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

8.12.2005

Übung zur Vorlesung „Data Mining in Datenbanken“

Übungsblatt 8[1]

Clusteranalyse
Abgabe am 15.2005

7.3 (zum 15.12.05) Implementieren Sie den k-Means Algorithmus mittels wiederholter SQL Anfragen. Hier sollen Sie auf die Ergebnisse aus 7.2 zur Initialisierung und zum Berechnen der neuen Repräsentanten zurückgreifen und diese mit `Insert` Statements verknüpfen.

Um den Algorithmus zu vervollständigen müssen Sie eine Anfrage zu Berechnung des Quantisierungsfehlers D entwickeln.

Führen Sie nach jeder Iteration einen Test durch ob, weniger als 3 Objekte einem Repräsentanten zugewiesen wurden und verschieben Sie diese in die Nähe des Repräsentanten mit dem großten Quantisierungsfehler. Dies können Sie mit einem `update` auf der Repräsentantentabelle tun. Falls der Test auf keinen Repräsentanten zutrifft, werden 0 Zeilen aktualisiert, es gibt aber keinen Fehler. Den Repräsentanten $\vec{x}$ in die Nähe des Repräsentanten $\vec{y}$ verschieben heißt, $\vec{x} = \vec{y} + \vec{\epsilon}$, wobei $\vec{\epsilon} = ((max_1 - min_1)/20, \ldots, (max_d - min_d)/20)$.

Beschreiben Sie die entwickelten SQL-Anfragen und geben Sie alle Anfragen für die Initialisierung und eine Iteration an.

8.2 Erstellen Sie eine Menge von SQL Anweisungen, die hintereinander ausgeführt werden, so dass die Repräsentanten der 20. Iteration berechnet werden. Starten Sie mit der Initialisierung mittels zufälliger Zuordnung aus 7.2.

Nutzen Sie die Iris-Daten für die Experimente. Berechnen Sie Clusterpartitionen für verschiedene Anzahlen von Clustern. Zeichnen Sie eine Kurve mit dem Quantisierungsfehler in Abhängigkeit von der Iteration für verschiedene Anzahlen von Repräsentanten k=2, 3, 4, 5. Beschriften Sie die Grafik! Diese Kurven sollen alle in dasselbe Diagramm gezeichnet werden.

Interpretieren Sie mit wenigen Sätzen die Kurven.

8.3 Berechnen Sie für jeden Cluster die Entropie für die Klassenverteilung, wichten diese mit der Clustergröße und mitteln die gewichteten Entropien über alle Cluster der jeweiligen Clusterpartition. Tun Sie dies für alle $k = 2, 3, 4, 5$ und zeichnen Sie eine Kurve der gewichteten mittleren Entropie in Abhängigkeit der Anzahl der Cluster k. Interpretieren Sie mit wenigen Sätzen die Ergebnisse.

[1] Achten Sie auch auf die Form Ihrer Lösungen, z.B. daß 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.