

Zusammenfassung

Das Forschungsfeld *Process Mining (PM)* hat sich seit der Erfindung des Alpha-Algorithmus (Ende der neunziger Jahre) stark entwickelt, wobei ein Großteil der Forschung im Bereich der unstrukturierten Prozesse stattfindet. Zu den unstrukturierten Prozessen gehören beispielsweise Behandlungen von Patienten im Krankenhaus.

Eines der Ziele, das mit PM verfolgt wird, ist es, Prozesse transparent zu machen. Bei strukturierten Prozessen spielt diese Problematik kaum eine Rolle, da Prozesse im Herstellungsbereich in der Regel standardisiert und dokumentiert sind. Deshalb wird in dieser Arbeit versucht, auf eine andere Weise mit PM einen Mehrwert im Herstellungsbereich zu generieren.

Um das oben genannte Ziel zu erreichen, wird PM in den *Cross Industry Standard Process for Data Mining (CRISP-DM)* integriert. Als Bestandteil dieses Standards wird versucht, mithilfe von verschiedenen PM-Techniken und Metriken, Merkmale aus dem Herstellungsprozess zu extrahieren. Danach wird auf Basis dieser Merkmale mit unterschiedlichen maschinellen Lernverfahren eine Vorhersage hinsichtlich der Freigabedauer einer Charge gemacht. Das bedeutet auch, dass im Gegensatz zu anderen Arbeiten in diesem Bereich, extrahierte Informationen aus einem Prozess (dem Herstellungsprozess) zur Vorhersage eines anderen Prozesses (des Freigabeprozesses) dienen. So wird in dieser Arbeit versucht, eine Brücke zwischen PM und dem Forschungsgebiet *Data Science* zu schlagen.

Das Verwenden der PM-Techniken führt auf zwei Wegen zur Merkmalsidentifizierung. Einerseits indirekt, da durch das Anwenden der PM-Techniken Domainwissen erarbeitet werden kann, das auf andere Weise nicht oder schwer möglich gewesen wäre. Andererseits direkt, da gefundene Prozessvarianten und Prozessmetriken direkt in Merkmale übersetzt werden können. Die Vorhersagen mit den maschinellen Lernverfahren waren nicht genau genug, da das beste Verfahren einen *Root Mean Squared Error (RMSE)* von 89 Stunden aufwies. Bei einer Gesamtdauer von 290 Stunden ist dieser Fehler relativ gesehen zu hoch.

Abstract

The research field of PM has advanced greatly since the development of the alpha algorithm in the late 1990s, with much of the research devoted to the area of unstructured processes. Unstructured processes include, for example, treatments of patients in hospitals.

One of the goals pursued with PM is to make processes transparent. This is rarely an issue in the case of structured processes, since processes in the manufacturing sector are usually standardized and well-documented. Therefore, this study seeks to generate added value in the manufacturing area in a different way with PM.

To achieve this goal, PM is integrated into the CRISP-DM. As part of this standard, an attempt is made to extract features from the manufacturing process using various PM techniques and metrics. Then, based on these features, different machine learning algorithms are used to make a prediction regarding the release time of a batch. This means that, unlike in any other work in this area, extracted information from one process (the manufacturing process) is used to predict the outcome of another process (the release process). Thus, this work attempts to explicitly combine PM and Data Science.

Using PM techniques leads to feature identification in two ways. On the one hand features are identified indirectly, since domain knowledge can be acquired by applying PM techniques. This would have been difficult or even impossible to obtain by other means. On the other hand, features are extracted by transforming process variants and process metrics directly into features. While acquiring potential features with PM techniques works well, the predictions made with the machine learning algorithms were not accurate enough. The best algorithm (Random Forest) achieved a *RMSE* of 89 hours. This is relatively high compared to the average total duration of a release process which is 290 hours. Therefore, there are no current practical applications for the models that were obtained during this study.