Einsatz des SQL Developer 3.0 und Oracle Data Miner 11g R2 beim Data Mining Cup 2011 -- Ein Erfahrungsbericht --

Enrico Bade und Rüdiger Steffan
Fakultät für Wirtschaftswissenschaften
Hochschule Wismar

Schlüsselworte:

Oracle 11gR2, Data Mining Cup, ODM, Data Miner, New Features, Workflow, Prozesse, Werkzeuge

Einleitung

In einem vorhergehenden Beitrag wurden Strategien und Erfahrungen beim Einsatz des Oracle Data Miner (OD Miner) beschrieben, insbesondere wenn ein erstes Data Mining-Projekt nur mit Grundkenntnissen in SQL und PL/SQL durchgeführt wird [1]. In diesem Beitrag werden zunächst die neuen Funktionalitäten von Oracle SQL Developer 3.0 und den jetzt darin integrierten OD Miner zusammengefasst [2]. Schwerpunkt ist die Integration von benutzerdefinierten Prozeduren in automatisierte Data Mining Prozesse, was vom neuen OD Miner durch Workflows unterstützt wird und beim Fallbeispiel Data Mining Cup 2011 von besonderer Bedeutung war. Ferner wird in diesem Zusammenhang die Verwendung der Java-Schnittstellen zum Oracle Data Mining aufgezeigt.

Beim Data Mining Cup 2011 (DMC) sollten Produkt-Items von Sessions eines Online-Shops vorhergesagt werden [3], was eigentlich einer klassischen Warenkorbanalyse entspricht. Die Schwierigkeit besteht jedoch darin, daß mehrere Items pro Session vorhergesagt werden müssen sowie in der zugrundeliegenden Datenmenge von über 9 Millionen Zeilen. In diesem Zusammenhang werden daher ebenso Performance-Aspekte und Strategien bei der Datenvorbereitung mit PL/SQL diskutiert. Es mußten eigene Prozesse für Build, Test und Score implementiert werden. Darüber hinaus sollte eine Java Anwendung zur dynamischen Verwendung der Data Mining-Modelle erstellt werden.

Schließlich werden die konkreten Ergebnisse von Studenten der Wirtschaftsinformatik an der Hochschule Wismar vorgestellt, die unter Verwendung des neuen ODMiner am Data Mining Cup 2011 teilgenommen haben [4]. Erfahrungen und hauptsächliche Probleme der Anwender werden zusammengefaßt und mit denen beim Einsatz von freien Open Source Produkten verglichen.

SQL Developer 3 und Data Miner 11gR2

Neben der Unterstützung verschiedener, mehr administrativer, Funktionalitäten zählt die vollständige Integration des Oracle Data Miners (OD Miner) und des Oracle Data Modelers mit zu den hauptsächlichen neuen Features des SQL Developers 3. Eng mit der Integration des OD Miners verknüpft ist die ebenso neue Möglichkeit, im SQL Developer mittels dem Paket DBMS_Scheduler, entfernte Datenbank-Jobs mit graphischer Unterstützung zu konfigurieren. Eine vollständige Liste aller neuen Funktionalitäten ist unter [3] zu finden.

Der freie Oracle Data Miner existierte als separates Java-Werkzeug für die Datenbankversionen 10.1 bis 11.1. Da das Werkzeug ausschließlich die Data Mining-Funktionalitäten im Datenbanksystem nutzt, kann ein Release des Werkzeugs auch nur mit einer bestimmten Datenbankversion verwendet

werden. Für Oracle 11.2 kann daher nur noch der im SQL Developer integrierte OD Miner verwendet werden. Eine Anmeldung an frühere Datenbankversionen ist nicht möglich. Das separate Werkzeug für die vorhergehenden Datenbankversionen ist jedoch nach wie vor verfügbar und heißt nun Oracle Data Miner Classic [2]. Für das Data Mining Repository existiert im Datenbanksystem der Benutzer `DMSYS` nicht mehr. Statt dessen werden die Metadaten im Schema `SYS` gespeichert.

Für die Verwendung des OD Miners werden im SQL Developer alle Datenbankverbindungen des Schema-Browsers zusätzlich in einem separaten Fenster angezeigt (siehe Abb. 1). Hier können jedoch nur die OD Miner-Funktionalitäten und keine Schemaobjekte aufgeklappt werden.

Data Mining-Projekt einrichten

Datenbankbenutzer, die für ein Data Mining-Projekt eingerichtet werden müssen, benötigen die Rolle `ODMRUSER` sowie weitere Objektprivilegien, hauptsächlich Execute-Rechte auf Scheduler-Programme im Schema `ODMRSYS`. Wird ein Benutzer im OD Miner Fenster ausgewählt, der die Rolle `ODMRUSER` nicht besitzt, können die Rechte zusammen mit Beispieldaten für Tutorials im SQL Developer eingerichtet werden. Dazu ist jedoch die Eingabe des `SYS`-Passworts erforderlich! Um dies im Rahmen einer Clientverbindung zu vermeiden und um mehrere Entwickler ohne graphische Oberfläche als Administrator vorab einzurichten, stehen die erforderlichen Skripte im Verzeichnis `sqldeveloper\odminer\scripts` zur Verfügung (z.B. `usergrants.sql`).

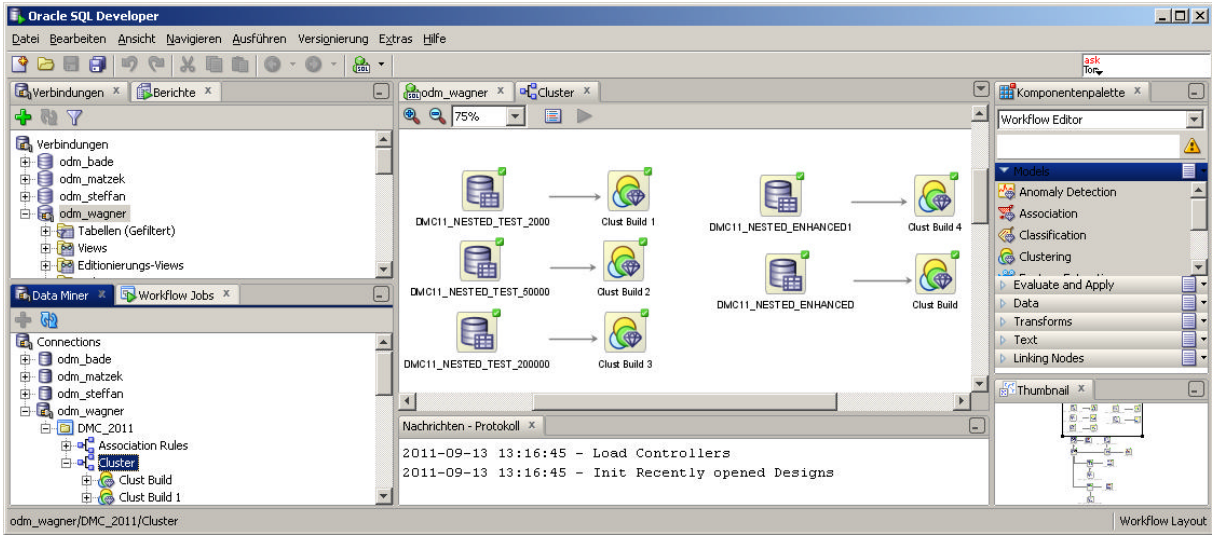

Abb. 1: SQL Developer 3 mit zwei getrennten Verbindungsfenstern (Schema Browser und Data Miner), welche jedoch dieselbe Verbindungsdefinition verwenden. In der Datenbank existieren insgesamt drei Sessions, für den Schema- und den Workflow-Browser sowie für die Workflow-Bearbeitung.

Die Sprache der OD Miner-Funktionalitäten ist noch nicht übersetzt. Für die Einarbeitung von Entwicklern mit Hilfe der Dokumentation ist ohnehin eine einheitliche englische Spracheinstellung empfehlenswert, die jedoch nach wie vor nicht über die graphische Oberfläche eingestellt werden kann. Hierzu sind in der Datei `sqldeveloper\ide\bin\ide.conf` folgende Einträge erforderlich:

```
AddVMOption -Duser.language=en
AddVMOption -Duser.country=US
```

Bereits in den Vorgängerversionen des SQL Developer 3 wurde die Funktion zum Daten laden überarbeitet, so dass nicht mehr der SQL*Loader einer lokalen Client-Installation dazu verwendet

wird. Die in [1] beschriebene Stolperfalle eines Versionskonflikts tritt daher nicht mehr auf, was beim zeitkritischen Projekt des Data Mining Cups zu signifikanten Verzögerungen führte.

Teilnahme am Data Mining Cup 2011 und Ergebnis

Der Data Mining Cup (DMC) ist ein jährlicher internationaler Wettbewerb für Studenten mit aktuellen und praxisnahen Aufgabenstellungen aus Unternehmen [4]. Der DMC 2011 hatte zwei Teilaufgaben, wobei die erste unabhängig von der zweiten gelöst werden konnte. Die Datenstruktur für die erste Aufgabe war relativ einfach und bestand aus zwei Dateien mit Train-Daten (rund 150 MB, 9.531.123 Zeilen) und Apply-Daten (rund 9 MB, 539.090 Zeilen). Trotz der großen Datenmenge war der Ladevorgang mit dem SQL Developer 3 problemlos. Auf der Basis einer Warenkorbanalyse sollten fehlende Produkt-Items vorhergesagt werden, so dass eine explizite Zielgröße fehlte. Die Train-Daten waren vollständig während bei den Apply-Daten bei jeder Warenkorbsitzung (Session) drei Items fehlten. Das Attribut TransactType gibt die Art der Warenkorbaktion wieder und mußte nicht vorhergesagt werden (0=anschauen, 1=auswählen, 2=kaufen). Beim zweiten Aufgabenteil sollte auf der Basis einer vorgegebenen Schnittstelle eine Java-Anwendung entwickelt werden, die in der Lage ist, neu hinzukommende Daten dynamisch mit einzubeziehen. Nachfolgend wird zunächst nur die Lösung der ersten Teilaufgabe unter Verwendung der neuer Funktionalitäten des SQL Developer 3 zusammengefasst.

```
SessionNo|ItemNo|TransactType
0|12864|0
0|8897|0
0|8897|1
1|11937|0
1|2208|0
1|11016|0
1|7721|0
1|12395|0
…
```

SESSIONNO	ITEMNO	TRANSACTTYPE
0	12864	0
0	8897	0
0	8897	1
1	11937	0
1	2208	0
1	11016	0
1	7721	0
1	12395	0

Abb. 2: Train-Daten beim DMC 2011 als Textdatei (links) und in eine Datenbanktabelle geladen (rechts).

Die Aufgabe stellte sich schwieriger heraus als zunächst angenommen. Dies wird rückblickend auch dadurch deutlich, dass von 103 angemeldeten Teams (aus 20 Ländern) lediglich 35 Ergebnisse für die erste Teilaufgabe eingereicht wurden, 15 für die zweite. Die Hochschule Wismar hatte mit zwei Teams teilgenommen. Das erste Team mit Master-Studenten arbeitete mit Open Source-Werkzeugen wie WEKA oder KNIME und belegte Platz 19. Die Ergebnisse des zweiten Teams bestehend aus Bachelor-Studenten mit Grundlagenkenntnissen in SQL und PL/SQL arbeiteten ausschliesslich mit Oracle und belegten Platz 24. Diese Ergebnisse bilden die Basis für den vorliegenden Beitrag [??].

Analysestrategien und Besonderheiten

Bereits beim Vergleich der Train- und Apply-Daten ist ein Unterschied deutlich geworden. Die Erkenntnis, dass die Spalten `SESSIONNO` und `ITEMNO` in den Apply-Daten als Primärschlüssel dienen, konnte nicht auf die Train-Daten angewandt werden. Dort wurden Items gemäß des Kaufverhaltens gespeichert (vgl. Abb.2, SessionNo 0, ItemNo 8897). Für die weiterführende Analyse wurde beschlossen, die Struktur der Train-Daten an die der Apply-Daten anzupassen. Dazu blieb nur die Zeile mit dem maximalen TransactType für ein Item bestehen, während die alle anderen Einträge des Items in der Session gelöscht wurden. Auf diese Weise konnten die Datensätze in den Train-Daten auf 6.615.839 reduziert werden.

Für die anschliessende Warenkorbanalyse kamen zunächst die Assoziationsverfahren in Frage, die als Ergebnis Entscheidungsregeln liefern. Anders als bei den Beispielen in den Tutorials können diese Modelle aber nicht unmittelbar mit dem Data Miner auf neue Daten angewendet werden (Apply). Vielmehr müssen zur Umsetzung der Regeln Prozeduren bzw. Funktionen erstellt werden, die dann auf neue Daten angewendet werden können.

Eine Analyse der gesamten Datenmenge scheiterte, da keine Regeln gefunden werden konnten. Daher wurde zunächst eine Clusteranalyse und dann für die einzelnen Datengruppen der Cluster separate Assoziationsanalysen durchgeführt. Zusätzlich wurde noch ein eigenes Entscheidungsmodell in Form eine PL/SQL-Prozedur entwickelt, das bei Sessions mit nur 1-3 Aktionen erfolgreicher war und daher für diese Art von Daten als Alternative verwendet wurde. Die Besonderheiten des resultierenden Gesamtmodells lassen sich wie folgt zusammenfassen:

- Die Warenkorbdaten erfordern die Verwendung von Nested Columns, die jetzt vom Oracle Data Miner 11gR2 unterstützt werden. Das neue Feature konnte erfolgreich eingesetzt werden.
- Mit Ausnahme des Clustermodells müssen die Entscheidungsmodelle in Form von Prozeduren angewendet werden (Apply). Ist dies mit Hilfe der neuen Workflows im Data Miner möglich?
- Im Workflow können Tabellen, basierend auf generierten Modellen erstellt werden. Dadurch würden die Assoziationsregeln schnell extrahiert, und für die individuellen Prozeduren aufbereitet werden.

Statt PL/SQL-Prozeduren in SQL*Plus für das Scoring (Apply) anzuwenden, sollen daher die neuen Features des SQL Developers 3 untersucht und genutzt werden.

Beispiel: Einsatz des Oracle Data Miner 11gR2 für Scoring (Apply)

Wie in Abb. 1 ersichtlich, werden bei dem neuen Oracle Data Miner die Data Mining-Berechnungen in Form von Workflows definiert, ähnlich wie bei dem Open Source-Werkzeug KNIME [6]. Für die Anwendung des Cluster-Modells ist eine Komponente explizit vorgesehen (siehe Abb. 3 `APPLY_CLUSTER`). Die Anwendung der Assoziationsregeln und des eigenen Vergleichsmodells erfolgt jedoch mittels PL/SQL-Prozeduren, für die es keine frei definierbare PL/SQL-Komponente gibt. Daher wurde statt dessen die Transform-Komponente verwendet, welches das Einbinden einer PL/SQL-Funktion als "Expression" erlaubt. Die somit erfolgreiche Implementierung des gesamten Modells als Apply-Workflow ist in Abb. 3 zu sehen.

Zunächst werden die Eingangsdaten (`DATA_INPUT`) für das eigene Modell und für das Cluster-Modell aufgeteilt, wobei die Daten für das Cluster-Modell mittels Nested Columns noch aufbereitet werden müssen. Die Vorhersagen des eigenen Modells (`WILKEN_OUT`) werden mit den Vorhersagen der einzelnen Assoziationsmodelle der Cluster zusammengeführt (`JOIN_SCORED_OUTS`). Handelt es sich bei der Berechnung um einen Testdurchlauf zur Kreuzvalidierung, dann liegen die Ergebnisse in einer separaten Tabelle vor (`DATA_EVAL`), mit denen die erwartete Punktzahl, analog zum Data Mining Cup, berechnet und gespeichert wird (`RESULT`).

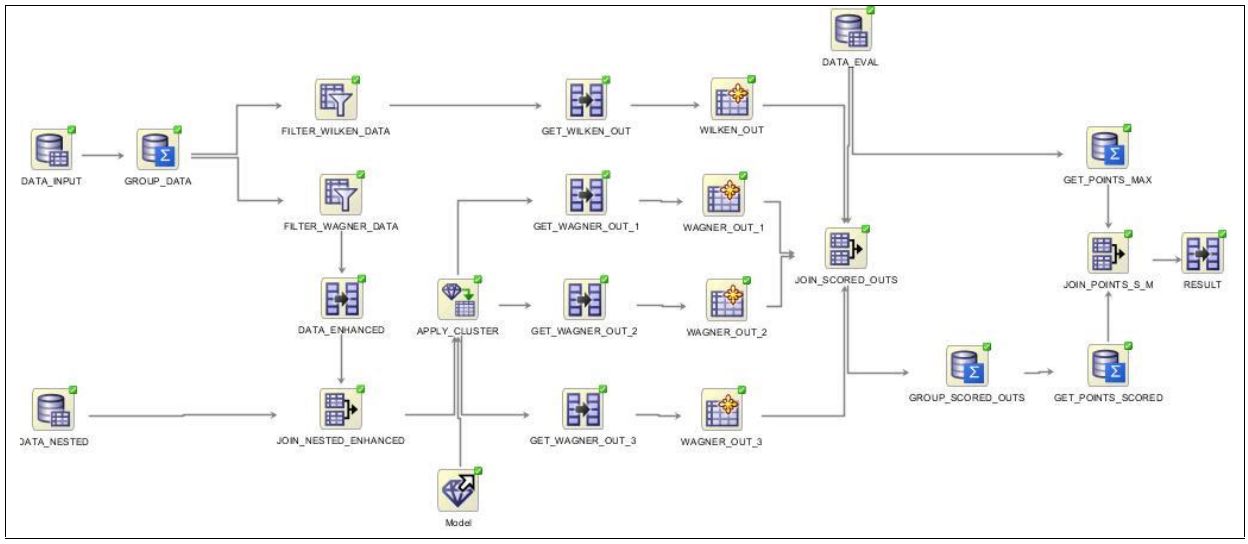

Abb. 3: Anwendung (Apply) des Data Mining Modells für den Data Mining Cup 2011 als Workflow.

Erfahrungen und Ausblick

Die Einarbeitung in das Projekt erfolgte mit dem Data Miner Classic. Während des Data Mining Cups wurde allerdings zunehmend der neue Data Miner eingesetzt, da dieser bezüglich Übersichtlichkeit und Verwendbarkeit enorme Vorteile aufweist. Die intuitive Workflow-Oberfläche ermöglicht es schnell, verschiedene Modelle zu erstellen und die Ergebnisse zu vergleichen. Diese Vorteile wurden vor allem bei der Diskussion verschiedener Cluster-Analysen im Team genutzt.

Ein automatisiertes Scoring (Apply) des Gesamtmodells wäre mit dem Data Miner Classic in dieser Form nicht möglich gewesen. Der Nutzer muss nur noch drei Aktionen vornehmen: Train- und Evaluierungs-Daten laden sowie den Workflow starten. Die Erstellung des Workflows erfordert jedoch einen relativ hohen Arbeitsaufwand zur Einarbeitung, so dass diese Vorteile in einem ersten Data Mining-Projekt unter Zeitdruck, wie im Fall des Data Mining Cups, nicht genutzt werden können. Zur einfacheren Erstellung individueller Workflows wären darüber hinaus flexible PL/SQL-Komponenten bis hin zu eigens definierbaren Modellkomponenten wünschenswert.

Des Weiteren werden die neuen Möglichkeiten zur Verwendung des Schedulers im SQL Developer 3 als Alternative zu den Workflows sowie die Java Schnittstelle zum Oracle Data Mining für die Lösung der zweiten Aufgabe des Data Mining Cups vorgestellt.

Quellenangaben

[1] Daniel Fritzler und Rüdiger Steffan: Strategien und Ergebnisse beim Einsatz des Oracle Data Miner 11g für den Data Mining Cup 2008, Proceedings 21. Deutsche ORACLE-Anwenderkonferenz (Hrsg.: DOAG e.V.), 2008.

[2] Oracle® Data Mining Administrator's Guide 11g Release 2 (11.2); Oracle Corporation Part No. E16807-06; May 2011.

[3] http://otn.oracle.com/developer-tools/sql-developer/rel3-featurelist-ea-189447.html

[4] www.data-mining-cup.de

[5] Enrico Bade, Diana Matzek, Martin Wagner, Martin Wilken; Projektarbeit im Studiengang Wirtschaftsinformatik (Bachelor); Hochschule Wismar, 2011.

[6] http://www.knime.org/

Kontaktadresse:

Rüdiger Steffan, Prof. Dr.-Ing.
Lehrgebiet Datenbank- und Datenkommunikationssysteme

Fakultät für Wirtschaftswissenschaften, Hochschule Wismar
University of Technology Business and Design
Phillip-Müller-Straße 14
23966 Wismar

Telefon: +49(0) 3841-753606
Fax: +49(0) 3841-7539606
E-Mail: ruediger.steffan@hs-wismar.de
Internet: www.wi.hs-wismar.de