Prof. Dr. Katharina Morik,
JProf. Dr. Uwe Ligges
Dipl.-Inform. Hendrik Blom,
M. Sc. Nadja Bauer,
Daniel Horn

Dortmund, 05.07.13
Abgabe: bis Fr, 12.07., 23:59 Uhr an
hendrik.blom@tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2013

Blatt 12

Wiederholung

- Was ist genau der Unterschied zwischen *überwachtem* und *unüberwachtem Lernen*? Wie unterscheidet sich Clustering von Klassifikation?
- Wofür stehen die Begriffe *Innerer Abstand Within* und *Zwischenähnlicher Abstand Between*? Wie sind diese Maße definiert?

Aufgabe 12.1 (5 Punkte)

Der in der Vorlesung vorgestellte Cluster-Algorithmus k-Means partitioniert eine Menge unklassifizierter Beispiele so, dass sich Objekte innerhalb von Clustern ähnlicher sein sollen als solche aus unterschiedlichen Clustern.

(a) Warum kann man allein anhand dieses Kriteriums den Parameter k nicht mit Hilfe einer herkömmlichen Parameter-Optimierung bestimmen?

Überlegen Sie sich, wie sich die Maße $W(C)$ (*Within*) bzw. $B(C)$ (*Between*) für die Extremfälle $k = 1$ bzw. $k \to N$ verhalten!

(b) Auf Beispielmengen, die Gruppen bilden lässt sich der k-Means Algorithmus gut anwenden. Was passiert, wenn die Beispiele im Extremfall gleichverteilt sind?

Betrachten Sie dazu den Datensatz unified.csv (von der Übungswebseite) in RapidMiner und wenden sie den Operator *K-Means* an. Probieren Sie verschiedene Werte für k aus und betrachten Sie die Cluster im Scatter-Plot der Datenansicht.

Interpretieren Sie das Ergebnis Ihrer Experimente!

Aufgabe 12.2 (5 Punkte)
In dieser Aufgabe soll eine Brücke zwischen Mehrklassenproblemen und Clustering geschlagen werden, indem wir die das Clustering von k-Means mit mit den *wahren* Labeln auf dem Datensatz vergleichen.

Gleichzeitig soll auch die Attribut-Gewichtung ausprobiert werden, die einen großen Einfluss auf das Distanzmaß hat.

(a) Betrachten Sie den *Iris*-Datensatz aus dem RapidMiner *samples*-Repository und wenden Sie den k-Means Algorithmus darauf an (Operator *K-Means*). Vergleichen Sie das *Label* Attribute mit dem *Cluster* Attribute, z.B. für zwei Attribute mit Hilfe des Scatter-Plots.

(b) Verwenden Sie den Operator *Map Clustering on Labels*, der das *Cluster*-Attribut zu einem *Prediction*-Attribut macht. Auf dem so umgeformten Datensatz können Sie jetzt mit Hilfe des *Performance (Classification)* Operators den "Clusterfehler" bestimmen.

(c) Als nächstes sollen Sie untersuchen, wie eine Gewichtung der Attribute das Clustering verbessert. Erweitern Sie ihr Exeriment um einen *Weight by SVM* und einen *Scale by Weights* Operator, mit denen die Attribute des Datensatzes zunächst gewichtet werden und die Daten dann mit diesen Gewichten skaliert werden.

Welche Effekt hat die Gewichtung auf den "Clusterfehler"?