Anomaly Detection for Environmental Data Using Machine Learning Regression

Fuqing Yuan and Jinmei Lu

Department of Technology and Safety, UiT the Arctic University of Norway, 9037
Tromsø, Norway
Email: yuan.fuqing@uit.no

Abstract. Environmental data exhibits as huge amount and complex dependency. Utilizing these data to detect anomaly is beneficial to the disaster prevention. Big data approach using the machine learning method has the advantage not requiring the geophysical and geochemical knowledge to detect anomaly. This paper using the popular support vector regression (SVR) to model the correlation between factors. From the residual of the regression, it develops a statistical method to quantify the extremity of some abnormal observed data. A case study is proposed to demonstrate the developed methods.

1. Introduction

Big data is a hot topic almost in all the engineering subjects [1]. The current challenges in engineering data crisis need a method can handle the data automatically to save analytic cost. In environmental engineering, the amount of data is unexceptionally huge. In Norway, the meteorological authorities collects all the temperature, humidity, wind, ice etc. data in their database, in order to utilize the data into helpful decision making, such as optimize the road traffic, give early operating warning for its big oil and gas industry, control its air traffic, the available data is a huge resources.

From the available to detect anomaly event such as abnormal high temperature, humidity, wind speed could herald nature disaster. In norther Norway, the temperature, wind speed correlates to the avalanche in the mountainous area. An abnormal temperature, eight too high or too low, could incur avalanche. On the road, unusual temperature in some seasons could also induce more traffic accidents. Since amount of data is huge and one are not able, perhaps not necessary to investigate the physical connection between the observed data and the correlated disaster. A method can correlate the environmental data and the disaster can benefit the society.

2. Methodology

Machine learning is key to solve big data problem. The development of machine learning is rapidly. For some machine learning, the abnormal pattern is detected based on the geometric distance between data sets. This method evaluates the similarity or dependency in terms of distances between data sets. This distance is not limited to the geometrical distance but also be the abstract distance, e.g. Euclidean distance, Riemannian distance, Mahalanobis distance, or Kullback-Leibler distance [2]. A simple machine learning method such as the K-Nearest Neighbor (KNN) is Euclidean distance based [3]. The Support Vector Machine (SVM) is Riemannian distance based [4]. The Artificial Neural Network (ANN) is diverse. In state of art, one can find all abstract distance based ANN.

The fundamental principle for most machine learning methods is general. In the problem this paper considering, the outlier or anomaly of some environmental measurement should be detected. For time-independent environmental factor, we can use machine learning method to define the boundary of the normal data, as shown in the left of Figure 1. For time-dependent factor, we can employ the machine learning regression model to remove the time effect. Thereafter we define the boundary for the normal data, the abnormal data then can be readily detected [5], as shown in the right of Figure 1. This paper uses the second one.

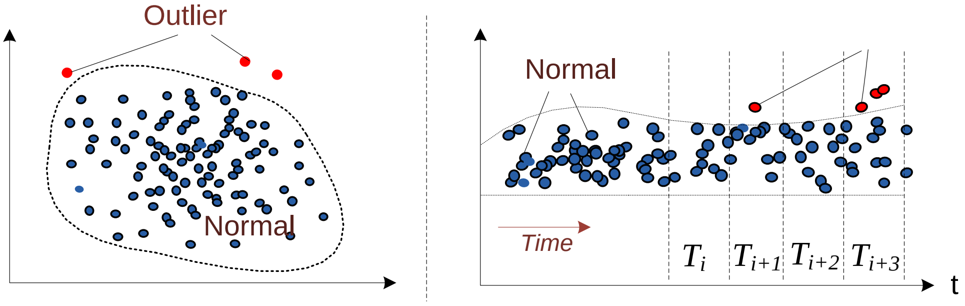


Figure 1. Approach to detect abnormal event

2.1. Machine Learning Regression

Support Vector Regression (SVR) is a nonlinear regression method. It can be realized by considering it as a support vector classifier. Suppose the desired regression function is $f(x)$. As shown in Figure 2, the data positioning in upper of the desired regression function $f(x)$ is considered as class 1, the data below $f(x)$ is considered as class 2. SVR converts the regression problem into a special classification problem. SVR uses soft margin to tolerate misclassification, i.e. the ε-insensitive loss function [6].

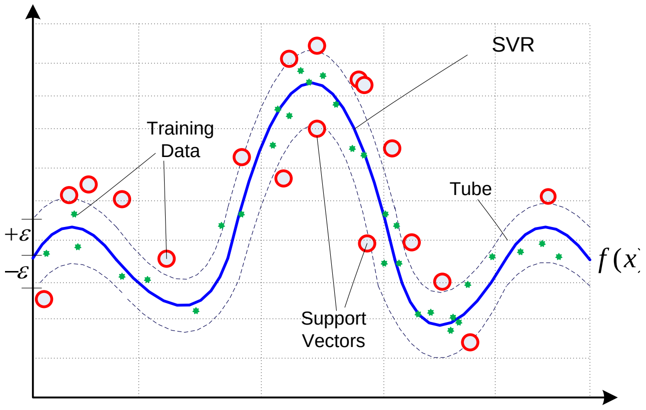


Figure 2. Support Vector Regression

Alike the support vector machine, the primal problem of SVR is to solve the problem:

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ &s.t.\ \ f(x_i) - y_i \le \varepsilon + \xi_i \\ &\quad\ \ y_i - f(x_i) \le \varepsilon + \xi_i^* \\ &\quad\ \ \xi_i \ge 0, \xi_i^* \ge 0.\ i = 1,2,3,....m. \end{aligned} \tag{1}$$

Then introducing Lagrangian multipliers, a dual problem of Formal (1) is formatted as:

$$
\begin{aligned}
&max \quad W(\alpha, a^*) = -\varepsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \left\langle x_i, x_j \right\rangle \\
&s.t. \quad \alpha_i \geq 0, \alpha_j \geq 0; \\
&\qquad\quad \alpha_i^* \leq C, \alpha_j^* \leq C; \\
&\qquad\quad \sum_{i=1,j}^{m} (\alpha_i^* - \alpha_i) = 0, \ i = 1, 2, ..., m
\end{aligned}
\tag{2}
$$

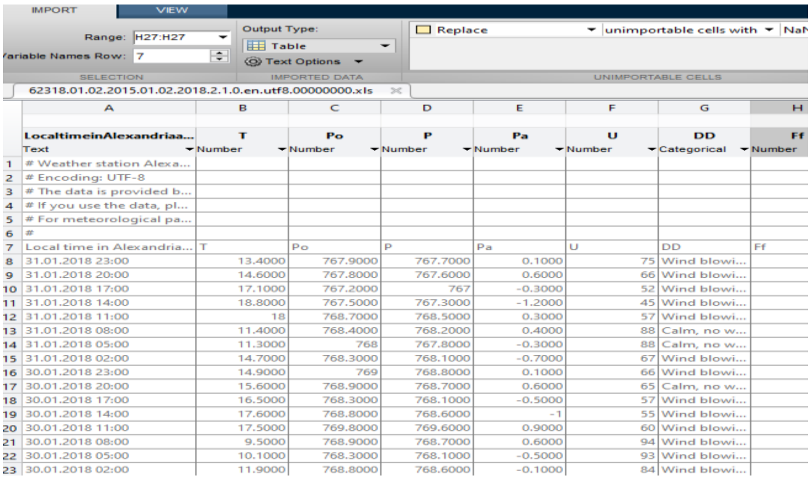
If we let the $<x_i,x_j>$ replaced by a kernel function $K<x_i,x_j>$. The desired function $f(x)$ is as follows:

$$
f(x) = \sum_{j=1}^{m} (\alpha_i^* - \alpha_i) K(x, x_j) + b \tag{3}
$$

The SVR is a black box method. In applying the method, we don't write the regression function as (3) explicitly, since it is not necessary as we solve the problem using software. In environmental engineering, it is not necessary to investigate the geophysical or geochemical background of these regression parameters α and b. There are data-driven and just adaptive to the data.

3. Case Study

The above mentioned approach is applied to detect the abnormal high or low temperature for environmental data collected in a Russian airport [7]. The data ranges from year 2016 to 2018. The January data is analysed in the case study. Original raw data for each day measured 8 times. We take the average of a day as this day's temperature. Totally 88 datasets is extracted from the raw data. Figure 3 shows screen shot of the original data.



Range: H27:H27
ariable Names Row: 7
Output Type: Table
Replace unimportable cells with NaN

62318.01.02.2015.01.02.2018.2.1.0.en.utf8.00000000.xls

	A	B	C	D	E	F	G	H
	LocaltimeinAlexandriaa...	T	Po	P	Pa	U	DD	Ff
	Text	Number	Number	Number	Number	Number	Categorical	Number
1	# Weather station Alexa...							
2	# Encoding: UTF-8							
3	# The data is provided b...							
4	# If you use the data, pl...							
5	# For meteorological pa...							
6	#							
7	Local time in Alexandria...	T	Po	P	Pa	U	DD	Ff
8	31.01.2018 23:00	13.4000	767.9000	767.7000	0.1000	75	Wind blowi...	
9	31.01.2018 20:00	14.6000	767.8000	767.6000	0.6000	66	Wind blowi...	
10	31.01.2018 17:00	17.1000	767.2000	767	-0.3000	52	Wind blowi...	
11	31.01.2018 14:00	18.8000	767.5000	767.3000	-1.2000	45	Wind blowi...	
12	31.01.2018 11:00	18	768.7000	768.5000	0.3000	57	Wind blowi...	
13	31.01.2018 08:00	11.4000	768.4000	768.2000	0.4000	88	Calm, no w...	
14	31.01.2018 05:00	11.3000	768	767.8000	-0.3000	88	Calm, no w...	
15	31.01.2018 02:00	14.7000	768.3000	768.1000	-0.7000	67	Wind blowi...	
16	30.01.2018 23:00	14.9000	769	768.8000	0.1000	66	Wind blowi...	
17	30.01.2018 20:00	15.6000	768.9000	768.7000	0.6000	65	Calm, no w...	
18	30.01.2018 17:00	16.5000	768.3000	768.1000	-0.5000	57	Wind blowi...	
19	30.01.2018 14:00	17.6000	768.8000	768.6000	-1	55	Wind blowi...	
20	30.01.2018 11:00	17.5000	769.8000	769.6000	0.9000	60	Wind blowi...	
21	30.01.2018 08:00	9.5000	768.9000	768.7000	0.6000	94	Wind blowi...	
22	30.01.2018 05:00	10.1000	768.3000	768.1000	-0.5000	93	Wind blowi...	
23	30.01.2018 02:00	11.9000	768.8000	768.6000	-0.1000	84	Wind blowi...	

Figure 3. Screenshot of raw data

SVR is applied to model the correlation between the date and the average, i.e. the input for the SVR is date and the corresponding output for SVR is the temperature. It is a rather simple case with one dimensional input data. The penalty C is 1000000 and Epsilon is 0.5 for soft margin. Nonlinear Gaussian function is chosen as the kernel function with parameter 0.1. As the regression is an expression of kernel functions, we ignore the regression expression here. By using the regression function for the date, a figure is plot as shown in Figure 4. The predicted data is centred at the real data. The error plot in the right, it shows the error centred around 0 and roughly follows Normal distribution. It implies the obtained regression function can capture the information quite well, so that we can use them for abnormal detect.

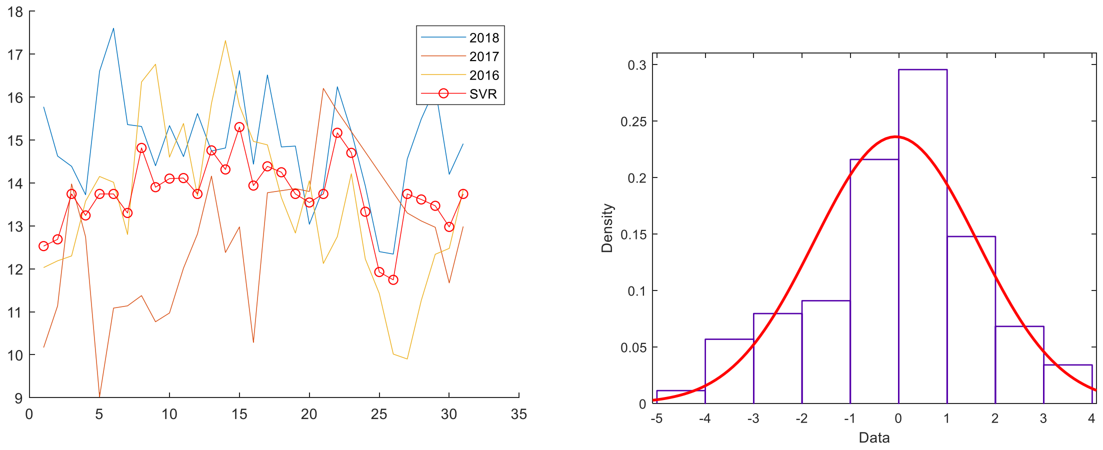


Figure 4. Predicted data and original data

We perform a residual analysis. The mean of the error is - 0.06234 and the standard deviation is 1.69. For a date with residual locates at extremes of the distribution can be considered as abnormal. As shown in Figure 5, on the right tail of the distribution, the date locates in the 2% area is abnormal.

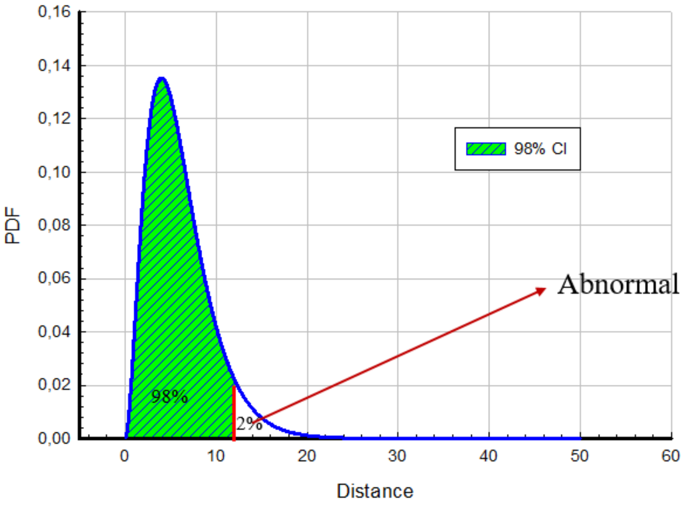


Figure 5. Abnormal Definition

For the data shown in Figure 3, if we defines the date locates within 1% of the extremum area as the abnormal date, we can find 2017.Jan. 05 is abnormal date. Above the level such as 0.999, we cannot find any abnormal for these three years. The corresponding extreme level and the abnormal dates are shown in Table 1. The extreme level 0.95 in the table are just for demonstration of the method. If we define the extreme level as 0.99, we found the 2017.Jan.5th has extreme temperature, i.e. it is an abnormal date.

Table 1. Abnormal Date

Extreme Level	**Abnormal Date**
0.95	*2017.jan.05; 2017.jan.08; 2017.jan.16; 2016.jan. 27; 2018.jan.06.*
0.99	2017.Jan. 05
0.999	None
0.9999	None
0.99999	None

4. Conclusions

The big data using machine learning method shows the feasibility for abnormal detect. SVR can successfully to capture the information of the correlation between date and the temperature. The abnormal date can be figured out from the residual analysis. One can generalize the approach to analyse humidity, wind speed or other environmental factors. The case study also shows the advantage of the machine learning method without requiring the geophysics or geochemical knowledge to detect the abnormal.

Acknowledgement

We appreciate the support from the EU Interreg Min-north project. The project provides the original idea of the method and the finance for the research.

References

[1] Jo J and Lee K W, "High-Performance Geospatial Big Data Processing System Based on MapReduce," *Isprs International Journal of Geo-Information*, vol. 7, Oct 2018.

[2] Patrick E A and Fattu J M, *Artificial intelligence with statistical pattern recognition*. Englewood Cliffs, N.J.: Prentice-Hall, Business and Professional Division, 1986.

[3] Theodoridis S and Koutroumbas K, *Pattern recognition*, 3rd ed. San Diego, CA: Academic Press, 2006.

[4] Amari S and Wu S, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks,* vol. 12, pp. 783-789, Jul 1999.

[5] Chen K Y, "Forecasting systems reliability based on support vector regression with genetic algorithms," *Reliability Engineering & System Safety,* vol. 92, pp. 423-432, Apr 2007.

[6] Schölkopf B and Smola A J, *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Cambridge, Mass.: MIT Press, 2002.

[7] Weather archive in Alexandria (airport) [Online]. Available: https://rp5.ru/Weather_archive_in_Alexandria_(airport)