Sarno *et al. J Big Data* (2020) 7:5
https://doi.org/10.1186/s40537-019-0277-1

Journal of Big Data

RESEARCH Open Access

Anomaly detection in business processes using process mining and fuzzy association rule learning

Riyanarto Sarno*, Fernandes Sinaga and Kelly Rossa Sungkono

*Correspondence:
riyanarto@if.its.ac.id
Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Abstract

Much corporate organization nowadays implement enterprise resource planning (ERP) to manage their business processes. Because the processes run continuously, ERP produces a massive log of processes. Manual observation will have difficulty monitoring the enormous log, especially detecting anomalies. It needs the method that can detect anomalies in the large log. This paper proposes the integration of process mining, fuzzy multi-attribute decision making and fuzzy association rule learning to detect anomalies. Process mining analyses the conformance between recorded event logs and standard operating procedures. The fuzzy multi-attribute decision making is applied to determine the anomaly rates. Finally, the fuzzy association rule learning develops association rules that will be employed to detect anomalies. The results of our experiment showed that the accuracy of the association rule learning method was 0.975 with a minimum confidence level of 0.9 and that the accuracy of the fuzzy association rule learning method was 0.925 with a minimum confidence level of 0.3. Therefore, the fuzzy association rule learning method can detect fraud at low confidence levels.

Keywords: Process mining, Anomaly detection, Fuzzy association rule learning

Introduction

Many corporations worldwide use an enterprise resource planning (ERP) system to manage their business process, which continuously changes due to dynamic business requirements [1]. Because the processes run continuously, ERP produces a considerable log of processes. Manual observation will have difficulty monitoring the sizeable log, especially detecting anomalies. It needs the method that can detect anomalies in the huge log.

Standard business processes are usually incorporated into standard operating procedures (SOP), which are used as a reference to find any deviations. Deviations or anomalies in the business process can be caused by variations or operation errors [2]; however, some of the anomalies may be the result of fraudulent behaviours [3]. Fraud can be committed in many ways and can lead to significant losses. In 2012, the Association of Certified Fraud Examiners (ACFE) reported that there had been 1.388 fraud cases in 96 countries, which have incurred US$1.4 billion in losses [4]. On average, organizations

have lost a gross profit of 7% per year to fraud [5]. This forces companies to instate strong security policies and an information system for fraud detection.

Analysis in the domain of process mining and data mining provides solutions for anomaly detection, which can be used for fraud detection. In previous research, we have investigated process mining for minimizing internal fraud in business processes [6]. In this research, several process mining methods were applied, such as conformance checker, dotted-chart analysis, social network miner, originator by task matrix and others, to investigate event logs of business processes [7–9].

Data mining analyses input data to construct a model or a pattern as output, which can be used to detect anomalies in the process under examination [10]. Several methods of data mining, such as decision tree, neural network, bayesian network and support vector machine have been implemented in previous researches [10–13] to identify cases of fraud. However, these methods still have weaknesses in detecting fraud since they are not able to analyse the behaviour of control flow in the business process. Another research supporting fraud detection used association rule learning (ARL) to extract association rules from transaction data, where ARL was applied to develop association rules related to fraudulent behaviours [14, 15].

In our previous research [6], only fraud with a high confidence level could be detected. In this paper, a method is proposed that can detect fraud with a low confidence level and a low intensity based on a certain threshold. The proposed method integrates process mining, fuzzy multi attribute decision making and fuzzy association rule learning to detect anomalies in a business process.

Related work

Types of Fraud

Fraud is a misuse of an organizational system [16]. The concept of the fraud triangle explains that frauds occur because of three things, i.e. pressure or coercion, opportunity and rationalization. When attempting to detect fraud in a business process, internal control can be used as a counter measure towards fraud that may occur [17]. The SOP for a business process should include a standard business process model, time record, resource, organization role and decision-making. The complete SOP can be used as a reference to detect anomalies in a running process and existing data that may contain fraud. Analyzing the anomalies in a business process can be done using process mining techniques [18–22].

Process-based fraud (PBF) refers to fraud occurring in business processes [6]. In a previous research concerning PBF, we have identified attributes and patterns in order to describe PBF [15]. The following six types of anomaly attributes or fraudulent behaviours in business processes can be distinguished.

Skipped activity

As its name implies, an anomaly is an activity that is skipped according to the SOP. Skipped Activity can be divided into two types, i.e. skipped sequence, for a skipped activity occurring in a sequence, and skipped decision, for a skipped activity occurring in a split decision activity.

Wrong throughput time

Wrong throughput time is a condition when an activity is performed faster or slower than the time limit as stated in the SOP. It is divided into two types: wrong throughput time min and wrong throughput time max.

Wrong resource

Wrong resource is a situation when an activity is not executed by an authorized employee in accordance with his or her role allowed by the SOP.

Wrong duty

Wrong duty is a condition when an employee performs two or more different activities in one running process. This type is divided into three types: Wrong duty sequence (occurring in sequence activity), wrong duty decision (occurring in decision activity) and wrong duty combine (occurring in sequence and decision activity).

Wrong pattern

Wrong pattern is a situation when a wrong activity sequence occurs that does not conform with the sequence of activities as stated in the standard business process.

Wrong decision

Wrong decision is a condition when a decision is made that does not conform to the decision-making process stated in the SOP.

To detect the anomaly attributes, four process mining analyses can be executed: control flow analysis, role resource analysis, throughput time analysis, and decision point analysis. Control flow analysis can be done using a manual analysis or with assistance of plug-ins in ProM. This analysis is crucial for the detection of fraud in the form of skipped activity and wrong pattern. Manually, the analysis is done by searching the event log using a process searching algorithm. fuzzy miner, which compares a fuzzy model to the standard business process, is the algorithm recommended for this searching process. However, this algorithm has a limitation in that it is relies heavily on the determination of the threshold value [23–25]. In addition to the manual method, this analysis can be done using the conformance checking plug-in in ProM, resulting in values for fitness, precision, and structure. These values can be used to measure the equality between a running process and a standard business process. The purpose of the control flow analysis is to measure the equality and difference between event logs resulted from a running process and a standard business process model. In this case, different parts in the running process can be suspected as anomalous. The result is in the form of fitness values revealing fraudulent behaviour.

Role resource analysis can be performed using the social network miner plug-in in ProM. Then, the role attribute of each event in a running process can be compared to the roles present in the SOP to obtain the probability of fraud occurrence in terms of its resources.

Throughput time analysis can be done by measuring the time interval between activities. This interval is measured from the start time stamp to the completed time stamp. The time of implementation of an activity can then be compared to the SOP in terms of the application time.

Decision point analysis is done by finding out the existence of a specific case as a result of decision-making in a business process. Detecting anomalies can be done by building a relational database and do a query for that specific case.

Case study and fraudulent issues

In this paper, we would like to provide an example of the occurrence of fraud in the credit application process in a bank as depicted in Fig. 1. The first executed activity is receive application, i.e. the activity of receiving the credit application. The data received are in the form of a rules and regulations document required for credit applications.

Check completeness is the activity of checking the completeness of the rules and regulations document provided by the creditor. If the rules and regulations document is not complete, the give info activity will be executed to give information to the creditor in order for him or her to complete the document. This activity can be executed repeatedly until the document is fully completed. Once it is completed the process moves on to the check SID activity.

The check SID activity is done by the system to check the credit application history of the creditor. If a creditor has ever submitted a credit application, the process moves on to loan decision and check loan type. Check loan type, collateral verification locate, collateral local Government, collateral office and complete verification are done to check the collateral owned by the creditor in accordance with the credit type proposed. Plafond estimation is used to estimate the amount of disbursed credit; this depends on the collateral check.

The check overrate activity is used to determine the further activities that need to be executed, in this case to determine whether the decision-making is executed by the director or to the leader. This depends on the amount of the credit applied. Further, loan decision is an activity that of division of the bank but in fact was done by a staff member. Another example is document validation, which is supposed to be done within 1 week but in fact was done after more than 1 week. Another example is in an activity that issues a branching output to more than one activity, such as in check overrate, where a

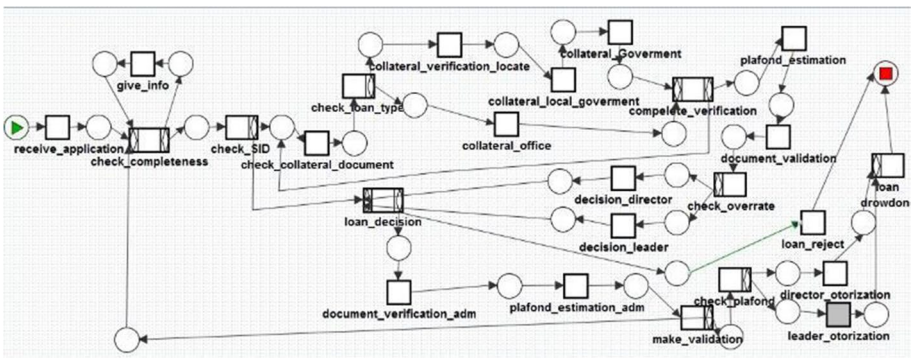

Fig. 1 Effects of selecting different switching under dynamic condition

credit application with an amount of over 500 million Rupiah should not be executed through director authorization but through leader authorization.

Proposed method

The training section in the proposed method is implemented in three steps, which are conformance checking, fuzzy multi attribute decision making and fuzzy association rule learning. The conformance checking, which is part of process mining, is applied to detect anomalies in the process business. The fuzzy multi attribute decision making is used to determine the anomaly rates. Finally, the fuzzy association rule learning develops rules which will be used to detect anomalies in the testing phase.

Skipped activity analysis

Activities that were wrongly executed may emerge in the event log, e.g. a skipped activity. This will lead to the presence of an anomaly in that activity. For example, the completeness verification is supposed to be done by the head out whether there is a skipped activity or any other activity not in line with the standard business process model. This analysis is done with the conformance checker plug-in in ProM, which was modified to give the number of skipped activities. The input of this analysis is the standard business process in the form of Petri nets and event logs. This analysis generates anomaly data for the skipped sequence and skipped decision attributes.

Wrong pattern analysis

In this part, pattern analysis of the event logs is done by comparing the sequence of activities to the standard business process model. If there is a case with an activity that is not done in line with the model, it will be marked as wrong pattern. This analysis is done with the conformance checker plug-in in ProM.

Wrong throughput time analysis

In this part, an analysis of the execution times of all activities in the event logs is made by comparing them with the execution time in the standard model. If the execution time is not in line with the standard model, being either too short or too long, it will be marked as an anomaly. This analysis is done with the conformance checker for attributes plug-in in ProM.

Wrong resource analysis

The analysis in this part is done for each actor who executes an activity recorded in the event logs using the conformance checker for attributes plug-in in ProM. If there is an activity that is executed by an unauthorized actor towards that activity according to the standard model, it will be marked as wrong resource.

Wrong duty analysis

The analysis in this part is done to see whether there is an actor who violates the segregation of duty as defined in the standard model. Anomalies have the form of two or more activities conducted by one actor at once. This analysis is done using the conformance checker for attributes plug-in in ProM.

Wrong decision analysis

In this part, activities involving decision-making or event branching are analysed. To conduct a wrong decision analysis, the event logs should first be changed into ontology-based event logs to facilitate doing a SPARQL ontology query. This analysis generates anomaly data for the wrong decision attribute.

All obtained anomalies are trained using fuzzy association rule learning. This procedure consists of two processes. It starts with the calculation of the anomaly rates for each case. This process is done by using fuzzy multi attribute decision making. The inputs for this process are the anomaly occurrences and expert assessments. There are two kinds of values that are calculated in this process. First, the importance weight of the anomaly attributes. This value shows the importance of the anomaly attributes according to assessment of experts. Second, the anomaly attribute occurrence rate. This value shows the occurrence rate of each anomaly attribute. Then, with these two values, all cases with an anomaly rate are trained using fuzzy association rule Learning. This process generates the association rules among the anomalous attributes.

Fuzzy multi attribute decision making

This method is used to determine the anomaly rates from a set of anomalies that occurred in a process. The determination of the anomaly rates is done using a combination of two concepts, i.e. fuzzification and multiple attribute decision making (MADM). MADM can be used to select one alternative out of a set of alternatives marked with several attributes [24]. However, MADM still has weaknesses in handling inaccurate or linguistic information. Hence, it is necessary to add fuzziness to support the handling of linguistic information.

Two data are required for determining the anomaly rates, i.e. an importance assessment of the PBF attributes by experts and the anomaly occurrences resulted from the conformance checking. Both data are converted into fuzzy numbers based on the table of importance weights and the level of membership. The importance weights of the anomaly attributes assessed by experts can be seen in Table 1, which uses the sample from the function of the anomaly attribute membership can be seen in Fig. 2.

The expert assessments used in this research are based on the expertise of an auditor of a bank, which are shown in Table 2. Based on Table 2, a weight measurement is done for each anomaly attribute. The weight value is divided into four parts: lower bound weight, middle weight 1, middle weight 2, and upper bound weight. Measuring the 4 weight values for each category can be done using Eqs. 1, 2, 3 and 4 where n is the number of experts and values

Table 1 Importance level of anomaly attributes

Levels	Fuzzy parameters				Scales
	a	b	c	d	
Very important (VI)	0.7	1	1	1	100%–70%
Important (I)	0.5	0.7	0.7	1	100%–50%
Fair (F)	0.2	0.5	0.5	0.8	80%–20%
Weak (W)	0	0.3	0.3	0.5	0%–50%
Very weak (VW)	0	0	0	0.3	0%–30%

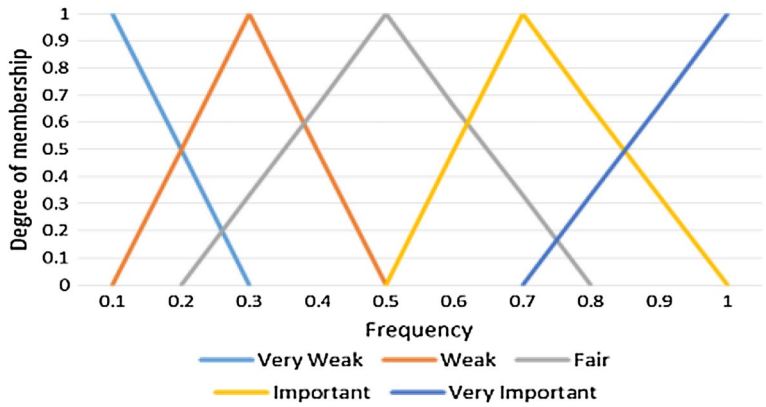

Fig. 2 Membership functions for levels of anomaly attributes

Table 2 Evaluation for each anomaly attribute

Anomaly attributes	Expert 1	Expert 2	Expert 3	Expert 4
Skipped				
Sequence	W	W	W	W
Decision	VI	VI	VI	VI
Wrong throughput time				
Min	F	I	I	I
Max	F	F	I	F
Wrong resource	W	W	W	W
Wrong duty	VW	VW	VW	VW
Sequence	W	W	W	W
Decision	W	W	W	W
Combine	W	W	W	W
Wrong pattern	W	W	W	W
Wrong decision	VI	VI	VI	VI

a, b, c and d are the values of vectors a, b, c and d in Table 1. In Table 1, the interval between 0 and 1 is divided accordingly into five categories to determine the membership function parameters a, b, c, and d. The result of the measurement can be seen in Table 3.

Furthermore, for the process anomaly data, the anomaly rates of the attribute occurrences are shown in Table 4 and their membership functions are shown in Fig. 3. The interval between 0 and 1 is divided accordingly into 9 categories to determine the membership function parameters a, b, c, and d. The formula for the membership functions are in Eq. 1 until Eq. 4 and the parameters of the membership functions are described in Table 5. In this formula, the value of x refers to the percentage of anomaly occurrences that is obtained from the calculation in Eq. 5.

$$\text{Lower bound} = \frac{\sum_{k=1}^{n} a_k}{n} = \frac{0+0+0+0}{4} = 0 \tag{1}$$

$$\text{Middle weight 1} = \frac{\sum_{k=1}^{n} b_k}{n} = \frac{0.3+0.3+0.3+0.3}{4} = 0.3 \tag{2}$$

Table 3 Weights of anomaly attributes

Anomaly attributes	L	M1	M2	U
Skipped				
Sequence	0	0.3	0.3	0.5
Decision	0.7	1	1	1
Wrong throughput time				
Min	0.425	0.65	0.65	0.95
Max	0.275	0.55	0.55	0.85
Wrong resource	0	0.3	0.3	0.5
Wrong duty				
Sequence	0	0	0	0.3
Decision	0	0.3	0.3	0.5
Combine	0	0.3	0.3	0.5
Wrong pattern	0	0.3	0.3	0.5
Wrong decision	0.7	1	1	1

where *L* Lower, *M1* Middle1, *M2* Middle2, *U* Upper

Table 4 Linguistic values of occurrence

Attribute	a	b	c	d
Very week (VW)	0	0	0.1	0.2
Between very weak and weak (BVW and W)	0	0.1	0.2	0.3
Weak (W)	0.1	0.2	0.3	0.4
Between weak and fair (BW and F)	0.2	0.3	0.4	0.5
Fair (F)	0.3	0.4	0.5	0.6
Between fair and strong (BF and S)	0.4	0.5	0.6	0.7
Strong (S)	0.5	0.6	0.7	0.8
Between strong and very strong (BS and VS)	0.7	0.8	0.9	1
Very strong (VS)	0.8	0.9	1	1

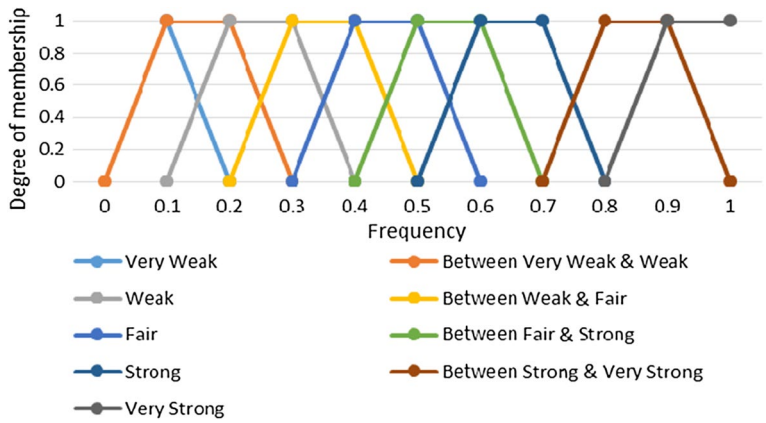

Fig. 3 Membership functions for linguistic values of occurrence

$$\text{Middle weight 2} = \frac{\sum_{k=1}^{n} c_k}{n} = \frac{0.3 + 0.3 + 0.3 + 0.3}{4} = 0.3 \tag{3}$$

$$\text{Upper bound} = \frac{\sum_{k=1}^{n} d_k}{n} = \frac{0.5 + 0.5 + 0.5 + 0.5}{4} = 0.5 \tag{4}$$

$$x = \frac{\text{Number of anomaly occurences for an attribute}}{\text{Maximum number of activities for an attribute}} \tag{5}$$

$$P = \frac{\left(\frac{C_a + C_b + C_c + C_d}{4}\right) + \left(\frac{D_a + D_b + D_c + D_d}{4}\right)}{2} \tag{6}$$

A probability measurement of the anomalies is done for each anomaly attribute. The result of this measurement is used in the phase of fuzzification into the membership class in Table 4. The next step is to calculate the anomaly attribute occurrence rate (anomaly attribute value). The data of the anomaly occurrences are obtained from the result of the conformance check. To calculate the occurrence rate of each anomaly attribute, the value from the conformance result and the importance weight from the experts adjusted with the resulted conformance are used. Each expert assessment is filled with the anomaly attribute occurrence rate adjusted with the value of the expert assessment and the conformance value. This adjustment is done with Eq. 6.

In Eq. 6, Ca, Cb, Cc and Cd are the values of vectors a, b, c and d in Table 4 according to the value of fuzzification of its anomaly. In addition, Da, Db, Dc and Dd are the values of vectors a, b, c and d in Table 1 in accordance with the importance weight given by the experts. The P value is the fuzzification into the anomalous class in accordance with the membership function of the anomaly attribute occurrence rate. For example, the result of the conformance check for the skipped sequence category is between very bad and bad, and the assessment of the first expert for the Equal category is weak. Thus, the adjustment value is given by Eq. 7. Furthermore, the P value = 0.2125 is the fuzzification into the anomaly class in accordance with the membership function of bad. The result of the adjustment from the example can be seen in Table 6.

Table 5 Parameters of membership functions for linguistic values of occurence

Membership function of VW		Membership function of BVW and W, W, BW and F, F, BF and S, S, BS and VS		Membership function of VS	
Degree	**Condition**	**Degree**	**Condition**	**Degree**	**Condition**
1	$a \leq x \leq c$	0	$x \leq a$	0	$x \leq a$
$(d-x)/(d-c)$	$c < x < d$	$(x-a)/(b-a)$	$a < x < b$	$(x-a)/(b-a)$	$a < x < b$
0	$x \geq d \,\|\|\, x < a$	1	$b \leq x \leq c$	1	$x \geq b$
		$(d-x)/(d-c)$	$c < x < d$		
		0	$x \geq d$		

Table 6 Assesment of occurrences for a case

Anomaly attributes	Expert 1	Expert 2	Expert 3	Expert 4	Linguistic occurences
Skipped					
Sequence	W	W	W	W	BVW and W
Decision	0	0	0	0	0
Wrong throughput time					
Min	0	0	0	0	0
Max	0	0	0	0	0
Wrong resource	0	0	0	0	0
Wrong duty	0	0	0	0	0
Sequence	0	0	0	0	0
Decision	0	0	0	0	0
Combine	0	0	0	0	0
Wrong pattern	BVW and W	BVW and W	BVW and W	BVW and W	VW
Wrong decision	0	0	0	0	0

Table 7 Evaluation of anomaly attribute occurrence rate for a case (S)

Anomaly attributes	L	M1	M2	U	L
Skipped					
Sequence	0.1	0.2	0.3	0.4	0.1
Decision	0	0	0	0	0
Wrong throughput time					
Min	0	0	0	0	0
Max	0	0	0	0	0
Wrong resource	0	0	0	0	
Wrong duty					
Sequence	0	0	0	0	0
Decision	0	0	0	0	0
Combine	0	0	0	0	0
Wrong pattern	0	0.1	0.2	0.3	0
Wrong decision	0	0	0	0	0

After the anomaly value has been adjusted with the importance assessment of the experts, the next step is calculating the evaluation of the anomaly attribute occurrences by using the same equations as for calculating the importance weight, i.e. Eqs 1, 2, 3 and 4, where n = the number of the expert + 1. Further, from the result of the calculation of the importance weight of the anomaly attributes and the anomaly attribute occurrence rate, calculation of the final rating is done to get the weight of lower bound, middle 1, middle 2, and upper bound by using Eq. 8.

$$\mathrm{P} = \frac{\left(\frac{0+0.1+0.2+0.3}{4}\right) + \left(\frac{0+0.3+0.3+0.5}{4}\right)}{2} = 0.2125 \tag{7}$$

$$Final_{Rating} = \frac{1}{k} x[(S_{c1} x W_{c1}) + \cdots + (S_{cn} x W_{cn})] \tag{8}$$

Table 8 Example of results of anomaly rate calculation

Anomaly attributes	Cases				
	1	2	3	4	5
Skipped					
Sequence	2	0	0	0	0
Decision	0	2	0	0	0
Wrong throughput time					
Min	0	0	4	0	0
Max	0	0	0	5	0
Wrong resource	0	0	0	0	
Wrong duty					
Sequence	0	0	0	0	
Decision	0	0	0	0	
Combine	0	0	0	0	0
Wrong pattern	0	2	2	0	0
Wrong decision	0	0	2	0	0
Rates of anomaly	0.055	0.393	0.107	0.086	0.033

Table 9 Parameter of membership function for percentage of anomaly occurrence

Membership function of low		Membership function of middle		Membership function of high	
Degree	Condition	Degree	Condition	Degree	Condition
0	$x < a$	0	$x < a$	0	$x < a$
$(x-a)/(b-a)$	$a \leq x < b$	$(x-a)/(b-a)$	$a \leq x < b$	$(x-a)/(b-a)$	$a \leq x < b$
1	$b \leq x <= c$	1	$b \leq x \leq c$	1	$x \geq b$
$(d-x)/(d-c)$	$c < x < d$	$(d-x)/(d-c)$	$c < x < d$		
0	$x \geq d$	0	$x \geq d$		

$$\text{Rates of anomaly} = \text{Final rating of lower bound} + \text{Final rating of middle 1} + \text{Final rating of middle 2} + \text{Final rating of upper bound} \quad (9)$$

where k refers to the number of categories, S refers to the value of the anomaly attributes in Table 7, W refers to the importance weight of the anomaly attributes in Table 3 and Cn refers to the n anomaly attributes. After the four weights of the final rating have been calculated, the anomaly rate of a case is calculated as the summation of all four weights of the final rating. A further elaboration can be found in Eq. 9. An example of the result of calculating the anomaly rates can be seen in Table 8. The calculation of the anomaly rates is performed for each process being examined.

Fuzzy association rule learning

Fuzzy association rule learning is the method for seeking the association rules between the occurred anomalies. The processed data are the anomalies that have occurrence rates for each process. The probability of the anomalies are calculated using Eq. 5. The probability values for each anomaly attribute are used to calculate the membership degree using parameters in Table 9.

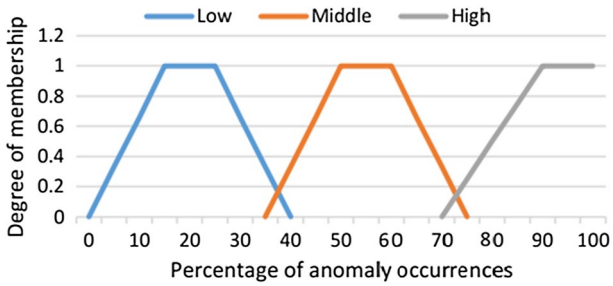

Fig. 4 Membership functions for percentage of anomaly occurrences

In seeking the association rule, each anomaly attribute is divided into three types: low, middle and high. Thus, the membership function used in the process of searching the association rule is divided into three types by dividing the attribute of skipped sequence into skipped sequence low, skipped sequence middle, and skipped sequence high. Its membership function is shown in Fig. 4, generated by using the parameter of its membership as shown in Table 9. In Fig. 4, the interval between 0 and 100% is divided accordingly into three categories (low, middle and high) to determine the membership function parameters a, b, c, and d.

$$\text{Support } (X, Y) = \left(\frac{\text{Transaction number contains X and Y}}{\text{Transaction number}} \right) \tag{10}$$

$$\text{Confidence } (X, Y) = \left(\frac{\text{Transaction number contains X and Y}}{\text{Transaction number contains X}} \right) \tag{11}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

The next step is to calculate the membership degree of each anomaly attribute towards the three types; low, middle and high. The value of the membership degree for each anomaly attribute that has been divided into three classes is used to seek the degree of association using fuzzy ARL.

n this research, the search of the association rules applied the a priori algorithm using the fuzzy data. The calculation of the support value is done using Eq. 10. Further, for each n-item set, the threshold value is determined to select the candidate of the item for the next items. The author determined the threshold value for each item set by considering the support value for each item. In the experiment, the threshold for 1 item set was determined as 0.01 and for the next n-item set as 0.005. The selected items are the items that have a higher support value or a support value equal to the determined threshold value. The chosen items are combined with the other items to make new item sets. The combination is done until no more combination between items can be made, or it can also be done by limiting the formation of item sets to k-item sets. In the experiment, the process of searching the association rule was limited to 5-item sets. A 5-item set means that 1 item set consists of a combination of 5 essential items.

The association rule taken as a result of this final task consists of the anomaly attributes in combination with the fraud attribute for each n-item set. The value of each association rule is the confidence value for the control calculated with Eq. 11. The confidence value shows to what extent an anomaly contributes to the fraud level. The result of this method is a set of association rules that describe a relationship between the anomaly attributes and the fraud used as data to detect fraud in other processes. Thus, to detect fraud in the different processes is done by matching the anomaly attributes with the existing association rules. A sample of the association rule result from the training data is shown in Table 10.

Detecting fraud with association rule data

This method is used as a testing phase to detect fraud in a business process by matching the anomaly attributes with the resulted association rules. If a case contains an association rule not marked as fraud, the situation is directly defined and categorized as not fraud, even if it has a high confidence value. Meanwhile, if a case contains a fraud association, the case is classified as fraud.

Results and discussion

Evaluation design

The evaluation in this research focuses on the following points: (1) finding the advantage of using the proposed fuzzy association rule learning method compared to using the association rule learning (ARL) method in the context of fraud detection, and (2) measuring the accuracy of both methods. The scenario and dataset used in this evaluation were the same for both methods. The experiment was based on a case study of credit applications in a bank. The dataset consists of a training dataset and a testing dataset generated by several distribution models as provided in [6].

By the analyses that have been done, anomalies were modelled against attributes using a Poisson distribution with the parameter set to 3. This parameter indicates that on average, there are 3 unusual cases each month. A Poisson distribution was used because its characteristics are in line with business process fraud behaviour. The number of exceptional cases for each attribute was generated randomly based on the Poisson distribution. Therefore, each attribute had a different number of exceptional cases for each month.

Furthermore, 50 credit applications were processed each month. The anomalies were spread among all credit applications based on a uniform (discrete) distribution. The aim was to randomly spread the anomalies over 50 credit applications a month, based on the

Table 10 Example of association rules

ARL	Support	Confidence	ARL	Support	Confidence
SkipSL-Fraud	0.056	0.369	SkipSL-wDutySecL-Fraud	0.021	0.928
SkipDL-Fraud	0.075	0.623	SkipSL-wDutyDecL-Fraud	0.030	1.314
TminL-Fraud	0.047	0.409	SkipSL-SkipDL-TmaxLFraud	0.011	2
TmaxL-Fraud	0.056	0.404	SkipSL-SkipDLwResourceL-Fraud	0.006	1
SkipSL-wResourceL-Fraud	0.023	1.027	SkipSL-Fraud	0.056	0.369

Table 11 The occurrences of anomaly attributes generated by a Poisson distribution

Anomaly attributes	Occurrences
Skipped	
Sequence	1
Decision	1
Wrong throughput time	
Min	7
Max	5
Wrong resource	6
Wrong duty	
Sequence	5
Decision	1
Combine	5
Wrong pattern	3
Wrong decision	3

Table 12 The distribution of anomaly attributes among cases using a uniform (discrete) distribution

Case	Skipped		Wrong throughput time		Wrong resources	Wrong duty			Wrong pattern	Wrong decision
	Sequence	Decision	Min	Max		Sequence	Decision	Combine		
1	0	0	0	0	0	0	0	1	0	0
2	0	0	1	0	1	1	0	0	0	1
3	0	0	0	0	1	0	0	0	0	1
4	0	0	0	0	1	1	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	1	1	0	0	0	0
7	0	0	0	1	1	0	0	0	0	1
8	0	0	0	0	1	0	0	1	0	0
9	0	0	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0

Table 13 Expert assessment for determining anomaly for ARL and fuzzy ARL training dataset

No	Fraud	Non-fraud
1	Case that contains skipped activity	Case that does not contain skipped activity
2	Case that violates wrong decision	Case that does not violate wrong decision
3	Case that violates wrong resource and wrong duty simultaneously	Case that violates wrong resource but does not violates wrong duty or vice versa
4	Case that contains an anomaly	Case that does not contain any anomalies
5	Case that violates more than one attribute	Case that violates maximum one attribute
6	Case with weight of fraud higher than or equal to 0.3	Case with weight of fraud lower than 0.3

number of anomaly occurrences for each attribute. An example of the generated data can be seen in Tables 11 and 12.

In all, 1200 cases were produced as experimental data. The experimental data were then divided into training data and testing data. In the training data, there were 1000 cases, consisting of 538 fraud cases and 462 legal cases. In the testing the data, there were 200 cases, consisting of 93 fraud cases and 107 legal cases.

To get the anomaly data from the event logs, we develop conformance checking for attributes plug-in. Next, the anomaly data were trained with an additional plug-into perform the ARL and fuzzy ARL methods (the author also developed this other ProM plug-in).

To compare the ARL and fuzzy ARL methods, training with ARL and training with fuzzy ARL used the same data; hence, the testing with ARL and fuzzy ARL also used the same data set. In the training process, expert opinion was used to define whether a case was fraud or legitimate. The specialist advice for ARL training and fuzzy ARL training is shown in Table 13.

Table 14 Item sets and the corresponding association rules resulted by the ARL training

Combination	Association rules	S	C
2-item set	Skip sequence-fraud	34	0.971
	Skip decision-fraud	40	0.976
	Throughput min-fraud	216	0.727
	Throughput max-fraud	166	0.738
	Wrong resource-fraud	196	0.754
	Wrong duty sequence-fraud	160	0.724
	Wrong duty decision-fraud	33	0.825
	Wrong duty combine-fraud	152	0.714
	Wrong pattern-fraud	87	0.696
	Wrong decision-fraud	118	1
3-item set	Skip sequence-throughput min-fraud	12	0.923
	Skip sequence-wrong duty sequence-fraud	12	1
	Skip sequence-wrong duty combine-fraud	14	1
	Skip decision-throughput min-fraud	12	1
	Skip decision-throughput max-fraud	10	0.909
	Skip decision-wrong duty combine-fraud	11	1
	Throughput min-throughput max-fraud	66	0.917
4-item set	Throughput min-wrong resource-wrong pattern-fraud	8	1
	Throughput min-wrong resource-wrong decision-fraud	4	1
	Throughput min-wrong duty sequence-wrong duty decision-fraud	3	1
	Throughput min-wrong duty sequence-wrong duty combine-fraud	10	0.833
	Throughput min-wrong duty sequence-wrong pattern-fraud	4	1
	Throughput min-wrong duty sequence-wrong Decision-fraud	6	1
5-item set	Throughput min-throughput max-wrong Resource-wrong duty sequence-fraud	6	1
	Throughput min-throughput max-wrong Resource-wrong duty combine-fraud	2	1
	Throughput min-throughput max-wrong duty Sequence-wrong duty combine-fraud	1	1
	Throughput min-wrong resource-wrong duty Sequence-wrong duty combine-fraud	2	1
	Throughput max-wrong resource-wrong Duty Sequence-wrong duty combine-fraud	3	1

Table 15 Item sets and the corresponding association rules resulted by the fuzzy ARL training

Combination	Association rules	S	C
2-item set	Skip sequence low-fraud	2.229	0.162
	Skip sequence medium-fraud	2.446	0.196
	Skip decision low-fraud	4.032	0.356
	Skip decision medium-fraud	5.542	0.504
	Throughput min low-fraud	24.023	0.265
	Throughput max medium-fraud	21.568	0.320
	Throughput max high-fraud	9.459	0.544
	Wrong resource low-fraud	16.960	0.230
	Wrong resource medium-fraud	17.652	0.207
	Wrong resource high-fraud	118	1
3-item set	Throughput min low-throughput max medium-fraud	2.719	0.445
	Throughput min low-wrong resource medium-fraud	4.157	0.483
	Throughput min low-wrong duty combine low-fraud	3.657	0.360
	Wrong resource low-wrong pattern low-fraud	1.306	0.214
	Wrong resource medium-wrong duty sequence low-fraud	1.612	0.213
	Wrong resource medium-wrong duty combine low-fraud	4.166	0.320
	Wrong duty sequence low-wrong duty combine low-fraud	3.120	0.290
4-item set	Throughput min low-wrong resource medium-wrong duty combine low-fraud	0.452	0.626
	Throughput min medium-throughput max low-wrong resource low-fraud	0.167	1
	Throughput min medium-throughput max low-wrong resource medium-fraud	0.444	1
	Throughput min medium-throughput max low-wrong duty sequence medium-fraud	0.560	0.577
	Throughput min medium-throughput max low-wrong duty combine low-fraud	0.605	0.473
	Throughput min medium-throughput max medium-Wrong resource low-fraud	0.615	0.481

where *S* Support, *C* Confidence

Table 16 Accuracy for the ARL Method

MC	TP	TN	FP	FN	A
0.1	93	31	76	0	0.62
0.2	93	31	76	0	0.62
0.3	93	31	76	0	0.62
0.4	93	31	76	0	0.62
0.5	93	31	76	0	0.62
0.6	93	31	76	0	0.62
0.7	93	31	76	0	0.62
0.8	93	86	32	0	0.90
0.9	88	107	0	5	0.98

where *MC* minimum confidence, *TP* true positive, *TN* true negative, *FP* false positive, *FN* false negative, *A* accuracy

The conducted training generated association rules between the anomalous attributes. The ARL method produced 95 standards, while the fuzzy ARL method produced 66 states. An example of the association rules provided by the ARL can be seen in Table 14 and of the fuzzy ARL in Table 15. S and C is Table 14 refer to support and confidence. The association rules contain rule, support value and confidence value. Further, the association rules, along with their confidence value, were used in the testing process to determine whether a case was fraud or legitimate.

Table 17 Accuracy for the fuzzy ARL Method

MC	TP	TN	FP	FN	A
0.1	93	31	76	0	0.62
0.2	93	31	76	0	0.62
0.3	93	31	76	0	0.62
0.4	93	31	76	0	0.62
0.5	93	31	76	0	0.62
0.6	93	31	76	0	0.62
0.7	93	31	76	0	0.62
0.8	93	86	32	0	0.90
0.9	88	107	0	5	0.98

Results

In measuring the accuracy of both methods, evaluation by receiver operating characteristic (ROC) framework analysis was done with Eq. 12. The results of the accuracy measurement for the ARL method are in Table 16 and for the fuzzy ARL method in Table 17.

From the results of the accuracy measurement, it can be seen that the accuracy of the ARL method is very high, reaching 0.975 at a minimum confidence value of 0.9. Meanwhile, the accuracy for the fuzzy ARL method reached 0.925 at a minimum confidence value of 0.3. Hence, the main difference between both methods lies in the minimum confidence level for each method. It can be said that the ARL method can detect fraud accurately if the confidence level of fraud is high, while the fuzzy ARL method can detect fraud accurately if the confidence level of fraud is lower. Hence, the fuzzy ARL method can be used to assist in identifying fraud with a lower confidence level of fraud than the ARL method. Furthermore, the accuracy of the fuzzy ARL method is better than that of the ARL method in previous related research [6] because the fuzzy ARL method can reduce the number of false positives.

Conclusion

According to the experimental results, it can be concluded that the integration of process mining with the ARL and fuzzy ARL method can be used to detect fraud in business processes. The process mining method can identify anomalies that occurred in a business process by doing conformance checking between the event logs and the SOP. The ARL method and fuzzy ARL method were trained using the same data to determine fraud in a running business process. The ARL method obtained an accuracy of 0.975 at a minimum confidence value of 0.9. This indicates that the ARL method can detect fraud accurately in cases with a high confidence level. The fuzzy ARL method, on the other hand, obtained an accuracy of 0.925 with a minimum confidence level of 0.3. This indicates that the fuzzy ARL method can detect fraud accurately at lower confidence levels. Thus, the fuzzy ARL method can be used to assist in identifying fraud in cases with a lower confidence level of fraud, so fraudulent claims with a little confidence level can be discovered more easily.

Abbreviations
A: accuracy; C: confidence; ARL: association rule learning; FP: false positive; FN: false negative; MC: minimum confidence; S: support; SOP: standard operating procedures; TP: true positive; TN: true negative.

Acknowledgements
Authors give a deep thank to Institut Teknologi Sepuluh Nopember, Program Bantuan Seminar Luar Negeri Ditjen Penguatan Riset dan Pengembangan, Kemenristekdikti, Direktorat Riset dan Pengabdian Masyarakat, and Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi dan Pendidikan Tinggi Republik Indonesia for supporting the research.

Authors' contributions
RS and FS discover anomaly detections by using process mining and fuzzy association rule learning. Then, KRS helps to compose the paper and do the experiment. All authors read and approved the final manuscript.

Funding
This research is partially funded by Indonesian Ministry of Technology and Higher Education under WCU Program, managed by Institut Teknologi Bandung.

Availability of data and materials
The used raw dataset in this research is not publicly available. Readers can contact the author if they want to access the data

Competing interests
The authors declare that they have no competing interests.

Received: 11 October 2019 Accepted: 23 December 2019
Published online: 09 January 2020

References

1. Sarno R, Djeni CA, Mukhlash I, Sunaryono D. Developing a workflow management system for enterprise resource planning. J Theor Appl Inf Technol. 2015;72:412–21.
2. Sarno R, Sari PLI, Ginardi H, Sunaryono D, Mukhlash I. Decision mining for multi choice workflow patterns. In: 2013 International conference on computer, control, informatics and its applications (IC3INA). p. 337–42.
3. Stoop J. A case study on the theoretical and practical value of using process mining for the detection of fraudulent behavior in the procurement process. Process mining and fraud detection. Netherlands: Twente University; 2012. p. 22–63.
4. Certified fraud examiners a. report to the nations on occupational fraud and abuse: 2016 global fraud study. association of certified fraud examiners 2016.
5. Goldmann P, Kaufman H. Anti-fraud risk and control workbook. Wiley online library; 2009.
6. Sarno R, Dewandono RD, Ahmad T, Naufal MF, Sinaga F. Hybrid association rule learning and process mining for fraud detection. Int J Comput Sci 2015;42(2):59–72.
7. Bernardi S, Alastuey RP, Trillo-Lado R. Using process mining and model-driven engineering to enhance security of web information systems. In 2017 IEEE European symposium on security and privacy workshops (EuroS&PW). IEEE; 2017, pp. 160–6.
8. Mans RS, van der Aalst WMP, Vanwersch RJB, Moleman AJ. Process mining in healthcare: data challenges when answering frequently posed questions. Process support and knowledge representation in health care. Berlin: Springer; 2012. p. 140–53.
9. Sarno R, Sungkono KR. A survey of graph-based algorithms for discovering business processes. Int J Adv Intell Inform. 2019;5:137–49.
10. Yee OS, Sagadevan S, Malim NH, Hassain A. Credit card fraud detection using machine learning as data mining technique. J Telecommun Electronic Comput Eng. 2018;10:23–7.
11. Tran PH, Tran KP, Huong TT, Heuchenne C, HienTran P, Le TMH. Real time data-driven approaches for credit card fraud detection. In: Proceedings of the 2018 international conference on E-Business and applications. Elsevier; 2018. p. 6–9.
12. Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: a comparative study. Decis Support Syst Elsevier. 2011;50:602–13.
13. Sungkono KR, Sarno R. Patterns of fraud detection using coupled Hidden Markov Model. 2017. In: 3rd international conference on science in information technology (ICSITech) [Internet]. Bandung: IEEE; 2017. p. 235–40. https://doi.org/10.1109/icsitech.2017.8257117.
14. Sánchez D, Vila MA, Cerda L, Serrano JM. Association rules applied to credit card fraud detection. Expert Syst Appl. 2009;36:3630–40.
15. Sarno R, Sinaga FP. Business Process anomaly detection using ontology-based process modelling and multi-level class association rule learning. In: International Conference on Computer, Control, Informatics and Its Applications (IC3INA). Bandung; 2015. p. 12–7.
16. Jans M, Lybaert N, Vanhoof K. Business process mining for internal fraud risk reduction: results of a case study. 2008
17. Jans M, Van Der Werf JM, Lybaert N, Vanhoof K. A business process mining application for internal transaction fraud mitigation. Expert Syst Appl Elsevier. 2011;38:13351–9.
18. Sungkono KR, Sarno R. Constructing control-flow patterns containing invisible task and non-free choice based on declarative model. In: International Journal of Innovative Computing, Information and Control (IJICIC). 2018; 14.

19. Sarno R, Sungkono KR, Johanes R, Sunaryono D. Graph-based algorithms for discovering a process model containing invisible tasks. Intell Netw Syst Soc. 2019;12:85–94.
20. Sarno R, Sungkono KR. Coupled Hidden Markov Model for process discovery of non-free choice and invisible prime tasks. Procedia Computer Science. Elsevier B.V.; 2018; 124:134–41.
21. Sungkono KR, Sarno R. CHMM for discovering intentional process model from event logs by considering sequence of activities. 2017. In: 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). Bandung: IEEE; 2017. p. 1–6.
22. Darmawan H, Sarno R, Ahmadiyah AS, Sungkono KR, Wahyuni CS. Anomaly detection based on control-flow pattern of parallel business processes. TELKOMNIKA. 2018;16:2808–15.
23. Folino F, Greco G, Guzzo A, Pontieri L. Mining usage scenarios in business processes: outlier-aware discovery and run-time prediction. Data Knowl Eng. 2010;70:1005–29.
24. Barreiros MP, Grillo A, Cruz-Machado V, Cabrita MR. Applying fuzzy sets for ERP Systems Selection within the Construction Industry. In: IEEE International conference on industrial engineering and engineering management (IEEM). 2010. p. 320–4.
25. Yang W-S, Hwang S-Y. A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl. 2006;31:56–68.

Publisher's Note