Improvements on diagnostic assessment questionnaires of Maturity Level Management with feature selection

Bruno Prece
State University of Londrina (UEL)
Londrina, Brazil
bprece@gmail.com

Edson Pacheco
State University of Londrina (UEL)
Londrina, Brazil
edsonpachec@gmail.com

Rodolfo Miranda Barros
State University of Londrina (UEL)
Londrina, Brazil
rodolfo@uel.br

Sylvio Barbon Jr.
State University of Londrina (UEL)
Londrina, Brazil
barbon@uel.br

ABSTRACT

In the last few years, several new tools addressing maturity level management have been proposed, e.g. diagnostic assessment questionnaires (DAQ). In practice, the usage of questionnaires presents some drawbacks related to subjectivity, time cost, and applicant bias. Moreover, the questionnaires may present a large number of questions, as well as part of them redundant. Another important fact of real-life application of DAQs concerns the usage of multiple questionnaires, increasing the shortcoming impacts. To pave the way to a more convenient tool to support and facilitate the achievement of organizational strategies and objectives, we proposed an intelligent reduction of DAQs by the use of single-label and multi-label feature selection. In this paper, we reduced four DAQs (Risk Management, Infrastructure, Governance and Service Catalogs) with our proposal in comparison to different feature selection algorithms (χ^2, Information Gain, Random Forest Importance and ReliefF). The reduction was driven by a machine learning prediction model towards ensuring the new subset of question grounded in the same obtained score result. Results showed that removing irrelevant and/or redundant question it was possible to increase the model fitting even reducing about one-third of the questions with the same predictive capacity.

CCS CONCEPTS

• **Applied computing** → **IT governance**; • **Software and its engineering** → **Capability Maturity Model**; • **Computing methodologies** → *Feature selection*;

KEYWORDS

Machine Learning, Information and Communication Technology, ICT Governance, Maturity Level

ACM Reference format:
Bruno Prece, Edson Pacheco, Rodolfo Miranda Barros, and Sylvio Barbon Jr.. 2019. Improvements on diagnostic assessment questionnaires of Maturity Level Management with feature selection. In *Proceedings of XV Brazilian Symposium on Information Systems, Aracaju, Brazil, May 20–24, 2019 (SBSI'19),* 8 pages.

1 INTRODUCTION

Information and communication technology (ICT) has been earning every day a strategic role within the corporate environment. ITC Management is the control and processing system required to achieve the strategic objectives settled by the organization's governing body, consisting of the appropriate use of resources (such as hardware, software, people, processes, practices) [11].

The Governance provides the direction for ICT. While the ICT plans, executes and controls operational tasks, the Governance controls the management. ICT governance creates controls in order to ICT cooperates transparently with stakeholders by aligning ICT with business processes. ICT Governance focuses on establishing processes to ensure the organization and control of compliance with strategic objectives [32].

In ICT Governance, the use of best practices guarantees the organization's infrastructure support and facilitate the achievement of organizational strategies and objectives [32]. To ensure better efficiency, two very important ideas - whose metrics help define maturity levels, certification, and international recognition - are on the agenda of managers' day. This is because the market demands high efficiency, operations quality and active vision throughout the ICT lifecycle.

The concepts that help in this walk are the Information Technology Infrastructure Library (ITIL) and the Control Objectives for Information and related Technology (COBIT). ITIL and COBIT are quite different from each other. COBIT has more affinity with process and control auditing, and ITIL is closer to managing IT services. While COBIT is primarily concerned in guiding organizations in the implementation, operation, and improvement of governance and IT management processes, ITIL provides good practice guidelines for managing and executing IT services from the perspective of business value creation. To assess the maturity of the organization, known methods such as ITIL and COBIT follow a similar and structured approach. They emphasize the need for developing processes

to improve product development and customer satisfaction and support the coordination of interdisciplinary activities related to the project [12][13].

The purpose of maturity assessment models is to identify and establish the capability and evolution of organizational processes. It is structured as a series of capability levels. Its maturity levels demonstrate stages of improvement in the implementation of organizational processes. The maturity levels indicate the achievement profile of the organization and orientate the organization to improve your processes to achieve the target profile [23].

Considering the Brazilian national scenario, in the last few years, some studies have proposed new diagnostic evaluation tools - diagnostic assessment questionnaires (DAQ) - for maturity models through the use of questionnaires, which suit the reality of software development in some regions of Brazil [4][7][10][28]. However, DAQ are composed of a large number of questions, many of them homologous and redundant inter and intra questionnaires. This makes it difficult and increases the costs of the processes of obtaining information from organizations and the evaluation of this information.

Through computational intelligence and based on the hypothesis that there are redundant questions in different DAQ from different areas of an organization, we propose to analyze four questionnaires (Risk Management [7], Infrastructure [10], Governance [4] and Service Catalog [28]) and to evaluate the most relevant questions to reduce the number from them guided by results obtained from the whole questions. It is important to highlight the fact of we are facing all of four DAQ at once, a challenge posed by a real-life scenario.

We proposed to tackle the question selection as a multi-label feature selection since four different questionnaires took place, at once. Different from traditional binary or single-label pattern recognition problem, the multi-label modeling is grounded in more than one expected target [36]. More precisely, each DAQ was dealt a given target and its value as an output. Our proposal, using multi-label feature selection [29], was compared to traditional feature selection techniques: Chi-Squared (χ^2) [34], RF Importance [8], Information Gain (IG) [18] and ReliefF [21]. All these techniques were compared over questionnaires collected from 20 organizations focusing on getting some information about processes of risk management, service catalog, infrastructure, and governance.

This work is organized as follows: Section 2 presents the maturity modeling questionnaire. Section 3 presents the multi-label learning, feature selection approaches, and evaluation metrics applied in the learning models. The materials and methods used in this work are presented in Section 4. Section 5 presents the results of the experiments and Section 6 concludes and presents the future works.

2 MATURITY MODELING QUESTIONNAIRE

The role of DAQs is to identify, through the responses provided by the applicant (interviewee), the maturity level with which an organization's Software Development Process (SDP) addresses the organizational processes of Risk Management, Service Catalog, Infrastructure, and Governance.

The structured questionnaire consists of an ordered set of closed-ended questions. The close-ended question limits the interviewee to a set of alternatives, which objectively translate the situations occurring into the organization daily to simplify the filling of the questionnaire by the interviewee. A coefficient (α) is assigned to each alternative that quantify their impact in relation to a given question. The coefficient (α) can take a range of values from -3 to +3. A value of 0 indicates that there is no influence, a value of 1 indicates low influence, a value of 2 indicates average influence and a value of 3 indicates high influence. While the signals "+" or "-" determine a positive or negative influence, as exemplified in Table 1. These factors are used to calculate the attendance rate (%) [4].

Table 1: Question example

Does the organization have well-defined criteria and parameters to identify the risks present in their projects?		
Alternative		α
(A)	Yes, the organization has well-defined criteria and parameters, which are known by everybody.	3
(B)	Yes, the organization has well-defined criteria and parameters, however they are not disclosed.	2
(C)	I don't know this information.	0
(D)	The organization has some criteria and parameters, which are not known by everybody.	-2
(E)	No, the organization does not have criteria and parameters to manage the risks.	-3

Besides the relationship between the questions and the alternatives, which are the coefficients (α), exemplified in Table 1, another important component of DAQs is the relationship between the questions and the services of the organizational processes, which is given by weights. Therefore, the same question can influence one or more services at the same time. Table 2 presents the relationship matrix between a question and the weights it exerts on each service [7].

Based on the information collected through the questionnaire following the question model exposed in the Tables 1 and 2, the result of the SDP assessment is obtained, which is services oriented. It is necessary to calculate the product between the weight of the question in the service and the coefficient (α) related to the selected alternative. The final score is obtained by the addition of these products for each service [7].

To calculate the attendance rate (%) on each service, this final score must be adjusted based on the extreme values of the questionnaire, which determine an interval between its highest and lowest possible value. The final score is positioned in the interval described, determining the percentage of attendance of each service. Therefore, the maturity level of the organization is defined based on the service with the lowest attendance rate and classified according to Table 3 for the approach of this paper [7].

3 MULTI-LABEL LEARNING

Supervised Machine Learning is based on assumption that given a set of instances and a finite set of labels, the learning algorithm's goal is to associate each instance with a unique label (single-label). However, in many problems, each instance can be associated with multiple labels (multi-label). For example, in text-categorization [1],

Table 2: Weighting example of the questions and services

Does the organization have well-defined criteria and parameters to identify the risks present in their projects?		
Service	**Justification**	**Weight**
Establish context	The parameters are present in all phases of the risk management process.	4
Identify risks	Parameters determine the scope of the risk management process.	3
Analyzing risks	The parameters define the methodologies that will be used to analyze the risks.	1
Assessing risks	The parameters define metrics to classify the risks.	2
Treating risks	The parameters establish how to proceed in the treatment.	1
Monitoring and control	The parameters indicate how to evaluate efficacy.	2
Communication and consultation	The parameters determine what must be reported.	1

Table 3: Conversion of attendance rate at maturity level

Attendance Rate (%)	**Level**
≤ 50	Immature
>50	Mature

each document may contain multiple subjects, image content may have multiple objects [24].

Considering X a dataset with N instances, and a set of labels $Y = \{y_1, y_2, y_3, ..., y_q\}$ containing q possible labels. The algorithm's goal is to learn a function $F : X \rightarrow 2^y$ from dataset X. Thus, for each new instance $\mathbf{x} \in X$, predicts a set of labels $\mathbf{y} \sqsubseteq Y$ which describes $\mathbf{x}$.

Multi-label leraning methods can be organized into two categories: problem transformation and algorithm transformation. In the first category, the multi-label problem is transformed into a single-label problem. Thus, is not required adaptations on the learning algorithm which enables the application of traditional learning algorithms such as Support Vector Machine (SVM) [6] and Random Forest (RF) [3]. Examples of problem transformation methods are Binary Relevance (BR) [2] and Label Powerset (LP) [31], both applied in this work and described in the following paragraphs. The second category consists of extended machine learning algorithms tailored to handle multi-label problems, for example, the kNN-based algorithm BRkNN [26] and multi-label naïve Bayes (MLBN) [35].

Binary Relevance [2] is a problem transformation approach that creates q binary problems, each for one different label in label space Y. Afterward, each training sample is used as a basis of induction of q binary learners, labeling as positive the relevant labels and negative for irrelevant. Thus, for new instance prediction, Binary Relevance execute the union of positive labels predicted by q independent models [36]. Although deals in a simple way with multi-label problems, the main drawback of BR is not to consider the potential correlation among the labels.

Label Powerset [31] considers that each distinct label set combinations in a multi-label problem as one label in a new single-label problem, then a single-label learning model can be inducted. For each unseen instance, LP outputs the set of labels liked with the class predicted with basis on single-label model prediction. Unlike BR, LP considers the correlation among the labels, but the numerous labels generated by problem transformation may be linked with a few training examples, and this would create an imbalanced single-label problem.

Problem transformation approaches makes possible to use traditional base learners in a multi-label problem. In this work the Random Forest and Support Vector Machines are used as base learners. Support Vector Machines is a machine learning algorithm proposed by [6], was developed according to statistical learning theory and structural risk minimization, to reduce the misclassification risk and improve the generalization potential by maximization margin criterion.

Random Forest algorithm builds an ensemble of learning models using the bootstrap aggregating (bagging) strategy, creating n new training sets by the random choice of training examples, where n is the number of examples in training dataset. This strategy improves the prediction accuracy by small changes in training samples. The label with the majority votes is the predicted result [3].

3.1 Feature Selection

Considering a search space $X = \{x_1, x_2, x_3, x_4, x_n...\}$, feature selection is intended to find a subset $X' \subseteq X$ with similar ability to describes the problem as the original dataset. Feature selection is an essential task in machine learning process improving the problems interpretability and the prediction accuracy, besides reducing storage requirements and training time [9].

Feature selection methods are classified such as filters, wrappers, and embedded approaches. Filter approach is independent of the learning algorithm and selects the most relevant features by ranking the information scores calculated for each feature, discarding the features with insufficient scores. Information Gain [18], Chi-squared [34] and ReliefF [21] are examples of filter approach. The first two are univariate and the last one multivariate method. Univariate methods measure the feature importance calculating the correlation between each feature and the labels based on entropy theory, therefore, correlations among the features are not considered. On the other hand, multivariate methods measure the features dependencies, which may result in a better selection, although, their scalability is lost [5].

Wrapper methods require specifics learning algorithms to evaluate and select the best features. Despite generally find better features for the specific learning algorithm, the approach is computationally expensive because of the learning algorithm is called for each set of attributes being evaluated. Genetic Algorithm it is an example of wrapper. Embedded methods are intended to reduce the computational complexity of the wrapper approach introducing feature selection task as part of the training process [5].

Chi-squared (χ^2) is a common statistical test that measures the independence degree between the features and categories by distribution expected. Given a feature t and the class c, where A is the number of times that t and c occur simultaneously. B is the number of times where the feature t occur without c, that is, t occur with another class. C is the number of class c occurrences with other features different from t. D is the number of examples in the dataset where neither t nor c occurs and, N is the total number of examples in the dataset [34].

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

In this work, the RF algorithm is used both to build the predictive model and to feature selection, known as RF Importance. In the feature selection task, features are ranked based on the mean of their Gini Index measured in all random trees generated during algorithm induction [27].

Information Gain (IG) is a filter approach that measures the dependencies between each feature and a single class label by entropy. The measure consists of evaluating the difference in entropy sum of all training examples in the dataset D with the entropy sum of each subset D_v, where these subsets are partitioned according to the feature value [25]. The higher the IG, the higher is the dependency between the feature X_j and the class label. This method has been used in both single-label and multi-label feature selection tasks.

$$IG(D, X_j) = entropy(D) - \sum_v \frac{|D_v| entropy(D_v)}{|D|} \quad (2)$$

ReliefF [16] is a multi-label method, able to deal with noisy and missing data, extension of a multivariate filter single-label algorithm Relief [15]. The algorithm's goal is ranking the features by estimating their qualities according to the identification of similar features values between a randomly selected example with their nearest different class examples, and different features values for the nearest different classes example [22].

The domain usage of feature selection approaches used in this works is shown in Table 4. Information Gain and ReliefF can be applied both in single-label problems and in multi-label by combining them with problem transformation methods. In this work, Information Gain and ReliefF are applied individually in single-labels and combined with problems transformation methods for multi-label.

Table 4: Feature selection methods and domain usages (single-label and multi-label)

Feature Selection	χ^2	RF Imp	IG	ReliefF
Single-label	✓	✓	✓	✓
Multi-label			✓	✓

3.2 Performance Evaluation

DAQ reduction proposed is grounded in supervised machine learning algorithms. In this way, it is required to states a model towards driving the feature selection algorithms and further comparisons with evaluation metrics.

Evaluation metrics employed in single-label problems such as accuracy, recall, precision, and F-measure cannot be employed in multi-label problems, because of each multi-label instance can be associated with more than one label [14]. In this work, multi-label models are evaluated using Hamming Loss and single-label models using traditional accuracy measure, both explained in the following paragraphs.

A misclassified label results from the association of a wrong label with an example (prediction error) or label absence (missing error). Hamming Loss measures the percentage of misclassified labels, where Δ is the difference between two sets [36].

$$hloss(h) = \frac{1}{p} \sum_{i=1}^{p} |h(\mathbf{x}_i) \Delta Y_i| \quad (3)$$

Accuracy can be described primarily as the percentage of predicted labels correctly divided by the total number of predictions. Assuming that there are positive and negative class. Where true positives (TP) is the number of positive examples labeled correctly. True negatives (TN) is the number of negative examples labeled correctly. False positives (FP) is the number of negative examples labeled as positive. False negatives (FN), is the number of positive examples labeled as negative [17].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Hamming Loss is the fraction of misclassified labels by the total number of labels and accuracy being the fraction of correct predictions by the total number of predictions [29]. Dealing with a binary problem, the accuracy in multi-label can be calculated by $1 - hloss(h)$.

4 MATERIALS AND METHODS

4.1 Questionnaires' Datasets

The information in the four DAQs was collected from organizations located in the northern region of the state of Paraná, Brazil, and the information began to be collected in 2012 [4]. All information is private and was provided by organizations members. The number of organizations providing information on each questionnaire is variant, thus, was selected 20 organizations that provided information in all four axes.

DAQs are composed of a different number of questions (features) and each example has a numeric *result* in a range of 0 to 100 calculated by a human evaluator. To make possible the single-label and multi-label models inductions and predictions, the *result* value was discretized in binary labels, 0 for a $result <= 50$ and 1 for a $result > 50$.

With *result* discretization in binary labels, single-label induction is possible to be performed. Then, for multi-label induction, the single-label problem was transformed into a binary multi-label problem with two labels.

For multi-label features selection purposes, the four original questionnaires ([7],[10],[4],[28]) were merged and the selectors were applied to it. Table 5 shows the number of questions (#Q) per questionnaire and its authors. We dealt with a total of 172 questions.

4.2 Algorithms Implementation

Experiments of multi-label feature selections and problem transformation were carried using the algorithms implemented in MULAN

Table 5: Questions index of each questionnaire in the merged dataset.

Questionnaire	# Q	Authors
Risk	48	Gaffo, F.H. and Barros, R.M. [7]
Infrastructure	25	Isique et al. [10]
Governance	51	Briganó, G.U. and Barros, R.M. [4]
Service Catalog	48	Taconi et al.[28]

[30] version 1.5 and the base learners implemented in MEKA [20], version 1.9. Single-label experiments were carried with WEKA [33] version 3.9. All bases learners were trained using 10-fold cross-validation with default hyperparameters. The comparison graph of the selected questions by each approach was generated using the R framework [19].

Based on assumptions that there may be redundant questions in different DAQ and the merging of these questionnaires lead to increase the complexity for obtaining the desired information, we employ the feature selection algorithm in two experimentation (Evaluation I and Evaluation II), as Figure 1 shows. First, two models were induced for each complete questionnaire, that is, all features in all questionnaires were taken into account. So, single-label feature selection approaches (χ^2, RF Importance, IG, ReliefF) and multi-label feature selection approaches (IG+BR, IG+BR, ReliefF+BR and, ReliefF+LP) were applied in each questionnaire.

Second, concerning the assumption that DAQ merging contributes to increasing the trade-off between questions correlated, questions redundant and answering drawbacks. The selected questions from each DAQ by each single-label features selector were merged. The first containing all selected questions by GA, the second selected by χ^2 and, the third selected by RF Importance from all questionnaires.

Both cases requested a problem conversion to adjust the dataset allowing the usage of feature selection algorithm of a different domain. Adjusts were made converting the problem cardinally by increasing (Evaluation I) or reducing (Evaluation II) the number of labels.

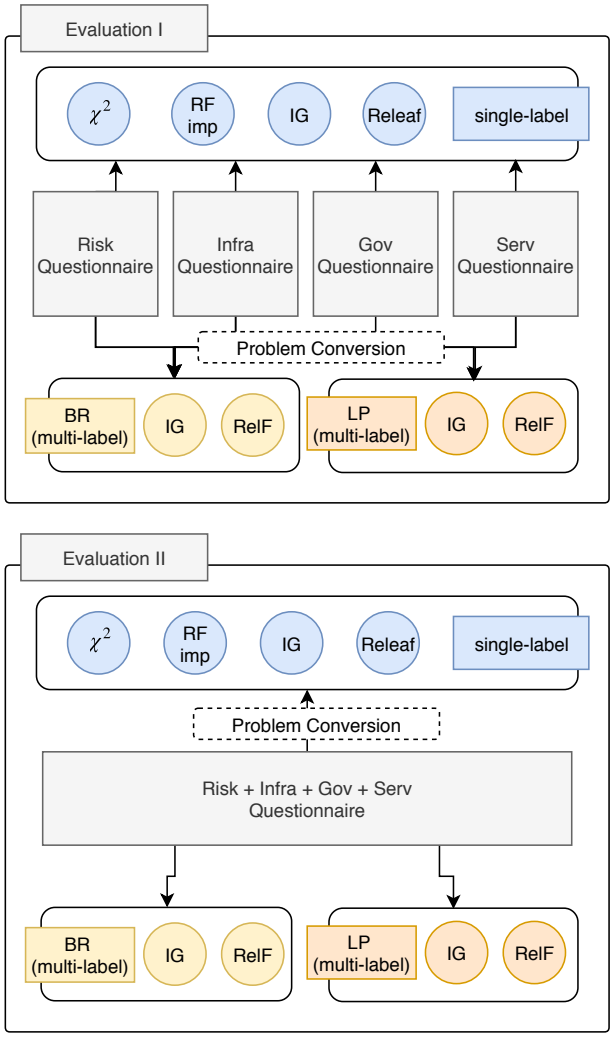

Figure 1: Experimentation overview: Evaluation I and Evaluation II

5 RESULT AND DISCUSSIONS

Result presentation, as stated in 4, was organized as Evaluation I and Evaluation II. In the first, each DAQ was reduced individually by all algorithms. Table 6 shows the number of selected questions (#*Q*) from each questionnaire (columns) according to each feature selection method (rows). Single-label feature selection algorithms obtained similar question reduction in comparison to multi-label. More precisely, χ^2 and IG (including coupled with multi-label methods) were able to reduce by approximately 1.5 times the original number, retaining only 28 of 48 from Risk, 20 of 51 from Governance, 18 of 48 from Service Catalog and 7 from 25 of Infrastructure DAQs. As can be observed, with ReliefF and IG when performed as single-label or coupled multi-label methods, were able to filter a similar number of questions even with different problem transformations methods (BR and LP). RF importance onbtained a regular performance.

The base learners, RF and SVM, were used to induce models to support the evaluation and implementation of feature selection algorithms. Comparing their accuracy in problem modeling, RF presented a slight advantage over SVM. It is important to observe that from 64 evaluations, being 4 questionnaires, 2 learning models and 8 feature selection approaches, 34.37% of evaluations showed accuracy decrease, 26.56% accuracy stability and 39.06% accuracy improvement. In other words, when reducing some questions, the modeling performance increase, reducing the bias between the interviewee answers. In Table 6 colorless cells show accuracy decrease or stability. The grey cells show evaluations where there was an accuracy improvement compared with the predictions using all questions.

A good ICT governance is directly linked to the Infrastructure and Service Catalog theme, as well as the Risk. The decision-making to a new ICT services implantation should be carefully analyzed by the Infrastructure and the current services, taking into account the risks involved, seeking to mitigate them and to develop products or

Table 6: Result for each questionnaire individually (Evaluation I) with original set of features over all algorithms

Algorithm	Risk			Governance			Service Catalog			Infrastructure		
	#Q	RF	SVM	#Q	RF	SVM	#Q	RF	SVM	#Q	RF	SVM
None *(Original)*	48	80%	75%	51	80%	75%	48	70%	60%	25	94%	94%
χ^2	28	85%	75%	20	80%	80%	18	70%	45%	7	100%	84%
RF Importance	31	80%	75%	34	75%	60%	30	65%	45%	15	84%	78%
IG	28	85%	75%	20	80%	80%	18	70%	45%	7	100%	84%
ReliefF	46	80%	75%	38	70%	55%	44	80%	65%	24	94%	89%
IG+BR	28	77% *	75% *	20	85% *	80% *	24	85% *	62% *	8	90% *	72% *
IG+LP	28	77% *	75% *	20	85% *	80% *	21	72% *	62% *	8	92% *	67% *
ReliefF+BR	46	80% *	75% *	38	75% *	80% *	42	77% *	72% *	24	85% *	77% *
ReliefF+LP *	46	80% *	75% *	38	75% *	80% *	45	80% *	70% *	24	85% *	75% *

* Multi-label accurary = $1 - hloss(h)$

services that help achieve the strategic objectives of the organization. In this sense, the maturity about risks and the growth of this maturity must be realized taking into account more specific aspects, such as the identification, answers, and strategies to manage the risks. This concern can be seen in the Risk DAQ, where was possible to reduce in 69.27% using single-label algorithms and 69.79% with multi-label algorithms. The low variance between single and multi-label shown greater expressiveness of relevant questions in Riks DAQ, making it possible for both approaches to select with more certainty the most relevant and to discard the irrelevant questions.

For Governance DAQ using single-label algorithms, it was possible to reduce the number of questions in 54.41%, but using multi-label approaches the reduction was 59%. The reduction for Service Catalog DAQ was 57.29% and 68.75, using respectively single-label and multi-label. And for Infrastructure DAQ the reduction using single-label was 53% and 71% with multi-label.

Merging the questionnaires (Evaluation 2) owing to investigate the correlated, redundant or noisy questions with all questionnaires at once, it was possible to achieve accuracy improvement or the maintenance from the original model. As Table 7 shows, from 16 evaluations, 50% presented accuracy improvement, 18.75% accuracy stability and 31.25% accuracy decrease.

Table 7: Result of the merged questionnaires (Evaluation II) with original set of features over all methods.

Algorithm	#Questions	Accuracy	
		SVM	RF
None *(Original)*	172	73%	73%
ReliefF+LP	171	71%	73%
ReliefF	152	71%	73%
ReliefF+BR	139	70%	75%
RF Importance	110	64%	67%
IG+BR	105	73%	82%
IG	73	76%	77%
χ^2	73	76%	77%
IG+LP	44	78%	79%

The combination of IG+BR for reduction achieved an improvement of RF modeling, reducing considerably 38.95% of the total number of questions with 9% accuracy increasing. If taking into account the accuracy average from SVM and RF predictions, the IG+LP has the best performance. Although it did not reach the best prediction result individually, IG+LP has the best accuracy with SVM and de second best accuracy with RF. Also, it did select the second least number of questions, only 44. This good performance could result from correlated questions evaluation of Label Powerset transformation. IG+LP obtained the best reduction, but in terms of modeling, the improvement was inferior to IG+BR coupled with RF.

ReliefF+LP and ReliefF have not achieved improvements, the two feature selection approaches have reduced a small number of questions, maintaining accuracy stability when coupled with RF and accuracy decrease when coupled with SVM. Reducing about 18% more questions than ReliefF+LP, ReliefF+BR achieved accuracy improvements with RF but has performance loss when coupled with SVM.

Like in Evaluation I, despite a considerable questions reduction RF importance onbtained a regular performance in SVM and RF predictions.

IG and χ^2 maintained similarities as in Evaluation I, selecting the same questions and thus obtained the same prediction performance. As shown in Table 7, the reduction of 57.55% in the number of questions provided performance improvements in the two learning models, but the improvement was inferior to IG+BR coupled with RF.

The Figure 2 shows the selected and the irrelevant questions following each feature selection algorithm, evidencing the gaps of removed questions shared by approaches such as IG+BR, χ^2 and IG+LP.

ReliefF+LP and IG+LP are extremities, selecting respectively the largest and the least number of questions, as Table 7 shows.

Among the sets of most selected (8 and 7 selections) and most irrelevant questions (1 and 2 selections) in Evaluation II, it is important to note questions of Governance DAQ. 23.65% of most selected questions and 85.71% of the irrelevant set belonging to Governance. It is shown that the questionnaire has a set of very significant questions and a set with very irrelevant questions, probably impairing maturity evaluation of this axis. This also can be observed in Table 6 with the significant number of increased accuracy in Governance DAQ with the reduction.

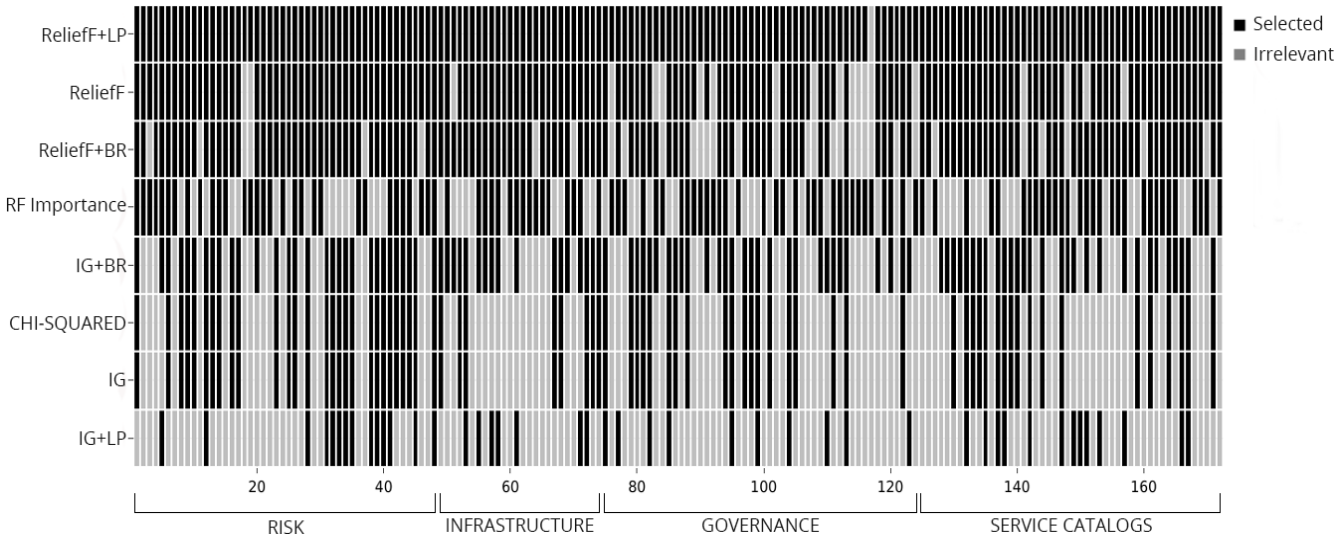

Figure 2: Selected features distribution by each feature selection approach.

Questions of Risk DAQ represents 46.80% of most selected questions and 14.28% of most irrelevant questions in the merged questionnaire. As opposed to results of governance performance, that show greater variability in Table 6 (Evaluation I), 75% of Risk DAQ evaluations had performance stability and demonstrating once again the greater expressiveness of relevant questions this DAQ.

Service Catalog DAQ is present in 21.27% of most selected questions and Infrastructure with only 4.26% and all Service Catalog and Infrastructure questions had at least 3 selections and are not present in the most irrelevant questions group. This shows a disagreement between the feature selection approaches mainly concerning the Infrastructure DAQ, where there is no question in which at least 7 selectors agreed on their irrelevance.

In this work, the goal was an intelligent question reduction on DAQs, as well as ensure a higher prediction performance from the induced models. The considerable reduction of 38.95% in questions generated by IG+BR with RF provides a significant improvement in obtaining data of companies on maturity evaluation process.

6 CONCLUSIONS

Currently, organizations have been engaged in maturity level management evaluation toward evolving. The diagnostic assessment questionnaires are important tools in this scenario. However, these questionnaires, even more, when applied at once could be optimized. This work addressed the problem of the amount of data (questions) required, reducing considerably the questions by merging questionnaires of four areas and applying feature selections algorithms. The induction of machine learning algorithms addresses the expensive evaluation process problem by the generation of models able to classify the companies maturity level in four axes. It was obtained promising results using IG+BR feature selection approach and RF as the learning model, making it possible to achieve 82% of prediction accuracy and reducing the number of questions in 38.95%. This performance was 9% better than the model inducted by the original problem, evidencing the feature selection benefits over DAQs and the usage of inducted models. As future work, we intend to explore additional DAQs in comparison to other feature selection, e.g. Genetic Algorithms.

REFERENCES

[1] A. M. Almeida, R. Cerri, E. C. Paraiso, R. G. Mantovani, and S. B. Junior. Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, 320:35–46, 2018.

[2] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] G. U. Briganó. Um framework para desenvolvimento de governança de tic. Master's thesis, Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2012.

[5] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[7] F. H. Gaffo and R. M. de Barros. Metodologia para avaliar o grau de maturidade da gerência de riscos. In *Anais do IX Simpósio Brasileiro de Sistemas de Informação*, volume 1, pages 242–253, 2013.

[8] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

[9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, Mar. 2003.

[10] L. H. R. Isique, R. M. de Barros, and B. B. Zarpelão. Gaia infrastructure: a framework for the management of information and communication technology infrastructure. In *2015 XLI Latin American Computing Conference (CLEI)*. CLEI, CLEI, 2015.

[11] Iso/iec 38500:2015 information technology – governance of it for the organization. Standard, International Organization for Standardization, Feb. 2015.

[12] I. ITGI. Cobit 4.1. *Framework Control Objective Management Guidelines Maturity Model*, 2007.

[13] U. ItSMF. An introductory overview of itil® 2011. In *The IT Service Management Forum UK, London*, 2011.

[14] S. Kashef, H. Nezamabadi-pour, and B. Nikpour. Multilabel feature selection: A comprehensive review and guiding experiments. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1240, 2018.

[15] K. Kira and L. A. Rendell. A practical approach to feature selection. In D. H. Sleeman and P. Edwards, editors, *Ninth International Workshop on Machine Learning*, pages 249–256. Morgan Kaufmann, 1992.

[16] I. Kononenko. Estimating attributes: Analysis and extensions of relief. In F. Bergadano and L. D. Raedt, editors, *European Conference on Machine Learning*, pages 171–182. Springer, 1994.

[17] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in*

computer engineering, 160:3–24, 2007.

[18] L. Li, H. Liu, Z. Ma, Y. Mo, Z. Duan, J. Zhou, and J. Zhao. Multi-label feature selection via information gain. In *International Conference on Advanced Data Mining and Applications*, pages 345–355. Springer, 2014.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[20] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes. MEKA: A multi-label/multi-target extension to Weka. *Journal of Machine Learning Research*, 17(21):1–5, 2016.

[21] O. Reyes, C. Morell, and S. Ventura. Scalable extensions of the relieff algorithm for weighting and selecting features on the multi-label learning context. *Neurocomputing*, 161:168–182, 2015.

[22] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[23] C. SEI. Cmmi for development, version 1.2. Technical report, CMU/SEI-2006-TR-008, ESC-TR-2006-008, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, 2006.

[24] M. S. Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18, 2010.

[25] N. Spolaôr, M. C. Monard, G. Tsoumakas, and H. D. Lee. A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomputing*, 180:3 – 15, 2016. Progress in Intelligent Systems Design.

[26] E. Spyromitros, G. Tsoumakas, and I. Vlahavas. An empirical study of lazy multilabel classification algorithms. In *Hellenic conference on artificial intelligence*, pages 401–406. Springer, 2008.

[27] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[28] L. H. Taconi, R. M. Barros, and B. Zarpelão. Proposal of a Maturity Model to Deploy a Service Catalog. In *IADIS International Conference on Applied Computing (AC)*. IADIS Press, 2013.

[29] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009.

[30] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12(Jul):2411–2414, 2011.

[31] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning*, pages 406–417. Springer, 2007.

[32] P. Weill and J. W. Ross. Conhecimento em ti: O que os executivos precisam saber para conduzirem com sucesso ti em suas empresas. *São Paulo: M. Books*, 2010.

[33] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[34] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, pages 412–420, 1997.

[35] M.-L. Zhang, J. M. Peña, and V. Robles. Feature selection for multi-label naive bayes classification. *Information Sciences*, 179(19):3218–3229, 2009.

[36] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.