Studierendenprogramm der 11. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2005), Karlsruhe, Germany.

Subspace Sequence Clustering

Datamining zur Entscheidungsunterstützung in der Hydrologie

Boris Glavic
boris.glavic@post.rwth-aachen.de

Zuordnung: Datamining, Mustererkennung, kartographische Daten

In einem bundesweiten Projekt wird aktuell an der Erfüllung der EU-Richtlinie 2000/60/EG [1] für Flußrenaturierung gearbeitet. Im Rahmen dieses Projektes wurden Daten über die Fließgewässer in NRW kartiert. Basierend auf diesen Daten beschäftigt sich das Lehr- und Forschungsgebiet für Ingenieurhydrologie der RWTH Aachen (Univ.-Prof. Dr.-Ing. Heribert Nacken) mit dem Erstellen von Renaturierungsmaßnahmen. In diesem Zusammenhang sollen inhärente Strukturen aus den Daten extrahiert werden. Aus dieser Aufgabenstellung entstand die Kooperation zwischen dem Lehrstuhl für Informatik IX (Datenmanagement und Exploration, Univ.-Prof. Dr. rer. nat. Thomas Seidl) und dem Lehr- und Forschungsgebiet Ingenieurhydrologie.

In einem ersten Schritt wurden Dataminingmethoden zur Clusteranalyse und Assoziationsregelbestimmung auf den Daten angewendet. Auf Basis der daraus gewonnen Ergebnisse und den dabei aufgetretenen Probleme wurde im Rahmen meiner Diplomarbeit ein neues algorithmisches Konzept entwickelt, welches Subspacecluster in mehrdimensionalen Sequenzen findet. Die Ergebnisse des Algorithmus sollen den Hydrologen genauere Kenntnisse der Daten vermitteln und sie beim Entwurf von Renaturierungsmaßnahmen unterstützen.

Zum besseren Verständnis der Problemstellung werden im Folgenden die Gewässerdaten genauer beschrieben. Für das Projekt wurden Strukturgütedaten von etwa 2000 Flüssen aus NRW erfasst [2]. Dabei wurden insgesamt ca. 120.000 Flussabschnitte von je 100 Meter Länge kartiert. Jeder Abschnitt wurde bezüglich 19 verschiedener Kriterien (funktionale Einheiten) bewertet und einer Güteklasse (1 bis 7) zugeordnet. Die Aufgaben der Experten in der Ingenieurhydrologie ist es, Maßnahmen zur Verbesserung der Strukturgüte zu entwerfen. Dabei werden bevorzugt Maßnahmen bzw. Maßnahmenkombinationen gesucht, die auf möglichst lange zusammenhängende Flussabschnitte anwendbar sind. Solche Maßnahmenkombinationen sind in der Umsetzung weniger kostenintensiv und erfordern weniger Organisations- und Zeitaufwand.

Voraussetzungen für die Anwendbarkeit einer Maßnahme sind meist mehrere unterschiedliche Eigenschaften eines Flußabschnitts, die sich typischerweise nicht eins-zu-eins in die Güteklassen der funktionalen Einheiten übertragen lassen. Die dadurch entstehende Unschärfe wird noch verstärkt durch die Tatsache, dass die Kartierung und Klassifikation durch unterschiedliche Personen vorgenommen wurde. Deshalb werden die Maßnahmenvoraussetzungen als Intervalle von Güteklassen einer oder mehrerer funktionaler Einheiten in Form von Regeln angegeben. Von Seiten der Ingenieurhydrologie ist man unter anderem an folgenden Fragestellungen interessiert: Wieviele Einzelabschnitte werden durch eine Regel abgedeckt? Welche längeren Abschnitte sind durch eine Regel abgedeckt? Über die vorgegebenen Regeln hinaus soll ein an der Struktur der Daten orientierter Ansatz verfolgt werden, um die Qualität der Regeln zu überprüfen und gegebenenfalls neue Regeln aufzustellen.

Aus Sicht der Informatik eignen sich insbesondere Methoden des Datamining, um Informationen in den Daten zu explorieren[3]. Die funktionalen Einheiten eines Flußabschnitts können als einzelne Dimensionen aufgefasst und zu einem mehrdimensionalen Punkt zusammengefasst wer-

den. Für die Ingenieurhydrologie sind Informationen über Korrelationen zwischen den einzelnen Dimensionen von Interesse. Viele Algorithmen zur Korrelationsbestimmung setzen einen kontinuierlichen Wertebereich bzw. arithmetische Operationen voraus. Wegen des nominalen Charakters der Daten, d.h. die Zahlenwerte der Güteklassen entsprechen lediglich Kategorien, wurden deshalb in einem ersten Schritt Assoziationsregelalgorithmen angewendet. Diese Algorithmen arbeiten auf einzelnen Transaktionen, was in unserem Fall einzelnen Flußabschnitten entspricht [4]. Durch die Zuordnung der Abschnitte zu den einzelnen Flüssen und die Fließrichtung dieser Gewässer sind Sequenzen von Einzelabschnitten festgelegt. Dieser Sequenzbezug bleibt bei den oben genannten Assoziationsregelmethoden unberücksichtigt. Es existieren zwar Assoziationsregelalgorithmen für Teilsequenzen [5], doch diese basieren auf einem lückenbehafteten Teilsequenzbegriff und sind deshalb für unsere Anwendung ungeeignet. Bei diesen Algorithmen müssen einzelne Elemente einer Teilsequenz in der Originalsequenz nicht direkt aufeinander folgen, sondern dürfen einen beliebigen Abstand voneinander besitzen. Da Maßnahmen für lange zusammenhängende Flußabschnitte gesucht werden und sich die Maßnahmen für einen Flußabschnitt auf die nachfolgenden Abschnitte auswirken, muß für unsere Anwendung ein lückenloser Teilsequenzbegriff gefordert werden.

Es besteht die Möglichkeit, diese Methoden so zu modifizieren, dass ihnen ein lückenloser Sequenzbegriff zugrunde liegt. Dabei bleibt ein grundlegendes Problem, das bei der Verwendung von Assoziationsregeln auf unseren Daten auftritt, ungelöst: Assoziationsregelalgorithmen können nur Häufigkeiten von Einzelsequenzen berechnen. Wegen der Unschärfe der Daten und der Art der Maßnahmenvoraussetzungen, ist man eher an Korrelationen zwischen Mengen von ähnlichen Sequenzen interessiert. Als eine Alternative bieten sich deshalb Clusteringalgorithmen an. Typische Clusteringalgorithmen arbeiten auf mehrdimensionalen Daten, sind aber nicht für das Bearbeiten von Sequenzen entworfen worden. Es ist möglich, Teilsequenzen einer konstanten Länge n zu clustern, wenn man sie als n-dimensionale Punkte auffasst. Dies führt zu Problemen, da für die Anwendung Cluster von möglichst langen Teilsequenzen gesucht werden sollen. Ein weiteres Problem ist, daß dichtebasierte Clusteringmethoden nicht für Nominaldaten geeignet sind, da auf diesen Daten kein Dichtebegriff definiert ist. Auch die wenigen für Nominaldaten entworfenen Algorithmen sind für unsere Daten nicht geeignet, da diese Methoden auf ungeordnete Nominaldaten zugeschnitten sind und somit die natürliche Ordnung der Güteklassen nicht ausnutzen/berücksichtigen.

Zur Lösung dieser Probleme wurde im Rahmen der Diplomarbeit ein neuer dichtebasierter Clusteringalgorithmus entworfen, der Cluster von Teilsequenzen findet. Dazu mußte ein sinnvoller Dichtebegriff für Nominaldaten definiert werden. Eine weitere Herausforderung bestand darin, die Sequenzstruktur der Daten in die Clustermethode zu integrieren. Da man auch an der Korrelation der Cluster untereinander interessiert ist, wurde eine zweite Phase des Algorithmus entworfen, in der durch geeignete Transformation und Reduktion der Daten diese Beziehungen effizient untersucht werden können.

Wie oben bereits erwähnt werden in einem ersten Schritt in jeder Dimension (funktionalen Einheit) Cluster von Sequenzen gesucht. Dieser Schritt wurde ausgehend von bekannten dichtebasierten Clusteralgorithmen [6] entwickelt. Im Laufe der Diplomarbeit wurden verschiedene neue Dichtebegriffe für Nominaldaten definiert und untersucht. Eine Sequenz ist dabei dicht, wenn die Summe aus der relativen Häufigkeit ihres Auftretens in der Datenbank (Support) und dem Support ihrer Nachbarn (mit Distanz kleiner ϵ) einen Schwellenwert α_{min} überschreitet. Dabei wird der Support der Nachbarn je nach Distanz zu der betrachteten Sequenz mit einer Gewichtungsfunktion σ unterschiedlich stark gewichtet. Mit diesem Dichtebegriff ist es möglich eindimensionale Sequenzen einer festgelegten Länge n zu clustern. Da Cluster für unterschiedliche Längen n gesucht werden, kann das Clusterverfahren in mehrere Iterationen mit verschiedenen Werten für n angewendet werden. Diese Art der Berechnung ist nicht sehr effizient, denn auf Grund des Sequenzbezugs sind die Ergebnisse der einzelnen Iterationen voneinander abhängig.

Wir konnten diese Abhängigkeit als neuen Monotoniebegriff formalisieren und nachweisen. Dies wird genutzt, um die Effizienz des Verfahrens zu steigern.

Um die Effizienz des Verfahrens auf der gegebenen großen Datenmenge zu gewährleisten, ist eine geeignete Indexstruktur unumgänglich. Daher ist ein Aspekt meiner Diplomarbeit der Entwurf eines Baumindex, welcher den effizienten Zugriff auf den Support einer Sequenz und der für die Dichtebestimmung relevanten Nachbarsequenzen ermöglicht. Der Index ist so aufgebaut, das die Monotonieeigenschaft genutzt werden kann, damit nicht bei jeder Iteration des Clusterverfahrens ein eigener Index neu aufgebaut werden muss. Das Ergebnis dieser ersten Phase sind dichtebasierte Cluster von verschiedenen langen Teilsequenzen für jede funktionalen Einheit.

Für die Maßnahmenerstellung sind neben den Sequenzclustern in den einzelnen funktionalen Einheiten auch die Beziehungen zwischen Clustern von Interesse. Deshalb wird in einer zweiten Phase nach Korrelationen zwischen den Clustern gesucht. Dazu werden die Daten folgendermaßen transformiert: Für jeden Abschnitt werden die Cluster gespeichert, welche eine Sequenz enthalten, die an diesem Abschnitt beginnt. In den transformierten Daten werden dann häufig auftretende Mengen von Clustern (Frequent Itemsets) gesucht. Als Grundprinzip wird hierbei die in [7] vorgestellten Methode verwendet. Bei der Anwendung auf die transformierten Daten speichert diese Methode redundante Informationen, da lange Cluster aus der Erweitung von kürzeren Clustern entstehen und somit dieselben Startpunkte haben. Im weiteren Verlauf der Diplomarbeit wird nach Möglichkeiten gesucht, die Berechnung von redundanten Frequent Itemsets zu vermeiden.

Das Endergebnis des Algorithmus sind Mengen von Clustern unterschiedlicher funktionaler Einheiten, die häufig zusammen in den Daten auftauchen. Die Cluster wiederum bestehen aus Teilsequenzen, die einander ähnlich sind und das oben beschriebene Dichtekriterium erfüllen. Somit ist es mit diesem neuen Algorithmus möglich Subspacecluster in mehrdimensionalen Sequenzen zu finden. Diese Clustermengen helfen den Hydrologen die Struktur der Flüße zu verstehen, da sie aufzeigen welche Art von Flußstrukturen häufig vorkommen. Dadurch können entworfene Maßnahmen überprüft werden indem getestet wird, ob diese Maßnahmen auf viele der vorhandenen Flußabschnittssequenzen anwendbar sind. Außerdem wird der Entwurf von neuen Maßnahmen unterstützt, wenn Clustermengen vorhanden sind auf denen keine der bisherigen Maßnahmen einsetzbar sind. Im weiteren Verlauf der Diplomarbeit wird der Einfluß der verschiedenen Parameter des Algorithmus (α_{min}, ϵ, σ) auf Effizienz und Qualität der Ergebnisse experimentell und analytisch evaluiert. Weiterhin wird die Qualität der Ergebnisse und die Effizienz dieser Methode für andere Anwendungsgebiete, wie z.B. Zeitreihen untersucht.

Literatur

[1] *Rat der EU: RICHTLINIE 2000/60/EG des europäischen Parlaments und des Rates*, 2000.

[2] *Strukturgütedaten NRW der kleinen und mittelgroßen Fließgewässer*, erstellt im Auftrag des LUA NRW, Stand 8.9. 2003.

[3] Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[4] Agrawal, R., Srikant, R.: *Fast Algorithms for Mining Association Rules*, Proc. VLDB'94, 487-499.

[5] Ayres, J., Flannick, J., Gehrke, J., Yiu, T.: *Sequential pattern mining using a bitmap representation*, Proc. SIGKDD'02, 429-435.

[6] Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Proc. KDD'96, 226-231.

[7] Han, J., Pei, J., Yin, Y.: *Mining frequent Patterns without Candidate Generation*, Proc. SIGMOD'00, 1-12.