ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse51832019.pdf
https://doi.org/10.30534/ijatcse/2019/51832019

Clustering Scholarship Programs Using Educational Data Mining Techniques

Gregorio Z. Gamboa, Jr.
Surigao State College of Technology, Surigao City, Philippines
greggy3659@yahoo.com

ABSTRACT

Major colleges and state universities assess students' quality and set different rewards for the various level in order to stimulate students' interest to study and participate in extracurricular activities. The main reward system that is used is the providing of financial incentives such as scholarship grants. In this paper, several data mining techniques such as clustering and time series analysis was integrated to discover and assess future outcomes and matters concerning scholarship offerings in Surigao State College of Technology (SSCT). The Student Financial Assistance Unit (SFAU) of SSCT holds all the records of scholarship grants and its grantees from June 2014 which was used as datasets. The study segmented every scholarship grants in SSCT to find patterns as to the scholarship grants and use it for the proliferation of data. It is suggested that the output of this study may be used as input and avenue for future researches using other data mining techniques.

Key words: ARIMA, EDM, Forecasting, Prediction

1. INTRODUCTION

Every university has its own methods to realize scholarship assessment. Scholarships are established by schools as a motivation for students that have outstanding and exceptional achievements. It is a form of incentive and benefit to inspire and embolden students and improve schools' learning principles [1]. From that point of view, an optimal way is considered to assess and distribute scholarship grants.

Data Mining is the process of extracting information from large data sets through the use of algorithms and methods drawn from the field of statistics and Database Management Systems [2]. Clustering, decision trees, genetic algorithms, Bayes classifiers, association rules, neural networks, and support vector machines to name some, are the algorithms and methods that can be used in data mining analysis that allow getting important information from the database [3].

One of the most popular data mining techniques is clustering. It represents an unsupervised learning method whose objective is to divide the data set so that the distance among the clusters should be minimal, whereas the inter-cluster distance should be maximal. One of the methods frequently used in data partition is the k-means data. The k-means algorithm is the easiest and the most common algorithm based on the squared error criterion [3]-[5].

The study clustered the indexed data of the different scholarship programs offered in Surigao State College of Technology, the only State College in the province of Surigao del Norte, Philippines as to the number of scholarship grantees using K-Means. The output of this study will provide the final groupings of scholarship programs that has similar traits with the others may it be due to number of grantees, or because of the profile of the grantees. This could be an avenue for future researches once a knowledge from groupings are extracted.

2. LITERATURE REVIEW

Data mining also termed as Knowledge Discovery in Databases (KDD), is a medium of discovering novel and potentially valuable information from large amounts of data [6]-[8]. The field of Educational Data Mining is fresh, new, and developing in the field of education sector which can also be applied in areas such as government, accounts, sports, transportation and a lot more [9]. This new emerging field concerns with developing methods that discover knowledge from data originating from educational environments. Data mining techniques such as decision trees, Naïve Bayes, K-Nearest neighbor, neural networks, K-means clustering and many others are instrumental in extracting data from the datasets [10].

Cluster analysis is considered as one of the most important techniques in data mining. It has attracted more and more attention in this big data era [11]. Data clustering is used to recognize patterns or clusters from a set of objects. In general, data clustering divides set of objects into groups or clusters where the objects in the same cluster or group are identical to each other than to objects from the other clusters [12].

A study analyzed crimes such as theft, homicide, and various drug offenses along with suspicious activities, noise complaints, and burglar alarm by using qualitative and quantitative approach [13]. Using K-means clustering data mining approach on a crime dataset from the New South Wales region of Australia, crime rates of each type of crimes and cities with high crime rates have been found.

Furthermore, the data mining techniques were implemented to understand specific trends and pattern of terrorist attacks in India. K-means clustering was used to determine the year wherein the terrorist groups were most active and also which terrorist group has affected the most [14].

On the other note, the literature on representations and distance measures for time-series, clustering, and classification is extensive [15]-[16]. Time series analysis method is a kind of data mining method, which is a sequence of data points, typically consisting of successive measurements made over a time interval. It is a method for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values [17].

Prediction can be made utilizing autoregressive integrated

moving average (ARIMA) algorithm that used historical data in predicting cases such as in education, society, climate, health, and others. In other countries, the use of ARIMA algorithm in forecasting incidence of hemorrhagic fever with renal syndrome in China was observed [18]. Meanwhile, the ARIMA model was also used in forecasting dengue hemorrhagic fever cases in Southern Thailand [19]. Further, the potential and effectiveness of using ARIMA modeling in the prediction of travel time to the urban roadway was also proven [20].

3. METHODOLOGY

3.1 Datasets

The data that were used are the following but are not limited as to wit: different scholarship grants that can be availed in SSCT, the gender of the grantees, and the number of grantees per scholarship grants per year.

3.2 Clustering

Clustering algorithms divide the group of objects into clusters, where objects in each cluster are similar to each other [21]. Indexed scholarship grants were grouped according to sex. Table 1 presents the datasets of indexed scholarship grants and its grantees according to sex from 2015 to 2017.

Table 1: Indexed scholarship grants in SSCT

	SCHOLARSHIP PROGRAM	MALE	FEMALE
1	Academic Scholarship	4	6
2	Athletic Scholarship	5	6
3	Barangay Scholarship	14	13
4	Choir	13	17
5	City Scholarship	147	247
6	CSSGP	7	15
7	Dance Troupe	5	5
8	ESGP-PA	61	167
9	LGU Basilisa	11	15
10	LGU Claver	11	19
11	PGMC (Platinum Group Metal Corporation)	7	7
12	Provincial Eskolaran	177	302
13	StuFAPS	9	21
14	Taganito Mining Corporation	27	32
15	Tulong Dunong 01	4	16
16	Tulong Dunong 02 (Cong. Bag-ao, Barbers/Romarate)	467	828
17	Tulong Dunong 02 (Cong. Matugas)	200	314

This algorithm is iterative and repeats for each object. It converges until the objects are stable. K-Means clustering is simple, and the necessary steps it follows are 1. Some clusters, K, is determined. 2. Assume a centroid or center of the K clusters. Any object can be randomly chosen and initialized as an initial centroid, or the first K objects can serve as the initial centroids. 3. The calculation of the distance of each object from each of the centroids. 4. Group the objects based on minimum distance (find the closest centroid for each object). Figure 1 presents the flowchart of the k-means algorithm.

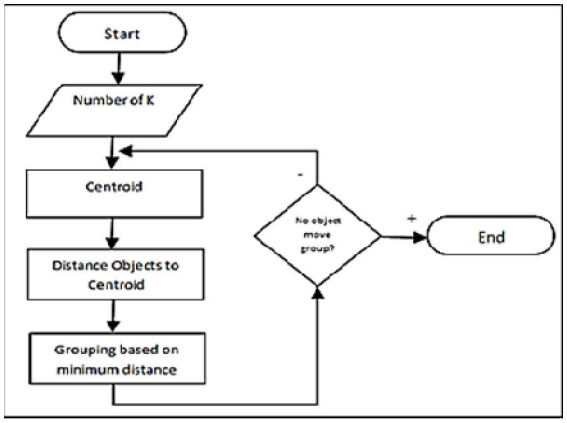


Figure 1: K-Means algorithm flowchart

Step 1: Initial value of centroids: Barangay Scholarship and ESGP-PA.

Let C1 and C2 denote the coordinate of the centroids, then C1= (14,13) and C2 = (61,167)

Step 2: Objects-Centroids distances denoted as OCD^0: The Euclidean distance is used to obtain the distance. The distance matrix at iteration 0 is showed in Table 2.

Table 2: Objects-Centroid distance: Iteration 0 denoted as $OCDI^0$

	SCHOLARSHIP PROGRAM	MALE	FEMALE
1	Academic Scholarship	12.2065556	170.7922715
2	Athletic Scholarship	11.4017543	170.4611393
3	Barangay Scholarship	0.0000000	161.0124219
4	Choir	4.1231056	157.4928570
5	City Scholarship	269.1560885	117.4563749
6	CSSGP	7.2801099	161.3071604
7	Dance Troupe	12.0415946	171.4059509
8	ESGP-PA	161.0124219	0.0000000
9	LGU Basilisa	3.6055513	160.0124995
10	LGU Claver	6.7082039	156.2177967
11	PGMC (Platinum Group Metal Corporation)	9.2195445	168.8668114
12	Provincial Eskolaran	331.7981314	177.9915728
13	StuFAPS	9.4339811	154.9838701
14	Taganito Mining Corporation	23.0217289	139.2156600
15	Tulong Dunong 01	10.4403065	161.4001239
16	Tulong Dunong 02 (Cong. Bag-ao, Barbers/Romarate)	932.4344481	775.7299788
17	Tulong Dunong 02 (Cong. Matugas)	353.8318810	202.3116408

Table 2 represents the indexed data. The third column refers to the distance of each scholarship grants with corresponding value to the first centroid. On the other hand, the last column is the distance of each object to the second centroid.

```
> results <- kmeans(mydatakmeans.features, 2)
> results
K-means clustering with 2 clusters of sizes 13, 4

Cluster means:
      MALE    FEMALE
1  13.69231  26.07692
2 247.75000 422.75000

Clustering vector:
 [1] 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 2

Within cluster sum of squares by cluster:
[1]  25065.69 287029.50
 (between_SS / total_SS =  67.5 %)
```

Figure 2: Clustering using K-Means algorithm in r studio

```
> results$cluster
 [1] 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 2
```

Figure 3: Result of their cluster function

Table 3: Object Clustering: Iteration 0 denoted as $OCLI^0$

SCHOLARSHIP PROGRAM	Group 1	Group 2
Academic Scholarship	1	0
Athletic Scholarship	1	0
Barangay Scholarship	1	0
Choir	1	0
City Scholarship	0	1
CSSGP	1	0
Dance Troupe	1	0
ESGP-PA	1	0
LGU Basilisa	1	0
LGU Claver	1	0
PGMC (Platinum Group Metal Corporation)	1	0
Provincial Eskolaran	0	1
StuFAPS	1	0
Taganito Mining Corporation	1	0
Tulong Dunong 01	1	0
Tulong Dunong 02 (Cong. Bag-ao, Barbers/Romarate)	0	1
Tulong Dunong 02 (Cong. Matugas)	0	1

Step 3: Objects clustering denoted as OCL0: Based on the result generated by K-Means function in R where each indexed scholarship grants was assigned based on the minimum distance, thus, Academic Scholarship, Athletic Scholarship, Barangay Scholarship, Choir, CSSGP, Dance Troupe, ESGP-PA, LGU Basilisa, LGU Claver, PGMC, StuFAPS, Taganito Mining Corporation, Tulong Dunong 01 were assigned to group 1. On the other hand, City Scholarship, Provincial Eskolaran, Tulong Dunong 02 (Cong. Bag-ao, Cong. Barbers/Romarate), Tulong Dunong 02 (Cong. Matugas) were assigned to group 2.

Step 4: Iteration – 1, determine centroids: Compute the new centroid of each group based on the new memberships. Group 1 has twelve members. Thus the centroid is the average coordinated among the twelve members: C1 = (117/12, 172/12). On the other hand, group 2 has five members. Hence the centroid is the average coordinated among the three members: C2 = (1052/5, 1858/5).

Step 5: Iteration 1, Objects-Centroid distances denoted as OCD1: Repeat the process in step number 2 to obtain the new distance matrix based on the new groupings. The results of the process are shown in Table 4.

Table 4: Objects-Centroid Distances: Iteration 1 denoted as $OCLI^1$

	SCHOLARSHIP PROGRAM	MALE	FEMALE
1	Academic Scholarship	6.7268120	280.2320467
2	Athletic Scholarship	6.0207973	279.7230773
3	Barangay Scholarship	5.5901699	269.1560885
4	Choir	7.6321688	266.1879036
5	City Scholarship	273.8179140	0.0000000
6	CSSGP	4.9244289	270.9686329
7	Dance Troupe	6.8007353	280.5851030
8	ESGP-PA	164.9128558	117.4563749
9	LGU Basilisa	4.9244289	268.9237810
10	LGU Claver	8.7321246	265.4806961
11	PGMC (Platinum Group Metal Corporation)	4.0311289	277.8488798
12	Provincial Eskolaran	336.4465039	62.6498204
13	StuFAPS	10.500000	264.8018127
14	Taganito Mining Corporation	28.0401498	246.2214450
15	Tulong Dunong 01	7.4330344	271.6799588
16	Tulong Dunong 02 (Cong. Bag-ao, Barbers/Romarate)	937.0540273	663.2955601
17	Tulong Dunong 02 (Cong. Matugas)	358.5990100	85.4283325

Step 6: Iteration-1, Objects Clustering denoted as OCL1. Assign each object based on the minimum distance. Based on the new distance matrix, the new group pattern presented in Table 5.

Table 5: Object Clustering: Iteration 1 denoted as $OCLI^1$

SCHOLARSHIP PROGRAM	Group 1	Group 2
Academic Scholarship	1	0
Athletic Scholarship	1	0
Barangay Scholarship	1	0
Choir	1	0
City Scholarship	0	1
CSSGP	1	0
Dance Troupe	1	0
ESGP-PA	1	0
LGU Basilisa	1	0
LGU Claver	1	0
PGMC (Platinum Group Metal Corporation)	1	0
Provincial Eskolaran	0	1
StuFAPS	1	0
Taganito Mining Corporation	1	0
Tulong Dunong 01	1	0
Tulong Dunong 02 (Cong. Bag-ao, Barbers/Romarate)	0	1
Tulong Dunong 02 (Cong. Matugas)	0	1

The results showed that $OCLI^1$ is equal to $OCLI^0$. Comparing the groups of the last iteration reveals that the objects are stable. Thus, the computation process of the k-mean clustering has reached its stability, and no more repetition is needed. The final grouping is presented in Figures 4-6.

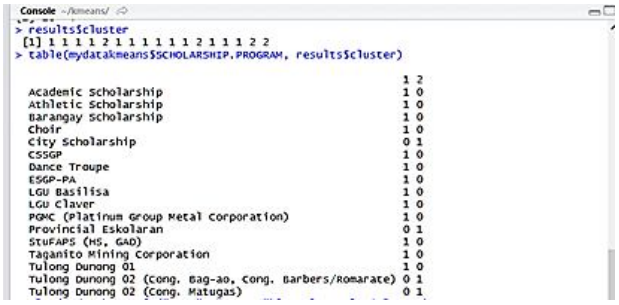


Figure 4: Clustered data

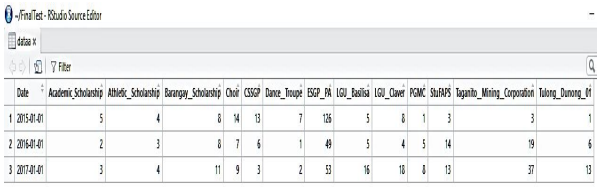
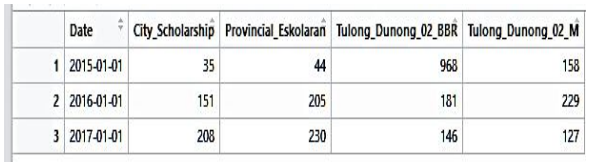


Figure 5: Indexed scholarship programs of cluster 1



	Date	City_Scholarship	Provincial_Eskolaran	Tulong_Dunong_02_BBR	Tulong_Dunong_02_M
1	2015-01-01	35	44	968	158
2	2016-01-01	151	205	181	229
3	2017-01-01	208	230	146	127

Figure 6: The indexed scholarship program of cluster 2

3.3 Trend Analysis

The R Language, which runs in RStudio software was utilized for the trend analysis of indexed data of the different scholarship grants offered in Surigao State College of Technology. Initially, the time series graph obtained in every cluster shows the behavior of indexed scholarship data. Figure 7

and 8 shows the time series graph of group one and two, respectively.

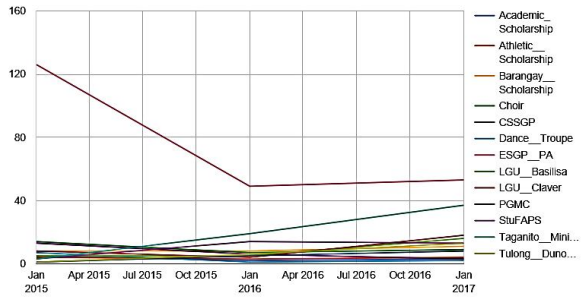


Figure 7: Time series plot for cluster 1

In group 1, ESGP-PA showed the highest number of grantees among the scholarship programs in Cluster 1. Moreover, it showed a very high mark during the year 2015 and showed a decreasing pattern starting from the year 2016 but has slightly increased in the year 2017.

On the other hand, Taganito Mining Corporation scholarship program showed an increasing pattern from the year 2015 to 2017 as shown in Figure 7 that makes it next to ESGP-PA as to the scholarship program with the most number of grantees in the cluster 1.

Furthermore, there is a declining pattern of many scholarship programs from 2015 to 2016 but then showed an increasing pattern from the year 2016 to 2017.

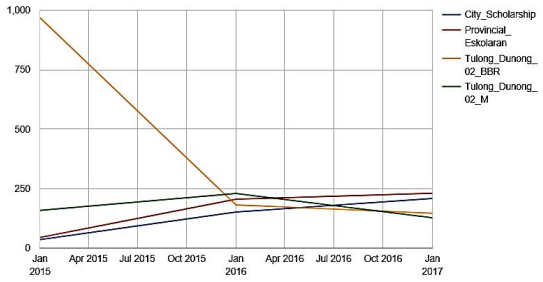


Figure 8: Time series plot for cluster 2

In group 2, Tulong Dunong 02 scholarship program by Cong. Bag-ao, Cong. Barbers and Cong. Romarate has the most number of grantees in the group in the year 2015. Later, it showed a decreasing pattern from 2015 to 2017.

On the contrary, City Scholarship and Provincial Eskolaran showed an increasing pattern from the year 2015 to the year 2017, which means an increase of grantees during that year was observed.

Moreover, Tulong Dunong 02 scholarship program under Cong. Matugas had an increasing pattern from the year 2015 to 2016 and a decreasing pattern from the year 2016 to 2017 as shown in Fig. 8.

4. CONCLUSION

Predicting grantees for each scholarship programs can be beneficial to the sponsoring agents since it will give them insight as to the number of their future grantees. This allows them to better prepare for budget allocation. However, in this study, the use of K-Means algorithm was used in finding patterns within scholarship programs that can be availed in SSCT. Simulation results showed that scholarship programs that belong to cluster one share the same traits that is why they are grouped as one. The same conclusion can be drawn in cluster 2. The results showed the effectiveness of the K-Means algorithm in segmenting educational datasets particularly scholarship programs. It is recommended that since groupings was already identified using K-Means algorithm, another knowledge extraction may be conducted using the same dataset as future research.

REFERENCES

[1] W. Wei, J. Han, J. Kong, and H. Xia, “Prediction of the Scholarship Using Comprehensive Development,” *Proc. - 4th Int. Conf. Enterp. Syst. Adv. Enterp. Syst. ES 2016*, pp. 183–188, 2017. https://doi.org/10.1109/ES.2016.30

[2] I. A. Khan and J. T. Choi, “An application of educational data mining (EDM) technique for scholarship prediction,” *Int. J. Softw. Eng. its Appl.*, vol. 8, no. 12, pp. 31–42, 2014.

[3] E. Susnea, “Using data mining techniques in higher education,” *High. Educ.*, vol. 1, no. 1, pp. 68–72, 1996.

[4] S. B. B, K. K. V, and A. N. Ahmed, “Semantically enriched Tag clustering and image feature based image retrieval system,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1, pp. 138–141, 2019.

[5] J. Goyal and B. Kishan, “Progress on Machine Learning Techniques for Software Fault Prediction,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 2, pp. 305–311, 2019. https://doi.org/10.30534/ijatcse/2019/33822019

[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[7] A. J. P. Delima, “An Experimental Comparison of Hybrid Modified Genetic Algorithm-based Prediction Models,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 1756–1760, 2019.

[8] A. J. P. Delima, A. M. Sison, and R. P. Medina, “Variable Reduction-based Prediction through Modified Genetic Algorithm,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 356–363, 2019. https://doi.org/10.14569/IJACSA.2019.0100544

[9] P. Kaur, M. Singh, and G. S. Josan, “Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector,” *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015. https://doi.org/10.1016/j.procs.2015.07.372

[10] S. K. Yadav and S. Pal, “Data Mining : A Prediction for Performance Improvement of Engineering Students using Classification,” *World Comput. Sci. Inf. Technol. J. WCSIT*, vol. 2, no. 2, pp. 51–56, 2012.

[11] C. Jinyin, L. Xiang, Z. Haibing, and B. Xintong, “A novel cluster center fast determination clustering algorithm,” *Appl. Soft Comput. J.*, vol. 57, pp. 539–555, 2017. https://doi.org/10.1016/j.asoc.2017.04.031

[12] G. Gan and M. K. P. Ng, “K-Means Clustering With Outlier Removal,” *Pattern Recognit. Lett.*, vol. 90, pp. 8–14, 2017. https://doi.org/10.1016/j.patrec.2017.03.008

[13] J. Agarwal, "Crime Analysis using K-Means Clustering," *Int. J. Comput. Appl.* , vol. 83, no. 4, pp. 975–8887, 2013.
https://doi.org/10.5120/14433-2579

[14] P. Gupta, A. S. Sabitha, and T. Choudhury, "Terrorist Attacks Analysis Using Clustering Algorithm," © *Springer Nat. Singapore Pte Ltd.*, pp. 317–328, 2018.
https://doi.org/10.1007/978-981-10-5547-8_33

[15] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks," *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '02*, p. 102, 2002.
https://doi.org/10.1145/775047.775062

[16] T. Warren Liao, "Clustering of time series data - A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.
https://doi.org/10.1016/j.patcog.2005.01.025

[17] W. Hongbin, D. Shuangyu, D. Jianzhuo, R. Ming, and D. Ming, "Study on Condition Pre-warning Method of Power Transformer based on Load Time Series Model *," pp. 223–227, 2015.
https://doi.org/10.1109/STA.2015.7505103

[18] Q. Liu, X. Liu, B. Jiang, and W. Yang, "Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model," 2011.
https://doi.org/10.1186/1471-2334-11-218

[19] S. Promprou, M. Jaroensutasinee, and K. Jaroensutasinee, "Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA Models," *Dengue Bull.*, vol. 30, pp. 99–106, 2006.

[20] D. Billings and Y. Jiann-Shiou, "Application of the ARIMA Models to Urban Roadway Travel Time Prediction-A Case Study. Systems, Man and Cybernetics, 2006. SMC'06," *IEEE Int. Conf.*, pp. 2529–2534, 2006.
https://doi.org/10.1109/ICSMC.2006.385244

[21] D. Hand, D. Hand, H. Mannila, H. Mannila, P. Smyth, and P. Smyth, *Principles of data mining*, vol. 30. 2001.