Proseminar Maschinelles Lernen und Experimentelles Design, UdS, SS11 Hausaufgaben für 9.6.2011

Caroline Sporleder

May 23, 2011

Aufgabe 1: Erste Schritte in Weka Startet Weka, ladet das **weather.nominal.arff** Datenset, und generiert einen **Decision Tree**, J48, (mit **66% Trainings- und 44% Testdaten**). Was für Ergebnisse bekommt ihr? Schaut euch den Entscheidungsbaum an, sieht er so aus, wie ihr es erwarten würdet?

Aufgabe 2: Naive Bayes vs. Decision Trees Wählt das Datenset **labor.arff**. Dieses Datenset beschreibt Arbeitsbedingungen, die in Verträgen festgelegt wurden. Ziel ist es, anhand der Bedingungen vorherzusagen, ob ein Arbeitsvertrag 'gut' oder 'schlecht' ist. Die Bedingungen werden durch die Attribute beschrieben. Die Attribute haben dabei z.T. numerische und z.T. nominale Wertebereiche. Klassifiziert dieses Datenset zuerst mit einem **Decision Tree Algorithmus (J48)** und dann mit **Naive Bayes**. Benutzt wieder einen **66%-44% Trainings-Test-Split**. Welcher Algorithmus gibt bessere Ergebnisse? Habt ihr eine Idee warum?

Aufgabe 3: Decision Trees: Pruning vs. kein Pruning Per Default generiert J48 'geprunte' Entscheidungsbäume. Wählt nochmal das Datenset **labor.arff**, aber generiert diesmal einen nicht-geprunten Baum. (Ihr könnt dies tun, indem ihr die Eigenschaften des Algorithmus verändert (siehe Weka-Folien) und "unpruned=True" setzt.). **Vergleicht den geprunten und den ungeprunten Baum**. Seht ihr einen Unterschied? Ist die Accuracy auf dem Testset für beide Bäume gleich? Wendet den `AddNoise`-Filter unter `Preprocess`→`Filter` an, **verrauscht die Ausgabeklasse** jeweils in 10%-Schritten (siehe die Weka-Folien), und generiert wieder einen geprunten und einen ungeprunten Baum. Bei welchem Prozentsatz verrauschter Daten, seht ihr einen Unterschied in der Accuracy der beiden Bäume?