# Rough Set based Rule Induction Package for R

Shusaku Tsumoto and Shoji Hirano
Department of Medical Informatics,
Shimane Medical University, School of Medicine,
Enya-cho Izumo City, Shimane 693-8501 Japan

**Abstract**

Rough set theory is a framework of dealing with uncertainty based on computation of equivalence relations/clases. Since a proability is defined as a measure of sample space, defined by equivalence classes, rough sets are closely related with probabilities in the deep level of mathematics. Furthermore, since rough sets are closely related with Demster-Shafer theory or fuzzy sets, this theory can be viewed as a bridge between classical probability and such subjective probabilities. Also, this theory is closely related with Baysian theories.

The application of this theory includes feature selection, rule induction, categorization of numerical variables, which can be viewed as a method for categorical data analysis. Especially, rough sets have been widely used in data mining as a tool for feature selection, extracting rules (if–then rules) from data. Also, this theory includes a method for visualization, called "flow graphs."

This paper introduces a rough set based rule induction package for R, including: (1) Feature selection: rough sets call a set of independent variables "reducts." This calculation is based on comparisons between equivalence classes represented by variables with respect to the degree of independence. (2) Rule Induction: rough sets provide a rule extraction algorithm based on reducts. Rules obtained from this subpackage are if–then rules. (3) Discretization (Categorization of Numerical Variables): discretization can be viewed as a *sequential* comparison between equivalence classes given in a dataset. (4) Rough Clustering: calculation of similarity measures can be also viewed as that of comparisons between equivalence classes. Rough clustering method gives a indiscernbility-based clustering with iterative refinement of equivalence relations. (5) Flow Graph: this subpackage visualizes a network structure of relations between given variables. Unlike bayesian networks, not only conditional probabilies but also other subjective measures are attached to each edge. (6) Rule Visualization with MDS: this subpackage gives a visualization approach to show the similar relations between rules based on multidimensional scaling. The usage of R gives the following advantages: (a) Rough set methods can be easily achieved by fundamental R-functions, (b) Combination of rough set methods and statistical packages are easily acheived by rich R-packages. In the conference, several aspects of this package and experimental results will be presented.

**Keywords:** Data Mining, Rough Sets