STATE-OF-THE-ART AND EVOLUTION IN PUBLIC DATA SETS AND COMPETITIONS FOR SYSTEM IDENTIFICATION, TIME SERIES PREDICTION AND PATTERN RECOGNITION

Joos Vandewalle, Johan Suykens, and Bart De Moor

Katholieke Universiteit Leuven, ESAT-SCD
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
Email contact: Joos.Vandewalle@esat.kuleuven.be

Amaury Lendasse

Helsinki Univ. Techn., Lab. Comp. and Inform. Sc.
P.O.Box 5400 FIN-02015 HUT, Finland
Email: lendasse@hut.fi

ABSTRACT

It is the aim of reproducible research to provide mechanisms for objective comparison of methods, algorithms, software and procedures in various research topics. In this paper, we discuss the role of data sets, benchmarks and competitions in the fields of system identification, time series prediction, classification, and pattern recognition in view of creating an environment of reproducible research. Important elements are the data sets, their origin, and the comparison measures that will be used to rank the performance of the methods. The issues are discussed, a comparison is made and recommendations are given.

***Index Terms*—** Identification, pattern recognition, prediction methods, time series

1. INTRODUCTION

The rise of information and communication technology has opened up various new mechanisms for cooperation and for pooling information in order to improve the quality of designs, systems, and processes. In a recent book of C. Sunstein [22] various important and recent sociological phenomena of the distributed production of knowledge are described and analyzed, like the self correcting mechanisms of wikis, the aggregation and synergy of information of market predictions, the large participation of contributors to technological developments using open source software, and the added value of aggregation of information without creating herd mentality.

In experimental research the typical role model is that of a researcher or a team of cooperating researchers that sets up an experiment to verify or falsify a certain concept, or design in the presence of a certain physical phenomenon. These researchers then describe their findings in a paper. Reviewers of that paper or competing researchers reading that paper then try to reproduce these experiments in order to verify the findings of that paper. However, there is often a lack of information on the experiment to reproduce it, thereby leading to frustration and a limited interest in the findings of that paper. Often also the experiment fails under slightly different circumstances, thereby reducing the value of the findings.

In this paper we discuss various forms of cooperation, interaction and competition among the different researchers in a domain such as benchmark problems, publicly available data sets, competitions, tournaments, and so on. In the same way as it has happened in various fields of sports such cooperations and competitions can lead to faster progress if a number of conditions of reproducibility and fairness are satisfied. Such mechanisms all fit very well in the whole idea of making research more reproducible and open access to knowledge in sciences and technology.

In the domains of time series prediction, classification, and pattern recognition, one has typically data sets of measurements that exhibit a wide range of ingredients and phenomena. During the design process of new methods the data set is split into three parts: the training set, the validation set and the test set. The training set is used in order to find the optimal parameters. The validation set is used to fix the meta-parameters or to select the best model during the design process. Finally the test set is used to compare the method with other methods. For a fair evaluation, the test set should not be used during the design. It is precisely at the test set that the pros and the cons of a competition versus a regular comparison can be distinguished. In a competition, the test results are not revealed to the participants during the design. These results are only presented publicly after the submission deadline, when they are compared during the oral or written performance analysis and when the different submitted methods are ranked. During a regular comparison, the test data is available at all times. The correctness of the comparison relies entirely on the honesty of the designers of the method. For example, the designers should refrain from using any information about the test data during the design of their method. They should also avoid choosing the test set in a biased way. This relies entirely on the honesty of the designers. For benchmark problems both in the case of a competition and in the case of a regular comparison the specifications of the system and the performance measures should be defined in advance and should be open to scrutiny and should have broad support in the scientific community.

This paper is organized as follows: In Section 2 we dis-

1-4244-0728-1/07/$20.00 ©2007 IEEE

cuss the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. In Sections 3, 4 and 5 we briefly discuss data sets in system identification, time series prediction and classification. In Section 6 recommendations for data sets selection and processing are given. Finally in Section 7, general conclusions are made.

2. RELATION TO OPEN ACCESS

The ideas and concepts of reproducible research fit very well in the general discussion on open access. It is worthwhile mentioning here the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities dated October 22 2003 [1], that has been signed by 164 organizations until now. The declaration points out that there are unique opportunities offered by the web and the internet, and that society can use these opportunities by making open access to data, software, methods, and writings.

Particularly relevant to this paper are the following quotes:

Definition of an Open Access Contribution:

"Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and meta-data, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material..."

Supporting the Transition to the Electronic Open Access Paradigm:

"Our organizations are interested in the further promotion of the new open access paradigm to gain the most benefit for science and society. Therefore, we intend to make progress by

- *encouraging our researchers/grant recipients to publish their work according to the principles of the open access paradigm...*
- *advocating the intrinsic merit of contributions to an open access infrastructure by software tool development, content provision, metadata creation, or the publication of individual articles."*

Along these lines the Organization for Economic Cooperation and Development (OECD) has recently drafted a recommendation [2] with similar statements.

3. DATA SETS FOR SYSTEM IDENTIFICATION

An initiative towards reproducibility of results in the area of system identification is the compendium of data sets on system identification called DAISY [3]. Its ideas include:

- Reproducibility of experimental results is one of the cornerstones of modern scientific research
- Cost-effectiveness: when many experimental datasets become publicly available, measurement set-ups do not need to be repeated
- Possibility for datasets to evolve into real benchmarks
- Stimulating interaction and collaboration between researchers active in system identification
- Standardized referencing to datasets in papers
- A fair and objective comparison of concepts, methods and algorithms
- Falsifiability: each theory should contain in itself the leverages by which it can be falsified. Data are instrumental in doing so.

The database is organized into several data categories such as for process industry systems, electrical/electronic systems, mechanical systems, biomedical systems, biochemical systems, econometric data, environmental systems, thermic datasets and others.

A benchmarking study organized in the area of nonlinear system identification is the Silver box case (NOLCOS 2004 special session, organizer J. Schoukens) with successful results obtained using nonlinear black-box techniques [18].

4. DATA SETS FOR TIME SERIES PREDICTION

Several challenging time-series competitions have been organized [5, 6, 7, 8] and time-series data sets have been collected, e.g. [4].

4.1. Santa Fe Time Series Competition

Six time series data sets were proposed: Data Set A within this competation: Laser generated data, Data Set B: Physiological data, Data Set C: Currency exchange rate data, Data Set D: Computer generated series, Data Set E: Astrophysical data, Data Set F: J. S. Bach's last (unfinished) fugue [5, 26]. The main benchmark of the competition was the Data Set A recorded from a Far-Infrared-Laser in a chaotic state. From this physical system 1,000 data points were given, and 100 points in the future had to be predicted by the participants. The winner of the competition was E.A. Wan using a finite impulse response neural networks for autoregressive time series prediction.

4.2. K.U. Leuven Time-Series Prediction Competition

The benchmark of the competition was a time series with 2,000 data [6, 23]. The competition data were generated from a computer simulated generalized Chua's circuit. The task was to predict the next 200 points of the time series. In total, 17 entries were submitted for the competition and the winning contribution was made by J. McNames (Fig.1). The strategy incorporated a weighted Euclidean metric and a novel multi-step cross-validation method to assess model accuracy. A nearest trajectory algorithm was proposed as an extension to fast nearest neighbor algorithms [20].

4.3. EUNITE: EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems classification competition

The problem to be solved here was the forecasting of maximum daily electrical load based on half an hour loads and average daily temperatures (time period 1997-1998). Also included were the holidays for the same period of time. The actual task of each participant was to supply the prediction of maximum daily values of electrical loads for January 1999 (31 data values all together). The advantages of this benchmark were the length (around 35,000 points) and that the real dataset allows to give further interpretation on the prediction result. The disadvantage was the specificity of the prediction with maximum of curves and the use of external inputs (temperatures). The winner of the competition was C.-J. Lin with a support vector machine method [17]. In total, 26 entries were submitted for the competition.

4.4. CATS Benchmark: Time Series competition

The proposed time series is the CATS benchmark, an artificial time series with 5,000 values [8]. The goal was the prediction of 5 blocks of 20 missing values. The advantage was that the set to be predicted was big enough and simultaneously the horizon of prediction was not too large (twenty step-ahead prediction). The disadvantage was that the problem is no longer a classical problem of time series prediction but a problem of determination of missing values in a temporal database. The winner of the competition was S. Sarkka using a Kalman smoother in order to perform the prediction [21]. In total, 25 entries were submitted for the competition.

5. DATA SETS FOR CLASSIFICATION

In the area of neural networks and machine learning it is currently common practice to test the design of new methods on data sets from e.g. UCI, Delve [9, 10]. Usually new techniques are being illustrated both on toy problems (or artificial data problems where one knows the true solution) and on real life data sets from repositories. As demonstrated e.g. in [24, 25] exhaustive benchmarking with comparisons between different methods on many different data sets can be very revealing. Although ’no-free-lunch’ theorems have been proven, certain techniques are able to become ranked consistently among the best results, while other techniques may sometimes perform excellently on certain types of data but break down on others. In this respect issues like scaling of data, removal of outliers and handling of different data types can be important. Challenging competitions have been organized e.g. on feature selection and on performance prediction [11, 12, 19]. Furthermore, the use of open source software is often stimulated [13, 14].

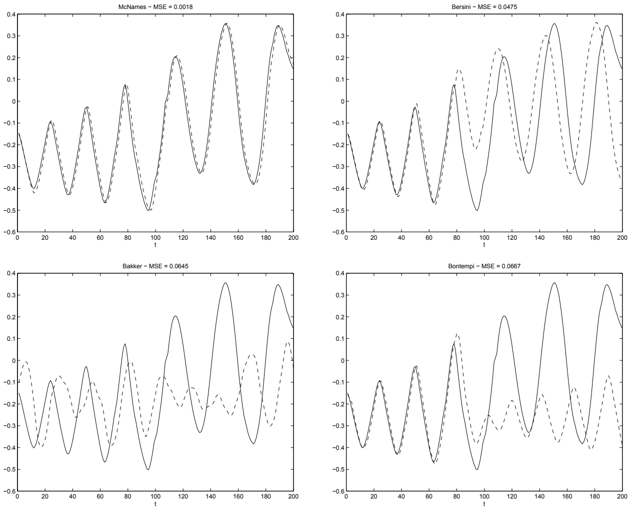

Fig. 1. *K.U. Leuven time-series prediction competition: illustration of the large variability in the results. Shown are 4 of the 17 submitted entries with the prediction of 200 future points in time (solid line: to be predicted, dashed line: prediction results).*

6. RECOMMENDATIONS FOR DATA SETS SELECTION AND PROCESSING

The design of benchmark problems and the selection of data sets involves many issues. First of all, it needs to be done objectively in order not to give any method unfair advantages. Moreover there is always a choice between breadth and depth. While broad coverage is desirable, it may not take into account the specificity of the concrete situation. Also the broad coverage avoids the design of methods that have a too narrow range of application, or even the design of methods that are tuned to a specific problem, and that do not work properly on other problems. The choice of problems can also range from toy problems to real applications. The toy problems have the advantage of being succinct, challenging and exciting the creativity, but may not convince the practitioner. In case of real applications the problems are often cluttered with so many details so that it is often quite tedious, but it is much more valuable for the users. So a delicate balance has to be struck in the design of benchmarks and the choice of data sets. Also one needs to select the data sets from different application domains in order to offer the user the opportunity to prove the broad validity of his/her methods.

7. CONCLUSIONS

This paper strongly encourages the development and broad distribution of benchmark problems and data sets and the organization of competitions for various relevant problems in signal processing, system identification, time series prediction, and classification. We argued that the wide availability will stimulate the quality of the new methods, and speed up the progress of the methods. Various participants in the research arena should contribute to make this process happen. Professional and research organizations like IEEE should endorse the data sets (see e.g. [15]), and benchmark problems that are designed and can widely distribute these among their members. The publishers of journals and organizers of conferences can stimulate their reviewers to devote special attention to the application of the methods on these benchmark problems or data sets. In education courses the design work of the students can be made more stimulating if they are invited to solve such benchmark problems or develop methods for public data sets.

Acknowledgements

Research supported by - Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, CoE EF/05/007 SymBioSys, IDO (Genetic networks), PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0407.02 (support vector machines), G.0197.02 (power islands), G.0141.03 (Identification and cryptography), G.0491.03 (control for intensive care glycemia), G.0120.03 (QIT), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0452.04 (new quantum algorithms), G.0499.04 (Robust SVM), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0211.05 (Nonlinear) , G.0226.06 (cooperative systems and optimization), G.0321.06 (Tensors), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); AWI: Bil. Int. Collaboration Russia; IWT: PhD Grants, GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), Eureka-Flite2, TAD-BioScope-IT, McKnow-E, Silicos; - Belgian Federal Science Policy Office: IUAP P5/22 (‘Dynamical Systems and Control: Computation, Identification and Modelling’, 2002-2006) ; - EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

8. REFERENCES

[1] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities
http://www.zim.mpg.de/openaccess-berlin/signatories.html.

[2] Draft OECD recommendation concerning access to research data from public funding, version for consultation, May 2006.

[3] DAISY: A Database for Identification of Systems
http://homes.esat.kuleuven.be/∼smc/daisy/.

[4] 800 well known time-series (Rob Hyndman)
http://www-personal.buseco.monash.edu.au/∼hyndman/TSDL/.

[5] Santa Fe Time Series competition
http://www-psych.stanford.edu/∼andreas/Time-Series/SantaFe.html.

[6] K.U. Leuven Time Series Prediction Competition
http://www.esat.kuleuven.ac.be/sista/workshop/.

[7] EUNITE: EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems classification competition
http://www.eunite.org/eunite/index.htm.

[8] The CATS Benchmark: Time Series competition
http://www.cis.hut.fi/∼lendasse/competition/competition.html.

[9] UCI Machine Learning Repository (University of California, Irvine)
http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[10] Delve Datasets
http://www.cs.toronto.edu/∼delve/data/datasets.html.

[11] Feature selection challenge NIPS 2003
http://clopinet.com/isabelle/Projects/NIPS2003/.

[12] Performance prediction challenge WCCI 2006
http://clopinet.com/isabelle/Projects/modelselect/.

[13] NIPS 2006 Workshop on Machine Learning Open Source Software
http://www.fml.tuebingen.mpg.de/raetsch/workshops/MLOSS06.

[14] Kernel machines website
http://www.kernel-machines.org/

[15] IEEE Computational Intelligence Society: Technical Activity Benchmark Repository
http://ieee-cis.org/standards/benchmarks/.

[16] No Free Lunch Theorems
http://www.no-free-lunch.org/.

[17] Chang M.-W., Chen B.-J., Lin C.-J., ”‘EUNITE Network Competition: Electricity Load Forecasting”’, *EUNITE competition*. Available: http://neuron.tuke.sk/competition/index.php.

[18] Espinoza M., Pelckmans K., Hoegaerts L., Suykens J.A.K., De Moor B., “A comparative study of LS-SVMs applied to the Silver box identification problem,” in *Proc. of the 6th IFAC Symposium on Nonlinear Control Systems* (NOLCOS 2004), Stuttgart, Germany, Sep. 2004.

[19] Guyon I., Alamdari A.R., Dror G., Buhmann J.M., “Performance prediction challenge,” *IEEE World Congress on Computational Intelligence* WCCI-IJCNN 2006, Vancouver, 2006.

[20] McNames J., “A nearest trajectory strategy for time series prediction,” *Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, July 8-10, 1998, K.U. Leuven Belgium, pp.112-128.

[21] Sarkka S., Vehtari A., Lampinen J., “Time series prediction by Kalman smoother with cross validated noise density,” In *Proc. International Joint Conferenece on Neural Networks* IJCNN 2004 Budapest, pp. 1653-1658, 2004.

[22] Sunstein C.R., *Infotopia: How many minds produce knowledge*, Oxford University Press, 2006.

[23] Suykens J.A.K., Vandewalle J., “The K.U.Leuven time-series prediction competition,” in Chapter 9 of *Nonlinear Modeling: advanced black-box techniques*, (Suykens J.A.K., and Vandewalle J., eds.), Kluwer Academic Publishers, 1998, pp. 241-253.

[24] Van Gestel T., Suykens J.A.K., Baesens B., Viaene S., Vanthienen J., Dedene G., De Moor B., Vandewalle J., “Benchmarking Least Squares Support Vector Machine Classifiers,” *Machine Learning*, vol. 54, no. 1, Jan. 2004, pp. 5-32.

[25] Pochet N., De Smet F., Suykens J.A.K., De Moor B., “Systematic benchmarking of micorarray data classification: assessing the role of nonlinearity and dimensionality reduction,” *Bioinformatics*, vol. 20, no. 17, Nov. 2004, pp. 3185-3195.

[26] Weigend A.S., Gershenfeld N.A., *Time Series Prediction: Forecasting the Future and Understanding the Past*, Reading, MA: Addison-Wesley, 1994.