Nuclear Engineering and Technology 54 (2022) 3996–4004

Contents lists available at ScienceDirect

Nuclear Engineering and Technology

journal homepage: www.elsevier.com/locate/net

Original Article

ADA: Advanced data analytics methods for abnormal frequent episodes in the baseline data of ISD

Biswajit Biswal [a, *], Andrew Duncan [b], Zaijing Sun [c]

[a] Department of Computer Science and Mathematics, South Carolina State University, United States
[b] Material Sciences and Technology, Savannah River National Laboratory, United States
[c] Health Physics and Diagnostic Sciences, University of Nevada, Las Vegas, United States

ARTICLE INFO

Article history:
Received 7 February 2022
Received in revised form
17 June 2022
Accepted 5 July 2022
Available online 8 July 2022

Keywords:
Advanced data analytics (ADA)
Abnormal frequent episode
Episode mining
ISD Sensor network testbed

ABSTRACT

The data collected by the In-Situ Decommissioning (ISD) sensors are time-specific, age-specific, and developmental stage-specific. Research has been done on the stream data collected by ISD testbed in the recent few years to seek both frequent episodes and abnormal frequent episodes. Frequent episodes in the data stream have confirmed the daily cycle of the sensor responses and established sequences of different types of sensors, which was verified by the experimental setup of the ISD Sensor Network Test Bed. However, the discovery of abnormal frequent episodes remained a challenge because these abnormal frequent episodes are very small signals and may be buried in the background noise of voltage and current changes. In this work, we proposed Advanced Data Analytics (ADA) methods that are applied to the baseline data to identify frequent episodes and extended our approach by adding more features extracted from the baseline data to discover abnormal frequent episodes, which may lead to the early indicators of ISD system failures. In the study, we have evaluated our approach using the baseline data, and the performance evaluation results show that our approach is able to discover frequent episodes as well as abnormal frequent episodes conveniently.

1. Introduction

A complete data analytics system gathers data first and then find information from the data and display the knowledge to the user. These systems involve many operations such as gathering, selection, preprocessing, transformation, data mining, evaluation, and interpretation [1,2]. However, the data analysis focused on this paper is responsible for finding the hidden patterns/rules/information from the data. Most data analytics research employ either statistical or machine learning algorithms—clustering, classification, association rules, and sequential patterns—to analyze the data and find the hidden information from the raw data [3]. Thus, one can modify these algorithms to enhance the performance of the data analysis.

Nowadays, the data that need to be analyzed are large, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, erroneous, and streaming in nature. These unique features of big data always challenge the statistical and data analysis approaches [4]. According to research results [5,6], and [7], most data analysis methods have issues of non-scalability, non-dynamic, and uniform data structure for big data. Because the traditional data analysis methods are not designed for large-scale, complex, and streaming data, redesigning and changing the data analysis methods are the new trends for big data analysis.

In the last few years, researchers at Savannah River National Laboratory (SRNL) have established an In-Situ Decommissioning (ISD) Sensor Network Test Bed, a unique, small scale, and configurable environment, for the assessment of prospective sensors on actual ISD system material, at a minimal cost [8,9]. During the years of 2011–2014, a large data set was collected in the process of testing and collecting baseline data [10]. All these data are time-specific, age-specific, and developmental stage-specific. This baseline data is quite big and ideal for incremental data analytics to validate ISD system performance and predict possible system failures (or future accidents) by detecting abnormal patterns (or frequent episodes) in the presence of background noise.

Frequent Episode Mining (FEM) is a widely accepted framework for discovering hidden patterns from time-series data, sequence data, and online stream data [11]. With the fast-growing data from

* Corresponding author.
E-mail address: bbiswaji@scsu.edu (B. Biswal).

https://doi.org/10.1016/j.net.2022.07.006

heterogenous sensors, in time-critical applications many episodes may become obsolete while new useful episodes keep emerging. So, there is always need of advanced, fast, and automatic solutions to discover all the hidden patterns, latest frequent episodes and abnormal frequent episodes from fast-growing data.

The traditional statistical and data analysis methods aren't ideal for large-scale, complex, and streaming data. Redesigning and changing the data analysis methods are the new trends in advanced data analysis. This paper introduces a new approach, advanced data analytics system, for real-time analysis of the baseline data while at-rest as well as in-motion by showing intermediate results as soon as they become available. This should allow the data analyst to take decisions in real-time. In this paper, we focused on data analytics, feature engineering, and machine learning methods of the advanced analytics engine. We introduce three new features in feature engineering: mean, median, and standard deviation to deal with background noise and improve the machine learning model accuracy. The machine learning models (i.e., baseline models) such as Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), etc., are used as advanced analytics to identify normal and abnormal frequent episodes.

The rest of the paper is structured as follows: Section II overviews the prior work on the concept of advanced analytics. Section III describes the proposed state-of-the-art advanced data analytics engine to achieve interactive analysis for baseline Data. Section IV discussed the results. Finally, Section V concludes the paper.

2. Prior work

In-Situ Decommissioning (ISD) Sensor Network Test Bed collected the baseline raw data during the year 2011–2014 [10]. It has 273,078 records, including battery information, strain information, temperature information from the 4 thermocouples epoxied inside the two concrete blocks, and tiltmeter data. The units of these data are Volt, Celsius, angle degrees, μstrains, etc. All the records were time-stamped in the format of "mm:dd:yyyy hh:mm" and collected in 5 min intervals. Thus, most of the data points were continuous except when the system was turned off for troubleshooting purposes.

Many real world applications were benefited from the temporal data mining (TDM) algorithm [12–14]. TDM is applied to discover the frequent patterns (or temporal episodes) [15], and these patterns are referred to as episodes that are ordered collections of events (i.e., abnormal sensor response). The work proposed by Sun et al. [10] is based on the temporal data mining (TDM) framework, which is a combination of the frequent episode discovery model and Hidden Markov Models (HMM). For instance, (RA → RB → RC) represents an episode where the events occurred in a timely order, i.e., first RA occurred, then RB, and then RC. According to the authors (Sun et al., 2018), the TDM method discovered some frequent episodes, but it did not "dig out" many abnormal frequent episodes. Also, it was not clear how often an abnormal frequent episode will lead to a certain significant "incident" in the ISD system. These confusions may be due to: (1) rare existence of the abnormal frequent episodes under normal conditions. It may be possible that the baseline data did not record too many points in sudden changes of system. (2) the fact that these abnormal frequent episodes are very small signals, and they may be buried in the background noise of voltage and current changes.

Xiang et al. [16] proposed a special-purpose algorithm to improve the local mining performance on Frequent Episode Mining (FEM) for “big” event sequences. The events were arranged in a predefined hierarchy i.e. the events from different levels of the hierarchy were partitioned based on event-centered and hierarchy-aware partitioning strategy before feed to local processes. Philippe et al. [17] worked on high utility episode mining that consists of high importance (e.g high profit) episodes in a sequence of events with quantities and weights. Their proposed algorithm named HUE-Span finds all patterns by taking into account all timestamps of minimal occurrences for utility calculations.

Maryam et al. [18] investigated a prediction model based on episode mining with the capability of online learning. Their proposed model considered the correlation between different resources and extracts behavioural patterns of applications independently of the fixed pattern length explicitly and their experimental results showed that the proposed model adapts to the behavioural changes of the application and learns the new behavioural patterns rapidly. Another online frequent episode mining method was proposed by Tao et al. [19]. They studied multiple continuous, unbounded and time-varying online data streams which were combined into a global data lattice based on sequence features. Then used the frequent episode tree to detect the expanding online serial episodes and parallel episodes from the data lattice and finally merge mixed episodes into existing serial and parallel frequent episodes.

All the above mentioned prior works were focused to determine frequent episodes and very specific to special applications in their domain. The limitation of most of the prior works were experimented with synthetic data and few were experimented with proprietary application specific dataset where there was no evidence of important information buried under the background noise of voltage and current changes. So, we revisited our prior work to continue further investigation.

In our prior work (Biswal et al. [20]), we used three statistical analytics methods: Linear Regression, Correlation, and Clustering, to determine the frequent episodes and abnormal frequent episodes. However, these methods failed to discover many abnormal frequent episodes, so we conducted further analysis called "peak" analysis [21] and with analysis, we were able to discover many abnormal frequent episodes. To determine if there is any pattern of abnormal frequent episodes (i.e., how often an abnormal frequent episode occurs), we looked at the distribution of the abnormal frequent episodes in the entire dataset. Unfortunately, the distribution didn't show any clear pattern of abnormal frequent episodes. So, we can't determine that the occurrence of abnormal frequent episodes will lead to a certain significant "incident" in the ISD system. The occurrence of abnormal frequent episodes may be due to weather changes such as rain or cloudy conditions or the occasional movement of sensors/equipment by technicians. Another possibility is that the abnormal frequent episodes are very small signals, and these may be buried in the background noise of voltage and current changes.

Compared to prior works, the approach introduced in this paper aims to deal with more frequent episodes discovery in the presence of background noise to detect the failure in the ISD system. We propose to build an Advanced Analytics Engine that will implement data analytics, feature engineering, and machine learning to discover frequent episodes in the baseline data.

3. Advanced analytics engine

We propose an advanced analytics architecture based on three main components (see Fig. 1): Data Layer, Advanced Analytics Engine, and Visualization Layer. The raw data are collected from the ISD Test Bed sensors. The data is time-series data (i.e., collected every 5 min interval, 24 by 7, and from years 2010–2014) and also includes outliers or noise due to the malfunctioning of many sensors. The raw form of data includes 72 columns and almost 600 MB (megabytes). Then, data preprocessing is performed to remove the noisy data caused by sensor malfunctioning. Below we discussed the steps to remove data outliers.

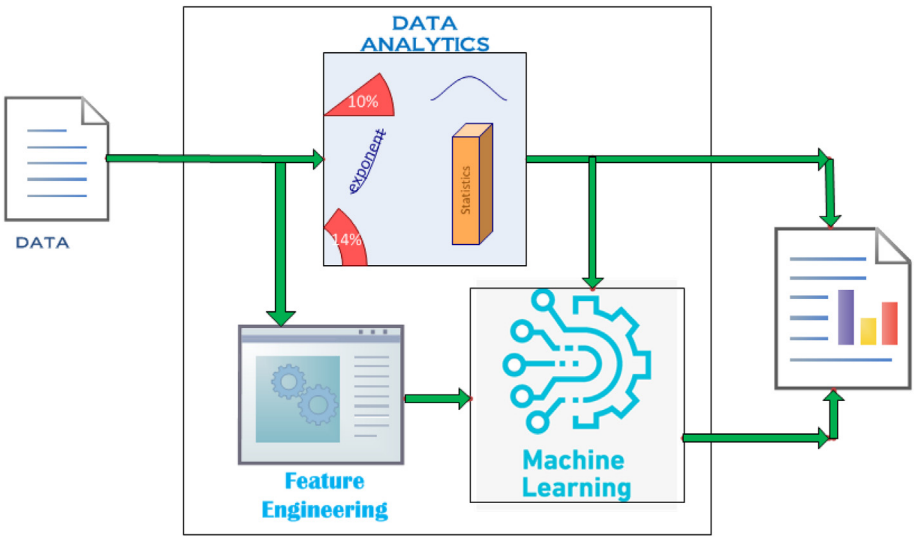

Fig. 1. Advanced analytics engine.

The following steps are carried out for the preprocessing of our data:

a) Load data to RStudio.
b) Look at the structure of the data to see if the data is of the correct type.
c) Looking at the result, everything is string. So, we need all as numeric types except TIMESTAMP.
d) Columns with −99999 or −88888 are outliers or noise. Determine which columns have outliers.
e) Replace columns with −88888 or −99999 with NA.
f) Delete columns with NA values.
g) Then, add three new columns (Year, Month, and WeekDay) to the clean data for the data analysis phase.

The Advanced Analytics Engine has three components: Data Analytics, Feature Engineering, and Machine Learning. In addition, the analytics engine can process data in two modes: Statistical Analytics and Advanced Analytics.

Statistical analytics are applied to identify frequent episodes in the data, which may lead to the early indicators of ISD system failures. Three statistical analytic methods: Linear Regression, Correlation, and Clustering are applied to the data, and their results are discussed in the result section. We choose these statistical methods because we would like see how well temperature sensors THT1, THT2, and THT4 follow the sensor THT3 values. Linear regression predicts a continuous value of an output variable (e.g. THT1, THT2, or THT4) as a linear function of input variable THT3. While Correlation measures the degree to which the input and output variables increase or decrease together (e.g. THT1 vs THT3, THT2 vs THT3, and THT4 vs THT3). Finally, Clustering analysis groups variables together based on the similarity of data points and data patterns.

Advanced Analytics mode, which includes feature engineering and machine learning. Advanced analytics are applied to discover abnormal frequent episodes in the baseline data, leading to the early indicators of ISD system failures. Feature engineering and machine learning methods aim to deal with more abnormal frequent episodes discovery in the presence of background noise. In the feature engineering method, we introduce three new features, mean, median, and standard deviation [22] to deal with the presence of background noise and to enhance the machine learning accuracy of frequent episodes discovery. Popular machine learning models (i.e., baseline models) such as Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), etc., are used to identify normal and abnormal frequent episodes from the heterogenous sensor data.

4. Results

4.1. Data visualization and data analysis

It is always ideal for visualizing data before we perform data analysis. Details on the heterogeneous sensor dataset can be found on the referenced paper Biswal et al. [20]. Fig. 2 shows the plot of the temperature sensors. First, we produce the frequent episodes for the temperature sensors in the order THT3→THT4→THT2→THT1 as reported by Sun et al. [10]. THT1 and THT2, correspond to temperature sensors on the front side (facing west) of the first block and trend each other. THT3 is the control sensor attached to the frame and is directly exposed to the sun and unshielded from the daily weather conditions. This sensor appears noisy because it is faster to respond to temperature fluctuations relative to the sensors insulated by the cement and epoxy. The THT4, is the temperature sensor facing east along with the THT3 sensor and grouted into 75–100 mm inside a 100-mm diameter hole compared to THT1 and THT2, which are cemented and epoxyed about 50 mm in holes that are approximately 75-mm deep and 25-mm wide. THT4 is supposed to respond last when compared to THT2 and THT1 but, THT4 sees a temperature increase first, which may be due to void spaces introduced during sensor placement.

In our preliminary work, we used three statistical analytic methods: Linear Regression, Correlation, and Clustering to determine the frequent episodes. In Linear Regression (LR) analysis THT3 is the control sensor, so we consider THT3 as the independent variable while the other three sensors as dependent variables. Code below provided the Linear Regression analysis of temperature sensors for a 6-month period of data from the year 2011. In this analysis, THT3 is considered as the reference sensor because it is facing east to the sun.

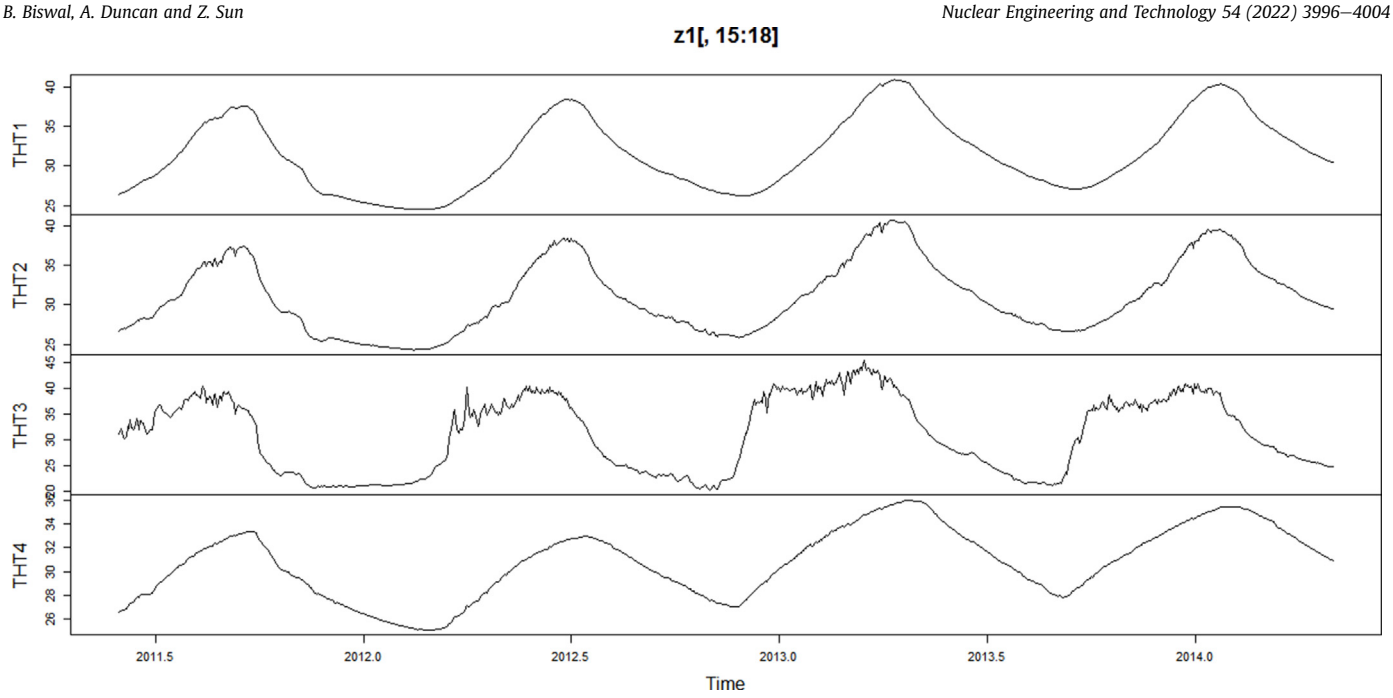

Fig. 2. Different temperature sensor responses.

```
> results <- lm(z1$THT4~z1$THT3,z1)
> summary(results)
Call:
lm(formula = z1$THT4 ~ z1$THT3, data = z1)
Residuals:
Min 1Q Median 3Q Max
-11.9155 -1.4793 0.4096 2.2281 5.6720
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.237564 0.065233 64.96 <2e-16 ***
z1$THT3 0.557129 0.004715 118.16 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.063 on 8926 degrees of freedom
Multiple R-squared: 0.61,Adjusted R-squared: 0.61
F-statistic: 1.396e+04 on 1 and 8926 DF, p-value: < 2.2e-16
```

Linear regression R code and analysis results.

From the above analysis results, the p-value is less than 0.05 which means with a 95% confidence level, the NULL hypothesis is rejected.

- Null hypothesis - > There exists no relationship between THT1 and THT3
- Alternate Hypothesis - > There exists a relationship between THT1 and THT3

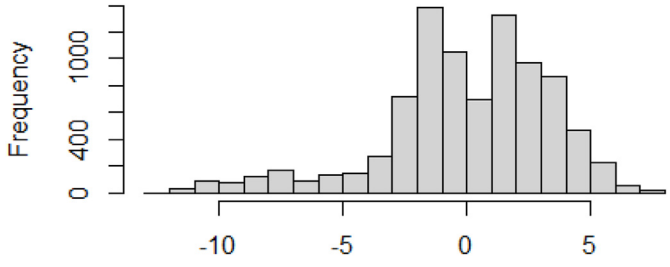

Fig. 3. Residual plot for THT4 and THT3.

However, the R-squared value is small compared to the maximum value of the R-squared value of 1.0, which means that the linear regression model cannot explain the relationship better (i.e., if the R-squared value approaches to 1, then the model explains the relationship better). So, we then looked at the residual plot between THT4 and THT3 as shown in Fig. 3, it should have proved the normality assumption (i.e., fit to bell curve), but in our case, it didn't fit the bell curve shape. Also, we have looked at other residual plots (THT2 vs. THT3), (THT1 vs. THT3), but none fit the bell curve. Looking at our analysis results, one can notice that the most frequent episodes are THT3→THT4→THT2→THT1. More case studies can be found in Biswal et al. [20].

To confirm the frequent episodes, we performed the Correlation method as the second analysis. Fig. 4 shows the correlation between THT1, THT2, THT3, and THT4 for the year 2011. Based on the coefficient values, one can see that the temperature variation sequence THT3→THT4→THT2→THT1. Furthermore, we performed clustering analysis to confirm frequent episodes. The clustering method "pvclust" is an R package for assessing the uncertainty in hierarchical cluster analysis [23]. For each cluster in hierarchical clustering, quantities called p-values are calculated via multiscale bootstrap resampling. The P-value of a cluster is a value between 0 and 1, which

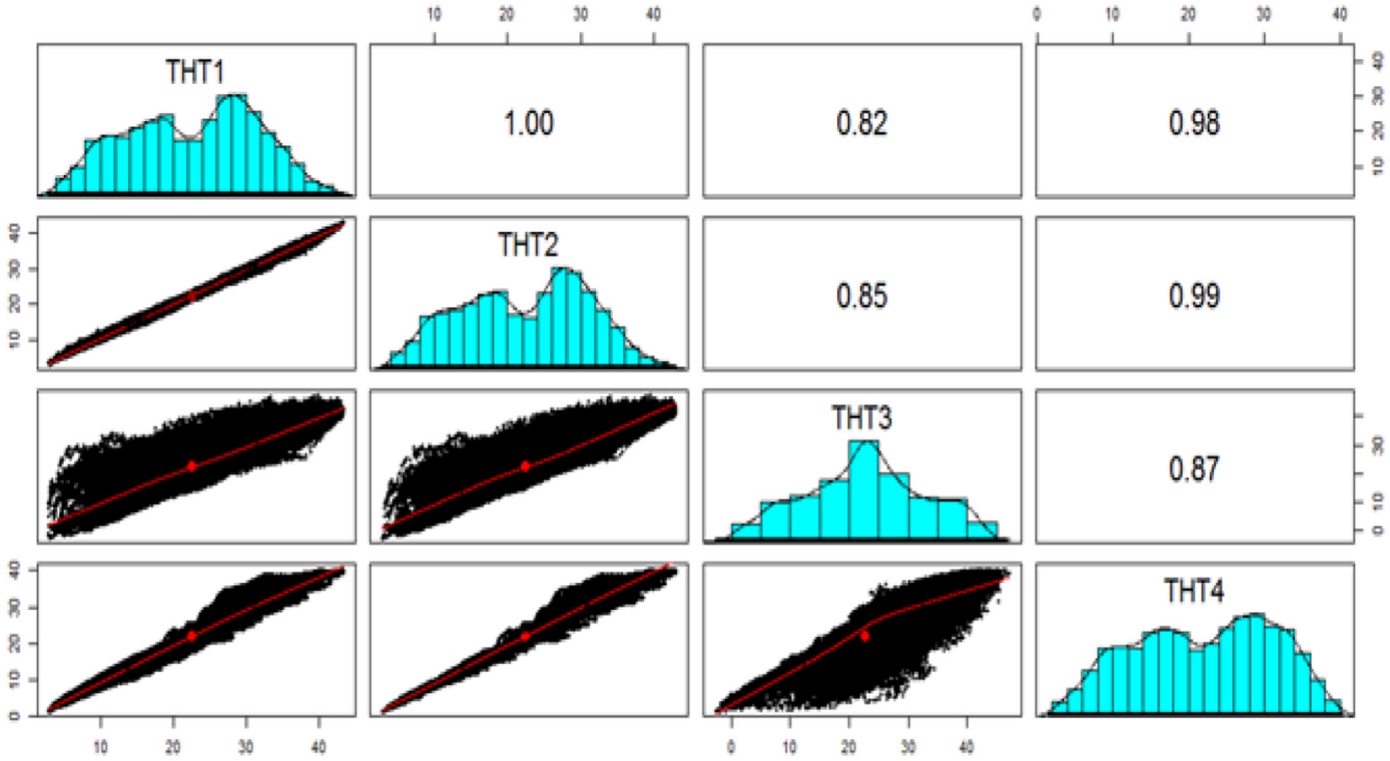

Fig. 4. Correlation Coefficient, Histogram plot, and data point distributions.

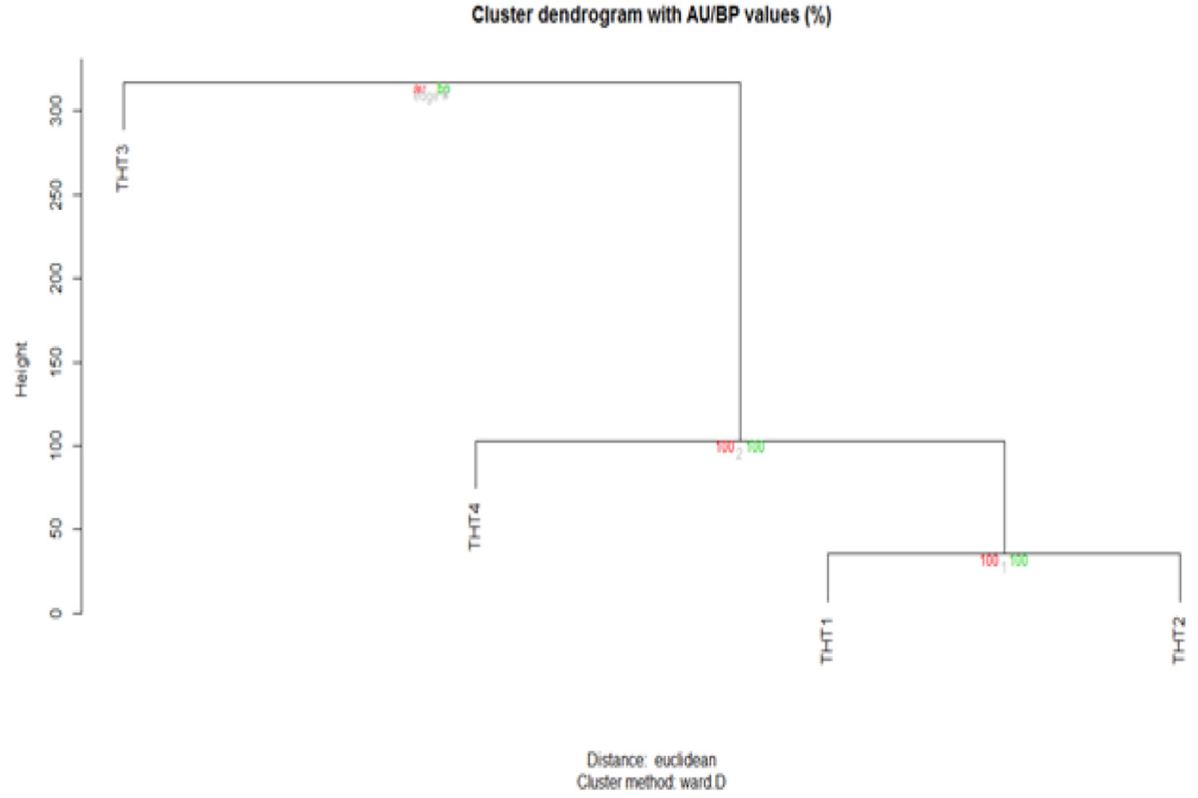

Fig. 5. Clustering between THT1, THT2, THT3 and THT4.

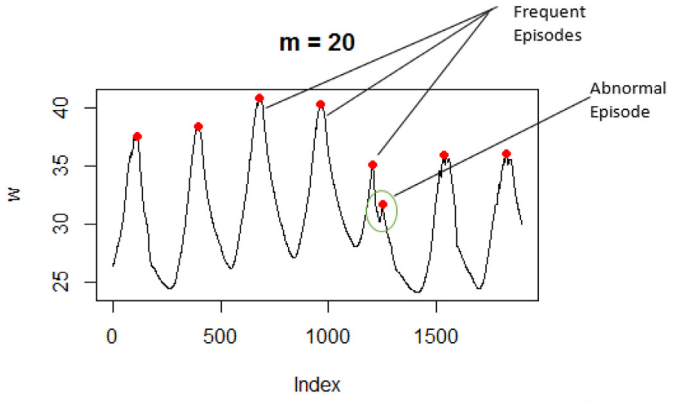

Fig. 6. A week data plot of June 2011.

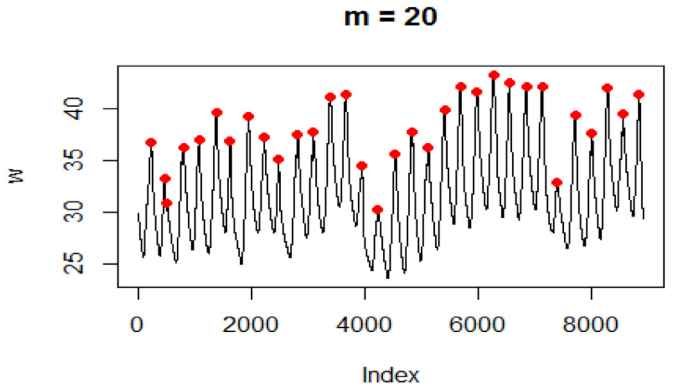

Fig. 7. A month data plot of July 2011.

indicates how strong the cluster is supported by data. The clustering function "pvclust" provides two types of p-values: AU (Approximately Unbiased) p-value and BP (Bootstrap Probability) value. AU p-value, which is computed by multiscale bootstrap resampling, is a better approximation to unbiased p-value than BP value computed by normal bootstrap resampling. Fig. 5 shows the clustering between THT1, THT2, THT3, and THT4. Thus, for a cluster with AU p-value > 0.95, the hypothesis that "the cluster does not exist" is rejected with a significance level of 0.05. All the above analytics methods that are used for the discovery of frequent episodes has demonstrated the existence of correlation among the sensors THT1, THT2, THT3, and THT4. However, these methods failed to discover many abnormal frequency episodes. So, we conducted further analysis called "peak" analysis using the local maxima method. This analysis is performed on a week data, and a month data to find out any abnormal frequent episodes, as shown in Fig. 6, and 7 respectively. The sensitivity of the peak detection procedure can be adjusted by a control parameter' m', larger m value results fewer peaks and smaller m value results more peaks. The detected number of abnormal frequent episode statistics are 19, 52, 56, and 3 for years 2011, 2012, 2013, and 2014 respectively using peak analysis. Only 3 abnormal frequent episodes are in the year 2104 because we have only less than a month of data (only 27 days of data). More case studies can be found in Biswal et al. [20].

To find out if there exists any pattern of abnormal frequent episodes (i.e., how often an abnormal frequent episode occurs), we looked at the distribution of the abnormal frequent episodes that are occurred in the entire dataset. Fig. 8 shows the distribution of abnormal frequent episodes for each month starting from the year 2011–2014. One can see that the distribution is random which

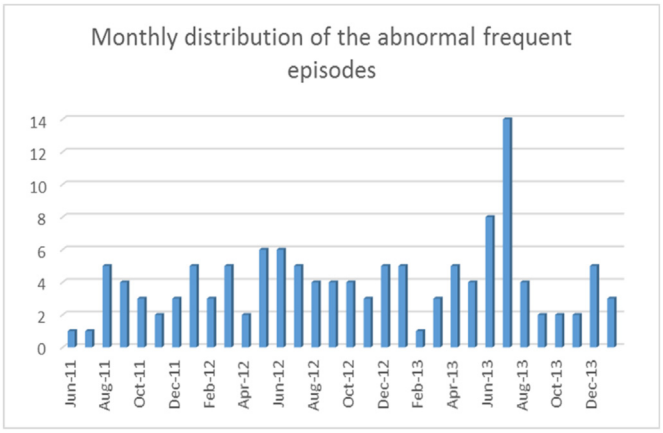

Fig. 8. Distribution of AFE (Monthly from 2011 to 2014).

means that there exists no clear pattern of abnormal frequent episodes. So, we 'can't be able to determine that the occurrence of abnormal frequent episodes will lead to a certain significant "incident" in the ISD system. The occurrence of abnormal frequent episodes may be either due to weather changes such as rain or cloudy conditions, or due to occasional movement of sensors/equipment by technicians. Another possibility is that the abnormal frequent episodes are very small signals, and these may be buried in the background noise of voltage and current changes. So, we'll be using the advanced analytics method using machine learning and feature engineering to discover any buried abnormal frequent episodes in background noise.

4.2. Feature engineering

SRNL site was identified by DOE-EM [24] to implement ISD Sensor Network Test Bed to provide an understanding of signal responses from the concrete blocks of the P-Reactor facility. A remote sensor monitoring system was deployed [25] to study the aging structures of concrete blocks, understand the likely changes over time in and around the ISD facility. The collected sensor data may have background noise voltage and current changes, which would hide certain significant "incidents" in the ISD system. So, we considered three new features such as mean, median, and standard

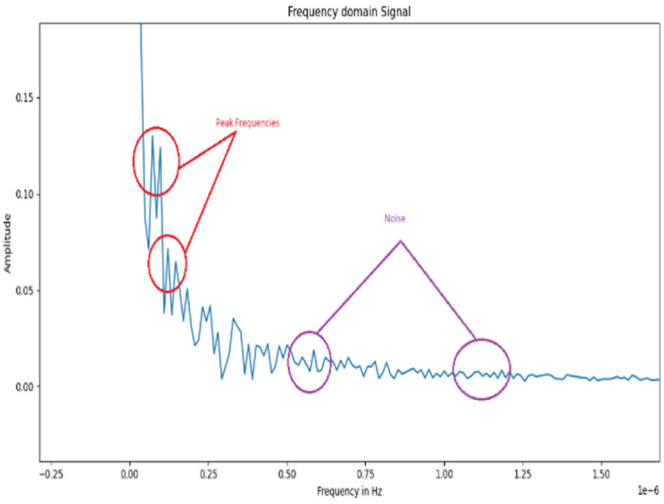

Fig. 9. FFT analysis.

deviation to filter out the noise and identify patterns. Our data is time-series data; the mean feature is calculated through the exponential moving average method [26] with a sliding window size of 3. We kept a smaller window size of 3 because of the nature of our data, and a smaller window size has increased sensitivity to changes. But, again, a larger window size would have suppressed some of the potential peaks buried under the small background noises. Thus, a smaller window size of 3 is preferred for our data. Similarly, we also computed moving median and standard deviation with a sliding window size of 3.

Through peak analysis we are able to discover many abnormal frequent episodes at random but it didn't show any occurrence of patterns. So, in this work we deployed machine learning models to discover more abnormal frequent episodes. We added a fourth feature, "fft-freq," to our dataset so that the machine learning model can have access to the most important frequency signal features. The "Date Time" column is converted to sine and cosine signals and processed through FFT (Fast Fourier Transform) [27] to calculate the "fft-freq" feature.

4.3. Presence of noise analysis

Noise analysis is mainly focused on dealing with background noise in data and discovering more abnormal frequent episodes. So, we performed a Fourier transform analysis to see the presence of noise [28] in the data. Fig. 9 shows the FFT analysis plot for THT3, and we can see there are many small-amplitude noise signals along with some distinctive "peak" frequency signals. The presence of noise signals may be due to external factors or due to the improper installation of sensors.

4.4. Training and testing dataset

We represent the discovery of abnormal frequent episodes as a problem of a binary classification problem. The baseline dataset didn't include any categorical information that could be easily used by our machine learning models for the training and testing process. So, we added a column to the dataset called "class" label to identify normal and abnormal frequent episodes. Based on our previous work, we identify an instance of a data as "abnormal" if a peak is detected; otherwise, it is noted as "normal". Numeric value 1 is used for "abnormal" and 0 for “normal”. Next, we looked at the distribution of sample examples in the dataset. Fig. 10 shows the distribution of samples in our dataset, and it is clear that we have a challenge of an imbalanced classification problem [29]. So, the "normal" instances are majority class, and the "abnormal" examples are minority class. The training and testing datasets are created with the most common split percentages 70%–30%.

Table 1
Evaluation results for imbalanced class data.

Model	Accuracy	Precision	Recall	F1-measure
LR	0.96	0.6	0.18	0.27
LDA	0.95	0.411	0.47	0.43
KNN	0.956	0.375	0.06	0.1
DT	0.934	0.23	0.21	0.22
NB	0.689	0.084	0.64	0.148
SVM	0.96	0.575	0.14	0.22

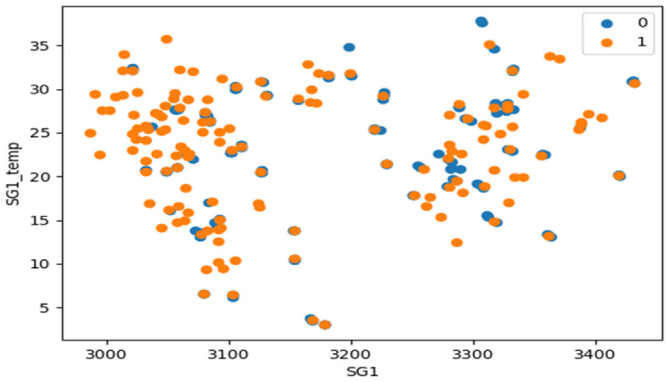

Fig. 11. Distribution of Balanced samples.

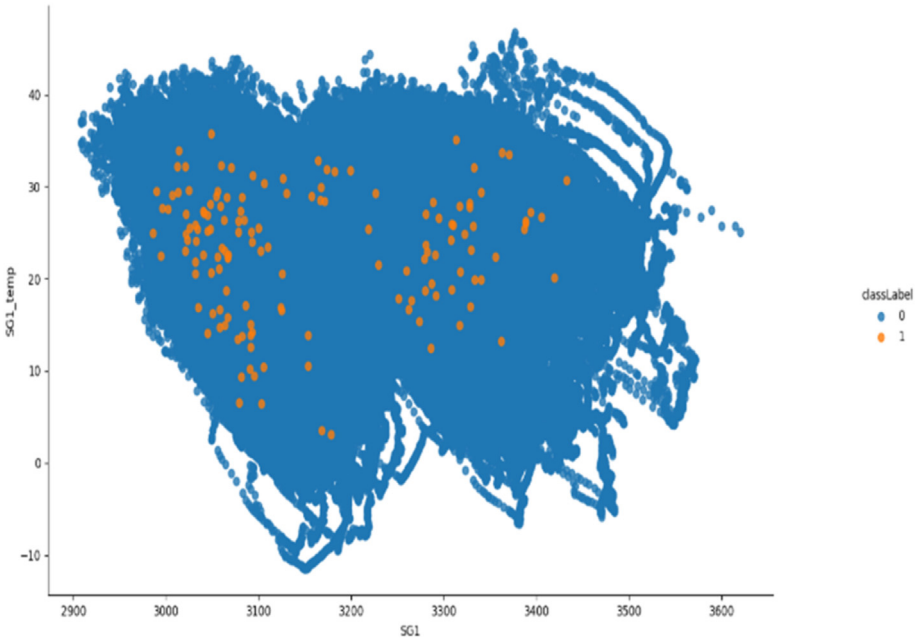

Fig. 10. Distribution of imbalanced samples.

Table 2
Evaluation results for balanced class Data.

Model	Accuracy	Precision	Recall	F1-measure
LR	0.59	0.602	0.554	0.573
LDA	0.59	0.606	0.564	0.578
KNN	0.63	**0.637**	0.605	**0.61**
DT	0.564	0.544	0.502	0.554
NB	0.545	0.534	**0.707**	0.601
SVM	0.565	0.559	0.585	0.566

4.5. Classification results

Table 1 shows the evaluation of models (prediction results) on the imbalanced data. Even though the prediction accuracy is high, it is inappropriate because of the imbalanced classification problem. So, the precision, recall, and F1-measure are used to evaluate the performance of the models on the minority class. The results are shown in Table 1 are based on the most commonly used 10-fold cross-validation. Next, we select the under-sampling method [30] to address the imbalanced classification problem because, in our dataset, the majority class instances are 272936, and minority class instances are 142. Fig. 11 shows the distribution of samples after under-sampling. Table 2 shows the evaluation results of the models on balanced data, and we can see that both the classes are emphasized with reduced accuracy.

In Table 2, the metric “accuracy” emphasizes the “normal” instances in the dataset. However, our focus is the “abnormal” instances and how well the models capture through labelling. So, we are more interested in metrics “Precision” and “Recall”. Precision tells how precise is that model (i.e. how many of the predicted “abnormal” instances are actually “abnormal”). Where as the Recall tells us which model to select as our best model. So, based on the metrics “Precision” and “Recall”, there are two models “KNN” and “NB”; but we prefer a balanced between Precision and Recall which is metric “F1-Score measure”. Based on the F1-score we select “KNN” is the best model for classification of “normal” and “abnormal” frequent episodes.

5. Conclusions and future work

In this paper, we introduce an advanced analytics engine (ADA) that is data analytics, feature engineering, and machine learning framework to determine abnormal frequent episodes. We evaluated the data analysis approach using the baseline data. The performance evaluation results show that our approach is able to determine more normal frequent episodes and abnormal frequent episodes.

For our future work, we'll be extending our advanced analytic engine framework to add more features, derived from the data. With more features available to our learning method, more abnormal frequent episodes may become visible in the presence of noise, and then it will be practical to estimate the likelihood of significant "incident" and possible system failure in the ISD system. Also, we will be applying our advanced analytic engine framework to newly collected data to discover abnormal episodes in the presence of background noise.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by the U.S. Department of Energy, Office of Environmental Management (EM) MSIPP program under TOA #T0000456309.

References

[1] M.K. Saggi, S. Jain, A survey towards an integration of big data analytics to big insights for value-creation, Inf. Process. Manag. 54 (2018) 758–790, https://doi.org/10.1016/j.ipm.2018.01.010.
[2] C.W. Tsai, C.F. Lai, H.C. Chao, A.V. Vasilakos, Big data analytics: a survey, J. Big Data 2 (2015) 1–32, https://doi.org/10.1186/s40537-015-0030-3.
[3] J.S. Dhoble, N. Shelke, Investigative research on big data: an analysis, available at, Int. J. Innov. Res. Sci. Eng. Technol. 4 (2015) 4476–4482. :, 06/13/2022, http://www.ijirset.com/upload/2015/june/58_12_Investigative.pdf.
[4] C. Ma, H.H. Zhang, X. Wang, Machine learning for big data analytics in plants, Trends Plant Sci. 19 (2014) 798–808, https://doi.org/10.1016/j.tplants.2014.08.004.
[5] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Waltham, MA, 2012.
[6] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (1999) 264–323, https://doi.org/10.1145/331499.331504.
[7] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Network. 16 (2005) 645–678, https://doi.org/10.1109/TNN.2005.845141.
[8] K.E. Zeigler, B.A. Ferguson, Development of an in-situ decommissioning sensor network test bed for structural condition monitoring, in: Waste Management 2012 Conference, Phoenix, AZ, USA, Feb 26 – March 1, 2012.
[9] K. Zeigler, B. Ferguson, D. Karapatakis, C. Herbst, C. Stripling, Development of a Sensor Network Test Bed for ISD Materials and Structural Condition Monitoring, Savannah River National Laboratory (SRNL), 2011, https://doi.org/10.2172/1018717. SRNL-STI-2011-00193.
[10] Z. Sun, A. Duncan, Y. Kim, K. Zeigler, Applying temporal data mining (TDM) on the baseline data acquired by the in-situ decommissioning (ISD) sensor network test bed, in: Waste Management 2018 Conference, Phoenix, AZ, USA, Mar 18-22, 2018.
[11] X. Ao, P. Luo, C. Li, F. Zhuang and Q. He, Online frequent episode mining, in: 2015 IEEE 31st International Conference on Data Engineering, pp. 891-902, https://doi.org/10.1109/ICDE.2015.7113342.
[12] P.S. Sastry, S. Laxman, K.P. Unnikrishnan, System and Method for Mining of Temporal Data, 2010 patent 7644078.
[13] D. Patnaik, S. Laxman, B. Chandramouli, N. Ramakrishnan, Efficient episode mining of dynamic event streams, Data Mining (ICDM), in: IEEE 12th International Conference, 2012, pp. 605–614, https://doi.org/10.1109/ICDM.2012.84.
[14] D. Patnaik, P. Sastry, K. Unnikrishnan, Inferring neuronal network connectivity from spike data: a temporal data mining approach, Sci. Program. 16 (2018) 49–77, https://doi.org/10.3233/SPR-2008-0242.
[15] S. Laxman, P.S. Sastry, K.P. Unnikrishnan, A fast algorithm for finding frequent episodes in event streams, in: KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 410–419, https://doi.org/10.1145/1281192.1281238.
[16] X. Ao, H. Shi, J. Wang, L. Zuo, H. Li, Q. He, Large-scale frequent episode mining from complex event sequences with hierarchies, ACM Trans. Intell. Syst. Technol. 10 (2019) 1–26, https://doi.org/10.1145/3326163.
[17] P. Fournier-Viger, P. Yang, J. Lin, U. Yun, HUE-span: fast high utility episode mining, in: 15th International Conference on Advanced Data Mining and Applications (ADMA 2019), Dalian, China, Nov 21-23, 2019, https://doi.org/10.1007/978-3-030-35231-8_12.
[18] M. Amiri, L. Mohammad-Khanli, R. Mirandola, An online learning model based on episode mining for workload prediction in cloud, Future Generat. Comput. Syst. 87 (2018) 83–101, https://doi.org/10.1016/j.future.2018.04.044.
[19] T. You, Y. Li, B. Sun, C. Du, Multi-source data stream online frequent episode mining, IEEE Access 8 (2020) 107465–107478, https://doi.org/10.1109/ACCESS.2020.2997337.
[20] B. Biswal, A. Duncan, Z. Sun, Applying advanced data analytics methods to the baseline data of ISD sensor network testbed for system failure detection, in: Waste Management 2021 Conference, Phoenix, AZ, USA, Mar 8-12, 2021.
[21] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, J. Jiao, Weakly supervised instance segmentation using class peak response, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3791–3800, https://doi.org/10.1109/CVPR.2018.00399.
[22] B. Biswal, S. Shetty, T. Rogers, Enhanced learning classifier to locate data in cloud data centres, Int. J. Metaheuristics (IJMHeur) 4 (2015) 141–158, https://doi.org/10.1504/IJMHEUR.2015.074248.
[23] R. Suzuki, H. Shimodaira, Pvclust: an R package for assessing the uncertainty in hierarchical clustering, J. Bioinfo. 22 (2006) 1540–1542, https://doi.org/10.1093/bioinformatics/btl117.
[24] C. Negin, C. Urland, A. Szilagyi, DOE-EM'S in-situ decommissioning strategy, in: Waste Management 2021 Conference, Phoenix, AZ, USA, Feb 24-28, 2008.
[25] N. Carino, V. Li, G. Heath, G. Song, C. Wang, P. Ziehl, Development of a Remote Monitoring Sensor Network for in Situ Decommissioned Structures, Savannah

River National Laboratory (SRNL), 2010. SRNL-RP-2010-01666.
[26] S. Hansun, A new approach of Brown's double exponential smoothing method in time series analysis, Balkan J. Electr. Comput. Eng. (2016) 75–78, https://doi.org/10.17694/bajece.14351.
[27] R. Benjamin, B. Pierre, T. Nicolas, G. Paulo, V. Pierre, Fourier could be a data scientist: from graph Fourier transform to signal processing on graphs, Compt. Rendus Phys. 20 (2019) 474–488, https://doi.org/10.1016/j.crhy.2019.08.003.
[28] S. Taixin, M. Yang, T. Jin, R.C.C. Flesch, Power harmonic and interharmonic detection method in renewable power based on Nuttall double-window all-phase FFT algorithm, IET Renew. Power Gener. 12 (2018) 953–961, https://doi.org/10.1049/iet-rpg.2017.0115.
[29] N. Samad, P. Hamid, F. Eshagh, Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification, Neurocomputing 276 (2018) 55–66, https://doi.org/10.1016/j.neucom.2017.06.082.
[30] J. Zhang, I. Mani, kNN approach to unbalanced data distributions: a case study involving information extraction, available at, in: Proceeding of International Conference on Machine Learning (ICML 2003), Washington DC, Aug 21, 2003. :, 06/13/2022, https://www.site.uottawa.ca/~nat/Workshop2003/jzhang.pdf.