Comparison of Classification Algorithms on Household Electricity Consumption Data

Brilian Putra Amiruddin[1*], Evanbill Antonio Kore[1], Dhiya Aldifa Ulhaq[2] and Auzan Widhatama[1]

[1] Electrical Engineering Department, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[2] Mechanical Engineering Department, Faculty of Industrial Technology, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
*e-mail: brilianamiruddin.17071@mhs.its.ac.id

Abstract. *The pattern of electricity consumption is one thing that is important to be known by a household, so it is essential to identify the type of intensity of electricity usage from the household's daily life. It can help determine how much electricity consumption of equipment so that efforts can be made to optimize electricity consumption further while saving costs. Due to that, the classification algorithms based on supervised learning is used. In this study, we compared several types of classification methods to determine the type of electricity usage patterns in a daily household life on Household Electric Power Consumption data obtained from Kaggle. The classification methods being compared are KNN, SVM, Decision Tree, and Logistic Regression. The accuracy of all methods is analyzed to find which method is best in identifying the intensity of electricity usage. From the results of this study, it was found that the Logistic Regression method was the most accurate in classifying the type of intensity of electricity consumption with an average accuracy value of 99%.*

Keywords: classification, comparison, household electricity, algorithms

Introduction

Nowadays, the need for electricity usage among the community is increasing, not only for daily needs but also for the needs of small and large scale industries. Analysis of the power flow or load flow is the essential thing to understand to analyze well the electric use intensity. Based on the definition, power flow is a process of channeling both active and reactive power from the source to load. Reactive and active power flows must have specific standards so as not to affect the distribution process in the electric power system. Reactive loads that are too high can cause a decrease in the power factor, causing a decrease in the efficiency of the distribution and transmission. This pattern of electricity usage can also affect the quality of electricity distribution. Therefore, assistance is needed to use the pattern of equipment in society against the power used. To achieved this, the classification algorithm is used, one of the supervised learning algorithms, to determine a set of objects into groups called classes based on the criteria of each class.

In terms of electricity supply companies, the pattern of electricity usage has great importance on energy that needs to be supported by electricity companies [1], high electrical energy requirements from the consumer side can use distressed electricity distribution networks [2], by using the regular electricity pattern of the user, the electricity distribution network will no longer be burdened, and the electricity distribution will be reasonable and efficient [3]. Related to the need for some patterns of electricity usage to overcome the distribution problem, we must well classify consumer behavior on electrical usage to know the pattern of the household. The methods used for classification in this study include the K-Nearest Neighbor, Support Vector Machine, Decision Tree, and Logistic Regression. After that, we compare the performance results of each algorithm that has been used to classify patterns in electricity usage data in the household, so we get the best algorithm for classifying the electrical intensity data.

Materials and Methods

The dataset used in this paper was data on the power use of electrical equipment in a household from Kaggle [4], from the correlation of data that has been analyzed with exploratory data analysis,

we finally determined the label or target of the data by using an average of 3 features namely sub_metering_1, sub_metering_2, and sub_metering_3, then we divided the existing data into three classes including 'HIGH_INTENSITY', 'MEDIUM_INTENSITY', and 'LOW INTENSITY', where 'LOW INTENSITY' showed the average intensity of customer usage less than or equal to 14 Wh, 'MEDIUM INTENSITY 'showed the average intensity of customer use more than 14 Wh and less than or equal to 28 Wh, 'HIGH INTENSITY' showed the average intensity of the customer electricity usage more than 28 Wh. After labeling our data, it turns out that our data has become imbalanced. The data scattered in each class, and it was not evenly distributed, such as follows: 'LOW INTENSITY' had data of 1008722, 'MEDIUM INTENSITY' had 33134 data, 'HIGH INTENSITY' had 2650 data. Therefore, data preprocessing was occurred using the random undersampling method [5], so evenly distributed data was gotten in each class following the lowest amount of data from the class that was 2650. In this paper, we only used 7 of the 9 features in the dataset global_active_power, global_reactive_power, voltage, global intensity, sub_metering_1, sub_metering_2, and sub_metering_3, which were numeric. Table 1 shows the details description of each data features and target

Table 1: Dataset Features and Target.

Name	Description
Date	dd/mm/yy
Time Format	hh:mm:ss
Global_active_power	Active power (kW)
Global_reactive_power	Reactive power (kVAR)
Global_intensity	Intensity of household current (Ampere)
Voltage	Average using of voltage (V)
Sub_metering_1	Power consumption in kitchen peripherals (Wh)
Sub_metering_2	Power consumption from laundry room (Wh)
Sub_metering_3	Power consumption from air conditioner and heater (Wh)
Target	Three types of intensity (LOW, MEDIUM, HIGH)

In this experiment, the conventional classification algorithms such as KNN [6], SVM [7], Decision Tree [8], and Logistic Regression [9] performance were compared. In this case, data were tested with four scenarios, namely test data (70% on the training set and 30% on the test set), k-fold cross-validation with k=10, (Leave-one-out-cross-validation) LOOCV, and random sampling. For comparison, the results were checked with the help of metrics such as evaluation of the confusion matrix to determine the precision, recall, and F1 as well the others metric, for instance, AUC scores, and the average of the four scenarios were taken to determine the best-performed algorithms. Various testing methods were used to find out the performance of the algorithm in more detail and precise because the evaluation parameters obtained would be more fruitful. For the software, we used Orange [10] to computed all of the algorithms, implementation workflow on the orange software shown in Figure 1

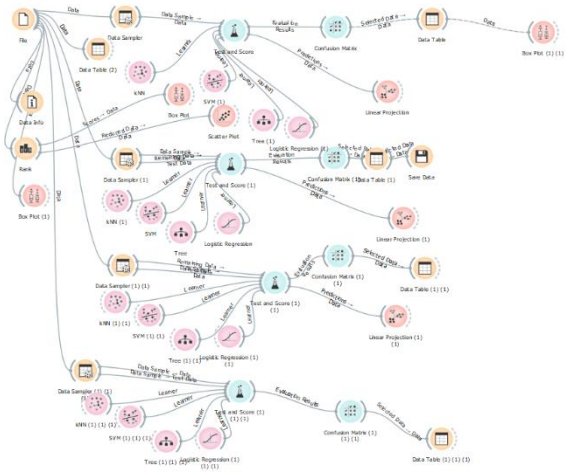

Figure 1: Implementation Workflow of The Classification.

Results and Discussion

The algorithms parameters used in this experiment were tuned to fit the dataset well before the classification task was performed. The KNN algorithm had a total neighbor number of 87 (K = 87), and the euclidean distance calculation method was applied. For the SVM, the Radial Basis Function (RBF) kernel was applied. The Logistic Regression was employed by the L2 regularization algorithm to elude the overfitting problem. On the Decision Tree, the depth of trees was set to 100.

After conducting experiments for all four test scenarios, namely by changing the proportion of the dataset distributed for training and test data, LOOCV, k-folds cross-validation, and random sampling, the results of each performance metrics was gotten, the average results of all four scenarios are shown in Table 2

Table 2: Algorithms Performance Metrics.

Algorithms	Precision	Recall	AUC	Accuracy	F1 Score
SVM	98.5%	98.5%	100%	98.5%	98.5%
KNN	97%	97%	99%	97%	97%
Logistic Regression	99%	99%	99.3%	99%	99%
Decision Tree	97.3%	97.3%	98%	97.3%	97.3%

In the aforementioned results, it was founded that the Logistic Regression algorithm, producing in excellent average classification accuracy of 99%, nearly reaching 100%, the method as well had superior Precision, Recall, and F1-Score compared to the other models or algorithms. The Support Vector Machine algorithm was the only algorithm which superior to Logistic Regression on one of the metrics, SVM performed well than the Logistic Regression on the AUC score. It scored 100%, which 1% higher than Logistic Regression. On the contrary, the Decision Tree algorithm performed better than K-Nearest Neighbour on almost all the metrics except the AUC score. However, both KNN and Decision Tree algorithms performance were beneath the Logistic Regression and SVM. From that, the Logistic Regression was the best-performed algorithms followed by SVM, Decision Tree, and KNN, respectively.

Conclusions

In conclusion, it was found that Logistic Regression was the method that had the most superior performance for classifying household electricity usage data when viewed from several performance evaluation parameters that have been done, Logistic Regression has an average accuracy of 99%. Also, it achieved superior performance in the other tested metrics. Nevertheless, the classification results also showed that the classification of the intensity of electricity use in households could be done well by the KNN, SVM, and Decision Tree algorithms because the results of the evaluation conducted for each evaluation parameter in all scenarios of the data test method had values above 95% so it could be said that the KNN, SVM, Decision Tree, and Logistic Regression algorithms are very suitable for classifying household electricity usage data.

References

[1] M. Manjunath, P. Singh, A. Mandal, and G. S. Parihar, "Consumer Behaviour towards Electricity –A Field Study," *Energy Procedia*, vol. 54, pp. 541–548, 2014, doi: 10.1016/j.egypro.2014.07.295.

[2] R. Miceli, "Energy Management and Smart Grids," *Energies*, vol. 6, no. 4, pp. 2262–2290, Apr. 2013, doi: 10.3390/en6042262.

[3] J. Ouyang, L. Gao, Y. Yan, K. Hokao, and J. Ge, "Effects of Improved Consumer Behavior on Energy Conservation in the Urban Residential Sector of Hangzhou, China," *J. Asian Archit. Build. Eng.*, vol. 8, no. 1, pp. 243–249, May 2009, doi: 10.3130/jaabe.8.243.

[4] "Household Electric Power Consumption (Version 1)." United States : UCI Machine Learning, Aug. 24, 2016, [Online]. Available: https://www.kaggle.com/uciml/electric-power-consumption-data-set.

[5] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data," in *2015 IEEE International Conference on Information Reuse and Integration*, San Francisco, CA, USA, Aug. 2015, pp. 197–202, doi: 10.1109/IRI.2015.39.

[6] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," vol. 3, no. 5, p. 6, 2013.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[8] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
[9] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: 10.1080/00220670209598786.
[10] J. Demšar and B. Zupan, "Orange: Data Mining Fruitful and Fun - A Historical Perspective," p. 6.