Paper published on eSEAM'97 Conference 1997

Knowledge Extraction from Data Using Self-Organizing Modeling Technologies

Frank Lemke

CONTENTS:

Abstract

Today, knowledge extraction from data (also referred to as Data Mining) plays an increasing role in sifting important information from existing data. Commonly, regression-based methods like statistics or Artificial Neural Networks as well as rule-based techniques like fuzzy logic and genetic algorithms are used.

This paper describes two methods working on the cybernetic principles of self-organization: Group Method of Data Handling (GMDH) and Analog Complexing. GMDH combines the best of both statistics and Neural Networks and creates adaptively models from data in the form of networks of optimized transfer functions (Active Neurons) in an evolutionary fashion of repetitive generation of populations of alternative models of growing complexity and corresponding model validation and survival-of-the-fittest selection until an optimally complex model has been created. Nonparametric models obtained by Analog Complexing are selected from a given variables set representing one or more patterns of a trajectory of past behavior which are analogous to a chosen reference pattern.

Both approaches have been developed for complex systems modeling, prediction, identification and approximation of multivariate processes, diagnostics, pattern recognition and clusterization of data samples and they are implemented in the KnowledgeMiner modeling software tool. They can be applied to problems in economy (macro economy, marketing, finance e.g.), ecology (water and air pollution problems e.g.), social sciences, medicine (diagnosis and classification problems) and other fields.

-

1. Introduction

Today, knowledge extraction from data is the key to success in many fields. Information technologies deliver a flood of data to decision makers (who doesn't make decisions?) and the question today is how to

get the most from your data without suffering information overload. Knowledge extraction techniques and tools can assist humans in analyzing the mountains of data and to turn information contained in the data into successful decision making.

Experience gained from expert systems, statistics, Neural Networks or other modeling methods has shown that there is a need to try to limit the involvement of modelers (users) in the overall knowledge extraction process to the inclusion of existing a priori knowledge, exclusively, while making the process more automated and more objective. Additionally, most user's interest is in results in their field and they may not have time for learning advanced mathematical, cybernetic and statistical techniques and/or for useing dialog driven modeling tools.

Self-organizing modeling is based on these demands and is a powerful way to generate models from ill-defined problems.

In a wider sense, the spectrum of self-organizing modeling contains regression-based methods, rule-based methods, symbolic modeling and nonparametric model selection methods.

a. regression-based methods

Commonly, statistically-based principles are used to select parametric models. Besides sophisticated methods of mathematical statistics there has been much publicity about the ability of Artificial Neural Networks to learn and to generalize. Sarle [Sarle, 1995] has shown that models commonly obtained by Neural Networks are overfitted multivariate multiple nonlinear (specifically linear) regression functions.

A second regression-based method for model self-organization is the Group Method of Data Handling (GMDH). GMDH combines the best of both statistics and Neural Networks while considering a very important additional principle: that of induction. This cybernetic principle enables GMDH to perform not only advanced model parameter estimation but, more important, to perform an automatic model structure synthesis and model validation, too. GMDH creates adaptively models from data in form of networks of optimized transfer functions (Active Neurons) in an evolutionary fashion of repetitive generation of populations (layers or generations) of alternative models of growing complexity and corresponding model validation and survival-of-the-fittest selection until an optimal complex model - not too simple and not too complex (overfitted) - has been created. Neither, the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron (transfer function of Active Neuron) are predefined. All this is adjusted during the process of self-organization by the process itself. As a result, an explicit analytical model representing relevant relationships between input and output variables is available immediately after modeling. This model contains the extracted knowledge applicable for interpretation, prediction, classification or diagnosis problems.

b. rule-based models in the form of binary or fuzzy logic

Rule induction from data uses genetic algorithms where the representation of models is in the familiar disjunctive normal form. A self-organizing fuzzy modeling may come to be more important for ill-defined problems using the mentioned GMDH algorithm.

c. symbolic modeling

Self-organizing structured modeling uses a symbolic generation of an appropriate model structure (algebraic formula or complex process models) and optimization or identification of a related set of

parameters by means of genetic algorithms. This approach assumes that the elementary components are predefined (model base) and suitably genetically coded.

d. nonparametric models

Known nonparametric model selection methods are: Analog Complexing and Objective Cluster Analysis.

Analog Complexing selects nonparametric prediction models from a given data set representing one or more patterns of a trajectory of past behavior which are analogous t o a chosen reference pattern. Analog Complexing is based on the assumption that there exist typical situations, i.e. each actual period of state evolution of a given multidimensional time process may have one or more analogues in history. If so, it will be likely that a prediction could be obtained by

transforming the known continuations of the historical analogues. It is essential that searching for analogous pattern is not only processed on a single state variable (time series) but on a set of representat ive variables simultaneously and objectively.

The purpose of Objective Cluster Analysis algorithm is to automatically subdivide a given data set optimally into groups of data with similar characteristics (classification). The optimal number of clusters, their width and their composition are selected automatically by the algorithm. It is described in more detail in [Madala/ Ivakhnenko, 1994].

This paper describes the GMDH and the Analog Complexing method and a software tool which has implemented these technologies for modeling and prediction of real -world problems.

2. Self-Organizing Modeling Technologies

2.1. GMDH-type Neural Networks

The traditional GMDH-algorithm was developed by A.G. Ivakhnenko in 1967 [Mada la/ Ivakhnenko, 1994]. Like Neural Networks the GMDH approach is based on

- the black-box method as a principle approach to analyze systems on the basis of input -output data samples and
- the connectionism as a representation of complex functions through net works of elementary functions.

Seperate lines of development starting from these scientific foundations are the theory of Neural Networks (NN) and the theory of Statistical Learning Networks (SLN). GMDH as the most important representative of SLN's was designed from a more cybernetical perspective and has implemented a stronger behavioristic power than NN's due to consideration of a third, unique priniciple: induction. This principle consists of:

– the cybernetic principle of self-organization as an adaptive creation of a network without subjective points given;

– the principle of external complement enabling an objective selection of a model of optimal complexity and

– the principle of regularization of ill-posed tasks.

To realize a self-organization of models on the basis of a finite number of input-output data samples the following conditions must exist to be fulfilled:

- a very simple initial organization (neuron) which enables through its evolution the description of a large class of systems (including the focussed object);
- a selection criterion for validation and measure of the usefulness of an organization relative to its intended task and
- a algorithm for mutation of the initial or already evolved organization of a population (network layer) (fig.1).

Fig. 1: Network at begin of modeling

In conjunction with an inductive generation of many variables a network is a function represented by a composition of many transfer functions. Typically , they are nonlinear functions of a few variables o r linear functions of many variables. Commonly, linear or second-order polynomial functions in two or three variables are used like

$$f(xi, xj) = a0 + a1xi + a2xj + a3xixj + a4xixi + a5xjxj.$$

To estimate the unknown parameters ai statistical techniques are u sed in distinction to the heuristical, global search algorithms of Neural Networks. More advanced GMDH algorithms perform a self-organization of the structure of the transfer function too as described in 3.1.a (optimized transfer function obtained by Active Neurons).

Components of input vector xi, i=1,2, ..., n may be independent variables, functional terms or finite difference terms. Lorentz and Kolmogorov have shown that any continuous function f(x1, ..., xd) on [0, 1]d can be exactly represented as a composition of sums and continuous one-dimensional functions. There exist continuous one-dimensional functions gj and hjk for j=1, 2, ..., 2d+1 and k=1, 2, ..., d so that

$$f(x1, ..., xd) = SUM\ gj\ SUM\ hjk(xk).$$

An important feature of GMDH is the use of an external selection criterion: the parameters of each neuron transfer function are estimated on a training set of observations to minimize the fit of the actual processed model to the desired final behavior (output variable) while a separate testing set is used to rank and select the best models of each generation (network layer). This approach, embedded in the overall modeling process, guarantees the objectivity of the knowledge extraction process and the avoida nce of overfitting.

The mutation process works basically as follows:

All possible pairs of the m inputs are generated to create and validate the transfer functions of the m*(m - 1)/2 neurons of the first network layer (fig.2). Then, a number of best models - consisting of a single neuron only in the first layer - are ranked and selected by the external selection criterion (fig.3). The selected intermediate models survive and they are used

Fig.2: Network after creation of all models of the 1st layer Fig.3: Net work after selection of best models

as inputs for the next layer to create a new generation of models while the nonselected models die (fig.4). This procedure of inheritance, mutation and selection stops automatically if a new generation of models

provides no further improvement. Then, a final, optimal complex model is obtained. In distinction to Neural Networks, the complete GMDH Network is, during modeling, a superposition of a large number of alternative networks living simultaniously. The final, optimal complex model represents a single network only while all others die after modeling (fig.5).

Fig. 4: after creation of all models of the 2nd layer

Fig. 5: Network after selection of an optimal model y*

The GMDH algorithm is based on adaptive network creation. In this approach the objective is to estimate networks of the right size (number of neurons and their transfer functions as well as number of layers) with a structure evolved during modeling process. The basic idea is, that once the neurons on a lower level are computed, the neurons of the next level may be computed on the lower level ones. One goal was to design an efficient learning algorithm which is regarded as a search procedure that correctly solves the modeling problem. Self-organization is considered while building the connections between the units through a learning mechanism to repeat discrete items. A QuickTime Animation is exemplary of the modeling process.

2.2. Analog Complexing

Almost all objects of recognition and control in economics, ecology, biology and medicine are undeterministic or fuzzy. They can be represented by deterministic (robust) parts and additional black boxes acting on each output of the object. The only information about these boxes is that they have limited values of output variables which are similar to the corresponding states of the object. According to Ashby and Beer the diversity of control systems is to not be smaller than

diversity of the object itself. This equal fuzziness of model and object is reached automatically if the object itself is used for forecasting. Forecasts are not calculated in the classical sense but selected from the table of observational data. This method is denoted as nonparametric, because there is no need to estimate parameters.

The main assumptions are:

- the system to be modeled is described by a multidimensional process;
- many observations of data sample are available (long time series);
- the multidimensional process is sufficiently representative, i.e., the essential system variables are included in the observations;
- it is possible that a past behavior might repeat in future.

If we succeed in finding one or more parts of the past (analogous pattern) which are analogous to the most recent part of behavior trajectory (reference pattern), the prediction can be achieved by applying the known continuation of these analogous pattern to the reference pattern. However, this relation, in this absolute form, is only true for non evolutionary processes.

Within the last years this method has been enhanced by an inductive, self-organizing approach and an advanced selection procedure to make it applicable to evolutionary processes, also. In conjunction with this

goal the question arises, whether it is permissible to apply such an approach in the fiel d of evolutionary processes. Besides this philosophical side of the question there has been a formal problem. For evolutionary processes stationarity as one important condition of this methodology is not fulfilled.

If it is possible to estimate the unknown trend (and perhaps seasonal effects) the difference between the process and its trend can be used for Analog Complexing. However, the trend is an unknown function of time and the subjective selection of an appropriate function is a difficult problem. A so lution to select the trend in a more objective way provides the GMDH algorithm through its extraction and transformation capabilities. In any case, however, the results of Analog Complexing would depend on the selected trend function.

Therefore, it was advisable to go a more powerful way which consists of 4 steps and which is described in more detail in [Lemke/Mueller, 1997]:

1. Generation of alternative patterns
2. Transformation of analogues
3. Selection of the most similar analogues
4. Combining forecasts.

A sliding window generates the set of possible patterns {Pi,k+1}, where Pi,k+1 = (xi, xi+1, xi+2, ... , xi+k) and k+1 is the length of the sliding window as well as the length of the patterns. The reference pattern is PRk = PN-k,k+1. For the given reference pattern PRk it is necessary to select the most similar patterns Pi,k+1, ieJ. Then, the continuations of these patterns are used, to compute the prediction of the reference pattern. Figure 6 illustrates this approach.

Fig.6: Selected analogous patterns Pk relative to a reference Pattern PR shown for a one-dimensional process

3. KnowledgeMiner - A Self-Organizing Modeling Software Tool

KnowledgeMiner is a powerful and easy-to-use modeling tool which was designed to support the knowledge extraction process on a highly automated level and which has implemented two advanced self-organizing modeling technologies:

1. GMDH

2. Analog Complexing.

3.1. GMDH Implementation

KnowledgeMiner has implemented 3 different GMDH-type self-organizing modeling algorithms to make knowledge extraction systematical, fast, successful and easy-to-use even for large and complex systems:

a. Active Neurons

KnowledgeMiner performs self-organization of an optimal complex transfer function of each created neuron (Active Neuron). Beginning at the simplest possible transfer function, f(xi, xj) = a0 , an optimal complex neuron is evolved by repetitive creation, validation and selection of different transfer function

representations. One important feature of Active Neurons is that they are able to select significant input variables themselves. As a result the synthesized network is a composition of different, a priori unknown neurons and their corresponding transfer function has been selected fro m all possible linear or nonlinear second-order polynomials. The algorithm ensures that essential independent variables will be selected at the lowest possible level already and supports in this way significantely that optimal complex networks will be created.

b. Network Synthesis (multi-input/single-output model)

Secondly, an algorithm for self-organization of multilayered networks of Active Neurons is implemented. It performs the creation of an optimal complex network structure (optimal number of neurons and number of layers) including cross-validation and selection of a number of best model candidates out of populations of competetive models. The algorithm ensures, for example, that even if creation of nonlinear models was chosen as permissable it really could be possible that a linear model only will be selected as optimal, finally. The implemented selection criterion subdivides the data set internally into training and testing data sets dynamically. This means, the user doesn't need to process data subdi vision in any way; the cross-validation criterion uses virtually the complete data set for training as well as for testing synthesized models. The result of the modeling process is an easily accessible and visible analytical model (model graph, model equation,model data output). All created models are stored in a model base and are immediately applicable for analysis and short- to long-term status-quo or what-if predictions.

c. Systems of Equations (multi-input/multi-output model)

One important feature of KnowledgeMiner is self-organization of an optimal, autonomous system of equations. This system has to be free of mathematical conflicts and can be viewed as a network of interconnected GMDH networks (3.1.b) which is visible through a system graph and applic able for long-term status-quo prediction of the whole system. It provides the only way to predict a set of input -output models autonomously, objectively and whithout additional efforts.

-

3.2. Analog Complexing Implementation

KnowledgeMiner provides an Analog Complexing algorithm for prediction of the most fuzzy processes like financial or other markets. It is a multi-dimensional search engine to select most similar, past system states relative to a chosen (actual) reference state. This means, searching for analogous patterns in the data set is usually not only processed on a single time series (column) but on a specified, representative set of time series simultaniously to extract significant hidden knowledge. Additionally, it is possible to let the algorithm search for different pattern lenght (number of rows a pattern consists of) whithin one modeling process. All selected patterns, either of the same or different pattern length, are then combined to synthesize a most likely prediction. KnowledgeMiner performs this in an objective way using a GMDH algorithm to find out the optimal number of patterns and their composition to obtain a best result. The reason for synthesizing models/predictions is that each model or pattern reflect only a specific behavior of reality. By combining several models it is more likely to reflect reality in a more complete and robust fashion.

-

4. Example Applications

The application field of self-organizing modeling is decision support in economy (analysis and prediction of economical systems, market, sales and financial predictions, balance sheet prediction), ecology (analysis and prediction of ecological processes like air and soil temperature, air and water pollution, growth of wheat, drainage flow, Cl- and NO3 -settlement, influence of natural position factors on harvest) and in all other fields with only small a priori knowledge about the system. The advantages of self-organizing modeling over the Neural Network approach as well as over statistics is that it works very fast, systematically and objectively since only minimal a priori information (pre-definitions) is required, it provides explicit models as an explanation component while making hidden knowledge visible and usable and the obtained results are in average as good as or better than results of other modeling techniques. Some times self-organizing modeling is the only way to get results for a problem at all.

The following examples are included in the downloadable KnowledgeMiner package.

-

4.1. National Economy

This example shows the prediction of 13 important characteristics of a national economy 3 years ahead along with their corresponding models.

Given are 27 yearly observations (1960 - 1986) of 13 variables like Gross Domestic Product, Unemployed Persons, Savings, Cash Circulation, Personal Consumption, State Consumption. Using this data set (13 columns, 27 rows) and a chosen system dynamic of up to 3 years, for each variable the invisible and normalized information basis for self-organization of a linear system of equations is constructed automatically. This means, for instance, that for x1 a model will be created out of this information basis:

x1,t = f(x2,t, x3,t, ... , x13,t, x1,t-1, x2,t-1, ... , x13,t-1, x1,t-2, ... , x13,t-3).

The information basis has this dimension: 51 input variables (columns) and 24 observations (rows). The task of the modeling process is now to evolve a specific instance of the function f considering a number of requirements to end up in a robust, optimal complex model. This is done not only for x1 but for all 13 variables automatically while avoiding conflicts between the unlagged variables. The result is an optimal, autonomous system of equations that can predict all 13 variables whithin one process and which is visible through a system graph.

-

4.2. Balance Sheet

Balance sheet analysis and prediction is an important part of the fundamental analysis in finance. Reliable information about status and evolution of a company is a key factor for success in that area. A lot of this information is contained in the data of balance sheet characteristics and the overall market and macroeconomic data (see 4.1.). A problem is the large amount of variables, the very small number

of observations for balance sheets and the unknown relationships and dynamics between these variables. Here, statistics as well as Neural Networks are practically not applicable. GMDH is.

In the balance sheet example shown here only balance sheet characteristics themselves are used to predict them 1 year ahead. Given are 13 characteristics for 7 years (1986 - 1992). Again, a dynamic, linear system

of equations was self-organized as described in 4.1. The average percentage error of the 1993 prediction was 16%.

-

4.3. COD concentration

This example describes a water pollution problem [Farlow, 1984]. COD stands for Chemical Oxygen Demand and is used as a proxy to measure water pollution. One difficulty here is that only a few characteristics like water temperature, salt concentration or transparency are measureable and that these measurements are very noisy. Many attempts were made to develop a mathematical model to predict COD levels for control of water pollution in compliance with the standards. Usually, the behavior of COD in a bay is calculated by the nonreaction-diffusion model. However, this model has some significant defects which are the reason for seeking alternative methods. One way for predicting COD provides GMDH. Using 6 characteristics with 40 monthly observations it is possible by creating a system of equations to predict COD 5 month ahead with satisfactory results. The obtained system graph is shown exemplary in fig.7.

Fig.7: System of equations obtained for the COD prediction problem

-

4.4. Flats (Best Buy Apartments)

Another kind of problem presents this example. It reflects a market analysis task searching for friendly or costly flat rates out of a given number of comparable objects. Given are 6 characteristics as input variables (location, type, requested equipment, extra equipment, number of rooms and m2/room) which are obtained by matrices of several, subjectively ranked subcriteria and which are expected to have influence on the variable of interest, rate/m2, in some way. In this case static linear and nonlinear models were created using the characteristics of 30 flats. Since the obtained models reflect different relationships they were combined to have a more robust decision basis. See the live example to get an idea on how this works.

You can find more examples in the demo package:

- Bread - a product order and delivery prediction problem;
- Computer - another market analysis example;
- Stock indexes - prediction of Dow Jones and S&P500 at the NYSE using daily close prices;
- XOR problem - modeling the XOR operator.

-

5. Conclusions

Self-organizing modeling technologies are a powerful approach to extract hidden knowledge from data serving for decision support of real-world problems. They are often an alternative choice to statistics, Neural Networks or NeuroFuzzy methods since they create optimal complex models automatically, fast and systematically and they provide an explanation component through explicit visible model descriptions. KnowledgeMiner is an advanced, easy-to-use self-organizing knowledge extraction and prediction tool

which can bring its capabilities to the desktops without the need of being an expert in modeling.

Actually, directions of further research and development for improvement of self-organizing approach are:

- development of an algorithm for automatic choice of an appropriate task related modeling method,
- further perfection of Analog Complexing,
- development of self-organizing fuzzy modeling,
- implementation of new (perhaps fuzzy) combining/ synthesis methods.

References

Farlow, S.J. (ed.) (1984): Self-organizing Methods in Modeling. GMDH

Type Algorithms. Marcel Dekker. New York, Basel

Lemke, F.; Mueller, J.A. (1997): Self-Organizing Data Mining for a Portfolio Trading System. Journal for Computational Intelligence in Finance, 5(1997)3

Madala, H.R.; Ivakhnenko, A.G. (1994): Inductive Learning Algorithms for Complex Systems Modelling. CRC Press Inc., Boca Raton, Ann Arbor,

London, Tokyo

Sarle, W.S. (1995): Neural Networks and Statistical Models. in: Proceedings of 19th Annual SAS User Group International Conference. Dallas. pp. 1538-1549