Alexander Hinneburg,
hinneburg@informatik.uni-halle.de

27.10.2005

Übung zur Vorlesung „Data Mining in Datenbanken“

Übungsblatt 3

Abgabe am 3.11.2005

Achten Sie auch auf die Form Ihren Lösungen, z.B. das jedes Bild eine Bildunterschrift hat. Bitte geben Sie keine losen Blätter ab. Um die Übungen in angemessener Zeit zu diskutieren, bereiten Sie neben Ihrer Lösung ein kleine Präsentation vor, mit der Sie dann Ihre Ergebnisse zeigen können.

Aufgabe 1 Konvertieren Sie die Daten genes-leukemia.csv (die Daten und eine Beschreibung finden Sie auf der Web-Seite) in eine Weka-Datei genes-a.arff. (a) Sie finden dazu ein Weka Kommando (nicht im Explorer sondern im Simple CLI) mit dem Sie die Datei automatisch von .csv nach .arff konvertieren können. (Falls Sie den Befehl nicht finden können Sie die Datei auch mittels Texteditor konvertieren, um die weiteren Aufgaben zu lösen, bekommen dann aber keine Punkte für 1.) (b) Beschreiben Sie den Befehl zum Konvertieren und diskutieren Sie Möglichkeiten viele Dateien automatisch per Skript zu konvertieren.

Aufgabe 2 Nutzen Sie den Algorithmus J48, um einen Entscheidungsbaum für genes-leukemia (Klassenattribute ist CLASS) zu erstellen. Nutzen Sie die Option „Use training set“. Beschreiben Sie kurz das Ergebnis und zeichnen Sie den erhaltenen Baum.

Aufgabe 3 Teilen Sie genes-leukemia.arff in zwei Teilmengen:

- genes-leukemia-train.arff, die ersten 38 Beispiele (s1 ... s38)
- genes-leukemia-test.arff, restlichen 34 Beispiele (s39 ... s72).

(a) Trainieren Sie J48 auf genes-leukemia-train.arff und spezifizieren Sie „Use training set“ als Test-Option. Welchen Entscheidungsbaum erhalten jetzt (bitte zeichnen)? Wie ist die Genauigkeit?

(b) Spezifizieren Sie dann genes-leukemia-test.arff als Testmenge. Zeichnen und vergleichen Sie den jetzt erhaltenen Entscheidungsbaum und dessen Genauigkeit mit den Ergebnissen aus Aufgabe 2. Diskutieren Sie kurz das Ergebnis.

Aufgabe 4 Löschen Sie das Attribut „Source“ so dass es nicht mehr vom Klassifikator genutzt werden kann und wiederholen Sie die Schritte aus Aufgabe 3b. Was beobachten Sie? Verbessert sich die Genauigkeit und falls das der Fall ist, was wäre eine Erklärung?

Zusatzaufgabe Welcher von den in Weka verfügbaren Klassifikatoren liefert die höchste Genauigkeit auf der Testmenge aus Aufgabe 3?