

## Praxisübung zur 3. Übung „Knowledge Discovery“

Wintersemester 2008/2009

### Clustering mit $k$ -Means

In der ersten Praxisübung haben Sie RapidMiner innerhalb der Eclipse-Entwicklungsumgebung installiert und das prinzipielle Vorgehen beim Entwickeln eines Plugins gelernt. Ziel dieser Übung ist das eigenständige Implementieren des Clustering-Algorithmus  $k$ -Means innerhalb der RapidMiner-Plattform.

Implementieren Sie folgende Variante des  $k$ -Means-Algorithmus:

- Die *initiale Clusterung* soll durch zufälliges Zuweisen einer jeden Instanz zu einem Cluster gebildet werden.
- Als *Distanzfunktion* nutzen Sie bitte die Manhattan-Distanz  $d(x, y) = |x - y|$ .
- In jedem Iterationsschritt soll der Algorithmus über alle Instanzen gehen und jede Instanz dem nächsten Cluster (Centroid) zuweisen. Jede Änderung soll eine *sofortige Neuberechnung der Centroide* zur Folge haben.
- Als *Abbruchbedingung* soll neben der maximalen Anzahl an Iterationen die Änderung des Fehlers zwischen zwei Iterationen beachtet werden.

Nutzen Sie die Datei KDDKMeans.java als Grundgerüst für Ihre Implementierung. Behalten Sie Klassen- und Paketnamen bei und implementieren Sie Ihren Algorithmus in der Methode `public ClusterModel createClusterModel(final ExampleSet exampleSet)`. Selbstverständlich können Sie ggf. weitere Methoden in der Klasse anlegen.

Beachten Sie bitte zusätzlich folgende Formalitäten:

- Abgabe der Praxisübung: bis **2.12.2008, 23:59 Uhr MEZ** per E-Mail an [jaeschke@cs.uni-kassel.de](mailto:jaeschke@cs.uni-kassel.de). Spätere Einreichungen werden nicht berücksichtigt!
- Alle benötigten Methoden, Attribute, etc. müssen sich in der Java-Datei `KDDKMeans.java` befinden, deren Grundgerüst Sie hier finden: <http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd/daten/KDDKMeans.java>.
- Am 4.12.2008 werden Sie uns Ihren Algorithmus kurz vorführen, erklären und ggf. Fragen beantworten. Termine (pro Teilnehmer ca. 5 Minuten) werden wir vorher vereinbaren.

- Für eine positive Bewertung muss ihr Code problemlos compilieren, einen Durchlauf durch den Iris-Datensatz aus der vorherigen Praxisübung ohne Fehler bestehen, eine nachvollziehbare Clusterung liefern und es muss im Gespräch erkennbar sein, dass Sie *Ihre Implementierung* des Algorithmus verstanden haben.

Sollten Sie Fragen haben, wenden Sie sich bitte an Robert Jäschke (jaeschke@cs.uni-kassel.de). Zusätzlich bieten wir Ihnen an, dass sie immer donnerstags vor und nach der Übung kurz Fragen stellen können. Weitere Hinweise und Dateien finden Sie auch auf der Webseite <http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd/uebung/rapidminer.html>.