An Instantiation of Hierarchical Distance-based Conceptual Clustering for Propositional Learning

A. Funes[1,2], C. Ferri[2], J. Hernández-Orallo[2], M. J. Ramírez-Quintana[2]

[1]Universidad Nacional de San Luis, San Luis, Argentina
afunes@unsl.edu.ar
[2]DSIC, Universidad Politécnica de Valencia, Valencia, España
{cferri, jorallo, mramirez}@dsic.upv.es

Agenda

- Motivation and objectives.
- Hierarchical Distance-based Conceptual Clustering (HDCC) algorithm.
- Analysis of consistency between distance and generalisation for propositional data.
- Experiments.
- Conclusions and future work.

Motivation

- Two different approaches for machine learning
  - Distance-based techniques
  - Model-based techniques

Motivation

- Distance-based techniques
  - Intuitive
  - Do not provide a description about a decision made for an individual.
    - Example: Clustering of molecules

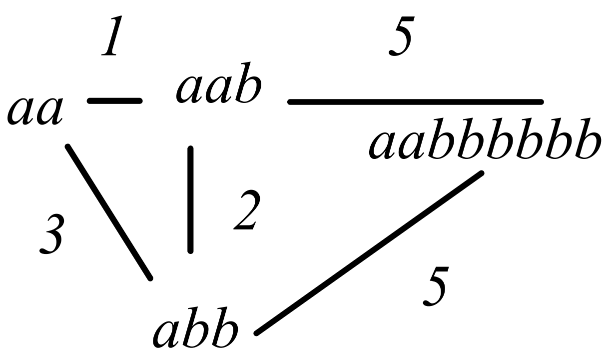
Objectives

- Combine both approaches on the basis of agglomerative hierarchical distance-based clustering.

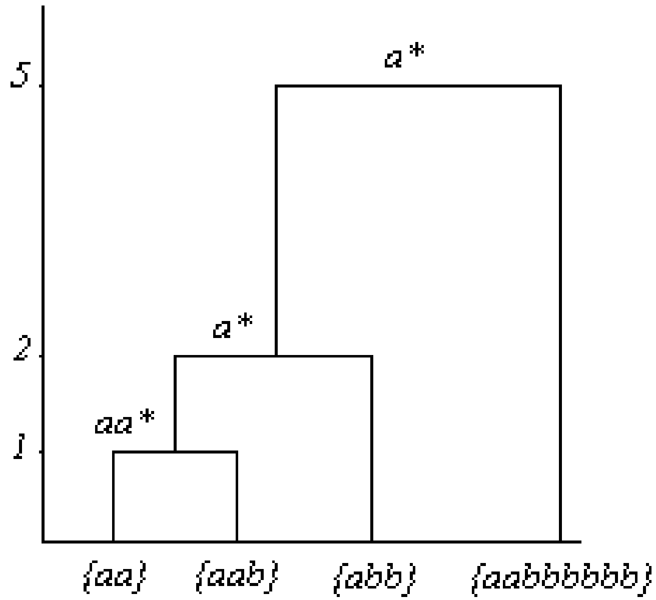
- Analyse the question:
  - *Are the elements in the clusters induced by a distance and the discovered patterns consistent?*
  - *Are all the elements covered by a pattern close w.r.t. the underlying distance?*

HDCC

- A first approach
  - On the basis of the traditional algorithm
    - Patterns are obtained either on-the-fly or as a post-process by using a *n*-ary generalisation operator.



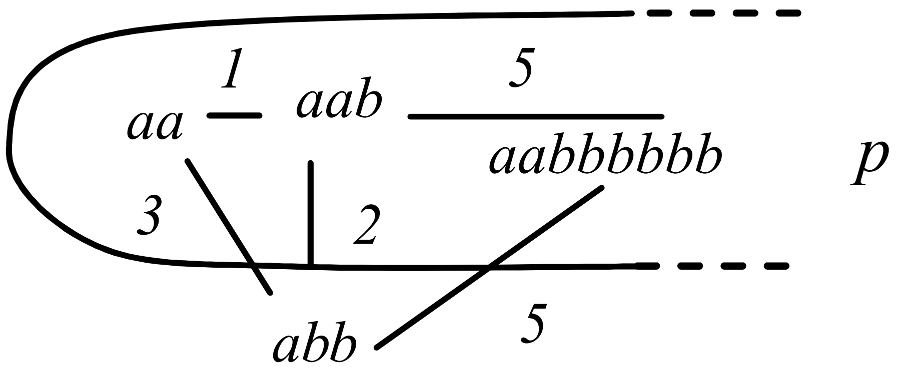
Four examples of lists in (Σ^*, *d*)



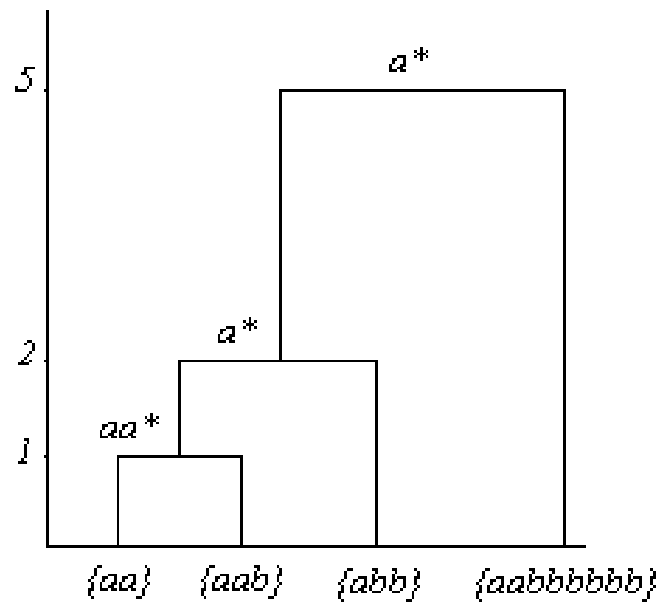
Dendrogram using single linkage distance

HDCC

- A first approach
  - Inconsistencies between the distance and the generalisation can arise.



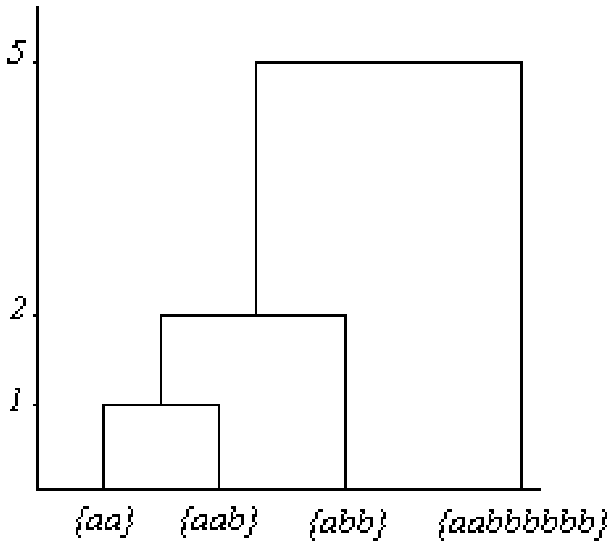
The coverage of pattern $p = aa*$



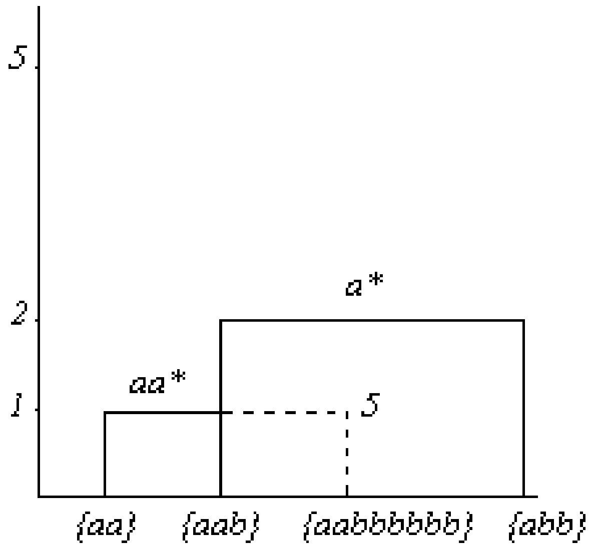
Dendrogram using single linkage distance

HDCC

- Our approach: Hierarchical Distance-based Conceptual Clustering (HDCC)
  - To overcome the inconsistency problem between the distance and the generalisation operator, HDCC performs at each iteration a coverage-reorganisation process.
    - Merge the two closest clusters according to the linkage distance.
    - ***Compute the pattern for the new discovered cluster using a pattern binary generalisation operator.***
    - ***Merge to the new discovered cluster all those clusters completely covered by the pattern*****.**



Traditional dendrogram



Conceptual dendrogram

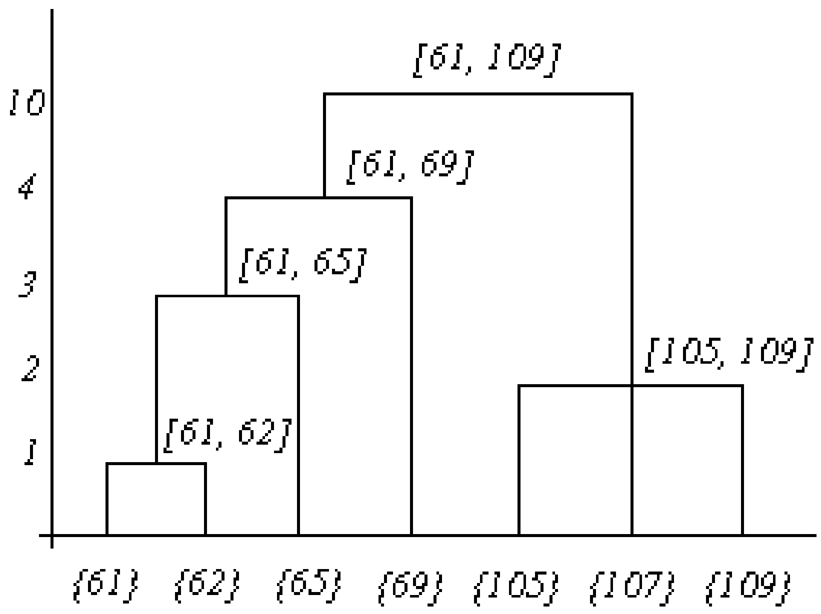
Consistency between Distances and Generalisation Operators

- We can observed that:
  - The dendrograms can differ considerably.
  - The shape of a conceptual dendrogram depends on:
    - the linkage distance d_L between clusters;
    - the distance d between elements in the metric space;
    - the generalisation operators used.
  - The more similar the dendrograms are, the more consistent the distance and the generalisations are.
- *We have defined different degrees of consistency between distances and generalisations on the basis of the similarity between a conceptual dendrogram and the traditional one.*
    - ***Equivalent to the traditional dendrogram***
    - ***Order-preserving***
    - ***Acceptable***

Consistency Levels

Equivalent Dendrograms

- A conceptual dendrogram is ***equivalent to the traditional dendrogram*** if for each cluster C all its children are linked at the same distance l.



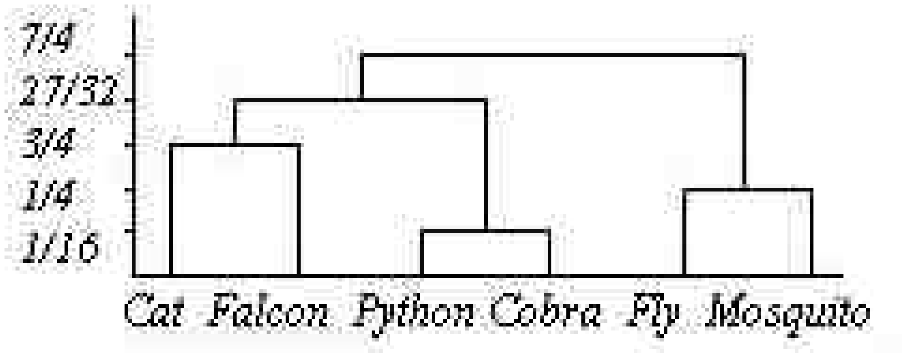
- Single linkage distance.
- Absolute difference.
- Closed intervals.

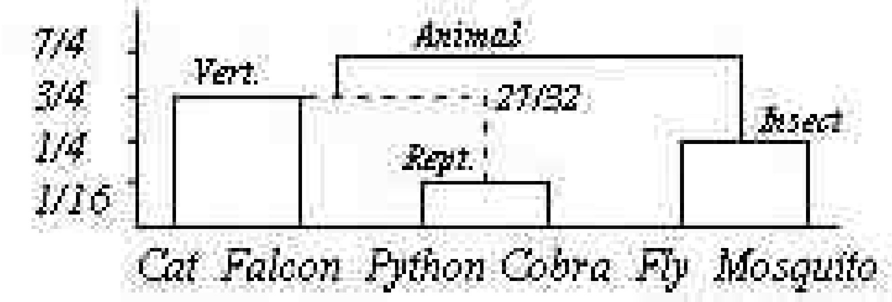
Consistency Levels

Order-preserving Dendrograms

- A conceptual dendrogram is ***order-preserving*** when the order in which its clusters are discovered is not swapped w.r.t. the traditional dendrogram.

- For any node (C, p, l) in the tree T, any child is linked at a same distance l, or it is linked by its pattern p at a linkage distance l' lower than the linkage distance from any other cluster not covered by the pattern.





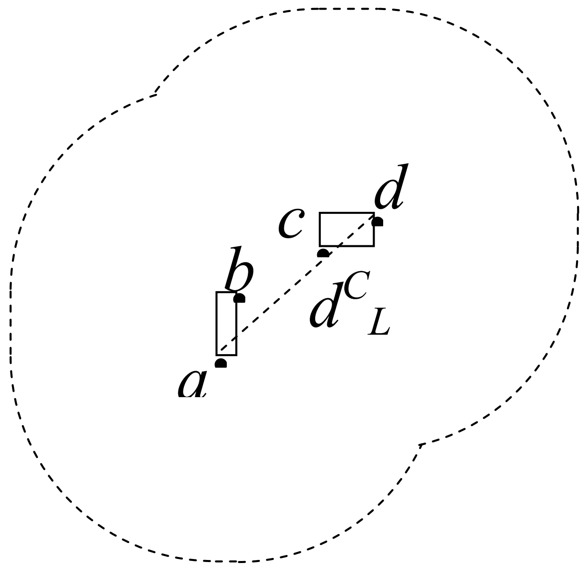
Consistency Levels

Acceptability Property

- A conceptual dendrogram is ***acceptable*** if it is the result of the use of an acceptable generalisation operator.

- Dendrograms can differ significantly.

- ***Acceptable operators***
  - A pattern should not cover elements whose distances to the old elements are greater than the maximum distance between the old elements.



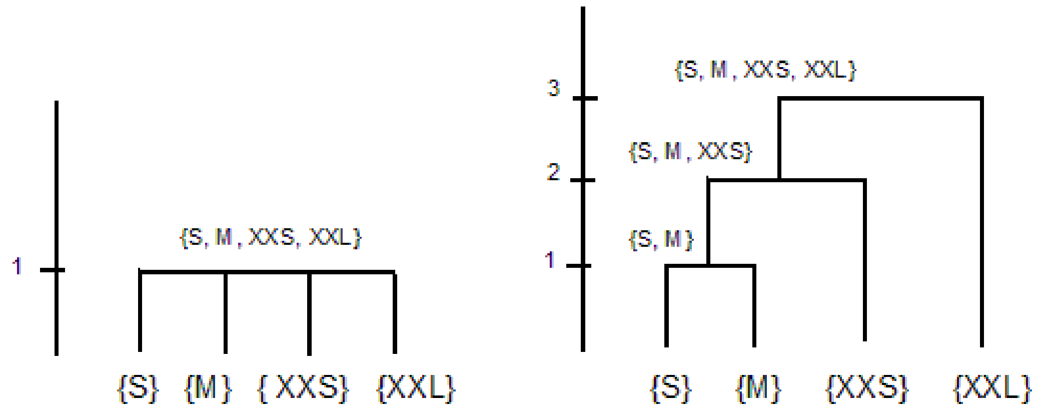
Instantiation for Propositional Clustering

- Data are expressed in terms of instances and attributes.
- We analise the consistence of these datatypes:
  - Numerical
  - Nominal
  - Tuples

Instantiation for Propositional Clustering

Nominal Data

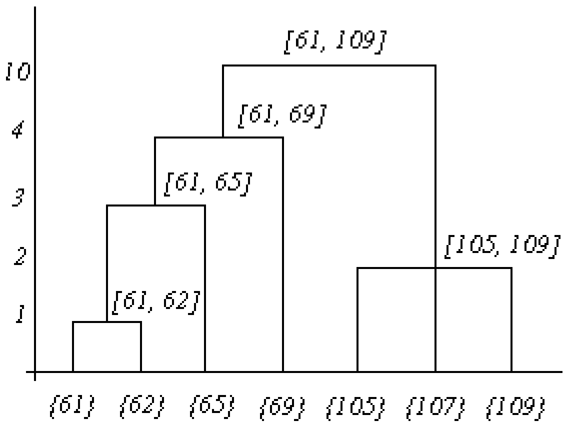
- Discrete Distance; User defined Distance.
- Generalisation: set union.
  - Equivalent dendrograms.



Instantiation for Propositional Clustering

Numerical Data

- Absolute Distance.
- Generalisation: minimum closed intervals.
  - Equivalent dendrograms.



Instantiation for Propositional Clustering

Tuples

- Distances:
  - Manhattan: $d(x,y)=\sum_{i=1}^{n} d_i(x_i,y_i)$
  - Euclidean: $d(x,y)=\sqrt{\sum_{i=1}^{n} d_i(x_i,y_i)^2}$
  - Chebysev: $d(x,y)=\max_{1\le i\le n} d_i(x_i,y_i)$
  - Weighted versions

Instantiation for Propositional Clustering

Tuples

- Generalisation:
  - The generalisation of two tuples *x* and *y* is defined as the tuple whose components are the generalisations of the respective components in *x* and *y*
- Coverage is defined in a similar way.

Instantiation for Propositional Clustering

Tuples

- Under the previous conditions:
  - The composability property of the generalisation can only be proved wrt. the complete linkage distance.

Experiments

Setting A

- Generalisation operators and distances for tuples that applied to HDCC under complete linkage distance produces equivalent conceptual dendrograms
  - Now they provide a description of each cluster in the hierarchy.

Experiments

Setting A

- Iris Dataset.
  - 150 instances, 3 classes
  - HDCC: complete and single linkage
  - Classes are not employed for learning

Experiments

Setting A

- Clusters:

		Pattern
C1	**Single**	([4.3,5.8],[2.3,4.4],[1.0,1.9],[0.1,0.6])
	Complete	([4.3,5.8],[2.3,4.4],[1.0,1.9],[0.1,0.6])
C2	**Single**	([4.9,7.7],[2.0,3.6],[3.0,6.9],[1.0,2.5])
	Complete	([4.9,6.1],[2.0,3.0],[3.0,4.5],[1.0,1.7])
C3	**Single**	([7.7,7.9],[3.8,3.8],[6.4,6.7],[2.0,2.2])
	Complete	([5.6,7.9],[2.2,3.8],[4.3,6.9],[1.2,2.5])

- Interpretation as a rule:
  - `(sepallength ≥ 4.3 AND sepallength ≤ 5.8 AND sepalwidth ≥ 2.3 AND sepalwidth ≤ 4.4 AND petallength ≥ 1.0 AND petallength ≤ 1.9 AND petalwidth ≥ 0.1 AND petalwidth ≤ 0.6)`

Experiments

Results

- Clustering quality S reflects the mean scattering over the k clusters
- Purity *P* can be interpreted as classification accuracy under the assumption that all the objects of a cluster are classified to be members of the dominant class for that cluster.

$$S = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\sum_{j=1}^{m}\sum_{l=j+1}^{m} d(\vec{x_j}, \vec{x_l})^2} \qquad P = \frac{1}{n}\sum_{i=1}^{k}\max_j \left(n_i^j\right)$$

- Values of S and P for the traditional and conceptual dendrograms under complete and single linkage distances

Linkage distance	$S_{\text{Traditional}}$	$S_{\text{Conceptual}}$	$P_{\text{Traditional}}$	$P_{\text{Conceptual}}$
Single (d^s_L)	46.56	46.56	0.68	0.68
Complete (d^c_L)	37.44	37.44	0.84	0.84

Experiments

Setting B

- If we use single linkage distance HDDC can produce different dendrograms

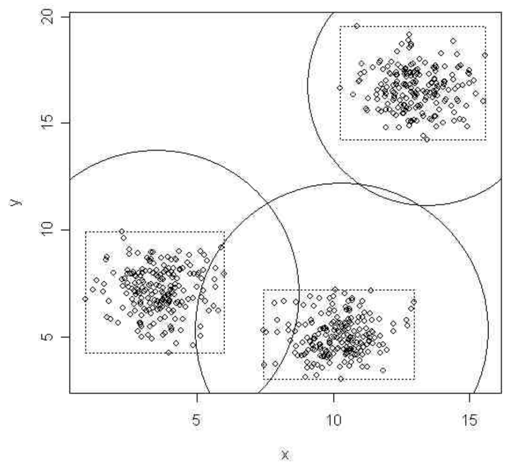
  - The new conceptual clustering does not undermine cluster quality when applied under single linkage.

Experiments

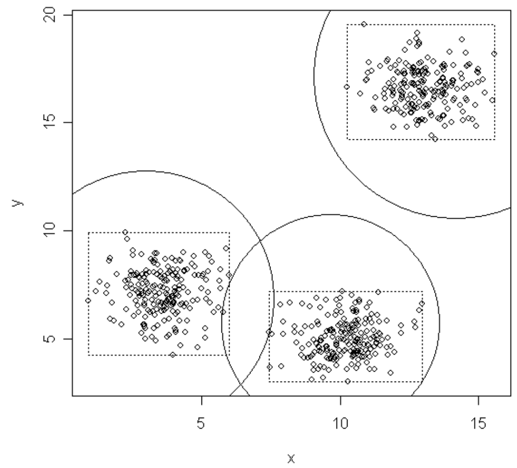
Setting B

- Compare HDCC against the traditional version of the hierarchical clustering algorithm.
  - 100 artificial datasets by drawing points from *k* (*k* = 3) Gaussian distributions in $\Re^2$. The centres are randomly located with a uniform distribution in [0,100] x [0,100].
  - Each dataset: 600 points (200 drawn from each of the 3 Gaussian distributions).
  - Four different experiments depending on the Gaussian distribution:
    - (i) av = 1 and sd in [0, 10] x [0, 10] ;
    - (ii) av = 1 and sd in[0, 200] x [0, 200];
    - (iii) av = 5 and sd in[0, 100] x [0, 100];
    - (iv) av = 5 and sd in [0, 200] x [0, 200].

Complete linkage distance



Single linkage distance



Experiments

Results

- The lower S is the better the clustering quality is.

	Trad. (i)	**Conc. (i)**	**Trad. (ii)**	**Conc. (ii)**	**Trad. (iii)**	**Conc. (iii)**	**Trad. (iv)**	**Conc. (iv)**
single	524,820	514,417	282,605	282,605	1830,421	1851,406	1607,842	1595,194
comp	285,622	285,622	282,605	282,605	1401,350	1401,350	1410,499	1410,499

- There is no difference in clustering quality.

Conclusions and Future Work

- Hierarchical distance-based conceptual clustering provides an integration of hierarchical distance-based clustering and conceptual clustering.
  - New graphical representation (conceptual dendrogram).
- Generally, for complex datatypes (sequences, graphs, etc.), HDDC builds different dendrograms.
  - Some pairs of distances and generalisation operators are compatible at some degree resulting in equivalent, order-preserving or acceptable conceptual dendrograms

Conclusions and Future Work

- With propositional data, and using the most common distances and generalisation operators the strongest properties hold
  - Integration of hierarchical distance-based clustering and conceptual clustering for propositional data is feasible.
  - Composability of tuples with complete linkage distance

- Our future work is focussed on finding operative pairs of distances and generalisation operators for common datatypes (graphs, sequences, etc..)

- Thanks for your attention!
- Questions?