ISBN: 978-981-18-1791-5
2021 the 11th International Workshop on Computer Science and Engineering (WCSE 2021)
doi: 10.18178/wcse.2021.02.009

Cluster Analysis of Myanmar Census Data using Hybrid Algorithm

Nway Yu Aung, Kyawt Kyawt San and Swe Zin Hlaing [+]

University of Information Technology (UIT), Yangon, Myanmar

Abstract. Clustering, one of the data-mining components is very useful for analyzing valuable information and has been applied in many application areas. The census dataset is collected from the 2014 Myanmar population and Housing Census and the key objective of the 2014 Census is to provide important information about the population to Government and other stakeholders in terms of demographic, educational, social, and economic characteristics, living conditions, and household amenities. To avoid the performance and cluster quality issues, this paper proposed the hybrid clustering algorithm that combine with Partition Around medoids (PAM), and one of the metaheuristics algorithms, Bat where the Bat is adjusted to solve the data cluster problem to locate multiple optimal medoids based on the multimodal search capability of the Bat. In order to handle large volume of data, Apache Spark parallel framework has introduced to run the proposed algorithms. Experimental results show that the proposed algorithm takes a significantly reduce time in computation with comparable performance against the PAM for large dataset. At the same time, the cluster quality of the proposed system is evaluated using silhouette validation, and observed that the proposed algorithm performed well.

Keywords: clustering, PAM, Bat, Apache Spark, Silhouette

1. Introduction

Data mining, interdisciplinary computer science, and statistics subfield is the overall objective of extracting information from datasets and converting information into a comprehensible structure for later use. The use of DM in the field of education is incipient and gives rise to a research area of Educational Data Mining (EDM). EDM is a paradigm for developing tasks, models, methods, and algorithms for exploring data from educational environments. In the field of EDM, various techniques have been developed to apply and test. Some of the useful techniques include association rule mining, statistical methods, data visualization, classification, and clustering [1]. Different clustering algorithms as applied to the EDM context. Clustering is a data mining technique that allows the aggregation of large amounts of data by creating meaningful groups or categories of objects. There are many variants of the clustering algorithms family: k-means, hierarchical, DBSCAN, spectral, gaussian, and birch are a few.

Many researchers research the clustering techniques in EDM. The clustering approach in data mining, which analyses the use of the $k-means$ algorithm to enhance student academic performance in higher education, and introduces the $k-means$ clustering algorithm as a simple and effective method for monitoring student performance progression [2]. Traditional $k-means$ clustering algorithm [3] and the Euclidean distance measurement of similarity were chosen to be used in the student scores analysis. The medoids-based algorithm is more stable and simpler to implement than other clustering algorithms [4]. But, one of the biggest problems using the traditional algorithm is time execution for large datasets. So, researchers proposed the clustering algorithm on parallel computing frameworks. The $k-medoids$ parallel algorithm based on MapReduce was introduced in the reference literature [5] and solved the efficiency problem of dealing with large data. A new algorithm proposed for the PAM extension of MapReduce. During the map phase, it assigns each object to the closest medoid and updates the current medoids in the reduce phase using pair-wise computation [6]. Another problem with the clustering algorithm based on medoids is the choice of the initial medoids [7]. Some researchers have tried to obtain accurate medoids of cluster algorithms. Metaheuristic algorithms such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) combining the clustering for optimization [8]. For solving

[+] Corresponding author.
E-mail address: nwayuaung@uit.edu.mm.

the execution time problem, this paper proposed a parallel algorithm applying on the Apache Spark [9]. The authors solved the initial medoids problem by the implementation of the hybrid (Partition around medoids) PAM-Bat algorithm [10]. In the hybrid algorithm, the Bat algorithm was finding the optimal initial medoids. These optimal medoids are used in the PAM algorithm. In this paper, this system aimed to implement the Myanmar Census Data on the hybrid method and test for cluster quality. The rest of this paper is arranged according to this. The second section presents the dataset description and hybrid algorithm in detail. And then, the following section discuss the experimental results. Finally, the system concludes to draw some conclusions.

2. Research Methodology

The proposed method involves three steps. During the first step, a pre-processing technique is adopted that transforms raw data. The pre-process dataset of the system before applying the hybrid algorithm. The hybrid method is suggested in the second step. The performance of the clustering algorithm can vary depending on the method used to select the initial medoids. The PAM algorithm selects the best initial medoids by helping the Bat Algorithm. The hybrid method applying the Apache Spark for parallelization. The final step validates the accuracy of the clustering's results.

2.1. Dataset Description

Myanmar has a population of about 51.5 million people in 15 regions of the state and is subdivided into 330 townships. The 2014 Myanmar population and Housing Census was conducted as the reference point at midnight on 29 March 2014. It is the first 30-year Census, the last was carried out in 1983. The former Ministry of Immigration and Population, now the Ministry of Employment, Immigration, and Population, oversaw the preparation and execution of this Census on behalf of the Government in compliance with the Population and Housing Census act 2013. The 2014 census was included questions about household assets for the first time in Myanmar history, according to the Ministry of Immigration and Population. The nationwide population and housing survey covered information on demography, fertility, education, employment, migration history, and household assets. Census data collection was achieved using scanning technology. 110,000 enumerators visited more than 12 million households to collect data and provide citizens and households. The system has been tightly optimized, with controls to guarantee the accuracy of results.

2.2. Hybrid Clustering Algorithm

One of the popular medoids-based partitioning algorithms is the Partition Around Medoids algorithm (PAM), which is simple to implement and more robust than other partitioning algorithms. The PAM algorithm, however, has two disadvantages. Time complexity is the first drawback, and the second drawback is initialized medoids randomly. The system is trying to improve the weak point of PAM. By combining the Bat optimization algorithm, the PAM-Bat method selects the best initial medoids. The first step of this algorithm is to define the bats number using objective function $f(x)$. These bats will be initialized as x_i and v_i. The x_i (position) of each virtual bat defines the medoids of a cluster. Define f_i (pulse frequency) at x_i also, initialize r_i (pulse rate) and A_i (loudness). The t (iterations) will be started from 1 and the maximum iteration value $iter_{max}$will be assigned. While the iteration value is less than the maximum iteration the function will get started to update the velocities, locations, and frequency of all the bats. Each artificial bat uses (1) to select a f_i in the range of frequency $[f_{min}, f_{max}]$.

$$f_i = f_{min} + (f_{max} - f_{min})\beta \tag{1}$$

where f_{min} and f_{max} are two integers in the range and also uses (2) to update velocity.

$$v_{ij}^{t} = v_{ij}^{t-1} + \left(x_{ij}^{t-1} - x^{*}\right)f_i \tag{2}$$

By applying the result in the (3) to calculate the next position.

$$x_{ij} = x_{ij}^{t-1} + v_{ij}^{t} \tag{3}$$

The k initial clusters medoids will be initialized after assigning each bat of x_i, where n (number of objects). Each cluster medoids updated that have been assigned as closest to it. By using (4) to (7), the Euclidean distance is calculated.

Case 1: O_i belongs to representative object O_j and $d(O_i, O_{i2}) < d(O_i, O_k)$:

$$C_{ijk} = d(O_i, O_{i2}) - d(O_{ik}, O_j) \quad (4)$$

Case 2: O_i belongs to representative object $O_j$and $d(O_i, O_k) < d(O_i, O_{i2})$:

$$C_{ijk} = d(O_i, O_k) - d(O_i, O_j) \quad (5)$$

Case 3: O_i belongs to representative object O_{i2} and $d(O_i, O_{i2}) < d(O_i, O_k)$:

$$C_{ijk} = 0 \quad (6)$$

Case 4: O_i belongs to representative object O_{i2} and $d(O_i, O_k) < d(O_i, O_{i2})$:

$$C_{ijk} = d(O_i, O_k) - d(O_i, O_{i2}) \quad (7)$$

The random value $(rand)$ of the other bats is greater than the pulse rate r_i. Selecting the solution among the best solutions generates a local solution around the selected best solution as follows in (8).

$$X_{new} = X_{old} + \varepsilon A_t \quad (8)$$

Where $rand$ is a random number $\in [0,1]$ and ε another random number $\in [-1,1]$, while $A_t =< A_t >$ is the average loudness of all the bats at the current generation. For each iteration of the algorithm, the loudness A_i and the emission r_i are updated using respectively (9) and (10).

$$A_i(t+1) = \alpha A_i(t) \quad (9)$$

$$r_i(t+1) = r_i(0)[1 - \exp(-\gamma t)] \quad (10)$$

Where α and γ are constants, α in [0,1] and $\gamma > 0$. After the random fly of the bats produces a new solution, the function ends. This value is less than the pulse rate, the new solution accepted. The pulse rate higher than the loudness rate and finally the best bat is found. Then the initial value is increased and the process continue until it gets the specified initial medoids.

2.3. Parallel Hybrid Algorithm Using Apache Spark

The hybrid algorithm is proposed to solve large data by distributing and parallel processing of data on different nodes. Bat movement and fitness calculations were implemented to improve the bat algorithm's ability for large data sets and population updates. Upon completion of the Bat movement and fitness, update each bat's information by combining the two output files and then send the bats to the next iteration. Both these operations are carried out in two phases. The driver program shall send the tasks to the executors in the first stage. Additionally, all particles are sent to the executors via a broadcast variable via a cluster manager. That executor reads the data record portion which is encapsulated in a Resilient Distributed Datasets (RDDs). It should be noted that for the next iterations, executors only read a portion of data instances and cache data instances in their memory. RDDs is stored the data and distributed in the computing cluster between the worker nodes. These RDDs make it possible to operations and transformations on the data in parallel. The data that is to be clustered in RDDs in the hybrid algorithm implemented. Setting the initial cluster medoids for each particle, a bat population is initialized randomly from the dataset. To compute a fitness, each cluster medoids in the particle needs to be compared to each data point.

3. Experiments and Results

Hybrid clustering algorithms work on the data where all attributes are either numeric or categorical data. But these methods are effective to numeric data. Also, Apache Spark is a highly scalable platform that accepts only numeric data for clustering.

Myanmar Census dataset includes over 2,400,000 different records and 44 attributes Dataset is complete, there is no missing value in any record. The dataset. have 44 attributes most of the attributes are nominal (categorical) features. Therefore, there may be a need to convert to a numeric. So, use the one-hot Encoding technique. One hot encoding creates a new variable for each level of a categorical feature. Take an example of that in Table 1 to better understand this. Suppose the dataset has different variables such as Currently Attending, Previously Attended, with a School Attendance attribute. After encoding, it has variables each representing a category in the feature. Now for each category that is present, we have 1 in the column of that category and 0 for the others.

Table 1: Data Transformation Using one-hot encoder

Index	School Attendance
0	Currently Attending
1	Currently Attending
2	Previously Attended
3	Currently Attending
4	Previously Attended

One-hot Encoder

Index	Currently Attending	Previously Attended
0	1	0
1	1	0
2	0	1
3	1	0
4	0	1

3.1. Performance Analysis

The Spark cluster comprises 4 worker nodes, one worker node that operates on the master node. The configuration of all 4 nodes is the same, i.e., Intel® Core™ i7-8565U CPU @ 1.8 GHz with 8 GB RAM and Ubuntu 14.04 LTS Operating System, see Table 2.

Table 2: Apache Spark Computational Environment

Workers		Executor		No: of core	Computational framework	Language used for coding	Distributed storage system
Number	4	Number	4	4	Apache Spark 2.0.0	Java Programming Language	Hadoop Distributed File System
Memory	2G	memory	2G				

Next, the system executes the time for the Hybrid algorithm by using different datasets size as shown in Fig. 1. As seen in the Fig. 2, the performance of the PAM and the hybrid method is very similar to each other. Although, the PAM slightly better than the proposed for small dataset size. But the computational time require for PAM, increases rapidly as the dataset size increases. Although, the proposed system takes about the nearly constant time.

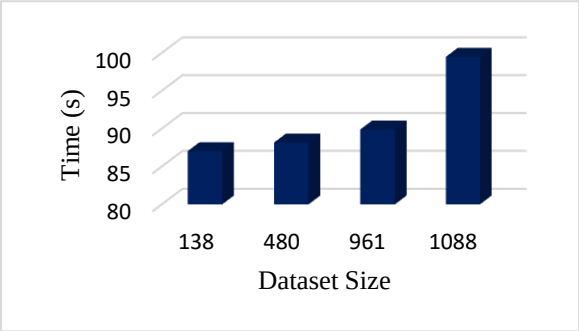

Fig. 1: Execution time of Hybrid with different dataset size

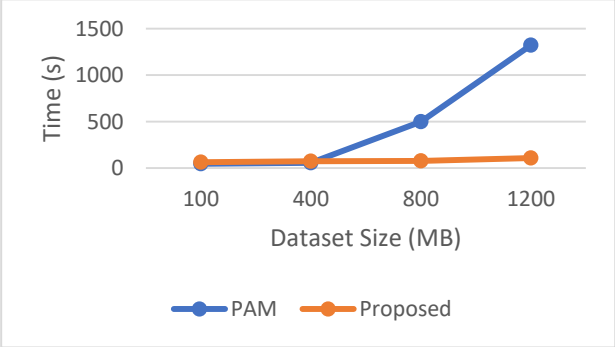

Fig. 2: Execution time of PAM and proposed Hybrid algorithm

3.2. Measuring Custer Quality

The proposed system used Silhouette as the cluster validation measure to test the clustering results. The silhouette approach offers a measure of how close the data is to the assigned cluster as compared to other clusters. For each data point, this is computed by calculating the silhouette value, and then averaging the result across the entire data set. The Silhouette Coefficient is calculated as the following equation (11).

$$S\_i = ((b\_i - a\_i))/\max(a\_i, b\_i) \quad (11)$$

The values of the silhouette between 0.7 and 1.0 indicate clustering results with excellent cluster separation. The PAM chooses initial medoids randomly and the proposed method is calculated the initial medoids itself see in Table 3. Table 4 illustrates the accuracy versus dataset comparison results. Based on the result, the proposed algorithm gives better accuracy for all the datasets, but standard PAM gives degrades accuracy for the large dataset.

Table 3: Parameters for Two Algorithms

Algorithm	Data	Number of clusters	Initial medoids
Standard PAM algorithm	Input by user	Input by user	Input randomly
Hybrid algorithm	Input by user	Input by user	Calculated by algorithm

Table 4: Comparison between PAM and Hybrid PAM-BAT using Silhouette score

Dataset Size (MB)	PAM	Hybrid PAM-Bat
500	0.6192	0.6587
1000	0.49930	0.6411
1500	0.4083	0.6833

4. Conclusions

In this paper, an algorithm that is designed for clustering large datasets, namely PAM-Bat with parallel, is a robust optimization technique. It is based on the behaviour of real Bats for the optimization aspect, especially for their more attractive collective intelligence. It also uses PAM clustering to create the clusters. The performance of the proposed algorithm and standard PAM algorithms have been compared using Myanmar Census data and observed that the proposed algorithm performed well, to achieve a good quality of clusters. The parallel algorithm using Apache Spark also resolves the execution time problem of the traditional PAM algorithm as a consequence of the experiment.

5. Acknowledgments

The authors would like to thank Dr. Swe Zin Hlaing and Dr. Kyawt Kyawt San for helping and giving valuable advice. The authors would also like to thank the reviewers and the Associate Editor for the extensive comments.

6. References

[1] Romero C, Ventura S, Garc á E. Data mining in course management systems: Moodle case study and tutorial. Computers & Education. 2008 Aug 1;51(1):368-84.

[2] Saxena PS, Govil MC. Prediction of student's academic performance using clustering. InNatl. Conf. Cloud Comput. Big Data 2009.

[3] Oyelade OJ, Oladipupo OO, Obagbuwa IC. Application of k Means Clustering algorithm for prediction of Students Academic Performance. arXiv preprint arXiv:1002.2425. 2010 Feb 11.

[4] Bhat A. K-medoids clustering using partitioning around medoids for performing face recognition. International Journal of Soft Computing, Mathematics and Control. 2014 Aug;3(3):1-2.

[5] Zhang X, Gong K, Zhao G. Parallel K-Medoids algorithm based on MapReduce. Journal of Computer Applications. 2013 Apr;33(4):1023-5.

[6] Xinxiang H, Henan X. A new data mining algorithm based on MapReduce and Hadoop. International Journal of Signal Processing, Image Processing and Pattern Recognition. 2014;7(2):131-42.

[7] Park HS, Jun CH. A simple and fast algorithm for K-medoids clustering. Expert systems with applications. 2009 Mar 1;36(2):3336-41.

[8] Inkaya T, Kayalıgil S, Özdemirel NE. Swarm intelligence-based clustering algorithms: A survey. In Unsupervised learning algorithms 2016 (pp. 303-341). Springer, Cham.

[9] Aung NY, Mon AC, Hlaing SZ. Performance Analysis of Parallel Clustering on Spark Computing Platform. In The 2nd International Conference on Advanced Information Technologies 2018.

[10] Aung NY, San KK, Hlaing SZ. Hybrid Partition Around Medoids Algorithm for Large Volume of Data. In2019 International Conference on Advanced Information Technologies (ICAIT) 2019 Nov 6 (pp. 280-285). IEEE.