Einführung in Data-Mining mit analytischen Funktionen und R

Vladimir Poliakov, Nürnberg

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

Laut Wikipedia [1] versteht man unter dem Begriff „Data-Mining" (Daten-Bergbau) die systematische Anwendung statistischer Methoden auf eine Datenbasis mit dem Ziel, neue Trends zu erkennen. Die Methoden dieses Verfahrens kommen aus der Statistik, dem maschinellen Lernen sowie 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. Die Aufgaben des Data-Minings sind:

- Ausreißer-Erkennung (Abweichungs-Analyse)
- Cluster-Analyse
- Klassifikation
- Assoziations-Analyse
- Regressions-Analyse
- Ein typischer Data-Mining-Prozess sieht dabei folgendermaßen aus:
- Frage für die Data-Mining-Analyse formulieren
- Daten vorbereiten (finden, bereinigen und ins Data-Mining-Tool laden)
- Die Verteilung der Daten analysieren
- Die Variablen auswählen und dann das Datamodell mit den Trainingsdaten bilden
- Ergebnisse der Berechnung überprüfen, interpretieren und auf die Testdaten anwenden

Mögliche Einsatzbereiche für das Data-Mining-Verfahren sind:

- *Bankwesen*
  Kreditwürdigkeit, Betrugserkennung
- *Handel*
  Warenkorb-Analyse
- *Industrie*
  Optimierung von Produktions- und Fertigungsprozessen (Optimierung der Stahlproduktion etc.)
- *Energieversorgung*
  Effizienz der Anlagen (etwa bei Windanlagen)
- *IT*
  Kapazitäts-Management, Ausfall der Server-Komponenten (wie Festplatten) abhängig von den Temperaturen im RZ
- *Medizintechnik, Technik allgemein*
  prognostizierbare Ausfälle der Geräte
- *Sonstiges*
  Vorhersage im Sportevent

Die letzte Behauptung klingt zunächst unrealistisch, aber das lässt sich dank vieler moderner Analyse-Tools relativ schnell überprüfen. Und das Thema ist eigentlich nicht neu, es gibt bereits sogar ein wissenschaftliches Buch über Statistik im Fußball [2]. Für diesen Artikel wurden die Daten zweier Eishockey-Ligen (NHL und DEL) aus den öffentlichen Datenquellen übernommen und analysiert. Dabei könnte die Data-Mining-Frage lauten: „Ist die Anzahl der Tore im Spiel von der Stärke der Mannschaften abhängig?" Da die Mannschaftsstärke durch die Anzahl der Tore beziehungsweise der Gegentore bestimmt werden kann, wird die Data-Mining-Frage jedoch letztendlich so lauten: „Ist die Anzahl der Tore im Spiel von der vor dem Spiel herrschenden Tordifferenz beziehungsweise von der Gegentor-Differenz beider Gegner abhängig?" Ist die Frage festgelegt und somit das Ziel der Analyse gesetzt, kann man mit der Datenvorbereitung (siehe oben) beginnen.

Das ist einer der wichtigsten Teile des Data-Mining-Analyse-Prozesses, weil diese

Vorbereitung bis zu 80 Prozent der Zeit des gesamten Data-Mining-Analyse-Prozesses einnehmen kann. Häufig wird diese Phase als „ETL" (Extract, Transform, Load) bezeichnet. Während dieser Phase werden die Daten aus den verschiedenen Datenquellen (operative Datenbanken, soziale Netzwerke, Log-Dateien etc.) geholt, bereinigt (beispielsweise ein Geburtsdatum liegt in der Zukunft oder die Geschlechtsdaten fehlen, obwohl die Anrede bekannt ist), zum einheitlichen Format transformiert (etwa das Datum wird in „DD.MM.YYYY"-Form abgespeichert oder die personenbezogenen Daten werden anonymisiert) und für die weitere Bearbeitung im Zielsystem (Data Warehouse, Hadoop Filesystem als Data Lake etc.) abgelegt. Ist die ETL-Phase abgeschlossen, lassen sich die Inputdaten in Analytical Records [3] zusammenfassen (etwa in Form einer View), falls das während der ETL-Phase nicht gemacht wurde.

Der Analytical Record ist ein Input, der von jedem Data-Mining-Tool (egal ob R oder ein anderes Tool) erwartet wird. Dabei stellen jede Zeile einen Fall, eine Spalte den vorherzusagenden Wert und die anderen Spalten die Eigenschaften des Falles, anders gesagt die Prädiktoren, darx.

Es handelt sich dabei um einige Vorbereitungsprozesse. Oft sind das die aufwändigen Manipulationen mit den operativen oder historischen Daten (Gruppieren, Summieren

```
LAG(GOALS,1,0) OVER (PARTITION BY TEAM_ID ORDER BY MATCHDATE, TEAM_ID) PREV_GOALS
SUM(PREV_GOALS) OVER (PARTITION BY TEAM_ID ORDER BY MATCHDATE, TEAM_ID) SUM_PREV_GOALS
```

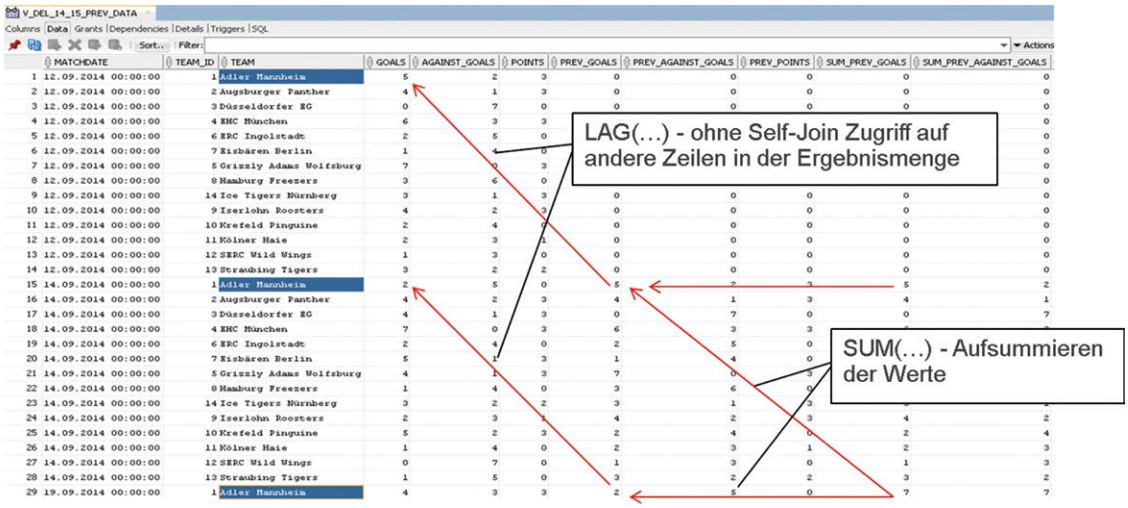

Abbildung 1: Die analytischen Funktionen im Einsatz

```
FUNCTION_NAME([Argumente]) OVER ([analytische Klause])
```

Listing 1

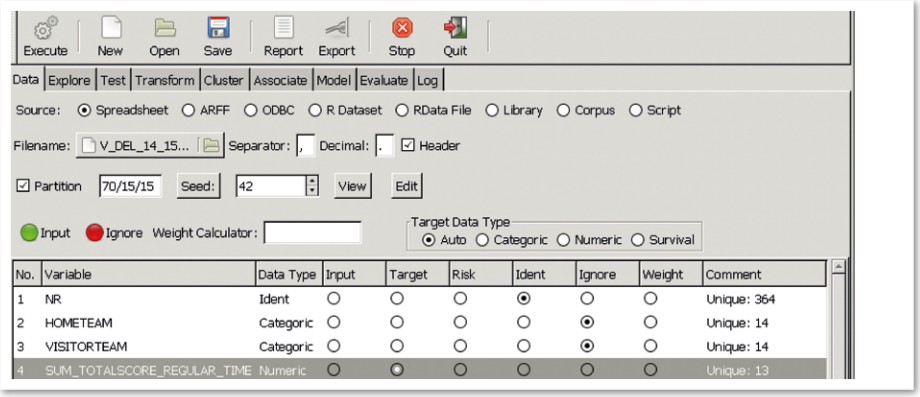

Abbildung 2: Definieren der Variablen in Rattle

etc.), die mithilfe der analytischen Funktionen (in allen Oracle Database Editions verfügbar) performant in der Analytical-Record-View zusammengefasst sind. *Listing 1* zeigt schematisch dargestellt die generelle Syntax des Aufrufs einer analytischen Funktion [4].

Die analytischen Funktionen fassen wie die Gruppen-Funktionen die Ergebnismenge zusammen, aber es wird für jede Zeile der Ergebnismenge einzeln ein Wert ausgegeben. Das Ausgabe-Format einer analytischen Funktion entspricht in Prinzip dem Analytical Record. Je nachdem können viele Funktionen als Gruppenfunktionen oder als analytischen Funktionen ausgeführt werden *(siehe Abbildung 1)*.

Ist die Analytical-Record-View vorbereitet, können die Daten in R geladen werden. R ist eine freie Programmiersprache, die für die Datenanalyse 1992 an der Universität Auckland in Neuseeland entwickelt wurde. Es ist eine Skriptsprache und wird nicht kompiliert, sondern in der R-Console interpretiert. Das ist jedoch kein K.o.-Kriterium. Es gibt bereits genug grafische Benutzeroberflächen, wie R-Studio, die einen Einstieg in die R-Welt ermöglichen. R-Studio kann die Daten zur Analyse importieren und hat einen Editor, die R-Console, und vieles mehr.

Die R-Tools (R-Studio, R Commander, R Data Miner etc.) präferieren per Default „comma separated values"-Files (CSV-Datei), können aber über Schnittstellen die Daten direkt aus der Oracle-Datenbank lesen. Das ist keine große Kunst, denn dafür sind lediglich der Oracle-Instant-Client sowie zusätzliche R-Bibliotheken je nach Zugriffsart (ROracle, RODBC, RJDBC etc.) erforderlich.

Der letzte Schritt vor dem Modell-Aufbau ist die Analyse der Häufigkeitsverteilung, 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 Artikels nicht näher beleuchtet werden. Es wird für das generalisierte lineare Modell (GLM) angenommen, dass die Daten der Poisson-Verteilung [2] unterliegen. Diese Hypothese ist der Grundstein für den Modellaufbau im R Data Miner.

R Data Miner oder kurz „Rattle" ist ein R-Paket, das die grafische Benutzeroberfläche hat und speziell für die Data-Mining-Analyse geschrieben ist. Es kann die Daten nicht nur aus den CSV-Dateien lesen, sondern auch aus Excel-Tabellen, ODBC-Quellen, R-Datasets und anderen Formaten. Wichtig ist, dass jede Aktion in Rattle explizit mit der Schaltfläche „Execute" bestätigt werden soll. Sind die Daten ins Tool geladen, kann man die Variablen identifizieren und mit dem Modellaufbau beginnen *(siehe Abbildung 2)*. Wie bereits erwähnt wurde, wird für das generalisierte lineare Modell (GLM) angenommen, dass die Daten der Poisson-Verteilung [2] unterliegen. Entsprechend dieser Aussage wird in Rattle das Modell aufgebaut *(siehe Abbildung 3)*.

Wie man sieht, sind die Input-Variablen für das Modell nicht signifikant, was bedeu-

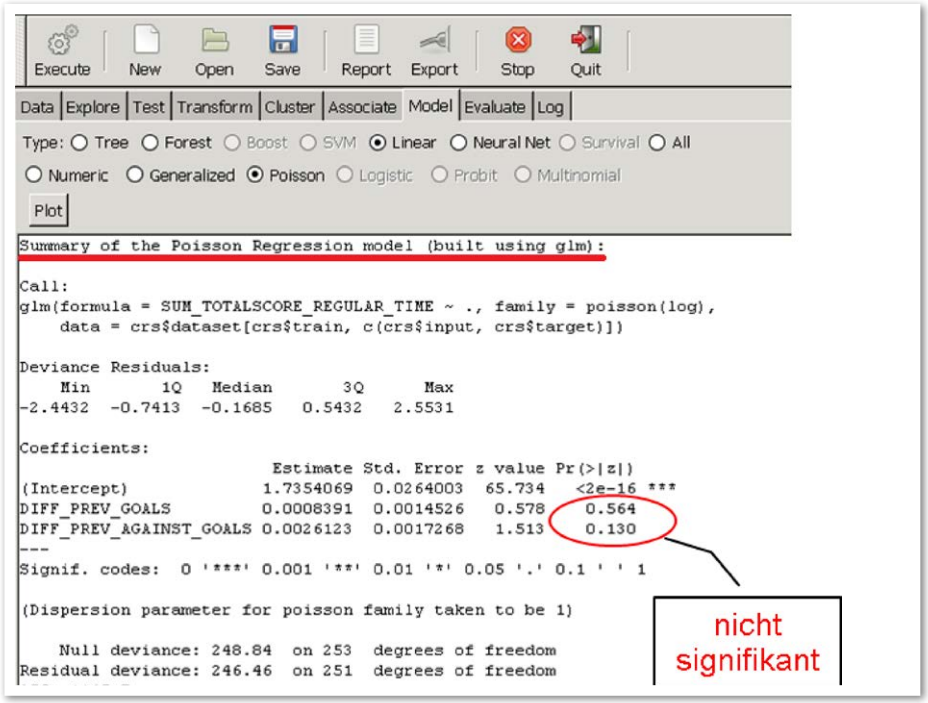

Abbildung 3: Das GLM für die Anzahl der Tore pro Spiel in der DEL Saison 2014/2015

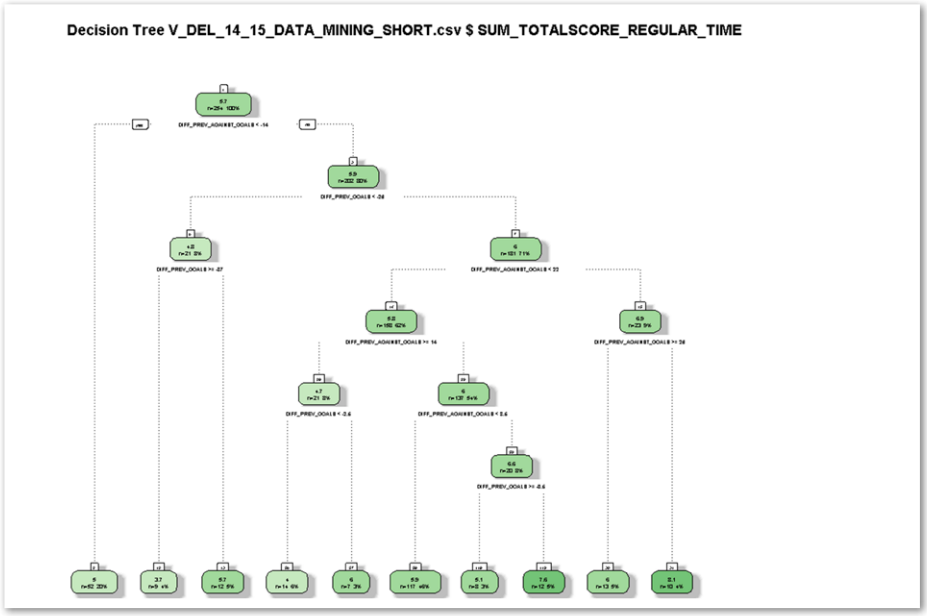

Abbildung 4: Das Entscheidungsbaum-Modell für die Anzahl der Tore pro Spiel in der DEL-Saison 2014/15

tet, dass es in der Deutschen Eishockey-Liga in der Saison 2014/15 keine Abhängigkeiten zwischen der Anzahl der Tore im Spiel und der Tordifferenz beziehungsweise der Gegentor-Differenz der Mannschaften vor dem Spiel gibt. Die Situation in der National Hockey League ist in der gleichen Saison anders [6]. Wahrscheinlich liegt es daran, dass in der DEL meistens nur freitags und sonntags gespielt wird, dagegen finden in der NHL die Spiele fast täglich statt und so langsam setzt sich das Gesetz der großen Zahlen durch.

Wenn die Input-Variablen für das Modell nicht signifikant sind, bedeutet das letztendlich, dass es einfach nicht die lineare Abhängigkeit gibt, die die Menschen so mögen. In diesem Fall können die Daten auf die anderen Modelle angewendet werden. Zum Beispiel auf das Entscheidungsbaum-Modell (Decision Tree Model), das Rattle defaultmäßig bietet. Dieses Modell ist leichter zu interpretieren und außerdem ist es robuster gegenüber Ausreißern. Jeder Zweig des Entscheidungsbaum-Modells präsentiert die zugrunde liegende Bedingung und das prozentuale Auftreten des Ereignisses. Die Entscheidungsregeln können entweder in Textformat im Rattle-Output oder graphisch im R Studio *(siehe Abbildung 4)* dargestellt werden.

Im Endeffekt sollen die bereits geleisteten Schritte auch leicht reproduziert werden. Das ist ohne großen Aufwand möglich, weil Rattle alle Aktionen im Log-Output protokolliert. Diese Log-Ausgabe ist nichts anderes als eine Reihe von R-Befehlen, die in der Console ausgeführt werden. Wenn der Data Scientist oder der Daten-Bergbauer mit den Ergebnissen seiner Arbeit zufrieden ist, kann alles als R-Skript abgespeichert und zum späteren Zeitpunkt mit den Test-Daten ausgeführt werden.

Fazit

Das Data-Mining-Verfahren hilft beim Untersuchen verschiedener 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 nichts zu tun haben, lohnt es sich, diese Funktionen kennenzulernen. Sie können gute Dienste im täglichen SQL-Leben leisten.

Dieser Artikel ist eine kleine Einführung in den Daten-Bergbau. Wie bereits am Anfang erwähnt, kann das Data-Mining-Verfahren in den verschiedensten Bereichen eingesetzt werden. Man muss aber tiefer in die Statistik, in die Verteilung der Daten einsteigen und sich mit der Signifikanz der Modelle auseinandersetzen. Die Ergebnisse der Analyse können dann verblüffend sein.

Referenzen

[1] Data-Mining in Wikipedia: https://de.wikipedia.org/wiki/Data-Mining

[2] Heuer, Andreas: Der perfekte Tipp – Statistik des Fußballspiels (Erlebnis Wissenschaft) 1. Auflage September 2012, Wiley-VCH, Weinheim Verlag ISBN 978-3-527-33103-1

[3] Zeitschrift iX Developer Big Data 2015 – Analytics Design Patterns

[4] Analytische Funktionen: http://www.muniqsoft.de/tipps/sql/sql-allgemein/analytische-funktionen.htm

[5] Dormann, Carsten F.: Parametrische Statistik – Verteilungen, maximum likelihood und GLM in R. Springer Spektrum, 2013. ISBN 978-3-642-32785-6

[6] Poliakov, Vladimir: Auf der Suche nach dem peRfekten Tipp: http://heise.de/-2776160

Vladimir Poliakov
v.poliakov@gmx.net