Einführung in Data-Mining mit analytischen Funktionen und R

Vladimir Poliakov
Nürnberg

Schlüsselworte
Analytics, Statistik, OLAP, Data-Mining, R, R Software, R Commander, RStudio, Rattle Package, analytische Funktionen

Einleitung
Durch die riesige Auswahl an Paketen für die Datenanalyse, Statistik und Visualisierungen ist die R Software mittlerweile zum Standardwerkzeug für Datenauswertungen geworden. Als Open Source Projekt macht R starke Konkurrenz zu den kommerziellen Produkten. Dank der vielen Schnittstellen kann R die Daten aus verschiedenen Datenquellen lesen, z.B. aus den CSV-Dateien oder aus einer Oracle Datenbank. Dabei sind die Oracle analytischen Funktionen ein mächtiges Tool für die Vorbereitung der Daten zur Analyse via R.

Der Vortag beschäftigt sich nicht mit der R Sprache innerhalb der Oracle Datenbank und ist eine Einführung ins Data-Mining Verfahren mit Hilfe der Software R. Es wird gezeigt wie R Software Desktop-Variante für die Datenbankzugriffe konfiguriert wird. Außerdem wird das R Paket Rattle präsentiert, das die grafische Oberfläche für einige Data-Mining Verfahren bietet.

Data-Mining
Laut Wikipedia [1] versteht man unter dem Begriff „Data-Mining“ (Daten-Bergbau) die systematische Anwendung statistischer Methoden auf einen Datenbasis mit dem Ziel, neue Trends zu erkennen. Die Methoden dieses Verfahrens kommen aus der Statistik, dem maschinellen Lernen, der klassischen Mustererkennung und wurden teilweise bereits vor Jahrzehnten entwickelt. Dieses „Data-Mining“ Analyseverfahren wird oft im Handel benutzt, aber im Prinzip kann man diese Methoden überall einsetzen, weil sie von der Herkunft der Daten unabhängig sind.

Aufgaben des Data-Mining:
1. Ausreißererkennung (Abweichungsanalyse)
2. Clusteranalyse
3. Klassifikation
4. Assoziationsanalyse
5. Regressionsanalyse

Mögliche Einsatzbereiche für das Data-Mining Verfahren:
- Bankwesen: Kreditwürdigkeit, Betrugserkennung
- Handel: Warenkorbanalyse
- Industrie: Optimierung Poduktions- und Fertigungsprozesse (z.B. bei Herstellung der Windrotoren etc.)
- Energieversorgung: Effizienz der Anlagen (z.B. Effizienz der Windanagen)
- IT: Kapazität Management, Ausfall der Serverkomponenten (z.B. Festplatten) abhängig von den Temperaturen im RZ

- Medizintechnik, Technik allgemein: prognostizierbare Ausfälle der Geräte
- Sonstiges: Vorhersage im Sportevent

Die letzte Behauptung klingt zuerst unrealistisch, aber das lässt sich dank vielen modernen Analyse-Tools relativ schnell prüfen. Im Rahmen der Ausarbeitung dieses Vortrags wurden die Daten zwei Eishockey Ligen (NHL und DEL) aus den öffentlichen Datenquellen genommen und analysiert. *Letztendlich lautete die Data-Mining Frage irgendwie so: Ist die Anzahl der Tore im Spiel von der vor dem Spiel herrschenden Tordifferenz beziehungsweise von der Gegentordifferenz beider Gegner abhängig?* Damit sollte das trockene Thema lebendiger gestaltet werden.

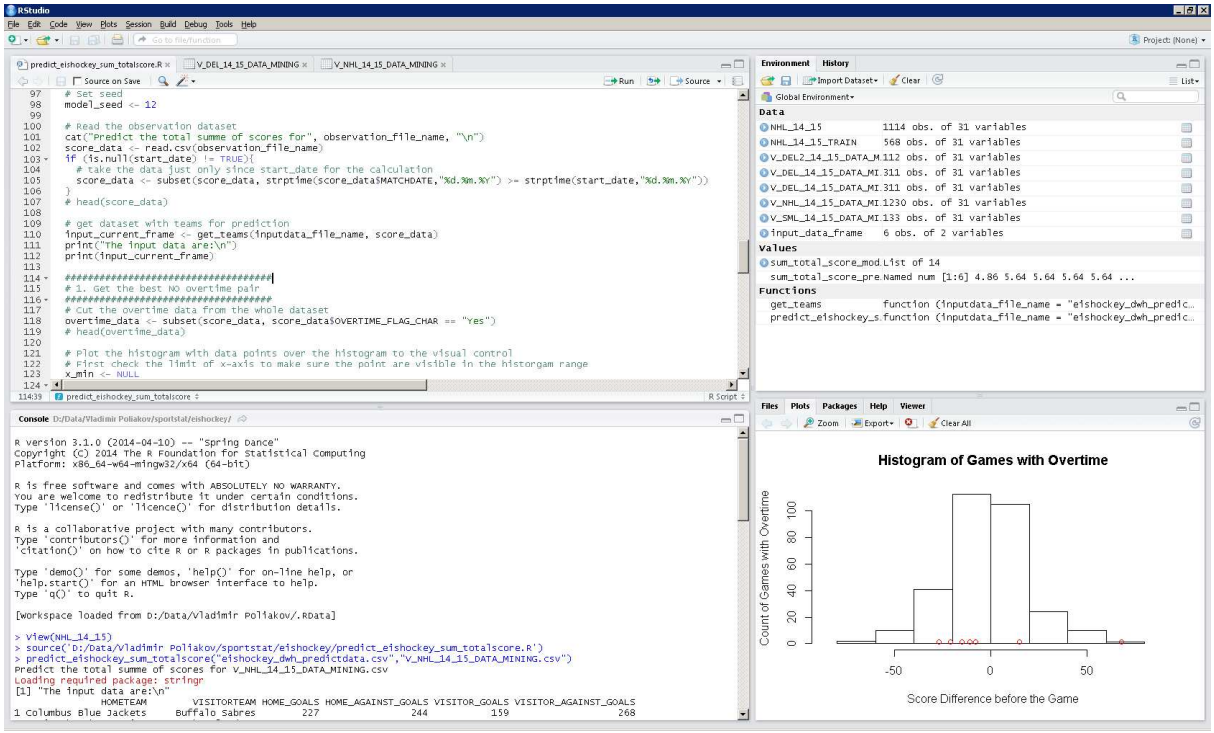
Ein typisches Data-Mining Prozess sieht folgendermaßen aus:
1. Frage für die Data-Mining Analyse formulieren
2. Daten vorbereiten (finden, bereinigen und ins Data-Mining Tool laden)
3. Die Verteilung der Daten analysieren
4. Die Variablen auswählen, das Datamodell mit den Trainingsdaten bilden
5. Ergebnisse der Berechnung überprüfen, interpretieren und auf die Testdaten anwenden.

R

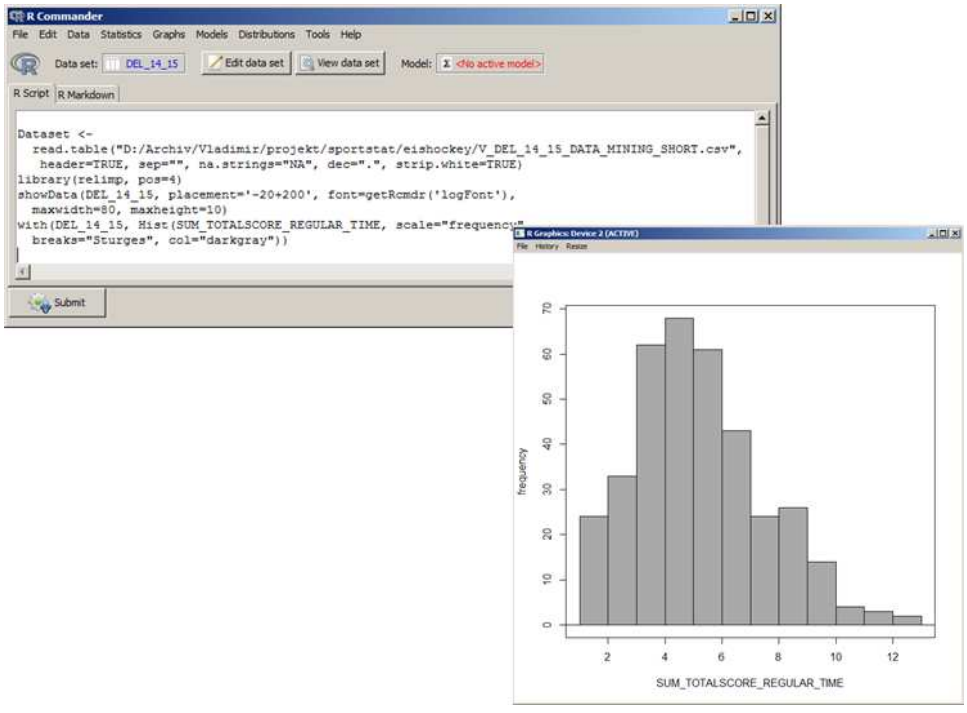
Zu den wichtigsten Analyse-Tools zählt ohne Zweifel R. Das ist eine freie Software, die auf Windows, MacOS, vielen Linux Plattformen läuft und eine Menge der wissenschaftlichen Paketen bzw. Bibliotheken für Statistik und Data-Mining Verfahren hat. Einige von ihnen wie Rcmdr (R Commander) oder Rattle (R Data Miner) haben grafische Benutzeroberflächen, was die Einarbeitung in die R-Sprache vereinfacht und die Arbeit mit den Daten erleichtert.

Bei diesem Vortrag wurden folgende graphische Benutzeroberflächen verwendet:

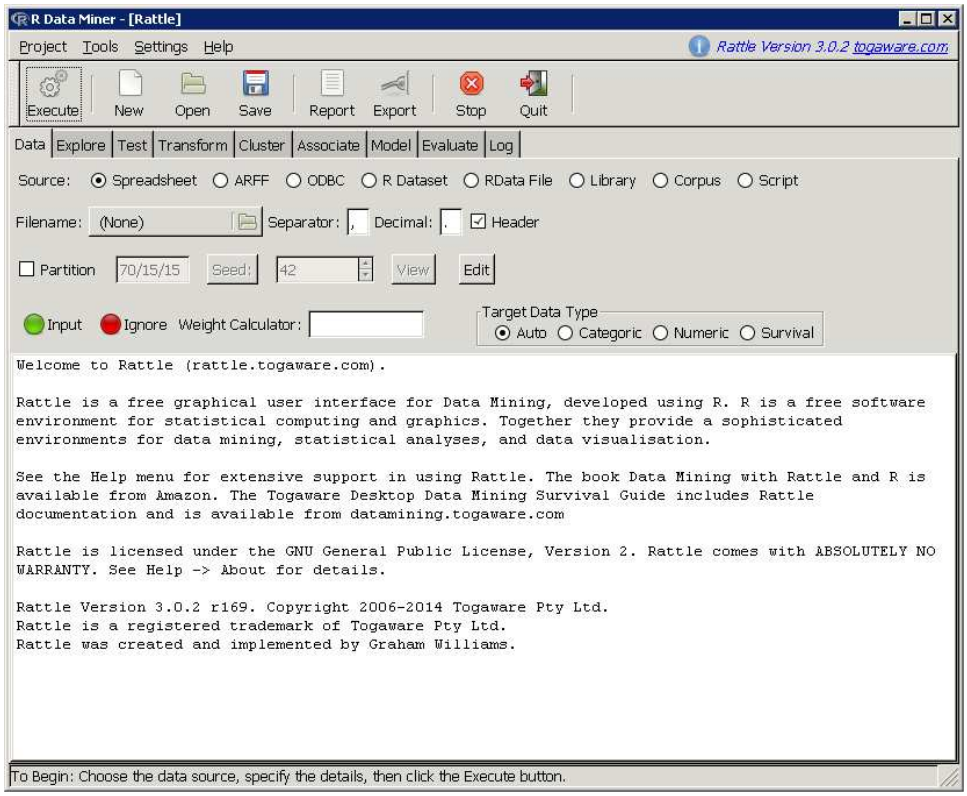
RStudio
R Commander
Rattle (R Data Miner)



RStudio



R Commander



Rattle (R Data Miner)

Data-Mining-Prozess Ablauf

Ist die Frage festgelegt bzw. ist das Ziel der Analyse gesetzt, kann man mit der Datenvorbereitung beginnen. Das ist ein der wichtigsten Teile des Data-Mining Analyse Prozesses, weil diese Vorbereitung bis zu 80% Zeit des gesamten Data-Mining Analyse Prozesses aufnehmen kann. Häufig wird diese Phase als ETL (Extract, Transform und Load) bezeichnet. Während dieser Phase werden die Daten aus den verschiedenen Datenquellen (operative Datenbanken, Sozialnetzwerke, Log-Dateien etc.) geholt, bereinigt (z.B. ein Geburtsdatum liegt in der Zukunft oder die Geschlechtsdaten fehlen, obwohl die Anrede bekannt ist), transformiert zum einheitlichen Format (z.B. das Datum wird im DD.MM.YYYY Form abgespeichert oder die personenbezogene Daten werden anonymisiert) und für die weitere Bearbeitung im Zielsystem (Datawarehouse, Hadoop Filesystem als Data Lake etc.) abgelegt. Ist die die ETL-Phase abgeschlossen, können die Inputdaten in Analytical Records [4] zusammengefasst werden, falls das während der ETL-Phase nicht gemacht wurde.

Der Analytical Record ist ein Input, der von jedem Data-Mining-Tool (egal ob R Software oder ein anderes Tool) erwartet wird. Dabei stellt jede Zeile ein Fall, eine Spalte der vorherzusagende Wert und die anderen Spalten die Eigenschaften des Falles, anders gesagt, die Prädiktoren, dar. Es handelt dabei um einige Vorbereitungsprozesse. Oft sind das die aufwendigen Manipulationen mit den operativen oder historischen Daten (Gruppieren, Summieren etc.), die mit Hilfe der analytischen Funktionen (in allen Oracle Database Edition verfügbar) performant im Analytical Record View zusammengefasst werden (s. unten):

- `FUNCTION_NAME(column|expression …) OVER (Order-by-Clause)`
- `LAG(…) OVER (PARTITION BY… ORDER BY…)`

- `SUM(…) OVER (PARTITION BY… ORDER BY…)`

Ist das Analytical Record View vorbereitet, können die Daten in R Software geladen werden. R-Tool präferiert per Default comma-separated-values File (CSV-Datei), kann aber über Schnittstelle die Daten direkt aus der Oracle Datenbank lesen. Dafür werden der Oracle Instant Client und zusätzliche R Bibliotheken (ROracle, RODBC, RJDBC etc.) benötigen. Das Konfigurieren des Datenbankzugriffes via ODBC wird im Vortrag näher eingegangen. Für die Basis-Statistik [2] ist ein anderes Tool - R Commander - sehr gut geeignet.

Der letzte Schritt vor dem Modellaufbau ist die Häufigkeitsverteilung Analyse, weil viele Modelle die Normalverteilung der Daten voraussetzen. Es gibt verschiedene Techniken und Verfahren in der Statistik zur Beschreibung der Verteilung der Daten [5], die im Rahmen dieses Vortrages nicht näher beleuchtet werden. Es wird für das Generalisierte Lineare Modell (GLM) angenommen, dass die Daten der Poisson Verteilung [3] unterliegen. Diese Hypothese ist der Grundstein für den Modellaufbau im R Data Miner. Außer GLM wird im R Data Miner das Entscheidungsbaum Modell (Decision Tree Model) präsentiert. Die mathematischen Grundlagen für die Lineare und Multiple Regression sowie für die Signifikanz der Modelle werden nur am Rande aus Zeitmangel erwähnt.

Fazit

Das Data-Mining Verfahren hilft beim Untersuchen verschiedenen Beziehungen im Datenbestand, was mit der Standard-Auswertung (Statistik und / oder OLAP) praktisch nicht möglich ist, da diese Verfahren nur bestimmte und vorher festgelegte Fragen beantworten. Die analytischen Funktionen helfen dabei, die Daten für die Analyse vorzubereiten. Auch für diejenigen, die mit dem Data-Mining nicht zu tun haben, lohnt es sich diese Funktionen kennenzulernen. Sie können gute Dienste im täglichen SQL-Leben leisten.

Der Vortrag ist eine kleine Crash-Einführung ins Data-Mining Verfahren und soll überwiegend als Denkanstoß dienen. Z.B. was kann ich aus meinen Daten gewinnen? Oder was für eine Auswirkung hat eine oder andere Input Variable? Dafür sollte man R Data Miner [6] besser kennenlernen und sich mit der Signifikanz der Modelle tiefer auseinander setzen. Zum Üben findet man die Rohdaten im Internet. Auf dem online Datenwissenschaft Portal Kaggle [7] gibt es genug davon. In der ersten Linie werden dort die Public Data-Mining Wettbewerbe ausgetragen, aber die Neulinge und Anfänger können dort auch ein paar gute Übungen mit der Anleitung (meistens nur auf Englisch) finden. Fühlt man sich später in R Sprache bzw. R Software sicherer, verzichtet man oft auf die manuelle Bearbeitung via R GUIs und beschleunigt man die Analyse mit Hilfe von R-Skripten.

Verwendete Quellen

1. Data-Mining in Wikipedia
   https://de.wikipedia.org/wiki/Data-Mining
2. Bernd Weiler, DOAG 2014, Einführung in die Statistik mit R
3. Heuer, Andreas
   Der perfekte Tipp: Statistik des Fußballspiels (Erlebnis Wissenschaft)
   1. Auflage September 2012, Wiley-VCH, Weinheim Verlag
   ISBN 978-3-527-33103-1
4. Zeitschrift iX Developer Big Data 2015 - Analytics Design Patterns
5. Dormann, Carsten F.
   Parametrische Statistik : Verteilungen, maximum likelihood und GLM in R
   Springer Spektrum, 2013
   ISBN 978-3-642-34785-6

6. Williams, Graham
Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R)
Auflage: 2011 (4. August 2011), Springer Verlag
ISBN 978-1-441-99889-7
7. Online platform Kaggle for predictive modelling and analytics competitions
http://www.kaggle.com

Kontaktdaten:

Vladimir Poliakov absolvierte 1995 die Hydrometeorologische Hochschule in St. Petersburg mit dem Schwerpunkt mathematische Modellierung und arbeitete im Forschungsinstitut für Arktis und Antarktis. Nach seiner Auswanderung nach Deutschland war er seit 1998 in verschiedenen Softwarehäusern als Entwickler tätig. Er ist Oracle Certified Professional und seit 2007 der führende Oracle DBA bei AREVA GmbH.

Telefon: +49 157 87611972
E-Mail: v.poliakov@gmx.net