DMLAB – MCFS-ID

MANUAL GUIDE V.2

DATA EDITOR

Data editor helps to create and modify text files that contain input data.

HOW TO RUN IT?

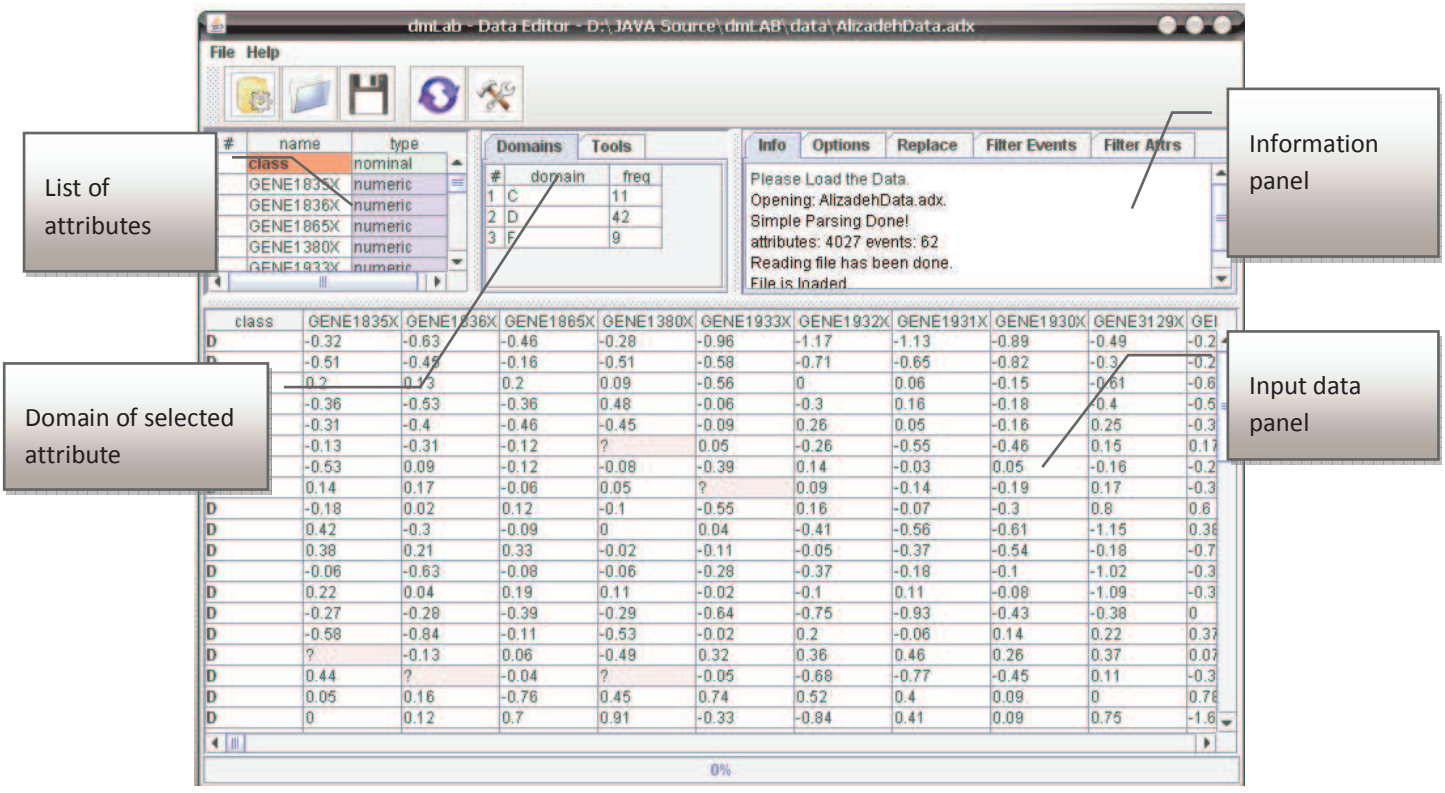
This software needs to have installed Java JRE (for more information please see http://java.sun.com).

Under Windows: run *editor.bat* file.

Under Linux: run *editor.bash* file.

HOW TO WORK WITH IT?

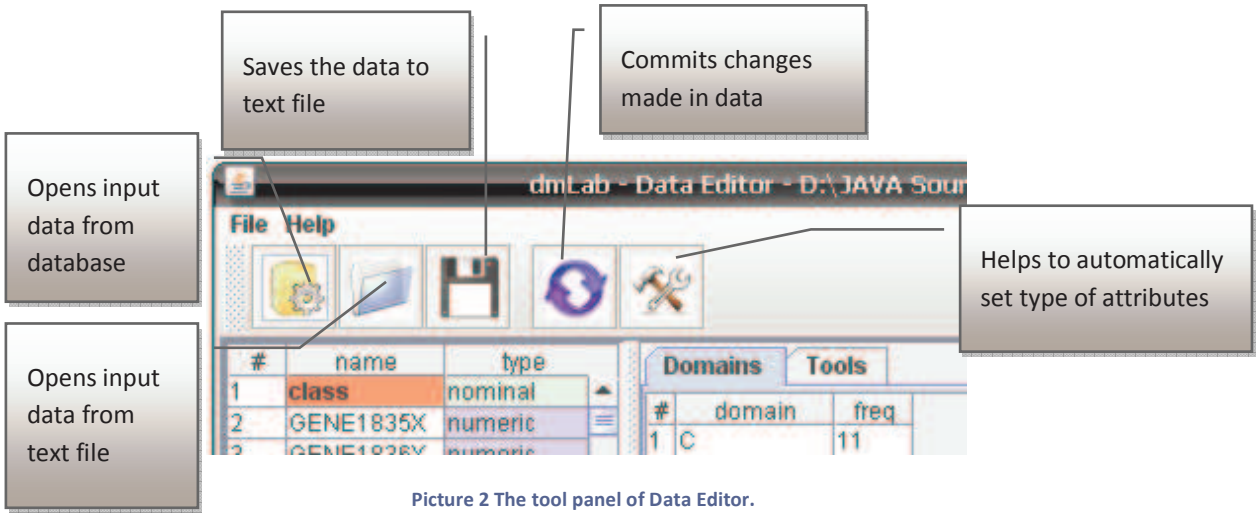
THE MAIN PANEL



Picture 1 The main panel of Data Editor.

The main panel shows all attributes and their values. All values can be modified and missing data are marked by character '?' (cell is in light red color). Currently selected row and column is marked by bold fonts. For each attribute, one can set/change its type and set it as decision attribute.

THE TOOL BAR

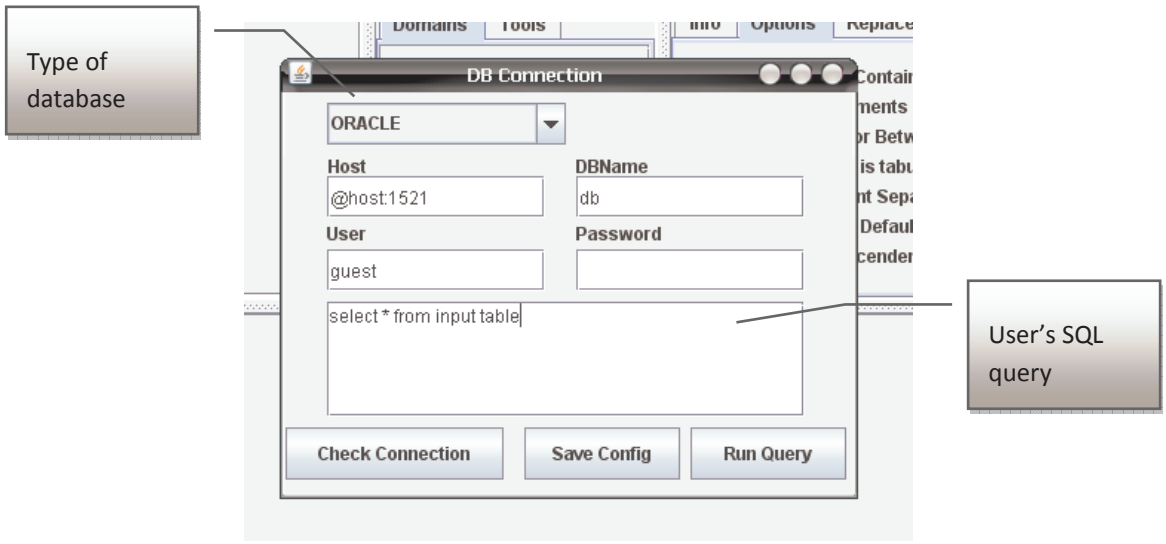


Picture 2 The tool panel of Data Editor.

Input file can be prepared in the following format:

- csv – comma separated values. However, input file does not need to keep all restrictions of csv. In the Options setting you can configure the following settings:
    - separator – it can be any character (also white char e.g. tabulation)
    - special treating of first line – the first line in your csv file can contain names of attributes
    - comment character – csv file also can be commented by # character
    - default name of attribute – if the first line does not contain names of attributes, they will be created automatically
    - treating of consequent separators – how to treat a few separators without any specified value (as one separator or as set of missing values)
- arff – native format of Weka data mining tool.
- adx – native format of dmLab.

DATABASE CONNECTION



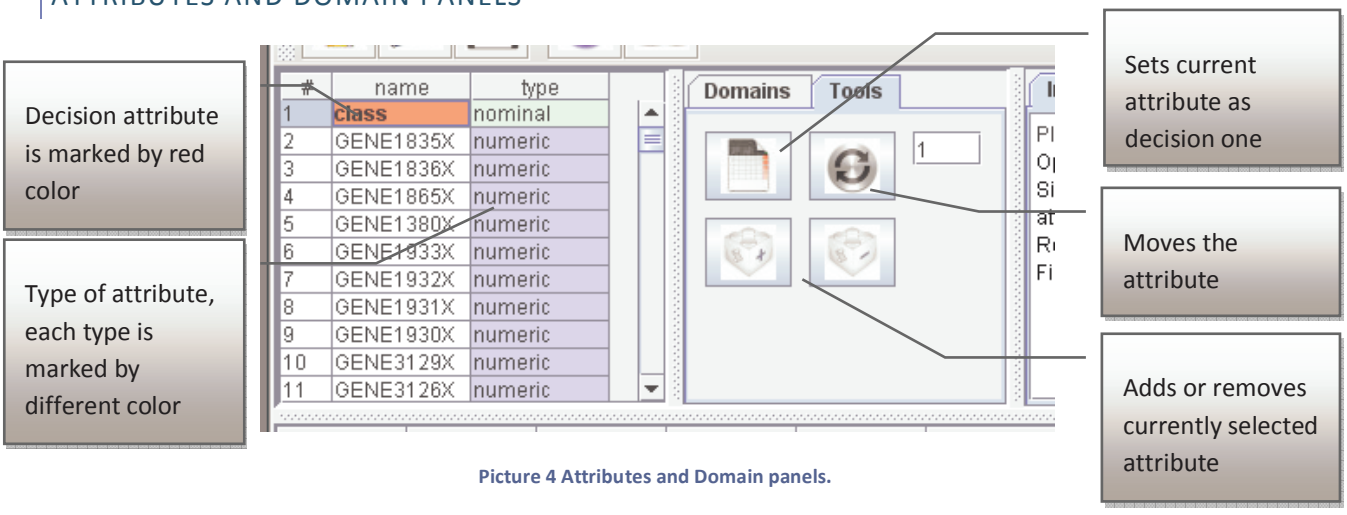
Picture 3 Opening data from Database.

Input data can be opened from database, dmLab Data Editor supports four types of databases:

- Oracle
- MS SQL
- MySQL
- Postgres

User has to set database name, configure the connection to the host, user's name and password. Selection of table and its columns and rows can be processed by typing appropriate SQL query.

ATTRIBUTES AND DOMAIN PANELS

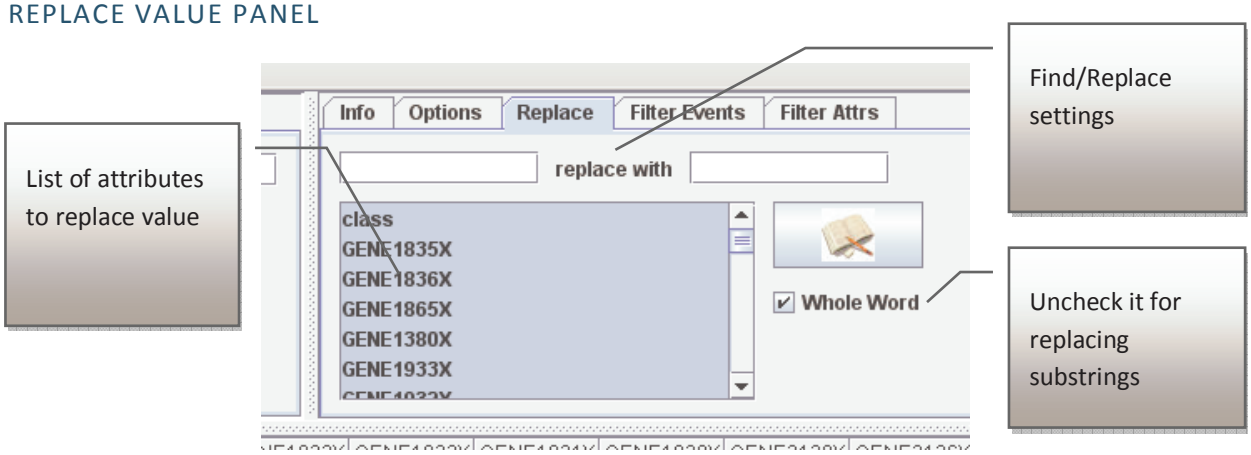


Picture 4 Attributes and Domain panels.

In these panels you can change the attribute's name or type, add or remove currently selected attribute, move the attribute by a specified number of positions (forward or backward) or set the attribute as decision

attribute. The Data Editor supports three types of attributes: nominal, numeric, integer. However for most of the supported formats integer is treated in the same way as numeric.

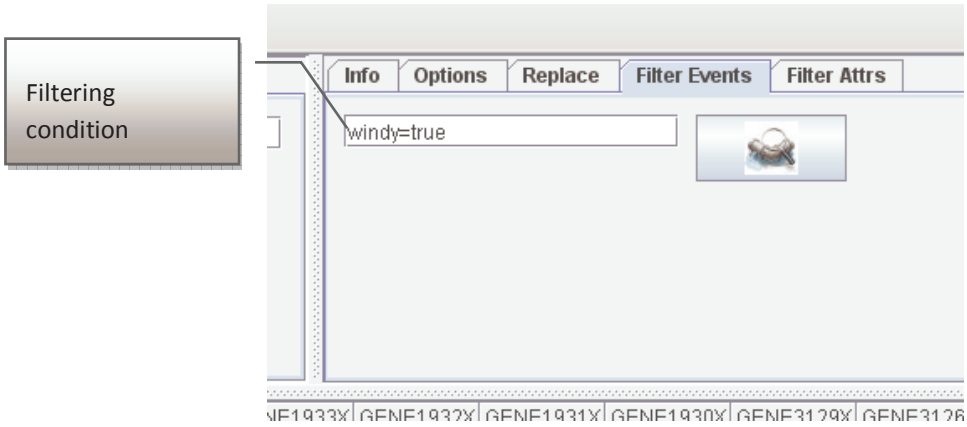
REPLACE VALUE PANEL



Picture 5 Replace value panel.

Here you can replace any string/substring for selected set of attributes. It is a good and fast way to replace many values for even huge input datasets.

FILTER EVENTS PANEL



Picture 6 Filter events panel.

This panel helps to remove rows that meet filtering condition. The expression is defined as following:

Attribute_name Operator Value

Possible operators are: =, <, >, <=, >=, != (not equal). For nominal attributes only = and != are allowed.

ADX FILE FORMAT

```
#comment starts with -> #
#attributes section
attributes
{
 outlook nominal
 temperature numeric ignore
 humidity numeric
 windy nominal
 play nominal decision(all)
}
#events section
#missing value is denoted by -> ?
events
{
 sunny,85,85,false,no
 sunny,80,90,true,no
 overcast,83,86,false,yes
 rainy,70,96,false,yes
 rainy,68,80,false,yes
 rainy,65,70,true,no
 overcast,64,65,true,yes
 sunny,72,95,false,no
 sunny,69,70,false,yes
 rainy,75,80,false,yes
 sunny,75,70,true,yes
 overcast,72,90,true,yes
 overcast,81,75,false,yes
 rainy,71,91,true,no
}
```

Example of ADX file (popular weather data)

Definition of attribute contains the name of attribute, type of attribute (*nominal* or *numeric*), role of attribute (*ignore* for ignoring the feature, *decision* if the feature defines classes).

MCFS-ID

MCFS (Monte **C**arlo **F**eature **S**election) is an algorithm to build the ranking of attributes/features which reflects their importance. Importance is measured based on attribute's prediction ability. If a given feature gained higher RI (relative importance) measure then the feature is more important. The algorithm is well described in the paper "Monte Carlo feature selection for supervised classification". If you are interested in MCFS algorithm please read the paper.

Next version of MCFS is called MCFS-ID and it also finds interdependencies between features. These interdependencies can be presented in a form of graph. Where nodes represent features and paths - strongest of interdependency between features.

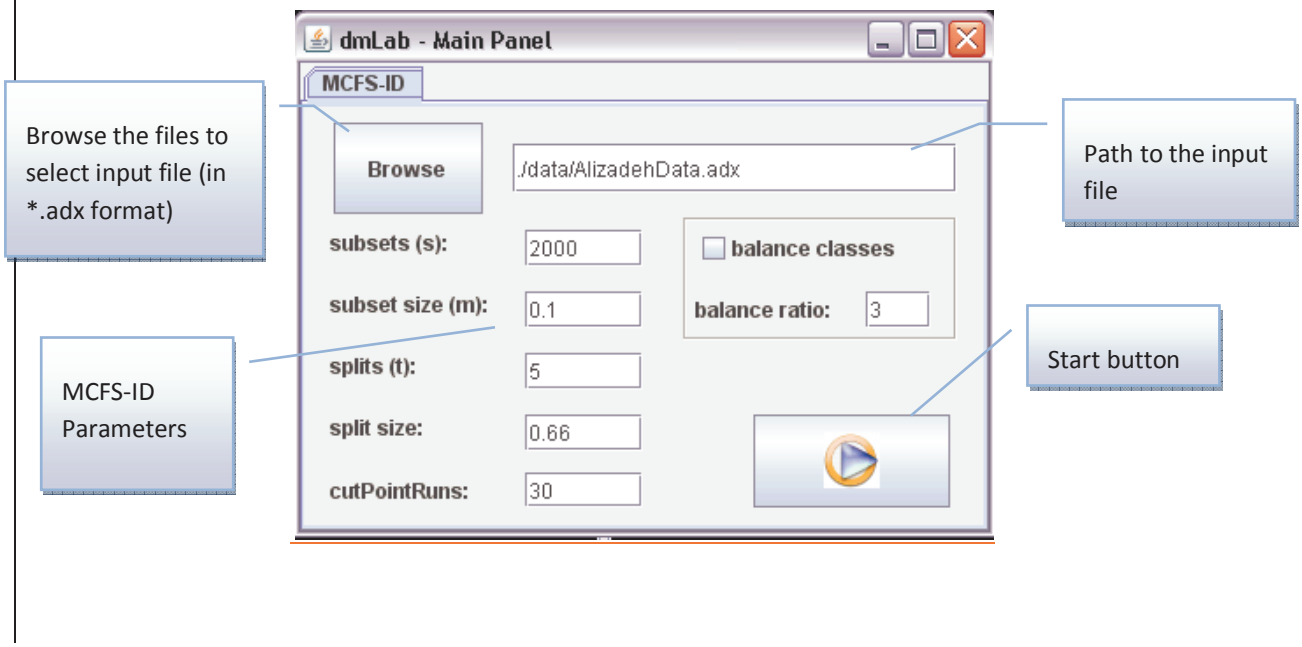
HOW TO RUN IT?

This software needs to have installed Java JRE (for more information please see http://java.sun.com).

Under Windows: run *dmLab.bat* file.

Under Linux: run *dmLab.bash* file.

HOW TO WORK WITH IT?



Subsets (s) – defines number of subsets (projections) with randomly selected features. This parameter should be set on a few thousands.

Subset size(m) – defines number of features in one subset. It can be defined by absolute number (e.g. 100 denotes 100 randomly selected features) or relatively by fraction of input attributes (e.g. 0.05 denotes 5% of randomly selected features)

Splits (t) – defines number of splits of each subset. Each subset is randomly divided into training and testing sets *t* times.

Split size – defines size of training set. This is the fraction of events in input subset.

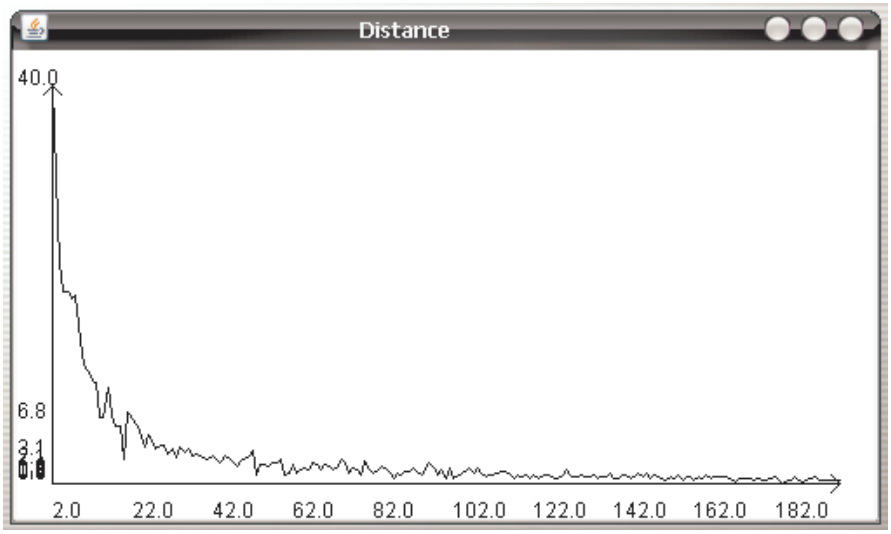
Balance classes – select it if input dataset contains heavily unbalanced classes. Each subset *s* will contain all events from the smallest class and randomly selected set of events from each of the rest classes. This function is helpful to select features that are important for discovering a relatively small class.

Balance ratio – defines possible maximal size of class related to the smallest one. If the ratio is set to 3, it means that each of the class within subset *s* will contain maximally 3 times more events than the smallest class.

cutPointRuns – defines number of additional experiments where decision variable is permuted. This operation helps to find the set of informative features as well as connections between them. It is recommended to process at least 30 additional experiments (with permutation) to have strong statistical result for both cut points.

OUTPUT

When MCFS is running, the distance measure is calculated and presented in the specified graph. The distance is calculated between the following rankings and it decreases with the number of iterations. If two rankings are identical then the distance between them is zero. If distance measure is stabilized and is close to zero, MCFS can be stopped. In the picture below, the distance after a sufficient number of subsets (iterations) is presented.



MCFS-ID produces also 6 types of output files (all of them are contained in *./resources/* directory):

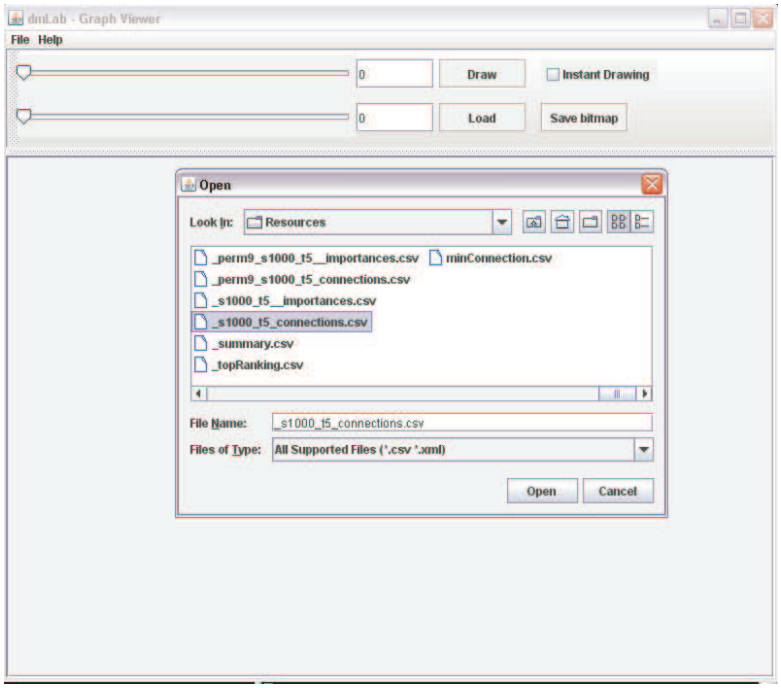
- _acc_wAcc.csv – contains accuracy (*acc*), weighted accuracy (*wAcc*) and number of incorrect classified of events (*errors*) for each single classification (*st*).
- _dist.csv – contains *distance* and *common part* measures calculated for 5% of top ranked features between two rankings. Rankings are compared every 10 subsets *s*.
- _s2000_t5__histogram.csv – contains all input features and their RI measure. The RI described in the paper is located in the last column. Features in this file are not ordered by RI, they are given in the same order as in the input file. The prefix of this file corresponds to parameters *s* and *t* that have been set for the experiment.
- _s2000_t5_ranking.csv – contains 5% of top estimated input features ordered by RI measure. The prefix of this file corresponds to parameters *s* and *t* that have been set for the experiment.
- _s1000_t5_connections.csv – contains all connections between features from original experiment. This file is an input for graph viewer.

- minConnection.csv – contains minimum value of important interdependency that is above of random interdependency.
- _perm1_*resultfile*.csv – set of results (same as above for original data) processed for dataset where decision attribute was permuted. The number after "_perm" prefix refers to the number of experiments.

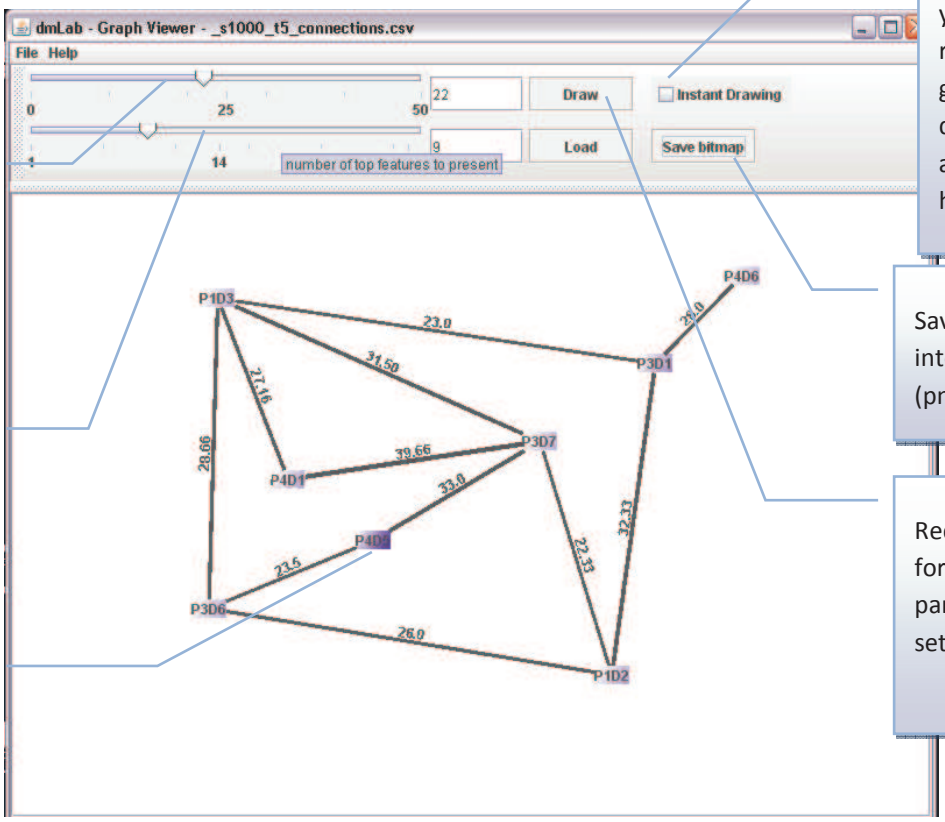
IMPORTANT NOTE: The output for original experiment does not have the prefix "*_perm*". The experiments with permuted decision attribute have alwas prefix "*_perm*".

GRAPH VIEWER

The graph viewer is a tool to present in graphical way statistically important interdependencies between features. The input file for the application comes from original experiment (with suffix "*_connections*").



Select "_connections.csv" file for the original experiment.



Minimum value of interdependency to be presented as edge.

Number of top ranking features to be presented as nodes.

Saturation of the label is related to the importance of the feature. This feature is the most important.

Check this option if you want instant redrawing the graph. It may slow down the application for huge graphs.

Saves the graph into bitmap file (png).

Redraws the graph for selected parameters (slides settings)

LICENCE

dmLab software is available for noncommercial use only.

Whenever the software is used, the original works of :

- *M.Dramiński, A.Rada-Iglesias, S.Enroth, C.Wadelius, J. Koronacki, J.Komorowski "Monte Carlo feature selection for supervised classification", BIOINFORMATICS 24(1): 110-117 (2008).*
- *Michał Dramiński, Marcin Kierczak, Jacek Koronacki, Jan Komorowski „Monte Carlo feature selection and interdependency discovery in supervised classification" in "Advances in Machine Learning II · Dedicated to the memory of Professor Ryszard S. Michalski", Koronacki, J., Ras, Z.W., Wierzchon, S.T., Kacprzyk, J. (Eds.), Vol. 263, 2010, ISBN 978-3-642-05178-4*

must be cited.

Modifications and redistribution of this software are not allowed.