Mining the National Career Assessment Examination Result Using Clustering Algorithm

M V Pagudpud¹, T T Palaoag² and L M Padirayon³

1 Quirino State University, Quirino, Philippines
2 University of the Cordilleras, Baguio City, Philippines
3 Cagayan State University, Cagayan, Philippines
mvpagudpud@gmail.com

Abstract. Education is an essential process today which elicits authorities to discover and establish innovative strategies for educational improvement. This study applied data mining using clustering technique for knowledge extraction from the National Career Assessment Examination (NCAE) result in the Division of Quirino. The NCAE is an examination given to all grade 9 students in the Philippines to assess their aptitudes in the different domains. Clustering the students is helpful in identifying students’ learning considerations. With the use of the RapidMiner tool, clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), k-means, k-medoid, expectation maximization clustering, and support vector clustering algorithms were analyzed. The silhouette indexes of the said clustering algorithms were compared, and the result showed that the k-means algorithm with k = 3 and silhouette index equal to 0.196 is the most appropriate clustering algorithm to group the students. Three groups were formed having 477 students in the determined group (cluster 0), 310 proficient students (cluster 1) and 396 developing students (cluster 2). The data mining technique used in this study is essential in extracting useful information from the NCAE result to better understand the abilities of students which in turn is a good basis for adopting teaching strategies.

1. Introduction

Education is a fundamental sector in any nation [1] for it improves the productivity and quality of peoples’ lives [2]. Furthermore, education plays a crucial role in securing national economic and social growth by equipping people with the relevant skills to fortify competitiveness and to prevent unemployment [3]. For these reasons, authorities continue to discover and proposed possibilities to strengthen education by exploring and understanding the information gathered from the educational environment.

Educational data mining (EDM) is concerned with developing methods for exploring the unique types of data that come from the educational setting and using those methods to better understand students and the settings which they learn in [4]. Within data mining, clustering is the most popular technique, and in EDM, clustering has been used in various contexts [5].

Clustering is an unsupervised learning technique which the main goal is to determine the intrinsic grouping in a set of unlabeled data using an algorithm [6][7]. It is the task of assigning a set of objects into clusters so that the objects in the same cluster are more similar to each other than to those in other clusters [8]. The clustering problem is defined as a problem of classifying a group of data points into some clusters without any prior knowledge about data structure, to produce a concise representation of

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
the data [9]. The method is vital in data mining, statistical data analysis machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [10]. Moreover, the application of various clustering algorithm has been applied in many cases to the educational dataset in diverse studies [11].

Furthermore, clustering students is one of the frequent tasks in educational data mining. The objective is to create groups of students according to their customized features, and personal characteristics [12] and the clusters of students obtained can be used by the education administration to build a personalized learning system [13].

In the Philippines, the Department of Education conducts an annual National Career Assessment Examination (NCAE) nationwide that produces ample data set. To this day, there are hardly any researchers conducted about the result of NCAE and since the Philippine educational system tries to apply educational innovation to promote national development and global competitiveness [14], this research tries to discover novel and potentially useful information from the NCAE data.

There are eleven abilities measured by the NCAE. The scientific ability, reading comprehension, verbal ability, mathematical ability, logical reasoning ability, clerical ability, non-verbal ability, visual manipulative skill, Humanities and Social Sciences Science (HUMMS) ability, Science and Technology, Engineering and Math (STEM) ability and Accountancy, Business and Management (ABM) ability. The examination is intended to assess the preferred track of the students basing from their abilities on the different domains tested which will guide the students in choosing the track they are going to enroll in their senior high school. Since NCAE is a good test for aptitude, the result of the said examination can be a good basis to help students enhance their different abilities on the various domain tested. Hence, the conduct of this undertaking is necessary.

The main objective of this study is to cluster students based on the result of their NCAE to recommend groups of students that would lead to the general improvement of their abilities in the different areas tested by the said examination. To be able to obtain the best group for the students, clustering algorithm using the RapidMiner tool was used. The dataset was tested with an internal validation measure to select which among Density-Based Spatial Clustering of Applications with Noise (DBSCAN), k-means, k-medoid, expectation maximization clustering, and support vector clustering is the most suited algorithm for the data set. The clustering model produced by the appropriate clustering algorithm can be a basic resource in identifying students’ learning considerations in the division of Quirino.

2. Methodology

The methodology used in this study is described in this section.

2.1. Data Set

The data used in this study is the NCAE 2015-2016 result obtained from the division of Quirino. The data contains 1183 instances with 11 attributes. The following are the attributes included:

- Scientific Ability (SA)
- Reading Comprehension (RC)
- Verbal Ability (VA)
- Mathematical Ability (MA)
- Logical Reasoning Ability (LRA)
- Clerical Ability (CA)
- Non-verbal Ability (NVA)
- Visual Manipulative Skills (VMS)
- Humanities and Social Sciences Science (HUMMS) Aptitude
- Science and Technology, Engineering and Math (STEM) Aptitude
- Accountancy, Business, and Management (ABM) Aptitude

They are real values ranging from 1 to 100 as shown in table 1.
2.2. Clustering
Clustering is one of the techniques often used in analyzing data sets where each cluster formed is a collection of data objects that are similar to another place within the same cluster but dissimilar to objects in other clusters [15]. This study makes use of clustering analysis to segment students into groups according to their aptitude scores in the different domains tested by the NCAE.

There are various existing clustering algorithms which may all produce different segments on the same set of data. In this paper, the clustering algorithms are evaluated using internal validity measures to choose which of the algorithms is most suitable for the NCAE dataset. Five clustering algorithms were subjected to the internal evaluation which includes DBSCAN, k-means, k-medoids, expectation maximization (EM) clustering and support vector clustering (SVC). The RapidMiner tool was used to carry out clustering and internal validation.

2.3. Internal Validation Measure
Internal validation measure was used to choose the best clustering algorithm and the optimal cluster number without any external information [16]. In this study, the average silhouette width of DBSCAN, k-means, k-medoids, expectation maximization and support vector clustering algorithms were compared. The following is the process of computing the silhouette index [17]:

For a given cluster, $X_j$ ($j = 1, \ldots, c$), the silhouette technique assigns to the $i$th sample of $X_j$ a quality measure, $s(i) = (i = 1, \ldots, m)$, known as the silhouette width. This value is a confidence indicator on the membership of the $i$th sample in the cluster $X_j$ and it is defined as:

$$s(i) = \frac{(b(i) - a(i))}{\max(a(i), b(i))}$$

(1)

where $a(i)$ is the average distance between the $i$th sample and all of samples included in $X_j$; $b(i)$ is the minimum average distance between the $it$ sample and all of the samples clustered in $X_k (k = 1, \ldots, c; k \neq j)$.

The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. If the silhouette has a high value, then the clustering configuration is appropriate.

3. Results and Discussion
After executing the process, the clustering algorithms produced clustering models and average silhouette values as shown in Table 2. The DBSCAN clustering produced three clusters, but its silhouette index is -0.063. A negative silhouette index is undesirable. The K-means algorithm where $k = 3$ has 0.196 silhouette index from three clusters. K-medoid ($k=3$) and EM have 0.178 and 0.090 silhouette indexes respectively, and SVC produced two clusters with 0.093 silhouette index. The silhouette indices computed were near 0 which indicates that some of the samples are very close to the decision boundary between two neighboring clusters.

Nevertheless, since the larger value of the silhouette index indicates a better clustering model, we can conclude that K-means provided the best model as compared to the other clustering algorithms tested. Cluster 0, cluster 1 and cluster 2 produced by the k-means algorithm were composed of 477, 310 and 396 students respectively.

The centroid plot produced by the k-means algorithm, where $k = 3$ is shown in Figure 1. The centroid plot is used to inspect the clustering results. It can be inferred from the figure that the students who belong to Cluster 0 (blue line) have average scores in almost all of the abilities tested as compared to

|   | SA  | RC  | VA  | MA  | LRA | CA  | NVA | VMS | HUMMS | STEM | ADM |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-------|------|-----|
| 1 | 80  | 66  | 14  | 36  | 36  | 82  | 54  | 93  | 85    | 36   | 54  |
| 2 | 60  | 18  | 51  | 43  | 83  | 38  | 86  | 88  | 36    | 48   | 35  |
| 3 | 60  | 79  | 62  | 51  | 11  | 87  | 95  | 83  | 56    | 80   | 60  |
the other two clusters. This group is called the ‘determined.’ They scored best on the clerical ability and non-verbal ability tests, but they scored low in mathematical ability and logical reasoning ability tests.

| Clustering Algorithm | Cluster 0 | Cluster 1 | Cluster 2 | Silhouette Index |
|----------------------|-----------|-----------|-----------|------------------|
| DBSCAN               | 1051      | 115       | 17        | -0.063           |
| K-Means (k = 3)      | 477       | 310       | 396       | 0.196            |
| KMedoid (k = 3)      | 241       | 414       | 528       | 0.178            |
| SVC                  | 373       | 810       | 0         | 0.093            |
| EM (k = 3)           | 788       | 151       | 244       | 0.090            |

Moreover, students who belong to cluster 1 (red line) had the best performances in all the abilities tested compared to the other clusters. Their group is called the ‘proficient.’ On the other hand, cluster 2 (green line) represents the students who had worst performances as compared to the other clusters. These students achieved the lowest result on clerical ability, HUMSS ability, and STEM ability tests. These students belong to the ‘developing’ cluster.

With this result, it would be better to apply separate learning strategies to the different clusters of students. Table 3 presents the summary of the result of the clustering technique and the students’ learning considerations recommended.

4. Conclusion
This study focuses on data mining by implementing clustering algorithm on the NCAE dataset using RapidMiner tool to come out with the clusters of students. By applying an internal validation measure which is the silhouette index measure, the result showed that the K-means algorithm, using k = 3, provided the best clustering model to cluster students having a value of 0.196. Three groups were formed having 477 students in the determined group (cluster 0), 310 proficient students (cluster 1) and 396 developing students (cluster 2). The groups presented in the clustering model was produced using the k-means clustering algorithm can be used by school administration in promoting students collaborative
learning, effective group training and personalized learning system in the division of Quirino. More importantly, the data mining technique used was essential in extracting valuable information from the NCAE result which is a good basis for building a learning system which can supplement student’s learning. It is recommended that students’ learning should consist of various learning content structure, more presentations, simulations, and exercises and should contain a reward mechanism for accomplishments to encourage student engagement.

**Table 3. Summary of the clustering result and the students’ learning considerations**

| Cluster      | Clustering Technique Result                        | Learning Considerations                                                                 |
|--------------|---------------------------------------------------|-----------------------------------------------------------------------------------------|
| Cluster 0    | - High in clerical                               | - More learning content structure in mathematics & logical reasoning                     |
| **Determined Group** | - High in non-verbal ability                       | - Extra learning content structure in clerical & non-verbal abilities                    |
|              | - Low in mathematical ability                     | - Supplementary learning content structure in other abilities                            |
|              | - Low in logical reasoning                        | - reward accomplishments                                                                |
|              | - Average in other abilities                      |                                                                                        |
| Cluster 1    | - Very high in Scientific Ability                 | - Supplementary learning content structure in other abilities                            |
| **Proficient Group** | - Very high in Clerical Ability                    | - reward accomplishments                                                                |
|              | - Very high in HUMMS                              |                                                                                        |
|              | - High in others abilities                         |                                                                                        |
| Cluster 2    | - Lowest in clerical ability                      | - More learning content structure in all areas                                          |
| **Developing Group** | - Lowest in HUMSS ability                          | - More presentations, simulations, and exercises learning contents in all areas         |
|              | - Lowest in STEM ability                          |                                                                                        |
|              | - Low in other abilities                           |                                                                                        |

For future works, the researchers intend to gather additional data on recent NCAE results to be able to build a model for forecasting students’ performance in the NCAE in the division of Quirino; and continue to pursue researches in data mining as applied to the educational context.

**Acknowledgement**

The authors extend their gratitude to the Quirino State University and the Department of Education, Division Office of Quirino for their support to the study providing all the necessary resources.

**References**

[1] Orji K E and M. Job 2013 The role of education in national development: Nigerian experience *European Scientific J.* 9(28).

[2] Taiwo R O A and Tunde A J 2016 Impact of education on national development *Int. J. of Advanced Academic Research* 2(1) (Springer Berlin Heidelberg) pp 323-37.

[3] Robert A T 2015 *Relevance of Vocational and Technical Education* [Online] Available: http://www.ghanaweb.com/GhanaHomePage/NewsArchive/Relevance-of-vocational-and-technical-education-353126. [Accessed 20 April 2017].

[4] Baker R S d 2010 Mining data for student models *Advances in Intelligent Tutoring Systems* (Springer Berlin Heidelberg) pp 323-37.

[5] Mohamad S K and Tasir Z 2013 Educational data mining: a review *The 9th Int. Conf. on Cognitive Science* (Malaysia).
[6] IndiraPriya P and Ghosh D K 2013 A survey on different clustering algorithms in data Int. J. of Modern Eng. Research 3(1) pp. 267-74.

[7] Feng Y 2012 Analysis on algorithm and application of cluster in data mining J. of Theoretical and Applied Inform. Technol. 46(1).

[8] Sharma N, Bajpai A and Lito R 2012 Comparison of the various clustering algorithms of weka Int. J. of Emerging Technol. and Advanced Eng. 2(5).

[9] Roseline R, Jenitha G and Amirhtaraj H 2014 Analysis and application of clustering techniques in data mining Int. J. of Computing Algorithm 03 pp. 910-2.

[10] Prasad J R 2014 Pattern recognition: possible research areas and issues Int. J. of Comp. Sci.and Network 3(5).

[11] Dutt A, Aghabozrgi S, Ismail M A B and Mahroeian H 2015 Clustering algorithms applied in educational data mining Int. J. of Information and Electronics Eng. 5(2).

[12] Vukićević M, Kirchner K, Delibašić B, Jovanović M, Ruhland J and Suknović M 2013 Finding best algorithmic components for clustering microarray data Knowledge and Information Syst. 35(1) pp 111-30.

[13] Goyal M and Vohra R 2012 Applications of data mining in higher education Int. J. of Computer Sci. Issues 9(2).

[14] Escudero A G 2013 House Bill No.2204 [Online] Available: http://www.congress.gov.ph/legisdocs/basic_16/HB02204.pdf. [Accessed 20 April 2017].

[15] Veeramuthu P, Periyasamy R and Sugasini V 2014 Analysis of student result using clustering Int. J. of Computer Sci. and Information Technol. 5(4).

[16] Liu Y, Li Z, Xiong H, Wu J, Gao X and Wu S 2013 Understanding and enhancement of internal clustering validation measures IEEE Transactions On Cybernetics 43(3).

[17] Rendon E, Abundez I M, Gutierrez C, Diaz S Z, Arizmendi A, Quiroz E and Arzate E 2011 A comparison of internal and external cluster validation indexes Applications of Math. and Computer Eng.