Identifying Irregularities Usage of Smart Electric Voucher using Support Vector Machines

Supriyadi La Wungo[1], Armin Lawi[2] and Hartono[3]
{supriyadi.la.wungo@gmail.com }

[1]Postgraduate Program of Electrical Engineering,Hasanuddin University, Makassar, Indonesia
[2]Department of Computer Science, Hasanuddin University, Makassar, Indonesia
[3]Department of Computer Sciences, STMIK IBBI, Medan, Indonesia

Abstract. This paper applies machine-learning approach in identifying the irregularities of customer behaviour to the purchase transaction of smart electric vouchers. The Support Vector Machine (SVM) is used as the classification machine learning to identify the irregularities usage into two classes. The performance of the classification system is evaluated using the 10-fold cross-validation technique. Validation results are measured using accuracy, precision, and recall values. Our implementation results showed the use of SVM method gives very good performances in classifying the electrical consumption behaviour. Experimental results with different amounts of data testing indicated that the SVM method has high degree of accuracy, precision, and recall of 99 to 100%.

Keywords: The Irregularities Usage, Smart Electric Voucher, Vector Machine

1 Introduction

Smart electric voucher is a more convenient and controlled product since the electricity consumption is fully controlled by the customer. The use of smart electric voucher in Indonesia is intensively encouraged in order to replace the postpaid electrical consumption in which customers will pay utility bills each month in accordance with their electricity consumption [1]. Despite the ease of smart electric vouchers that have been provided by energy providers, there are still many customers found and indicated to have electricity theft in the field.For instance, the customer changed the wiring pattern on kWh meter indicator and connected the power cable without going through the indicator[2], [3].

Supervision of the customer is very minimal done by the company. This happens because the company no longer conducts the recording and manual checking services of each month.In order to monitor the misuse of electrical energy usage transactions, the electric energy company filters the token/voucher purchase transaction period. However, this method still has the disadvantage such as the nominal shortage can still be considered as a reasonable use, e.g., if the customer only make token/voucher purchase transaction with a nominal value of IDR 20,000 and the transaction amount is only once a month even though the target of IDR 100,000 usage in a month is not fulfilled then the customer is considered to have reasonable transaction by the company. Therefore, to overcome this problem, this research proposes to also filter the number of transactions and nominal purchase of token/voucher of the customers.

WMA-1 2018, January 20-21, Indonesia

DOI 10.4108/eai.20-1-2018.2281928

There have been many related works on the abuse of electrical energy usage has been proposed by previous researchers as follows. Babu, et al., in[4]used FCM to investigate non-technical loss detection by monitoring the profile of irregular customer consumption in power distribution systems. Their Fuzzy based classification method detected non-technical losses with accuracy of 80%.Depuru, et al.,in [5] proposes the detection of electricity theft through energy consumption patterns of some customers involved in the theft. They used classification method using Support Vector Machine (SVM) with kernel rbf, and the resulted accuracy was 98.4%.The further research in [6] proposed a comprehensive top-down scheme based on the Decision Tree (DT) and SVM to detect and locate real-time power theft at any power level in transmission and distribution. DT is used to calculate the energy consumption that has been used by the customer based on the value of the attribute. Then the results of this calculation will be used as input on SVM to classify normal or abnormal customers in the use of electrical energy. The combination of both methods has an accuracy rate of 92.5% and a very low false positive of 5.12%.Another computational technique for fraudulent classification in electricity consumption through a power consumption profile has been posed in [7]. The method used Fuzzy C-Means (FCM) to group the customers who have the same pattern from within database and then process the customer classification that is not normal. Classification based on Fuzzy C-Means was classified using 2 metrics namely Assertiveness and Sensitivity. The accuracy of both metrics is 0.745 and 0.100, respectively.

2 Literature Review

2.1 Customer Electricity Usage Behaviour Problems

Currently, the supervision system of electricity usage by electrical energy companies to customers is very minimal because it no longer uses the service of recording and manual checking each month. Thus, the electric energy company conducts monitoring on the customer indicated to abuse the transactions of electrical energy usage that is considered unfair by filtering the token/voucher purchase transaction period. However, the filters that the company does still have many weaknesses and the manual checks are time inefficiently and costly [6].

Some actions of unreasonable customer behavior are putting the magnet above the kWh meter to reduce disc rotation, the customer connecting the power cord without going through the kWh meter and changing the wiring pattern at kWh meter[2], [3], [5]so the counter on kWh does not work and the kWh meter keeps working even though the remaining credit token/voucher has run out.

The impact of the electrical energy misuses for consumers are; the power shared by other consumers is reduced, power outage is often happened, and fire may occur due to short circuit. The influence for the company is financial loss reaches millions and even billions of rupiah due to power loss [5].

2.2 Identification of relevant variables

In case of transaction history of purchase pulses, there are several variables are used to identify irregularities of electricity usage of each customer as follows.

- *Customer ID* identifies how many customers make credit purchase transactions in a month.
- *Installed Power* (kWh) identifies the power consumption of customer's use based on their installed power.
- *Total credit in voucher* is thenumber of kWh quota in the voucher by the customertransaction.
- *Total power consumption* is the difference between maximum and minimumelectricity utilizationsof the amount usage in a month.
- *Duration of voucher usage* is the duration of voucher usage by the customer during a certain period.
- *Status* is the dependent (response) variable defines proper or improper usage classification of the power consumption usage of customers utilizes vouchers.

These five independent variables (or predictors) are considered to represent each customer in identifying usage patterns in order to support the analysis of electrical data as well as to detect the morbidity of existing systems. The usage data is one-month span data which consist of 10,000 customers in order to minimize the losses in the company since the electricity payment by the customer is done monthly and the transaction of the customer's credit purchase by the company is done also per month. Therefore the determination of this time span is very relevantfor the process[7], [4].

2.3 Identification of relevant variables

Identification of irregularities is a technique of searching data to find an object that does not meet certain criteria when compared with other objects. These criteria can be known through the attribute value of an object. An object is said to be unnatural if the value of an attribute possessed by that object does not meet certain criteria compared to the attribute values of another object that meets the criteria. Therefore, to determine the imperfection of an object must select the appropriate attributes [3].

Identification of irregularities can be done by detecting a fraud or theft. In the case of electric theft, irregularities can be identified by looking at the usage patterns of electricity consumption for customers who have above average usage of actual usage targets. Customer behavior patterns of electricity consumption can be seen in the historical data of energy consumption. In large-scale companies, historical data is used to identify customer usage patterns that are considered to have an unnatural transaction in electricity consumption. Based on the morbidity, the company can indicate the theft or electric fraud committed by the customer[8], [9].

2.4 Classification techniques

Support Vector Machine (SVM) can process data sets with high dimensions and it can also use kernel technique to map original data from its original dimension to another relatively higher dimension. The data on the contributing SVM is called support vector.Based on the training data vectors $x_i \in R^n, i = 1, 2, \dots, l,$ in two classes of vectors $y \in R^l$as$y_i \in \{1, -1\}$, support vector classifier is used to solve the following primal optimization problems:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

Subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i$$

$\zeta_i \geq 0$, i = 1, …, l, where$\phi(x_i)$is a kernel maps$x_i$into the high-dimensional space and $C > 0$.

Equation$y_i[w^T\phi(x_i) + b] \geq 1$consists of two constraints, the first is$w^T\phi(x_i) + b \geq +1$if$y_i = +1$, and the second is$w^T\phi(x_i) + b \leq -1$ if$y_i = -1$ where$\zeta_i$is a slack variable that may occur during a misclassification of an unequal set[2].

Data training which is used when constructing the SVC model in reality mostly cannot be separated linearly. To resolve the classification problem linearly, some slack variables will be mapped into high dimensions using the RBF and Polynomial kernels in SVM so that errors and disturbances during dataset training can be reduced[2]. Linear, RBF (radial basis function) and Polynomials kernels which are used in this paper as follows[10].

- Linear:$K(x_i, x_j) = x_i^T x_j$.
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
- Gaussian:$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0$
- Sigmoid: $K(x_i, x_j) = \tanh(\gamma . x_i^T + r)$

where$x_i$is support vector and$x_j$is the data value of the attribute and$\gamma$is the kernel parameter. The main purpose of this kernel is to make the best decision limit when doing the classification of training sets into two parts[6].

2.5 Performance Evaluation

In general, the confusion matrix is used to measure the performance of the classification method. The confusion matrix is a table used to display performance results from the classification of normal and abnormal data. Confusion matrix also called contingency table can be seen in Table I[11]

Table I. Confusion matrix

		Actual class prediction	
Actual class		Normal	Abnormal
	Normal	TP	FN
	Abnormal	FP	TN

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where, TP is the amount of normal data that is correctly predicted as a normal class, FN is the amount of normal data that is predicted as an abnormal class, TN Is the amount of abnormal data that is correctly predicted as an abnormal class and FP is the amount of abnormal data that is predicted as a normal class.

Accuracy is a comparison of the correct classification results of the total classification. Precision is the correct comparison of normal data classification of the total classification of mismatch. The recall is a classification comparison detecting an irregularity to total impropriety. If the model shows a high recall value, then this model is reliable. This is because the model rarely makes a mistake in diagnosing normal data to an abnormal class. Whereas if the recall value shows a high value, then the model's performance is very good because the model has a high accuracy in diagnosing normal data of morbidity.

3 Results

The first paragraph after a heading is not indented (Bodytext style)In this study, using the dataset history of purchase transaction pulses obtained from the company PT. PLN (State Electricity Company) Mamuju, West Sulawesi, Indonesia. The transaction history dataset of the pulse has a real-type attribute with a total of 10000 lines of data consisting of 7783 fair transactions and 2217 unusual transactions and has 4 attributes and 2 classes. The dataset will be divided into 10 sections using k-fold cross validation as data validation, consisting of 9 sections will be used for training data and 1 part will be used for data testing.Experimental research results are processed using an ASUS laptop with an Intel® Core i5 processor CPU M 460 @ 2.53 GHz, 4.00 GB of RAM, and the Linux operating system 64-bit Ubuntu version 6.10. While the software to develop applications is Python version 2.7.The classification method to be used in this research is Support Vector Machine.

Kernel	Performance	Percentage of data testing			
		20	40	60	80
Linear	Accuracy	1.00	1.00	1.00	1.00
	Precision	1.00	1.00	1.00	1.00
	Recall	1.00	1.00	1.00	1.00
Rbf	Accuracy	1.00	1.00	1.00	0.99
	Precision	1.00	1.00	1.00	0.99
	Recall	1.00	1.00	1.00	1.00
Polynomial	Accuracy	1.00	1.00	1.00	1.00
	Precision	1.00	1.00	1.00	1.00
	Recall	1.00	1.00	1.00	1.00

Table 1. Accuracy, precision, recall using SVM

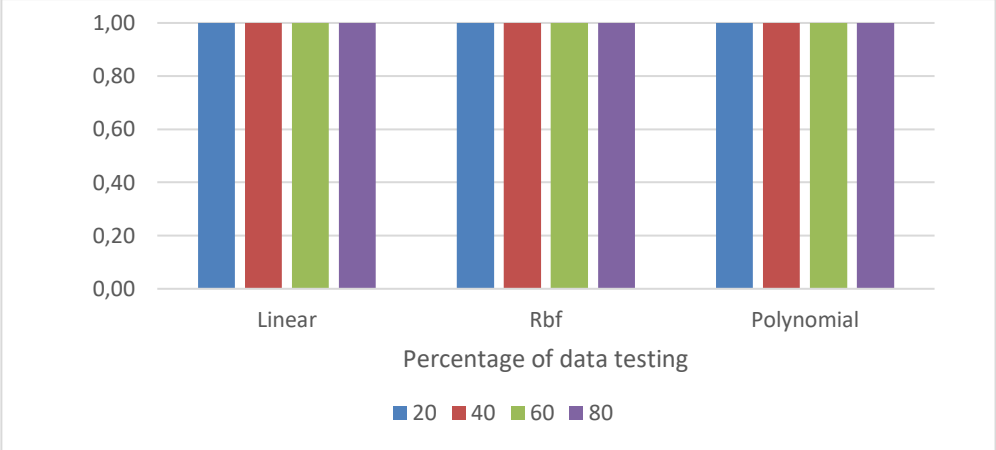

Figure 1. Result of accuracy Linear, Rbf, and Polynomial of SVM

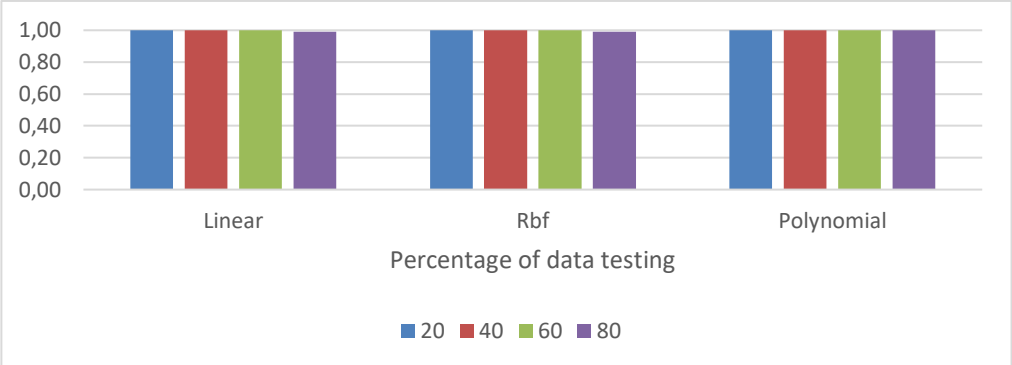

Figure 2. Result of precision Linear, Rbf, and Polynomial of SVM

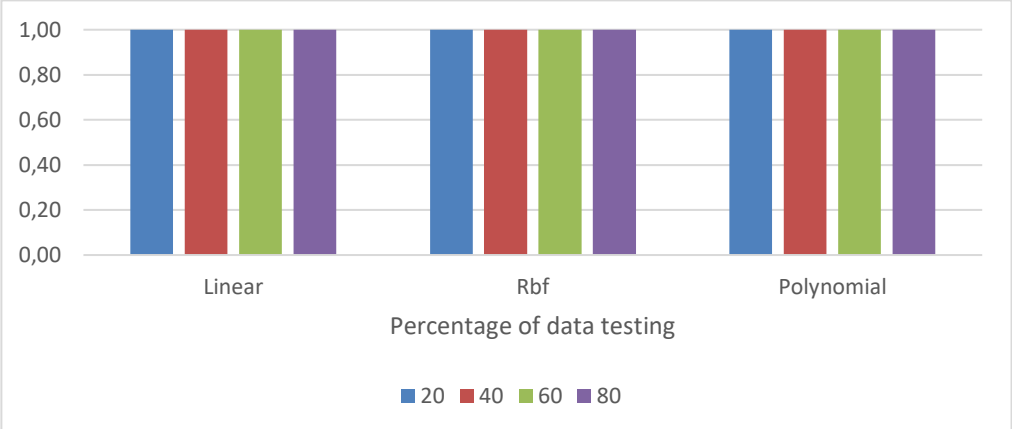

Figure 3. Result of recall Linear, Rbf, and Polynomial of SVM

Table I, Figures 1, Figures 2 and Fig. 3, show that in predicting the morbidity of purchase transactions of electric vouchers, the amount of test data used does not affect the accuracy,

precision and recall value of Linear and Polynomial kernels. The value of accuracy, precision, and recall obtained are all 100%. This shows that with the addition of test data, the value of False Positive and False Negative remains 0. Thus, in each addition of precision value test data, accuracy and recall do not change. However, in the RBF kernel, the accuracy and precision value decreases as the number of test data increases. This is because the value of TN has decreased. While the recall value does not change that is equal to 100%. It can be said that the SVM method can predict morbidity accurately.

4 Conclusions

In this paper, the proposed model uses machine learning techniques. The classification method that has been used to measure the level of classification accuracy is the Support Vector Machine (SVM) method. The results showed that the SVM method using Linear and Polynomial kernels has 100% accuracy, precision and recall. Kernel RBF has accuracy and precision of 99 up to 100%, while the recall value of 100%. However, in order to know which method with the best accuracy between the two methods, further research is needed with more data simulations.Based on the experiment results, we found that the accuracy by using SVM in classifying the behavior of smart electric voucher customers through the purchase transaction history has a good degree of accuracy.

References

[1] S. Sen and V. Agarwal, "Advanced Metering Infrastructure Analytics -A Case Study," pp. 3–8, 2014.

[2] B. Dangar and S. K. Joshi, "Electricity theft detection techniques for metered power consumer in GUVNL, GUJARAT, INDIA," *2015 Clemson Univ. Power Syst. Conf.*, pp. 1–6, 2015.

[3] V. Ford, A. Siraj, and W. Eberle, "Smart Grid Energy Fraud Detection Using Artificial Neural Networks," pp. 0–5, 2014.

[4] T. V. Babu, T. S. Murthy, and B. Sivaiah, "Detecting unusual customer consumption profiles in power distribution systems - APSPDCL," *2013 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2013*, vol. 3, 2013.

[5] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Enhanced encoding technique for identifying abnormal energy usage pattern," *2012 North Am. Power Symp. NAPS 2012*, 2012.

[6] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid,"*IEEE Trans. Ind. Informatics*, vol.12, no.3, pp.1005-1016,2016.

[7] E. W. S. Dos Angelos, O. R. Saavedra, O. A. C. Cortes, and A. N. De Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Trans. Power Deliv.*, vol. 26, no. 4, pp. 2436–2442, 2011.

[8] C. Cody, V. Ford, and A. Siraj, "Decision tree learning for fraud detection in consumer energy consumption," *Proc. - 2015 IEEE 14th Int. Conf. Mach. Learn. Appl. ICMLA 2015*, pp. 1175–1179, 2016.

[9] J. No, S. Y. Han, Y. Joo, and J.-H. Shin, "Conditional abnormality detection based on

AMI data mining," *IET Gener. Transm. Distrib.*, vol. 10, no. 12, pp. 3010–3016, 2016.
[10] C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, "A Practical Guide to Support Vector Classification," *BJU Int.*, vol. 101, no. 1, pp. 1396–400, 2008.
[11] T. Fawcett, "An introduction to ROC analysis," *Irbm*, vol. 35, no. 6, pp. 299–309, 2006.
[12] D Abdullah, Tulus, S Suwilo, S Effendi and Hartono, "DEA Optimization with Neural Network in Benchmarking Process", IOP Conference Series: Materials Science and Engineering, vol. 288. 2018.
[13] Abdullah D, Tulus, Suwilo S & Efendi S, Data Envelopment Analysis with Upper Bound on Output to Measure Efficiency of Department in Malaikulsaleh University. J. Phys.: Conf. Ser. 890, 2017.