Fighting Terrorism: Semi-Supervised Outlier Detection of Electrical Power Consumption

Liyang Tang[1, a], Hongjian Gu[2, b], Liwei Pan[1, c], and Ming Lei[1, d]

[1]China Electronic Technology Group Corporation 38th Research Institute, Key Laboratory of Public Safety Emergency Information Technology of Anhui Province, Hefei 230088, China

[2]Special Investigation Team of Xinjiang Public Security Bureau, Xinjiang 830000, China.

[a]Corresponding author: tangliyang921@gmail.com

[b]759946466@qq.com, [c]panliwei0813@163.com, [d]buaalei001@163.com

Keywords: Electrical power consumption; Outlier detection; Semi-supervised method; Key knowledgeable personnel; Counter-terrorism.

Abstract: Among years of experiences in counter-terrorism investigative practices, it has been found that terrorist organizations and their activities typically exhibit an exceptional demand for electrical power. For example, making tools for criminal purpose such as knives and bombs, or performing underground preaching, concealed gatherings and other illegal activities, might contribute to unusual fluctuation of power consumption. Intriguingly, facilitated by monitoring key knowledgeable personnel behavior, learning anomalies from massive electrical power consumption data helps to timely obtain predictive clues and warnings for fighting terrorism. Along this line, in this paper we propose a semi-supervised approach incorporating power consumption and key personnel monitoring to draw insights on counter-terrorism. Specifically, we: (1) extract features differentiating normal and abnormal power consumption patterns; (2) construct a semi-supervised method, including unsupervised stage to discover suspicious exceptional power consumption from massive data, and supervised stage leveraging key knowledgeable personnel behavior to refine targets; and (3) deploy a prototype system in Xinjiang China, collaborated with local authorities, which practically testifies the interesting association between power consumption and counter-terrorism. Besides, experiments demonstrate satisfactory performance of proposed approach. Indeed, the innovation of combining power consumption and key knowledgeable personnel behavior in perceiving terrorism in advance has profound theoretical and practical significance.

1. Introduction

Terrorism as a social phenomenon, has posed a serious threat to peace and security of mankind. Since the definition of terrorism is controversial, here we consider the implication in Global Terrorism Index (GTI): terrorism refers to “the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation”, while terrorist activity is “an intentional act of violence or threat of violence by a non-state actor” [16]. In fact, China has been struggling with the infiltration of terrorism and secessionism, especially forces from violent terrorists, ethnic secessionists and religious extremists. For example, East Turkestan Islamic Movement (ETIM) in Northwest China such as Xinjiang and Tibet plotted and carried out terrorist bombings, assassinations and seditions [8]. Currently, terrorist activities in China mainly include ideological terrorist activities by ethnic secessionists and religious extremists, socially aggressive terrorism with extreme violence, vicious terrorist criminal activities of individual interests, and terrorist activities by gangs and triads.

Consequently, fighting terrorism is extremely significant for national security and public interest, especially through targeting key knowledgeable personnels. Indeed, criminologists have stated that 80% of crimes are committed by 20% of criminals [19,7], known as the Pareto principle or 80/20 rule. Similarly, the vast majority of illegal and criminal activities related to social stability and public security are associated with only a few persons (e.g. violent terrorists, ethnic secessionists

and religious extremists), which therefore should be especially prioritized and monitored by public security authorities, termed as *key knowledgeable personnel* or *key personnel* in short. By extensively surveillance and control against key personnels, counter-terrorism intelligence from different aspects and various sources could be correlated and integrated, and thus providing early warning information before the terrorist attacks.

Meanwhile, in the provision of energy infrastructure, one of the most straightforward ways to investigate suspicious activities is leveraging electrical power consumption intelligence. Indeed, there are many efforts on diagnostic analysis on power consumption. For example, Aleem *et al.* [1] reviewed the literature on advanced application of fault diagnosis in power systems, with emphasis on reliable fault detection and classification of power system faults. Fontugne *et al.* [6] constructed a model exploring usage pattern of devices within large buildings for operational efficiency, by uncovering relationships between devices. Capozzoli *et al.* [2] performed energy fault detection analysis to detect abnormal consumption in buildings, in order to reduce and optimize energy usage. Khan *et al.* [11] detected abnormal building lighting energy consumption through a neural networks ensemble approach and statistical pattern recognition techniques. Fan *et al.* [5] analyzed the operation of metering devices via power consumption information collection system, in order to screen out abnormal operations, metering failures or electricity stealing activities. Moreover, Rose *et al.* [17] tried to tie electric power system with terrorist attacks by exploring the economic losses from electricity outages. However, their work emphasized on the impact assessment *after* the blackout. Even though fault diagnosis and outlier detection in electric power infrastructure is extensively studied, the attempt of integrating power consumption and key personnel in counter-terrorism is innovative.

To this end, we propose a semi-supervised outlier detection method to incorporate key knowledgeable personnel behavior into exceptional power consumption discovery. Specifically, it consists of two stages: (1) detecting abnormal power consumption behavior based on unsupervised clustering algorithm; and (2) screening out suspicious targets supervised by key personnel information. Furthermore, we deploy the proposed framework in Xinjiang China, collaborated with local authorities, which practically testifies the interesting association between power consumption and counter-terrorism with satisfactory performance.

The remainder of this paper is organized as follows. We present the proposed method thoroughly in Section 2, including overall framework, feature extraction, clustering and outlier detection. Then extensive experiments on a real-world dataset are presented in Section 3. Finally, we conclude the whole paper in Section 4.

2. Proposed Method

In this section, we demonstrate proposed semi-supervised method for targeting person of interest from power consumption data.

Figure 1 illustrates the overall framework. First, we devise a power consumption data collection system for acquiring both historical records and real time data stream. As shown in Figure 2, principle components in the power consumption information collection system include: host, terminals, and metering devices. The host compiles and assigns tasks to terminals, which are responsible for preserving power consumption historical records from electrical metering devices; in addition, host can directly read data through smart metering devices [4]. Then, perform data preprocessing such as cleaning and normalization to prepare data for subsequent analysis.

The first stage of our semi-supervised method is narrowing down to abnormal power consumption behavior from massive power consumption data records, using unsupervised clustering technique. The second stage is further refining suspicious targets within the reduced inspection scope from Stage 1, under the supervision of key personnel information. In this way, we effectively detect suspicious targets from power consumption data together with key personnel information, to provide early warning before terrorist activities or attacks, and therefore prevent and reduce the consequences of terrorism.

In subsequent sections, we present the method in details.

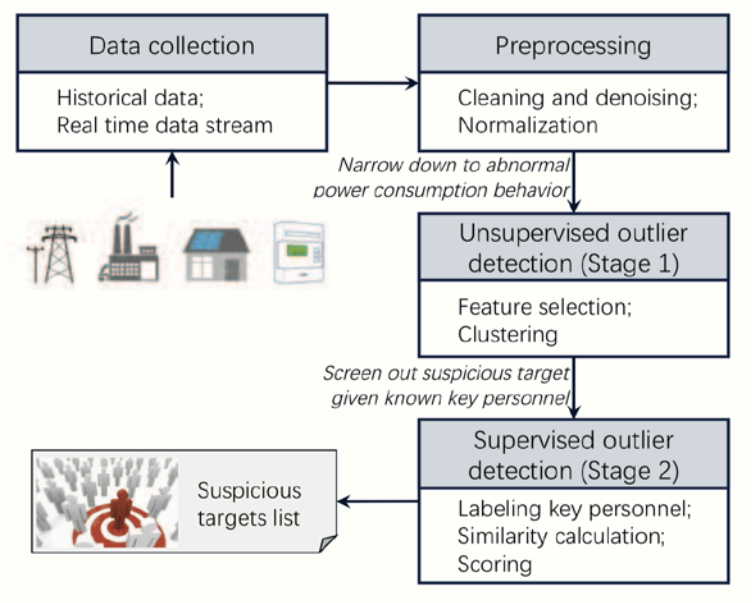


Figure 1. Overall framework of proposed semi-supervised method.

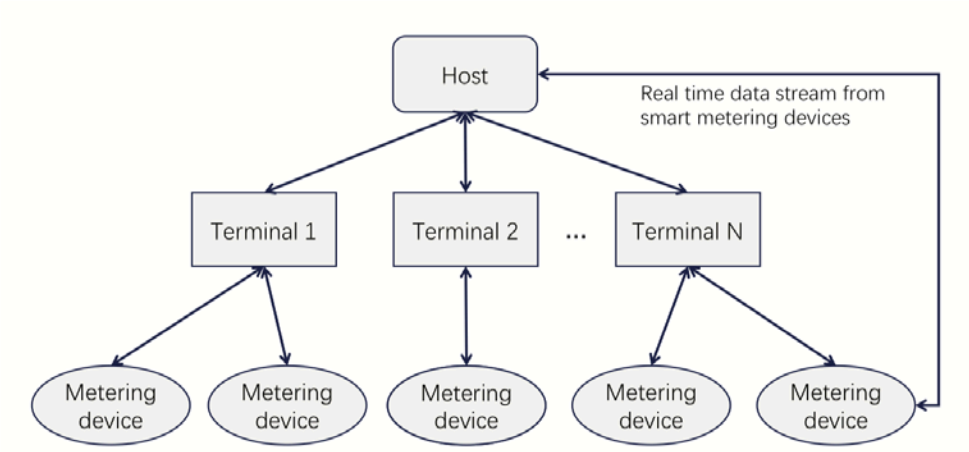


Figure 2. Power consumption information collection system.

2.1 Extracting features from power consumption data

Before performing unsupervised clustering, we first need to do data preprocessing and feature extraction, in order to characterize power consumption and differentiate normal and abnormal behaviors. In this paper, we use a real-world dataset collected from a town in Xinjiang, China.

Figure 3(a) illustrates the distribution of daily power consumption, where x axis denotes the daily power consumption during the whole month (i.e. *today_total*), and y axis denotes the ratio within the dataset. We can easily observe that a large proportion of power consumption is less than 100, which is significantly lower than the average in developed cities (e.g. Beijing). The reason could be that a large number of people live in scarcely populated village, and household appliances are relatively fewer compared to residents in cities. On the other hand, there also exist some heavily consumption of power (e.g. >2000), which might be caused by single metering device shared among multiple residents, or power consumptions of factories, enterprises, and other public places. Thus, the intuition here is that abnormal behavior could be spotted from the power consumption data.

Typically, as for normal power consumption behavior, the ratio of peak time is highest, while the ratio of valley time is lowest. On the contrary, terrorism activities usually happen in late night or early morning (i.e. valley time). Therefore, analysis of power consumption distribution helps to discover anomalies.

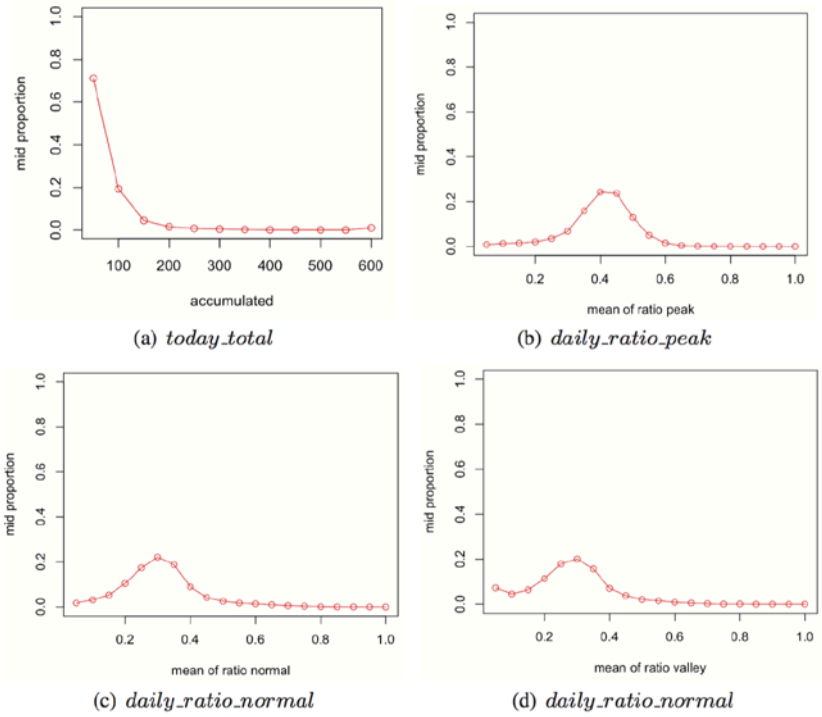
Define the ratios of daily peak/normal/valley power consumption to the total during the day respectively as following:

$$daily\_ratio\_peak = daily\_peak / daily\_total, \quad (1)$$

$$daily\_ratio\_normal = \text{daily\_normal} / daily\_total, \quad (2)$$

$$daily\_ratio\_valley = daily\_valley / daily\_total. \quad (3)$$

Figure 3(b)(c)(d) show the distribution of average ratios in Equations (1) ~ (3), where the medians are 0.45, 0.3 and 0.25 respectively. We can observe that regularly power consumption satisfies *peak* > *normal* > *valley*; however, change or mutation happens in unusual circumstances, which might aid in forecasting terrorism.



(a) *today_total* (b) *daily_ratio_peak*

(c) *daily_ratio_normal* (d) *daily_ratio_normal*

Figure 3. Distribution of daily power consumption.

Another significant category of features is related to the fluctuation of power consumption. Generally, the power consumption behavior of normal residents remains relatively stable within a specified time period; suddenly increasing could indicate additional electricity appliances or a growing population, while sharply decreasing might suggest a possible family migration. In view of that, exceptional behavior could be detected from the fluctuation of power consumption data.

Define *cv_today_total* as the fluctuation coefficient of power consumption of the day:

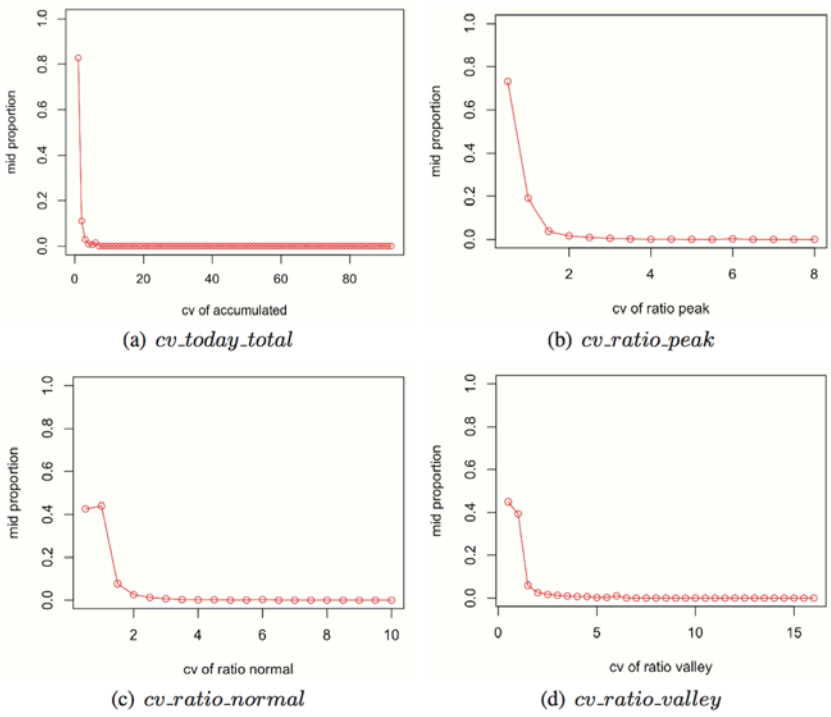
$$cv\_today\_total = \frac{\sqrt{\sum(today\_total - (\sum today\_total/cnt\_date))^2 / cnt\_date}}{\sum today\_total/cnt\_date}, \quad (4)$$

where *cnt_date* denotes the total number of dates.

Larger fluctuation coefficient means greater difference between daily consumptions, and thus more obvious fluctuation. Figure 4(a) plots the distribution of fluctuation of daily power consumption. We can observe that the value of *cv_today_total* is less than 1 for 80% residents, meaning the majority of residents have a stable power consumption pattern. Besides, for more than 99% residents, *cv_today_total* is less than 5, indicating that large fluctuation is abnormal behavior.

Similarly, we calculate *cv_ratio_peak*, *cv_ratio_normal* and *cv_ratio_valley* to capture the fluctuation during the peak/normal/valley time, as shown in Figure 4(b)(c)(d) respectively. Figure 4(b) shows that *cv_ratio_peak* is less than 1 for more than 90% residents, less than 3 for 99% residents, and less than 6 for 99.9% residents; Figure 4(c) shows that *cv_ratio_normal* is less than 1 for more than 80% residents, less than 4 for 99% residents, and less than 6 for 99.9% residents;

Figure 4(d) shows that *cv_ratio_valley* is less than 1 for more than 85% residents, less than 5 for 99% residents, and less than 6 for 99.9% residents. Above observations indicate that large values in variance coefficients reflect exceptional behavior.



(a) *cv_today_total* (b) *cv_ratio_peak*

(c) *cv_ratio_normal* (d) *cv_ratio_valley*

Figure 4. Fluctuation coefficient in different times.

Table 1 summarizes the features extracted from power consumption data, which is further used for clustering.

2.2 Detecting abnormal power consumption using unsupervised clustering

The objective of unsupervised clustering of power consumption data is to discover similar power consumption behavior together with exceptional pattern.

We employ FCM (Fuzzy C-Means) algorithm for clustering, which have been successfully applied in various applications [10]. For example, Liu *et al.* [13] used FCM for network traffic anomaly detection, Kumarage *et al.* [12] employed FCM for anomaly detection in industrial wireless sensor networks, and Yang *et al.* [20] reviewed the usage of FCM in electric load classification. Apparently, FCM outperforms K-Means by allowing one data sample belonging to multiple clusters and introducing the degree of memberships [14].

FCM can be described as Algorithm 1. Given dataset $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$ is the feature vector composed in Section 2.1, x_{iv} is the v-th feature, d is the number of features (i.e. dimensions), and n is the number of residents (i.e. instances). Suppose u_{ij}^m is the membership degree of x_i belonging to cluster j, where m is the fuzzy coefficient, subjected to $\sum u_{ij}^m = 1$. The objective of FCM is to minimize the following function:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m} \left\| x_i - c_j \right\|^2, \tag{5}$$

where c_j is the centroid of cluster j, and $\|\cdot\|$ is the distance function.

Update u_{ij}^m and c_j as following:

$$u_{ij}^{m} = \left(\sum_{q=1}^{k} \left(\frac{\|x_i - c_j\|}{\|x_i - c_q\|} \right)^{2/(m-1)} \right)^{-1}, \tag{6}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}}. \tag{7}$$

Inspired by the usage in Support Vector Machines (SVM) [18], we introduce kernel function to deal with the nonlinearity and high dimension problems. The basic idea of kernel function is to map X into a higher dimension space using a non-linear mapping φ, that is, $\varphi(X) = \{\varphi(x_1), \varphi(x_2), \dots, \varphi(x_n)\}$. Suppose function $K: X \times X \to R$ satisfies:

$$\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \langle \varphi(\mathrm{x_i}), \varphi(\mathrm{x_j}) \rangle, \forall \mathrm{x_i}, \mathrm{x_j} \in \mathrm{X}, \tag{8}$$

where $\langle \cdot \rangle$ denotes dot product, and K is called kernel function.

Perceived from the fundamentals of geometry, cosine measure between vectors can be used to represent the similarity between data samples. Here we use kernel function to simplify distance calculation:

$$\mathrm{d}(\mathrm{x_i}, \mathrm{x_j}) = \frac{\mathrm{x_i} \cdot \mathrm{x_j}}{\|\mathrm{x_i}\| \cdot \|\mathrm{x_j}\|} \approx \mathrm{K}(\mathrm{x_i}, \mathrm{x_j}). \tag{9}$$

Apply RBF kernel [21], we have:

$$\mathrm{d}(\mathrm{x_i}, \mathrm{x_j}) = \mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \exp\left(-\frac{\|\mathrm{x_i} - \mathrm{x_j}\|^2}{2\sigma^2}\right), \tag{10}$$

where $\|\cdot\|$ denotes the Euclidean distance.

Substitute distance calculation in Equations (5)(6) with Equation (10).

ALGORITHM 1. FCM algorithm.

Input: the number of clusters k; fuzzy coefficient $m (1 \leq m \leq 2)$.
Output: a set of clusters $\{C_1, C_2, \dots, C_k\}$.
1: Initialize membership matrix $U = [u_{ij}^m]$;
2: **repeat**
3: **for** cluster j in $\{1, 2, \dots, k\}$ **do**
4: Calculate centroid of cluster C_j as Equation (7);
5: **for** data sample i in $\{1, 2, \dots, n\}$ **do**
6: Reassignment x_i by updating u_{ij}^m as Equation (6);
7: **end for**
8: **end for**
9: $t = t + 1$;
10: **until** J in Equation (5) converges or maximum number of iterations reached
11: **return** $\{C_1, C_2, \dots, C_k\}$

2.3 Refining targets supervised by key personnel

After clustering on dataset X, we have several data clusters, where residents with similar power consumption behavior are grouped together. Accordingly, abnormal power consumption behavior is clustered as well, denoted as suspect S. However, we cannot assert that S is the set of suspected terrorists for sure. The reason is that, even though the power consumption behavior is deviated from ordinary residents, many other factors could contribute to that discrepancy. For example, power consumption of public places and infrastructures typically goes beyond the standard level. Therefore, we need to further screen targets out from S, denoted as T. The objective of supervised outlier detection is to further refine T out of S with known key personnel pattern, as illustrated in Figure 5, where L is the key personnel data.

In this stage, we augment original dataset X with known key personnel data L. Specifically, we have a key personnel database tracking routines of known terrorists and criminals, maintained by local authorities. Similarly, we collect power consumption data by monitoring key personnels, i.e. L. Note that L is labeled dataset, in which each sample is labeled as 1: $L = \{(x_j, 1),\ j = 1,2, \dots, p\}$, where p is the size of key personnels. On the contrary, there is no priori in the cluster with abnormal power consumption behavior: $S = \{(x_i, l_i), i = 1,2, \dots, s\}$, where s is the size of suspected cluster, and label $l_i \in \{1,0\}$ means x_i should be regarded as potential terrorist ($l_i = 1$) or not ($l_i = 0$). Therefore, the task here is to learn l_i for each $x_i \in S$ given L, and thus the positive labeled instances are emitted as targets T.

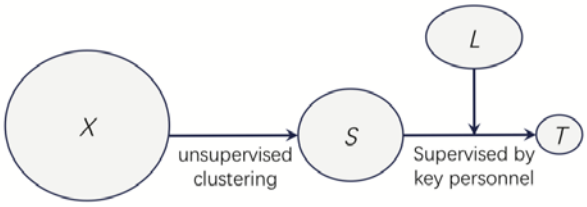


Figure 5. Illustration of dataset evolution.

The basic idea is ranking relevance between suspected power consumption and known anomalies of key personnels. We employ learning-to-rank technique [9] in this stage. Figure 6 demonstrates the basic idea of learning-to-rank framework. Note that since key personnel dataset L is limited, we use the whole labeled L and part of unlabeled S for training; the remaining S is used as test set. Specifically, we utilize Ranking Support Vector Machines (SVM) using scikit-learn toolkit (sklearn) [15].

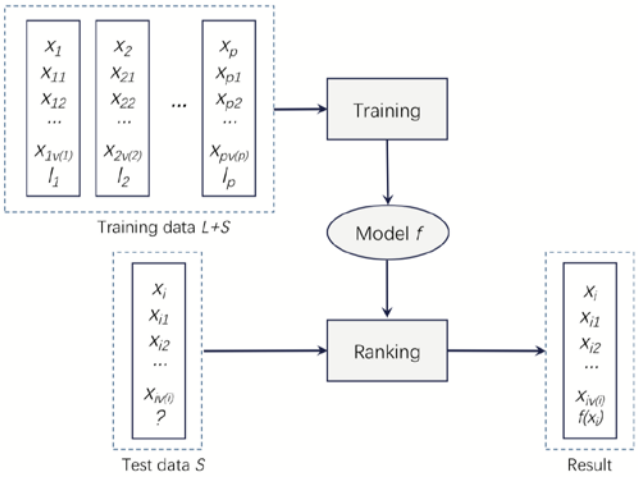


Figure 6. Illustration of learning-to-rank framework.

3. Experiments

We deploy the power consumption information collection system for a whole month and prepare power consumption data during the March, where the computing and storage is built upon Hadoop. The dataset contains 252,321 metering devices and 7,821,951 records in total during 31 days. After removing unreasonable data as following: a) more than 7 days of consecutive failures, or b) negative value in power consumption, we have 49,063 metering and 203,258 records. We preprocess data using Gaussian smoothing to eliminate fluctuations and noises, and all data fields in Table 1 are normalized within [0,1].

Stage 1. We empirically select No. 38~44 features in Table 1 for clustering. Figure 7 demonstrates the clustering result in a two-dimensional plot, where axes denote features of training data and different colors denote 5 clusters. The following two clusters exhibit exceptional behaviors:

Type (a) Accumulated power consumption is large, which has been verified as enterprises, factories and other public places;

Type (b) Fluctuation during the valley time is significant, which obviously goes beyond the normal pattern. This exceptional power consumption behavior cluster is further investigated in Stage 2.

As shown in Table 2, proposed clustering outperforms others. The metrics used to evaluate the performance of clustering are as follows.

(1) SSE (Sum of Squared Errors):

$$\mathrm{SSE} = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i - c_j \right\|^2, \tag{11}$$

which measures the total error. The smaller, the better.

(2) The ratio of Intra- and Inter- Clusters (IIC), defined as:

$$\mathrm{IIC} = \frac{\sum_{i,j}^{k}\|c_i - c_j\|^2}{\sum_{r=0}^{k}\sum_{i=0}^{n_r}\|x_i - c_r\|^2}, \tag{12}$$

where c_i, c_j, c_r denote the centroid of clusters, n_r is the size of cluster r, k is the number of clusters, x_i denote each element within cluster, and $\|\cdot\|$ is the Euclidean distance. The numerator measures the distance between different clusters, and the denominator indicates the differences within specific cluster. Larger IIC means better clustering.

(3) Dunn index [3]:

$$\mathrm{DI} = \frac{\min_{m,n\in K}\left\{\min_{\forall x_i\in C_m,\forall x_j\in C_n}\|x_i - x_j\|\right\}}{\max_{m\in K}\max_{\forall x_i,x_j\in C_m}\|x_i - x_j\|}, \tag{13}$$

where C denotes cluster, $m,\ n$ denote the index of cluster, k is the number of clusters, and x_i denote each element within cluster. The larger DI is, the better clustering is.

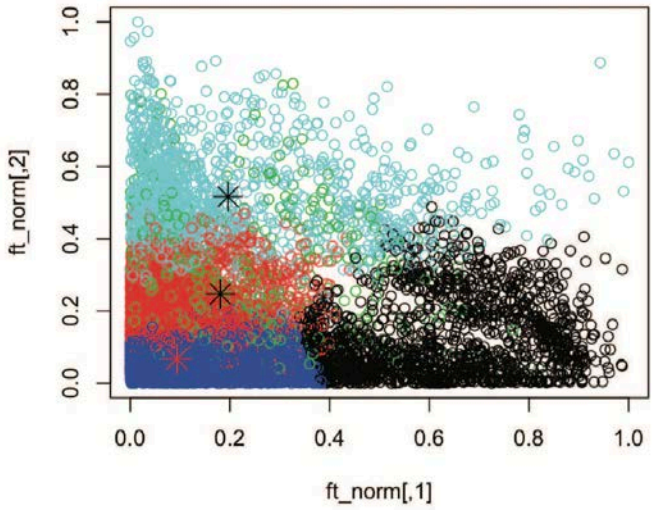
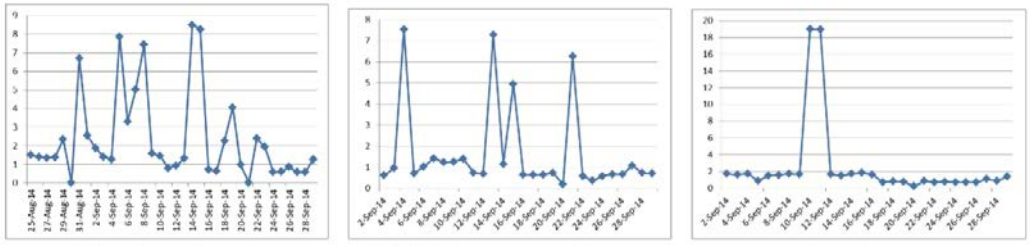


Figure 7. Clustering results.

Table 2. Evaluation of clustering algorithms.

Metrics	SSE	IIC	DI
K-Means	486.625	5.543	6.132
FCM	364.732	7.448	7.997
Ours	210.921	8.962	9.083

Stage 2. We also monitor the power consumption behavior of key personnels. For example, Figure 8 shows the power consumption behavior of criminals. Obviously, all three cases peak at nights, which is normally the valley time instead.



(a) making bombs (b) high power appliances (c) gathering

Figure 8. Power consumption peaks during nights of criminals.

We validate proposed outlier detection method for discovering potential terrorism through an

implementation in a town in Xinjiang, China. Figure 9 (left) shows a screenshot of our predictive warning implementation, where red spots represent locations of exceptional power consumption learned by proposed method. Besides, predicted targets are ranked in Figure 9 (right), represented by the intensity of color. The detection rate is approximately 87.9%, and the false alarm rate is about 9.2%. The performance is satisfactory in practical application indeed.

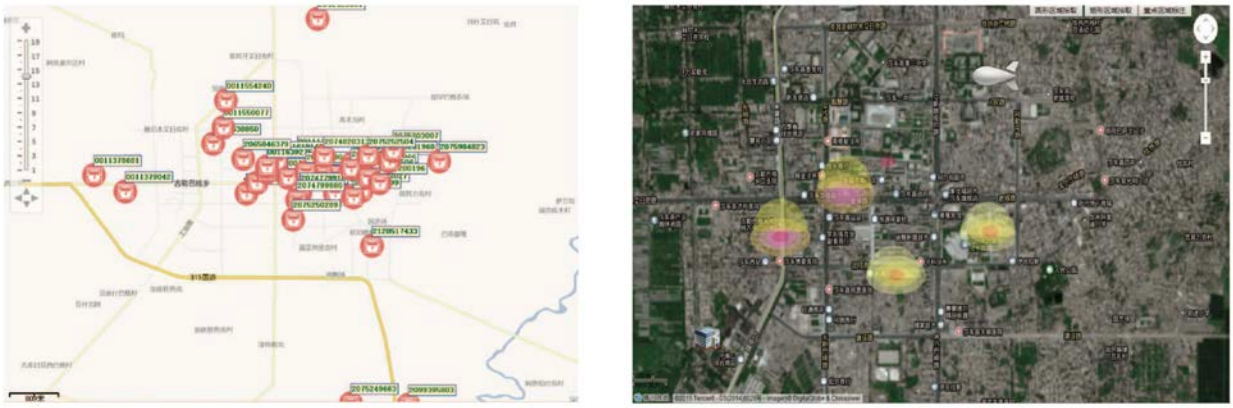


Figure 9. Screenshots of prototype implementation for discovering potential terrorism

4. Conclusion

In this paper, we attempt to leverage electrical power consumption data analysis for discovering potential terrorism targets. The idea is first learning from original power consumption data to find abnormal behavior, and then introducing key personnel information from public security domain to refine the detected outliers. Note that we fusion power consumption data from the power energy system and key personnel data from public security authorities. Our work reveals the insight that counter-terrorism efforts could be achieved through data analytics from various domains, especially those related to public infrastructures.

However, we utilize power consumption data only, without consideration of smart appliances, sensors or signals. In future work, we try to integrate various sources of data from smart grid with the operational data from public security authorities. Moreover, we would consider optimize the implementation in practical promotion.

Acknowledgments

This research work was partly supported by National Key Research and Development Program of China (Grant No. 2016YFC0800100), Major Research Program of the National Natural Science Foundation of China (Grant No. 91546103), and Anhui Provincial Natural Science Foundation (Grant No. 1708085QG162).

References

[1] Aleem, S.A., Shahid, N., Naqvi, I.H., “Methodologies in power systems fault detection and diagnosis”. *Energy Systems* 6(1), 85-108 (2015).

[2] Capozzoli, A., Lauro, F., Khan, I., “Fault detection analysis using data mining techniques for a cluster of smart office buildings”. *Expert Systems with Applications* 42(9), 4324-4338 (2015).

[3] Dunn, J.C., “A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters”. (1973).

[4] Fadel, E., Gungor, V., Nassef, L., Akkari, N., Maik, M.A., Almasri, S., Akyildiz, I.F., “A survey on wireless sensor networks for smart grid”. *Computer Communications* 71, 22-33 (2015).

[5] Fan, J., Chen, X., Zhou, Y., “An intelligent analytical method of abnormal metering device based on power consumption information collection system”. (in Chinese). *Electrical Measurement*

and Instrumentation 50(11), 4-9 (2013).

[6] Fontugne, R., Ortiz, J., Tremblay, N., Borgnat, P., Flandrin, P., Fukuda, K., Culler, D., Esaki, H., "Strip, bind, and search, a method for identifying abnormal energy consumption in buildings". *Proceedings of the 12th international conference on Information processing in sensor networks*, pp. 129-140. ACM (2013).

[7] Groff, E., McCord, E.S., "The role of neighborhood parks as crime generators". *Security journal* 25(1), 1-24 (2012).

[8] Hu, A., Ma, W., Yan, Y., "silk road economic belt, Strategic connotation, orientation and implementation path". (in Chinese). *Journal of Xinjiang Normal University (Philosophy and Social Sciences)* 1 (2014).

[9] HUANGZhen-Hua, ZHANGJia-Wen, TIANChun-Qi, SUNSheng-Li, Yang, X., "Survey on learning- to-rank based recommendation algorithms". (in Chinese). *Journal of Software* 27(3), 691-713 (2016).

[10] Izakian, H., Pedrycz, W., "Anomaly detection and characterization in spatial time series data, A cluster- centric approach". *IEEE Transactions on Fuzzy Systems* 22(6), 1612-1624 (2014).

[11] Khan, I., Capozzoli, A., Lauro, F., Corgnati, S.P., Pizzuti, S., "Building energy management through fault detection analysis using pattern recognition techniques applied on residual neural networks". *Italian Workshop on Artificial Life and Evolutionary Computation*, pp. 1-12. Springer (2014).

[12] Kumarage, H., Khalil, I., Tari, Z., Zomaya, A., "Distributed anomaly detection for industrial wireless sensor networks based on fuzzy data modelling". *Journal of Parallel and Distributed Computing* 73(6), 790-806 (2013).

[13] Liu, D., Lung, C.H., Lambadaris, I., Seddigh, N., "Network traffic anomaly detection using clustering techniques and performance comparison". *Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on*, pp. 1-4. IEEE (2013).

[14] Nayak, J., Naik, B., Behera, H., "Fuzzy c-means (fcm) clustering algorithm, a decade review from 2000 to 2014". *Computational Intelligence in Data Mining*-Volume 2, pp. 133-149. Springer (2015).

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-hofer, P., Weiss, R., Dubourg, V., et al., "Scikit-learn, Machine learning in python". *Journal of Machine Learning Research* 12(Oct), 2825-2830 (2011).

[16] Ratcliffe, J.H., "Intelligence-led policing". *Routledge* (2016).

[17] Rose, A., Oladosu, G., Liao, S.Y., "Business interruption impacts of a terrorist attack on the electric power system of los angeles, customer resilience to a total blackout". *Risk Analysis* 27(3), 513-531 (2007).

[18] Scho¨lkopf, B., Smola, A.J., "Learning with kernels, support vector machines, regularization, optimization, and beyond". *MIT press* (2002).

[19] Sherman, L.W., "The power few, experimental criminology and the reduction of harm". *Journal of Experimental Criminology* 3(4), 299-321 (2007).

[20] Yang, S.l., Shen, C., et al., "A review of electric load classification in smart grid environment". *Renew- able and Sustainable Energy Reviews* 24, 103-110 (2013).

[21] Zhou, S.K., Chellappa, R., "From sample similarity to ensemble similarity, Probabilistic distance measures in reproducing kernel Hilbert space". *IEEE transactions on pattern analysis and machine intelligence* 28(6), 917 (2006).

Table 1. List of features of power consumption.

ID	Data field	Explanation
1	*measure_id*	terminal id
2	*cnt_date*	length of metering period in days
3	*resident_id*	householder metering device id
4	*resident_address*	address of householder metering device
5	*sum_pc_total*	sum total of power consumption of the day
6	*mean_pc_total*	average of power consumption of the day
7	*var_pc_total*	variance of power consumption of the day
8	*max__pc_total*	maximum power consumption of the day
9	*min_pc_total*	minimum power consumption of the day
10	*sum_pc_peak*	sum of power consumption during the peak time
11	*mean_pc_peak*	mean of power consumption during the peak time
12	*var_pc_peak*	variance of power consumption during the peak time
13	*max__pc_peak*	maximum power consumption during the peak time
14	*min_pc_peak*	minimum power consumption during the peak time
15	*sum_pc_valley*	sum of power consumption during the valley time
16	*mean_pc_valley*	mean of power consumption during the valley time
17	*var_pc_valley*	variance of power consumption during the valley time
18	*max__pc_valley*	maximum power consumption during the valley time
19	*min_pc_valley*	minimum power consumption during the valley time
20	*sum_pc_normal*	sum of power consumption during the normal time
21	*mean_pc_normal*	mean of power consumption during the normal time
22	*var_pc_normal*	variance of power consumption during the normal time
23	*max__pc_normal*	maximum power consumption during the normal time
24	*min_pc_normal*	minimum power consumption during the normal time
25	*mean_ratio_peak*	average ratio of peak to total
26	*var_ratio_peak*	variance in ratio of peak to total
27	*max_ratio_peak*	maximum ratio of peak to total
28	*min_ratio_peak*	minimum ratio of peak to total
29	*mean_ratio_normal*	average ratio of normal to total
30	*var_ratio_normal*	variance in ratio of normal to total
31	*max_ratio_normal*	maximum ratio of normal to total
32	*min_ratio_normal*	minimum ratio of normal to total
33	*mean_ratio_valley*	average ratio of valley to total
34	*var_ratio_valley*	variance in ratio of valley to total
35	*max_ratio_valley*	maximum ratio of valley to total
36	*min_ratio_valley*	minimum ratio of valley to total
37	*sum_pc_total_log*	log of *sum_pc_total*
38	*cv_pc_total*	coefficient of variance of power consumption of the day
39	*cv_pc_peak*	coefficient of variance of power consumption during the peak time
40	*cv_pc_normal*	coefficient of variance of power consumption during the normal time
41	*cv_pc_valley*	coefficient of variance of power consumption during the valley time
42	*cv_ratio_peak*	coefficient of variance of ratio during the peak time
43	*cv_ratio_normal*	coefficient of variance of ratio during the normal time
44	*cv_ratio_valley*	coefficient of variance of ratio during the valley time
45	*ratio_max_pc_total*	ratio of maximum power consumption to daily total
46	*ratio_max_pc_peak*	ratio of maximum power consumption during the peak time to daily total
47	*ratio_max_pc_normal*	ratio of maximum power consumption during the normal time to daily total
48	*ratio_max_pc_valley*	ratio of maximum power consumption during the valley time to daily total