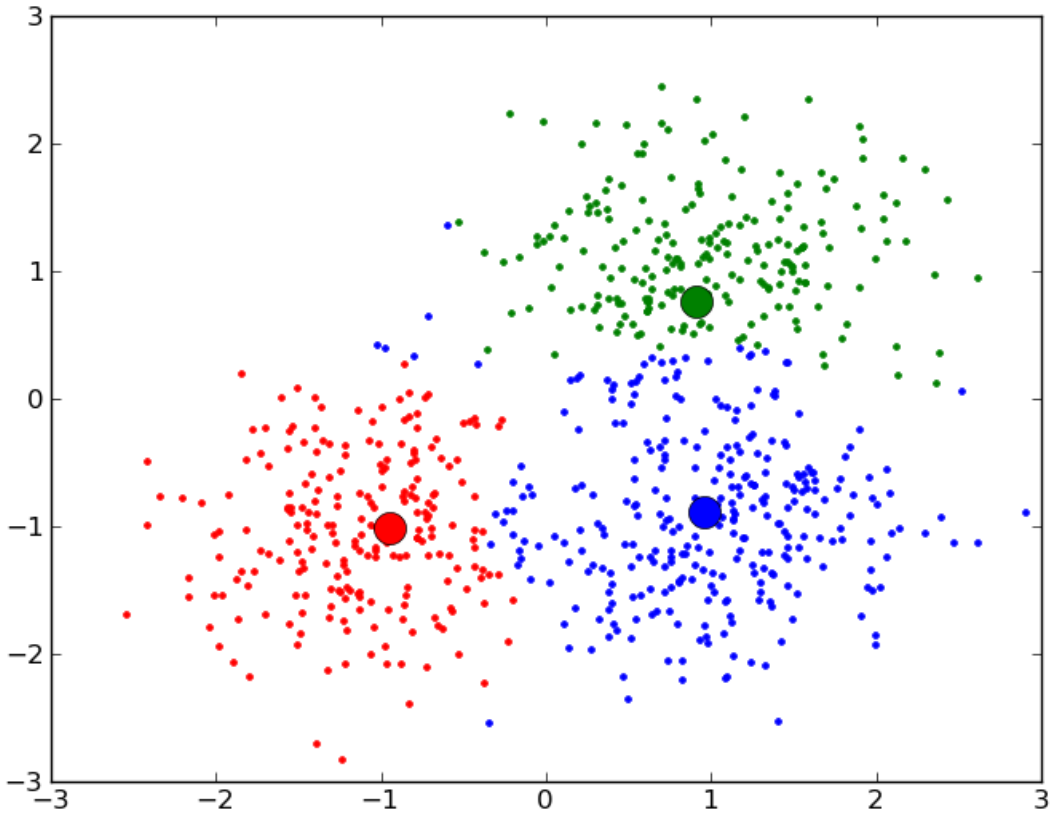
Handbuch: 7.2.4. Clustering

Quantitative Daten können oft in qualitative Gruppen (Cluster) unterteilt werden. Wenn wir beispielsweise eine Gasturbine betreiben, können wir sie im Leerlauf betreiben, mit Halblast oder mit Volllast. Diese Konzepte sind für einen menschlichen Ingenieur sinnvoll, der weiß, was darunter zu verstehen ist. Auch für einen Computer lässt sich dies definieren, indem strikte Regeln angewendet werden, die auf Messgrößen beruhen, wie etwa die Rotationsrate. Allerdings gehen mit der Erstellung solcher Regeln auch einige Probleme einher: (1) Die Regeln müssen notwendigerweise strikt und einfach sein, (2) ein menschlicher Ingenieur muss sich die Zeit nehmen, diese Regeln zu definieren, (3) die Regeln müssen mit Hilfe irgendeines Systems aufrecht erhalten und regelmäßig auf ihre Genauigkeit überprüft werden, weil sich die Maschinerie oder die Anlage im Laufe ihrer Lebenszeit verändert, (4) die Dateneingabe ist fehleranfällig, wenn wir den Umfang solcher Definitionen berücksichtigen, die für eine industrielle Anlage mit hunderten von Maschinen und jeweils hunderten oder gar tausenden von Messungen nötig sind.

Alle diese Fragen können aber gelöst werden, wenn wir die vielen Punkte der Betriebsanlage auf der Grundlage einer Datenanalyse zusammen gruppieren, statt dass wir Regeln nach menschlichen Gesichtspunkten entwerfen. Diese Vorgehensweise nennen wir Clustering (Gruppierung). Dabei versuchen wir die Zahl der Cluster und deren Mitglieder mittels einer historischen Sammlung von Beobachtungen zu bestimmen, abhängig von zwei Kriterien: (1) Zwischen den Mitgliedern eines Clusters sollte es nur wenige Unterschiede geben, wohingegen es (2) zwischen Mitgliedern unterschiedlicher Cluster große Unterschiede geben sollte.



Wenn wir es mit quantitativen Daten zu tun haben, können wir einen Cluster im Wesentlichen als eine Kugel verstehen. Mit anderen Worten: Ein Cluster ist ein Punkt

im Raum mit einem bestimmten Radius. Jede historische Beobachtung, die innerhalb dieses Radius zu liegen kommt, rechnen wir dem Cluster zu, und alle Punkte außerhalb dieser Kugel rechnen wir ihm nicht zu. Dabei ist klar, dass dieser Ansatz mit überlappenden Kugeln rechnen muss sowie mit Messpunkten, die in keine Kugel passen. Gleichwohl ist dieser Clustering-Ansatz der populärste.

In der Praxis stellen wir die Frage, wie viele Kugeln wir benötigen, wo sich diese befinden und wie groß sie sein müssen, damit sie den beiden Kriterien genügen. Dafür gibt es Methoden, mit deren Hilfe wir praktisch jeden Datensatz in sinnvolle Cluster aufteilen können. Unglücklicherweise bedürfen diese Methoden eines hohen Aufwands an erforderlichen Rechnerzeiten, aber immerhin können sie gut auf die Daten von Industrieanlagen angewandt werden.

Der Vorteil dieser Methoden ist es, dass man mit ihrer Hilfe einen automatisierten Ereignisrahmen (event framework) definieren kann. Wenn dann die Maschinen ihre Aufgabe erledigen, können wir leicht sagen, in welchem Zustand sich die Maschinen befinden bzw. welche Ereignisse gerade stattfinden. Jedesmal, wenn die Maschine ihren Zustand oder ihr Ereignis verändert, kann das aufgezeichnet werden. Beide Eigenschaften zusammengenommen können dazu verwendet werden, für die Maschine eine Berechnung für äquivalente Betriebsstunden zu betreiben, da die Übergänge von einem Zustand zum nächsten zwar zeitlich rasch erfolgen, aber durchaus einige Lebenszeit der Maschinerie in Anspruch nehmen.