Application of Dynamic Programming Approach to Optimization of Association Rules Relative to Coverage and Length

Beata Zielosko

Institute of Computer Science, University of Silesia
39, Będzińska St., 41-200 Sosnowiec, Poland
beata.zielosko@us.edu.pl

In the paper, an application of dynamic programming approach to optimization of approximate association rules relative to the coverage and length is presented. It is based on the extension of dynamic programming approach for optimization of decision rules [1] to the case of inconsistent decision tables.

Applications of rough sets theory to the construction of rules for knowledge representation or classification tasks are usually connected with the usage of decision table as a form of input data representation. In such a table one attribute is distinguished as a decision attribute and it relates to a rule's consequence. However, in the last years, associative mechanism of rule construction, where all attributes can occur as premises or consequences of particular rules, is popular. Association rules can be defined in many ways. In the paper, a special kind of association rules is studied, i.e., they relate to decision rules.

In the considered approach, an information system I with attributes $f_1, \ldots,$ f_{n+1} is transformed into a decision table for each attribute f_i, $i = 1, \ldots, n+1$. The column f_i is removed from I and a table with n columns labeled with attributes $f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_{n+1}$ is obtained. Values of the attribute f_i are attached to the rows of the obtained table denoted by I_{f_i}. Values of the attribute f_i are interpreted as values of a decision attribute. It is possible that the decision table I_{f_i} is inconsistent, i.e, contains equal rows with different decisions.

This paper is an essential extension of the paper [1] in which only consistent decision tables are considered that do not contain equal rows with different decisions. When association rules for information systems are studied and each attribute is sequentially considered as the decision one, inconsistent tables are often obtained. So, the approach from [1] was extended to the case of inconsistent decision tables. It required changes in definitions, algorithms (new uncertainty measure for decision tables), proofs of algorithm correctness, and, especially, in the software.

The aim of the paper is to create a research tool which is applicable to medium sized decision tables and allows one to construct approximate association rules with minimum length or maximum coverage. Exact rules can be overfitted, i.e., dependent on the noise or adjusted too much to the existing examples. If rules are used for knowledge representation, then instead of exact rules with many attributes it is more appropriate to work with approximate ones having relatively good accuracy and containing smaller number of attributes.

To work with approximate association rules, an uncertainty measure P is used, where $P(T)$ is the number of unordered pairs of different rows with different decisions in a decision table T. Then (β, f_i)-association rules are defined that localize rows in subtables of I_{f_i} $i = 1, \ldots, n+1$, with uncertainty at most β. The union of sets of (β, f_i)-association rules, $i = 1, \ldots, n+1$, is considered as the set of β-association rules for I.

In the paper, coverage and length are used as measures that allow one to make optimization of association rules. The rule coverage is important to discover major patterns in the data. Short rules are more understandable and easier for interpreting by experts. Construction of short rules is connected with the Minimum Description Length principle. Unfortunately, the problems of construction of rules with maximum coverage or minimum length are NP-hard.

The most part of approaches, with the exception of brute-force and, in some sense, Apriori algorithm, cannot guarantee the construction of shortest rules or rules with the maximum coverage. The proposed approach allows one to construct optimal rules. To this end, a directed acyclic graph is constructed, for each decision table from the set $\{I_{f_1}, \ldots, I_{f_{n+1}}\}$. Nodes of this graph are subtables of decision table I_{f_i} given by sets of descriptors "attribute = value". The partitioning of a subtable is finished when its uncertainty is at most β. The parameter β controls the computational complexity and makes the algorithm applicable to more complex problems.

The constructed graph $\Delta_\beta(I_{f_i})$, $i = 1, \ldots, n+1$, allows one to describe the whole set of so-called irredundant (β, f_i)-association rules. Then, based on the graph $\Delta_\beta(I_{f_i})$, $i = 1, \ldots, n+1$, it is possible to make local optimization relative to coverage, i.e., all (β, f_i)-association rules with the maximum coverage can be described. After that, global optimization relative to coverage is made, i.e., among the considered sets of (β, f_i)-association rules, $i = 1, \ldots, n+1$, only these sets are selected, where the value of coverage is maximum. The union of selected sets of (β, f_i)-association rules forms the set of β-association rules with the maximum coverage for information system I. Similarly, β-association rules with the minimum length for information system I can be obtained.

It was proved that by removal of some descriptors from the left-hand side of each (β, f_i)-association rule, $i = 1, \ldots, n+1$, that is not irredundant and by changing the decision on the right-hand side of this rule it is possible to obtain an irredundant (β, f_i)-association rule which coverage is at least the coverage of initial rule and the length is at most the length of initial rule. It means, that not only optimal rules among irredundant (β, f_i)-association rules, $i = 1, \ldots, n+1$, are considered but also optimal rules among all (β, f_i)-association rules.

Experimental results connected with the maximum coverage and minimum length of β-association rules for some information systems I and the size of the directed acyclic graphs relative to the parameter β will be presented.

References

1. Amin, T., Chikalov, I., Moshkov, M., Zielosko, B.: Dynamic programming approach to optimization of approximate decision rules. Inf. Sci. **221** (2013) 403–418