Business insight

TIMi suite

integrated data mining solutions

An automated predictive datamining tool
Data Preparation – File Format v1.05

March 2015

Data preparation - Introduction

When using **TIM*i***, it is strongly suggested to accumulate the highest number of information (rows and columns) about the process to predict: Don't reduce arbitrarily the number of rows or columns: always keep the "full datatset" (even if the target size is less than one percent). **TIM*i*** will use this extra information (that is not available to other "classical" datamining softwares that require sampling in order to work) to produce great predictive models that, 99% of the time, outperform the predictive models constructed with any other datamining software. When using classical tools (such as SAS, SPSS, KXEN, etc.), we know that it's a very common miss-practice to perform a strong sampling on the dataset, this is why we insist here that you don't perform any kind of sampling.

Of course, to be completely rigorous, you should also not forget to "let aside" a TEST dataset that will be used to really assert the quality of the delivered predictive models (to be able to compare in an objective way the models constructed with different predictive datamining tools). See this web page that explains the importance of the TEST set:
http://www.business-insight.com/html/intelligence/bi_test_dataset.html

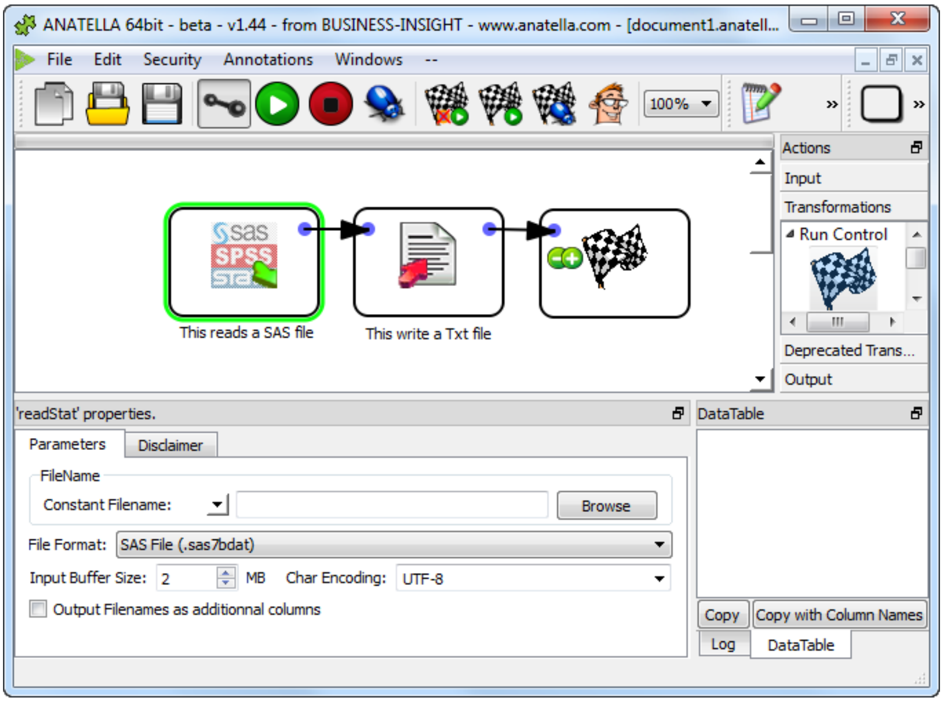
Creation/Learning dataset file format

TIMi can read datasets from many data sources: simple "txt" or ".csv" flat files, SAS files, ACCESS files, ODBC & OleDB links to any databases (Teradata, Oracle, SQLServer,etc.). **But** for a first "quick benchmark", it's suggest to store the ***creation dataset*** inside a simple "txt" or ".csv" flat file (in order to prevent any inter-operability errors).

The "txt" or ".csv" flat file should follow the following format:

- The ***creation dataset*** is a "txt" or ".csv" file where the separator is a dot-comma ';'. The first line of the file must contain the column names.

> WARNING: "txt" files exported from SAS have a size limitation: one line cannot exceed 65535 characters. If you encounter this bug in SAS, the easiest solution is the following: Convert the .sas7bdat SAS file to a simple "txt" file using Anatella (or even better: convert to a .gel_anatella file!): Use the following Anatella-data-transformation-graph:



'readStat' properties.

FileName

Constant Filename:

File Format: SAS File (.sas7bdat)

Input Buffer Size: 2 MB Char Encoding: UTF-8

☐ Output Filenames as additionnal columns

- Column names must be unique.

  <u>WARNING:</u> TIMI is case IN-SENSITIVE (as is SQL)

- Column names are NOT within quotes.

- The data in the columns are NOT within quotes (never).
- The field separator character (here ' ;') is not allowed (neither in the data, neither in the column names).

- The ***creation dataset*** contains <u>one</u> unique primary key.

- The decimal character is a dot and not a comma (Standard English notation or Scientific notation for numbers).

- If The ***target*** column (the column to predict) is:
  - <u>Binary:</u> then it must contains only '0' and '1' values (and the "one's" are the value to predict and must be the **minority case**).
  - <u>Continuous:</u> then it should not contain any "missing value".

- Missing values must always be encoded as empty values ("").

- <u>OPTIONAL:</u> The ***creation dataset*** should not contain any "consequence columns". If the dataset nevertheless contains some "consequence columns", it's good to know their name in advance. However, you can always use **TIMi** to find all the "consequence columns" easily.

- <u>OPTIONAL:</u> the flat file can be compressed in RAR (.rar), GZip(.gz), Winzip(.zip)

- OPTIONAL: all the columns that represent a “True/False” information may contain only two different value: ‘0’ (for false) or ‘1’ (for true) or are empty (“”) if the value is missing.

- OPTIONAL: all the columns that represent either:
  - a number
  - an information that can be ordered

  … should be encoded as pure number. For example:

number of cats
missing
no cat
one cat
2 cats
3 or more

number of cats
0
1
2
3

Social class
missing
poor
middle
rich

social class
0
1
2