Published in: A. Gnauck (Hrsg.), 2006:
Modellierung und Simulation von Ökosystemen - Workshop Kölpinsee 2005.
ASIM Mitteilung AMB 98, ISBN 3-8322-5477-3, pp. 173-182

Ein webbasiertes Data-Mining-Werkzeug zur Analyse ökologischer Daten

Michael Stadler[1], Michael Sonnenschein[1]

Zusammenfassung

Für die Analyse von Daten aus einer Datenbank über Merkmale von Pflanzenarten Nordwesteuropas wurde ein webbasiertes Werkzeug entwickelt. Dieses ermöglicht es, verschiedene Data-Mining Algorithmen – insbesondere symbolische Lernverfahren für Klassifikatoren – auf den Datenbestand anzuwenden und so Muster in den Daten zu erkennen und zu beschreiben. Das besondere Augenmerk bei der Entwicklung des Werkzeugs bestand in der einfachen Benutzerführung sowie in seiner Unabhängigkeit von einem Einzelproblem.

1 Das LEDA Projekt

Im EU-Projekt LEDA (Knevel et al. 2003, Stadler und Sonnenschein 2005) wurde eine Datenbank (Traitbase) mit Daten zu etwa 30 Merkmalen (Traits) von ca. 2000 Pflanzenarten Nordwesteuropas entwickelt. Ziel war es, erstmalig die verstreuten Datenbestände über einzelne Traits, einzelne Artengruppen und einzelne Regionen aus Datenbanken und Literatur zusammenzuführen und durch Messungen im Projekt zu ergänzen.

Der Datenbestand soll unter anderem Naturschutzbehörden die Vorhersage der Reaktionen von Nutzpflanzen auf Bewirtschaftungsmethoden und Eingriffe in die Umwelt ermöglichen, als Basis für die Vorhersage und Modellierung von Veränderungen in der Vegetation dienen und eine Grundlage für die Ermittlung funktionaler Pflanzentypen im Rahmen der vergleichenden funktionalen Ökologie sein.

Für diesen Zweck sind Werkzeuge zur Datenanalyse erforderlich, von denen eines, der DIONE Data Miner im Rahmen des Projekts durch eine studentische Projektgruppe entwickelt und realisiert wurde.

[1] OFFIS e.V., Escherweg 2, D-26121 Oldenburg; e-mail {stadler, sonnenschein}@offis.de

2 Anforderungen und Funktionsumfang

Data Mining bezeichnet ein in vielen Anwendungsbereichen verwendetes Bündel von Methoden zur Analyse von – meist umfangreichen – Datensätzen auf darin enthaltene inhaltliche Zusammenhänge. Eine Einführung in diese Methoden, die vielfach dem Bereich des maschinellen Lernens zuzuordnen sind, geben etwa (Han und Kamber 2000, Witten und Frank 2000). In der Ökologie finden sich zwar etliche Anwendungen von Clustering-Methoden oder künstlichen neuronalen Netzen zur Klassifikation (Recknagel 2001), jedoch bisher eher wenige Beispiele zur Anwendungen von symbolischen Lernverfahren (Džeroski 2001), wobei uns dies im Wesentlichen auf einen mangelnden Bekanntheitsgrad der Verfahren für diesen Anwendungsbereich zurückzuführen zu sein scheint.

So wurde das Data-Mining Werkzeug DIONE in erster Linie zur Beantwortung wissenschaftlicher Fragestellungen in ökologischen Zusammenhängen durch symbolische Verfahren entwickelt. Typische Fragestellungen an die Datenanalyse im gegebenen Zusammenhang sind etwa die Abhängigkeit der Ausprägungen von Pflanzenmerkmalen von Umweltbedingungen, die Folgerung der Ausprägung eines Pflanzenmerkmales aus den Werten anderer Pflanzenmerkmale sowie die Identifikation von Arten, die für einen gegebene Menge von Pflanzenmerkmalen ähnliche Ausprägungen haben und somit einem gemeinsamen sogenannten funktionalen Pflanzentyp angehören. Während die letztere Fragestellung etwa durch Clustering-Methoden beantwortet werden kann, lassen sich die vorherigen Fragestellungen mit Hilfe von Klassifikatoren beantworten, die durch induktive Lernverfahren generiert werden. Ein Klassifikator ist in diesem Zusammenhang eine Menge von Regeln oder ein Entscheidungsbaum, die ein Konzept beschreiben. Ein solches Konzept ist eine intensionale Beschreibung der Elemente einer Klasse, die sich für die Zuordnung von Datensätzen zu dieser Klasse eignet.

Das entwickelte Data Mining Werkzeug eignet sich sowohl für die Identifikation von Mengen ähnlicher Datensätze (Clustering) als auch für das Lernen von Regeln und Entscheidungsbäumen aus Datensätzen. Des weiteren ist auch das Erlernen von Regeln oder Bäumen zur operationalen Beschreibung von Clustern möglich. Zu jedem erkannten Cluster ist es somit prinzipiell möglich, einen Klassifikator zur Zuordnung weiterer Datensätze zu generieren.

Neben der Berechnung von Clustern, Regeln und Bäumen, bietet das vorgestellte

Werkzeug die Möglichkeit, Klassifikatoren anhand von Testmengen auf Ihre Genauigkeit hin zu überprüfen. Schließlich ist es möglich, neue Datensätze anhand einmal erzeugter Klassifikatoren zu klassifizieren und so das gelernte Wissen anzuwenden.

Eine gute Erlernbarkeit und Bedienbarkeit sowie die kostenfreie Verfügbarkeit stand beim Entwurf des Werkzeugs ebenso im Vordergrund wie der einfache Zugriff sowohl auf Daten aus der Traitbase als auch auf Daten aus Dateien mit kommaseparierten Datensätzen. Des Weiteren sollte das Werkzeug in Form einer Web-Anwendung realisiert werden, um den Aufwand für Wartung und Distribution klein zu halten. Dies ist ein wichtiges Unterscheidungsmerkmal zu kommerziell verfügbaren Data-Mining-Werkzeugen wie etwa SPSS Clementine (Clementine), die in der Regel als Desktop- oder Client-Server-Anwendungen konzipiert sind. DIONE-Benutzer benötigen nur einen Web-Browser um das Programm nutzen zu können.

3 Systemarchitektur

Der DIONE Data Miner ist im Wesentlichen ein Programm zum Web-basierten Zugriff auf Algorithmen aus vorhandenen, frei verfügbaren Data-Mining Bibliotheken, das über Funktionen zur Übernahme, Organisation, Weitergabe und Ausgabe von Daten verfügt. Die Systemarchitektur ist in Abbildung 1 dargestellt. DIONE, wie auch die anderen Komponenten von LEDA, nutzt die J2EE-Technologie, d.h. insbesondere Java Server Pages (JSP) und Java Beans zur Realisierung.

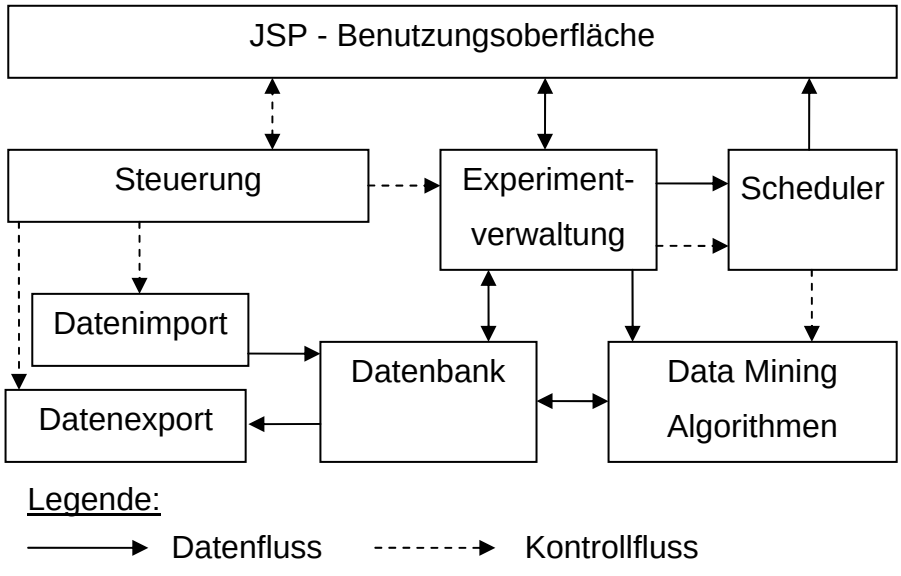

Abbildung 1: Funktionsblöcke des DIONE Data-Miners und ihre Interaktion

Der zentrale Begriff zum Verständnis der Abläufe im Programm ist das Experiment. Ein Experiment besteht aus einer Menge von gleich aufgebauten Eingabedatensätzen – auch Traitmatrix genannt, Konfigurationsdaten zur Auswahl und Parametrisierung eines Data Mining Algorithmus sowie ggf. einem Klassifikator oder einer Menge von Clustern als Ergebnis. In der Traitmatrix ist jeder Datensatz markiert und zwar entweder als Trainingsdatensatz zur Berechnung des Ergebnisses, als Testdatensatz mit Klassenzugehörigkeit zum Test eines Klassifikators oder als zu klassifizierender Datensatz, dem mit Hilfe eines Klassifikators eine Klasse zugeordnet werden soll.

4 Verfügbare Data-Mining Algorithmen

Der DIONE Data Miner bietet verschiedene Algorithmen aus den Data-Mining Bibliotheken Weka (Witten und Frank 2000, Weka) und der Open-Source-Version von Xelopes (Xelopes) an. Insgesamt handelt es sich um zwei Algorithmen zur Generierung von Entscheidungsbäumen, zwei Algorithmen zur Erzeugung von Klassifikationsregeln, einen Algorithmus zur Erzeugung von Assoziationsregeln sowie um das K-Means Clusteringverfahren und eine Variante davon. Zusätzlich wurde ein Entscheidungsbaumalgorithmus so modifiziert, dass er schrittweise ausführbar ist und der Benutzer nach jedem Konstruktionsschritt in die Lage versetzt wird, in den Algorithmus einzugreifen, wozu der entstehende Baum und die verfügbaren Entscheidungen zur Konstruktion visualisiert werden. Bei den Clustering-Verfahren besteht noch der Ergänzungsbedarf um eine hierarchische Methode.

5 Datenquellen

Nach dem Anlegen eines Experiments muss eine Menge von Datensätzen gleicher Struktur zum Import in die Traitmatrix festgelegt werden. Dem Benutzer werden hierzu mehrere Datenquellen angeboten:

- Für die Abfrage von Daten aus der LEDA Traitbase wurde eine webbasierte Abfragekomponente entwickelt. Diese ist an den DIONE Data Miner angebunden und ermöglicht den komfortablen Zugriff auf Daten aus der Traitbase.
- Daten können aus einem anderen Experiment kopiert werden.

- Es besteht die Möglichkeit, mit Hilfe eines komfortablen Wizards Daten aus Dateien mit kommaseparierten Datensätzen zu importieren, wie sie etwa mit gängigen Tabellenkalkulationsprogrammen erzeugt werden können (Vgl. Abschnitt 6.2).

6 Benutzeroberfläche und Bedienungsabläufe

Die Benutzeroberfläche des DIONE Data Miners wird im Internet Browser des Benutzers dargestellt. Nach der Anmeldung mit Benutzerkennung und Paßwort, hat der Benutzer die Wahl, ob er ein neues Experiment anlegen möchte oder mit bereits vorhandenen Experimenten zu arbeiten. Letztere können entweder von ihm selber in früheren Sitzungen angelegt worden sein oder es kann sich um Experimente handeln, die ein anderer Benutzer angelegt und veröffentlicht hat, wobei von dritten veröffentlichte Experimente erst kopiert werden müssen, bevor sie verändert werden können. Eine Statusanzeige gibt jederzeit über die zuletzt bearbeiteten bzw. zur Ausführung gebrachten Experimente Auskunft (Abbildung 2).

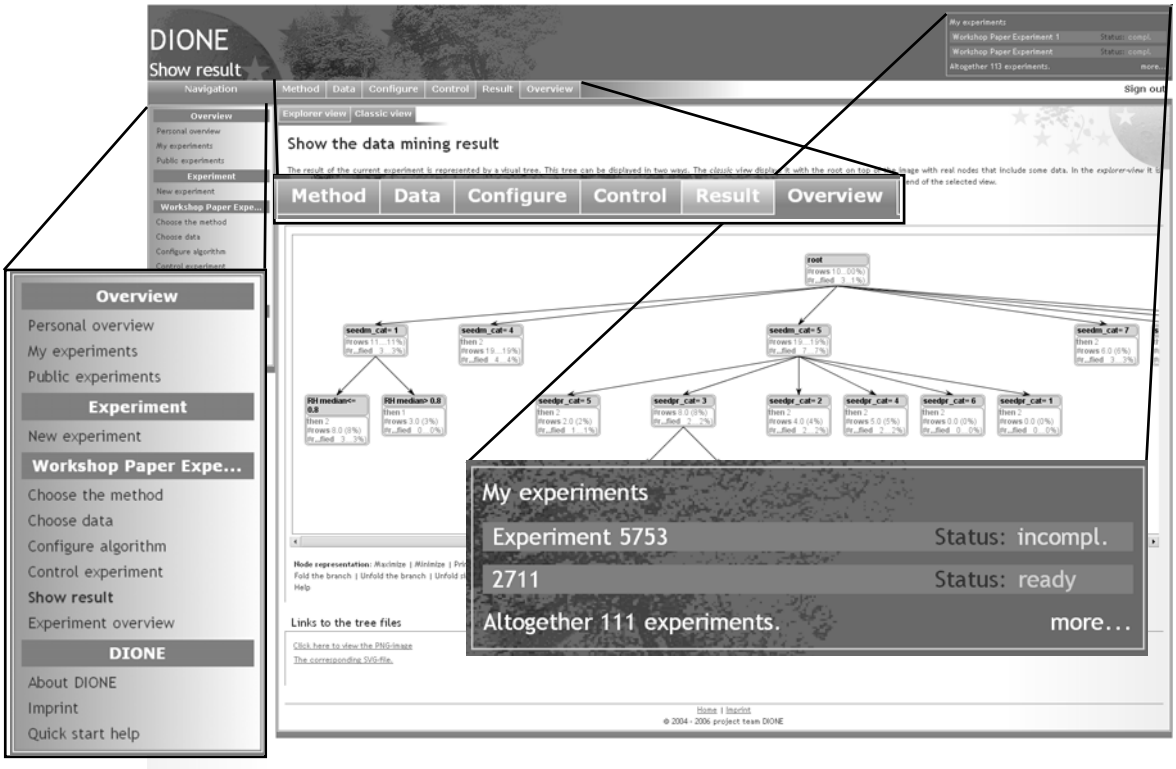

Abbildung 2: Elemente zur Auswahl von Experimentkategorien und zur Statusanzeige

Nachdem der Anwender zwischen den Experimenttypen Data Mining (Model Creation), Test (Model Verification) und Klassifizierung (Prediction) ausgewählt hat,

wird er schrittweise durch die Konfiguration eines Experiments, dessen Ausführung und die Ansicht der Resultate geführt. Die einzelnen Schritte sind Auswahl eines Algorithmus, Auswahl von Daten, Konfiguration des Algorithmus, Start des Experiments und Resultatsansicht (Abbildung 2). Bei fertigen Experimenten können Veränderungen in beliebiger Reihenfolge vorgenommen werden.
Startet der Benutzer die Auswertung eines Experiments, so wird es an einen Scheduler zur Verwaltung der (unter Umständen zeitintensiven) Data-Mining-Tasks weitergegeben, um im Hintergrund ausgeführt zu werden. Dies ermöglicht das ununterbrochene Arbeiten, auch während Experimente zur Ausführung kommen und gewährleistet bei einem hohen Aufkommen an Jobs eine faire Aufteilung von Ressourcen zwischen den Anwendern des Systems.

6.1 Visualisierung von Ergebnissen

Nach Abschluß eines Experiments wird dies dem Benutzer im Statusfenster angezeigt, wobei aufgrund der Verwendung von Standard Browsern als Anzeigemedium das Neuladen der Seite erforderlich ist, um die Statusmeldung zu aktualisieren. Experimentresultate werden je nach Experimenttyp und verwendeten Algorithmus auf verschiedene Weise dargestellt.
Entscheidungsbäume werden graphisch wiedergegeben, wobei die Knoten jeweils Information über die Entscheidungen und die Blattknoten zusätzlich Information über die Klassen, die sie repräsentieren, enthalten. Es existieren zwei alternative Ansichten für Entscheidungsbäume, von denen eine Ansicht den Baum in Form einer durch Einrückungen strukturierten Liste darstellt und sich somit auch zur Visualisierung großer Bäume eignet. Die andere, „klassische“ Baumansicht ist in Abbildung 2 im Hintergrund zu erkennen. Beide Baumdarstellungen ermöglichen das Ausblenden aller Elemente unterhalb von Knoten, womit es möglich ist, die Anzeige auf Teile des Baumes zu beschränken.
Die aus dem Clustering hervorgegangenen Konzepte werden als vollständige Liste der Datensätze mit Clusterzuordnungen angezeigt. Zusätzlich erfolgt eine statistische Analyse, deren Kennwerte für jedes Cluster in einer Tabelle wiedergegeben werden (vgl. Abbildung 3).

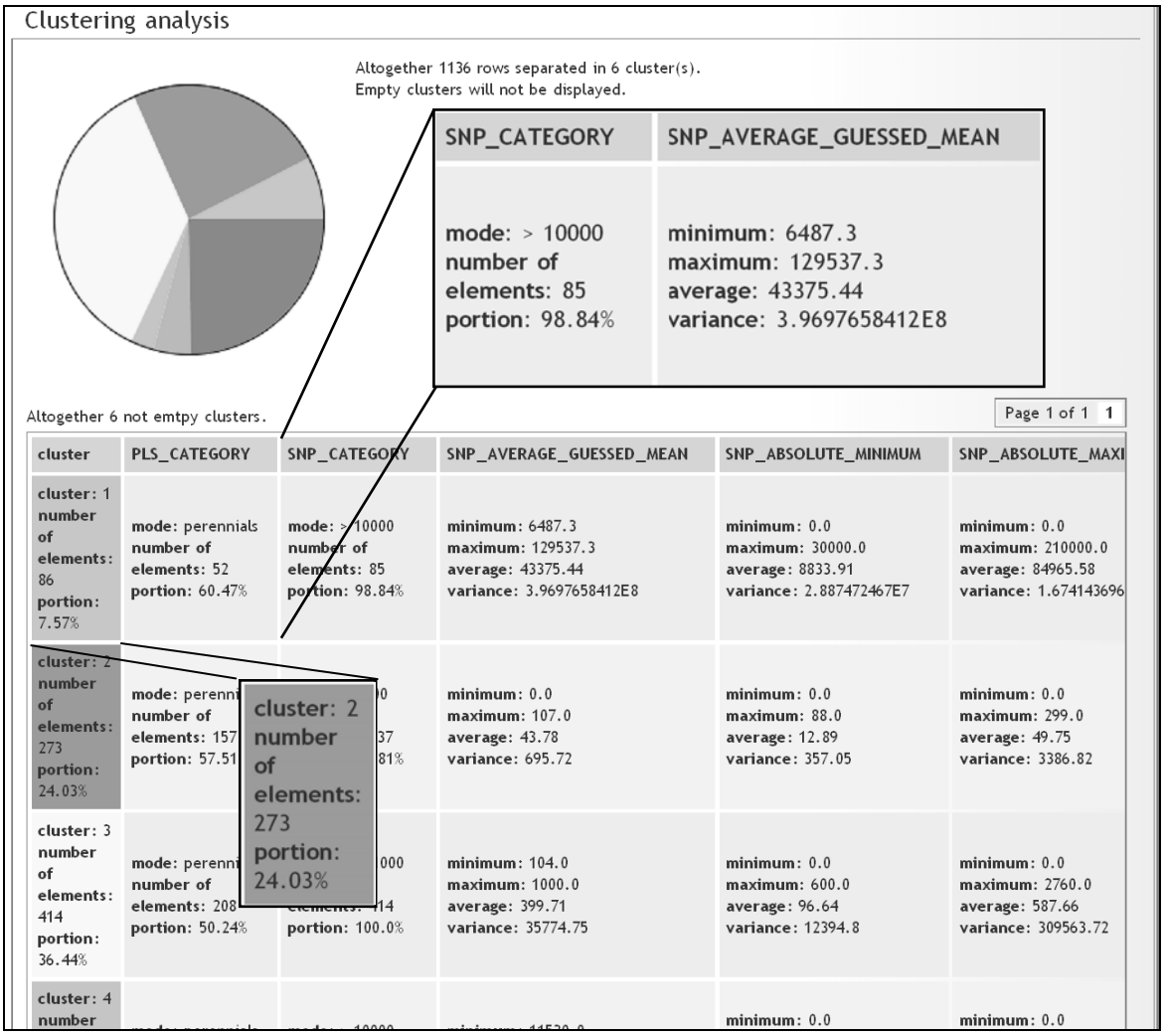

Abbildung 3: Tabellarische Wiedergabe statistischer Kenngrößen einer Clusteranalyse

Die Resultate von Verifikationsexperimenten werden in zwei Matrizen angezeigt, welche die absoluten und relativen Häufigkeiten der richtigen und falschen Zuordnungen der Testdatensätze zu jeder Klasse wiedergeben. Zusätzlich wird die prozentuale Anzahl korrekter Klassifikationen hervorgehoben, so dass eine schnelle Vorabschätzung des Testergebnisses möglich ist (Vgl. Abbildung 4).

Die Ergebnisse eines Klassifikationsexperiments schließlich werden tabellarisch ausgegeben, wobei jedem Eingabedatensatz eine Klasse zugeordnet ist.

Es besteht die Möglichkeit, die Ergebnisse eines Experiments in Form eines Berichts als PDF-Datei zu exportieren. Diese Berichtsdatei enthält neben Meta-Daten zu den analysierten Datensätzen und dem verwendeten Algorithmus sämtliche Parameterwerte zur Ausführung des Data-Mining-Verfahrens und natürlich eine Darstellung des Ergebnisses, die ähnlich der Darstellung auf dem Bildschirm ist.

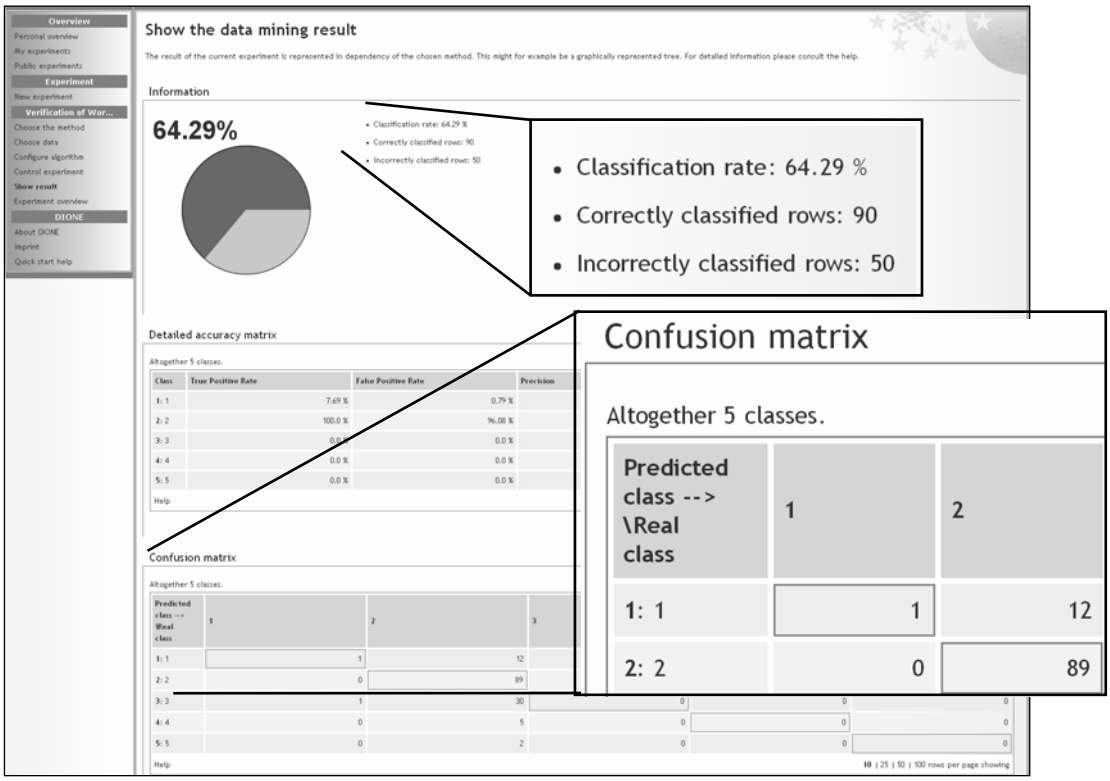

Abbildung 4: Ergebnis eines Verifikationsexperiments

6.2 Import von Dateien mit kommaseparierten Werten

Die Daten für die Erstellung von Klassifikatoren sowie etwa zu klassifizierende Daten liegen in der Praxis in den unterschiedlichsten Datei- und Datenbankformaten vor. Ein Speicherformat, das jedoch von allen gängigen Datenbankmanagement-systemen und Tabellenkalkulationsprogrammen unterstützt wird beruht auf komma-separierten Werten (CSV). Jeder Datensatz hat dabei eine konstant vorgegebene Anzahl von Einträgen, die durch Kommata oder ähnliche Trennzeichen voneinander abgegrenzt sind. Leere Einträge werden durch zwei direkt aufeinanderfolgende Trennzeichen angegeben.

Der DIONE Data Miner bietet eine komfortable Unterstützung des Imports kommaseparierter Daten in Form eines Wizards. Hiermit können die notwendigen semantischen Informationen wie der Datentyp einer Zelle, Spaltenbeschriftungen sowie die Art der Behandlung von Werten bei der Datenanalyse (Interpretation als rein beschreibender Wert, nominaler Wert, numerischer Wert oder als textueller Wert) festgelegt werden, bevor die Übertragung der Werte in eine Traitmatrix erfolgt.

Im Zusammenhang mit dem Import von Daten für die Klassifikation ist es möglich,

durch DIONE eine kommaseparierte Datei generieren zu lassen, welche die notwendigen semantischen Informationen bereits enthält und nur noch – etwa mit einem Tabellenkalkulationsprogramm - ausgefüllt werden muß um dann mit wenigen Bedienungsschritten importiert werden zu können. Dieses Vorgehen ist sinnvoll, weil bei der Klassifikation bereits feststeht, welche Struktur die zu klassifizierenden Datensätze haben müssen.

Eine Standardsituation, die beim DIONE Data Miner den Import von Daten erfordert, ist die Erzeugung von Klassifikatoren zur Erklärung bzw. Klassifikation von Clustern, die durch eine vorherige Datenanalyse entdeckt wurden. Hierzu ermöglicht das System die Ausgabe des Clustering-Ergebnisses in eine Datei mit kommaseparierten Werten. Diese Datei kann anschließend als Eingabe für einen Entscheidungsbaum oder Regelmengen erzeugenden Algorithmus importiert werden.

7 Bisherige Ergebnisse

Der DIONE Data Miner wurde durch Ökologen und Biologen aus dem LEDA Projekt bereits mehrfach mit Erfolg für die Datenanalyse genutzt. Unter anderem konnte eine existierende Analyse zur Relevanz bestimmter Pflanzeneigenschaften für die Gefährdung von Pflanzenarten verfeinert werden und bei einer Analyse von Daten zur Überdauerungsfähigkeit von Samen im Boden sowie zu Samen- und Fruchtgewichten konnte eine Erklärung für eine Unregelmäßigkeit im sonst geltenden Zusammenhang zwischen Fruchtgröße und Überdauerungsfähigkeit der Samen im Boden gefunden werden. Weiterhin konnte eine Anzahl von bekannten Zusammenhängen zwischen Pflanzenmerkmalen bestätigt werden, was wichtig war, um die prinzipielle Eignung von Methoden des Data Mining für die Datenanalyse im Kontext der Pflanzenökologie gegenüber Biologen nachzuweisen. Für eine genauere Aufschlüsselung der Ergebnisse sei auf (Bekker und Kwak 2005, Stadler et al. 200x) verwiesen.

8 Verfügbarkeit des DIONE Data Miners

Ein öffentlicher Zugriff auf den DIONE Data Miner ist zum Zeitpunkt der Abfassung dieses Berichts noch nicht möglich, wird aber im Zusammenhang mit der Veröffentlichung der LEDA Traitbase unter www.leda-traitbase.org ab Herbst 2006 geplant. Die Software ist in diesem Kontext frei nutzbar. Prinzipiell kann die DIONE-Software auch für andere Anwendungen im Bereich der Ökologie und Umwelt-

wissenschaften verwendet werden, jedoch wird zum Betrieb derzeit ein Oracle Datenbanksystem oder eine Microsft SQL Server Installation benötigt, die nur kommerziell erhältlich sind. Für die Anpassung an weitere Datenbanksysteme sowie andere Erweiterungen wird der Quellcode auf Anfrage unter den Bedingungen der GPL gerne zur Verfügung gestellt.

Danksagung

Das EU-Projekt LEDA wurde im fünften Rahmenprogramm unter Contract No. EVR1-CT-2002-40022 gefördert.

Wir danken besonders den Mitgliedern der studentischen Projektgruppe, die DIONE entwickelt hat: B. Bensien, R. Hackelbusch, S. Heisecke, N. Henze, R. Hilbrands, H. Kraef, J. Künnemann, P. Kuhn, D. Meyerholt, F. Postel und H. Tschirner.

Literatur

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

Clementine (WWW link): http://www.spss.com/de/clementine/

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

Knevel, I.C., Bekker, R.M., Bakker, J.P., Kleyer, M., 2003. Life-history traits of the Northwest European flora: The LEDA database, Journal of Vegetation Science 14: 611-614

Han, J., Kamber, M., 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers

Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecological Modelling 146: 303-310

Stadler, M., Sonnenschein, M., 2005. LEDA Traitbase - Eine Datenbank funktionaler Merkmale von Pflanzen Nordwest Europas. In: A. Gnauck (Hrsg.): Modellierung und Simulation von Ökosystemene - Workshop Kölpinsee 2003. ASIM Mittelung AMB 87, ISBN 3-8322-4560-X, pp. 144-156

Stadler, M., Ahlers, D., Bekker, R.M., Finke, J., Kunzmann, D., Sonnenschein, M., 200x. Web-basaed tools for data analysis and quality assurance on a life-histroy trait database of plants of Northwest Europe. Environmental Modelling & Software, in print

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

Witten, I.H., Frank, E., 2000. Data Mining, Morgan Kaufmann Publishers

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