Wnioskowanie z danych z brakującymi wartościami atrybutów
Reasoning from data with missing attribute values

Rafał Latkowski

Abstract

The dissertation concerns the reasoning on data with missing attribute values, with particular emphasis on methods and algorithms based on the rough set theory. The content of the dissertation, which consists of a series of papers, presents several methods of reasoning on data with missing values along with their experimental evaluation, as well as related works related to effective parallelization of experiments and to increase of computational complexity of an algorithm used in comparison.

The work presents a method allowing reasoning on data with missing values by bypassing the problem, using the decomposition of the initial data (information systems in the Pawlak sense) into information subsystems that do not contain missing values, and then applies reasoning on complete information subsystems. The application of this method with the use of various classification algorithms is presented: decision rules induced by the rough set methods, decision trees compatible with C4.5 and methods for decision rule joining that generate effective and compact classifiers. In addition we present also significant increase of the computational complexity of C4.5 compliant algorithms resulting from a large number of missing values.

In the further part we present a generalization of the indiscernibility relation, which enables expressing the semantics of missing values within the rough set theory. For the algorithmic search for the optimal semantics of the missing values (that is defined by the optimal indiscernibility relation), two parametric families of such relations are presented, as well as their capability to represent some data imputation algorithms and an experimental evaluation in simplified case.

We also present a software that allows to perform massive experiments in parallel using grid-computing paradigm, as well as the constantly developed library Rseslib 3 that contain open object-oriented implementation of the rough set theory algorithms that enable the use of any (parametric) indiscernibility relation.