Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

6.11.2008

2. Übung „Knowledge Discovery“

Wintersemester 2008/2009

Arbeiten mit dem KDD-Programm RapidMiner

Die erste Aufgabe ist mit dem Programm RapidMiner zu lösen. Auf der Webseite `http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd/uebung/rapidminer.html` finden Sie Hinweise zur Installation von RapidMiner sowie die benötigten Datensätze. Folgen Sie den Anleitungen auf der Seite, um RapidMiner innerhalb von Eclipse zu installieren.

1. Laden Sie den Datensatz iris.xrff mit Hilfe des Operators XrffExampleSource in RapidMiner ein. Starten Sie das Experiment und wechseln Sie bei der Ausgabe (im Reiter DataTable) auf die Ansicht Plot View. Mittels der Felder x-Axis sowie y-Axis können Sie Attribute des Datensatzes gegeneinander abbilden.

   Wählen Sie verschiedene Attribute und visualisieren Sie diese mit verschiedenen Plottern, die Sie im Feld Plotter auswählen können. (Achtung: bei Plots/Color/... sollte class eingestellt werden.)

   Welche zwei Attribute reichen aus, um die Daten der jeweils gleichen Klassen des Datensatzes bestmöglich voneinander zu trennen? Welchen Plotter haben Sie benutzt, und wie sah das resultierende Bild aus, welches Sie zu Ihrer Entscheidung bewogen hat?

2. Für die folgende Aufgabe benötigen Sie zusätzlich zum Datensatz aus der vorangegangenen Aufgabe den Testdatensatz irisTest.xrff.

   Erweitern Sie den ersten Versuch, indem Sie nach Einlesen des ersten Datensatzes ein Lernverfahren anwenden (z. B. den /Learner/Trees/DecisionTree). Anschließend müssen Sie den Testdatensatz mit einem XrffExampleSource-Operator einlesen. Daraufhin benutzen Sie den ModelApplier-Operator, um das gelernte Verfahren anzuwenden.

   Starten Sie das Experiment, und wechseln Sie bei der Ausgabe (im Reiter DataTable) auf die Ansicht data view.

3. Laden Sie sich den Datensatz mushrooms.xrff herunter. Erstellen Sie ein Experiment, das lediglich aus einem XrffExampleSource besteht, und lesen Sie somit die Daten ein. Welche Ausprägung des Attributes ring-type tritt im Datensatz am häufigsten auf? Wie oft tritt diese Ausprägung im Datensatz auf? Wie heißen die beiden Klassen des Datensatzes?

4. Fügen Sie nun den Regellerner RuleLearner als weiteren Operator in das Experiment ein, und lassen Sie eine Entscheidungsregel mit Standardeinstellungen (d. h. Sie müssen keine Einstellungen ändern) lernen. Welche Regel wurde gelernt?

5. Zur Evaluierung der Güte des Lernverfahrens auf dem Datensatz erstellen Sie nun bitte ein Experiment mit einer SimpleValidation. Lernen Sie auf einer Trainingsmenge von 70% (split_ratio = 0.7) und evaluieren Sie die Maße „Accuracy“ und „Precision“ auf den verbleibenden 30% der Daten. Hierfür ist es notwendig, das Model mit einem ModelApplier anzuwenden und mit einem PerformanceEvaluator (hier müssen Sie accuracy und precision auswählen) zu evaluieren. Beachten Sie, dass Sie die benötigten Operatoren in den Operator SimpleValidation „einhängen“ müssen. SimpleValidation erlaubt nur zwei Unteroperatoren. Einen für die erste Datenmenge, den anderen für die restliche Datenmenge. Benutzen Sie den Operator OperatorChain, um mehr als einen Operator für die jeweilige Datenmenge zu benutzen.