Ludwig-Maximilians-Universität München
Institut für Informatik
Priv. Doz. Dr. Matthias Schubert

München, den 07.06.2010

Knowledge Discovery in Databases II

SoSe 2010

Übungsblatt 5: Paralleles und Verteiltes Data Mining

Besprechung am Donnerstag, 10.6.2010

Aufgabe 5-1 *Parallele Klassifikation*

Überlegen Sie sich einen Algorithmus, der das Training eines Naive-Bayes Klassifikators auf verteilten Datenbeständen ermöglicht!

Welche Eigenschaften des verwendeten Klassenmodells sind dabei nützlich ?

Würde eine parallele Variante Ihres Algorithmus die Laufzeit des Trainings reduzieren (die Daten müssen erst noch aufgeteilt werden)?

Aufgabe 5-2 *Privacy Preservation in Standard-Klassifikatoren*

Gegeben seien folgende vier Klassifikatoren: Entscheidungsbäume, Nächste-Nachbarn-Klassifikation, Support-Vector-Machines und Naive-Bayes.

- Untersuchen Sie die Frage, ob bereits trainierte Klassifikatoren an Dritte weitergegeben werden dürfen, ohne dass dabei Teile der Trainingsmenge offengelegt werden!
- Wie könnte man versuchen, etwaige Probleme bei den einzelnen Klassifikatoren zu lösen?

Aufgabe 5-3 *Parallele Assoziationsregeln*

Überlegen Sie sich die Vorteile und Nachteile einer horizontalen und einer vertikalen Verteilung bei der nebenläufigen Berechnung von Assoziationsregeln!