OPEN ACCESS
Indo-JC Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 6, Issue. 1, April 2021. pp. 53-60
doi:10.34818/indojc.2021.6.1.547

Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J

Arief Chaerudin[1], Danang Triantoro Murdiansyah[2], Mahmud Imrona[3]

[1, 2, 3]*School of Computing, Telkom University*
Bandung, Indonesia

[1] ariefch@students.telkomuniversity.ac.id
[2]danangtri@telkomuniversity.ac.id
[3]mahmudimrona@telkomuniversity.ac.id

Abstract

In the era of data and information, data has become one of the most useful and desirable things. Data can be useful information if the data is processed properly. One example of the results of data processing in business is by making customer segmentation. Customer segmentation is useful for identifying and filtering customers according to certain categories. Analysis of the resulting segmentation can produce information about more effective target market, more efficient budget, more accurate marketing or promotion strategies, and much more. Since segmentation aims to separate customers into several categories or clusters, a clustering algorithm can be used. In this research, customer segmentation is carried out based on the value of income and value of expenditure. The categorization method that will be used for this research is to use the K-Means ++ algorithm which is useful for determining clusters of the given data. In this study, the implementation of K-Means ++ is carried out using Neo4J. Then in this research, a comparison of K-Means ++ and K-Means is carried out. The result obtained in this study is that K-Means ++ has a better cluster than K-Means in term of silhouette score parameter.

Keywords: K-Means++, K-Means, Neo4J, customer segmentation.

Abstrak

Di era data dan informasi saat ini, data menjadi salah satu hal yang paling berguna dan paling diminati. Data dapat menjadi informasi yang berguna apabila dilakukan pemrosesan data dengan tepat. Salah satu contoh hasil pengolahan data dalam bisnis yaitu dengan membuat segmentasi pelanggan. Segmentasi pelanggan bermanfaat untuk mengenali dan memfilter pelanggan sesuai dengan kategori tertentu. Analisis dari segmentasi yang dihasilkan dapat menghasilkan informasi target pasar yang lebih efektif, anggaran dana yang lebih efisien, strategi marketing atau promosi yang lebih akurat, dan masih banyak lagi. Karena segmentasi bertujuan untuk memisahkan pelanggan ke beberapa kategori atau *cluster,* maka algoritma *clustering* dapat digunakan. Pada tugas akhir ini akan dilakukan segmentasi pelanggan berdasarkan nilai pendapatan dan nilai belanja. Metode kategorisasi yang digunakan untuk penelitian ini adalah dengan menggunakan algoritma K-Means++ yang berfungsi untuk menentukan *cluster* dari data yang diberikan. Dalam penelitian ini implementasi K-Means++ dilakukan dengan menggunakan Neo4J. Kemudian pada penelitian ini juga dilakukan perbandingan K-Means++ dengan K-Means. Hasil yang didapatkan pada penelitian ini adalah bahwa K-Means++ memiliki *cluster* yang lebih baik dibandingkan K-Means jika dilihat dari parameter silhouette *score*.

Kata Kunci: K-Means++, K-Means, Neo4J, segmentasi pelanggan

Received on January, 2021. Accepted on Mei, 2021

I. Introduction

Nowdays, in information's era, data are one of the things that are created and needed by everyone. Most companies treat data as their most important assets, these data are used for many things, including decision making, CRM (Customer Relationship Management), generating statistical reports, understanding customers, and much more. Data must be processed before can be used to get the information needed. One way to process data is to use the clustering method, which is to divide the data given into several groups (clusters). One of the popular clustering algorithms is K-Means. K-Means makes clusters by calculating the distance between each data object with the middle value after each iteration [1]. In the customer segmentation process, the customer data will be filtered, and clustering all customers into several clusters according to characteristics of specified categories. By doing the segmentation, the coverage of products sold will be in accordance with the target customers.

By using graph modeling, it is easier to see relationship between data. Graph database uses graph modeling. In graph database, nodes represent objects, and edges represent relations for connecting nodes. By using graph database, relationship between each data is easier to analyze, each node has its own relationship. Graph database technology is an effective tool for data modeling when the focus of relationships between entities is the goal in data model design [2]. One of the most frequently used graph databases is Neo4j. It is claimed that Neo4j can process nodes as many as 34 billion nodes. Neo4j uses a query command called cypher. In this study, the K-Means ++ algorithm will be implemented to process customer data which are composed of income data and shopping scores, then silhouette score will be calculated to determine whether the clusters results obtained are good or not.

II. Literature Review

Research related to the implementation of K-Means using Neo4J has been conducted by K. Lavanya, Rani Kashyap, S. Anjana and Sumaiya Thasneen entitled "An Enhanced K-Means MSOINN Based Clustering Over Neo4j with an Application to Weather Analysis" [1]. The research carried out was to carry out K-Means MSOINN clustering based on the SOINN algorithm using the Delhi weather report dataset in 2016, with 3 samples of weather types, fog, clear, smoke. The results obtained from this study are that Neo4J is most suitable for representing connected data sets, the data set which is grouped and analyzed can predict weather patterns, and each node tends to have better accuracy using K-Means MSOINN compared to K-Means with a distance. Euclidean.

Then the K-Means research uses customer segmentation data without using Neo4J written by Musthofa Galih Pradana and Hoang Thi Ha entitled "Maximizing Strategy Improvement in Mall Customer Segmentation using K-Means Clustering" [3]. In this study it was proved that customer segmentation can be done by doing machine learning which results are very profitable in the industry. This study proves that machine learning can be implemented in the shopping district industrial segmentation, with, although machine learning can be grouped with accurate accuracy, the authors claim that humans can learn, change habits or change their shopping spending patterns.

III. K-Means and K-Means++

K-Means [4] is an unsupervised clustering algorithm, K-Means aims to group data points that are the same or similar and then look for patterns between data. The K-Means algorithm calculates the Euclidean distance (e.g., formula 1 below), to calculates the distance between each data object [1].

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

The value of q is the data and the value of p is the centroid, and q_i, p_i is the starting point of the value. The first step to grouping K-Means is to determine the number of clusters that will be used (K value), then select data points for a specified number of clusters, these data points are referred to as centroids. After that, calculate the distance from the first data with the centroid. Compare data distance with centroid and data group with centroid that has the closest distance. After all the data have been grouped, calculate the mean value of each cluster, then calculate and classify each data with the mean distance of each cluster. If the data cluster does not change until the last iteration, the process is complete. The number of clusters (K) is determined manually, but you can also use The Elbow Method [5] or The Silhouette Method [6] to find the optimum number of K.

Because the selection of the centroid of K-Means is random, this is a weakness of the clustering process. Therefore, to overcome these shortcomings, the K-Means ++ algorithm [7][8] was created, K-Means ++ ensures that the initialization of the centroid is done more intelligently to improve the quality of the grouping. Apart from the selection of other centroids the process is carried out the same as the standard K-Means process.

The difference between K-Means and K-Means ++ is in the initial selection of the centroid. In K-Means, the selection of centroids is carried out randomly as many as the specified K value, but in K-Means++, 1 centroid of random selection is carried out and then the calculation of the centroid distance is carried out with other data. The greatest value means having the farthest distance and the highest probability to become a new centroid, the process will be carried out as much as the specified K value.

III. Neo4J

Neo4J is an open source DBMS (Database Management System), released in February 2010, developed by Neo4j, Inc. The use of graphs is limited to node and edge designs, to help facilitate data storage and retrieval, the relationship between data and the value of the data. claims a maximum of 34 billion nodes. Neo4j only has one server and one database that does not need to define a database name. Neo4j can help developers to import data into graphs, business analysis to explore data easily, and make decisions for data science based on analysis results. Neo4j does not rely on a data relational layout, but a network storage model that natively stores nodes, relationships and attributes [9].

Neo4J uses the concept of graph [10], therefore the components used by Neo4J are similar to graph theory. Nodes represent data attributes in Neo4j, a node can have relations that connect relations with other data, to provide a description of the relationship, the type of relation must be added. A node can also be assigned labels to describe the data, a node can have multiple labels. Neo4J provides consistent and real time performance for multi-hop querying on a big connected datasets, is high availability, and is developer friendly [11]. In Neo4J, labels are used to represent the roles used by nodes on the graph. Since nodes can have different roles in a graph, with Neo4j it is possible to add more than one label to a node [9]. Neo4j is generally used for various use cases, examples of which are in the problem: 1. fraud detection, 2. knowledge graph, 3. recommendation system, 4. master data management, 5. social network graph.

Neo4j uses a query language called cypher. Cypher is a declarative graph query language that allows for expressive and efficient querying and updating of graphs. Cypher is made to be easy to read and understand by developers, cypher also allows users to give commands to the database to find data that matches the specified pattern [12]. Cypher is inspired by the query SQL language.

III. Research Method

This section describes the process of the system to segment the store customers.

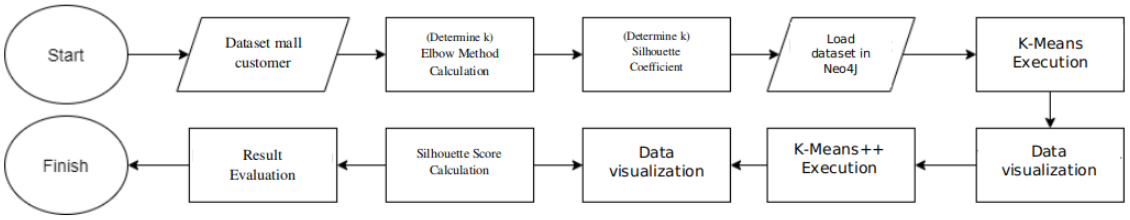

Fig. 1. Flowchart of process of the system for segmenting store customers

Steps of the system to segment store customers are: 1. The dataset is taken from Kaggle with the title 'Mall Customer Segmentation Data', 2. Elbow Method calculation is performed to find the most optimum K value, the experiment is carried out as many as 2 to 12 clusters and the optimum K value is taken from a linear distortion decrease, 3. Silhouette Coefficient calculation is performed to find the most optimum K value, the experiment is carried out as much as 2 to 12 clusters and the optimum K value is taken from the largest Silhouette value, 4. Load the dataset using Neo4J, 5. K-Means execution using Neo4J is done in the annual income and spending score column, 6. Visualization of the K-Means Neo4J clustering results, 7. The K-Means ++ execution using Neo4J is carried out in the annual income and spending score columns, 8. Data visualization of the K-Means ++ Neo4J clustering results, 9. Load the dataset using Python, 10. The K-Means execution using Python is carried out in the annual income and spending score columns, 11.Visualization of K-Means Python clustering data, 12. The K-Means ++ execution using Python is carried out in the annual income and spending score columns, 13. Visualization of K-Means ++ Python clustering results, 14. The Silhouette Score calculation is performed to determine the quality of the cluster, 15. Comparison between the Silhouette Score and time is carried out, and the clustering results are analyzed.

When the K-Means execution is carried out on Neo4J, 6 queries will be executed with each query aimed to: 1. Load the dataset and allocate the annual income and spending scores to x and y, 2. Determine the number of clusters and create a random centroid, 3. Give each cluster cluster a Number so they can be distinguished, 4. Declare the coordinates as x and y, 5. Measure the distance with the distance function, group the data with clusters, recalculate the location of the data with the mean of each cluster, 6. The previous steps are repeated until there is no data moving clusters, determine the total limit of iterations to be performed (if needed).

When the execution of K-Means ++ is performed on Neo4J, 7 queries will be executed with each query aimed to: 1. Load the dataset and allocate the annual income and spending scores to x and y, 2. Select 1 centroid randomly, 3. Perform K-Means ++ calculations to find the best centroid, 4. Declare the coordinates as x and y, 5. Give each cluster cluster a Number so they can be differentiated, 6. Measure the distance with the distance function, group the data with clusters, recalculate the location of the data with the mean of each cluster, 7. The previous steps are repeated until there is no data moving clusters, determine the total limit of iterations that will be carried out (if needed).

IV. Results And Discussion

The optimal number of clusters is determined by using the Elbow Method [5]. The results of the Elbow Method calculations carried out on the annual and spending data are as shown in Fig. 2. The optimum cluster value is taken 5 because at point 5 is the point where the inertia begins to decrease gradually. linear (much like an elbow).

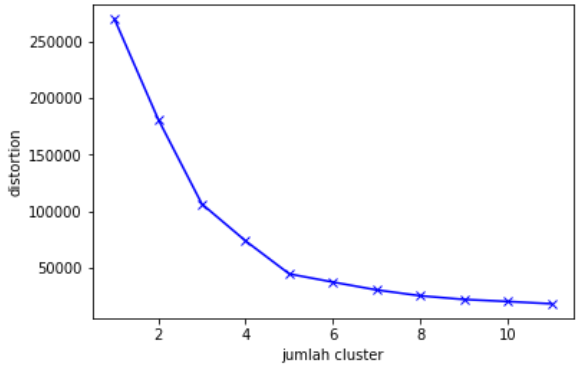

Fig. 2. Results of Elbow Method for Determining k Value.

Testing the optimal value of k can also be done using the Silhouette Coefficient [6] where the optimal value is taken from the largest silhouette score of each number of clusters that are tried. The Silhouette Coefficient value ranges from -1 to 1, where the value close to 1 indicates that the grouping of each data is in the correct group and the value close to -1 indicates that the data grouping is in the wrong group. The Silhouette Coefficient can be calculated using the formula 2 below.

$$S(i) = \frac{b(i) - a(i)}{Max(a(i), b(i))} \quad (2)$$

The value of a (i) is the average distance of point i from all other points in the cluster, and b (i) is the closest average distance from i to points in another cluster. The silhouette score results obtained in the testing of this final project with a range of 2 to 12 can be seen in table 1, that just like using the elbow method, the optimum K value is 5 (five).

TABLE I
THE AVERAGE VALUE OF THE SILHOUETTE SCORE AGAINST THE NUMBER OF CLUSTERS

K (*Cluster)*	Mean s*ilhouette score*
2	0.2968969162503008
3	0.46761358158775435
4	0.4931963109249047
5	0.553931997444648
6	0.5379675585622219
7	0.5288104473798049
8	0.45732611752686836
9	0.45819645551960536
10	0.45056557470336733
11	0.43560008750473395
12	0.42635706431613235

After obtaining the optimum value of k, then the clustering process is carried out using K-Means and K-Means ++, the results of clustering can be seen in Table 2, Fig. 3, and Table 3.

TABLE 2
K-Means and K-Means++ Clustering Results

	K-Means	K-Means++
Number of nodes in blue cluster	39	39
Number of nodes in brown cluster	23	23
Number of nodes in yellow cluster	80	81
Number of nodes in green cluster	36	35
Number of nodes in red cluster	22	22

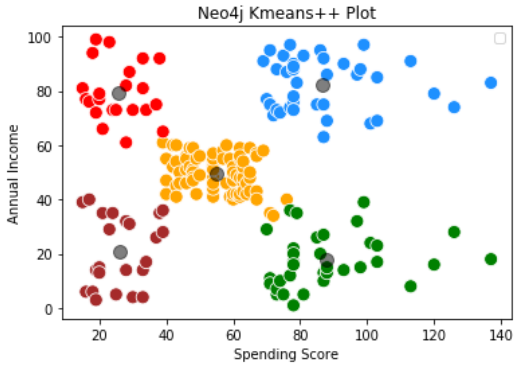

Fig. 3. Plots of results of K-Means++ clustering using Neo4J

TABLE 3
Silhouette score of clusters of K-Means and K-Means++ clustering results

Clustering Algorithm	*Silhouette Score*
K-Means++	0.553931997444648
K-Means	0.553217610757543

The cluster has a silhouette score of 0.5532176107575425 for K-Means and 0.553931997444648 for K-Me ans ++. Although the results are not near 1, if compared with the silhouette distance coefficient which has ran ge of -1 to 1, the value of 0.55 is satisfactory and indicates that the quality of the cluster is good.

The Table 4 and Table 5 below show the time of each queries performed by K-Means and K-Means++ resp ectively, this time data retrieval is retrieved using the "Neo4J Query Log Analyzer". Obtained average time, m inimum time, maximum time, maximum compile, and average execution.

TABLE 4
K-Means query cypher test time

Query	**Avg Time (ms)**	**Min Time (ms)**	**Max Time (ms)**	**Max Compile (ms)**	**Avg Execution (ms)**
Query 1	205.544	37.356	496.929	68.856	190.983
Query 2	22.032	2.004	61.094	1.977	21.025
Query 3	29.405	1.216	67.884	66.307	11.972

Query 4	7.561	2.634	17.36	1.972	6.45
Query 5	72.955	23.654	153.16	33.377	64.342
Query 6	597.637	355.906	1027.065	33.469	582.348
Total	935.134	422.770	1.653.194	205.958	870.721

TABLE 5
K-MEANS++ QUERY CYPHER TEST TIME

Query	Avg Time (ms)	Min Time (ms)	Max Time (ms)	Max Compile (ms)	Avg Execution (ms)
Query 1	212.705	49.827	501.741	1.804	211.309
Query 2	183.914	1.615	700.705	30.051	175.499
Query 3	403.682	107.407	702.685	25.338	399.047
Query 4	14.13	4.781	23.478	4.104	11.618
Query 5	29.779	29.779	29.779	27.473	2.306
Query 6	18.821	5.626	86.298	24.785	15.614
Query 7	89.351	12.689	425.964	36.515	80.051
Total	938.279	211.724	2.470.650	150.070	895.444

Based on the Table 4 and Table 5, the comparison between the performance given between the K-Means and K-Means ++ cypher, shows that the avg time given is not much different, for K-Means is 0.935134 seconds and for K-Means ++ is 0.938279 seconds. The time to run K-Means ++ is faster than K-Means. Query 1 takes a relatively large amount of time because the query loads and preprocesses data. Query 6 in K-Means has a large time value because in query 6 the repeated process is carried out. Meanwhile, in K-Means ++, query 7 has a faster time value than query 6 in K-Means, but query 3 (the centroid selection process) in K-Means++ has a large time value.

Analysis of Test Results

From the results, the analysis carried out, it can be studied customer spending habits from monthly income and shopping scores. Table 7 shows information that can be concluded from the results of clustering of K-Means and K-Means++.

TABLE 7
CLUSTER RESULTS AND CUSTOMER PRIORITIES IN MARKETING.

Clusters Color	Marketing Priorities	Information
Blue	1	*High income, high spending*
Red	2	*High income, low spending*
Green	3	*Low income, high spending*
Yellow	4	*Medium income, medium spending*
Brown	5	*Low income, Low spending*

Let's focus on Table 7. Blue group is a customer who has high income and high shopping value, customers in this group are the main target because it is likely that the customers of this group will buy something offered to him and most likely are repeat customers [3]. The red group is a customer who has low income and high shopping value, the customer in this group is the type of customer who is wasteful and will almost buy the product offered. Customers in this group are good targets in addition to customers in the blue group because of the very high purchase value [3]. Green group is customers who have high income and low shopping value, customers in this group are customers who are careful in shopping or it can be assumed that

these customers are not satisfied with the facilities provided or are not regular customers [3]. The yellow group are customers who have average income and average expenditure value. It can be assumed that customers in this group are customers who have high shopping preferences but with average income, so the shopping value is low [3]. The brown group is a customer who has low income and low shopping value. Customers in the group generally will not make excessive purchases and will shop carefully and choose products that are durable due to the low shopping value. Customers in this group are the most recent target in product offerings [3].

V. Conclusion

The experiment shows that K-Means++ can be implemented by using Neo4J. In this study K-Means++ is implemented for store customers segmentation. The optimal value of K of K-Means++ are determined by using Elbow method. The experiment results of K-Means++ are compared to K-Means. In term of average time, K-Means++ achieved a bit faster than K-Means. The average time of K-Means++ and K-Means is 938.279 ms and 935.134 ms respectively. The quality of clustering (segmentation) of K-Means++ and K-Means is measured by silhouette score. In term of silhouette score, K-Means++ achieved better than K-Means. The silhouette score of clusters of K-Means++ and K-Means is 0.553931997444648 and 0.553217610757543 respectively. The clusters results are analyzed to get customers segmentation based on the value of spending and income. We can infer customers priority, there are customers in the top priority and customers in last priority for sales and product offerings.

References

[1] K. Lavanya, R. Kashyap, S. Anjana, and S. Thasneen, "An Enhanced K-Means MSOINN Based Clustering Over Neo4j with an Application to Weather Analysis," 2020.
[2] C. Pascal, S. Ozuomba, and C. kalu, "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services," *Int. J. Adv. Res. Artif. Intell.*, 2015.
[3] M. G. Pradana and H. T. Ha, "Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering," *J. Appl. Data Sci.*, vol. 2, no. 1, pp. 19–25, 2021.
[4] D. Steinley, "K-means clustering: A half-century synthesis," *Br. J. Math. Stat. Psychol.*, 2006.
[5] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, 2013.
[6] T. Van Craenendonck and H. Blockeel, "Using internal validity measures to compare clustering algorithms," *Icml*, 2015.
[7] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.
[8] A. Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms," in *3rd IEEE International Conference on* , 2017.
[9] D. Dominguez-Sal, P. Urbón-Bayes, A. Giménez-Vañó, S. Gómez-Villamor, N. Martínez-Bazán, and J. L. Larriba-Pey, "Survey of graph database performance on the HPC scalable graph analysis benchmark," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010.
[10] D. T. Murdiansyah and Adiwijaya, "Computing the metric dimension of hypercube graphs by particle swarm optimization algorithms," in *Advances in Intelligent Systems and Computing*, 2017.
[11] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, 2015.
[12] B. Sasaki, "Graph Databases for Beginners: Why Graph Technology Is the Future," *Neo4J*. 2018.