RSESLIB 3: Rough Sets and Machine Learning Open Source in Java

Arkadiusz Wojna
Security On-Demand
wojna@mimuw.edu.pl

Agenda

- Library Overview
- Contents
- Rseslib 3 in Weka
- Rough set classifier
- Architecture
- Graphical Interface
- Computing in cluster
- Library usage
- Future work
- Summary

Motivation

- Deliver library of rough set methods in Java
  - Open source
  - Easily extensible
  - Easily modifiable
- Speed-up research & development of new machine learning algorithms
  - Reduce development effort
  - Additive implementation
    - Increase reusability of code
    - Increase inheritance of available algorithms
  - Code organization
- Speed-up experiments
  - Multi-platform executables – Java
  - Grid Computing / Network of Workstations
- Didactic framework
  - Research of new algorithms
  - Applications

Rseslib 3: Overview

- Java Library
- Open Source (GNU GPL)
- Collection of Rough Set and other Machine Learning algorithms
- Modular Component-Based Architecture
- Available in Weka
- Graphical Interface

Library Content

- Classification
- Reduct calculation
- Rule calculation
- Metric induction
- Filtering
- Sampling
- Clustering
- Sorting
- Transformation
- Discretization
- Missing value completion
- Genetic algorithm
- Logic algorithms
- Principal component analysis

Classification: Unique Implementations

- Rough Set Rule Classifier
- K Nearest Neighbours / RIONA
- K Nearest Neighbours with Local Metric Induction

Classification: Classics

- Decision tree C4.5 (Quinlan)
- Rule Classifier AQ15 (Michalski et al)
- Neural Network
- Naive Bayes
- Support Vector Machine
- PCA classifier
- Local PCA classifier
- Metaclassifiers
  - Bagging
  - AdaBoost

Algorithms (1)

- Reducts
  - …
- Discretization
  - ...
- Rule calculation
  - From global reducts
  - From local reducts
  - AQ15
- Metric induction
  - City + Hamming
  - City + Simple Value Difference (SVDM)
  - Interpolated Value Difference (IVDM) + SVDM
  - Density-Based Value Difference (DBVDM) + SVDM
  - Attribute weighting: distance-based, accuracy-based, perceptron

Algorithms (2)

- Filtering
  - Wilson Editing
  - Reduction Technique Editing (RT)
  - Missing Values Filter
  - General Boolean Function Filter
- Sampling
  - With repetitions
  - Without repetitions
  - With given class distribution
- Clustering
  - K Approx Centers
- Sorting
  - Attribute-Based
  - Distance-Based

Algorithms (3)

- Transformation
  - Discretizations
  - Attribute selection
  - Scaling
  - Radial
  - Perceptron
  - Arithmetic
- Missing value completion
  - Non-invasive data imputation (Gediga, Duentsch)
- Genetic algorithm
  - General scheme
- Logic
  - 2 algorithms generating prime implicants from CNF
- Principal Component Analysis
  - OjaRLS algorithm

Data formats

- ARFF (Weka)
- CSV + Rseslib header
  - header file apart
  - header and data in one file
- RSES 2.x

Weka

- Platform for machine learning
- Written in Java
- Developed since 1997
- One of the two most popular machine learning Java platforms in the world
  - 1.5 million downloads a year
- 4 graphical interfaces
- 1 command line interface

Rseslib 3 in Weka

- Official registered package
  - Available in Weka Package Manager
  - requires Weka 3.8.0 or later
- 3 classifiers available now in Weka
  - Rough Set Rule Classifier
  - K Nearest Neighbours / RIONA
  - K Nearest Neighbours with Local Metric Induction

Rough Set Rule Classifier

- Uses discretization
- Generates reducts and rules from reducts
- Handles missing values
- Handles inconsistent data (Generalized Decision)

Discretizations

- Equal Width
- Equal Frequency
- 1R
- Entropy Minimization Static
- Entropy Minimization Dynamic
- Chi Merge
- Maximal Discernibility Heuristic Global
- Maximal Discernibility Heuristic Local

Discretization: Entropy Minimization

$$Ent(S) = -\sum_{i=1}^{k} \frac{P(C_i,S)}{|S|} \log\left(\frac{P(C_i,S)}{|S|}\right)$$

Minimize:

$$E(A_i,T,S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

S - data set

C_i – decision class

$P(C_i,S)$ – number of records from decision class C_i in S

S_1, S_2 – data set S split by a value T on an attribute A_i

Discretization: ChiMerge

Merge neighbouring pair of intervals with minimal:

$$\chi^2(S_1, S_2) = \sum_{i=1}^{k} \frac{\left(P(C_i, S_1) - E(C_i, S_1)\right)^2}{E(C_i, S_1)} + \sum_{i=1}^{k} \frac{\left(P(C_i, S_2) - E(C_i, S_2)\right)^2}{E(C_i, S_2)}$$

S_1, S_2 - data sets from neighbouring intervals

C_i – decision class

$P(C_i,S)$ – number of records from decision class C_i in S

$E(C_i,S)$ – expected number of records from decision class C_i in S

Discretization: Maximal Discernibility

Maximize:

$$\left|\left(x_i, x_j\right) \in S_1 \times S_2 : dec\left(x_i\right) \neq dec\left(x_j\right)\right|$$

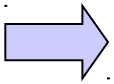
Reduct Algorithms

- All Global
- All Local
- One Johnson
- All Johnson
- Partial Global
- Partial Local

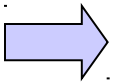
All Reducts

- Data Table → Discernibility Matrix → Prime Implicants → Decision Rules

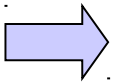
a	b	c	dec
1	2	3	1
1	3	4	2
2	1	1	1
2	2	1	2



	x1	x2	x3	x4
x1		bc		ac
x2	bc		abc	
x3		abc		b
x4	ac		b	



Prime Implicants
Φ → ψ
CNF →DNF



(a=1)∧(b=2) → (dec=1)
(a=1)^(b=3) → (dec=2)
(a=2)^(b=1) → (dec=1)
(a=2)^(b=2) → (dec=2)
…

- Global & Local
- Advanced algorithm finding prime implicants

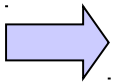
Global reducts: {a, b}, {b, c}
Local reducts for x1: {a, b}, {c}

Johnson Reduct

- Repeat
  - Find most frequent attribute a in discernibility matrix
  - Remove all fields with a from discernibility matrix
  - Add a to R
- until discernibility matrix is empty

Partial Reducts

a	b	c	dec
1	2	3	1
1	3	4	2
2	1	1	1
2	2	1	2



	x1	x2	x3	x4
x1		bc		ac
x2	bc		abc	
x3		abc		b
x4	ac		b	

R is an α-reduct if:
discerns ≥ (1 – α) of non-empty fields of discernibility matrix
none subset of R satisfies the above property

{b} is 0.25-reduct but is not 0.2-reduct
{a,c} is not 0.25-reduct because {c} is 0.25-reduct

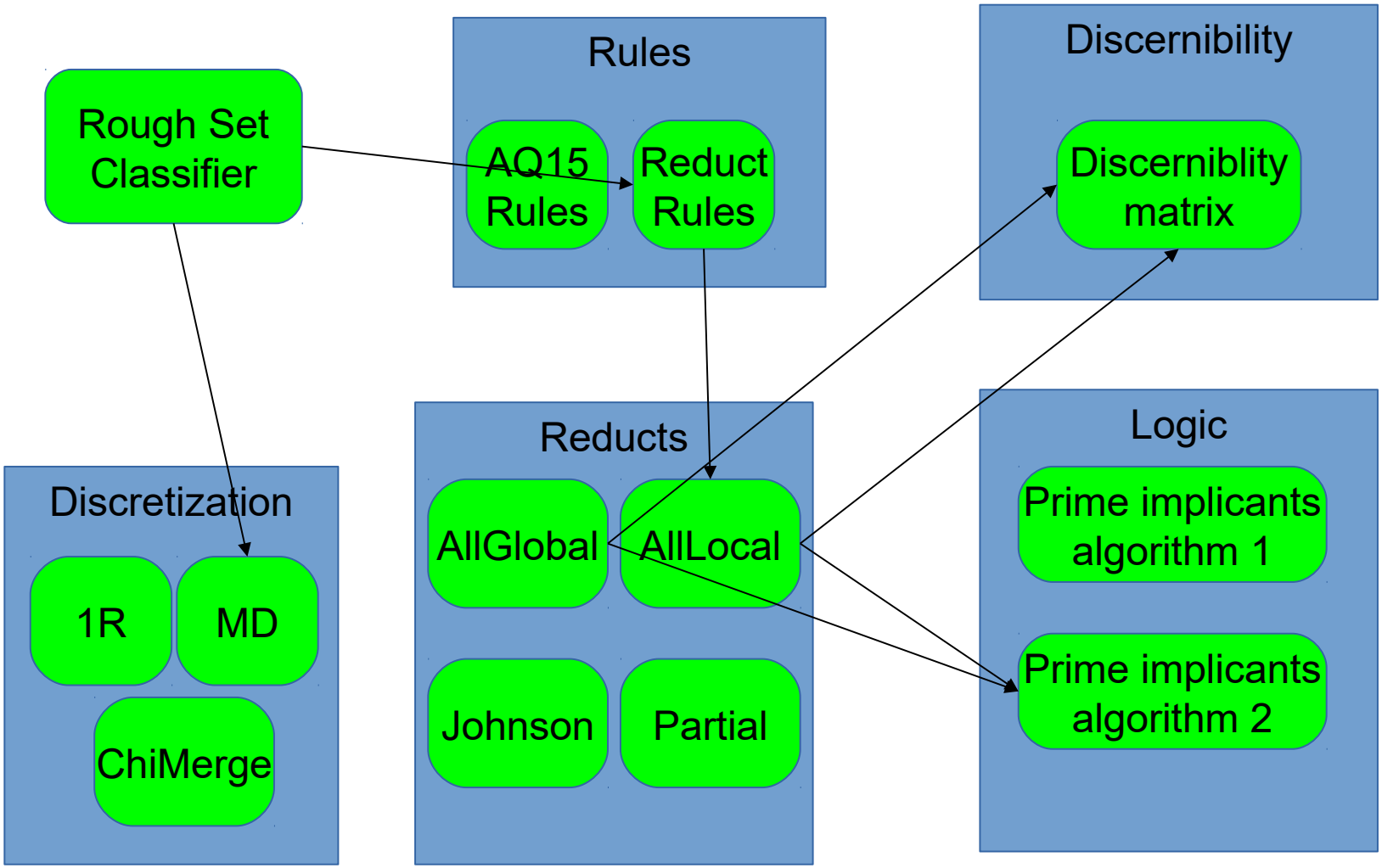
Modularity

- Modules
- Interfaces
- Isolated elementary mathematical objects
- Isolated processing algorithms

Mathematical Objects

- Basic:
  - attribute, data header, data object, boolean data object, numbered data object, data table, nominal attribute histogram, numeric attribute histogram, decision distribution
- Boolean functions/operators:
  - attribute equality, attribute interval, attribute value subset, binary discrimination, metric cube, negation, conjunction, disjunction
- Real functions/operators:
  - scaler, perceptron, radius, multiplication, addition
- Integer functions:
  - discrimination (discretization, 3-value cut)
- Decision distribution functions
  - nominal to dec distr, numeric to vicinity-based dec distr, numeric to interpolated dec distr
- Vector, linear subspace, PCA subspace, vector function
- Linear order
- Indiscernibility relations
- Metrics:
  - City + Hamming, City + SVDM, IVDM, DBVDM, metric-based indexing tree
- Rules:
  - boolean function rule, equality descriptors rule, partial matching rule
- Probability
  - guassian kernel function, hypercube kernel function, m-estimate

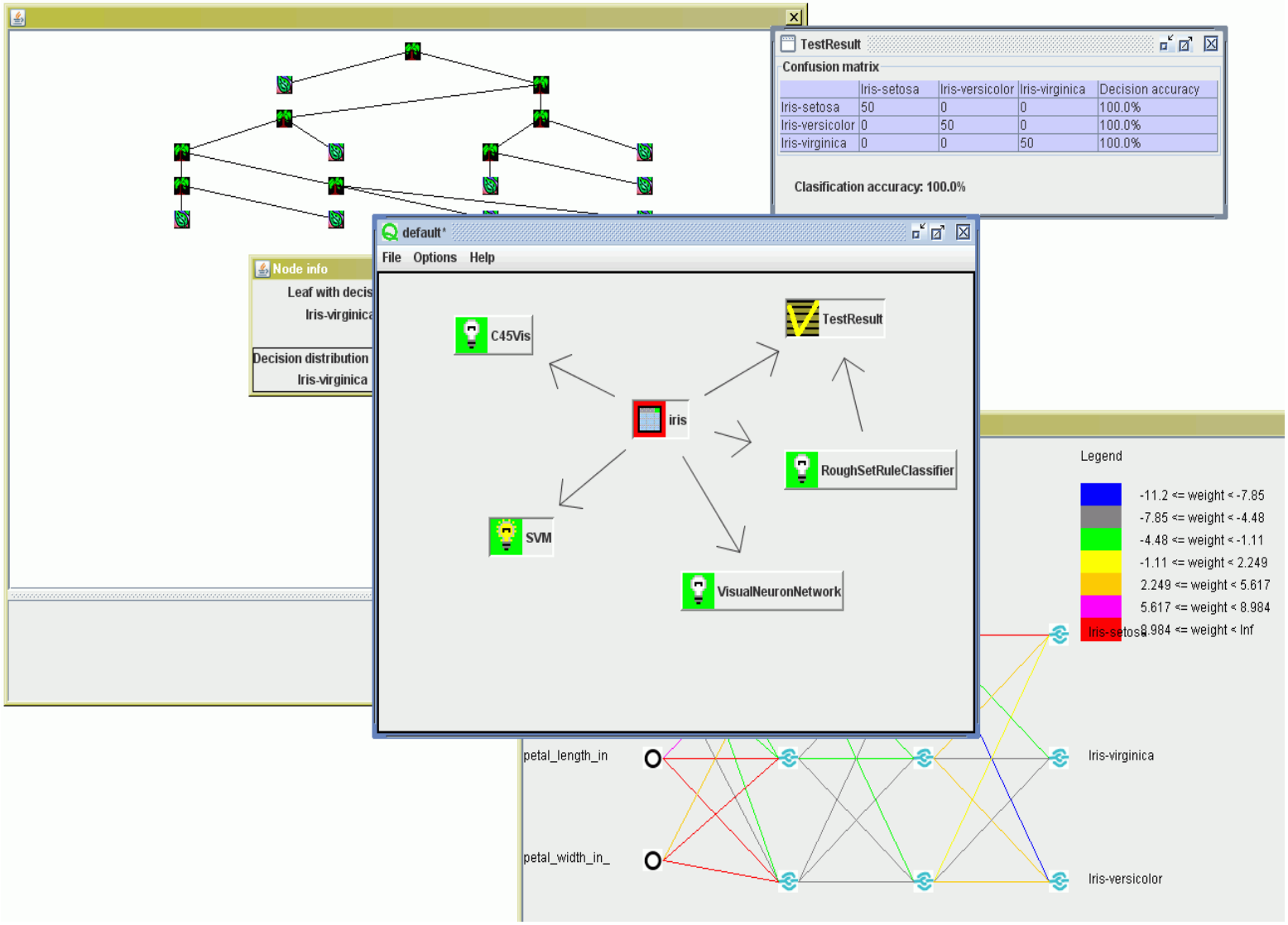
Modularity examples (1)



Modularity examples (2)

- Attribute weighting in metric
  - Perceptron as one of weighting methods
- Estimate of value probability at given decision
  - Probability defined by k nearest neighbours

Qmak – graphical interface for Rseslib

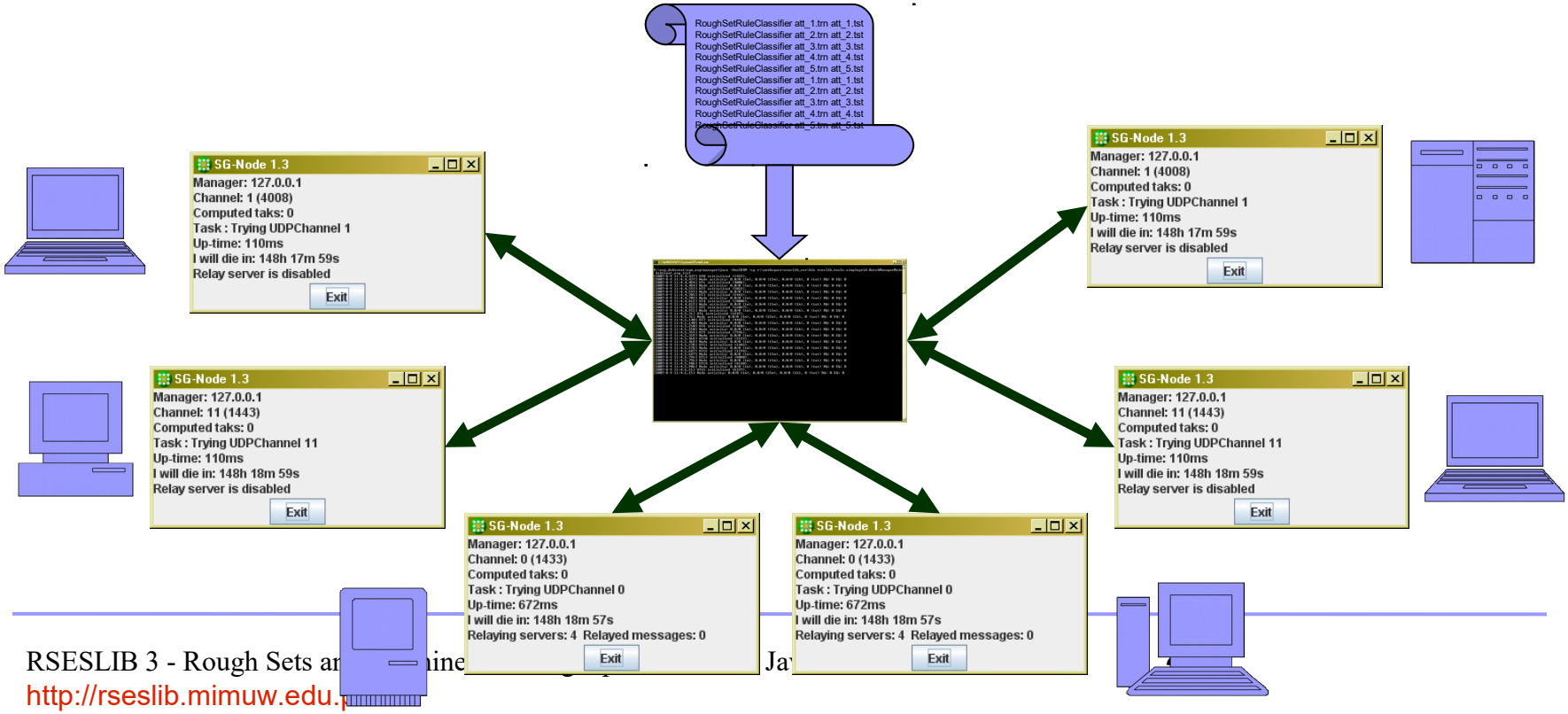


Qmak functionality

- Visualization of
  - classifiers
  - classification
  - data
- Classifier modification (interactive)
- Classification of test data
  - shows misclassified objects
- Experiments
  - Cross-validation
  - Multiple cross-validation
  - Multiple random split
- New classifiers including visualization
  - can be added within GUI or in the configuration file
  - do not require changes in Qmak

Simple Grid Manager

- Motivation:
  - many experiments, accessible network of PC's
  - network subsystem dead-lock free
  - auto-integration with Rseslib 3 algorithms
- Distribution of any Rseslib 3 experiments over grid of computers
- Heterogeneous network (Internet/Intranet), platform (Java)



Simple Grid Manager features

- Skipping completed jobs
- Resuming failed jobs
- Communication protocol UDP
- Firewall/Packet filter by-passing
- Intranet break-through (message relaying)
- Many jobs in message
- One server thread per one worker node
- Negligible communication overhead

Rseslib 3 used in

- mahout-extensions
  - attribute selection extensions to Mahout (an extensible programming environment and framework for building scalable algorithms in machine learning)
  - running on Spark (an open-source cluster computing framework based on Hadoop)
- DMEXL
  - parallel feature selection algorithm based on rough sets and particle swarm optimization
- TunedIT
  - system for automated evaluation, benchmarking and comparison of data mining and machine learning algorithms
- Research
  - 3 Phd (+1 in progress)
  - 6 MSc
  - 10+ BSc
- Teaching
  - Many lab projects

Future work

- Qmak 1.0
- GitHub
- RIONIDA
- Discretizations in Weka package
- Maven
- RapidMiner (?)
- Other algorithms...

Contributors: Rseslib

Jan Bazan, Rafał Falkowski, Grzegorz Góra, Marcin Jałmużna, Łukasz Kosson, Łukasz Kowalski, Michał Kurzydłowski, Rafał Latkowski, Łukasz Ligowski, Michał Mikołajczyk, Krzysztof Niemkiewicz, Dariusz Ogórek, Marcin Piliszczuk, Maciej Próchniak, Jakub Sakowicz, Sebastian Stawicki, Cezary Tkaczyk, Arkadiusz Wojna, Witold Wojtyra, Damian Wójcik, Beata Zielosko

From:
University of Warsaw, University of Rzeszów, Silesian University, Wrocław University of Technology

Contributors: Rseslib tools

- Graphical interface Qmak
  Katarzyna Jachim, Damian Mański, Michał Mański, Krzysztof Mroczek, Robert Piszczatowski, Maciej Próchniak, Tomasz Romańczuk, Piotr Skibiński, Marcin Staszczyk, Michał Szostakiewicz, Leszek Tur, Arkadiusz Wojna, Damian Wójcik, Maciej Zuchniak
- Simple Grid Manager
  Rafał Latkowski

Summary

- Ready to use Open Source Java Library
- Broad collection of Rough Set & Machine Learning algorithms
- Ease to use & implement own algorithms
- Visit project home page:
  - http://rseslib.mimuw.edu.pl/

RSESLIB 3: Rough Sets and Machine Learning Open Source in Java

Questions

http://rseslib.mimuw.edu.pl/