the dataset, @ATTRIBUTES listing the data attributes and @DATA for the data collected. An example of the ARFF data used in the research can be seen in Figure 2.

```
@relation 'preferences measurement'

@attribute Study {Informatics,Broadcasting}
@attribute Student_knowledge_level {Intermediate,Master,Low}
@attribute Gender {Male,Female}
@attribute Interest_in_image_object {Yes,No}
@attribute Preferences {Visual,Verbal}

@data
Informatics,Intermediate,Male,Yes,Visual
Informatics,Intermediate,Male,No,Visual
Informatics,Master,Male,Yes,Visual
Informatics,Intermediate,Male,Yes,Visual
Informatics,Intermediate,Male,Yes,Visual
Informatics,Low,Male,Yes,Verbal
Informatics,Intermediate,Male,No,Verbal
Informatics,Intermediate,Male,Yes,Verbal
Informatics,Intermediate,Male,No,Visual
Informatics,Master,Male,No,Visual
Informatics,Low,Male,Yes,Visual
Informatics,Low,Male,No,Visual
Informatics,Low,Male,Yes,Visual
```

Figure 2: Data relation, attributes and student data.

The attributes are: study, level of student knowledge, gender, interest in drawing object and preferences. The student data for each attribute then follows. The grouping of student data was based on similarity patterns of the attribute data.

RESULTS AND DISCUSSION

The k-means clustering process consists of three stages:

1) mapping student data;
2) determining the number of clusters;
3) iterative method to allocate data to clusters.

Student data are grouped by distance from the centroid of a cluster. The iteration continues until there is no change in the group allocation of student data. The result of data mining analysis of student attributes is shown in Figure 3.

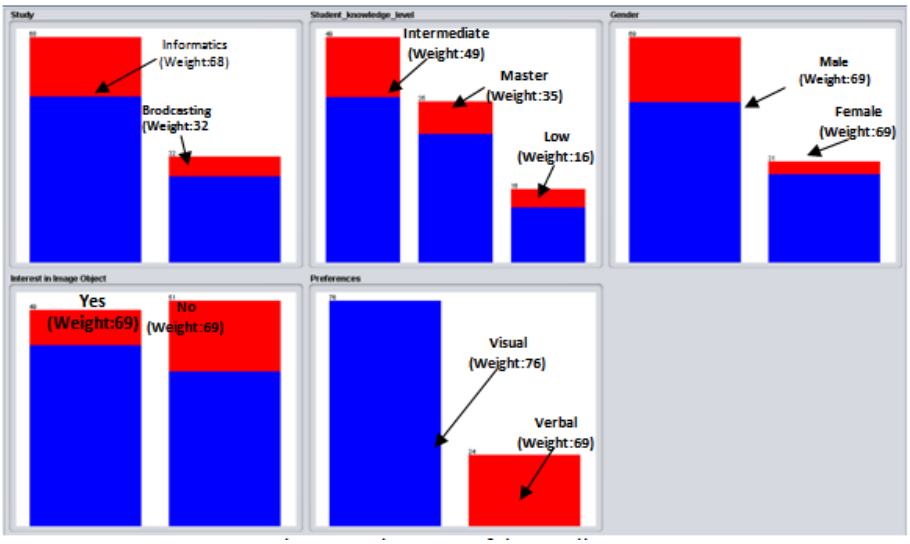

Figure 3: Histogram of data attributes.

Each parameter has a spread on the visual-verbal preference. The training data used in this research had a preferences relation (data file *measurement preferences.arff*), which has attributes study, student knowledge level, gender, drawing interest, preferences. The study attribute has the dimensions of informatics and broadcasting; the student knowledge level attribute has dimensions low, intermediate, master; gender has the dimensions of man and women; drawing interest has dimensions yes and no; preferences have visual and verbal dimensions. In this study, the iteration process ran six times to achieve unchanging clusters. The results of the process are shown in Figure 4.

Figure 4 shows cluster 0 has a data distribution with similar attribute data of 30%; cluster 2 has a data distribution with similar attribute data of 22% and Cluster 1 has a data distribution with attribute similarity of 21%. The other groups did not show significant similarity. Cluster 0, Cluster 1 and Cluster 2 can be used as reference models of the students. The results show that each cluster has distribution based on verbal-visual preferences. Distribution of cluster data on verbal-visual preferences is shown in Figure 5.

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 81.0

Initial starting points (random):

Cluster 0: Informatics,Intermediate,Male,No,Verbal
Cluster 1: Broadcasting,Intermediate,Male,No,Verbal
Cluster 2: Informatics,Master,Male,No,Visual
Cluster 3: Informatics,Master,Male,Yes,Visual
Cluster 4: Broadcasting,Intermediate,Male,No,Visual

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1           2           3           4
                             (100.0)      (30.0)      (21.0)      (22.0)      (17.0)      (10.0)
=============================================================================================================
Study                     Informatics  Informatics Broadcasting  Informatics  Informatics Broadcasting
Student_knowledge_level  Intermediate Intermediate      Master      Master Intermediate Intermediate
Gender                           Male         Male        Male        Male      Female        Male
Interest in Image Object           No           No          No         Yes         Yes          No
Preferences                    Visual       Verbal      Visual      Visual      Visual      Visual
```

Figure 4: Clusters of student attributes.

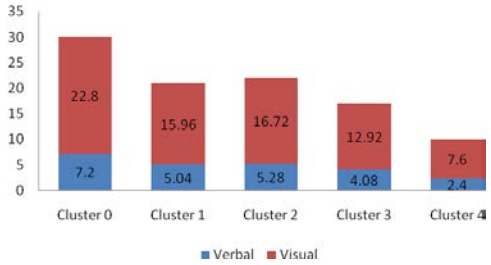

Figure 5: Distribution of cluster data on verbal-visual preferences.

In this study three clusters were identified that can be used as reference models. Student modelling divided clusters based on similarity attributes, as shown in Figure 6.

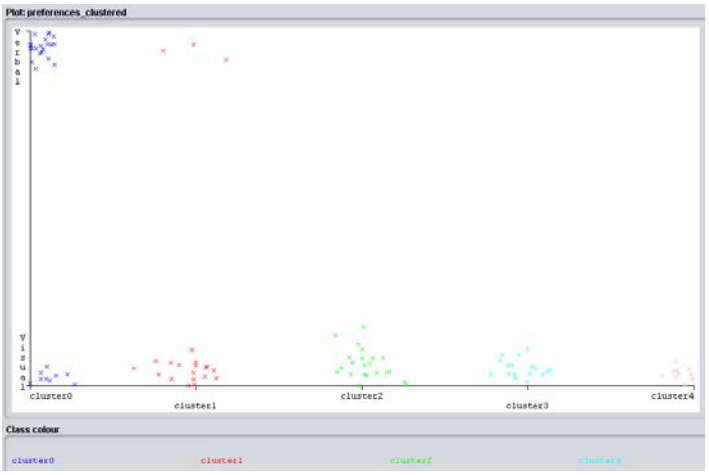

Figure 6: Scatter plot clusters against visual-verbal preferences.

Figure 6 shows the distribution of data in groups on verbal-visual preferences. Cluster 0 shows the distribution of data on verbal preferences, whereas Cluster 1 shows the distribution of data on visual preferences and Cluster 0 also shows the distribution on visual preferences.

CONCLUSIONS

Student attribute data were modelled in this research by using the k-means clustering method. The results identified attribute data that had the closest distance from the centroid. The k-means clustering method identified groups having similar dataset patterns. It was found that the k-means clustering positively aided the grouping process based on attribute data in the analysis of the student data. Hence, k-means analysis can be used to model student data.

REFERENCES

1. Esichaikul, V. Lamnoi, S. and Bechter, C., Student modelling in adaptive e-learning systems. *Knowledge, Manage. E-Learning an Inter. J.*, 3, **3**, 342-355 (2011).

2. Castillo, G., Gama, J. and Breda, A.M., *An Adaptive Predictive Model for Student Modeling, in Online and Distance Learning: Concepts, Methodologies, Tools and Applicatio*ns. Tomei, L.A. (Ed), Information Science Reference/IGI Global, 1-19 (2008).
3. Nakic, J., Granic, A. and Glavinic, V., Anatomy of student models in adaptive learning systems : a systematic literature review of individual differences from 2001 to 2013. *Educ. Comput. Research*, 51, **4**, 459-489 (2015).
4. Sharma, P., Comparative analysis of various clustering algorithms using WEKA. *Inter. Research J. of Engng. Technol*., 2, **4**, 107-112 (2015).
5. Jia, B., Yang, Y. and Zhang, J., Study on student modeling in adaptive learning system. *J. of Comput*., 7, **10**, 2585-2592 (2012).
6. Baukal, C.E. and Ausburn, L.J., Learning strategy and verbal-visual preferences for mechanical engineering students. *American Society for Engng. Educ*., 121 (2014).
7. Benton, S., Altemeyer, B. and Manning, B., Behavioural prototyping: making interactive and intelligent systems meaningful for the user. *Intelligent Adaptation & Personalization Techniques*, 195-210 (2012).
8. Trevino, A., Introduction to K-means Clustering, Data Science (2016), 1 March 2018, https://www.datascience.com/blog/k-means-clustering
9. Xu, R. and Wunsch, D.C., *Clustering*. New Jersey: IEEE Press a, 10 (2009).
10. Richardson, A., verbalizer-visualizer: a cognitive style dimension, *J. of Ment. Imag.*, 1, 109-126 (1977).
11. Kirby, J.R., Moore, P.J. and Schofield, N.J., Verbal and visual learning styles *Contemp. Educ. Psychol*.,13, **2**, 169-184 (1988).

BIOGRAPHIES

Citra Kurniawan is a postgraduate student at the State University of Malang. He is also currently teaching electrical engineering at Sekolah Tinggi Teknik Malang, Malang, East Java, Indonesia. His research interests are data mining, computer networking and learning engineering.

Punaji Setyosari is a Professor in the Department of Education and Technology at the State University of Malang, East Java, Indonesia. His research interests include research methodologies, evaluation and assessment, instructional media, problem-based learning and collaborative learning.

Waras Kamdi is a Professor in the Department of Mechanical Engineering Education at the State University of Malang, East Java, Indonesia. His research interests include project-based learning, vocational education, instructional technology and learning strategies.

Saida Ulfa is a lecturer in the Department of Educational Technology at the State University of Malang, East Java, Indonesia. Her research interests include mobile learning, instructional media and learning engineering.