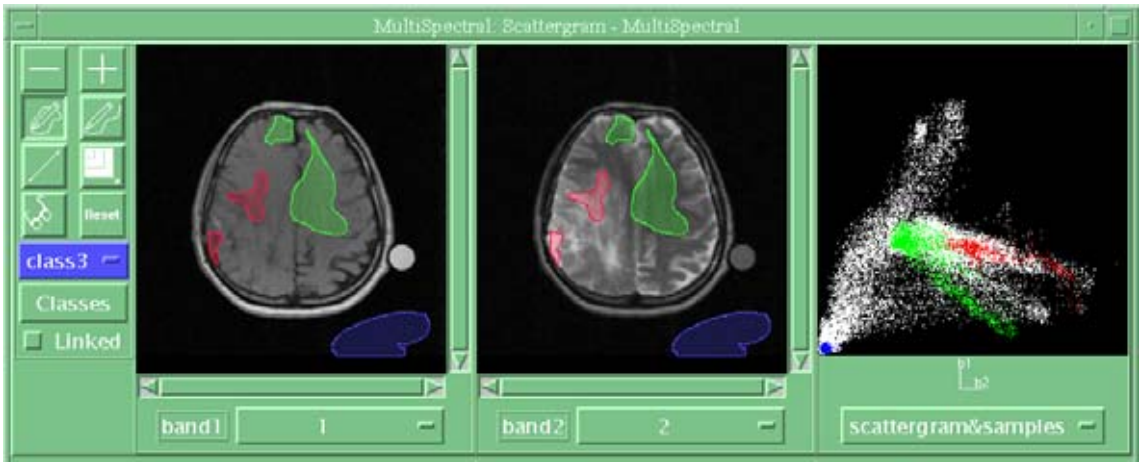

Abbildung 5.2: Multidimensionale Klassifizierung, Markierung der Stichproben in den Komponentenbildern und dem Scattergramm

im eindimensionalen Fall. Haben zwei Komponenten in der Stichprobe keinen Zusammenhang – sie sind nicht korreliert – so ist der entsprechenden Koeffizient in der Covarianzmatrix gleich 0. Mit steigender Ähnlichkeit steigt ebenfalls der Covarianzkoeffizient.

Für eine Segmentierung müssen nun die den Gewebsklassen zugeordneten multivariaten Verteilungen gefunden werden. Dazu müssen sorgfältig für jedes Gewebe Stichproben gesammelt werden. Dies geschieht meist interaktiv und kann mit einfachen graphischen Tools durchgeführt werden. Für jeden Gewebetyp g werden die Parameter μ_g, Σ_g der Verteilungen bestimmt und anschließend wird folgender Algorithmus durchgeführt:

1. Berechne für alle Bildpunkte die Wahrscheinlichkeit der Zugehörigkeit zu einem der definierten Gewebstypen.
2. Ordne das Pixel jenem Gewebe zu, für welches es die höchste Wahrscheinlichkeit besitzt.
3. Überprüfe ob die Anzahl der veränderten Pixel ein bestimmtes Limit übersteigt, wenn nein, beende den Algorithmus, wenn ja fahre fort.
4. Berechne die Parameter $\vec{\mu}$ und Σ anhand der Pixel, die einem Gewebe zugeordnet wurden.
5. Beginne bei Schritt 1

Die Definition von Stichproben und die anschließende Segmentierung aufgrund der Stichproben ist in Abbildung 5.2 und 5.3 dargestellt.

5.4 Clusteranalyse

Eine weitere Möglichkeit Gruppen, d.h. in unserem Fall, Gewebstypen, zu klassifizieren ist die Clusteranalyse. Ein Cluster ist eine Gruppe von Datenpunkten

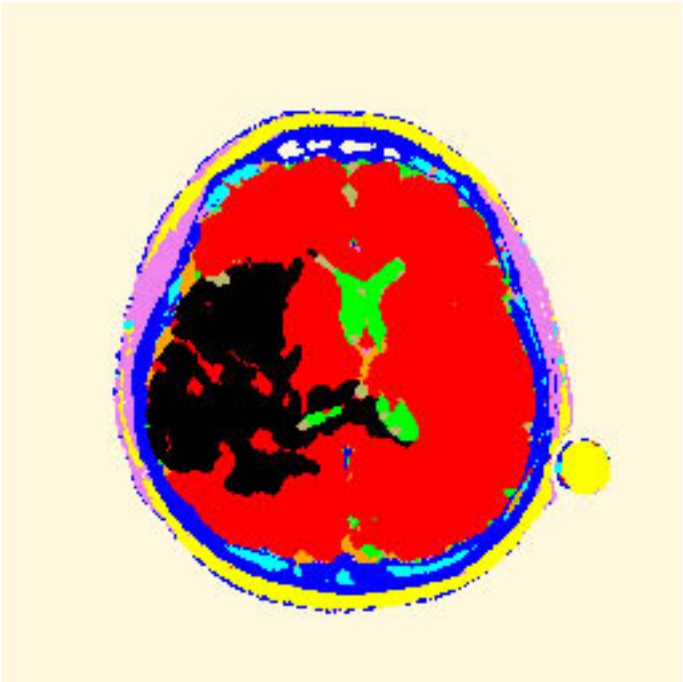

Abbildung 5.3: Klassifizierung von 9 Gewebstypen anhand multispektraler MRI Bilddaten

in einem Raum, die durch bestimmte Ähnlichkeiten ausgezeichnet sind. Diese Ähnlichkeiten können die örtliche Nähe in Bezug auf eine bestimmte Metrik sein oder etwa andere Merkmale, die sich Mithilfe eine Metrik formulieren lassen. Hier wäre wieder das Beispiel der Farben zu verwenden, wenn sich zwei Farben ähnlich sind, so liegen auch ihre Repräsentanten im RGB-Raum nahe beieinander. Alle Farbpunkte einer gleichfarbigen Fläche würden einen Punkthaufen im RGB-Raum bilden, dessen Ausdehnung durch die Variationen der Farbgebung bestimmt wird.

Grob lassen sich die Clustering-Methoden in zwei Kategorien einteilen, die *hierarchischen* Methoden und die *partitionierenden* Methoden. Die Wahl der Methode beruht grundsätzlich auf dem Typ der Daten. In jeder dieser Gruppe existieren eine Vielzahl von Algorithmen, von denen meist keiner *a priori* als geeignetster bezeichnet werden kann. Durch die Verschiedenartigkeit der Problemstellungen werden in der Praxis meist einige Algorithmen ausprobiert und nach der Beurteilung der Ergebnisse wird eine Entscheidung für einen Algorithmus getroffen. Diese Vorgehensweise ist in der *explorativen* Datenanalyse üblich, da mit diesen Methoden der Informationsgehalt in den Daten geprüft werden soll und nicht wie im Gegensatz zu den statistischen Methoden eine Hypothese durch Datenmaterial belegt werden soll. Bei den hierarchischen Algorithmen wird keine Einteilung der Daten vorgenommen, sondern es wird eine Klassifizierung der Daten in Baumstrukturen vorgenommen, wie es bei Stammbäumen üblich ist. Eine weitere Verbreitung in der Medizinischen Bildverarbeitung haben die partitionierenden Methoden. Dabei werden k Cluster konstruiert, d.h.

die Daten werden in k Gruppen unterteilt, die folgende Bedingungen erfüllen

- Jede Gruppe muß mindestens ein Element enthalten
- Jedes Objekt muß mindestens zu einer Gruppe gehören

Als repräsentativen Algorithmus wird der *k-means* Algorithmus detailliert beschrieben. In dieser Methode wird ein Cluster durch seinen Centroiden beschrieben, für jede Dimension der Objekte einer Datenmenge gilt

$$\bar{x}_m(v) = \frac{1}{N_v} \sum_{i \in C_v} x_{im} \quad , \tag{5.18}$$

wobei der Centroid der Schwerpunkt der Koordinaten in jeder Dimension ist, der aus allen Objekten, die im Cluster C_v enthalten sind, gebildet wird. Der Centroid muß somit kein Objekt der Datenmenge sein. Ein Maß für die Dichte des Clusters ist die Summe der Abstände aller Clusterobjekte vom Centroiden

$$D(C_v) = \sum_{i \in C_v} \sum_{m} (x_{im} - \bar{x}_m(v))^2 \quad . \tag{5.19}$$

Die gesamte Abstandssumme über den klassifizierten Datensatz ist

$$D = \sum_{v} D(C_v) \quad . \tag{5.20}$$

Ein Kriterium für die Klassifizierung ist die Einteilung der Daten in Cluster, sodaß der Gesamtabstand D minimal wird. Eine solche Methode wird auch als Minimierung der Varianz bezeichnet. Der k-means Algorithmus läßt sich durch folgende Schritte beschreiben:

1. Auswahl von k Centroiden und Aufteilung der Daten in k Cluster, wobei jedes Objekt dem nähesten Centroiden zugeordnet wird, oder Durchführung einer *a priori* Partitionierung der Daten in k Cluster.
2. Berechnung der Centroiden der Cluster.
3. Für jedes Objekt der Daten wird getestet, ob der Centroid eines andern Clusters w näher ist als der Centroid des Clusters v, zu dem das Objekt gehört. Wenn ja, wird die Clusterzugehörigkeit vertauscht und die Centroiden der beiden Cluster werden neu berechnet.
4. Hat sich der Gesamtabstand nach Durchlaufen aller Objekte im Datensatz um mehr als einen vordefinierten Betrag verringert, wird die Prozedur in Schritt 3 fortgesetzt, ansonsten ist die Clusterung abgeschlossen.

Dieser Algorithmus hat den Vorteil, daß keine leeren Cluster entstehen können, denn wird aus einem Cluster mit zwei Objekten ein Objekt entfernt, so wandert der Centroid in das Objekt, wodurch das Objekt nicht mehr aus dem Cluster entfernt werden kann (Abstand 0). Da der Wechsel von einem Cluster zum anderen nur durchgeführt wird, wenn das Objekt näher zum Centroiden des anderen Clusters liegt, wird in jedem Schritt der Gesamtabstand vermindert, wodurch die Konvergenz des Algorithmus gegeben ist. Weiters wird durch die Neuberechnung der Centroiden der Gesamtabstand abermals verringert, da der Centroid jener Punkt mit minimalem Abstand zu allen Objekten im Cluster ist.