Application of Clustering Methods for Anomaly Identification in Efforts to Improve P2TL Quality

Ardik Crisdianto - 23522312
Program Studi Magister Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail (gmail): ardikcrisdianto@gmail.com

Abstract*— This study evaluates the effectiveness of two clustering methods namely K-Means and DBScan in detecting anomalies in Automatic Meter Reading (AMR) data. The data used includes voltage and current parameters and adds frequency as additional data. The AMR data used are 3-phase indirect measurement customers with 53kVA to 197kVA power. The clustering process is performed using K-Means with k3 and k4 and DBScan using epsilon 0.05 and 0.1. The value of k is known based on the elbow method and epsilon is known by k-distance. Evaluation of clustering results is done using Silhouette Score and Davies-Bouldin Index. The results show that DBScan with epsilon 0.1 gives the best results with the highest Silhouette Score of 0.6736 and the lowest Davies-Bouldin Index of 0.9078 indicating this method is most effective in clustering the data and detecting anomalies. This research provides valuable insights into the selection of optimal clustering methods for detecting anomalies in AMR data.

***Keywords*— clustering, k-means, DBbscan, anomaly detection**

I. Introduction

There are various ways that electric energy providers can overcome challenges in the electricity sector, one of which is the Penertiban Pemakaian Tenaga Listrik (P2TL) to deal with anomalies in customer electricity usage both in terms of kWh meter errors and illegal use of electrical energy by users. One of the techniques in determining the P2TL operation target is by analyzing anomalies in customer meter readings. One approach that can be used is the clustering method, which makes it possible to group customers based on current and voltage readings, current and voltage by tariff designation, energy consumption patterns, and other relevant factors. The most important thing in clustering is determining the number of groups (Messinis and Hatziargyriou (t.t.)) so that the groups representing the anomalies can more accurately determine whether the meter is problematic or not.

A. Research Objectives

The main objective of this research is to develop the potential application of clustering methods in the identification of problem customer groups in P2TL. Some specific objectives include:

1. Determine the most appropriate clustering method to group customers based on the available data.
2. Identifying behavior patterns or special characteristics that characterize anomalous.

B. Problem Constraint

In this research, there are limitations that need to be considered, among others:

1. The data used in this study is limited to Automatic Meter Reading (AMR) customer meter reading data during the period January 2024 to March 2024 for all tariffs R, S, B, P and I which have a power range of 53 kVA to 197 kVA. The unit studied is 5114.
2. This clustering model uses two algorithms namely K-Means and DBScan.

II. Literature Review

A. Main and Supporting Theories

Machine learning (ML) techniques have been used in various ways to solve problems. ML is used to teach machines how to handle data more efficiently (Mahesh, 2018). ML can do several things that can make it easier for humans to research something. Among other things, (a) pattern recognition whether it is data, images, text, sound and other data structures, (b) classification and prediction based on the data given, (c) grouping data into groups that have similar characteristics, (d) recommendations based on previous preferences, and several other things. In accordance with this research, ML is used in the case of clustering customer anomalies.

B. *Unsupervised Learning*

Unsupervised learning is one of the approaches in machine learning where the model is given data that does not have predefined labels or target information. Unsupervised learning is very useful for clustering algorithms (Dike et al., 2018) so that groups of data with similar characteristics can be clustered without the need for labels or target information beforehand

C. *Clustering*

The application of clustering in customer data analysis allows PT PLN (Persero) to better understand customer needs and behavior. According to research (Qi et al., 2017) clustering is one of the most important things in data analysis, allowing companies to develop more effective strategies. Clustering is the process of grouping data into several clusters or groups so that the data in each group has high similarity and between groups has low similarity (Mughnyanti et al., 2020).

D. *K-Means Algorithm*

The K-Means algorithm requires determining the number of clusters as well as the initial state. The method of K-Means is to randomly select k objects (number of clusters) from D (dataset containing n objects) as initial cluster centers. Then group the objects by reallocating each object to the cluster where it is most similar based on the average value (mean) of the cluster objects. Next, the average of the objects for each cluster is calculated again, and the step is repeated until there is no change (convergent). In finding the value of k, several ways are used, including using the Davies-Bouldin Index (DBI) and the Elbow Method

E. *DBScan Algorithm*

DBScan is one of the most popular and widely used Density-based algorithms. DBScan uses Nϵ(p) and a threshold called minPts (minimum number of points) to detect dense regions and can classify points in a dataset into Core, Border and Noise (Hahsler et al., 2019). MinPts is a parameter used to determine whether a point can be considered a core, if the number of its neighbors within a certain radius (defined by the epsilon parameter) is at least equal to minPts.

F. *Previous Research*

In (Deng, 2020) DBScan is used to explore and detect anomalies in data related to network security. Details about the dataset are not mentioned in the paper. K-Means clustering is used as a comparison in detecting anomalies. The steps involved setting the epsilon (ε) and setting the minimum points (MinPts) and then clustering the data points based on density and identifying outliers as anomalies.

Parameter name	Numerical value
Experimental tools	PyCharm
eps	10
Min_samples	2
MinPoints	6
n-clusters	2

Table II.1 Parameter Setting DBScan

The final result of the research is that DBScan is more effective in detecting anomalies in network data compared to K-Means clustering. However, the paper does not mention more details regarding the clustering results or the evaluation matrix.

III. SYSTEM DESIGN

Types of anomalies that can be identified from AMR meter readings include:

1. Voltage Drop (Under Voltage)
2. Voltage Rise (Over Voltage)
3. Current Unbalance
4. CT (Current Transformer) Saturated
5. Wiring Error on the Meter
6. PLTS (Solar Power Plant) customers who do not have permission to On-Grid to PLN

The main features analyzed are Current and Voltage. These two features are influenced by several things, including tariffs. So the tariff is also a supporting feature that is used so that the research results are better.

The system can only detect anomalies based on existing data and is unable to detect anomalies caused by external factors that are not detected by AMR meters.

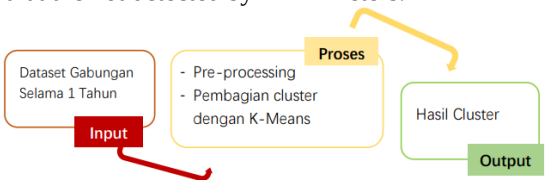

Fig III.1 System Design

IV. SYSTEM ANALYSIS AND TESTING

The nominal voltage of the low voltage network is 220V. while the standard service voltage in accordance with SPLN 1: 1978 service voltage has a maximum value of 5% (230V) and a minimum of 10% (198V) of the nominal voltage.

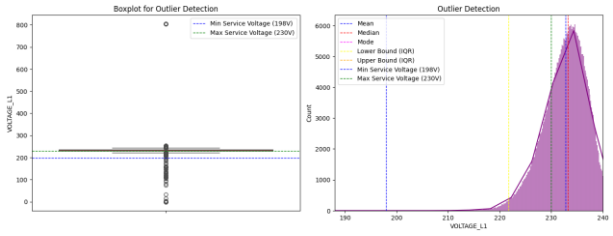

Fig III.1 Univariate Analysis Voltage

The following is a bivariate analysis using two variables, namely location_code and voltage:

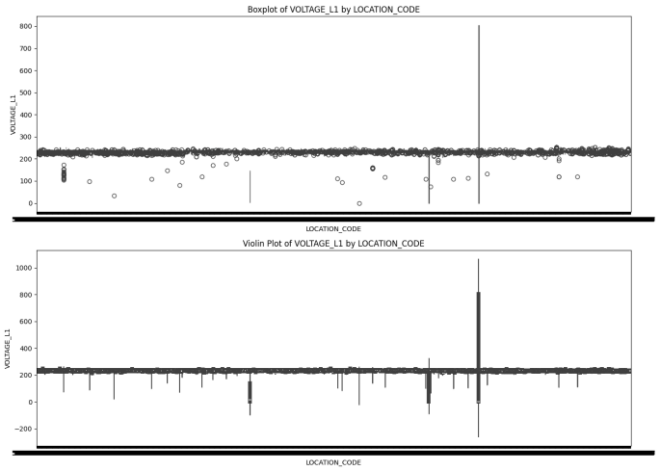

Fig III.2 Bivariate Analysis Voltage

Here is the correlation between features using voltage, current and power:

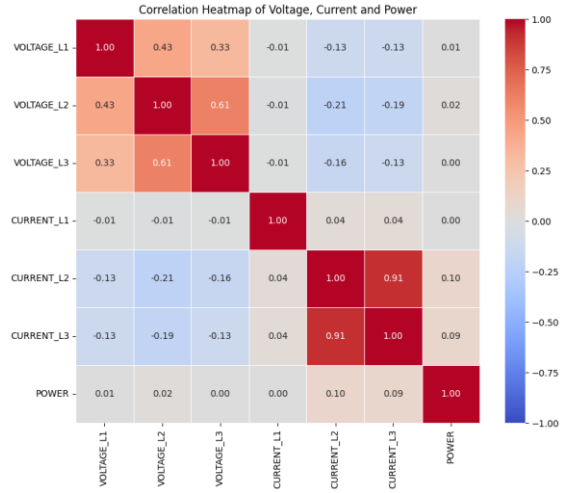

Fig III.3 Correlation Heatmap Current, Voltage dan Power

A. K-Means

The value of k is around 3 and 4. This is determined by the point at which the decrease in inertia starts to slow down significantly. In this case, both values of k were tried. Next, clustering with k-K-Means was done based on tariffs. Tariffs were separated between R, B, P, S and I to get a deeper look at the anomalies. PCA was used to accommodate visualization with features VOLTAGE_L1, VOLTAGE_L2. VOLTAGE_L3 and CURRENT_L1, VOLTAGE_L2 and VOLTAGE_L3.

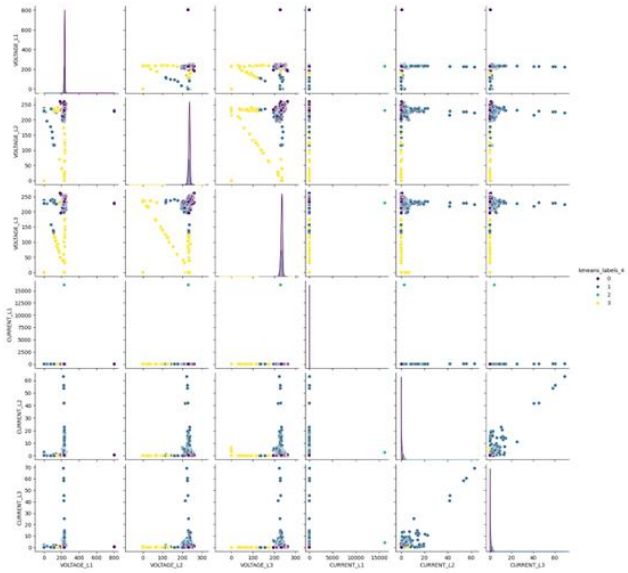

Fig III.4 Pairplot K-Means k=4

B. DBScan

By using k-distance, obtained epsilon 0.05 and 0.1:

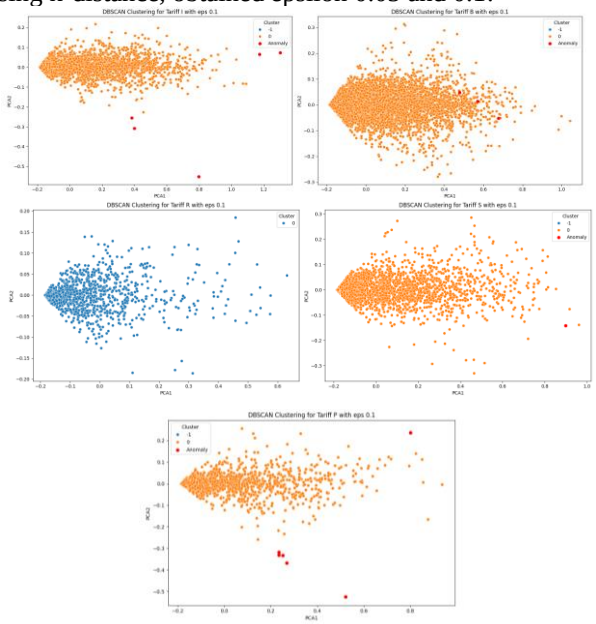

Fig III.5 DBScan dengan eps=0.1 MinPts=5

C. Analysis and Evaluation of Test Results

From the research results, DBScan with eps = 0.1 has better results than K-Means. Here are the details of the results:

NO	EVALUATION	K-Means		DBScan	
		k3	k4	eps 0.05	eps 0.1
1	Silhouette Score	0.48564	0.48215	0.55569	0.67357
2	Davies-Bouldin Index	0.63088	0.63259	1.71569	0.90776

Table III.1 Evaluation Result

V. Conclusions And Suggestions

From the research results, the following conclusions can be drawn:

1. Evaluation results using Silhouette Score and Davies-Bouldin Index show that DBScan has better results than K-Means. This means that DBScan with eps 0.1 is able to categorize data well, especially in detecting anomalies.
2. The evaluation results can illustrate that clustering has successfully performed data clustering even to detect anomalies in the data.

To get the best results, the clustering process can be used as a first step. Furthermore, it can be combined with other machine learning algorithms to get more accurate anomaly detection results.

References

[1] Deng, D. (2020): Research on Anomaly Detection Method Based on DBSCAN Clustering Algorithm, Proceedings - 2020 5th International Conference on Information Science, Computer Technology and Transportation, ISCTT 2020, Institute of Electrical and Electronics Engineers Inc., 439–442. https://doi.org/10.1109/ISCTT51595.2020.00083

[2] Dike, H. U., Zhou, Y., Deveerasetty, K. K., and Wu, Q. (2018): Unsupervised Learning Based On Artificial Neural Network: A Review, 2018 IEEE International Conference on Cyborg and Bionic Systems, CBS 2018, Institute of Electrical and Electronics Engineers Inc., 322–327. https://doi.org/10.1109/CBS.2018.8612259

[3] Hahsler, M., Piekenbrock, M., and Doran, D. (2019): Dbscan: Fast density-based clustering with R, Journal of Statistical Software, 91. https://doi.org/10.18637/jss.v091.i01

[4] Huang, H., Wei, B., Dai, J., and Ke, W. (2020): Data Preprocessing Method for the Analysis of Incomplete Data on Students in Poverty, Proceedings - 2020 16th International Conference on Computational Intelligence and Security, CIS 2020, Institute of Electrical and Electronics Engineers Inc., 248–252. https://doi.org/10.1109/CIS52066.2020.00060

[5] Mahesh, B. (2018): Machine Learning Algorithms-A Review, International Journal of Science and Research. https://doi.org/10.21275/ART20203995

[6] Messinis, G. M., and Hatziargyriou, N. D. (n.d.): Unsupervised Classification for Non-Technical Loss Detection.

[7] Mughnyanti, M., Efendi, S., and Zarlis, M. (2020): Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation, IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, 725. https://doi.org/10.1088/1757-899X/725/1/012128

[8] Qi, J., Yu, Y., Wang, L., Liu, J., and Wang, Y. (2017): An effective and efficient hierarchical K-means clustering algorithm, International Journal of Distributed Sensor Networks, 13(8), 1–17. https://doi.org/10.1177/1550147717728627

[9] Wijaya, Y. A., Kurniady, D. A., Setyanto, E., Tarihoran, W. S., Rusmana, D., and Rahim, R. (2021): Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities, TEM Journal, 10(3), 1099–1103. https://doi.org/10.18421/TEM103-13

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 8 Juni 2024
Ttd
Ardik Crisdianto dan 23522312