Ludwig-Maximilians-Universität München
Institut für Informatik
Priv. Doz. Dr. Matthias Schubert, Dr. Arthur Zimek
Erich Schubert

München, den 24.6.2009

Knowledge Discovery in Databases II

SoSe 2009

Übungsblatt 7: Multirepräsentiertes Data Mining

Besprechung am Donnerstag, 25.6.2009

Aufgabe 7-1 *Verteilte Multiclass-Klassifikation*

Ihr Datenbestand (Trainings- und Testdaten) ist über n Rechner verteilt und Sie möchten verteilte Multiclass-Klassifikation auf diesen Rechnern betreiben. Welche Vorteile bzw. Nachteile treten auf, wenn Sie dafür

- Entscheidungsbäume,
- Nächste-Nachbarn-Klassifikation,
- Support-Vector-Maschinen oder
- Naive Bayes

verwenden?

Aufgabe 7-2 *Komplementarität von Klassifikatoren*

Gegeben seien zwei binäre Klassifikatoren f_1 und f_2, die auf je einer Repräsentation der Objekte eines Datensatzes D mit Klassen $\{0, 1\}$ arbeiten. Entscheiden Sie, ob in den folgenden Fällen eine Kombination der Klassifikatoren sinnvoll ist:

(a) $f_1(x) = f_2(x)$ für alle $x \in D$

(b) $f_1(x) = 1 - f_2(x)$ für alle $x \in D$

Aufgabe 7-3 *Abhängigkeitsmaß*

Gegeben sei ein Maß h, welches die Abhängigkeit zwischen zwei Kernelmatrizen K und K' misst. Anschaulich heißt das, dass $h(K, K')$ groß ist, wenn die zugehörigen Kernels k und k' dieselben Objekte als ähnlich und als unähnlich betrachten. Wenn sie die Ähnlichkeit derselben Objekte unterschiedlich bewerten, sei $h(K, K')$ niedrig.

Seien nun ein Datensatz D mit einem Klassenlabel und r Repräsentationen pro Objekt gegeben. Wir berechnen eine Kernelmatrix K_i für jede der r Repräsentationen und eine Kernelmatrix L auf den Klassenlabels. Überlegen Sie sich, wie man mittels h eine Linearkombination der K_i bestimmen kann, die die Ähnlichkeit der Klassenlabels möglicht gut widerspiegelt.