EASE 2012

16th International Conference on
Evaluation & Assessment in Software Engineering
Escuela Superior de Informática, Universidad de Castilla-La Mancha
Ciudad Real, 14-15 May, 2012
http://alarcos.esi.uclm.es/ease2012

Proceedings

Edited by:

Teresa Baldassarre
Marcela Genero
Emilia Mendes
Mario Piattini

Organised by:

In cooperation with:

Published by:

The Institution of Engineering and Technology

(ISBN 978-1-84919-541-6)

A study of the effectiveness of two threshold definition techniques

Laura Sánchez-González, Félix García, Francisco Ruiz
Instituto de Tecnologías y Sistemas de Información,
University of Castilla La Mancha,
Ciudad Real, España
{laura.sanchez | felix.garcia | Francisco.ruizg}@uclm.es

Jan Mendling
Wirtschaftsuniversität Wien
Augasse 2-6 1090
Vienna Austria
Jan.mendling@wu.ac.at

Abstract—**Background**: Measurement is a technique that is widely-used to quantify quality of process models. Evaluation of measurement results implies comparison against limit values, called thresholds. Determining thresholds is no trivial task and it requires the application of complex techniques. There are several techniques that have been published to date, proposing different approaches for threshold extraction. Two of the most prominent techniques are ROC curves and the Bender method. Although they come from different fields, both use logistic regression analysis as a discriminator function. **Aim**: For this reason, the main hypothesis is that thresholds obtained by both of those techniques are equally efficient in classifying the measurement results. **Method**: To check the hypothesis, we obtained thresholds for a group of empirically-validated measures for business process models, by applying both techniques. Then we checked the accuracy of the results**. Results**: The results indicate that the hypothesis should be rejected. **Conclusions:** ROC curves obtained more accurate thresholds for measurement evaluation.

Keywords- threshold, ROC curves, Bender method, Business Process models

I. Introduction

Measurement activities provide a good means for obtaining important information and for helping us to plan and track improvement efforts, communicate goals and convey reasons for improvements [1]. Improving efficiency in any organization generally requires better process control. Since a process is a complex entity which describes a lifecycle, some authors have affirmed that processes should be improved starting at the design stage, because “more than half the errors that occur during process developments are requirements errors” [2] and those errors are easier to eliminate early on than they are in post-implementation stages [3].

Even though the measurement of process models is considered to be very useful in obtaining information on potential improvement directions [1], the evaluation of measurement is no trivial task. Evaluation of measurement results implies “having an alarm which occurs whenever the value of the specific measure exceeded some predetermined value” [4]. This value is called a threshold.

Definition of thresholds requires a theory and practical base and it should meet certain requirements. It should fulfill the following conditions: it should not be based on expert opinion, but on measurement data, it should respect the statistical properties of the measure, such as measure scale and distribution and be resilient against outlier values, and finally, it ought to be repeatable, transparent and easy to carry out [5]. Some authors have worked on different techniques for threshold definition. However, a relevant number of authors used ROC curves [6-8] and the Bender method [9-13] for threshold determination. ROC curves is a technique from signal detection theory to select possible optimal models and to discard suboptimal ones independently of the cost context [14]. On the other hand, Bender [9] defined a method for quantitative risk assessment in epidemiological studies. Both methods follow a two-step approach: firstly, the estimation of the discriminator function and secondly, the determination of thresholds. The first step is the same for both methods: the discriminator function is based on the logistic regression; it should then be considered whether both techniques generate similar results or not. All this being so, we address the following research question in this paper: which threshold determination technique will obtain more accurate threshold values? To resolve this question, we have applied both techniques on previously-defined experimental data, in order to obtain a threshold for business process model measures. Threshold values are then validated using the recall and precision measures [15].

The remainder of this paper proceeds as follows. Section II provides the background of this research work by introducing business process model measures and techniques for threshold extraction. In Section III, the threshold determination techniques chosen are described in detail and then, in Section IV, these are applied to extract thresholds for business process measures. In Section V, the effectiveness of both techniques is checked and results obtained are discussed, along with threats to validity. Section VI concludes the paper with a summary and an outlook on future research.

II. BACKGROUND

A. *Business process model measures*

A systematic literature review concerning business process models was published in [16] and updated in [11]. In these documents, several measurement proposals for business process models were selected. These measures are about structural aspects of these models: for example, the control-flow complexity, or the number of elements of a specific design element. The most important aspect is that the measures should be supported by some kind of empirical validation, which makes them reliable and which facilitates the establishment of a more objective relationship between them and external quality characteristics. Some works about empirical validation of measures have been published in [11, 17, 18]. Although external quality of models can be discussed from different perspectives, most authors investigate understandability, since process models are typically used as a communication vehicle between stakeholders. Using the model therefore signifies being able to understand the semantics represented and to then adapt those to the new business requirements [19].

In the context of this paper, therefore, we use empirically-validated measures found in literature which have demonstrated their ability to predict the understandability of business process models. These measures are depicted in Table I. Although most of these measures can be applied on business process models independently of the notation, in this paper they are applied on models represented in BPMN [20]. Measurement results do not report significant benefits if they cannot be contrasted against limit values, however which is why thresholds have to be defined. The next section discusses techniques for threshold definition.

TABLE I. BUSINESS PROCESS MODEL MEASURES

Mendling [11, 21]	
Nº nodes: number of activities and routing elements	GM: sum of gateway pairs that do not match with each other
Diameter: the length of the longest path from a start to an end node	GH: different types of gateways that are used in the model
Density: ratio of the total nº of arcs	Sequentiality: degree to which the model is constructed out of pure sequences of tasks
AGD: average of the nº of incoming and outgoing arcs of gateways	Separability: nº of cut-vertex
MGD: maximum the nº of incoming and outgoing arcs of gateways	TS: max. nº of paths that may be concurrently activated
CNC: ratio of the total nº of arcs to its total nº of nodes	Cyclicity: nº of nodes in a cycle to the sum of all nodes
Rolón [17]	
TNSF: Total nº of sequence flows	NID: number of inclusive decisions
TNE: total nº of events	NPF: number of parallel forking
TNG: total nº of gateways	NP: number of pools
NSFE: nº of sequence flows from events	TNA: total number of activities
NMF: number of message flows	NCD: nº of complex decisions
NSFG: number of sequence flows from gateways	NEDDB: nº of exclusive gateways based on data
CLP: connectivity level between participants	NEDEB: nº of exclusive gateways based on events
NDO: number of data objects	
Cardoso [18, 22]	
CFC: control-flow complexity	

B. *Related works on threshold determination techniques*

Several proposals on threshold determination have been published to date. For example, Erni and Lewerentz [23] used mean and standard deviation to extract thresholds for software measures, specifically for class and method complexity, coupling and cohesion. French [24] also uses mean and standard deviation to extract thresholds for some software measures, but in addition using Checbyshev's inequality theorem. This technique requires data to follow a normal distribution, which is rarely applicable for model measure values, and it is sensitive to a large number of outliers. On the other hand, Benlarbi [25] defined thresholds for Chidamber and Kemerer measures using a linear regression analysis. There was no empirical evidence supporting that model, however. Rosenberg [26] extracted thresholds for object-oriented software measures in order to check the error-probability using histogram analysis, but there was no clear evidence of how these values are associated with error-probability. Other authors used techniques from the Artificial Intelligence field, for example Herbold et al. [27], who used a machine learning-based method, but this only produced a binary classification. Yoon et al. [28] used a k-means cluster algorithm, but it required an input parameter that affects both the performance and the accuracy of the results.

Several authors proposed different techniques for threshold determination, but a few of them agree on using ROC curves and the Bender method. For example, Shatnawi [10] used the Bender method for Chidamber and Kemerer measures. Likewise, Sanchez-Gonzalez et al. [11, 13], use the same method for business process model measures, in particular, for a group of structural complexity measures. Perez-Castillo et al. [12] also used the Bender method for measures related to business process mining. Others authors agree upon the use of ROC curves, including Shatnawi [6] and Catal et al. [8] works, in which ROC curves were used to obtain thresholds for software measures. Mendling et al. also use that technique to define thresholds for business process models [29]. Since both of these techniques are the two used most extensively in the literature; they are explained with more detail below.

III. THRESHOLD DETERMINATION TECHNIQUES

ROC curves and the Bender method involve a two-step approach. The first step is about estimating the discriminator function, and the second is the determination of thresholds. For both methods, the logistic regression is utilized to estimate a discriminator function. Logistic regression is a statistical model for estimating the probability of binary choices [30]. In this paper, the binary variable *understandability* can take the values of understandable/non-understandable. The idea of a logistic regression is that this probability can be represented by the odds. This is the ratio of considering the model as understandable, divided by the probability of considering it as non-understandable. The logistic regression estimates the odds based on the logit function, which is $Logit(p_i) = \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki}$ (1), where α is called the intercept and β_1, β_2, β_3 and so on, are called the regression coefficients of independent variables x_{1i}, x_{2i}, x_{3i} respectively. In our case, we will consider *k* business process model measures as input variables and observations from *i* business process models.

This part is common for both techniques: they require a logistic regression equation; in the case of the Bender method, to obtain α and β, which are needed to perform the operations; for ROC curves, it indicates the relationship between input and dependent variables. That means that ROC curves evaluate the ability of the logistic model to distinguish between the two states: understandable/ non-understandable. Having described the step in common, then, we will go on to explain how the technique obtains threshold values.

A. *The Bender method*

The Bender method assumes that the risk of an event happening is constant below a specific value (i.e. the threshold) and increases according to a logistic equation. By defining acceptable levels for the absolute risk, the corresponding benchmark values of the risk factor can be calculated by means of nonlinear functions of the logistic regression coefficients. Generally, a benchmark value is a characteristic point of the dose-response curve at which the risk of an event rises very steeply. The difficulty is to define what is meant by "very steeply". At first, a benchmark can be defined as the "Value of an Acceptable Risk Level" (VARL) defines as equation 2.

$$VARL = \frac{1}{\beta}\left(\ln\left(\frac{p_0}{1-p_0}\right)-\alpha\right) \qquad (2)$$

In Equation (2), p_0 is the probability of an event occurring. This value is indicated by the engineer who is applying that method and it can vary from 0 to 1. For example, $p_0 = 0.6$ indicates that there is a probability of 0.6 the measures to be considered as appropriate. On the other hand, α and β are coefficients of a logistic regression equation, as was indicated in (1). The independent variable in the logistic regression model is the measure or measures of which we want to determine the thresholds. The dependent variable must be a binary variable.

B. *ROC curves*

Receiver Operating Characteristics (ROC) curves provide a pure index of accuracy by demonstrating the limits of a test's ability to discriminate between alternative states [14]. For the definition of a ROC curve, we need two variables: one binary and another that is continuous. Each point in the ROC curve represents a pair of sensitivity and 1-specificity. In this way, it represents the classification performance of any potential threshold.

The test performance is assessed using the Area Under the ROC Curve (AUC). AUC is a widely-used measure of performance of classification [31]. Ranging between 0 and 1, it can be used to assess how good threshold values are at discriminating between groups. There are rules of thumb for assessing the discriminative power of measures based on AUC [30]. An AUC < 0.5 is considered no good, poor if AUC < 0.6, fair if AUC < 0.7, acceptable if AUC < 0.8, excellent if AUC < 0.9 and outstanding if AUC <1. The standard error or p-value is estimated using a 95% confidence interval. The test checks if the AUC is significantly different from 0.5. For those measures that are found to be valid according to the AUC value, we can determine a threshold based on the ROC curve. We need a criterion to choose a threshold value for a measure (sensitivity, 1-specificity pair) to balance benefits and costs. The purpose is to maximize both values, i.e. sensitivity and specificity, while at the same time [30] minimizing false-positive and false-negative. As was indicated in [7], we assume sensitivity and specificity to be of equal importance. The best threshold can then be selected by finding the point on the curve that maximizes both sensibility and specificity. This is the point with the greatest distance from the 0.5 diagonal.

IV. ANALYSIS OF THE EFFECTIVENESS OF THRESHOLD DETERMINATION TECHNIQUES

After describing the threshold determination techniques in detail, these will be used to extract thresholds, and after that, those thresholds will be validated to detect which technique obtains the most accurate threshold values.

A. *Hypothesis*

The previous discussion gives us reason to assume that threshold determination techniques based on logistic regression equation obtain threshold values with similar accuracy in classifying models. The main hypothesis is:

H_0: Thresholds obtained by both techniques are equally efficient in classifying the measurement results

To check this hypothesis, we calculate thresholds for a set of measures, which are able to predict understandability through the application of ROC curves and Bender method.

B. *Experimental settings*

To check the hypothesis we have used the experimental data obtained in two families of experiments. The **first family of experiments** was conducted by Rolón et al. [17] and included one experiment and two replicas (to see Figure 1). The **experimental material** was composed of 15 BPMN models which included a group of questions about the understandability of the model. We collected the efficiency of understandability tasks carried out by each subject in each model, which was calculated by dividing the number of correct answers by the time spent. Therefore, the **dependent variable** is the understandability and the **independent variables** the measures about structural properties of the model (number of nodes, control-flow complexity, etc.).

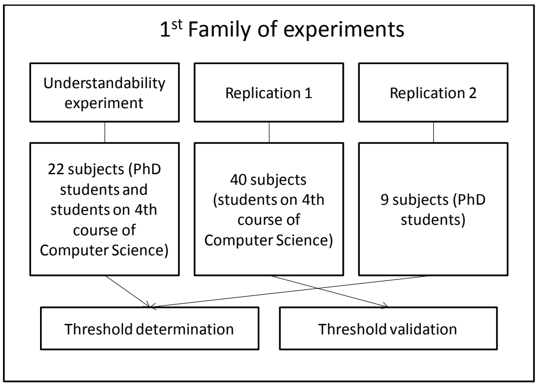

1st Family of experiments
Understandability experiment
Replication 1
Replication 2
22 subjects (PhD students and students on 4th course of Computer Science)
40 subjects (students on 4th course of Computer Science)
9 subjects (PhD students)
Threshold determination
Threshold validation

Figure 1. Description of the first family of experiments

The second **family of experiments** included one experiment and one replica (to see Figure 2). The

experimental material was composed of 10 BPMN models which included some understandability questions. As in the first family of experiments, the **dependent variable** was the understandability measured through the efficiency and the **independent variables** were the measures of structural properties of the models.

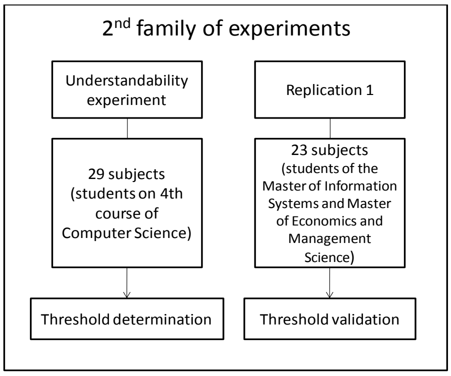

2nd family of experiments
Understandability experiment
Replication 1
29 subjects (students on 4th course of Computer Science)
23 subjects (students of the Master of Information Systems and Master of Economics and Management Science)
Threshold determination
Threshold validation

Figure 2. Description of the 2nd family of experiments

The mean (μ) and the standard deviation (σ) of the independent variables in the two families of experiments are summarized in Table II. Cells with a '-' indicate that the measure does not vary in that family of experiments.

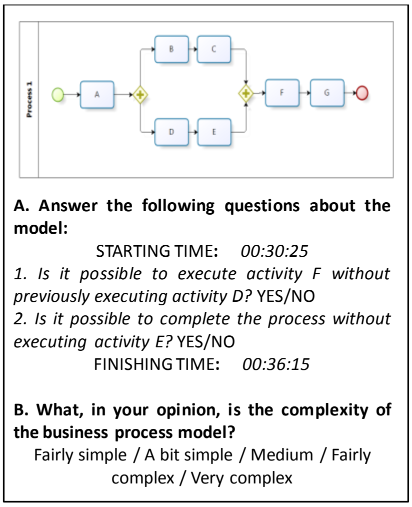
TABLE II. MEAN AND STANDARD DEVIATION OF INDEPENDENT VARIABLES IN THE 1ST AND 2ND FAMILY OF EXPERIMENTS

M	1st family of exp.		2nd family of exp.		M	1st family of exp.		2nd family of exp.	
	μ	*σ*	*μ*	*σ*		*μ*	*σ*	*μ*	*σ*
NEDDB	2.55	1.66	8.5	1.02	TNSF	36.2	23.8	64	8.44
NEDEB	0.66	1.2	-	-	CFC	10.8	11.1	32.3	16.9
NID	0.67	1.25	2.3	1.18	Nodes	43.6	23.4	-	-
NCD	0.73	1.24	-	-	Diam	12.2	5.01	13.6	5.91
NPF	0.53	0.96	4.5	1.18	Densi	0.10	0.21	0.03	0.005
NSFG	11.7	12.5	60.8	23.5	AGD	2.78	1.22	3.87	0.35
NP	2.7	1.3	-	-	MGD	3.33	1.85	6.4	2.01
NMF	7.13	5.86	-	-	GM	11.4	10.8	13.3	7.04
NSFE	3.8	2.78	5.6	2.54	GH	0.28	0.37	0.76	0.28
TNG	5.13	5.27	15.3	5.29	Sequent.	0.49	0.26	0.29	0.12
CLP	2.21	1.6	-	-	Separ	0.38	0.23	0.46	0.19
TNE	7.4	4.46	-	-	CNC	0.89	0.31	1.39	0.18
TNA	21.9	13.4	28.7	23.5	TS	0.13	0.34	1.2	0.87

In Figure 3 is included an excerpt of the experimental material, which was similar in both families of experiments.

C. Thresholds calculation

The obtained data were divided into two groups. One group was used to define the thresholds and the other to validate them. For the first family of experiments, the first experiment and the 2nd replication were used for threshold definition (because subjects have similar background) and in the second family of experiments, the experiment is used for threshold determination and the replication for validating them. Thresholds are calculated by applying ROC curves and the Bender method on the experimental data described previously. A prerequisite for calculating thresholds is that measure values must vary enough for the results obtained with the threshold determination techniques to be significant. This is the reason why some measures such as NEDEB or TNE and others were not used for threshold definition in the second family of experiments.


Process 1
A
B
C
D
E
F
G

A. Answer the following questions about the model:

STARTING TIME: *00:30:25*

1. Is it possible to execute activity F without previously executing activity D? YES/NO

2. Is it possible to complete the process without executing activity E? YES/NO

FINISHING TIME: *00:36:15*

B. What, in your opinion, is the complexity of the business process model?

Fairly simple / A bit simple / Medium / Fairly complex / Very complex

Figure 3. An example of experimental material

TABLE III. THRESHOLDS FOR BUSINESS PROCESS MODEL MEASURES

M	1st family of experiments				2nd family of experiments			
	Bender			*ROC*	*Bender*			*ROC*
	0.6	*0.7*	*0.8*		*0.6*	*0.7*	*0.8*	
NEDDB	3.16	3.87	4.72	1.5	9.12	9.8	10.6	9.5
NEDEB	1.59	2.62	3.87	0.5	-	-	-	-
NID	1.36	2.17	3.15	0.5	2.91	3.55	4.34	3.5
NCD	1.40	2.18	3.13	0.5	-	-	-	-
NPF	1.02	1.6	2.3	0.5	6.16	7.98	10.2	7.5
NSFG	16.9	23.2	30.9	4.5	74,7	88.9	106	85.5
NP	3.37	9.7	12.7	1.5	-	-	-	-
NMF	10	12.2	17.1	1	-	-	-	-
NSFE	6.17	8.74	11.8	2.5	2.39	-	-	6.5
TNG	7.20	9.71	12.7	3.5	18.1	21.1	24.8	20.5
CLP	2.97	3.8	4.8	0.5	-	-	-	-
TNE	9.30	11.5	14.2	4	-	-	-	-
TNA	26.3	31.3	37.4	12	25.8	22.7	19.1	23.5
TNSF	42.2	50.2	60.1	24.5	69.1	74.4	80.9	72.5
CFC	15.5	21.1	27.9	6	40.3	49.5	60.7	47.5
Nodes	50.6	58.1	67.2	41.5	-	-	-	-
Diam	10.8	7.92	5.17	8.5	7.16	0.36	-	9
Densi	0.13	0.2	0.28	0.03	0.06	0.04	0.04	0.03
AGD	2.38	1.82	1.13	1	4.08	4.28	4.53	4.17
MGD	2.43	1.42	0.19	4.5	7.53	8.75	10.2	8.5
GM	5.76	-	-	7.5	17.1	21.1	26.1	19
GH	0.08	-	-	0.54	1.12	1.16	1.4	0.93
Sequent.	0.58	0.7	0.85	0.56	0.16	0.01	-	-
Separ	0.53	0.71	0.92	0.48	0.31	0.15	-	0.53
CNC	0.65	0.37	0.03	1.06	1.5	1.61	1.75	1.57
TS	-	-	-	0.5	3.59	6.18	9.35	0.5

Obtaining thresholds by the application of Bender method and ROC curves requires the definition of the input variables. As we mentioned, those techniques have two steps: the logistic regression analysis and the extraction of thresholds. The logistic regression analysis is common in both techniques; it requires a binary variable (dependent variable) and a continuous one (independent variable). Continuous variables are the measures. The binary variable requires the dichotomization of the dependent variable: in this case, the efficiency of understandability. In the family of experiments this variable is not binary, because it ranges between 0 and 1. However, it can be converted into a dichotomous one, signifying that it would be 1 when it was higher than the median and 0 when it was lower [32]. The median was also used for dichotomizing variables in [13]. As regards the second step in threshold determination, ROC curves do not require any configuration. The Bender method requires the definition of p_0, however. That value is used to indicate the probability of considering the model as non-understandable. For example, if p_0 is 0.9, the probability of considering the model non-understandable is about 90%. Since there is no consensus about what value of p_0 is the most suitable, we chose three possible options: p_0 = 0.6, p_0 = 0.7, and p_0 = 0.8. We believe that a model with a 60%, 70% or 80% percentage of being deemed non-understandable should be submitted to the redesign process.

Thresholds obtained in the first and second family of experiments for each measure are depicted in Table III. The symbol '-' signifies that the threshold for that measure could not be calculated, or that the threshold value is out of the variable domain.

D. Thresholds validation

In this section, we present findings from applying the threshold in experimental data for validity. We approached the validation of threshold from an information retrieval perspective. In this field, true and false positives, as well as true and false negatives, are used as the basis for calculating precision, and recall measures are employed for assessing the quality of a search results [15]. Precision is the ratio of true positives to the sum of true and false positives. In terms of understandability, this is the ratio of correctly-found non-understandable models, based on a threshold value in relation to the sum of all error predictions. Recall is the ratio of true positives to the sum of true positives and false negatives. In other words, recall is the ratio of correctly-found non-understandable models to the sum of all non-understandable models.

The next section compares results obtained for each technique.

V. Discussions

In this section, we discuss the derived threshold for each technique, along with their validation, in order to check which technique obtained the most accurate threshold values.

A. Bender method vs. ROC curves

The results of precision and recall of threshold are set out in Table IV, Table V and Table VI. In Table IV and Table V there are some cells in gray, to highlight the most suitable values (the higher the values of recall or precision, the more suitable the result is).

According to precision results of thresholds, 60% of the most accurate precision results were obtained by the Bender method in the first family of experiments and 42% in the second one. That means thresholds obtained by any technique can classify models with similar precision. As regards recall results, 68% of the most accurate results were obtained by ROC curves in the first family of experiments and 47% in the second one. That means that the classification of models by thresholds obtained with ROC curves is more comprehensive than with the Bender method.

TABLE IV. Validation of the Bender method in the first family of experiments

	Precision			Recall			F-measure		
	0.6	***0.7***	***0.8***	***0.6***	***0.7***	***0.8***	***0.6***	***0.7***	***0.8***
NEDDB	0.80	0.80	0.67	0.43	0.43	0.18	0.55	0.55	0.28
NEDEB	0.81	0.88	0.77	0.32	0.23	0.10	0.45	0.36	0.17
NID	0.78	0.83	0.67	0.31	0.22	0.09	0.44	0.34	0.15
NCD	0.81	0.83	0.67	0.32	0.22	0.09	0.45	0.34	0.15
NPF	0.78	0.78	0.67	0.31	0.31	0.09	0.44	0.44	0.15
NSFG	0.43	0.44	0.45	0.41	0.31	0.22	0.41	0.36	0.29
NP	0.41	0.49	0.49	0.45	0.11	0.11	0.42	0.17	0.17
NMF	0.38	0.43	0.46	0.35	0.21	0.08	0.36	0.28	0.13
NSFE	0.39	0.43	0.42	0.30	0.21	0.11	0.33	0.28	0.17
TNG	0.42	0.41	0.41	0.27	0.20	0.20	0.32	0.26	0.26
CLP	0.39	0.41	0.47	0.47	0.28	0.15	0.42	0.33	0.22
TNE	0.45	0.45	0.46	0.34	0.34	0.11	0.38	0.38	0.17
TNA	0.44	0.46	0.46	0.52	0.34	0.34	0.47	0.39	0.39
TNSF	0.44	0.45	0.45	0.53	0.34	0.22	0.48	0.38	0.29
CFC	0.44	0.44	0.46	0.42	0.42	0.22	0.42	0.42	0.29
Nodes	0.44	0.44	0.48	0.43	0.43	0.25	0.43	0.43	0.32
Diam	0.38	0.38	0.36	0.75	0.93	0.98	0.50	0.53	0.52
Densi	0.36	0.35	0.34	0.93	0.94	0.99	0.51	0.51	0.50
AGD	0.35	0.36	0.36	0.89	0.98	0.98	0.50	0.52	0.52
MGD	0.35	0.36	0.36	0.89	0.98	0.98	0.50	0.52	0.52
GM	0.38	-	-	0.83	-	-	0.52	-	-
GH	0.40	-	-	0.55	-	-	0.46	-	-
Seq	0.35	0.36	0.36	0.81	0.98	0.98	0.48	0.52	0.52
Separ	0.35	0.34	0.34	0.86	0.99	0.99	0.49	0.50	0.50
CNC	0.36	0.33	0.33	0.93	0.95	1	0.51	0.48	0.49

TABLE V. Validation of Bender method in the second family of experiments

	Precision			Recall			F-measure		
	0.6	***0.7***	***0.8***	***0.6***	***0.7***	***0.8***	***0.6***	***0.7***	***0.8***
NEDDB	0.83	0.83	0	0.33	0.33	0	0.47	0.47	0
NID	0.62	0.83	0	0.5	0.33	0	0.55	0.47	0
NPF	0.64	0.83	0	0.38	0.33	0	0.47	0.47	0
NSFG	0.64	0.83	0	0.38	0.33	0	0.47	0.47	0
NSFE	0	-	-	0	-	-	0	-	-
TNG	0.64	0.83	0.83	0.38	0.33	0.16	0.47	0.47	0.46
TNA	0.64	0.83	0	0.38	0.33	0	0.47	0.47	0
TNSF	0.64	0.83	0	0.38	0.16	0	0.47	0.26	0
CFC	0.64	0.83	0.83	0.38	0.33	0.16	0.47	0.47	0.26
Diam	0.55	0	-	1	0	0	0.70	0	0
Den	0.83	0	0	0.33	0	0	0.47	0	0
AGD	0.64	0.83	0.83	0.38	0.33	0.16	0.47	0.47	0.46
MGD	0.64	0.83	0.83	0.38	0.33	0.16	0.47	0.47	0.46
GM	0.64	0.83	0.83	0.38	0.33	0.16	0.47	0.47	0.46
GH	0.64	0.83	0.83	0.38	0.33	0.16	0.47	0.47	0.46
Seq	0	0	-	0	0	-	-	-	-
Separ	0.56	0	-	0.45	0	-	0.49	0	-
CNC	0.62	0.83	0	0.5	0.16	0	0.55	0.26	0
TS	0	0	0	0	0	0	0	0	0

Interpreting both measures in an isolated manner does not report as many advantages as when this is done with them in conjunction. A way to combine precision and recall is the harmonic mean, which is typically called F-measure. It ranges between 0 and 1. This measure provides a single measurement for the relationship we are validating.

TABLE VI. VALIDATION OF ROC CURVES IN THE FIRST AND SECOND FAMILY OF EXPERIMENTS

	Precision		Recall		F-measure	
	1st f.exp	*2nd f.exp*	*1st f.exp*	*2nd f.exp*	*1st f.exp*	*2nd f.exp*
NEDDB	0.81	0.83	0.32	0.33	0.45	0.47
NEDEB	0.78	-	0.41	-	0.53	-
NID	0.78	0.83	0.41	0.33	0.53	0.47
NCD	0.69	-	0.46	-	0.55	-
NPF	0.75	0.83	0.28	0.33	0.40	0.47
NSFG	0.38	0.83	0.76	0.33	0.50	0.47
NP	0.37	-	0.95	-	0.53	-
NMF	0.36	-	0.73		0.48	-
NSFE	0.36	0.51	0.72	0.62	0.48	0.55
TNG	0.4	0.58	0.5	0.73	0.44	0.64
CLP	0.37	-	0.90	-	0.52	-
TNE	0.37	-	0.89	-	0.52	-
TNA	0.39	0.83	0.94	0.33	0.55	0.47
TNSF	0.4	0.83	0.89	0.33	0.55	0.47
CFC	0.4	0.58	0.71	0.73	0.51	0.64
Nodes	0.41	-	0.84	-	0.55	-
Diam	0.38	0.46	0.93	0.83	0.53	0.59
Densi	0.4	0.83	0.89	0.33	0.55	0.47
AGD	0.36	0.58	0.98	0.73	0.52	0.64
MGD	0.42	0.58	0.19	0.73	0.26	0.64
GM	0.39	0.58	0.71	0.73	0.50	0.64
GH	0.43	0.58	0.5	0.73	0.46	0.64
Seq	0.37	-	0.8	-	0.50	-
Separ	0.36	0.58	0.85	0.82	0.50	0.67
CNC	0.41	0.83	0.47	0.33	0.43	0.47
TS	-	0.56	-	0.56	-	0.56

Moreover, we need a statistical technique to compare the results of precision and recall measures formally. Since we cannot assume that the sample follows a normal distribution, the comparison requires a non-parametric test. The test of Mann-Whitney [33] is used to check the heterogeneity of two ordinal samples, so it will indicate to us whether there is a significant difference between the two techniques, based on precision and recall measures. The result of this test is shown in Table VII. Precision, recall and the harmonic mean of them (F-measure) is displayed, and we have highlighted the significant results in grey.

In Table VII, the test of Mann-Whitney reveals that the precision of both techniques is not significantly different (p-value > 0.05); this means that in classifying models there is no prevalence between the precision of thresholds obtained by either of the techniques. On the other hand, there are significant differences between recall values of the two techniques; mainly for high values of p_0 (p_0 is the input value of the Bender method). Finally, the F-measure is significantly different for both techniques in most of the cases. In conclusion, then, the comparison of these techniques presented in this paper is valid, because the differences between them are meaningful.

Since the Mann-Whitney test indicates that there are significant differences between recall and F-measure values, it interests us to know which of those techniques is the most suitable; in other words, to find out the technique that can extract thresholds with higher values of recall and F-measure. As it can be observed when comparing the values of the F-measure of both techniques in the two families of experiments (Table IV, Table V and Table VI), more suitable values are obtained by ROC curves (68% in the first family and 52% in the second one), which indicates that ROC curves are a more suitable technique for threshold determination. This is illustrated in Figure 4 and Figure 5. Those figures show a comparison of the F-measure between the ROC curves and the Bender method for each family of experiments. Those charts reflect that the F-measure associated with ROC curves is higher in most of the cases.

TABLE VII. DIFFERENCES BETWEEN TECHNIQUES

P_0	F.Exp	Measure	U Mann-Whitney	p-value
0,6	1	Precision	284	0.579
		Recall	215	0.065
		F-measure	153	0.002
	2	Precision	0.89	0.049
		Recall	116	0.259
		F-measure	111	0.109
0,7	1	Precision	208	0.23
		Recall	187	0.090
		F-measure	125	0.002
	2	Precision	144	0.425
		Recall	59	0.000
		F-measure	103	0.036
0,8	1	Precision	233	0.494
		Recall	167	0.032
		F-measure	86	0,000
	2	Precision	96	0.056
		Recall	0	0.000
		F-measure	0	0.000

B. *Threats to validity*

With regards to the **conclusion validity**, the size of the sample data for performing the calculations is about 71subjects for the first family of experiments and 52 subjects for the second one. The number of measures used to obtain thresholds is 25, which is the number of comparisons between precision and recall values, and this is considered significant to allow us to obtain conclusion validity.

Construct validity is about reflecting our ability to measure what we want to measure. The comparison of the techniques (which are the independent variables of the study) is done by the measures precision and recall (which are the dependent variables of the study). Those measures are commonly accepted in related works, so we considered that making a comparison between techniques based on precision and recall measures provides an objective comparison.

Internal validity concerns whether the effect measured is due to changes caused by the researcher, or from some other unknown cause, in other words, if there is a causal relationship between treatment and outcome. To answer this question we have to highlight some points. First of all, we examine the threats to the experimental data used for threshold determination. In both families of experiments, all subjects had roughly the same knowledge about modeling, because all the students were close to finishing their degree, or had recently finished it.

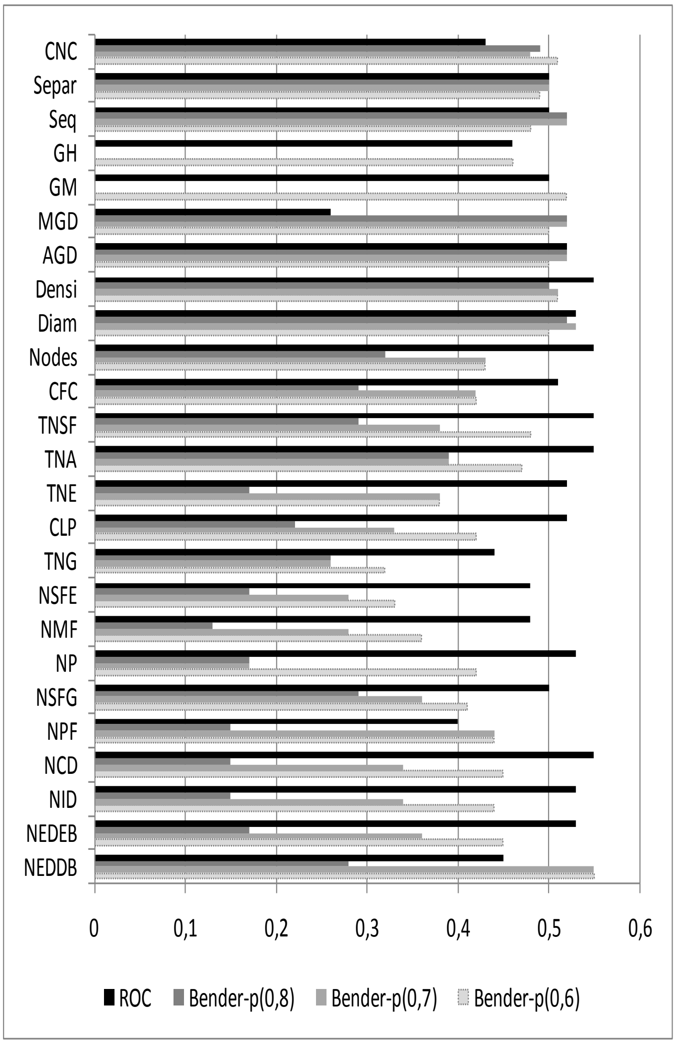


Figure 4. Comparison of the F-measure between the two techniques in the first family of experiments

Subject motivation effects are discarded, because all our subjects received extra marks, and fatigue effects were mitigated by conducting the experiments on different days. Since we consider that most of the typical threats to internal validity are not real threats for the experiments, we are confident that there are no major risks. More details about threats to validity in the experiments are described in [17]. On the other hand, there are threats related to the techniques used for threshold determination. The first is that logistic regression requires a binary variable, and dichotomization can imply a loss of information. This can result in the thresholds not being very accurate. Another point is that one particular curve may have a larger AUC (which is apparently better), even though the alternative may show superior performance over almost the entire range of values of the classifications threshold. This fact indicates that sometimes ROC curves can offer not very accurate thresholds. Moreover, application of the Bender method requires the subjective definition of p_0 and that directly affects the accuracy of thresholds. Despite these limitations, we believe that the comparison between both of the techniques studied in this paper offers strong evidence for ROC curves being more accurate in threshold determination.

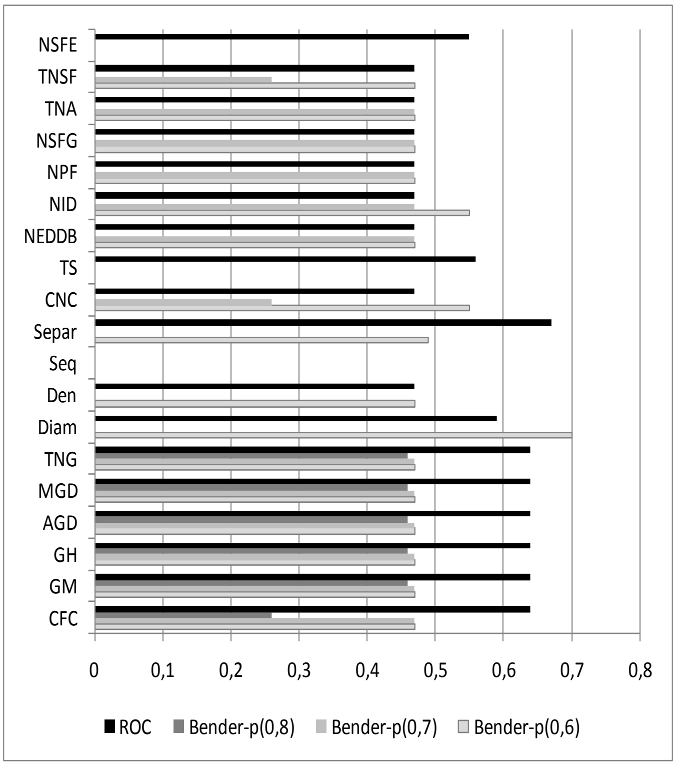


Figure 5. Comparison of the F-measure between the two techniques in the second family of experiment

In relation to **external validity**, some characteristics of the experiments could limit the applicability in reality. In our case the generalization of results to other studies means that ROC curves will always obtain more accuracy thresholds than the Bender method. This can depend on the p_0 chosen: a p_0 that is different for each measure obtains more accuracy results than the same p_0 for all. For this reason, we selected three different values, in order to cover as many choices as possible. We believed that determining thresholds with a 60%, 70% or 80% chance of considering the model as non-understandable is sufficiently valid.

VI. Conclusions and Future work

In this paper, we analyzed two techniques for threshold determination, in order to demonstrate which of them obtain more accurate thresholds. The techniques are ROC curves and the Bender method, both of which use the logistic regression as a part of their calculations. For this reason, the hypothesis assumed that both techniques obtain similar threshold values, and to prove this, we extracted thresholds with them, using the experimental data of two families of experiments. For each threshold obtained, the precision and recall measure, and the harmonic mean of them were calculated.

The differences between both techniques are checked through the application of a statistical test, the Mann-Whitney test. This test detected that there are significant differences between the F-measure values, and the charts revealed that the better results corresponded to ROC curves.

The main finding of this paper is that both techniques are able to obtain thresholds with similar precision (precision means that all the non-understandable models were classified using thresholds, although others which were considered understandable were also selected). However, only when the p_0 is higher than 0.6 are the differences of recall values between the techniques significant, and ROC curves are again the most suitable technique. The F-measure, however, considers both the precision and the recall to compute the score. We can thus conclude that ROC curves obtain more accurate thresholds in classification tasks than the Bender method.

As future work, we propose the comparison of the threshold obtained by other techniques, in order to select the most effective one for threshold determination. Moreover, more empirical validation is needed to generalize the results shown in this paper.

ACKNOWLEDGMENTS

This work was partially funded by the following projects: ALTAMIRA (Junta de Comunidades de Castilla-La Mancha, Fondo Social Europeo, PII2I09-0106-2463), INGENIOSO (Junta de Comunidades de Castilla La Mancha, PEII11-0025-9533) and PEGASO/MAGO (Ministerio de Ciencia e Innovación MICINN and Fondo Europeo de Desarrollo Regional FEDER, TIN2009-13718-C02-01).

REFERENCES

[1] Park, R.E., W.B. Goethert, and W.A. Florac, *Goal-Driven software Measurement: A Guidebook.* HANDBOOK CMU/SEI-96-HB-002, 1996.

[2] Enders, A. and H.D. Rombach, *A Handbook of Software and Systems Engineering: Empirical Observations.* Laws and Theories, Addison-Wesley, Reading, MA,, 2003.

[3] Boehm, B.W., *Software Engineering Economics.* Prentice-Hall, Englewood Cliffs, 1981.

[4] Henderson-Sellers, B., *Object-Oriented Metrics: Measures of Complexity.* Prentice-Hall, 1996.

[5] Alves, T.L., C. Ypma, and J. Visser, *Deriving metric thresholds from benchmark data*, in *Proceedings of the 2010 IEEE International Conference on Software Maintenance*. 2010, IEEE Computer Society. p. 1-10.

[6] Shatnawi, R., et al., *Finding Software Metrics Threshold values using ROC Curves.* Sofware Maintenance and Evolution: Research and Practice, 2009.

[7] Mendling, J., et al., *Thresholds for Error Probability Measures of Business Process Models.* International Journal of Systems and Software, 2011. **pending of publication**.

[8] Catal, C., O. Alan, and K. Balkan, *Class noise detection based on software metrics and ROC curves.* Information Sciences, 2011. **181**(21): p. 4867-4877.

[9] Bender, R., *Quantitative Risk Assessment in Epidemiological Studies. Investigating Threshold Effects.* Biometrical Journal, 1999. **41**(3): p. 305-319.

[10] Shatnawi, R., *A Quantitative Investigation of the Acceptable Risk levels of Object-Oriented Metrics in Open-Source Systems.* IEEE Transactions on Software Engineering, 2010. **36**(2): p. 216-225.

[11] Sánchez-González, L., et al., *Quality Assessment of Business Process Models Based on Thresholds.* CoopIS 2010 - 18th International conference on Cooperative Information Systems, 2010: p. 78-95.

[12] Perez-Castillo, R., et al., *Obtaining Thresholds for the Effectiveness of Business Process Mining.* ESEM 2011, 2011: p. 453-462.

[13] Sánchez-González, L., et al., *Towards Thresholds of Control Flow Complexity Measures for BPMN Models.* 26th Symposium On Applied Computing SAC 10, 2011: p. 1445-1450.

[14] Zweig, M. and G. Campbell, *Receiver-Operating Characteristic (ROC) Plots: A fundamental evaluation tool in clinical medicine.* Clinical Chemistry, 1993. **39**(4): p. 561-577.

[15] Baeza-Yates, R.A. and B.A. Ribeiro-Neto, *Modern Information Retrieval.* ACM Press / Addison Wesley, 1999.

[16] Sánchez-González, L., et al., *Measurement in Business Processes: a Systematic Review.* Business process Management Journal, 2010. **16**(1): p. 114-134.

[17] Rolon, E., et al., *Evaluation of BPMN Models Quality. A Family of Experiments.* ENASE - International Conference on Evaluation of Novel Approaches to Software Engineering, 2008.

[18] Rolón, E., et al., *Analysis and Validation of Control-Flow Complexity Measures with BPMN Process Models.* The 10th Workshop on Business Process Modeling, Development, and Support, 2009.

[19] Seffah, A., et al., *Usability measurement and metrics: A consolidated model.* Software Quality Control, 2006. **14**(2): p. 159-178.

[20] OMG. *Business Process Model and Notation (BPMN), Version 2.0.* 2011; Available from: http://www.omg.org/spec/BPMN/2.0/.

[21] Mendling, J., *Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness*. 2008: Springer Publishing Company, Incorporated.

[22] Cardoso, J., *Process control-flow complexity metric: An empirical validation.* SCC '06: Proceedings of the IEEE International Conference on Services Computing, 2006: p. 167--173.

[23] Erni, K. and C. Lewerentz, *Applying Design-metrics to Object-Oriented Frameworks.* Proceedings of METRICS, 96, 1996: p. 64-74.

[24] French, V.A., *Establishing software metric thresholds.* International wotkshop on software measurement, 1999.

[25] Benlarbi, S., et al., *Thresholds for Object-Oriented Measures.* Institute for Information Technology, National Research Council Canada, 2000.

[26] Rosenberg, L., *Applying and interpreting object oriented metrics.* Software Technology Conference, 1998.

[27] Herbold, S., J. Grabowski, and S. Waack, *Calculation and optimization of thresholds for sets of software metrics.* Empirical Software Engineering, 2011.

[28] Yoon, K.A., O.S. Kwon, and D.H. Bae, *An approach to outlier detection of software measurement data using the K.means clustering method.* IEEE computer society, 2007: p. 443-445.

[29] Mendling, J., et al., *Thresholds for Error Probability Measures of Business Process Models.* International Journal of Systems and Software, 2012. **85**(5): p. 1188-1197.

[30] Hosmer, D. and S. Lemeshow, *Applied Logistic Regression (2nd edn).* Wiley-InterScience, 2000.

[31] Hand, D., *Measuring Classifier Performance: a Coherent Alternative to the Area Under the ROC curve.* Machine Learning, 2009. **77**(1): p. 103-123.

[32] Royston, P., G.A. Douglas, and W. Sauerbrei, *Dichotomizing continuous predictors in multiple regression: a bad idea.* Statistics in Medicine, Wiley InterScience, 2005. **25**: p. 127-141.

[33] Siegel, S. and J. Castellan, *Nonparametric statistics for the behavioral sciences.* London MacGraw-Hill, 1988.