Journal of King Saud University – Computer and Information Sciences 33 (2021) 11–20

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Anomaly detection using control flow pattern and fuzzy regression in port container handling

Dewi Rahmawati [a,b], Riyanarto Sarno [b]

[a] Department Software Engineering, Faculty of Information Technology and Industry, Institut Teknologi Telkom Surabaya, Surabaya 60231, Jawa Timur, Indonesia
[b] Department of Informatics, Institut Teknologi Sepuluh Nopember, Jl. Teknik Kimia, Gedung Teknik Informatika, Kampus ITS Sukolilo, Surabaya 60111, Jawa Timur, Indonesia

ARTICLE INFO

Article history:
Received 5 June 2018
Revised 5 November 2018
Accepted 13 December 2018
Available online 11 January 2019

Keywords:
Anomaly
Control flow patterns
Fuzzy regression
Multiple linear regression

ABSTRACT

Deviations in port container handling can be detected by many factors. One of them is anomalies in the process model. Several studies have proposed anomaly detection methods. However, these methods do not accommodate verbal judgments of experts. These methods treat instances with low deviation as containing anomalies while in reality not all instances with low deviation contain anomalies. Considering this, a method was developed for detecting anomalies in port container handling using fuzzy regression in order to accommodate verbal expert judgments on the rate of anomaly (ROA). First, a control flow pattern is built to form an anomaly pattern that will be used for detecting wrong patterns. Then, five anomaly attributes were declared, i.e. skip sequence, wrong throughput time (max), wrong throughput time (min), wrong patterns and wrong decisions. In the experiment, the rate of anomaly was found using three methods, namely fuzzy regression (FR), support vector regression (SVR) with radial basis function (RBF) kernel, and multiple linear regression (MLR). The results showed that fuzzy regression was better at detecting anomalies than multiple linear regression and support vector regression. The experimental validation showed that fuzzy regression combined with control flow pattern was able to reduce false positives and false negatives. The sensitivity, specificity and accuracy of the proposed method were 96%, 97% and 99%, respectively.

1. Introduction

Anomalies in a business process can be interpreted as processes or subprocesses that are not carried out as usual and are not in accordance with standard operation procedures (SOP) (Amara et al., 2013). If an accurate anomaly detection system and a properly functioning early warning system are implemented, several problems can be prevented, ranging from fraud to inefficiency. Recently, a case of extortion in port container handling in Samarinda, East Borneo was uncovered. It is assumed that this has caused a total loss of about 2 trillion rupiah or about 9% of annual revenue. It occurred without being detected from 2010 to 2016 (Rahmah). Another example is, if work time deviations from standard operating procedures (SOPs) are identified timely, companies can change the pattern of work time intensity to minimize the possibility of transgressions.

Several studies have proposed anomaly detection methods however, they do not accommodate verbal expert judgments on the rate of anomaly (ROA). These methods treat instances with low deviation as anomalies, while in reality not all instances with low deviation are anomalies. Low deviation can be caused by ambiguity in determining the anomaly attribute values and low frequency of occurrence (Huda et al., 2016; Huda et al., 2015).

Here, we propose a method of detecting anomalies that accommodates verbal expert judgments on the rate of anomaly based on deviation from standard operating (SOP) procedures. SOP deviations are procedural errors that can cause anomalies. We propose the use of fuzzy regression and verbal expert judgments on the rate of anomaly to correctly detect fraudulent activities. We hypothesized that the weight of the anomaly attributes can affect the rate of anomaly. Finally, the anomaly rate is used to determine whether a deviation indicates high anomaly, medium anomaly, low anomaly or no anomaly. The major scientific contribution of this work is to reduce false positives and false negatives by improving the following indicators: skip sequence, skip decision, wrong throughput time (max), wrong throughput time (min), and wrong pattern.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

E-mail address: dewi16@mhs.if.its.ac.id (D. Rahmawati)

https://doi.org/10.1016/j.jksuci.2018.12.004
1319-1578/

A previous research (Correia, 2012) has proposed a standard business procedure with 43 variations of business process patterns from many fields called control flow patterns. In order to build an anomaly pattern model for detecting anomalies, especially wrong patterns, rules to build the anomaly patterns are needed. In this research, we built anomaly patterns and then used graph pattern matching to detect wrong patterns.

Graph isomorphism (Cordella et al., 2004; Nabti and Seba, 2016) is a method that can be added to graph pattern matching. It detects a subgraph model from a full graph model in the form of a business process model. The algorithm uses several graph matching rules, matching the degree of the nodes and the number of nodes and arcs. Here, we propose to build control flow patterns to detect wrong patterns by using graph isomorphism. However, the proposed graph isomorphism algorithm can only detect one of the anomaly attributes, i.e. wrong pattern. To detect the other 4 anomaly attributes, i.e. skip sequences, wrong throughput time (min), wrong throughput time (max), and wrong decision, detection rules are needed. Thus, this research declares rules for detecting 4 anomaly attributes.

In the proposed method anomalies are detected based on SOP deviations caused by skip sequence, wrong throughput time (min), wrong throughput time (max), wrong decision, and wrong pattern. The attribute values of the anomalies are discrete, i.e. *high anomaly*, *medium anomaly*, and *low anomaly*. Therefore, fuzzy regression is used to investigate the predicted anomaly rate (i.e. high anomaly, medium anomaly, and low anomaly) based on the attribute importance weights. We hypothesized that the importance weight of the attributes can provide the weight of attribute anomalies. Finally, the weight of the attribute anomalies can be used to determine whether an anomaly is high, medium, or low. The performance of fuzzy regression was compared with that of multiple linear regression and support vector regression with radial basis function (RBF) kernel to get the most accurate rate of anomaly predictions.

This research evaluated the accuracy of the proposed method in anomaly detection. This was done based on two aspects, i.e. sensitivity and specificity (Buijs et al., 2012).

The rest of this paper is structured as follows: Section 2 presents a literature review, Section 3 describes the proposed method for composing the control flow patterns and defining the anomaly attribute values, Section 4 reports the results and analysis of the experiment, Section 5 presents the conclusions of the research.

2. Material and methods

Rates of anomaly in port container handling in the form of discrete data were determined based on verbal expert judgments. Regression analysis is more appropriate for producing continuous output and not for discrete output as needed in the classification problem in this study. Some years ago regression-based classification has been introduced as an alternative method for dealing with classification problems. An interesting approach is classification based on support vector regression producing higher accuracy than multiple linear and non-linear regression. Regression-based classification has been successfully applied in face recognition (Liu et al., 2013) and data streaming (Osojnik et al., 2017) in real cases. However, while regression algorithms have widely been applied, problems may arise in some circumstances (Shapiro, 2005):

- Uncertainty between input and output variables (Chandola et al., Jul. 2009)
- Inadequate number of observations (Saia et al., 2017)
- Difficulty in verifying distribution assumptions (Mahapatra and Chandola, 1804)
- Ambiguity of incidence or extent of occurrence (Saia, 2017)
- Inaccuracies and distortions introduced by linearization (Akoglu et al., 2013)

From the description above, statistical regression is problematic. In this study, a fuzzy regression method was therefore developed to detect anomalies, accommodating verbal expert judgments on the rate of anomaly. Several regression techniques were evaluated to determine the regression model. The regression output was decoded corresponding to the class label (high anomaly, medium anomaly and low anomaly). Three regression methods, MLR, SVR RBF kernel and fuzzy regression (FR) were compared to get the most accurate rate of anomaly prediction. The details of the regression methods are given below.

2.1. Multiple linear regression

Algorithms to model the relationship between two or more independent and dependent variables or responses operate by applying a linear equation to the observed data. Each value of the independent variable x is related to a value of the dependent variable y (DeForest et al., 2017). In the present case of detecting anomalies in port container handling, multivariate data of attribute anomalies were used to predict the ROA. In an experiment, the ROA values (y_i) were predicted by five predictors $(x_1, x_2, \ldots, x_5)$ as expressed by the equation as shown in Eq. (1).

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \quad (1)$$

where $\beta_1 - \beta_5$ are independent variables; α is an dependent variable; and variables x_1 to x_5 explain the anomaly attributes as shown in Table 1. Variable x_1 designates skip sequences, variable $x_2$designates wrong decision, variable x_3 designates wrong pattern, variable x_4 designates wrong throughput time (min), variable x_5 designates wrong throughput time (max) and variable y_i designates the rate of anomaly.

The data used in the experiment described a total of 65,000 instances in 4 months. An expert in port container handling divided the data into 4 classes, namely High Anomaly, Medium Anomaly, Low Anomaly and No Anomaly. Also, the expert specified the following thresholds: below 0.6 means No Anomaly, between 0.6 and 0.7 means Low Anomaly, between 0.7 and 0.8 means Medium Anomaly, and between 0.8 and 1 means High Anomaly. With these thresholds, we found 680 instances containing anomalies and 64,320 instances not containing anomalies. The 680 instances with anomalies were used as training and testing data: 60% of the total anomaly instances were used as training data and 40% were used as testing data, as shown in Table 3. On average, there were 5 instances with anomalies every day (Table 4).

The training data used to build the multiple linear regression models with x consisted of 5 attribute values with y as the rate of anomaly. Then, with the Minitab tool, the multiple linear regression model was built as formulated in Eqs. (2), (3) and (4):

$$y_H = 0.220 + 0.858\,x_1 + 1.03\,x_3 + 0.587x_4 \quad (2)$$

$$y_M = 0.229 + 0.196\,x_1 + 1.01\,x_3 + 0.594x_4 + 0.666x_5 \quad (3)$$

$$y_L = 0.545 + 0.405\,x_1 + 0.304\,x_3 - 0.115x_4 + 0.196x_5 \quad (4)$$

where y denotes the output of the MLR model.

2.2. Support vector regression

SVR is based on the same principle as SVM, with some small differences. The main difference is that is produces a continuous output instead of a discrete output corresponding to the real numbers

Table 1
Anomaly Attribute Values and Rate of Anomaly.

Case	Skip Sequences (x_1)	Wrong Decision (x_2)	Wrong Patterns (x_3)	Wrong Throughput Time MIN (x_4)	Wrong Throughput Time MAX (x_5)	Rate of Anomaly (y_i)
1	0	0.07	0.07	0.5	0.33	0.65
2	0.22	0	0.93	0.03	0	0.51
3	0	0.03	0	0.5	0.64	0.90
4	0.03	0.02	0.02	0.5	0.45	0.72
...	...	...	...	...	...	...
65,000	0	0	0	0	0.03	0.003

Table 2
Ranges of Rate of Anomaly used for Classification.

Class Label	Rate of Anomaly Range	Number of Cases
High Anomaly	>0.8–1	80
Medium Anomaly	>0.7–0.8	280
Low Anomaly	>0.6–0.7	320
No Anomaly	0–0.6	64,320

Table 3
Number of Cases for Training and Testing Log Data.

Class Label	Number of Cases for Training (60%)	Number of Cases for Testing (40%)
High Anomaly	48	32
Medium Anomaly	168	112
Low Anomaly	192	128
No Anomaly	38,592	25,728

Table 4
Metrics for Regression Analysis.

Metrics	Equation	Descriptions
RMSE	$RMSE(a.p) = \sqrt{\frac{\sum_{i=1}^{n}(a-p)^2}{n}}$	RMSE measures the errors between the actual and the predicted values. A lower RMSE indicates few prediction errors.
Metrics	Equation	Descriptions
R^2	$R^2(a.p) = 1 - \frac{\sum_{i=1}^{n}(a_1-p_i)^2}{\sum_{i=1}^{n}(a_1-\bar{p})^2}$	R-squared designates the fitting between the actual and the predicted vectors, i.e. a good fit is represented by high R^2 and vice versa. Usually, the range of R^2 is between 0 and 1.
AdjustedR^2	$\bar{R}^2(a.p) = R^2(a.p) - \frac{(1-R^2(a.p))k}{n-k-1}$	Adjusted R^2 is a variant of R^2 to reduce the bias caused by the addition of predictors.

in the regression problem. Epsilon needs to be set as a tolerance margin in SVM because it is difficult to predict continuous information with unlimited possibilities. In the regression analysis, the main idea is always to minimize the error considering the tolerance margin (Estelles-Lopez et al., 2017). In this experiment, the RBF kernel function can be expressed as in Eq. (5), where $||x_i - x_j||^2$ is the square Euclidean distance and σ is a free parameter:

$$k(x_i, x_j) = exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \quad (5)$$

2.3. Fuzzy regression

In previous studies, such as (Shapiro, 2005), fuzzy regression (FR) was used to predict the association of risk factors with work-related injuries. According to (Uejima and Asai, 1982), the general form of the model is as shown in Eq. (6).

$$\bar{Y} = \bar{B}_0 + \bar{B}_1 x_1 + \bar{B}_2 x_2 + \bar{B}_3 x_3 + \bar{B}_4 x_4 + \bar{B}_5 x_5 \quad (6)$$

where $\bar{Y}$ is the fuzzy output, $x_1 - x_5$ are non fuzzy input vectors, $\bar{B}_1 - \bar{B}_5$ are independent variables, and $\bar{B}_0$ is a dependent variable. The membership function (MF) $\mu_a(\alpha)$, as shown in Fig. 1, uses a trapezoidal membership function for better results in most of the cases. The selection criteria of a, b, c,and d are used for rate of anomaly (ROA) membership.

Based on Fig. 1, $\mu_a(\alpha)$ is a mode. The basic idea of the fuzzy regression approach is to minimize the model's ambiguity. Two common methods for developing a fuzzy regression models are: (1) models in which the relationship between the variables are fuzzy; and (2) models in which the variables themselves are fuzzy (Sánchez and Gómez, 2004; Andrés-Sánchez, 2016; Bell and Heng, 1997; Pedregosa et al., 2011). Here, we focused on models in which the data are discrete and the relationships between the variables are uncertain.

In the experiment, the ROA values $\left(\bar{Y}\right)$ consisted of High ROA $\left(\bar{Y}_H\right)$, Medium ROA $\left(\bar{Y}_M\right)$ and Low ROA $\left(\bar{Y}_L\right)$, predicted by five predictors ($x_1, x_2, \ldots, x_5$). To build the fuzzy regression models, the method differs from multiple linear regression. The first step is to build the membership of the 5 attributes, i.e. Skip Sequence, Wrong Throughput Time (Min), Wrong Throughput Time (Max), Wrong Decision, Wrong Pattern and Rate of Anomaly. The next step is to divide the anomalies into low, medium and high anomalies to find the fuzzy rate of each criterion. Then multiple linear regression is used to obtain the equation for fuzzy low anomaly, fuzzy medium anomaly and fuzzy high anomaly for the training process. Equations (7), (8) and (9) are the resulting regression equations for fuzzy low anomalies, fuzzy medium anomalies, and fuzzy high anomalies, respectively. Thus, the FR model can be formulated as shown in Eq. (7), Eq. (8) and Eq. (9).

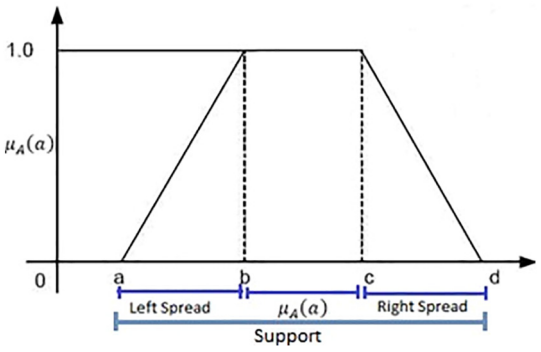

Fig. 1. Fuzzy Coefficient.

$$\bar{Y}_H = 0.566 + 0\,x_1 + 0.315\,x_2 + 0.0692\,x_3 + 0.0096\,x_4 + 0.0398\,x_5 \quad (7)$$

$$\bar{Y}_M = 0.727 + 0\,x_1 + 0\,x_2 + 0.0014\,x_3 + 0.151\,x_4 + 0.00106\,x_5 \quad (8)$$

$$\bar{Y}_L = 0.664 + 0\,x_1 + 0\,x_2 + 0.0737\,x_3 + 0.050\,x_4 + 0.035\,x_5 \quad (9)$$

Scikit-learn was used as the machine learning library for the regression analysis (Rahmawati et al., 2017). Further, three metrics for measuring the performances of the different regression methods were used: root mean squared error (*RMSE*), R-squared $\left(R^2\right)$, and adjusted R-squared (adjusted R). Table 10 shows the details of these metrics. The symbols *a, p, k, n*, denote actual vector, predicted vector, number of predictors, and total samples, respectively. In this research, actual vector designates the observed rate of anomaly and prediction vector designates the result of the regression analysis.

Before the performance comparison, the continuous output produced by the best regression model needs to be converted to a discrete output in accordance with the labels to be used based on the three thresholds. The thresholds were determined by an expert on port container handling. A comparison of these thresholds (Threshold 1, 2 and 3) is shown in Fig. 2 using the receiver operating characteristics (ROC) space. The ROC space was created by plotting the true positive rate (TPR), or sensitivity, against the false positive rate (FPR), or specificity, at various threshold settings. Here, the correct thresholds were found by looking at the diagonal (random guess) dividing the ROC space. Points above the diagonal represent good classification results (better than random), while points below the line represent bad results (worse than random).

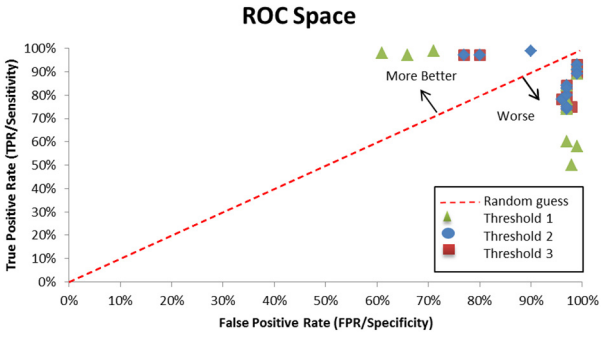

Fig. 2. ROC Space of threshold.

Decoding was done with the following thresholds:

$$class(\bar{Y}) = \begin{Bmatrix} low, & 0.6 \le x \le 0.7 \\ medium, & 0.7 \le x \le 0.8 \\ high, & 0.8 \le x \le 1 \end{Bmatrix} \quad (10)$$

where *low, medium,* and *high* indicate a low, medium, and high anomaly, respectively, and x is an anomaly instances in the event log. $\bar{Y}$ is the predicted ROA.

2.4. The accuracy of detecting an anomaly

The fuzzy regression method proposed in this paper was evaluated by measuring its accuracy of anomaly detection. This research used sensitivity, specificity and accuracy as the metrics (Liu et al., 2013). Sensitivity measures how many positive anomaly instances were correctly identified, specificity measures how many negative anomaly instances were correctly identified, and accuracy measures the level of proximity of the measurement results with the actual anomaly values. The equations for calculating sensitivity, specificity and accuracy are Eq. (11), Eq. (12), Eq. (13). The confusion matrix is shown in Table 5. In Eq. (11), the variable *True Positive* stores the number of correctly identified positive anomaly instances and *False Negative* stores the number of incorrectly identified positive anomaly instances. Here, we did not consider false alarm performance. In Eq. (12), the variable *True Negative* stores the number of correctly identified negative anomaly instances

Table 5
Confusion Matrix for Calculation of Sensitivity Fuzzy Regression.

Classifier	Confusion matrix				
		true-high (*TH*)	true-med (*TM*)	true-low (*TL*)	true-no (*TNa*)
Fuzzy Regression	pred-high	*TH*	*FM*	*FL*	*FNa*
	pred-med	*FH*	*TM*	*FL*	*FNa*
	pred-low	*FH*	*FM*	*TL*	*FNa*
	pred-no	*FH*	*FM*	*FL*	*TNa*
	Sensitivity	$\frac{TH}{TH+FH}$	$\frac{TM}{TM+FM}$	$\frac{TL}{TL+FL}$	$\frac{TNa}{TNa+FNa}$
	Specificity	$\frac{TM+TL+TNa}{M+L+Na}$	$\frac{TH+TL+TNa}{H+L+Na}$	$\frac{TH+TM+TNa}{H+M+Na}$	$\frac{TH+TM+TL}{H+M+L}$
	Accuracy	$\frac{TH+TM+TL+TNa}{H+M+L+Na}$			

Table 6
Method of Composing Anomaly Patterns.

Composing Process Model with YAWL Petri net	**Composing Process Model with YAWL Petri net**	**Build Control Flow Patterns with Anomaly Pattern**	
	Input:Transformation results in 65000 log data (December 2015-March 2016)	Input:A set of rules of anomaly patterns and the process model	Build
	Output:Process model	Output:Control flow pattern in the form of anomaly patterns	
	Control Flow Patterns with Anomaly Pattern		
Input: Transformation results in 65,000 log data (December 2015-March 2016)	Input: A set of rules of anomaly patterns and the process model		
Output: Process model	Output: Control flow pattern in the form of anomaly patterns		

and *False Positive* stores the number of incorrectly identified negative anomaly instances.

$$Sensitivity = \frac{True\ Positive(TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad (11)$$

$$Specificity = \frac{TrueNegative(TN)}{TrueNegative(TN) + FalsePositive(FP)} \quad (12)$$

$$Accuracy = \frac{TruePositive(TP) + TrueNegative(TN)}{TP + TN + FP + FN} \quad (13)$$

3. Results

The proposed method consists of two steps. The first step is composing the control-flow patterns (CFP) with anomaly pattern rules. The second step is defining the attribute values of the anomalies.

Table 7
Rules for Anomaly Patterns.

Anomaly Rule	Description
no_anomaly_patterns (a,c)	If A is recorded, then the next recorded activity is C (No anomaly)
anomaly_patterns (a, c)	If C is recorded, then the next recorded activity is A (Anomaly)

3.1. Composing control flow patterns

For composing the anomaly patterns, several rules are proposed. The role model that is used in this research contains *no_anomaly_patterns (a, c)* and *anomaly_patterns (a, c)*. The method of composing the anomaly patterns is shown in Table 6. The anomaly patterns are directly formed based on anomaly rules as shown in Table 7, which contains *Sequence Patterns* and *Exclusive choice patterns* as examples. Based on Fig. 3, the patterns are formed by considering the *no_anomaly_patterns* and *anomaly_patterns* rules. The process model was built with Yet Another Workflow Language (YAWL) (Huda et al., 2015; Rahmawati et al., 2017).

After composing the anomaly patterns, the control flow patterns are converted into string rules in a Java program to be used for detection of anomalies by finding wrong patterns in the event log. The method for detecting wrong patterns with anomaly pattern rules is shown in Fig. 4.

After getting the anomaly pattern rules, the degrees of the nodes are calculated, which are used for the parameters to detect wrong pattern anomalies using graph isomorphism (graph pattern matching). Wrong patterns can only be detected when the graphs are isomorphic. The requirement of graph isomorphism is fulfilled when the degrees of the nodes, and the number of arcs and nodes in both graphs are the same.

If these three requirements are met but the graphs are still not isomorphic, another additional factor is needed. A dependencies matrix is proposed to detect the similarity of two graphs by so-called graph pattern matching. Examples of the application of graph isomorphism with a dependencies matrix in a instance containing anomalies is shown in Fig. 5.

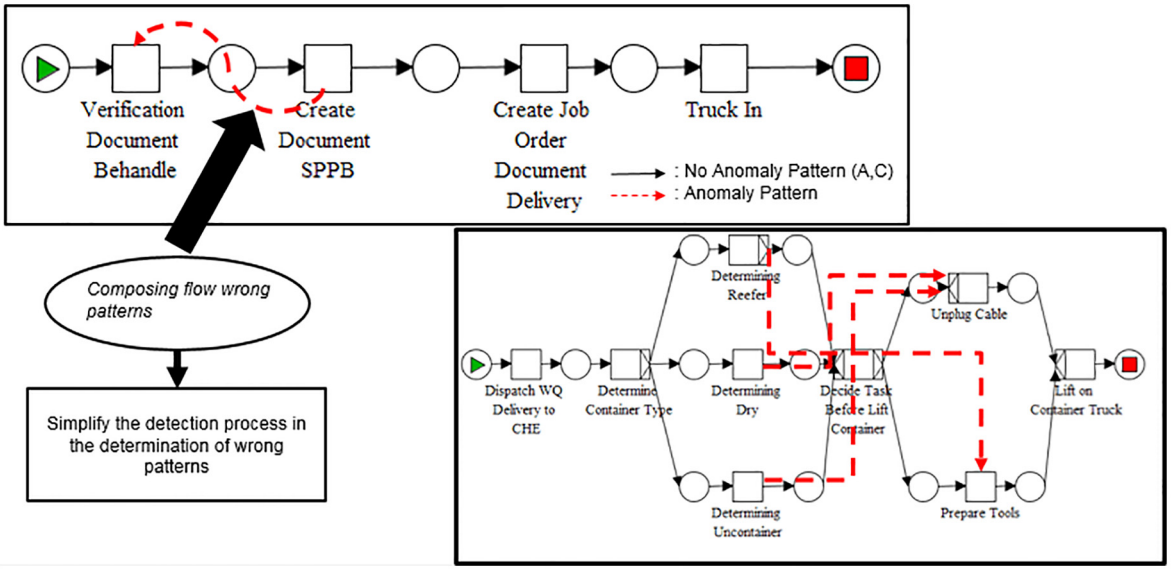

Fig. 3. Composing anomaly patterns with control flow patterns.

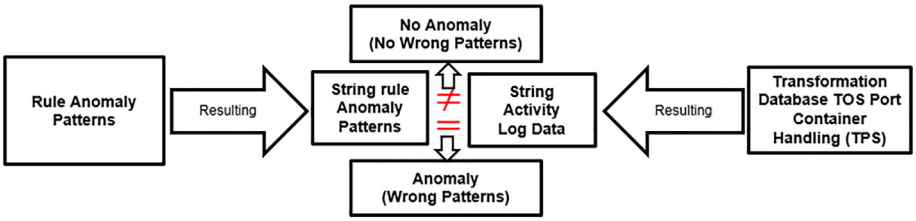

Fig. 4. Method of detecting wrong patterns with anomaly pattern rules.

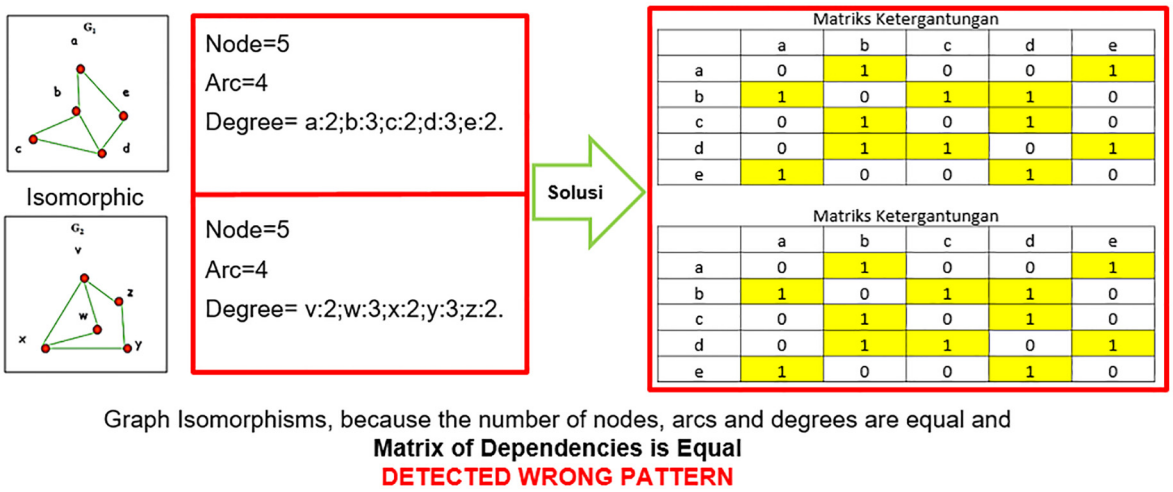

Fig. 5. Wrong pattern with dependencies matrix.

3.2. Defining attribute values of anomaly

The attribute values of the anomalies are obtained from the output values of the anomaly detection in the Java program. Pseudocode for detecting anomalies in a log of port container handling is shown in Table 8.

After the detection has been done automatically by the Java program, the output of the detection process contains values for each anomaly attribute. Table 9 shows the 5 anomaly attribute values.

4. Discussion

This research aimed to find the best algorithm for detecting anomalies, especially low anomalies. Previous methods treat instances of low anomaly as anomalies, but not all instances with low anomaly are anomalies. To deal with this issue, we propose a method for correctly detecting anomaly instances. We hypothesized that fuzzy regression is a good algorithm for detecting anomalies, especially low anomalies, because it can accommodate verbal expert judgments on rate of anomaly. We tested this hypothesis by measuring the specificity, sensitivity and accuracy of anomaly detection by the proposed method.

The proposed method has already been implemented and evaluated in a program in port container handling from December 2015 to March 2016. An expert in port container handling specified the thresholds to distinguish the No Anomaly, Low Anomaly, Medium Anomaly and High Anomaly classes.

In 65,000 port container handling data logs (December 2015-March 2016), the system found anomalies in 680 data logs and no anomalies in 64,320 data logs (see Table 2). Subsequently, 60% of the anomaly data were used as training data and 40% were used as testing data, as shown in Table 3. A screenshot of the pro-

Table 8
Rule of Anomaly Patterns.

Attribute Anomaly	Pseudocode
Skip Sequences	1. for i = 0 to activity in one case_SOP 2. if activity_sop not the same as activity_log 3. && not decision_activity
Attribute Anomaly	Pseudocode
	4. skip_activity ++ 5. return skip_activity / activity in log case
Throughput Time Maximal or Minimal	1. for i = 0 to activity in one case SOP 2. if time_cost_activity_log lower than 3. standard_time 4. throughputMin++ 5. if time_cost_activity_log higher than 6. throughputMax++ 7. return throughputMin/total_activity and 8. throughputMax/total_activity
Wrong Pattern	1. for i = 0 to activity in one case_SOP 2. if activity_sop not same as activity_log 3. patternSop[] <- activity_sop 4. patternLog[] <- activity_log 5. if patternSop not same as patternLog 6. wrongPattern += total_wrong_index 7. return wrongPattern / patternLog size
Wrong Decision	1. for i = 0 to activity in one case_SOP 2. if type_container is dry 3. if yard_block_log not same as SOP 4. wrongDecision++ 5. if yard_slot not same asanSOP 6. wrongDecision ++ 7. return wrongDecision / 2

Table 9
Anomaly Patterns Rules.

Attribute Anomaly	Number of Attribute Values	Score	Attribute Value
Skip Sequence	1	31/31	1
	30	30/31	0.96
	...	...	...
	0	0/31	0
Wrong	*Throughput Time (Min)*		
	$-.(StandardDeviation)$	$-.(StandardDeviation)$	
	$-1/2.(StandardDeviation)$	1 $-1/2.(StandardDeviation)$	...
Attribute Anomaly	Number of Attribute Values	Score	Attribute Value
	$\bar{x}$	*0*	*0*
Wrong	*Throughput Time (Max)*		
	$+.(StandardDeviation)$	$+.(StandardDeviation)$	1
	$+1/2.(StandardDeviation)$	$+1/2.(StandardDeviation)$	...
	$\bar{x}$	0	0
Wrong Decision	2	2/2	1
	1	½	0.50
	0	0/2	0
Wrong Pattern	30	30/30	1
	...	...	...
	1	1/30	0.03
	0	0	0

Table 10
The Performance Comparison of regression analysis.

Metrics	Multiple Linear Regression				Support Vector Regression with RBF kernel				Fuzzy Regression			
	High	Medium	Low	Overall	High	Medium	Low	Overall	High	Medium	Low	Overall
RMSE	27.23	35.14	45.37	40.17	29.42	23.17	40.32	30.17	0.12	5.32	3.32	4.12
R^2	0.30	0.12	0.1	0.5	0.3	0.013	0.6	0.04	1	0.51	0.98	0.98
AdjustedR^2	0.2	0.1	0.1	0.5	0.2	0.03	0.3	0.04	1	0.46	0.98	0.98

gram for detecting Wrong Throughput Time (Min/Max) is shown in Fig. 6, Skip Sequence, Wrong Decision and Wrong Pattern are shown in Fig. 7.

The performance of the proposed method using MLR, SVR with RBF kernel and FR was evaluated to determine the best regression method for ROA prediction. Table 10 shows a visual comparison of MLR, SVR with RBF kernel and FR. Fig. 8 indicates that the ROA can be predicted well by MLR, despite some errors. The plot between the actual and the predicted ROA shows that the majority of predictions follow the line of equity ($y = x$), with a margin of 0.30. However, some predictions of high anomalies differed greatly from the actual ROA. A high RMSE confirms that an instance belongs to the High Anomaly" class (27.23). The values of R^2 and adjusted R^2 were only 0.3 and 0.2, respectively. They indicate that the MLR model could only correctly predict less than half of the actual ROA variance. The performance of SVR with RBF kernel is shown in Fig. 9. It showed better performance than MLR, especially for Low Anomaly and Medium Anomaly predictions. The overall RMSE value of 30% shows that it performed better than MLR but was less accurate than FR. FR, shown in Fig. 10, had the best performance against MLR and SVR with RBF kernel as shown by having the lowest RMSE and the largest portion of variance that could be correctly predicted by the model (R^2 = 0.98 and adjusted R^2 = 0.98).

Based on Fig. 10 and Table 10, fuzzy regression (FR) was used to classify the rate of anomaly after the output was decoded. Table 10 shows a performance comparison of MLR, SVR with RBF kernel and FR in predicting the rate of anomaly. The prefixes 'true' and 'pred' denote the actual and predicted values, respectively. The proposed fuzzy regression method showed the best performance against MLR and SVR with RBF kernel. It could recognize High Anomaly, Medium Anomaly, Low Anomaly and No Anomaly with an accuracy of 99%. SVR with RBF kernel recognized High Anomaly, Medium Anomaly, Low Anomaly and No Anomaly with an accuracy of 97%.

MLR and SVR with RBF kernel had difficulty detecting low anomalies. This is indicated by their low sensitivity, specificity and accuracy in low anomaly prediction. These results confirm that low anomalies are hard to recognize by ordinary classifiers. Fuzzy regression achieved the highest sensitivity, specificity and accuracy for the Low Anomaly class. The measurements of sensitivity, specificity and accuracy are presented in Table 11 and Figs. 11, 12 and 13, respectively. Fig. 11 presents a comparison of the sensitivity values of MLR, SVR with RBF kernel and FR., Fig. 12 shows a comparison of the specificity values of MLR, SVR with RBF kernel and FR, and Fig. 13 shows a comparison of the accuracy values of MLR, SVR with RBF kernel and FR.

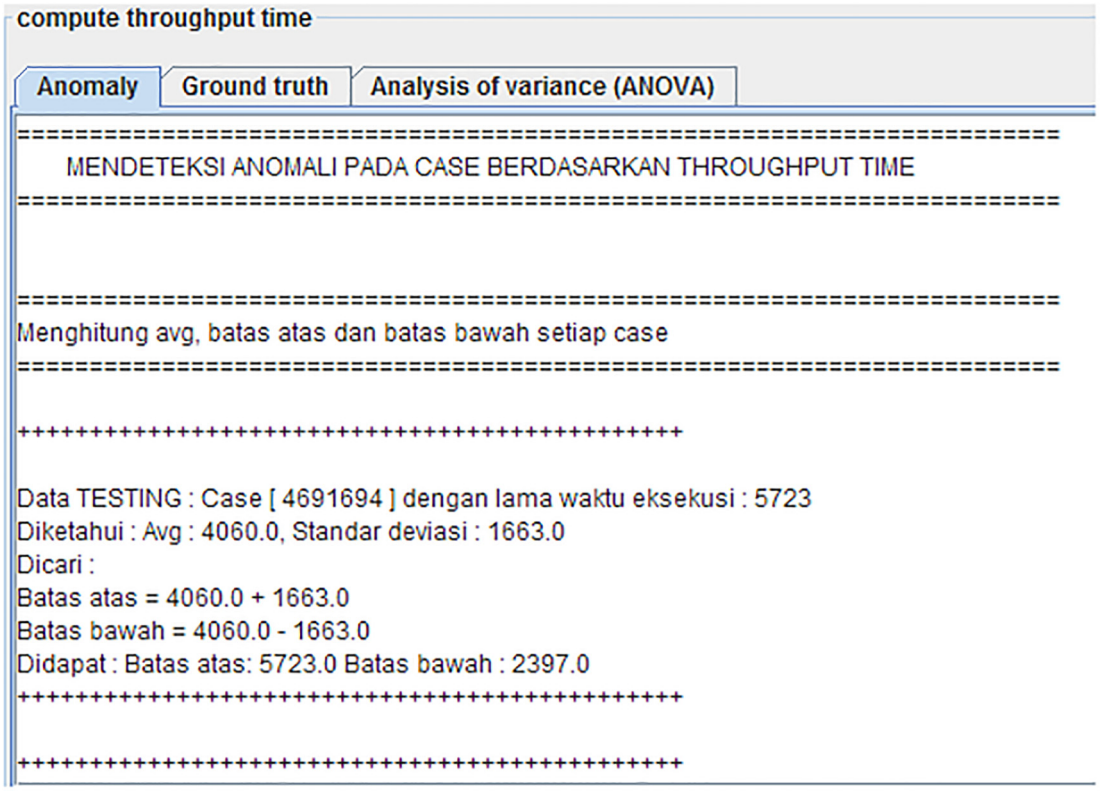

Fig. 6. Result of detecting Wrong Throughput Time (Min and Max).

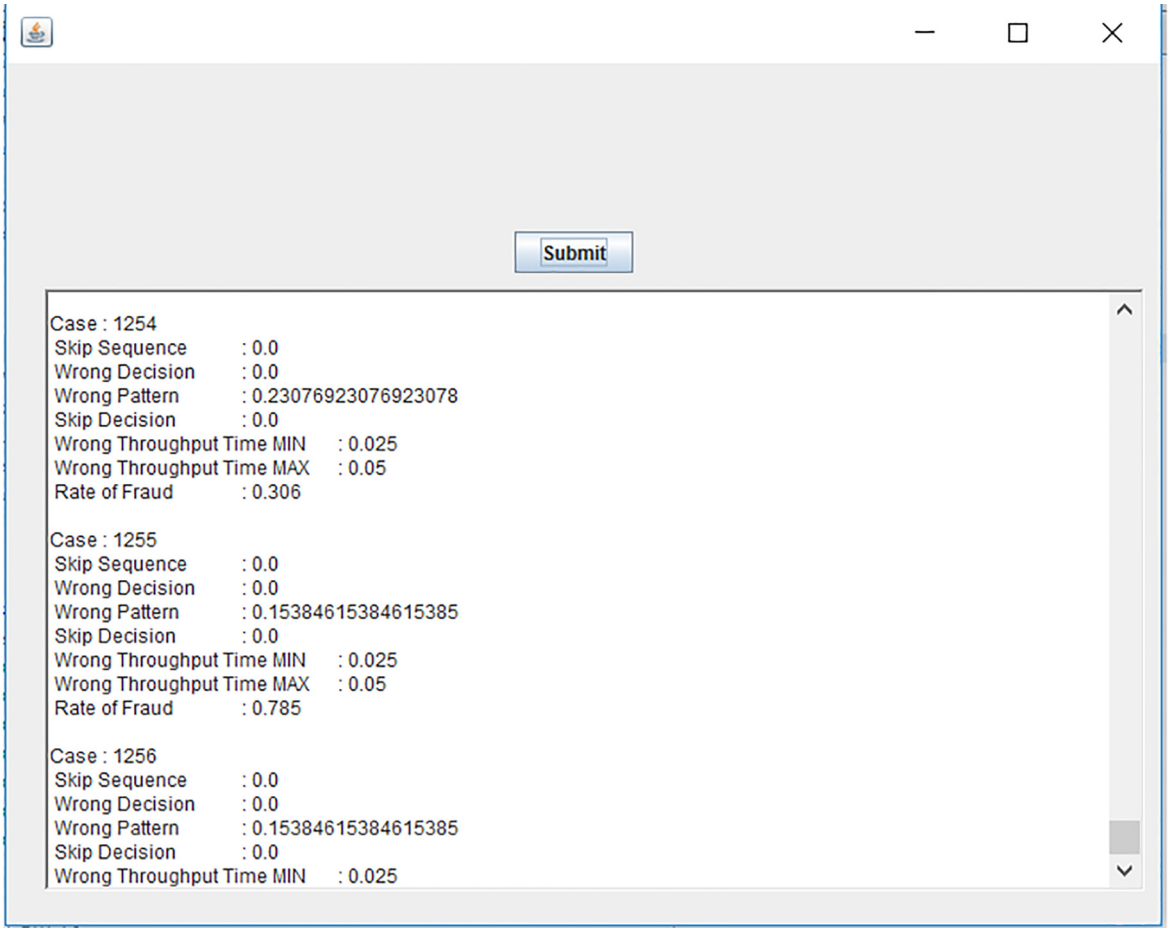

Fig. 7. Result of detecting Skip Sequence, Wrong Pattern and Wrong Decision.

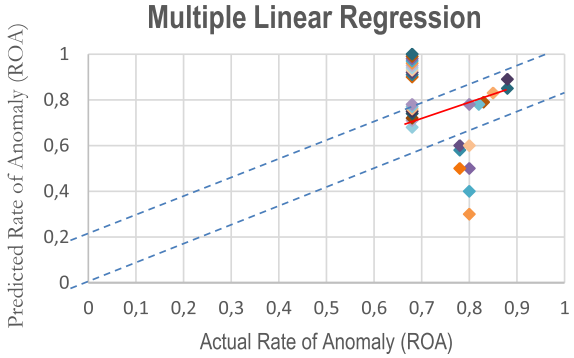

Fig. 8. Results of regression analysis to predict ROA using multiple linear regression.

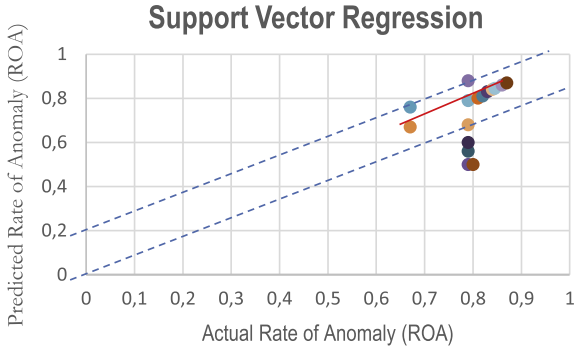

Fig. 9. Results of regression analysis to predict ROA using SVR with RBF kernel.

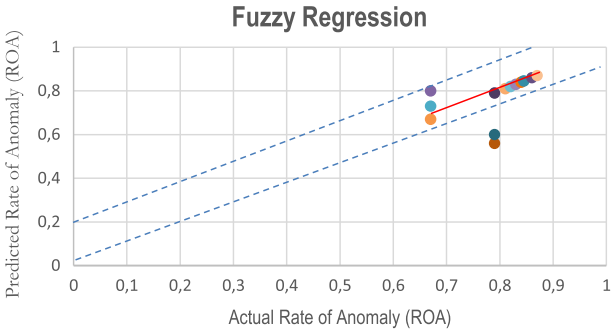

Fig. 10. Results of regression analysis to predict ROA using fuzzy regression.

5. Conclusions

In this study, a method for detecting low anomalies in container port handling that accomodates verbal expert judgments on rate of anomaly was developed. It was found that fuzzy regression performed better as part of the method compared to support vector regression with radial basis function kernel and multiple linear regression. Fuzzy regression achieved the highest sensitivity, specificity and accuracy for the Low Anomaly class. The result of the experiment confirmed that low anomalies are hard to recognize by ordinary classifiers, as proved by the sensitivity, specificity and accuracy values of the FR method, which were 96%, 97% and

Table 11
Confusion Matrix of Regression Analysis.

Classifier	Confusion matrix					Sensitivity	Specificity	Accuracy
		true-high	true-med	true-low	true-no			
Multiple Linear Regression (MLR)	pred-high	25	30	0	28	78%	97%	96%
	pred-med	5	145	15	200	60%	97%	
	pred-low	2	40	95	372	74%	97%	
	pred-no	0	25	18	25,000	97%	66%	
Support Vector Regression (SVR) with RBF kernel	pred-high	27	35	11	0	84%	97%	97%
	pred-med	3	120	10	100	50%	98%	
	pred-low	2	70	97	372	75%	97%	
	pred-no	0	15	10	25,128	98%	61%	
Fuzzy Regression (FR)	pred-high	30	50	0	30	93%	99%	99%
	pred-med	2	140	5	10	58%	99%	
	pred-low	0	50	115	10	89%	99%	
	pred-no	0	0	8	25,550	99%	71%	

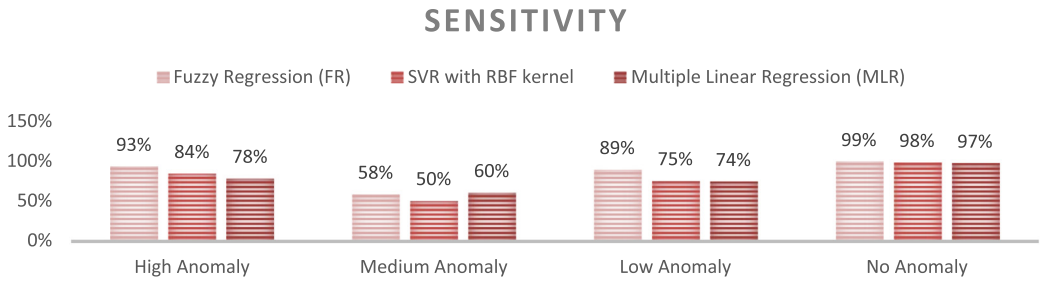

Fig. 11. Sensitivity of detecting anomalies using FR, SVR, and MLR.

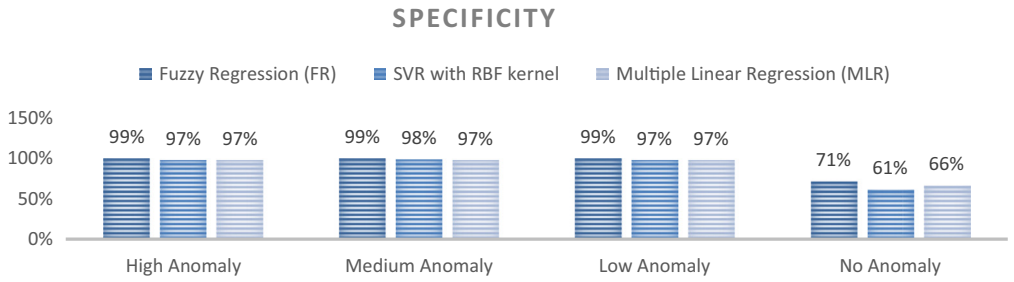

Fig. 12. Specificity of detecting anomalies using FR, SVR, and MLR.

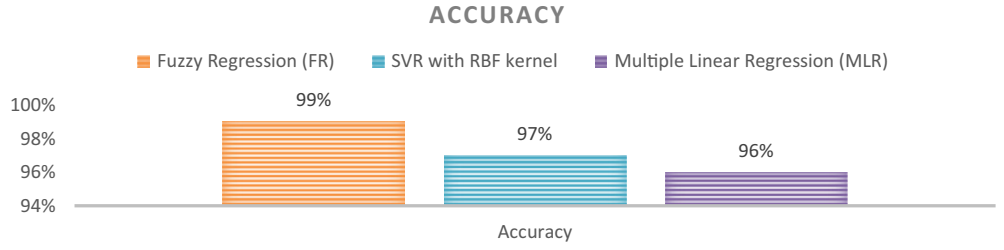

Fig. 13. Accuracy of detecting anomalies using FR, SVR, and MLR.

99%, respectively. Further exploration of optimization to determine the membership function of interval type-2 fuzzy sets is needed. Future work also includes the effect of sigmoid and fuzzy type-2 membership functions to increase the specificity, specificity and accuracy of detecting low anomalies.

Acknowledgements

The authors would like to thank Institut Teknologi Sepuluh Nopember and the Ministry of Research, Technology and Higher Education, Indonesia for supporting this research. The authors also appreciate the supportive comments and suggestions by the reviewers for improving this paper.

References

Amara, I., Amar, A.B., Jarboui, A., 2013. Detection of Fraud in Financial Statements: French Companies as a Case Study. Int. J. Acad. Res. Account. Fin. Manage. Sci. 3 (3), 44–55. https://doi.org/10.6007/IJARAFMS/v3-i3/34.
Rahmah, G., Pungli Pelabuhan Samarinda, Polri: Komura Terima Rp 2 Triliun.

Huda, S., Sarno, R., Ahmad, T., 2016. Increasing accuracy of process-based fraud detection using a behavior model. Int. J. Software Eng. Appl. 10(5), 175–188. https://doi.org/10.14257/ijseia.2016.10.5.16.

Huda, S., Sarno, R., Ahmad, T., 2015. Fuzzy MADM Approach for Rating of Process-Based Fraud. J. ICT Res. Appl. 9(2), 111–128. https://doi.org/10.5614/itbj.ict.res.appl.2015.9.2.1.

Correia, A.C., 2012. Workflow Patterns Group. Encyclopedia of Database Syst., 1–12 https://doi.org/10.1007/springerreference_63860.

Cordella, L., Foggia, P., Sansone, C., Vento, M., 2004. A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs. IEEE Trans. Pattern Anal. Mach. Intell. 26 (10), 1367–1372. https://doi.org/10.1109/TPAMI.2004.75.

Nabti, C., Seba, H., 2016. Subgraph Isomorphism Search in Massive Graph Databases. In: Proceedings of the International Conference on Internet of Things and Big Data, pp. 204–213. https://doi.org/10.5220/0005875002040213.

Buijs, J.C., Dongen, B.F.V., Aalst, W.M.V.D., 2012. On the role of fitness, precision, generalization and simplicity in process discovery, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7565 LNCS (Part 1), 305–322. https://doi.org/10.1007/978-3-642-33606-5_19.

Liu, G., Yan, Y., Wang, H., 2013. Robust modular linear regression based classification for face recognition with occlusion. In: Proceedings 2013 7th International Conference on Image and Graphics, ICIG 2013, pp. 509–514. URL 10.1109/ICIG.2013.108

Osojnik, A., Panov, P.S.D., 2017. Multi-label classification via multi-target regression on data streams. Mach. Learn. 106(6), pp. 745–770. https://doi.org/10.1007/s10994-016-5613-5.

Shapiro, A.F., 2005. Fuzzy regression models.

Chandola, V., Banerjee, A., Kumar, V., Jul. 2009. Anomaly detection. ACM Comput. Surv. 41 (3), 1–58.

Saia, R., Carta, S., 2017. A Frequency-domain-based Pattern Mining for Credit Card Fraud Detection. Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security.

Mahapatra, S., Chandola, V., 2018. Learning Manifolds from Non-stationary Streaming Data, arXiv preprint arXiv:1804.08833.

Saia, R., 2017. A Discrete Wavelet Transform Approach to Fraud Detection. Springer, Cham.

Akoglu, L., Faloutsos, C., 2013. Anomaly, event, and fraud detection in large network datasets. Proceedings of the sixth ACM international conference on Web search and data mining. ACM.

DeForest, D.K., Brix, K.V., Tear, L.M., Adams, W.J., 2017. Multiple Linear Regression (MLR) Models for Predicting Chronic Aluminum Toxicity to Freshwater Aquatic Organisms and Developing Water Quality Guidelines. Environ. Toxicol. Chem. 37, 80–90. https://doi.org/10.1002/etc.3922.

Estelles-Lopez, L., Ropodi, A., Pavlidis, D., Fotopoulou, J., Gkousari, C., Peyrodie, A., Panagou, E., Nychas, G.J., Mohareb, F., 2017. An automated ranking platform for machine learning regression models for meat spoilage prediction using multi-spectral imaging and metabolic profiling. Food Res. Int. 99, 206–215. https://doi.org/10.1016/j.foodres.2017.05.013.

Uejima, S.T.H., Asai, K., 1982. Linear Regression Analysis with Fuzzy Model. IEEE Trans. Syst. Man Cybern. 12(6), 903–907. URL 10.1109/TSMC.1982.4308925.

Sánchez, J.D.A., Gómez, A.T., 2004. Estimating a fuzzy term structure of interest rates using fuzzy regression techniques. Eur. J. Oper. Res. 154, 313–331. https://doi.org/10.1016/S0377-2217(02)00854-8.

Andrés-Sánchez, J.D., 2016. Fuzzy regression analysis: an actuarial perspective. Stud. Fuzziness Soft Comput. 343, 175–201. https://doi.org/10.1007/978-3-319-39014-7_11.

Bell, P.M., Heng, W., 1997. Fuzzy linear regression models for assessing risks of cumulative trauma disorders. Fuzzy Sets Syst. 92 (3), 317–340. https://doi.org/10.1016/s0165-0114(96)00178-9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, Édouard, 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. 12, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.

Rahmawati, D., Sarno, R., Fatichah, C., Sunaryono, D., Oct. 2017. Fraud detection on event log of bank financial credit business process using Hidden Markov Model algorithm. In: 2017 3rd International Conference on Science in Information Technology (ICSITech). https://doi.org/10.1109/icsitech.2017.8257082.