Rulex’s Logic Learning Machines successfully meet biomedical challenges.

Rulex is a predictive analytics platform able to manage and to analyze big amounts of heterogeneous data. With Rulex, it is possible, by means of machine learning methods, to create models that describe the data provided by the user, without any prior information. The models generated by Rulex can be used to forecast the behavior of the system in future situations. Besides standard statistical and machine learning methods, Rulex incorporates innovative approaches, named Logic Learning Machines (LLMs): the peculiarity of LLMs consists in the possibility of generating intelligible rules about the problem at hand. As a matter of fact, models produced by standard machine learning methods are able to forecast the future behaviors but, being described by complex equations, cannot provide to the user an insight of the studied system. On the contrary, the LLMs approach generates a set of threshold rules that can be easily understood by a human being. Consider, for example, the case of medical diagnosis. While standard machine learning methods provide the state of a patient (e.g. ill or healthy) as a complex function of the inputs (e.g. blood_pressure, markers etc…), LLMs produce rules in the form:

IF “blood_pressure > x”
AND “lymphoid_cell_concentration > y”
AND “…..”
then “Patient is sick”

where \( x \) and \( y \) are thresholds automatically determined by the algorithm. The number of conditions in a rule is variable and depends on the complexity of the considered problem. The rules generated by Rulex-LLMs can be interpreted by a physician and, in case, adjusted on the basis of her experience. Moreover, Rulex-LLMs, with no additional effort, a ranking of the most relevant features in determining, for example, if a patient is sick or healthy. Thus, it is possible to discover if a variable is not important and, in case, to remove it from the list of the quantities to be monitored.
Rulex-LLMSs has been employed to solve problems related to different application fields. In particular, several applications in the biomedical area have been carried out successfully: some examples are reported below.

1. Extraction of rules for pleural mesothelioma diagnosis

Malignant pleural mesothelioma (MPM) is a rare highly fatal tumor, whose incidence is rapidly increasing in developed countries due to the widespread past exposure to asbestos in environmental and occupational settings. The correct diagnosis of MPM is often hampered by the presence of atypical clinical symptoms that may cause misdiagnosis with either other malignancies (especially adenocarcinomas) or benign inflammatory or infectious diseases (BD) causing pleurisies. Cytological examination (CE) may allow to identify malignant cells, but sometimes a very high false negative proportion may be encountered due to the high prevalence of non-neoplastic cells. Moreover, in most cases a positive result from CE examination only does not allow to distinguish MPM from other malignancies [3].

Many tumor markers (TM) have been demonstrated to be useful complementary tools for the diagnosis of MPM. In particular, recent investigations analyzed the concentrations of three tumor markers in pleural effusions, namely: the soluble mesothelin-related peptide (SMRP), CYFRA 21-1 and CEA, and their association with a differential diagnosis of MPM, pleural metastasis from other tumors (MTX) and BD. SMRP showed the best performance in separating MPM from both MTX and BD, while high values of CYFRA 21-1 were associated to both MPM and MTX. Conversely, high concentrations of CEA were mainly observed in patients with MTX. Taken together, these results indicate that information from the three considered markers and from CE might be combined together in order to obtain a classifier to separate MPM from both MTX and BD.

In this context, Rulex has been applied for the differential diagnosis of MPM by identifying simple and intelligible rules based on CE and TM concentration. The results have been compared to those obtained by other supervised methods showing that Rulex outperforms all the competing approaches (Decision Trees, K-Nearest Neighbors and Artificial Neural Networks).

2. Extraction of a simplified gene expression signature for neuroblastoma prognosis.

Cancer patient’s outcome is written, in part, in the gene expression profile of the tumor. In this study, a 62-probe sets signature (NB-hypo) to identify tissue hypoxia in neuroblastoma was previously identified and showed to stratify neuroblastoma patients in good and poor outcome. It was important to develop a prognostic classifier to cluster patients into risk groups benefiting of defined therapeutic approaches.

Novel classification and data discretization approaches can be instrumental for the generation of accurate predictors and robust tools for clinical decision support. In this paper, Rulex was applied to gene expression data; in particular the Attribute Driven Incremental Discretization technique for transforming continuous variables into simplified discrete ones was employed as a pre-processing step for rule extraction by means of Logic Learning Machine. The application of Rulex-LLMs produced 9 rules utilizing mainly two conditions of the relative expression of 11 probe sets. These rules were very effective predictors, as shown in an independent validation set, demonstrating the validity of the Rulex-LLMs applied to microarray data and patients’ classification. Rulex-LLMs performed as efficiently as Prediction Analysis of Microarray and Support Vector Machine, and outperformed other learning algorithms such as C4.5. Rulex carried out a feature selection by selecting a new signature (NB-hypo-ll) of 11 probe sets that turned out to be the most relevant in predicting outcome among the 62 of the NB-hypo signature. Rules are easily interpretable as they involve only few conditions.

3. Extraction of intelligible rules concerning the prognosis of neuroblastoma.

Neuroblastoma is the most common pediatric solid tumor. About fifty percent of high risk patients die despite treatment making the exploration of new and more effective strategies for improving stratification mandatory. Hypoxia is a condition of low oxygen tension occurring in poorly vascularized areas of the tumor associated with poor prognosis. The aim of this study was the development of a prognostic classifier of neuroblastoma patients’ outcome blending existing knowledge on clinical and molecular risk factors with the prognostic NB-hypo signature. Classifiers outputting explicit rules, that could be easily translated into the clinical setting, are particularly interesting in this context.

Rulex-LLMs exhibited a good accuracy and promised to fulfill the aims of the work. This algorithm was utilized to classify NB-patients on the bases of the following risk factors: Age at diagnosis, INSS stage, MYCN amplification and NBhypo. The algorithm generated explicit classification rules in good agreement with existing clinical knowledge. Through an iterative procedure, the examples causing instability in the rules were identified and removed from the dataset. This workflow generated a stable classifier, very accurate in predicting good and poor outcome patients. The good performance of the classifier was validated in an independent dataset. NB-hypo was an important component of the rules with a relevance similar to that of tumor staging.

4. Validation of a new classification for Multiple Osteochondromas patients.

Multiple osteochondromas (MO), previously known as hereditary multiple exostoses (HME), is an autosomal dominant disease characterized by the formation of several benign cartilage-capped bone growth defined osteochondromas or exostoses. Various clinical classifications have been proposed but a consensus has not been reached. The aim of this study was to validate (using a machine learning approach) an “easy to use” tool to characterize MO patients in three classes according to the number of bone segments affected, the presence of skeletal deformities and/or functional limitations. The proposed classification has been validated (with a highly satisfactory mean accuracy) by analyzing 150 different variables on 289 MO patients through Rulex LLMs. This approach allowed us to identify ankle valgism, Madelung deformity and limitation of the hip extra-rotation as “tags” of the three clinical classes. In conclusion, the proposed classification provides an efficient system to characterize this rare disease and is able to define homogeneous cohorts of patients to investigate MO pathogenesis.

5. Benchmarking Rulex-LLMs performances on standard biomedical datasets.

In this study, Rulex-LLMs were applied to three benchmark datasets regarding different biomedical problems. The datasets, are taken from the UCI archive, a collection of data for machine learning benchmarking, and include:

- **Diabetes**: it regards 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.

- **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.

Rulex-LLMs’ performances were compared to those of other supervised methods, namely Decision Trees (DT), Artificial Neural Networks (ANN), Logistic Regression (LR) and K-Nearest Neighbor (KNN). These tests showed that Rulex-LLMs results are better than those of ANN, DT (that produce rules) and KNN, and are comparable with those of LR.


**Bibliography**


