A performance study of anomaly detection using entropy method

A.A. Waskita∗†, H. Suhartanto†, L.T. Handoko§¶
∗Center for Technology and Safety of Nuclear Reactor, National Nuclear Energy Agency,
Kawasan Puspiptek Serpong, Tangerang 15310, Indonesia
Email : adhyaksa@batan.go.id
†Faculty of Computer Science, University of Indonesia,
Kampus UI Depok, Depok 16424, Indonesia
Email : heru@cs.ui.ac.id
§Group for Theoretical and Computational Physics, Research Center for Physics, Indonesian Institute of Sciences,
Kawasan Puspiptek Serpong, Tangerang 15310, Indonesia
Email: laksana.tri.handoko@lipi.go.id
¶Department of Physics, University of Indonesia,
Kampus UI Depok, Depok 16424, Indonesia
Email: handoko@fisika.ui.ac.id

Abstract—An experiment to study the entropy method for an anomaly detection system has been performed. The study has been conducted using real data generated from the distributed sensor networks at the Intel Berkeley Research Laboratory. The experimental results were compared with the elliptical method and has been analyzed in two dimensional data sets acquired from temperature and humidity sensors across 52 micro controllers. Using the binary classification to determine the upper and lower boundaries for each series of sensors, it has been shown that the entropy method are able to detect more number of out ranging sensor nodes than the elliptical methods. It can be argued that the better result was mainly due to the lack of elliptical approach which is requiring certain correlation between two sensor series, while in the entropy approach each sensor series is treated independently. This is very important in the current case where both sensor series are not correlated each other.

Index Terms—anomaly detection, elliptical method, entropy method

I. INTRODUCTION

Detecting anomaly, especially in a safety critical system is very important to mitigate any system failures in the near future [1]. In some systems, such failures could lead to the tremendous environmental disasters. Therefore, those systems are always equipped with robust monitoring system based on the either wireless or wired sensor network (WSN). The network should involve various types and ranges of sensors which transmit the acquired data to the central unit. In some cases, the sensors are embedded in the nearby cascade controller prior to the main unit relatively far away from the monitoring node in the field. In a more complex system, it could consist of several tiers from the main system till the end monitoring nodes.

Some examples of the systems with tremendous environmental influences are the so-called landslide early warning system (LEWS) involving micro-electromechanical system (MEMS) based sensors, fiber optic strain sensing and GPS tracking system to monitor the ground motion related to earthquakes or volcanic activities [2], [3]; the forest fire detection and monitoring system [4].

Some techniques to detect the anomalies have originally been developed for cyber security, in particular to mitigate the cyber attacks. For instance, the intrusion detection system (IDS) or intrusion prevention system (IPS) was worked out by [5]–[8]. In term of cyber security, those methods are complement to the signature approaches. It should be noted that the signature based IDS performs better in detecting the well known patterns of intrusion, while the anomaly based ones suits for the unknown patterns [9].

In contrast with the unpredictable "pattern" in cyber attacks which require the training procedure to define the "normal" patterns, in most cases any systems under monitoring through WSN have been constructed based on the pre-defined rules or model with certain parameter sets. These parameter sets consequently govern the allowed ranges of all sensor nodes within the system. However, the model is perfect under a presumption that all sensor and controlling nodes are working well without any failures. Concerning any potential failures during the operation, it is considerable to put the so-called early anomaly detection system (EADS) prior to the main processing unit. The EADS should lightweight, and not overburden the whole system. It is not necessarily accurate, but it should be able to provide, at least preliminary, information of any partial failures in advance. Actually in our previous works, the anomaly detection system has also been investigated using the statistical approaches. In the approaches a kind of interactions among single or cluster sensor nodes within the system has been modeled through the weighted "relationships" among the nodes [11], [12]. Unfortunately, the model is quite exhaustive and requires huge computing power.

From this point of view, some approaches based on the
"previous" pattern as adopted in the cyber security might not be appropriate. It would be better to set up more deterministic approaches like the entropy method [10]. The paper attempts to apply the entropy based method for the EADS in sensor network. The sensor nodes could be homogeneous or hybrid with various characteristics without assuming any interactions among them. The method only measures the level of irregularities in the system based on the predefined allowed ranges of each node following its specifications. The irregularities at certain degrees within a cluster or the whole system are interpreted as anomalies. As already argued in some other previous works [13] and references therein, the entropy based method requires light computing power and fast enough for anomaly detection. These natures are suitable for our purpose in the present case.

The paper is organized as follows. After this section, the entropy method is briefly explained in Sec. II. Sec. III deals with the experiment using the real data set, and followed with discussion on the comparison with another methods in the previous work by Rajasegarar et.al.. The paper is ended with the summary.

II. ENTROPY METHOD

Following the seminal work of Shannon [14], the entropy is defined as the level of irregularities occur, or in another word a measure of disorder in a system under consideration. It can be calculated using the master formula [10],

\[
H = - \sum_{k=1}^{K} p_k \log p_k ,
\]

where,

\[
p_k = \frac{a_k}{\sum_{j=1}^{K} a_j},
\]

is the elements of probability \( P = \{p_1, p_2, \cdots, p_K\} \) of \( D_k \). \( D_k \) is the elements of accumulated state set, \( S_A = \{D_1, D_2, \cdots, D_K\} \) with \( K \leq M \), and it is composed of all non-repetitive states in \( S \). On the other hand, \( a_k \) is the elements of \( A = \{a_1, a_2, \cdots, a_K\} \) which is representing the number of repetitions of \( D_k \).

Now, these procedure can be applied to investigate the real data and to perform a comparison with another methods.

III. EXPERIMENT

The experiment was conducted by taking the real data from the Intel Berkeley Research Laboratory (IBRL) data set during the acquisition period of March 1st, 2004 from 00:00 to 03:59. This period was chosen following the work by Rajasegarar et.al. [15], [16]. Only temperature and humidity sensors were taken into consideration over 54 MICA2DOT microprocessors, where each of them actually consisted of 4 sensor nodes: temperature, humidity, light and voltage.
As already mentioned in the preceding section, the entropy method itself does not require the correlation between temperature and humidity sensors. Therefore, one can determine independently the normal ranges for each sensor series using its manufacture specifications, and also took into account the fact that all data from the 37th node and a part of the 14th node were considered anomaly. Hence, the normal boundary condition for all of them is $x_b = 0$.

Further, each sensor series was divided into a smaller interval of time, namely 10 minutes, to have a set of data being calculated using Eqs. (1) and (2). This procedure was taken to enable the entropy analysis at the node level. Further calculation is illustrated in Fig. 1. Each acquisition in the 10 minutes interval should be evaluated against the normal boundary condition to determine whether the data acquisition is normal or not. The combination of normal-abnormal of the 10 minutes interval data acquisition establishes a cumulative value.

The following is the illustration for determining the entropy from the experiment in Fig. 1. For the first 10 minutes interval of the figure, there are 7 data acquisition for temperature and humidity parameters. At this interval, all temperature obtained exceeds the normal boundary. Based on the 10, the value of $A$ for all of them is 1 and construct the cumulative value array of $S_A$ as $\{1, 2, 3, 4, 5, 6, 7\}$. With different cumulative value elements of cumulative value array of $S_A$ as $\{1, 1, 1, 1, 1, 1, 1\}$. Then, each of the $S_A$ elements has the probability value of $\frac{1}{7}$. This leads to $H = 0.85$. On the other hand, all of the first humidity acquisition data from Fig. 1 are inside the boundary then the value of $A$ for all of them is 0 and construct the cumulative value array of $S_A$ as $\{0, 0, 0, 0, 0, 0, 0\}$. There is only one cumulative value ($A = \{7\}$) that produce the probability value into 1. This leads to $H = 0$.

The algorithm 1 describes the step-by-step of the procedure illustrated in Fig. 1.

Algorithm 1 Evaluate the interaction

Require: $x_1, \ldots, x_n$ (The data acquisitions from a sensor in a 10 minutes time interval, the number of $n$ as much as the data captured)

Require: $x_{b(1)}, x_{b(2)}$ (A normal boundary for a physical parameter, $x_{b(1)} =$ lower boundary, $x_{b(2)} =$ upper boundary,)

Require: $S_A$ (An array with $n$ elements of cumulative value)

Require: $A$ (An array listed a number of different cumulative values produced. If only one cumulative value exist, its value should equal to the number of data captured from the sensor in a certain 10 minutes time interval)

Require: $P$ (An array of the probability for each different cumulative value)

1: for $i = 1 \rightarrow n$ do
2: if $x_{b(1)} \leq x_i \leq x_{b(2)}$ then
3: $S_A(i) = 0 + S_A(i-1)$
4: else
5: $S_A(i) = 1 + S_A(i-1)$
6: end if
7: end for
8: construct the array of $A$ (Its element is the number of different cumulative value occurred ($K$))
9: calculate the array of $P$ (Its element is a probability of different cumulative value occurred in a certain 10 minutes time interval)
10: $H = -\sum_{k=1}^{K} p_k \log p_k$.

One should note that the 5th and 15th nodes were discarded since the data were missing in the data set. The calculated results for each sensor series are plotted in Fig. 2.

In the next section, the result is be compared with another methods done in some previous works.

IV. DISCUSSION AND SUMMARY

The result using entropy method in Fig. 2 shows the anomalies scattered over the area, while the normal data are on both horizontal and vertical axis. In particular, one should notice that all data coming from the 37th node (green triangles) are completely outranged, while only partial part of the 14th node (red triangles) are recognized as anomalies as expected. On the other hand, the entropy has successfully detected the data anomalies coming from various nodes.

One can compare the current result with the previous ones done by Rajasegarar et.al [16] using the elliptical method. The method is based on the elliptical curves to determine the "normal" region over the data as illustrated in Fig. 3. In the figure, the 90% CL (confidence level) curve is shown by the blue curve generated from the correlation between the data from temperature and humidity sensor nodes. Then, one can easily...
describe more curves with lower CLs to exclude the should-
be outlier data. However it is not trivial to fit the curve to
accomodate the allowed and anomalous regions.

More detailed study was conducted in the paper by Moshtaghi et.al. using the fractional elliptical method. It also dealt
with the same data set, but different period of time, to accomo-
date the inaccuracies in Rajasegarar et.al. [17]. The fractional
elliptical method is able to detect better the outlier data in
between the curves. Unfortunately we cannot provide the one
by one direct comparison due to the different period of data
set.

Finally, the present paper has shown the result of EADS
using the entropy method, and its comparison with the previous
results using the elliptical method. The comparison has been
conduct in two dimensional space based on the entropies
calculated from the data series of temperature and humidity
nodes. Each value of entropies have been calculated using the
data set of 10 minutes interval along the whole period under
consideration. It is argued that the entropy method is able to
detect the scattered anomalies across the space, regardless its
pattern in contrast with, for instance, the elliptical method.

ACKNOWLEDGMENTS

AAW thanks the Indonesian Ministry of Research and Tech-
nology for financial support, and the Group for Theoretical and
Computational Physics, Research Center for Physics LIPI for
warm hospitality during the work. LTH thanks to the Abdus
Salam ICTP for hospitality when the initial part of this work
was done. LTH is funded by Riset Unggulan LIPI in fiscal year
2016 under Contract no. 11.04/SK/KPPI/II/2016.

REFERENCES

[1] T. Ahram, W. Karwowski, D. Schmorrow, R. L. Boring, K. D.
Thomas, T. A. Ulrich, and R. T. Lew, “6th international conference
on applied human factors and ergonomics (ahfe 2015) and the
affiliated conferences, ahfe 2015 computerized operator support
systems to aid decision making in nuclear power plants,” Procedia
Manufacturing, vol. 3, pp. 5261 – 5268, 2015. [Online]. Available:
http://www.sciencedirect.com/science/article/pii/S23519789150006058.

[2] D. Hanto, B. Widiyatmoko, B. Hermanto, P. Puranto, and L. T. Handoko,
“Real-time inclinometer using accelerometer MEMS,” in Proceeding
of the International Conference on Physics and Its Applications for
Environmentally Friendly Technology and Disaster Management, 2010.

[3] R. W. R. Tu, T. R. W. M. Ge, M. Ramatschi, C. Milkeret, D. Bindi,
and T. Dahm, “Cost-effective monitoring of ground motion related to
earthquakes, landslides, or volcanic activity by joint use of a single-
frequency GPS and a MEMS accelerometer,” Geophysical Research
Letters, vol. 40, pp. 3825–3829, 2013.

[4] Y. E. Aslan, I. Korpeoglu, and Ø. Ulusoy, “A framework for use of
wireless sensor networks in forest fire detection and monitoring,”
Computers, Environment and Urban Systems, vol. 36, pp. 614–625, 2012.

[5] M. Moshtaghi and M. Suresh, “Processing Massive Data Streams to Achieve
Anomaly Intrusion Prevention,” in Fourth International Conference on
Computational Intelligence and Communication Networks, 2012, pp.
948–952.

[6] A. G. Fragkiadakis, V. A. Siris, N. E. Petroulakis, and A. P. Traganitis,
“Anomaly-based intrusion detection of jamming attacks, local versus col-
laborative detection,” Wireless Communications and Mobile Computing,
vol. 15, no. 2, pp. 276–294, 2015.

[7] S.-W. Lin, K.-C. Ying, C.-Y. Lee, and Z.-J. Lee, “An intelligent algorithm
with feature selection and decision rules applied to anomaly intrusion
detection,” Applied Soft Computing, vol. 12, no. 10, pp. 3285–3290, 2012.

[8] W. Xiong, H. Hu, N. Xiong, L. T. Yang, W.-C. Peng, X. Wang, and Y. Yu,
“Anomaly secure detection methods by analyzing dynamic characteristics
of the network traffic in cloud communications,” Information Sciences,
vol. 258, pp. 403–415, 2014.

[9] S. Elhag, A. Fernández, A. Bawakid, S. Alshmromani, and F. Herrera,
“On the combination of genetic fuzzy systems and pairwise learning for
improving detection rates on Intrusion Detection Systems,” Expert
Systems with Applications, vol. 42, no. 1, pp. 193–202, 2015.

[10] A. A. Waskita, H. Subartanto, and L. T. Handoko, “Entropy based method
for early anomaly detection in sensor network,” submitted to Information
Sciences, 2016.

[11] A. A. Waskita, H. Subartanto, Z. Akbar, and L. T. Handoko, “Exhaustive
search-based model for hybrid sensor network,” in 4th International
Conference on Intelligent and Advanced Systems, 2012, pp. 557–561.

[12] A. A. Waskita, H. Subartanto, P. D. Persadha, and L. T. Handoko, “A
simple statistical analysis approach for intrusion detection system,” in
Systems, Process Control (ICSPC), 2013 IEEE Conference on, Dec 2013,
pp. 193–197.

[13] Z. Jian-Qi, F. Feng, Y. Ke-xin, and L. Yan-Heng, “Dynamic entropy based
DoS attack detection method,” Computers & Electrical Engineering,
vol. 39, no. 7, pp. 2243–2251, 2013.

[14] C. E. Shannon, “A Mathematical Theory of Communication,” ACM
SIGMOBILE Mobile Computing and Communications Review, vol. 5,
no. 1, pp. 3–55, Jan. 2001.

[15] S. Rajasegarar, J. C. Bezdek, C. Leckie, and M. Palaniswami, “Analysis
of anomalies in ibrl data from a wireless sensor network deployment,”
in Sensor Technologies and Applications, 2007. SensorComm 2007.
International Conference on, Oct 2007, pp. 158–163.

[16] ——, “Elliptical anomalies in wireless sensor networks,” ACM Trans.
Sen. Netw., vol. 6, no. 1, pp. 7:1–7:28, Jan. 2010. [Online]. Available:
http://doi.acm.org/10.1145/1653760.1653767.

[17] M. Moshtaghi, S. Rajasegarar, C. Leckie, and S. Karunasekera, “Anomaly
detection by clustering ellipsoids in wireless sensor networks,” in Intel-
ligent Sensors, Sensor Networks and Information Processing (ISSNIP),
2009 5th International Conference on, Dec 2009, pp. 331–336.