Ira Assent,
Ralph Krieger,
Thomas Seidl

Evolutionäre Clustering-Algorithmen

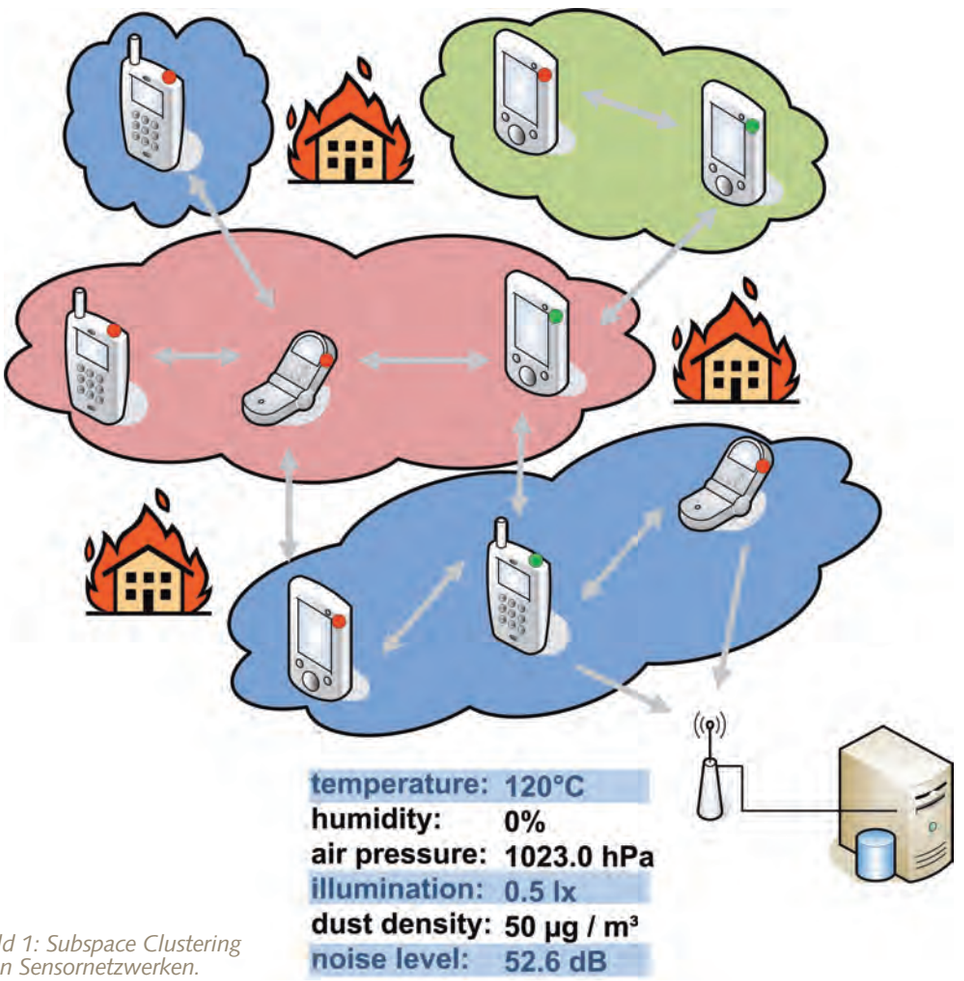


Bild 1: Subspace Clustering von Sensornetzwerken.

In vielen Bereichen der technisch-naturwissenschaftlichen Forschung sowie in zahlreichen alltäglichen Anwendungen entstehen zunehmend große und multidimensionale Datenmengen. Diese Datenmengen sind auf Grund ihrer Größe und Komplexität nicht mehr manuell zu durchsuchen oder zu analysieren. Für Genomdatenbanken stellen sich etwa Fragen wie „Welche Gene sind für ähnliche Funktionen zuständig?“ oder „Unter welchen Bedingungen werden ähnliche Gene aktiviert?“ In medizinischen Bilddatenbanken beispielsweise treten Fragen nach dem Vorkommen von Gruppen ähnlicher pathologischer Fälle oder nach Behandlungserfolgen bei Therapien mit ähnlichen Problembefunden auf. Auch im Internet trifft man auf vergleichbare Fragestellungen, etwa bei der Suche in Videoportalen: „Welche Filme sind ähnlich zu dem, den ich gerade angesehen habe?“ Die automatisierte Zusammenfassung von komplexen Objekten wie Gene, Bilder oder Videos in Gruppen ähnlicher Objekte wird als Clustering-Problem bezeichnet: In den gefundenen Clustern sollen ähnliche Objekte zusammengefasst werden, während Objekte aus verschiedenen Clustern möglichst unterschiedlich sein sollen.

Klassische Datenbankanfragen finden Objekte an Hand von beschreibenden Merkmalen, etwa Datum oder Ort einer Bildaufnahme oder die Angabe von Temperaturen, pH-Werten und Wirkstoffkonzentrationen bei Daten über Genexpressionsexperimente. Diese Merkmale, auch Attribute genannt, werden von den Benutzern für die Suche bestimmter Objekte vorgegeben. Heutige Datenbanktechniken haben kein Problem damit, klar spezifizierte Anfragen auch in großen Datenbeständen schnell zu bewältigen. Im Gegensatz dazu fällt die oben skizzierte Clustersuche in die Klasse der explorativen Anfragen, des so genannten Data Mining. Die Benutzer geben in der Regel keine klaren Beschreibungen ihrer Anfrage ab, sondern suchen nach Mustern und Regelmäßigkeiten in der Datenbank, ohne diese im Vorhinein genau beschreiben zu können. Vielfach sind sie auch an Abweichungen von häufigen Mustern interessiert, die auf interessante Sonderfälle schließen lassen. Konkrete Anwendungen hierzu stellen die Betrugserkennung bei Kreditkarten oder die Mobilkommunikation dar.

Die explorative Suche im Sinne des Data Mining ist Teil eines Prozesses, der in der Literatur als Wissensentdeckung in Datenbanken – englisch „Knowledge Discovery in Databases“ – bezeichnet wird. Hier werden die zu analysierenden Daten zuerst aufbereitet und bereinigt, um sie für die Analyse in geeigneter Form vorliegen zu haben. Das eigentliche Data Mining sucht nach neuen, bisher unbekannten Mustern in den Daten. Abschließend werden diese Muster für eine benutzerorientierte Darstellung visuell aufbereitet, so dass die Fachexperten oder auch die Alltagsbenutzer vertiefte Einsichten über in den Daten bislang verborgene Schätze gewinnen können.

Eine wichtige Aufgabe des Data Mining ist das bereits erwähnte Clustering: Ähnliche Objekte sollen zu Gruppen zusammengefasst werden, wobei die Objekte aus verschiedenen Gruppen sich deutlich unterscheiden sollen. Beispielsweise illustriert Bild 1 ein Clustering auf Sensoren für die Brandbekämpfung. Um Energie zu sparen, werden diejenigen Sensoren automatisch zu Clustern zusammengefasst, die die gleichen Umgebungsbedingungen messen. Aus diesen Clustern wird dann abwechselnd nur ein repräsentativer Sensor genutzt, und die übrigen können in den Bereitschaftsmodus geschaltet werden.

Weitere wichtige Anwendungsgebiete für Clustering sind die automatische Kundensegmentierung, also beispielsweise die Feststellung, ob es Kundengruppen gibt, die bezüglich ihres Konsumverhaltens Ähnlichkeiten aufweisen. Clustering ist auch in der Biologie von zentraler Bedeutung für die Auswertung von Experimenten, um Vorschläge für die Gruppenbildung sich ähnlich verhaltender Gene oder Proteine zu erzeugen. In all diesen Anwendungsbereichen hat Clustering zu neuen Erkenntnissen für die Benutzer geführt. Die konkrete Aufgabe des Clustering stößt bei Datenbanken mit zahlreichen Attributen schnell an ihre Grenzen. In der Informatik wird aktiv daran geforscht, wie Clusteralgorithmen für den Einsatz auf hochdimensionalen Daten funktionieren können. Von hochdimensionalen Daten spricht man, wenn sehr viele Attribute zur Objektbeschreibung vorliegen. Beispielsweise sind Genexpressionsexperimente durch pH-Werte, Temperaturen, Konzentrationen von

Bionik-Methoden für Knowledge Discovery in Datenbanken

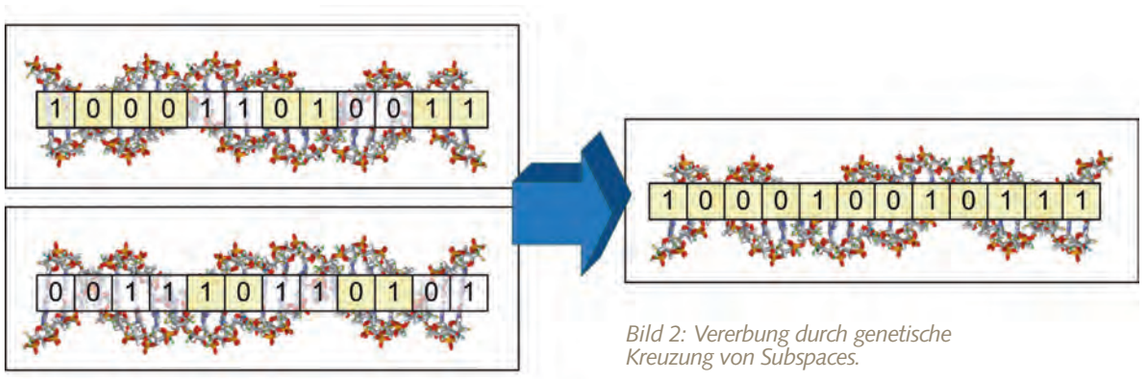


Bild 2: Vererbung durch genetische Kreuzung von Subspaces.

Stoffen der Zelle sowie von externen Wirkstoffen und eine Fülle weiterer Parameter beschrieben. Kundendaten können neben persönlichen Angaben auch bisherige Konsumparameter oder weitergehende soziodemographische Merkmale beinhalten. In diesen hochdimensionalen Daten tritt ein Effekt auf, der in der Informatik als „Fluch der Dimensionalität" bezeichnet wird. Dies bedeutet, dass traditionelle Clusteringalgorithmen keine Muster mehr finden können, da zu viele Unterscheidungen möglich werden.

Es ist eben bei Weitem nicht klar, welche Attribute und Merkmalskombinationen dazu führen, dass Objekte in Gruppen zusammenfallen. Die Suche beginnt nun damit, die für die Cluster bestimmenden Attribute überhaupt erst festzustellen. Man versucht daher die Menge der betrachteten Attributkombinationen einzuschränken. Die Suche nach relevanten Merkmalen wird in der Literatur als Subspace Clustering bezeichnet, da durch die Beschränkung auf wenige Attribute nur Unterräume des ursprünglichen Raumes betrachtet werden.

Ein solches Subspace Clustering ist in Bild 1 dargestellt. Für den untersten Subspace Cluster sind lediglich die drei blau unterlegten Merkmale Temperatur (temperature), Lichtverhältnisse (illumination) sowie Geräuschpegel (noise level) relevant. Die übrigen Merkmale variieren sehr stark in ihren Werten und spielen für diesen Subspace Cluster daher keine Rolle.

Die Frage beim Subspace Clustering ist also, in welchen der vielen Unterräume sich signifikante Clusterstrukturen finden lassen. Eine einfache Rechnung zeigt, dass bei zehn ursprünglichen Merkmalen schon über Tausend Unterräume existieren, bei zwanzig und dreißig gegebenen Attributen gibt es über eine Million beziehungsweise eine Milliarde mögliche Unterräume. Es ist offensichtlich, dass auf Grund dieses exponentiellen Wachstums eine vollständige Überprüfung aller Unterräume schlichtweg zu lange dauert. Während in diesem Umfeld schon einige nützliche heuristische und approximative Ansätze entstanden sind, geht die Suche nach effizienten Verfahren unvermindert weiter, um auch große Parameterräume bewältigen zu können.

Diese Situation ist wie geschaffen für einen bionikorientierten Ansatz. Die Natur hat an vielen Stellen mit großen Parameterräumen zu tun, für die sie in vielfältiger Weise optimale Lösungen hervorgebracht hat. Der Schlüssel hierzu ist in den Mechanismen der Evolution zu finden, die wir uns auch bei der Unterraumsuche zu Nutze machen. Anstatt nach einer Methode zu suchen, mit der alle Subspaces berechnet werden, imitieren wir den evolutionären Prozess der Auslese. In unserer Simulation evolutionärer Prozesse wird eine Population möglicher Lösungen bezüglich ihrer „Fitness" einer Selektion für die Fortpflanzung unterworfen. Nach mehreren „Generationen" kristallisieren sich gemäß dem Prinzip des „survival of the fittest" geeignete Lösungen heraus. Wir modellieren Attributkombinationen als Individuen einer Population und messen ihre Fitness als Grad der Neigung, Cluster zu bilden. Im Ergebnis erhalten wir diejenigen Unterräume, die die Datenbank in möglichst gute Gruppen zusammenfassen.
Der zentrale Modellierungsschritt für das Subspace Clustering besteht darin, Unterräume als Individuen für den evolutionären Prozess zu kodieren. Dazu werden mehrdimensionale Bitvektoren verwendet, in denen jede einzelne Stelle anzeigt, ob ein bestimmtes Attribut dem individuellen Unterraum angehört oder nicht, ob das Attribut also für die Clustersuche an- oder abgeschaltet ist. Die Länge der Bitvektoren entspricht damit der Dimension des ursprünglichen, vollständigen Raums. Im zweiten Modellierungsschritt wird eine für die evolutionäre Modellierung unabdingbare Fitnessfunktion gewählt. Diese bestimmt in unserem Fall die Eignung eines Individuums, also eines Unterraums, gute Clusterstrukturen auszubilden. Hier sind große Unterschiede in den Unterräumen zu beobachten. Attribute, die zum Untersuchungsziel nur Rauschen beitragen, werden durch eine niedrige Clusterfitness entlarvt. Attribute hingegen, die häufig zu Clusterstrukturen beitragen, bekommen eine hohe Fitness zugesprochen. Unterräume mit solchermaßen geeigneten Attributen verbleiben häufiger in der Population als Unterräume mit stark rauschenden Merkmalen. Sie werden entsprechend häufig mit anderen ebenfalls geeigneten Individuen gekreuzt, wodurch sich gute Clusterattribute dauerhaft durchsetzen können.

Bild 2 zeigt die Kreuzung von Unterräumen. Die einzelnen Unterräume sind genetisch als so genannte Bitcodes modelliert, die in der Grafik als Folge von Nullen und Einsen dargestellt sind. Eine Eins kodiert dabei ein für das Clustering relevantes Merkmal, Nullen zeigen irrelevante Merkmale an. Bei der Kreuzung wird eine Kombination der genetischen Codes der Ausgangsindividuen gebildet. Dabei entsteht ein neues Individuum mit entsprechendem Bitcode, der eine neue Auswahl relevanter Merkmale kodiert. In unserer Darstellung erbt das neue Individuum die gelb markierten Teile im Bitcode seiner Eltern. Durch diese Kreuzung und ebenfalls erlaubte Mutationsvorgänge entstehen nacheinander neue Generationen von Unterräumen, bei denen sich im evolutionären Modell, diejenigen mit der höchsten Fitness durchsetzen.

Evolutionäre Unterraumsuche begeht auch Nebenwege und untersucht weitere Möglichkeiten. Ausgehend von einer zufällig gewählten Menge von Unterräumen kristallisieren sich auf diese Art und Weise mit der Zeit gute Lösungen in dem riesigen Suchraum heraus.

Am Lehrstuhl für Informatik 9 (Datenmanagement und -exploration) werden diese Verfahren derzeit mit prototypischen Implementierungen evaluiert. Erste Ergebnisse auf synthetisch erzeugten Clustern in mehrdimensionalen Daten zeigen den Erfolg des evolutionären Ansatzes, weitere Untersuchungen sind auf Daten aus realen Anwendungen geplant. Da gleichzeitig auch die klassischen algorithmischen Lösungen in aktiven Forschungsprojekten weiterentwickelt werden, bleibt der Vergleich in diesem Wettbewerb auch in der Zukunft spannend. In jedem Fall profitieren die Anwendungen von den Fortschritten, die auf beiden Seiten erzielt werden.

www-i9.informatik.rwth-aachen.de

Autoren:
Dipl.-Inform. Ira Assent und Dipl.-Inform. Ralph Krieger sind Wissenschaftliche Mitarbeiter am Lehrstuhl für Informatik 9 (Datenmanagement und -exploration).
Univ.-Prof. Dr. rer. nat. Thomas Seidl ist Inhaber des Lehrstuhls für Informatik 9 (Datenmanagement und -exploration).