The Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)

Proceedings

Bled, Slovenia, September 27 - October 1

Edited by Sašo Džeroski,
Bernard Ženko, and Marko Debeljak

The Fourth International Workshop on Environmental Applications of Machine Learning EAML 2004

Proceedings

Bled, Slovenia
September 27 – October 1, 2004

Edited by

Sašo Džeroski
Bernard Ženko
and
Marko Debeljak

Jožef Stefan Institute
Ljubljana, Slovenia

CIP-Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

004.85:574(063)(082)
004.85:504(063)(082)

INTERNATIONAL Workshop on Environmental Applications of Machine Learning
(4 ; 2004 ; Bled)
Proceedings / The Fourth International Workshop on Environmental Applications of Machine Learning – EAML 2004, Bled, Slovenia, September 27–October 1, 2004 ; edited by Sašo Džeroski, Bernard Ženko and Marko Debeljak. – Ljubljana : Jožef Stefan Institute, 2004

ISBN 961-6303–59–7
1. Džeroski, Sašo

215423744

Editors:
Sašo Džeroski
Bernard Ženko
Marko Debeljak

Production, design, and type provided by the conference organization

Cover image, photograph of Bled, Slovenia, courtesy of Marjan Smerke,
Jožef Stefan Institute, Ljubljana, Slovenia

Publisher:
Jožef Stefan Institute,
Jamova 39,
SI-1000 Ljubljana, Slovenia
http://www.ijs.si

Printed in Slovenia by GRAFIKO, d.o.o., Grosuplje
Ljubljana, September 2004

Foreword

The Fourth International Workshop on Environmental Applications of Machine Learning (EAML) and the Fourth European Conference on Ecological Modelling (ECEM) were organized jointly in Bled, Slovenia, during the week of September 27 - October 1, 2004. The aim of these tievents was to bring together researchers from the areas of ecology, ecological modelling and environmental sciences, on one hand, and the areas of data analysis, data mining and machine learning, on the other. In many respects, they were treated as a single event, namely ECEM/EAML 2004.

Ecological modelling is concerned with the development and use of mathematical, computer and simulation models of ecosystems. It is a relatively young scientific discipline, which is rapidly gaining importance, especially because of the use and usefulness of ecological models for the management of natural resources. The European Conference on Ecological Modelling is the premiere European scientific event in the area of ecological modelling and regularly attracts an international audience.

As environmental concerns grow and information technology develops, more and more data on the different aspects (physical, chemical, biological, ecological) of the environment are gathered. There is an increasing need to analyse the collected environmental data for different purposes, which include the support for environmental management decisions. The International Workshop on Environmental Applications of Machine Learning provided a forum for presenting recent advances in applying machine learning and data mining techniques for the analysis of environmental data.

EAML 2004 covered all topics related to the application of data mining and machine learning methods to environmental data. An indicative, but non-exhaustive list of topics is given below:

- Analysis of environmental data with:
  - computational scientific discovery;
  - decision and regression tress;
  - evolutionary computing, e.g., genetic algorithms and programming;
  - statistical learning, e.g., kernel methods and SVMs;
  - neural networks, e.g., MLPs and SOMs;
  - probabilistic methods, e.g., Bayesian networks;
  - relational learning methods;
  - rule induction methods.

- Data mining and machine learning for:
  - modelling of different types of ecosystems, e.g., agricultural ecosystems, aquatic ecosystems, grassland ecosystems, forest ecosystems;
  - modelling different aspects of ecosystems/ecological processes, e.g., biodiversity changes, habitat suitability, population dynamics;
  - analysis of different types of environmental data, e.g., monitoring data (air/soil/water samples; chemical, physical, biological), remote sensing data (e.g., on atmosphere, geology, vegetation), spatial data (e.g., GIS data on land cover), temporal data (e.g., time series data on pollution levels);
  - various environmental applications of machine learning, e.g., for decision support in environmental management, for earthquake prediction, for environmental risk assessment, in environmental epidemiology, in meteorological/atmospheric sciences, in predictive ecotoxicology.

A total of 110 abstracts were submitted and 90 were accepted after a review process. Each submitted abstract was sent to three members of the Program Committee for review. The Program Committee members themselves did the majority of reviews, assisted by a few additional reviewers appointed by the Committee members. Authors of accepted abstracts (for oral or poster presentation) were invited to submit full versions of their papers. These were reviewed separately by the members of the Program Committee and other reviewers for inclusion in a special issue of the journal Ecological Modelling which was to be published after the conference.

The program of EAML 2004 consisted of three invited talks by Joseph C. Coughlan (NASA Ames, USA), Cesare Furlanello (ITC-irst, Trento, Italy), and Jacqueline McGlade (European Environment Agency), as well as oral and poster presentations of accepted contributions.

Many people contributed in various ways to the EAML 2004 workshop. We would especially like to thank:

- The authors of submitted contributions who made the workshop possible by presenting their work;
- The three invited speakers: Joseph C. Coughlan, Cesare Furlanello, and Jacqueline McGlade;
- The Advisory Committee for their suggestions regarding the Program Committee, invited speakers and encouraging remarks;
- The members of the Program Committee for their efforts in evaluating the submitted abstracts and papers;
- The members of the Organizing Committee;
- The sponsors for their generous support;

- The management and staff of the Albatros Congress Turist Agency, Bled, and the Center for Knowledge Transfer in Information Technologies, Jožef Stefan Institute, Ljubljana for their support;

- The Jožef Stefan Institute for providing the organizational infrastructure.

Sašo Džeroski, Bernard Ženko and Marko Debeljak
EAML 2004 Program Co-chairs

Contents

Acknowledgements

We acknowledge the support of the following institutions:

Jožef Stefan Institute, Ljubljana, Slovenia

Nova Gorica Polytechnic, Nova Gorica, Slovenia

Slovenian Ministry for Education, Science and Sport, Ljubljana, Slovenia

ISEM, The International Society for Ecological Modelling

KD–net, The Knowledge Discovery Network

PASCAL – Pattern Analysis, Statistical Modelling and Computational Learning, The network of excellence

Workshop Chairs

Sašo Džeroski

Bernard Ženko

Marko Debeljak

Organizing Committee

Tina Anžič

Jana Babič

Mili Bauer

Damjan Demšar

Peter Ljubič

International Advisory Committee

Ivan Bratko

Thomas G. Dietterich

Cesare Furlanello

Pat Langley

Sovan Lek

Donato Malerba

Friedrich A. Recknagel

Marko Slokar

William J. Walley

Program Committee

Nataša Atanasova
Vladan Babović
Marko Bohanec
Ivan Bratko
Broder Breckling
Bert Bredeweg
Ulises Cortes
Marko Debeljak
Thomas G. Dietterich
Yannis Dimopoulos
Sašo Džeroski
Nicola Fohrer
Tony Fountain
Cesare Furlanello
Jasna Grbović
Francois Guerrin
Sven E. Jørgensen
Michio J. Kishi
Andrej Kobler
Alexander S. Komarov
Boris Kompare
Branko Kontič
Vipin Kumar
Tetsuya Kusuda
Bill Langford
Pat Langley
Tarzan Legović
Sovan Lek
Li Bai-Lian
Horst Malchow
Donato Malerba
Joao Carlos Marques
Robert I. McKay
Felix Mueller
Søren N. Nielsen
Mark O'Connor
Luca Palmeri
Bernard C. Patten
Friedrich A. Recknagel
Miquel Sanchez-Marre
Dragan Savić
Mark Schwabacher
Masahiko Sekine
Cosimo Solidoro
Yuri M. Svirezhev
Ljupčo Todorovski
Sergio Ulgiati
Tanja Urbančič
Alexey Voinov
Hans Voss
Monica Wachowicz
William J. Walley
Peter A. Whigham
Franz Wotawa
Bernard Ženko

Organizational Support

The Albatros Congress Tourist Agency, Bled, Slovenia
Center for Knowledge Transfer in Information Technologies,
Jožef Stefan Institute, Ljubljana, Slovenia

Invited Talks

An overview of ecological modelling and machine learning research within the U.S. National Aeronautics and Space Administration

Joseph C. Coughlan

NASA Ames, USA

In the early 1980's NASA began research to understand global habitability and quantify the processes and fluxes between the Earth's vegetation and the biosphere. This effort evolved into the Earth Observing System Program which current encompasses 18 platforms and 80 sensors. During this time, the global environmental research community has evolved from a data poor to a data rich research area and is challenged to provide timely use of these new data. This talk will outline some of the data mining research NASA has funded in support for the environmental sciences in the Intelligent Systems project and will give a specific example in ecological forecasting, predicting the land surface properties given nowcasts and weather forecasts, using the Terrestrial Observation and Prediction System (TOPS).

GIS-based predictive models for ecology

Cesare Furlanello
Joint work with S. Merler, S. Menegon, M. Neteler,
S. Fontanari, R. Blažek, A. Rizzoli, and C. Chemini

ITC-irst, Trento, Italy

This talk will discuss how machine learning methods may be integrated within a Geographical Information System (GIS) for the development of new approaches in ecology research. As needed in tasks in landscape epidemiology and wildlife management, it is now possible to develop unified environments in which methods of statistical learning and spatial statistics are combined and applied to feature vectors derived from GIS analysis of digital maps, or from relational databases with embedded GIS capabilities. For example, multitemporal predictive maps may be obtained by modelling with classification trees or with Breiman's Random Forest in R, analyzing geodata and digital maps in the GRASS GIS, and managing biodata samples and climatic data in PostgreSQL. Overall, connecting a working data notification and management system to the predictive models is of crucial importance for a practical use of machine learning on ecological data.

We first describe a risk mapping system for tick-borne diseases, applied to create a multitemporal risk model of exposure to Lyme borreliosis and TBE in Trentino, Italian Alps. The system input features include vegetation data derived from the Forest register and multitemporal climatic data, also from remote sensing. As a second example, a predictive risk model for deer-vehicle collisions will be presented, considering variables as distance from urban areas and from waters, wildlife population density maps, vector line analysis (road curvatures) and traffic data. Features for the model include a multiscale accident site characterization based on the integrated use of orthophoto landuse classification and morphometrical analysis of the digital elevation model. The methodology has been applied at mesoscale (6200 km^2) for the predictive modelling of deer-vehicle collisions, a project for the Wildlife Management and Road Transportation Services of Trentino. We will present methods for variable importance analysis, classification with combined models and the resulting roe deer-vehicle accident risk maps.

Spatial assessments of Europe's environment

Jacqueline McGlade

European Environment Agency, Denmark

Integrated environmental and ecosystem health assessments rely on combining information from local and global attributes derived from surveys and case studies. In order to properly examine issues such as the impact of climate change, loss of biodiversity, environmental threats to human health or the long-term effects of infrastructure development on Europe's landscapes, the European Environment Agency (EEA) needs to be able to analyse changes across a range of scales and media (water, air, soil etc.). But despite extensive monitoring and research, the current situation in Europe is that we cannot meet the challenge of supporting consistent environmental and sectoral policies at a European, national and regional levels.

The knowledge needed will not be obtained solely through the accumulation of observations on individual systems but, will require such in situ data to be integrated within overall frameworks of models and data analysis to generalise their information content. In relation to the demands of understanding changes in Europe's environments, using spatially distributed data and information on ecosystems and human activities is a key factor, as they:

- can help identify where conflicts in use of the territory take place, and under which type of pressure;
- contribute to the stratification of data and knowledge from existing monitoring networks and research programmes;
- help in designing efficient sampling schemes for new monitoring networks as well as targeting research programmes to priority needs;
- provide important input to modelling, in particular when very heterogeneous information from the bio-physical, social and economic realms need to be integrated and
- can be up- and downscaled to the appropriate levels of decision making of the various public and private bodies.

In this context, land accounts for Europe are being implemented by the EEA. The purpose of land accounts is to observe, qualify and quantify the cover of land resulting from ecosystem and land use. Stocks of land cover are described as well as their change. A first set of land cover accounts is under construction using CORINE land cover data from 1990 and 2000. Within this accounting framework, assessments of ecosystem condition has been produced; for example, in the case of European wetlands, spatial data on the change in extension, fragmentation, connectivity and neighbourhoods can provide insights into the possible destruction and stress. These first variables are being supplemented by data on

flora and fauna and by quantitative and qualitative data on water. Other spatial data to be included, are land use in agriculture, urban development and transport infrastructure which will help to identify the sources of stress. These spatial data will be supplemented, at a more aggregated level, with social and economic statistics, from the perspective of the development of land use accounts to show how social and economic activities influence our environment.

The results of land accounts will be fully made available on the EEA website with the aim of facilitating access to these data and approaches to a range of users, including researchers and the wider public.

Contributed Abstracts

Accounting for different sources of temporal and spatial autocorrelation in non-linear habitat preference models

Geert Aarts[1], Jason Matthiopoulos[2], Bernie McConnel[3], Monique MacKenzie[4] and Mike Fedak[5]

[1] Sea Mammal Research Unit, Gatty Marine Laboratory, University of St. Andrews, St. Andrews, Fife KY16 8LB, Scotland, UK, ga15@st-andrews.ac.uk, +44-1334-462656, +44-1334-462632
[2] Sea Mammal Research Unit, jm37@st-andrews.ac.uk
[3] Sea Mammal Research Unit, bm8@st-andrews.ac.uk
[4] School of Mathematics and Statistics, University of St. Andrews, St. Andrews, Fife KY16 9SS, Scotland, UK, monique@mcs.st-and.ac.uk
[5] Sea Mammal Research Unit, maf3@st-andrews.ac.uk

Key words: Habitat preference, satellite telemetry, independence, mixed-effects models, marine predators

Introduction

Animals need resources to meet their requirement for survival, growth and reproduction. Conservation and management requires that we document animal preference for resources and the conditions at which they occur. Habitat selection studies deal with these questions (Manly *et al.* 1993).

The technology for remotely tracking individual animals has been rapidly developed over the last decade. The resulting data on the locations of individual animals have been widely used for constructing habitat selection functions. However, analysis of such data has not taken advantage of recent developments in the field of statistical modelling.

Model framework

Testing of hypotheses about the habitat selection at the population, sub-population or individuals level, based on animal locations, is complicated by the presence of non-independence of the sampling unit used at each of those levels. Non independence may result from the temporal and spatial auto-correlation of successive animal locations (Aebischer *et al.* 1993) , the repetition of individual behaviour and intra-specific social attraction or repulsion.

Given the complexity of most animals' responses to their environment, problems of non-independence must be addressed within a framework that is also capable of dealing with non-linearity.

Mixed-effect models (Pinheiro and Bates 2000) simultaneously model the habitat preference of the population (fixed-effects) and variations around the parameter estimates due to individual or sub-population stochasticity (random effects). To allow for nonlinear covariate relationships, parametric spline functions are used. The estimates of the model parameters and their robust variances are calculated using generalised estimating equations, which account for the non-independence in the responses.

Case study

We apply our methodology to data from a marine mammal, the grey seal (Halichoerus grypus) and a sea bird, the gannet (Morus bassanus). Both species are central-place

foragers that make forage trips to areas many kilometres offshore (McConnell *et al.* 1999, Matthiopoulos *et al.* 2004, Hamer *et al.* 2000). We relate these locations obtained from satellite relay data loggers with environmental variables such as sediment type, bathymetry and the accessibility of each point in space from the animals' central places. We propose biological interpretations for the variability measured within different groupings of the data.

Acknowledgements
This work was funded by NERC and DSTL. The environmental data on sediment type was provided by BGS.

References

Manly, B. F. J., McDonald, L.L. , and Thomas, D. L. *Resource selection by animals: statistical design and analysis for field studies*. Chapman and Hall, London, 1993.

Aebischer, N. J., Robertson, P. A., and Kenward, R. E.. 1993. Compositional Analysis of Habitat Use from Animal Radio- Tracking Data. *Ecology* 74:1313-1325, 1993.

Pinheiro, J. C., and Bates, D. M. *Mixed-Effects Models in S and S-PLUS.* Springer, New York, 2000.

McConnell, B. J., Fedak, M. A., Lovell, P., and Hammond, P. S. Movements and foraging areas of grey seals in the North Sea. *Journal of Applied Ecology*, 36:573-590, 1999.

Matthiopoulos, J., McConnell, B. J., Duck, C., and Fedak, M. A. Using satellite telemetry and aerial counts to estimate space use by grey seals around the British Isles. *Journal of Applied Ecology*, 41:476-491, 2004.

Hamer, K. C., Phillips, R. A., Wanless, S., Harris, M. P., and Wood, A.G. Foraging ranges, diets and feeding locations of gannets Morus bassanus in the North Sea: evidence from satellite telemtry. *Marine Ecology Progress Series,* 200:27-264, 2000.

Computational Revision of Ecological Process Models

Nima Asgharbeygi,[1] Pat Langley,[1] Stephen Bay,[1] and Kevin Arrigo[2]
[1]Computational Learning Laboratory, CSLI, Stanford University, Stanford, CA 94305
[2]Department of Geophysics, Stanford University, Stanford, CA 94305

Most ecological models are developed manually by scientists, who decide on their basic structure, tune their parameters, compare them against available data, and refine them in response. In contrast, most work on computational scientific discovery has emphasized the automated generation of models from data and background knowledge. In this abstract, we describe an approach to model revision that incorporates ideas from both traditions. We believe that computational tools for model revision offer great practical value to scientists by decreasing the time required to search for models while letting them retain control over the search space.

Our approach involves the modification of *quantitative process models*, a representation of knowledge that constrains search while remaining interpretable. In this framework, a model consists of a set of processes, each of which specifies one or more differential or algebraic equations that represent causal relationships among variables. Processes can include threshold conditions on variables that characterize when they are active. A variable may be labeled as observable, meaning it is present in the data, or play the role of a theoretical term that serves to link processes. Each variable is also marked as either exogenous, in that it influences other variables but is not influenced in return, or as endogenous, which means it is causally dependent on other variables.

For example, we have developed a process model for the aquatic ecosystem of the Ross Sea based on Arrigo et al.'s (2003) earlier model, which is cast as a set of differential equations. The new version incorporates four observable variables: the available light, the amount of ice, and the concentrations of phytoplankton and nitrate. The model includes processes for the loss of phytoplankton from natural causes and for its growth as a function of current concentration and growth rate. A third process specifies the decrease in nitrate associated with its update by phytoplankton, and another indicates that the growth rate is a product of the unconstrained rate and the minimum of two theoretical terms, nitrate-rate and light-rate, which determine the fraction of the unconstrained rate achievable when the available nitrate or light are limited. Two final processes specify parameters that occur across processes and describe the variable light as a function of time. We can utilize a process model of this sort, together with initial values, to simulate its behavior over time and thus predict values for each endogenous variable.

In previous work (Langley et al., 2003), we developed an initial algorithm, called IPM, to address the task of inducing process models like the one described above from time-series data and from knowledge about the domain. We cast this background knowledge as a set of *generic* processes which are distinguished from specific processes in that they do not commit to particular variables or parameter values. However, they can contain constraints, such as types for generic variables and intervals for parameter values. Although IPM produced encouraging results, it had drawbacks that limited its applicability: the space of explored models could still be large, and it provided no way to guide the search toward models a scientist might find more plausible.

In response, we have developed a new system, IPM/R, which adopts a revision approach to process model induction. This algorithm requires the user to specify four inputs. These include an initial model that encodes the user's beliefs about the processes that are most likely involved, a set of possible changes that specify which initial processes can be removed or have their parameters altered, a set of generic processes that may be added to the initial model, and observations to which the revised model should be fit. The possible changes, combined with the candidate processes for addition, guide IPM/R's search toward parts of the model space that are consistent with the user's knowledge about the domain.

IPM/R generates a set of revised models that are sorted by their distance from the initial model and presented to the user with their mean squared errors on the data. This output format lets the user observe the trade-off between the performance of the revised models and their similarity to the initial model, in order to determine the best compromise when selecting among the candidate revisions. We applied both IPM and IPM/R to the problem of modeling phytoplankton population dynamics in the Ross Sea, using the initial model described above and alternative generic processes that included mechanisms for zooplankton grazing on phytoplankton, nitrate remineralization, and residue loss. Our input data consisted of 188 daily measurements of sunlight, ice amount, phytoplankton concentration, and nitrate concentration in the Ross Sea.

Our runs revealed that IPM/R found revised models which reduced error substantially by making only a few ecologically plausible revisions to the original model, including the addition of processes for nutrient remineralization and zooplankton grazing. In contrast, IPM generated models with comparable error after a much longer execution time, and these were very different from the initial model and less comprehensible. These results demonstrate that IPM/R can produce accurate and comprehensible models that make contact with existing domain knowledge. Although some earlier work has utilized machine learning to revise quantitative models (Todorovski et al., 2003; Whigham & Recknagel, 2001), we have adapted this approach to the improvement of dynamical process models, which seem especially appropriate for fields like ecology.

References

Arrigo, K. R., Worthen, D. L. & Robinson, D. H. (2003). A coupled ocean-ecosystem model of the Ross Sea: 2. Iron regulation of phytoplankton taxonomic variability and primary production. *Journal of Geophysical Research, 108,* C7, 3231.

Langley, P., George, D., Bay, S., & Saito, K. (2003). Robust induction of process models from time-series data. *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 432–439). Washington, DC: AAAI Press.

Todorovski, L., Dzeroski, S., Langley, P., & Potter, C. (2003). Using equation discovery to revise an Earth ecosystem model of carbon net production. *Ecological Modeling, 170,* 141–154.

Whigham, P. A. & Recknagel, F. (2001). Predicting Chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. *Ecological Modelling, 146,* 243–251.

Modelling spatial distribution of Croatian marine benthic habitats

Tatjana Bakran-Petricioli[1], Oleg Antonić[2], Dragan Bukovec[3], Donat Petricioli[4], Ivica Janeković[2], Josip Križan[4], Vladimir Kušan[4], Sandro Dujmović[4]

[1] Department of Biology, Faculty of Science, University of Zagreb, Rooseveltov trg 6, 10000 Zagreb, Croatia, e-mail address: tatjana.bakran-petricioli@zg.htnet.hr, tel: +385 (0)91 5410 912

[2]Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia

[3]Croatian Natural History Museum, Demetrova 1, 10000 Zagreb, Croatia

[4]Oikon Ltd. Institute for Applied Ecology, Vlade Prekrata 20, 10020 Zagreb, Croatia

Key words: digital bathymetrical model, infralittoral, Landsat ETM+, neural networks, raster-GIS

Introduction

The Ministry of Environmental Protection and Physical Planning of the Republic of Croatia had financed the project "Mapping the habitats of the Republic of Croatia". Three-year project was finished in the spring 2004, resulting in the multi-layer spatial database about Croatian habitats, needed for the application of new Croatian Law on Nature Protection and for large number of other practical applications. On the land part of Croatian territory data source for mapping were classified and interpreted Landsat ETM+ satellite images (following standard supervised and unsupervised image processing procedures, e.g. Lillesand and Kiefer, 1994) with minimum mapping area (MMU) of 9 ha, as well as results of intense fieldwork. Existing data for sea bottom mapping, as well as new data that could have been collected during the project, were not sufficient to cover large spatial variability of marine habitats in Croatia, so a different methods were applied.

Supralittoral and mediolittoral habitats were mapped as a function of spatial distributions of basic classes of coastal lithology and basic levels of human impact. Circalittoral and bathyal habitats were mapped by overlapping and reinterpretation of existing small-scaled lithological maps in the framework of the raster-GIS.

Infralittoral was mapped on the basis of spatial modelling (with spatial resolution of 30 x 30 m), using neural networks (NN) as a modelling tool. This tool was chosen according to its flexibility for the solving a complex regression and classification ecological problems with incompletely known physical, chemical and biotic background, and according to the previous experience of the project team (e.g. Antonić et al., 2003). Basic benthic habitat types (*Posidonia* meadows, photophilic algae habitat, fine sand habitats and muddy sand habitats), determined by SCUBA diving on 972 infralittoral locations along the eastern Adriatic coast, represented dependent variable. Final list of independent variables (after sensitivity analysis and selection of significant estimators) included: 1) Euclidean distance from the coast, 2) median of Euclidean distances from the coast in the circle of 1 km around each unit of spatial resolution (30 x 30 m), 3) the second channel of Landsat ETM+ satellite image, 4) magnitude of spring sea bottom current, 5) magnitude of sea bottom current induced by strong wind forcing from south and north-east directions, 6) spring and summer sea bottom temperatures, 7) winter sea bottom salinity (see details about 4-7 in the parallel contribution of *Janeković et al.*) and 8) latitude and longitude. Prediction model was derived using the feedforward NN with multilayer perceptrons, which is appropriate for classification problems (Patterson,

1996). Logistic function was used as activation function and back-propagation method was used for the network training. The finally chosen NN architecture had one hidden layer with 7 neurons.

Results and discussion

The final model has total classification correctness of 77.83 % for the training dataset and 70.87 % for the independent verification dataset. The better results were achieved for *Posidonia* meadows and biocoenosis of photophilic algae (77.09 % and 73.15 % of correctly classified cases, respectively) in relation to the results for fine sand and muddy sand habitats (67.24 % and 66.01 %, respectively). According to this result, and after the preliminary application of the described original model (using the spatial distributions of independent variables) in the real space (where muddy sand habitats almost disappeared), it seemed reasonable to join fine sand and muddy sand habitats into one widely understood type (infralittoral sand habitats). This was done without the building of new NN, only results of described finally chosen NN have been reinterpreted in sense of joining two mentioned types. Resulting agreggated type has classification correctness of 78.08 %, which *a posteriori* increases total classification correctness on 76.60 %. Interpretation of remaining unexplained variability could be addressed to: 1) errors in mapping of localities with known habitat types, 2) errors in mapping of independent variables, 3) use of discrete and general habitat types, while natural boundaries between types are often blurred, 4) impact of other potential spatial predictors (especially lithological substratum) and 5) the model error.

The mentioned model was used for the spatial prediction of three main infralittoral habitat types for the entire Croatian maritory, in the frame of raster-GIS (using spatial resolution of 30 x 30 m). Yielded results were spatially generalized (using 3 x 3 focalmajority filter and omission of all homogeneous groups with minimum mapping unit less than 2.25 ha), and included in the final cartographical products arised from the project.

Conclusions

Developed model explains a significant part of the total variability of the main infralitoral habitat types recognized during the field work and it is usable for the construction of preliminary spatial distribution of main infralittoral habitat types for entire Croatian maritory. This preliminary distribution is the first and the only spatial database that cover entire infralittoral area of the Croatian part of the Adriatic Sea. Moreover, it is generally expected that this kind of models could give significant support to the nature protection and sea management purposes, especially in the sense of generalization of existing data, which are usually scarce due to the complexity and expensiveness of submarine fieldwork.

Acknowledgements

This work was supported by the Ministry of Environmental Protection and Physical Planning of the Republic of Croatia and by OIKON Ltd. - Institute for Applied Ecology.

References

Antonić, O., Pernar, N. and Jelaska S.D. Spatial distribution of main forest groups in Croatia as a function of basic pedogenetic factors. *Ecological Modelling* 170. 363-371, 2003.

Patterson, D. *Artificial Neural Networks*. Prentice Hall, Singapore, 1996.

Lillesand, T.M. and Kiefer, R.W. *Remote sensing and image interpretation*. John Wiley & Sons, Inc. New York, 1994.

Detecting density-dependence in ecological time series via VC-theory

Giorgio Corani[1], Marino Gatto[1]

[1] Dipartimento di Elettronica ed Informazione, Politecnico di Milano
Via Ponzio, 34/5- 20133 Milano, Italy
e-mail: corani@elet.polimi.it,
Phone: ++0039 02 2399 3562 Fax: ++0039 02 2399 3412

[2] Dipartimento di Elettronica ed Informazione, Politecnico di Milano
e-mail: gatto@elet.polimi.it

Key words: VC dimension, Structural Risk Minimization, model selection, demographic time series, density-dependence

Introduction

To recognize whether a population is growing in a density-dependent (DD) or independent way (DI) is of great practical importance in the design of proper policies for sustainable management and exploitation of populations. In fact, statistically distinguishing density-dependent from independent time series is of paramount interest for understanding the mechanisms regulating the species demography and predicting future population abundances. A widely adopted approach in recent works is the use of information criteria (IC) to choose the best from a suite of alternative models, including both density-independent and density-dependent demographies. IC's include the Final Prediction Error (FPE) and in particular the Schwartz Information Criterion (SIC) (see for instance Taper and Gogan, 2002). ICs were derived by using asymptotic arguments -which hold just for large dataset- and under restricting hypotheseses, such as the linearity of the unknown target function to be discovered; however, they are very often applied even if their constitutive assumptions are not met.

As a viable alternative to classical IC's, we propose the use of a model selection framework developed within Statistical Learning Theory (SLT) and called Structural Risk Minimization (SRM). SLT is the result of the joint work of Vapnik and Chervonenkis (Vapnik, 1995), and is built around the idea of VC-dimension *h*, a complexity index for classes of functions. Differently from traditional IC's, SLT is derived under more general hypotheses, such as finite sample settings and non-linear estimation. In the case of linear estimators, VC-dimension actually corresponds to the number of free parameters, and this allows the straitghforward application of SRM. Otherwise, it must be calculated with suitable methods (Vapnik et al, 1994).

Here, we conduct some theoretical experiments to investigate the effectiveness of SRM in detecting the correct demography from available ecological data; we generate noisy artificial time series, both DI and DD, and use FPE , SIC and SRM to recognize the model underlying the data, from among a suite made up of four different models, both density-dependent and independent. We show that SRM significantly overperforms traditional approaches in recognizing both density-dependence and independence.

Results and discussion

Our experimental methodology consists of three steps:

1. *simulation:* we simulate four different models (the drift or Malthusian (*M*), the Ricker (*R*), the Ricker with one environmental covariate (*RI*) and the Ricker with two covariates (*RII*)). Among them, the Malthusian is the only density-independent one. Each simulation is characterized by a *simulation setting*, given the parameters of the simulated model, the noise levels used to corrupt the data and the time series length. 500 experiments are performed for each simulation setting;
2. *identification:* from noisy simulated time series we identify the four models above;
3. *model selection:* we apply FPE, SIC and SRM in order to choose the best model from among the four candidates identified at step 2.

Table 1 summarizes the results of our experiments reporting the average recognition proportion. The average is taken over 500 simulations x 3 parameters setting x 4 noise levels x 4 time series lengths. The Malthusian model is almost always recognized by SIC and SRM (99% and 98%), while FPE has a signicant number of failures, selecting a density-dependent demography about 20% of the times. As for the density-dependence recognition, the Ricker model is selected with very high percentages just by SRM (92%), which overperforms FPE and SIC by about 20 points. The Ricker I model is recognized 58% of times by SRM and FPE, while SIC provides lower performances. Finally, the *RII* model is the only one for which FPE works better than SRM and SIC.

	Recognized model (FPE)				Recognized model (SIC)				Recognized model (SRM)			
	M	*R*	*RI*	*RII*	*M*	*R*	*RI*	*RII*	*M*	*R*	*RI*	*RII*
Simulated model												
M	**81%**	9%	5%	5%	**99%**	1%	0%	0%	**98%**	2%	0%	0%
R	0%	**72%**	15%	13%	19%	**78%**	2%	1%	4%	**92%**	3%	1%
RI	0%	24%	**58%**	18%	19%	29%	**49%**	3%	3%	36%	**58%**	3%
RII	0%	7%	21%	**72%**	6%	24%	18%	**51%**	2%	17%	24%	**57%**

Table 1: *Average ability of FPE, SIC and SRM in recognizing the model underlying the artificially generated time series. Percentages in bold refer to the model that generated the data (the higher, the better).*

Conclusions

According to our experimental findings, FPE tends to choose overparametrized models, and SIC tends to select parsimonious demographies. SRM clearly appears as the best balanced approach, as it provides the best outcome in three cases out of four.

References

Vapnik, V. N. *The Nature of Statistical Learning Theory*.Springer-Verlag,1995.

Taper, M. and Gogan, P., The northern yellowstone elk: density dependence and climatic conditions. *J. Wild. Management,* 66(1):106-122, 2002.

Vapnik, V., Levin, E. and LeCun, Y., Measuring the VC-dimension of a Learning Machine. *Neural Computation*, 6(5): 851-876, 1994.

Characterizing vertical forest stands structure using data mining methods

Marko Debeljak[1] and Jana Babič[2]

[1] Department of Knowledge Technologies, "Jožef Stefan" Institute, Jamova 39, SI-1000 Ljubljana, Slovenia, Email: Marko.Debeljak@ijs.si, Tel.: +386 1 4773307, Fax: ++ 386 1 4251038
[2]Nova Gorica Polytechnic, School of Environmental Sciences, Vipavska 13, Nova Gorica, Slovenia, E-mail: jana.babic@p-ng.si,

Key words: forest ecosystem, forest stand development, vertical structure, machine learning, decision trees

Introduction

The vertical structure of tree crowns is affected by a variety of factors ranging from the spatial position of the nearest trees, stand density, arrangement of branches on trees, to the stand dynamics which can be influenced by natural forest development processes (cyclical stand development, successions) or silviculture treatments (Bončina, 2000).

Structures and processes are inseparable, where the structure is reflecting processes, since the information about the vertical structure of tree crowns is very important in ecosystem based forest management. Therefore the study of the vertical structure of tree crowns may make a significant contribution to the knowledge about growth and developmental processes of forest ecosystems (Jørgensen et al., 2000). In order to improve ecosystem based forest management properties of vertical structures in stand dynamic process for both natural and managed forests have to be identified.

Many methods for quantifying vertical structures are arbitrarily defined and don't represent natural stratification patterns of forest stands (Latham et al., 1998), they are too time consuming for large scale analysis, and they can be too expensive (LiDAR) (Zimble et al., 2003).

To overcome those shortages we conducted a very detailed study of the vertical stand structure in both managed and virgin forest. Study plots were selected in the virgin forest remnant Rajhenavski Rog, Slovenia and lightly managed forest in its vicinity. The selection of study plots was restricted to the most dominant forest plant community *Omphalodo-Fagetum omphalodetosum* which belongs to the group of high karst Dinaric forests *Omphalodo-Fagetum* (Marinček et al., 1992) with silver fir *(Abies alba* Mill.) and beech (*Fagus sylvatica* L.) as the most dominant tree species.

Stand dynamics was described by four indicative forest cyclical developmental phases (Leibundgut, 1982), which have distinctive vertical and horizontal stand structures. The research has considered stands going through a juvenile phase, an optimal phase, a mixed phase, and a regeneration phase.

Four research plots (35m by 35m) were randomly selected within each developmental phase in both managed and virgin forest. All living trees from the research plots were described by the following attributes: tree species, diameter at breast high, tree height, tree volume, tree biomass, layer, depth of the crown, width of the crown, social position, vitality.

Results and discussion

To find structural properties for particular developmental phases in both managed and virgin forest, data mining methods were used. We developed a vertical forest stands models by automated data analysis using machine learning techniques with classification trees. Classification trees (Breiman et al., 1984), often called also decision trees (Quinlan, 1986), predict the value of a discrete dependent variable with a finite set of values (called class) from the values of a set of independent variables (called attributes), which may be either continuous or discrete. Decision trees have been induced as Top-Down Induction of Decision Trees (TDIDT) (Quinlan, 1986). Data mining analysis was performed by the Weka machine learning package. We used J4.8 algorithm, which is Weka's implementation of C4.5 decision tree algorithm (Quinlan, 1993) known as one of the most qualified and often used decision tree system.

Decision trees were induced for each of the four developmental phases for both manage and virgin forest. They were evaluated qualitatively and quantitatively. Developmental phases in virgin forest have very characteristic structure while in managed forest the differences between developmental phases are not very large.

Conclusions

Silviculture performs thinning practice in all forest stand develomept phases. Suppressed and weakened trees are removed from stands. The concurrence for growing place is much lower as in natural forest, and natural mortality of selected trees is thus reduced. This facilitate very homogeneous vertical crown structures of the stands in managed forest. We assumed that forest management measures in general and forest thinning practices in particular are the main reasons for the homogenisation of the vertical structure in managed forest. Differences in vertical structure between natural and managed forest should be used as arguments to make some changes in exisitng silviculture practices if we want to achive sustainable forest management.

References:

Bončina A., 2000. Comparison of structure and biodiversity in the Rajhenav virgin forest remnant and managed forest in Dinaric region of Slovenia. Global Ecology and Biogeography, 9: 201-211.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont.

Jørgensen S.E., Patten B.C., Straškraba M., 2000. Ecosystem emerging: 4. growth. Ecological Modelling, 126: 249-284.

Latham P., Zuuring H.R., Coble D.W., 1998. A method for quantifying vertical forest structure. Forest Ecology and Management, 104: 157-170.

Leibundgut H., 1982. Europäische Urwälder der Bergstufe. Bern und Stuttgart, Verlag Paul Haupt: 306 pp.

Marinček L., Mucina L., Zupančič M., Poldini L., Dakskobler I., Accetto M., 1992. Nomenklatorische Version der Illirischen Buchenwälder (Verband *Aremonio-Fagion*). Studia Geobotanica, 12: 121-135.

Quinlan J.R., 1986. Induction of Decision Trees. Machine Learning, 5: 239-266.

Quinlan J.R., 1993. Programs for Machine Learning. Morgan Kaufmann, San Mateo CA.

Zimble D.A., Evans D.L., Carlson G.C., Parker R.C., Grado S.C., Gerard P.D., 2003. Characterizing vertical forest structure using small-footprint airborne LiDAR, Remote Sensing of Environment, 87: 171-182.

Using multiobjective classification to model communities of soil microarthropods

Damjan Demšar[1], Sašo Džeroski[1], Paul Henning Krogh[2], Thomas Larsen[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia, {damjan.demsar, saso.dzeroski}@ijs.si
[2] Department of Terrestrial Ecology, National Environmental Research Institute, Roskilde, Denmark, {phk, thl}@dmu.dk

Key words: multiobjective classification, modelling, soil microarthropods,

Introduction

With the final task of designing a decision support system for managing farms, we start on a low level, trying to model effects of different farming practices on community of soil microarthropods. The impact of anthropogenic sources on the soil environment is almost exclusively assessed for chemical factors only, although in agriculture other mechanical factors like tillage and biological factor like crops have large impacts also (Steen, 1983). And since farming systems consist of a combination of a certain temporal sequence of interdependent events of different type and duration it is imperative to handle farming systems as a whole in order to rank its environmental benefits and impacts. One way to do this ranking is to collect information about the agricultural events and the soil biological parameters reflecting those events and relate the sequence of agricultural events to the biological parameters. Since the collection of data is in progress we were able to use the already collected data as an input to machine learning algorithms with the primary task of constructing empirically based models useful for predicting the soil quality in terms of quantities describing the soil microarthropod community from agricultural measures. And with the final task of constructing a decision support system in mind we had additional prerequisites.

We wanted to identify the most important factors like in (Demšar et al., 2003) by preferring small and simple models to bigger and more complex models while limiting the performance costs of small models in terms of decreased accuracy. Identifying the most important factors for the community of soil microathropods can guide us in further experimentation and data collecting where we would pay more attention to identified factors. Additional to the discovery of new knowledge we wanted to "rediscover" knowledge useful for the decision support system that the experts already know, but it is hard to transfer because of the gap between different branches of science and can be difficult to put down in writing. Since we prefer the machine learning tools that produce descriptive models, we can use those as a source of questions, that otherwise could not be asked without a lot of background knowledge in the domain of agriculture. The answers form the experts will be used (along with other sources of knowledge) to construct the decision support system. For both reasons we wanted as simple (without sacrificing too much accuracy) we tried to model community of with tools for multiobjective classification. Thereby we can produce one model for all agricultural measures we want to model. This one model is not only simpler as several models usually needed to model several agricultural measures, but also help us to understand different effects of same actions on different aspects of soil microathropods community.

We combined the two available datasets: The first dataset describes four experimental farming systems (Foulum experimental station, Denmark) in the years from 1989 to 1993, allocated to 15 fields, with pesticide use in a conventional system and in two integrated

farming systems and no pesticide use on the other (organic) fields, with 530 microarthropod samples collected (Krogh, 2004). The second dataset describes several organic farms (Foulum and Flakkebjerg experimental stations plus various farms in Jutland) in the year 2002. 430 samples were collected. To those datasets we added newly available data from 2003, giving us a total of 1330 samples, of which 1138 were suitable for predicting all class variables.

The datasets available for the study include the agricultural measures (attributes), for example, packing, tillage, fertilizer and pesticide use, crops planted and cattle grazing. The history of crops and grazing for the last 3 years is also available. The datasets also contain environmental variables describing the circumstances of the samples where community data on soil microarthropods have been produced. The transformations used on some attributes (different forms of tillage) were used to simulate the occasional non linear reducing impact of tillage (different powers simulate differently steep curves of impact). The dataset also includes measured species. Some species were grouped into acari group (mites), the rest of the measured species belong into collembola group (springtails) and all were used to calculate biodiversity.

We used multiobjective classifiers like CLUS to predict several measures at once. With this approach we can additionally simplify the models, since we do not have multiple models, but one model. Further more with one model for several measures we can compare different effects of same actions on different aspects of community of soil microathropods. For measuring the predictive performance of the model, we evaluated the correlation coefficient and several error measures using ten-fold cross-validation. We evaluated mean average error, root mean square error, relative average error and root relative square error.

Conclusions

We tried to model community of soil microathropods with machine learning methods from the data describing chemical, biological and mechanical actions on the fields. We then used so produced models to identify the most important parameters for soil mites, springtails and biodiversity of soil microathropods. By preferring small and simple models to bigger and complex models, we discovered that the most important factor for community of soil microathropods are soil type, previous crops grown in the observed field, and the different forms of tillage. We also identified the different effects of one action on several agricultural measures. We identified actions that have positive effect on one part of community of soil microathropods and negative effect on another part. Furthermore we used the models as a source of questions for the domain experts. We gained knowledge that will help us in further modeling and building decision support system for the management of farms. We have shown that the machine learning models can be used in multiple ways from predicting new values, to gaining new knowledge about the relation between the attributes and the dependent variable, to extracting knowledge from the domain experts.

Acknowledgements

This work was supported by ECOGEN funded by the Fifth European Community Framework Programme: Quality of Life and management of living resources contract no QLK5-CT-2002-01666 and DARCOF, Nature quality in organic farming.

References

Aha, D., and D. Kibler (1991) "Instance-based learning algorithms", Machine Learning, vol.6, pp. 37-66.

Demšar, D., Džeroski, S., Krogh, P. H. and Larsen T. (2003) Identifying the most important ag-ricultural factors for the soil community of microathropods, Proceedings of the International Electrotechnical and Computer Science Conference, Ljubljana, Slovenia

Krogh, P. H. (1994). Microarthropods as bioindicators. A study of disturbed populations. PhD thesis Ministry of the Environment and Energy. National Environmental Research Institute, Silkeborg.

Steen, E. (1983). Soil animals in relation to agricultural practices and soil productivity. Swedish J. agric. Res. 13, 157-165.

The Use of Data Mining for the Monitoring and Control of Anaerobic Waste Water Treatment Plants

[1]Julian Gallop, [2]Maurice Dixon, [3]Simon Lambert, [4]Laurent Lardon, [4]Jean-Phillipe Steyer

[1]Business and Information Technology Department,
CCLRC Rutherford Appleton Laboratory,
Chilton, Didcot, Oxon. OX11 0QX, UK

[2]London Metropolitan University, UK

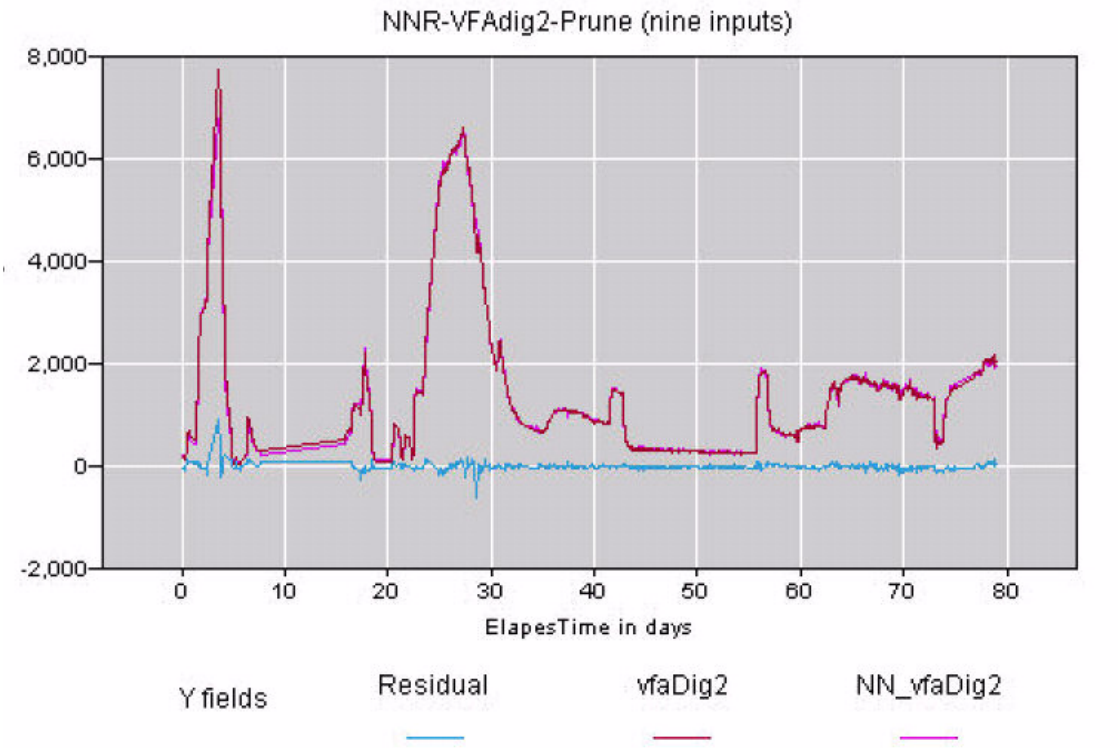
[3]CCLRC Rutherford Appleton Laboratory

[4]Laboratoire de Biotechnologie de l'Environnement - INRA, France

Key words: waster water treatment, data mining

This paper describes an approach using data mining to solve problems in the monitoring and control of anaerobic waste water treatment plants.
The anaerobic process possesses well understood advantages, but a digester based on these principles may enter an unstable state, leading to a breakdown of the process, requiring a shutdown and lengthy restart process. For reasons of caution therefore, an anaerobic digester is typically operated with a low throughput and the ideal would be to run a plant at higher efficiency, while maintaining stability.
Within the EU-funded project TELEMAC, several approaches are taken. One approach is to remotely monitor multiple plants through a Telecontrol Centre (TCC), providing expert support for organizations managing small plants with no local expertise. Another is to strengthen the expertise available at a TCC through a Decision Support System (DSS). Towards this end, investigation into plant simulators and into fault detection, isolation and diagnosis is undertaken. Measurements of critical physical and chemical variables are taken manually and by sensors. Procedures for detecting sensor faults, estimating critical measurements in the absence of a sensor (whether faulty or never present) and understanding which chemical variables are most critical to the determination of the digester state are investigated with a view to incorporating this understanding into the DSS. Although the TELEMAC project focuses on waste water from the production of alcoholic beverages, there is, between plants, a diversity of digester volume, underlying biological principles, and of sensor sets.
A fault detection, isolation and diagnosis procedure has already been designed on the basis of existing expert knowledge. For data mining, the task is whether a better understanding of the current and imminent state of the digester can be gained and which sensors are required for this. The aims of using data mining for digester plants in Telemac are to determine: the most useful set of sensors; whether some sensors can be omitted; to be able to predict imminent problems. Data mining results are made available to the Decision Support System which controls the treatment plants.. Several data mining techniques have been chosen. These

are neural nets, for analyzing prediction risk, cluster analysis and rule induction and work on principal component analysis is beginning.
The paper presents each of the chosen data mining techniques, its application to the problems identified in the management of anaerobic waste water treatment plants and some results. The paper also presents the role of visual analysis in relation to data mining in this application.



Grid-based Data Analysis of Air Pollution Data

Moustafa Ghanem, Yike Guo , John Hassard , Michelle Osmond and Mark Richards

Imperial College London, 180 Queens Gate, London SW7 2BW, UK,
Tel: +44 20 7594 8357, Fax: +44 20 7581 8024
{m.ghanem, y.guo, j.hassard, m.osmond, mark.richards}@imperial.ac.uk

Key words: remote sensing, grid-based knowledge discovery, distributed data sources, data mining, monitoring spatio-temporal data

Introduction

In this paper, we present a distributed infrastructure based on grid technology and data integration and mining tools to address the main informatics challenges that arise when a high-throughput sensor network is constructed to address real environmental challenges, such as real-time urban air pollution monitoring and mapping. Our efforts are part of the Discovery Net project (Curcin, 2002), a UK e-Science project developing grid-based knowledge discovery environments to support real-time processing, interpretation, integration, visualisation and mining of massive amounts of time-critical data. One of the application areas of Discovery Net is the analysis of data generated by high-throughput GUSTO sensors for monitoring urban air pollution. GUSTO is an acronym for Generic Ultraviolet Sensors Technologies and Observations. Based on open-path DUVAS^{TM} (differential ultraviolet absorption spectroscopy) technology, GUSTO measures and transmits the volume mixing ratios (at ppb levels) of key urban pollutants in real-time, providing exceptional temporal and spatial resolution. GUSTO is unique in the sense that it combines the accuracy of spectroscopic measurement techniques with a fast reliable retrieval algorithm - derived from satellite remote sensing principles (Martin, 2002).

Knowledge Discovery Challenges for Environmental Data

The wealth of data produced by a GUSTO sensor network allows a variety of analysis, modelling and visualisation tasks to be conducted, including the respective spatiotemporal variation of multiple pollutants, and their correlation with third-party data, such as weather, health or traffic data. Such data sets typically reside on remote databases, are stored in a variety of formats, and must be integrated with a multitude of data analysis components. This adds a new layer of complexity for data management and analysis considerations, and therefore it is vital that a flexible infrastructure is in place that allows full exploitation of the available data sets and which can incorporate new analysis components as determined by varying end user analysis requirements, (e.g. city planners vs. environmental organizations). Furthermore, such data lends itself to real-time environmental decision-making capabilities as hazardous pollution levels can be identified quickly.

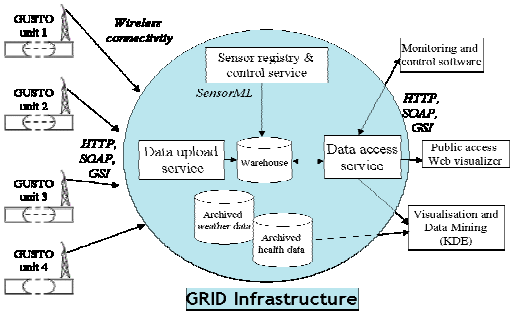
Grid-based Knowledge Discovery Infrastructure

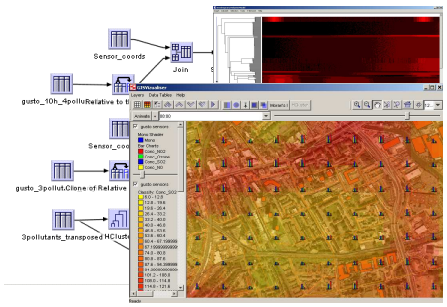
The diversity, physical distribution and heterogeneity of environmental data makes it impractical to use `closed' data mining systems that assume a centralised database or a data warehouse where all the data required for an analysis task can be materialised locally at any time, before feeding them to data mining algorithms which themselves have been predefined at the configuration stage. The Discovery Net architecture provides a platform for open data mining allowing the integration of distributed data sources and distributed tools in knowledge discovery activities, providing an application layer for grid-based knowledge discovery

services. Whereas current research into fundamental Grid technologies, such as Globus (Foster, 1997) has concentrated on the provision of protocols, services and tools for creating coordinated, transparent and secure globally accessible systems, Discovery Net allows scientists to create and manage complex knowledge discovery workflows that integrate data and analysis routines. The architecture allows scientists to store, share and execute these workflows remotely, as well as publish their workflows as new services, using a web service interface for access and integration by other remote applications. Discovery Net provides a higher level of abstraction for knowledge discovery activities on the Grid, separating the end-users from resource management issues handled by existing and emerging Grid standards.

By extending Discovery Net's use of grid services to support a distributed sensor network, each sensor may in effect become a grid service, with its capabilities published in a registry allowing the sensor's data to be accessed and retrieved using standardised protocols. A computer in each GUSTO unit analyses the sensor readings, generating a measurement of concentration for each pollutant every 2 seconds, and uploading the data at intervals to a remote Grid service, which manages the centralised warehousing of data. The sensor network may be monitored and controlled using similar technology.

We have developed components for data visualisation using GIS technology, and data mining components to work with archived data, including various forms of automatic data clustering techniques to study the correlation between different pollutants across time and space, components to identify trends within and across different time series at multiple levels of resolution, and components to find correlations with external data sources. There is further potential for conducting knowledge discovery processes that integrate standard data mining with both text and image mining activities (Ghanem, 2002).





Acknowledgements

This work was supported by the EPSRC under the UK e-Science Programme.

References

Foster, I. and Kesselman,C. Globus: a metacomputing infrastructure toolkit. *Int. J. Supercomputer Appl.*, 11 (2), 115–128, 1997.

Curcin, V., Ghanem, M., Guo, Y., Kohler, M., Rowe, A., Syed, J., and Wendel, P. Discovery Net: Towards a Grid of Knowledge Discovery. *Proceedings of KDD-2002. The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2002.

Ghanem, M., Guo, Y., Lodhi, H. and Zhang, Y. Automatic Scientific Text Classification Using Local Patterns: KDD CUP 2002 (Task 1). *SIGKDD Explorations*. Vol. 4. No.2 2002.

Martin, P. N. Measurement of Atmospheric Trace Gases Using Open Path Differential UV Absorption Spectroscopy for Urban Pollution Monitoring, PhD Thesis, University of London 2002.

A new multiobjective strategy to support model selection for environmental modeling

Orazio Giustolisi[1], Angelo Doglioni[1], Dragan A. Savic[2] and Francesco di Pierro[2]

[1] Civil and Environmental Engineering Department, Technical University of Bari, Engineering Faculty of Taranto, V.le del Turismo n. 8, 74100 Taranto, ITALY.
E-mail: a.doglioni@poliba.it. Tel. +390994733210 – Fax +390994733230.

[2] Centre for Water Systems, Department of Engineering, School of Engineering, Computer Science and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QX, UK. E-mail d.savic@ex.ac.uk.

Key words: Data-driven, Multiobjective, Evolutionary polynomial regression, Pareto front, Robust modelling.

Introduction

Data-driven techniques are commonly considered among the most powerful approaches to analysis of environmental phenomena. Their increased popularity can also be attributed to the increase in availability of monitoring data. Among the data-driven techniques, non-linear symbolic modelling represents a paradigm that enables the user to make a comprehensive analysis of the problem. In particular, a symbolic expression usually allows the analyst to achieve two important goals: (i) to predict the behaviour of environmental systems under different scenarios; (ii) to gain physical insight about the environmental phenomenon.
A well known, powerful method for performing symbolic regressions is based on Genetic Programming (GP) (Koza, 1992; Babovic and Keijzer, 2000). However, such methods exhibit certain limitations, as described by Davidson et al. (2003). Hybrid methods combining evolutionary algorithms with traditional numerical regression models, such as Least Squares (LS) method, were subsequently introduced. A recently developed hybrid method, the Evolutionary Polynomial Regression (EPR), see Giustolisi and Savic (2004a), has recently been tested on several different environmental problems (Giustolisi and Savic, 2004a, 2004b; Giustolisi et al, 2004a; Giustolisi et al, 2004b) showing encouraging results. In particular, EPR offers the following advantages: (a) it returns simple symbolic expressions; (b) it is robust, i.e. it always returns a consistent result; (c) it returns multiple symbolic expressions, each one aimed at investigating a particular aspect of the problem considered. The last feature is important as reported because it allows a multi-model approach to problem solving (Giustolisi et al. 2004c), which is seeking the best model for a particular purpose. Therefore, the user can get the best interpretable model for physical insight, the best model for data augmentation and the best models for each prediction horizon he/she wants to forecast. In this scenario, the aspect related to the selection of the best set of models comes up as a serious problem. The EPR multi-model approach produces a number of models and then the selection of the best models is made by the user according to the post-processing results. The user chooses the best models for the problem according to the performance measure (objective) and according to his/her physical insight (Giustolisi et al., 2004c).

Results and discussion

An improvement to EPR is reported in this paper, named Multi-Objective EPR (MO-EPR), which represents a development of the classical EPR for the robust choice of the best models in environmental modelling. MO-EPR finds the set of symbolic expressions (models) that perform well according to two or more conflicting criteria considered simultaneously: the level of agreement between simulated and observed measurements, and structural parsimony

of the expressions obtained. The three objective functions used are: (a) maximization of the fitness (forecasting), (b) minimize the number of inputs selected by the modelling strategy (see Giustolisi and Savic 2004a) and (c) minimize the length of the model expression. MO-EPR is tested on a case study dealing with the construction of the relationship between rainfall height and groundwater levels measured in a sampling well. The available data present some missing samples inside the data record. The solutions, i.e. the symbolic models, searched are ranked according to the Pareto dominance criterion (Van Veldhuizen and Lamont, 2000). Finally, the user is presented with the expressions (models) that dominate others in the population of solutions. The Pareto set solution are likely to be the best set of expressions required for the analysis of the problem, note that the proper choice of the objective functions plays an important role in this modelling framework. The MO-EPR models range from the simplest model, i.e. the average value of the training parameter, to the highly non-linear complex structured model. The set of non-dominated solution found by MO-EPR is used for a comprehensive analysis of the phenomenon.

Conclusions

The results obtained in the case study show the power and the potentialities of the MO-EPR methodology, in particular as decision support technique. MO-EPR proves to (1) provide symbolic interpretable models, (2) support the robust selection of models, (3) be computationally fast and (4) provide a comprehensive analysis of the phenomenon. Finally, the described MO-EPR is not customised at all for the case study, therefore it has great potentialities in the study of a wide range of environmental problem.

References

Babovic V., Keijzer M. Genetic programming as a model induction engine. *Journal of Hydroinformatics*, 2(1) 35–61, 2000.

Davidson, J.W., Savic D.A., and Walters G.A. Symbolic and Numerical Regression: Experiments and Applications. *Information Sciences*, 150 (1/2), pp. 95-117, 2003.

Giustolisi, O., and Savic, D.A. A Symbolic Data-driven Technique Based on Evolutionary Polynomial Regression, *Journal of Computing in Civil Engineering*, ASCE, in review, 2004a.

Giustolisi, O., and Savic, D.A. A novel strategy to perform genetic programming: Evolutionary Polynomial Regression. Proceedings of 6th International Conference on Hydroinformatics, in press, Singapore, 2004b.

Giustolisi, O., Savic, D.A. Decision Support for Water Distribution System Rehabilitation Using Evolutionary Computing. Proceedings of ACTUI seminar, pages 76-83, *CWS*, Exeter, UK, 2004c.

Giustolisi, O., Savic, D.A., and Doglioni A. Data Reconstruction and Forecasting by Evolutionary Polynomial Regression. Proceedings of 6th International Conference on Hydroinformatics, in press, Singapore, 2004a.

Giustolisi, O., Savic, D.A., Doglioni A., and Laucelli, D. Knowledge discovery by Evolutionary Polynomial Regression. Proceedings of 6th International Conference on Hydroinformatics, in press, Singapore, 2004b.

Giustolisi, O., Doglioni, A., and Savic, D.A. A multi-model approach to analysis of environmental phenomena. Proceding of iEMSs 2004, Biennial meeting of the International Environmental Modelling and Software Society, in press, Osnabrück, Germany, 2004c.

Koza J.R. Genetic *Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

Van Veldhuizen, D. A., and Lamont, G. B. Multiobjective Evolutionary Algorithms Analyzing the State-of-the-Art. *Evolutionary Computation*, Vol. 8(2), pp. 125-144, 2000.

Using Wavelets for the Classification of Hyperspectral Images

Hector Jasso[1], Peter Shin[1], Tony Fountain[1], Deana Pennington[2]

[1]San Diego Supercomputer Center, University of California San Diego.
9500 Gilman Drive, MC 0505, La Jolla, CA 92093-0505, USA.
{hjasso, fountain, kg}@sdsc.edu, Phone: 858-534-5000, Fax: 858-534-5152
[2]LTER Network Office, University of New Mexico Dept. of Biology.
MSC03 2020, Albuquerque, NM 87131-0001, USA. penningd@lternet.edu

Key words: hyperspectral, wavelets, classification, landscape, ecology

Introduction

A Jet Propulsion Lab (JPL) Aviris (Airborne Visible InfraRed Imaging Spectrometer) (http://aviris.jpl.nasa.gov) 300 x 300 pixel hyperspectral image from the Sevilleta National Wildlife Refuge in New Mexico (http://sev.lternet.edu) was collected, and a small subset (673 pixels out of 90,000) was manually labeled with eight possible land covers: river, riparian, agriculture, arid upland, semi-arid upland, barren, pavement, or clouds (see figure 1). The labeled data has been used to automatically estimate the land cover of the complete image using machine learning algorithms like Maximum Likelihood, Naive Bayes Classifier, Minimum Distance, and Support Vector Machine (SVM) (Tooby, 2003). The classification accuracies for these algorithms were of 96.4%, 90.9%, 88.4%, and 77.6%, respectively (Pennington, Jasso, Shin & Fountain, 2004).

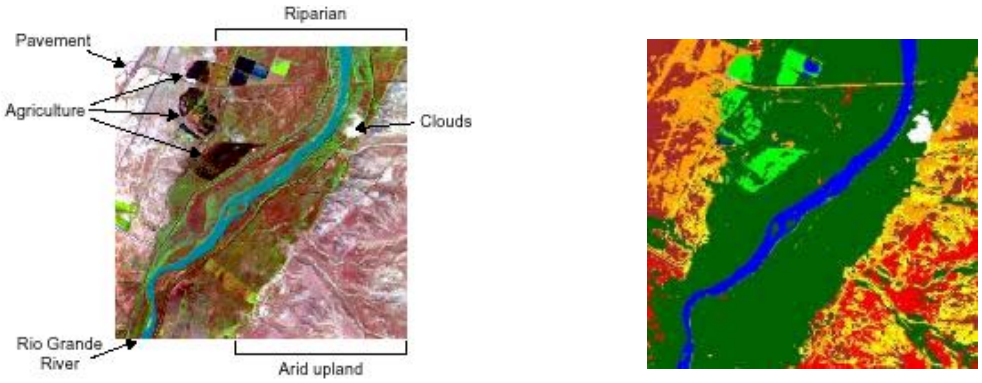


Figure 1. *False composite of the 300 x 300 pixel (6 km^2) study area (left), and classification results using Support Vector Machines (right).*

Results and discussion

Hyperspectral images sample frequencies ranging from ultraviolet to infrared, resulting in 201 values per pixel. Although the algorithms used are well suited for this high-dimensional data, the extra information gained by sampling new frequencies makes less evident certain characteristics that can help increase the classifier accuracy. Specifically, the general continuity in the values of adjacent frequencies produces specific patterns across the whole

spectrum that give a good indication of the pixel's land cover type (see figure 2). Wavelet analysis makes these global patterns explicit, by breaking down the signal into variable-sized windows, where long time windows capture low-frequency information and short time windows capture high-frequency information. High frequency information translates to information among close neighbors while low frequency information displays the overall trend of the features. We preprocessed the data using different families of wavelets, increasing the performance of the Naive Bayesian Classifier and SVM to 94.2% and 90.1%, respectively.

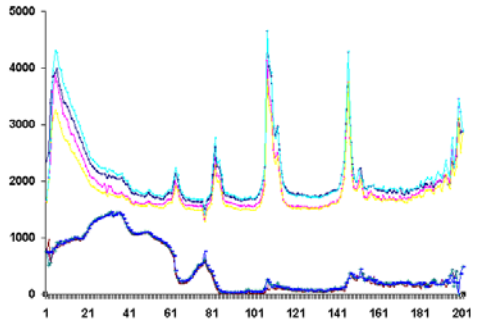


Figure 2. *Hyperspectral values for four cloud pixels (above) and four river pixels (below)*

Classification accuracy with SVM was further increased by modifying the mechanism by which multi-class is achieved using basic two-class SVMs. The winner-take-all SVM scheme was replaced with a one-against-one scheme, with a resulting classifier accuracy of 97.1%. One-against-one works by building *k(k-1)* binary classes for a *k* class problem, one binary class for every pair of classes. New data points are classified according to the binary class with the highest number of votes.

Conclusions

By preprocessing the Sevilleta hyperspectral data with wavelets and using a one-against-one scheme for multiclass SVMs we have increased the accuracy of previous classifiers. High accuracy classifiers are important because resulting images are used as the basis of further scientific analyses like the study of the biodiversity of plant species across time, the estimation of forest fire risk, and the construction of models of future climate change. This work is part of a joint effort between the Long Term Ecological Research (LTER) Network Office and the San Diego Supercomputer Center (SDSC) for the application of data mining algorithms to classification analysis of voluminous remotely sensed imagery.

Acknowledgements

This project was funded by the National Science Foundation through the National Partnership for Advanced Computational Infrastructure, NSF Cooperative Agreement ACI-9619020.

References

Tooby, P. Discovering knowledge in massive data collections. *EnVision* V19, No.3, 2003. (http://www.npaci.edu/enVision/v19.3/grid_computing.html)

Pennington, D., H. Jasso, P. Shin, & T. Fountain. The effect of landscape heterogeneity on classification accuracy: a comparison of classifier prediction in sub-opotimal sampling conditions. *Seventh Workshop on Mining Scientific and Engineering Datasets, 2004 SIAM International Conference on Data Mining (SDM 2004)*, Orlando, Florida, 2004.

Application of machine learning methods to palaeoecological data

Marjeta Jeraj [1], Sašo Džeroski [2], Ljupčo Todorovski [2], and Marko Debeljak[2]

[1]Department of Botany, University of Wisconsin
430 Lincoln Drive, WI-53706 Madison, USA
Email: jeraj@wisc.edu, Phone: (+1) 608 262 2279, Fax: (+1) 608 262 7509
[2] Department of Knowledge Technologies, "Jožef Stefan" Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
Emails: Saso.Dzeroski@ijs.si, Ljupco.Todorovski@ijs.si, Marko.Debeljak@ijs.si

Key words: palaeoecology, vegetation dynamics, machine learning, classification trees, equation discovery, hierarchical clustering

Introduction

Palaeoecology is a study of past environment, including palaeoclimate, geomorphology, past hydrology, soil development, paleovegetation, palaeofauna and human settlement history (Bell and Walker, 1992). Various palaeoecological techniques provide a reconstruction of former environmental conditions and dynamics, and allow insight into their causes and relationships. Pollen analyses in particular trace the dynamics of vegetation as well as temporal and spatial changes in its composition. They also indicate climate changes and human disturbance of the environment (Birks and Birks, 1980). In order to find potential regularities in pollen records, revealing dependencies among co-existent plant species through time, different machine learning methods were applied to pollen data from the southwestern Ljubljana Moor. Pollen records from two sites, Bistra and Hočevarica, were used in the analyses (Jeraj, 2004). The data consisted of the relative frequencies of the most common plant taxa/groups at different depths, which represent different ages in the Early and Mid Holocene period.

Results and discussion

Initially, pollen data from the southwestern Ljubljana Moor were analysed using different machine learning tools such as classification/regression trees, Naïve Bayes and the nearest neighbour method, as implemented in the WEKA toolkit (Witten and Frank, 2000). However, predictive models did not produce meaningful results, since they emphasized binary data and lower pollen values, and are as such not recommended for palaeoecological applications.

The LAGRANGE system, that performs equation discovery (Džeroski and Todorovski, 1995), was further applied to find relationships between plant species, as well as plant species and depth at Bistra. Some of the correlations found can be adequately interpreted from an ecological point of view. Higher positive correlation coefficients ($R>0.75$) were found for the pairs *Pteridophytes* (ferns)-*Pinus* (pine), *Quercus* (oak)-*Corylus* (hazel), *Aquatics* (aquatic plants)-*Carpinus* (hornbeam), *Cerealia* (cereals)-*Cyperaceae* (sedges), and higher negative correlations for the pairs *Quercus*-*Pinus* and *Cerealia*-depth. For example, cereals and sedges, which show the strongest positive correlation ($R = 0.85$), both indicate human appearance and activities in the area, largely associated with early agriculture and forest

clearing. A strong positive correlation of *Fagus* (beech) with depth (R = 0.72) was found, however, it was expected to be even stronger since beech appeared to be one of the first major species entering Holocene forests after the Late Glacial.

Hierarchical clustering, using a correlation-based distance between time series (Todorovski et al., 2002), was performed on pollen data from Bistra and Hočevarica. In the case of Bistra, the cluster dendrogram comprises four distinct clusters. The only homogenous pattern of occurrence was found among species from a cluster with *Quercus*, *Corylus*, *Alnus* (alder) and *Betula* (birch), which reached their maximum appearance in the middle of the observed period, i.e., in the Mid Holocene. Similarly, in the cluster dendrogram for Hočevarica, homogenous appearance of plants (*Cerealia*, *Chenopodiaceae* and *Poaceae*) was detected only in one of three clusters. In both dendrograms, the correlation-based distances between plant types range between 0.6 and 1.6 (the corresponding correlations range between 0.82 and –0.28). The lowest distances (between 0.6 and 1.0, correlations between 0.82 and 0.5) correspond to the pairs *Pinus*-ferns, *Poaceae* (grasses)-*Aquatics*, and *Corylus*-*Quercus*, as well as *Pinus*/ferns-*Cyperaceae*, *Corylus*/*Quercus* - *Alnus* and *Abies*-cereals for Bistra and to the pairs *Cerealia*-*Chenopodiaceae* (goosefoot), *Alnus*-*Corylus*, *Carpinus*-*Aquatics*, and *Pinus*-*Picea* (spruce) for Hočevarica. They suggest similar patterns and simultaneous changes of particular species over time to the ones revealed in the equations discovered by LAGRANGE.

Conclusions

In conclusion, machine learning techniques, such as equation discovery and hierarchical clustering, applied for pollen data from southwestern Ljubljana Moor, seem to be useful and successful to some extent. With both approaches, we were able to find similar relationships among particular plant types, which are known to grow in the same plant communities because of their similar tolerance to specific ecological factors. However, some of the previously known and observed correlations were not detected. Additional correlations among plants may be found if more data from nearby locations were analysed, and/or some other data mining approaches were used.

References

Bell, M., Walker, M. J. C. Late Quaternary Environmental Change, *Physical and Human Perspectives*, 273 p., London, 1992.

Birks, H.J.B., Birks, H.H. *Quaternary palaeoecology*. University Park Press, Baltimore. 289 p., 1980.

Džeroski, S., Todorovski, L. Discovering dynamics: from inductive logic programming to machine discovery. *Journal of Intelligent Information Systems* 4: 89-108, 1995.

Jeraj, M. *Archaeobotanical and palaeoecological reconstruction of the southwestern Ljubljana Moor, Slovenia.* Doctoral Dissertation, Nova Gorica, 136 p., 2004.

Todorovski L., Cestnik B., Kline M., Lavrac N., and Džeroski S. Qualitative clustering of short time series: A case study of firms reputation data. In *Proc. Fifth International Multi-Conference Information Society, Volume A*, pages 143-146. Jozef Stefan Institute, Ljubljana, Slovenia, 2002.

Witten, I., Frank, E., WEKA, Machine Learning Algorithms in Java. In: Witten I., Frank E. (eds.) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, 265-320, 2000.

Analysis and Spatial Modelling of Winter and Annual Habitats of the Red Deer (*Cervus Elaphus* L.) in the Dinaric Forests of South-Western Slovenia with Decision Trees in a Raster GIS Environment

Klemen Jerina*, Miha Adamič

University of Ljubljana, Biotehnical Faculty, Department of Forestry, Večna pot 83, 1000 Ljubljana, Slovenia

*Corresponding author. Tel.: +386-1-423-11-61 ext. 536; fax: +386-1-257-11-69.
E-mail address: klemen.jerina@uni-lj.si (K. Jerina)

Key words: Habitat, Model, Decision trees, Red deer, Geographic Information System

Introduction

Decision trees (Quinlan, 1986) were used to analyse winter and annual habitat selection of red deer (*Cervus elaphus* L.) in the Dinaric forests of south-western Slovenia and to elaborate the spatial habitat model for red deer for winter period and for the entire year.

The data on spatial distribution of red deer were collected through radio-telemetry. Fourteen adult hinds and ten adult stags were equipped with VHF transmitters and monitored from December 1997 to November 2002. Using the standard triangulation technique, their locations were recorded at least once per week, during the daytime. In total, more than 2,300 spatially and temporarily referenced locations were identified. In order to minimize the effect of noise caused either by random excursions of the animals or by imprecise location readings, the home range, based on pooled data for all traced animals, was used in habitat selection analyses instead of raw telemetry locations. The yearlong and winter (from mid-November to the end of March) total home ranges were computed using Biotas software which employs the fixed kernel method (Worton, 1989). For the purpose of habitat analyses, pixels inside the home range were considered positive, and pixels outside the range were treated as negative examples.

In the raster GIS environment (spatial resolution 200×200 meters), we elaborated 20 layers, presumed to be important for the red deer, which presented vegetation (share of forest, forest development phase, distance from the nearest forest edge), topography (altitude, aspect, slope), presence of large predators (density of wolf observations), infrastructure (distance to the nearest main/forest roads and settlements) and other environmental variables (distance to the nearest feeding place, intensity of solar radiation). It was assumed that the selected variables condition or reflect the food availability, availability of security, thermal and snow interception cover, thermal environment, human and predator activity, snow depth, water availability and expenditure, etc., and thus correspond well with the majority of known basic ecological variables which enable or limit the survival of and habitat selection of the red deer.

Simple bivariate comparisons between the use and availability of the analysed factors and multivariate, decision trees, classification analyses were performed to evaluate the winter and annual habitat selection of the red deer. For the induction of decision trees, we used the C.45 (Quinlan, 1993) system, incorporated in the J48.J48 module in Weka 3.2 software; the constructed trees were pre- and post-pruned. Half of the data were used for the training and the other half for estimating the accuracy of the decision tree. The training and testing datasets contained an equal number of positive and negative samples (points in or out of the home range).

Results and discussion

The annual habitat selection of red deer was most strongly influenced by the following variables: distance to the nearest main road, distance to the nearest feeding place, altitude, percentage of forest area and percentage of young stands. Monitored animals were almost absolutely forest dwelling. They also markedly avoided the vicinity of main roads (< 700 meters) and preferred to use the forest types offering good security cover and more food (young stands) in the areas close to the main roads. At an altitude of more than 1300 meters, the use of space started to decrease sharply. In wintertime two types of habitat were preferred: nearly pure conifer stands (share of conifers > 90 %), generally situated in the lower parts of the study area, or the areas near the feeding places (distance < 1400 meters).

If the estimated habitat selection patterns are interpreted in view of red deer energy balance, we can establish that they select a living space which allows minimization of unnecessary energy expenditure (e.g. avoidance of vicinity of main roads where disturbances and also flights would be frequent; avoidance of the highest parts of research area where the energy expenditure for thermo-regulation would increase; selection of security cover in proximity of main roads; selection of snow interception cover – conifer stands in winter period) and maximization of energy intake (selection of areas near feeding places in wintertime).

The results of the present study are also important for the management of the red deer and its environment. As it is, spatial distribution of feeding places strongly influences the distribution of red deer, so we can expect that in the proximity of feeding places negative influences (new forest growth browsing intensity, bark peeling) of red deer on forests are more pronounced than elsewhere; the presence of main roads or other forms of human induced disturbances evidently diminishes the area usable for red deer. Consequently, red deer tend to concentrate in smaller areas, which may trigger additional difficulties in natural forest regeneration and affect the development dynamics of the entire forest ecosystems. On the other hand, the spatial distribution of red deer and also the impact strength of red deer populations on forests can easily be manipulated by distributing the feeding places and by closing some of the less important roads.

The accuracy of the decision tree classifier built for the annual home range was 74 %, and 84 % for the winter home range. Induced decision trees were linked with GIS thematic layers for subsequent habitat/non-habitat classification of the entire research area and a map of winter and annual red deer habitat was elaborated.

Conclusions

Spatial distribution of red deer in the Dinaric forests of south-western Slovenia is most strongly influenced by the remoteness of main roads, feeding places, altitude and forest structure. Habitat selection patterns based on decision tree classifiers were rather hard to interpret; however, due to relatively high classification accuracy and the ability to reveal non-linear associations between variables, the method proved to be useful in this type of analysis.

Acknowledgments

The study, performed in cooperation with the Slovenian Forestry Service, is a part of the project »Conservation of endangered and other wildlife species V4-017597«. The authors are especially grateful to Anton Marinčič and Valentin Vidojevič for doing all the field work.

References

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1986), pp. 81–106.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning., Morgan Kaufmann, San Francisco, CA (1993).

Worton, B.J., 1989. Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70 1, pp. 164–168.

Using neural networks and GIS to predict the spatial occurrence of freshwater fish and decapods

Michael K Joy[1], Russell G Death[1],
[1]Massey University, Box 11-222, Palmerston North, New Zealand, M.K.Joy@massey.ac.nz,
Phone + 64 6 3505799 ext 7363, Fax 64 6 350 5623

Key words: artificial neural networks, diadromous fish, freshwater fish, GIS, New Zealand, presence/absence, prediction maps

Introduction

Species-environment models have been crucial in many fields of freshwater ecology including: species conservation, assessment of the impacts of flow regulation, biological quality assessment and the prediction of aquatic macrophyte and diatom distribution. Artificial neural networks (ANNs) offer a promising alternative to traditional statistical approaches for predictive modeling when non-linear patterns exist. ANNs also have advantages over traditional modeling methods because they are not dependent on particular functional relationships, need no assumptions regarding underlying data distributions and no *a-priori* understanding of variable relationships (Olden & Jackson 2001). Thus, this independence from assumptions makes ANNs a powerful option for exploring complex potentially non-linear relationships such as the associations between stream fish and their environment.

We used fish and decapod spatial occurrence data and associated these data with their corresponding geospatial landuse, geomorphologic, climatic, and spatial data in a GIS to model fish occurrence in the Wellington Region, New Zealand. To predict the occurrence of each species at a site from a common set of predictor variables we used a multi-response, artificial neural network, to produce a single model to predict the entire fish and decapod assemblage in one procedure. The predictive model was then extended to fill in the gaps between the surveyed sites using a GIS river network to give a spatial map of species probability of occurrence for the entire region.

Results and discussion

The species compositions of fish assemblages predicted by the neural network were very similar to the observed assemblages. There was on average an 88 % similarity between the observed and generated crossvalidated assemblages measured using the simple matching coefficient. Considering the species individually the ANN exhibited very high levels of correct classification. The least accurate predictions were for trout and redfin bullies at 69 and 72 % correctly classified, respectively. Sensitivity or the true positive prediction rate was moderate with an average of 54 %; the lowest value was 22 % for torrentfish and the highest was 85 % for the most common taxon, the longfin eel. The average true negative prediction rate (specificity) was higher at 86 %; the minimum was 41 % for the most common taxon, the longfin eel, while the highest (99 %) was for the equal rarest taxon the shortjaw kokopu. The average Cohen's kappa value for all the species was 0.44, the average kappa *z*-score was 8.8 and all species predictions were significant at the $P < 0.0001$ level. This measure of prediction success taking into account chance prediction was highest for the non-migratory

bullies and lowest for the longfin eel. The average area under curve value was also high at 0.78 and all of the taxa had values greater than 0.7.
The results of the assessment of the relative contribution of the predictor variables using connection weights showed that no particular variable type was more important than another with variables from geology, climate, landuse and longitudinal position all having high relative contributions. The highest-ranking variables were total annual catchment rainfall and site latitude. The rainfall variable was a product of catchment area and thus is a proxy for stream size. Stream order, average catchment slope, elevation, and air temperature were the next ranked variables. The next variables in the ranking were the first of the landuse variables with proportions of native, scrub and pastoral landuse. The variables related to longitudinal position, distance from the coast and elevation were also close in the ranking. The first of the geological variables appeared next in the ranking headed by greywacke, then loess and windblown material.
To illustrate the use of the model we displayed the probabilities of occurrence using redfin bully as an example. The map showed a mainly coastal distribution for this taxon but extending farther inland on the southern and western coasts where indigenous forest occurs close to the coast. The large central-eastern area with no or low probability of occurrence was the part of the region dominated by pastoral farming. Maps can be produced for all species using a range of prediction categories.

Conclusions

The model we have presented here revealed a high predictive power given the scale of the predictor variables and number of species being modeled. However, further research is needed to survey sites to validate the model predictions away from areas where input data was available for model construction. The extension of the model predictions using GIS data to the entire river network opened up a range of potential uses. Examples of potential uses include; monitoring and predicting temporal changes caused by human activities and shifts in climate, the elucidation of areas in need of protection, biodiversity hotspots, and areas for the reintroduction of endangered or rare species. From a management perspective, the model provides easily accessible species distribution information in a simple format that can be linked to species-specific biological information. Finally, this model conjugates the power of both GIS and ANNs to formalise the link between species and their habitat.
.

Acknowledgements

This work was supported by the sustainable management fund (New Zealand Ministry for the Environment grant No. 5099 “Nga Ika Waiora”: stream health evaluation tool) and the Wellington Regional Council and a Massey University PhD scholarship to MKJ.

References

Olden J. D. & Jackson D. A. (2001) Fish-habitat relationships in lakes: Gaining predictive and explanatory insight by using artificial neural networks. *Transactions of the American Fisheries Society*, 130, 878-897.

Habitat Mapping of an Ikonos Satellite Image Using Kernel-based Reclassification Enhanced with Machine Learning

Andrej Kobler[1], Sašo Džeroski[2], Iphigenia Keramitsoglou[3]
[1]Slovenian Forestry Institute, Večna pot 2, Ljubljana, Slovenia, andrej.kobler@gozdis.si
[2]Jozef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
[3] Department of Physics, University of Athens, Greece

Key words: habitat mapping, satellite images, kernel-based reclassification, machine learning, decision trees

Introduction

The conservation of biotopes is a high priority issue in the environmental policy of many countries, including the EU member states. Monitoring the natural environment and detecting changes in it is thus of increasing importance, as is habitat mapping which can be used for this purpose. A cost effective way to generate habitat maps is the classification of satellite images, which are of ever increasing spatial resolution.

Conventional methods for pixel-based classification were developed for images of low to medium spatial resolution. They attribute a classification label to one particular pixel without taking into account the vicinity of the pixel. Recently, as very high spatial resolution (VHSR) images have become available, the performance of these classification schemes has decreased. The reason for this is the fact that on a VHSR image an individual object (e.g. tree crown) is represented by multiple image pixels reflecting its uneven color or illumination, while the low- to middle resolution satellite sensors (e.g. Landsat TM with a 30 m resolution) provide a spectrally smoother image as several neighboring objects can contribute to a single image pixel. Therefore, when applying a pixel based classification scheme, the output will be rather noisy in the former case and quite homogeneous in the latter.

Kernel-based reclassification (Barnsley and Barr 1996) overcomes this problem with noise by considering the spatial variation of a pre-classified image. The pre-classification can be done using any pixel-based classification technique. The basic assumption of the algorithm is that individual habitat classes are characteristic mixtures of spectrally distinct classes. During the training stage average class adjacency frequencies of different spectrally distinct classes within a kernel are recorded in an adjacency event matrix (AEM) for each habitat class, based on ground truth data. During the reclassification stage the habitat class for a given pixel is determined by comparing its AEM to a set of class-related template AEMs and by identifying the class with the highest similarity (i.e. class membership) value. However, with this approach the classification result can be ambiguous where two or more classes have comparable similarity values.

In our contribution, we extend the original kernel-based reclassification method: The decision on reclassification takes into account the entire set of similarity values at each pixel, rather than the similarity value for the most similar class. The reclassification rules, that actually make the decision, are learned by the machine learning approach of decision tree induction. As a further improvement, the entire sets of similarity values for two pixel-based classifications (instead of a single classification) are taken into account. The decision trees are then used for reclassification instead of the original kernel-based reclassification. Comparing the two approaches, the results of the decision tree reclassification are better (in terms of classification accuracy/Kappa statistic estimated on an independent control sample).

Experiments

Kernel based reclassification has been used for urban environments before (Kontoes et al. 2000). In the present work, which is a part of the SPIN project (SPIN), an attempt was made to use this technique for habitat mapping using VHSR satellite imagery of natural ecosystems. The proposed extension of kernel-based reclassification with machine learning was applied to a VHSR Ikonos satellite image to perform habitat mapping. The habitat nomenclature used for the classification was the European Nature Information System (EUNIS), which was developed by the European Environment Agency (EEA) to facilitate a comprehensive pan-European and harmonized description and collection of data. The EUNIS nomenclature covers all types of habitats from natural to artificial, from terrestrial to freshwater and marine habitats types.

The 19,52 sq. km test site in SW Slovenia extends across a spectrum of vegetation successions – from the cultivated alluvial bottom of the Pivka valley across mostly abandoned karstic pastures and meadows towards the forested slopes of the Javorniki ridge. This site was selected to be representative of the landscape in the process of spontaneous afforestation, which is the main shaping factor of recent land cover change in most of rural Slovenia. The site also includes the dry lake-bed of the Palško intermittent karstic lake, one of several such lakes in the Postojna area. The classification was based on an Ikonos satellite image (panchromatic 1 m resolution + multispectral 4 m resolution) acquired on October 14, 2001. The reference map prepared for this classification contained 2.166 sample polygons covering 12 classes of the EUNIS nomenclature (levels 2 and 3). The reference map was created by delineating a sample of polygons using image segmentation, subsequently the EUNIS classes of the polygons were identified by a combination of field checks and stereoscopic photointerpretation of aerial imagery. Finally the polygons were rasterized and pixels were randomly subsampled with equal representation of classes.

Results

We used two per-pixel classification approaches, one unsupervised (ISODATA) and one supervised (MINDIST). We performed kernel-based reclassification with kernels of size 3, 5, 7, and 9 for each of these, as well as for a homogeneity image (HOMOGEN) enhanced using histogram equalization. The machine learning method of decision tree induction was then applied on the class membership values for MINDIST as well as ISODATA+HOMOGEN and appropriate reclassification was performed with the learned decision trees.

Taking in consideration the spatial context considerably improves the classification accuracy. Our results for an autumn Ikonos image of a semi-natural submediterranean landscape show that re-classification using decision trees (DT) can increase the classification accuracy in comparison to the kernel-based reclassification (KRC) approach. Looking at kernel 7 x 7, the least accurate is the KRC of the HOMOGEN image, followed by KRC of ISODATA image. The highest accuracy with a 7 x 7 kernel size is achieved by DT, which is simultaneously taking into account similarity values related to both HOMOGEN and ISODATA images.

References

Barnsley, M.J., and Barr, S.L., 1996. Inferring urban land use from satellite sensor images using kernel-based spatial reclassification, Photogrammetric Engineering and Remote Sensing, 62, 949-958.

Kontoes C. C., V. Raptis, M. Lautner and R. Oberstadler, 2000. The potential of kernel classification techniques for land use mapping in urban areas using 5m-spatial resolution IRS-1C imagery, International Journal of Remote Sensing, 21: 3145-3151.

SPIN. A research project supported by the European Commission under the Fifth Framework Programme and contributing to the implementation of the Key Action "Development of generic earth observation satellite technologies" of Energy, Environment and Sustainable Development. Contract n°: EVG 1-CT-2000-019; www.spin-project.org

A New Synergetic Paradigm in Environmental Numerical Modeling: Hybrid Models Combining Deterministic and Machine Learning Components

Vladimir M. Krasnopolsky and Michael S. Fox-Rabinovitz

Earth System Science Interdisciplinary Center, University of Maryland, MD, USA
vladimir.krasnopolsky@noaa.gov , (301)-763-8000 ext. 7262

Key words: environmental numerical modeling, machine learning, neural networks

Tremendous developments in numerical modeling and in computing capabilities during the last decades contributed dramatically to scientific and practical significance of interdisciplinary ecological numerical modeling. One of the main problems of development and implementation of these high-quality high-resolution environmental numerical models (ENMs) is the complexity of physical, chemical and biological processes involved. Parameterizations of model physics (chemistry, etc.) are approximate, adjusted to model resolution and computer resources, schemes based on simplified physical process equations and empirical data. Still the parameterizations are so time-consuming, even for most powerful modern supercomputers, that some of them have to be calculated less frequently than model dynamics. This may negatively affect the accuracy of model physics calculations and its temporal consistency with model dynamics that may lead to a significant reduction of the accuracy of ecological simulations and predictions. For example, calculation of a model physics package in a typical moderate (a few degrees) resolution climate GCM (General Circulation Model) like NCAR (National Center for Atmospheric Research) CAM (Community Atmospheric Model) takes about 70% of the total model computations. Higher uniform and variable model resolutions (e.g. Fox-Rabinovitz et al., 2002) and more frequent model physics calculations, desirable for temporal consistency with model dynamics, would increase the percentage to more than 90%.

Such a situation is an important motivation for looking for alternative, faster, and most importantly, very accurate ways of calculating model physics, chemistry and biology. The approach discussed in this talk introduces a new paradigm which is based on a synergetic combination of deterministic numerical modeling with machine learning techniques for approximating atmospheric physics and chemistry processes. A traditional statistical technique based on representation of input/output relationship as an expansion of hierarchical correlated functions has been successfully used in some atmospheric chemistry applications (see Schoendorf et al., (2003) and references there). However, current GCMs require much higher accuracy of approximation and a better flexibility of approximation methods than those provided by traditional techniques.

During the last decade a new emerging approach based on machine learning neural network (NN) approximations has found the variety of applications in different fields and, more specifically, for accurate and fast approximation of model physics processes (Krasnopolsky and Chevallier, 2001, 2003), and for atmospheric radiative processes and satellite retrieval procedures (Krasnopolsky and H. Schiller, 2003). Recently, the NN approach has been successfully applied for developing a fast (8 times

faster than the original parameterization) and accurate long-wave (LW) radiation parameterization for the ECMWF (European Centre for Medium-range Weather Forecasting) model (Chevallier et al., 2000). This LW radiation parameterization is being used, operationally within the ECMWF 4-DVAR data assimilation system, since October 2003. The NN approach has been also used by the authors for approximations of model physics in ocean and atmospheric numerical models (Krasnopolsky et al. 2002, 2004) where acceleration of calculation from 10 to 10^5 times has been achieved, as compared with original parameterizations.

In this talk, we discuss a conceptual and practical possibility of an efficient synergy or an optimal combination of the traditional deterministic modeling like that of GCMs and statistical learning techniques for accurate and fast approximation of atmospheric model physics. We introduce a new type of ENMs, hybrid environmental numerical models (HENMs) based on combining deterministic modeling and machine learning components. Using several climate and oceanic HENMs, developed by the authors, we discuss in our talk the key questions of the new approach: (i) are developed HENMs close enough to the original ENMs so that their use (instead of the original ENM) allows us to preserve all richness/complete integrity and all the detailed features of environmental processes, (ii) are HENMS fast enough to significantly accelerate model calculations, (iii) are these statistical/machine learning techniques able to successfully coexist with deterministic components of HENMs so that their combination can be efficiently used for accurate and fast climate simulations without any negative impacts on their quality, and (iv) is there a real/productive synergy here, in other words, does this new combination of deterministic and statistical learning approaches lead to new opportunities in environmental simulation and weather prediction.

References

Chevallier, F., J.-J. Morcrette, F. Chéruy, and N. A. Scott, "Use of a neural-network-based longwave radiative transfer scheme in the EMCWF atmospheric model", *Quarterly Journal of Royal Meteorological Society*, *126*, 761-776, 2000

Fox-Rabinovitz, M. S., L. L. Takacs, and R. C. Govindaraju, "A variable-resolution stretched-grid general circulation model and data assimilation system with multiple areas of interest: Studying the anomalous regional climate events of 1998", *J. Geophys. Res*., 107(D24), 4768, doi:10.1029/2002JD002177, 2002.

Krasnopolsky, V.M., and F. Chevallier, Some neural network applications in environmental sciences. *ECMWF Technical Memorandum No. 359*, 2001

Krasnopolsky, V.M., and F. Chevallier, "Some neural network applications in environmental sciences. Part II: Advancing computational efficiency of environmental numerical models", *Neural Networks*, *16*, 335-348, 2003

Krasnopolsky, V.M. and H. Schiller, "Some neural network applications in environmental sciences. Part I: Forward and inverse problems in satellite remote sensing", *Neural Networks*, *16*, 321-334, 2003

Krasnopolsky, V.M., D.V. Chalikov, and H.L. Tolman, "A neural network technique to improve computational efficiency of numerical oceanic models", *Ocean Modelling*, *4*, 363-383, 2002

Krasnopolsky, V.M., M.S. Fox-Rabinovitz, and D.V. Chalikov, "Using neural networks for fast and accurate approximation of the long wave radiation parameterization in the NCAR community atmospheric model: evaluation of computational performance and accuracy of approximation", 84th AMS Annual Meeting, Seattle, Washington, 11-15 January 2004, *Proceedings, 15th Symposium on Global Change and Climate Variations,* P1.20, 2004

Schoendorf, J., H. Rabitz, and G. Li, A fast and accurate operational model of ionospheric electron density, *Geophys. Res. Lett., 30*, 1492-1495, 2003

CARROTAGE, a software for mining land-use data

F. Le Ber[1,2], J.-F. Mari[2], M. Benoît[3], C. Mignolet[3], and C. Schott[3]

[1] ENGEES, 1 quai Koch, BP 1039, F-67070 Strasbourg CEDEX
fleber@engees.u-strasbg.fr, phone: 33 388248230, fax: 33 388248284
[2] UMR 7503 LORIA, BP 239, F-54506 Vandœuvre-lès-Nancy CEDEX
jfmari@loria.fr
[3] INRA SAD, Domaine du Joly, 662 avenue Louis Buffet, F-88500 Mirecourt
{benoit,mignolet,schott}@mirecourt.inra.fr

Keywords: Data Mining, Markov Models, Land Use, Cropping Patterns, Crop Rotation.

Introduction

Mining sequential patterns is an active area of research in artificial intelligence. One basic problem in analyzing a sequence of items is to find frequent episodes, i.e. collections of events occurring frequently together. We rely on new numerical algorithms, based on high-order stochastic models – the second-order hidden Markov models (*HMM2*) – that were initially specified for speech recognition purposes (Mari *et al.*, 1997). We have shown that, with minor changes, they can extract spatial and temporal regularities that can be explained by human experts and may constitute elements of a *knowledge discovery process* (Mari *et al.*, 2002).

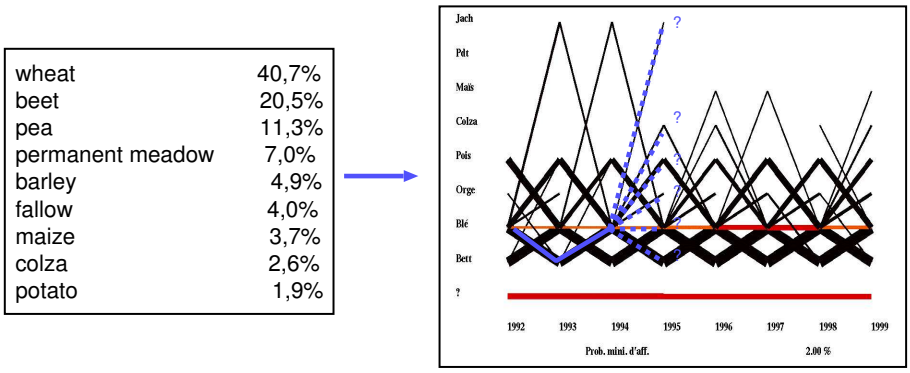
The *HMM2*'s are based on the probabilities and statistics theories. They are implemented with an unsupervised training algorithm, the EM algorithm (Dempster *et al.*, 1977), that allows to estimate a model parameters from a corpus of observations and an initial model. The resulting model is capable to segment each sequence in stationary and transient parts and to build up a classification of the data together with the *a posteriori* probability of this classification. This characteristic makes the *HMM2*'s appropriate to discover temporal and spatial regularities. For all these reasons the *HMM2*'s have been chosen as a basis for the land-use data mining software CARROTAGE (`http://www.loria.fr/~jfmari/App/`).

Results and Discussion

CARROTAGE has been developed for studying the cropping patterns of an agricultural territory. It uses therefore a french agricultural database, named *Ter-Uti*, which records every year the land-use category of about 500,000 sites regularly spaced on french territory (Ledoux & Thomas, 1992). One *Ter-Uti* site represents roughly 100 hectares. The collected land-use categories (wheat, corn, potato, forest, ...) are logged in a matrix in which the rows are the sites and the columns the time slots. CARROTAGE takes this matrix as an input and builds a partition together with its *a posteriori* probability.

We work within an interdisciplinary research program which aims to develop a tool for forecasting water quality in the Seine river watershed, based on assumptions upon agricultural changes. Thus, we analyse the agricultural activities in the watershed, their dynamics and their spatial organisations, focusing on the crop (temporal) rotations and the associated agricultural practices that are able to explain a part of the risk of nitrate loss (Mignolet *et al.*, 2004). The

data mining software CARROTAGE has been used on *Ter-Uti* data from the Seine watershed. As shown in the figure below for a small agricultural district from the north-east of France, CARROTAGE allows to compute the crop distribution in a given periode (here from 1992 to 1999, left), and to view the annual transitions between crops (right).



These results are analysed, with respect to the main transitions between crops. For example the dashed lines represent the possible transitions between the triple *wheat-beet-wheat* and the other crops: beet, pea, wheat, barley, colza or fallow. The same analysis has been done for all small agricultural districts in the Seine watershed. The districts are then clustered according to their main crop rotations (and their evolutions), for modelling water pollution risks wrt spatial characteristics of the Seine watershed.

Conclusion

The *HMM2*'s have proven to be appropriate to discover temporal and spatial regularities. Furthermore, the CARROTAGE software, based on *HMM2*'s, has proven to be useful for analysing spatial and temporal cropping patterns. The models and visualisation tools have been designed within a collaboration between agronomists and computer scientists. CARROTAGE has been used successfully for studying the link between nitrate contamination of groundwater and surface water in the Seine watershed and the evolution of agricultural activities. It has also been used for helping satellite images interpretation, whith an aim of irrigation needs forecast. Other applications could be developed, regarding ecological or environmental spatio-temporal data.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum-Likelihood From Incomplete Data Via The EM Algorithm. *Journal of Royal Statistic Society, Ser. B*, 39:1 – 38, 1977.

Ledoux, M., and Thomas, S. De la photographie aérienne à la production de blé. *AGRESTE, la statistique agricole*, (5), 1992.

Mari, J.-F., Haton, J.-P., and Kriouile, A. Automatic Word Recognition Based on Second-Order Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 5:22–25, 1997.

Mari, J.-F., Le Ber, F., and Benoît, M. Segmentation temporelle et spatiale de données agricoles. *Revue Internationale de Géomatique*, 12(4):439–460, 2002.

Mignolet, C., Schott, C., and Benoît, M. Spatial dynamics of agricultural practices on a basin territory: a retrospective study to implement models simulating nitrate flow. The case of the Seine basin. *Agronomie*, 24, 2004. To be published.

Learning to Predict Channel Stability using Biogeomorphic Features

Stephanie L. Moret
Louisiana State University
School of the Coast and Environment
Department of Environmental Studies
Baton Rouge, LA 70803

William T. Langford
National Center for Ecological Analysis and Synthesis
University of California at Santa Barbara
735 State Street
Santa Barbara, CA 93101-3351

Dragos D. Margineantu*
The Boeing Company
Mathematics & Computing Technology, Adaptive Systems
P.O. Box 3707, M/S 7L-66
Seattle, WA 98124-2207

Current human land use activities are altering many components of the river landscape, resulting in unstable channels. The instability of river channels may have serious negative consequences for water quality, aquatic, and riparian habitat, and for river-related human infrastructure such as bridges and roads.

Land use practices such as farming, grazing, forestry, urban development, dams, and mining increase flow velocity and decrease sediment storage. Increased velocity scours the channel bed which in turn causes riparian degradation. Maintaining a healthy riparian zone is important because the riparian zone is responsible for filtering recharge water, providing food for aquatic organisms, providing wildlife habitat, providing large woody debris, cooling stream temperature, adding roughness (velocity resistance), and stabilizing banks. Erosion caused by channel instability also produces sediment, which contributes to the decline of salmon and other aquatic organisms. In addition to compromising water quality and causing aquatic and riparian habitat damage, channel instability can undermine bridge supports and expose pipelines or other structures buried within the riverbed and can cause road and trail damage as well as instigate or increase the intensity and frequency of mass wasting and flooding, resulting in large-scale events such bank failures and landslides.

In order to restore equilibrium for an adversely impacted system and have naturally functioning channel stability, it is important to be able to identify river locations where the channel is likely to be unstable. Resource management agencies such as the U.S. Forest Service have developed rapid bioassessment surveys to help do this in a fast and cost-effective way. One such method is the Stream Reach Inventory and Channel Stability Evaluation (SRICSE) which has been used in over 60% of the national forests in the United States. In the SRICSE, a series of inventory items about a location are assessed using maps, field observations, and field measurements. Channel stability is then evaluated by assigning a score from a rating sheet to various channel stability attributes. A higher numerical score corresponds to a poorer rating. While this assessment can be done for a single site fairly rapidly, it is still time-consuming and expensive to apply over large watersheds. Since public agencies are short on staff and funding, channel assessment activities must be prioritized.

We have employed map data that are commonly available as input for our learning algorithms for predicting the relative channel stability of different locations in a watershed. The output was then used to prioritize where the efforts of resource personnel are most needed to increase safety, decrease cost, and improve habitat. This research is based on a survey of fifty-eight third and fourth order rivers at high elevations in the Upper Colorado

* ***Corresponding author***: **Dragos D. Margineantu**, The Boeing Company, M&CT, P.O. Box 3707, M/S 7L-66, Seattle, WA 98124-2207. E-mail: dragos.d.margineantu@boeing.com, tel: 425-957-5057, fax: 425-865-2964.

River Basin - in Colorado, USA. Mapped information, taken from both paper maps and a Geographic Information Survey (GIS), was used to characterize site-specific information.

The data we collected and used in the experiments reported here has eleven attributes: seven discrete and four continuous, describing features such as sinuosity, topographic gradient, elevation, land use and land cover, and geology. In the original data, each example had a stability factor (a real-value between 55 and 117.5) associated with it, representing the output variable. Higher values of the stability factor mean higher instability of the corresponding channel. We have transformed the problem from its original regression format (the targets, represented by the stability factors, were real valued numbers) into a two-class classification problem in which the classes are *stable* and *unstable*.

The original stability values were computed using a table-based protocol (also called the Rapid Assessment Protocol, or RAP) provided by the US Department of Agriculture (USDA). The scores in the table were calculated based on expert observations in the Rocky Mountain region, but exact ("true") threshold θ_0 between stable and unstable stability factor values is unknown. To handle this situation, we have defined different classification problems by assigning the stable-unstable threshold θ different values from the [80,95] interval. These threshold values were suggested to us by the environmental sciences researchers studying this problem.

We employed two learning algorithms that have been proven to compute good class probability estimates: Bagged Lazy Option Trees (B-LOTs), described in [4], and Bagged Probability Estimation Trees (B-PETs) [5].

We first tested the algorithms using leave-one-out cross-validation [3], because of the small size of the available data sample. We have measured the area under the ROC curve (AUC) [2] for the two algorithms, for different values of the θ parameter. The B-LOTs had better overall performance than the B-PETs for all values of θ except for the [90,93] range.

Next, we did a cost-sensitive analysis based on the class probability estimates computed by the two algorithms. Given that the most costly mistake is to classify unstable channels as stable (a value in the order of millions - representing the potential losses of human lives and property that can be caused by an unstable channel that was not remediated), we analyzed the number of misclassified stables when all unstables were classified correctly. The results show that if the decision threshold is set optimally, the B-PETs give lower costs when $\theta < 90$ and the B-LOTs are better for larger values of θ. It is important to observe that larger values of θ correspond to fewer unstable examples in the data, creating a class imbalance. This suggests that lazy learning was able to handle larger class imbalance better than the B-PETs by focusing on the individual examples.

References

[1] Bonneau P. R., and Snow R. S., Character Headwaters Adjustment to Base Level Drop, Investigated by Digital Modeling, *Geomorphology*, 5, pp.475-487, 1992.

[2] Bradley A. P., The use of the Area under the ROC curve in the Evaluation of Machine Learning Algorithms, Pattern Recognition, 30, pp.1145-1159, Elsevier Science, 1997.

[3] Kearns M., Ron D., Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation, *Neural Computation*, 11:6, pp.1427-1453, 1999.

[4] Margineantu D. D., Dietterich T. G., Improved Class Probability Estimates from Decision Tree Models, in *Nonlinear Estimation and Classification*, D.D. Denison, C.C. Holmes, M.H. Hansen, B. Mallick, and B. Yu (Eds.), Springer Verlag, Lecture Notes in Statistics 171, 2002.

[5] Provost F., Domingos P., Tree Induction for Probability-based Rankings, *Machine Learning*, 52:3, 2002.

ARTIFICIAL NEURAL NETWORKS AND TIME SERIES MODELLING OF A FORESTED WATERSHED

Mohamed H. Nour[1], Daniel W. Smith[1], Mohamed Gamal El-Din[1], and Ellie E. Prepas[2]

[1] Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2M8, E-mail: mnour@ualberta.ca, Phone: 1(780)492-3441, Fax: 1(780)492-8289

[2] Faculty of Forestry and the Forest Environment, Lakehead University, Thunderbay, ON, Canada P7B 5E1, and Department of Biological Sciences, Biological Sciences Building, University of Alberta, Edmonton, AB, Canada T6G 2E1

Key words: Watershed, modelling, flow, phosphorus, boreal, forest, artificial neural networks, time series, multi-slab hidden layer, hystereses

Abstract

Historically, management strategies in Canada's boreal forest have focused on forest polygons and terrestrial biodiversity to address ecological considerations in forest management. However, efficient management strategies in Canada's boreal forest must consider ecological processes from a watershed perspective. The Boreal Plain ecozone of the Canadian boreal forest is exemplified by low topographic relief, alkaline, phosphorus-rich soils developed from sedimentary bedrock. Thus, soil sediment during snowmelt and rain events, when soil is more susceptible to erosion, represents the highest threat with respect to phosphorus migration to water bodies. Particulate phosphorus, being the dominant phosphorus form in the region, contributes largely to nutrient enrichment of receiving streams. The area is currently experiencing both natural (mainly wildfires) and anthropogenic (primarily forest harvesting activities) watershed disturbance. The associated accelerated rate of watershed disturbance threatens to destabilize aquatic ecosystems of the region as a result of possible dissolved oxygen depletion, cyanobacteria growth and toxin production. Hence, a science-based decision support tool that can predict the impact of future forest activities on water quantity and quality, particularly phosphorus concentration in receiving streams, is critical to sustainable development of a forested ecosystem.

This study represents a building block of the required decision support tool. It examines the applicability of using artificial neural network (ANN) in modelling stream flow and water-phase phosphorus concentration from a watershed perspective. A three-layer feed forward multi layer perceptron ANN trained with error back propagation algorithm was used to model both the flow and the daily change in the total phosphorus concentration (ΔTP) of a small forested watershed, Northern Alberta, Canada. Flow was first modelled using Environment Canada Whitecourt airport weather station data (rainfall, snowfall, and mean, minimum, and maximum daily air temperatures). ΔTP was then modelled utilizing the produced modelled flow time series and the Environment Canada Whitecourt airport weather station data. The capabilities of time series analysis in quantifying autocorrelation and cross-correlation statistics was utilized to identify possible model lagged inputs prior to model development

addressing the time dependency of the modelled variables. The concept of multi-slab hidden layer was adopted, in which the hidden layer was divided into three slabs with different activation functions (sigmoid, Gaussian, and Gaussian complement), to simulate the conceptual differences between base flow, snow melt, and storm event behaviours. Spectral analysis was also used to identify the dominant frequency (ν) in the studied time series. Two additional inputs namely, $\sin(2\pi\nu t)$ and $\cos(2\pi\nu t)$ were then added to deal with the hystereses effect in the modelled processes.

The present study highlighted the capabilities of ANN in modelling such complex ecosystem and provided a key block in the targeted decision support tool. Results showed that ANN was successful in modelling both time series achieving a coefficient of multiple determination (R^2) of about 0.8 and 0.75 for flow and ΔTP, respectively. It appears that introducing the periodicity components in the input layer of both flow and ΔTP models has helped in addressing data hystereses and has significantly increased the forecasting accuracy.

Using Regression Trees to Estimate Surface Water Runoff and Soil Erosion for Rangelands

Yakov Pachepsky[1], Frederick Pierson[2], Kenneth E. Spaeth[3], and Mark Weltz[3]

[1]USDA-ARS Environmental Microbial Safety Laboratory, Beltsville, MD, 20705, ypachepsky@anri.barc.usda.gov
[2]USDA-ARS Northwest Watershed Research Center, Boise, Idaho, 83712 USA, fpierson@nwrc.ars.usda.gov
NW Watershed Research Center, 800 Park Blvd., Suite 105, Boise, ID 83712, USA
kspaeth@nwrc.ars.usda.gov
[3]USDA-ARS National Program Staff, Beltsville, MD, 20705, USA, maw@ars.usda.gov

Key words: regression tree, surface runoff, rangelands, soil erosion

Introduction

Estimates of surface runoff and soil erosion are needed for both rangeland management and rangeland productivity evaluation. Such estimates are made using various models ranging from purely empirical to mechanistic. Mechanistic runoff/erosion models require parameters that are difficult to measure and impractical to obtain for large scale projects. Estimating such parameters from more easily obtainable data introduces an error. Our hypothesis was that the direct use of more easily obtainable data in machine-learning predictive tools can be a viable option.

Results and discussion

The database comprised data from 442 rainfall event simulations at 26 rangeland locations in the Western United States. Total of 81 variables were measured at each experimental plot. They included parameters of rainfall, runoff, infiltration and sediment amount, ground cover, above ground and below ground biomass for main types of vegetation, basic soil properties, and random roughness. All of the plots were set at about 6% slopes and had dimensions of 10 m x 3 m. The rainfall simulations have been done at each plot at three initial soil water content distributions labeled as “dry”, wet”, and “very wet”. Regression trees were developed with SPLUS software. The decrease in deviance after splits was used to select grouping variables in the course of binary partitioning of the database. The jackknife cross-validation was used to trim the trees. Regression tress were developed for runoff:rainfall ratio and for sediment:runoff ratio (a) with input variables that could be obtained from public sources and (b) with input variables that had to be measured on-site. The public access data included parameters of rainfall, soil texture, organic matter content. The on-site information included the public access data, surface roughness, ground cover, and plant biomass. Ten terminal nodes provided the root-mean-squared error in the jackknife test datasets that did not increase with further increase in the terminal node number. About 60 % of the variability in runoff could be explained using soil basic properties whereas the stepwise linear regression explained about 48% of the variability. The top splitting variables in regression trees were rain intensity, organic matter content for low rain intensities and silt content for high rain intensities. Further grouping occurred with total rainfall, clay content, and coarse fragment

content as splitting variables. Adding on-site measured values of ground cover and surface roughness in the list of predictors helped to explain about 70% of the variability whereas the stepwise regression explained just 55% of the variability. The top splitting variables in regression trees were rain intensity, organic matter content for low rain intensities and surface random roughness for high intensities. Further grouping occurred with total rainfall, clay content, and coarse fragment content as splitting variables. Using details of plant biomass distributions did not improve the accuracy of the trees developed to estimate runoff:rainfall ratio. Predictions of the sediment yield were slightly worse as only 60% of variation could be explained. Overall, the accuracy of regression trees was the same or better than that of available mechanistic runoff/erosion models with parameters estimated from basic soil and vegetation properties.

Conclusions

Regression trees were easy to interpret, and grouping had a clear physical meaning in most cases. Outliers were quite easily identified. Regression trees appeared to be an extremely efficient technique of its ability to discover relationship structures specific for subsets of the whole database, transparency of results, and ability to select the most influential input variables.

Neural network modelling for the analysis of forcings/temperatures relationships at different scales in the climate system

Antonello Pasini[1], Massimo Lorè[2], and Fabrizio Ameli[3]

[1]CNR - Institute of Atmospheric Pollution, via Salaria km 29.300
I-00016 Monterotondo Stazione (Rome), Italy
Email: pasini@iia.cnr.it, Phone: +39 06 90672274, Fax: +39 06 90672660
[2]CNR - Institute of Atmospheric Pollution, via Salaria km 29.300
I-00016 Monterotondo Stazione (Rome), Italy, Email: lore@iia.cnr.it
[3]INFN - National Institute of Nuclear Physics, P.le Aldo Moro 2, I-00185 Rome, Italy
Email: fabrizio.ameli@roma1.infn.it

Key words: neural networks, climate change, forcings, circulation patterns

Introduction

It is well known that Atmosphere-Ocean General Circulation Models (AOGCMs) are the standard tools for grasping the complexity of climate system and simulating its behaviour, in the past as well as in future scenarios. In particular, they allow us to reconstruct and forecast the climate at large scale. Nevertheless, this dynamical approach also shows some limits in the simulations at regional and local scales and, at a more fundamental stage, in the balance among the relative strength of feedbacks and the various parametrisation routines.
In this framework, a non-dynamical approach, which is able to catch non-linear relationships among several variables in the climate system, can be useful in order to "weight" the magnitude of different causes on a single effect (like the temperature variations) and to assess the relative importance of global forcings and regional circulation patterns, even on regional mean variables.

Results and discussion

Here we consider a multi-layer perceptron and apply it to the reconstruction of observed annual and seasonal values of temperature at global and regional scales (targets), starting from data of variables (inputs) which have well-known influence on them. In the case studies presented here, the available target data are: annual global temperature anomalies and a series of Central England Temperatures (CET) during extended winters (December to March) in the last 140 years. The available input data are: solar irradiance anomalies, stratospheric aerosol optical thickness, global concentration of CO_2, sulfate emissions at global level, Southern Oscillation Index (SOI) related to El Niño Southern Oscillation (ENSO) and a monthly North Atlantic Oscillation (NAO) index.
The neural model used in this investigation has been previously developed and applied to prognostic problems in the boundary layer (Pasini and Ameli, 2003; Pasini et al., 2001). Recently, also a preliminary attempt at analysing climatic data has been performed (Pasini et al., 2003).
The neural networks considered in this work are feed forward and endowed with a backpropagation training rule with both gradient descent and momentum terms; a usual quadratic cost function is chosen. A particular attention is paid to the form of transfer

functions: we choose sigmoids whose arguments are normalised with respect to the number of connections converging to a single neuron of the input and hidden layer, respectively, as in (Pasini et al., 2001), where this choice has shown its validity in order to avoid problems in cases of networks with many neurons. An early stopping method is also used to furtherly prevent overfitting. Finally, a particular attention is devoted to the learning procedure in cases of our "historical" data. For more details on this model, see (Pasini and Potestà, 1995; Pasini and Ameli, 2003) and, for a modified version, (Pasini et al., 2001).
In the first case study we consider solar irradiance and stratospheric optical thickness as indices of natural forcings, CO_2 concentration and sulfate emissions as anthropogenic ones. We estimate their influence on annual global temperature during the last 140 years by means of neural modelling. Our results show that we are not able to correctly reconstruct the temperature trend if we consider only the inputs related to natural forcings; on the other hand, anthropogenic forcings are necessary to estimate the correct temperature behaviour. Furthermore, the introduction of an input related to ENSO allows us to better catch the inter-annual variability of global temperature.
In the second case study we analyse the influence of the NAO circulation pattern on regional winter temperatures in Europe. It is well known that NAO index correlates linearly quite well with winter temperatures in Europe: now we show that a fully non-linear neural network is able to catch non-linearities hidden in these data and performs better in order to reconstruct the temperature behaviour at these scales. Furthermore, we compare the "weights" of NAO and global forcings on regional winter mean temperatures and find that the driving force of temperature time pattern in central England is just the NAO behaviour.

Conclusions

In conclusion, this neural approach allows us to obtain simple assessments about the magnitude of different causes on the same effect (here, temperature change) in a complex environment like the climate system: in particular, we are able to reconstruct the global temperature behaviour only if we take anthropogenic forcings into account. Furthermore, the recognition of the influences of global forcings and regional circulation patterns (ENSO and NAO) at global and regional scales suggests to use these results in order to identify the fundamental elements to be considered for overcoming some limitations of AOGCMs, e.g. allowing us to achieve a successful downscaling of AOGCMs, not only on climate reconstructions in the past, but even on future scenarios.

References

Pasini, A., and Ameli, F., Radon short range forecasting through time series preprocessing and neural network modeling. *Geophys. Res. Lett.*, 30(7): 1386, doi:10.1029/2002GL016726, 2003.

Pasini, A., Ameli, F., Lorè, M., Pelino, V., and Žujić, A., Application of a neural network model to the analysis of climatic observations at global, regional and local scales. *Proceedings of the First Italian IGBP Conference*, F. Miglietta and R. Valentini eds., pages 185-187, 2003.

Pasini, A., Pelino, V., and Potestà, S., A neural network model for visibility nowcasting from surface observations: Results and sensitivity to physical input variables. *J. Geophys. Res.*, 106(D14):14951-14959, 2001.

Pasini, A., and Potestà, S., Short-range visibility forecast by means of neural-network modelling: a case study. *Nuovo Cimento*, 18C(5):505-516, 1995.

Radon forecasting in the low atmosphere by neural network modelling and estimation of the stable layer depth

Antonello Pasini[1], Fabrizio Ameli[2], and Massimo Lorè[3]

[1]CNR - Institute of Atmospheric Pollution, via Salaria km 29.300
I-00016 Monterotondo Stazione (Rome), Italy
Email: pasini@iia.cnr.it, Phone: +39 06 90672274, Fax: +39 06 90672660
[2]INFN - National Institute of Nuclear Physics, P.le Aldo Moro 2, I-00185 Rome, Italy
Email: fabrizio.ameli@roma1.infn.it
[3]CNR - Institute of Atmospheric Pollution, via Salaria km 29.300
I-00016 Monterotondo Stazione (Rome), Italy, Email: lore@iia.cnr.it

Key words: neural networks, boundary layer, natural radioactivity, short-range forecasting, box model

Introduction

Physical characterisation and forecasting in the atmospheric environment is a very difficult task, if we adopt an approach with complete dynamics. This is even more critical if we deal with the physics of the Planetary Boundary Layer (PBL), that is the lowest layer of the atmosphere, where many non-linear phenomena interact and the complexity of the system has been easily recognised. In this framework a non-dynamical approach, which is able to catch non-linear relationships between several variables in the system, can be useful in order to characterise it or to model its evolution in the near future.
In this paper we identify a physical quantity (radon concentration) and an index (stable layer depth) which allow us to summarise our knowledge of the PBL state. Then we forecast the future values of this index by neural network modelling + application of a box model.

Results and discussion

The relevance of radon detection for the characterisation of PBL properties was recognised since the late 70s, while, more recently, the link between the radon concentration and the depth of the stable layer in nocturnal (stable) situations was clearly recognised (see Allegrini et al., 1994, and references therein). This led to the use of a simple box model in order to estimate this depth, which allows us to calculate the volume available for the dispersion of primary pollutants in the low layers of the atmosphere.
In what follows, we present the application of a neural network model to the short range forecast of radon concentration. This leads us to know more on the future state of the PBL, as far as its dispersion properties are concerned. Furthermore, in nocturnal stable situations, these predicted values can be used in the cited box model and allow us to obtain reliable estimations of future values of the stable layer depth. In doing so, we stress the peculiarities of both our neural networks and pre-processing method.
The neural networks considered in this work are multi-layer perceptrons endowed with a backpropagation training rule with both gradient descent and momentum terms; a usual quadratic cost function is chosen. A major attention is paid to the form of transfer functions: we choose sigmoids whose arguments are normalised with respect to the number of

connections converging to a single neuron of the input and hidden layer, respectively, as in (Pasini et al., 2001), where this choice has shown its validity in order to avoid problems in cases of networks with many neurons. An early stopping method is also used to furtherly prevent overfitting. Finally, a moving window training, extensively discussed in (Pasini et al., 2001), is used for our "historical" data. For more details on this model, see (Pasini and Potestà, 1995; Pasini and Ameli, 2003) and, for a modified version, (Pasini et al., 2001).
In previous papers (Pasini and Ameli, 2003; Pasini et al., 2003), some approaches to this prediction problem were explored: first of all a standard time series approach from radon data only, then attempts at forecasting radon concentration starting from several meteorological data at a certain time t_0 (a so called synchronous pattern approach), finally the application of a pre-processing method to the radon time series in order to extract the known periodicity linked to the day-night cycle and to leave to the network the forecasting activity on the hidden dynamics in a residual series of radon data. This last approach showed the best results. Here, after the presentation of an improved version of the box model and an accurate description of our neural model, we stress the importance of a pre-processing activity and present new results, showing that meteorological conditions lead to improvements in catching non-linearities when periodic contributions to the PBL dynamics are subtracted by pre-processing in the time series of each variable. The final results allow us to improve the forecast performance of the neural model on the residuals, giving rise to improved final forecasts for radon concentration as well as for values of the stable layer depth.

Conclusions

In conclusion, a neural network approach to a problem of short range forecasting in the atmosphere has shown its potentialities in catching complex non-linear relationships in the PBL, especially when the driving and overwhelming day-night cycle is "detrended" and the training of the networks is performed on residual series of radon and meteorological data. Even if these residuals look like noise, our neural model is able to recognise a hidden dynamics therein. Furthermore, the results obtained by the joint application of pre-processing and neural model (+ box model in nocturnal stable cases) are significant and very relevant for air pollution forecasts, because they permit to characterise the future physical state of the PBL at short range with a high time resolution.

References

Allegrini, A., Febo, A., Pasini, A., and Schiarini, S., Monitoring of the nocturnal mixed layer by means of particulate radon progeny measurement. *J. Geophys. Res.*, 99(D9):18765-18777, 1994.
Pasini, A., and Ameli, F., Radon short range forecasting through time series preprocessing and neural network modeling. *Geophys. Res. Lett.*, 30(7):1386, doi:10.1029/2002GL016726, 2003.
Pasini, A., Ameli, F., and Lorè, M., Mixing height short range forecasting through neural network modeling applied to radon and meteorological data. *Third Conference on applications of artificial intelligence to the environmental science: Proceedings CD-ROM*, paper 3.5, American Meteorological Society (83rd Annual Meeting), 2003.
Pasini, A., Pelino, V., and Potestà, S., A neural network model for visibility nowcasting from surface observations: Results and sensitivity to physical input variables. *J. Geophys. Res.*, 106(D14):14951-14959, 2001.
Pasini, A., and Potestà, S., Short-range visibility forecast by means of neural-network modelling: a case study, *Nuovo Cimento*, 18C(5):505-516, 1995.

Analysis of radon concentration in Slovenian thermal waters for earthquake prediction

Andreja Popit, Ljupčo Todorovski, Boris Zmazek, Janja Vaupotič,
Sašo Džeroski and Ivan Kobal

Jožef Stefan Institute, SI-1001, Ljubljana, P.O.B. 3000, Slovenia
andreja.popit@ijs.si , ljupco.todorovski@ijs.si, boris.zmazek@guest.arnes.si,
janja.vaupotic@ijs.si, saso.dzeroski@ijs.si, ivan.kobal@ijs.si

Key words: earthquake prediction, regression trees, model trees, linear regression, instance based regression

Introduction

Regression methods from the area of machine learning have recently been applied in earthquake prediction from radon data measured in soil gas (Zmazek et al., 2003, Džeroski et al., 2003). The obtained results show that in periods without seismic activity, the correlation between radon concentration and environmental parameters is significantly higher as compared to the correlation in periods with seismic events. In the present work, we apply machine learning regression methods to radon data measured in thermal springs.

Results and discussion

Geothermal waters are in contact with deep crustal rocks. This is the reason why spring gases might be more representative of the local environment than soil gases. Spring gases are much richer in deep gases and only slightly contaminated by atmospheric gases, and have been proved to be better earthquake precursors (Toutain and Baubron 1999). Our monitoring stations are installed in three thermal springs in West Slovenia (Zatolmin, Bled and Hotavlje), which are known to have deep groundwater circulation. These thermal waters raise from deep groundwater reservoirs through tectonic faults.

Radon concentrations were predicted from the measured geophysical, hydrological and meteorological data using different machine learning regression methods, implemented within the WEKA data mining suite (Witten and Frank, 2000); regression trees (rt), model trees (mt), linear regression (lr) and instance based regression (ib). We took radon concentration as the dependent variable, while the independent variables included the daily averages of water temperature, air temperature, the difference between air and water temperature, water pressure, atmospheric pressure, rainfall, and electrical conductivity. In addition, the one-day gradients of water pressure and atmospheric pressure were used as independent variables.

In the first part of our analyses, regression methods were tested as to how accurately they predict radon concentration from geophysical, hydrological and meteorological data. Their accuracy was measured in terms of the correlation coefficient (r), indicating the level of

correlation between the measured and predicted values of radon concentration, as well as with the root mean squared error (RMSE). In the second part, our study was continued with the most accurate regression method. A model for predicting radon concentration was built from data obtained during periods without seismic activity. Here, our hypothesis stated that the predictive performance for these periods would be better than for periods with seismic activity.

Conclusions

Our preliminary study using data from Bled and Hotavlje has shown that better results were obtained with model and regression trees than with the other regression methods. Comparing the predictive performance in periods without seismic activity with periods with seismic activity, we find the RMSE is lower in the first case, which confirms our hypothesis. The correlation is slightly higher, but this may be an artefact of the small dataset. Further data analyses of the geophysical and environmental parameters measured at the aforementioned thermal springs are planned. At Bled and Zatolmin, data are available not only from May 2002, but from 2000 on. We plan to analyze this data with the methodology outlined above. Earthquakes potentially responsible for strain effects and consecutive geochemical or geophysical anomalies in the investigated area will be considered.

Acknowledgements

This work was supported by the Ministry of Education, Science and Sport of Slovenia.

References

Džeroski, S., Todorovski, L., Zmazek, B., Vaupotič, J. and Kobal, I. Modelling soil radon concentration for earthquake prediction. *In Proc. Sixth International Conference on Discovery Science*, pages 87-99. Springer, Berlin, 2003.

Toutain, J.P. and Baubron, J.C. Gas geochemistry and seismotectonics: a review. *Tectonophysics*, 304: 1-27, 1999.

Witten, I.H. and Frank, E. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, San Farncisco, 2000.

Zmazek, B., Todorovski, L., Džeroski, S., Vaupotič, J. and Kobal, I. Application of decision trees to the analysis of soil radon data for earthquake prediction. *Appl. Rad. Isot*, 58: 697-706, 2003.

Improvement of emergency vaccination strategies against rabies in red fox (*Vulpes vulpes*) populations using a combination of Cellular Automata and Evolutionary Algorithms

Thomas Selhorst

Federal Research Centre for Virus diseases of Animals
WHO Collaborating Centre for Rabies Surveillance and Research
Department of Epidemiology
Seestrasse 55, D-16868 Wusterhausen, Germany
Email: thomas.selhorst@wus.bfav.de

Key words: Rabies eradication, fox population dynamics, Cellular Automate, Evolutionary Algorithm

Introduction

Rabies is an infectious, zoonotic disease whose lethality for humans necessitates world-wide efforts to combat it (Charlton, 1988). In Europe, rabies occurs in a sylvatic cycle. The red fox (Vulpes vulpes) is the main vector and the major host (Wandeler et al., 1974). Sylvatic rabies entered the European fox population in the mid-20th century, spreading from the east throughout the whole continent (Macdonald and Voigt, 1985; Steck and Wandeler, 1980). Despite various management policies in rabies control, the first notable results were only achieved by the use of oral vaccination the main vector (Steck et al., 1978; Winkler and Bögel, 1992, Stöhr and Meslin, 1996).

Long-term, large-scale oral vaccination as applied in Europe (Barrat and Aubert, 1993; Schlüter and Müller, 1995; Masson et al., 1996) continuously immunized about 70% of the fox population (Barrat and Aubert, 1993), resulting in a drastic decrease of rabies incidence in Europe. Today, most parts of Europe are free of rabies (some small rabies foci remain in the western part of Germany) and the European authorities now start to develop emergency vaccination strategies to be applied in cases of re-emerging or re-introduction of rabies into the European Community.

Because rabies is a zoonosis, it is clear that the conduction of field trials intended to find optimal vaccination strategies is forbidden. Therefore, a computer simulation model on spread of rabies in space and time is used in combination with an optimization algorithm in order to support European and National authorities to legislate for rabies emergency vaccination strategies.

The model used is based on that one described in detail in Tischendorf et al. (1998). It combines the spatial (two-dimensional) concept of a standard epidemic model with an individual-based approach for individual movements depending on the disease status and season. The model describes the spread of a single vector disease (i.e. rabies) in a host population (i.e. the fox population), which is characterized by separated social groups, referred to as “infection communities” (IFCs). According to epidemiological modelling, the state of an IFC can be one of the possible states: susceptible, infectious or empty. State transitions depend on: the fox population dynamics including dispersal, the characteristic of the disease, and the influence of man (immunization). Immunization varies the status of

infection communities and, consecutively, the rabies dynamics. In order to find optimal vaccination strategies the simulation program described above is combined with an evolutionary algorithm (Selhorst, 2000). The algorithm alters the immunization status of each IFC. A strategy is an optimum strategy, on condition that it is cost efficient and robust to environmental fluctuations. Robustness of the chosen strategy is assessed with the help of a threshold analysis in which strategy dependencies on the model parameters are analysed.

Results and discussion

Results of the analysis indicate that a strategy aiming at the rapid establishment of a stable vaccination coverage is optimal. The area to be vaccinated depends on the dispersal distance of rabid foxes and can be adjusted according to the environmental conditions.

Conclusions

Improvement of wildlife disease management strategies is always complicated by the fact that it is nearly impossible to conduct field trials. Hence computer simulation models in combination with optimization algorithms will gain increasing importance.

References

Barrat, J., Aubert, M.F., 1993. Current status of rabies in Europe. Onderstepoort J. Vet. Res. 60, 357-363.

Charlton, K.M., 1988. The pathogenesis of rabies. In: Campball, J.B. (Ed.), Rabies. Kluwer Academic Publishers, Boston, pp. 101-150.

Macdonald, D.W., Voigt, D.R., 1985. The biological basis of rabies models. In: Bacon, P.J. (Ed.), Population Dynamics of Rabies in Wildlife. Academic Press, London, pp. 71-108.

Steck, F., Wandeler, A., 1980. The epidemiology of rabies in Europe. Epidemiol. Rev., 2, 72-96.

Masson, E., Aubert, M.F., Barrat, J., Vuillaume, P., 1996. Comparison of the efficacy of the antirabies vaccines used for foxes in France. Vet. Res., 27, 255-266.

Schlüter, H., Müller, T. 1995. Tollwutbekämpfung in Deutschland. Ergebnisse und Schlußfolgerungen aus über 10-jähriger Bekämpfung. Tierärztl. Umschau 50, 748-758.

Selhorst, T., 2000. Improving the oral immunization of foxes (Vulpes vulpes) against rabies with the help of an evolutionary algorithm. Ecol. Mod. 129 2-3, 297-305

Steck, F., Wandeler, A., Bichsel, P., Capt, S., Schneider, L.G., 1982. Oral immunization of rabies against rabies. Zentralbl. Veterinärmed. 29, 372-396.

Stöhr, K., Meslin, F.M., 1996. Progress and setbacks in the oral immunization of foxes against rabies in Europe. Vet.Rec. 139, 32-35.

Tischendorf, L., Thulke, H.H., Staubach, C., Müller, M.S., Jeltsch, F., Goretzki, J., Selhorst, T., Müller, T., Schlüter. H., Wissel, C., 1998. Chance and risk of controlling rabies in large-scale and long-term immunized fox populations. Proc. R. Soc. Lond. B 265, 839-846.

Wandeler, A.,Wachendörfer, G., Förster, U., Krekel, H., Schale, W., Müller, J., Steck, F., 1974. Rabies in wild carnivores in central Europe. I. Epidemiological Studies. Zentralbl. Veterinärmed. 21, 735-756.

Winkler, W.G., Bödel, K., 1992. Control of rabies in wildlife. Sci. Am. 266, 56-62.

Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees

Mika Sulkava, Jarkko Tikka and Jaakko Hollmén

Helsinki University of Technology, Laboratory of Computer and Information Science, P.O.Box 5400, FIN-02015 HUT, Finland, Mika.Sulkava@hut.fi, tikka@mail.cis.hut.fi, Jaakko.Hollmen@hut.fi Tel. +358-9-451 3647, Fax. +358-9-451 3277

Key words: Linear sparse regression, prediction, foliar nutrition

Background and data

Analysis of foliar nutrient concentrations is an important part of environmental monitoring. Understanding and predicting the development of nutrient concentrations based on measurement data of the forest are challenging tasks. In this study sparse regression models were used to represent the relations between different measurements.

The nutrient data used in the analysis consist of needle mass (NM) and 12 element concentrations: Al, B, Ca, Cu, Fe, K, Mg, Mn, N, P, S and Zn. These 13 measurements were made to needles of foliar age classes C and $C + 1$ (the needles that were grown in the measuring year and in the previous year, respectively) in 16 Norway spruce and 20 Scots pine stands located in different parts of Finland between years 1987–2000.

In addition, there were 9 additional measurements available for the stands, namely the geographic coordinates, the total N and S deposition, the average temperature and total precipitation, the deviations of average temperature and precipitation from their long term averages and the age of the forest. All the measurements were done annually.

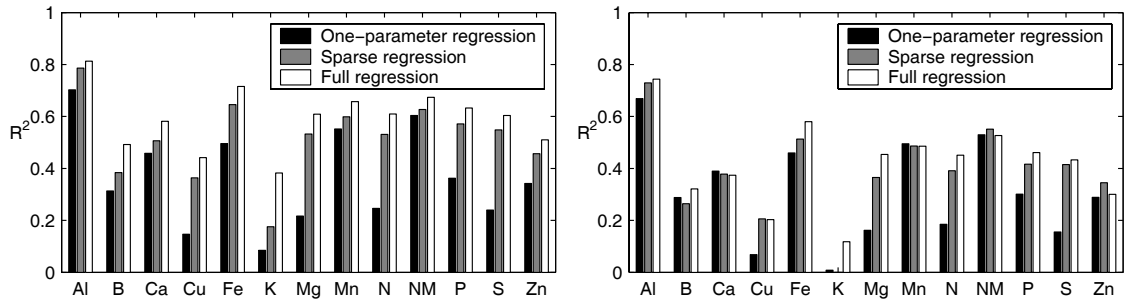
The problem at hand is to predict the nutrient concentrations and needle mass of $C + 1$ needles in year t using the measurements of C needles in year $t - 1$ and the additional measurements in year t. That is, we want to model the effect of the environment and nutrients to the aging of the needles. Also, the models should give an understandable description of the process. The purpose is to use only a few significant regressors of total 22 for each response. The most significant regressors are selected separately for each response, so that differences in dependency relations between the response and regressors in different models can be observed more easily.

Methods

Different multiple linear regression models are used for prediction. The use of linear models is justified by their interpretability and the fact that over short ranges, any process can be well approximated by a linear model. In a linear sparse regression model there are $k < K$ nonzero regression coefficients, where K is the number of regressors in a full model.

Using a sparse regression model instead of the full model is convenient, because reducing the number of coefficients makes the model easier to interpret and at the same time less prone to overfitting. In the models used in this study the most significant regressors are found using the Least Angle Regression model selection algorithm (Efron, 2004). An initial value of k is selected based on the Minimum Description Length information criterion. Subsequently, the final k is obtained by setting statistically insignificant coefficients to zero.

The sparse model is compared to the full linear regression model and a simple linear



(a) Training set (b) Validation set

Figure 1: *Average R^2-values of the measurements from pine for one-parameter regression, sparse regression and full regression obtained using cross-validation. Results for both (a) training and (b) validation sets.*

one-parameter model that tries to predict the value of a $C+1$ measurement in year t by only using its C value in year $t-1$.

Results

The sparse model was found to be more suitable for the problem than the two other models. The quality of prediction was studied using cross-validation. The prediction accuracy of the different models was measured with the coefficient of determination R^2.

The results for pine are shown in Figure 1 for both the training and validation sets. In the right panel of Figure 1 it can be seen that usually the sparse model outperforms the simple one-parameter model, and its results are mainly comparable to the full model. However, the number of parameters in the sparse model is much lower: on an average $k = 5$ coefficients ($K = 22$). This is an important advantage of the sparse models, because it helps finding the important dependencies between the different measurements.

The sparse model fits rather well to the data without any noticeable signs of overfitting. The difference between the R^2 values of the training and validation sets was constantly smaller with the sparse regression model than with the full model.

The quality of the models for spruce is similar, but the dependencies between the measurements were slightly different for the two tree species. The linear sparse regression model proved to be capable of providing rather good and reliable predictions of the development of foliage with a relatively small number of parameters.

Using a permutation test, it was found that virtually always the best possible regressors were chosen to the sparse models. That is, given the number of coefficients, it is extremely difficult to construct a linear model that would better characterize the relations between the measurements.

In addition, using cross validation, relative importance of the regressors was computed, that reveal the strength of the connections between different measurements. The values of relative importance can also be regarded as a discrete probability distribution, that shows, which regressors are likely to be included in the model. Usually, a measurement naturally has the strongest connection to its previous year value. Also other, more interesting dependencies were found between the measurements.

References

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, April 2004.

PERPEST: an expert model to Predict the Ecological Risks of PESTicides at the ecosystem level

Paul J. Van den Brink[1] and Colin Brown[2]

[1]Alterra, Wageningen University and Research centre, P.O. Box 47, 6700 AA Wageningen, The Netherlands. Email: paul.vandenbrink@wur.nl, Phone: +31-317-474338, Fax: +31-317-419000
[2]Cranfield University, Silsoe, Bedford, MK45 4DT, UK. Email: c.brown@cranfield.ac.uk

Key words: Effect model, Aquatic community, Ecological risk assessment, Pesticides, Case-Based Reasoning

Introduction

An important topic in the registration of pesticides and the interpretation of monitoring data is the estimation of the consequences of a certain concentration of a pesticide for the ecology of aquatic ecosystems. Solving these problems requires predictions of the expected response of the ecosystem to chemical stress. Up until now, a dominant approach to come up with such a prediction is the use of simulation models or safety factors. The disadvantage of the use of safety factors is that it lacks an adequate scientific basis and it does not provide any insight into the processes. For the modelling approach, all available quantitative information on important processes is combined in a mathematical model. However, these models have serious drawbacks, such as: 1) they are often very complex and lack transparency, 2) their implementation is expensive and 3) there may be a compilation of errors, due to uncertainties in parameters and processes.

The use of Case Based Reasoning in risk assessment

Case-based reasoning is a problem-solving paradigm that is able to utilise the specific knowledge of previously experienced, concrete problem situations (cases) for solving new problems. CBR is an approach that enables incremental, sustained learning since new experience is retained, making it immediately available for future problems (Aamodt and Plaza, 1994). A very important feature of case-based reasoning is its ability to learn. By adding present experience into the case base, improved predictions can be made in the future.
The Wageningen University and Alterra have developed a Case Based Reasoning methodology for the prediction of pesticide effects on freshwater ecosystems (Van den Brink et al., 2002). This methodology is called PERPEST (Prediction of the Ecological Risks of PESTicides) and is incorporated into a user-friendly interface (Van Nes and Van den Brink, 2003). It predicts the effects of a certain concentration of a pesticide on various (community) endpoints simultaneously. The database containing the “experiences from the past” was constructed by performing a review of freshwater model ecosystem studies evaluating the effects of insecticides and herbicides. This review was performed to assess the effects on various endpoints (e.g. community metabolism, phytoplankton, and macro-invertebrates) and to classify these effects according to their magnitude and duration. This literature review resulted in a database containing the effects of 22 herbicides and 24 insecticides. In total 104 experiments (49 herbicide, 55 insecticide) were evaluated, resulting in 421 cases (208

herbicide, 213 insecticide). The PERPEST model searches for analogous situations in the database based on relevant fate characteristics of the compound, exposure concentration and type of ecosystem to be evaluated. A prediction is provided by using weighted averages of the effects reported in most relevant literature references. PERPEST results in a prediction showing the probability of no, slight or clear effects on the various grouped endpoints (see Figure 1). The PERPEST model is described in Van den Brink et al. (2002) and Van Nes and Van den Brink (2003) and available at **www.perpest.alterra.nl**.
A key advantage of PERPEST over single species/safety factor analyses is that it removes the need to extrapolate to the ecosystem level. Also it encompasses chronic effects without the need to manipulate exposure concentrations in some artificial way (e.g. TWA concentrations) because one can just use the instantaneous concentration and compare it to the dosing concentration used in the experiments summarised in PERPEST.
In the full paper the way to integrate the PERPEST model with predicted and measured concentrations will be presented. Predicted concentrations will be combined with PERPEST predictions using Joint Probability Curves, which display the relationship between the probability of occurrence of a concentration and the probability of occurrence of a clear effect given that concentration. These probabilities can also be summarised in an overall frequency that clear effects would result. Monitored concentrations will be translated into ecological risks using PERPEST, which also is able to estimate the ecological risks of mixtures.

Outlook

This approach also has some obvious drawbacks: 1) often only very few really comparable cases are available and 2) specific cases are easily generalised too much. This led us to the idea that it would be good to seek the best of both worlds by using case-based reasoning as a mimic of the experts' approach and subsequently fine-tuning the results with the aid of simple ecological models. Branting et al (1997) called this integration of Case-Based Reasoning and model-based reasoning ‘model-based adaptation’ and described an example involving a system for rangeland grasshopper management. The latter part of this approach is new in the field of predictive modelling and will yield an approach that better estimates the effects of chemical stress and management on the ecology of aquatic ecosystems and also a modelling concept that can be used to tackle a variety of problems. In the light of the tiered approach that has been adopted in risk assessment and the availability of models, this integration looks promising for the field of ecological risk assessment of pesticides for their registration on the European market.

Acknowledgements

We would like to thank Egbert van Nes for his help developing the model and this work was supported by the Dutch Ministry of Agriculture, Nature and Food Safety (DLO/PO research programme 416).

References

Branting, L.K., Hastings J.D. and Lockwood J.A. Integrating cases and models for predictions in biological systems. *AI Applications* 11:29-48, 1997

Van den Brink, P.J., J. Roelsma, E.H. Van Nes, M. Scheffer and Brock, T.C.M. PERPEST, a Cased-Based Reasoning model to predict ecological risks of pesticides. *Environ. Toxicol. Chem.* 21: 2500-2506, 2002

Van Nes, E.H. and Van den Brink, P.J. PERPEST version 1.0, manual and technical description. A model that Predicts the Ecological Risks of PESTicides in freshwater ecosystems. Alterra-Report 787, Wageningen, The Netherlands, 2003.

Segmentation of paleoecological spatio-temporal count data

Kari Vasko[1], Hannu Toivonen[2] and Atte Korhola[3]

[1]CSC – Scientific Computing ltd., P.O.Box 405, FIN-02101, Finland, Kari.Vasko@csc.fi,
Phone: +358-9-4572734, Fax: +358-9-4572302
[2]Department of Computer Science, University of Helsinki, Finland,
Hannu.Toivonen@cs.helsinki.fi
[3]Department of Hydrobiology, University of Helsinki, Finland, Atte.Korhola@helsinki.fi

Key words: Spatio-temporal data analysis, segmentation, analysis of compositional data

Segmentation analysis addresses the following data analysis problem: given a time series, find a partitioning of the sequence to segments that are internally homogenous with respect to the desired pattern language and cost function. We will consider applications of segmentation analysis towards analysis of paleoecological spatio-temporal time series data. Our emphasis is both on computational and model building issues.

We outline a probabilistic framework for the spatio-temporal segmentation problems that occurs in paleoecology and discuss computational issues that arise in this setting. To this end, there has been no solid theoretic framework behind the zonation task. For instance, the current methods for numerical zonation of biostratigraphic sequences, e.g., broken stick, are limited since they do not fully specify local and global likelihoods of the data and, thus, they do not provide explicit assumptions concerning the data generating mechanism [1].

We introduce as an application Dirichlet-Multinomial Bayesian segmentation model for spatio-temporal count data that occurs frequently in paleoecological data analysis. A typical example of paleoecological time series count data is a sediment core data or a set of sediment cores that consists of abundances of species. The most probable segmentation model for species count data can be used to identify environmental changes if the species composition is known to be sensitive to the environmental changes. For instance, an organism called chironomid can be used to identify likely changes in air temperature, since chironomids are known to be sensitive with respect to the air temperature. As a simple example suppose we know that species A prefers warm conditions and species B prefers colder conditions. Further, suppose we collect the data represented in Table 1. The data indicates that it is more probable that 3000 years before present (BP) the environmental conditions have been clearly warmer than 1000 to 2000 years BP. A reasonable guess could now be that there are two zones in the data set illustrated in Table 1: one that covers time points 1000 and 2000 years BP and another one that covers 3000 years BP.

We will discuss computational issues that are related to the determination of the number segments using the probabilistic approach we adopt. We will introduce an efficient technique to compute an approximation of the marginal likelihood of the Dirichlet-Multinomial segmentation model for paleoecological spatio-temporal count data, which is needed in the determination of the number of segments.

The framework we introduce is capable of analyzing multiple data sets, e.g. data sets that consist of several time series that are possibly collected in different spatial locations, unlike existing methods used by paleoecologists. This feature can be used to identify local vs. global

changes as follows. Suppose we have two sediment cores A and B that were collected from two different spatial locations. Further, suppose that the both cores consist of abundances of the same organism, which, in turn, is known to be sensitive to the changes of the environment. Assume that at some time point there is a probable segment boundary in the core A but which is not probable in the core B. Then it can be argued that the corresponding indirect indication of the environmental change is not probably global one.

Layer id	***Species A***	***Species B***	***Time before present***
1	24	76	1000
2	15	85	2000
3	78	22	3000

Table 1: *Example count data for a segmentation task.*

We demonstrate using synthetic data that the proposed approximation gives more accurate predictions than Bayesian information criteria (BIC) altough the proposed approximation has the same time complexity as BIC. We will also consider and demonstrate performance of the state-of-the-art frequentist techniques [2,3]. Our experiments indicate that the Bayesian approach to zonation analysis is more suitable than the frequentist one, in particular, when short zones exist. Finally, we give demonstrations using real data.

References

[1] Bennett, K. Determination of the number of zones in a biostratigraphical sequence. *New Phytologist,* 132, 155-170, 1996.

[2] Vasko, K. and Toivonen, H. Estimating the number of segments in time series data. *Proceedings of The 2002 IEEE International Conference on Data Mining (ICDM'02)* pages 466-473.

[3] Vasko, K. *Computational methods and models for paleoecology*, Department of Computer Science Series of Publications A, Report A-2004-3 (PhD Thesis).

Modelling Lake Glumsø with Q^2 Learning

Daniel Vladušič[1], Boris Kompare[2] and Ivan Bratko[1]

[1]Faculty of Computer and Information Science, Ljubljana, Slovenia,
daniel.vladusic@fri.uni-lj.si, +386 1 4768 299
[2]Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia

Key words: Automated model building, Machine learning, Qualitative reasoning, Learning qualitative models, Numerical regression

Introduction

In this paper we describe an application of Q^2 learning, a recently developed approach to machine learning in numerical domains (Šuc *et al*., 2003), to the automated modelling of an aquatic ecosystem from measured data. We modelled the time behaviour of phytoplankton and zooplankton in Danish lake Glumsø using data collected by S.E. Jorgenson (Jorgenson *et al*., 1986). A number of researchers have applied various machine learning approaches to this data in the past. The novelty of Q^2 learning is in its paying attention to the *qualitative* correctness of induced numerical models. Experiments suggest that this property also leads to more accurate *quantitative* predictions and good interpretability of induced models.

Lake Glumsø: learning data and related work

Lake Glumsø (Jorgenson *et al*., 1986) is situated in a sub-glacial valley in Denmark. It is shallow with average depth of about 2 m and its surface area is 266,000 m^2. For several years, it was receiving mechanically-biologically treated waste water from a community with 3,000 inhabitants and a surrounding area which was mainly agricultural. The high nitrogen and phosphorus concentration in the treated waste water caused hypereutrophication. Concentrations of phytoplankton, zooplankton, soluble nitrogen, soluble phosphorus and temperature were considered relevant for modelling the phytoplankton growth. These variables were measured at 14 distinct time points over a period of two months. The amount of measured data itself is too small for automated modelling, so additional data was obtained as follows in earlier experiments with machine learning in this domain (Kompare, 1995). First, graphs with the measurements in time were given to three experts to interpolate a curve that, in their own opinion, best described the dynamic behaviour of the observed variable between the measured points. Such expert curves can be regarded as a way of introducing expert's domain knowledge, resulting in reliable additional data. These curves were then smoothed with Bezier splines. The sampling from these splines yielded three new, larger data sets. Additionally, a new (fourth) data set was obtained by applying a more sophisticated smoothing method to the curves plotted by the first expert. Other applications of machine learning to this data include Todorovski et al. (1998) and Todorovski (2003) using the LAGRAMGE equation discovery system. LAGRAMGE allows the introduction of expert's domain knowledge essentially in the form of domain-specific equation fragments. Q^2 learning, applied in the present paper, is a completely different approach that exploits monotonicity constraints present in the domain.

Q^2 learning method

The learning problem addressed by Q^2 method is as follows. Given is a set of numerical examples *S*, (observations, measurements), where each example gives the values of a set of independent variables and a set of dependent variables. The problem is to find a numerical

function f_i for each dependent variable, for predicting the value of the *i*-th dependent variable given the values of the independent variables. Q^2 learning solves this problem in two stages: (1) Construct qualitative constraints *QC* that hold in the modelled domain, and (2) Construct numerical functions f_i so that these functions (a) respect the constraints *QC*, and (b) fit the data *S*. The resulting numerical functions constitute a numerical model of the domain. The intermediate qualitative constraints *QC* are also part of the overall model, useful for the understanding and interpretation. The two stages above can be carried out in various ways. In this paper we used program QUIN (Bratko, Šuc, 2001) that induces qualitative constraints in the form of qualitative trees, and program QCGrid (Vladušič *et al*., 2004) that performs piece-wise linear regression respecting a given qualitative tree.

Q^2 learning with Glumsø data

A number of experiments with Q^2 learning using Glumsø data were performed and the results compared with those obtained with other representative methods for numerical machine learning. In particular, among the competing state-of-the-art methods, the learning of regression trees with M5 and LWR (locally weighted regression) were studied. Each of the four time series was cut into ten pieces. In a procedure similar to cross-validation, we used nine pieces for learning and the remaining piece for testing. Overall, we performed 40 experiments. The average prediction error, measured by RMSE (root mean squared error), over all 40 experiments is 7.69, 17.11 and 40.42 for Q^2, LWR and M5, respectively. The overall conclusion is that Q^2 convincingly outperformed regression trees and LWR in terms of predictive accuracy. The obtained differences over all 40 experiments were found significant (t-test, $p<0.05$). As the learning data was obtained in such a peculiar fashion, we were interested in "qualitative" comparison of the expert views on the original data. We used the Q^2 approach to induce corresponding qualitative models and reify them into the numerical ones. These models were then used to predict the time behaviour of phytoplankton on the remaining data sets, so we in a sense performed cross-validation. As it turns out, the data from every expert can be successfully used to predict the phytoplankton on other data sets.

Conclusions

We applied the Q^2 learning to the automated modelling of an aquatic ecosystem from measured data. The prediction accuracy obtained with Q^2 learning is significantly better than those obtained with the competing machine learning methods. The results suggest that Q^2 learning can be successfully used for explanation and prediction in ecological domains.

References

Bratko, I., Šuc, D. (2003) Learning qualitative models. *AI Magazine*, Vol. 24, No. 4, 107-119.

Jørgensen, S.E., Kamp-Nielsen, L., Chirstensen, T., Windolf-Nielsen, J. & Westergaard, B., 1986: Validation of a prognosis based upon a eutrophication model. *Ecological Modelling*, 32: 165-182.

Kompare, B. (1995) *The use of artificial intelligence in ecological modelling*. PhD Thesis, Royal Danish School of Pharmacy, Copenhagen, Denmark.

Šuc, D., Vladušič, D., Bratko, I. (2003) Qualitatively faithful quantitative learning. *Proc. IJCAI'03*, Acapulco, Mexico, August 2003

Todorovski, L. (2003) Using Domain Knowledge for Automated Modelling of Dynamic Systems with Equation Discovery. Ph.D. Thesis, Univ. of Ljubljana.

Vladušič, D., Šuc, D., Bratko, I. (2003) *Q2Q software for inducing quantitative models from designer's vague specifications*. Clockwork project report Rest-57, Univ. of Ljubljana.

Q^2 Prediction of Ozone Concentrations

Jure Žabkar[1], Rahela Žabkar[2], Daniel Vladušič[1], Danijel Čemas[3], Dorian Šuc[1], and Ivan Bratko[1]
[1]Faculty for Computer and Information Science, Tržaška 25, 1000 Ljubljana, Slovenia,
jure.zabkar@guest.arnes.si, {ivan.bratko, daniel.vladusic}@fri.uni-lj.si, +386 1 4768 299
[2]Faculty of Mathematics and Physics, Jadranska 19, 1000 Ljubljana, Slovenia,
zabkarr@fmf.uni-lj.si
[3]Environmental Agency of the Republic of Slovenia, Vojkova 1b, 1000 Ljubljana, Slovenia,
danijel.cemas@gov.si

Keywords: Ozone concentration prediction model, Air pollution, Qualitative modelling

Introduction

In this paper we describe an application of Q^2 learning (Šuc et al., 2003), to the automated prediction of ozone concentrations in the cities of Ljubljana and Nova Gorica. Ozone prediction model is being developed for the purpose of issuing public alerts to avoid exposure to high ground-level ozone concentrations. Input data are meteorological (temperature, relative humidity, wind speed and direction, solar radiation, precipitation) and air quality measurements (O_3, NO, NO_2) as well as predictions of the ALADIN (Aladin, 1997) model. Measurements and ALADIN data were provided by Environmental Agency of the Republic of Slovenia (ARSO). The predictions are made at 08:00 for 15:00 at the same day. The learning set consists of data collected from the years 2001-02, while the data from April to October 2003 are used as a test set. We are primarily focused on building a well interpretable qualitative model that explains the complex nature of chemical and physical background. Field experts find our model consistent with their understanding of the process. Numerical accuracy of the model's predictions are compared to classical machine learning methods - regression trees (M5) and linear regression (LR), both implemented in Weka (Witten, 2000).

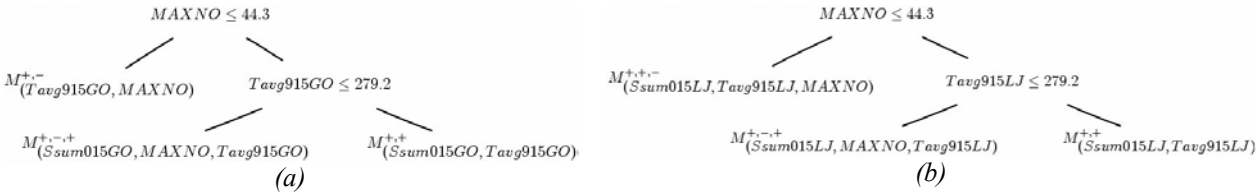
Results and discussion

Q^2 learning is carried out in two stages. At first, a qualitative tree is induced from training data using program QUIN (Bratko, 2003). The second stage, qualitative-to-quantitative (Q2Q) transformation, is carried out using program QCgrid (Vladušič, 2003) that performs piece-wise linear regression respecting the qualitative constraints, given by the qualitative tree. The result of the first stage, the qualitative tree, is used for model explanation while the second stage enables numerical predictions, respecting the qualitative constraints.

The attributes used in the learning process were built from the ALADIN predictions at the model grid points, neighbouring the meteorological station point in both cities. At that point, the meteorological measurements were performed. The attributes used are: *MAXNO* (max. concentration of NO in the last 36 hours before the prediction is made), *Tavg915GO /LJ* (avg. of the Aladin's predictions of temperature from 09:00 to 15:00) in Nova Gorica (*GO*) and Ljubljana (*LJ*) and *Ssum015GO /LJ* (the sum of Aladin's predictions of solar radiation from 00:00 to 15:00 for each city). The attributes were chosen and built by field experts.

The qualitative trees for both cities are shown on Fig.1. The structure of both trees is very similar and schematically shows that the ozone concentration is positively correlated to the temperature and solar radiation while negatively correlated to the concentration of NO. The concentration of NO in the roots of the qualitative trees evaluates the dominating mechanisms of the ozone cycle. Higher NO concentrations occur during nighttime hours with low ozone concentration (right branch). On the contrary, high ozone concentration as a result of photochemical formation prevents high NO concentration (left branch).

Temperature and solar radiation are from statistical point of view highly coupled so model can choose each of the variables to describe the presence and intensity of photochemical reactions in the atmosphere. During night time and cloudy days without solar radiation usually results in lower temperatures. In our case temperature showed better statistical correlation to ozone concentration which resulted in the second tree root. This proves that highest ozone concentrations occur during daytime in summer in hot sunny and dry weather. No wind dependence can be found, which indicates that local sources of ozone precursors in the city have an important role in ozone formation.



(a) *(b)*

Figure 1: Qualitative trees for Nova Gorica (a) and Ljubljana (b).

Numerical accuracy of induced models is compared to LR and M5. Table 1 shows the RMSE achieved on the test set. Q^2 turns out to be superior to LR and M5 although not significantly.

Table 1: Comparison of the numerical accuracy of the competing methods.

RMSE on test set	LR	M5	Q^2
LJ	21.63	22.94	19.9
GO	20.98	19.92	19.81

Conclusions

We introduced the qualitative models for prediction of ozone concentration in the cities of Ljubljana and Nova Gorica. The experts found the models explanatory and consistent with their understanding of the complex process.

Acknowledgements

We are grateful to dr. Matevž Pompe for evaluation of our qualitative models from the chemical point of view.

References

Bratko, I., Šuc, D., *Learning qualitative models*. AI Magazine, Vol. 24, No. 4, 2003.

Šuc, D., Vladušič, D., and Bratko, I., *Qualitatively faithful quantitative prediction*, in: Proceedings of the eighteenth International Joint Conference on Artificial Intelligence, pp. 1052-1057, San Francisco: Morgan Kaufmann Publishers, 2003. Acapulco, August, 2003.

Vladušič, D., Šuc, D., and Bratko, I., *Q2Q software for inducing quantitative models from designer's vague specifications*. Clockwork project report Rest-57, Univ. of Ljubljana, 2003.

Witten, I., and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, 2000.

Aladin International Team, *The ALADIN project: Mesoscale modelling seen as a basic tool for weather forecasting and atmospheric research*, WMO Bull., 46, pp. 317-324, 1997.

Environ International Corporation, *Comprehensive Air Quality Model with Extensions*, version 4, User's Guide, www.camx.com, Novato, California, 2004.

Automatic construction of concept hierarchies: The case of foliage-dwelling spiders

Martin Žnidaršič[1], Aleks Jakulin[2], and Sašo Džeroski[1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia, martin.znidarsic@ijs.si, +386 1 477 3366
[2] Artificial Intelligence Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Key words: data-based hierarchy construction, interaction analysis, feature construction, spiders, field margins

Introduction

Ecological domains are complex, with interdependent variables and hidden relations that are difficult to explain. These characteristics indicate that the ecology experts might benefit from the use of machine learning methods. Machine learning can be used to confirm hypotheses or to discover new relations, thus gaining insight into vast amounts of data. However, the hardest part, evaluation and explanation, has to be done by experts.

In this paper we present the hierarchical structures that were learned from raw data with the use of two machine learning techniques. Some information is given about these techniques and the way they were used to construct hierarchies of variables from the data. The dataset we used contains the measurements of variables that might influence the density of foliage-dwelling spiders in field margins. This dataset was used before by domain experts [2] to construct a fuzzy qualitative model of hierarchical variable dependency. The model was mainly constructed manually with some use of data analysis techniques.

Results and Discussion

We automatically constructed hierarchical models of this dataset in two ways, using interaction analysis and function decomposition. First we applied *interaction analysis* [1], a set of tools for identifying interactions among the variables in data. Interactions are dependencies between variables that deserve closer investigation. In prediction tasks, we are especially interested in 3-way interaction between two independent variables (such as *education* or *age*) and the result (*salary*). There are two types of 3-way interactions: the

two variables may be synergistic in the sense that controlling for both of them unlocks an otherwise hidden pattern. On the other hand, the two variables may be redundant if they both provide the same information. The interaction dendrogram, which summarizes the 3-way interactions found in data, matches the structure of the model that was built by domain experts exactly. It also provides some clues about which variables are equally appropriate to be used in the same place of the model structure, as well as additional clues about variable dependencies (some of which can be due to noise).

The second method we used, *function decomposition*, is a member of a larger family of constructive induction methods, which focus on discovering novel concepts in data. We employed the hierarchy induction tool HINT [3]. This method is also able to create new concepts and rules to compute their values, not only the hierarchical structure. However, it tends to be sensitive to noise in the training data. Its results did not exactly match the results of the experts, but the constructed concepts could be interesting to domain specialists nevertheless.

Conclusions

The results we obtained with interaction analysis are encouraging as the automatically constructed hierarchy is very similar to the one that was manually constructed by domain experts. This indicates that the method could be useful for construction of preliminary sketches of such models, thus saving experts some valuable time. It is also possible that both methods can help us identify complex concepts that would be hard to discover manually.

The results suggest that integrating the approaches of interaction analysis and function decomposition is a promising direction for further work: i.e., one might extract the hierarchy with interaction analysis and then use function decomposition, i.e., HINT to find the rules in its internal nodes.

References

[1] Jakulin, A., Bratko, I.: Analyzing Attribute Dependencies. Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003), Cavtat-Dubrovnik, Croatia, September 22-26, 2003. N. Lavrac, D. Gamberger, H. Blockeel, L. Todorovski (Eds.) Lecture Notes in Artificial Intelligence, Vol. 2838, Springer, pp. 229-240, 2003.

[2] Kampichler, C., Barthel, J., Wieland, R.: Species density of foliage-dwelling spiders in field margins: a simple fuzzy rule-based model. Ecological Modelling 129: pp. 87-99, 2000.

[3] Zupan, B., Bohanec, M., Demsar, J., Bratko, I.: Learning by discovering concept hierarchies. Artificial Intelligence 109: pp. 211-242, 1999.

Author Index