HOM-E/O-MINING
Data Mining
für zu Hause, für die Wirtschaft und für die Politikberatung?!

LÁSZLÓ PITLIK, GÖDÖLLÖ - UNGARN

Abstract

On the market for analysis, offers are to be found exclusively for large-scale enterprises. The conception of HOM-E/O-MINING offers the possibility, to study the too mystical Data Mining projects in Excel (for everyone) itself. A lot of studies on the field of market and business analysis refer on it, that training and learning processes (on the base of white box methods) can be helpful to understand the potential of the Data Mining technologies.

1 Einführung

Ähnlich, wie im Fall der integrierten betrieblichen Informationssysteme gibt es auch am Markt der neuesten Analysewerkzeuge (Data Mining Tools) einen spürbaren Mangel an preiswerten, jedoch leistungsfähigen/flexiblen und dazu noch transparenten (white box) Angeboten. Die Konzeption (Methodik) von HOM-E/O-MINING bietet die Möglichkeit, die oft zu mystisch dargestellten Data Mining Prozesse in Excel selber zu studieren. Zahlreiche Fallstudien in den Bereichen der Politikberatung und Wirtschaftsprognosen sprechen dafür, die Ausbildung und Weiterbildung der Analyseexperten, sowie die Stärkung ihrer Kreativität und ihres Verständnisses über die Data Mining Prozesse viel ausgeprägter ist, wenn transparente, induktive „white box" Techniken und passende Kurse zum Selber-Lernen/„Basteln" vorhanden sind. In dieser Darstellung werden zunächst unter Punkt 2 gewisse Projekte beschrieben, deren Ergebnisse zur Idee der HOM-E/O-MINING führten.

2 Projekte

FKFP (1995-1999)

Vor 5 Jahren wurde in Kooperation deutscher und ungarischer Experte ein Fragebogen ausgearbeitet. Der umfangreiche Fragebogen (als Leitfaden für Interviews) versucht Schritt für Schritt abzutasten, wie die potentiellen Entscheidungsträger, bzw. Berater über die systematische Vorbereitung der Entscheidungen denken. Hierbei wurden insgesamt 775 Fragebögen ausgefüllt (1996) und ausgewertet (1997-1999). Die detaillierten Ergebnisse sind unter http://miau.gau.hu/miau/16/fkfp.doc zu finden. Das Projekt wurde vom Landwirtschaftsministerium, sowie vom TEMPUS und OTKA finanziert. Die wichtigsten Aussagen lauteten wie folgt:

- die Befragten sind mit sich selbst bezüglich der Güte der Entscheidungsvorbereitung und -findung zufrieden, obwohl sie kaum auf die Ressourcen der Informatik (Daten, Modelle, IT) zurückgreifen und nur ab und zu ihre Prognosen über die Zukunft notieren, um sie der Wirklichkeit gegenüber später kontrollieren zu können.
- Somit besteht nicht die Möglichkeit, für die Befragten nachzuweisen, in welchen Bereichen sie relativ unsichere Prognosen erstellen (wenn überhaupt) und daher falsche Entscheidungen treffen. Diese Tatsache schließt beinahe aus über alternative Wege (z.B. Dato Mining) , begründet Angebote zu formulieren.
- Es gehört leider zum Gesamtbild, sowohl die Anzahl und Qualität der zugänglichen Datenbasen als auch die routinemäßig eingeübten Analysetechniken Defizite aufweisen und somit

die Verhaltensmuster der Befragten viel mehr als eine Zwangsbahn als bewusst gewählte Strategie bezeichnet werden kann.

STOCKNET (1997-1998)

Um zu zeigen, das Niveau der Technologie bereits viel mehr erlaubt, als man denkt und kennt, wurde das Online Börsenanalyse-Programm STOCKNET erstellt. Hierbei kann der Client online bestimmen, welche Börsendaten er weiterverarbeiten will. Die Daten werden zwar abgefragt, jedoch nicht an den Clienten heruntergeladen. In ähnlicher Weise kann der Client unter Methoden und Einstellungen wählen. Die Modelle werden ebenso auf der Server-Seite aktiviert. Der Client erhält jedoch die gewünschten Ergebnisse (Tabellen, Abildungen) darüber, wie sich die Kurse der ausgewählten Papiere in dem vorgegebenen Zeitraum ändern werden. Solche Techniken könnte man u.a. dafür verwenden, die Landwirte oder Berater die Zeitreihen des Market Information Systems nach eigenen Wünschen weiter verarbeiten. Ähnliche Analysen sind auch für die Testbetriebsdaten zu realisieren. Weitere Informationen sind unter http://miau.gau.hu/miau/ zu finden. Das Projekt wurde von einer Börsen-Consultingfirma (EcoControl GmbH) finanziert.

OTKA (1999-2002)

Die Ergebnisse mit den Börsendaten bei Stocknet sowie die parallel laufenden Analysen zu den EU-Beitrittsverhandlungen Ungarns wiesen darauf hin, dass das Analysepotential der neuen Techniken auch im Bereich der EU-Datenbasen zu testen ist. Hierfür wurde ein Projekt bei dem Ungarischen Forschungsfonds (OTKA) beantragt. Das Ziel des Projektes ist es, anhand von den SPEL-Daten und aufgrund von modernen Analysemethoden zu zeigen, welche Fragestellungen (diverse Prognosen für Mengen, Flächen, Preise, Verwendungen, etc.) mit einem akzeptierbaren Niveau zu handhaben sind. Die detaillierten Ergebnisse des ersten Jahres sind unter http://miau.gau.hu/miau/19/otkastudy.doc zu lesen. Hervorzuheben sind folgende Aussagen:

- die sog. „bekannten statistischen Methoden" sind immer mit den neuen Ansätzen (WAM - http://miau.gau.hu/miau/19/marcius.html) zu „besiegen".
- Falls nur halbautomatische Suchstrategien verwendet werden, können die Analyseexperten ihre eigene Hypothesen prüfen und per Hand verfeinern.
- Die Ergebnisse (induktive Expertensysteme) sind stabil, transparent sowie robust und erreichen meistens bereits akzeptierbare Trefferquoten auch im Test.
- Die mit WAM festgelegten Zusammenhänge sind auch als Online-Expertensystem anzubieten: http://miau.gau.hu/oszr/index.html

IKTABU (2000-2001)

Im Jahre 1999 wurde ein Konsortium gebildet, in dem sich die Teledatacast GmbH und die AgroConsult GmbH (Universität Gödöllö) vertreten sind. Die Aufgabe des Konsortiums ist es: Anhand von primären Daten solche Online-Analysetools und Navigationssysteme auszuarbeiten, welche die tagtägliche Arbeit der SAPARD-Microregionen, sowie der potentiellen Experten unterstützen. Die geplante Dienstleistung nennt sich ikTAbu. Im Wortspiel ist zunächst die Abkürzung des Auftraggebers (OMFB IKTA – Anwendungsprojekte für Kommunikations- und Informationstechnologie beim ungarischen Entwicklungsfond OMFB) zu identifizieren, ausserdem ein Hinweis auf das Thema: Anwendung von Data Mining Technologien für die Entwicklung ländlicher Räume. Das Systemplan wurde Anfang Sommer 2000 erstellt und angenommen. Die grobe Struktur der geplanten Dienstleistungen ist unter http://miau.gau.hu/iktabu/iktabu2.html zu lesen. Wichtige Informationen sind hierbei:

im Projekt werden keine primären Daten veröffentlicht, da die juristische Lage der Wiederverwendung von Rohdaten unsicher ist,

- es werden ausschließlich Analysen angeboten (Ähnlichkeiten, Rangfolgen, Prognosen, Empfindlichkeitsanalysen, DEA-Analysen, Online-Expertensysteme, etc.)
- die Dienstleistung wird als Prototyp vom MIVIR (s. unten) betrachtet.
- die Dienstleistung basiert auf Oracle Technologien und Web-basierter Kommunikation.

MIVIR (2000-)

Parallel den Hintergrundprojekten hat sich das Landwirtschaftsministerium entschlossen, ein integriertes Informationssystems (MIVIR) für die Landwirtschaft, sowie für den Bereich Entwicklung Ländlicher Räume zu entwickeln. In der Konzeption sind folgende Thesen zu lesen:

- Der Zugang zu den gemeinnützigen Daten und somit die Möglichkeit zu deren Analysen müssen juristisch und technisch neu konzipiert werden.
- Die Qualität der Datenbasen muss anhand von Konsistenzanalysen gesteigert werden (vgl. SPEL).
- Es müssen Kataloge entstehen die das gesamte Datenvermögen transparent machen.
- Für das öffentliche Informationssystem müssen die bereits stabil funktionierenden Muster (DWH, OLAP, etc.) der Informationssysteme bei den Grossunternehmen zugrunde gelegt werden.

Eine deutschsprachige Zusammenfassung der Konzeption ist unter der Adresse http://miau.gau.hu/miau/20/asa2.doc zu finden

IDARA (2000-2003)

Im Vorjahr wurde bei der Kommission der EU ein 6-Länder Projekt beantragt, welches ab 2000 für drei Jahre bewilligt wurde. Das Projekt hat das Ziel, im Bereich Entwicklung ländlicher Räume mannigfaltige Analysen zu erstellen, u.a. im Bereich der SPEL-basierten Simulationen. Hierfür werden zunächst die verfügbaren statistischen Daten mit Expertenschätzungen ergänzt und einem Konsistenzcheck unterzogen, um anschließend die Folgen verschiedener Scenarien ableiten zu können. Die IDARA Internet-Seite befindet sich unter http://www.agp.uni-bonn.de/agpo/rsrch/idara/idara_e.htm

3 HOM-E/O-MINING

Damit kommen wir zu dem Kernpunkt der Darstellung. Bevor die rätselhafte Abkürzung erklärt wird, sollte zunächst kurz die Kernidee beschrieben werden: in Zukunft werden immer mehr Online-Datenbasen zugänglich sein. Parallel dazu werden sicherlich immer mehr Online-Analyse-Dienste entstehen. Die Entscheidungsträger können diese beiden Ressourcen gut verwenden, da sie bislang auch ohne gut strukturierten Datenbasen und ohne raffinierte, ausgereifte Methoden zurecht gekommen sind, indem sie heuristisch oder intuitiv nach Ähnlichkeiten in ihren abstrakten Welten der Erfahrungen gesucht haben. Dank der impulsiven Entwicklungsphasen im Bereich der KI-s besteht allmählich die effiziente Möglichkeit, die spontanen Entscheidungen systematischer zu machen.

Hierbei sollte endlich die vielleicht „komisch" erscheinende Abkürzung HOM-E/O-MINING näher erläutert werden: Ein Teil des Begriffes kommt eindeutig aus dem Begriff DATA MINING. Die übrigen Teile (HOM-E/O) tragen drei Bedeutungsschichten in sich:

HOME steht für zu Hause und symbolisiert eine Technik/Methodik, die so leicht, transparent und verständlich ist, dass man sie ohne weiteres zu Hause verwenden kann. Und wofür: bei der Auswahl von Autos/Immobilien, etc. (Musterfrage: Weist ein Preis auf eine über- oder unterbewertete Situation hin?)

HOMO steht für das Mensch und symbolisiert, dass die Methodik soweit übersichtlich ist, dass jeder seine eigenen Ideen/Hypothesen sowohl bei der Fragestellung als auch bei der Lösung wiederfinden kann und dazu noch ständige Lerneffekte angeboten sind.
HOMEO steht für das Gleichgewicht, welches zwischen dem Mensch und der Maschine gefunden werden sollte, um effizient kooperieren können. Das Mensch ist improvisativ, assoziativ und evt. heuristisch. Im Gegensatz dazu ist eine Maschine monoton, schnell und systematisch. Sobald jemand ein Problem lösen möchte, steht vor einem Antagonismus: Darf ich improvisieren oder sollte viel mehr systematisch vorgehen? Es gibt natürlich keine exakte Antwort auf diese Frage. Es gibt jedoch unterschiedliche menschliche Charakterzügen, die eine Kooperation mit den Maschinen gut oder weniger erfolgreich unterstützen.
Unter http://miau.gau.hu/miau/21/njszt.doc ist eine realistische Vision beschrieben. Die ungarische Gesellschaft für Informatik hat die Mitglieder aufgefordert, eine Situation zu beschreiben, in der wir uns alle in 20 Jahre befinden werden.
Laut der Idee der HOM-E/O-MINING sollten in den nächsten Jahren immer mehr Daten online erreichbar sein und parallel dazu auch Online-Analysetools angeboten werden. Falls inzwischen die Ausbildung in Richtung Analysetechniken immer stärker ausgeprägt wird, kann erwartet werden, dass die potentiellen Anwender die Vorteile der Online Analysen Schritt für Schritt wahrnehmen. Dazu sind zunächst solche „white box Methoden" notwendig, die leistungsstark, jedoch bereits für die Anfänger übersichtlich sind und keine zusätzlichen Programme erfordern, sondern z.B. in Excel reproduzierbar sind. So eine Methode ist das Modell WAM. Die Erfahrungen weisen deutlich darauf hin, dass ein Teil der Studenten solche latente Fähigkeiten besitzt die ihnen ermöglichen, anhand von größeren Datenmengen und einigen Methoden effizient und erfolgreich Prognosen erstellen zu können. Nach dieser Lernphase entsteht eine Art Bereitschaft, online angebotene Datenbasen und Methoden zu nutzen. Die Online-Tools haben u.a. den Vorteil, dass sie vermutlich nie auf den Software-Schwarzmarkt kommen und somit keinerlei Copyrights-Probleme verursachen. Andererseits können diese Methoden in einem ständigen Wettbewerb weiterentwickelt werden. Im Wettbewerb kann notiert werden, anhand welcher Daten und Methodeneinstellungen welche Ergebnisse erreicht worden sind.
Es ist jedoch von Anfang an klar, dass ein idealisierter GPS (General Problem Solver) vermutlich nie existieren wird und im Bereich der Problem+Lösung-Paare die Zusammenhänge viel zu chaotisch sind, um über dem Optimum sprechen zu können. Es gibt jedoch immer mehr Daten und typische Fragestellungen, in denen es sich nicht mehr lohnt, die klassische menschliche Vorgehensweise (d.h. Improvisation) zu wählen. Um aber mit den Maschinen und mit den Methoden kooperieren zu können, braucht der Mensch eine Art Trainingsphase, in der jeder selbst sehen kann, ob die Fähigkeiten (spezielle IQ) gerade bei ihm vorhanden sind oder nicht.
In 20 Jahren also kann man evt. über eine deutliche Spaltung der Gesellschaft sprechen. Auf der einen Seite stehen diejenigen die bereit und fähig sind, mit Daten und Methoden zu jonglieren, auf der anderen Seite sind diejenigen, die viel mehr an die Intuition glauben. Falls hierbei sich eine Art spontanes Gleichgewicht einstellt, dann kann man über eine neue, positive Erscheinung der Informationsgesellschaft sprechen, da durch eine Symbiose (Mensch+Maschine) immer mehr Fehlentscheidungen vermieden werden können.

4 Literatur - s. http-Einträge im Text