Extracting knowledge from biomedical data through Logic Learning Machines and Rulex

Marco Muselli✉
Institute of Electronics, Computer and Telecommunication Engineering, National Research Council, Genoa, Italy

Motivation and Objectives

When dealing with biomedical data concerning a giving problem, usually experts are required to infer specific conclusions about a pathology or a biological phenomenon of interest starting by a sample of previously collected observations. Besides conventional statistical techniques that allow to retrieve important indications about the characteristics of the system, machine learning methods have revealed to be very effective in predicting its behavior in cases different from those included in the observations at hand.

Among this last group of techniques rule generation methods build models described by a set of intelligible rules, thus permitting to extract important knowledge about the variables included in the analysis and on their relationships with the output attribute. Two different paradigms have been proposed in literature to perform rule generation: decision trees (Duda et al., 2001), which adopt a divide-and-conquer approach for generating the final model, and methods based on Boolean function reconstruction (Boros et al., 2000; Muselli and Ferrari, 2011), which follow an aggregative procedure for building the set of rules.

Available commercial software, such as SAS, SPSS or STATA, allows to employ a wide range of statistical techniques for the analysis of real world data, allowing also the application of some machine learning algorithms, among which neural networks and decision trees. However, the focus of these suites is more centered on conventional statistics rather than on machine learning and consequently it is difficult for a non expert to successfully extract knowledge from its own data by employing advanced techniques offered by commercial packages.

This is even more true when considering freely available software tools, such as Weka (www.cs.waikato.ac.nz/ml/weka), Orange (orange.biolab.si), or R (www.r-project.org); in these cases a wider range of machine learning approaches is generally made available, but the level of experience needed to achieve a satisfying result is often too high to allow their use by a non expert.

To overcome these difficulties a new suite for extracting knowledge from real world data has been developed by Impara srl (www.impara-ai.com); it is named Rulex (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.

Besides standard techniques, such as decision trees (DT), neural networks (NN), logistic (LOGIT), and k-nearest-neighbor (KNN), 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). 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 application of Rulex to the analysis of biomedical datasets included in the Statlog benchmark (Michie et al., 1994) permits to appreciate the good characteristics of this new analysis software. In particular, it is shown how different algorithms can be easily employed to extract knowledge from data at different levels of intelligibility, comparing results produced by corresponding models.

Methods

Although conventional statistical techniques or standard machine learning approaches allow to retrieve important indications about the characteristics of a system or of a phenomenon of interest, starting from a sample of observations regarding historical data, a deeper insight into the relationships among the considered variables can only be obtained by adopting rule generation methods. These techniques are capable of

constructing models described by a set of intelligible rules having the following form:

if **premise** then **consequence**

where **premise** is the conjunction of conditions on the input variables, whereas **consequence** contains information about the output of the model.

For instance, in a diagnosis problem rule generation techniques produce not only the subset of variables actually correlated with the pathology of interest, but also explicit intelligible conditions that determine a specific diagnosis. As a consequence, relevant thresholds for each input variable are identified, which represent valuable information for understanding the phenomenon at hand.

Most used rule generation techniques belong to the following two broad paradigms: decision trees and methods based on Boolean function synthesis. The approach adopted by the first kind of algorithms divides iteratively the training set into smaller subsets according to a divide and conquer strategy: this gives rise to a tree structure from which an intelligible set of rules can be easily retrieved. It is important to observe that the divide and conquer strategy leads to conditions and rules that point out differences among examples of the training set belonging to different output classes. In this sense we can say that the DT approach implements a discriminant policy: differences between output classes are the driver for the construction of the model.

In contrast, methods based on Boolean function synthesis adopt an aggregative policy: at any iteration some patterns belonging to the same output class are clustered to produce an intelligible rule. Suitable heuristic algorithms (Boros et al., 2000; Muselli and Ferrari, 2011) are employed to generate rules exhibiting the highest covering and the lowest error; a trade-off between these two different objectives has been obtained by applying the Shadow Clustering (SC) technique (Muselli and Ferrari, 2011), which generally leads to final models exhibiting a good accuracy.

The aggregative policy allows to retrieve intelligible rules that better characterize each output class with respect to approaches following the divide-and-conquer strategy. As a matter of fact, clustering examples of the same kind permits to extract knowledge regarding similarities about the members of a given class rather than information about their differences. This is very useful in many applications and often leads to models showing a higher generalization ability.

LLM and DT represent two of the techniques available in Rulex for the analysis of real world data. In fact, Rulex can efficiently approach and solve supervised (classification, regression) and unsupervised (clustering) machine learning problems by allowing the creation of complex analysis processes through the composition of elementary tasks. A simple but powerful GUI permits to manage datasets providing advanced interactive visualization as well as complete control on the various computational phases. The software suite is in rapid evolution; therefore, the number and the functionalities of available tasks increase every day.

Results and Discussion

The functionalities of Rulex have been verified by considering three biomedical datasets included in the Statlog benchmark (Michie et al., 1994) and concerning as many classification problems:

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

Heart: it deals with the detection of heart disease from a set of 13 input variables concerning patient status; the total sample of 270 elements is formed by 120 cases and 150 controls.

Table 1: Results obtained by the application of five classification algorithms on biomedical datasets included in the Statlog benchmark.

	LLM			DT			NN	LOGIT	KNN
	Accuracy	# Rules	# Cond.	Accuracy	# Rules	# Cond.	Accuracy	Accuracy	Accuracy
Diabetes	76.52%	16	3.75	76.09%	42	4.77	75.65%	76.52%	68.70%
Heart	75.31%	19	4.26	64.20%	17	4.18	72.84%	74.07%	51.85%
Dna	91.98%	19	6.84	90.04%	67	6.26	87.09%	92.57%	40.68%

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.

Five different classification algorithms have been considered for each dataset (LLM, DT, NN, LOGIT and KNN) and their results are compared both in terms of accuracy of the retrieved solution and of quantity of knowledge extracted from the dataset of examples at hand. For evaluating this last aspect the intelligibility of the rule set, measured by the number of rules and by the average number of conditions for each of them, has been taken into account.

Table 1 shows these two values for the rule sets produced by DT and LLM in the three considered datasets. To evaluate the quality of the resulting models the accuracy obtained by each method on an independent test set including 30% of data has also been reported.

Acknowledgements

This work has been partially supported by the Italian MIUR Flagship Project "InterOmics".

References

1. Boros E et al. (2000) An implementation of logical analysis of data. IEEE Transactions on Knowledge and Data Engineering, 12(2):292-306.
2. Duda RO, Hart PE, Stork DG (2001) Pattern Classification. New York: John Wiley & Sons.
3. Michie D et al. (1994) Machine Learning, Neural, and Statistical Classification. London: Ellis-Horwood.
4. Muselli M (2006) Switching neural networks: A new connectionist model for classification. In WIRN/NAIS 2005, vol. 3931 of Lecture Notes in Computer Science. Eds. Apolloni B et al., Berlin: Springer-Verlag, 23-30.
5. Muselli M, Ferrari E (2011) Coupling Logical Analysis of Data and Shadow Clustering for partially defined positive Boolean function reconstruction. IEEE Transactions on Knowledge and Data Engineering 23(1):37-50.