A data mining process for building recommendation systems for agricultural machines based on big data

Recommendation system for agricultural machinery application

Mohamed Altaleb[1,2], Henning Deeken[1,2] and Joachim Hertzberg[1]

Abstract: There is a potential expansion in the agricultural machinery industry by using the collected data from different years. Big data is already being used in other industries like e-commerce to improve decision-making processes. There are several existing process models to lead through the generic processes of data mining. The common factor between the process models that have attained dominant public position is that they are domain-agnostic frameworks. This paper proposes a method to extend the CRoss-Industry Standard Process for Data Mining (CRISP-DM) to focus on the agricultural domain and give guidelines on how to handle and structure the agricultural data and processes to reach defined data mining goals. The paper provides a walk-through for a use case to build a recommendation system.

Keywords: agricultural machinery, data mining, process model, recommendation system

1 Introduction

In the last two decades, different researchers, institutes, and companies have introduced several data mining process models to structure data science projects. e.g., ASUM-DM by IBM, SEMMA by SAS, and CRISP-DM by several companies, for comprehensive description refer to [Ma21]. They describe the generic process models to guide through the standard phases and steps of data mining. These process models ensure transparency in the communication within the project, and they help to plan and achieve structured results.

In the modern agricultural industry, it is common to collect process and machinery data. Using this data with the help of data mining can bring potential growth by exploiting the valuable information contained in that data. Other industries like e-commerce use big data, e.g., to build recommendation systems. These systems propose purchasable items as recommendations to potential customers based on historical data and exploit insights into the behavior of similar customers. In the agricultural machinery industry, likewise, big data is collected and can be used to build recommendation systems, e.g., to propose better ma-

[1] Osnabrück University, Institute of Computer Science, Wachsbleiche 27, 49090 Osnabrück, henning.deeken@uni-osnabrueck.com, joachim.hertzberg@uni-osnabrueck.de
[2] CLAAS E-Systems GmbH, Advanced Engineering, Sommerkämpen 11, 49201 Dissen am Teutoburger Wald, mohamed.altaleb@claas.com, henning.deeken@claas.com

chine settings using historic telemetry. Similarly, data mining can improve current optimization processes through a better understanding of the common traits of agricultural machinery across different operational and environmental contexts.

2 Problem Statement

In the agricultural machinery industry, the optimization of machines is a crucial task. For many agricultural processes, onboard optimization systems that use closed-loop approaches to configure the machines using internal telemetry have been proposed [Es20]. While these systems are beneficial, they are usually limited to local optimizations of a single machine based on the logged data during the ongoing field operation.

Information transfer from previous operations and analysis across multiple machines is rarely seen in the design of systems targeted to optimize machine settings, e.g., for tractors and harvesters. The exchange of information across different fleets of machinery requires better understanding for the different involved entities, machines, environment, and quality measures. This is where data mining can play a major role as the right tool to be used in order to convert the big data from just recorded telemetry signals into useful information that describes how different entities are associated and connected.

CRISP-DM is a uniform framework and an open standard made for industrial data science projects [Sh00]. As such, it was developed to work with any data mining tools and to structure any data mining problems. It describes six phases, each with sub-steps and tasks, to perform business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Although CRISP-DM has achieved a dominant position by public acceptance [Ma09], it is sometimes considered too generic within the data mining community [Ni15; Hu19]. At CLAAS, CRISP-DM is used to structure data science projects. Since it is a domain-agnostic framework, we realized the need to extend it with an extra layer of process description that directly relates to the agricultural domain. The contribution of this paper is a supplementary data mining process coupled with CRISP-DM and targeted to data mining in the agricultural machinery domain.

3 Materials and Methods

The six phases of CRISP-DM are shown in Figure 1.a. They capture the possible routines of a standard data mining process. Therefore, the process does not include any domain-relevant information. While performing data mining in a specific domain, however, the specifics of this domain are always subject to some considerations. In the agricultural machinery domain, machine optimization is a key aspect, which relates to three major characteristics, quality metrics, environment conditions, and machine variance. These terms

are the particular components in any agricultural machinery application, e.g., a machine uses certain settings under specific environmental conditions when particular process quality is required.

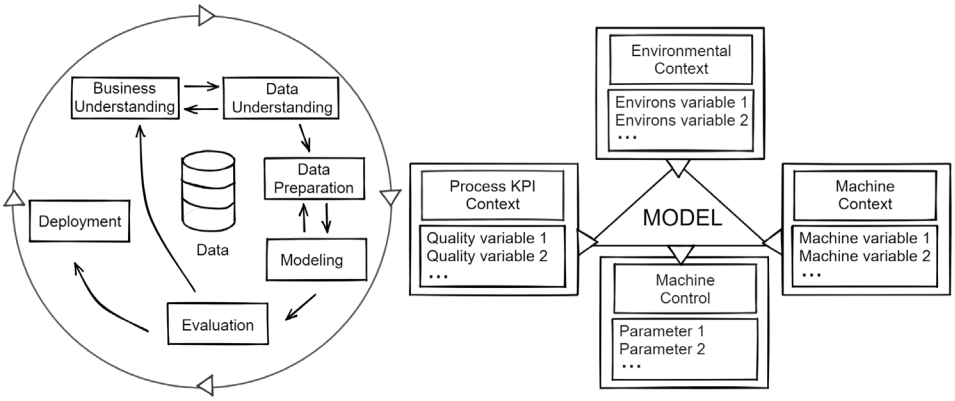

Fig. 1: Since agricultural machinery optimization relates to machine, environment, and process quality, the paper proposes a model that extends the generic phases of CRISP-DM, data understanding, data preparation, and modeling by the context relevant information.

An example of a use case for combine harvesters is shown in Figure 1.b, where the process of optimization configures the machine settings (i.e., 'Machine Control') based on the environmental conditions (i.e., 'Environmental context') and the quality process requirements (i.e., 'Process KPI context'), where these considerations hold merely between similar machines (i.e., 'Machine Context'). Consequently, the configuration of an agricultural machinery application can be described with the Agricultural Operational Context (AOC), under similar operational conditions, similar machine settings can be used.

The AOC can be defined as the combination of three pieces of the context information, environmental, machinery, and agrarian process quality. Each of this information can be described as a list of agriculture-related machine variables. The AOC might be defined based on the selected decision variables within the lists, so it depends on the studied use case. On the input of the data mining process, these lists contain candidate decision variables, and on the outputs, the lists contain the significant related decision variables for the desired model.

In any agricultural machinery system, the machine has to be configured optimally because this directly affects the profit. Therefore, finding the correct decision variables is an essential step, as the decision variables will be the decision criteria that express how the system should behave. Data-driven models need decision variables to control the way the desired model behaves. The CRISP-DM considers these variables implicitly in the first

four phases. This paper proposes a data mining model for the agricultural machinery industry that explicitly refers to the decision variable. It defines a scheme on how to deal with the different measurement signals in an agricultural machinery dataset.

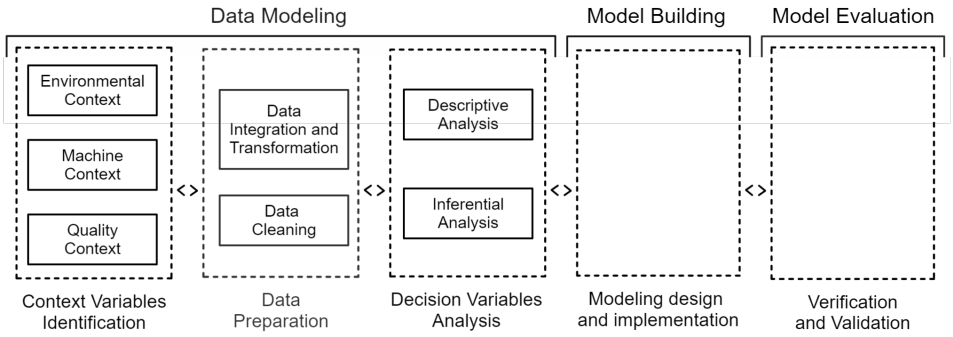

Fig. 2: The proposed data mining process model for the agricultural machinery domain contains three phases, data modeling, model building, and model evaluation, and accounts for the specific data categories relevant to machine optimization, i.e., machine, environment and quality contexts.

The proposed data mining process model is shown in Figure 2. It comprises three phases, data modeling, model building, and model evaluation. The contribution in this paper relates solely to the data modeling phase.

The first phase, 'data modeling', determines the potential decision variables through a sequence of descriptive and inferential analyses to identify the involved variables that should be part of the model. The phase consists of three steps:

The first step, 'context variables identification', starts with forming hypotheses that comprise potential decision variables. The types of variables are environmental, machinery, or quality variables. The combination of the three variables together defines operational context. The hypotheses should cover the three variable types. The hypotheses are made by domain experts, background research, and data observations. This step is essential due to its effect on the resulting system quality at the output; therefore, it is crucial to carefully gather the hypotheses and potential variables. The decision variables are the variables that correlate with the optimization of the agricultural process.

The second step, 'data preparation', is to integrate the required data from the information assets. The information assets may exist in different forms, databases, or flat documents. After integrating all required data listed as potential decision variables, data cleaning is essential to ensure the quality of all following steps and phases. The data cleaning may include identification and treatment of incomplete, inaccurate, or incorrect data. Additional data preparation might be required based on the use case and modeling method.

The third step, 'decision variables analysis', uses descriptive and inferential analyses to identify the valid hypotheses and possible correlations between the system and the relevant

output variables. The descriptive analysis summarizes and visualizes the data to identify possible patterns, correlations, or helpful observations, bringing the data into use. The inferential analysis verifies the hypotheses that are formed in the first step based on statistical test tools. This step is vital as the analysis drives the decision variables, which any data-driven model depends on.

Our data mining process model corresponds with CRISP-DM in its phases and steps. The 'context variables identification' and 'decision variables analysis' steps in our process model are implicitly mentioned in the 'business understanding', 'data understanding', and 'modeling' phases of CRISP-DM. While the data preparation step in our process model matches with 'data preparation' phase in CRISP-DM. As in the CRISP-DM, the different phases and steps may also be iterated several times, backward and forward, until a stable definition for the operational context is achieved.

4 Results and Discussion

Next, we illustrate how to apply the first phase of our process model. Our example is based on the use case of building a recommendation system for combine harvester settings. The results are shown in Figure 3.

In the first step, hypotheses are written based on experts, literature, and research. The hypotheses are converted into a list of potential variables. The variables in the list are integrated into the dataset, and the required data cleaning is performed.

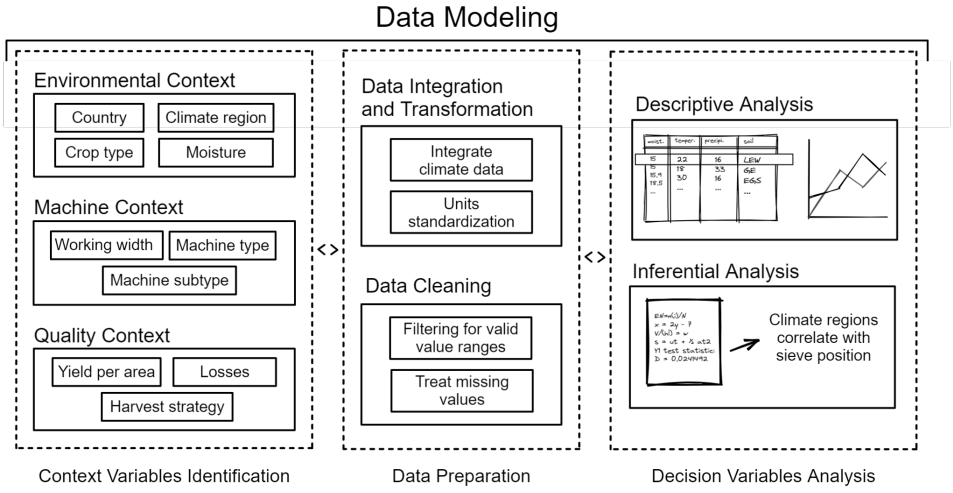

Fig. 3: The first phase of the proposed process model that illustrates the steps to build a recommendation system for the combine harvester settings.

A statistical summary is established and analyzed. The second part of the analysis deals with the hypotheses, and as a result, a new decision variable is attached or dismissed. The distribution of machine settings under the current definition of the operational context indicates how mature and stable the definition is. Therefore, there is backward and forward during the first phase. The model's accuracy within the different operational contexts may vary depending on the available data within each context.

5 Conclusion and Outlook

The proposed process model helped us to structure the development of the data-driven model. The process works based on the operational context of the agricultural machine. Our process model uses the first phase to determine the context variables, while the context variables are not directly considered in CRISP-DM, because it is designed to work with any problem and for any domain.

We focus on the data modeling phase within this paper. However, we plan to extend the other mentioned two phases (i.e., model building and model evaluation) to define an agricultural-specific scheme for data mining in the agricultural machinery domain.

References

[Ma21] Martínez-Plumed, F. et. al.: CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. In IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3048-3061, 2021.

[Es20] Esau, T. et. al.: Development and Evaluation of a Closed-Loop Control System for Automation of a Mechanical Wild Blueberry Harvester's Picking Reel. In AgriEngineering, vol. 2, no. 2, pp. 322-335, 2020.

[Sh00] Shearer, C.: The CRISP-DM model: the new blueprint for data mining. In Journal of data warehousing, vol. 5, no. 4, pp. 13-22, 2000.

[Ma09] Marbán, O. et. al.: Toward data mining engineering: A software engineering approach. In Information Systems, vol. 34, no. 1, pp. 87-107, 2009.

[Ni15] Niaksu, O.: CRISP Data Mining Methodology Extension for Medical Domain. In Baltic J. Modern Computing, vol. 3, no. 2, pp. 92-109, 2015.

[Hu19] Huber, S. et. al.: DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. In Procedia CIRP, vol. 79, pp. 403-408, 2019.