Does Your Accurate Process Predictive Monitoring Model Give Reliable Predictions?

Marco Comuzzi[1(✉)], Alfonso E. Marquez-Chamorro[2], and Manuel Resinas[2]

[1] Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
mcomuzzi@unist.ac.kr
[2] Universidad de Sevilla, Sevilla, Spain
{amarquez6,resinas}@us.es

Abstract. The evaluation of business process predictive monitoring models usually focuses on accuracy of predictions. While accuracy aggregates performance across a set of process cases, in many practical scenarios decision makers are interested in the reliability of an individual prediction, that is, an indication of how likely is a given prediction to be eventually correct. This paper proposes a first definition of business process prediction reliability and shows, through the experimental evaluation, that metrics that include features defining the variability of a process case often give a better prediction reliability indication than metrics that include the probability estimation computed by the machine learning model used to make predictions alone.

Keywords: Business process · Predictive monitoring · Reliability

1 Introduction

The ubiquitous support of information systems and the emerging availability of Internet-of-Things (IoT) technology enable the collection of large amount of data during process execution for process analysis, design and enhancement. Data collected during process execution, usually in the form of *event logs*, are the input of process predictive monitoring, which aims at predicting specific aspects of interests regarding cases currently executing, e.g., which activities are going to be executed next, when a case will terminate, or the value of specific process performance indicators.

Research on business process predictive monitoring recently has focused intensely on the adaptation of existing machine learning techniques to solve new predictive monitoring problems with higher accuracy [4]. From a practical standpoint, however, when making decisions within the scope of an individual process

This work has partially received fundings from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 645751 (RISE BPM), grants TIN2015-70560-R (MINECO/FEDER, UE) and P12-TIC-1867 (Andalusian R&D&I program), and NRF Korea Project Number 2017076589.

case, decision makers such as process owners or users are not only concerned with the accuracy of a prediction model, but also, and often most importantly, with having a means to gauge the *reliability* of an individual prediction. For instance, when deciding whether to renegotiate an agreement with a client to extend the service completion due date or assign more resources to this client to meet the agreed due date, a service provider may rely on a due date predictive model of their internal processes which is 80% accurate on average. However, because these decisions are taken on a per-case basis, a service provider clearly requires a measure to understand to what extent they can rely on, or *trust*, a specific prediction made for each particular client.

In this paper, we focus on predicting business process *outcomes* at the level of individual process instances. Process outcomes, for instance, can be the satisfaction of violation of specific SLA properties, such as process instance execution time being below a certain threshold negotiated by a service provider with customers, or the fulfillment or violation of specific constraints, e.g., regarding order of activities, during process execution.

Accuracy of a predictive model of process outcomes is calculated by aggregating prediction results across a test set of previous cases. As such, it does not give an indication of how much decision makers can *trust* an individual prediction based on new data or, in other words, about the likelihood that a new individual prediction is eventually correct. Machine learning models often define specific metrics for the reliability of predictions, such as the classification probability in decision trees. Other model-independent reliability measures can be defined, which can be based on sensitivity of a prediction or on *transduction*. With sensitivity analysis, a prediction is considered more reliable if the variability of predictions made for similar input data is limited. With transduction reliability is assessed by comparing predictions using models trained with and without a particular new example [1]. However, these measures are based only on the training data and do not take into account features the system generating data used to learn a model.

We argue that, besides the data collected and the chosen learning model, the reliability of an individual prediction may depend on a variety of other factors characterising the *variability* of the *system* generating the data, that is, in our case, the business process. A prediction, for instance, is likely to be more reliable when a process case is almost complete or, more generally, when the choices available to complete a case are limited. Variability may also be associated with the time elapsed to execute a specific case, e.g., the longer a case has been executing, the closer it may be to its termination and, therefore, the fewer the possible choices available for its completion. Part of this knowledge about variability of process cases while making predictions may be already embedded by a predictive model in the learning phase, particularly in the case of complex non linear models, such as neural networks. It is practically impossible, however, to disentangle this knowledge from the internal functioning of a training algorithm in order to obtain a measure of reliability for an individual prediction [1].

The aim of this paper is to put forward the issue of the reliability of individual prediction in business process outcomes predictive monitoring. In order to do so, in Sect. 2 we provide an initial definition of a measure of process prediction reliability that combines prediction probabilities available for the chosen machine learning model with features of a process case, such as its expected completion time or expected number of activities to be executed before case completion. Then, in Sect. 3, we evaluate the proposed metric using a real world business process event log and state of the art predictive monitoring techniques. The results show that metrics that include terms capturing process variability provide a better indication of the reliability of an individual prediction, than any intrinsic reliability metric available for the chosen learning model alone. Conclusions are briefly drawn in Sect. 4. We argue that the issue identified by this paper will spark a new area of research in the field of process predictive monitoring focusing on the definition of prediction reliability metrics well suited for the scenario of business process execution.

2 Model

As a general case of business process outcomes prediction, we consider the scenario of predicting the value of process performance indicators (PPIs) for process cases that are currently executing. Let $\mathcal{P}$, $\mathcal{I}$, and $\mathcal{T}$ represent the universe of processes, PPIs, and the time domain, respectively. Hence, a process $P \in \mathcal{P}$ is associated with M process performance indicators $PPI_m \in \mathcal{I}$. Each indicator $PPI_m \in \mathcal{I}$ assumes value within a domain D_m, which can be numerical or categorical and possibly infinite.

Let $\mathcal{C}_P$ be the universe of cases of a process P. We define the *value* v and *predicted value* $\hat{v}$ of a PPI for a case as follows:

- $v : \mathcal{C}_P \times \mathcal{I} \times \mathcal{T} \rightarrow D_m \cup \{\perp\}$, written $v_j^t(PPI_m)$, mapping a case $j \in \mathcal{C}_P$ and an indicator PPI_m onto a value in the domain D_m at a given time instant t. Note that the undefined value $\perp$ is used when the value of PPI_m cannot be calculated at time t. For instance, the execution time of a case can only be calculated after a case has completed;
- $\hat{v} : \mathcal{C}_P \times \mathcal{I} \times \mathcal{T} \rightarrow D_m \cup \{\perp\}$, written $\hat{v}_j^t(PPI_m)$, mapping a case $j \in \mathcal{C}_P$ and an indicator PPI_m onto a value a *predicted* value in the domain D_m at a given time instant t. A predicted value is obtained using some prediction model trained with data generated during process execution. The undefined value $\perp$ is used when a predicted value cannot be calculated based on available data.

Note that, for a given case j and indicator PPI_m, at any given time t, only one of $v_j^t(PPI_m)$ and $\hat{v}_j^t(PPI_m)$ is available. If for a case j the value $v_j^t(PPI_m)$ cannot be calculated and it is not yet possible to generate a predicted value $\hat{v}_j^t(PPI_m)$, then both value and predicted value are undefined.

The objective of this paper is to define a metric to measure the reliability of predicted PPI values $\hat{v}_j^t(PPI_m)$. In order to be a reliability metric, the proposed metric (i) must assume only values between 0 and 1, with 1 signifying that there

is a 100% likelihood that the predicted value of a PPI is eventually correct, (ii) it must assume the value 1, i.e., 100% reliability, when the actual value of a PPI $v_j^t(PPI_m)$ becomes available, and (iii) it should increase as the likelihood of a predicted value to be correct increases. Properties (i) and (ii) are guaranteed by design in the proposed definition given below. Property (iii) drives the design of the proposed reliability metric and its achievement is demonstrated by the experimental evaluation.

The problem of defining a reliability metric for SLA prediction is the problem of defining a function $r : \mathcal{C}_P \times \mathcal{I} \times \mathcal{T} \rightarrow [0, 1]$, written as $r_j^t(PPI_m)$, to indicate the reliability of an individual predicted value of an indicator PPI_m for the j-th case of process P at time t.

In this paper, we propose the following definition of $r_j^t(PPI_m)$:

$$r_j^t(PPI_m) = w_1 \cdot adv_j^t(PPI_m) + w_2 \cdot time_j^t(PPI_m) + w_3 \cdot pred_j^t(PPI_m).$$

That is, with $\sum_w = 1$, the proposed reliability metric is comprised of the weighted sum of the following 3 terms: (i) $adv_j^t(PPI_m)$ considering the advancement in execution of the j-th instance at time t; (ii) $time_j^t(PPI_m)$ considering the time elapsed since the j-th instance has started; and (iii) $pred_j^t(PPI_m)$ refers to a value of probability estimate of the prediction as defined by the prediction technique in use, e.g., prediction probability in decision tree-based classification. Note that this value may not be available when the prediction technique in use does not provide any kind of prediction reliability.

Regarding the first term $adv_j^t(PPI_m)$, let l_j be the number of activities executed thus far in the j-th case. We assume that an estimate of the remaining number of activities to be executed in the j-th case $\hat{l}_j$ is available. This estimate can be obtained is several ways, e.g., naively by considering the average number of remaining activities in all previous cases that matched the execution thus far of the current case, using a predictive monitoring technique [4], by matching the current execution trace with the most similar previous case, or by considering the average number of activities on all possible paths to complete a process execution, possibly weighted by the probabilities of taking specific paths, if available.

Then, $adv_t^j = f(l_j, \hat{l}_j)$, where f is a monotonic increasing *activation function* with values between 0 and 1 and $\lim_{\hat{l}_j \rightarrow 0} f(l_j, \hat{l}_j) = 1$, e.g., $f(l_j, \hat{l}_j) = \frac{l_j}{l_j + \hat{j}_l}$.

The second term $time_j^t(PPI_m)$ also relies on an estimate of the remaining time to complete the execution of a process case. This can also be calculated in several ways, e.g., using predictive monitoring or by averaging the remaining execution time of previous similar cases available in an event log.

Let t_j^{ex} be the time elapsed from the start of case j and $\hat{t}_j$ the estimate of the remaining time required to complete case j, then $time_t^j = f(t_j, \hat{t}_j)$, where f is a monotonic increasing *activation function* with values between 0 and 1, e.g., $f(t_j, \hat{t}_j) = \frac{t_j}{t_j + \hat{t}_l}$.

3 Evaluation

This section presents a simple experimentation to assess the validity of the proposed reliability metric definition. In particular, our aim is to show that higher prediction reliability is achieved when the weights of the *adv* and *time* terms in the reliability metric definition are not zero, to show that including information about case variability improves reliability of predictions.

A real-life event log from the IT Department of a regional public administration was used in the experimentation. This dataset represents an incident management log. In this scenario, a service level agreement (SLA) is established considering certain key performance indicators (KPIs). This SLA determines the penalties derived from the under-fulfillment of a threshold for each of the KPIs. Thus, predictive monitoring is necessary to anticipate the possibility of violating the SLA and, therefore, incurring into penalties. We consider one specific KPI (named K20), which indicates abnormal idle time during the resolution of an incident. An incident management case, in fact, may remain idle due to a variety of reasons, such as unavailability of personnel or scheduling errors. Idle time is clearly unproductive and should be avoided. The KPI K20 states that idle time should not exceed a certain threshold in any given case.

This event log consists of 174.989 process cases, each of them with 15 attributes. Beside standard attributes of events in event logs, such as activity name, timestamp, and resource, each event contains additional information about, for instance, the type of incident, its priority or the service center to which it was assigned[1]).

For our experimentation, we have divided this dataset in three parts: 60% as training set, 20% for validation and 20% as test set. We have trained the model with the training set and estimated the different parameters related to the reliability using the validation set. As encoding we have considered a sliding window of 2 events, since empirical evaluation has showed this is a good window size. Each feature vector is composed by the different attributes of the 2 events of the event window, while the last position corresponds to the class, which indicates a value of the KPI to be predicted. Each attribute can be nominal or a real number. More detailed information of the encoding is provided in [3].

The reliability of predicted values of indicator K20 is $rel_j^t(K20) = w_1 \cdot adv_j^t(K20) + w_2 \cdot time_j^t(K20) + w_3 \cdot pred_j^t(K20)$. To compute the term $adv_j^t(K20)$, we use the activation function described above using the average number of activities in all previous cases as an estimate of the remaining number of activities. A similar method has been adopted for $time_j^t(K20)$, i.e., we have considered the elapsed time from the beginning of the case for each activity. Then, the average total execution time of the cases is calculated, and this is used as an estimate of the remaining execution time for each validation case. A decision tree algorithm has been used as model. Then, the third term $pred(K20)$ is the predicted class probability for each case of the validation set.

[1] Description of the attributes can be found at https://goo.gl/ye68ei.

We have computed the values of reliability $rel_j^t(K20)$ for each case j in the event log and each possible combination of the weights w_1, w_2, w_3, sampling weights values at intervals of 0.1. To assess the validity of the reliability metric, we have divided the predictions according to their calculated reliability value in intervals of size 0.1. For instance, the reliability interval $(0.3, 0.4]$ contains all predictions for which $0.3 < rel_j^t(K20) \leq 0.4$. Then, we have defined a sensitivity-based estimation of prediction errors for all the intervals as follows. For each interval, we first obtain the number of correct (P) and incorrect (NP) predictions to determine the sensitivity ($Sens$) of the prediction, with $Sens = P/(P+NP)$. Then, we determine the deviation of this value $Sens$ from the center of the interval. If the proposed reliability metric is valid, for instance, this means to assume that the interval $(0.3, 0.4]$ should contain approximately 35% of correct predictions, 45% for the next interval and so on. Finally, we have obtained the average error (avg_err) for all deviations and all possible values of weights. An extract of the results showing the combinations of weights values with lowest average error is shown in Table 1.

Table 1. Experimental results.

w_{adv}	w_{time}	w_{pred}	avg_err	P_corr	p_value
0.0	0.0	1.0	0.2164	0.6162	0.1926
0.1	0.0	0.9	0.2167	0.1126	0.8318
0.2	0.0	0.8	0.1784	0.6332	0.1771
0.3	0.0	0.7	**0.1532**	0.8342	0.0389
0.2	0.1	0.7	**0.1544**	0.9269	0.0078
0.3	0.2	0.5	0.1633	0.4258	0.3997
0.5	0.0	0.5	0.1533	0.6184	0.1906

We can appreciate how the average error decreases when considering the terms *adv* and *time* in the reliability definition. For instance, for $w_{pred} = 1$ the average error is 0.2164, while the best results are achieved for $w_{adv} = 0.3$ and $w_{pred} = 0.7$ ($avg\_err = 0.1532$) or $w_{adv} = 0.2$, $w_{time} = 0.1$ and $w_{pred} = 0.7$ ($avg\_err = 0.1544$). An improvement of 5% points is achieved in these cases by including the variability terms *adv* and *time*.

The fifth column of Table 1 shows the Pearson correlation coefficient between the sensitivity measure and the center of intervals. These two values should be correlated for an accurate estimation of the reliability [2]. The significance level ($\alpha \leqslant 0.05$) of the correlation is reported in the sixth column. As we can see, the positive correlation has statistical significance for the best combinations of weights cited above.

To summarise, even if very limited, our experimentation shows that a reliability metric that includes terms capturing process variability is more likely to estimate correctly the probability that a prediction will be eventually correct than considering only a reliability parameter typical of the chosen machine learning model, i.e, predicted class probability of decision trees in our case.

4 Conclusions

The objective of this paper has been to signal the need to define metrics of reliability of individual process predictive monitoring predictions and outline some preliminary ideas on how to face it. The empirical results, in particular, highlight that prediction reliability is higher when terms capturing case variability are included.

Future work should look at refining the initial definition provided in this paper and at assessing the impact of a high quality reliability metric on real world process predictive monitoring use cases.

References

1. Bosnić, Z., Kononenko, I.: An overview of advances in reliability estimation of individual predictions in machine learning. Intell. Data Anal. **13**(2), 385–401 (2009)
2. Bosnić, Z., Kononenko, I.: Estimation of individual prediction reliability using the local sensitivity analysis. Appl. Intell. **29**(3), 187–203 (2008)
3. Márquez-Chamorro, A., Resinas, M., Ruiz-Cortés, A., Toro, M.: Run-time prediction of business process indicators using evolutionary decision rules. Expert Syst. Appl. **87**(Supplement C), 1–14 (2017)
4. Marquez-Chamorro, A.E., Resinas, M., Ruiz-Cortes, A.: Predictive monitoring of business processes: a survey. IEEE Trans. Serv. Comput. **11**, 962–977 (2017)