



Generalization of Hierarchical Crisp Clustering Algorithms to Fuzzy Logic

Mathias Bank
mathias.bank@uni-ulm.de

Faculty for Mathematics and Economics
University of Ulm

Dr. Friedhelm Schwenker
friedhelm.schwenker@uni-ulm.de

Institute of Neural Information Processing
University of Ulm

Fuzzy Logic

Question:

Is there a salami sandwich in the refrigerator?

Answer:

0.5

If probability:

then there is or isn't, with probability one half

If measure:

then there is half a salami sandwich there

If fuzzy:

then there is something there, but it isn't really a salami sandwich. Perhaps it is some other kind of sandwich, or salami without the bread...

Motivation for Hierarchical Fuzzy Clustering

Fuzzy

Documents deal with more than one topic → overlapping topic categories

Hierarchical

Topic Analysis on different abstraction levels

Precondition for different Clustering Approach

Hierarchical agglomerative Clustering

Deterministic

Global Optimum (based on similarity)

Different Linkage Metrics available



Hierarchical agglomerative Fuzzy-Clustering

Deterministic

Global Optimum (based on similarity)

Different Linkage Metrics reusable

Overlapping topics per document → Multi-Assignment

Generalized agglomerative Clustering Algorithm

	D1	D2	D3	D4	D5	D6
D1	1	S(2,1)	S(3,1)	S(4,1)	S(5,1)	S(6,1)
D2		1	S(3,2)	S(4,2)	S(5,2)	S(6,2)
D3			1	S(4,3)	S(5,3)	S(6,3)
D4				1	S(5,4)	S(6,4)
D5					1	S(6,5)
D6						1

Group Clusters C_i and C_j with highest similarity

Generalized agglomerative Clustering Algorithm

	D1	D2	D3	D4	D5	D6
D1		$S(2,1)$	$S(3,1)$	$S(4,1)$	$S(5,1)$	$S(6,1)$
D2			$S(3,2)$	$S(4,2)$	$S(5,2)$	$S(6,2)$
D3			1	$S(4,3)$	$S(5,3)$	$S(6,3)$
D4				1	$S(5,4)$	$S(6,4)$
D5					1	$S(6,5)$
D6						1

Group Clusters C_i and C_j with highest similarity

Delete all similarity values of C_i and C_j

Generalized agglomerative Clustering Algorithm

	D1	D2	D3	D4	D5	D6
D1	1	0	S(3,1)	S(4,1)	S(5,1)	S(6,1)
D2		1	S(3,2)	S(4,2)	S(5,2)	S(6,2)
D3			1	S(4,3)	S(5,3)	S(6,3)
D4				1	S(5,4)	S(6,4)
D5					1	S(6,5)
D6						1

Group Clusters C_i and C_j with highest similarity

~~Delete all similarity values of C_i and C_j~~

Delete similarity value of C_i and C_j

Generalized agglomerative Clustering Algorithm

	D1	D2	D3	D4	D5	D6	C1
D1	1	0	S(3,1)	S(4,1)	S(5,1)	S(6,1)	0
D2		1	S(3,2)	S(4,2)	S(5,2)	S(6,2)	0
D3			1	S(4,3)	S(5,3)	S(6,3)	S(C1,3)
D4				1	S(5,4)	S(6,4)	S(C1,4)
D5					1	S(6,5)	S(C1,5)
D6						1	S(C1,6)
C1							1

Group Clusters C_i and C_j with highest similarity

~~Delete all similarity values of C_i and C_j~~

Delete similarity value of C_i and C_j

Insert new cluster $C_i \cup C_j$ into similarity matrix

Linkage metrics

Cluster – Document Similarity

Well known Linkage metrics reusable

Single Linkage, Complete Linkage, ...

Except to documents $d_i \in C_j$:

$$S(d_i, C_j) = 0$$

Subgraph property

Cluster – Cluster Similarity

Well known Linkage metrics reusable

Single Linkage, Complete Linkage, ...

Except to clusters $C_i \in C_j$:

$$S(C_i, C_j) = 0$$

Subgraph property

How to calculate $S(C_i, C_j)$, if $C_i \cap C_j \neq \emptyset$?

Use common documents for similarity calculation?

Linkage metrics

Cluster – Document Similarity

Well known Linkage metrics reusable

Single Linkage, Complete Linkage, ...

Except to documents $d_i \in C_j$:

$$S(d_i, C_j) = 0$$

Subgraph property

Cluster – Cluster Similarity

Well known Linkage metrics reusable

Single Linkage, Complete Linkage, ...

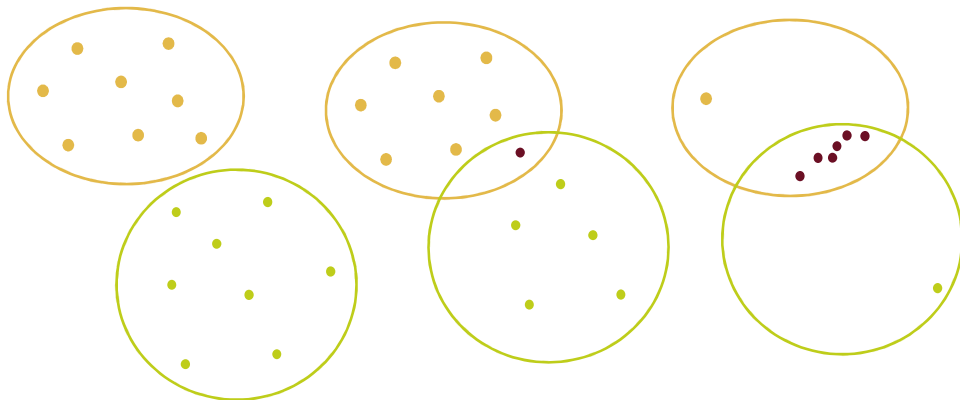
Except to clusters $C_i \in C_j$:

$$S(C_i, C_j) = 0$$

Subgraph property

How to calculate $S(C_i, C_j)$, if $C_i \cap C_j \neq \emptyset$?

Use common documents for similarity calculation?



Linkage metrics

Cluster – Document Similarity

Well known Linkage metrics reusable

Single Linkage, Complete Linkage, ...

Except to documents $d_i \in C_j$:

$$S(d_i, C_j) = 0$$

Subgraph property

Cluster – Cluster Similarity

Well known Linkage metrics reusable

Single Linkage, Complete Linkage, ...

Except to clusters $C_i \subset C_j$:

$$S(C_i, C_j) = 0$$

Subgraph property

How to calculate $S(C_i, C_j)$, if $C_i \cap C_j \neq \emptyset$?

Generalized subgraph property:

don't use common data points for similarity calculation.

Fuzzyness specification

	D1	D2	D3	D4	D5	D6
D1	1	0	f^* $S(3,1)$	f^* $S(4,1)$	f^* $S(5,1)$	f^* $S(6,1)$
D2		1	f^* $S(3,2)$	f^* $S(4,2)$	f^* $S(5,2)$	f^* $S(6,2)$
D3			1	$S(4,3)$	$S(5,3)$	$S(6,3)$
D4				1	$S(5,4)$	$S(6,4)$
D5					1	$S(6,5)$
D6						1

Fuzzifier f specify fuzziness

$$f \in [0; 1]$$

Applied to selected rows / columns

$$\text{If } f * S(C_1, C_2) < \varepsilon \rightarrow S(C_1, C_2) = 0$$

Definable for each level separately

(level counted by document level)

→ Crisp level definition possible

Generalized agglomerative Clustering Algorithm

While (!empty(SimilarityMatrix))

Group Clusters C_i and C_j with highest similarity

~~Delete all similarity values of C_i and C_j~~

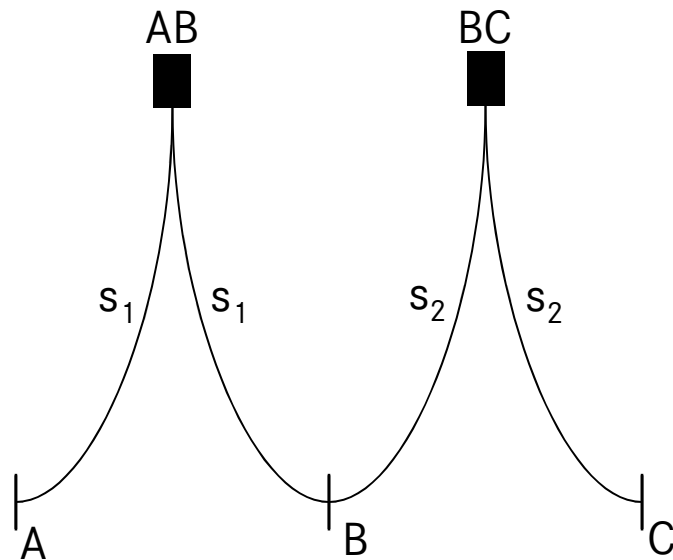
Delete similarity value of C_i and C_j

Apply fuzzifier f to similarity values fo C_i and C_j

Insert new cluster $C_i \cup C_j$ into similarity matrix

End

Fuzzy Membership Calculation



$$\mu_{A,AB} = \frac{S_1}{S_1}$$

$$\mu_{B,AB} = \frac{S_1}{S_1 + S_2}$$

$$\mu_{C_i, C_j} = \frac{S(C_i, C_j)}{\sum_{C_l=p(C_j)} S(C_i, C_l)}$$

p: Parents

Generating Topic Groups

Hierarchy size

Binary merging process

different fuzziness

→ Different hierarchy levels

→ No adequate topic representation

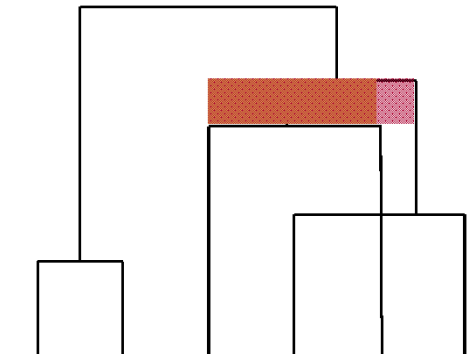


Shrinking process

Set h maximum distance between all nodes in the directed cluster graph

Merge cluster levels with smallest similarity difference until h reaches a predefined level

Interpretable hierarchy size $h \in [5,10]$ possible



Generalized agglomerative Clustering Algorithm

While (!empty(SimilarityMatrix))

Group Clusters C_i and C_j with highest similarity

~~Delete all similarity values of C_i and C_j~~

Delete similarity value of C_i and C_j

Apply fuzzifier f to similarity values fo C_i and C_j

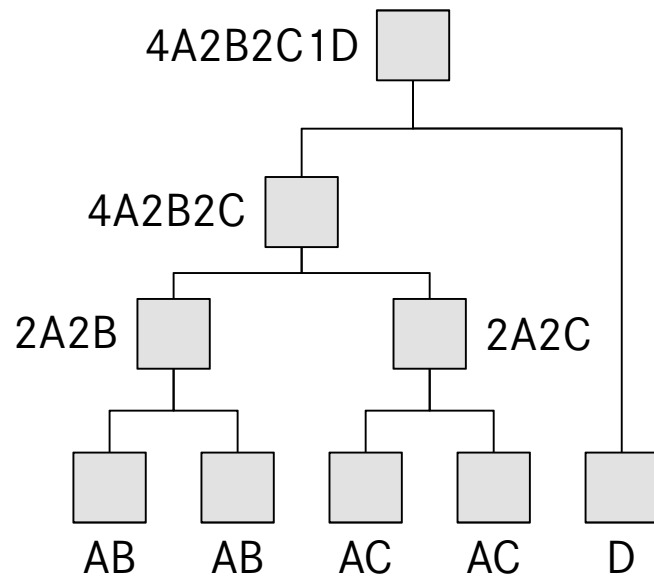
Insert new cluster $C_i \cup C_j$ into similarity matrix

End

Calculate Fuzzy Membership $\mu_{ij} \forall i, j$

Shrink cluster level size h to predefined level size

Evaluation measure for external labeled data



COS-Quality-Index

Propagate document labels via membership

Cluster quality defined as statistics of all cos similarities of document label vector and cluster label vector

Fuzziness Statistics

Recall in fuzzy mode difficult to define

➔ Fuzziness analysis for recall information

Data

7 randomly chosen document collections

Based on..

RCV1 (26501, 23309, 10148 docs)

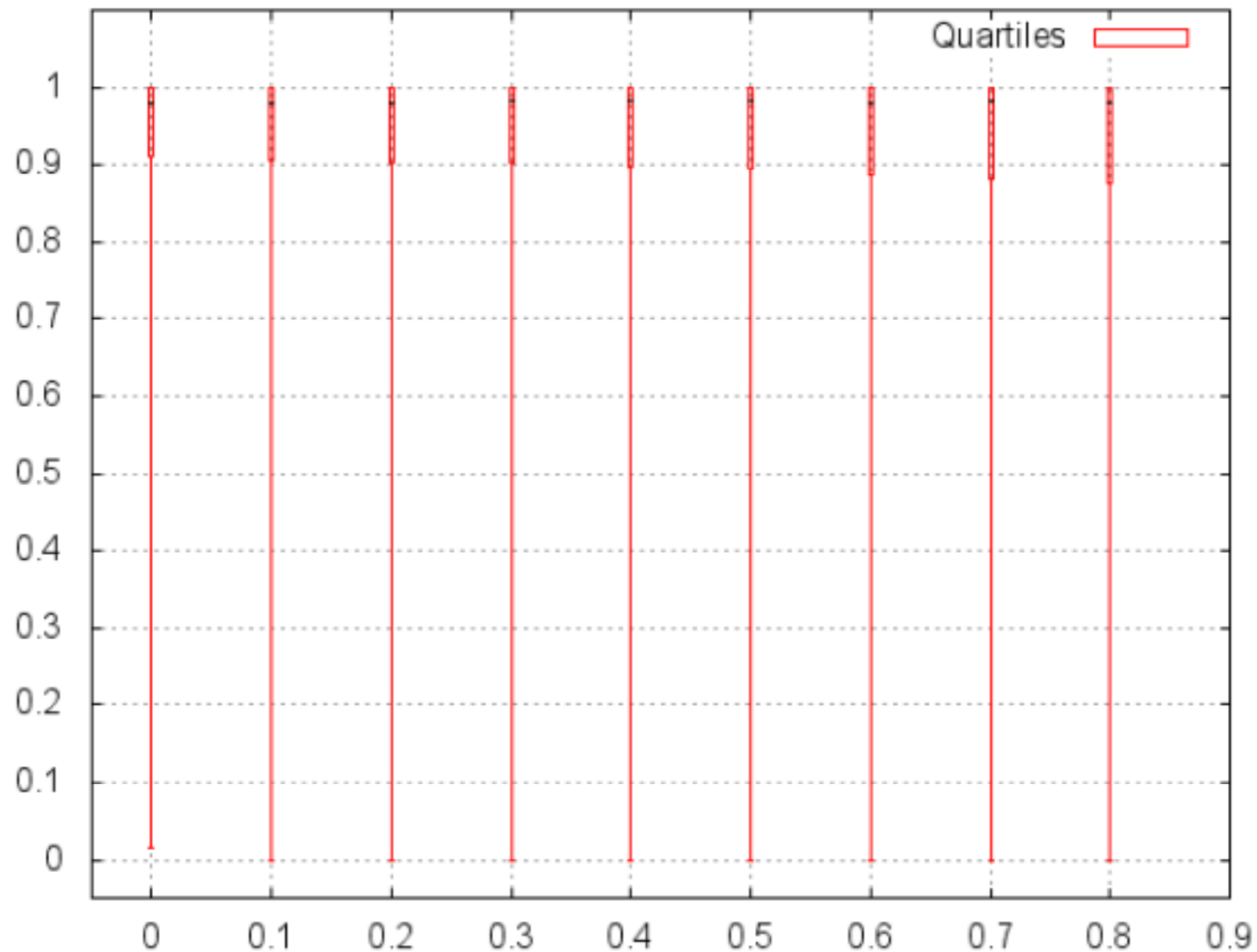
RCV2-german (16604, 13146, 11600, 9350 docs)

Fuzzyfier $f \in \{0.0, 0.1, \dots, 0.8\}$ for document level, 0.0 else

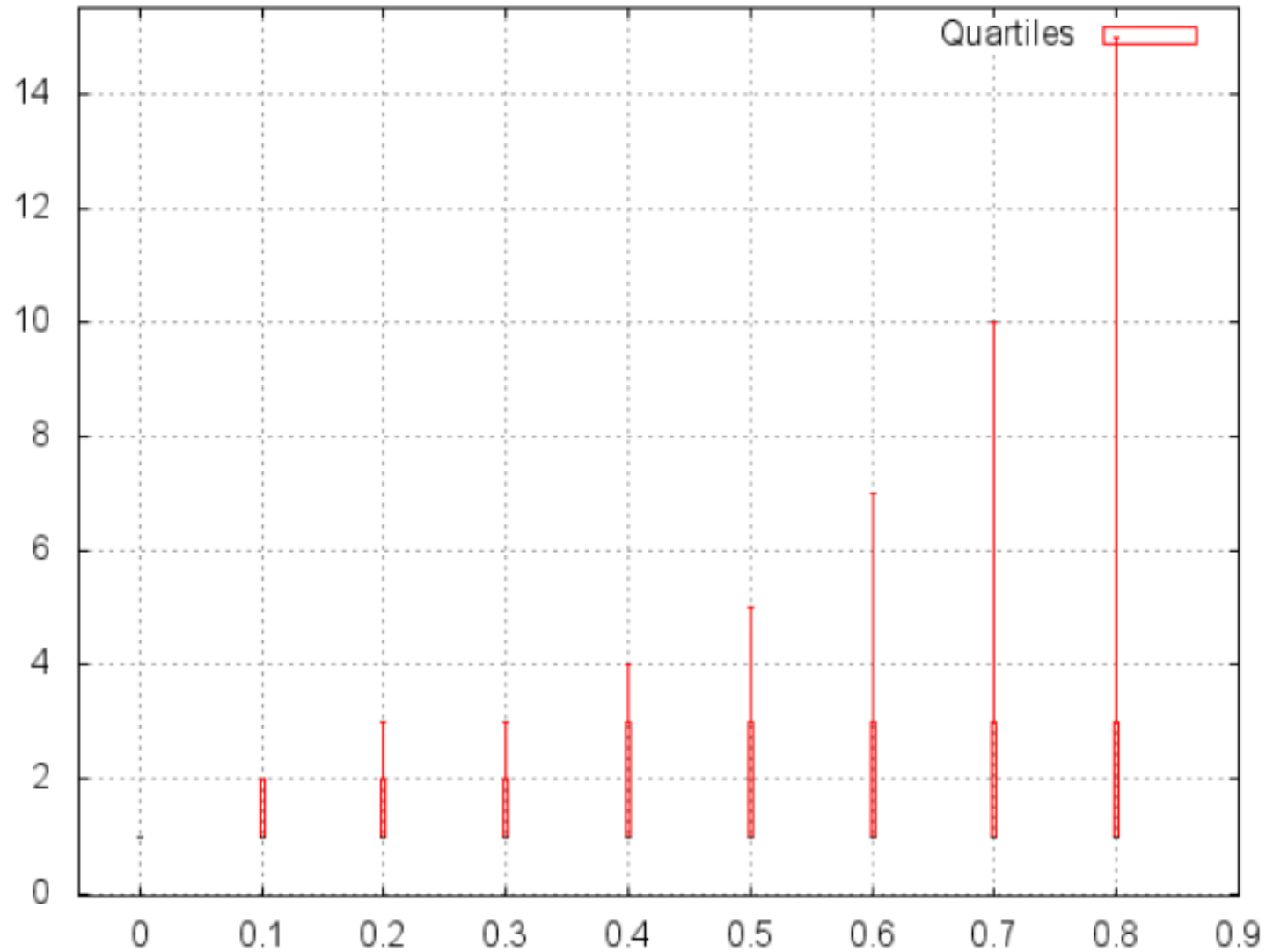
Hierarchy level $h = 5$

➔ Quality analysis for document parents necessary

Evaluation Results – COS-Quality for document parents



Evaluation Results – Fuzziness per document



Conclusion

Agglomerative Fuzzy Cluster Algorithm

Agglomerative Crisp Cluster algorithm generalizable to Fuzzy Logic

→ Deterministic hierarchical fuzzy cluster algorithm

→ Global optimal cluster based on similarity function

Topic Groups per level shrinkage possible (crisp & fuzzy)

Quality Results

Fuzziness does not highly influence cluster quality

25% - 50% of random chosen textual data affected by multi assignment

Discussion