Detection of untrustworthy IoT measurements using expert knowledge of their joint distribution

Ilia Nouretdinov, Salaheddin Darwish, and Stephen Wolthusen

Information Security Group, RHUL, Egham, TW20 OEX, United Kingdom
{i.r.nouretdinov,salaheddin.darwish,stephen.wolthusen}@rhul.ac.uk
http://isg.rhul.ac.uk/

Abstract. The aim of this work is to discuss abnormality detection and explanation challenges motivated by Medical Internet of Things. First, any feature is a measurement taken by a sensor at a time moment, so abnormality detection also becomes a sequential process. Second, an anomaly detection process could not rely on having a large collection of data records, but instead there is a knowledge provided by the experts.

Keywords: anomaly explanation, untrustworthy data, Internet of Things

This work was initially motivated by some security challenges in Medical Internet of Things (MIoT). An individual instance (data record) in this context is presented as a sequence of measurements generated by multiple sensors. Abnormality of a record usually becomes a reason to produce an alert to doctors, reporting a suspected critical health state of the patients. However, the task is to separate a real health alarm from threats and vulnerabilities of the MIoT system. The principal question is which of the measurements (features) are less trustworthy than the others. The key assumption is that some knowledge of the joint feature distribution is available before having the measurements. The information about feature dependencies may be extracted from data analysis or obtain from experts. Involving experts in data analysis is very desirable. This is discussed e.g. in [5] where a Bayesian causal network for diagnostic is provided with elements of human feedback. Expert knowledge may also include some prior knowledge collected from earlier research on different data sets (e.g. connection between pulse pressure and coronary heart disease in [6]). Therefore, we assume that prior knowledge comes in the form of elements of probabilistic model. It is important to mention that we rely neither on collected historical data nor on regular quick feedback. The work [1] develops a feature-related anomaly explanation approach *providing user with information about the combination of dimensions (an attribute subset) in which an outlier shows the greatest deviation.* This might suit our needs, but the solutions in [1] require a sufficient quantity of instances to learn about normal and abnormal ones. In MIoT modelling, the features appear to follow a sequential form, the output has to be updated on each step. This has something in common with on-line machine learning [3], but we have to interpret new measurements as features, not as instances. Sequential feature explanation is also addressed in [2] but unlike our setting, the order of features is not fixed.

1 Data and prior knowledge

Let the data *record* for a patient has the form $D = (d_1, \dots, d_m)$ where d_j is j-th *feature (measurement)*, m is the overall number of measurements. In general, j-th feature is a measurement taken at time moment t_j from a sensor $s_j \in \{1, \dots, q\}$.

The prior information known from the experts can be of the following types:

1. Information about joint distribution of the features.
2. Information about exceptions: explainable deviations from typical behaviour of the system. It may also include the recommended reaction:
   - **Ignoring**: to continue without any change.
   - **Deleting**: the feature(s) from the record.
   - **Closing**: as a compromise, it may be used for training, but not considered as an abnormality.
   - **Correcting**: to eliminate the contribution of an external factor.
   - **Switching** to the new pattern, with deleting/closing of the prehistory.

Example. We propose that each sensor has its own stochastic schedule. The distribution of the time between the measurements of a Sensor is exponential with variance $\lambda = 1$ for Sensor 1, $\lambda = 0.5$ for Sensor 2. We assume that the data record is the sum of two stochastic components. The first 'proper' one is related to the measured values themselves generated by a natural multi-dimensional Gaussian distribution. The second 'noisy' component reflects an influence of Sensor 1 on Sensor 2. Let the joint distribution of measurements from Sensor 1 be Gaussian with mean 0, variance 1 and covariance e^{-t} where t time between the measurements. Similar parameters for Sensor 2 are 0, 1.5 and $1.5e^{-2t}$. The covariance between a measurement of Sensor 1 and a measurement of Sensor 2 is $-e^{-4t}$. If the last measurement of Sensor 1 was done in less than $t_1 = 0.5$ time then the measurement of Sensor 2 is enlarged by a random noisy component distributed uniformly on $[0, \lambda]$. We assume that $l_0 = 0.1 \leq \lambda \leq 0.5 = l_1$.

Two types of exceptions are included into the model:

1. An individual exception: The noisy influence of Sensor 1 on Sensor 2 may sometimes disappear (as if $\lambda = 0$).
2. A temporary shift: Assume that the time is measured in days, and the non-integer part of the time stamp is below 0.5 at night. Within the time intervals $(0, 0.5)$, $(1, 1.5)$ etc., λ may raise temporarily to its maximal value 0.5.

We use the following basic settings for simulation experiments:

1. Low noise: $\lambda = 0.1$ ended with a sensor fault at $t = 15$.
2. Medium noise: $\lambda = 0.3$ ended with a system fault at $t = 15$.
3. High noise: $\lambda = 0.5$ ended with a critical health state at $t = 15$.
4. Attack simulation (within-range negative shift) at a time point: $\lambda = 0.3$, changed to $\lambda = 0.1$ at $t = 15$.
5. Attack simulation (out-of-range negative shift) at the origin, $\lambda = 0.05$.
6. Attack simulation (out-of-range positive shift) at the origin, $\lambda = 0.6$.
7. Attack simulation (wave shift) at the origin, variable noise: $\lambda(t) = 0.3e^{sin(t)\sqrt{t}}$.

2 Testing for abnormalities

We consider a data record as anomalous if it is anomalous even for the best fitting distribution of the distributions that agree with the prior knowledge. Therefore, we split the task of anomaly detection into two stages: (2) estimating of the best fitting distribution $\hat{P}$ (in Bayesian or Maximum Likelihood sense); (2) testing the data on agreement with this distribution.

The first testing question is whether any abnormality is caused by the last measurement added to the system. The statistical test related to this particular feature can be based on the residual i.e. the difference between the true value of the last measurement and one expected from $\hat{P}$.

What we also need is some form of accumulation of abnormality reflected in the sequential features. Therefore, we apply the second type of testing based on the elements of machine learning and ranking: try to predict each measurement from $\mathcal{P}$ and the remaining measurements, and to calculate the *probabilistic residuals* i.e. p-values measuring how likely the *true* value of this feature looks according to the predictive model, and to apply *i.i.d.* testing (in assumed way of measurements we can expect them to be nearly *i.i.d.* in a normal situation).

There still may happen that none of $P \in \mathcal{P}$ fits the data at a satisfactory level, but this is not explainable by abnormality of the last feature or a group of them. Let us imagine that the knowledge is two-level: there exists a hard model $\mathcal{P}$ and an extended model $\overline{\mathcal{P}}$. In that case, it is possible to compare best fitting $\hat{P} \in \mathcal{P}$ with the best fitting $\hat{\overline{P}} \in \overline{\mathcal{P}}$. The alert is produced if the difference in fitting degree is essential.

Example (continued). Let $\hat{\lambda}$ be the maximum likelihood solution. The following cases may appear:

1. $\hat{\lambda} < 0$: we reduce this case to $\hat{\lambda} = 0$, and go the point 2.
2. $0 \leq \hat{\lambda} < l_0$: the most likely parameter value is out of range; this may be a possible reason for a special alert, if the ratio of likelihoods at $\hat{\lambda}$ and l_0 is above the pre-selected threshold ϕ; otherwise we just reduce $\hat{\lambda}$ to l_0.
3. $l_0 \leq \hat{\lambda} \leq l_1$: we make further steps of analysis in assumption of this value;
4. $l_1 < \hat{\lambda}$: the most likely parameter value is out of range; this may be a possible reason for a special alert, if the ratio of likelihoods at $\hat{\lambda}$ and l_1 is above the pre-selected threshold ϕ; otherwise we just reduce $\hat{\lambda}$ to l_1.

Let $D = (d_1, \ldots, d_m)$ be the observed vector of measurements at a step m (with 'deleted' ones). It is needed to calculate p-values:

$$p_i = \hat{P}\{\tilde{d} : \delta_i(\tilde{d}) \leq \delta_i(d_i) | d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_m\}.$$

If $p_m < \varepsilon'_1$ (strict alarm level) we detect an individual measurement error.

Then, for each sensor i, we consider the sequence $\left(p^i_1, \ldots, p^i_{m_i}\right)$ of measurements from the sensor i. Let $\tilde{p}^i_{(h)}$ $(h = 1, \ldots, m_i - 1)$ be the p-value produced by Mann-Whitney-Wilcoxon 'ranksum' test on $(p^i_1, \ldots, p^i_h)$ and $(p^i_{h+1}, \ldots, p^i_{m_i})$. The group error is reported if $\min_h \tilde{p}^i_{(h)} < \varepsilon_2$.

3 Explanation

The aim of explanation is to analyse the visible contradiction between the data and the model. In our example, we assume that abnormal health state or system fault is reflected as group error of more than one sensors, unlike a fault of one of the sensors.

Example (continued). To check type 1 exception, exclude the latest measurement from the 'affected' list (temporary set $z_k = 0$), re-run the process of anomaly detection and check whether the alert is reproduced. Type 2 can be checked similarly: if error disappears if λ is changed to 0.5 for the same night.

The items in the following list are defined based on order of priority: each of them is used only if none of the preceding ones is applicable.

- *Normal work or detected exception.*
  Action for type 1 exception: 'closing' the latest feature.
  For type 2 exception: 'closing' all 'affected' features from the same night.
- *Measurement mistake.* An individual error (strict alert).
  Action: 'deleting' the measurement.
- *Alarm A/B: health state or system fault.* A group error for both sensors.
  Action: 'deleting' the measurements after the earlier splitting point.
- *Alarm C: sensor fault.* A group error for only one of the sensors.
  Action: 'deleting' the measurements of this sensor after the splitting point.
- *Measurement mistake.* An individual error (moderate alert).
  Action: 'closing' the measurement.
- *Special alert: information bias.* A general shift.
  Action: no immediate actions, just marking for investigation.

4 Experiments and evaluation

Scenarios 1–7 follow Sec.1. The measurement mistakes and exceptions of types 1–2 are imputed at arbitrary moments (8, 18-19, 32, 39-41, 49, 54, 52). We are modelling mistakes leading to out-of-range or rare measurements with big absolute values: 1 for Sensor 1 and 2 for Sensor 2 as examples.

In Tab.1, we apply the methodology of untrustworthy measurement detection to the data records created in 7 scenarios. The star mark (*) in the table means the setting is selected for graphical representation in Fig. 1–3. For **measurements mistakes**, the evaluation criterion is the number of recognised vs. missed measurement mistakes. For **group errors**, we observe quickness of reaction: how many steps passed before an alert was produced. The change point is $t = 15$ (step 44) for scenarios 1–4, and the origin (step 1) in scenarios 5–7. In scenarios 1–4 we consider any alert as an evidence of reaction, while in 5–7 we are waiting for an alert of the proper type ('special alert'). Also, we have taken a note (in brackets) of the number of produced special alerts, and the amount of **false alerts** reported without any real causes. For true alerts, we check **accuracy of explanation**. For a measurement mistake and we calculate the number of examples with this (correct) explanation (+) vs. the others (-).

Sc.	$\phi = 10, \varepsilon_2 = 0.05$		1.Meas.error	2.Group error	3. False	4.Explanation
	ε_1	ε_1'	+/-	delay	alerts	accuracy
1	0.001	0.001	-10	-6	-8	N/A; gr+(S1)
	0.05	0.01	+4/-6	-4	-7	+3/-1; gr±(S2)
2 (*)	0.001	0.001	+4/-6	0	-11	+4; gr-(MM)
	0.05	0.01	+2/-8	-4	-13	+1/-1; gr-(S2)
3	0.001	0.001	+6/-4	0	-11	+4/-2; gr-(MM)
	0.05	0.01	+2/-8	0	-13	-1/-1; gr-(S1)
4 (*)	0.001	0.001	+4/-6	0	-11	+4; gr-(MM)
	0.05	0.01	+4/-6	-4	-13	+2/-2; gr-(S2)
5	0.001	0.001	-10	-11(2)	-8	N/A
	0.05	0.01	+4/-6	-11(1)	-10	+3/-1
6	0.001	0.001	+4/-6	-11(9)	-6	+1/-3
	0.05	0.01	+4/-6	-11(17)	-12	+1/-3
7	0.001	0.001	+2/-8	-11(3)	-9	+2
(*)	0.05	0.01	+4/-6	-11(6)	-9	+2/-2

Table 1. Effectiveness of the algorithms

An explanation of a group mistake may be correctly(+) or wrongly(-) assigned to one of the types: sensor fault; health state; global mistake. We consider it as partially right (±) if a sensor fault is recognised with a wrong sensor.

In our experiments, only a part of measurement mistakes is recognisable but this may be due to insufficient amount of data for analysis. Typically, some amount of collected features is needed for sensitivity of measurement mistake. On the other hand, elimination of suspicious measurement mistakes is useful for better detection of group errors. Group errors are recognised in some of scenarios/settings as individual measurement mistakes. This may hopefully be resolved by using extra 'strict' significance level for individual mistake. Group errors are likely to be recognised in this setting. However, detection of the exact cause appears to be harder; the easiest one for recognition is 'positive shift'.

5 Conclusion

In this work, we shed light on the problem of untrustworthy measurement detection motivated by MIoT. The proposed solution is based on its interpretation as a form of anomaly detection and explanation. We take into account specific challenges: lack of reliable historical data and feedback, working only with the general experts' knowledge and one actual record, sequential addition of features. We have validated the approach using a synthetic data sample that includes imputed scenarios of measurement mistakes, exceptions, faults and attacks. It appears that individual mistakes are hardly recognisable at first steps of the work, but improves with growth of the amount of collected measurements. The group errors are quickly recognisable by statistical analysis, but detection of the exact cause may be not so easy. The reaction of the system attacks (global errors) to intentional attacks is promising. The prior task for the future work

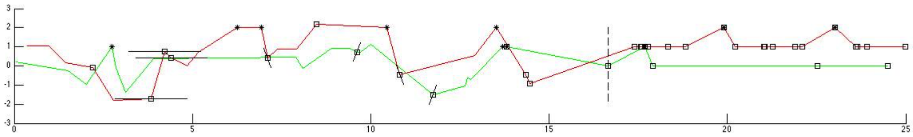

Fig. 1. Scenario 2: S1-2 fault

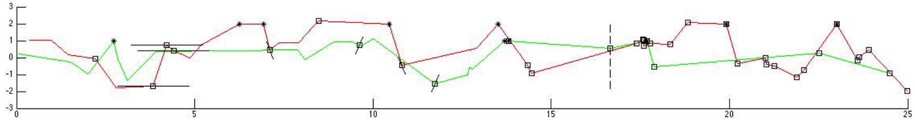

Fig. 2. Scenario 4: Late shift (decrease)

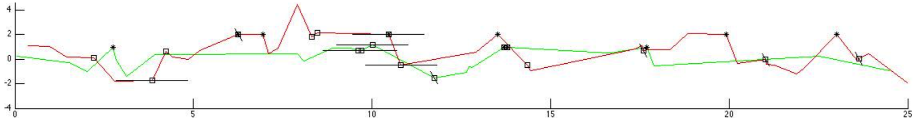

Fig. 3. Scenario 7: Early shift (wave)

The markers on these figures mean: a star is for an imputed measurement mistake; a square is for a produced alert (measurement mistake if no lines are attached); one oblique vertical line crossing the square is for sensor fault alert; two oblique vertical lines are for system fault or healthy state alert; a horizontal line is for a special alert.

may be developing some kind watermarking as in [4], with elements of active learning in investigation.

Acknowledgements This work was supported by Technology Integrated Health Management (TIHM) project awarded to the School of Mathematics and Information Security at Royal Holloway as part of an initiative by NHS England supported by InnovateUK, and by European Union grant 671555 ("ExCAPE").

References

1. B. Micenkova, R. T. Ng, X.-H. Dang, I. Assent, Explaining outliers by subspace separability. In Data Mining (ICDM), IEEE, pp. 518–527, 2013.
2. M. A. Siddiqui, A. Fern, Th. G. Dietterich, W.-K. Wong. Sequential Feature Explanations for Anomaly Detection. arXiv:1503.00038 [cs.AI], 2015.
3. L. Bottou. Online Algorithms and Stochastic Approximations. Cambridge University Press, 1998. ISBN 978-0-521-65263-6
4. Y. Mo, S. Weerakkody, B. Sinopoli. Physical Authentication of Control Systems. IEEE Control Systems Magazine, vol. 35, no. 1, pp. 93-109, 2015.
5. A. Zagorecki, P. Orzechowski, K. Holownia. A System for Automated General Medical Diagnosis using Bayesian Networks. MEDINFO, 2013. C.U. Lehmann et al. (Eds.) doi:10.3233/978-1-61499-289-9-461
6. S. S. Franklin, Sh. A. Khan, N. D. Wong, M. G. Larson, D. Levy. Is Pulse Pressure Useful in Predicting Risk for Coronary Heart Disease? The Framingham Heart Study. Circulation, July 27, 1999, Volume 100, Issue 4.