COSMIC: Conceptually Specified MI-Clusters

Grundidee:

- dichtebasiertes hierarchisches Clustering für MI-Objekte
- Konzepte werden durch dichte Bereiche im Instanzraum beschrieben. (robust gegenüber Parameter Wahl)

  => Clustering der Instanzen mit OPTICS-ähnlichem Verfahren
- MI-Objekte bestehen dann wieder aus Realisierungen dieser Konzepte.
- MI-Cluster werden über Menge von Objekten beschrieben die Instanzen in den gleichen Konzepten haben.

  => Cluster können überlappen.

  => Cluster können in Ober- und Teilcluster von einander sein, wenn Beschreibung nur eine Teilmenge der Konzepte enthält.

Konzept Gitter

- Ein Object *O* wird wird durch die Menge aller enthaltenen Konzepte *Desc(O)*$\subseteq A$ beschrieben.(A = Menge aller Konzepte)
- Die binäre Relation *I* bildet dann Konzepte auf Objekte ab: $I \subseteq DB \times A$

Cluster: 1. $C = \{o \in DB \mid \forall a \in Desc(C) : (o,a) \in I\}$

2. $Desc(C) = \{a \in A \mid \forall o \in DB : (o,a) \in I\}$

Beispiel: Gegeben eine Bilddatensatz *DB*
und die Konzepte *A: {Haus, Dach}* $\subseteq A$

- *C = {alle Bilder die Häuder mit Dächern darstellen}*
- *Desc(C) ={Haus,Dach)*

$\Rightarrow$ Es existiert kein Bild $i \in DB$, das ein Haus mit einem Dach darstellt, das nicht Element von *C* ist.

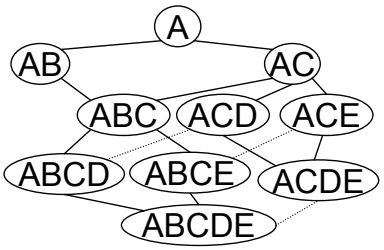
$\Rightarrow$ Es existiert kein weiteres Konzept k', das ebenfalls alle Bilder des Clusters C beschreiben würde.

Konzeptionelles Dichtebasiertes MI-Clustering

1. Gegeben $O = \{o_1,..,o_n\}$: bilde $o_k \in IR^d$ auf das diskrete Attribut A ab
   => Clustering der Instanzen

2. Bilde alle möglichen MI-Cluster
   => Konzept Gitter

3. Jeder Cluster beschreibt eine Mengevon MI-Objekten mit einer Menge von ähnlichen Instanzen
   - ⇒ die Cluster überlappen
   - ⇒ die Größe von *Desc(C)* beschreibt die Stärke der Verbindung
   - ⇒ Das Konzept Gitter beschreibt also alle Möglichkeiten um Cluster zu bilden



Beispiel für ein Konzept Gitter

Übersicht über COSMIC

1. Bilde Instanz-Menge und verwende angepassten OPTICS-Algorithmus um Reachability Plot abzuleiten.
2. Durchlaufe Reachability Plot mit einem Top-Down-Sweepline Algorithmus, der Konzepte und Konzept Gitter expandiert.

Herausforderungen:

Sehr viele dichtebasierte Instanz-Cluster

=> Auswahl der Instanz-Cluster zur Einschränkung der Konzepte

Kriterien:

1. Ist Instanzcluster allgemein genug für Konzept
2. Ist Konzept notwendig um MI-Cluster zu unterscheiden

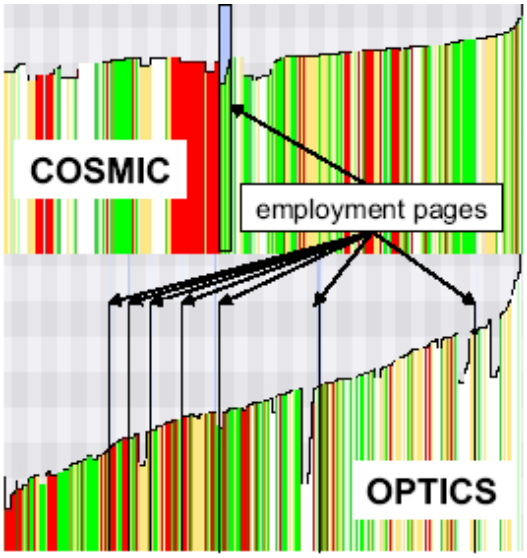
Clustering der Instanzen

Problem: Ein Instanz-Cluster stellt kein Konzept dar, falls nur Instanzen einer oder zu weniger Objekte enthalten sind.
(z.B. Konzept beschreibt nur 1 MI-Objekt)

Lösung: *Erweitere Prädikat*
Für jedes Objekt *O* zählt bei der Bestimmung der Kern-Distanz nur noch die nächst gelegene Instanz anderer Objekte.

⇒ Kerndistanz hängt von mindestens MinPts Objekten ab, anstatt von MinPts Instanzen.

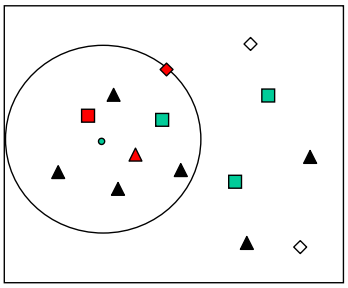
⇒ Bei der Erreichbarkeit werden wie bisher alle Objekte gezählt.



Beispiel Websites: Nur die Berücksichtigung der Objektzugehörigkeit erzeugt nützliche Konzepte.

Clustering der Instanzen

Definition: *Concept-Core-Distance*
Sei $MinObjs \in IN$, $\varepsilon \in IR^+$ und DB eine Menge von MI-Objekten. $I_{DB} = \bigcup_{DB} o$. Als *MinObjs* nächste-Nachbarn einer Instanz i bezeichnet man die kleinste Menge $N^{MI}_{MinObjs}(i) \subseteq I_{DB}$ so dass folgende Bedingungen gelten:



$$(1) \forall p \in N^{MI}_{MinObs}(i), \forall q \in DB \setminus N^{MI}_{MinObs}(i) : d(p,i) < d(q,i)$$

$$(2) \left| \left\{ MiObj(x) \middle| x \in N^{MI}_{MinObs}(i) \right\} \right| \geq MinObs$$

Dann ist $d_{MinObs}(i) = \max\left\{ d(i,q) \middle| q \in N^{MI}_{MinObs}(i) \right\}$ und die Concept-Core-Distance ist definiert durch:

$$ConceptCoreDist^{\varepsilon}_{MinObs}(i) = \begin{cases} d_{MinObs}(i) & : d_{MinObs}(i) \leq \varepsilon \\ \infty & d_{MinObs}(i) > \varepsilon \end{cases}$$

Definition: Concept-Reachability-Dinstance

$$\text{ConceptReachDist}^{\varepsilon}_{\text{MinObs}}(i,j) = \max\left\{ ConceptCoreDist^{\varepsilon}_{MinObs}(i), d(i,j) \right\}$$

Ableiten von Konzepten

Gegeben: Reachability-Plot mit ConceptReachDist erstellt.

Probleme:

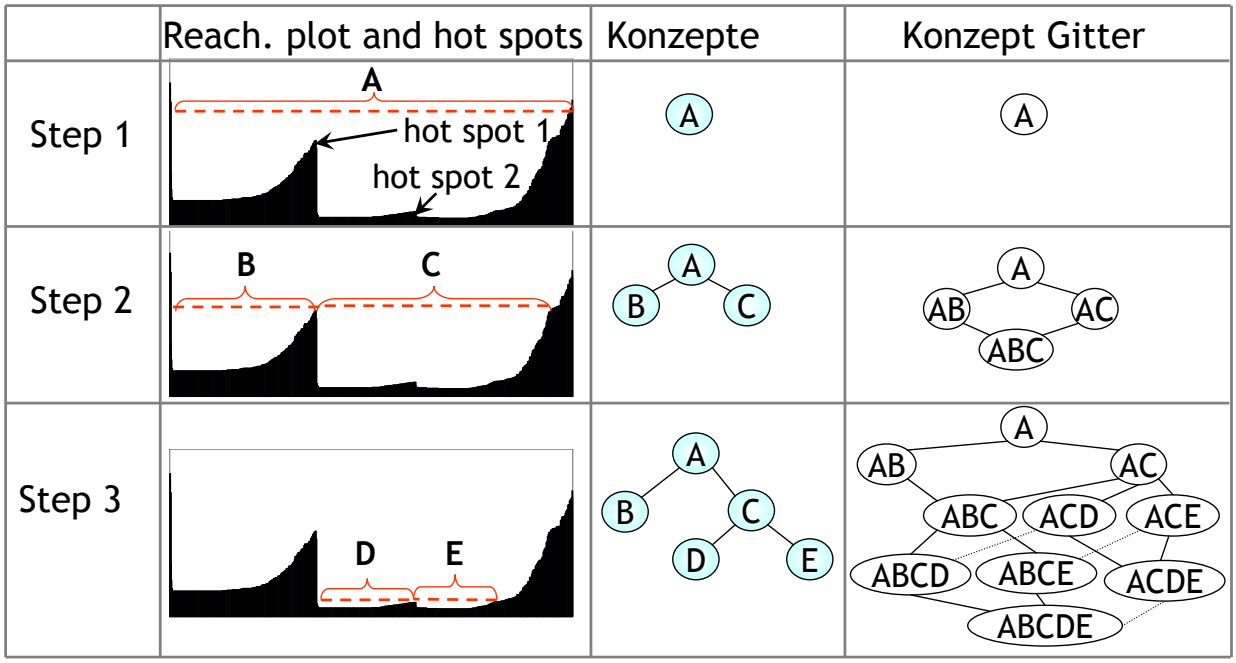
- Konzepte sind nur implizit in Tälern vorhanden.
  => Ableiten der Konzepte.
- Nicht alle Konzepte beschreiben einen ausreichend großen Cluster => Ableiten von überflüssigen Konzepten vermeiden.

Ideen:

- allgemeine Konzepte werden eher benötigt um MI-Cluster zu beschreiben=> top-down
- Leite keine Konzepte ab deren Vaterkonzept nicht Teil einer Clusterbeschreibung ist.

Bestimmung von Konzepten und Clustern

Beispielablauf für das Ableiten von Konzepten und Konzept-Gittern



Hot Spots im Reachability Plot

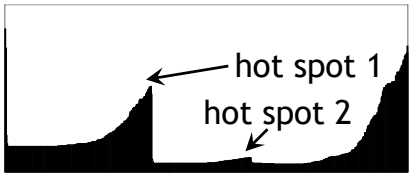
Extraktion der Konzepte aus dem Plot:

- Hot-Spot: Punkt an dem 2 Cluster im Plot unterschieden werden

  (1) $reach(i) > reach(i+1)$

  (2) $\exists l \in IN : reach(i-l) < reach(i) \wedge \forall k : (i-l) < k < i : reach(i) = reach(k)$

  reach(i) := Erreichbarkeitsdistanz auf Position *i* im Plot
- Hot-Spots werden während des Clusterings der Instanzen erkannt und Priority-Liste bzgl. der max. Erreichbarkeitsdistanz abgelegt.
- Hot-Spots trennen Cluster voneinander.



Bestimmung von Konzepten und Clustern

- COSMIC durchläuft die Liste der Hot Spots
- Jeder Hot Spot erzeugt mind. 1 neuen Konzept-Kandidaten
- Ist das Vater-Konzept des Kandidaten nicht in der Konzepthierarchie, dann braucht Kandidat nicht weiter betrachtet zu werden.
- Falls Vater in Hierarchie, teste ob mit dem neuen Konzept neue MI-Cluster gebildet werden können.
- Falls ja wird das Konzept zur Konzepthierarchie hinzugenommen und das Konzept Gitter um die neuen Cluster erweitert.

Fazit COSMIC

- COSMIC leitet alle möglichen MI-Cluster bzgl. eines Reachability Plots ab.
- Kann man die Konzepte (Instanz Cluster) mit einem Cluster-Modell gut approximieren (z.B. Centroid, Gauß-Kurve), dann ergibt die Menge der Konzept Beschreibungen eine Beschreibung der MI-Cluster.
- Obwohl COSMIC die Kardinalität der einzelnen Konzept nicht betrachtet, können die resultierenden Cluster diesbezüglich noch untersucht werden
- Overhead für die Analyse des Instanz-Plots ist verschwindend gering im Vergleich zum Clustering der Instanzen

Literatur

- Kriegel H.-P, Pryakhin A., Schubert M.: ***An EM-Approach for Clustering Multi-Instance Objects,*** Proc. 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2006), Singapore, 2006.
- Kriegel H.-P., Pryakhin A., Schubert M., Zimek A.:***COSMIC: Conceptually Specified Multi-Instance Clusters*** in proc. 6th int. Conference on Data Mining **(ICDM 2006)**, Hong Kong, China
- Dietterich T.G., Lathrop R.H., Lozano-Perez T. : ***Solving the Multiple Instance Problem with Axis-Parallel Rectangles,*** Artificial Intelligence, vol. 89, num.1-2, Seiten 31-71, 1997
- Weidmann N., Frank E., Pfahringer B.: ***A Two-Level Learning Method for Generalized Multi-instance Problems***. ECML 2003: S. 468-479
- Gärtner T., Flach P.A., Kowalczyk A., Smola A.j.: ***Multi-Instance Kernels***, *Proceedings of the 19th International Conference on Machine Learning*, p. 179-186, 2002
- Zhang Q., Goldman S.:***EM-DD: An improved multiple-instance learning technique***. Neural Information Processing Systems 14, 2001.
- Eiter T., Mannila H.: ***Distance Measures for Point Sets and Their Computation***. Acta Informatica, 34(2):103-133, 1997.
- Brecheisen S, Kriegel H.-P., Kröger P., Pfeifle M., Schubert M..: ***Using Sets of Feature Vectors for Similarity Search on Voxelized CAD Objects*** Proc. ACM SIGMOD Int. Conf. on Management of Data (**SIGMOD'2003**), San Diego, CA, 2003