Self Organizing System
Übung 2 WS2010

von Benjamin Bachhuber (1028430), Patrick Marschik (0625039)

Thema: Dendogram Visualisierung von hierarchischen Clusteringverfahren

How To Start:
Aufruf wie den normalen SOM Viewer.
Dann ein hierarchisches Clustering-Verfahren auswählen.

Task - Dendrogramm Visualisierung

Das Ergebnis eines hierarchischen Clusteringverfahrens (single, complete, ward's linkage) ist ein Baum, der die Anzeige von Clustern auf unterschiedlichen Levels erlaubt.
Implementieren Sie eine Visualisierung des Dendograms, wobei Sie gerne auf existierende, open-source Lösungen zurückgreifen können ((LGPL, Apache License, etc. bevorzugt, GPL eher nicht).
Der Benutzer sollte über das Dendogram auch die Anzahl der auf der Karte abgebildeten Cluster auswählen können (als Alternative zum momentan vorhanden Spinner).

Screenshot

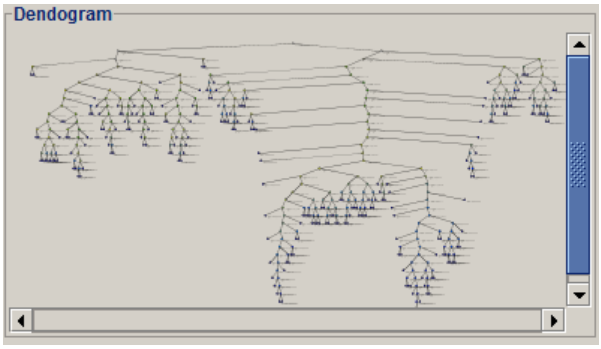

Abb. 1: Dendrogram Control

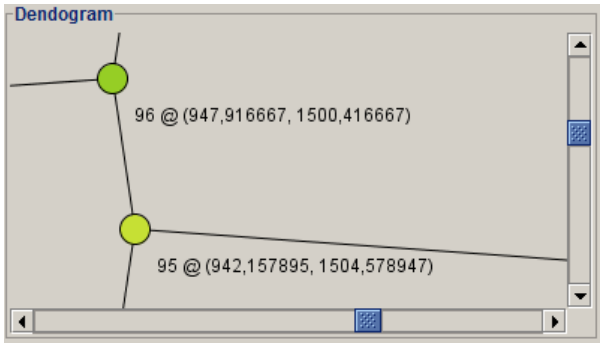

Abb. 2: Dendrogram Control (Zoom)

Implementierung

Die Basis für die Visualisierung eines Dendrogramms stellt die Verwendung einer Java Bibliothek zum Zeichnen von Graphen dar. Die Entscheidung fiel schlussendlich auf die Verwendung der JUNG Bibliothek (nachdem zuerst JGraph evaluiert wurde). Da diese Bibliothek unter der BSD Lizenz verfügbar ist, eignet sie sich für die Verwendung in der SOMToolbox.
Die Implementierung der einzelnen hierarchischen Clusteringverfahren erfordert bereits die Abspeicherung eines Clusterbaums. Dieser Clusterbaum wird dann mithilfe eines mit JUNG erstellten Dendrogramms visualisiert.
Zu erwähnen ist außerdem, dass das Dendrogramm Control eine zusätzliche Zoom Funktionalität besitzt, die es mithilfe des Mausrads ermöglicht das Dendrogramm zu zoomen.
Eine weitere Interaktionmöglichkeit ist das Auswählen eines bestimmten Clusterknotens des Dendrogramm, dass Einfluss auf den Grad des dargestellten Clusterings auf der SOM hat. Zudem wird der im Dendrogramm ausgewählte Cluster farblich in der Karte hervorgehoben.

Der zu den Baumknoten dargestellte Text zeigt die Anzahl der Units in diesem Cluster sowie die Koordinaten des Cluster-Zentroiden auf der Map dar.

Evaluierung

Die Verwendung des Dendrogramm Controls bietet einige Vorteile gegenüber der Verwendung des bisher vorhandenen Spinner Controls:

- Besseren Überblick über die Struktur des Clusterings.
- Selektion eines bestimmten Clusters.
- Einblick in die Unterschiede der einzelnen Clusteringalgorithmen.
- Tooltipps ermöglichen die Anzeige von zusätzlichen Informationen über die einzelnen Cluster.

Als möglicher Nachteil muss aufgeführt werden, das die Navigation durch einen mitunter großen Baum nicht immer ganz einfach ist. Durch die Möglichkeit mit dem Mausrad zu zoomen wurde versucht die Navigation zu vereinfachen.

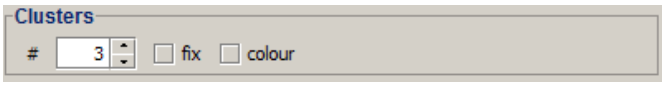

Abb. 3: Steuerung der Clusterdarstellung mithilfe des alten Spinner Controlelements

Task - Darstellung der Clusterdistanzen

Zeichnen Sie zwischen den einzelnen Knoten Verbindungslinien in unterschiedlicher Stärke, die die Distanz im Baum darstellen, ein.

Screenshot

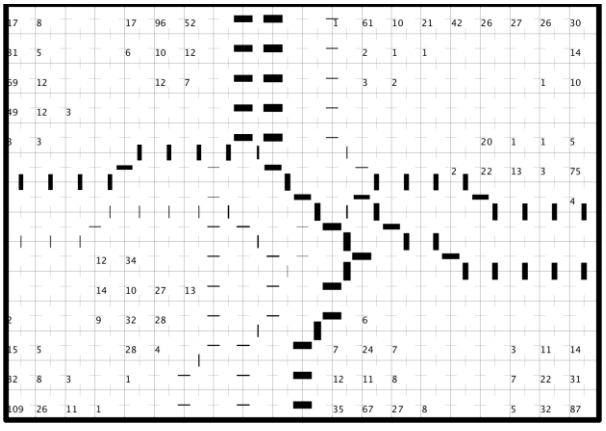

Abb. 4: Darstellung der Clusterdistanzen (ward's linkage - expiremental)

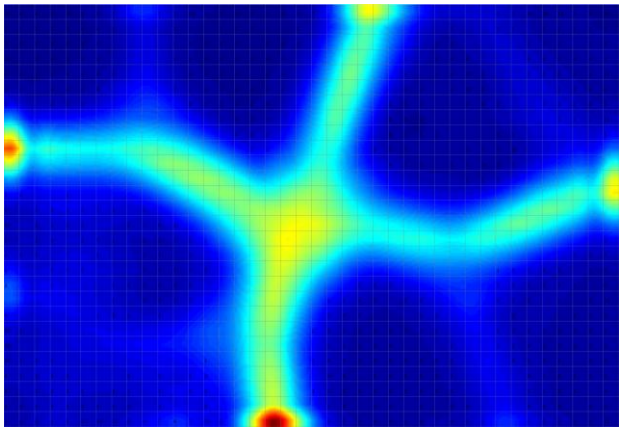

Abb. 5: D-Matrix Visualisierung

Implementierung

Um die Cluster-Distanzen darzustellen wurde jede SOM Unit als Leaf-Node im Clusterbaum betrachtet. Zwischen jeweils zwei adjazenten SOM Units wurde nun eine Linie (ähnlich der ClusterConnections Visualisierung) gezeichnet. Die Dicke dieser Linie ist abhängig von den Merge Kosten der Nodes. Die Merge Kosten wurden mit Min-Max Scaling auf eine Dicke zwischen 1 und 40 Pixel skaliert.

Evaluierung

Die Distanzen der einzelnen Knoten im Dendrogramm ist ein wesentlicher Anhaltspunkt für die Grenzen der wichtigsten Cluster der Karte. Abb. 4 zeigt die Distanzen der einzelnen Knoten im Baum für eine Beispielkarte. Abb. 5 zeigt die D-Matrix Visualisierung derselben Self Organizing Map. Auf den ersten Blick fällt sofort die Ähnlichkeit der beiden Abbildungen auf. Ein Grund dafür ist sicherlich die beidseitige Fokussierung auf die Distanzen der einzelnen Knoten.
Die Darstellung der Distanzen der einzelnen Knoten in der SOM stellt somit eine nützliche zusätzliche Möglichkeit dar, die Clustergrenzen herauszufinden.

Task - Darstellung der G*üte der einzelnen Cluster*

Verwenden Sie die Information aus dem Baum, um die "Güte" der einzelnen Cluster auf der Karte darzustellen. Dies könnte z.b. durch eine Farbcodierung geschehen, durch die Cluster mit ähnlich hohen Werten im Dendogram identifizierbar sind.

Screenshot

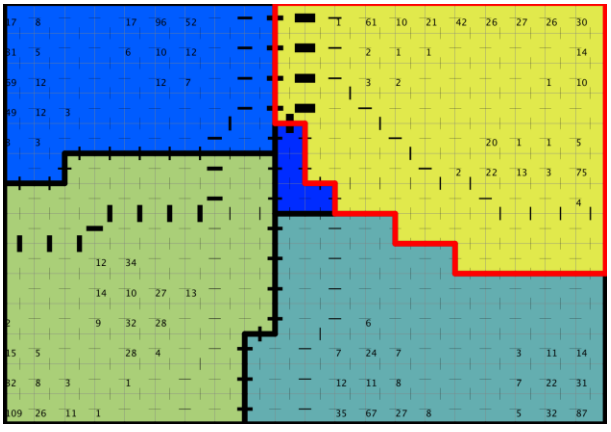

Abb. 6: Güte-Darstellung der Cluster (5 Cluster)

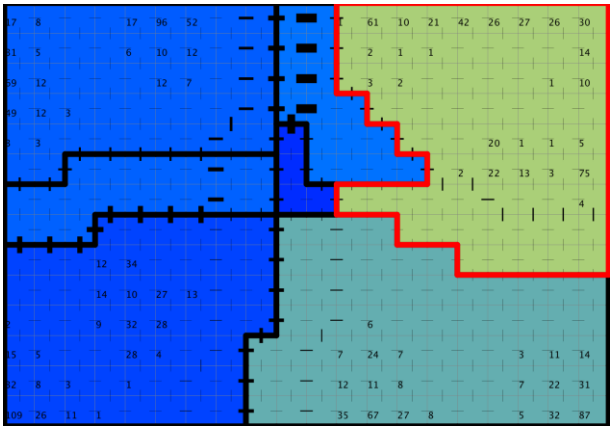

Abb. 7: Güte-Darstellung der Cluster (7 Cluster, Child-Cluster aus Abb. 6)

Implementierung

Um die “Güte” eines einzelnen Clusters darzustellen wurden die Merge Kosten der Cluster betrachtet. Die Merge Kosten ist die Distanz zwischen zwei Clustern, diese ist abhängig vom jeweiligen Verfahren. Vom Clusteringverfahren werden jeweils die zwei Cluster, zwischen denen die geringsten Merge Kosten festgestellt wurden, gemerged.
Die Merge Kosten eines Clusters können somit im Verhältnis zu den Merge Kosten der anderen Cluster im Baum betrachtet werden. Unter Verwendung von Min-Max Scaling werden die Merge Kosten der Cluster auf die aktuell ausgewählte Palette gemapped und dementsprechend eingefärbt.

Evaluierung

Im Gegensatz zur bisher vorhandenen “zufälligen” Farbcodierung der Cluster (aufteilen der Palette unter den Clustern), bietet die Codierung der Güte anhand von Farben einen Informationsgewinn.
Es ist somit auf einfache Weise möglich dieses zusätzliche Qualitätsmaß abzulesen und Vergleiche zwischen unterschiedlichen Clusteringlevels herauszufinden.
Ein weiterer Vorteil ist die Möglichkeit herauszufinden, ab wann das Clustering ausreichend und eine weitere Unterteilung in Sub-Cluster abgebrochen werden kann. Dies ist gut auf den Abbildungen 6 & 7 zu sehen. Die Cluster auf der linken Seite haben in Abbildung 7 bereits relativ niedrige Merge Kosten weshalb es vermutlich nicht mehr Sinn macht über 5 Cluster (wie in Abbildung 6) darzustellen.