Publishing journal: Sensors

Disguised Missing Values

Clarification

It may seem from the title of [2] that we have chosen the incorrect paper and therefore we should clarify our thinking that led us to this paper. From our understanding of [1], disguised missing values are any values that appear to be true but are in fact false or indirectly express empty values. In [1], researchers gave an example from data used to diagnose diabetes. This data contains a column for blood pressure in patients, and the column contains several zero values, but it is impossible for a person with this characteristic to exist, so the zero values in this case are disguised missing values. However, some people may not know this information and thus will not notice the presence of disguised missing values, which will negatively affect any procedures on the data. Therefore, methods must be developed to detect these values. Researchers have classified these methods into two categories: a category for detecting extreme values (outlier detection), which is based on the assumption that disguised missing values are outside the data distribution, and the other category is local extreme data (inlier detection), which is considered a more difficult problem to solve, as disguised missing values have acceptable values, but if they are studied well, they may be classified as false values.

Based on the above, the choice of the paper was successful - or so we hope - because the values of any sensors are considered disguised missing values if the values are outside the range of acceptable values (there is a downside to this assumption that will be explained in the section on limitations and downsides).

Method

The researchers proposed an anomaly detection and repairing method in an air quality monitoring system, where sensors measure four gases O3, CO, NO2, NO, temperature and humidity for two channels, a primary channel and a secondary channel. The sensors output in millivolts (raw output), and the detection and correction process is done at this

stage, and then the data is used to train a calibration model to convert the raw measurements into the concentrations of each gas in the air. The researchers used a time series to model this problem.

1. Anomaly detection

The anomaly detection process is done by a majority voting (MV) system consisting of three algorithms. We will talk about each algorithm separately.

a. Sliding window anomaly detection (SWAD)

This algorithm is applied to a set of k measurements. The arithmetic mean, standard deviation, upper limit and lower limit are calculated using the equations:

$$IQR = q_3 - q_1$$
$$lower\ bound = q_1 - q * IQR$$
$$upper\ bound = q_3 - q * IQR$$

q_1 = first quartile
q_3 = third quartile
q = 6

When a new measurement x_{t+1} is received, the following equation is applied:

$$|x_{t+1} - x_t| > \varepsilon$$

The value of ε is determined by experts or using statistical methods. If the inequality is satisfied, x_t is classified as an outlier, but the disadvantage of this method is that it cannot detect successive outliers. To compensate for this disadvantage, this value is normalized by the following equation if the previous equation cannot detect the anomaly:

$$z_{t+1} = \frac{x_{t+1} - mean}{standard\ deviation}$$

If z_{t+1} is greater than the upper bound or less than the lower bound, it will be classified as an outlier. Each time a new point x_{t+1} is added, the point x_{t-k} is removed from the list and the values of the arithmetic mean,

standard deviation, upper bound and lower bound are updated and the previous steps are repeated.

b. Forgetting factor iterative data capture anomaly detection (FFIDCAD)

The following algorithm is applied to the primary and secondary channels for each gas separately and for temperature and humidity. The hyper-ellipsoid is defined by the following equation:

$$ell_k(m_k, S_k^{-1}, t) = \{x\ in\ R^d \ \big|(x - m_k)^T S_k^{-1}(x - m_k) \leq t^2$$

Where m_k is a matrix containing the arithmetic mean of the features, x is a point, t^2 is the confidence space of the data distribution, S_k^{-1} is the inverse of the covariance of the data distribution, and it can also be known as the accuracy matrix.

In the first stage of this algorithm, the arithmetic mean of the first two points is calculated, and the size of the matrix S is 2x2, where it contains the sensor values of two consecutive points x_k, x_{k+1} for the primary and secondary channels. The matrix S is updated when a new point arrives with the following equation:

$$S_k^{-1} = \frac{kS_k^{-1}}{k-1}[I - \frac{(x_{k+1} - m_k)(x_{k+1} - m_k)^T S_k^{-1}}{\frac{k^2-1}{k} + (x_{k+1} - m_k)^T S_k^{-1}(x_{k+1} - m_k)}$$

The arithmetic mean is updated with the following equation:

$$m_k = \lambda\, m_{k-1} + (1 - \lambda)x$$

Where $\lambda \in (0, 1)$.

The point x_{k+1} is considered anomalous if the following inequality is true:

$$(x_{k+1} - m_k)^T S_k^{-1}(x_{k+1} - m_k) > bound$$

Where bound is calculated by the percent point function by choosing an appropriate value for p using the equation:

$$p = 1 - 10^{-i}$$

Where i = 16.

c. Temperature and humidity-based anomaly detection (THAD)

The researchers studied the relationship between the features and found that the values of NO2 and NO are related to temperature values, while the values of O3 are related to humidity values. Then, the values of NO2 and NO were divided into six groups according to temperature values, and the values of O3 were divided into five groups according to humidity values. Then, they calculated the arithmetic mean, standard deviation, upper limit and lower limit for each group separately using the IQR algorithm. The difference between this algorithm and the previous algorithms is that the values of the arithmetic mean, standard deviation, upper limit and lower limit are not updated when new values are received. Therefore, the quality of this algorithm depends on the quality of the training data, which covers a full year of measurements.

In the anomaly detection stage, a value is assigned to one of the groups, then it is normalized and compared with the upper and lower limits for the four gases. As for the temperature and humidity values, anomalies are detected by applying this algorithm to each data group.

2. Anomaly repairing

The researchers used the vector autoregressive model (VAR) for this stage, which is given by the following equation:

$$\gamma_t = \beta + \alpha_1 \gamma_{t-1} + \cdots + \alpha_p \gamma_{t-p} + \varepsilon_t$$

Where γ_t is a vector consisting of T sample and k variable, β is a vector kx1 that represents the intercept, ε_t is an error that follows the normal distribution. The VAR model is based on two assumptions:

1. That each variable in the time series is affected by the other variables, and to test this hypothesis, we need to apply Granger causality.
2. The time series must be stationary. To test this hypothesis, we need to apply the augmented Dickey-Fuller test.

To choose the variable p, the model is tested using the Akaike information criterion (AIC), which is given by the equation:

$$AIC = 2k - 2\ln(L)$$

k: the number of variables.
L: the maximum value of likelihood.

To apply this method, there must be at least 10 previous samples that do not contain empty or outlier values (or one non-empty sample in the previous five values). If the time series does not meet the stationarity condition, researchers perform a differentiation of it once or twice. If it does not meet the condition, the empty or outlier values will not be filled.

To evaluate the performance of this algorithm, the researchers used mean absolute percentage error (MAPE), which is given by the equation:

$$MAPE = 100 * \left|\frac{F - R}{R}\right|$$

Where F is the expected value and R is the true value, the arithmetic mean of the MAPE is calculated for each variable. If the percentage is less than 50%, the model prediction is considered correct. If it is greater than that, the prediction is ignored, and another round is conducted to correct for null or outlier values by making the null values equal to the values that precede them. Then the MAPE is calculated. If it is less than 50%, the prediction is considered correct.

3. Calibration model

An LSTM autoencoder model was used to convert the raw measurements into the concentrations of each gas in the air. No further details are given.

Dataset

To test the anomaly detection algorithms, the researchers collected sensor measurements over the course of about a year. They used an outlier generator called Agots to generate two types of outliers: extreme outliers and variance outliers.

In order to test the anomaly detection algorithms with correction algorithms, sensor measurements were collected over a period of approximately one year.

The data used is available at this link: https://drive.google.com/drive/folders/1LqZSVXA_2A1Hk_7fk9UwDOYEda-J6qvG

Tests and Results

In the anomaly detection stage only, the results of the three algorithms were compared individually and together with the LSTM model, which consists of six layers, the LSTM layer, the dropout layer, repeat vector layer, LSTM layer, dropout layer and time-distributed layer. The first three layers represent encoder, the last three layers represent the decoder. The researchers used the mean absolute error function and the Adam algorithm. The results were as follows:

		NO					NO_2					O_3				
		SWAD	FFIDCAD	THAD	MV	LSTM	SWAD	FFIDCAD	THAD	MV	LSTM	SWAD	FFIDCAD	THAD	MV	LSTM
Extreme	R	0.15	0.3	0.2	0.42	0	0.34	0.64	0.25	0.95	0	0.14	0.5	0.3	0.97	0
	P	1	0.34	1	0.33	0.68	1	0.83	1	0.81	0	1	0.71	1	0.97	0
	F1	0.26	0.32	0.34	0.37	0	0.5	0.72	0.4	0.87	0	0.24	0.59	0.46	0.97	0
Variance	R	0.25	0.4	0.3	0.3	0	0.35	0.98	0.4	0.65	0.38	0.23	1	0.55	0.5	0.53
	P	1	0.31	1	0.35	0.82	1	0.81	1	0.83	0.82	1	0.98	1	0.73	0.68
	F1	0.4	0.35	0.46	0.32	0	0.52	0.89	0.57	0.73	0.52	0.37	0.99	0.71	0.59	0.6

The MV algorithm outperformed most of the time in both cases.

The final evaluation process was done through three experiments:

The 1st experiment: It includes taking the arithmetic mean of the measurements over a 10-minute field without detecting or correcting the anomaly, then the calibration model was used.

The 2nd experiment: It includes detecting anomalies using the MV algorithm, then the arithmetic mean was calculated over a 10-minute field, then the calibration model was used.

The 3rd experiment: It is similar to the second experiment except that anomalies were corrected using the VAR model, then the calibration model was used.

The results were as follows:

		RMSE			ACCURACY		
Gas	Sensor	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
NO	4003	5.24	3.53	**3.18**	0.99	0.99	0.99
	4005	2.82	**2.34**	2.94	1	1	1
	4006	**2.53**	2.59	2.71	1	1	1
	4007	**2.74**	3.15	4.5	0.99	0.99	0.99
	4008	93.24	4.37	**2.62**	0.99	0.99	1
	4010	3.8	**2.48**	5.18	0.99	0.99	0.99
	4011	4.16	**3.92**	4.23	0.99	0.99	0.99
	4013	26.24	**2.2**	2.73	0.99	0.99	0.99
	4014	2.39	2.81	**2.05**	1	1	1
	M	**15.91**	**3.04**	**3.35**	**0.99**	**0.99**	**0.99**
NO_2	4003	8.25	15.07	**6.53**	0.97	0.98	0.98
	4005	**12.2**	12.44	12.44	0.95	0.95	0.95
	4006	9.55	**8.81**	9.09	0.97	0.97	0.97
	4007	10.76	9.83	11.89	0.95	0.95	0.95
	4008	347.03	9.38	**6.62**	0.96	0.96	0.97
	4010	7.78	7.58	**7.11**	0.98	0.98	0.98
	4011	64.54	8.89	**8.58**	0.98	0.98	0.98
	4013	8.01	8.2	**7.96**	0.98	0.98	0.98
	4014	10.05	9.33	**7.82**	0.96	0.97	0.96
	M	**53.13**	**9.95**	**8.67**	**0.97**	**0.97**	**0.97**
O_3	4003	20.03	18.51	**18.05**	0.53	0.56	**0.58**
	4005	19.9	**17.7**	18.78	0.65	**0.75**	0.73
	4006	19.41	**17.67**	17.87	0.61	**0.65**	**0.65**
	4007	14.62	**12.23**	13.41	0.8	0.82	**0.85**
	4008	**19.79**	22.49	22.66	**0.61**	0.57	0.57
	4013	25	**22.52**	22.68	0.48	0.5	**0.54**
	4014	222.53	86.27	**14.64**	**0.71**	0.52	0.66
	M	**48.75**	**28.20**	**18.30**	**0.63**	**0.62**	**0.65**

Where accuracy and root mean square error (RMSE) were used as comparison criteria. We can see the superiority of the results of the third experiment.

Limitations and downsides

These previous algorithms may not be able to detect inliers or local outliers. An example of this problem is that a sensor may give a "normal" value but it is wrong. It may be useful to study the relationships between gases to correct the wrong values. The researchers did not mention this drawback and it may be that such a problem never occurs or the probability of its occurrence is low, so developing a solution to this problem may not be appropriate.

References

[1] Pearson, R. K. (2006). The problem of disguised missing data. SIGKDD Explorations, 8(1), 83–92. https://doi.org/10.1145/1147234.1147247

[2] Rollo, F., Bachechi, C., & Po, L. (2023). Anomaly detection and repairing
for improving air quality monitoring. Sensors, 23(2), 640. https://doi.org/10.3390/s23020640