

Nature-Inspired Methods for Knowledge Generation from Data in Real-Time

Plamen Angelov
Department of Communication Systems
InfoLab21, Lancaster University
Lancaster, LA1 4WA, UK
e-mail: p.angelov@lancaster.ac.uk

Challenge

Over the last two decades, the fast development of information technology has not only led to an enormous increase of the speed of computers, but also to **enormous amount of data** produced. In addition, prices for sensors have dropped and electronic data recording has become a routine in almost every field of industry, economy and science. Today, however, many organizations have more than just very large databases: they have databases that **grow continuously over time, without limit** and possibly **very fast**. Examples of data overload are the Internet, genome information, complex industrial processes etc. [14,16,17]. Such high-volume, non-stationary data streams offers unique opportunities, but also bring new challenges. In particular, data streams cannot be analysed in a batch mode, since storing the complete data is often practically impossible. Instead, systems have to be developed that analyse the data in an **on-line** manner in **real-time**. This have accelerated the interest in the field of *knowledge generation from data streams in real-time* (signals, images, multi-media etc.) [5]. According to a widely accepted definition, *knowledge generation* or *discovery* refers to the non-trivial process of identifying valid and understandable/interpretable *structure* in the data [4].

At the same time the Human Brain is an example of a tool that is able to process complex information and make decisions in real time, and example of cognition and reasoning that inspired creation and fast development of such areas of computational intelligence as fuzzy logic (dealing with approximate reasoning, decision making under uncertainties), neural networks (dealing with classification and prediction). A number of other emerging techniques stem from the Nature (such as swarm intelligence and ant colonies [18], evolutionary algorithms [19], artificial immune systems [20], bio-memetic algorithms [21] etc.) and are candidates for generating efficient new methods for knowledge extraction from data.

Conventional evolutionary algorithms (genetic algorithms [19], genetic programming [22] etc.) mimic the evolutionary processes that take place in **populations of individuals** and use 'crossover', 'mutation', 'selection', and 'recombination' of chromosomes as mechanisms of adaptation. The emerging *evolving* systems [1,23-26] borrow from the **evolution of individuals** as such in nature, especially humans: learning through experience, inheritance, 'gradual' change, knowledge generation from routine operation, rule extraction from the data are the mechanisms and techniques used in emerging *evolving* fuzzy and neuro-fuzzy systems [1, 23-26].

The main challenges that needs to be addressed are:

- a) designing effective and **computationally efficient learning methods**;
- b) **interpretability of the knowledge generated** by such techniques;
- c) problems of **cooperative knowledge generation** between different individuals/systems.

Examples of **application areas** of real-time Knowledge Generation from Data include, e.g., autonomous robotic systems (automatic adaptive target recognition, moving obstacle avoidance, cooperative multi-model real-time classifiers, etc.), bio-informatics (knowledge extraction from genomic data, reverse engineering, proteomics etc.), customers behaviour analysis (marketing and Internet user data, communication, mobile workforce related problems, etc.). In customer relationship management, for example, large amounts of data about customers are collected every day. The analysis of such data can inform about changes in customer behaviour, for instance the conditions (special offers, advertisements, etc.) under which a customer tends to buy certain products. Another important issue concerns the prediction of customer preferences. Such predictions may have an immediate influence on the production, especially for products like cars that offer a great variety of extra equipment and devices. Generating interpretable knowledge that evolves over time by incremental learning from continuously arriving data with adaptive dimensionality reduction, feature selection is very attractive and desirable. Even though most of these desiderata have already been addressed individually, **learning systems satisfying all of them are not yet available**.

Most conventional learning methods, such as support vector machines [11,15], discriminant analysis, decision trees [13], neural networks [6] are designed to work in *batch mode*, that is, by using the complete data that has been observed. For stationary processes this may be enough. For complex non-stationary processes, however, methods and techniques for updating the induced models in an efficient way are needed. In order to avoid starting from scratch every time, these techniques must be able to learn in an *incremental* way, i.e., to adapt the current model by using only the new data, but without referring to the old data (in a recursive manner). Moreover, corresponding methods should be able to react to changes in the underlying function to be learned, either gradual ones (*concept drift*) or even abrupt ones (*concept shift*). Another important issue is the knowledge and data integration [1], that is the adaptation and evolution of the knowledge (model/rule-base) that accommodates the information brought by the new data and reconciles this with the previously existing knowledge.

Even though first steps in this direction have already been made (e.g. [1,2, 3,10]), this research direction is still in its infancy and more focused efforts are needed. In particular, many approaches do simply perform “adaptive tuning”, that is, they permanently re-estimate the parameters of a model [2]. Quite often, however, it will also be necessary to adapt the *structure* of the rule-base e.g., to delete rules from a rule-based classifier or add additional ones [1].

A brief summary of the open problems in this area and some examples of on-going work by the author in collaboration with industrial or academic partners:

- 1) **Changing concepts and patterns.** This is an acute problem in fields like process industries and robotics (e.g., non-stationary signal processing, moving images, dynamic non-stationary processes).
Examples: collaboratives work with Ford R&D, BAE Systems.
- 2) **Very large and heterogeneous data sets.** Even if the data set to be analysed is static, it may be too large to process in batch mode. Moreover, analysing a data set as a whole might be problematic if the data is heterogeneous. In this case, good models might be found for different subsets of the data (locally valid sub-models) but not for the complete data set.
Examples: market-basket analysis (on-going work with Retail Analytics Ltd.), mass-spectrometry data analysis in proteomics (a collaborative work with Prof. F. Klawonn, Germany) etc.
- 3) **Scalability and model adaptation.** In many applications, the knowledge model may soon become outdated. Instead of designing a completely new system, a viable alternative is to adapt the existing one. In fact, a gradual evolution of an existing system has several advantages. For example, quickly adapting the rule-base to new operating conditions will improve process security. This legacy problem has implications, e.g., to decision support in the bio-medical area, speech processing, or hand-written character recognition. Moreover, if a solution has to be designed for a new but similar application, costs can be saved by taking over large parts of an existing system.
Examples: Error concealment in VoIP communication; collaboration with Nokia
- 4) **Hybrid systems.** In many applications, one disposes not only of empirical data, but also of knowledge coming from human experts. Problems of Knowledge and Data integration are subject to a collaboration with Prof. N. Kasabov (Auckland, New Zealand)
- 5) **Speed of learning and computational simplicity.** In many applications, for example in the field of robotics or traffic management, measurements are recorded in real-time. Correspondingly, these applications require real-time model evolution and, hence, incremental learning techniques with especially high computational performance.
Examples: collaborative work with Ford R&D, BAE Systems.
- 6) **Adaptive feature selection.** In industrial applications, the number of features extracted from objects like images or signals is typically large. As learning in high-dimensional input spaces is difficult and irrelevant features may impede the learning, the problems of dimensionality reduction and feature selection become crucial. The data space dimensionality may change over time, these techniques have to be used in an on-line mode. The questions of how to perform feature selection incrementally and how to adapt a model to a modified input space pose challenging problems.

Statements:

- ✓ Methods based on the way humans deduct knowledge from experience, this knowledge gradually evolves and is inherited will be useful and efficient tools to cope with the enormous amount of information that surrounds us;

- ✓ evolving systems (in the sense of evolution of an individual, not a population of individuals as in conventional evolutionary algorithms) are potentially very useful and powerful concept that can address problems of high level adaptation of non-linear complex systems to non-stationary environment and to internal changes in these systems;
- ✓ interpretability of the knowledge (usually in the form of rule-base) generated automatically from the data in real-time may be vitally important in such application as evolving decision support systems in medicine, market analysis, social systems etc.; further research is needed in this direction;
- ✓ the problems of real-time co-operation in knowledge generation from data between two or more systems needs further research. The spatially-local overlapping models can contribute to build the 'bigger picture' more effectively. Conflicts resolution and aggregation of the partial knowledge will borrow from group decision making techniques exercised by humans.

Key words: real-time, rule-base, knowledge discovery, evolving

References

- [1] Angelov, P., N. Kasabov (2005) Evolving Computational Intelligent Systems, *Proc. I Workshop on Genetic Fuzzy Systems*, Granada, March 17-19, 2005, pp. 76-82
- [2] Chai, K. M. A., H. T. Ng, H. L. Chieu (2002) Bayesian On-line Classifiers for Text Classification and Filtering, *Proc. SIGIR'02*, August 11-15, 2002, Tampere, Finland, pp. 97-104
- [3] Domingos P., G. Hulten (2001) Catching Up with the Data: Research Issues in Mining Data Streams, *Workshop on Research Issues in Data Mining and Knowledge Discovery*, Santa Barbara, CA.
- [4] Duda, R.O., P.E. Hart, D.G. Stork (2000) *Pattern Classification – 2nd Edition*. Wiley-Interscience, Chichester, UK.
- [5] Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth (1996) From Data Mining to Knowledge Discovery: An Overview, *Advances in Knowledge Discovery and Data Mining*, MIT Press.
- [6] Haykin S. (1999) *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Prentice Hall, NJ, USA.
- [7] Hand, D., H. Mannila, and P. Smyth (2001) *Principles of Data Mining*. MIT Press, MA, USA.
- [8] Haralick R.M., L.G. Shapiro (1993) *Computer and Robot Vision Vol. II*. Addison-Wesley, USA.
- [9] Hastie, T., R. Tibshirani, and J. Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag, Heidelberg, Germany.
- [10] Jin R., G. Agrawal (2003) Efficient Decision Tree Construction on Streaming Data. *Proc. of ACM SIGKDD*.
- [11] Kecman, V. (2001) *Learning and Soft Computing*, MIT Press, Cambridge, MA, USA.
- [12] Mitchell, T. M. (1997) *Machine Learning*, McGraw-Hill Education, UK.
- [13] Quinlan, J. R. (1987) Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234.
- [14] Simani, S., C. Fantuzzi, R.J. Patton (2002) *Model-based Fault Diagnosis in Dynamic Systems Using Identification Techniques*. Springer Verlag, Berlin Heidelberg.
- [15] Vapnik, V. N. (1998) *Statistical Learning Theory*, Springer, London, UK.
- [16] O. Zamir, O. Etzioni, O. Madani and R. M. Karp. Fast and intuitive clustering of Web documents. In *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 287-290, 1997.
- [17] C.S. Moeller-Levet, F. Klawonn, K.-H. Cho, O. Wolkenhauer: Clustering of Unevenly Sampled Gene Expression Time-Series Data. *Fuzzy Sets and Systems* 152 (2005), 49-66
- [18] Bonabeau, E., M. Dorigo, G. Theraulaz (1999) *Swarm Intelligence: From Natural to Artificial Systems*, Santa Fe Institute Studies in the Sciences of Complexity, USA. ISBN 0195131592
- [19] Michalewicz, Z. (1996) *Genetic Algorithm+Data Structures=Evolution Programs*, 3rd ed. Springer-Verlag: NY, US
- [20] de Castro, L. N. and Timmis, J. I. (2002) *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer-Verlag, London.
- [21] Bentley, P. (1999) *Evolutionary Design by Computers*, Morgan Kaufmann, ISBN: 155860605X
- [22] Koza J. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, ISBN 0-262-11170-5
- [23] Kasabov N., Q. Song (2002) DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-Series Prediction, *IEEE Trans. on Fuzzy Systems* 10(2) 144-154.
- [24] Lin, F.-J., C.-H. Lin, P.-H. Shen (2001) Self-constructing fuzzy neural network speed controller for permanent-magnet synchronous motor drive, *IEEE Trans. on Fuzzy Systems* 9 (5) 751-759.
- [25] Angelov, P., D. Filev, "An approach to on-line identification of evolving Takagi-Sugeno models", *IEEE Trans. on Systems, Man and Cybernetics, part B*, vol.34, No.1, 2004, 484-498
- [26] Angelov P. P. (2002) *Evolving Rule-based Models: A Tool for Design of Flexible Adaptive Systems*, Springer-Verlag, Heidelberg, New York, ISBN 3-7908-1457-1.