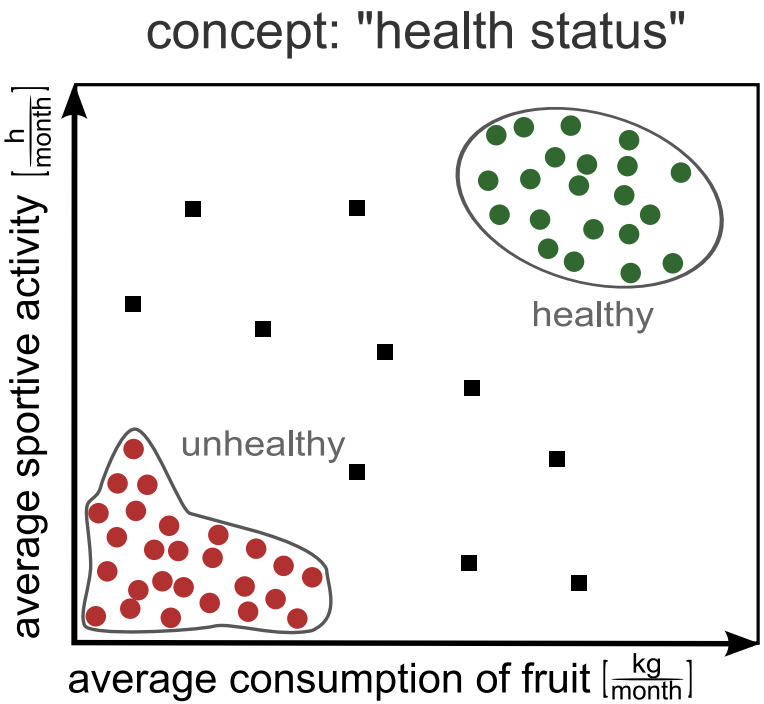
ASCLU – Alternative Subspace Clustering

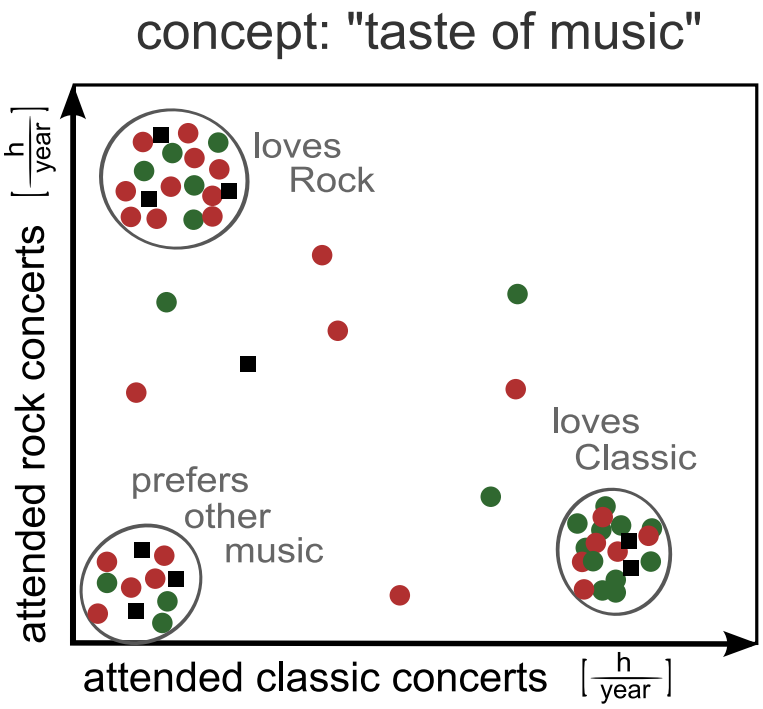
Stephan Günnemann Ines Färber Emmanuel Müller Thomas Seidl

Data management and data exploration group
RWTH Aachen University, Germany

MultiClust at KDD 2010
July 25, 2010

Why Subspace Clustering?



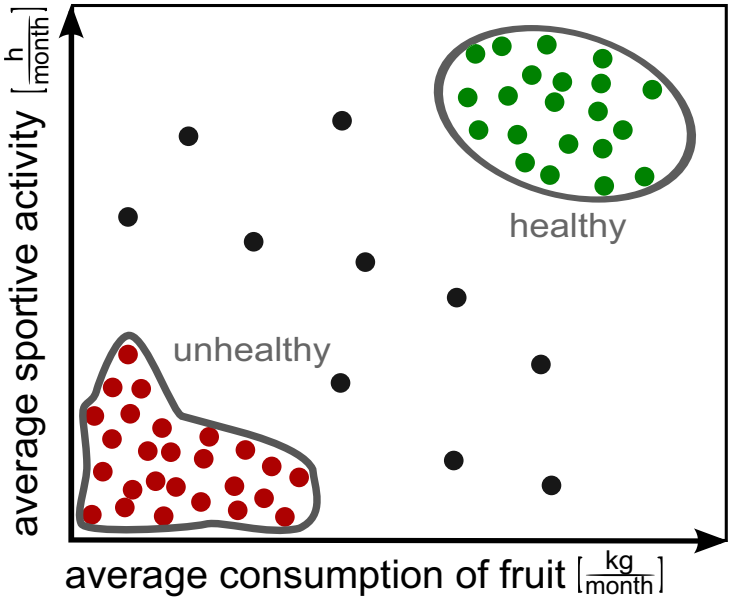


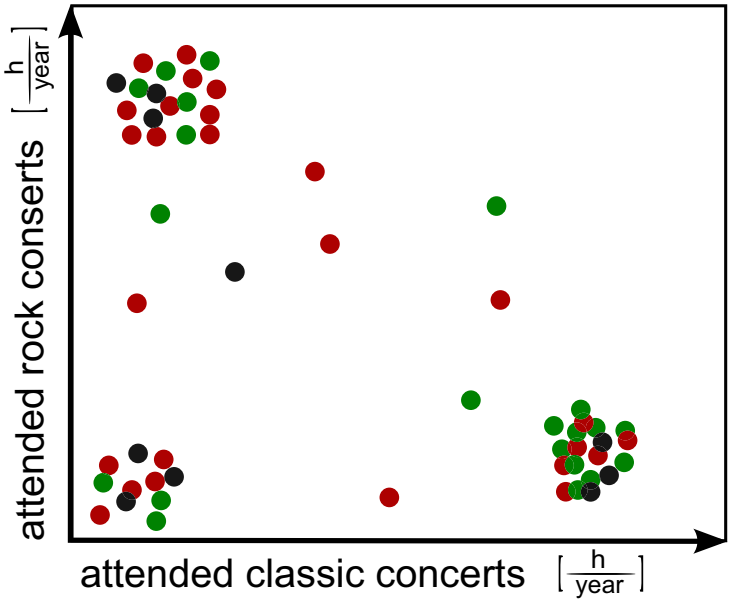
- data allows to be clustered through different perspectives
- each object in various groupings based on different attributes

⇒ multiple views due to locally relevant dimensions of clusters

⇒ subspace clustering

Why Alternative Clustering?





- often trivial groupings or already detected clusters given
- user not satisfied with previous results

⇒ aiming for alternative, yet comparable good groupings

⇒ avoid re-detection of already known clusters

→ combination of subspace clustering and alternative clustering

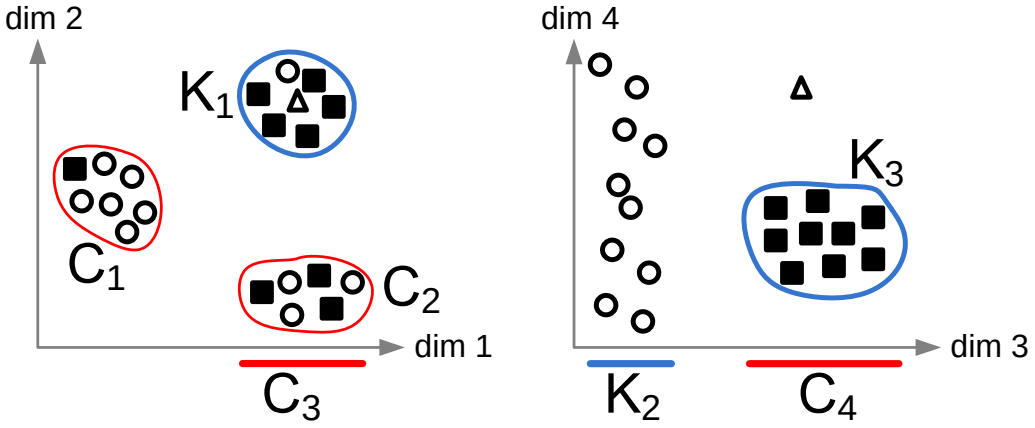
What To Do? – The General Picture

input: subspace clustering $Known = \{K_1, \ldots, K_m\}$

Cluster selection approach

set of possible subspace clusters *All*
$\rightarrow$ select optimal subset $Res \subseteq All$ fulfilling specific properties

- avoid redundancy
  $\rightarrow$ w.r.t. known clusters
  $\rightarrow$ among novel clusters
- select alternative clusters



Alternative Subspace Clustering Model – I

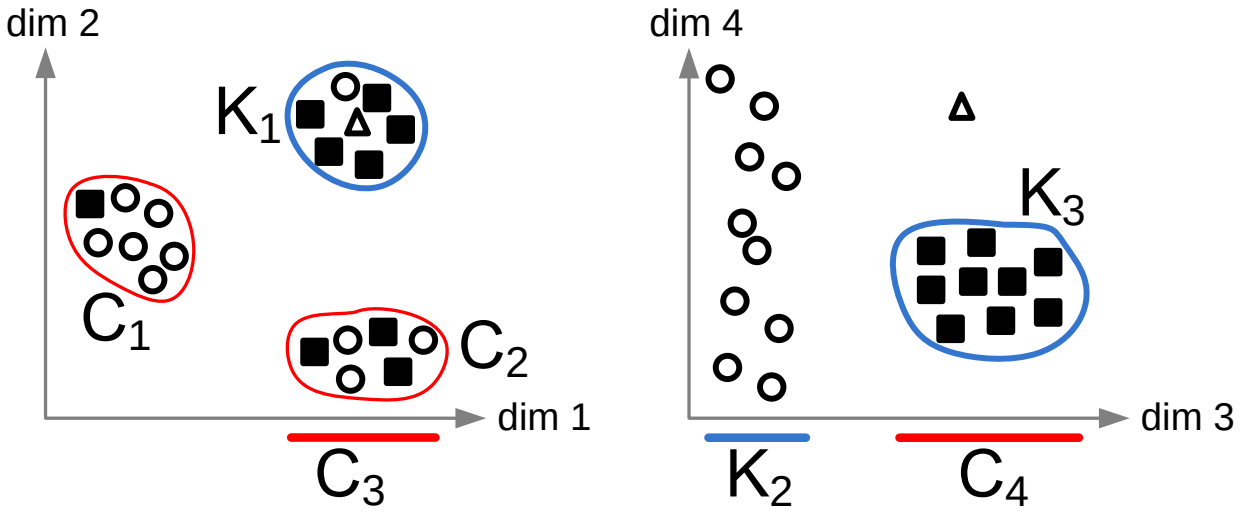
Each cluster $C \in Res \subseteq All$ should deviate from *Known*

1. deviating w.r.t. subspaces
   - known clusters in **alternative** subspaces are already different enough

$$InAlterSubspace(Known, C) = \{(O_i, S_i) \in Known \mid |S \cap S_i| < \beta \cdot |S|\}$$

2. deviating w.r.t. objects
   - known clusters in **similar** subspaces should cover different objects

$$\frac{|O \setminus CoveredInSimilar(Known, C)|}{|O|} \geq \alpha$$



Is C_1 an alternative to *Known*?

Alternative Subspace Clustering Model – I

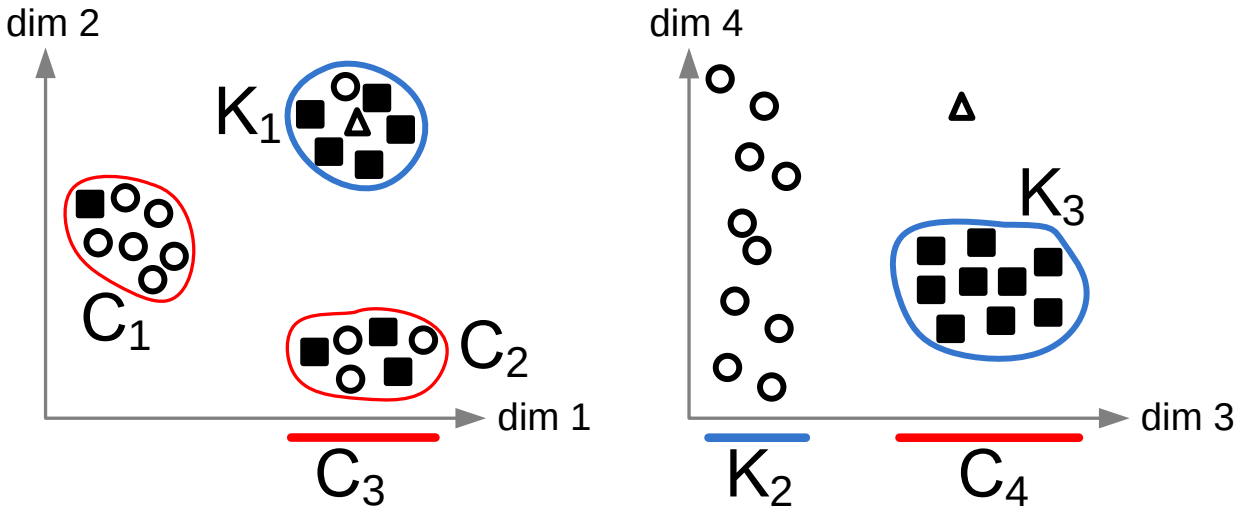
Each cluster $C \in Res \subseteq All$ should deviate from *Known*

1. deviating w.r.t. subspaces
   - known clusters in **alternative** subspaces are already different enough

$$InAlterSubspace(Known, C) = \{(O_i, S_i) \in Known \mid |S \cap S_i| < \beta \cdot |S|\}$$

2. deviating w.r.t. objects
   - known clusters in **similar** subspaces should cover different objects

$$\frac{|O \setminus CoveredInSimilar(Known, C)|}{|O|} \geq \alpha$$

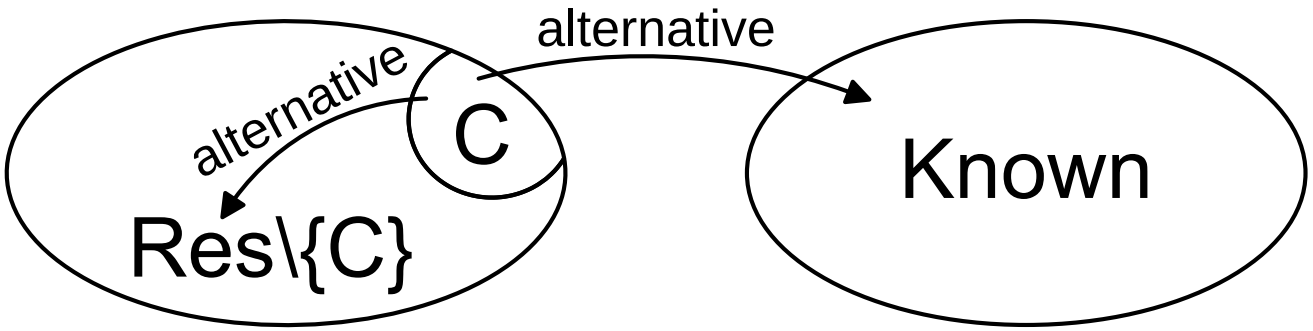


C_1 is a ***valid alternative*** to *Known*!

Alternative Subspace Clustering Model – II

Avoiding redundancy

- so far: $C \in Res$ different to given clusters $\Rightarrow$ non-redundant w.r.t. *Known*
- redundancy between $X, Y \in Res$ still possible
- solution: $C \in Res$ valid alternative to remaining novel clusters $Res \setminus \{C\}$



Alternative Subspace Clustering Model – II

Avoiding redundancy

- so far: $C \in Res$ different to given clusters $\Rightarrow$ non-redundant w.r.t. *Known*
- redundancy between $X, Y \in Res$ still possible
- solution: $C \in Res$ valid alternative to remaining novel clusters $Res \setminus \{C\}$

Optimal alternative subspace clustering

Given previous clustering *Known* and set of possible subspace clusters *All*, choose $Res \subseteq All$ such that

1. $\forall C \in Res : C$ is a ***valid alternative*** to *Known*
2. $\forall C \in Res : C$ is a ***valid alternative*** to $Res \setminus \{C\}$
3. *Res* is the most interesting clustering fulfilling 1 & 2

Conclusion

ASCLU – Alternative Subspace Clustering

- ASCLU detects alternatives based on
  $\longrightarrow$ deviating subspaces
  $\longrightarrow$ deviating object sets
- ASCLU avoids redundant clusters

Conclusion

ASCLU – Alternative Subspace Clustering

- ASCLU detects alternatives based on
  $\longrightarrow$ deviating subspaces
  $\longrightarrow$ deviating object sets
- ASCLU avoids redundant clusters

Thank you for your attention.

Questions?