impara
intelligent machines

rulex
rule generation

Extracting Knowledge from Biomedical Data through Logic Learning Machines and Rulex

Marco Muselli

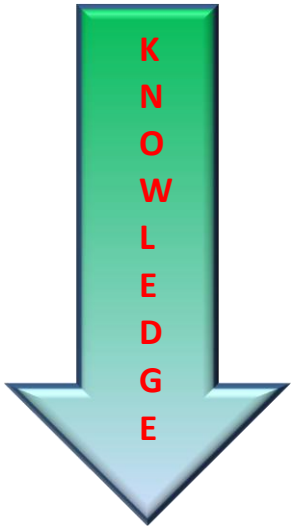
Institute of Electronics, Computer and Telecommunication Engineering
National Research Council of Italy, Genova, Italy
marco.muselli@ieiit.cnr.it

Extracting knowledge from data

Basic problem: Infer some knowledge about a biological phenomenon of interest starting from a sample of data.

Type of knowledge:

- Correlation, statistical measures
- Feature ranking, analysis of relevance
- Prediction, clustering, risk analysis
- Intelligible model (rules)

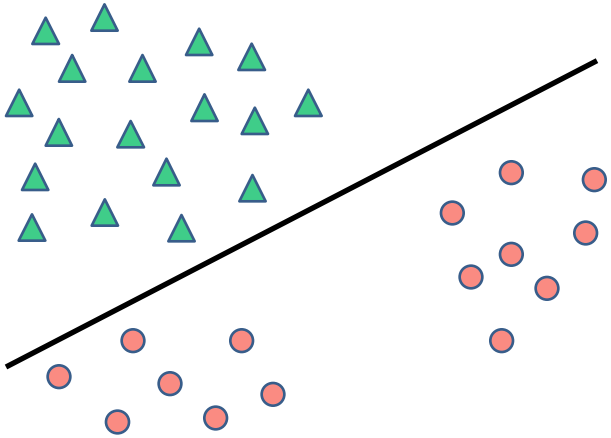


Rule generation methods

Extract models described by a set of intelligible rule in **if-then** form

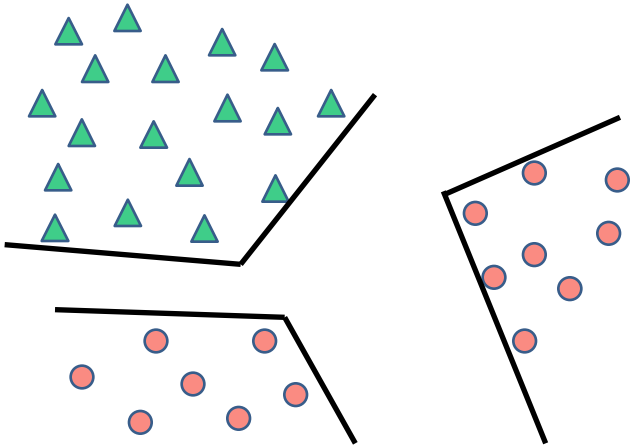
If Pressure > 115 **and** Heart_rate < 100 **then** Disease = Yes

Divide-and-conquer approach



Emphasis on differences!

Aggregative approach



Emphasis on similarities!

Statistical vs. Machine learning methods

Statistical methods

- Simpler to be used with huge experience
- Plenty of commercial and free tools available
- Limited quantity of knowledge extracted
- A priori hypotheses on probability distributions

Machine learning methods

- Their application is not straightforward and experience is not so big
- Commercial tools are often extensions of statistical packages; free programs are not so friendly
- Relevant quantity of knowledge extracted
- No a priori hypothesis is required

Machine learning software

Commercial software

- SAS Enterprise Miner (www.sas.com/technologies/analytics/datamining/miner)
- IBM SPSS Statistics Software (www-01.ibm.com/software/analytics/spss/products/statistics)
- Salford Systems Data Mining Suite (www.salford-systems.com)
- Statistica Data Miner (www.statsoft.com/products/data-mining-solutions)

Free Software

- WEKA (www.cs.waikato.ac.nz/ml/weka)
- RapidMiner (rapid-i.com)
- Orange (orange.biolab.si)
- Machine Learning & Statistical Learning in R language (cran.r-project.org/web/views/MachineLearning.html)

RULEX® Suite

The suite RULEX® (contraction of RULe Extraction) developed by Impara Srl (www.impara-ai.com), a spin-off of the National Research Council of Italy, offers a new simple and powerful tools for extracting knowledge from real world data.

The name RULEX is the contraction of RULe Extraction since it is especially devoted to generate intelligible rules, although a wide range of statistical and machine learning approaches will be made available.

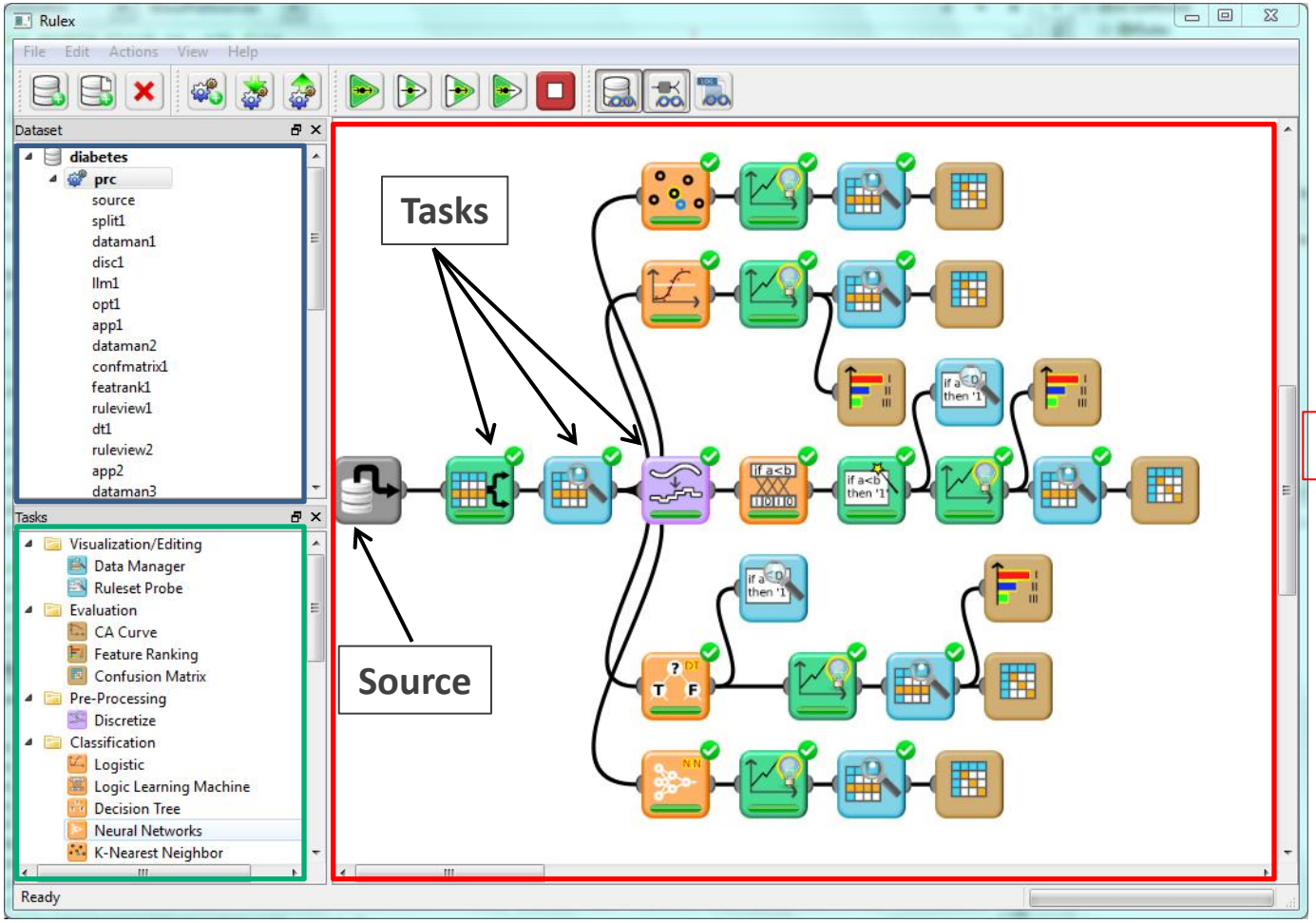
An intuitive graphical interface allows to easily apply standard and advanced algorithms for analyzing any dataset of interest, providing solution to classification, regression and clustering problems.

The software suite is in rapid evolution; therefore, the number and the functionalities of available tasks increase every day.

RULEX GUI

Dataset panel

Component panel



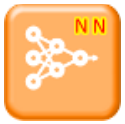
Stage

Logic Learning Machine

Besides standard techniques, such as:

Decision trees

Logistic



Neural networks

K-nearest-neighbor

Rulex offers the possibility of applying an original proprietary approach, named

Logic learning machine (LLM)

which represents an efficient implementation of the switching neural network model (Muselli, 2006).

Logic Learning Machine

LLM allows to solve classification problems producing sets of intelligible rules capable of achieving an accuracy comparable or superior to that of best machine learning methods.

The approach of LLM is based on monotone Boolean function synthesis (Shadow Clustering) and adopts an aggregative policy: at any iteration some patterns belonging to the same output class are clustered to produce an intelligible rule.

Since the training process occurs in a binary projected space, the application of LLM must be preceded by a discretization task that finds proper cutoffs for ordered (continuous and discrete) input variables.

An application in biomedical analysis

The functionalities of Rulex have been verified by analyzing three biomedical datasets included in the Statlog benchmark:

Diabetes: it concerns the problem of diagnosing diabetes starting from 8 input variables; all the 768 considered patients are females at least 21 years old of Pima Indian heritage: 268 of them are cases and 500 are controls.

Dna: it has the aim of recognizing acceptors and donors sites in a primate gene sequences with length 60 (basis); the dataset consists of 3186 sequences, subdivided into three classes: acceptor, donor, none.

Heart: it deals with the detection of heart disease from a set of 13 input variables concerning patient status; the total sample of 250 elements is formed by 120 cases and 150 controls.

An application of Rulex (results)

Five classification algorithms have been considered: LLM, DT, NN, LOGIT, and KNN.

Results obtained on an independent test set including 30% of data has been compared both in terms of accuracy and of quantity of knowledge extracted (number of rules and average number of conditions).

	LLM			DT			NN	LOGIT	KNN
	Accuracy	# Rules	# Cond.	Accuracy	# Rules	# Cond.	Accuracy	Accuracy	Accuracy
Diabetes	77.40%	14	3.00	73.04%	56	4.02	75.22%	77.23%	69.13%
Dna	94.01%	64	10.86	90.04%	67	6.26	88.69%	92.57%	40.68%
Heart	85.19%	19	5	81.48%	18	3.67	80.25%	83.95%	80.25%

Conclusions

A new suite, called Rulex, for the analysis of biomedical datasets through conventional and advanced machine learning techniques has been presented. It is able to solve classification, regression and clustering problems.

An intuitive graphical interface allows to construct complex analysis processes through the composition of elementary tasks. Facilities for displaying and managing datasets are also provided.

Besides standard methods, like logistic, k-nearest-neighbor, neural networks and decision trees, Rulex makes available a new approach, logic learning machines (LLM), whose models are described by intelligible rules.

Results obtained for the analysis of three biomedical datasets belonging to the Statlog benchmark point out the good quality of LLM, which achieves an excellent accuracy while providing understandable knowledge about the problem at hand.

Work in progress

Version 2.0 of Rulex is currently under beta testing. Several features have been added with the intent of giving researchers a simple but powerful tool for analyzing their own datasets.

Functionalities are continuously added to Rulex to improve the versatility of the suite. Suggestions arising from researchers are extremely important, since they allow us to offer a product satisfying the real needs of users.

To this aim, we are searching for researchers interested to try the Rulex suite, signaling bugs and providing us advices for improving each part of the product.

If you are interested to test Rulex for your specific application, please send me an email (m.muselli@impara-ai.com) and we will provide you a fully functional copy of Rulex.

impara
intelligent machines

rulex
rule generation

Thanks for your attention!

www.impara-ai.com