Aufgabe 5.2: Clustering (P)_Artischocken

Zentraler Bestandteil der für das Wirkstoffdesign verwendeten Screening- und Analysemethoden ist das Clustering. Das Ziel dieser Methodik ist es, aus einer großen Substanzbibliothek einen deutlich kleineren Datensatz zusammenzustellen, der Substanzen enthalten soll, die sich in bestimmten Eigenschaften möglichst ähnlich bzw. möglichst unähnlich sind. Die Kriterien, nach denen geclustert wird, hängen stark von der beabsichtigen Verwendung des gewünschten Datensatzes ab. So sollte eine initiale Screeningbibliothek nach Möglichkeit aus Verbindungen bestehen, die sich in bestimmten Eigenschaften möglichst unähnlich sind, wohingegen die Ergebnisse eines Screeningdurchlaufs sinnvollerweise zu Clustern möglichst ähnlicher Verbindungen zusammengefaßt werden sollten.

Nr.	x	y
1	1	1
2	2	4
3	2	5
4	2	7
5	4	4
6	4	6
7	7	1
8	8	2
9	8	3
10	9	1

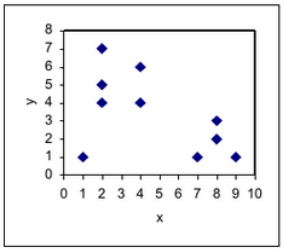

Abbildung 2: Abbildung des zu verwendenden Datensatzes.

1. Beschreiben Sie zunächst die konzeptionellen Unterschiede zwischen hierarchischen und partitionalem Clustering. Wo sehen Sie Vor- und Nachteile der einzelnen Ansätze, bzw. für welche Art von Daten sind diese jeweils geeignet?

2. Zeichnen Sie zu dem in Abbildung 2 gezeigtem Datensatz die Dendrogramme für single-linkage, complete-linkage und average-linkage Cluster-Analyse. Verwenden Sie dazu als Abstandsmaß die Euklidische Distanz. Interpretieren Sie die Unterschiede in den Dendrogrammen.

1.) Konzeptionelle Unterschiede zwischen hierarchischem und partitionalem Clustering
Clustering beschreibt die Gruppierung einer Datenmenge gemäß eines Ähnlichkeitsmaßes.[1]

Beim hierarchischen Clustering kann agglomerativ oder divisiv vorgegangen werden, wobei ein Baum entweder bottom-up oder top-down erstellt wird.[2]
Beim hierarchischen agglomerativen Clustering wird durch sukzessives Vereinigen von Clustern eine Baumstruktur über der Datenmenge aufgebaut. Die Grundidee des hierachischen Clustering ist die, dass nah beieinander liegende Punkte näher miteinander verwandt sind als weiter entfernt voneinander liegende Punkte.[3] Die verschiedenen Algorithmen des hierarchischen agglomerativen Clusterings unterschieden sich dabei durch die angewandten Distanzfunktionen.[2] So werden beim single-linkage/nearest neighbour Clustering immer die Objekte mit geringster Distanz zu einem Cluster zusammengefasst, beim complete-linkage/furthest neighbour werden die Objekte mit maximaler Distanz, bei der Methode mit group average wird der Mittelwert der Punkte gewählt und bei der centroid cluster analysis wird ein centroid der Cluster bestimmt und für das Distanzmaß verwendet.[1]

Bei partitionalem Clustering, wie z.B. dem k-means-Algorithmus, wird zuerst die Anzahl der Cluster festgelegt, dann werden iterativ die Zentren der einzelnen Cluster bestimmt bis die Zuordnungen der Punkte zu den einzelnen Clustern konstant sind. Dabei wird eine Fehlerfunktion minimiert. [4] Beim k-means-Algorithmus werden dabei die Objekte initial zu den vorher durch die Clusteranzahl bestimmten Clustern zugeteilt. Mittels einer Clustering-Gütefunktion werden dann iterativ die Objekte in den einen oder anderen Cluster verschoben bis die Zuordnung nicht weiter optimiert werden kann. [1]

---

1 Vorlesung Prof. Rarey, Kapitel 3 Molekulare Ähnlichkeit, S. 38-48
2 https://en.wikipedia.org/wiki/Hierarchical_clustering (27.06.2020)
3 https://en.wikipedia.org/wiki/Cluster_analysis#Connectivity-based_clustering_(hierarchical_clustering) (27.06.2020)
4 https://de.wikipedia.org/wiki/Clusteranalyse#Partitionierende_Clusterverfahren (27.06.2020)

Tabelle 1: Vor- und Nachteile von hierarchischem vs. partitionalem Clustering

Hierarchisches Clustering	Partitionales Clustering
Vorteile	
-Anzahl der Cluster wird nicht zu Anfang des Clustering bestimmt, sondern ergibt sich im Laufe	-Zuordnungen von Objekten zu Anfang des Clustering können sich im Laufe des Algorithmus ändern[4]
-Untersch. Distanzfunktionen können verwendet werden[2]	-Laufzeitkomplexität O(nkdi) – i.d.R für heuristische Version des k-means (Loyd's) (n=Anz. Dimensionen, k=Anz. Cluster, i=Anz. Iterationen bis Konvergenz)[6]
Nachteile	
-Objekte, die zu Anfang des Algorithmus bestimmten Clustern zugeordnet wurden, ändern ihre Zuordnung im Lauf des Algorithmus nicht mehr	-Anzahl der Cluster muss zu Anfang festgelegt werden[4]; schwierige auszuwählen[6]
-nicht sehr robust gegenüber Ausreißern („chaining phenomenon")[5]	-Bestimmung der initialen Zuordnung ist mehr oder weniger wahllos und die Verbesserung durch Mittelwertbildung ist fragwürdig; u.A. weil die Bewertung und das Resultat auf der ursprünglichen Annahme einer best. Anzahl an Clustern basiert
-Laufzeitkomplexität i.d.R.: $O(n^3)$- agglomerativ $O(2^{n-1})$- divisiv[5] -Speicherplatz: $O(n^2)$-agglomerativ[2]	-Ausreißer sind nicht erkennbar[6] -keine Auswahl zwischen mehreren Distanzfunktionen[6]
→i.d.R. zu langsam für große Datensätze[5]	→ Verwendung für große Datensätze, besonders bei Verwendung von Heuristiken; oft auch Verwendung als Vorverarbeitung von Daten[6]

5 https://en.wikipedia.org/wiki/Cluster_analysis#Connectivity-based_clustering_(hierarchical_clustering) (27.06.2020)

6 https://de.wikipedia.org/wiki/K-Means-Algorithmus (27.06.2020)