Project Title:	Human Brain Project
Sub-Project Title:	Medical Informatics Platform (SP8)

Document Title:	MIP Tools – User Guideline V1.0 Public release

Summary:	Step-by-step guidelines of all the Data-Mining and Machine Learning tools of the MIP Audience: any users of Medical Informatics Platform

Versions, changes and contributors:

Authors	Version	Date	Status	Change Details
Elia Sbeiti (UNIGE / FBF)	V1	29/01/2016	Draft 1	Initial draft
Alberto Redolfi (UNIGE / FBF)	V1.1	29/01/2016	Draft 1.1	To be reviewed before the release

TABLE OF CONTENTS

The Medical Informatics platform offers a wide list of Data mininig and Machine learning tools for patients classification exploiting a bottom up approach (from data to diagnosis). The main tools are reported hereinafter:

- Semi-supervised rule-based clustering algorithm:

  The rule-based algorithm aims to explain the variability between individuals, and describes a population by a group of “local over-densities”. These are defined as subspaces over combinations of variables. The algorithm performs an exhaustive search of the data space to predict the outcome variables. A typical use case consists of the health status of each subject in terms of the presence or absence of AD. In our experiment, the predictive variables are the 90 brain region volumes, age, gender, and individual subject global volumes.

- Informatics-based Model - Enriched Automated Diagnostic Tools:

  This is an automated classifier from a set of MRI scans that came from deceased, pathologically diagnosed individuals. This classifier provides prognostic value on clinically categorised living people.

- Informatics-based Model - Deep Learning for Automated Features Extraction:

  The increasing calculation power of computers has led to a rising interest in complex machine learning methods. In particular, the investigation of artificial neural networks with many hidden layers continuously results in promising new applications. These include image and face recognition (1, 2, 3), speech recognition (1, 2) and signal processing (1). Very recently, these deep learning networks have also been used in the classification of AD patients versus healthy control subjects, resulting in accuracies of up to 95% (Suk, Heung-Il; Shen, Dinggang; Deep learning-based feature representation for AD/MCI classification Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013 583-590,2013).

Informatics-based Model - Rasch Model and Factor Analysis for Learning Disease Severity:

We propose to extract an index or a latent variable from the neuroimaging data to quantify the disease severity for each subject and regional vulnerability by applying factor analysis. Since the atrophy pattern correlates with the loss of neurons, this severity has a biological meaning and is independent of symptoms. We aim to test if the estimated severity significantly associates with clinical diagnosis and identify the regions highly weighted in calculating the severity.

Informatics-based Model - Bi-Clustering Applied to Gene Expression and Brain Volumetric Data:

The Ping-Pong Algorithm (PPA) is a bi-clustering method used to compare two datasets with one common dimension. The algorithm uses a random weighted set of genes as a starting point (called a seed). It then selects subjects in which these genes deviate from the mean across subjects. Using these subjects, it then selects brain regions whose volumes deviate from the mean in these subjects. Using these regions, it selects a second vector of subjects for which the volumes of these regions deviate from the mean across subjects. Finally, it selects a new set of genes whose expression in these subjects deviates from the norm. This process is repeated until convergence is reached (i.e. the gene, subject, and region sets do not change from one iteration to the next).

Informatics-based Model - Bayesian Causal Model:

We aimed at designing, equations derivation and implementing the causal Bayesian model similar to the General Linear Model (GLM) for distributed Data. GLM is one of the most used models to estimated dependences between clinical, neuropsychological and Neuroimaging variables. In our case the data is distributed in different Hospitals and it is not possible to move them to a unique Federation node where the GLM could be computed in a classical way. Therefore special equations should be developed that allow us to have reliable GLM estimations under this condition. The Bayesian Formalism provides us the necessary armamentarium to deal with it and offers general sophisticated ways to extend to other models and managing high dimensional and Multimodal Data.

- Disease subtypes signatures - Big Medical Data Strategy (3-C):

We offer a 3C strategy (Categorize, Cluster & Classify) that starts from the medical knowledge, categorizing the available set of features into three types: the patients' assigned disease diagnosis, clinical measurements and potential biological markers, proceeds to an unsupervised learning process targeted to create new disease diagnosis classes, and finally, classifying the newly proposed diagnosis classes utilizing the potential biological markers. Our strategy, developed as part of the medical informatics work package at the EU Human Brain flagship Project strives to connect between potential biomarkers, and more homogeneous classes of disease manifestation that are expressed by meaningful features.

Label Propagation Framework is a MIP function providing a Feature Extraction Framework. Data mining is to be based on a number of brain structure volume features, which are automatically extracted from patient MRI scans.

- Multi-Target Regression on Data Streams is a MIP function that implements the FIMT-DD and iSOUP-Tree algorithms for learning decision trees from data streams, the former for single-target prediction and the latter for multi-target prediction. Both of these algorithms produce models in the form of a model tree, i.e., a decision tree, which uses linear functions in the leaves to achieve better performance.

- Predictive Clustering Trees combine aspects from both predictive modeling and clustering. Predictive clustering trees (PCTs) partition the set of examples into subsets in which examples have similar values of the target variable, while clustering produces subsets in which examples have similar values of the descriptive variables. The task of predictive clustering is to find clusters of examples which have similar values of both the target and the descriptive variables.

- The Rule Ensembles function implements the FIRE algorithm, which employs the rule ensembles approach for solving multi-target regression and time series problems. We can improve the accuracy of the rule model by adding simple linear functions to the ensemble, which results in a model that is a combination of global linear functions and rules.

- The Feature Ranking for Structured Targets implements two algorithms for feature ranking for structured targets: (1) RF-RANK exploits the random forests mechanism and (2) GENIE3 exploits the variance reduction at each tree node from the ensemble. For the latter method, the ensemble could be random forest or an ensemble of extra trees. The both methods use predictive clustering trees as base predictive models.

- Subgroup Discovery from Multi-Resolution Data is a tool that can use ontological domain knowledge (RDF graphs) in the learning process. Subgroup descriptions contain terms from the given domain knowledge and enable potentially better generalizations.

- The Subgroup Discovery from Heterogeneous Data performs network propositionalization on a heterogeneous network by first deconstructing the network into several homogeneous networks. The homogeneous networks are constructed using user-supplied meta-paths in the heterogeneous network (for example, in a network consisting of papers and their authors, we can construct a homogeneous network of papers where two papers are connected if they share an author). The result of this function is a set of feature vectors, one for each node of the target type. Recently, the function was updated so that it can accept not only heterogeneous networks, but also standard data instances (with feature vectors) as input. In that case, the function constructs a proximity network of instances and performs network propositionalization on the resulting network.

- Brainspan Co-Expression Clustering is a multi-dimensional co-expression analysis method for extracting disease signatures. We used BrainSpan human transcriptome database for our experiments.

- BH-tSNE is a non-linear dimensionality reduction (DR) algorithm to visualize high-dimensional data. The main advantage of non-linear DR algorithms in comparison to linear DR algorithms such as PCA is that non-linear DR algorithms can represent neighbouring samples in the high-dimensional space better than linear DR algorithms in the lower dimensional space such as 2D or 3D. This advantage is crucial to visualize similarities between the features of samples of the data (such as gene expressions, disease phenotypes), hence to observe possible correlations between the samples and classify them in the same cluster.