Streszczenie pracy autorstwa mgr inż. Cezary Maszczyk zatytułowanej „Zastosowanie głębokiej eksploracji danych, w tym dyskretnego głębokiego uczenia, w indukcji reguł logicznych zawierających złożone warunki elementarne” napisane w języku angielskim

This work addresses the topics of machine learning, rule-based systems, and data mining, with a particular focus on the induction of complex dependencies expressed in the language of descriptors (elementary conditions). The main goal was to discover deeper relationships that standard rule induction algorithms, based on heuristics, are unable to detect. In certain situations, this can contribute to improving the interpretability of the resulting rule-based models, making it possible to obtain more concise and understandable descriptions of the data. The author presents three rule induction methods that enable the discovery of complex dependencies in data. Each of them handles three types of data: classification, regression, and survival. The described methods fit within the concept of deep discrete learning, described in Chapter 5. Unlike the currently popular subsymbolic algorithms, such as deep neural networks, the methods proposed in this work focus on creating human-readable descriptions of data in the form of sets of decision rules. Each of the three methods — **ComplexConditions**, **MofNRules**, and **DeepRules** — presents a different approach to the problem. All of them, however, focus on searching for complicated dependencies in data, whose representation using simple conditions (e.g., age > 30 or color = red), as used by traditional rule-based algorithms, would be overly verbose and illegible. A common feature of all developed methods is their operation based on the popular sequential covering strategy, where each new rule is induced on a set of previously uncovered examples. This minimizes rules redundancy, allowing for the creation of concise sets of rules where each rule provides a unique value.

The **ComplexConditions** algorithm represents a refreshed approach to the idea of data-driven constructive induction. Building on the well-described RuleKit [1] method, it extends it with a wide range of complex conditions, such as attribute relationship conditions (e.g., height > weight) or internal alternatives (e.g., color ∈ {red, green, yellow}). This enables the creation of more concise data descriptions compared to widely used rule-based algorithms. As a result, it can lead to a reduction in the complexity of the resulting model, thereby facilitating its interpretation.

The second method, **MofNRules**, is an extension of the first one, introducing a new type of condition: M-of-N. They allow for the description of dependencies in data that would require the use of many rules with simple conditions. This algorithm operates in accordance with the idea of hypothesis-driven constructive induction, generating rules in a process of two consecutive induction passes, thereby expanding the original feature space. A unique feature of the **MofNRules** method is also that it is one of the few currently described in the literature that allows for the creation of rules containing, in addition to M-of-N conditions, other types of conditions in their premises.

The third and final method described, **DeepRules**, presents a different perspective on the problem. Instead of exhaustively searching the space of complex conditions, as was the case with **ComplexConditions**, this algorithm builds elaborate logical expressions in the form of CNF (conjunctive normal form) and DNF (disjunctive normal form). These expressions can then be simplified, replacing their individual fragments with logically equivalent complex conditions. This algorithm thus makes it possible to obtain rules containing all types of descriptors used in the previous two methods (with certain limitations), while offering lower computational complexity.

The key conclusion from the conducted experiments is that none of the methods is universally best for all types of problems. The results show that their effectiveness is largely dependent on the characteristics of the data. In the case of datasets where complex patterns do not occur, the proposed approaches may sometimes lead to the generation of overly complicated rules. However, the conducted research has shown that the described methods have the potential to generate sets of rules with less complexity than the reference algorithms, and their predictive results generally remain at a similar level. This proves that they are able to describe the data more concisely, while being equally accurate. Based on this, it can be suspected that they are able to discover certain deeper dependencies in data that are overlooked by the reference algorithms. This is clearly visible in the case studies of the "MONK" problems. All of them can be described using a set of rules in DNF form. Despite this, the RuleKit algorithm does not find the correct representations of the positive class in their case. This shows that although a given dependency can be described using rules with simple conditions, it does not mean that it will be discovered in practice by the algorithm. This is because rule learning methods usually operate on the basis of heuristics, not analyzing the full solution space, which means that some of them may be overlooked. The developed algorithms, by focusing on the search for complex patterns, increase the chance of their detection, often leading to more concise and potentially more interpretable data descriptions.

[1] Gudyś, A., Sikora, M., Wróbel, Ł. RuleKit: A comprehensive suite for rule-based learning. Knowledge-Based Systems, 194:105480, 01 2020.