EnviroInfo 2005 (Brno)
Informatics for Environmental Protection - Networking Environmental Information
 ISBN: 80-210-3780-6

Using Data Mining Techniques for Exploring the key Features of Plant Dynamics Upon a Newly Built Plant Trait Database

Michael Stadler[1], Renée M. Bekker[2], Jens Finke[3], Dierk Kunzmann[4] and Michael Sonnenschein[3]

Abstract

Most data mining techniques have rarely been used in ecology. To address the specific needs of scientists analysing data from a plant trait database developed during the LEDA project, a web-based data mining tool has been developed. This paper presents the DIONE data miner and the project it has been developed in. It addresses the nature of plant trait data from a data mining perspective and points out problems that arise when preparing data for being mined. Furthermore it contains information about data mining results on plant trait data achieved so far and conclusions towards the applicability of data mining to the fields of ecology, and the newly emerging field of research recently addressed as ecoinformatics.

1. The LEDA Project

The LEDA project (Life-history traits of the Northwest European Flora - a database), funded by the European union (contract no. EVR1-CT-2002-40022) currently builds up a European-wide database with information on 30 plant characteristics for about 3000 plant species of Northwest Europe. Data comes from scattered database initiatives, from literature and from new measurements and observations conducted during the project and is collected in the so called *LEDA Traitbase* [Knevel 2003]. For the first time in history, plant trait data covering a substantial part of a continent will be available from a single database accessible through the WWW.

The LEDA Traitbase carries the potential of generating a huge amount of knowledge on species grouping, assembly rules of communities and the understanding of species' mobility through the landscape resulting in explaining invasiveness or endangerment of species. Its web interface consists of a portal application and sub-applications for the purposes of online data input, uploading data, performing online queries, online reviewal of data [Ahlers 2005], administrative tasks, editorial tasks, and for the purpose of data mining. The web-applications will be publicly accessible through the LEDA web portal (http://www.leda-traitbase.org) by November 2005. Registration is necessary to use the data-mining application.

2. Data analysis in ecology

To analyse data, ecologists most often use statistical methods as e.g. regression or ordination techniques. For classification purposes hierarchical classification (a clustering method) is commonly used [Lepš 2003]. Only in rare cases researchers from the field of ecology make use of symbolic machine learning methods [Džeroski 2001], e.g. to explain classes in terms of their instances' attribute values.

[1] OFFIS Research Institute, Escherweg 2, D-26121 Oldenburg, Germany, Email: stadler@offis.de
[2] Community and Conservation Ecology Group, University of Groningen, P.O. Box 14, 9750AA Haren, The Netherlands
[3] Carl von Ossietzky University of Oldenburg, Dept. for Computer Science, D-26111 Oldenburg, Germany
[4] Carl von Ossietzky University of Oldenburg, Landscape Ecology Group, P.O. Box 2503, D-26111 Oldenburg, Germany

However, due to the nature of data in the LEDA Traitbase, there are several reasons calling for application of data-mining methods either proposing classifications for previously unclassified data or describing classes of data by the data's properties:

1. Quite many traits are difficult to measure or observe. Thus, biologists are interested in knowing whether there is a set of other, easier accessible plant traits that imply certain values of the trait under question. Descision trees and decision rules are well suited to produce answers to questions of this kind.
2. Plant traits may differ not only between species but also between different environmental conditions, habitats, altitudes, regions or even latitudinal zones. Clustering methods are a promising means of analysing these interdepedencies.
3. Combinations of plant traits may be indicative for so called plant functional types. Therefore, it is useful to use clustering to search the Traitbase for sets of species similar in terms of groups of plant traits which than can be tested for their functional response to changes in the environment.

3. Mining plant trait data with DIONE

To allow ecologists and biologists to mine data from the LEDA Traitbase using methods of symbolical machine learning and clustering, a tool was needed, having direct access to data from the LEDA traitbase without lengthy configuration and being accessible over the WWW by use of a simple browser. For this purpose the DIONE data miner has been designed and developed.

3.1 DIONE data miner architecture

The data mining application has a web-based user-interface. It allows data-mining methods to be applied to LEDA Traitbase data (direct interface) or to arbitrary data uploaded via a csv file import option. There are three kinds of experiments: *model creation experiments* for learning intensionally described concepts from example data, *model verification experiments* for verifying learned concepts based on known data, and *prediction experiments* allowing data to be classified according to learned concepts. Users are guided through the subsequent phases of constructing and processing data mining experiments by a sequence of tabs on the upper part of the web-pages. At each point in the experimenting process only those functions available due to the state of the experimentation process are accessible. The remaining tabs are grayed-out until all data needed for successfully carrying out the respective steps are available. Users can publish data mining experiments to make them accessible to other users, who can analyse them or copy them and then continue to work on them. Data mining results can be downloaded as PDF-documents. Data enhanced by a class membership as outcome of either a clustering process or a prediction experiment can be downloaded as csv file.

In order for reducing system load data mining processes experiments are run one at a time by a scheduler on a first come first serve basis. No need was felt for more sophisticated scheduling algorithms, as only a couple of users are expected to use the data mining tool at any given time. DIONE has been developed as a web-application using J2EE. It runs within a standard-conformant J2EE web application container.

3.2 Supported methods of data mining

Currently, eight data mining algorithms are supported. Two algorithms for inductive learning of decision trees, two algorithms for center-based clustering, two algorithms for constructing decision rules and one

algorithm for learning association rules. One of the decision tree algorithms is available in two flavours: The standard flavour for automatic descision tree induction and the vanilla flavour for step-wise, user-interactive and system guided construction of decision trees. Most of these methods have been implemented using the algorithms provided by the [Xelopes] data-mining library. [Witten 2000] gives detailed information on data mining algorithms.

3.3 Data Structure, data quality and data preparation

Each record of plant trait data stored in the LEDA Traitbase consists of multiple values that together reflect a trait characteristic. The values may either denote categories, being nominal or ordinal values or they are numeric. Values may either describe a measurement or observation per se or they might describe the context of a measurement, e.g. a geographical location, a habitat or the protocol followed for collecting and measuring data.

Due to the large amount of different data sources, data may be both of very different statistical and ecological data quality [Stadler 2004] and it may be incomplete. Such factors have to be accounted for during the data preparation phase when assembling data mining sets by aggregation and domain mapping. In the context of mining plant trait data, the following problems were found to occur frequently:

1. Information not covered by data. If information relevant to a problem is not included in the data to be mined, decision tree and decision rule algorithms produce classifiers with a low rate of correct classifications due to an overfitting to training data. Therefore, e.g. if data from trees and herbs is included in a set of data about seed dispersal and seed mass, an attribute enabling differentiation between these kinds of plants should be included.
2. Handling of missing values. Only few data mining algorithms can cope with missing values. Therefore heuristics for handling missing values have to be applied. These consist of either dropping records or attributes with missing values or of deriving missing values from other values. However, the latter solution gives no additional information to the data mining algorithm, so the former method should be preferred.
3. Cardinality of classes in training set. All objects from the training set assigned to the same class constitute a training subset. If cardinalities of training subsets differ too much from one another, subsets of high cardinality are likely to be described with less precision than subsets with low cardinality. To avoid this, the training set has to be carefully chosen.

4. Exemplary Problems and Results

About ten different problems have so far been addressed by LEDA project members using the DIONE data miner. For most of these meaningful results have been obtained. A selection of data mining results and their implications will be presented in this section.

4.1 Analysing plant traits as indicators of rarity

[Bekker and Kwak 2005] contains a study in that, for 135 species, the values plant traits *life span, clonality, breeding system, seed production, seed dispersal,* and *soil seed bank longevity* are each mapped to a number between 0.1 and 0.9 indicating the vulnerability of a species for extinction by the measured plant trait value. E.g. if a species only produces a few seeds per year, it is vulnerable to factors either destroying seeds or rendering them inable to germinate. It is thus assigned a high vulnerability w.r.t. the seed produc-

tion trait. Rarity of a species is deduced from a version of the red list of endangered species covering the area where trait measurements were taken from.
The study presents a way of deducing the trait combinations indicating rarity of a species from the trait-related vulnerability indices.

Using the DIONE data miner with the J4.8 algorithm from the [WEKA] project, a decision tree with species rarity as target attribute was generated in order to reproduce results from the study. It showed that not only the study results could be reproduced, but also more detailed information in terms of ranges of-trait values indicating the rarity of species could be found.

4.2 Analysing correlation between regeneration traits and dispersal traits

During an experiment based on the WEKA J4.8 decision tree induction algorithm, data for about 450 species was mined in order to analyse the dependency of plant height from seed mass, seed production and plant life span. The known fact, that a species' height primarily depends upon its life span could be reproduced. Also, the secondary correlation between seed mass and seed production on one hand and plant height on the other hand shows in the decision tree generated and the fact that plants with heavier seeds produce fewer seeds within a given plant height category also showed in the tree. Finally, the resulting tree also shows that relatively small plants produce a lower number of seeds than relatively high plants.

Especially during this experiment, it showed, that data quality has an important influence on the expressiveness of decision trees.

4.3 Analysing seed bank data against seed mass and fruit mass data

Two series of experiments trying to discover known relationships between data about seeds surviving in the soil (soil seed bank data) on the one hand and seed mass / fruit mass on the other hand were conducted. The first series was focused on the relationship between seed mass and soil seed bank characteristics. It could be shown, that seeds with lower mass survive in the soil for a longer period of time than seeds with a high seed mass. This is an ecologically well-known interrelation. The second series was used to analyse the relationship between fruit mass and soil seed bank characteristics. Here also, the already known relationship between fruit size on one hand and species soil seed bank longevity could be shown. But, in addition a class of species was identified, for which the commonly known relationship between fruit size and seed bank longevity does apparently not hold, and blurred earlier correlations: These species have large fruits, but their survival capacity in the soil seed bank is extraordinarily high. This special cases involve all species with large fruits that contain (many) small seeds. This analyses helps refining datasets and produces issues for further investigation.

The approach used was a combination of clustering and decision tree construction supported by DIONE through its capability of using clustering results as input to decision tree construction algorithms. In a first step, potentially meaningful clusters of species in terms of the plant traits under examination were detected using clustering algorithms. After this first step domain experts could assess the importance, meaning and usefulness of clusters based both on statistical parameters delivered by the data miner and on the mapping of records to clusters. As a means of better understanding the meaning of clusters, decision trees explaining the clusters were generated with the cluster name as target attribute. The single steps were repeated several times while readjusting data mining algorithm parameters.

5. Judging the usability of data mining for ecological data analysis

Domain experts rated the general usability of the DIONE data mining as very well, especially due to the intuitive guidance of users through the entire experiment creation and processing flow. Users were especially enthusiastic about the easily understandable and explanatory nature of class descriptions generated by decision tree induction algorithms and algorithms producing decision rules.

The data mining experiments on real data conducted so far suggest, that data mining techniques can both reproduce known relationships between plant traits, point out new interrelationships between plant traits and point to exceptions from known relationships between plant traits that have to be further analysed. Ecologists and biologists involved in the LEDA project were enthusiastic about the relatively small amount of work involved in producing results with a data mining tool compared to statistical tools they are used to. However, it was felt that due to the relative newness of the data mining techniques used in DIONE in the field of ecological research, results would always have to be validated with well-known statistical methods before they could be published.

6. Further Work

It is planned to make the DIONE data mining tool available under open source (GPL license). For information, please contact info@eco-software.org. Development will be continued in terms of adding additional algorithms, smoothing the workflow, offering properties of generic coupling to mining data sources, exporting generated classifiers and adding additional statistical evaluation possibilities for mining results.

7. Acknowledgements

We thank the European Union for funding the LEDA project. The DIONE data miner has been developed, designed and implemented by a computer-science master student project group during one year. The students involved are: B. Bensien, R. Hackelbusch, S. Heisecke, N. Henze, R. Hilbrands, H. Kraef, J. Künnemann, P. Kuhn, D. Meyerholt, F. Postel and H. Tschirner.

Bibliography

Ahlers, D. (et al) (2005) A web-based reviewing process guidance system for an ecological database of plant traits, Proc. EnviroInfo 2005

Bekker, R.M. and Kwak, M.M. (2005). Life history traits as predictors of plant rarity, with particular reference to hemiparasitic Orobanchaceae. Folia Geobotanica 40: 231-242

Džeroski, S. (2001) Applications of symbolic machine learning to ecological modelling, Ecological Modelling 146 pp. 263-273

Knevel, I.C. (et al) (2003) Life-history traits of the Northwest European flora:The LEDA database, Journal of Vegetation Science 14 pp. 611-614

Lepš, J. and Šmilauer, P. (2003) Multivariate Analysis of Ecological Data using CANOCO, Cambridge University Press

Stadler, M. (et al) (2004) Data Quality, Abstraction and Aggregation in the LEDA Traitbase, Proc. EnviroInfo 2004, Vol 1, pp. 515-524, Geneva, Switzerland.

WEKA (WWW link) http://www.cs.waikato.ac.nz/~ml

Witten, I.H. and Eibe, F. (2000) Data Mining, Morgan Kaufmann Publishers

Xelopes (WWW link) http://www.prudsys.com/Software/Algorithmen/Xelopes