SAT-DM: Satisfiability Data Mining for Classification Problems

Vincent Gardeux[1], Lars Magnus Hvattum[2], Fred Glover[3,*]

[1] University of Arizona, Bio5 Institute, Tucson, Arizona 85721, USA
[2] Dept. of Ind. Eco. and Tech. Manage., Alfred Getz veg 3, N-7491 Trondheim, Norway
[3] University of Colorado Boulder, Colorado 80309-0419, USA
[*]Corresponding Author: fred.gover@colorado.edu

Abstract

SAT-DM is a new method for binary data classification problems, based on generating a collection of logical clauses, or equivalently a collection of inequalities in zero-one variables, for each group of points representing a given classification. A point with unknown membership is classified as belonging to a particular group based on comparing the number or proportion of the inequalities it satisfies for that group versus the number or proportion it satisfies for other groups. We make use of a fundamental observation which states that inequalities are satisfied by a subset of elements of a particular group (and correspondingly violated by a subset of elements from a complementary group) if and only if these inequalities correspond to feasible solutions to a special variant of a satisfiability problem. Based on this, we propose a method for generating membership-defining systems of inequalities that provide a filter for segregating points lying in different groups.

SAT-DM may be viewed as a procedure for generating multiple hyperplanes that segregate points of different groups by isolating their logical properties. The inequalities produced by SAT-DM capture classification regions in feature space that are more varied and complex than those derived from hyperplane separating procedures such as those used in support vector machines (SVMs) and related procedures based on linear programming and convex analysis. A particularly useful feature is the ability to generate the collections of segregating inequalities (complementary half-spaces) in a highly efficient manner, allowing the approach to handle large data sets without difficulty. The underlying processes can also be used for feature selection, or more generally attribute selection, to isolate a subset of attributes from large data sets that yield a high classification power while reducing the time and complexity of classification.

Although SAT-DM is primarily designed for handling binary attributes, it has been extended to any representation using a binarization method, inspired by the IDEAL algorithm. SAT-DM is a perfect example of the "Meta-Analytics" concept. It uses metaheuristics, through the resolution of satisfiability problems, to enhance machine learning classification process.