# Clustering Approach to Analyzing Student Data by using K-Means Algorithm

Khin Su Su Wai, Myat Myat Min
*Computer University, Taung-Ngu*
khinsusuwai@gmail.com, myatiimin@gmail.com

## Abstract

*Clustering is the process of grouping data into classes of clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. K-means clustering is a partitioning method. . K-means clustering algorithm is used to cluster the student data. The proposed system finds the relationship between students' government technology high school (G.T.H.S) entrance examination results and their success using cluster analysis. Euclidean distance measure also used to calculate the closest centroids for each object.*

*Keywords: Clustering Approach, K-means Algorithm, and Euclidean distance.*

## 1. Introduction

Data mining, also popularly referred to as knowledge discovery in databases (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored in large database, data warehouses, and other massive information repositories. It is a technology used in different disciplines to search for significant relationship among variables in large data sets. Data mining is mainly use in commercial applications. In this system, data mining stressed in an education environment [7].

K-means clustering satisfies the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group. K-means algorithm defines a random cluster centroid according to the initial parameters. Each consecutive case adds to the cluster according to the proximity between the mean value of the case and the cluster centroid. The clusters are re-analyzed to determine the new centroid point. This procedure repeated for each data object. K-means algorithm used to cluster the student data from student database [8].

Students' data clustering is the grouping of students data based on the principal of maximizing intra-cluster similarity and minimizing inter-cluster similarity. The major challenge of clustering is efficiently identify meaningful groups that concisely annotated [6].

In the paper, related work presents in section 2. In section 3, we describe theory of clustering and proposed system overview in section 4. The experimental result of the system shows in section 5. Finally, the conclusion included in section 6.

## 2. Related Works

K-means clustering used to reduce run time to complexity by accepting k value from the user. Student data placed into student database. Dissimilarity between two students has measured by Euclidean distance. For case study, we used the student data from student database as inputs for clustering.

Z. Senol Erdogan, Meltepe University and T. Mehpare, Istanbul University has proposed "A Data Mining Application in a Student Database". In this paper, university students grouped according to their characteristics, forming clusters [9].P. Jeyanthi, Sathyabama University and V. Jawahar Senthil Kumar, Anna University has showed "Image Classification by k-means clustering". In this paper, they propose to use k-means clustering for the classification of feature set obtained from the histogram [3]. C.R. Palmer, J.Pesenti and other, Informedia project, Carnegie Mellon University that demonstrates the clustering of search results from Carnegle Mellon's Informedia database, a large video library that supports indexing and retrieval with automated general descriptions, has described "Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results" [1].

## 3. Data Mining

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning,

visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines applied, such as neural networks, fuzzy and rough set theory, knowledge representation, inductive logic programming, or high performance computing. Depending on the kinds of data to be mined or on the data mining application. The data mining system integrate techniques from spatial data analysis, information retrieval, pattern recognition and image analysis [2].

## 3.1. Clustering and Cluster Analysis

Clustering is the process of grouping the data into classes or clusters. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

Cluster analysis is a multivariate method which aims to classify objects on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group [4].

### 3.2. K-Means Clustering Algorithm

The most well known and commonly used partitioning method is k-means and its variations. It is an algorithm to classify or to group your objects based on attributes/features into k number of group.

The k-means algorithm takes the input parameter k that is positive integer number. The grouping of data present by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of k-means clustering techniques, also refer to as reallocation methods, at attempt to optimize a given clustering by repeatedly reassigning objects to the cluster to which they are most similar.

The biggest advantage of the k-means algorithm in data mining application is its efficiency in clustering large data sets. However, its use is limited to numerical values. The computational complexity of the algorithm is O (T n k), where n is the total number of objects, k is the number of clusters, and T is the number of iterations [2].

**K-Means Algorithm is as follows:**

Input: Number of k clusters and database containing n objects.
Output: A set of k clusters.
Method:

(1) arbitrarily choose k objects as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the objects is the most similar, based on the mean value of the objects in the cluster;
(4) Update the cluster means, i. e. , calculated the mean value of the objects for each cluster;
(5) Until no change.

### 3.3. Euclidean Distance Measure

The Euclidean distance measure is frequently used as a distance measure, and is easy to use in two-dimensional planes. As the number of dimensions increases, the calculatibility time also increases.

$$d(i,j) = \sqrt{\left(\left|x_{i1}-x_{j1}\right|^2+\left|x_{i2}-x_{j2}\right|^2+...+\left|x_{ip}-x_{jp}\right|^2\right)}$$

The formula defines data objects j, with number of dimension and I equal to p. The distance between the two data objects d (i, j) is expressed in formula. $x_{ip}$ is the measurement of object i in dimension p [2].

## 4. Proposed System Overview

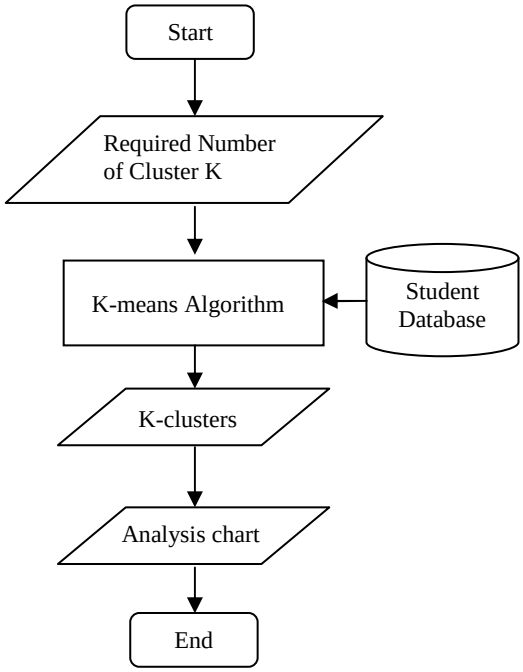


**Figure 1. Overview of the proposed system**

Figure 1 shows an overview of the proposed system. In the database of the system, the data of students from government technology high school are stored. The system takes an initial centroid for each cluster. The system calculates the distance of each record in the database by Euclidean distance and put each record in the cluster that has the minimum distances. Moreover, the new centroid for each cluster derived by calculating the means of all centroid value in each cluster.

This process continued until the distance of the old centroid and new centroid is equal. Finally, the system outputs the detail information of each cluster. If the user wants to analyze each cluster, the system can show the analysis chart of each cluster by marks.

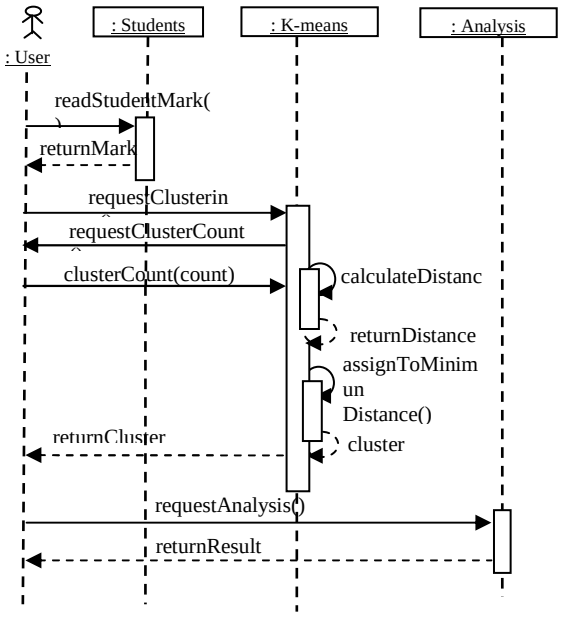


**Figure 2. Sequence diagram of the proposed system**

## 5. Experimental Results

In this section, we present the computing steps of algorithm and the experimental results of the system. The system implements by k-means algorithm, which the data mining task is to cluster the following eight points with (BT, AMT, MT, ET, ECT, BST) representing marks into three clusters.

There are six majors in GTHS. They are BT, AMT, MT, ET, ECT and BST. BT represents Building Technology major. AMT represents Auto Mechanics Technology major. MT represents Machining Technology major. ET represents Electrical Technology major. ECT represents Electronics Technology major. Finally, BST represents Building Services major respectively.

**Table 1. Example of student database**

| Roll-No | BT | AMT | MT | ET | ECT | BST |
|---|---|---|---|---|---|---|
| **1** | **60** | **47** | **68** | **75** | **81** | **71** |
| **2** | **50** | **61** | **81** | **60** | **76** | **80** |
| **3** | **65** | **63** | **75** | **55** | **57** | **85** |
| **4** | **55** | **52** | **80** | **49** | **43** | **54** |
| **5** | **45** | **48** | **74** | **82** | **64** | **75** |
| **6** | **70** | **74** | **51** | **65** | **72** | **80** |
| **7** | **50** | **63** | **55** | **63** | **49** | **55** |
| **8** | **75** | **58** | **45** | **41** | **68** | **63** |

Table 1 shows an example of student database as input data. Initially, we assign Roll-No 4, 6 and 8 as the center of each cluster. The following distance function is Euclidean distance for iterations of each centroid. Let SQRT be square root.

Iteration (1)

$d\ (4,1)=SQRT\ [(55-60)^2+(52-47)^2+(80-68)^2+(49-75)^2+(43-81)^2+(54-71)^2]$
$= 51.02$

$d\ (6,1)=SQRT\ [(70-60)^2+(74-47)^2+(51-68)^2+(65-75)^2+(72-81)^2+(80-71)^2]$
$= 25.69$

$d\ (8,1)=SQRT\ [(75-60)^2+(58-47)^2+(45-68)^2+(41-75)^2+(68-81)^2+(63-71)^2]$
$=47.58$

$d\ (4,2)=SQRT\ [(55-50)^2+(52-61)^2+(80-81)^2+(49-60)^2+(43-76)^2+(54-80)^2]$
$=44.64$

$d\ (6,2)=SQRT\ [(70-50)^2+(74-61)^2+(51-81)^2+(65-60)^2+(72-76)^2+(80-80)^2]$
$=38.86$

$d\ (8,2)=SQRT\ [(75-50)^2+(58-61)^2+(45-81)^2+(41-60)^2+(68-76)^2+(63-80)^2]$
$=51.42$

$d\ (4,3)=SQRT\ [(55-65)^2+(52-63)^2+(80-75)^2+(49-55)^2+(43-57)^2+(54-85)^2]$
$=37.93$

$d\ (6,3)=SQRT\ [(70-65)^2+(74-63)^2+(51-75)^2+(65-55)^2+(72-57)^2+(80-85)^2]$
$=35.67$

$d\ (8,3)=SQRT\ [(75-65)^2+(58-63)^2+(45-75)^2+(41-55)^2+(68-57)^2+(63-85)^2]$
$=42.73$

$d\ (4,5)\ =SQRT\ [(55-45)^2+(52-48)^2+(80-74)^2+(49$

$-82)^2+ (43-64)^2+ (54-75)^2]$
$=46.08$

d (6,5) $=SQRT [(70-45)^2+(74-48)^2+(51-74)^2+(65-82)^2+ (72-64)^2+ (80-75)^2]$
$=46.99$

d (8,5 ) $=SQRT [(75-45)^2+(58-48)^2+(45-74)^2+(41-82)^2+ (68-64)^2+ (63-75)^2]$
$=60.68$

d (4,7) $=SQRT [(55-50)^2+(52-63)^2+(80-55)^2+(49-63)^2+ (43-49)^2+ (54-55)^2]$
$=31.69$

d (6,7) $=SQRT [(70-50)^2+(74-63)^2+(51-55)^2+(65-63)^2+ (72-49)^2+ (80-55)^2]$
$=41.17$

d (8,7) $=SQRT [(75-50)^2+(58-63)^2+(45-55)^2+(41-63)^2+ (68-49)^2+ (63-55)^2]$
$=40.73$

**Table 2. Final three clusters**

| Cluster no | 1 | 2 | 3 |
|---|---|---|---|
| Roll no | 4,5,7 | 6,1,2,3 | 8 |
| BT | 50 | 61.25 | 75 |
| AMT | 54.33 | 59.75 | 58 |
| MT | 69.67 | 68.75 | 45 |
| ET | 64.67 | 63.75 | 41 |
| ECT | 52 | 71.5 | 68 |
| BST | 61.33 | 79 | 63 |

Table 2 shows the final three clusters as output. In this processing, cluster 1 is the most suitable with MT major and the least suitable with BT major. Cluster 2 is the most appropriate with BST major and the least appropriate with AMT major. Finally, cluster 3 is the most suitable with BT major and the least suitable with ET major.

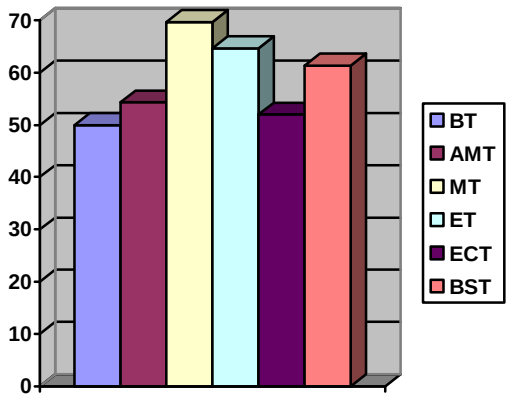


**Figure 3. Bar chart of cluster 1**

Figure 3 shows the results of cluster 1 by students' examination marks.

In this system, we can yearly test entrance exam marks of first year students to decide appropriate majors in Government Technology High School.

The person who is responsible for the management of educational plan can use the system to manage the students. The users can read the students 'database, can request the clustering of students and can check the bar chart of each cluster. When the user requests to cluster the students, the system asks the user to input the number of clusters. After giving the number of clusters, the system takes the random initials according to the number of cluster and starts the clustering process. After all, the system shows the resulted clusters to the user. If the user wants to analyze each cluster, the system shows the corresponding bar chart.

## 6. Conclusion

The major challenge of clustering is efficiently meaningful groups that concisely annotated. Students' data clustering is the automatic organization of students into cluster or groups so that students within a cluster have high similarity in comparison to one another, but are very dissimilar to students in other clusters. In this system, cluster analysis and k-means algorithm use in the field of education. Clustering the student's data are according to their exam marks. The system can analyze the student's data and output the analysis results. The proposed system provides for any government technology high school, which has the relationship between students' entrance examination result and their success.

## 7. References

[1] C.R.Palmer, J.Pesenti, R.E. Valdes-Perez, M.G. Christel, A.G. Hauptmann, D.Ng and H.D. Wactlar, Informedia Project, Carnegie Mellon University, Pittsburgh, PA 15213, valdes @ cs.cmu.edu.

[2] H. Jiawei and K. Micheline, "Data Mining Concepts and Technique", Second Edition.

[3] http://www. ripublication.com/acst.htm.

[4] Khaled Alsabti, Sanjay Ranka and Vineet Singh "An Efficient K-Means Clustering Algorithm".

[5] L. Robert Andrews, "Cluster Analysis", April 2005. Conference on Machine Learning, Pittsburgh, PA, 2006.

[6] N. V. Anand Kumar and G. V. Uma, "Improving Academic performance of Students by Applying Data Mining Technique", European Journal of Scientific

Research ISSN 2450-216X vol. 34 No. 4(2009), pp. 526-534, Inc. 2009.

[7] Rok Podgornik, Marijan Zafred and Anja Pajtler, “A Study of k-means Method where Starting Condition are Changed: Simulation Study”, vol. 1, No. 1, 204, 75-97.

[8] Rosie Cornish, “Cluster Analysis”, Mathematics Learning Support Center, 2007.

[9] Z. Senol Erdogan and T. Mehpare, “A Data Mining Application in a Student Database”, Journal of Aeronautics and Space Technologies, July 2005 vol. 2 Number 2(53-57).