Abstract. Developed in 1968 by A.G. Ivakhnenko, the self-organising modelling approach introduces the idea of external information into modelling and a class of algorithms which inductively, systematically and autonomously evolve an optimal complex model from noisy observation data by employing parameter and structure identification. Later, many important analytical results were obtained forming the theory of noise immunity modelling. It was shown that inductively self-organised “non-physical” models of optimal complexity that match the unknown noise-to-signal ratio are usually more accurate for interpolation and forecasting purposes than “physical” models obtained deductively. Recent research and development resulted in Networks of Active Neurons, parallel implementations of diverse algorithms, and in multi-level self-organisation for modelling high-dimensional data sets. Current efforts are targeting at self-organising modelling for decision support of complex real-world problems. We think that integration of soft systems approaches, learning from data, and inductive modelling paradigms will lead to a new framework of inductive modelling for modelling, predicting and controlling complex systems and processes in near real-time.

Keywords

1 On the History of GMDH-based Self-organizing Modelling Technologies

The idea of developing models from noisy data in a self-organizing way has a long history. In 1968 the Ukrainian scientist A.G. Ivakhnenko based on statistical learning, perceptron, and pattern recognition concepts developed the first version of the Group Method of Data Handling (GMDH) algorithm as the core of self-organizing modelling [1, 2, 3]. He introduced the idea of external information into modelling by subdividing a dataset into training and testing data sets for model evaluation, and he used linear or second-order polynomial functions in two or three variables as elementary neuron transfer functions which inductively, systematically, and autonomously evolve to an optimal complex model by employing parameter and structure identification. An optimal complex model is a model that optimally balances model fit on a given learning data set and its generalization power on new, not previously seen data with respect to the data's noise level and the task of modelling (prediction, classification, identification, etc.). It thus solves the basic problem of experimental systems analysis of systematically avoiding over-fitted models based on the data's information only. Additionally, the self-organised model is available analytically in an explicit form of algebraic or difference equations.

From the beginning the GMDH approach was a computer-based method so that a set of computer programs and algorithms were the primary practical results of research and new theoretical findings. Thanks to the author's policy of open source code sharing the method was quickly adopted by a number of scientific institutes worldwide. In result, considerable enhancements were introduced in the 1970s and 1980s by versions of the Polynomial Network Training algorithm by Barron and the Algorithm for Synthesis of Polynomial Networks by Elder when concepts of Adaptive Learning Networks and GMDH were flowing together [4, 5].

In the past more than 40 years of its history GMDH-based self-organizing modelling technologies have been improving continuously. So the problem of modelling noisy data under incomplete information has been solved. Multi-criteria model selection and utilization of well-known a priori information was proposed. Experiments have shown that in this way models can be reliably self-organised from data sets that show a noise-to-signal ratio of up to 10. In the 1980s many important analytical results were obtained forming the theory of noise immunity modelling [6, 7, 8, 9, 10]. It was shown, for example, that “physical models” couldn’t be used for long-term forecasting on noisy data. It was proven that self-organised “non-physical” models of optimal complexity that match the unknown noise-to-signal ratio are usually more accurate for interpolation and forecasting purposes than physical models whose parameters are...
estimated by regression analysis [8, 9, 11]. Non-parametric modelling algorithms for describing fuzzy objects have been developed.

In the 1990s an original and efficient concept of neural networks with active neurons was introduced [12]. The core idea here is in self-organising not only the network structure but, simultaneously, the individual optimal transfer function of each elementary neuron. This twice self-organised network of active neurons also results in various network topologies with heterogeneous neurons. They are based on competing types of transfer functions like those used in ANNs by choosing the optimal activation function for every neuron [13, 14]. First results have shown that this is a very promising approach to developing flexible hybrid networks with automatic evolution of self-organising networks.

Recent research and development resulted in parallel implementations of diverse algorithms [15, 16], inclusion of Genetic Algorithms and other optimization techniques [17], multi-level self-organisation for modelling high-dimensional data sets of many thousand input variables and for modelling interdependent systems of equations (multi-input/multi-output systems), cost-sensitive modelling, and new model evaluation techniques to improve reliability and applicability of models to the user [18].

Today, self-organising modelling is a proven and highly efficient knowledge extraction technology and there exist different implementations in open source, free, or commercial software that has been applied in various fields from image recognition over biomarker detection and QSAR modelling, wastewater management and reuse questions to Global Warming and micro and macro-economic forecasting problems.

2 The Self-organising Modelling Approach

A.G. Ivakhnenko described the traditional inductive approach of self-organising modelling 45 years ago [1]. This initial GMDH method has been further developed since then. However, all GMDH variants share one principle that makes it different from other well known modelling and data mining methods: that of induction. The concept of induction is composed of three ideas:

- The principle of self-organisation for adaptively evolving a network model without subjective points given;
- The principle of external information to allow objective selection of a model of optimal complexity, and
- The principle of regularization of ill-posed tasks.

The GMDH algorithm is based on adaptive networks. Self-organisation is considered in identifying connections between the network units by a learning mechanism to represent discrete items. For this approach, the objective is to estimate networks of relevant and sufficient size with a structure evolving during the estimation process. A process is said to undergo self-organisation if identification emerges through the system's environment.

To realise a self-organisation of models from a finite number of input-output data samples the following conditions must exist to be fulfilled:

First condition: There is a very simple initial organisation that enables the description of a large class of systems through the organisation's evolution.

A common class often used is that of dynamic systems, which can be described by Volterra functional series. Discrete analogues of the Volterra functional series describing systems with a finite memory are higher-order polynomials of the Kolmogorov-Gabor form.

Second condition: There is an algorithm for mutation of the initial or already evolved organisations of a population.

Genetic Algorithms are working on more or less stochastic mutations of the model structure by means of crossover, stochastic changes of characteristic parameters, and others. In the GMDH approach, a gradual increase of model complexity is used as a basic principle. The successive combination of many variants of mathematical models with increasing complexity has proven to be a universal solution in the theory of self-organisation presenting mutation in a way like that in biological selection processes. To apply this principle, a system of basic functions (first condition) is needed. Their appropriate choice, and the way the elementary models are combined to more complicated models decide the success of self-organisation. In most self-organising modelling algorithms, pairwise combination of \( M \) inputs is used to develop model candidates of growing complexity.

Third condition: There is a selection criterion for measuring and validating the usefulness of a model relative to the intended task of modelling.
According to this condition, several best model candidates are ranked and selected by an external selection criterion. The selected models survive, and they are in turn used as inputs for the following layer(s) to develop a new, more complex generation of models while the non-selected models die.

The principle of selection is closely linked to the principle of self-organisation in biological evolution; it is very important for the evolution of species. In the case of self-organising modelling, it is applied when the number of all possible model candidates in one generation is going to become too large for a complete induction. Using a threshold value, those model candidates are selected that are best in the sense of a given quality function, and they are stored as inputs for the next generation's model evolution.

The overall procedure of inheritance, mutation and selection stops automatically if a new generation of models provides no further model quality improvement. Then, a final optimal complex analytical model is obtained.

The process of self-organising modelling is summarised in fig. 1.

**Fig. 1.** Self-organisation of a Network of Active Neurons.
3 Modelling and Decision Support

Decision support, whatever the field of human endeavour, requires formulation and a good understanding of what the problem is. To predict what may happen to a system under certain circumstances is often very difficult even for the simplest of systems, especially if they are not man-made. Humans have for centuries been seeking proxies for real processes. A substitute that can generate reliable information about a real system and its behaviour is called a model and they form the basis for any decision.

The world around us is getting more complex, more interdependent, more connected and global. We can observe it, but we cannot understand it because of its complexity and the myriad interactions that are impossible to know let alone foresee. Uncertainty and vagueness, coupled with rapid developments radically affect humanity. Though we observe these effects, we most often do not understand the consequences of any actions, the dynamics involved and the interdependencies of real-world systems in which system variables are dynamically related to many others, and where it is usually difficult to differentiate which are the causes and which are the effects.

There are many cases in practice where it is impossible to create analytical models using classical theoretical systems analysis since there is incomplete knowledge of the processes involved. Environmental, medical and socio-economic systems are but three examples. We are facing complex problems, which do need decision-making, but the means – the models – for understanding, predicting, simulating, and where possible controlling such systems are absent. This is an increasingly common situation in many real-world problems. To fill this increasing gap, new and appropriate inductive self-organising modelling methods have been theoretically and practically developed as powerful tools in revealing the missing implicit relationships within complex systems.

Mathematical modelling is at the core of many decision support systems. However, many problems in economics, ecology, biology, biochemistry, sociology, and life sciences, to name but a few, are ill-defined and can be characterized by:

- Insufficient a priori information about the system for adequately describing the inherent system relationships,
- Possessing a large number of variables, many of which are unknown and/or cannot be measured,
- Noisy data available in small data sets, only,
- Vague and fuzzy objects whose variables have to be described adequately.

Common to all modelling problems this means to:

- Apply a systematic, holistic approach to modelling,
- Take into account the incompleteness and inadequacy of a priori information about the real-world system,
- Describe the vagueness and uncertainties of variables and, consequently, uncertainty of results and
- Handle very small to large sets of noisy data.

For ill-defined systems the classical hard approach that is based on the assumption that the world can be understood objectively and that knowledge about the world can be validated through empirical means needs to be replaced by a soft systems paradigm which can better describe vagueness and imprecision. This approach is based on the observation that humans only have an incomplete and rather vague understanding of the nature of the world but nevertheless are able to solve unexpected problems in uncertain situations.

4 Sample Application: Global Warming Prediction

This sample application, which is still on going, is to confirm the above statements and to demonstrate the advantages of inductive modelling over traditional theory-based approaches in case of ill-defined modelling problems on a well-recognised real-world modelling problem.

Our atmosphere is a complex system where the system variables are interconnected in a not completely known and understood way with unknown dynamics, building a complex relationship pattern where it is hard to tell cause from effect. This missing a priori knowledge is a major problem for deductive climate modelling based on physical principles, which all Global Circulation Models (GCMs) used by the Intergovernmental Panel on Climate Change (IPCC) show, and it grows a lot of assumptions and often non-holistic approaches to model formulation that introduce considerable subjectivity into modelling and into results. On the other hand, there is an increasing pool of observational
data available, and essential information about the complex behaviour of the atmosphere is hidden in this data. This information about the system can be extracted appropriately from the data by inductive modelling to describe and predict the atmospheric system in short-term in an automated and objective way.

Based on monthly observational data (October 1988 – April 2011) of 6 variables of the atmospheric system - ozone concentration, aerosol index, radiative cloud fraction, CO$_2$ concentration, and global mean temperature as endogenous variables and sun activity as exogenous variable of the system – a non-linear dynamic system with lags of up to 120 months has been self-organised from over 700 potential inputs by high-dimensional modelling of the Insights GMDH tool.

The final self-organised system model consists of all provided initial variables except CO$_2$ concentration, i.e., CO$_2$ concentration has not been self-selected by inductive modelling as a relevant system variable [19]. A model ensemble that implicitly describes model uncertainty and that provides a prediction range of low, most likely, and high values, represents each model of the selected 5 system variables.

As of March 2013, the ex-ante prediction accuracy of the most likely prediction (solid red line) of the Insights system model is 65%, and the accuracy relative to the prediction range (pink area) is 98% (fig. 2).

Compared to this model, the GCMs which the IPCC AR4 projections are based on and which simplistically rely on atmospheric CO$_2$ as major climate driver show an ex-ante prediction accuracy of only 21% for the time period 2007 (the year of publication) till today. The tight connection between temperature projection (yellow) and CO$_2$ concentration projection (gray) is clearly visible for the forecast horizon, too, as well as the growing gap between IPCC projection and observed temperatures (fig. 3).

The considerably higher predictive power of the self-organised Insights system model is clearly a result of the unique ability of inductive modelling to autonomously and reliably extract more relevant information from the noisy observational data for modelling the internal workings of this ill-defined climate system than theory-driven modelling approaches can achieve based on incomplete and uncertain human knowledge.
The idea of developing models from noisy data in a self-organising way has a long history. Starting in 1968 by Ivakhnenko, based on statistical learning, perceptron, and pattern recognition concepts, research in Inductive Modelling formulated, defined, and proved a mathematical framework that contains key concepts of modern data mining and machine learning methods such as the idea of external information, the link between model complexity and noise dispersion leading to optimal complex models by parameter and structure identification, the development of a unique noise immunity modelling theory to systematically avoid over-fitted models and to increase model reliability and validity, or self-organisation of active neurons.

Decision support requires formulation and a good understanding of what the problem is. Humans have for centuries been seeking proxies for real processes. A substitute that can generate reliable information about a real system and its behaviour is called a model and they form the basis for any decision.

Systems can be modelled through deductive logical-mathematical methods or by inductive modelling methods. Deductive methods have been used to advantages in cases of well-understood problems and that obey well-known principles. The spectacular results in aerospace are prime examples of this approach. Here, the theory of the object being modelled is well known and obeys known physical laws.

In contrast, inductive methods are used when macroscopic models are the only alternative. These models are derived from real physical data and represent the relationships implicit within the system without knowledge of the physical processes or mechanisms involved.

In the real world there is a vast treasure trove of data that is being continuously amassed that contains useful information about the behaviour of systems. This is priceless information, which only needs to be trawled and suitably mined so as to transform it into useful knowledge. Theory-driven approaches to modelling are unduly restrictive to this end because of insufficient a priori knowledge, complexity and the uncertainty of the objects, as well as the exploding time and computing demands.

Recent research using new inductive learning modelling and knowledge mining technologies based on self-organising concepts shows breakthrough results and leads to an intelligent inductive modelling technology which is...
fundamental and key for the long-term vision. The concepts and results of this effort it is hoped will lead to new lines of research that should provide answers to many open questions like:

- Why a system behaves in a certain manner?
- Which causes lead to the observed effects?
- What will happen if certain causes are subject to changes?
- How are system variables inter-related?
- How can an object be manipulated to obtain desired effect?

It is clear that reliable answers to these questions can only be obtained if the resulting model has sufficient validity, fidelity, transparency, and reproducibility to adequately explain observed phenomena and predict them over a finite future horizon.

We expect that integration of soft systems approaches, learning from data, computational intelligence, and inductive modelling paradigms will lead to a new synergetic framework of new theories, methodologies, algorithms and technologies for modelling, describing, predicting and controlling complex systems and interrelated processes in near-real-time.

References