UNIVERSITÄT POTSDAM

Andreas Nastansky (Hrsg.)

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 55

Andreas Nastansky

Gruppierung von Daten:
Topologische Verfahren vs. Clusteranalyse

Potsdam 2022

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 55

Andreas Nastansky

Gruppierung von Daten: Topologische Verfahren vs. Clusteranalyse

Autoren: Prof. Dr. Andreas Nastansky, Hochschule für Wirtschaft und Recht (HWR) Berlin, Email: andreas.nastansky@hwr-berlin.de

Herausgeber: Prof. Dr. Andreas Nastansky, Hochschule für Wirtschaft und Recht (HWR) Berlin, Professur für Quantitative Methoden und Mathematik, Email: andreas.nastansky@hwr-berlin.de
2022

Danksagung: Ich danke Herrn M.Sc. Dariusz Lesniowski für seine Vorarbeiten und wertvolle Unterstützung.

Online veröffentlicht auf dem Publikationsserver der Universität Potsdam:
https://doi.org/10.25932/publishup-57272

Zusammenfassung

Dieser Beitrag beinhaltet einen Vergleich zwischen den Methoden der Topologischen Datenanalyse (TDA) und statistischen Clusterverfahren bei der Gruppierung von Daten. Es werden Gemeinsamkeiten und Unterschiede bei der Bildung der Cluster und Zuordnung der statistischen Einheiten identifiziert. Hierzu werden zwei empirische Datensätze aus der Biologie und Medizin herangezogen.

Zusammengefasst haben sich die Verfahren der TDA als ein praktikables Werkzeug bei der Gruppierung von Objekten erwiesen. Vor allem mit dem Mapper-Algorithmus konnten adäquate Cluster erkannt werden. Beim *Iris Flower*-Datensatz hat die TDA ähnliche Ergebnisse wie die Clusteranalyse erzielt. Der *Heart Disease*-Datensatz war schwieriger zu behandeln. Die genutzten clusteranalytischen Verfahren waren nicht geeignet, die beiden Gruppen von Patienten korrekt zu identifizieren. Im Vergleich zu den Standardverfahren der Clusteranalyse zeigte sich eine leichte Überlegenheit der topologischen Verfahren.

Abstract

This paper includes a comparison between Topological Data Analysis (TDA) methods and statistical clustering methods in grouping data. Similarities and differences in the formation of clusters and assignment of statistical units are identified. Two empirical data sets from biology and medicine are used for this purpose.

In summary, the procedures of TDA have proven to be a viable tool in grouping objects. Especially with the mapper algorithm adequate clusters could be detected. For the *Iris Flower*-dataset, TDA produced similar results to cluster analysis. The *Heart Disease*-dataset was more difficult to deal with. The used cluster analytic techniques are not capable of correctly identifying the two groups of patients. Compared with the standard cluster analysis methods, the topological procedures showed a slight superiority.

1 Einleitung

Bei der Analyse von höherdimensionalen Daten kann deren gegenseitige räumliche Anordnung im von den Variablen (Merkmalen) aufgespannten Raum wichtige Informationen über den Datensatz liefern. Mit Hilfe der Topologie ist es möglich, gleiche Strukturen in verschiedenen Räumen zu entdecken. Bei einer gegebenen Punktwolke, die aus einem unbekannten topologischen Raum ausgewählt wurde, versucht die Topologische Datenanalyse (TDA) den ursprünglichen Raum zu rekonstruieren. Die Topologische Datenanalyse, d. h. ein Gebiet der Mathematik das topologische Methoden zur Untersuchung reeller Daten nutzt, geht u.a. auf Carlsson (2009) zurück und hat in den letzten Jahren an Bedeutung gewonnen und findet z. B. bei der Gruppierung von Daten als Alternative zu klassisch statistischen Verfahren wie der Clusteranalyse Anwendung. So nutzten Lum et al. (2013) Methoden der TDA bei der Clusterung von Daten aus den Bereichen Medizin, Politik und Sport [9]. Dabei stellten sie zum Teil große Abweichungen zwischen den statistischen und topologischen Gruppierungen fest.

Dieser Beitrag beinhaltet einen Vergleich zwischen den Methoden der TDA und statistischen Clusterverfahren bei der Gruppierung von Daten. Es werden Gemeinsamkeiten und Unterschiede bei der Bildung der Cluster und Zuordnung der statistischen Einheiten identifiziert. Hierzu werden zwei empirische Datensätze aus der Biologie und Medizin herangezogen. Das erste Anwendungsbeispiel beinhaltet Informationen über drei Schwertlilien-Gattungen. Es wird untersucht, wie viele Klassen von Blumen im Datensatz existieren. Das zweite Anwendungsbeispiel enthält Patientendaten - mit dem Ziel der Vorhersage einer Herzkrankheit.

Nastansky (2019) gibt eine Einführung in die zentralen Konzepte der Topologischen Datenanalyse: Persistente Homologie und Mapper [10]. Die Persistente Homologie ist eines der Standardwerkzeuge in der TDA. Sie findet ihre Anwendung beispielsweise in den Bereichen Formerkennung und -beschreibung. Der Mapper-Algorithmus als zweites wichtiges Konzept der TDA wandelt umfangreiche, höherdimensionale Datensätze in Simplizialkomplexe um und kann dadurch geometrische und topologische Eigenschaften der Daten bestimmen. Des Weiteren ist die Mapper-Methode ein brauchbares Werkzeug zur Clusterung und Visualisierungen von mehrdimensionalen Daten.

Die Topologie beschäftigt mit der Form eines Objektes. Dadurch kann sie dem Nutzer quantitative Informationen über die Form des Datensatzes liefern. Die Topologie repräsentiert hierbei ein mathematisches Fachgebiet, das u.a. verschiedene topologische Räume zu klassifizieren versucht. Solche Räume besitzen gleiche topologische Eigenschaften und lassen sich stetig aufeinander abbilden. Mit Hilfe der Topologie ist es möglich, gleiche Strukturen in verschiedenen Räumen zu entdecken. Bei der Arbeit mit höherdimensionalen Daten ist es elementar, in den großen und häufig schwer überschaubaren Datensätzen Muster zu finden. Ein Ziel der TDA besteht in der Darstellung von Daten als topologischer Raum. Das schließt die Erkennung von Wegzusammenhangskomponenten ein, die als homogene Cluster interpretiert werden können.

Topologische Methoden der Datenanalyse weisen gegenüber klassischen statistischen Verfahren mehrere Vorteile auf: Auf der einen Seite erfordert die quantitative Analyse von Daten zunächst eine Vorstellung über die qualitative Struktur der Daten und topologische Methoden liefern gerade diese qualitativen Aspekte geometrischer Objekte. Auf der anderen Seite sind Daten üblicherweise als Punktmengen im Raum gegeben, wobei es keine theoretische Rechtfertigung einer konkreten Wahl von Koordinaten oder einer Metrik gibt. Die Topologie untersucht die geometrischen Eigenschaften, die bei stetigen Transformationen gleich bleiben. Dadurch können gegebene Daten transformiert werden, sodass sie besser zu verstehen und auszuwerten sind. Überdies liefert die Topologie mittels der Teilung eines Objekts in Wegzusammenhangskomponenten eine natürliche Klassifikation in Klassen. Distanz oder Krümmung spielen in der Topologie eine weitaus geringere Rolle als zum Beispiel in der Geometrie. Dadurch können einfacher Muster in Datensätzen gefunden werden, in denen kein natürliches Konzept der Metrik existiert. Topologische Methoden sind zudem relativ invariant gegenüber der Anwendung unterschiedlicher Koordinaten oder Abstandsmaße.

Der Beitrag ist wie folgt gegliedert: In Kapitel 2 werden mehrere Standardmethoden der Custeranalyse (K-Means, drei hierarchische Verfahren, DBSCAN) skizzert. Anschließend werden die Ergebnisse der Clusterung zweier Datensätze der statistischen Verfahren mit den Methoden der TDA verglichen. Der Beitrag endet mit einer kritischen Diskussion der Methodik.

2 Clusteranalyse

Die Clusteranalyse repräsentiert eine Strukturen-entdeckende multivariate Analysemethode und strebt eine Bündelung von Objekten an (siehe u.a. [3, 4, 12, 15]). Das Ziel der Clusteranalyse besteht darin, Objekte eines Datensatzes S zu Gruppen (Cluster) mit ähnlichen Eigenschaften zusammenzufassen. Die Objekte innerhalb der Gruppen sind möglichst homogen und die Gruppen untereinander sind möglichst heterogen. In der Literatur existiert eine Vielzahl von Clusterverfahren, die für unterschiedliche Anwendungsgebiete entwickelt wurden. Dabei werden die folgenden Verfahrenstypen unterschieden: partitionierende, hierarchische und dichte-basierte Verfahren. Die ersten beiden Verfahrenstypen zählen zu den klassischen Clustermethoden; während dichtebasierte Verfahren eher neueren Datums sind.

Partitionierendes Verfahren

Das K-Means-Verfahren oder auch iteriertes Minimaldistanzverfahren ist das wichtigste partitionierende Clusterverfahren [12]. Das Verfahren erfordert vorab die Festlegung der Anzahl der Cluster und ordnet die Objekte entsprechend der zur Clusterung herangezogenen metrisch skalierten Variablen den Gruppen zu. Hierzu werden Clusterzentren benutzt, um K Cluster in einem Datensatz S zu finden. Das Vorgehen kann wir folgt skizziert werden:

1. Der Algorithmus startet mit K Startwerten/Startzentren c_i.
2. Jeder Punkt aus einem Datensatz wird einem Zentrum zugewiesen, zu dem die Distanz am kleinsten ist. So entstehen K Cluster C_i.
3. Im jeden Cluster wird das Zentrum c_i neu berechnet.
4. Hat sich ein Zentrum c_i verändert, dann wird zu 2. zurückgegangen.

Dabei werden die Zentren so berechnet, dass die Summe der quadratischen Abstände bzw. Abweichungen SSE

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, c_i)^2 = \sum_{i=1}^{K} \sum_{x \in C_i} (x - c_i)^2.$$

minimal ist mit

$$c_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

wobei n_i die Anzahl der Punkte im Cluster c_i ist [15]. Die Zentroid-Berechnung beim K-Means-Verfahren als Mittelwert der Cluster-Elemente entspricht einer lokalen Varianzminimierung [12]. Zur Distanzberechnung wird die einfache euklidische Distanz herangezogen.

Die Startwerte können auf verschiedener Art gewählt werden. Die ersten K Zentren können komplett zufällig aus dem Datensatz ausgesucht sein. Da die Konstellation der Startzentren den Output des Algorithmus beeinflusst, kann das zu unbrauchbaren Ergebnissen führen. Eine Lösung für dieses Problem besteht darin, den Algorithmus mehrfach zu wiederholen. Eine Alternative dazu, ist das erste Zentrum zufällig zu wählen. Folglich stellt jedes nächste Zentrum den Punkt dar, der am weitesten von den anderen Zentren entfernt liegt. Hier wiederum entsteht die Gefahr, dass ein Ausreißer als Zentrum ausgesucht wird, was zu leeren Clustern führen kann. Das kann vermieden werden, indem zunächst eine Stichprobe aus dem Datensatz gezogen wird und im Anschluss daran das Suchvorgehen implementiert wird.

Ein Vorteil des K-Means-Verfahrens repräsentiert seine hohe Effizienz - insbesondere bei großen Datenmengen [12]. Wie bei Methoden der Streuungsminimierung unter Verwendung von Mittelwerten üblich, reagiert das Verfahren empfindlich auf Ausreißer im Datensatz. Sie sollten möglichst vorher eliminiert werden. Ein weiteres Problem stellt die ex ante Vorgabe der Anzahl der Cluster dar. Wenn K unbekannt ist, muss das Verfahren mit verschiedenen K angewandt werden und die Ergebnisse müssen auf

Plausibilität geprüft werden. Außerdem können während des Algorithmus leere Cluster entstehen (zum Beispiel durch eine unzureichende Wahl der Startwerte), die dann auch bis zum Ende leer bleiben werden. Bei einer vorab festgelegten Anzahl der Cluster ist das eher kontraproduktiv. Außerdem erkennt das Verfahren kugelförmige Cluster besser, d. h. die Form der Cluster kann eine Rolle spielen.

Hierarchische Verfahren

Hierarchische Clusterverfahren unterteilen sich in zwei Gruppen: Bei den divisiven (absteigenden) Verfahren wird am Anfang der komplette Datensatz als ein Cluster betrachtet und dann sukzessiv in kleinere Cluster unterteilt [12]. Die agglomerativen (aufsteigenden) Methoden funktionieren umgekehrt. Zuerst wird jeder Punkt als ein Cluster betrachtet und mit jedem Schritt werden die Punkte iterativ zu größeren Clustern verschmolzen.

Das Ergebnis einer hierarchischen Analyse wird gewöhnlich in einem Dendrogramm (Hierarchie von Clustern) dargestellt (siehe Abbildung 1). Neue Cluster einstehen bei der Distanz d, die aus dem Dendrogramm abgelesen werden kann; gemeinsam mit den Punkten, die den neuen Cluster bilden.

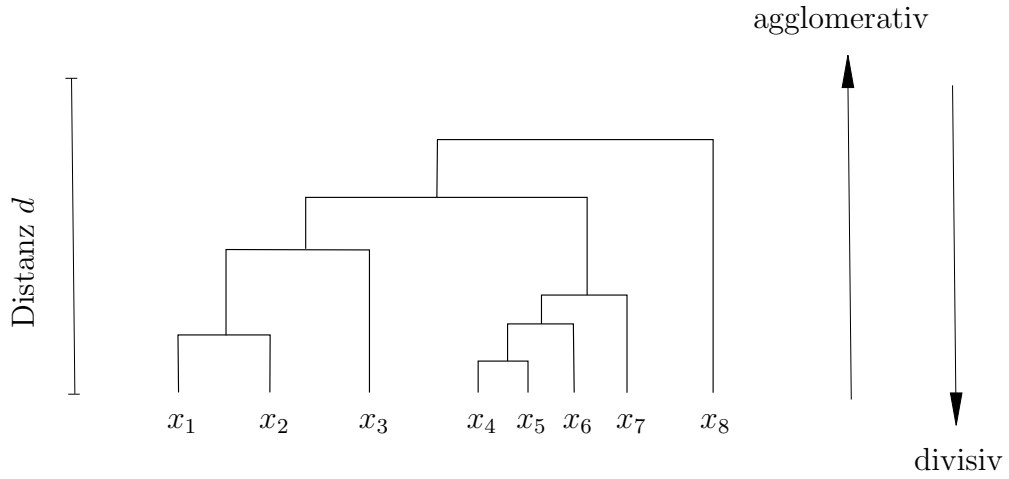


Abbildung 1: Dendrogramm der hierarchischen Clusteranalyse

Für die Praxis sind die agglomerativen Verfahren von größerer Bedeutung. Deren Ablauf lässt sich wie folgt skizzieren [4]:

1. Der Algorithmus startet mit der feinsten Partition $C_i = x_i$.
2. Eine Distanzmatrix D wird berechnet, welche die Abstände $d(C_i, C_j)$ aller Cluster zueinander enthält:

$$D_{n\times n} = \begin{pmatrix} 0 & d(C_1, C_2) & \dots & d(C_1, C_n) \\ d(C_2, C_1) & 0 & \dots & d(C_2, C_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(C_n, C_1) & d(C_n, C_2) & \dots & 0 \end{pmatrix}.$$

3. Die Cluster C_k und C_l werden so bestimmt, dass deren Distanz d am kleinsten ist, d.h.
$$d(C_k, C_l) = \min_{i \neq j} d(C_i, C_j).$$

4. C_k und C_l bilden zusammen neue Cluster C'_k.

5. Existieren mehr als zwei Cluster, dann wird zu 2. zurückgegangen.

Für die agglomerative Methoden ist nicht nur die Distanz zwischen den Punkten, sondern auch zwischen den einzelnen Clustern notwendig. Der Output des Algorithmus hängt stark mit der Definition dieser Distanz zusammen. Die populärsten agglomerativen Methoden und deren Distanzen sind:

Single-Linkage-Verfahren: $d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x_i, x_j)$,

Complete-Linkage-Verfahren: $d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x_i, x_j)$,

Average-Linkage-Verfahren: $d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i, y \in C_j} d(x_i, x_j)$.

Das Single-Linkage-Verfahren (Nächstgelegener Nachbar) basiert auf minimalen Abständen. Es erkennt relativ gut zusammenhängende Komponenten; weist aber Probleme mit Ausreißern und Clustern mit verschiedenen Dichten auf. Das Complete-Linkage-Verfahren (Entferntester Nachbar) reagiert weniger empfindlich auf Ausreißer; tendiert aber zur Konstruktion von kleinen Gruppen und kann dadurch große Cluster zerreißen. Beim Average-Linkage-Verfahren ist die Distanz zwischen zwei Gruppen als Durchschnitt aller paarweisen Abstände definiert.

Die hierarchische Clusteranalyse kann auch beim Auftreten von extremen Beobachtungen eine robuste Gruppierung liefern. Ausreißer verschmelzen bei agglomerativen Verfahren erst sehr spät mit anderen Clustern und werden dadurch adäquat getrennt. Bei Clustern mit kleiner Dichte kann das wiederum zu Problemen führen, da die Verfahren „lokal“ arbeiten. Ein weiterer Nachteil ist die hohe Rechenintensivität dieser Algorithmen. Aus diesem Grund wird in der Praxis häufig zweistufig vorgegangen: zuerst eine Partition mit hierarchischen Verfahren bestimmen und anschließend mit einem anderen Verfahren (zum Beispiel K-Means-Verfahren) die Cluster bilden.

Hinzu tritt das Problem, dass die hierarchische Clusteranalyse als Ergebnis nicht nur eine Partition des Datensatzes liefert, sondern eine ganze Reihe. Die Wahl, welche Unterteilung am sinnvollsten ist, wird dem Anwender überlassen. Am Ende bleibt nur die Betrachtung des Dendrogramms und die Entscheidung nach „Gefühl und vorhandenem Wissen“. Eine Auswertung des Dendrogramms in Abbildung 1 würde zum Beispiel vier mögliche Cluster $C_1 = \{x_1, x_2\}$, $C_2 = \{x_3\}$, $C_3 = \{x_4, x_5, x_6, x_7\}$ und $C_4 = \{x_8\}$ suggerieren; eine andere Kombination mit nur zwei Cluster wäre aber auch denkbar.

Dichtebasiertes Verfahren

DBSCAN steht für Density-Based Spatial Clustering of Applications with Noise (Dichtebasierte räumliche Clusteranalyse mit Rauschen)[6]. Die dichtebasierte Clusterverfahren versuchen Regionen mit vielen Punkten zu einem Cluster zusammenzufassen (große Dichte) und nicht stark besiedelte Regionen zu eliminieren (Rauschen). Dabei kann die Dichte, genau wie „Ähnlichkeit“ der Elemente, unterschiedlich formuliert werden. Das DBSCAN Verfahren verwendet die zentrenbasierte Methode.

Definition 2.1: Seien S eine Punktwolke, $\epsilon \in \mathbb{R}_+$, $MinPts \in \mathbb{N}$. Außerdem sei für einen Punkt $x \in S$ die Anzahl der Punkte $y \in S$ für die gilt $d(x, y) < \epsilon$ als m bezeichnet. x heißt **Kernobjekt**, wenn $m \geq MinPts$ gilt. Gilt dagegen $m < MinPts$, aber gleichzeitig existiert ein Kernobjekt y, sodass $d(x, y) < \epsilon$, dann wird x ein **Dichte-erreichbare Objekte** genannt. Sonst ist x ein **Rauschpunkt**.

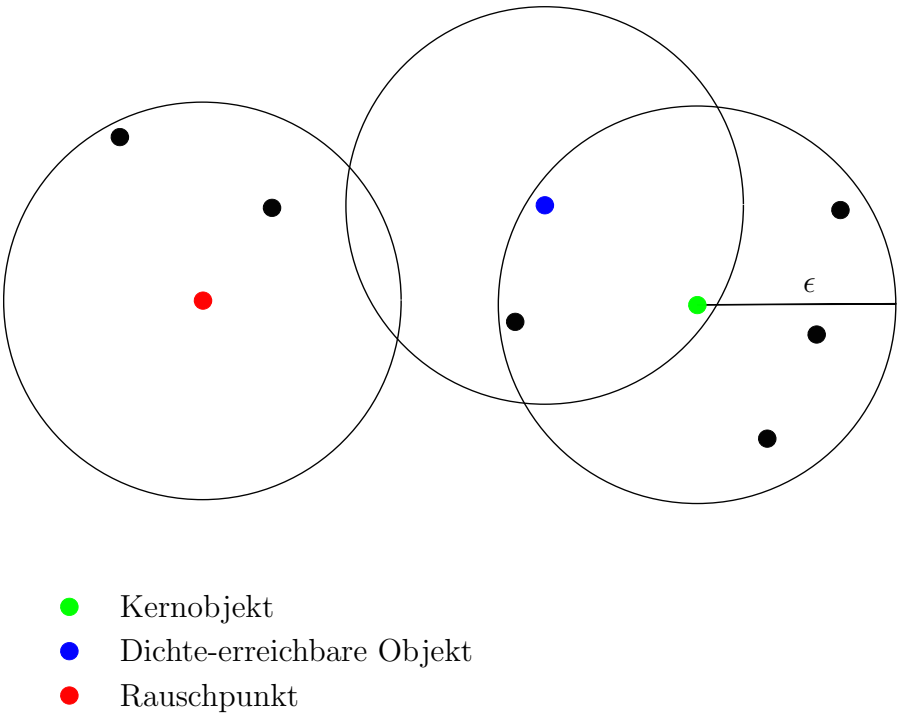


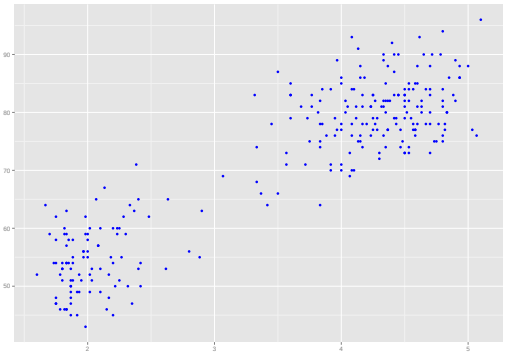
Abbildung 2: Arten von Punkten im DBSCAN Algorithmus ($MinPts = 6$)

Abbildung 2 zeigt eine graphische Veranschaulichung für Kernobjekte, dichte-erreichbare Objekte und Rauschpunkte. Der Verlauf des DBSCAN hängt von den beiden Parametern ϵ und $MinPts$ ab und sieht wie folgt aus:

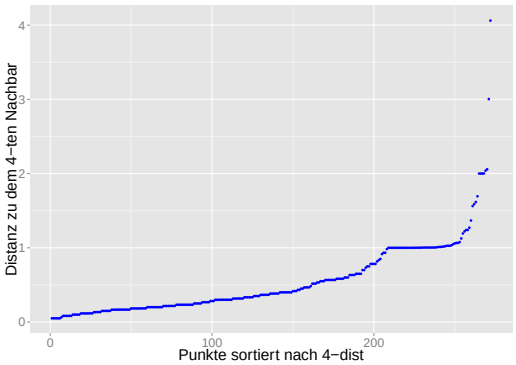
1. Kernobjekte, dichte-erreichbare Objekte und Rauschpunkte werden bestimmt.
2. Rauschpunkte werden eliminiert.
3. Zwei Kernobjekte x und y werden zu einem Cluster zugeordnet, wenn $d(x, y) < \epsilon$.

4. Ein dichte-erreichbares Objekt x und ein Kernobjekt y werden zu einem Cluster zugeordnet, wenn $d(x, y) < \epsilon$. Kann x zu mehreren Clustern zugeordnet werden, muss hier eine Entscheidung getroffen werden.

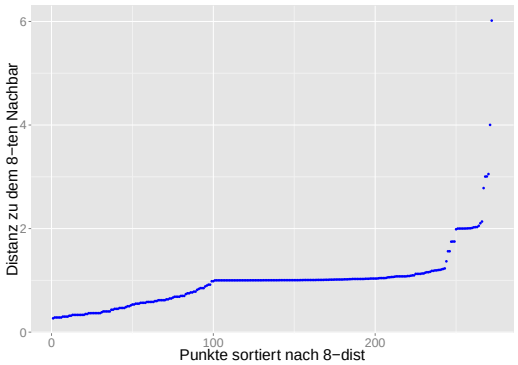
Für die Güte des dichtebasierten Verfahren DBSCAN ist die Wahl von $\epsilon \in \mathbb{R}$, $MinPts \in \mathbb{N}$ zentral. Ein üblicher Ansatz repräsentiert die Betrachtung des Verhaltens der Distanzen k-$dist$ eines Punktes zu seinem k-Nächsten-Nachbarn. Normalerweise wird k-$dist$ klein sein, wenn k kleiner als die Größe eines Cluster ist. Für die Rauschpunkte wird sie wiederum groß sein. Dabei soll die Distanz aller Punkte zu deren k-Nächsten-Nachbarn berechnet, sortiert und in einem Koordinatensystem abgebildet werden. Es wird eine starke Änderung des Verlaufes bei geeigneter Distanz erwartet. Diese Distanz k-$dist$ sollte dann als ϵ und k als $MinPts$ gewählt werden. Damit werden alle Punkte mit k-$dist < \epsilon$ als Kernobjekte markiert und alle anderen werden den zwei anderen Gruppen zugeordnet [6]. Obwohl ϵ hier mit Hilfe von k bestimmt wird, ändert es sich bei verschiedenen k nur schwach (siehe Abbildung 3).



(a) Datensatz S



(b) 4-$dist$ Plot



(c) 8-$dist$ Plot

Abbildung 3: Ein Datensatz S und zwei dazugehörige $k - dist$ Plots

Wenn k zu klein gewählt wird, werden die Rauschpunkte fälschlicherweise einem Cluster zugeordnet. Ein großes k verursacht, dass Cluster mit weniger als k Punkten als Rauschen bezeichnet werden.

Aufgrund der Dichtebasierung reagiert DBSCAN unempfindlich auf Ausreißer im Datensatz. Das Verfahren eignet sich auch bei Clustern verschiedener Größen und Formen,

wohingegen die partitionierenden Verfahren wie K-Means damit Probleme haben. Ein Nachteile des Verfahrens entsteht beim Auftreten von Clustern verschiedener Dichten. Darüber hinausgehend ist DBSCAN rechenintensiv.

Neben den vorgestellten Grundverfahren der Clusteranalyse existiert eine Vielzahl weiterer Algorithmen zur Gruppierung von Daten (siehe u.a. [3, 4, 12, 15]). Wichtig bleibt anzumerken, dass eine „gute Clusterung“ subjektiver Natur ist. Denn zuerst werden Cluster mit einem mathematischen Verfahren gebildet. Im Anschluss werden die gebildeten Gruppen betrachtet und möglichst fachlich fundiert interpretiert, indem Ähnlichkeiten in den geclusterten Objekten gesucht werden, die in der Lage sind, die Homogenität zu erklären.

3 Vergleich Gruppierung mit TDA und Clusteranalyse

Im Folgenden wird auf der Basis von zwei Datensätzen aus der Biologie und Medizin die Eignung topologischer Methoden bei der Gruppierung von Objekten untersucht und mit den Ergebnissen klassischer statistischer Verfahren verglichen. Hierbei werden Gemeinsamkeiten und Unterschiede bei der Bildung der Cluster und Zuordnung der statistischen Einheiten identifiziert. Das erste Anwendungsbeispiel beinhaltet Informationen über drei Schwertlilien-Gattungen. Es wird untersucht, wie viele Klassen von Blumen im Datensatz existieren. Das zweite Anwendungsbeispiel enthält Patientendaten - mit dem Ziel der Vorhersage einer Herzkrankheit.

Zur Berechnung der Persistenten Homologie wurde die Software JavaPlex genutzt. Auf Basis der Arbeit von [16] für BioMapper wurden in Anlehnung an [10] in R Modifizierungen, Fehlerbehebung und Anpassungen zur Berechnung von $Mapper(f^{-1}(\mathcal{V}), C)$ vorgenommen. Als Clusteralgorithmus C wird das Single-Linkage-Verfahren angewandt (siehe Kapitel 2). Die statistischen Berechnungen zur Clusteranalyse werden ebenfalls mit R durchgeführt.

3.1 Iris Flower-Datensatz

Der *Iris Flower*-Datensatz S (auch *Fisher-Anderson's Iris Flower*-Datensatz) stellt einen klassischen Datensatz für die Untersuchung der Funktionsfähigkeit von Methoden der Clusteranalyse und weiterer Verfahren z. B. des Maschinellen Lernens dar. Mit Hilfe des Datensatzes von Fisher und Anderson kann die Problematik der Identifikation von homogenen Gruppen beispielhaft demonstriert werden.

Der Datensatz beinhaltet 150 Beobachtungen von Schwertlilien (lat. Iris). Dabei wurden drei Gattungen untersucht: Iris Setosa, Iris Virginica und Iris Versicolor. Gemessen wurden die Länge und die Breite des Kelchblattes (Sepalum) sowie des Kronblattes (Petalum). Die Werte wurden vom Biologen Edgar Anderson [2] erhoben. Populär wurde der Datensatz aber durch den Bio-Statistiker Ronald A. Fisher, der die Güte seiner diskriminanzanalytischen Methode daran getestet hat [7]. Abbildung 4 zeigt den Datensatz mittels paarweiser Streudiagramme.

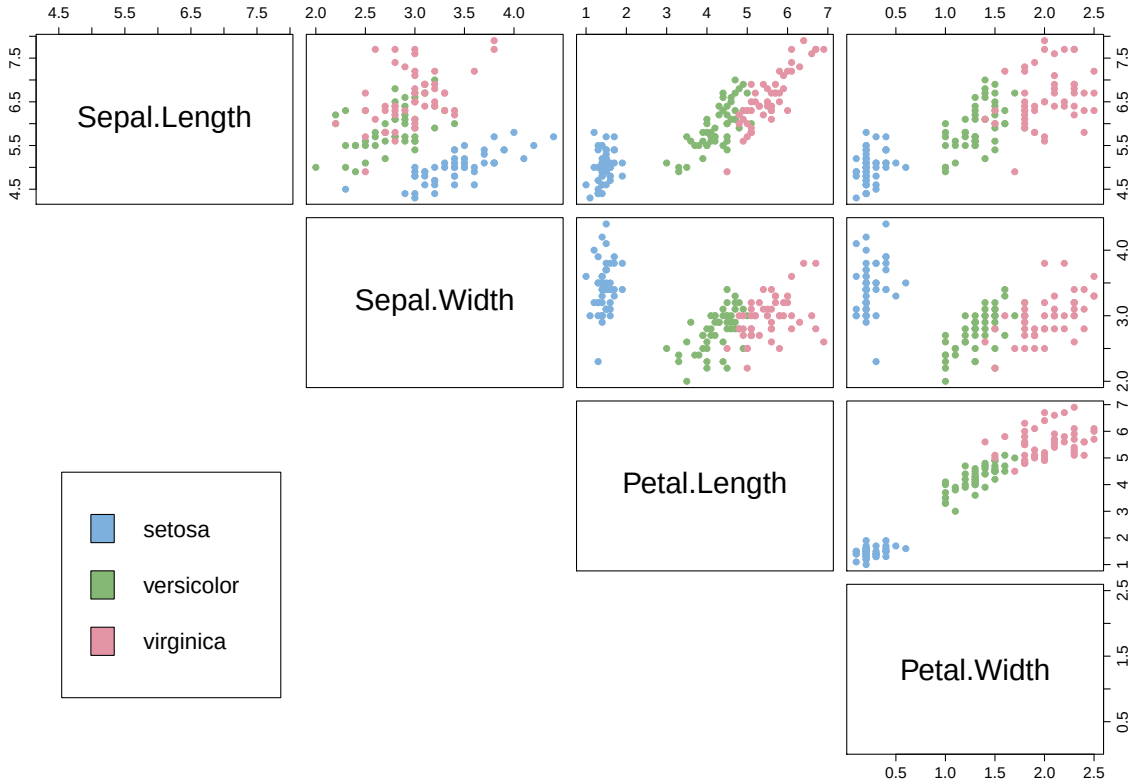


Abbildung 4: Iris Flower-Datensatz

Aus der Visualisierung wird ersichtlich, warum der *Iris Flower*-Datensatz clusteranalytische Verfahren vor große Herausforderungen stellt. Die Messungen der Breite und Länge der Gattungen Versicolor und Virginica liegen sehr nah beieinander und überschneiden sich teilweise. Demgegenüber hebt sich die Gattung Setosa deutlich von den beiden anderen Arten ab.

Partitionierendes Verfahren

Abbildung 5 zeigt, die mit dem K-Means-Verfahren gefundenen Cluster. Dieser Algorithmus liefert brauchbare Ergebnisse und weist für den *Iris Flower*-Datensatz eine Trefferquote von ca. 90%. Ein Vorteil von K-Means ist, dass die Anzahl der Cluster von Anfang an vorgegeben ist. Bei Vorgabe für den Parameter $K = 2$ verschmelzen Versicolor und Virginica zu einem Cluster. Für $K > 3$ ergeben sich keine sinnvollen Gruppen.

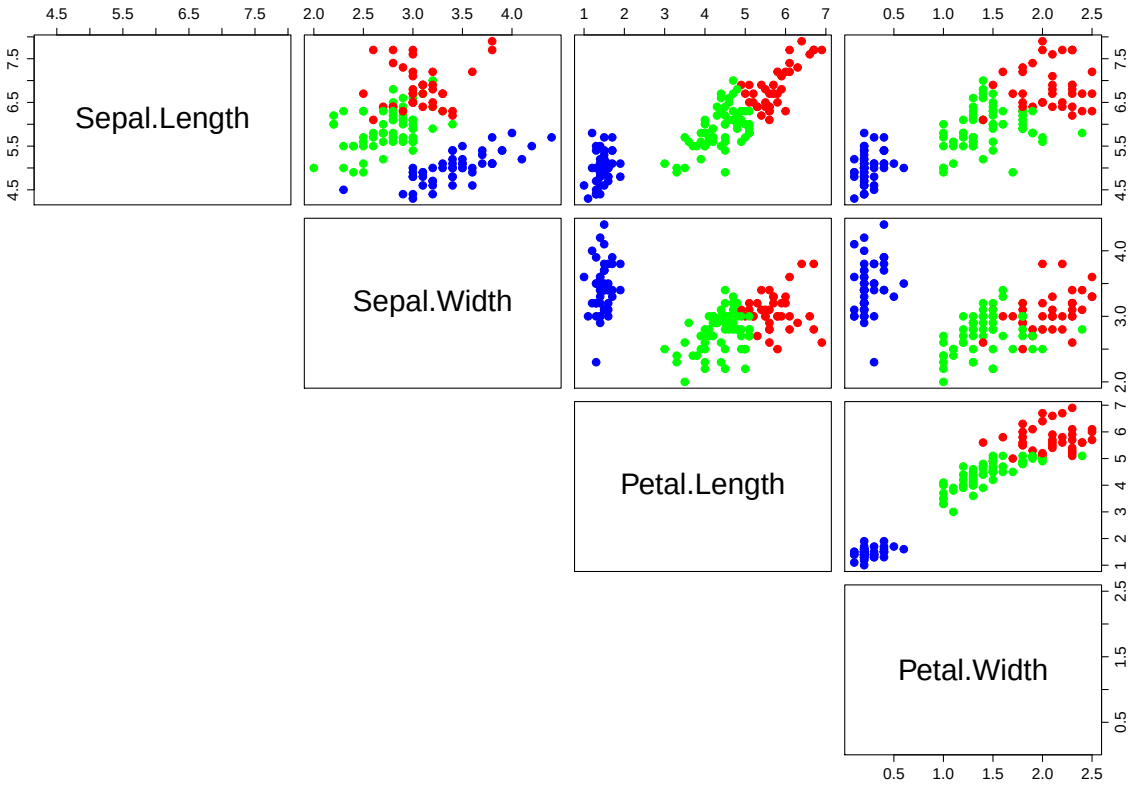
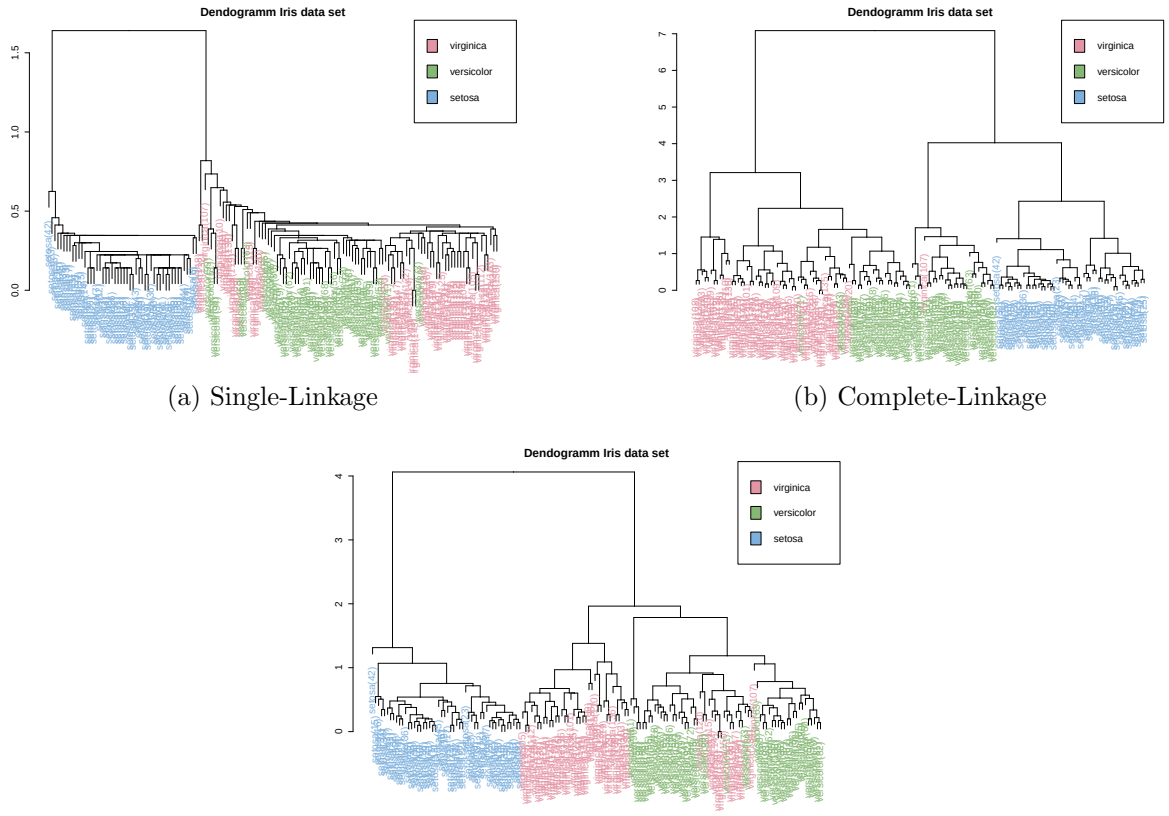


Abbildung 5: K-Means für den Iris Flower-Datensatz mit $K = 3$

Hierarchische Verfahren

Als Nächstes werden die drei folgenden hierarchischen Clusterverfahren auf den *Iris*-Datensatz angewandt: Single-Linkage, Complete-Linkage und Average-Linkage. Die Dendrogramme wurden einmal üblich (Abbildung 6) und wegen der besseren Lesbarkeit in der Kreisform (Abbildung 7) dargestellt. Das Dendrogramm des Single-Linkage-Verfahrens suggeriert zwei Cluster. Hingegen präferieren das Complete-Linkage und das Average-Linkage-Verfahren drei Cluster, wobei auch zwei Cluster als plausibel erscheinen. Bei drei Clusten vermischen sich die Gattungen Versicolor und Virginica und können nicht von adäquat voneinander getrennt werden (siehe Abbildung 7).

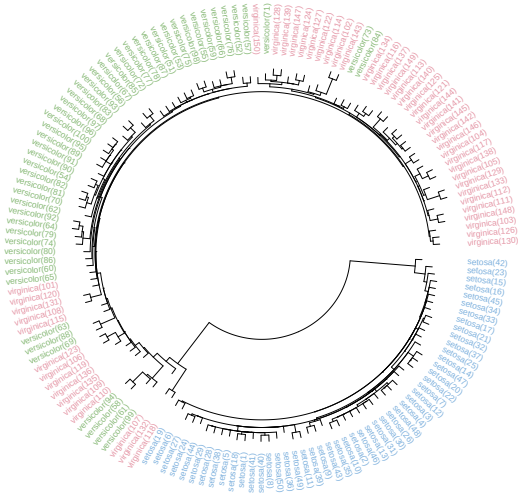


(a) Single-Linkage

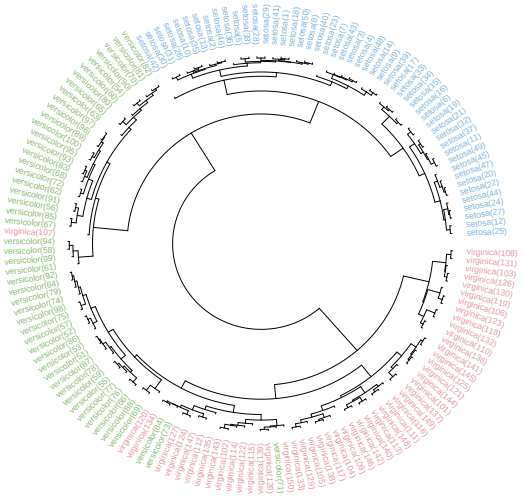
(b) Complete-Linkage

(c) Average-Linkage

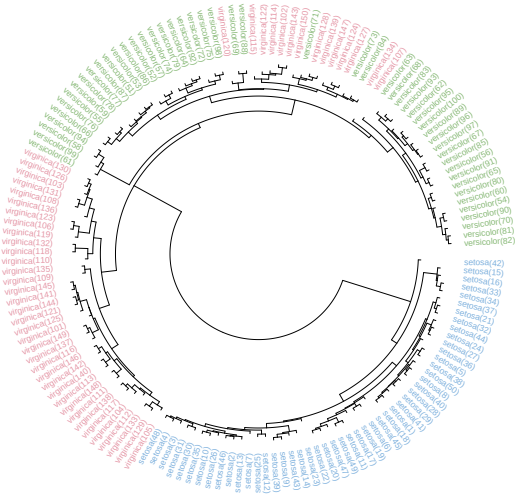
Abbildung 6: Dendrogramme der hierarchischen Verfahren



(a) Single-Linkage



(b) Complete-Linkage



(c) Average-Linkage

Abbildung 7: Kreisförmige Dendrogramme der hierarchischen Verfahren

Dichtebasiertes Verfahren

Die Parameter ϵ und $MinPts$ für DBSCAN wurden mit der k-$dist$-Methode bestimmt. Der Radius ϵ lag stets zwischen 0.5 und 1. Das dichtebasierte Clusterverfahren wurde wiederholt mit verschiedenen Parametern getestet - mit dem identischen Ergebnis von lediglich zwei Clustern. DBSCAN kann schlecht zwischen Virginica und Versicolor unterscheiden. Zwischen den beiden Gattungen liegt keine Region mit kleinerer Dichte und infolgedessen konnten beide Blumenarten nicht voneinander getrennt werden. Demnach liefert DBSCAN ähnliche Resultate wie die drei betrachteten hierarchischen Clusterverfahren.

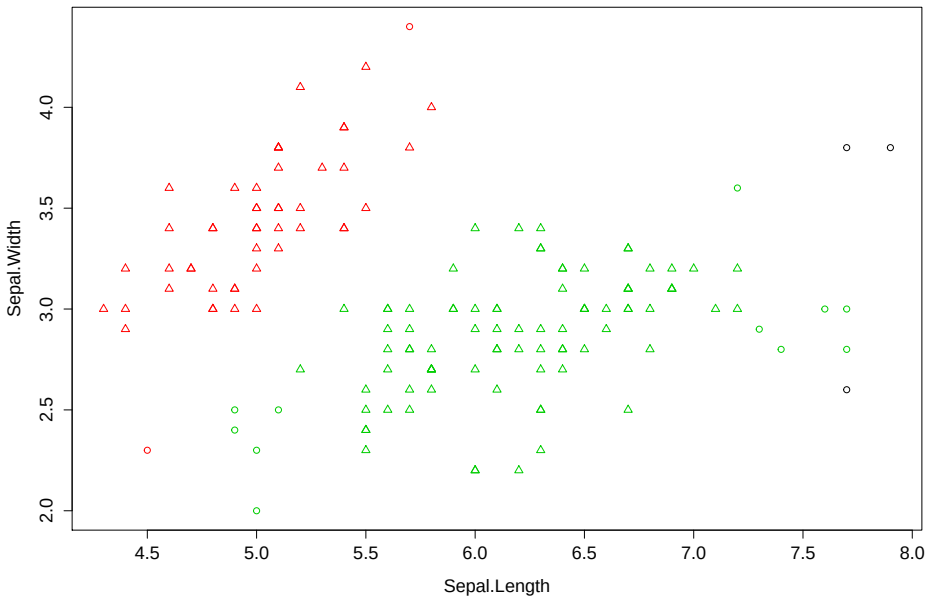


Abbildung 8: Beispiel Output des DBSCAN für den Iris Flower-Datensatz

In Abbildung 8 stellen Dreiecke die Kernobjekte und farbige Kreise die Dichte-erreichbaren Objekte dar. Schwarze Punkte bilden die Rauschpunkte ab.

Topologische Verfahren

Die Ergebnisse der statistischen Clusterverfahren werden im Weiteren der Gruppierung mit Hilfe der beiden topologischen Verfahren Persistente Homologie und Mapper gegenübergestellt. Als Erstes wird die Persistente Homologie des *Iris Flower*-Datensatzes berechnet. Dazu werden die Witness-Komplexe benutzt. Aus dem Datensatz werden zufällig 50 Landmarken L gewählt. Dabei gilt: $\epsilon = \max(w, l),\ w \in W,\ l \in L$ und die benutzte Filtration ist $W_{\frac{1}{30}\epsilon} \subseteq \ldots \subseteq W_{\frac{29}{30}\epsilon} \subseteq W_\epsilon$. In Abbildung 9 wurde die Persistente Homologie als Barcode dargestellt.

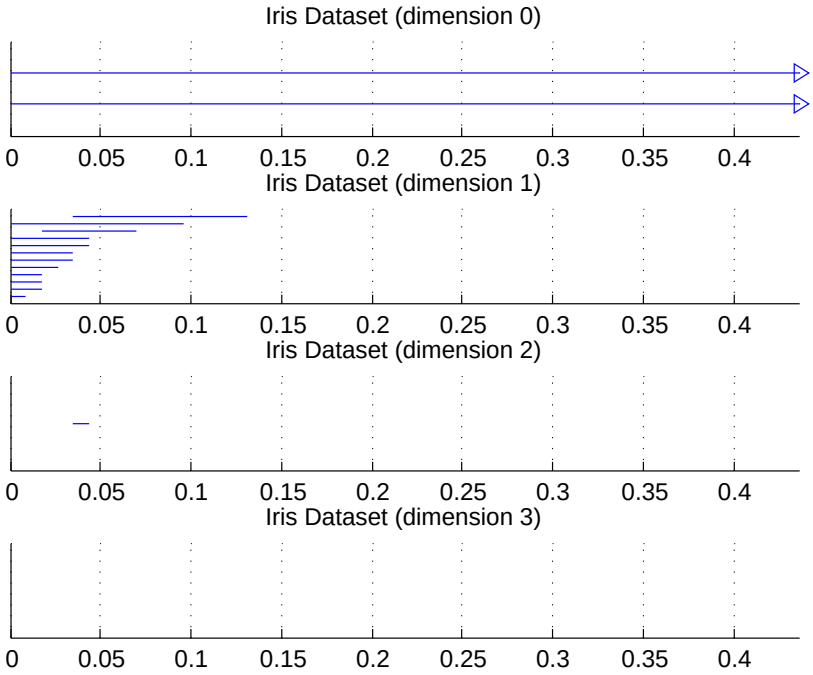


Abbildung 9: Persistente Homologie für den Iris Flower-Datensatz

Im Diagramm fällt die Lebensspanne von zwei Homologiegruppen in der Dimension Null auf. Ansonsten existieren in der Dimension Eins zwei Klassen, die relativ lang am Leben bleiben. Auch wiederholte Experimente mit anderen Parametern und anderen Komplexen liefern sehr ähnliche Resultate. Die Persistente Homologie deutet damit auf folgende Homologie für S hin:

$$H_n(K) = \begin{cases} \mathbb{Z} \oplus \mathbb{Z}, & n = 0, \\ \mathbb{Z} \oplus \mathbb{Z}, & n = 1, \\ 0, & \text{sonst.} \end{cases}$$

Zu beachten ist, dass $\epsilon \approx 0.45$ beträgt, was im Vergleich mit der Breitspanne des Datensatzes klein ist. Dementsprechend sind die eindimensionalen Löcher des Datensatzes S minimal. Dabei beinhaltet der Datensatz noch wenige Beobachtungen. Es ist möglich, dass die beide Klassen in $H_1(S)$ nur ein Rauschen darstellen und wenn n größer würde, sie auch verschwinden würden. Um das zu überprüfen, müssten mehr Beobachtungen gesammelt worden sein. Die zwei Klassen in der nullten Homologie (interpretiert als zwei Cluster) scheinen auf jeden Fall plausibel zu sein.

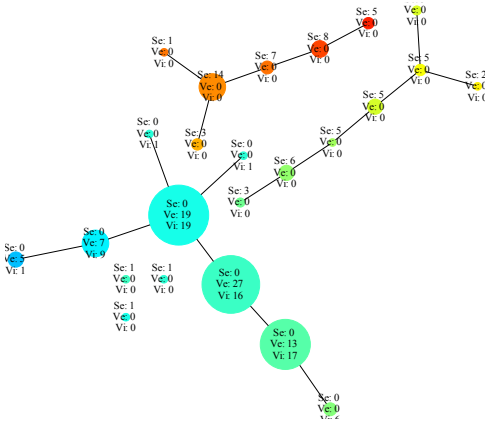
Für die Analyse mit dem Mapper-Algorithmus muss zuerst ein Filter gewählt werden. Die Persistente Homologie suggeriert mindestens zwei Cluster im Datensatz S; berücksichtigt jedoch nicht die Dichte der Punkte. Es ist durchaus vorstellbar, dass zum Teil noch stark verdichtete Regionen existieren, obwohl alle Punkte aus Wegzusammenhangskomponenten stammen. Das wäre ein Zeichen für weitere Gruppen. Der Filter $E_1(x)$ kann sich bei der Visualisierung von Daten als nützlich erweisen (siehe [10]) und erfasst den „Verlauf“ des Datensatzes vom „Zentrum“ nach Außen; ignoriert aber die Dichte der Punkte. Um die Informationen von Dichte und Exzentrizität zu kombinieren, wird ein gemischte Filter $f(x) = E_1(x) \cdot dens_k(x)$ verwendet. Als Parameter wird normalweise $k \approx \sqrt{n}$ empfohlen [13]. Als Label wird die Anzahl der verschiedenen Gattungen ausgegeben.

Um einen Überblick über die Daten zu erhalten, lohnt sich der Verlauf des Mapper beim wachsenden Parameter n zu betrachten. Es bilden sich mehr Abzweigungen und dadurch werden mehr Strukturen erkannt.

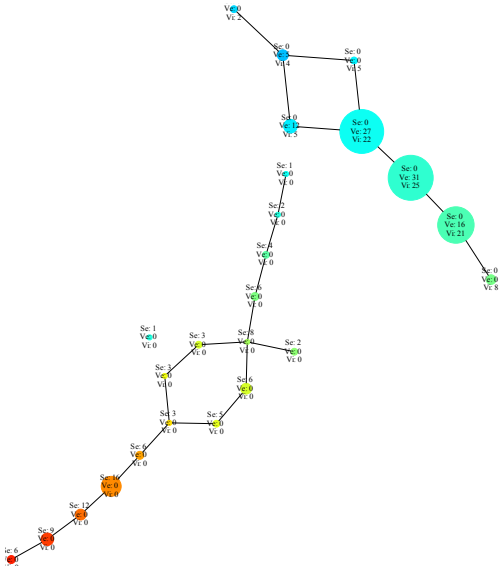
Der Mapper-Algorithmus wurde mit verschiedenen Einstellungen angewandt, unterteilt den Datensatz fast immer in zwei Hauptgruppen und trennt die Gattung Setosa von Virginica und Versicolor. Aus Abbildung 10(a) wird deutlich, dass in der Setosa-Gattung eine starke Filter-Wertänderung auftritt, was auf die wechselnde Dichte der Gruppe zurückzuführen ist. Setosa wird dadurch in zwei Cluster geteilt, was auf die gewählten Parameter zurückgeführt werden kann. Die andere Hauptgruppe beinhaltet nur die Gattungen Virginica und Versicolor. In den größten Knoten sind immer noch beide Sorten vorhanden. Das zeigt, dass sehr viele Ausprägungen von Virginica und Versicolor identische Messungen haben. Daraus kann der Schluss gezogen werden, dass beide Gattungen mittels topologischer Verfahren nur schwer voneinander unterschieden werden können.

Mit Erhöhung des Überlappungsparameters p verbleiben nur noch zwei Hauptgruppen (siehe Abbildung 10(b), (c), (d)). Da die Reichweite des Filters für Sepota sehr breit ist und die äußeren Knoten kaum Messungen beinhalten, deutet dieser Befund auf Ausreißer im Datensatz hin. Der Datensatz hat relativ wenig Beobachtungen. Deswegen kann der Parameterraum bei einem kleinen p nicht mit zu vielen Intervallen überdeckt werden. Sonst entstehen, wie in Abbildung 10(a), immer mehr getrennte kleine Komponenten. Da diese Knoten relativ klein sind und farbig nicht zu den großen Clustern passen, verbergen sich dahinter vermutlich Ausreißer. In Abbildung 11 wurde eine Überlappung von 95% gewählt. Dadurch kann der Parameterraum auf kleinere Bereiche unterteilt und mehr Strukturen erkannt werden. Das Verhalten vom Mapper bleibt aber ähnlich. Interessant ist die Beobachtung, dass die Gruppe Setosa am Ende wieder in zwei Teile gespalten wird (siehe 10(c)). Das deckt sich mit der Beobachtung aus Abbildung 10(a).

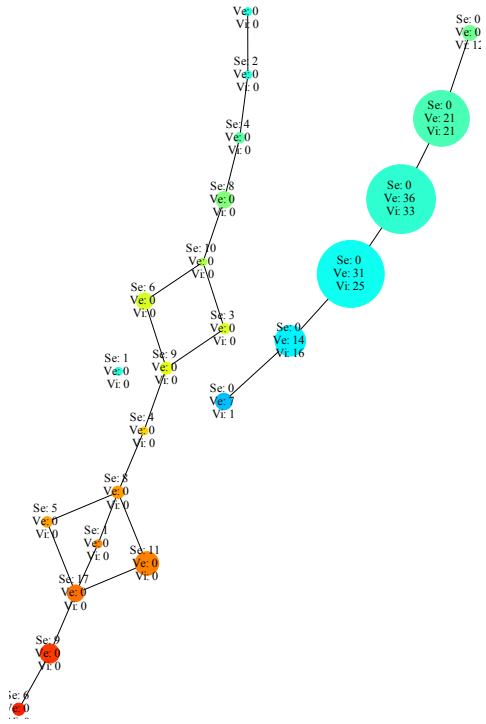
Die Gruppierung mittels der TDA deutet auf das Vorhandensein von lediglich zwei Clustern im *Iris Flower*-Datensatz hin, obwohl in der Realität drei Gattungen existieren. Weder die Persistente Homologie noch der Mapper-Algorithmus deuten auf eine Trennung zwischen Virginica und Versicolor an. Die Merkmalsausprägungen sind zu ähnlich. Die Länge und Breite des Kelch- und des Kronblattes reichen nicht aus, um diese Gattungen voneinander zu unterscheiden. Mehr Beobachtungen wären dafür notwendig. Damit liefern die Verfahren der Topologischen Datenanalyse identische Resultate wie die unterschiedlichen Verfahren der Clusteranalyse.



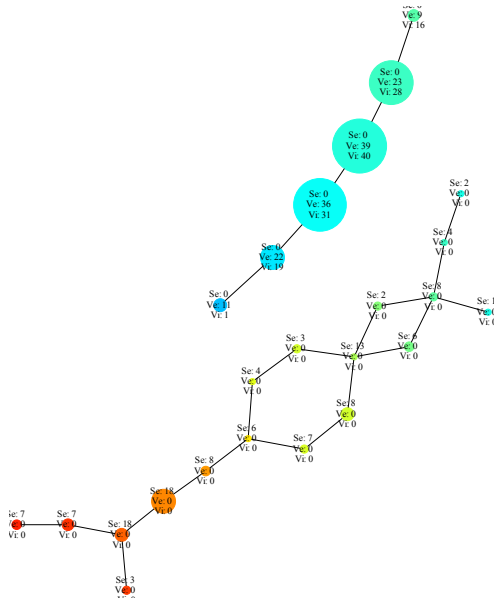
(a) $Mapper(f^{-1}(\mathcal{V}(15, 0.25), \mathcal{C}))$



(b) $Mapper(f^{-1}(\mathcal{V}(15, 0.45), \mathcal{C}))$

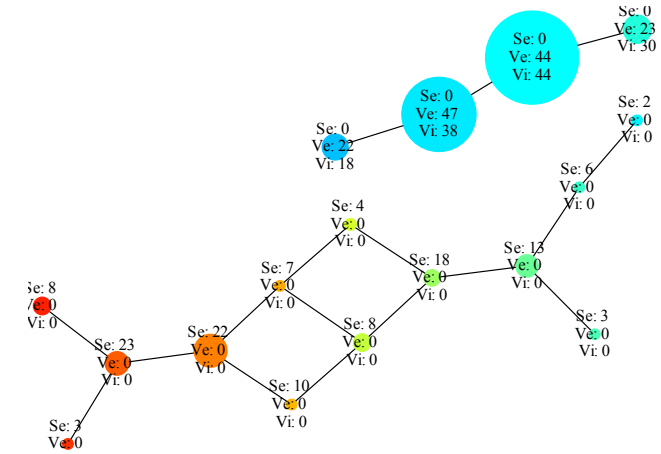


(c) $Mapper(f^{-1}(\mathcal{V}(15, 0.65), \mathcal{C}))$

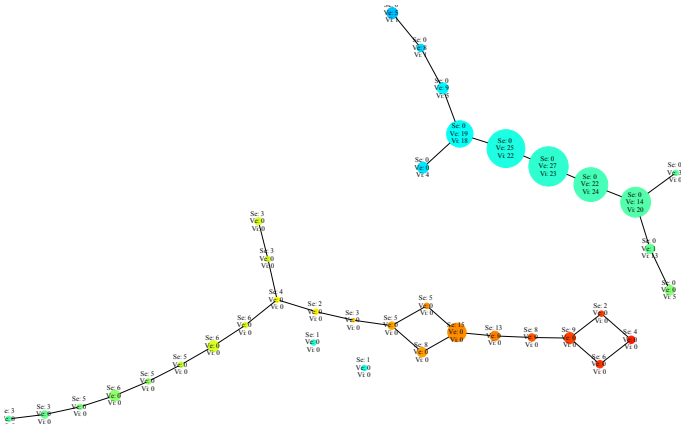


(d) $Mapper(f^{-1}(\mathcal{V}(15, 0.95), \mathcal{C}))$

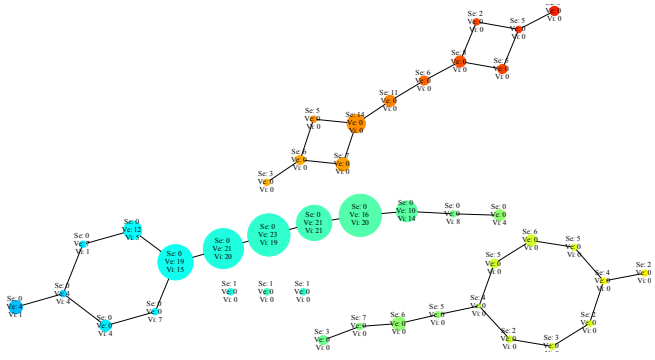
Abbildung 10: Mapper-Algorithmus für den Iris Flower-Datensatz



(a) $Mapper(f^{-1}(\mathcal{V}(10, 0.95), \mathcal{C}))$



(b) $Mapper(f^{-1}(\mathcal{V}(25, 0.95), \mathcal{C}))$



(c) $Mapper(f^{-1}(\mathcal{V}(30, 0.95), \mathcal{C}))$

Abbildung 11: Mapper-Algorithmus für den Iris Flower-Datensatz

3.2 Heart Disease-Datensatz

Als zweites Beispiel für den Vergleich der Gruppierung von Objekten wird der *Heart Disease*-Datensatz herangezogen. Dieser beinhaltet Beobachtungen von 270 Patienten. Anhand der folgenden fünf Merkmalen soll erkannt werden, ob eine Herzkrankheit eines Patienten vorliegt:

- Ruheblutdruck
- Cholesterinspiegel
- Maximale Herzrate
- Auftreten einer *Angina pectoris* (Schmerz in der Brust) bei Übungen
- ST-Senkung (Auffälligkeit im EKG) - verursacht durch Übungen.

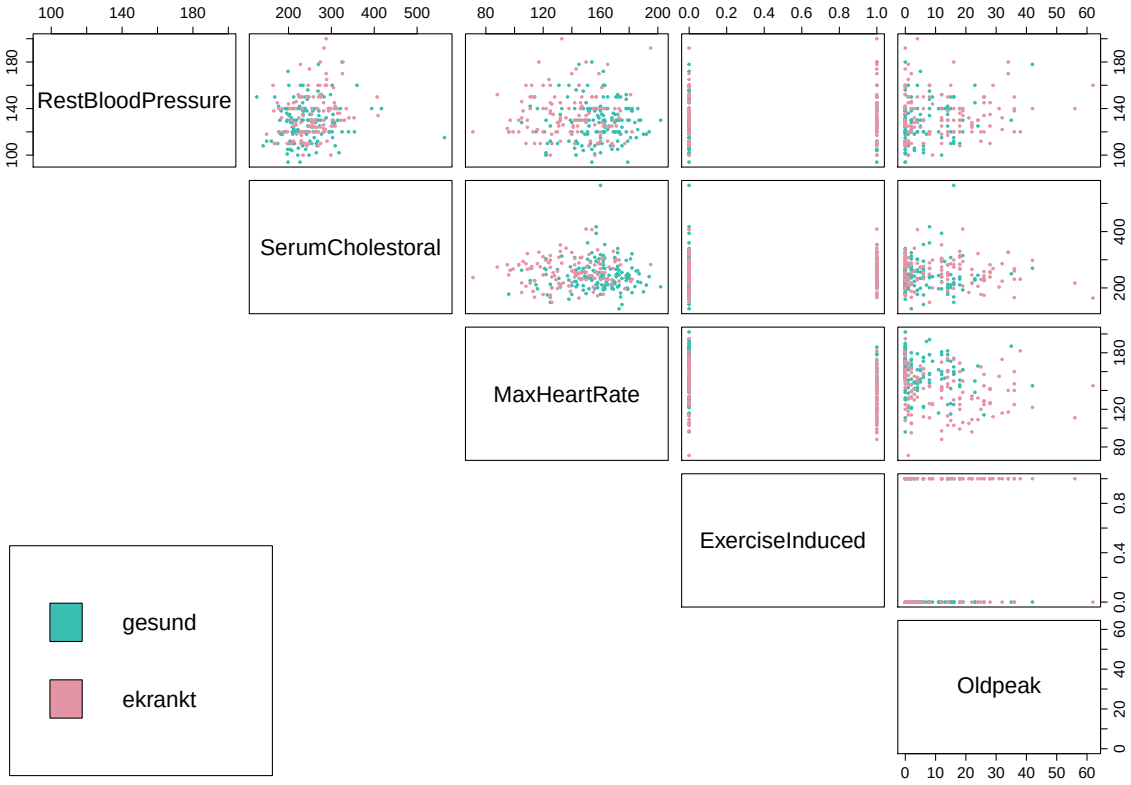
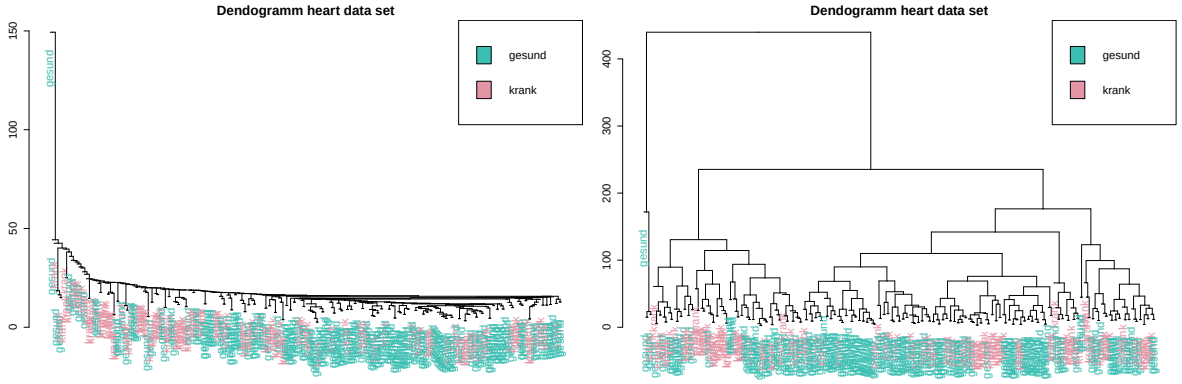


Abbildung 12: Heart Disease-Datensatz

Aus den paarweisen Streudiagrammen in Abbildung 12 lässt sich aufgrund der Vermischung der Punktwolken die Vermutung ableiten, dass beide Gruppen von Patienten (Herzkranke vs. Nicht-Herzkranke) nur schwer voneinander zu trennen sind. Als Nächstes wird der Gruppierung der Clusteranalyse mit den Verfahren der TDA verglichen.

Clusteranalytische Verfahren

Die verschiedenen Verfahren der Clusteranalyse schneiden beim *Heart Disease*-Datensatz relativ schlecht ab. Das K-Means-Verfahren erreicht lediglich eine Trefferquote von ca. 41%. Auch das dichtebasierte Verfahren DBSCAN und alle drei hierarchischen Verfahren liefern wenig zufriedenstellende Gruppierung der Objekte. DBSCAN findet sogar nur einen Cluster. Die Dendrogramme der Single-Linkage und Complete-Linkage sind in Abbildung 13 zu finden. Es wird deutlich, dass beide hierarchische Verfahren erkrankte und gesunde Patienten nur schlecht voneinander unterscheiden können. Die statistischen Verfahren der Clusteranalyse sind dementsprechend wenig geeignet, um im Hinblick auf das Herzinfarktrisiko eine trennscharfe Gruppierung zu bilden.



(a) Single-Linkage (b) Complete-Linkage

Abbildung 13: Dendrogramme der hierarchischen Verfahren

Topologische Verfahren

Im Folgenden wird die Eignung der beiden topologischen Verfahren geprüft. Für die Berechnung der Persistenten Homologie werden Witness-Komplexe mit $|L| = 50$, $\epsilon = \max(w, l), w \in W,\ l \in L$ und die Filtration $W_{\frac{1}{50}\epsilon} \subseteq \ldots \subseteq W_\epsilon$ genutzt. Aus Abbildung 14 lässt die folgende Homologie vermuten:

$$H_n(K) = \begin{cases} \mathbb{Z}, & n = 0, \\ \mathbb{Z} \oplus \mathbb{Z}, & n = 1, \\ \mathbb{Z}, & n = 2, \\ 0, & \text{sonst.} \end{cases}$$

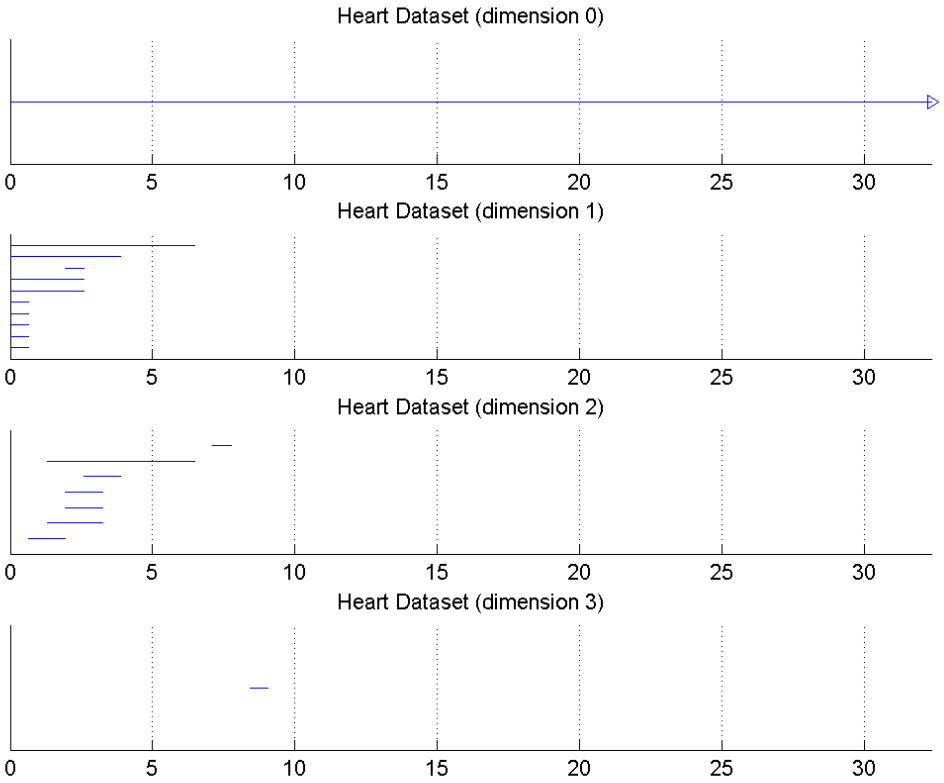
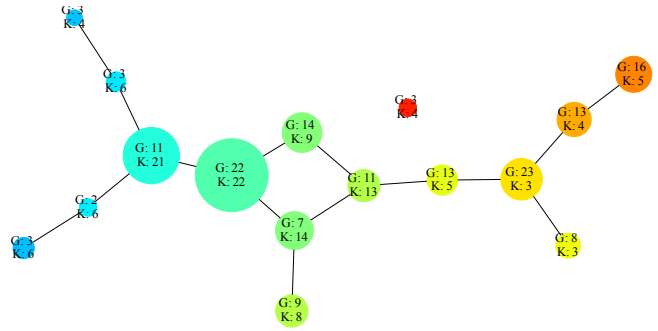


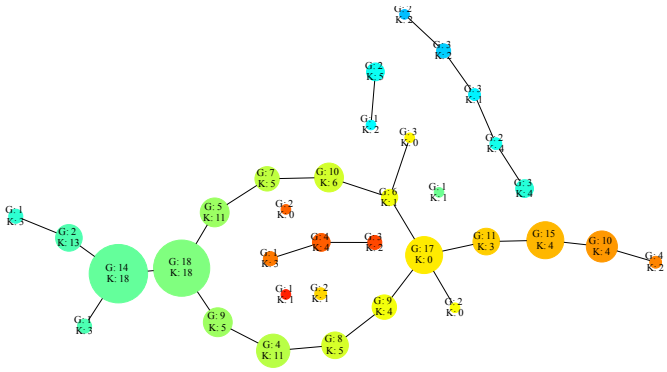
Abbildung 14: Persistente Homologie für den Heart Disease-Datensatz

Im *Heart Disease*-Datensatz befinden sich erneut nur relativ wenige Beobachtungen. Die Problematik stellt sich ähnlich wie beim *Iris Flower*-Datensatz dar, d. h. die nullte Homologie ist $H_0(S) = 0$.

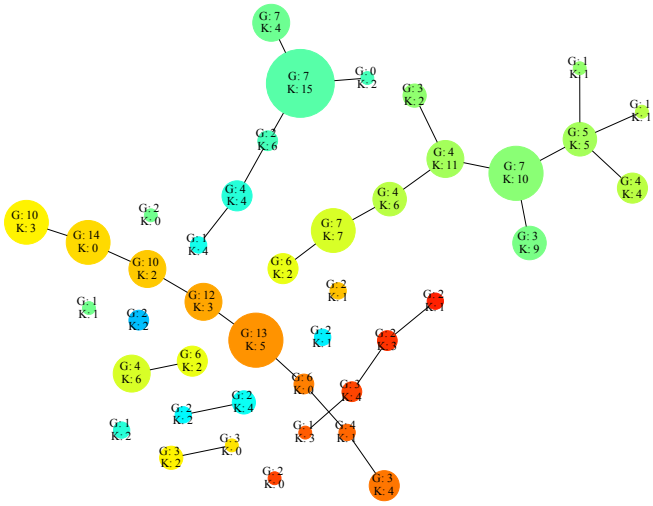
Für den Mapper-Algorithmus wird der Filter $f(x) = \frac{dens_k(x)}{E_1(x)}$ und $k \approx \sqrt{n}$ benutzt. Als Beschriftung wird die Anzahl gesunder und erkrankter Patienten angezeigt. Abbildung 15 zeigt den Mapper-Output.



(a) $Mapper(f^{-1}(\mathcal{V}(15, 0.15), \mathcal{C}))$



(b) $Mapper(f^{-1}(\mathcal{V}(20, 0.15), \mathcal{C}))$



(c) $Mapper(f^{-1}(\mathcal{V}(25, 0.15), \mathcal{C}))$

Abbildung 15: Mapper-Algorithmus für den Heart Disease-Datensatz

In (a) lässt sich nur erkennen, dass sich am rechten Ende sehr viele gesunde Patienten befinden. Mit wachsendem Parameter n etablieren sich neue Strukturen. In (b) sind zum Beispiel im großen Skeleton zwei Häufungen von Punkten auf gegenüberliegenden Enden zu sehen. Wie (c) zeigt, wird der Datensatz in drei große Gruppen geteilt. In der ersten Gruppe befinden sich hauptsächlich die gesunden Patienten; in der zweiten Gruppe die erkrankten Patienten und in der dritten Gruppe ist die Anzahl von gesunden und erkrankten Patienten ausgeglichen. Der Mapper-Algorithmus unterteilt die Beobachtungen in drei Gruppen: wahrscheinlich erkrankte Patienten, wahrscheinlich gesunde Patienten und Patienten, über die keine eindeutige Aussage getätigt werden kann. Das ist das beste Resultat, dass bis jetzt geliefert wurde. Insgesamt weist auch der Mapper-Algorithmus große Probleme bei der korrekten Gruppierung der Beobachtungen im *Heart Disease*-Datensatz auf.

3.3 Kritische Diskussion

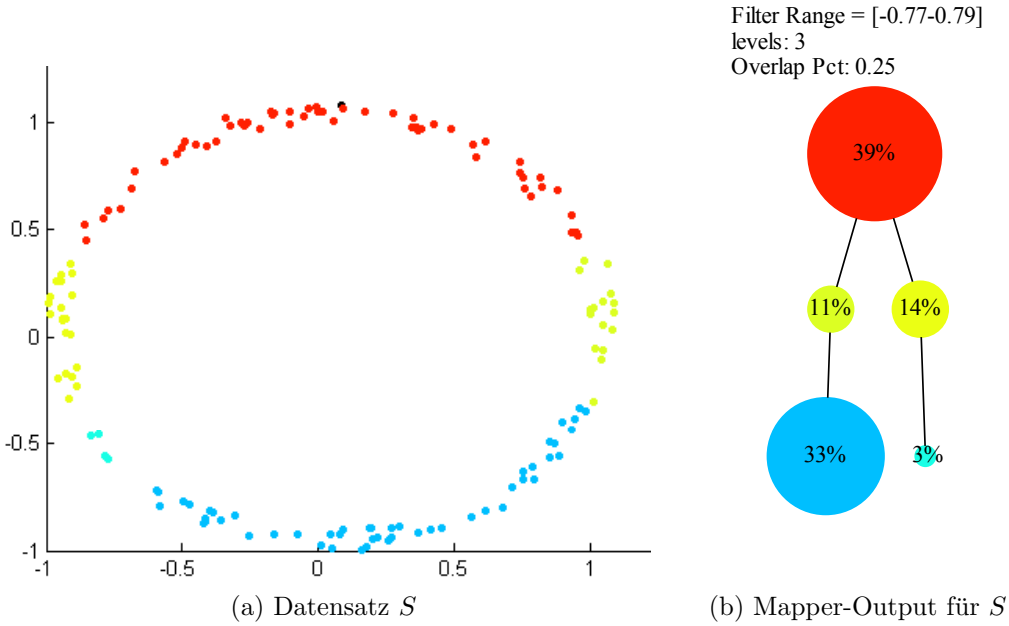
Aus der vorangegangenen Analyse bleibt festzuhalten, dass der Mapper-Algorithmus großes Potenzial bei der Gruppierung von Objekten besitzt. Bezogen auf den *Heart Disease*-Datensatz hat dieser Ansatz das beste Ergebnis geliefert. Komplett befriedigend waren diese aber nicht. Um eine korrekte Clusterung zu erzielen, benötigt die Mapper-Software als Input die nachfolgenden möglichst adäquat gewählten Parameter:

- Filter f,
- Anzahl Klassen für das Histogramm k,
- Anzahl Intervalle für die Überdeckung n,
- sowie deren prozentuale Überlappung p.

An dieser Stelle werden einige Probleme und Erweiterungen des Mapper-Algorithmus angesprochen. Für den Output $Mapper(f^{-1}(\mathcal{V}), C)$ ist ein Clusteralgorithmus C notwendig. Für die Berechnungen wurde das Single-Linkage-Verfahren genutzt. Hierarchischen Verfahren liefern jedoch mehrere Möglichkeiten einen Datensatz zu gruppieren. Um die Wahl zu automatisieren, wird wie folgt vorgegangen:

1. Es wird ein gewichteter Graph aufgebaut - mit der Distanz zwischen den Punkten als Kantengewicht.
2. Der minimale Spannbaum des Graphen wird gefunden (dieser entspricht einem Dendogramm).
3. Es wird ein Histogramm der Gewichte mit k Klassen konstruiert.
4. Es wird die Schwelle $t \in \mathbb{R}$ gesucht. Als t wird die Mitte der ersten leeren Klasse, die links vom Balken mit den längsten Kanten liegt, gespeichert. Sollte keine leere Klasse existieren, wird als Schwelle t die längste Kante herangezogen.
5. Der Graph wird anhand der Schwellen t partitioniert und die einzelne Komponenten werden als Cluster interpretiert.

Diese Vorgehensweise hängt von der Anzahl der Klassen k ab und die Mapper-Software wurde so geschrieben, dass der Parameter als Input übergeben werden muss. Die empirische Arbeit mit der Software zeigte, dass k relativ robust gegenüber Änderungen ist. Jedoch wurde in Abbildung 16 eine Reproduktion der Abbildung 16 aus Nastansky (2019) vorgenommen, wobei der Parameter k bewusst zu groß gewählt wurde. Das Problem ist schnell zu identifizieren. Punkte, die eigentlich zusammengehören, werden auf zu viele Cluster geteilt und dadurch fehlen die Überlappungen in diesem Cluster. Überlappungen existieren nur am Rande eines Intervalls. Dadurch kann der Mapper-Algorithmus sie nicht mehr verbinden. Eine denkbare Lösung für dieses Problem wäre, den Algorithmus rekursiv bei jeder Clusterung abzurufen. Der Nachteil ist, dass der Prozess rechenintensiver wird. Außerdem müssten beim wiederholten Abruf des Programms alle Parameter automatisch neu und, soweit es möglich ist, korrekt gewählt werden. Der Mapper reagiert mit einer fehlerhaften Gruppierung auf eine falsch gewählte Anzahl der Klassen k im Histogramm.



(a) Datensatz S (b) Mapper-Output für S

Abbildung 16: Reproduktion der Abbildung 16 (vgl. Nastansky (2019), S. 27) mit der Mapper-Software. Als Filter dient die Funktion $f(x, y) = y$. Der Homotopietyp von $\mathbb{X}$ wird hier nicht erkannt.

Darüber hinaus sind andere Erweiterungen denkbar. Bislang wurde nur das Single-Linkage-Verfahren zur Clusterung angewandt. Prinzipiell kann ein beliebiger Clusteralgorithmus verwendet werden.

4 Fazit

Zusammengefasst haben sich die Verfahren der Topologischen Datenanalyse als ein praktikables Werkzeug bei der Gruppierung von Daten erwiesen. Vor allem mit dem Mapper-Algorithmus konnten adäquate Cluster erkannt werden. Beim *Iris Flower*-Datensatz hat die TDA ähnliche Ergebnisse wie die Clusteranalyse erzielt. Der *Heart Disease*-Datensatz war schwieriger zu behandeln. Die genutzten clusteranalytischen Verfahren waren nicht geeignet, die beiden Gruppen von Patienten korrekt zu identifizieren. Mit Mapper wurden drei Klassen relativ gut voneinander getrennt. Im Vergleich zu den Standardverfahren der Clusteranalyse zeigte sich eine leichte Überlegenheit der topologischen Verfahren.

Als Ausblick könnte ein Verfahren für Maschinelles Lernen für den Mapper-Algorithmus entwickelt werden. Mit Hilfe von Trainingsdatensätzen könnte die Gruppierung von Objekten deutlich verbessert werden.

Literatur und Quellen

[1] H. Adams, A. Tausz, M. Vejdemo-Johansson, *JavaPlex: A research software package for persistent (co)homology.*
`http://appliedtopology.github.io/javaplex/`

[2] E. Anderson, *The species problem in iris.* Annals of the Missouri Botanical Garden 23 (3): 457–509, 1936.

[3] J. Bacher, A. Pöge, K. Wenzig, *Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren.* 3. Auflage, Oldenbourg Wissenschaftsverlag, München-Wien-Oldenbourg, 2010.

[4] K. Backhaus, B. Erichson, W. Plinke, R. Weber, *Multivariate Analysemethoden.* 14. Auflage, SpringerGabler, Berlin-Heidelberg, 2016.

[5] G. Carlsson *Topology and data.* Bulletin of the American Mathematical Society (New Series) vol. 46, no. 2, s. 255–308, 2009.
`http://www.ams.org/journals/bull/2009-46-02/S0273-0979-09-01249-X/S0273-0979-09-01249-X.pdf`

[6] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996.

[7] R. A. Fisher, *The use of multiple measurements in taxonomic problems.* Annals of Eugenics 7 (2): 179–188, 1936.

[8] K. Jänich, *Topologie.* 8. Auflage, Springer-Verlag, Berlin-Heidelberg, 2005.

[9] P.Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, G. Carlsson *Extracting insights from the shape of complex data using topology.* Scientific Reports 3 No. 1236, 2013.

[10] A. Nastansky, *Topologische Datenanalyse: Eine Einführung in die Persistente Homologie und Mapper.* Statistische Diskussionsbeiträge (Nr. 53), 2019.

[11] R Development Core Team, *R: A Language and Environment for Statistical Computing.*
`https://www.r-project.org/`

[12] R. Schlittgen, *Multivariate Statistik.* Oldenbourg Wissenschaftsverlag, München, 2009.

[13] B. W. Silvermann, *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[14] G. Singh, Y. Yao, *Bio Mapper v1.*
`https://simtk.org/project/xml/downloads.xml?group_id=362`

[15] M. Steinbach, P.-N. Tan, V. Kumar, *Introduction to Data Mining*. Pearson Higher Ed USA, 2005.

[16] Y. Yao, G. R. Bowman, G. Carlsson, L. J. Guibas, X. Huang, M. Lesnick, V. S. Pande, G. Singh und J. Sun *Topological methods for exploring low-density states in biomolecular folding pathways*. The Journal of Chemical Physics 130, 2009.

UNIVERSITÄT POTSDAM

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 1	1995	Strohe, Hans Gerhard: Dynamic Latent Variables Path Models - An Alternative PLS Estimation -
Nr. 2	1996	Kempe, Wolfram. Das Arbeitsangebot verheirateter Frauen in den neuen und alten Bundesländern - Eine semiparametrische Regressionsanalyse
Nr. 3	1996	Strohe, Hans Gerhard: Statistik im DDR-Wirtschaftsstudium zwischen Ideologie und Wissenschaft
Nr. 4	1996	Berger, Ursula: Die Landwirtschaft in den drei neuen EU-Mitgliedsstaaten Finnland, Schweden und Österreich - Ein statistischer Überblick
Nr. 5	1996	Betzin, Jörg: Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten
Nr. 6	1996	Berger, Ursula: Die Methoden der EU zur Messung der Einkommenssituation in der Landwirtschaft - Am Beispiel der Bundesrepublik Deutschland
Nr. 7	1997	Strohe, Hans Gerhard / Geppert, Frank: Algorithmus und Computerprogramm für dynamische Partial Least Squares Modelle
Nr. 8	1997	Rambert, Laurence / Strohe, Hans Gerhard: Statistische Darstellung transformationsbedingter Veränderungen der Wirtschafts- und Beschäftigungsstruktur in Ostdeutschland
Nr. 9	1997	Faber, Cathleen: Die Statistik der Verbraucherpreise in Rußland - Am Beispiel der Erhebung für die Stadt St. Petersburg
Nr. 10	1998	Nosova, Olga: The Atractiveness of Foreign Direct Investment in Russia and Ukraine - A Statistical Analysis
Nr. 11	1999	Gelaschwili, Simon: Anwendung der Spieltheorie bei der Prognose von Marktprozessen
Nr. 12	1999	Strohe, Hans Gerhard / Faber, Cathleen: Statistik der Transformation - Transformation der Statistik. Preisstatistik in Ostdeutschland und Rußland
Nr. 13	1999	Müller, Claus: Kleine und mittelgroße Unternehmen in einer hoch konzentrierten Branche am Beispiel der Elektrotechnik. Eine statistische Langzeitanalyse der Gewerbezählungen seit 1882
Nr. 14	1999	Faber, Cathleen: The Measurement and Development of Georgian Consumer Prices
Nr. 15	1999	Geppert, Frank / Hübner, Roland: Korrelation oder Kointegration – Eignung für Portfoliostrategien am Beispiel verbriefter Immobilenanlagen
Nr. 16	2000	Achsani, Noer Azam / Strohe, Hans Gerhard: Statistischer Überblick über die indonesische Wirtschaft
Nr. 17	2000	Bartels, Knut: Testen der Spezifikation von multinominalen Logit-Modellen
Nr. 18	2002	Achsani, Noer Azam / Strohe, Hans Gerhard: Dynamische Zusammenhänge zwischen den Kapitalmärkten der Region Pazifisches Becken vor und nach der Asiatischen Krise 1997
Nr. 19	2002	Nosova, Olga: Modellierung der ausländischen Investitionstätigkeit in der Ukraine
Nr. 20	2003	Gelaschwili, Simon / Kurtanidse, Zurab: Statistische Analyse des Handels zwischen Georgien und Deutschland
Nr. 21	2004	Nastansky, Andreas: Kurz- und langfristiger statistischer Zusammenhang zwischen Geldmengen- und Preisentwicklung: Analyse einer kointegrierenden Beziehung
Nr. 22	2006	Kauffmann, Albrecht / Nastansky, Andreas: Ein kubischer Spline zur temporalen Disaggregation von Stromgrößen und seine Anwendbarkeit auf Immobilienindizes
Nr. 23	2006	Mangelsdorf, Stefan: Empirische Analyse der Investitions- und Exportentwicklung des Verarbeitenden Gewerbes in Berlin und Brandenburg
Nr. 24	2006	Reilich, Julia: Return to Schooling in Germany
Nr. 25	2006	Nosova, Olga / Bartels, Knut: Statistical Analysis of the Corporate Governance System in the Ukraine: Problems and Development Perspectives
Nr. 26	2007	Gelaschwili, Simon: Einführung in die statistische Modellierung und Prognose
Nr. 27	2007	Nastansky, Andreas: Modellierung und Schätzung von Vermögenseffekten im Konsum
Nr. 28	2008	Nastansky, Andreas: Schätzung vermögenspreisinduzierter Investitionseffekte in Deutschland

UNIVERSITÄT POTSDAM

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 29	2008	Ruge, Marcus / Strohe, Hans Gerhard: Analyse von Erwartungen in der Volkswirtschaft mit Partial-Least-Squares-Modellen
Nr. 30	2009	Newiak, Monique: Prüfungsurteile mit Dollar Unit Sampling – Ein Vergleich von Fehlerschätzmethoden für Zwecke der Wirtschaftsprüfung: Praxis, Theorie, Simulation –
Nr. 31	2009	Ruge, Marcus: Modellierung von Stimmungen und Erwartungen in der deutschen Wirtschaft
Nr. 32	2009	Nosova, Olga: Statistical Analysis of Regional Integration Effects
Nr. 33	2009	Mangelsdorf, Stefan: Persistenz im Exportverhalten – Kann punktuelle Exportförderung langfristige Auswirkungen haben? -
Nr. 34	2009	Kbiladze, David: Einige historische und gesetzgeberische Faktoren der Reformierung der georgischen Statistik
Nr. 35	2009	Nastansky, Andreas / Strohe, Hans Gerhard: Die Ursachen der Finanz- und Bankenkrise im Lichte der Statistik
Nr. 36	2009	Gelaschwili, Simon / Nastansky, Andreas: Development of the Banking Sector in Georgia
Nr. 37	2010	Kunze, Karl-Kuno / Strohe, Hans Gerhard: Time Varying Persistence in the German Stock Market
Nr. 38	2010	Nastansky, Andreas / Strohe, Hans Gerhard: The Impact of Changes in Asset Prices on Real Economic Activity: A Cointegration Analysis for Germany
Nr. 39	2010	Kunze, Karl-Kuno / Strohe, Hans Gerhard: Antipersistence in German Stock Returns
Nr. 40	2010	Dietrich, Irina / Strohe, Hans Gerhard: Die Vielfalt öffentlicher Unternehmen aus der Sicht der Statistik - Ein Versuch, das Unstrukturierte zu strukturieren
Nr. 41	2010	Nastansky, Andreas / Lanz, Ramona: Bonuszahlungen in der Kreditwirtschaft: Analyse, Regulierung und Entwicklungstendenzen
Nr. 42	2010	Dietrich, Irina / Strohe, Hans Gerhard: Die Vermögenslage öffentlicher Unternehmen in Deutschland - Statistische Analyse anhand von amtlichen Mikrodaten der Jahresabschlüsse.
Nr. 43	2010	Ulbrich, Hannes-Friedrich: Höherdimensionale Kompositionsdaten – Gedanken zur grafischen Darstellung und Analyse -
Nr. 44	2011	Dietrich, Irina / Strohe, Hans Gerhard: Statistik der öffentlichen Unternehmen in Deutschland – Die Datenbasis
Nr. 45	2011	Nastansky, Andreas: Orthogonale und verallgemeinerte Impuls-Antwort-Funktionen in Vektor-Fehlerkorrekturmodellen
Nr. 46	2011	Dietrich, Irina / Strohe, Hans Gerhard: Die Finanzlage öffentlicher Unternehmen in Deutschland - Statistische Analyse amtlicher Mikrodaten der Jahresabschlüsse -
Nr. 47	2011	Teitge, Jonas / Nastansky, Andreas: Interdependenzen in den Renditen DAX-notierter Unternehmen nach Branchen
Nr. 48	2011	Dietrich, Irina: Die Ertragslage öffentlicher Unternehmen in Deutschland - Statistische Analyse amtlicher Mikrodaten der Jahresabschlüsse -
Nr. 49	2011	Kauper, Benjamin / Kunze, Karl-Kuno: Modellierung von Aktienkursen im Lichte der Komplexitätsforschung
Nr. 50	2011	Nastansky, Andreas / Strohe, Hans Gerhard: Konsumausgaben und Aktienmarktentwicklung in Deutschland: Ein kointegriertes vektorautoregressives Modell
Nr. 51	2014	Nastansky, Andreas / Mehnert, Alexander / Strohe, Hans Gerhard: A Vector Error Correction Model for the Relationship between Public Debt and Inflation in Germany
Nr. 52	2019	Kauffmann, Albrecht / Nastansky, Andreas: Explorative Analyse der Preise von Einfamilienhäusern und Eigentumswohnungen in Deutschland
Nr. 53	2019	Nastansky, Andreas: Topologische Datenanalyse: Eine Einführung in die Persistente Homologie und Mapper
Nr. 54	2022	Kauffmann, Albrecht / Nastansky, Andreas: Regionale Mieten in Deutschland: Explorative Analyse der Mieten in der Wiedervermietung
Nr. 55	2022	Nastansky, Andreas: Gruppierung von Daten: Topologische Verfahren vs. Clusteranalyse