Enhancing the Quality of Predictions in Predictive Business Process Monitoring

Jongchan Kim[1]

Ulsan National Institute of Science and Technology (UNIST), 50, UNIST-gil, Ulsan 44919, Republic of Korea
jckim@unist.ac.kr

Abstract. The aim of this thesis is to develop and evaluate methods to enhance the quality of predictions for predictive business process monitoring, focusing on the accuracy and stability of predictions. Three different approaches to increase the accuracy and stability of predictions are suggested. Firstly, to improve the accuracy of predictions on minority classes, different resampling techniques are applied to data samples in event logs and the accuracy of the predictions from these resampling techniques are compared, and new metric that considers different weights of activities is developed. Secondly, the stability of predictions of different data distribution-resampling technique pairs is compared. Lastly, the stability of predictions of different case types in event logs is compared, and a new performance metric that considers different case types and provides balanced predictions is suggested. Besides evaluation using publicly available event logs, a case study is also conducted using a real-life event log from a large hospital in South Korea.

Keywords: Business process · predictive monitoring · stability.

1 Introduction

The goal of predictive business process monitoring is to predict various aspects of interest of business processes, such as the next event in a case [2], process outcomes [4], or remaining time in a case [5], by enabling real-time prediction of the features based on an incomplete event log. Predictive business process monitoring is of paramount importance since not only does it help decision makers be better prepared for what is going to happen in the future, but it also prevents them from facing huge costs due to highly uncertain business environments. In order to maximize the advantage of predictive business process monitoring, it is widely acknowledged to increase the prediction accuracy. Hence, all approaches in the literature are evaluated considering the accuracy of predictions on a set of real world and/or synthetic event logs. Accuracy improvements can be achieved either by modifying the data preprocessing scheme or training algorithms. Firstly, incomplete traces can be manipulated using different encoding techniques in a preprocessing step [10]. As different encoding techniques are implemented, the way in which the traces are characterized changes, thereby leading to better

representation of traces for classifiers to train. Secondly, different classifiers can be employed to fit the training data followed by the rising number of advanced classifiers proposed in machine learning field that show superior performances [8]. In recent years, growing number of researchers have intensively conducted research on predictive business process monitoring, applying various techniques to enhance the accuracy of predictions in predictive business process monitoring.

While techniques proposed in the literature are thoroughly evaluated from the point of view of performance, e.g., accuracy of f-score of predictions, we argue that little emphasis has been put on the practical applicability of the prediction results. Therefore, it remains unclear to what extent predictive monitoring can improve the business processes in practice. Specifically, the data mining and machine learning literature stresses that other dimensions of *quality* of the results achieved by a model, besides accuracy, should be considered in order to achieve a comprehensive evaluation of a predictive model. In particular, the accuracy and stability of the predictions have not been examined enough in predictive business process monitoring, while they are deemed important in data mining and machine learning [3].

For a clearer understanding, let us consider the following two examples. First, let us assumes that a manufacturing company needs to detect anomalous traces in a process as early as possible in order to minimize the cost of defects resulting from anomalous process execution. Since anomalous traces usually do not account for the majority of traces, enough data samples for anomalous traces may not be available, which hinders the capability of a classifier of learning a model of decent quality for fitting the anomalous traces. Assuming that the anomalous traces account for 10% of traces, and the prediction accuracy measured by f-score is 0.9 and 0.1 for the normal traces and anomalous traces, respectively, the average accuracy of this classifier is 0.82, even though the prediction accuracy of the anomalous traces is only 0.1. To conclude, the gap between the prediction accuracy of the minority class and the average prediction accuracy is immense, thereby making the classifier virtually unusable in practice. In the second example, let us assume that an insurance company wants to predict the next activity in running processes in order to efficiently handle the ongoing workload. In this case, predictive monitoring would be useless in practice if predictions and performance are not stable across a variety of dimensions, such as time, i.e., predictions are different with every different trial, or distribution of case types, i.e., predictions are very accurate for specific case types and highly inaccurate for others. While research has analysed the temporal stability of predictions in outcome-based predictive monitoring [9], a more general analysis of quality of predictions across different dimensions of interest remains largely undiscovered.

In this context, this thesis aims to bridge the gap between academia and practice by proposing techniques for enhancing the accuracy and stability of predictions in predictive business process monitoring. This thesis will focus on the prediction of next activity in a case and process outcomes. As the prediction of next activity is a sequential multi-class classification task, while the prediction of the overall outcome is a non-sequential binary classification task, preprocessing

schemes for these two types of predictions will be different. For the data, we will consider publicly available event logs and also real-life event log from a large hospital in South Korea. A case study will be presented using the real-life event log in order to enhance the practical applicability of the results of the thesis.

This thesis was scheduled to be completed in three years, where approximately two years are remaining for now. During the first year, the literature on accuracy and stability of predictions in predictive business process monitoring, resampling techniques in machine learning, and trace clustering techniques in process mining were thoroughly investigated. In addition, different resampling techniques for imbalanced learning were partly implemented. For the second year, the remaining analysis of the first approach and a case study will be conducted. For the case study, the collection of real-life event log from the hospital will be finished and the evnet log will be preprocessed. Especially in the process of collecting the real-life event log, domain experts in the hospital such as doctors and nurses will be interviewed for the advice of different weighting schemes for different activities. In addition, probability distributions will be fitted to either augmented or reduced data for the second approach. In the last year, stability of the prediction from each "data distribution - resampling technique" pair will be calculated to figure out the optimal conditions for stable predictions. Finally, various trace clustering techniques will be applied and the stability of the prediction from each trace clustering technique will be calculated for the analysis of third approach. After the experiments, the limitation of the paper will be unveiled and concrete scope of the future research will be determined based on the assessed limitations.

2 Approach

At the current stage of the research, three existing issues in accuracy and stability of the predictions in predictive business process monitoring have been identified. These three issues concerns predictive monitoring use cases that are solved using classification techniques, i.e., prediction of next activity in a running trace and prediction of process outcomes.

2.1 First approach: Comparison of the accuracy of predictions of different resampling techniques for imbalanced learning

The imbalance in class ratio of features in an event log decreases accuracy of the predictions. Most popular classification techniques show poor performance when they have imbalanced data as input data [7]. Therefore, this thesis compares different resampling techniques to balance the class ratio of features and shows how resampling techniques improve the accuracy of the predictions, and the detailed framework is presented in Fig. 1. The resampling techniques to be used in the thesis are composed of over-sampling techniques and under-sampling techniques. For performance measurement, a new metric that calculates prediction accuracy based on modified weights of activities is developed, where the

weights are modified with the help of domain experts. In a nutshell, this new accuracy metric gives larger penalty to the wrong prediction of a class that has higher weight.

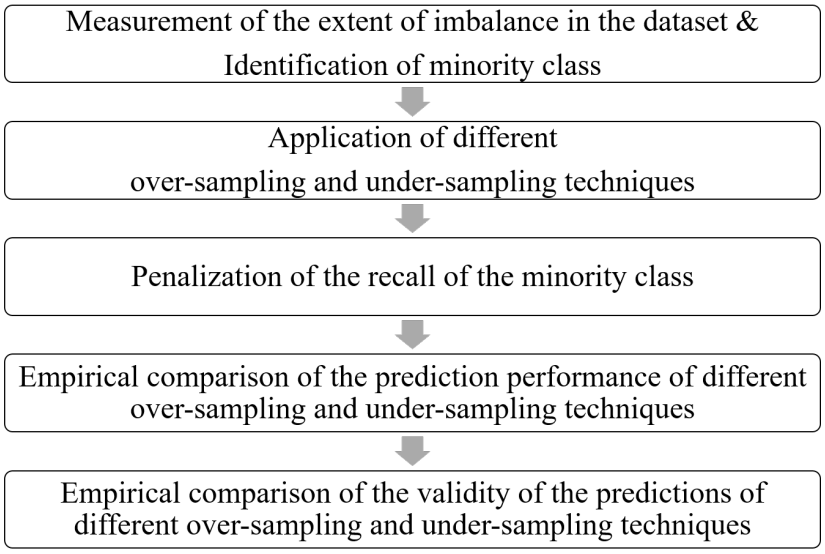

Fig. 1. The framework of the first approach.

Over-Sampling Over-sampling techniques augment the number of samples in the minority class by sampling from the minority class. Three existing oversampling techniques will be compared.

Under-Sampling Under-sampling techniques delete samples from the majority class. Ten existing undersampling techniques are compared.

In practice, the problem of imbalanced class ratio becomes more evident when the minority class is considered to be a critical task in the operation of the process. In this case, wrong prediction of minority class may bring huge damage to those who run the corresponding process even though the classifier is good at predicting the majority class. In order to account for the importance of predicting the minority class, the penalty is given to the recall of the minority class, thereby modifying the f-score. To better demonstrate the effectiveness of resampling techniques, a case study will be presented using a real-life event log from a large hospital in Korea.

2.2 Second approach: Comparison of the stability of predictions of different "data distribution-resampling technique" pairs

Different data distribution may affect the stability of the predictions. Once the resampling technique is applied, the data will either be augmented or deleted. In addition, the way in which the data is augmented or deleted varies depending on which resampling technique is applied. In other words, the distribution after the application of resampling techniques can be totally different from the distribution of the data where different resampling techniques are applied, even though the initial data distribution was identical. Therefore, this thesis compares prediction performance and the stability of different "data distribution-resampling technique" pairs as in Fig. 3. In a bimodal distribution where the value of one mode is slightly higher than the other mode, for instance, one oversampling technique may augment the number of minority class by focusing only on increasing the samples that belong to the mode that has higher value, whereas the other oversampling technique may augment the number of minority class by increasing the samples that belong to both modes. In this case, the stability of the classifiers of two oversampling techniques is expected to be different. Moreover, there may exist specific types of classifiers that can generate more solid rules for fitting the data depending on different distribution of the input data. Therefore, it is expected that the best classifier for each resampling technique will be revealed after experiments.

In addition, effects of different sizes of the data modified by resampling techniques on accuracy and stability can be an another subtopic of interest in this context. For example, over-sampling techniques may relatively show greater accuracy and stability on huge dataset, but not in a case where the majority class has 10 samples and minority class has only 1 sample. In this case, under-sampling techniques may relatively show greater accuracy and stability where the over-sampling technique has to over-sample based on only 1 sample, thereby resulting in a poorer accuracy and stability. Here, we will use case-based stratified sampling that samples data based on case-level, which means that the distribution of the data is maintained by deleting cases, not by deleting individual events.

2.3 Third approach: Comparison of the stability of predictions of different case types

Different case types may affect the stability of the predictions, which means that the prediction accuracy may fluctuate at times depending on different case types. Therefore, this thesis compares prediction performance and the stability of different case types as in Fig. **??**. There have been several research regarding grouping the cases based on features [6, 1], which is frequently called as trace clustering. The idea of comparing the stability of predictions of different case types stem from the fact that different case types have different variations in their features. In addition, consideration of different case types can be helpful in practice, since the classifier can offer tailored prediction results for each case type, thereby providing more stable predictions for each case type. If not, the

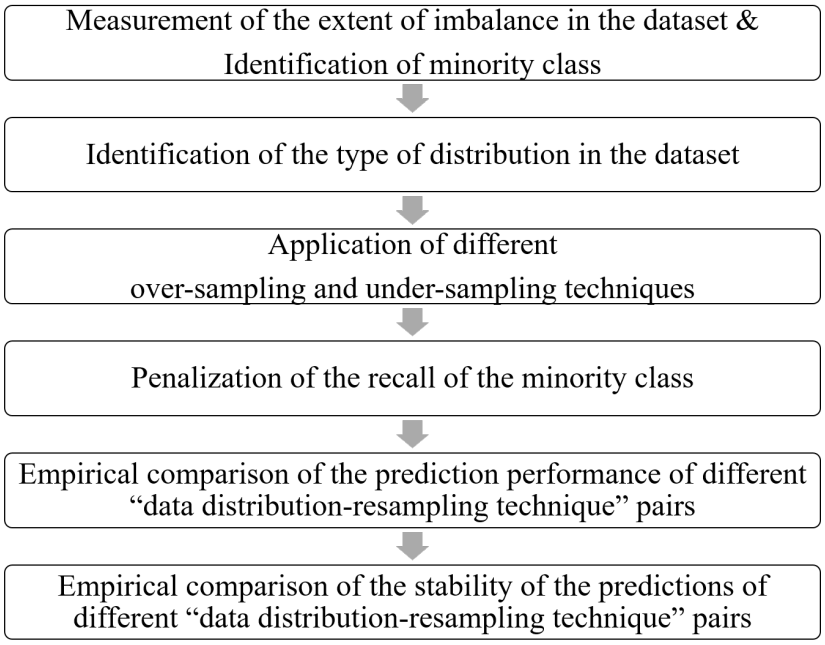

Fig. 2. The framework of the second approach.

prediction accuracy may fluctuate across each trial depending on the distribution of different case types in a training set. The case types will be identified by trace clustering techniques suggested in BPM literature and other clustering techniques suggested in traditional machine learning literature. After the identification of different case types, new performance metric that account for balanced predictions across case types will be devised.

3 Discussion and Conclusion

It is expected in the first approach that the prediction accuracy of classifiers with resampling techniques calculated by new metric would be higher than that of a baseline classifier without any resampling techniques. Beyond these results, it is expected in the second approach that specific distribution of event logs whose samples are either augmented or deleted may show stable prediction accuracy with specific classifiers. Similarly, in the third approach, it is expected that clusters of cases from some trace clustering techniques may exhibit stable prediction accuracies across every trial. For future works, four improvements can be made. Firstly, further efforts can be made on developing ways to give penalty to the wrong predictions of the minority class. Throughout this paper, it was suggested that f-score should be modified by penalizing the recall of the minority class, which means that the penalty is imposed after the classifier is fitted and

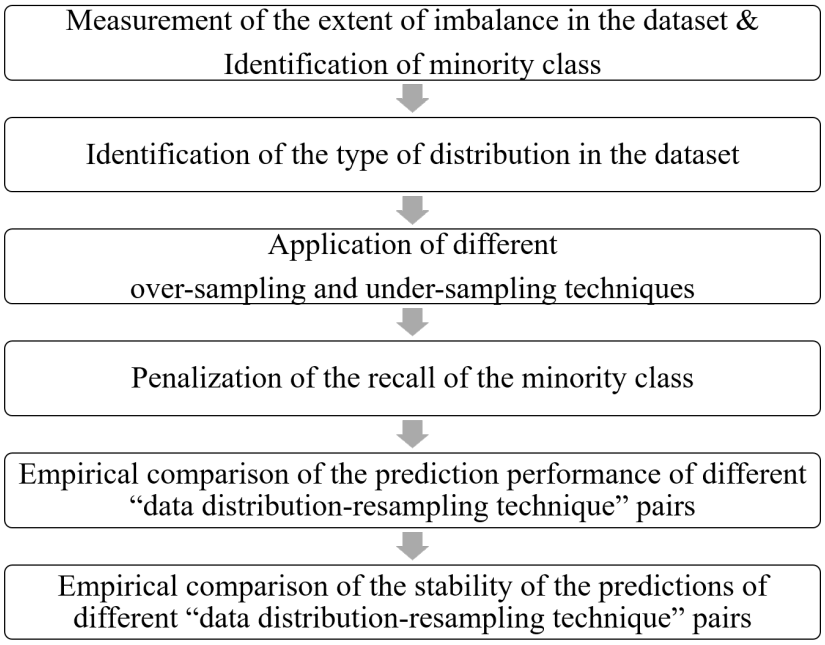

Fig. 3. The framework of the second approach.

completes the prediction. However, the penalty can be given not only by modifying the f-score but by modifying the way the classifier calculates the error in learning procedure. In case of neural network-based classifiers, cost function can be modified, while in tree-based classifiers, cost complexity or entropy can be modified. In these cases, conventional performance metrics will be applied without any modification. Secondly, better ways to define "important" minority class can be developed. This is because in real-life cases, there can be a huge number of minority classes, while some of these cases are not "important" classes but merely minor or deviant classes. Thirdly, other advanced quality metrics of predictions from the literature in machine learning except for the stability would be implemented to improve the quality of predictions for predictive business process monitoring. Lastly, generative models can further be applied in oversampling the minority class since not enough over-sampling techniques are considered in this thesis compared to under-sampling techniques.

References

1. R. J. C. Bose and W. M. Van der Aalst. Context aware trace clustering: Towards improving process mining results. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 401–412. SIAM, 2009.
2. J. Evermann, J.-R. Rehse, and P. Fettke. Predicting process behaviour using deep learning. *Decision Support Systems*, 100:129–140, 2017.

3. M. Kukar and I. Kononenko. Reliable classifications with machine learning. In *European Conference on Machine Learning*, pages 219–231. Springer, 2002.
4. A. Metzger, P. Leitner, D. Ivanović, E. Schmieders, R. Franklin, M. Carro, S. Dustdar, and K. Pohl. Comparing and combining predictive business process monitoring techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(2):276–290, 2015.
5. A. Rogge-Solti and M. Weske. Prediction of remaining service execution time using stochastic petri nets with arbitrary firing delays. In *ICSOC*, pages 389–403. Springer, 2013.
6. M. Song, C. W. Günther, and W. M. Van der Aalst. Trace clustering in process mining. In *BPM*, pages 109–120. Springer, 2008.
7. Y. Sun, A. K. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009.
8. N. Tax, I. Verenich, M. La Rosa, and M. Dumas. Predictive business process monitoring with lstm neural networks. In *International Conference on Advanced Information Systems Engineering*, pages 477–492. Springer, 2017.
9. I. Teinemaa, M. Dumas, A. Leontjeva, and F. M. Maggi. Temporal stability in predictive process monitoring. *Data Mining and Knowledge Discovery*, 32(5):1306–1338, 2018.
10. I. Teinemaa, M. Dumas, M. L. Rosa, and F. M. Maggi. Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):17, 2019.