Data-Mining in sozialen Online-Netzwerken

Bianca Böckelmann
Robotron Datenbank-Software GmbH
Dresden

Schlüsselworte

Data-Mining, Oracle Data Mining, Knowledge Discovery in Databases, indirektes Bestimmen privater Informationen, soziale Online-Netzwerke, Facebook, Schutz der Privatsphäre

Einleitung

In sozialen Online-Netzwerken kommen naturgemäß mit der Zeit viele personenbezogene Daten zusammen. Durch die zusätzliche Bekanntgabe von Freundschaftsbeziehungen entsteht ein sozialer Graph vernetzter privater Attribute. Im Rahmen einer Masterarbeit an der Brandenburgischen Technischen Universität Cottbus in Kooperation mit der Robotron Datenbank-Software GmbH und der IHP GmbH wurde 2012 eine „Untersuchung von Data-Mining-Algorithmen zur indirekten Bestimmung privater Attribute unter Berücksichtigung graphenbasierter Strukturen“ durchgeführt. Hierbei wurde der Frage nachgegangen, inwieweit sich in sozialen Graphen durch Verknüpfung von öffentlichen Informationen der Freunde Rückschlüsse auf die privaten Daten eines Nutzers ziehen lassen, ohne dass diese Daten explizit offengelegt wurden.

Mit dem Einsatz von Oracle Data Miner (ODMr) wurden Data-Mining-Algorithmen auf Daten freiwilliger Facebook-Nutzer angewendet, mit dem Ziel Risiken für die Privatsphäre eines Einzelnen aufzudecken. Im Folgenden wird zuerst ein Überblick über die Schritte des zugrunde gelegten Knowledge Discovery in Databases (KDD)-Prozesses gegeben. Nach einer kurzen Einführung in Oracle Data Mining wird der Prozess praxisnah am Beispiel des Forschungsthemas beschrieben. Neben der Anwendung der eigentlichen Data-Mining-Algorithmen für die Prognose beinhaltet der KDD-Prozess die vorherige Vorverarbeitung und Transformation der Nutzerdaten, welche zeitlich einen Großteil am Gesamtprozess einnahmen. Abschließend werden die Ergebnisse vorgestellt, welche die Vermutungen bekräftigen, dass auch die Betrachtung der Freundesinformationen wichtig ist, um seine Privatsphäre zu schützen.

Der KDD-Prozess

Knowledge Discovery in Databases (KDD) ist der (semi-) automatische [1] nicht triviale Prozess des Identifizierens von gültigen, bisher unbekannten, potentiell nützlichen und verständlichen Mustern in Daten [2]. Der KDD-Prozess besteht aus mehreren Phasen, die iterativ durchlaufen werden können, wie es die Abbildung 1 zeigt. Zu Beginn wird bei der Selektion das Ziel festgelegt und die Daten für die Wissensextraktion herangezogen und bezüglich ihrer Qualität beurteilt [1]. Im nächsten Schritt werden die Daten für das Erzielen von Integrität, Konsistenz und Vollständigkeit vorverarbeitet. Anschließend folgt die Transformation der vorverarbeiteten Daten in eine für den Analyseschritt geeignete Repräsentation. Darunter zählt unter anderem die Auswahl relevanter Attribute. In der eigentlichen Data-Mining-Phase werden die Algorithmen für die Extraktion von Mustern angewendet. Abschließend werden die erzielten Ergebnisse interpretiert und hinsichtlich des festgelegten Ziels bewertet. Sind die gefundenen Modelle gut, kann das hieraus gewonnene Wissen beispielsweise gezielt für das Ableiten von Handlungsdirektiven eingesetzt werden.

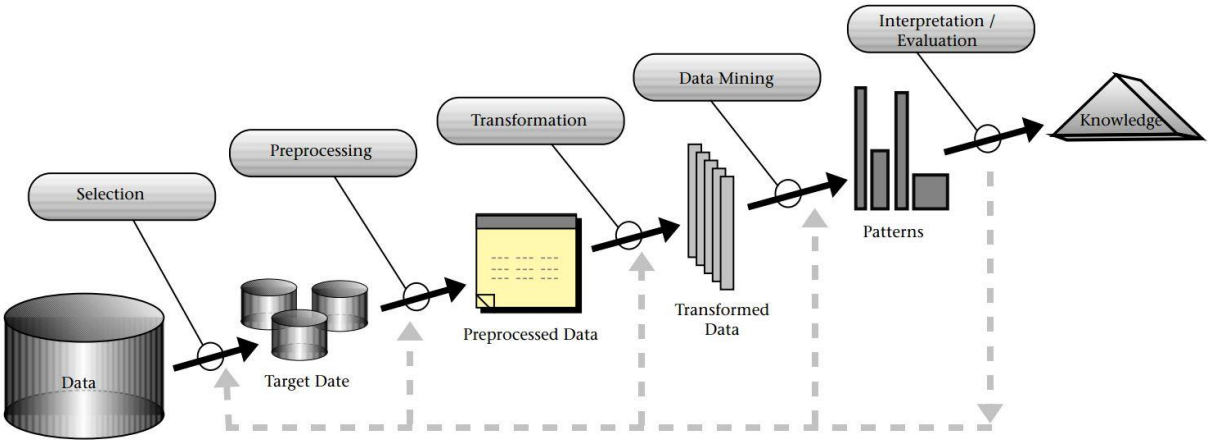

Abbildung 1: Schritte im KDD-Prozessmodell (entnommen aus **[3]***)*

Oracle Data Mining

Oracle bietet mit Oracle Data Mining (ODM) als Teil der Oracle Advanced Analytics Option die Möglichkeit, Data-Mining-Algorithmen direkt in der Oracle Datenbank auszuführen. Für einen Zugriff auf die Quelldaten, die Modelle sowie die Analyseergebnisse ist somit kein zusätzlicher Datentransport nötig. Auf die ODM-Funktionalitäten kann über PL/SQL, SQL und R zugegriffen werden [4]. Des Weiteren wird eine grafische Benutzeroberfläche namens Oracle Data Miner (ODMr) bereitgestellt, die eine Erweiterung des SQL Developers ist. Mittels Workflows, die sich aus Knoten und Verbindungen zusammensetzen, kann der gesamte KDD-Prozess abgebildet werden. Ein einfacher Workflow ist in der Abbildung 2 zu sehen.

Neben einfachen Statistiken, die einen Überblick über die Daten erlauben, können die Daten geeignet gefiltert, aggregiert und transformiert werden. Hierfür stehen beim ODMr verschiedene Knoten im Workfloweditor zur Auswahl. Grundlegend werden bei ODM zwei Arten von Daten unterschieden, die numerischen und kategorischen [5]. Anhand dieser Differenzierung werden die Daten in verschiedenen Data-Mining-Algorithmen anders behandelt. Eine Zuordnung findet entweder automatisch über den Datentyp oder manuell statt.

Für das Lernen von Data-Mining-Modellen unterstützt ODM unter anderem die Klassifikation, die Regression, das Clustering, die Assoziationsanalyse und die Anomalieerkennung. Zu den bei der Klassifikation angebotenen Algorithmen zählen der Entscheidungsbaum, Naive Bayes, Support Vector Machine (SVM) und das Generalisierte Lineare Modell (GLM), und zwar hierbei die logistische Regression [4]. Bei der Regression kann zwischen der linearen Regression und SVM gewählt werden. Cluster werden entweder anhand des k-Means oder des O-Cluster-Algorithmus gefunden und Ausreißer anhand One-Class SVM. Für die Erstellung von Assoziationsregeln wird der Apriori-Algorithmus eingesetzt. Bei ODM wird für das Anwenden der Algorithmen vorausgesetzt, dass alle Daten in genau einer Tabelle oder Sicht vorliegen, der sogenannten Falltabelle. Einstellungen zu den Algorithmen können entweder standardmäßig belassen oder individuell zugeschnitten auf den jeweiligen Anwendungsfall vorgenommen werden. Für eine Bewertung der gelernten Modelle stehen in Abhängigkeit vom Verfahren diverse Evaluierungsmaße und Visualisierungen bereit. Beispielsweise kommt bei der Klassifikation die Gesamtklassifikationsgenauigkeit zum Einsatz. Diese ist der Anteil der korrekt klassifizierten Objekte in der gesamten Menge.

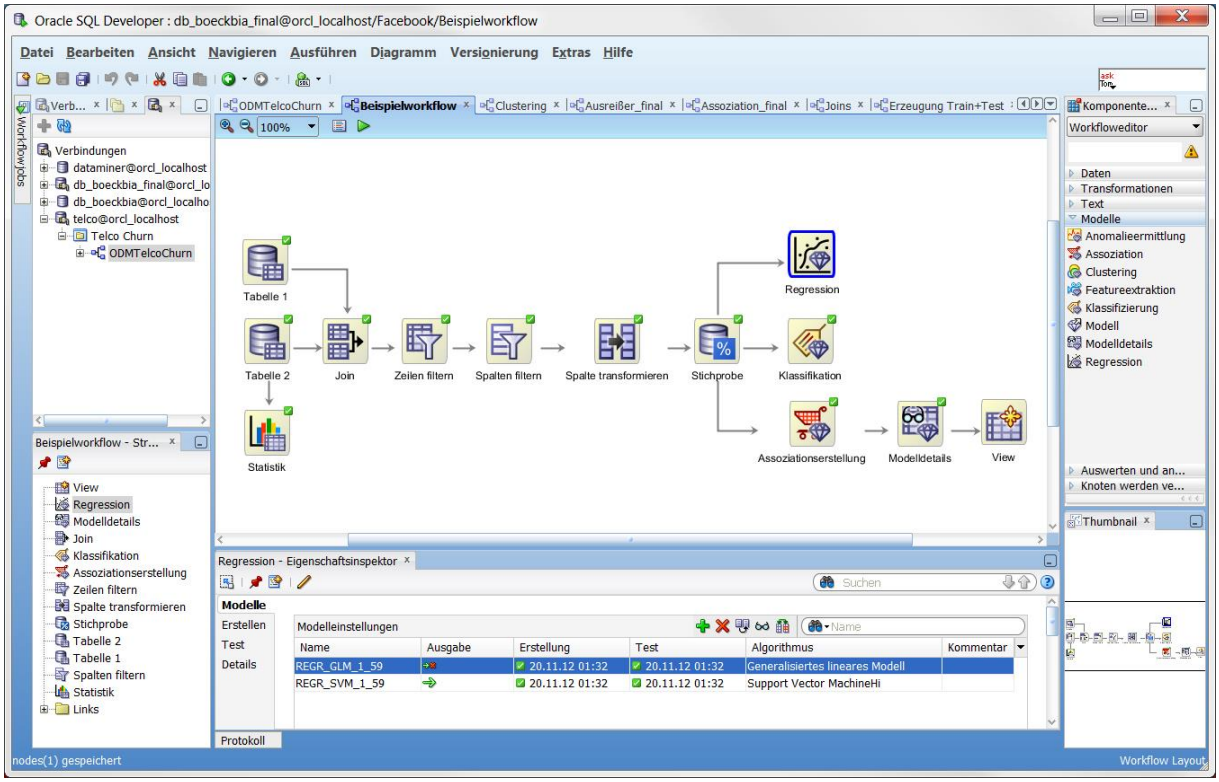

Abbildung 2: Benutzeroberfläche des Oracle Data Miners mit der Anzeige eines Beispielworkflows

Indirekte Bestimmung privater Informationen

In der Masterarbeit wurde untersucht, ob es in sozialen Online-Netzwerken mittels Data-Mining möglich ist, nur anhand der öffentlichen Informationen der Freunde auf die privaten Daten einer Person selbst zu schließen. Müssen also neben den eigenen Profilinformationen auch die Freundschaftsbeziehungen vor der breiten Öffentlichkeit privat gehalten werden, um personenbezogene Daten vor dem Ausspähen durch Fremde zu schützen?

Selektion

Zu Beginn wurde hierfür die Facebook-Anwendung „Hidden Profile“ in PHP entwickelt, mit dessen Hilfe knapp 700 freiwillige Nutzer gewonnen werden konnten. Neben Profilinformationen wie Geburtstag, Geschlecht, Wohn- und Heimatort, „Gefällt mir“-Angaben (unter anderem TV, Film, Musik), Informationen zur Arbeit und Bildung wurden auch Freundschaftsbeziehungen erfasst. Hierbei wurden nur diejenigen Freunde in der Datenbank gespeichert, die auch explizit der Anwendung zugestimmt haben.

Vorverarbeitung

Im Schritt der Vorverarbeitung wurden zum einen fehlende Werte behandelt und zum anderen die Daten bereinigt. Während einige fehlende Daten wie der Wohn- und Heimatort anhand anderer im eigenen Profil zu findenden Informationen abgeleitet werden konnten, war dieses für andere Attribute nicht möglich. Bei letzterem wurde der Median beziehungsweise der Modus unter allen Teilnehmern eingesetzt, um fehlende Werte zu entfernen. Selten angegebene Attribute wurden jedoch von der weiteren Betrachtung ausgeschlossen, da zu viele künstliche Werte die Analyse verfälschen könnten. Des Weiteren war eine Bereinigung der Daten notwendig gewesen, da in Facebook aufgrund von bei-

spielsweise Rechtschreibfehlern und Mehrsprachigkeit viele unterschiedliche Bezeichnungen für das semantisch gleiche Objekt existieren. Für die Erstellung eines hierfür notwendigen einheitlichen Vokabulars wurden DBpedia[1], Freebase[2] und andere frei verfügbare Datenquellen herangezogen. Die Profilinformationen aus Facebook wurden automatisch jeweils dem Wort aus den zuvor genannten Datenquellen zugeordnet, welches die maximale Jaro-Winkler-Ähnlichkeit besitzt und einen gewissen Schwellwert überschreitet. Im Anschluss daran wurde eine manuelle Nachkorrektur aufgrund von Synonymen und Homonymen vorgenommen.

Transformation

In der Transformationsphase wurde neben der Normalisierung von metrischen Attributen und Datentypkonvertierungen auch eine Kategorisierung vorgenommen, um durch die Reduzierung der Kardinalität von Attributen möglicherweise kompaktere Modelle zu erzielen. Hierfür wurden vorwiegend die zuvor erwähnten Datenquellen genutzt. Beispielsweise wurde eine Ortshierarchie aufgebaut sowie TV-Sendungen ihren Genres beziehungsweise das Alter mittels Quantil-Binning Altersgruppen zugeordnet. Des Weiteren wurden mit dem ODMr Ausreißer anhand One-Class SVM identifiziert und entsprechend behandelt. Als Ausreißer gelten Nutzer mit seltenen, von der Mehrheit abweichenden Informationen. Beispielsweise wurden unrealistische Altersangaben im Bereich von 100 Jahren geglättet.

Ein weiterer wichtiger Punkt ist die probate Abbildung der Freundschaftsbeziehungen für die Prognose. Nur ein Teil der knapp 700 teilgenommenen Facebook-Nutzer konnte für die Vorhersage genutzt werden, da nur bei wenigen auch genügend Freunde der Anwendung zugestimmt haben und somit deren Daten vorlagen. Während beispielsweise 241 Nutzer mindestens 3 teilgenommene Freunde besaßen, waren es bei mindestens 5 Freunden nur noch 150 Facebook-Mitglieder. Außerdem wurden solche Personen gänzlich von der Analyse ausgeschlossen, die mehr als 700 Leute in ihrer Facebook-Freundesliste pflegten. Dieses liegt in der Vermutung begründet, dass neben engen Freundschaften auch viele Bekanntschaften enthalten sind, die eine Prognose wahrscheinlich verschlechtern würden. Des Weiteren war es vorstellbar, neben den Personen aus der eigenen Freundesliste zusätzlich die indirekten Freunde einzubeziehen, also die Freunde der Freunde, um möglicherweise bessere Modelle zu erhalten. Da die Anzahl der Freunde über verschiedene Personen hinweg variiert und ODM genau eine Falltabelle voraussetzt, ist eine Aggregation der Freundeswerte notwendig gewesen. Es wurde deshalb für jeden Nutzer ein Freundesvektor für jedes Attribut erstellt, der dem gewichteten relativen Vorkommen eines jeden Attributwerts unter seinen Freunden entspricht. ODM stellt für die Speicherung von solchen verschachtelten Daten die Datentypen `DM_NESTED_NUMERICALS` und `DM_NESTED_CATEGORICALS` bereit [5]. Da die Anzahl der indirekten Freunde viel größer ist als die der direkten, wurde der relative Anteil für die zwei Freundespartitionen separat berechnet. Um den direkten Freunden eine größere Bedeutung zukommen zu lassen als den indirekten, wurde ihnen bei der anschließenden Summierung der Anteile ein größeres Gewicht zugewiesen.

Irrelevante und korrelierte Attribute verfälschen ein Modell. Aus diesem Grund wurden zum einen mit dem ODMr irrelevante Attribute für die Vorhersage bestimmter Merkmale automatisch mittels Minimum Description Length (MDL) gefiltert. Zum anderen wurden verschiedene Modelle ausschließlich auf den Facebook-Originaldaten, nur auf den bereinigten oder nur auf den kategorisierten Daten gelernt, um den zusätzlichen zeitlichen Aufwand für die Bereinigung und die Kategorisierung mit einer vermuteten besseren Modellqualität rechtfertigen zu können.

[1] http://dbpedia.org/About
[2] http://www.freebase.com/

Data-Mining

Das Ziel der Analyse war eine Prognose der Attributwerte eines Facebook-Nutzers ausschließlich anhand der Angaben seiner Freunde. Der vorherzusagende Nutzer stellte somit eine Blackbox dar. Die Klassifikation und die Regression waren für eine solche Aufgabenstellung geeignet. Für eine exemplarische Prognose von Geschlecht und Wohnort wurden die Klassifikationsalgorithmen SVM und Naive Bayes eingesetzt. Aufgrund einer binären Zielvariablen kam beim Geschlecht zusätzlich die logistische Regression zum Einsatz. Da der Entscheidungsbaum beim ODM in der aktuellen Version 11.2 nur nicht verschachtelte Daten verarbeiten kann, wurde hierfür beim Freundesvektor ein repräsentativer Wert ausgewählt. Für die Vorhersage des metrischen Alters wurden bei der Regressionsanalyse die lineare Regression und SVM angewendet. Eine Evaluierung der verschiedenen Modelle war durch eine Aufteilung der Objektmenge in 2/3 Trainings- und 1/3 Testmenge gegeben.

Des Weiteren existiert neben den klassischen Prognoseverfahren auch die Möglichkeit anhand von Assoziationsregeln Gemeinsamkeiten zwischen allen Nutzern, nicht nur zwischen seinen Freunden, für die Prognose genau eines Merkmals einzusetzen. Dieses setzt voraus, dass der Nutzer selbst für ein Attribut mindestens einen Wert öffentlich hält, damit anhand von Regeln weitere private Angaben vorhergesagt werden können. Die Assoziationsanalyse wurde exemplarisch für TV-Sendungen durchgeführt. Gehört beispielsweise „The Big Bang Theory“ zu den Lieblingsfernsehsendungen eines Nutzers, so kann man anhand der gefundenen Regel „The Big Bang Theory“ → „How I Met Your Mother“ die nicht in seinem Profil öffentlich einsehbare Serie „How I Met Your Mother“ zu seinen favorisierten Sendungen zählen.

Evaluierung

Die Ergebnisse bekräftigen die Vermutung, dass die Daten der Freunde für das Ableiten privat gehaltener Profilinformationen eines Nutzers gut geeignet sind. Während der Wohnort bei 7 von 10 Personen richtig vorhergesagt werden konnte, war dieses sogar mit 86 % Gesamtklassifikationsgenauigkeit für das Bundesland des Wohnorts möglich. Das Geschlecht war in 65 % der Prognosen richtig. Während bei Wohnort und Geschlecht alle Werte der Freunde verwendet wurden, war dieses beim Ableiten des Alters nicht nötig. Nur mit dem Wissen, das Alter der Freunde zu kennen, weichte das vorhergesagte Alter um durchschnittlich ±2, 05 Jahre vom tatsächlichen ab. MDL hatte sich in der Masterarbeit nicht unbedingt als hilfreich herausgestellt, um relevante Attribute aufzudecken. Der Entscheidungsbaum hatte im Allgemeinen schlechte Ergebnisse geliefert. Dieser ist im Gegensatz zu anderen Modellen verständlicher und leichter interpretierbar. Es konnte festgestellt werden, dass für die Vorhersage des Wohnorts Naive Bayes am besten geeignet ist. Beim Geschlecht war es hingegen SVM.

Eine Bereinigung und Kategorisierung der Attribute führte nur beim Bestimmen des Wohnorts zu besseren Ergebnissen. Dieser Aufwand ist demzufolge bei der Vorhersage des Geschlechts und des Alters nicht nötig. Das Einbeziehen der indirekten Freunde führte nicht zu einer Verbesserung der Modelle. Diese können demzufolge bei zukünftigen Analysen ignoriert werden. Des Weiteren konnte festgestellt werden, dass die Ergebnisse bei mindestens 5 direkten Freunden besser sind als bei mindestens 3. Es konnte jedoch keine allgemeine Aussage hieraus abgeleitet werden, da die Testmengen unterschiedlich groß waren.

Es wurden außerdem Assoziationsregeln auf TV-Sendungen gelernt. Hierbei konnten 21 Regeln gefunden werden, die einen Support von mindestens 10 %, eine Konfidenz von mindestens 40 % und einen Lift größer 1 besitzen. Es konnten keine Regeln extrahiert werden, die sowohl hohe Support-, Konfidenz- als auch Liftwerte besitzen.

Aus diesen Erkenntnissen lässt sich schlussfolgern, dass neben privaten Profilinformationen auch die Freundschaftsbeziehungen vor der Öffentlichkeit verborgen werden sollten, um die eigene Pri-

vatsphäre zu schützen. Hierbei müssten jedoch beide Freunde ihre Privatsphäreeinstellungen in Facebook ändern, um Fremden nicht die Möglichkeit zu geben, die Beziehung indirekt über die andere Person herausfinden zu können.

Literaturverzeichnis

[1] M. Ester und J. Sander, Knowledge Discovery in Databases - Techniken und Anwendungen, Springer, 2000.

[2] U. M. Fayyad, G. Piatetsky-Shapiro und P. Smyth, „From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA, USA, American Association for Artificial Intelligence, 1996, S. 1-34.

[3] U. M. Fayyad, G. Piatetsky-Shapiro und P. Smyth, „From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, Bd. 17, Nr. 3, S. 37-54, 1996.

[4] K. L. Taylor, „Oracle Data Mining Concepts, 11g Release 2 (11.2), E16808-06," Juli 2011. [Online]. Available: http://docs.oracle.com/cd/E11882_01/datamine.112/e16808.pdf. [Zugriff am 25.03.2013].

[5] K. L. Taylor, „Oracle Data Mining Application Developer's Guide, 11g Release 2 (11.2), E12218-07," Juli 2011. [Online]. Available: http://docs.oracle.com/cd/E11882_01/datamine.112/e12218.pdf. [Zugriff am 21.03.2013].

Kontaktadresse

Bianca Böckelmann
Robotron Datenbank-Software GmbH
Stuttgarter Straße 29
01189 Dresden
Deutschland

Telefon:	+49 351 25859-2463
Fax:	+49 351 25859-3699
E-Mail	bianca.boeckelmann@robotron.de
Internet:	www.robotron.de