Determination of the Senior High School Department Using the Expectation Maximization Algorithm

I Made Gede Yudiyana[1*], Andrew Sumichan [2], Muhammad Ridwan Satrio[3], Antonius Ibi Weking[4]

[1,2,3,4]Departement Electrical and Computer Engineering, Post Graduate Program, Udayana University
Department of Electrical Engineering, Udayana University
*yudiyana.made@baledigital.com [1]

Abstract - High school is an educational institution that has directed students to concentrate or learn more in the field of science in accordance with the academic abilities of the students concerned. Generally the top high school has 3 department intended for students when climbing levels from grade 1 to grade 2, namely majors in Science, Social Sciences and Language. Nowadays with a growing human population, the number of students enrolled in a school is quite large, with the number generally in public schools 30 to 40 students per class. With the large number of students this will provide difficulties to determine the direction of each student if done manually. With this problem the researcher implements an EM (Expectation Maximization) method, this method is chosen because of its ability to get the highest probability value and can find the optimal parameters even though the available available map is incomplete or lacking. Using this method will make it easier for a school institution to determine the right direction for students.

.

***Index Terms*— High School Department, EM (Expectation Maximization), Data Collection, Student**

I. INTRODUCTION

THE decision of department in major high school is very important to the students. The problem of choosing majors for high school students can be solved by various methods in the realm of computer science [1], [2], [3]. It is possible that unrecorded data in making departmental decision making for high school students arises because of the large amount of data processed. The missing value problem can be overcome by the EM (Expectation Maximation) algorithm [4], [5]. In the research of L. R. Sirait and A. P. Kurniati (2013) the use of the EM algorithm is able to handle data well which has missing value. This can be seen from the value of the NRMSE performance [4]. From these studies, GAP was found to utilize the Expectation Maximization algorithm in the problem of selecting high school student majors. The contribution of this study is to introduce the Expectation Maximization algorithm in solving the problem of choosing majors in high schools. This study consisted of 5 sections, research method and discussion is applied in sections III and IV until concluding with the conclusions of this study in section V.

II. LITERATURE

A. Clustering Algorithm Overview

Clustering can be considered as the most important learning problem that is not found, and

1

(p-issn: 2579-5988, e-issn: 2579-597X)

like every other problem of this kind it aims to find structures (intrinsic groupings) in data sets that are not labeled. Therefore clusters are a collection of 'similar' objects between each other and 'different' with objects belonging to other clusters. Another type of grouping is conceptual grouping where two or more objects are considered to belong to the same cluster if they define a general concept for all these objects. That is, objects are grouped according to their suitability with descriptive concepts, not according to the simple measure of similarity. An important question is how to decide what is a good grouping, because it is generally recognized that there is no absolute 'best' criterion that will be independent of the final goal of grouping. As a result, it is users who must provide the criteria that best suits their particular needs, and the results of the grouping algorithms can be interpreted differently[6].

B. EM (Expectation Maximization) Method

Definition (Hogg, McKean and Craig, 2005) The EM algorithm was first introduced by Dempster, Laird, and Rubin in 1977. Broadly speaking, the EM algorithm is an algorithm for estimating a parameter in a function using MLE, where the function contains incomplete data. EM algorithm is a process that is divided into two steps, namely :

- Expectation (E-step)
  Search for expectation values for the likelihood function based on observed variables.
- Maximization (M-Step)
  Search for MLE from parameters by maximizing the likelihood expectation generated from the E-step.

The parameters generated from the M-step will be reused for the next step, and this step will be repeated until it gives a convergent value and is an estimator of a parameter[8].

III. RESEARCH METHODS

A. Data Source

Sources of data in the implementation of this study are dummy data created by researchers who adjust the actual state of the assessment of students' academic abilities to the direction that will be determined from the school.

B. Research Method

The method and research flow in this study can be described as follows:

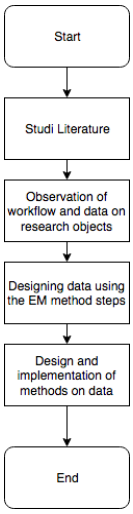

Figure 1. Research Method Flowchart

The study began with a literature study of the theories supporting the application of the EM method with 3 steps to process data. The implementation of the 3 steps of the EM method on the object data then the final results are obtained with the accuracy of determining the student's majors.

IV. IMPLEMENTATION AND RESULTS

A. Implementation Expectation Maximization Method

The implementation of the EM method is divided into 3 steps, with the workflow process in Fig.2, where the initial process is done by inputting all student values.

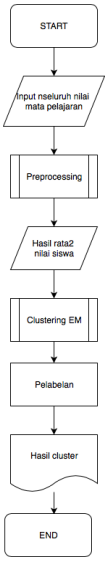

Figure 2. Steps of EM Method

1. Preprocessing
   At this stage is the stage of searching the average value of students for the whole subject and grouping them in 3 types of values according to the available departments, namely the value of science, social studies and language.

Table 1. Average score of students

Student	Science	Social Science	Language
Student A	50	60	70
Student B	65	80	73
Student C	72	70	65
Student D	83	65	80
Student E	40	82	73
Student F	95	71	85
Student G	60	74	96
Student H	75	75	92
Student I	83	55	70
Student J	91	60	65

2. Clustering
   At the stage of trapping, students will be grouped based on the average value obtained at the preprocessing stage. The results of this process are only temporary groupings based on average scores.

3. Labeling
   At this stage will be grouped using the EM method, students will be grouped based on the average value obtained, with the highest standard probability of 70 each department. This probability standard is determined based on the standard of acceptance of majors in general in public schools. So that in this study the probability value is static.

The following are some of the grouping processes in the EM method :

1. Guess Model Parameter
   Determine the probability value of data for a cluster. To determine the qmk and αk parameter values.

$$r_{nk} = \frac{\alpha_k(\prod_{t_m \in d_n} q_{mk})(\prod_{t_m \notin d_n}(1-q_{mk}))}{\sum_{k=1}^{K} \alpha_k(\prod_{t_m \in d_n} q_{mk})(\prod_{t_m \notin d_n}(1-q_{mk}))}$$

2. Expectation
   Determines the total probability value of the cluster and then determines Frequency Counts.

$$\sum_{n=1}^{N} r_{nk} I(t_m \in d_n)$$

3. Maximization
   The probability value term m to a cluster where the term m is a member of a document n.

$$q_{mk} = \frac{\sum_{n=1}^{N} r_{nk} I(t_m \in d_n)}{\sum_{n=1}^{N} r_{nk}}$$

Then repeat steps 2 and 3 until the Convergence / cluster probability data values are Convergence.

Table 2. Result clustering using EM method

Science (7)			
Student D	83	65	80
Student I	83	55	70
Student K	92	91	55
Student L	76	80	59

Student Q	82	50	80
Student R	81	65	88
Student T	77	71	55

Social Science (6)			
Student B	65	80	73
Student C	72	70	65
Student E	40	82	73
Student O	63	79	69
Student P	58	93	76
Student S	76	74	70

Language (7)			
Student A	50	60	70
Student F	95	71	85
Student G	60	74	96
Student H	75	75	92
Student J	91	60	65
Student M	75	65	74
Student N	74	76	89

V. CONCLUSION

From the results of the implementation of the EM method on the determination of high school majors two conclusions can be drawn. The first is based on the standard value of majors by 70, obtained by the level of accuracy of the data grouping of 95% and based on the grouping of the magnitude of student scores on majors, the use of the EM algorithm has a grouping accuracy rate of 65%.

ACKNOWLEDGMENT

We, as the authors, would like to express our first gratitude to the Almighty God who made this research complete smoothly. then to the lecturer who guides us and to all parties and colleagues who support us in doing this research.

REFERENCES

[1] M. Rahmayu and R. K. Serli, "Sistem Pendukung Keputusan Pemilihan Jurusan Pada Smk Putra Nusantara Jakarta Menggunakan Metode Analytical Hierarchy Process (Ahp)," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 9, no. 1, pp. 551–564, Apr. 2018.

[2] "ANALISIS KEPUTUSAN MENENTUKAN JURUSAN PADA SEKOLAH MENENGAH KEJURUAN DENGAN METODE SIMPLE ADDITIVE WEIGHTING," no. 1, p. 8, 2017.

[3] E. W. Sulystiyawati and I. E. Purwaningsih, "PERAN HASIL TES PENJURUSAN STUDI TERHADAP PEMILIHAN JURUSAN PADA SISWA SMA," *JS*, vol. 5, no. 1, p. 35, Apr. 2017.

[4] L. R. Sirait and A. P. Kurniati, "Analisis Algoritma Expectation Maximization (EM) Dalam Penanganan Missing Value," p. 6, 2013.

[5] "Algoritma Expectation-Maximization(EM) Untuk Estimasi Distribusi Mixture - PDF." [Online]. Available: https://docplayer.info/40900789-Algoritma-expectation-maximization-em-untuk-estimasi-distribusi-mixture.html. [Accessed: 27-May-2019].

[6] Yong Gyu Jung, Min Soo Kang and Jun Heo "Clustering performance comparison using K- means and expectation maximization algorithms", Biotechnology & Biotechnological Equipment, 2014.

[7] Law Rencus Sirait, Shaufiah and Angelina Prima Kurniati"Analisis Algoritma Expectation Maximization (Em) Dalam Penanganan Missing Value" , Telkom University.

[8] Tomy Angga Kusuma and Suparman, "Algoritma Expectation-Maximization(EM) Untuk Estimasi Distribusi Mixture" urnal Konvergensi Vol. 4, No. 2, 2014 .

[9] Paul S. Bradley, Usama M. Fayyad and Cory A. Reina, "Scaling EM (Expectation-Maximization) Clustering to Large Databases" Microsoft Research Microsoft Corporation One Microsoft Way Redmond, WA 98052.

[10] Osama Abu Abbas, "Comparasions Between Data Clustering Algorithm," The International Arab Journal of Information Technologi Vol 5 No. 3, 2008.