Automatische Parameterbestimmung durch Gravitation in Subspace Clustering

Jiwu Zhao
Institut für Informatik
Datenbanken und Informationssysteme
Heinrich-Heine-Universität Düsseldorf
D-40225 Düsseldorf, Germany
zhao@cs.uni-duesseldorf.de

Zusammenfassung

Im Vergleich zu den traditionellen Clusteringverfahren ermöglicht Subspace Clustering die Suche nach Clustern in den Unterräumen (Subspaces) der Daten. Man unterscheidet zwei Hauptarten des Subspace-Clustering-Verfahrens: Top-Down- und Bottom-Up-Verfahren. Die Algorithmen des Top-Down-Verfahrens verkleinern die Suchbereiche von hohen zu niedrigen Dimensionen. In dem Bottom-Up-Verfahren suchen die Algorithmen nach Clustern in einer umgekehrten Reihenfolge.

Die Bestimmung der Parameter in den meisten Subspace-Clustering-Verfahren ist nicht einfach, was die Anwendung der Verfahren erschwert. Daher wird in dieser Arbeit ein Verfahren zur automatischen Parameterbestimmung diskutiert.

Keywords: *Subspace Clustering, Parameterreduzierung, Gravitation, Bottom-Up-Verfahren*

1 Grundlagen

Die traditionellen Clustering-Verfahren wie z.B. K-means, DBscan usw. suchen nach Clustern normalerweise im gesamten hochdimensionalen Raum. Subspace Clustering stellt eine Erweiterung des traditionellen Clustering dar, das die Suche nach Clustern in den Unterräumen ermöglicht. Es bietet mehr relevante Informationen, da nur die Kombination von manchen Datenattributen sinnvoll ist. Die *Abbildung 1* zeigt ein Beispiel, in dem drei Cluster auf jeweils zwei, aber nicht drei Dimensionen liegen.

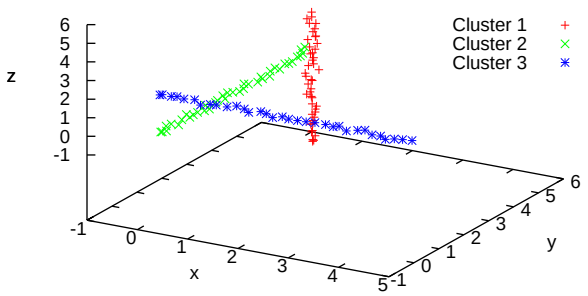

Abbildung 1: Subspace Clustering

Es gibt zwei Hauptarten des Subspace-Clustering-Verfahrens: Top-Down- und Bottom-Up-Verfahren[1]. Zu den Top-Down-Verfahren zählen PROCLUS[2], ORCLUS[3], FINDIT[4], σ-Clusters[5], COSA[6] usw. Die bekannten Beispiele für Bottom-Up-Verfahren sind CLIQUE [7], ENCLUS [8], MAFIA [9], CBF [10], CLTree [11], DOC [12] usw.

Das Problem der meisten Subspace-Clustering-Verfahren ist die Parameterbestimmung. Die *Tabelle 1* veranschaulicht die Parameter der oben genannten Verfahren.

Verfahren	**Kurze Beschreibung**	**Parameter**
PROCLUS	K-Medoid	K
ORCLUS	Abstand zweier Punkte im Unterraum	1. Clusteranzahl 2. Größe des Unterraumes
FINDIT	Medoids mit eigener Abstandfunktion	1. Minimale Punktanzahl im Cluster 2. Minimaler Abstand zwischen den Clustern
σ-Clusters	Coherence (Kohärenz)	1. Clusteranzahl 2. Clustergröße
COSA	Gewichtung der Punkte im Unterraum, K-nächste Nachbarn	1. Grenzwert des Gewichts 2. K
CLIQUE	Gruppierung durch Raster	1. Raster-Größe 2. Grenzwert der Dichte
ENCLUS	Entropie	Größe des Rasterintervalls
MAFIA	Histogramm	1. Grenzwert der Dichte 2. Grenzwert für Merging
CBF	Cell-basiert	1. Grenzwert der Sektion 2. Minimale Dichte
CLTREE	Decision Tree	1. Minimale Anzahl in einem Bereich 2. Dichte für Merging
DOC	Dichtebasiert	1. Maximale Länge 2. Minimale Punktanzahl in einem Cluster

Tabelle 1: Vergleich von Parametern in Subspace-Clustering-Verfahren

Die Tabelle zeigt, dass die meisten Verfahren Parameter benötigen müssen, die sehr schwer zu bestimmen sind. Daher wird in der vorliegenden Arbeit ein Verfahren mit automatischer Parameterbestimmung diskutiert.

2 Verfahren

2.1 Definition von Subspace Clustern

Eine Datenbank enthält Objekte und Attribute. Die letzteren können als Dimensionen betrachtet werden, wobei die Objekte die Punkte in diesen Dimensionen sind. Eine Datenbank kann man deswegen als einen Raum $\mathcal{D}$ ansehen, der eine Kombination von einem Attributraum $\mathcal{A}$ und einem Objektraum $\mathcal{O}$ ist:

$$\mathcal{D} = (\mathcal{A}, \mathcal{O})$$

Der Attributraum $\mathcal{A}$ ist definiert als die Menge aller Attribute $\mathcal{A} = \{a_1, a_2, \cdots, a_n\}$, und der Objektraum $\mathcal{O}$ ist die Menge aller Objekte $\mathcal{O} = \{o_1, o_2, \cdots, o_p\}$. Ein Subspace Cluster $\mathcal{S}$ ist ein Unterraum von $\mathcal{D}$:

$$\mathcal{S} = \widetilde{\mathcal{D}} = (\widetilde{\mathcal{A}}, \widetilde{\mathcal{O}})$$

wo $\widetilde{\mathcal{A}} \subset \mathcal{A}$, $\widetilde{\mathcal{O}} \subset \mathcal{O}$ und $\mathcal{S}$ eine in jedem Verfahren unterschiedlich definierte Bedienung $\mathcal{C}$ erfüllt.

Die Anzahl der Objekte in $\mathcal{S}$ wird mit $|\mathcal{S}|$ bezeichnet und $|\mathcal{S}| = |\widetilde{\mathcal{O}}|$. Die Schnittemenge von zwei Subspace Cluster wird wie folgt definiert:

$$\mathcal{S}_1 \cap \mathcal{S}_2 = (\widetilde{\mathcal{A}}_1 \cup \widetilde{\mathcal{A}}_2, \widetilde{\mathcal{O}}_1 \cap \widetilde{\mathcal{O}}_2)$$

2.2 Eindimensionales Gravitationsverfahren (Schritt 1)

Für hochdimensionale Daten ist es schwer, die Anzahl der Cluster vorher einzuschätzen, deswegen ist die Durchführung eines K-medoid-basierten Verfahrens problematisch. Aus diesem Grund basiert der Algorithmus auf Bottom-Up-Verfahren des Subspace Clustering. Der erste Schritt des Gravitationverfahrens wird in eindimensionalen Räumen durchgeführt.

Nehmen wir an, dass jedes Objekt ein wahres Objekt mit Gewicht ist, der Datenbankraum $\mathcal{D}$ geographische Eingenschaften hat und die Objekte in einer Dimension Gravitation (Anziehungskraft) erzeugen können.

Die Gravitationsfunktion ist definiert als $G = \mathcal{G} \cdot \frac{m_1 m_2}{r^2}$. Für die Vereinfachung nehmen wir die Gravitationskonstante heraus, dann erfolgt die Bestimmung der vereinfachten Gravitation zwischen zwei Objekten o_a, o_b folgendermaßen:

$$G_{ab} = \frac{m_a m_b}{r_{ab}^2}$$

Es sei hervorgehoben, dass das Gewicht von einem Objekt $1/N$ ist, wobei N die Anzahl der Objekte ist und der Abstand zwischen o_a, o_b als $r_{ab} = \frac{l_{ab}}{L/(N-1)}$ definiert ist, der eine Proportion des wahren zu dem durchschnittlichen Abstand $L/(N-1)$ bezeichnet, wo L die Länge einer Dimension ist. Dann kann die Funktion des Gravitationsverfahrens für eine Dimension so aussehen:

$$G_{ab} = \frac{1/N^2}{\left(\frac{l_{ab}}{L/(N-1)}\right)^2} = \frac{L^2}{l_{ab}^2 N^2 (N-1)^2}$$

Falls $l_{ab} = 0$ ist, dann werden a und b ein neues Objekt mit Gewicht $m_a + m_b$ bilden.

Die Gravitation eines Objektes und die globale Gravitation in einer Dimension werden durch folgende Maße berechnet:

- Die Gravitation eines Objektes o in der Dimension A: Summe der Gravitation mit allen Objekten, $G_o = \sum_{\forall p \in A} G_{op}$
- Die globale Gravitation einer Dimension A ist der Durchschnittswert der Gravitationen aller Objekte, $\overline{G} = \frac{1}{N} \sum_{\forall o \in A} |G_o|$

Die Gravitation in einem eindimensionalen Raum zeichnet sich durch folgende Eigenschaften aus:

- Ein Objekt in der Mitte hat einen größeren Wert als eines am Rand
- In einem kleinen Bereich liegen die Objekte mit einem großen Gravitationswert näher zu den Nachbarobjekten als die mit einem kleineren Gravitationswert.

Im Vergleich zu $\overline{G}$ kann das Objekt o mit einer sehr kleinen G_o schon als ein Outlier entfernt werden. Es sei betont, dass ein Objekt mit kleiner Gravitation in der Lage ist, wie ein Separator die Clusterobjekte zu trennen. Daher bekommen wir die Clusterinformationen in einem eindimensionalen Unterraum.

In der *Abbildung 2* kann man deutlich sehen, dass die Clusterobjekte im Vergleich zu den Nicht-Clusterobjekten größere Werte der Gravitation haben.

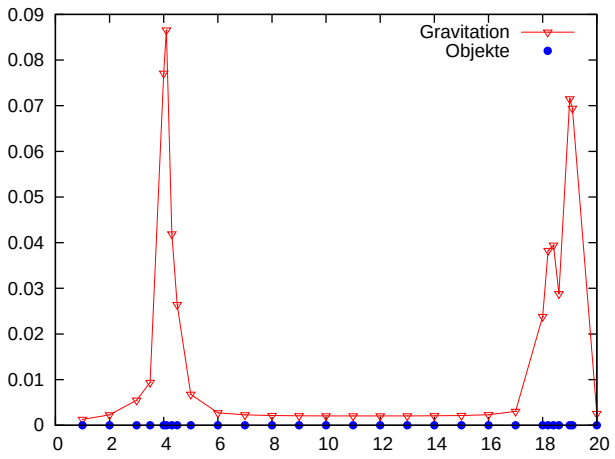

Abbildung 2: Gravitation der Objekte im eindimensionalen Raum

2.3 Suche nach hochdimensionalen Subspace Clustern (Schritt 2)

Nachdem die Cluster in jeder Dimension gefunden worden sind, müssen sie miteinander in hohen Dimensionen kombinieren werden, damit neue Subspace Cluster gebildet werden können. Die Kombination wird durch das Prinzip der Schnittmenge durchgeführt. Zum Beispiel wird für Subspace Cluster $\mathcal{S}_1$ und $\mathcal{S}_2$ die Schnittmenge $\mathcal{S}_1 \cap \mathcal{S}_2$ überprüft. Falls $\mathcal{S}_1 \cap \mathcal{S}_2$ die Bedienung $\mathcal{C}$ erfüllt, ist die neue Menge auch ein Subspace Cluster. Dieser Prozess wird durchgeführt, bis kein neuer Subspace Cluster gefunden wird. Dann ist das Subspace Clustering abgeschlossen.

Das Überprüfen ist einfach: Die Anzahl der Objekte in $\mathcal{S}_1 \cap \mathcal{S}_2$ wird berechnet, und falls $|\mathcal{S}_1 \cap \mathcal{S}_2| > M$ ist, bildet $\mathcal{S}_1 \cap \mathcal{S}_2$ einen neuen Subspace Cluster. M ist die minimale Anzahl von Objekten in einem Cluster. Da $|\mathcal{S}_i \cap \mathcal{S}_j| \leq |\mathcal{S}_i|$ ist, haben die neugebildeten Cluster nicht mehr Objekte als die älteren, deswegen soll M nicht groß sein. Wenn M auf 0 gesetzt wird, dann ist es möglich ganz kleine Cluster zu finden, die später entfernt werden können. M ist trotzdem ein Parameter, den man durch eine grobe Bestimmung im Verfahren verwenden kann.

3 Schlussfolgerung

In dieser Arbeit wurde die Idee für Subspace Clustering mit einer Gravitationsfunktion vorgestellt, um das Problem der schweren Parameterbestimmung zu beseitigen.

Die vereinfachte Gravitationsfunktion wird auf eindimensionale Objekte verwendet. Dank den guten Eigenschaften der Gravitationsfunktion ist es möglich Cluster und Outlier zu trennen. Hochdimensionale Subspace Cluster können durch die Kombination von Objekten in niedrigen Unterräumen gebildet werden.

Eine spätere Arbeit wird sich der Beziehung von Objekten durch die Gravitation in mehreren Dimensionen widmen. Es wird versucht, die vereinfachte Gravitationsfunktion noch zu verfeinern, damit sie an hohe Dimensionen angepasst werden kann und die Trennung der Cluster- von Nicht-Clusterobjekten besser erfolgt.

Literatur

[1] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: A review. *Sigkdd Explorations*, 6:90–105, June 2004.

[2] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park. Fast algorithms for projected clustering. In *Proceedings of the 1999 ACM SIGMOD international*

conference on Management of data, pages 61–72, Philadelphia, Pennsylvania, United States, May 31-June 03 1999.

[3] Charu C. Aggarwal and Philip S. Yu. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 70–81, Dallas, Texas, United States, May 15-18 2000.

[4] K.-G. Woo and J.-H. Lee. *FINDIT: a Fast and Intelligent Subspace Clustering Algorithm using Dimension Voting*. PhD thesis, Korea Advanced Institute of Science and Technology, Taejon, Korea, 2002.

[5] J. Yang, W. Wang, H. Wang, and P. Yu. δ-clusters: Capturing subspace correlation in a large data set. In *Proceedings of the 18th International Conference on Data Engineering*, page 517, February 26-March 01 2002.

[6] J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2002.

[7] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, Seattle, Washington, United States, June 01-04 1998.

[8] Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 84–93, San Diego, California, United States, August 15-18 1999.

[9] S. Goil, H. Nagesh, and A. Choudhary. Mafia: Efficient and scalable subspace clustering for very large data sets. Technical report cpdc-tr-9906-010, Northwestern University, June 1999.

[10] Du-Seok Jin Jae-Woo Chang. A new cell-based clustering method for large, high-dimensional data in data mining applications. In *Proceedings of the 2002 ACM symposium on Applied computing*, pages 11–14, Madrid, Spain, March 2002.

[11] Bing Liu, Yiyuan Xia, and Philip S. Yu. Clustering through decision tree construction. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 20–29, McLean, Virginia, United States, November 06-11 2000.

[12] Cecilia M. Procopiuc, Michael Jones, Pankaj K. Agarwal, and T. M. Murali. A monte carlo algorithm for fast projective clustering. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, Madison, Wisconsin, June 03-06 2002.