Clustering for multi-dimensional data set: a case study on educational data

M R Yudhanegara* and K E Lestari
Department of Mathematics Education, Universitas Singaperbangsa Karawang, Karawang, Indonesia

*Corresponding author’s email: mridwan.yudhanegara@staff.unsika.ac.id

Abstract. This paper presents a two-step cluster method to handle data sets in educational data mining. It can handle multi-dimensional metric data points especially in complex data sets. For a case study, we use tracer study data in order to assign a clustering of alumni based on their profile. The result of clustering will help the school to evaluate and improve the quality of its graduates.

1. Introduction
At present, there are many ways to develop research in education through data mining. Usually, the methods are called educational data mining (EDM). One of the current promising methods in EDM is clustering. Clustering is a technique for grouping data points based on similarity [1]. Besides, the derived clusters can be visualized more efficiently and effectively than the origin data set [2]. Dominguez et al. used clustering for student cases [3], Shih et al. employed clustering for learning tactics [4], for curriculum planning [5], for estimating skill [6].

Determining the number of clusters of data is an important problem [7]. It is a difficult task if not known beforehand [8]. On the other hand, maybe we had been known the number of clusters from the data. Although there is no perfect way to determine the number of clusters, there are some statistics that can be analyzed to help in the process [9] [10]. The step in cluster analysis can be showed in Figure 1.

Figure 1. Step in cluster analysis [7].
Generally, there are two types of attributes involved in educational data to be clustered: continuous and categoric. The hierarchical clustering methods are only for small data sets [11], k-means should only be used on continuous attributes [7] and two-step clustering can be used all type of attributes simultaneously. Therefore, for a case study, we use two-step clustering to handle a tracer study data.

This paper is organized into the following sections. In the first section, the algorithm undertakes a pre-cluster step uses a sequential clustering approach by constructing a modified cluster feature (CF) tree [12]. Section 2 describes two-step clustering methods (procedure, distance measure, and determine the number of clusters). Section 3 shows the application of two-step clustering on educational data set. The data obtained from a tracer study data that contain a profile of alumni in one of senior high school in Bandung. Furthermore, we investigate the behavior of alumni to survive at university based on alumni profiles. An investigation is done by comparing each two pair of attributes with the GPA. Our conclusion is also in this section.

2. Methods

The two-step cluster analysis developed by Chiu et al. [13] has been specifically designed to handle mixed variables measured on different scale levels. The two-step clustering is based on two-stage approached, i.e. pre-cluster and second cluster the sub-clusters resulting.

2.1. Pre-clustering

The pre-cluster step scans the data records one by one and decides if the current record should be merged with the previously formed a new cluster based on the distance criterion.

The cluster feature $CF_j$ of a cluster $C_j$ is

$$ CF_j = \left\{ N_j, s_{Aj}, s_{Aj}^2, N_{Bj} \right\}, \tag{1} $$

where $N_j$ is the number of data records in $C_j$, $s_{Aj}$ is the sum of continuous attributes of the $N_j$ data records, $s_{Aj}^2$ is the sum of square continuous attributes of the $N_j$ data records, and $N_{Bj} = (N_{Bj1}, N_{Bj2}, \ldots, N_{BjkB})$ is a $\sum_{k=1}^{K_B} (L_k - 1)$-dimensional vector where the $k$-th sub-vector is of $(L_k - 1)$ dimension, given by $N_{Bjk} = (N_{jkl1}, N_{jkl2}, \ldots, N_{jklL_k-1})$ in which $N_{ijkl}$ is the number of data record in $C_j$ whose $k$-th categorical attribute takes the $l$-th category, $l = 1, 2, \ldots, L_k - 1$.

When two clusters $C_j$ and $C_r$ are merged, it means that the two corresponding sets of data points are gathered together to from a union. In this case, the $CF_{(j,r)}$ for the merged cluster $C_{(j,r)}$ can be calculated by adding the corresponding entries in $CF_j$ and $CF_r$, that is,

$$ CF_{(j,r)} = \left\{ N_j + N_r, s_{Aj} + s_{Ar}, s_{Aj}^2 + s_{Ar}^2, N_{Bj} + N_{Br} \right\}. \tag{2} $$

This cluster feature is an efficient way of representation of data records[13].

2.2. Clustering

In this stage, we use an agglomerative hierarchical clustering methods. The log-likehood based distance measure is used in this step. The distance between clusters $C_j$ and $C_r$ is defined as

$$ d(j, r) = \xi_j + \xi_r - \xi_{(j,r)}, \tag{3} $$

where $\xi_v = -N_v \left( \sum_{k=1}^{K_A} \frac{1}{2} \log(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2) + \sum_{k=1}^{K_B} E_{vk} \right)$, and $E_{vk} = -\sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v}$.

$K_A$ is the total number of continuous attribute and $K_B$ is the total number of categorical attribute. $\hat{\sigma}_k^2$ is the estimated variance of the continuous attribute $k$ and $\hat{\sigma}_{vk}^2$ is the estimated variance of the continuous attribute $k$, in cluster $v$. 


If \( \hat{\delta}^2_k \) is ignored in the expression for \( \xi_s \), the distance between clusters \( C_j \) and \( C_r \) would be exactly the decrease in log-likelihood when the two clusters are combined. The \( \hat{\delta}^2_k \) term is added to solve the problem caused by \( \hat{\delta}^2_k = 0 \), which would result in the natural logarithm being undefined. This would occur, for example, when a cluster has only one case.

We used BIC (Schwarz’s Bayesian Information Criterion) to determine the number of clusters. For \( J \) cluster, the BIC is as

\[
BIC (J) = -2 \sum_{j=1}^{J} \xi_j + m_j \log N ,
\]

where \( m_j = J \left[ 2K + \sum_{k=1}^{K} (L_k - 1) \right] \). \( N \) is the number of data records. \( L_k \) is the number of categories for the \( k \)-th categorical attribute.

The ratio of change in BIC at each successive merging relative to the first merging determines the initial estimate. Let \( dBIC (J) \) be the difference in BIC between the model with \( J \) clusters and that with \( (J + 1) \) clusters, \( dBIC (J) = BIC (J) - BIC (J + 1) \). Ratio for \( J \) is

\[
R_\xi (J) = \frac{dBIC (J)}{dBIC (1)} ,
\]

If \( dBIC (1) < 0 \), then the number of clusters is set to 1, and the second stage is omitted. Otherwise, the initial estimate for the number of clusters \( s \) is the smallest number for which \( R_\xi (J) < 0.04 \). This is done as follows

1. Model \( C_s \) indicated by the BIC criterion, and then take the ratio of minimum inter-cluster distance, the purpose is for that model and the next larger model \( C_{s+1} \), previous model in the hierarchical clustering procedure

\[
R_2 (s) = \frac{d_{\min} (C_s)}{d_{\min} (C_{s+1})} ,
\]

where \( C_s \) is model with \( s \) clusters and \( d_{\min} (C_s) \) is the minimum inter-cluster distance for cluster model \( C_s \), given by

\[
d (x_i, x_j) = \frac{\sum_{n=1}^{p} \delta_i^{(n)} d_{ij}^{(n)}}{\sum_{n=1}^{p} \delta_i^{(n)}} ,
\]

where \( d (x_i, x_j) \) is the dissimilarity between two instances, containing \( p \) attributes, the indicator \( \delta_i^{(n)} = 0 \) if one ofthe value is missing. The contribution of attribute \( n \) to the distance between the two objects \( d_{ij}^{(n)} \) is computed according to its type as follow

a. When attribute is categorical,

\[
d_{ij}^{(n)} = \left\{ \begin{array}{ll}
0, & \text{if } x_i = x_j \\
1, & \text{otherwise}
\end{array} \right. ,
\]

b. When attribute is continuous,

\[
d_{ij}^{(n)} = \frac{|x_{in} - x_{jn}|}{\max_{h} x_{hn} - \min_{h} x_{hn}} ,
\]

where \( h \) runs over all non-missing objects for attribute \( n \)

c. If the attribute is ordinal, the standardized values of the attribute are computes first and then \( z_{ln} \) is treated as continuous-valued.

2. From model \( C_{s-1} \), compute the same ratio with model \( C_s \). Repeat it until we have the ratio \( R_2 (2) \).

3. Compare the two largest \( R_2 \) ratios.
3. Results and Discussion
The data set that has been used in the case study has been obtained from a tracer study data that contain a profile of alumni in one of senior high school in Bandung, Indonesia. The data set has 70 record and presented in Table 1. This table contain information about the attribute, type of attribute and category. Overall data analysis using the most widely available statistical software–SPSS. In Table 2, we present the descriptive statistics of each attribute.

| Attribute                      | Type of attribute | Category                                      |
|--------------------------------|-------------------|-----------------------------------------------|
| National exam score (NES)      | Continuous        | -                                             |
| Grade point average (GPA)      | Continuous        | -                                             |
| University                     | Nominal           | 1 = ITB, UI, UGM, University of Southampton   |
|                                |                   | 2 = Unpad, Undip                              |
|                                |                   | 3 = Unpar, UKM                                |
|                                |                   | 4 = elsewhere                                 |
| Student admission test (SAT)   | Nominal           | 1 = SBMPTN                                    |
|                                |                   | 2 = SNMPTN                                    |
|                                |                   | 3 = SMMPTN                                    |
|                                |                   | 4 = elsewhere                                 |
| Skills                         | Ordinal           | 1 = weak                                      |
|                                |                   | 2 = fair                                      |
|                                |                   | 3 = good                                      |
|                                |                   | 4 = strenght                                  |
| Critical thinking (CT)         | Ordinal           | 1 = weak                                      |
|                                |                   | 2 = fair                                      |
|                                |                   | 3 = good                                      |
|                                |                   | 4 = strenght                                  |
| Competitiveness                | Ordinal           | 1 = weak                                      |
|                                |                   | 2 = fair                                      |
|                                |                   | 3 = good                                      |
|                                |                   | 4 = strenght                                  |

Table 1. Summary of source data. It contains a multi-dimensional metric of 70 data records. The categorical attributes consist several categories.

| Attribute      | Type of attribute | Category                                      |
|----------------|-------------------|-----------------------------------------------|
| NES            | Continuous        | -                                             |
| GPA            | Continuous        | -                                             |
| Maximum        | 59                | 3.96                                          |
| Minimum        | 34                | 2                                             |
| Range          | 25                | 1.96                                          |
| Mean           | 50.53             | 3.40                                          |
| Std. deviation | 4.605             | 0.350                                         |
| Frequency      |                   | 1                                             |
| SAT            |                   | 39                                            |
| Skills         |                   | 31                                            |
| CT             |                   | 0                                             |
| Competitiveness|                   | 0                                             |

Table 2. Descriptive statistics of data set. It shows that mostly the level of skill, critical thinking, and competitiveness of alumni are good.
3.1. Two-step clustering on data sets

This data set contains 5 categorical attributes and 2 continuous attributes. The categorical attributes are standardized by default and we use log-likelihood for distance measure, given in (3).

Table 3. According to the BIC criterion, the optimal numbers of clusters are three, which has the smallest coefficient of BIC and the largest distance ratio.

| Number of cluster | BIC     | Ratio  |
|-------------------|---------|--------|
| 1                 | 893.621 |        |
| 2                 | 824.233 | 1.465  |
| 3                 | 798.427 | 1.720  |
| 4                 | 811.874 | 1.013  |
| 5                 | 826.017 | 1.277  |
| 6                 | 851.825 | 1.193  |
| 7                 | 884.441 | 1.299  |
| 8                 | 925.198 | 1.046  |
| 9                 | 967.147 | 1.037  |
| 10                | 1010.027| 1.119  |
| 11                | 1055.581| 1.052  |
| 12                | 1102.239| 1.039  |
| 13                | 1149.694| 1.212  |
| 14                | 1200.732| 1.074  |
| 15                | 1252.933| 1.040  |

The results of clustering the data using two-step clustering are represented in Figure 2.

![Figure 2](image)

Figure 2. Result of two-step clustering. The cluster quality weight be determined into three clusters, it’s a fairly good solution.

Furthermore, the composition of each attribute in each cluster is represented in Figure 3. The first cluster, which fills 21.4%, contains mostly university type 4 (such as UPI, UIN, Unjani elsewhere), the average of national exam score is 49.27 and GPA 3.44, student admission test through SMMPTN (60%), the critical thinking level is fairly good (60%), level of skills and competitiveness is good, respectively 33.3% and 46.7%. Cluster 2 fills 27.1%, contains mostly university type 2 (Unpad and Undip), the average of national exam score is 52.00 and GPA 3.54, student admission test through SBMPTN (100%), the level of critical thinking, skills and competitiveness is good, respectively 57.9%, 42.1% and 73.7%.
The important cluster is the third. It is the largest (51.4%) containing university type 1 (ITB, UI, UGM, and University of Southampton), the average of national exam score is 50.28 and GPA 3.32, student admission test through SNMPTN (58.3%), the level of critical thinking, skills and competitiveness is good, respectively 72.2%, 63.9% and 55.6%.

![Figure 3. Composition of each clusters. Each cluster contains a different attribute composition.](image)

The comparison of all clusters simultaneously and shows the predictor importance in clusters. The predictor importance view shows the relative importance of each attribute in estimating the cluster. Furthermore, we investigate the behavior of alumni to survive at university based on alumni profiles. An investigation is done by comparing each two pair of attributes with the GPA.

4. Conclusion

Based on this case study, we conclude that two-step clustering can be used determines number of clusters with optimal. Two-step clustering can be applied in educational data mining, which use large data sets to find hidden patterns. Since most the educational data contain a multi-dimensional metric, the two-step clustering methods can work efficiently.

We identified alumni profile. The most important profile contains the alumni whose study at the best university with a good level of skill, critical thinking, and competitiveness, but the lower index of GPA. This clustering is useful to help the school to evaluate and improve the quality of its graduates.

5. References

[1] Trivedi S, Pardos Z A, Sarkozy G N, and Heffernan N T 2011 Spectral Clustering in Educational Data Mining Proc. Int. Conf. on Educational Data Mining vol 4 (Netherlands: Semantic Scholar) 138

[2] Lee R C T 1981 Clustering analysis and its applications: Advances in Information System Science vol 8, ed J T Toum (New York: Plenum Press) 292

[3] Dominguez A K, Yacef K and Curran JR 2010 Data Mining for Individualized Hints in eLearning Proc. Int. of Educational Data Minin 3 (Pittsburgh: Pittsburgh Science of Learning Center) 100

[4] Shih B, Koedinger K R and Scheines R 2010 Unsupervised Discovery of Student Learning Tactics Proc. Int. Conf. on Educational Data Mining 3 (Pittsburgh: Pittsburgh Science of Learning Center) 210

[5] Maull K E, Salvidar M G and Sumner T 2010 Online Curricular Planning Behavior of Teachers Proc. Int. Conf. on Educational Data Mining 3 (Pittsburgh: Pittsburgh Science of Learning Center) 130

[6] Nugent R, Dean N and Ayers E 2010 Skill-set Profile Clustering: The Empty K-Means Algorithm with Automatic Specification of Starting Cluster Center Proc. Int. Conf. on Educational Data Mining 3 (Pittsburgh: Pittsburgh Science of Learning Center) 160

[7] Mooi E and Sarstedt M 2011 A Concise Guide to Market Research (Verlag Berlin: Springer)
[8] Agarwal P, Alam M A and Biswas R 2011 Issue, Challenges and Tool of Clustering Algorithms *J. Computer Science* 8 528

[9] Milligan G W, and Cooper M C 1985 An Examination of Procedures for Determining the Number of Clusters in a Data Set *J. Psychometrika* 50 179

[10] Kaufman L and Rousseeuw P J 2005 *Finding Groups in Data: An Introduction to Cluster Analysis* (New Jersey: John Wiley & Sons)

[11] Zhang T, Ramakrishnan R and Livny M 1996 BIRCH: An Efficient Data Clustering Method for Very Large Database *Proc. Int. Conf. on Management of Data* 25 (Canada: ACM SIGMOD)

[12] Chiu T, Fang D, Wang Y and Jeris C 2001 A Robust and Scalable Clustering Algorithm for Mixed Type Attribute in Large Database Environment *Proc. Int. Conf. in Knowledge Discovery and Data Mining* 7 (San Francisco: Association for Computing Machinery) 268

[13] Șchiopu D 2010 Applying TwoStep Cluster Analysis for Identifying Bang Customers’ Profile *J. Stiințe Economice* 62 75