Does My Model Reflect a Valid Relationship?

Frank Lemke

KnowledgeMiner Software

http://www.knowledgeminer.eu

frank@knowledgeminer.eu

Abstract

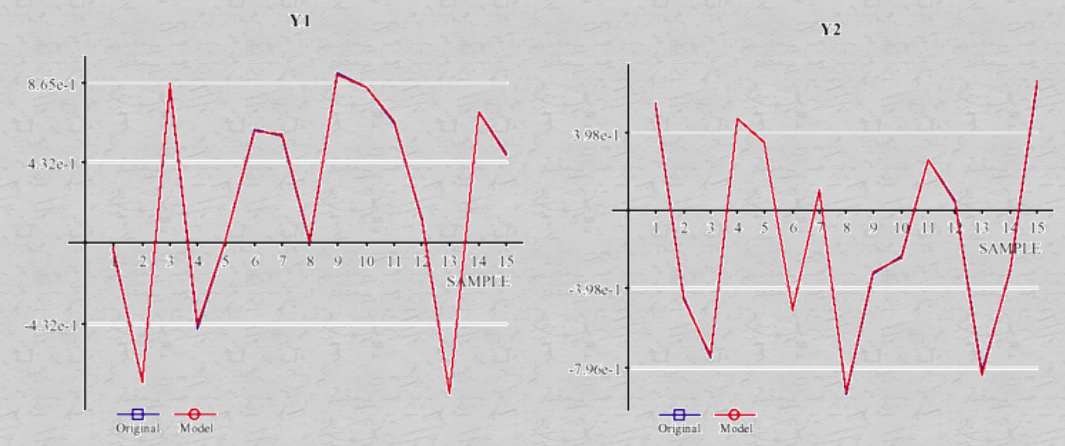
One original and powerful feature implemented in **Insights** (formerly KnowledgeMiner) is additional external evaluation of self-organized linear and nonlinear analytic models. This document is about to show how this new model evaluation approach actively supports answering the above question. Also, a new model quality measure that takes into consideration the noise filtering power of the modeling algorithm and model complexity is introduced: **Descriptive Power**.

The Problem

A key problem in data mining is final evaluation of developed models. This evaluation process is an important condition for deployment of data mining models. By learning from a finite set of data, only, it is hardly possible to decide whether the developed model reflects a valid relationship between input and output or if it's just a stochastic model with non-causal correlations. Model evaluation needs, in addition to a properly working noise filtering procedure for avoiding overfitting the learning data, some new external information to justify a model's quality, i.e., both its predictive and descriptive power.

Why

Let's have a look at this example: Based on an artificial data set of 2 outputs, 4 inputs, and 15 samples KnowledgeMiner **Insights** self-organizes an analytical model for each output variable, Y1 and Y2 (fig.1).



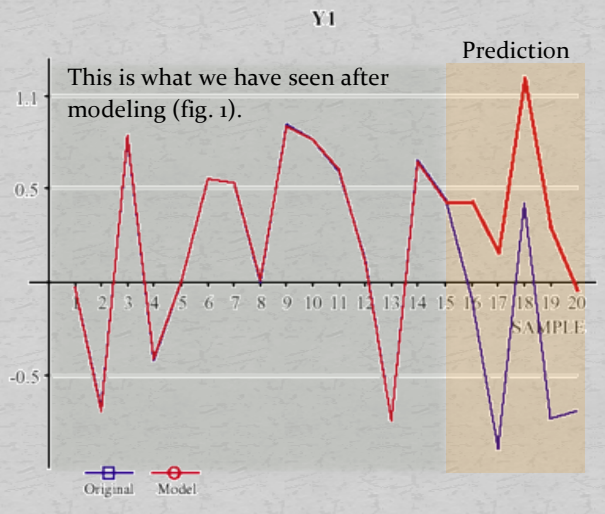
a) Model 1: $Y_1=f_1(\underline{\mathbf{x}})$ *b) Model 2: $Y_2=f_2(\underline{\mathbf{x}})$*

Figure 1. *Model (red) vs. actual (blue, overlaid) graph of the two models.*

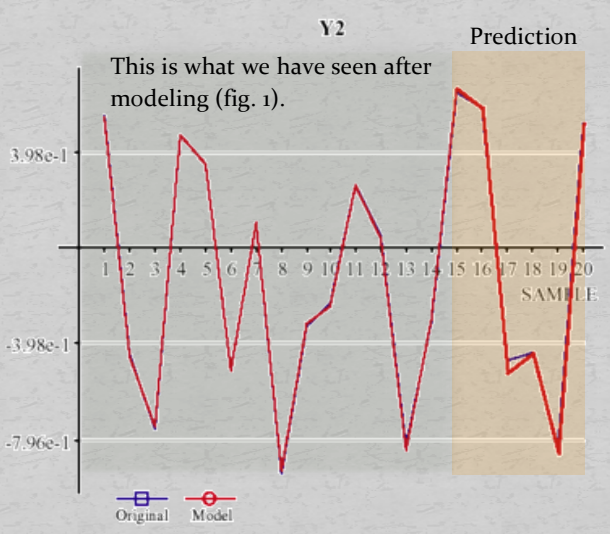
For model 1, a model quality *Q* of 0.9998 (with 1.0 as the best possible and zero as the worst model quality) is reported, while model 2 shows a model quality of 0.9997. Concluding from this model quality and from the graphs in fig. 1 there is no obvious reason to not consider both models as "true" models that reflect a causal relation between input and output. Also, taking into account that **Insights**, compared to the vast majority of data mining tools, is implementing in its inductive self-organizing

model synthesis a powerful noise filtering procedure, already (see also "Self-Organizing Data Mining" book, section 3.2), this seems to underline the above assumption.

However, the person who created the data set for this example states that only one model actually describes a causal relationship while the other model simply reflects some stochastic correlations, because output and inputs are completely independent (random numbers). Even with this information given - which is usually not the case for real-world knowledge and data mining problems - the modeler (you) cannot decide from the available information which of the two models is the true model. Only applying (predicting) the models on some new data - which adds new information - will turn out the true model (fig. 2):



a) Model 1: invalid



b) Model 2: valid

Figure 2. *Prediction of samples 16 to 20 by the two models for Y1 and Y2.*

This example clearly shows that any "closeness-of-fit" measure is not sufficient to evaluate a model's predictive and descriptive power. Recent research has shown that model evaluation requires a two-stage validation approach (at least):

First Level

Noise filtering to avoid overfitting the learning data based on external information (hypothesis testing) not used for creating a model candidate (hypothesis) as an integrated part of the "Model Learning" process. A corresponding tool that is used in **Insights** from the beginning within "Model Learning" is leave-one-out cross-validation.

Second Level

A characteristic that describes the noise filtering behavior of the "Model Learning" process to justify model quality based on external information not used in the first validation level. This noise-filtering characteristic is implemented in KnowledgeMiner **Insights** for the first time for linear and nonlinear analytical models. This characteristic was obtained by running Monte Carlo simulations many times. In this way, new and independent external knowledge is available that *any* model has to be adjusted with.

Figure 3 shows a detail of the characteristic for linear analytical models.

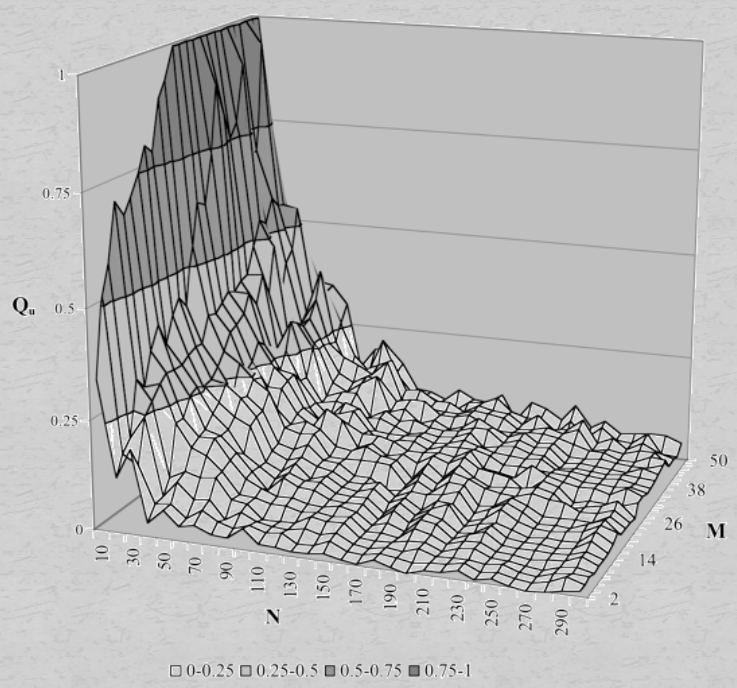


Figure 3. *Noise filtering characteristic*

M: number of inputs; N: number of samples; Q_u: virtual quality of a model
Q_u = 1: noise filtering does not work at all; Q_u = 0: ideal filtering

The reason for a second level validation is (1) that noise filtering implemented in level 1 is not an ideal noise filter and thus is not working properly in every case (see this example) and (2) to get a new model quality measure that is adjusted by the noise filtering power of the modeling algorithm.

This noise-filtering characteristic expresses a virtual model quality Q_u that can be obtained when using a data set of M potential inputs of N random samples. It is virtual model quality, because, by definition, there is not any causal relationship between stochastic variables (true model quality $Q = 0$), but there are actually and usually models of quality $Q > 0$ (see example above), which just reflect stochastic correlations. In result, given any number of potential inputs M and number of samples N, a threshold quality $Q_u = f(N, M)$ is calculated by **Insights** that *any* model of quality Q must exceed to be considered valid with respect to describing a relevant relationship between input and output. Otherwise, a model of quality $Q \leq Q_u$ is assumed invalid, since its quality Q can also be obtained when simply using random variables, which means that this certain model's quality does not significantly differ from a chance model. It has to be considered unreliable and invalid.

In addition to deciding if a model appears being valid or not, the noise filtering characteristic is also a tool for quantifying to which extent the data is described by a relevant relationship between input and output. This introduces a new, noise filtering and model complexity adjusted model quality measure: **Descriptive Power (DP)**, which is defined as:

$$DP = \begin{cases} 0 & Q \leq Q_u(N,L) \\ \dfrac{Q - Q_u(N,L)}{1 - Q_u(N,L)} & Q > Q_u(N,L), Q_u(N,L) < 1 \end{cases}$$

with Q as the measured quality of the evaluated model and $Q_u(N, L)$ as the reference quality calculated from the number of samples N the model was built on and from the number of input variables L the model is actually composed of (i.e., selected relevant inputs), with $L \leq M$. This means that Descriptive Power is a chance-correlation-adjusted quality measure, which is independent from the data set dimension used to develop the model. For example, two models M_1 and M_2 show the same quality $Q = Q_1 = Q_2$, but M_1 uses more inputs than M_2 to get that quality Q. So, with $L_1 > L_2$, the Descriptive Power of M_2 is higher than that of M_1.

The bottom line

KnowledgeMiner **Insights** evaluates a developed model by calculating its Descriptive Power after modeling on the fly. You don't have to care about it. **Insights** will provide all information in the model report to make you more effective and successful in your knowledge mining efforts.

Back to our example above, **Insights** shows this evaluation information after modeling in the report for the two models (fig. 4):

MODEL EVALUATION: INVALID

The requested noise immunity could not be applied for the chosen sample length. Instead, VERY POOR noise immunity was used for modeling, only. To get the requested noise immunity, increase the number of samples to at least 21.

The model seems not reflecting a valid relationship. The likelihood that the data used for modeling is actually random data with no existing input-output relationship is 33%.

Keep in mind, however, that the model was built using VERY POOR noise immunity. This makes evaluation of the model more uncertain.

a) Report of Model 1 --> status: invalid

MODEL EVALUATION: VALID

The requested noise immunity could not be applied for the chosen sample length. Instead, VERY POOR noise immunity was used for modeling, only. To get the requested noise immunity, increase the number of samples to at least 21.

The model seems to establish a valid relationship. The Descriptive Power of the model relative to a chance model is 42% for the actually used noise immunity.

b) Report of Model 2 --> status: valid

Figure 4. *Reported evaluation results of the two models*

This means, the modeler (you) knows instantly that model 2 does well indeed with a Descriptive Power of 42% while model 1 is seen invalid to 33%. Following the recommendation given in the report of model 1, increasing the number of samples to 21, in a second modeling run **Insights** now comes up with this report (fig. 5):

MODEL EVALUATION: INVALID

The model seems not reflecting a valid relationship. The likelihood that the data used for modeling is actually random data with no existing input-output relationship is 67%.

The model was generated by self-organizing high-dimensional modeling.

Figure 5. *Evaluation result of model 1 after remodeling --> status: invalid with increased likelihood of chance model.*

Insights now reports an increased certainty of 67% that this model is just a chance model and therefore has to be rejected. Interesting to note is also that this tiny modeling problem has been identified as high-dimensional modeling task, which sounds strange, first. However, "high-dimensional" has to be seen not only in absolute but also in relative terms: *every* modeling problem with a high number of inputs-to-samples ratio is a high-dimensional modeling task, actually, with respect to model building and validation and has to be handled as such.

The two-stage model validation approach implemented in KnowledgeMiner **Insights** allows for the first time in a data mining software to get active decision support in model evaluation for minimizing the risk of false interpreting a model's quality and power and using invalid models for prediction and classification tasks that in fact just reflect a chance correlation. In combination with our original Live Prediction Validation technology, it gives you the highest degree of reliability about your data mining models you get from software available on the market today.