Machine Learning and Artificial Intelligence
J.-L. Kim (Ed.)

doi:10.3233/FAIA220434

Improving Anomaly Detection in IoT-Based Solar Energy System Using SMOTE-PSO and SVM Model

Liang-Sian LIN[1], Zhi-Yu CHEN, Yu WANG and Li-Wei JIANG
Department of Information Management, National Taipei University of Nursing and Health Sciences, Ming-te Road, Taipei 112303, Taiwan, R.O.C

Abstract. A large number of sensors based on Internet of Things (IoT) technology are now widely deployed in artificial intelligence, health care monitoring, air quality monitoring, and other fields. The sensors require high power consumption for real-time monitoring data. Some studies have suggested using solar energy for the primary power source to operate sensors. However, due to uncertain climate change, solar energy supply cannot always provide sufficient voltage to operate sensors. Consequently, some abnormal behavior events frequently occur in the IoT system using solar energy. Abnormal detection is a typical imbalanced learning problem due to the very rare amount of abnormal events. Under such data with skewed class distribution, classic classification models fail to provide reliable classification results with abnormal events. Under this condition, in this paper, deploying solar energy supply, we developed an IoT-based system using Arduino Microcontroller and Banana Pi, in which, the SMOTE-PSO algorithm is utilized to improve classification accuracies on abnormal event data in our system. Finally, two types of SVM kernel functions are used to verify classification capability in the developed IoT system.

Keywords. IoT-based system; solar energy supply; imbalanced dataset; abnormal event

1. Introduction

Over the past decade, Internet of Things (IoT) technology has been widely applied in industry and mechanical systems. IoT-based systems often consume high electricity for continuous 24-hour operation. To save electricity cost, some studies have suggested using renewable energies as a main power source for operating IoT sensors [1-3]. Ichikawa et al. [1], for example, recommended running an IoT-based system by solar energy. However, IoT devices operating on solar energy may be unreliable due to uncontrollable weather or environmental factors. Abnormal events may thus frequently occur in IoT-based systems, such as low-voltage or voltage interruption events reducing sensor lifespan and performance [4]. In order to cope with this problem, some studies have suggested machine learning models and classification models to detect solar panel faults [5, 6]. However, the amount of normal events is always very large, and abnormal

[1]Corresponding Author: Liang-Sian Lin, Department of Information Management, National Taipei University of Nursing and Health Sciences, Ming-te Road, Taipei 112303, Taiwan R.O.C; E-mail: lianghsien@ntunhs.edu.tw.

events are very rare. Hence, it is difficult to obtain satisfactory detection accuracy using existing learning models due to this imbalanced data distribution.

1.1. Learning with an imbalanced dataset

Imbalanced data distribution problems occur when the majority class examples largely outnumber the minority class examples [7]. In this circumstance, traditional classifiers fail to provide excellent classification accuracy on minority class [8]. Some studies have suggested a sampling-level approach to balance class distribution [9, 10]. These sampling-level methods can be split into under-sampling and over-sampling methods. The under-sampling method aims at eliminating instances in the majority class. However, it may ignore important information for majority class classification. By contrast, the over-sampling method is used to generate synthetic instances of minority class to increase leaning performance on the minority class. Among the oversampling methods, the Synthetic Minority Over-Sampling Technique (SMOTE) proposed by Chawla et al. [9] is one typical oversampling method. It creates synthetic minority class instances by the linear interpolation method between original minority class instance $x_{original}$ and nearest neighbor $x_{nearest}$ of it, as $x_{synthetic} = x_{original} + \delta \cdot (x_{nearest} - x_{original})$, where δ is a random number between 0 and 1. Some studies have developed the SMOTE method to increase fault detection accuracies in electrical power systems. Li et al. [11], for example, proposed an improved SMOTE method for processing skewed electric charge data sets. Furthermore, Cai et al. [12] utilized the SMOTE method to increase forecasting accuracies of the support vector machine (SVM) model on energy consumption prediction. The experience of these studies demonstrates the SMOTE can effectively improve classification capability for imbalanced datasets. Nevertheless, in order to find the better synthetic instances of minority class, Cervantes et al. [13] proposed the SMOTE-PSO method to evolve synthetic instances of support vectors in the minority class by the SMOTE algorithm to find representative new instances by the particle swarm optimization (PSO) algorithm. On the other hand, some studies have suggested extracting information from sensing data as representative features for anomaly detection in IoT-based systems [14-16].

1.2. Motivation

Based on studies mentioned above, we use real sensing data and sensor power consumption as input features for constructing a SVM model for fault detection of solar power supply in our simulated IoT system. To address imbalanced datasets, we employ the SMOTE-PSO method to balance sensing data between classes. In this paper, on the SVM model, we compared three oversampling methods: the random oversampling (ROS), the SMOTE, and the SMOTE-PSO method. In addition, the classification results are evaluated by G-mean and F-measure. According to our experimental results, on the SVM model, the SMOTE-PSO method can obtain better classification performance than the ROS method in items of G-mean and F-measure. Furthermore, we use Banana Pi M2+ as an edge computing device to save sensor data into a SQLite database.

The remaining part of this paper is organized as follows: Chapter 2 explains evaluation metrics for imbalanced data classification and a brief introduction to the SVM model used in our system; Chapter 3 illustrates the proposed IoT-based system and

provides experimental results on the SVM model; Chapter 4 concludes and discusses future research.

2. Background

In this section, we introduce evaluation criteria for imbalanced data classification. In addition, the SVM model used to detect abnormal events of solar power supply in our system is also explained in the following.

2.1. Evaluation metrics

Classification accuracy is a common metric for assessing classification performance. However, for imbalanced data classification, Han et al. [17] argued accuracy rate (ACC) cannot reflect the prediction performance on minority class. Hu et al. [18] suggested that the G-measure (G-mean) and F-measure (F1) are better evaluation metrics. In this paper, we used G-mean and F1 to assess classification performance for an imbalanced dataset. The confusion matrix is listed in Table 1.

Table 1. Confusion matrix.

		Prediction labels	
		Positive	**Negative**
Actual labels	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

- Accuracy (ACC) expresses the ratio of correct classification as Eq. (1).

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{1}$$

- Precision positive is the number of positive class examples which is classified correctly as Eq. (2).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- Recall (Specificity) metrics are the true positive (negative) rate as Eq. (3) and (4), respectively.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

Based on the above metrics, we conduct G-mean and F1 as Eq. (5) and (6).

$$Gmean = \sqrt{\text{Recall} \times \text{Specificity}} \tag{5}$$

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{6}$$

2.2. Support vector machine

In 1995, based on statistical theory, the support vector machine (SVM) was proposed by Cortes and Vapnik [19]. The SVM model aims at minimizing an upper bound of the generalization error via maximizing the margin with a separating hyperplane, as shown in Figure 1.

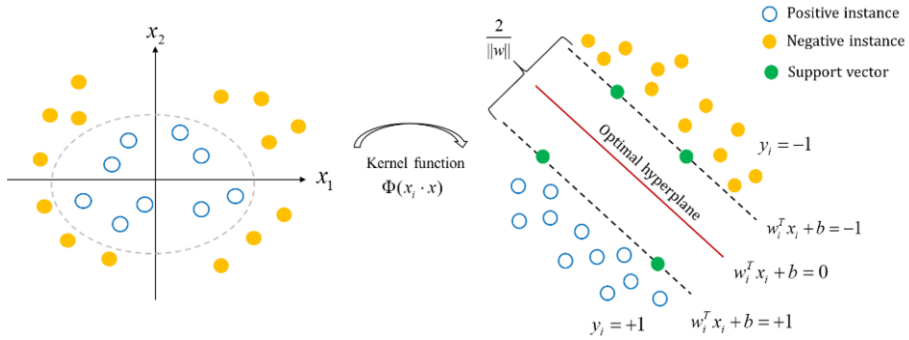

Figure 1. Hyperplane for SVM classification.

The maximum margin hyperplane gives the maximum separation between binary classes. In order to maximize the margin $\frac{2}{||w||}$, instances will be kept from the hyperplane. Given a training set of n samples $(\vec{x}_i, y_i)$, $i = 1, 2, ..., n$ where $\vec{x}_i$ is input variables of ith example, $y_i = +1$ (positive instance) or $y_i = -1$ (negative instance) is output of ith example. Based on the concept of Cortes and Vapnik, the separating hyperplane is defined as $w_i^T x_i + b = 0$, as Eq. (7).

$$f(x) = \sum_{i=1}^{n} w_i^T x_i + b \tag{7}$$

where w is support vector and b is constant. The training instances can be classified by satisfying the constraint, as

$$y_i = \begin{cases} +1, \text{ if } y_i(w_i^T x_i + b) \geq 0 \\ -1, \text{ if } y_i(w_i^T x_i + b) < 0 \end{cases} \tag{8}$$

y_i is the classification label. In addition, Mercer [20] proposed the kernel function $\Phi(x_i \cdot x)$ to map original data onto a high-dimensional space [21] for non-linear data classification, as:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i \cdot x) + b \tag{9}$$

where $\alpha_i \geq 0,\ i = 1, 2, ..., n$ are the Lagrange multipliers and $\Phi(x_i \cdot x)$ is the kernel function.

3. The simulated IoT system and results

In this section, we introduce the proposed IoT-based solar energy system. In addition, the data acquisition and experimental findings are discussed.

3.1. The proposed IoT architecture

The proposed IoT architecture is illustrated as shown in Figure 2. In our system, the Lipo Rider Pro, one solar panel, is used as the primary power supply, and lithium batteries are used as the standby power supply. The Lipo Rider Pro can convert solar energy into 5V power which is provided to run a microcontroller (Arduino Uno), and five sensors. The five sensors include DHT11 sensor used to measure temperature and humidity, GP2Y1010AU0F and SGP30 sensors used to measure PM2.5 and carbon dioxide (CO2) in per million (ppm), respectively, grove-sunlight sensor used to analyze visible (VIS) light, UV light, and infrared (IR) light, and INA219 sensor for measuring current and voltage of these sensors. In Table 2, we summary the information of used sensors and IoT devices.

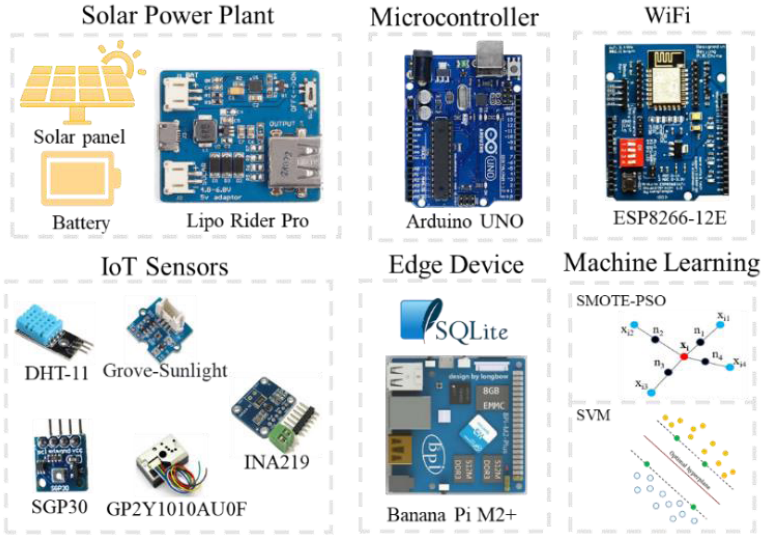

Figure 2. Proposed IoT-based solar energy system.

Table 2. Sensors and IoT devices.

Sensors/IoT	Feature	Specification
DHT11	Temperature detection Humidity detection	Supply voltage: 5 V Temperature range: 0-50°C error of ±1°C Humidity range: 20-90% RH ±1%
GP2Y1010AU0F	PM2.5 detection	Supply voltage: 4.5~5.5 V Concentration range: 0~3000 (ppm)
SGP30	Carbon dioxide (CO_2) detection	Supply voltage: 1.62~1.98 V Concentration range: 0~1000 (ppm)
Grove-sunlight	VIS, UV, and IR spectrum detection	Supply voltage: 3.0~5.5 V Spectrum detection range: 280~950 (nm)
INA219	Current and voltage measurement	Supply voltage: 3.0~5.5 V Bus voltage range: 0V to 26V Current range: ±3.2A
Lipo rider pro	Convert solar energy into 5V and provide stable supply voltage through either lithium battery or solar power	Power supply: 5V
Solar panel	Primary power supply	Typical voltage: 5.5 V
Lithium battery	Standby power supply	Power supply: 3.7 V
Arduino UNO	A microcontroller to operate sensor and capture sensor data via analog and digital pins	Supply voltage: 5 V Digital I/O pins:14; Analog input pins:6
ESP8266-12E	An ESP8266 based WiFi module	Supply voltage: 3.0~5.5 V
Banana Pi M2+	A microcomputer based on ARM 32-bit processor	GPU: Mali400MP2 GPU @600 MHz, Supports OpenGL ES 2.0; CPU: H3 Quad-core Cortex-A7 H.265/HEVC 4K; Memory: 1GB DDR3; I/O pins: 40

In the proposed IoT-based solar energy system, sensor data is transmitted to Banana Pi M2+ through a Wi-Fi module (ESP8266-12). On the Banana Pi M2+, we installed the SQLite database to save sensor data. The operating system is Ubuntu 16.04 on the Banana Pi. If one of these sensors is restarted at least once every 5 mins, it is regarded as an abnormal event in solar energy supply, such as low-voltage, voltage sag, and voltage interruption circumstances. The actual operation of the proposed system is illustrated in Figure 3.

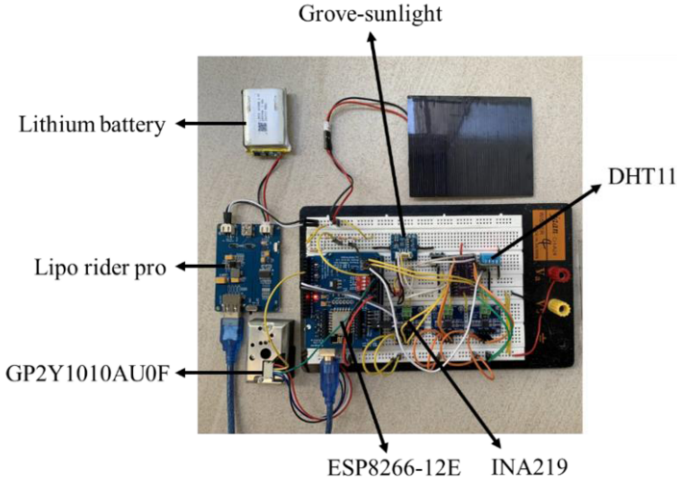

Figure 3. Actual IoT-based solar energy system.

3.2. Sensor data acquisition

In this paper, we extracted the values on the sensors as input features for constructing the SVM model to detect faults of solar energy power. We calculated the average value of sensor data every 5 minutes, as shown in Table 3. The operational status of solar energy supply is defined as binary classes: normal events are labeled as class 0 and abnormal events are labeled as class 1. The abnormal event represents low-voltage or voltage interruption events and normal events represent solar power supply stabilization. In Table 3, the first 12 examples are abnormal event class, and the last 109 examples are normal event class. temperature (Temp), humidity (Hum), CO_2 $PM_{2.5}$, VIS, IR, UV, total power consumption (TPC), and solar power voltage (SPV) extracted are set as input features for building the SVM model.

Table 3. The acquisition of sensor data.

Input features											Output
id	**date**	**Temp**	**Hum**	**CO_2**	**$PM_{2.5}$**	**VIS**	**IR**	**UV**	**TPC**	**SPV**	**class**
Abnormal event											
1	2022/8/2 11:55	23.34	57.00	360.78	3334.11	269.78	295.11	0	199.42	5.91	1
2	2022/8/2 14:00	27.14	71.22	414.78	3253.89	267.00	285.00	0	184.96	0.86	1
…	…	…	…	…	…	…	…	…	…	…	…
12	2022/8/12 10:45	31.60	57.00	0.00	3993.00	347.00	883.00	0	172.20	7.62	1
Normal event											
1	2022/8/12 11:10	25.03	68.00	509.92	3047.92	265.50	291.17	0	172.55	6.15	0
2	2022/8/12 11:15	25.98	67.85	533.08	3050.00	266.69	288.85	0	172.63	6.06	0
…	…	…	…	…	…	…	…	…	…	…	…
109	2022/8/12 11:30	30.14	56.80	400.00	3030.40	333.80	788.00	0	170.52	7.61	0

3.3. The experimental results

In this section, the experiment is implemented with Python 3.8.10 to perform data pre-processing and construct the SVM model. In our experiment, we randomly draw 80 percent of data from an original dataset as a training dataset and the remaining data was set as a testing dataset according to the imbalanced ratio: 109/12. The 30 experiments are implemented to verify classification effectiveness of two types of SVM kernel functions: linear kernel (SVM_linear) and polynomial kernel function (SVM_poly). Additionally, we compare the classification performance using three methods for imbalanced datasets, as IMB (using imbalance dataset), ROS (using random oversampling method), SMOTE, and SMOTE-PSO method. These experimental results are shown in Table 4. In Table 4, the values in bold show the best classification performance in terms of G-mean and F1. For example, on the SVM_poly model, SMOTE-PSO method was improved on IMB method from 0.797 to 0.825 in the item of G-mean.

Table 4. The experimental results.

Model	SVM_linear					
Methods	**ACC**	**Precision**	**Recall**	**Specificity**	**G-mean**	**F1**
IMB	0.926	0.140	0.535	1.000	0.729	0.527
ROS	0.852	0.304	0.673	0.886	0.765	0.627
SMOTE	0.842	0.252	0.650	0.879	0.748	0.601
SMOTE-PSO	0.924	0.417	0.635	0.979	**0.782**	**0.642**
Model	**SVM_poly**					
Methods	**ACC**	**Precision**	**Recall**	**Specificity**	**G-mean**	**F1**
IMB	0.931	0.467	0.658	0.983	0.797	0.670
ROS	0.936	0.583	0.691	0.983	0.818	0.713
SMOTE	0.932	0.506	0.674	0.981	0.807	0.687
SMOTE-PSO	0.939	0.600	0.700	0.984	**0.825**	**0.723**

4. Conclusion and future work

In this paper, we constructed the SVM model to detect abnormal events in solar power supply in our developed IoT system. However, when the ratio of normal events to abnormal events is highly imbalanced, it is difficult to obtain satisfactory detection accuracies on abnormal events using the traditional SVM model. In order to handle this problem, we employed the SMOTE-PSO oversampling method to improve SVM classification on abnormal events in solar power supply. In future research, we consider two study directions: One is considering the temporal variables as the inputs for detecting abnormal events. The other is expanding deployment of more low-cost power sensors for monitoring solar energy use.

Acknowledgement

This study is supported by National Taipei University of Nursing and Health Sciences, Taiwan. The National Science and Technology Council, Taiwan financed this study pursuant to contract number MOST 110-2222-E-227-001-MY2.

References

[1] Ichikawa H, Yokogawa S, Kawakita Y, Sawada K, Sogabe T, Minegishi A, Uehara H. An approach to renewable-energy dominant grids via distributed electrical energy platform for IoT systems. 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm); IEEE; 2019.

[2] Jack KE, Usoro A, Udofa ES, Johnson LA. Real time energy data monitoring model for integrated renewable energy system with other collaborative energy supply. J Electr Electron Eng. 2020 15(6): 33-40.

[3] Eltamaly AM, Alotaibi MA, Alolah AI, Ahmed MA. IoT-based hybrid renewable energy system for smart campus. Sustainability. 2021;13(15):8555.

[4] Sinsel SR, Riemke RL, Hoffmann VH. Challenges and solution technologies for the integration of variable renewable energy sources—a review. Renew Energ. 2020; 145: 2271-2285.

[5] Katoch S, Muniraju G, Rao S, Spanias A, Turaga P, Tepedelenlioglu C, Banavar M, Srinivasan D. Shading prediction, fault detection, and consensus estimation for solar array control. 2018 IEEE Industrial Cyber-Physical Systems (ICPS). IEEE; 2018.
[6] Badr MM, Hamad MS, Abdel-Khalik AS, Hamdy RA, Ahmed S, Hamdan E. Fault identification of photovoltaic array based on machine learning classifiers. IEEE Access. 2021; 9: 159113-159132.
[7] Napierala K, Stefanowski J. Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst. 2016; 46(3): 563-597.
[8] Ganganwar V. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering. 2012; 2(4): 42-47.
[9] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell. 2002; 16: 321-357.
[10] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. 2018; Springer.
[11] Wu L, Yu B, Ju X, Jiang L. An improved SMOTE algorithm for processing unbalanced electric charge data sets. The 2nd International Conference on Computing and Data Science; 2021; p. 1-5.
[12] Cai H, Shen S, Lin Q, Li X, Xiao H. Predicting the energy consumption of residential buildings for regional electricity supply-side and demand-side management. IEEE Access. 2019; 7: 30386-30397.
[13] Cervantes J, Garcia-Lamont F, Rodriguez L, López A, Castilla JR, Trueba A, PSO-based method for SVM classification on skewed data sets. Neurocomputing. 2017; 228: 187-197.
[14] Manco G, Ritacco E, Rullo P, Gallucci L, Astill W, Kimber D, Antonelli M. Fault detection and explanation through big data analysis on sensor streams. Expert Syst Appl. 2017 87: 141-156.
[15] Benkercha R, Moulahoum S. Fault detection and diagnosis based on C4. 5 decision tree algorithm for grid connected PV system. Sol Energy. 2018 173: 610-634.
[16] Praveenchandar J, Vetrithangam D, Kaliappan S, Karthick M, Pegada NK, Patil PP, Rao SG, Umar S. IoT-Based harmful toxic gases monitoring and fault detection on the sensor dataset using deep learning techniques. Sci Program. 2022.
[17] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International Conference on Intelligent Computing; 2005; Springer; p. 878-887.
[18] Hu Y, Guo C, Ngai E, Liu M, Chen S. A scalable intelligent non-content-based spam-filtering framework. Expert Syst Appl. 2010 37(12): 8557-8565.
[19] Cortes C, Vapnik V. Support-vector networks. Mach. Learn. 1995 20(3): 273-297.
[20] Widodo A, Yang BS. Support vector machine in machine condition monitoring and fault diagnosis. Mech Syst Signal Process. 2007 21(6): 2560-2574.
[21] Noble WS. What is a support vector machine? Nat Biotechnol. 2006 24(12): 1565-1567.