Knowledge Extraction From High Dimensional Data Using Multileveled Self-organization

Frank Lemke

Abstract

Introduction

Analysis of mass spectral data in order to find peptides responsible for a certain disease goes beyond data mining: Knowledge extraction is necessary. Extracting new, generalizable knowledge from a finite set of data is an ill-posed problem. It is characterized by (1) inadequate a priori (domain) knowledge about the object; (2) large number of potential, often immeasurable variables that may describe the object's behavior; (3) noisy and few data samples; and (4) fuzzy objects [1]. To solve this problem in an objective way, Inductive Learning is required [2, 1]. An inherent part of Inductive Learning is regularization and here, in turn, one concept is making the learning process task-driven.

Fayyad [3] describes knowledge discovery from data (KDD) as an iterative and complex process of several subtasks, roughly divided into preprocessing, data mining, and evaluation and application of results. This implies, consequently, that *every* subtask of the whole process needs some inductive approach, in order to fit the initial goal, i.e., extracting knowledge from data.

One approach in that direction is self-organizing data mining (SODM) [1, 4] that employs self-organizing modeling technologies at different levels of the KDD process, but also within each process subtask as well. In this way, data transformation, dimension reduction, variables selection (preprocessing), data mining, and validation of models [4, 5] flow together into a single, autonomously running process.

Methods

SODM is built on the Group Method of Data Handling (GMDH) approach [1, 6] and can be composed of different self-organizing modeling technologies. We mainly used GMDH here, too, to generate parametric regression models, but also Fuzzy Rule Induction (FRI) for creation of rules of linguistic terms and Analog Complexing based classification (AC class) for providing a similar cases view were applied on reduced data sets. In addition to the black-box method and the connectionist structure of Neural Networks, SODM implements principles of induction: Self-organization, regularization of ill-posed tasks, and integration of external information. The point here is that, during the learning process, *every* synthesized model candidate (hypothesis) is validated (hypothesis tested) based on some information that was not yet used for generating that model/hypothesis. In this way it is possible to self-organize an optimal complex model, objectively, that does not overfit the learning data and that is able to describe relevant relations in the data. A hypothesis testing tool that provides good predictive power and that is integrated in our GMDH implementations [7] is the PESS leave-one-out cross-validation criterion. In our analysis, we used a prototype of KnowledgeMiner [7] that takes high dimensional data to synthesize models composed of most relevant input variables, directly, and that implements another stage of model validation.

Results

We analyzed the two PROB2PDS data sets. Following the notes at the conference site's problem description, all peak data was normalized by a fraction's maximum peak value for each specimen separately. Molecular masses that occurred more than once due to fractionation were averaged together to a single variable. This preprocessed data form the data set *D1*. We planned the following approach:

A. Working on *D1* data
   - A.1 Modeling using the entire *D1* data (m/z and peak data, about 22,000 inputs, 1 model);
   - A.2 Modeling each fraction of *D1* separately (about 1100 inputs each, 20 models, multileveled approach up to a final model);

B. Working on *D2* data
   - B.1 Modeling using the entire *D2* data (m/z and peak data, about 2,000 inputs, 1 model);
   - B.2 Modeling each fraction of *D2* separately (about 400 inputs each, 20 models, multileveled approach up to a final model);

The data set *D2* is a lower resolution version of *D1*, first because we realized that the molecular masses are very high resolved and secondly, the idea was to find "hot regions", first, and then zooming into that regions more deeply using higher resolved data (Note: This might not be reasonable from a biological point of view, but it seems to make sense to us non-experts in peptidomics.). So we started mining the *D2* data. Due to time constraints (start: early July), however, it was not possible to do some modeling on the *D1* data until now.

<u>B. Working on *D2* data</u>

From the available 41 specimen, we always used 39 for learning (including leave-one-out cross-validation, see above) and hold out the remaining two samples (A24 and B1) for ex-post prediction.

1. Level data

All molecular masses are rounded to full tens (1972,699 is rounded to 1970, for example), and equal masses are then combined to a single input variable.

<u>B.1</u>

Running a multileveled self-organization on the obtained 2307 less resolved variables, this linear regression model is created autonomously out of many hundred thousand model candidates:

[GMDH1]

X1 = 1.085444 -54.352837X2248 -13.512579X155 -27.053354X437
-17.812988X579 -11.188010X2095 + 0.209273X35 -15.980230X1796
-25.227913X1789 -65.822487X2212 -13.070695X1504
+ 34.639774X2099 -10.229363X434

Coefficient Of Determination (R^2): 0.9757
Descriptive Power (DP): 81 %
ROC Analysis:
TPR: 1.0 FPR: 0.0 Area: 1.0
Prediction error (false classification rate on test samples A24 and B1): 0%

X1 is the target variable, where a 1 is assigned to all samples of class A and a –1 to all cases of B. Each descriptor variable X_i represents the peak values of a given molecular mass.
A fuzzification of the 12 selected descriptors of the GMDH1 model into 3 linguistic terms each (negative N_ (read: small peak value), zero ZO_ (medium), and positive P_ (high)) builds the data base for a fuzzy rule induction, and the following best rule is obtained:

[FRI1]
IF P_X35 **OR** N_X437 **&** N_X1789 **& NOT**_P_X2095 **&** N_X2212 **&** N_X155
& NOT_P_X1796 **&** N_X2248
THEN Has_Disease

Mean Absolute Percentage Error (MAPE): 3.90 %
ROC Analysis:
TPR: 1.0 FPR: 0.0 Area: 1.0
Prediction error: 50%

B.2
20 models are self-organized, one for each fraction. Models of 4 fractions (fraction f7, f11, f12, f15) show a ROC area value of 1.0 *and* a prediction error of zero. From the descriptors selected by the 20 fraction models (about 220), in a second modeling run, this GMDH model is synthesized (from approximately a million model candidates overall):

[GMDH2]
X1 = 1.000457 -6.499664X192 -10.947432X2 -95.786530X141 -10.608231X31
+ 1.263484X7 -6.528339X20 -38.333191X58 -4.832493X125
-4.897674X117 -12.802607X67 -2.636172X98 -130.640793X101
-0.876323X100 + 4.424458X39 -5.826747X75 + 6.399203X200

R^2: 0.9730
DP: 79 %
ROC Analysis:
TPR: 1.0 FPR: 0.0 Area: 1.0
Prediction error: 0%

A corresponding fuzzy model is:

[FRI2]
IF P_X2 **OR** N_X3 **&** N_X5 **&** N_10 **&** N_X11
& NOT_P_X12 **&** N_X15 **&** N_X23 **& NOT**_P_X31
& N_X48
THEN Has_Disease

MAPE: 0.29 %
ROC Analysis:
TPR: 1.0 FPR: 0.0 Area: 1.0
Prediction error: 0%

2. Level data

Based on the results of running the first level data, all corresponding selected descriptors ("hot regions") are then higher resolved in rounding to plain integer values (a molecular mass of 1972,699 is now rounded to 1973, for example). This new data set was then mined, again by running a multileveled GMDH and FRI process:

Continuing B.1 approach

[GMDH3]

X1 = 0.567877 -9.479472X9 + 1.071493X2 -3.564071X49 -19.072380X16
-21.077286X31 -8.697422X14 -44.868900X24 -30.435816X56
-17.136744X21 -3.084588X44 -23.033760X60

R^2: 0.7395

DP: 37 %

ROC Analysis:

TPR: 1.0 FPR: 0.12 Area: 0.97

Prediction error: 0%

[FRI3]

IF **NOT**_P_ X118 **&** **NOT**_P_ X73 **&** N_ X92 **&** N_ X101 **&** N_ X71
& **NOT**_P_ X97 **&** **NOT**_P_ X88
THEN Has_Disease

MAPE: 9.61 %

ROC Analysis:

TPR: 1.0 FPR: 0.12 Area: 0.932

Prediction error: 50%

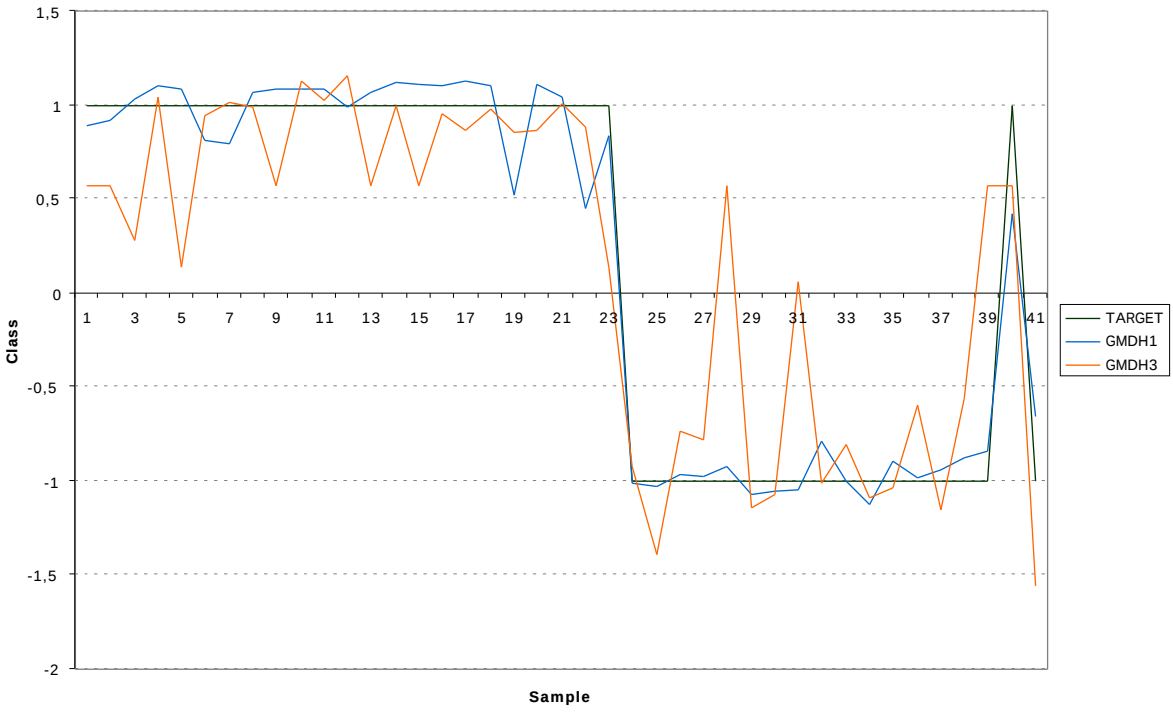

Figure 1: Plot of GMDH1 and GMDH3 models

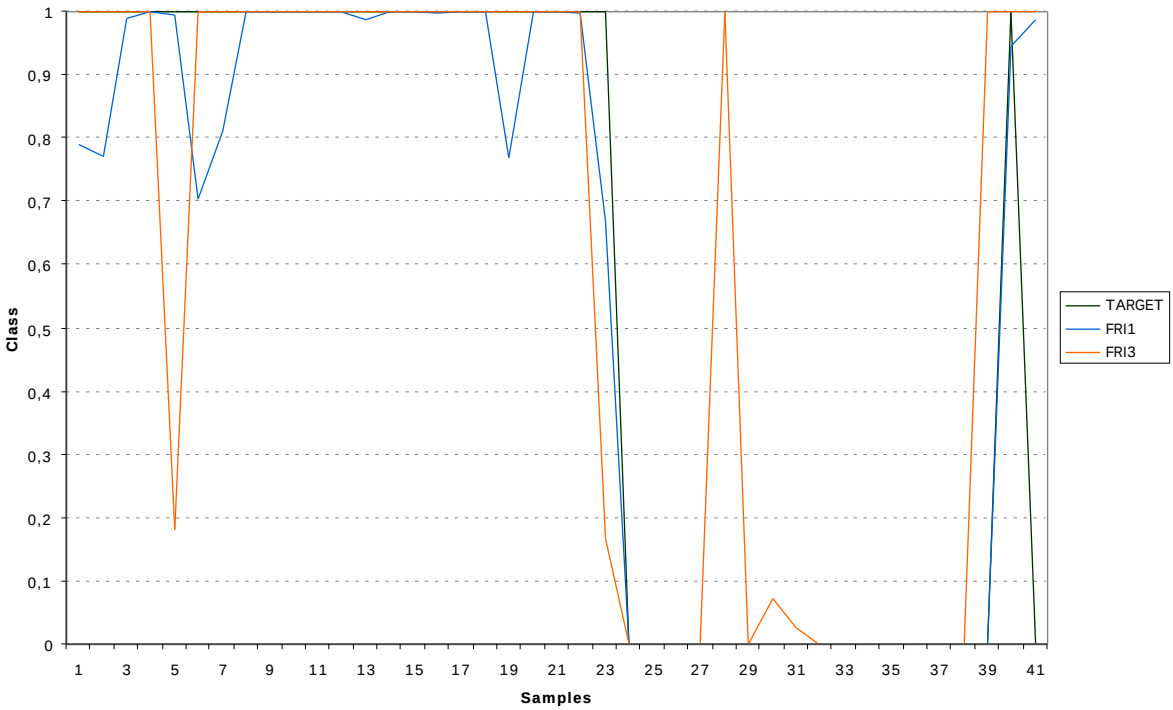

Figure 2: Plot of FRI1 and FRI3 models

Continuing B.2 approach

[GMDH4]

X1 = 0.697969 + 1.554865X4 + 8.081791X56 -104.838165X47 + 1.574077X5
-37.690151X60 -727.811890X45 -24.953989X57 -49.218437X17
-4.013736X67 -15.265388X23 -90.668419X79

R^2: 0.7110

DP: 34 %

ROC Analysis:

TPR: 1.0 FPR: 0.12 Area: 0.954

Prediction error: 0%

[FRI4]

IF P_m6830 **OR** ZO_m1981 **OR** **NOT**_P_m1493 **&** N_m3078
& **NOT**_ZO_m5336 **&** N_m8626 **&** N_m147934 **&** N_m3080 **&** N_m6833
& N_m17431 **&** N_m8631 **&** **NOT**_P_m8873 **&** **NOT**_ZO_m3079
& N_m8630
THEN Has_Disease

MAPE: 3.27 %

ROC Analysis:

TPR: 1.0 FPR: 0.0 Area: 1.0

Prediction error: 50%

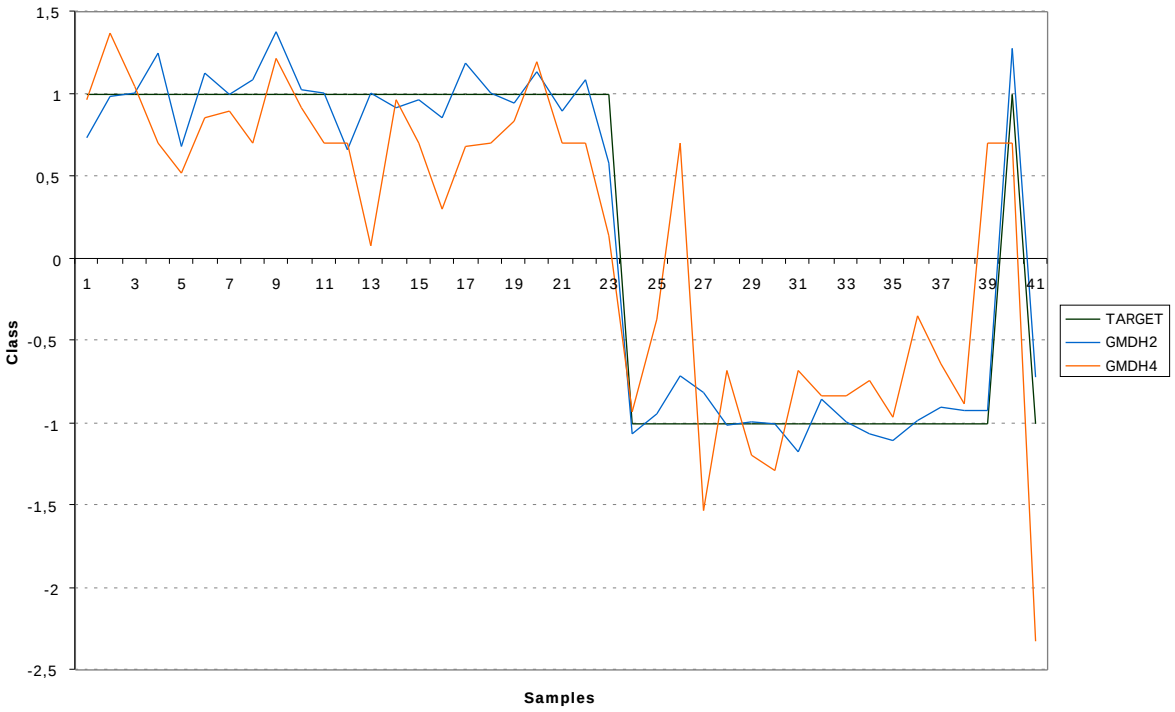

Figure 3: Plot of GMDH2 and GMDH4 models

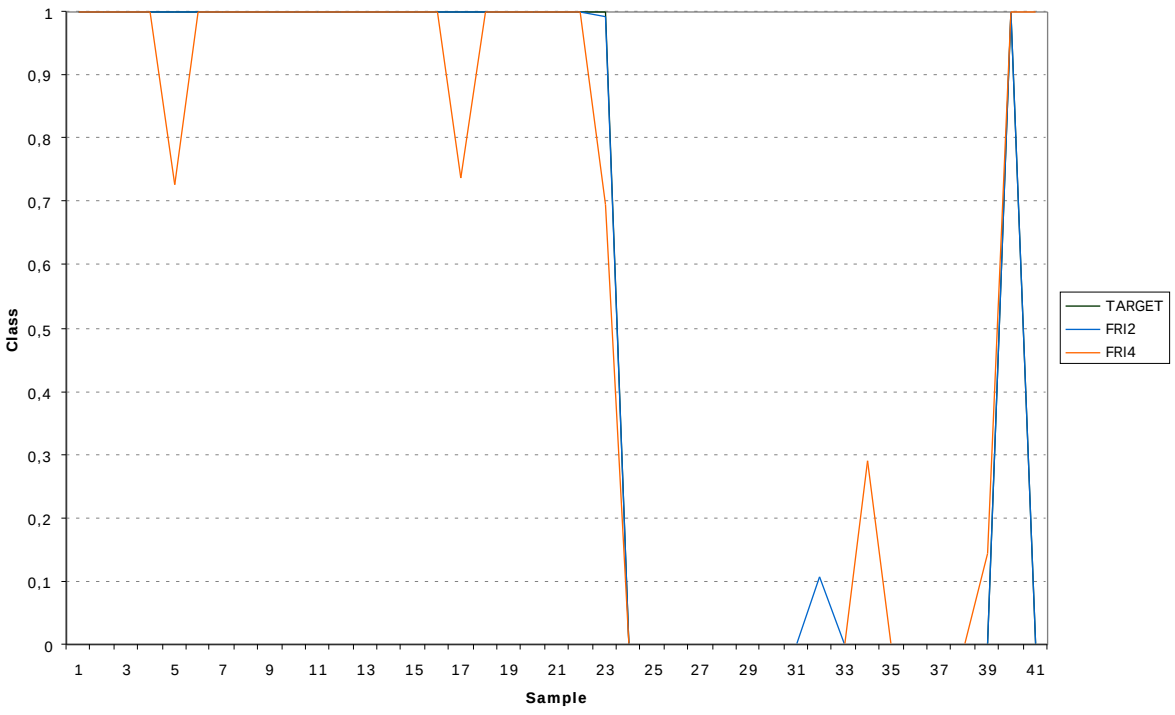

Figure 4: Plot of FRI2 and FRI4 models

Conclusions

It is possible by means of a multileveled self-organized knowledge extraction to generate models that classify to 100% lower resolved mass spectral data. When increasing the spectral resolution, however, it turns out that both classification accuracy and descriptive power of the models decreases sufficiently. Several aspects may be of relevance here:

- Validation. In order to get valid models, the inputs-to-samples ratio is of concern: For linear models and 39 samples, only about 14 variables can be considered in a single model [5] to keep the algorithm's noise filtering work sufficiently well. Maybe, a good classifier of highly resolved spectral data is really a composite of a higher number of descriptors (> 20). In this case, however, more samples are required to get valid models. The same is true for nonlinear models.
- Noise. Highly resolved spectral data may also have a high noise-to-signal ratio or may show some measurements jitter. This would underline the need of a well working noise filtration in data mining. But is the noise-to-signal ratio small enough to extract some relevant information from the data?
- Missed descriptors. It is not impossible that an analysis of the *D1* data set may find some relevant descriptors that our approach B missed during the selection process.

Summarizing, our tests give a first, promising insight into the problem. There is still much potential in this approach. Feedback of medical experts will help to go ahead.

References

[1] Müller, J.-A., Lemke, F.: Self-Organising Data Mining. Extracting Knowledge From Data. BoD Hamburg 2000.

[2] Cherkassky, Mulier: Learning from Data. J. Wiley & Sons. New York 1998.

[3] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. In: Fayyad et al: Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Menlo Park 1996, pp. 83-116

[4] Lemke, F., Müller, J.-A.: Self-Organising Data Mining. Conference paper Chengdu, China, 2002

[5] Lemke, F., Müller, J.-A.: Validation in Self-organising Data Mining. ICIM 2002, Lvov, Ukraine.

[6] Madala, H.R., Ivakhnenko, A.G.: Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, Boca Raton 1994

[7] http://www.knowledgeminer.net