Environmental Modelling & Software 25 (2010) 412–420

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A new library to combine artificial neural networks and support vector machines with statistics and a database engine for application in environmental modeling

Ralf Wieland [a,*], Wilfried Mirschel [a], Bernd Zbell [a], Karin Groth [a], Alison Pechenick [b], Kyoko Fukuda [c]

[a] Leibniz-Center for Agricultural Landscape Research, Institute of Landscape Systems Analysis, Eberswalder Street 84, 15374 Muencheberg, Germany
[b] College of Engineering and Mathematical Sciences, University of Vermont 351 Votey Hall, Burlington, VT 05405, USA
[c] Department of Mathematics and Statistics, Department of Computer Science and Software Engineering, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 27 November 2008
Received in revised form
2 November 2009
Accepted 14 November 2009
Available online 5 December 2009

Keywords:
Artificial neural networks
Support vector machine
Database management
Statistics
Open source software
Yield modeling
Spectral analysis

ABSTRACT

SADATO (SAMT DAta TOol) is an open source software library presenting new possibilities in modeling based on artificial neural networks and support vector machines. The main advantage of SADATO is its central data management based on Sqlite3 or MySQL and the statistical functions inherited from the APOPHENIA software. SADATO can be used for modeling as well as in large simulations. Modeling is demonstrated with two examples of artificial neural networks and support vector machines. The use of SADATO in simulation is supported by its very high computation speed. The highly aggregated functions in SADATO keep the software simple and easy to maintain. This allows the scientist experienced in software development easy access to all methods provided by SADATO. Additionally, an easy-to-use graphical user interface was developed to support scientists in developing models without any special knowledge in computer science.

1. Introduction

A wide variety of methods can be used to describe ecological dependencies. When these dependencies are linear, classical statistical methods such as regression methods or principal component analysis (PCA) can be used. However, in ecology the relationships are predominantly non-linear; for these cases, artificial neural networks (ANN) and more recently support vector machines (SVM) are widely used. Examples of applications using ANNs can be found in Kalteh et al. (2008) for water resources analysis, in Adoloye (2009) for predicting the capacity of water supply reservoirs, in Wieland and Mirschel (2007) for spatial grain yield estimation of winter cereals, in May and Sivakumar (2009) for prediction of urban stormwater quality, in Chenard and Caissie (2008) for stream temperature modeling, and in Vellido et al. (2007) for stream ecosystem modeling. Applications using SVMs are described in Lu and Wang (2005) for forecasting ambient air pollution trends and in Solomatine and Ostfeld (2008) for new model approaches in hydrology. A wide range of software has been developed to implement ANNs and SVMs. There are different open source software implementations available for ANN including Fast Artificial Neural Network Library (FANN)[1] and the Stuttgart Neural Network Simulator (SNNS)[2] as well as for SVM[3] Commercial tools like NeuroSolutions[4] or implementations in MATLAB are also common. The scientist who wants to share data and models with other scientists has to transform this data into a format compatible with others' software. There is no common database for these different software packages. The usage of the models underlying the trained ANN or SVM in simulations also depends on the software. Sometimes additional runtime libraries must be purchased (for MATLAB); sometimes the code must be transferred from the modeling tool to the simulation tool. In the case of closed source software the scientist has no opportunity to modify the software. These constraints handicap the transfer of scientific ideas and the development of new approaches in ANN and SVM.

The open source implementations of ANN and SVM within the statistical software R [Stevens and Henry (2009); Dimitriadou et al.

* Corresponding author: Tel./fax: +49 33432 82337.
E-mail address: rwieland@zalf.de (R. Wieland).

[1] http://leenissen.dk/fann/index.php.
[2] http://www.ra.cs.uni-tuebingen.de/SNNS/.
[3] http://www.csie.ntu.edu.tw/cjlin/libsvm.
[4] http://www.neurosolutions.com/.

1364-8152/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2009.11.006

(2008)] and the free MATLAB counterpart Octave[5] can help to cope with this situation. R also provides access to databases such as MySQL[6] and Sqlite3[7]. This allows the use of a common database for all data, statistical analyses of that data, and the trained SVM and ANN models provided by R itself, i.e., all components necessary for a scientist to work with ANNs and SVMs. This is an important advantage over closed source solutions. However, R is handicapped by being too slow for complicated (Monte Carlo) simulations, and lacks a comfortable user interface. These constraints apply to MATLAB as well.

As an alternative, a fast and reliable toolbox should be implemented in C++ with the following features:

- be open source to provide the user transparency and adaptability,
- contain different ANN types, as well as different training algorithms for ANNs and SVMs (including cross-validation),
- provide a unified access to the data sets required for training and validation,
- contain statistical functions (correlation analysis, PCA etc.) for data analysis and analysis of the modeling results and
- wrapped in a graphical user interface to support the scientist with an accessible tool.

The new methodology of such an approach is that all relevant functions should be included in the software library. This opens a "toolbox" for the scientist to combine and to compare different methodical approaches. Higher aggregated functions use basic functions from the underlying libraries (e.g., GSL). This provides an easy-to-use toolbox as well as a powerful implementation.

The aim of this paper is to explain a new methodical approach called SADATO (SAMT DAta TOol), SAMT: (Wieland et al., 2006). SADATO implements a new highly aggregated interface to open source software libraries and creates a high level software library preferred for ecological modeling. The library itself implements many hundreds of functions that can be used to combine what have been separate approaches to a new method. For example, the combination of wavelets with statistical methods such as multiple linear regression, or more powerfully with ANNs and SVMs, are interesting for the handling of spectral data as shown below. For actual implementation, it is important to show the efficiency in terms of runtime. A lengthy calculation often limits the interactive use of the software, and confounds the scientist's attempts at method optimization.

In this paper, two examples underline the new methodical approach SADATO. The focus of this paper is not to discuss an ideal solution for either example.

2. Method

2.1. Implementation of SADATO

SADATO (Fig. 1) has been implemented in order to provide an open source library combining two modern soft computing tools, SVM and ANN, with fast statistical and numerical power (Klemens (2008), Gough et al. (2009)) and a database engine in the background (Sqlite3 or MySQL). The user may incorporate this library into developed projects or wrap it in a graphical user interface, as discussed below. The C++ implementation provides a uniform interface to the SVM-Library and to FANN, making different data

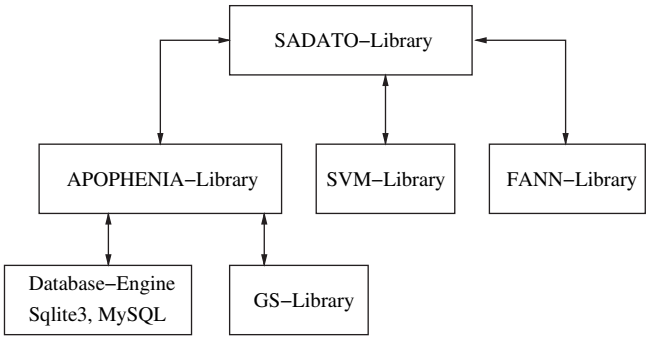

Fig. 1. Structure of SADATO Library.

formats obsolete (FANN uses a full input followed by the output, each on a separate line; SVM uses sparse matrix format). Using SADATO, the user can access all data via a database engine. This engine is the central point for storing data, and for accessing data during calculations. This ensures that an application implemented using SADATO has to provide only one interface for data import/export. For example, the "Spatial Analysis and Modeling Tool" (SAMT, SADATO's main application) (Wieland et al., 2006), can export samples of raster map grid points to database tables. This allows the direct use of satellite images, elevation models, outputs from spatial simulations, etc. as inputs to an ANN or SVM. Additionally, the user can add tables in the form of comma-separated value files to the database. The database engine in SADATO organizes the access to all these different data types, and ensures their integrity.

APOPHENIA-library (Klemens, 2008) provides a rich set of statistical routines:

- Maximum likelihood estimators for probit, exponential, gamma, Waring, Yule-Walker, Zipf, etc. estimators
- OLS and family, discrete choice models, kernel density estimators, and other common models
- Moments, percentiles, and other basic statistics utilities
- *t*-tests, *F*-tests, etc.
- Monte Carlo simulation

APOPHENIA does not re-implement basic matrix operations, random generators or numerical solvers. Instead, it builds upon the numerical software GS library and uses Sqlite3 as its database engine. The statistics routines in SADATO, provided by APOPHENIA, can be used for exploratory data analysis before initiating SVM or ANN training. For example, a correlation analysis helps to find dependencies between inputs; a PCA (Brumelis et al., 2000) reduces the number of inputs and ensures a simpler neural network structure. Modern methods such as GSL's implementation of wavelet transformations (Hubbard, 2006) are also included. The statistical routines can also be used to evaluate the trained neural networks or SVM. But this is not the only example for using statistical methods. In a real application involving ANN or SVM, statistical analysis often plays a key role. For example, in a research project such as "LandCaRe 2020" (LandCaRe2020, 2008) the impact of climate change on agriculture should be investigated. The relevant climate database contains data such as temperature, precipitation, solar radiation, etc. for many stations over many years. This data set is derived from the global climate simulation (ECHAM5, 2008) and depends on different climate scenarios. The user of this project needs information about the change in the yield caused by climate change from a basic time period (1971–2000) to a target time period (2021–2050). Statistics are used in this project to separate the long-term climate information from the short-term

[5] http://www.gnu.org/software/octave/.
[6] http://www.mysql.com/.
[7] http://www.sqlite.org/.