Dottorato di Ricerca in Ingegneria dell'Informazione

Data Mining and Soft Computing

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI^{2}S)
Dept. of Computer Science and A.I.
University of Granada, Spain

Email: herrera@decsai.ugr.es
http://sci2s.ugr.es

http://decsai.ugr.es/~herrera

DECSAI
Universidad de Granada

Summary

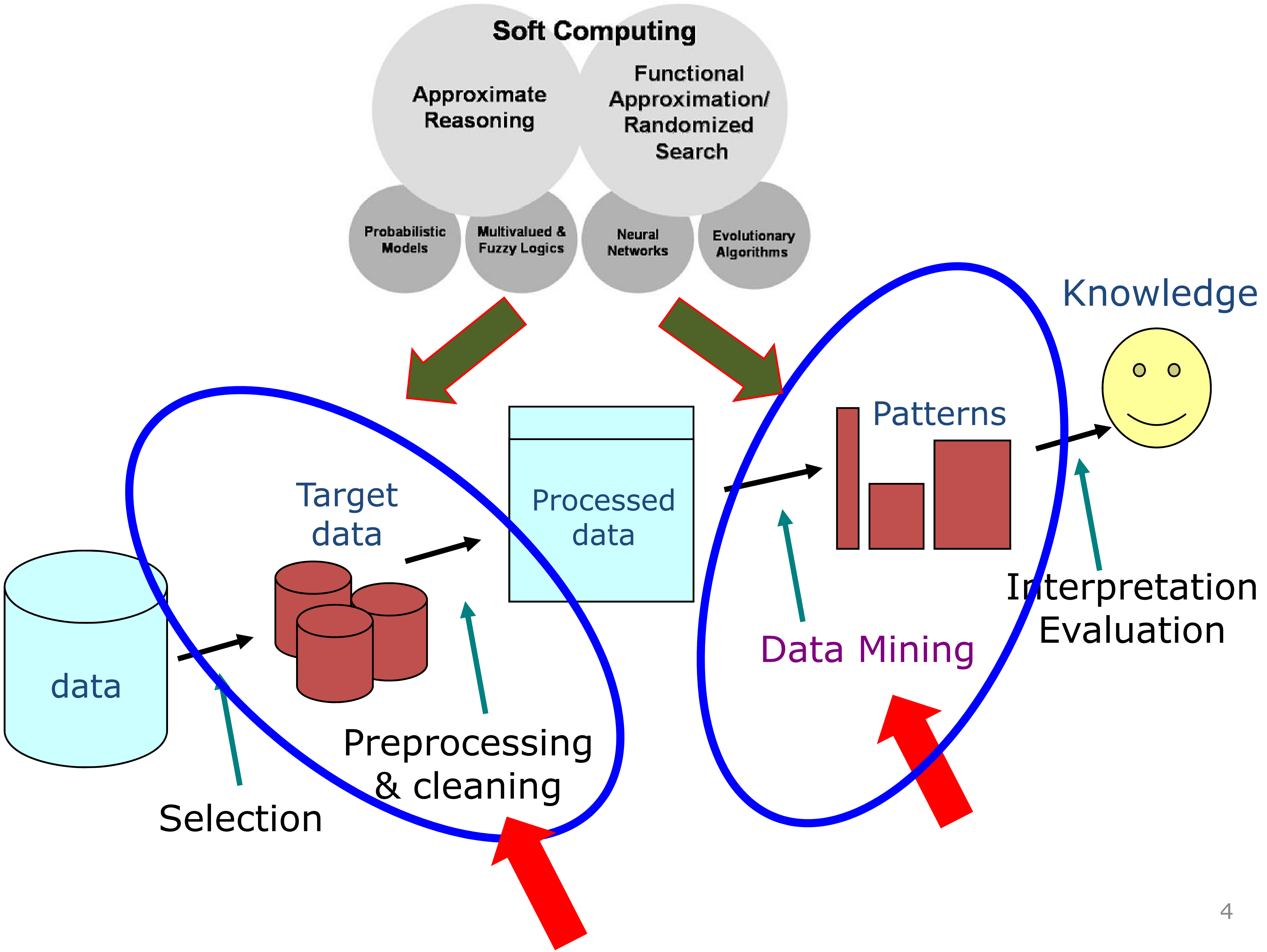
1. **Introduction to Data Mining and Knowledge Discovery**
2. **Data Preparation**
3. **Introduction to Prediction, Classification, Clustering and Association**
4. **Data Mining - From the Top 10 Algorithms to the New Challenges**
5. **Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation**
6. **Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning**
7. **Genetic Fuzzy Systems: State of the Art and New Trends**
8. **Some Advanced Topics I: Classification with Imbalanced Data Sets**
9. **Some Advanced Topics II: Subgroup Discovery**
10. **Some advanced Topics III: Data Complexity**
11. **Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.**

Outline

- ✓ Introduction: Soft Computing Techniques in Data Mining
- ✓ Fuzzy Data Mining
- ✓ Evolutionary Data Mining
- ✓ Concluding Remarks

Outline

- ✓ Introduction: Soft Computing Techniques in Data Mining
- ✓ Fuzzy Data Mining
- ✓ Evolutionary Data Mining
- ✓ Concluding Remarks



Soft Computing
Approximate Reasoning
Functional Approximation/ Randomized Search
Probabilistic Models
Multivalued & Fuzzy Logics
Neural Networks
Evolutionary Algorithms
Knowledge
Patterns
Target data
Processed data
data
Data Mining
Interpretation Evaluation
Preprocessing & cleaning
Selection

Role of Fuzzy Sets

— Modeling of imprecise/qualitative knowledge

—Transmission and handling uncertainties at various stages

— Supporting, to an extent, human type reasoning in natural form

Fuzzy Sets in Data Mining

- Classification/Regression/ Clustering
- Discovering association rules (describing interesting association relationship among different attributes)
- Data summarization/condensation (abstracting the essence from a large amount of information).

Role of ANN

- Adaptivity, robustness, parallelism, optimality
- Machinery for learning and curve fitting (Learns from examples)

➢ Initially, thought to be unsuitable for — black box nature — no information available in symbolic form (suitable for human interpretation)

➢ Recently, embedded knowledge is extracted in the form of symbolic rules ⇨ making it suitable for *Rule generation*.

ANNs provide Natural Classifiers/prediction based models having

- *Resistance* to Noise,
- *Tolerance* to Distorted Patterns /Images (Ability to Generalize)
- *Superior Ability* to Recognize Overlapping Pattern Classes or Classes with Highly Nonlinear Boundaries or Partially Occluded or Degraded Images
- *Potential* for Parallel Processing
- Non parametric

Role of Genetic Algorithms

- ❖ Robust, parallel, adaptive search methods — suitable when the search space is large.

- ❖ GAs : Efficient, Adaptive and robust Search Processes, Producing near optimal solutions and have a large amount of Implicit Parallelism

- ❖ GAs are *Appropriate* and *Natural Choice* for problems which need – Optimizing Computation Requirements, and Robust, Fast and Close Approximate Solutions

GAs in Data Mining

- Many tasks involved in analyzing/identifying a pattern need *Appropriate Parameter Selection* and *Efficient Search* in complex spaces to obtain Optimal Solutions

Why Soft Computing in Data Mining?

Relevance of FL, ANN, GAs **Individually** to Data Mining problems is established.

The hybrid methods provide a more power tool for data mining incorporating representation, learning and optimization features in the data mining model.

Outline

- ✓Introduction: Soft Computing Techniques in Data Mining
- ✓Fuzzy Data Mining
- ✓Evolutionary Data Mining
- ✓Concluding Remarks

Relevance of Fuzzy Sets in DM

- Representing linguistically phrased input features for processing
- Representing multi-class membership of ambiguous patterns
- Generating rules & inferences in linguistic form
- Extracting ill-defined image regions, primitives, properties and describing relations among them as fuzzy subsets

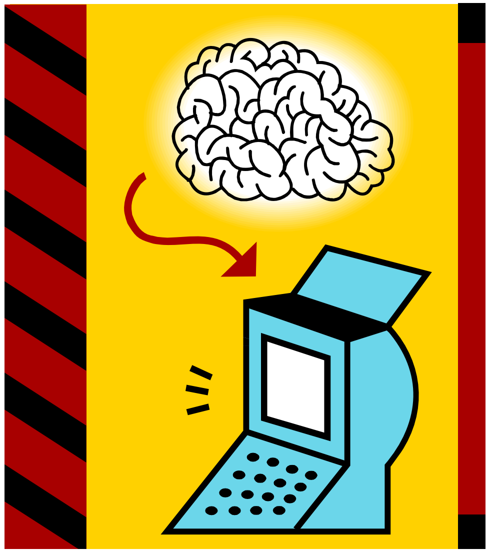
Fuzzy Data Mining

- Fuzzy set theory found in almost every area of data mining
- Fuzzy representation very appropriate
  – Humans perceive a great deal of uncertainty
- One approach:
  – Partition data into categories to create fuzzy grids

FUZZIFYING DATA DOES NOT SEEM TO MAKE LESS ACCURATE

Data Mining Use of Fuzzy Sets

- Fuzzy Rule Based Systems
  - Regression
  - Pattern Classification
- Fuzzy Clustering
- Fuzzy Association Rules



Fuzzy rule based systems

- Fuzzy inference systems based on
  - Fuzzy set theory
  - Fuzzy-if-then rules
  - Fuzzy reasoning
- Application
  - Automatic control
  - Data classification
  - Decision analysis
  - Expert systems
  - Time series prediction
  - Robotics
  - Pattern recognition
- Fuzzy inference system is known as
  - Fuzzy-rule-based system
  - Fuzzy expert system
  - Fuzzy model
  - Fuzzy associative memory
  - Fuzzy logic controller
  - Fuzzy system
- Components
  - A rule base
  - Database (dictionary)
  - A reasoning mechanism

Fuzzy rule based systems: Regression

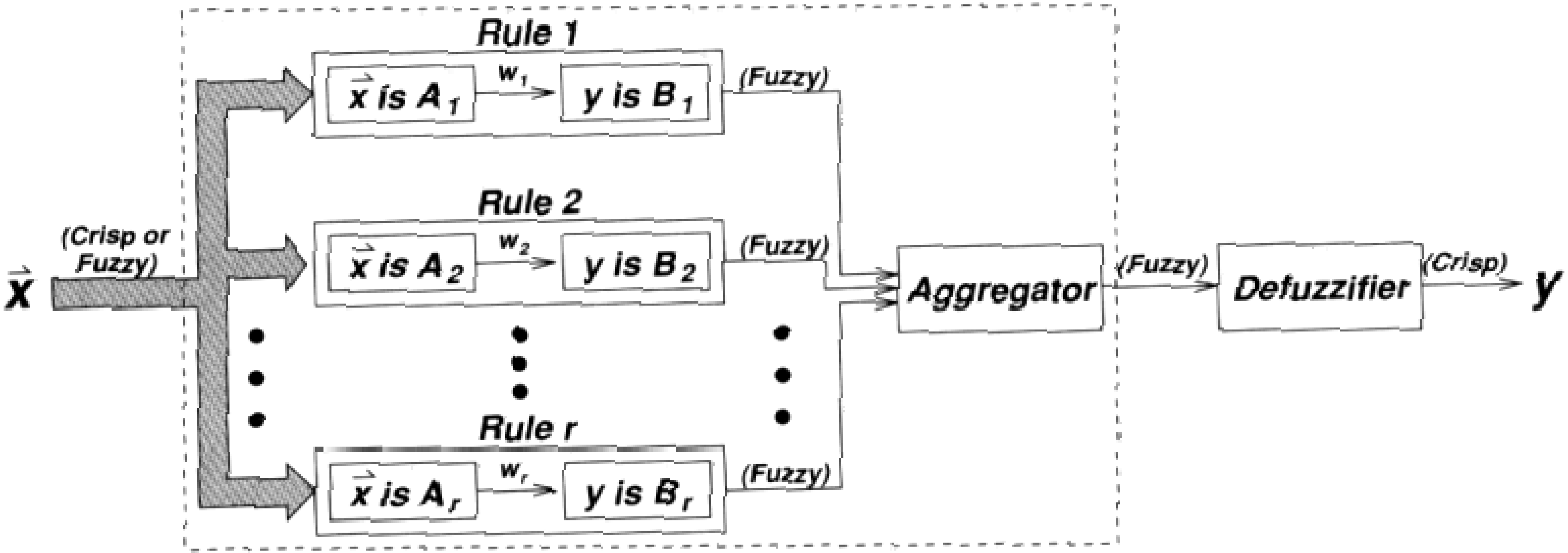
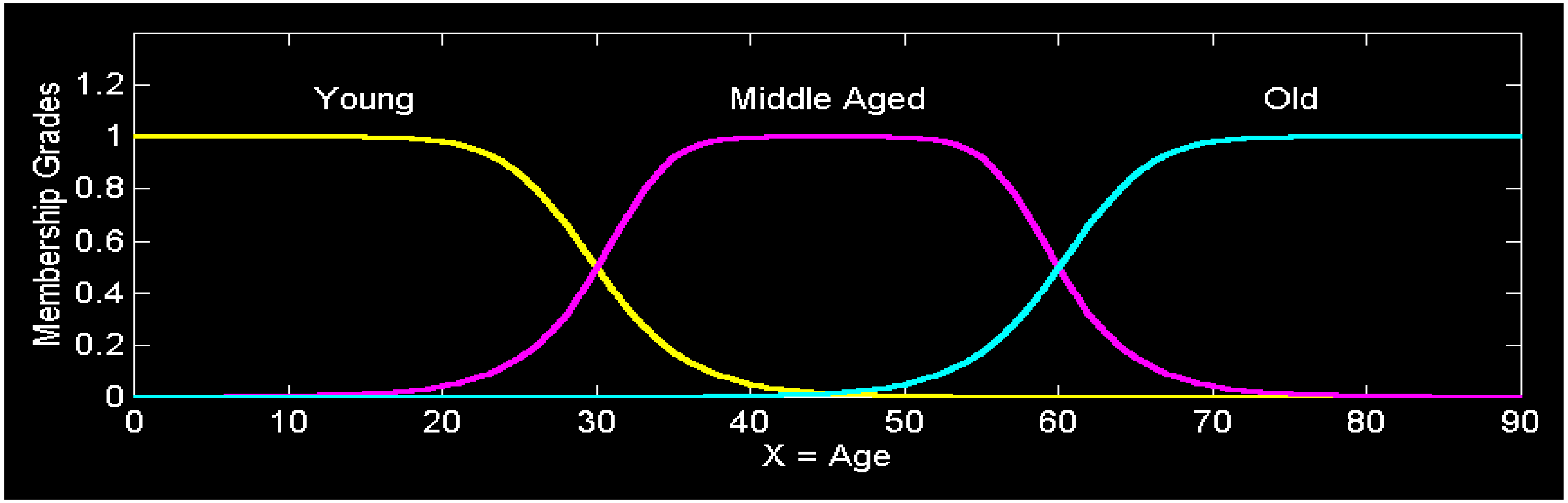


Figure 4.1. *Block diagram for a fuzzy inference system.*

Linguistic variables

L. Variable = Age
terms, fuzzy sets : { young, middle aged, old}



Fuzzy rules

Commonsense knowledge may sometimes be captured in an natural way using fuzzy rules.

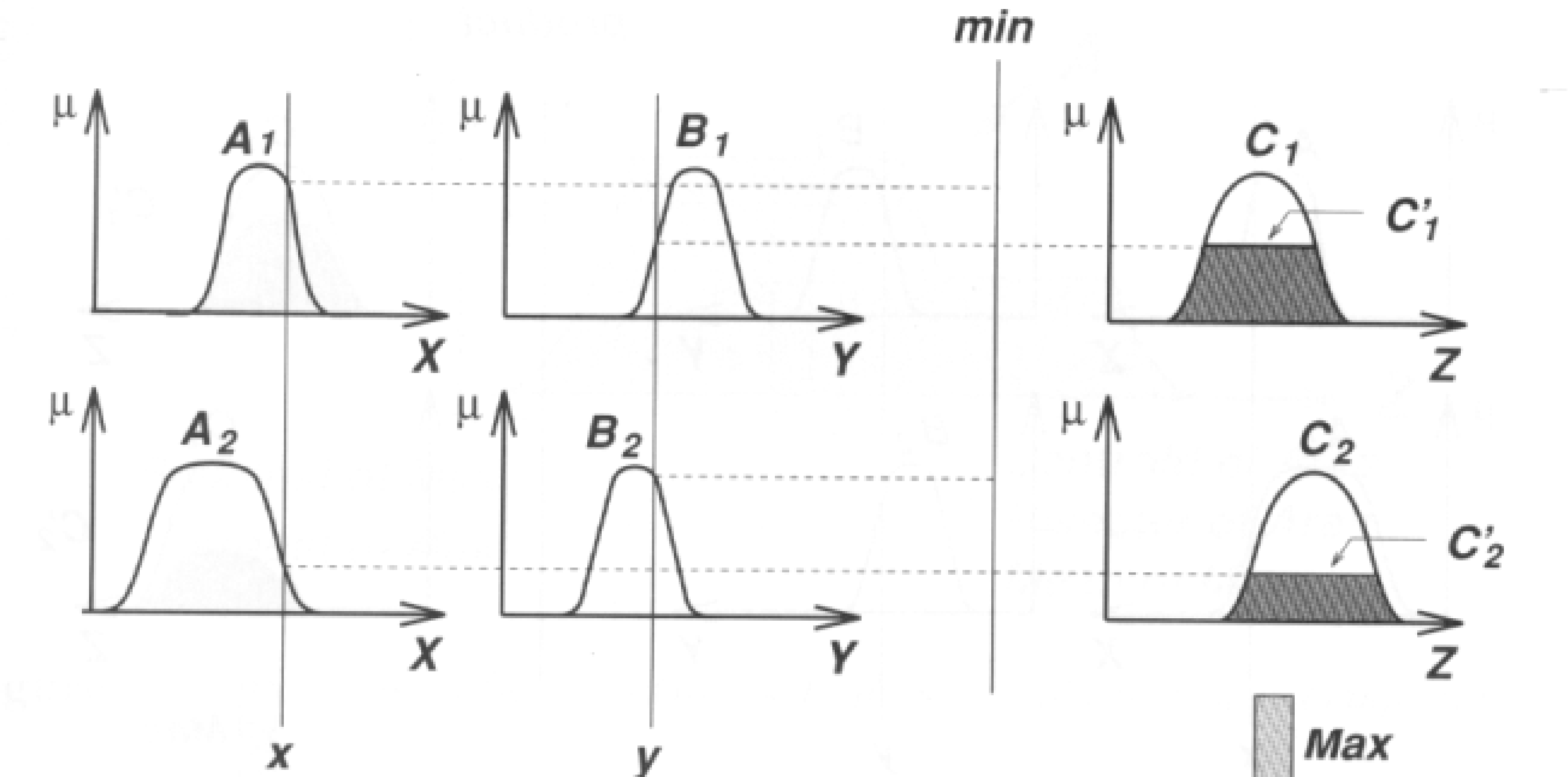
IF X_1 is high and X_2 is good
THEN Y is small

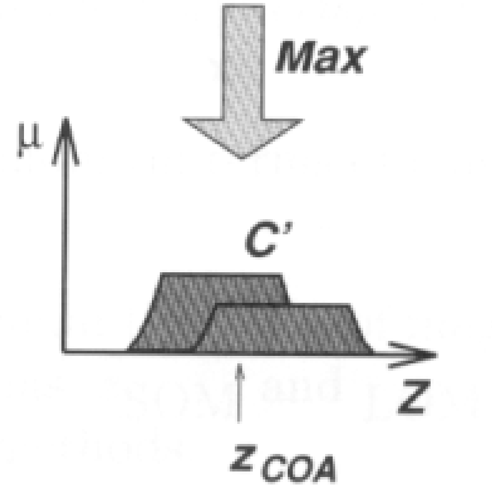
What does it mean for fuzzy rules:

IF *x* is A then *y* is B ?

Fuzzy implications provide us the meaning of the rule

Mamdani Fuzzy Model





A two-rule Mamdani fuzzy inference system

Derive the overall output z when subjected to two crisp inputs x and y

T-norm operator: Min

T-conorm operator: Max

Example: Rules base

Heating Price	freezing	Temperature cold	chilly
cheap	full	full	medium
so-so	full	medium	weak
expensive	medium	weak	no

IF Temperatura=chilly and Heating-price=expensive THEN heating=no

IF Temperature=freezing and Heating-price=cheap THEN heating=full

1. Fuzzification

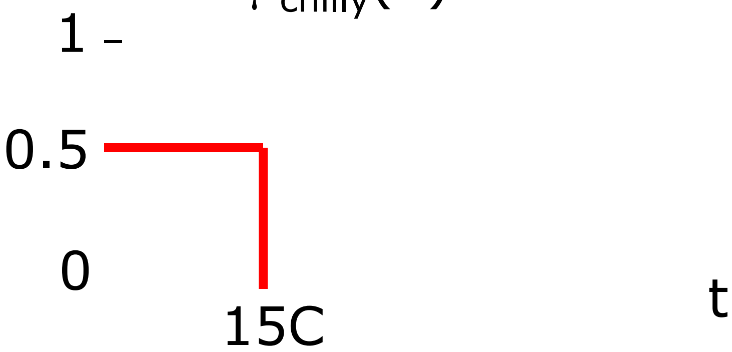
Fuzzification: from measured values to MF:

Determine membership degrees for all fuzzy sets (linguistic variables):

Temperature: T=15 C

Heating-price: p=48 Euro/MBtu

$\mu_{chilly}(T)=0.5$



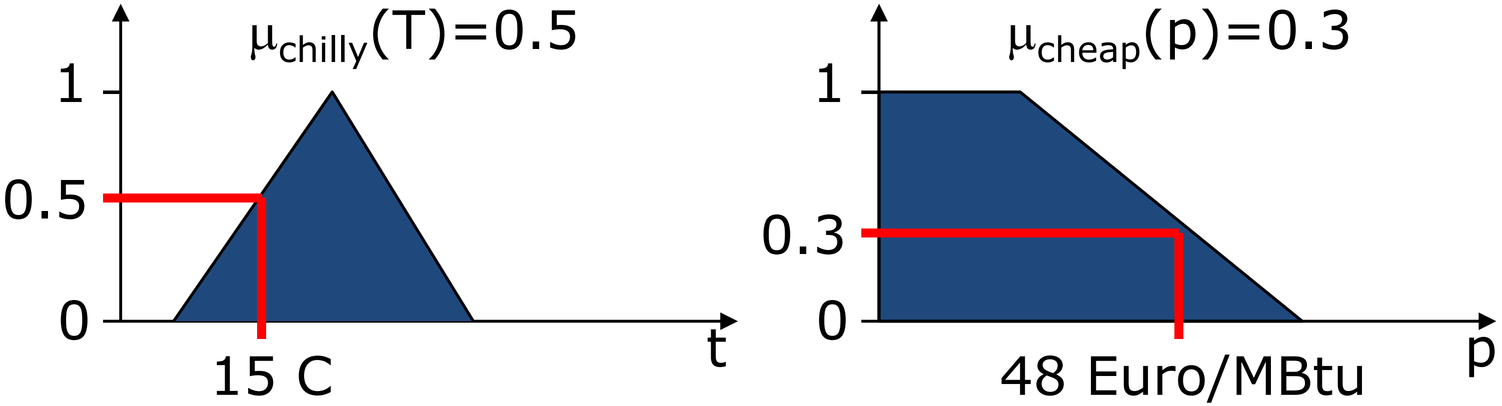
$\mu_{cheap}(p)=0.3$

IF Temperature = chilly

and Heating-price = cheap...

2. Matching degree

Calculate the degree of rule fulfillment for all conditions combining terms using fuzzy AND, ex. MIN operator.



IF Temperature=chilly and Heat-price=cheap...

$\mu_A(\mathbf{X}) = \mu_{A1}(X_1) \wedge \mu_{A2}(X_2) \wedge \mu_{AN}(X_N)$ for rules R_A

$\mu_{all}(\mathbf{X}) = \min\{\mu_{chilly}(t), \mu_{cheap}(p)\} = \min\{0.5, 0.3\} = 0.3$

3. Inference

Calculate the degree of truth of rule conclusion: use T-norms such as MIN or product to combine the degree of fulfillment of conditions and the MF of conclusion.

$\mu_{full}(h)$ $\mu_{conclusions}(h)$

1

... $\mu_{cond}=0.3$

0

h

Inference MIN

$\mu_{concl}=\min\{\mu_{cond}, \mu_{full}\}$

THEN Heating=full

$\mu_{mocno}(h)$ $\mu_{konkl}(h)$

1

... $\mu_{cond}=0.3$

0

h

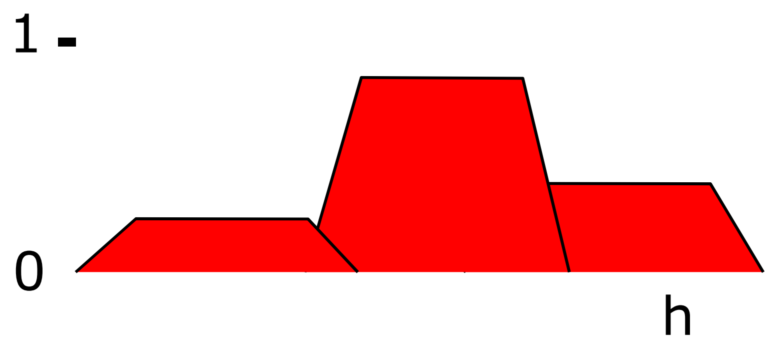
Inference

$\mu_{concl.} = \mu_{cond} \bullet \mu_{full}$

4. Aggregation

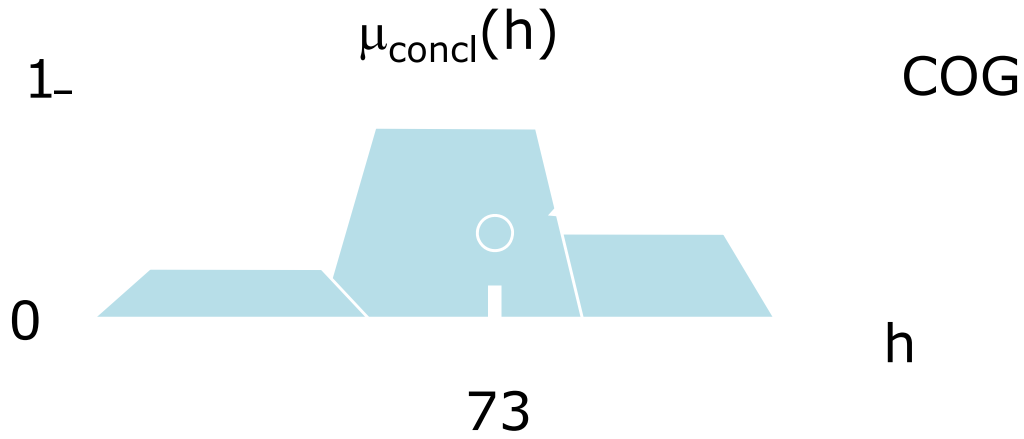
Aggregate all possible rule conclusion using MAX operator to calculate the sum.

THEN Heating=full
THEN Heating =medium
THEN Heating =no



5. Defuzzification

Calculate crisp value/decision using for example the "Center of Gravity" (COG) method:



For discrete sets a „center of singletons", for continuous:

$$h = \frac{\Sigma_i \ \mu_i \bullet A_i \bullet c_i}{\Sigma_i \ \mu_i \bullet A_i}$$

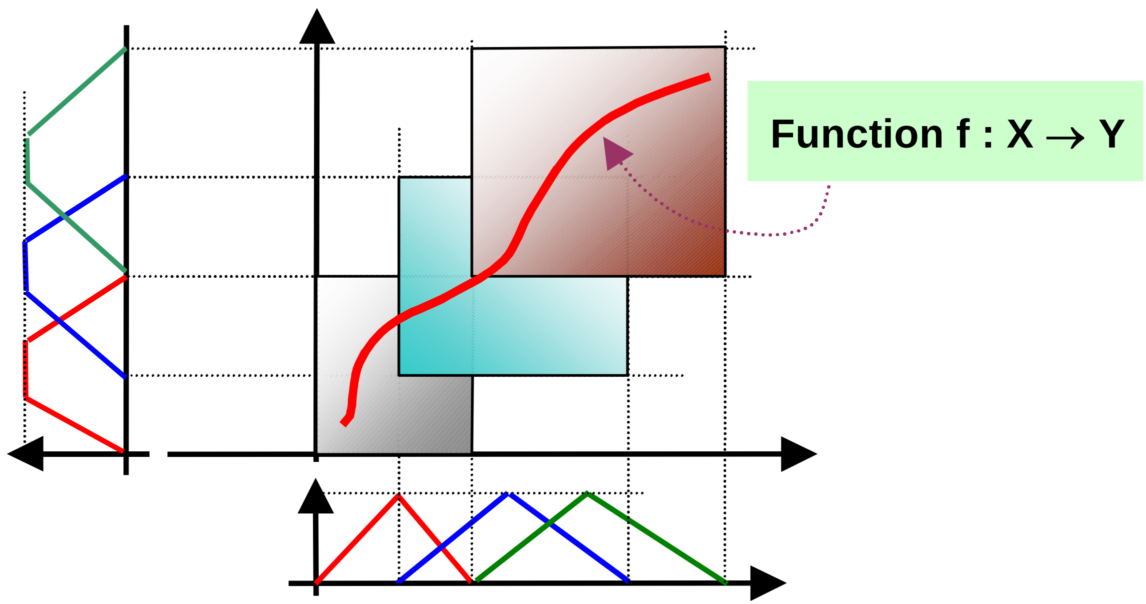
μ_i = degree of membership in *i*
A_i = area under MF for the set *i*
c_i = center of gravity for the set *i*.

Defuzzification

- To extract a crisp value from a fuzzy set as a representative value.
- Five methods for defuzzifying a fuzzy set A of a universe of discourse Z.
  - Centroid of area z_{COA}
  - Bisector of area z_{BOA}
  - Mean of maximum z_{MOM}
  - Smallest of maximum z_{SOM}(not used so often)
  - Largest of maximum z_{LOM}(not used so often)

Fuzzy approximation

- Fuzzy systems F: $\Re^n \rightarrow \Re^p$ use m rules to map vector x on the output F(x), vector or scalar.



Singleton model: R_i: IF x is A_i Then y is b_i

Sugeno Fuzzy Models

- TSK fuzzy models
  - Proposed by Takagi, Sugeno and Kang(1985, 1988)
  - if x is A and y is B then z=f(x,y),
    - A,B are fuzzy sets in the antecedent, while z=f(x,y) is a crisp function in consequence
    - f(x,y) is usually a polynomial.
  - A first-order Sugeno fuzzy model
    - f(x,y) is a first-order polynomial.
  - A zero-order Sugeno fuzzy model
    - f(x,y) is a constant.
    - A special case of the Mamdani fuzzy inference system

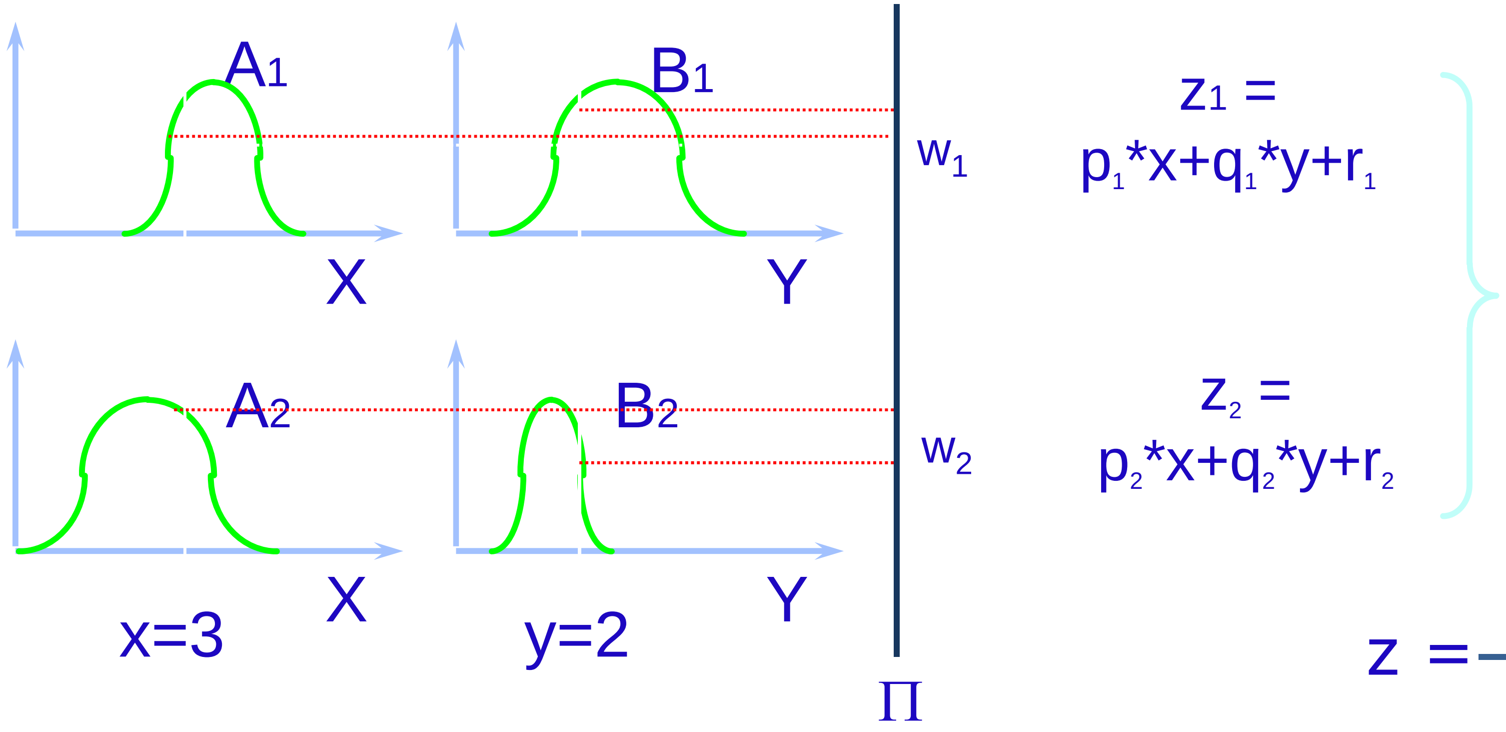
First-order TS FIS

- Rules

  IF X is A_1 and Y is B_1 then $Z = p_1*x + q_1*y + r_1$

  IF X is A_2 and Y is B_2 then $Z = p_2*x + q_2*y + r_2$

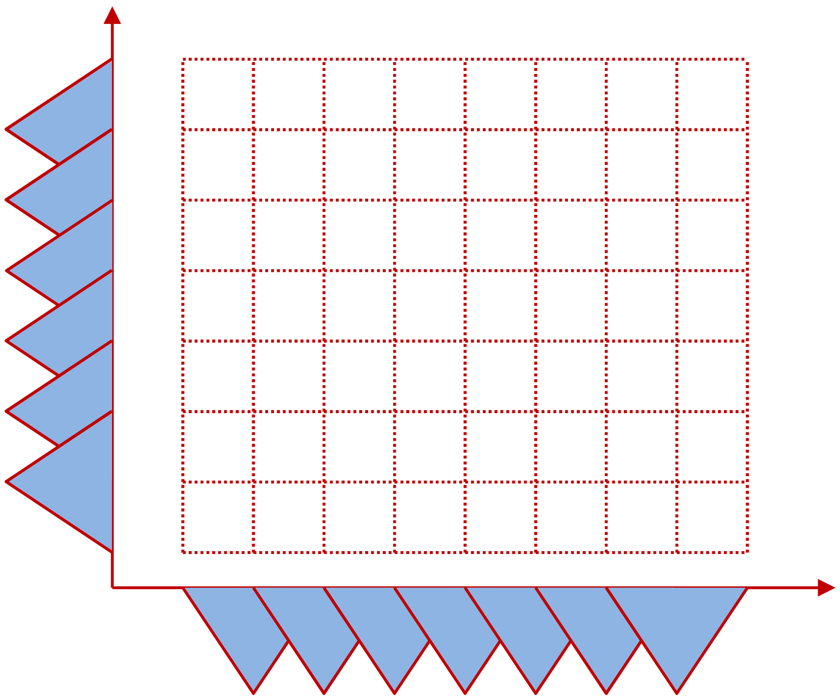
- Fuzzy inference



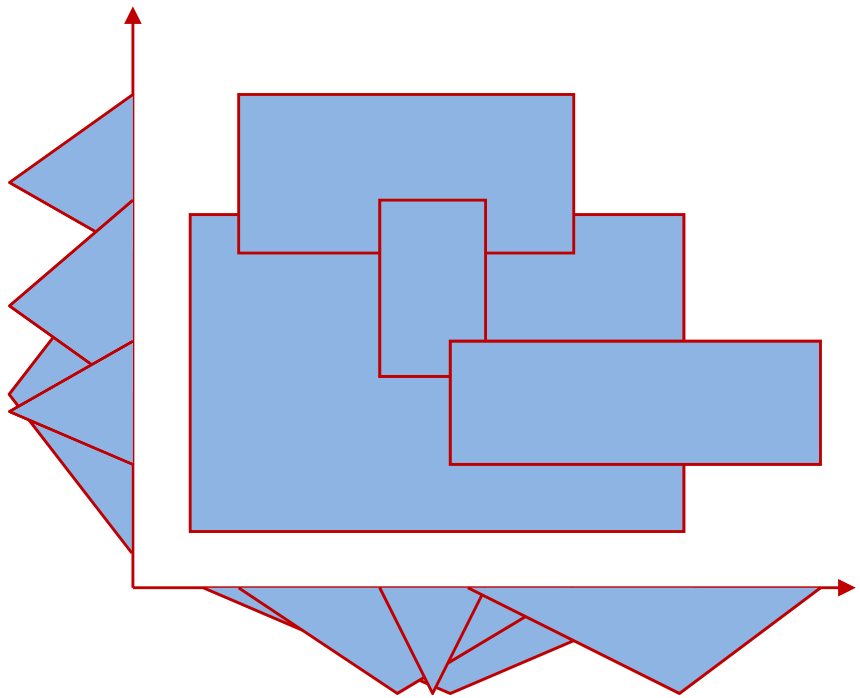
$$z = \frac{w_1*z_1+w_2*z_2}{w_1+w_2}$$

Feature space partition

Regular grid



Independent functions



MFs on a grid

- Advantage: simplest approach
- Regular grid: divide each dimension in a fixed number of MFs and assign an average value from all samples that belong to the region.
- Irregular grid: find largest error, divide the grid there in two parts adding new MF.
- Mixed method: start from regular grid, adapt parameters later.

- Disadvantages: for *k* dimensions and *N* MFs in each N^k areas are created !
  Poor quality of approximation.

Optimized MFs

- Advantages: higher accuracy, better approximation, less functions, context dependent MPs.

- Optimized MP may come from:
- Neurofuzzy systems – equivalent to RBF network with Gaussian functions (several proofs).
  FSM models with triangular or trapezoidal functions.
  Modified MLP networks with bicentral functions, etc.
- Genetic fuzzy systems
- Fuzzy machine learning inductive systems.
-
- Disadvantages: extraction of rules is hard, optimized MFs are more difficult to create.

Induction of fuzzy rules

Choices/adaptive parameters in fuzzy rules:

- The number of rules (nodes).
- The number of terms for each attribute.
- Position of the membership function (MF).
- MF shape for each attribute/term.
- Type of rules (conclusions).
- Type of inference and composition operators.
- Induction algorithms: incremental or refinement.
- Type of learning procedure.

Fuzzy rule based classification systems

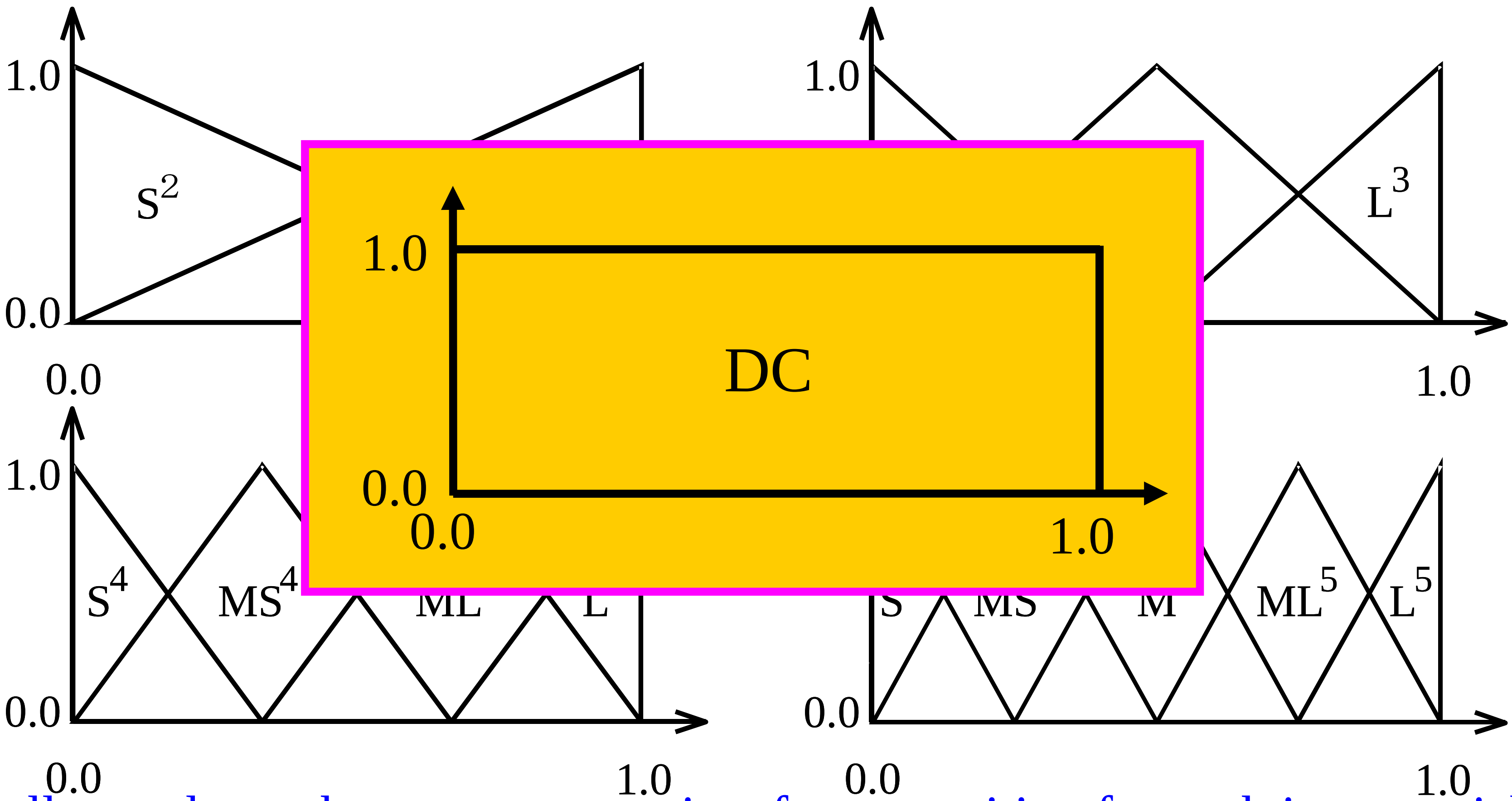
Fuzzy Rules for *n*-dimensional Problems

If x_1 is A_1 and … and x_n is A_n
then Class C with CF

A_i : Antecedent fuzzy set
Class C : Consequent class
CF : Rule weight (Certainty factor or others)

Fuzzy rule based classification systems

Antecedent Fuzzy Sets (Multiple Partitions)

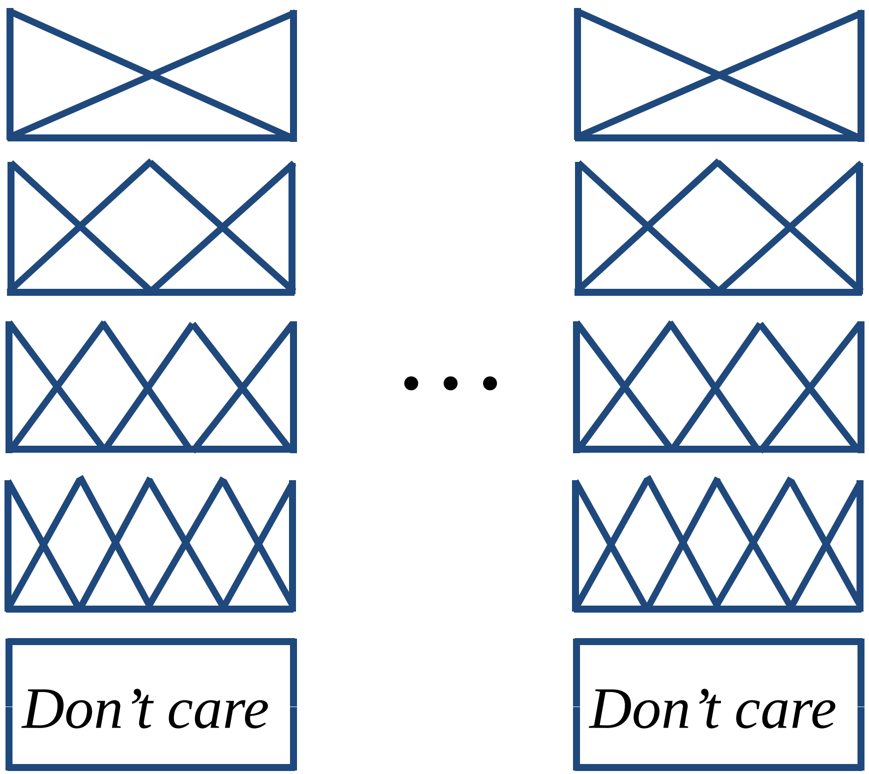


Usually we do not know an appropriate fuzzy partition for each input variable.
Usually, authors use 5 or 7 labels.

Fuzzy rule based classification systems

Possible Fuzzy Rules

Total number of possible fuzzy rules
Selection/Learning method



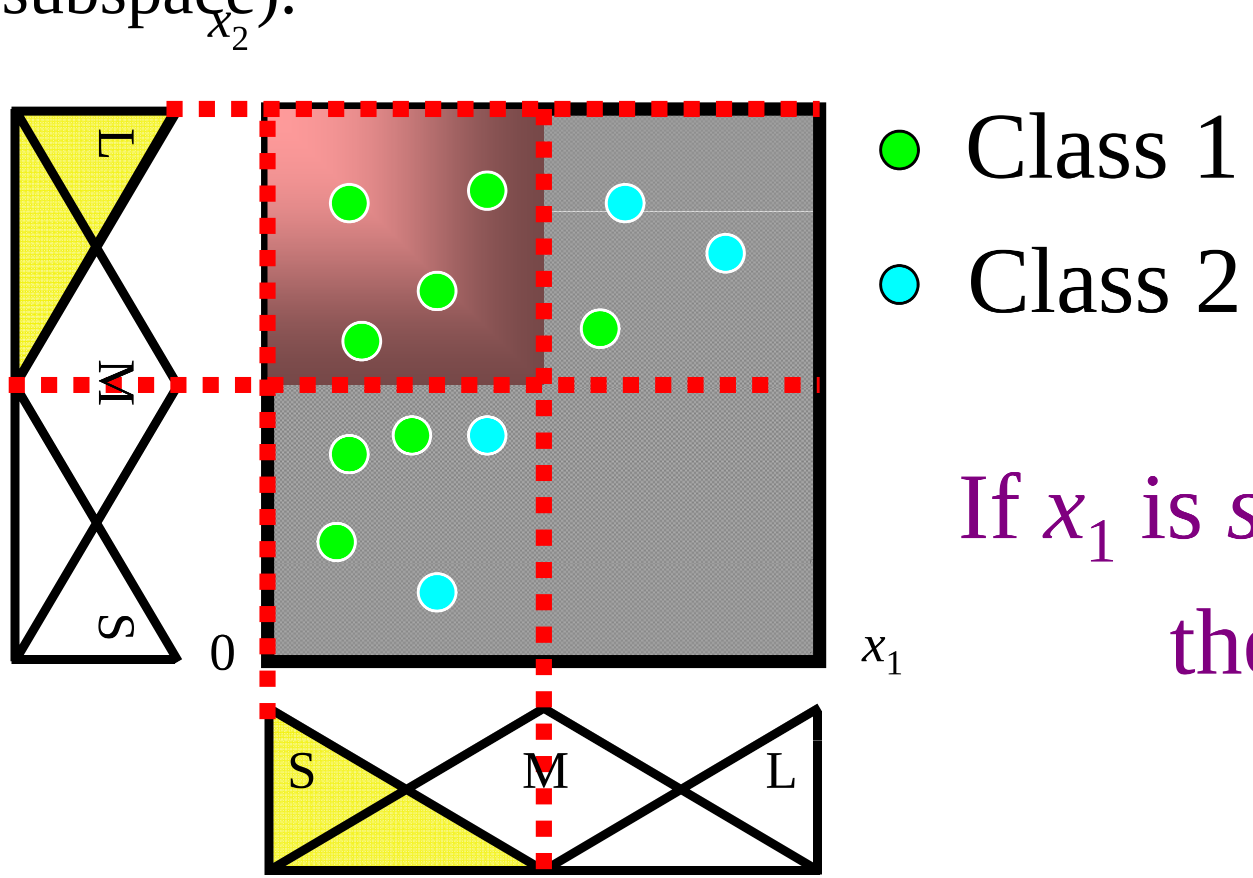
$$x_1 \qquad\qquad x_n$$

$$(14+1)\times \;\cdots\; \times(14+1)=(15)^n$$

Fuzzy rule based classification systems

Consequent Class

The consequent class of each fuzzy rule is determined by compatible training patterns (i.e., the dominant class in the corresponding fuzzy subspace).

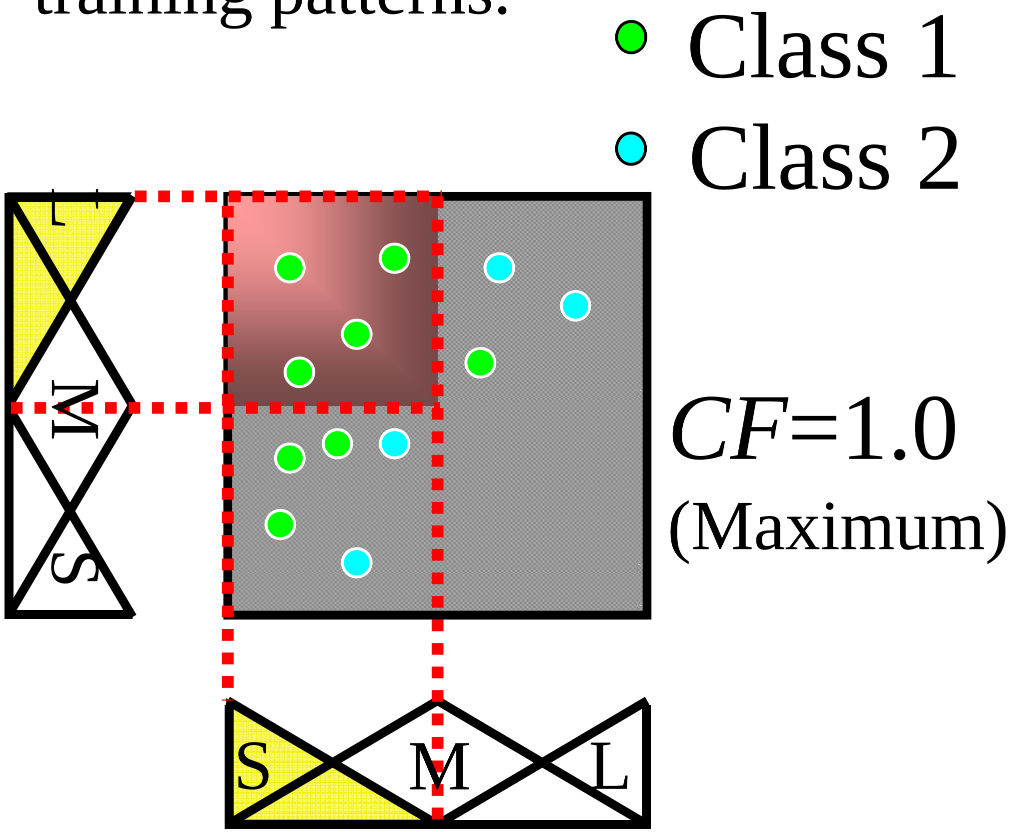


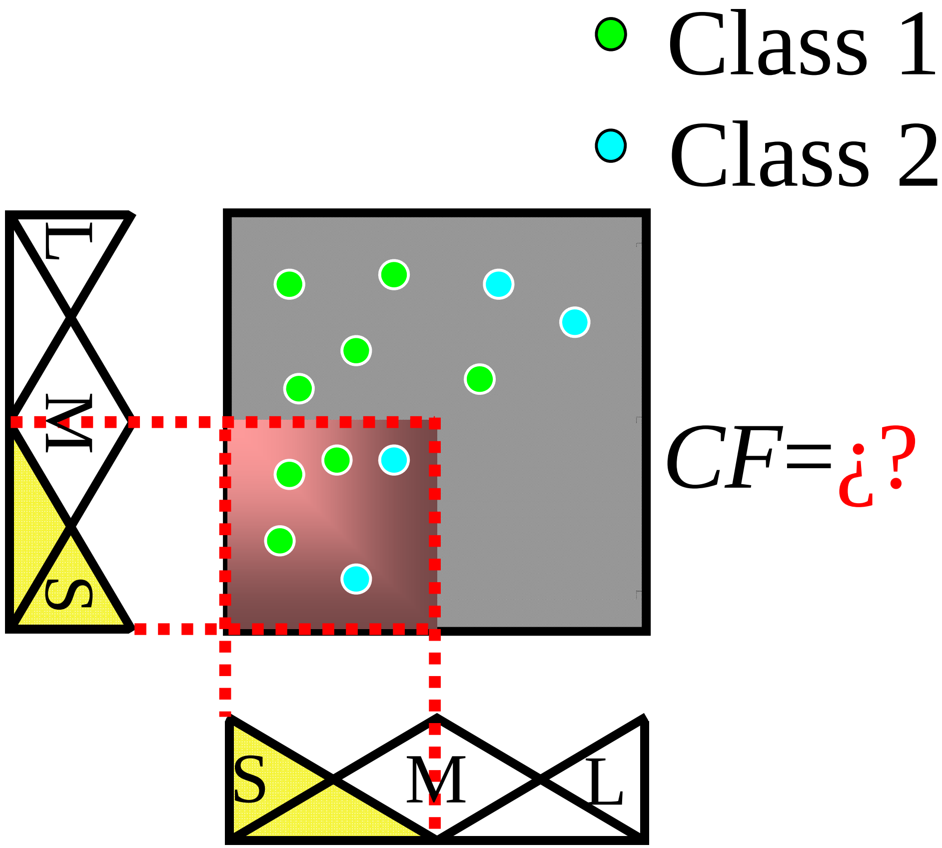
If x_1 is *small* and x_2 is *large*
then Class 1 with 1.0

Fuzzy rule based classification systems

Rule Weight (Certainty Factor)

The rule weight *CF* of each fuzzy rule is calculated from compatible training patterns.





Fuzzy rule based classification systems

Rule Weight (Certainty Factor)

R_k: If x_1 is A_1^k and ... and x_N is A_N^k then Y is C_j with r^k,

where r^k is the certainty degree of the classification in the class C_j for a pattern belonging to the fuzzy subspace delimited by the antecedent. This certainty degree can be determined by the ratio

$$\frac{S_j^k}{S^k},$$

where, considering the matching degree as the compatibility degree between the rule antecedent and the pattern feature values,

- S_j^k is the sum of the matching degrees for the class C_j patterns belonging to the fuzzy region delimited by the antecedent, and
- S^k the sum of the matching degrees for all the patterns belonging to this Fuzzy subspace, regardless its associated class.

O. Cordón, M.J. del Jesus, F. Herrera, A Proposal on Reasoning Methods in Fuzzy Rule-Based Classification Systems. *International Journal of Approximate Reasoning Vol. 20 (1999), 21-45*

Fuzzy rule based classification systems

Rule Weight

$$\mu_{\mathbf{A}_q}(\mathbf{x}_p) = \mu_{A_{q1}}(x_{p1}) \cdot \ \ldots \ \cdot \mu_{A_{qn}}(x_{pn}),$$

where $\mu_{A_{qi}}(\cdot)$ is the membership function of A_{qi}.

$$c(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{\sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}{\sum_{p=1}^{m} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}.$$

Confidence of the fuzzy rule

The consequent class C_q is specified by identifying class with the maximum confidence:

$$c(\mathbf{A}_q \Rightarrow \text{Class } C_q) = \max_{h=1,2,\ldots,M} \{c(\mathbf{A}_q \Rightarrow \text{Class } h)\}.$$

H.Ishibuchi, T.Yamamoto: Rule Weight Specification in Fuzzy Rule-Based Classification Systems, IEEE Trans. on Fuzzy Systems, Vol.13, No.4, pp.428-435 (2005, Aug.).

Fuzzy rule based classification systems

Rule Weight

$$CF_q = c(\mathbf{A}_q \Rightarrow \text{Class } C_q) - \sum_{\substack{h=1 \\ h \neq C_q}}^{M} c(\mathbf{A}_q \Rightarrow \text{Class } h).$$

H.Ishibuchi, T.Yamamoto: Rule Weight Specification in Fuzzy Rule-Based Classification Systems, IEEE Trans. on Fuzzy Systems, Vol.13, No.4, pp.428-435 (2005, Aug.).

OTHER MODELS:
E.G. Mansoori, M.J. Zolghadri, S.D. Katebi, A weighting function for improving fuzzy classification systems performance. Fuzzy Sets and Systems, 2007.

Fuzzy rule based classification systems

Inference Process

Pattern: $e=(e_1, ..., e_n)$ and Rule Base $\{R_1, ..., R_L\}$:

1. Matching degree with R_i:

$$h_i = T\,(\mu_{A^i_1}(e_1), \mu_{A^i_2}(e_2), \ldots \mu_{A^i_n}(e_n)) \quad ; \quad i = 1, \ldots, L$$

Usually Minimum t-norm

$$h_i = \mathrm{Min}\,(\mu_{A^i_1}(e_1), \mu_{A^i_2}(e_2), \ldots \mu_{A^i_n}(e_n))$$

or product t-norm

$$\mu_{\mathbf{A}_q}(\mathbf{x}_p) = \mu_{A_{q1}}(x_{p1}) \cdot \ \ldots \ \cdot \mu_{A_{qn}}(x_{pn}),$$

where $\mu_{A_{qi}}(\cdot)$ is the membership function of A_{qi}.

Fuzzy rule based classification systems

Inference Process

2. Association degree between the pattern and the classes C_j, j=1, ..., M:

$$b^j_i = g\,(h_i, r_i) \quad ; \quad j = 1, ..., M \quad ; \quad i = 1, ..., L$$

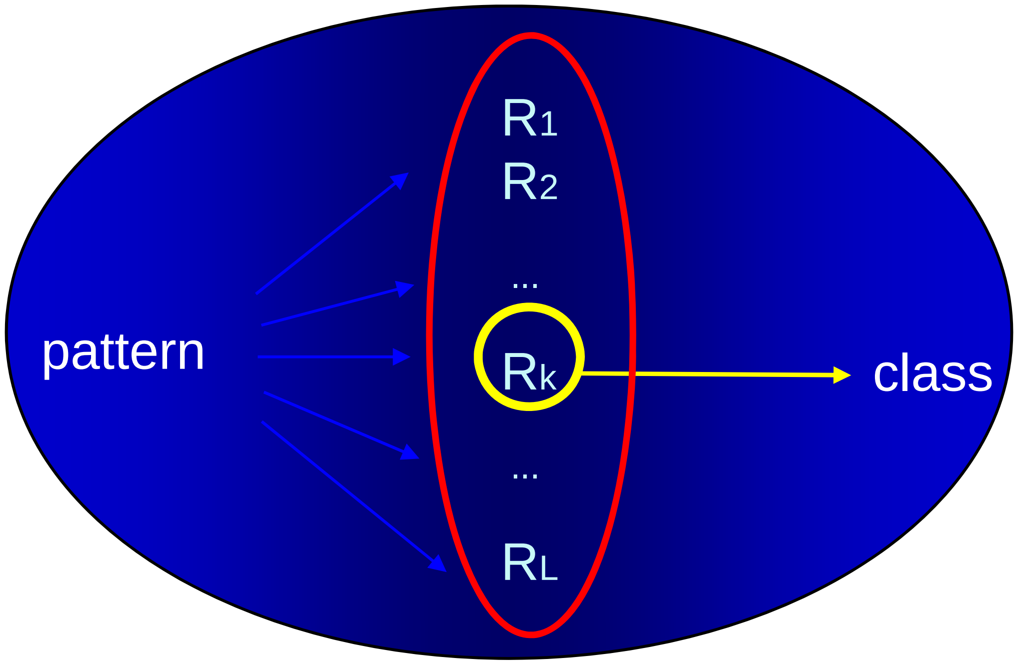
g = Min

g = Producto

3. Classification degree for each class: aggregation of the associatin degreess for each class. Two classical modelos: winning rule (Max of association degrees) and Sum based voting model (adding the association degrees)

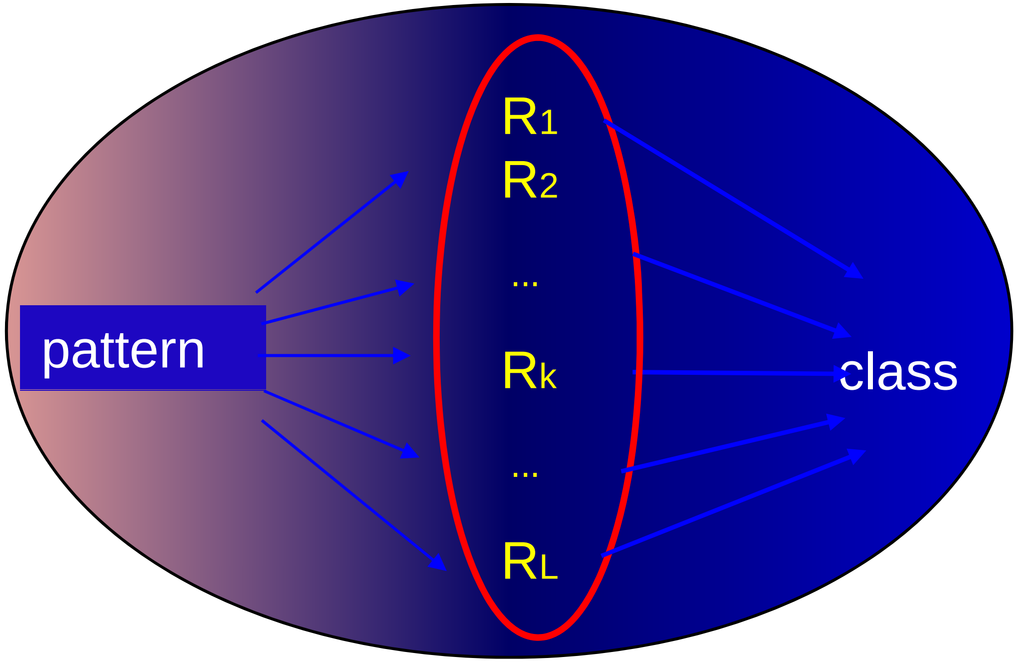
Fuzzy rule based classification systems

- Classification via the rule with maximun association degreee



Fuzzy rule based classification systems

- Classification via a voting model

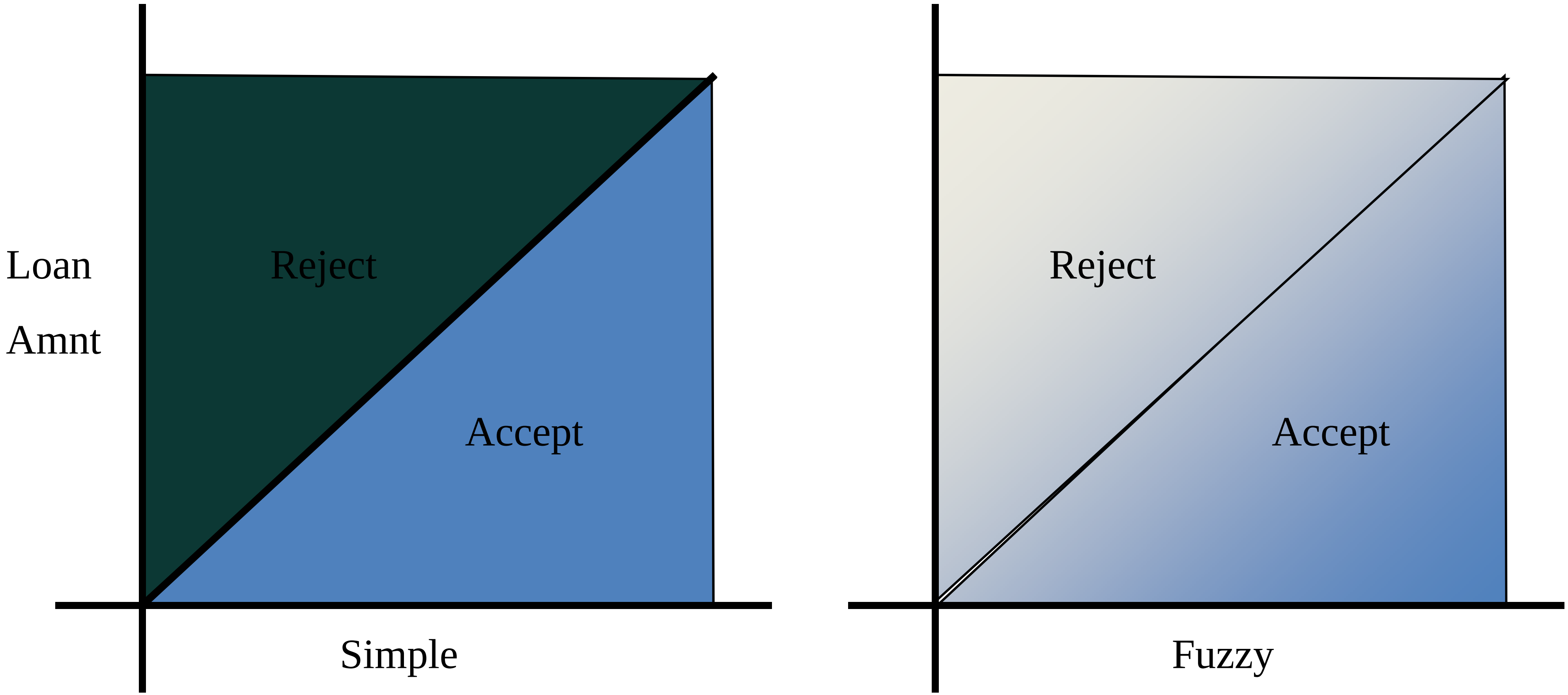


[24] O. Cordon, M. J. del Jesus, and F. Herrera, "A proposal on reasoning methods in fuzzy rule-based classification systems," *International Journal of Approximate Reasoning*, vol. 20, no. 1, January 1999, pp. 21-45.

[25] H. Ishibuchi, T. Nakashima, and T. Morisawa, "Voting in fuzzy rule-based systems for pattern classification problems," *Fuzzy Sets and Systems*, vol. 103, no. 2, April 1999, pp. 223-238.

Fuzzy rule based systems: Final Comments

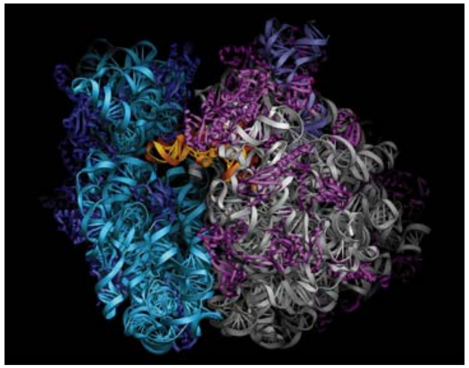
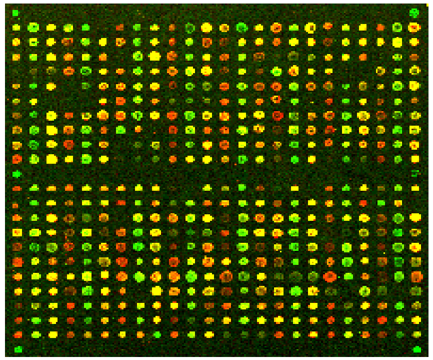
DM: Prediction and classification are fuzzy.



Introduction to Fuzzy Clustering

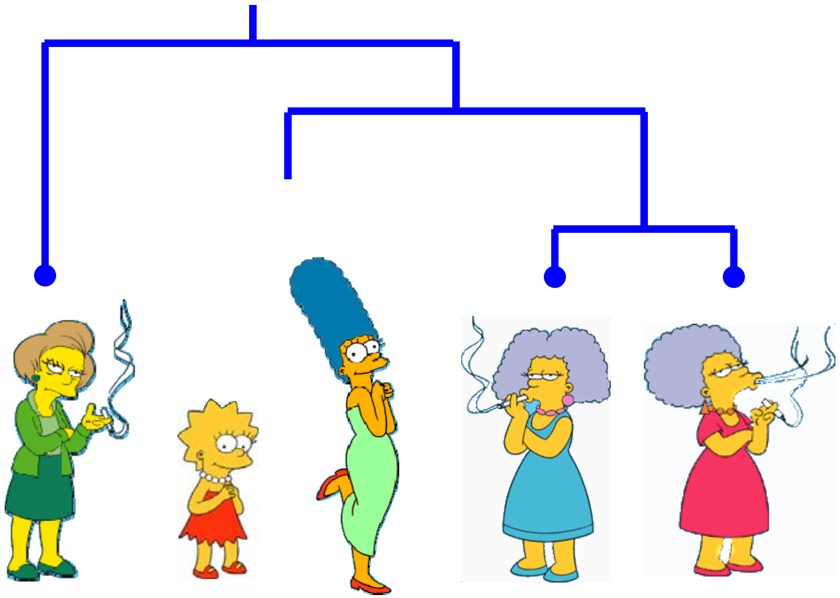
The aim of cluster analysis is to classify objects based on similarities among them.

- Motivation:
  - Why do we need clustering?
  - Why do we need fuzzy clustering?

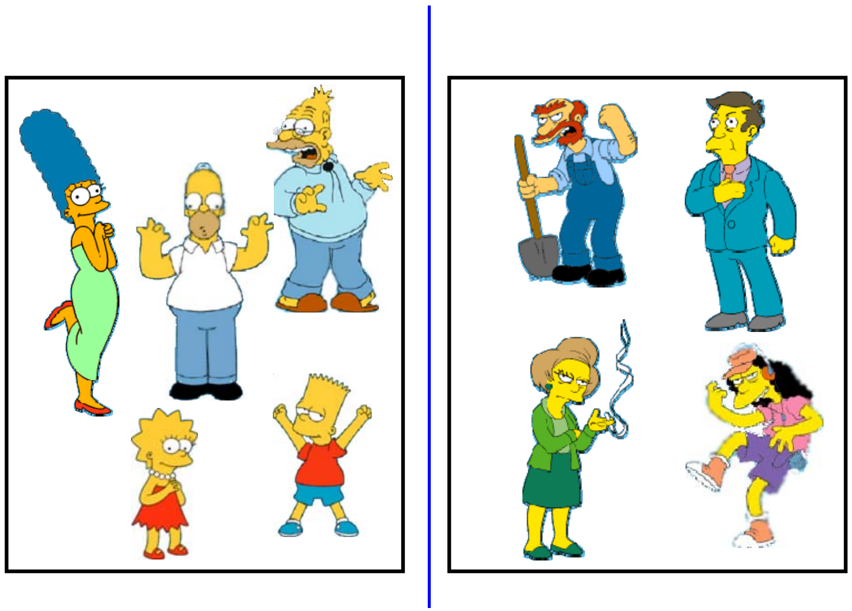




Why do we need clustering?

Hierarchical



Partitional



Fuzzy Clustering

- With fuzzy sets, how could clustering be performed to take into consideration:
  - Overlapping of clusters, and
  - To allow a record to belong to different clusters to different degrees.

Fuzzy C-Means algorithm is the most popular objective function based fuzzy clustering method, it is also the common base for most of the newly developed objective function based fuzzy clustering methods.

Why do we need Fuzzy Clustering?

- The mean height value for cluster 2 (short) is 5’3” and cluster 3 (medium) is 5’7”.
- You are just over 5'5” and are classified "medium".

- Fuzzy clustering: A membership value of each observation to each cluster is determined.
- User specifies a fuzzy MF.
- **A height of 5'5'' may give you a membership value of 0.4 to cluster 1, 0.4 to cluster 2 and 0.1 to cluster 3.**

Objective Function Based Fuzzy Clustering Methods

- An objective function measures the overall dissimilarity within clusters
- By minimizing the objective function we can obtain the optimal partition

Classical Clustering

Membership matrix **M**

data point *k*

cluster centre *i*

cluster centre *j*

$$m_{ik} = \begin{cases} 1 \; if \; \|\mathbf{u}_k - \mathbf{c}_i\|^2 \leq \|\mathbf{u}_k - \mathbf{c}_j\|^2 \\ 0 \; otherwise \end{cases}$$

distance

c-partition

All clusters *C* together fills the whole universe *U*

Clusters do not overlap

$$\bigcup_{i=1}^{c} C_i = U$$

$$C_i \cap C_j = \emptyset \quad for\,all\,i \neq j$$

$$\emptyset \subset C_i \subset U \quad for\,all\,i$$

$$2 \leq c \leq K$$

A cluster *C* is never empty and it is smaller than the whole universe *U*

There must be at least 2 clusters in a c-partition and at most as many as the number of data points *K*

Objective function

Minimise the total sum of all distances

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, \mathbf{u}_k \in C_i} \left\| \mathbf{u}_k - \mathbf{c}_i \right\|^2 \right)$$

Fuzzy membership matrix **M**

Point *k*'s membership of cluster *i*

Fuzziness exponent

$$m_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(q-1)}}$$

Distance from point *k* to current cluster centre *i*

Distance from point *k* to other cluster centres *j*

$$d_{ik} = \left\| \mathbf{u}_k - \mathbf{c}_i \right\|$$

Fuzzy membership matrix **M**

$$m_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(q-1)}}$$

$$= \frac{1}{\left(\frac{d_{ik}}{d_{1k}} \right)^{2/(q-1)} + \left(\frac{d_{ik}}{d_{2k}} \right)^{2/(q-1)} + \cdots + \left(\frac{d_{ik}}{d_{ck}} \right)^{2/(q-1)}}$$

$$= \frac{\frac{1}{d_{ik}^{\;2/(q-1)}}}{\frac{1}{d_{1k}^{\;2/(q-1)}} + \frac{1}{d_{2k}^{\;2/(q-1)}} + \cdots + \frac{1}{d_{ck}^{\;2/(q-1)}}}$$

Gravitation to cluster *i* relative to total gravitation

Fuzzy c-partition

All clusters C together fill the whole universe U. *Remark: The sum of memberships for a data point is 1, and the total for all points is K*

$$\sum_j \mu_{c_j}(x_i) = 1 \quad \forall x_i \in X$$

Not valid: Clusters do overlap

$$\bigcup_{i=1}^{c} C_i = U$$

$$C_i \cap C_j = \emptyset \quad for\, all\, i \neq j$$

$$\emptyset \subset C_i \subset U \quad for\, all\, i$$

$$2 \leq c \leq K$$

A cluster C is never empty and it is smaller than the whole universe U

There must be at least 2 clusters in a c-partition and at most as many as the number of data points K

Fuzzy c-Means

- Dunn defined a fuzzy objective function:

$$J_D(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^2 \left\| x_j - v_i \right\|^2$$

v_i is cluster center of *i* set

- Bezdek extended it to:

$$J_m(U,V;X) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^m \left\| x_j - v_i \right\|^2, 1 \le m \le \infty$$

$\left\| x_k - v_i \right\|^2$ represents the deviation of data with . The number *m* governs the influence of membership grades.

…Fuzzy C-Means

- The objective function of Fuzzy C-Means clustering can be defined as following:

$$J_{FCM} = \sum_{i=1}^{C} \sum_{j=1}^{n} \left(\mu_{ij}\right)^{m} \left\| x_j - z_i \right\|^2, \quad 1 < m < \infty$$

$$\textit{where} \quad (1) \;\; 0 \le u_{ij} \le 1, \quad j = 1, \ldots, n; i = 1, \ldots, C$$

$$(2) \sum_{i=1}^{C} u_{ij} = 1, \quad j = 1, \ldots, n;$$

$$(3) \; 0 < \sum_{j=1}^{n} u_{ij} < n, \quad i = 1, \ldots, C$$

…Fuzzy C-Means

- The FCM algorithm can be summarized as following steps:

1). Choose an initial partition membership matrix U^0;

2). Set the stop condition as follows:

$$\sum_{j=1}^{C} \left\| \overline{v}_j^{t} - \overline{v}_j^{t-1} \right\| < \varepsilon \qquad (1)$$

3). While not stop at condition (1) do

3-1). Compute reference vectors for each part family using Eq.(2)

$$\overline{v}_j^{*} = \sum_{i=1}^{P} \left(\mu_{ij}^{*}\right)^{m} \overline{a}_i \Big/ \sum_{i=1}^{P} \left(\mu_{ij}^{*}\right)^{m} \qquad (2)$$

...Fuzzy C-Means Clustering

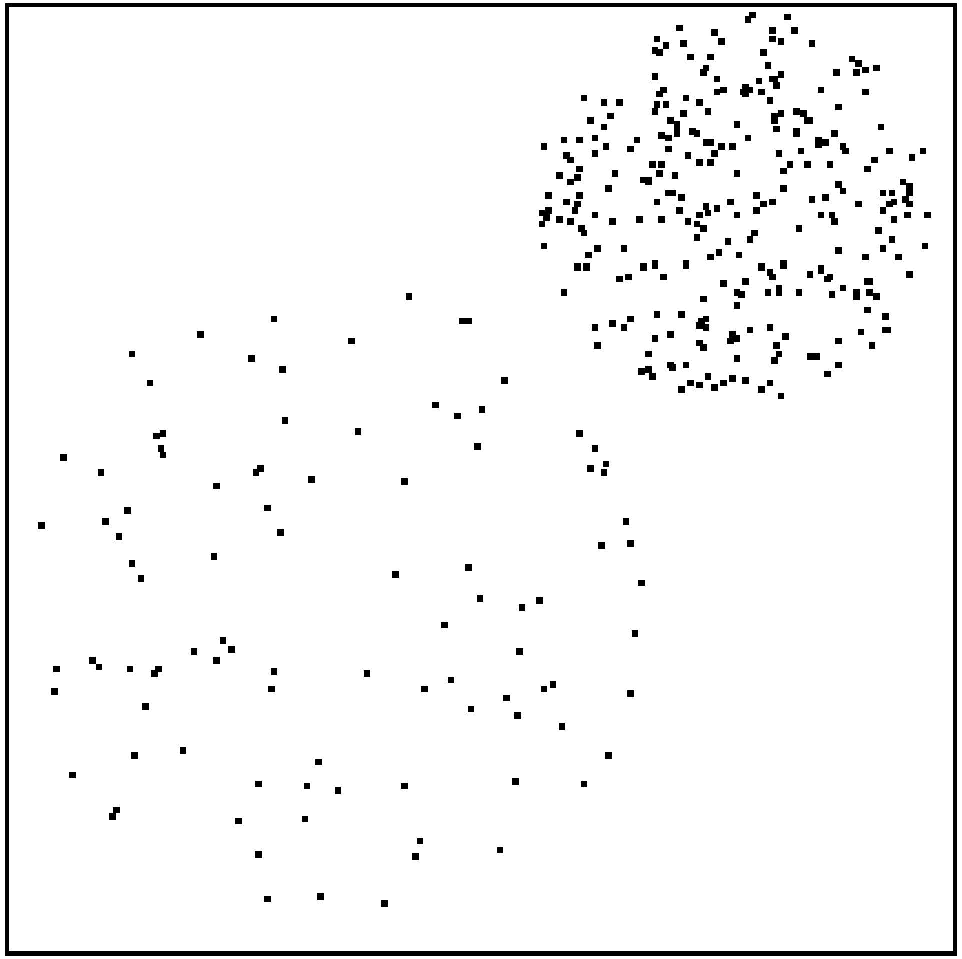
3-2). Update the part membership matrix according to (3)

$$\mu_{ij}^{*} = \left[\frac{1 / \left\| \overline{a}_i - \overline{v}_j^{*} \right\|}{\sum_{j'=1}^{C} \left(1 / \left\| \overline{a}_i - \overline{v}_{j'}^{*} \right\| \right)} \right]^{2/(m-1)} \qquad (3)$$

3-3). Evaluate $J_m(U,V)$ using (4). (4)

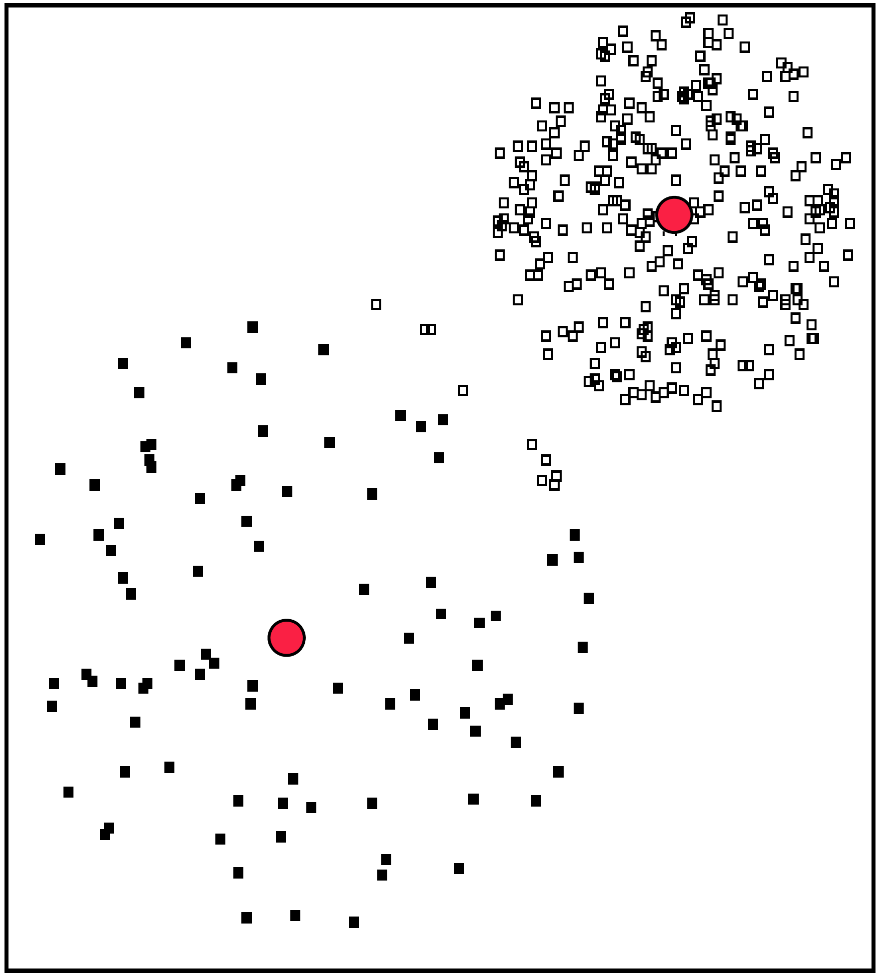
$$J_m(U,V) = \sum_{i=1}^{P} \sum_{j=1}^{C} \left(\mu_{ij} \right)^m \left\| \overline{a}_i - \overline{v}_j \right\|^2$$

K-means vs. Fuzzy *c*-means

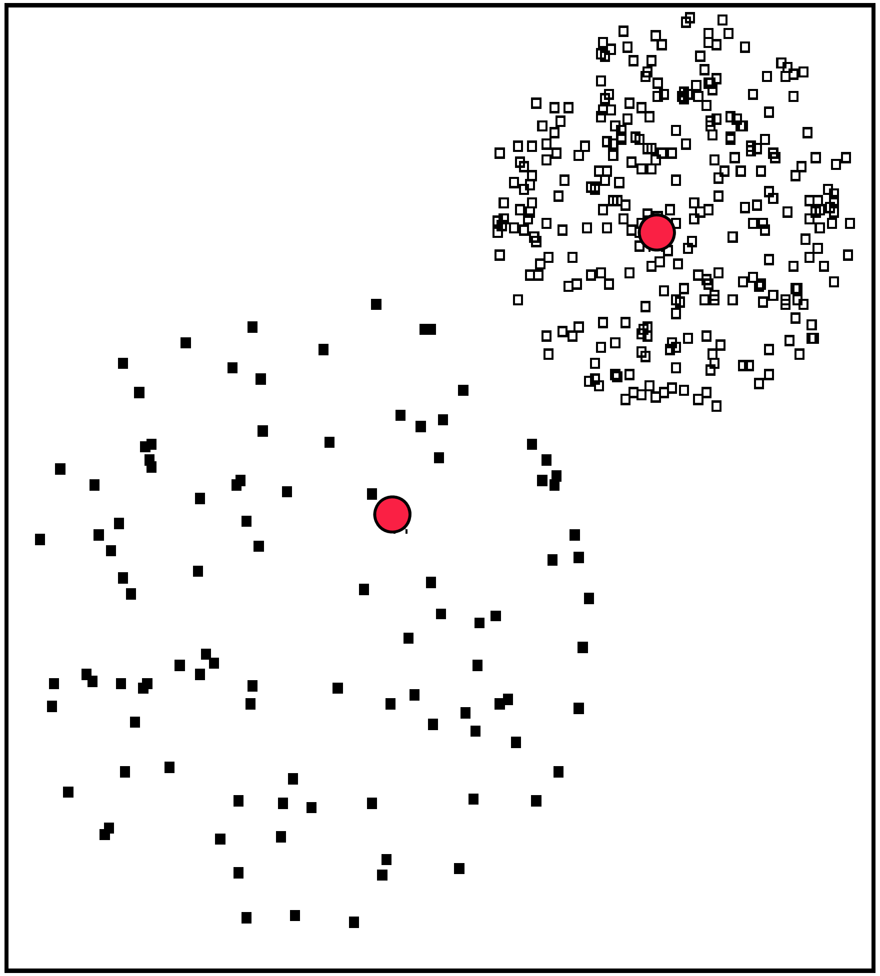


Sample Points

K-means vs. Fuzzy *c*-means

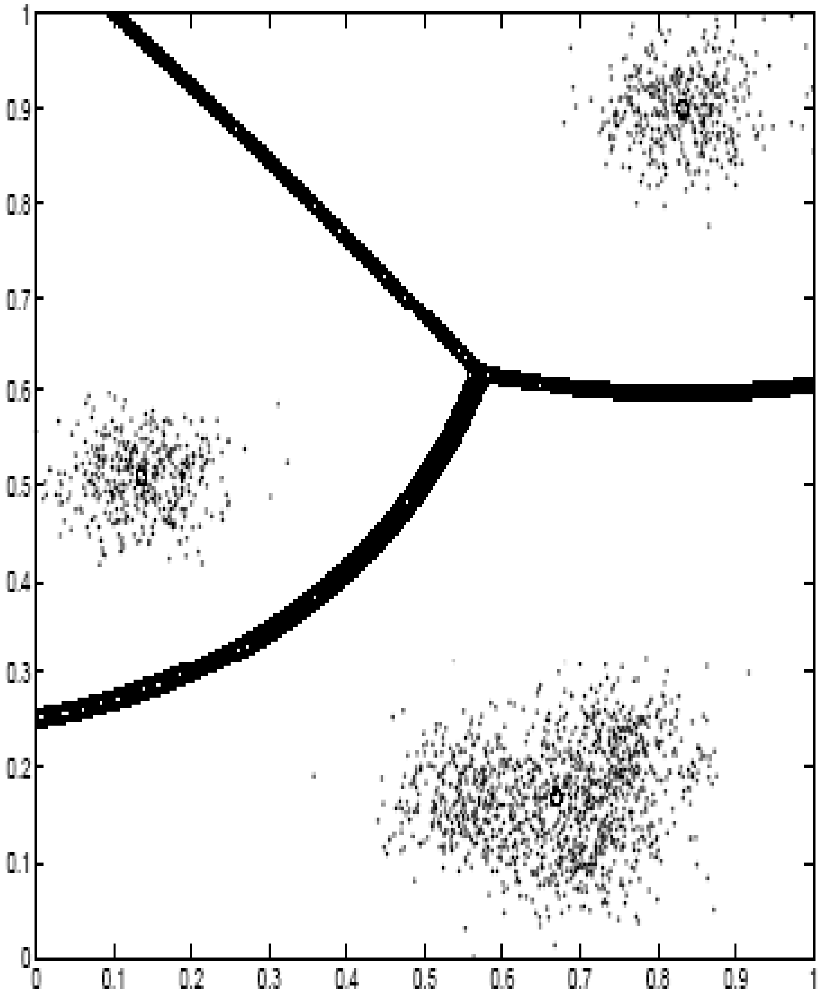


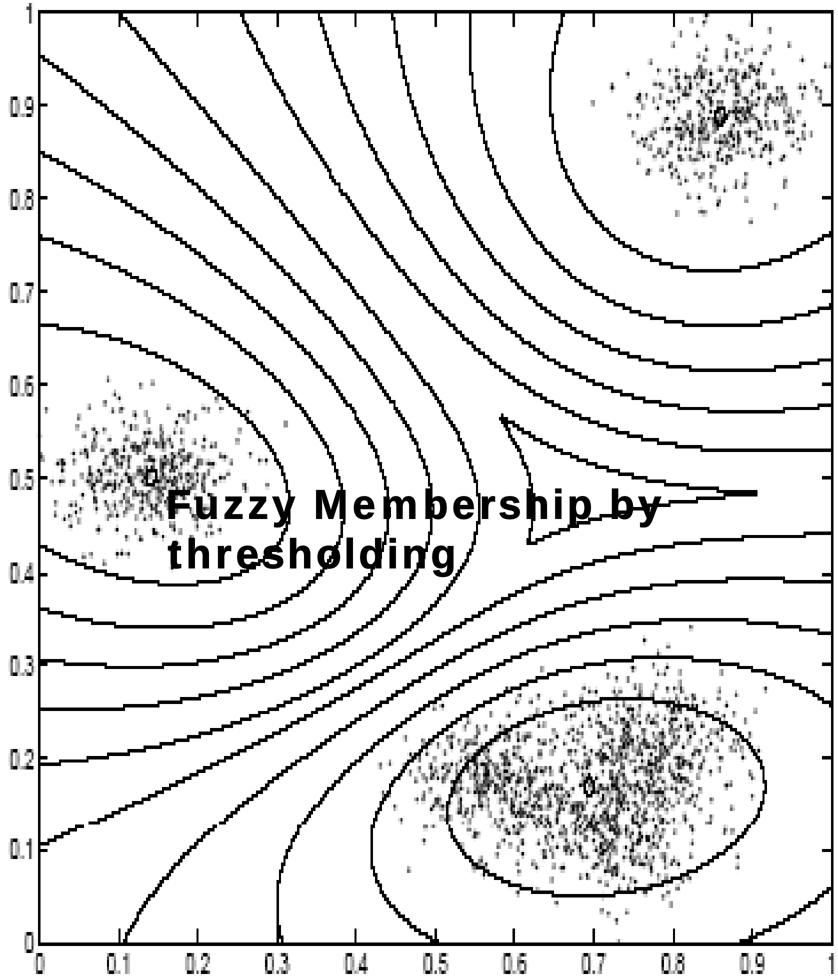
K-means



Fuzzy c-means

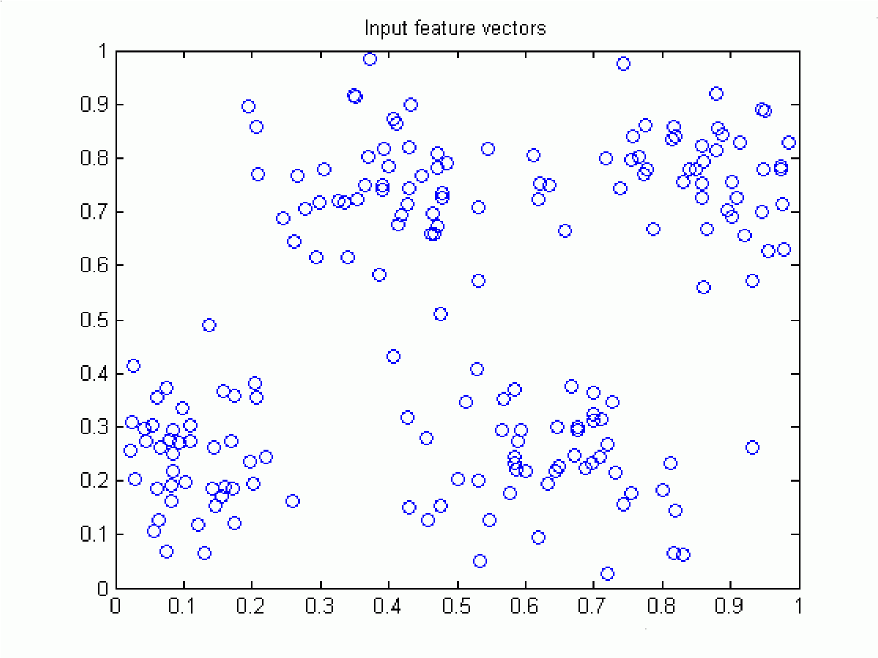
Clustering: hard/fuzzy methods

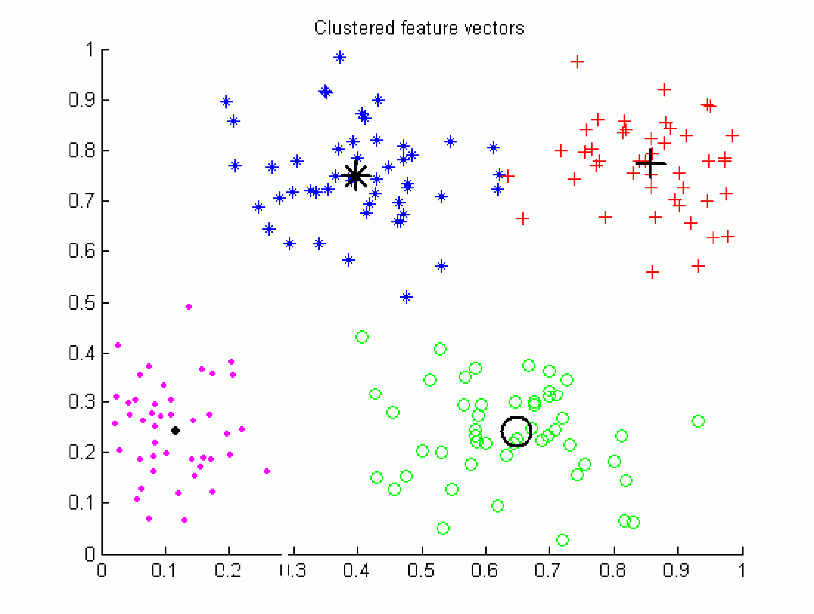




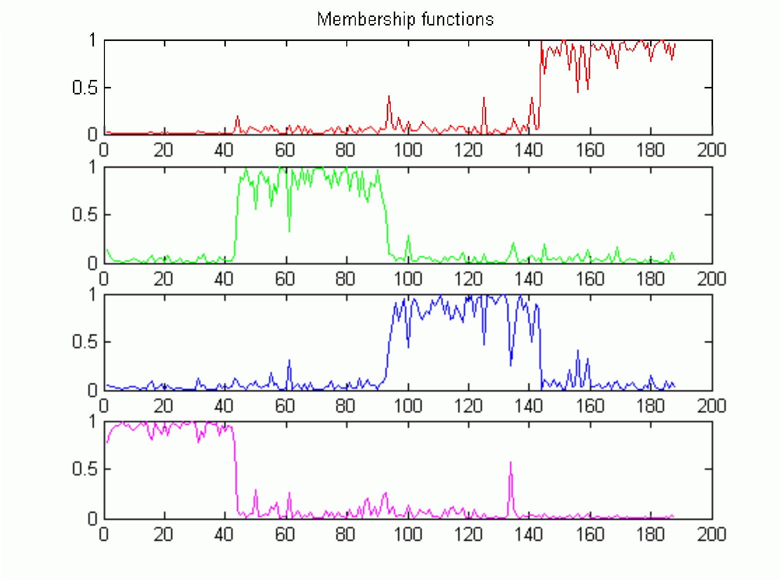
Fuzzy c-Means Clustering

Given c= # of groups





Membership Values



Example: Classify cracked tiles

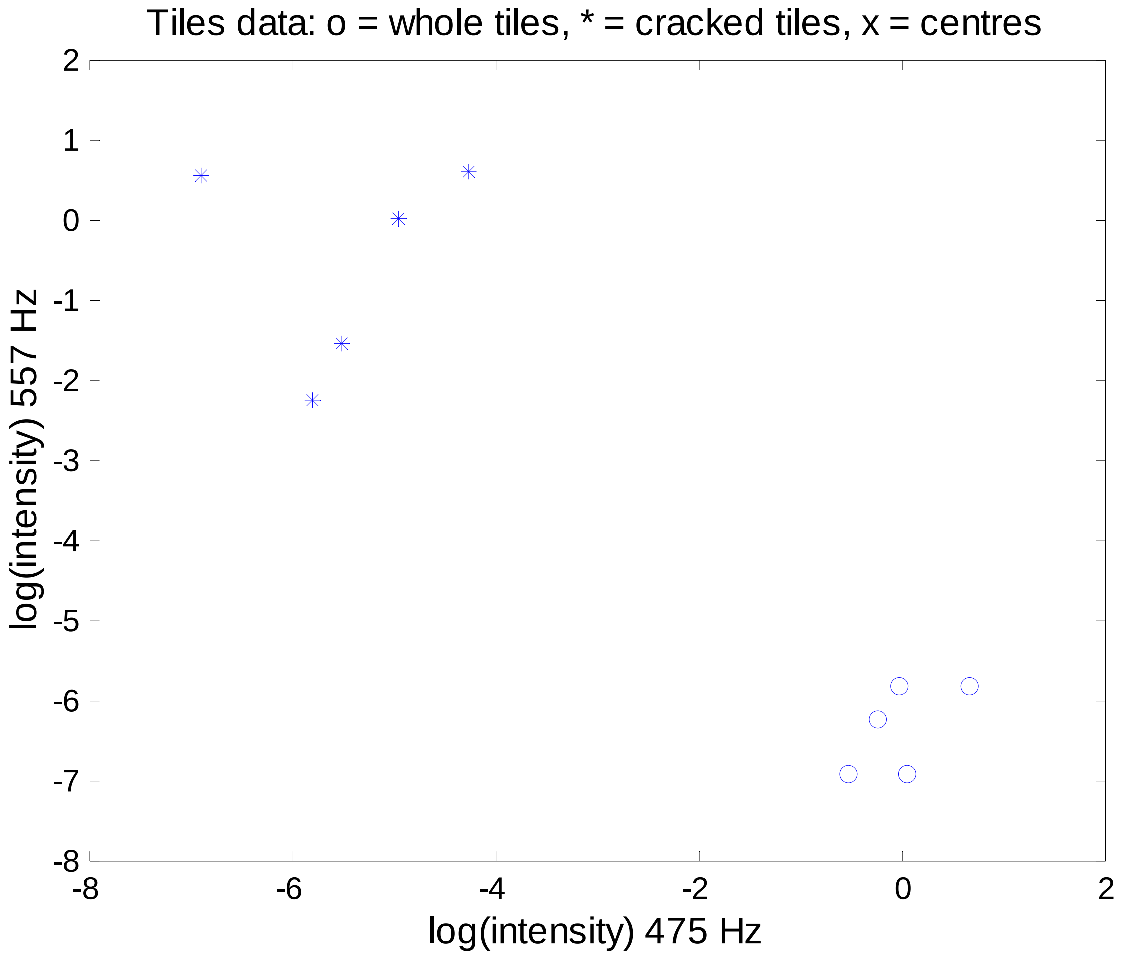
Algorithm: hard c-means (HCM)

(also known as k means)

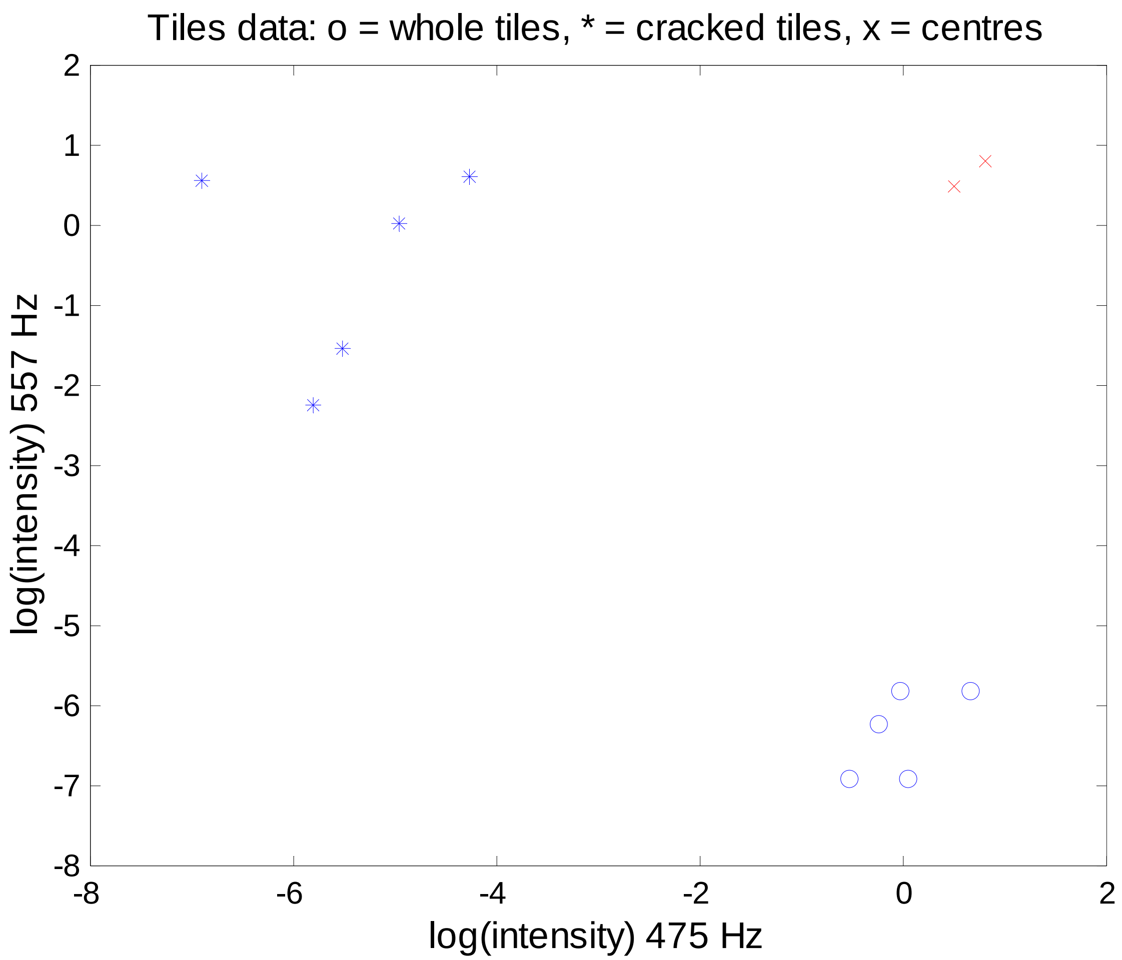
475Hz	557Hz	Ok?
0.958	0.003	Yes
1.043	0.001	Yes
1.907	0.003	Yes
0.780	0.002	Yes
0.579	0.001	Yes
0.003	0.105	No
0.001	1.748	No
0.014	1.839	No
0.007	1.021	No
0.004	0.214	No

Table 1: frequency intensities for ten tiles.

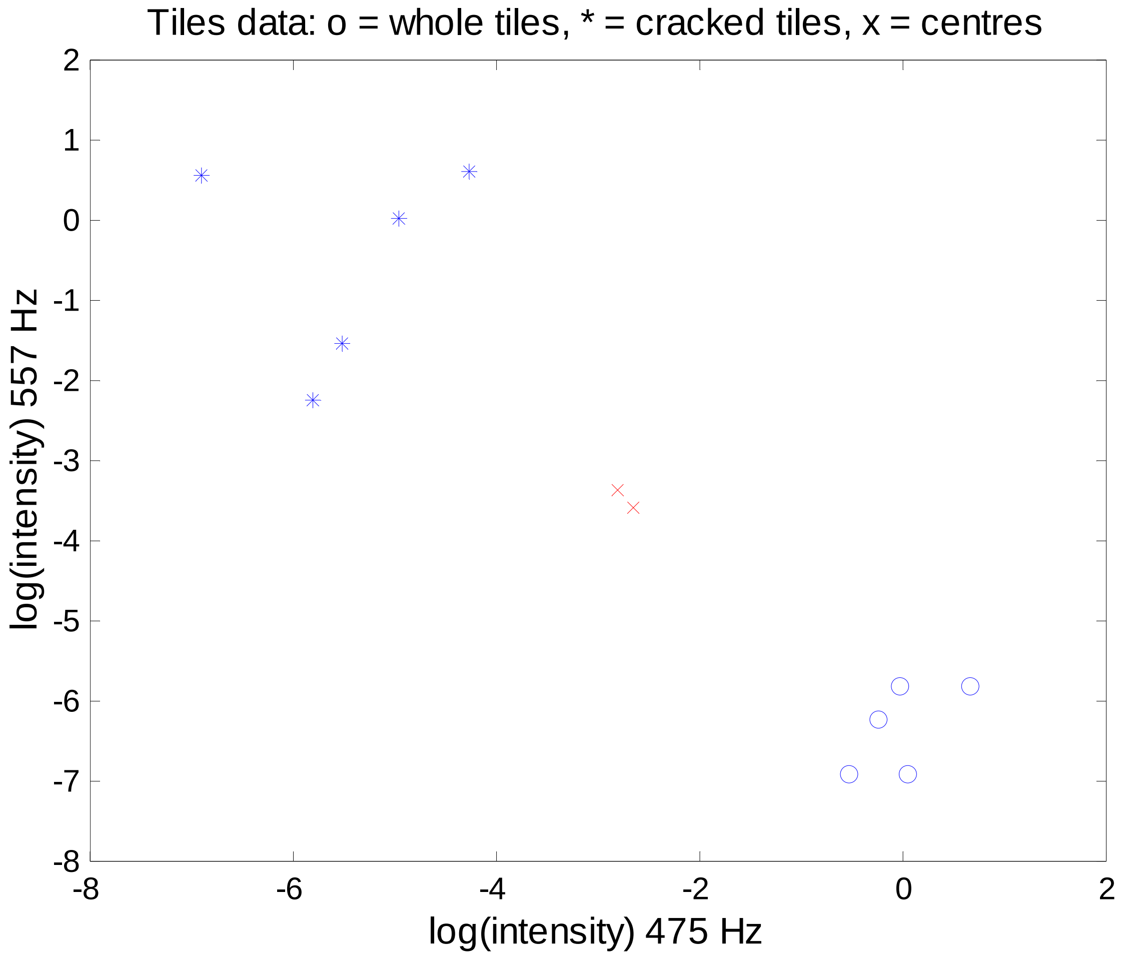
Tiles are made from clay moulded into the right shape, brushed, glazed, and baked. Unfortunately, the baking may produce invisible cracks. Operators can detect the cracks by hitting the tiles with a hammer, and in an automated system the response is recorded with a microphone, filtered, Fourier transformed, and normalised. A small set of data is given in TABLE 1 (adapted from MIT, 1997).



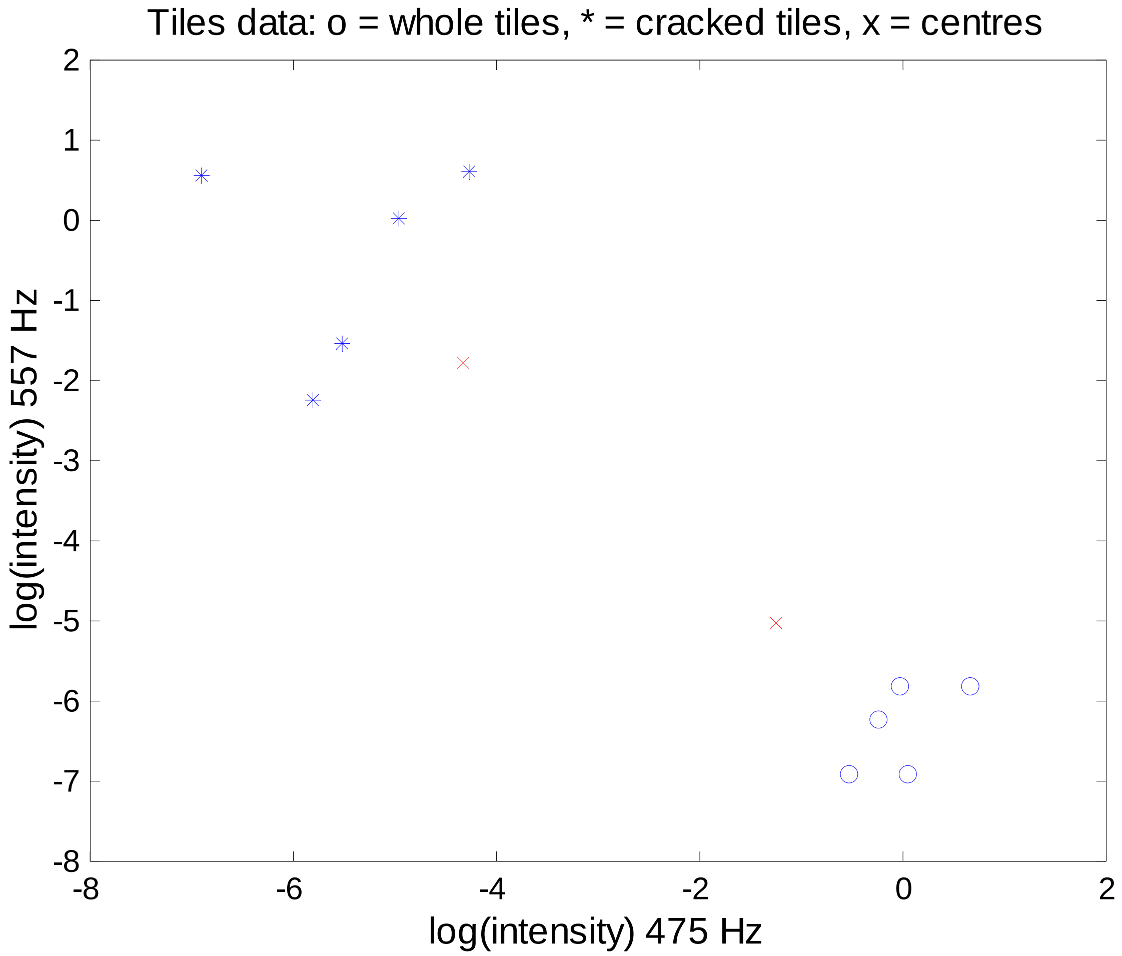
Plot of tiles by frequencies (logarithms). The whole tiles (o) seem well separated from the cracked tiles (*). The **objective** is to find the two clusters.



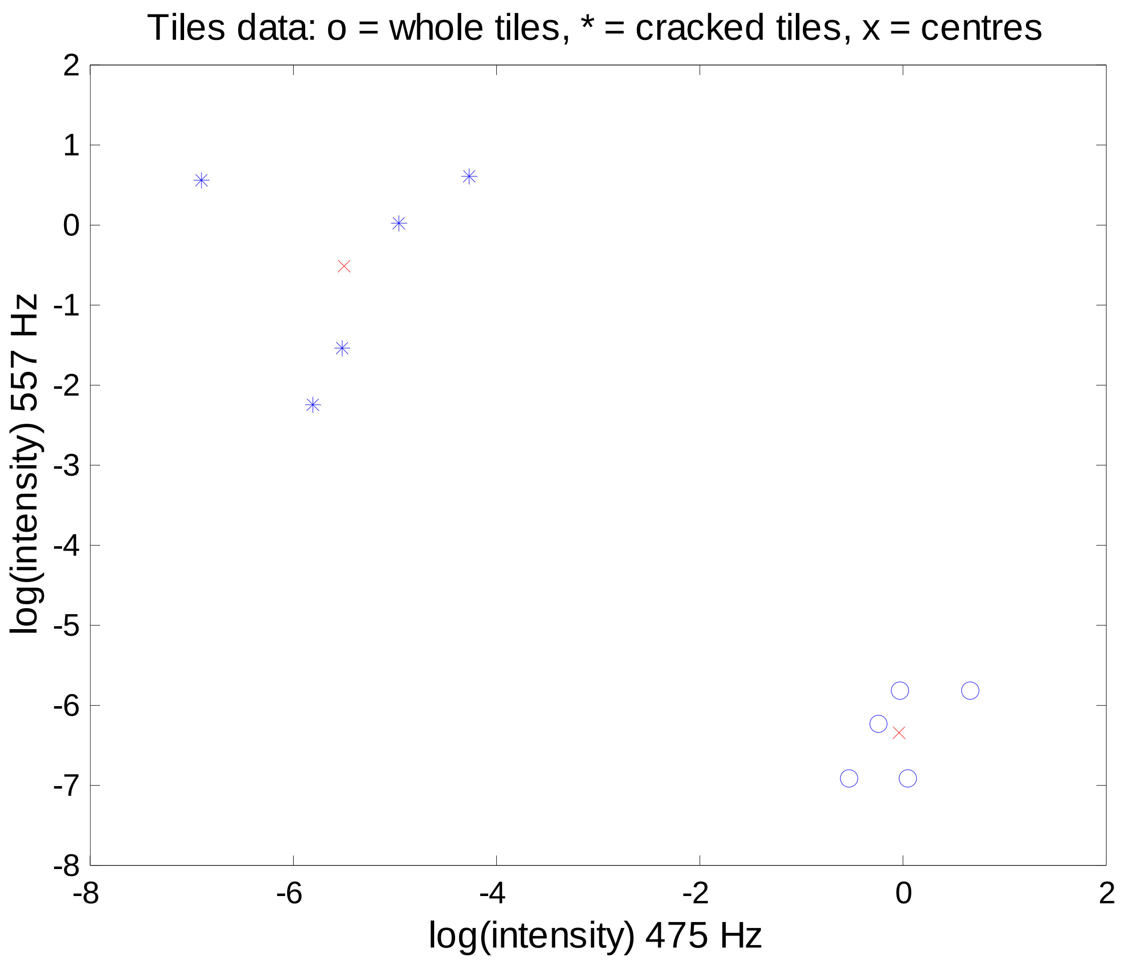
1. Place two cluster centres (x) at random.
2. Assign each data point (* and o) to the nearest cluster centre (x)



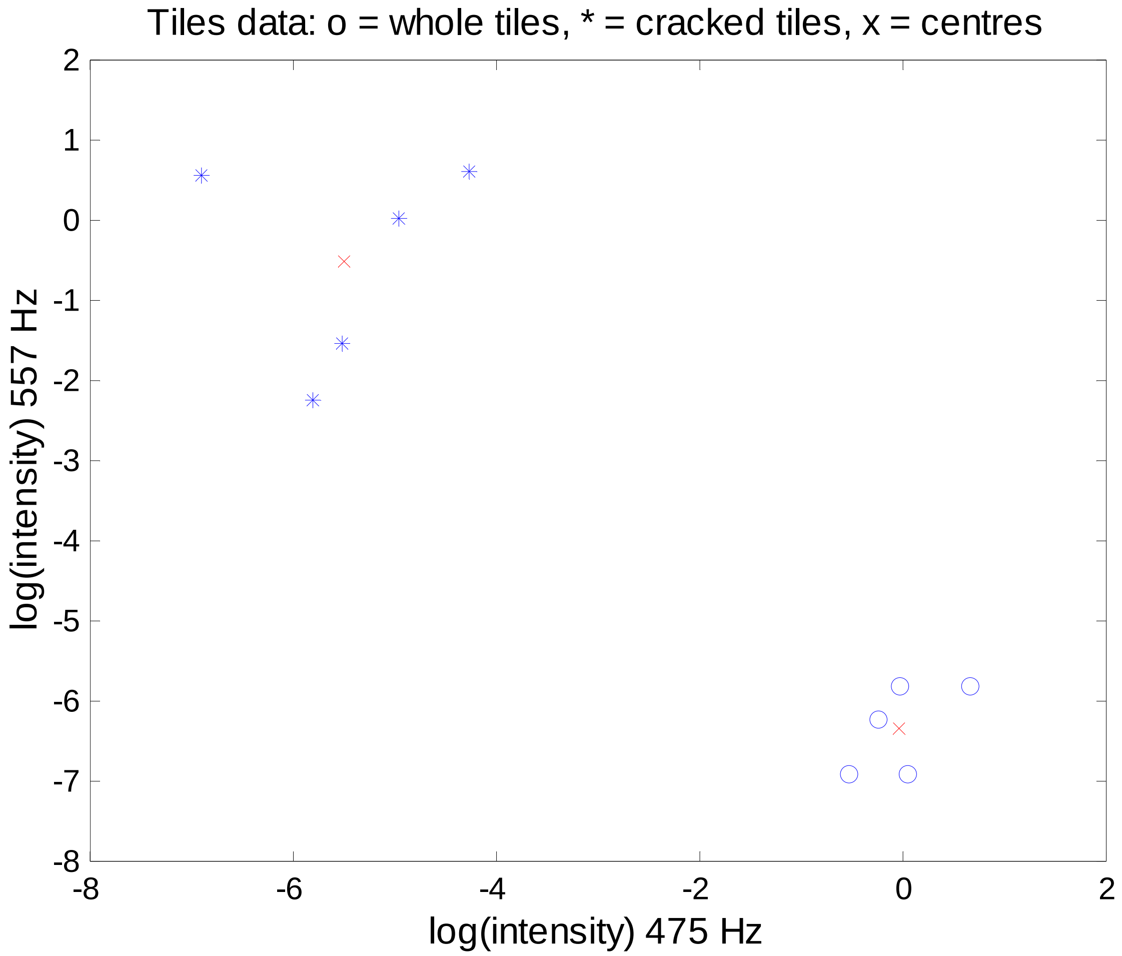
1. Compute the new centre of each class
2. Move the crosses (x)


Tiles data: o = whole tiles, * = cracked tiles, x = centres
log(intensity) 557 Hz
log(intensity) 475 Hz
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-8
-6
-4
-2
0
2

Iteration 2


Tiles data: o = whole tiles, * = cracked tiles, x = centres
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
log(intensity) 557 Hz
-8
-6
-4
-2
0
2
log(intensity) 475 Hz

Iteration 3



Iteration 4 (then stop, because no visible change)
Each data point belongs to the cluster defined by the nearest centre

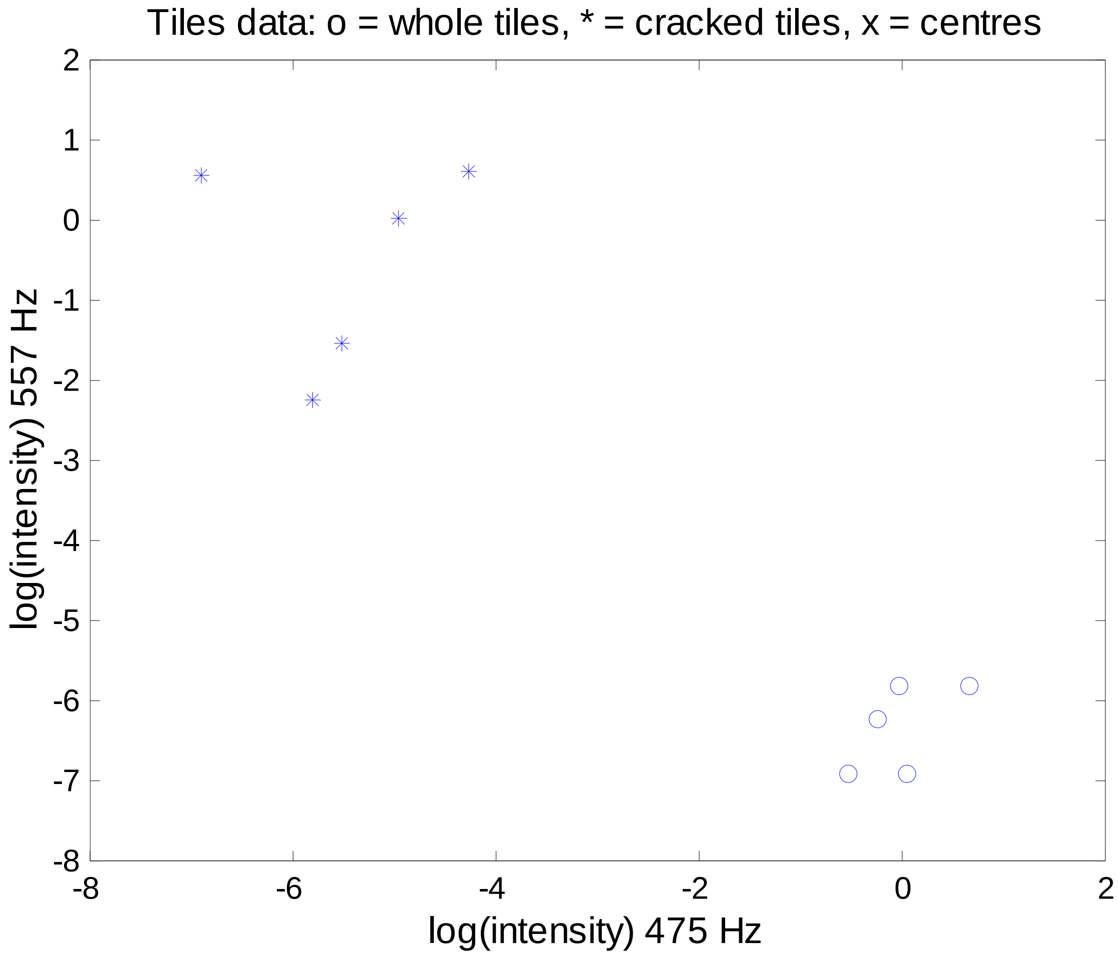
```
M =

    0.0000    1.0000
    0.0000    1.0000
    0.0000    1.0000
    0.0000    1.0000
    0.0000    1.0000
    1.0000    0.0000
    1.0000    0.0000
    1.0000    0.0000
    1.0000    0.0000
    1.0000    0.0000
```

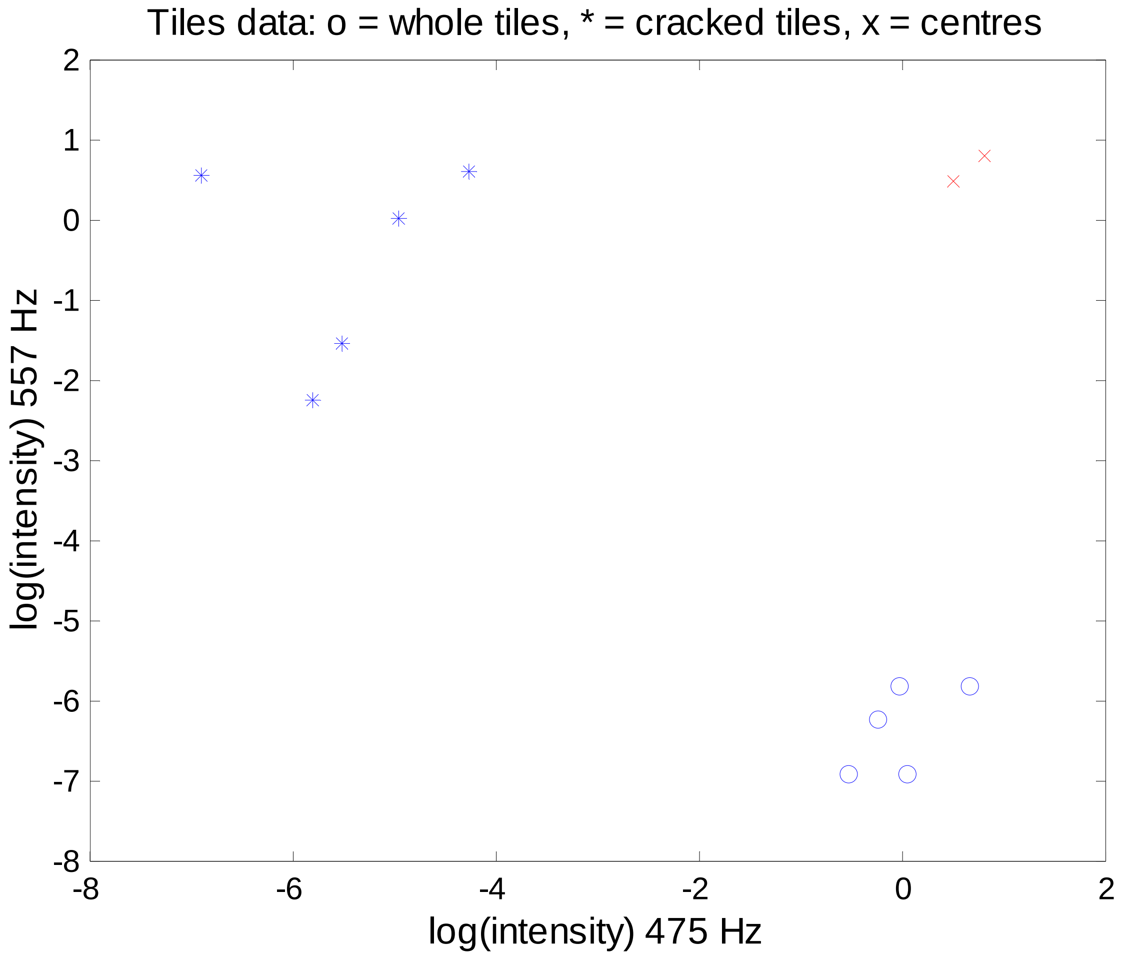
The membership matrix M:

1. The last five data points (rows) belong to the first cluster (column)
2. The first five data points (rows) belong to the second cluster (column)

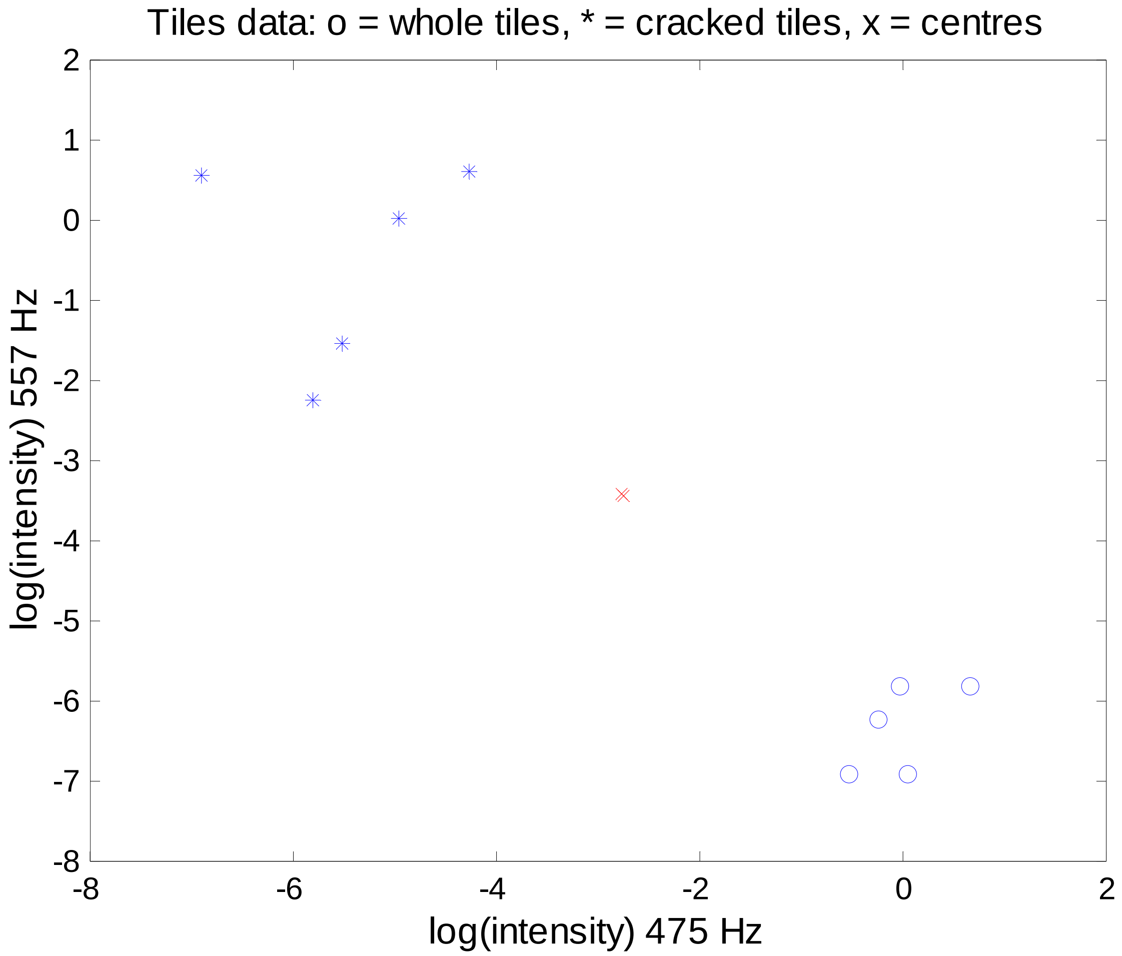
Algorithm: fuzzy c-means (FCM)



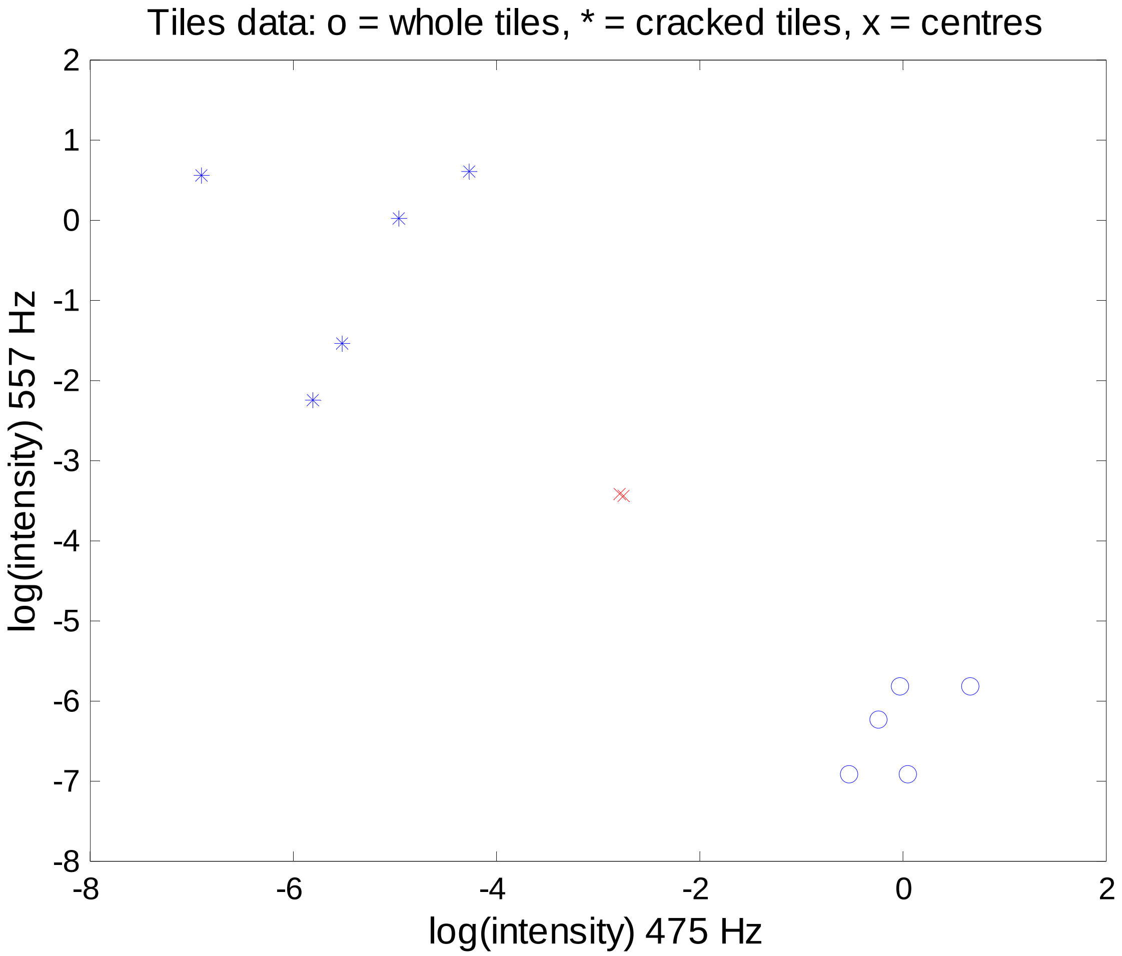
Each data point belongs to two clusters to different degrees



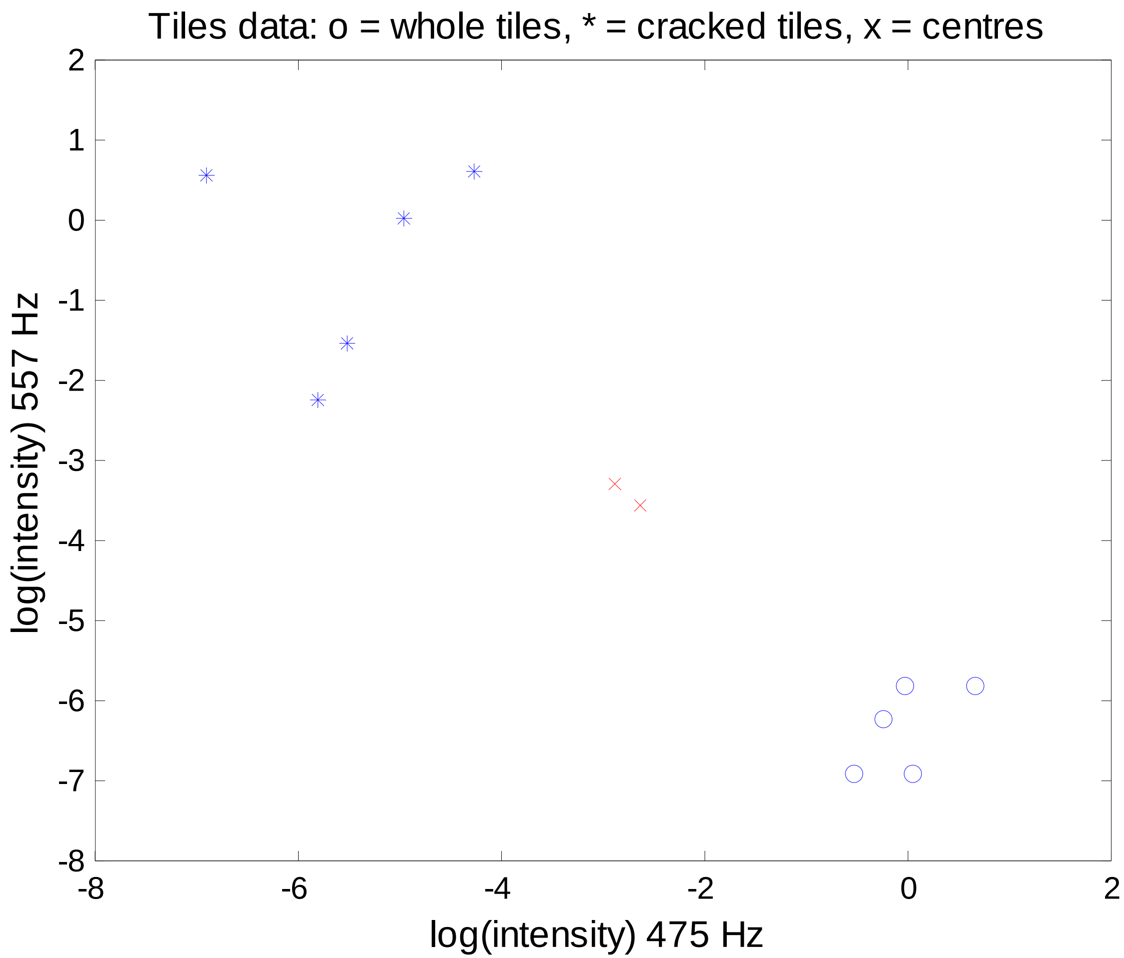
1. Place two cluster centres
2. Assign a fuzzy membership to each data point depending on distance



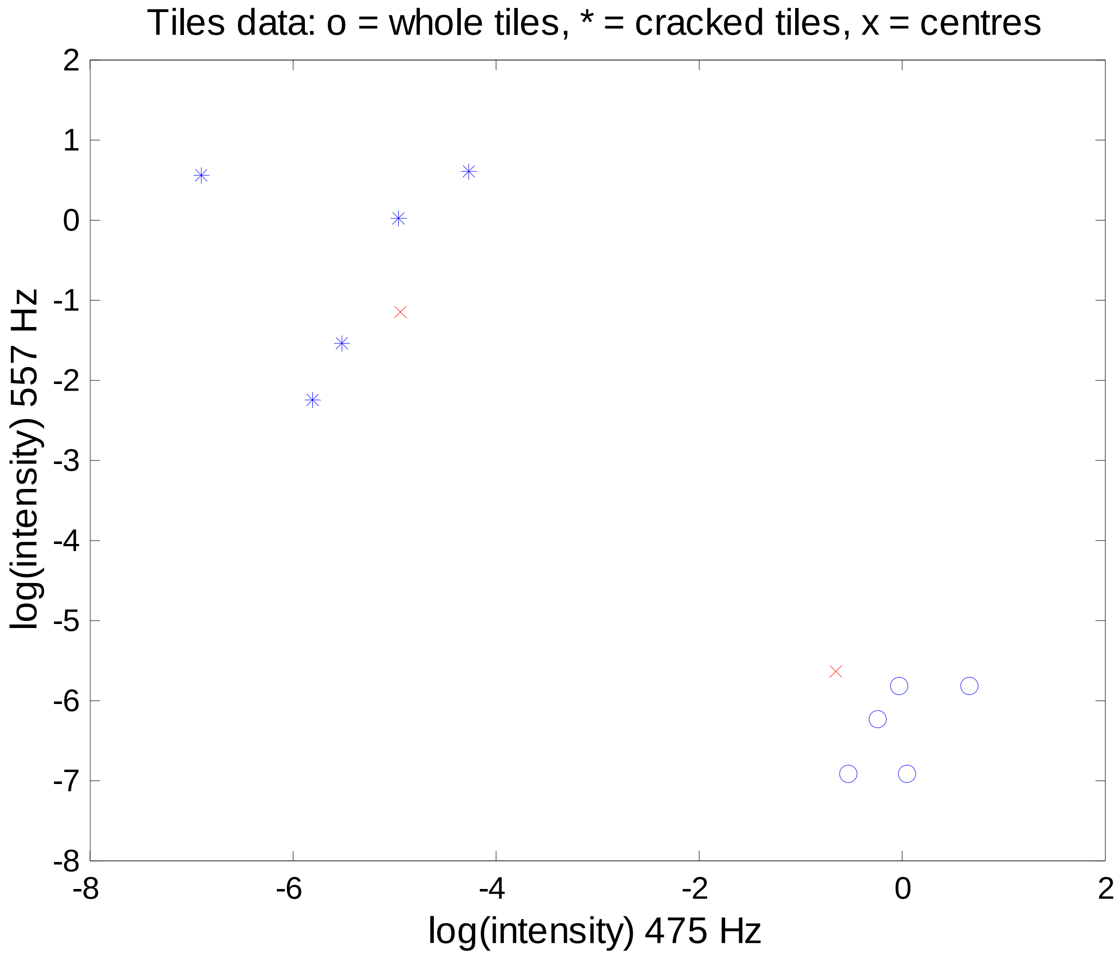
1. Compute the new centre of each class
2. Move the crosses (x)


Tiles data: o = whole tiles, * = cracked tiles, x = centres
log(intensity) 557 Hz
log(intensity) 475 Hz
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-8
-6
-4
-2
0
2

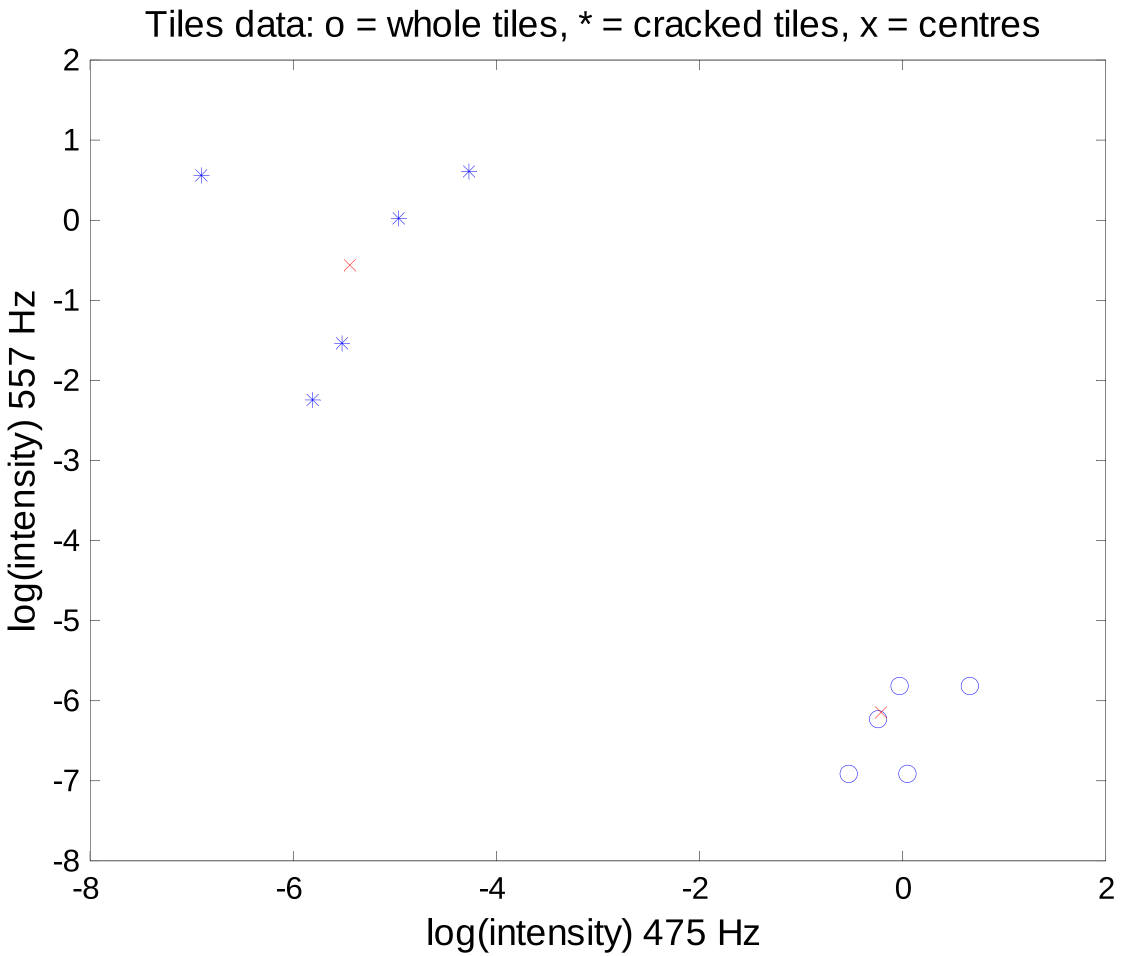
Iteration 2


Tiles data: o = whole tiles, * = cracked tiles, x = centres
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-8
-6
-4
-2
0
2
log(intensity) 557 Hz
log(intensity) 475 Hz

Iteration 5



Iteration 10



Iteration 13 (then stop, because no visible change)
Each data point belongs to the two clusters to a degree

```
M =

    0.0025    0.9975
    0.0091    0.9909
    0.0129    0.9871
    0.0001    0.9999
    0.0107    0.9893
    0.9393    0.0607
    0.9638    0.0362
    0.9574    0.0426
    0.9906    0.0094
    0.9807    0.0193
```

The membership matrix M:

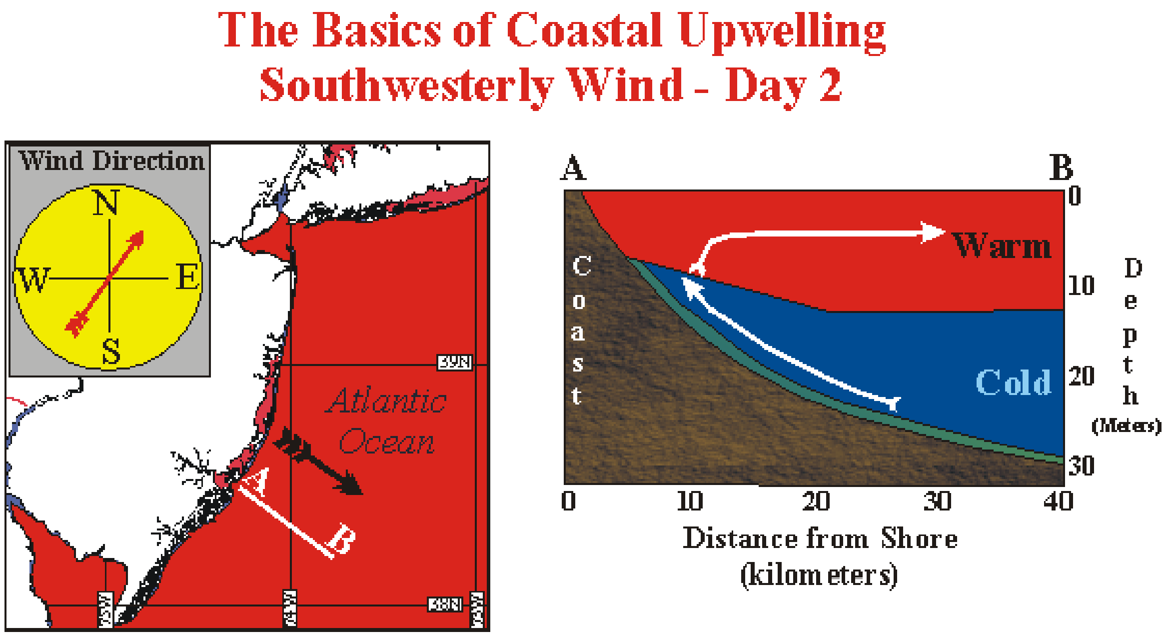
1. The last five data points (rows) belong mostly to the first cluster (column)
2. The first five data points (rows) belong mostly to the second cluster (column)

Example: Applicability of Fuzzy Clustering for the Identification of Upwelling Areas on Sea Surface Temperature Images (Nascimiento et al., 2005)

- **What is Upwelling?**

➢ It is a mass of deep, cold, and nutrient-rich seawater that rises close to the coast.

➢ Upwelling occurs when winds parallel to the coast induce a net mass transport of surface seawater in a 90º direction, away from the coast, due to the Coriolis force. Deep waters rise in order to compensate the mass deficiency that develops along the coastal area.



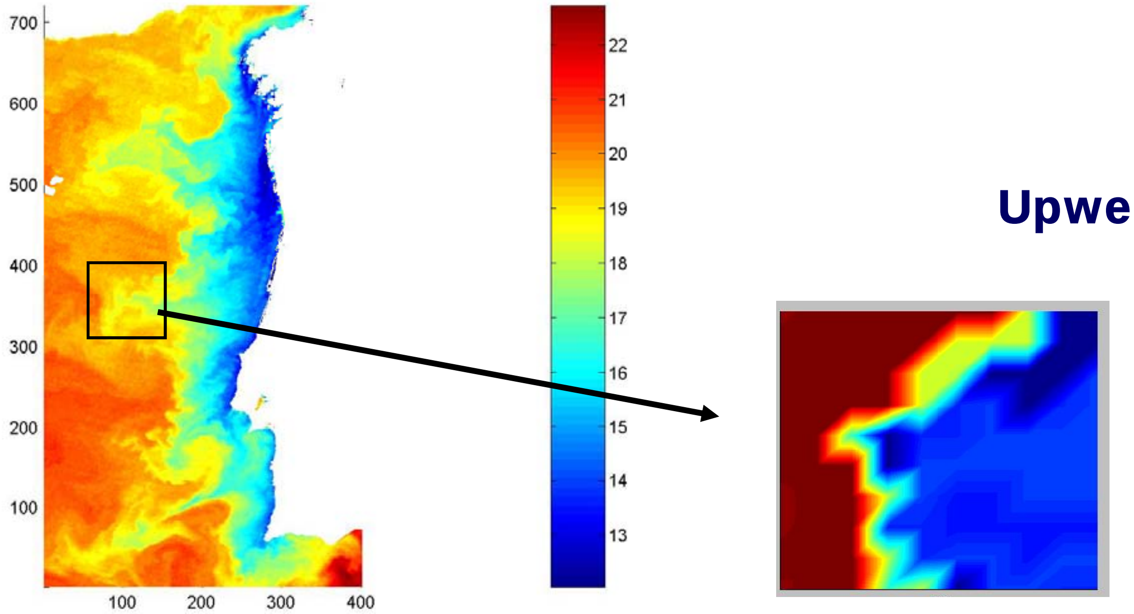
Fuzzy Membership by thresholding

Why is Upwelling so important?

- ✓ Brings nutrient-rich deep waters close to the ocean surface, creating regions of high biological productivity.
- ✓ Strong impact on fisheries, and global oceanic climate models

Why SST Image Segmentation by Fuzzy Clustering?

- Nature of the problem is Fuzzy

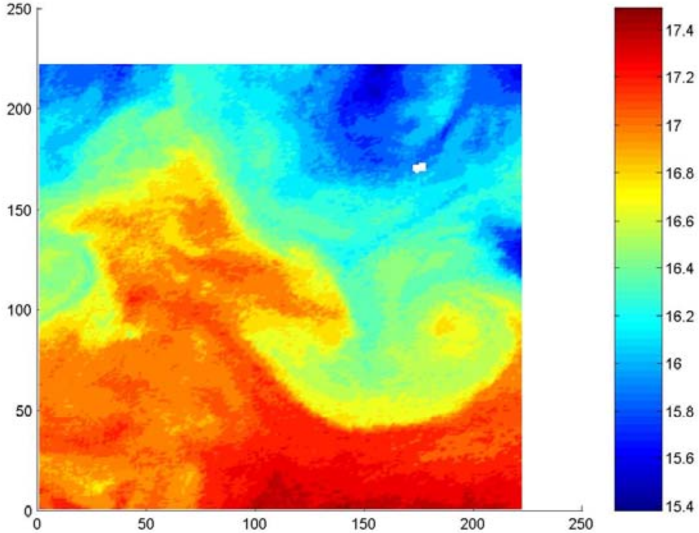


- Unsupervised segmentation does not require training data.

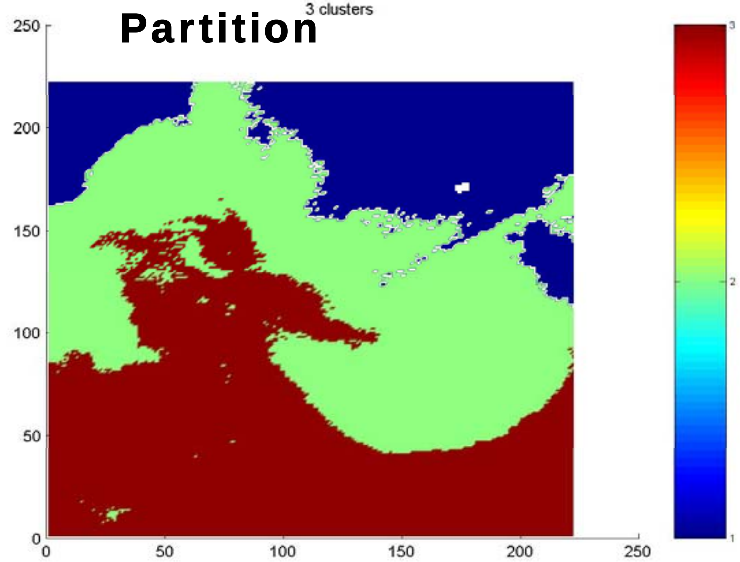
- Expert´s can take advantage of visualization skills and interpretability of fuzzy membership values.

Example: Applicability of Fuzzy Clustering for the Identification of Upwelling Areas on Sea Surface Temperature Images (Nascimiento et al., 2005)

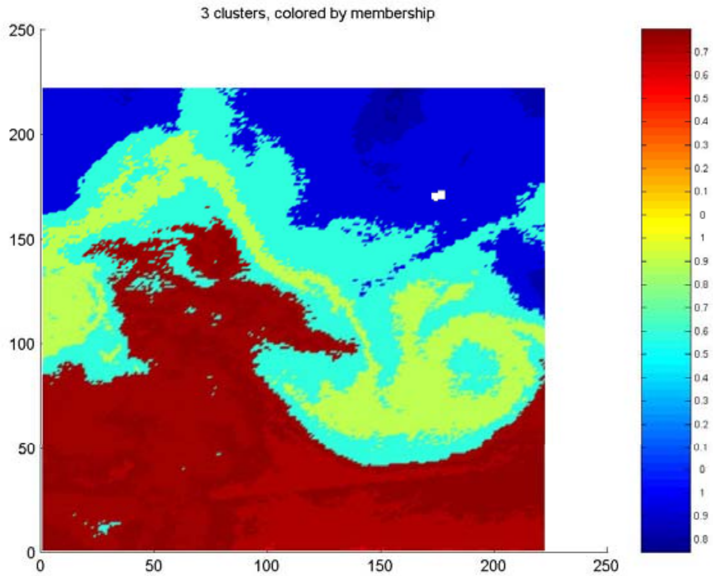
Original Image



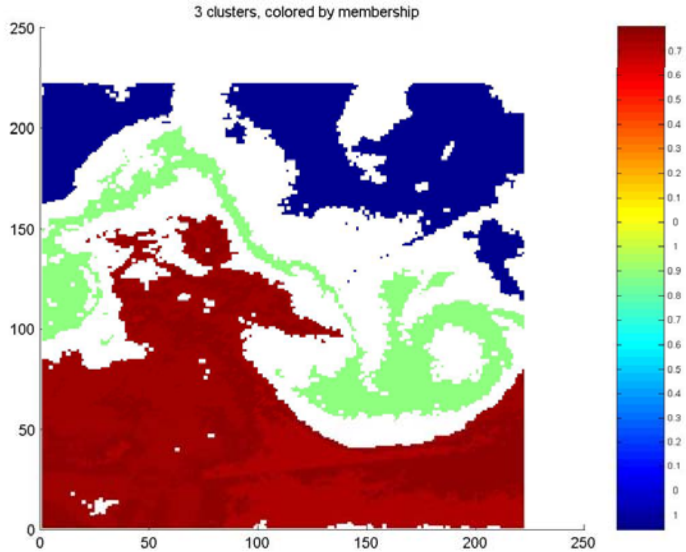
Defuzzified Partition



Max Fuzzy Membership Partition



Fuzzy Membership by thresholding



Fuzzy c-Means

Limitation:

- it needs to know the number of clusters.
- How to find an optimal number of clusters?.
- A lot of validity indexes.

Cluster validity measures

Partition coef.: $$PC = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{C}\mu_{ik}^2$$

Baker's measure: $$B = 1 - \frac{C}{C-1}[1 - PC] = \frac{1}{C-1}\sum_{j=i+1}^{C}\sum_{i=1}^{C-1}\left[\frac{1}{n}\sum_{k=1}^{n}\left(\mu_{ik} - \mu_{jk}\right)^2\right]$$

Partition entropy: $$PE = -\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{C}[\mu_{ik}\ln(\mu_{ik})]$$

Partition linear index: $$PLI = \sum_{i=1}^{n}\min\{(1 - \max_C(\mu_i)), \max_C(\mu_i)\}$$

Amount of overlap: $$OL = \frac{1}{nC}\sum_{j=i+1}^{C}\sum_{i=1}^{C-1}\sum_{k=1}^{n}\min\left(\mu_{ik}, \mu_{jk}\right)$$

Partition stability: std. of prototypes found for dif. *m*, const. *C*.

Fuzzy c-Means: Final Comments

- The aim of cluster analysis is to classify objects based on similarities among them.
- With fuzzy sets, clustering performs taking into consideration:
  - Overlapping of clusters, and
  - To allow a record to belong to different clusters to different degrees.

Fuzzy Association Rules

- Many based on Apriori algorithm
- Treat all attributes (or at least linguistic) as uniform

Some approaches:

  - T.P. Hong, Ch.S. Kuo, S.S. Tseng, Mining association rules from quantitative data, Intelligent Data Analysis 3 (1999) 363-376
  - T.P. Hong, K.Y. Lin, S.L. Wang, Fuzzy Data Mining for Interesting Generalized Association Rules, Fuzzy Sets and Systems 138 (2003) 255-269

The Apriori Algorithm — Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

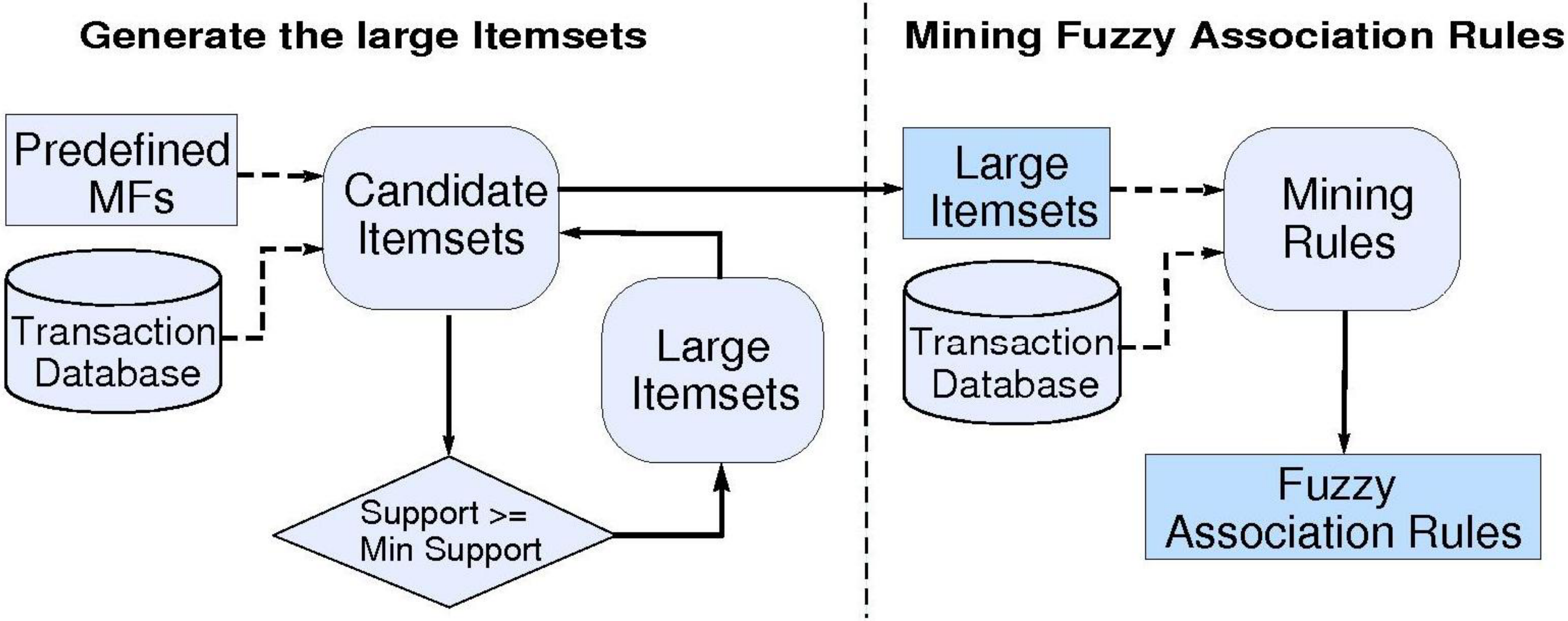
itemset
{2 3 5}

Scan D

L_3

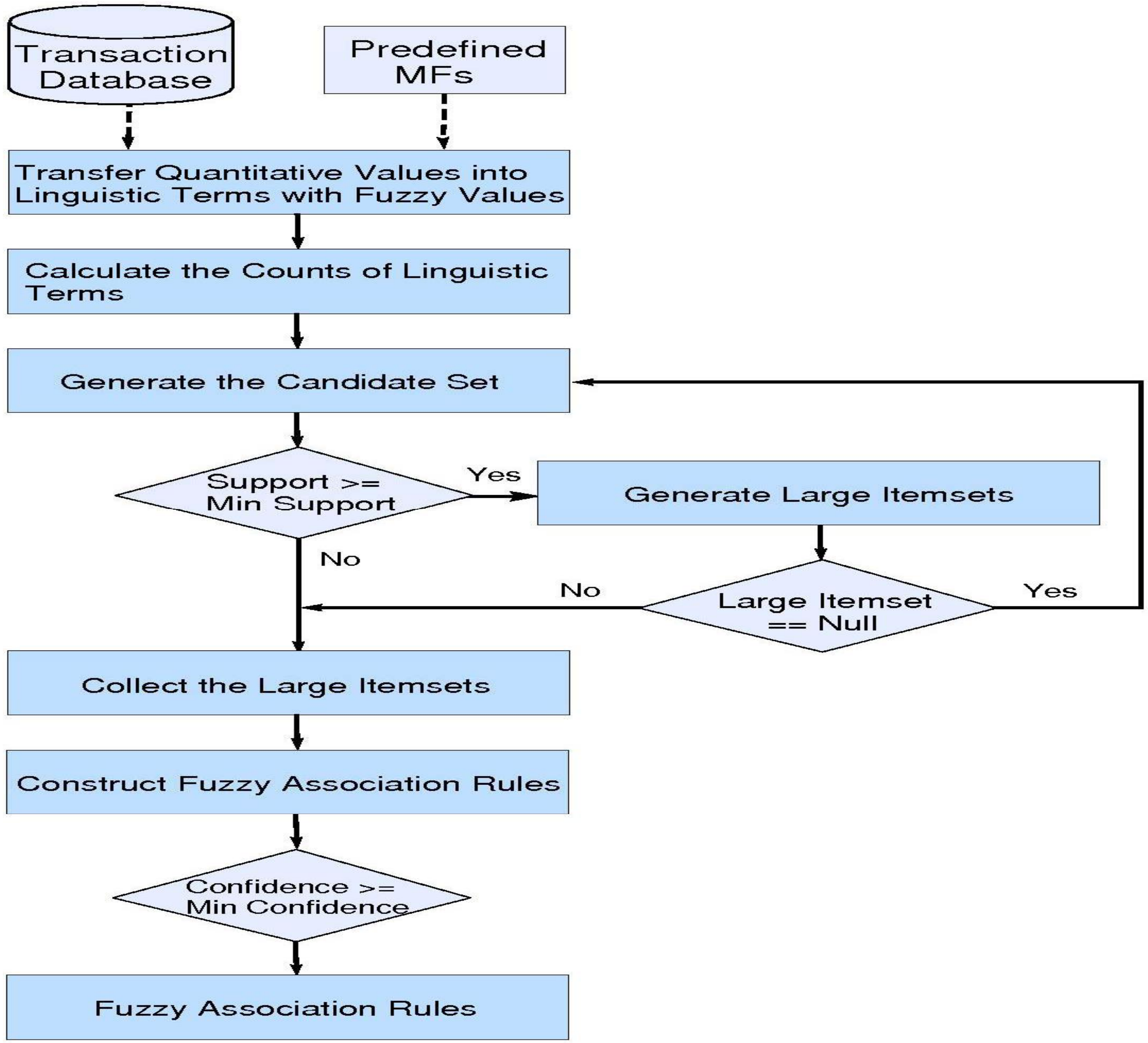
itemset	sup
{2 3 5}	2

FuzzyApriori



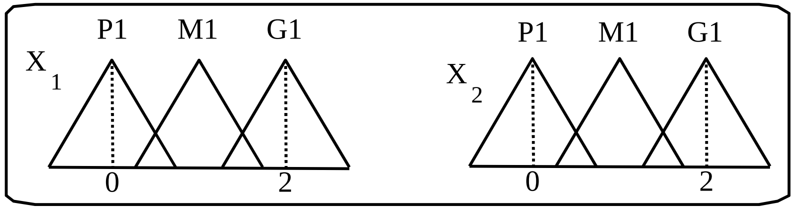
T.P. Hong, Ch.S. Kuo, S.S. Tseng, Mining association rules from quantitative data, Intelligent Data Analysis 3 (1999) 363-376

FuzzyApriori



FuzzyApriori

Data Base



Transactions

t_1= (0.2 , 1.0)

t_2= (0.4 , 0.8)

t_3= (1.0 , 1.2)

→ P1: (0,7 + 0,55 + 0) / 3 = 0,42
→ M1: (0 + 0,2 + 1.0) / 3 = 0,4
→ G1: (0 + 0 + 0) / 3 = 0
→ P2: (0 + 0 + 0) / 3 = 0
→ M2: (1,0 + 0,8 + 0,8) / 3 = 0,87
→ G2: (0 + 0 + 0) / 3 = 0

Minumum Support = 0,4

Frequent itemsets size 1

P1 M1 M2

Itemsets size 2

P1, M2

M1, M2

P1,M2: (0,7 + 0,55 + 0) / 3 = 0,42 ←
M1,M2: (0 + 0,2 + 0,8) / 3 = 0,33 ←

Minimum support = 0,4

Frequent Itemsets size 1

P1,M2

Candidate rules

If X1 is P1 them X2 is M2

If X2 is M2 them X1 is P1

→ Conf: 0,42 / 0,42 = 1,0
→ Conf: 0,42 / 0,87 = 0,48

Minimum confidence = 0,8

Fuzzy association rules

If X1 is P1 them X2 is M2

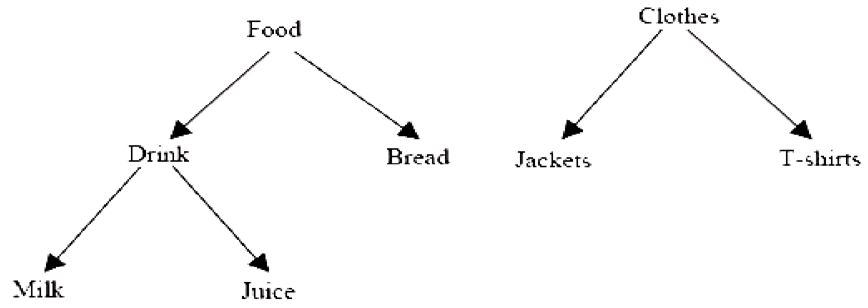
Tzung-Pei Hong, Kuei-Ying Lin, Shyue-Liang Wang
Fuzzy Data Mining for Interesting Generalized Association Rules
Fuzzy Sets and Systems 138 (2003) 255-269

- **Association rules** discovers relationships among items .
- Designing a sophisticated fuzzy data-mining algorithm able to deal with quantitative data under a given **taxonomy**.
- Transform quantitative values in transactions into **linguistic terms**,then finds interesting fuzzy rules by modifying **Srikant and Agrawal's method**

R. Srikant, R. Agrawal, Mining generalized association rules. International Conference on Very Large Databases. Zurich (Switzerland, 1995) 407-419.

Srikant and Agrawal's mining method (1/2)

1. Ancestors of items in each given transaction are added according to the predefined **taxonomy**.



2. Candidate itemsets are generated. The number of an itemset appearing in the transactions is larger than a predefined threshold value **(minimum support)**, the itemset is considered a **large itemset**.

Srikant and Agrawal's mining method (2/2)

3. Induced from the large itemsets found in the second phase with calculated **confidence values** larger than a predefined threshold **(minimum confidence)** are kept.

4. Association rules are pruned away and output :

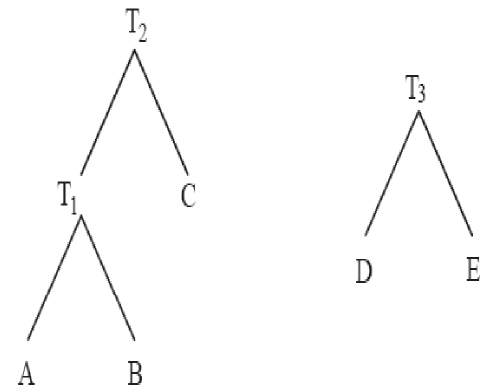
- Rules have no ancestor rules .
- Support value of a rule is R-time larger than the expected support values .
- Confidence value of a rule is R-time larger than the expected confidence values.

The fuzzy generalized mining algorithm — *Example*

The ancestors of appearing items are added.

Six transactions in this example

Transaction ID	Items
1	(Milk, 3) (Bread, 4) (T-shirt, 2)
2	(Juice, 3) (Bread, 7) (Jacket, 7)
3	(Juice, 2) (Bread, 10) (T-shirt, 5)
4	(Bread, 9) (T-shirt, 10)
5	(Milk, 7) (Jacket, 8)
6	(Juice, 2) (Bread, 8) (Jacket, 10)



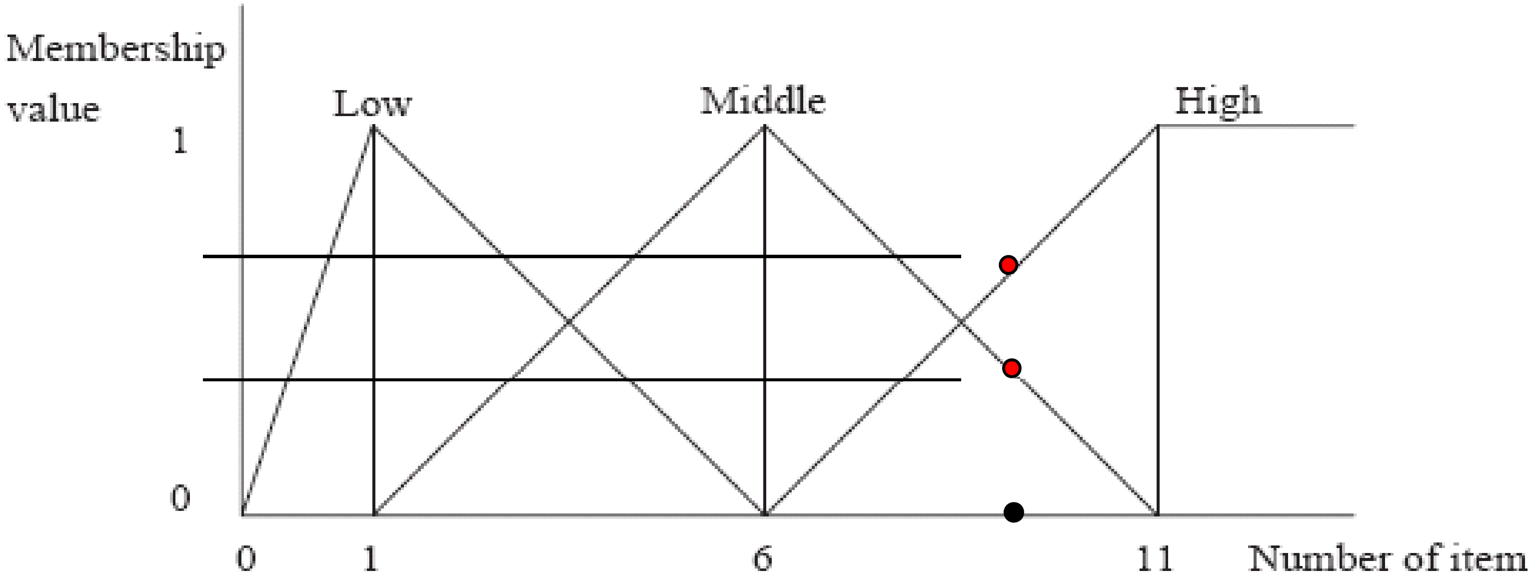
The new representation of the given taxonomy in this example.

The expanded transactions

Transaction ID	Expanded items
1	$(A,3)(C,4)(E,2)(T_1,3)(T_2,7)(T_3,2)$
2	$(B,3)(C,7)(D,7)(T_1,3)(T_2,10)(T_3,7)$
3	$(B,2)(C,10)(E,5)(T_1,2)(T_2,12)(T_3,5)$
4	$(C,9)(E,10)(T_2,9)(T_3,10)$
5	$(A,7)(D,8)(T_1,7)(T_2,7)(T_3,8)$
6	$(B,2)(C,8)(D,10)(T_1,2)(T_2,10)(T_3,10)$

The fuzzy generalized mining algorithm — *Example* (cont.)

The quantitative values of the items are represented using fuzzy sets.



(C,9) => C (0 / low , 0.4 / middle , 0.6 / high)

fuzzy set (0.0/Low + 0.4/Middle+0.6/High)

The fuzzy generalized mining algorithm — *Example* (cont.)

Fuzzy region in the transactions is calculated as the count value.

The fuzzy sets transformed from the data in Table 2

C.High as an example
cardinality=(0.2+0.8+0.6+0.4)=2.0

TID	Level-1 fuzzy set
1	$(\frac{0.6}{A.Low}+\frac{0.4}{A.Middle})(\frac{0.4}{C.Low}+\frac{0.6}{C.Middle})(\frac{0.8}{E.Low}+\frac{0.2}{E.Middle})(\frac{0.6}{T_1.Low}+\frac{0.4}{T_1.Middle})(\frac{0.8}{T_2.Middle}+\frac{0.2}{T_2.High})(\frac{0.8}{T_3.Low}+\frac{0.2}{T_3.Middle})$
2	$(\frac{0.6}{B.Low}+\frac{0.4}{B.Middle})(\frac{0.8}{C.Middle}+\frac{0.2}{C.High})(\frac{0.8}{D.Middle}+\frac{0.2}{D.High})(\frac{0.6}{T_1.Low}+\frac{0.4}{T_1.Middle})(\frac{0.2}{T_2.Middle}+\frac{0.8}{T_2.High})(\frac{0.8}{T_3.Middle}+\frac{0.2}{T_3.High})$
3	$(\frac{0.8}{B.Low}+\frac{0.2}{B.Middle})(\frac{0.2}{C.Middle}+\frac{0.8}{C.High})(\frac{0.2}{E.Low}+\frac{0.8}{E.Middle})(\frac{0.8}{T_1.Low}+\frac{0.2}{T_1.Middle})(\frac{1.0}{T_2.High})(\frac{0.2}{T_3.Low}+\frac{0.8}{T_3.Middle})$
4	$(\frac{0.4}{C.Middle}+\frac{0.6}{C.High})(\frac{0.2}{E.Middle}+\frac{0.8}{E.High})(\frac{0.4}{T_2.Middle}+\frac{0.6}{T_2.High})(\frac{0.2}{T_3.Middle}+\frac{0.8}{T_3.High})$
5	$(\frac{0.8}{A.Middle}+\frac{0.2}{A.High})(\frac{0.6}{D.Middle}+\frac{0.4}{D.High})(\frac{0.8}{T_1.Middle}+\frac{0.2}{T_1.High})(\frac{0.8}{T_2.Middle}+\frac{0.2}{T_2.High})(\frac{0.6}{T_3.Middle}+\frac{0.4}{T_3.High})$
6	$(\frac{0.8}{B.Low}+\frac{0.2}{B.Middle})(\frac{0.6}{C.Middle}+\frac{0.4}{C.High})(\frac{0.2}{D.Middle}+\frac{0.8}{D.High})(\frac{0.8}{T_1.Low}+\frac{0.2}{T_1.Middle})(\frac{0.2}{T_2.Middle}+\frac{0.8}{T_2.High})(\frac{0.2}{T_3.Middle}+\frac{0.8}{T_3.High})$

The fuzzy generalized mining algorithm — *Example* (cont.)

The fuzzy region with the highest count for each item is found.

The counts of the fuzzy regions

Item	Count	Item	Count	Item	Count	Item	Count
A.Low	0.6	*C.Low*	0.4	*E.Low*	1.0	T_2*.Low*	0.0
A.Middle	1.2	*C.Middle*	2.6	*E.Middle*	1.2	T_2*.Middle*	2.4
A.High	0.2	*C.High*	2.0	*E.High*	0.8	T_2*.High*	3.6
B.Low	2.2	*D.Low*	0.0	T_1*.Low*	2.8	T_3*.Low*	1.0
B.Middle	0.8	*D.Middle*	1.6	T_1*.Middle*	2.0	T_3*.Middle*	2.8
B.High	0.0	*D.High*	1.4	T_1*.High*	0.2	T_3*.High*	2.2

The fuzzy generalized mining algorithm — *Example* (cont.)

The count of any region selected in Step 4 is larger than 1.5 put in L1

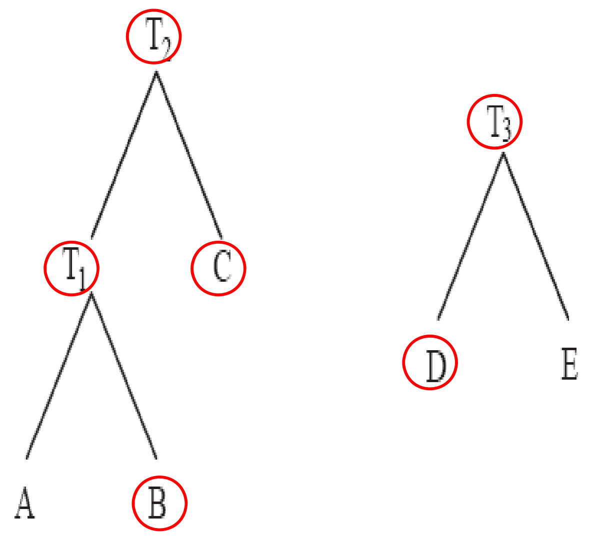
The set of large 1-itemsets in this example

Itemset	Count
$B.Low$	2.2
$C.Middle$	2.6
$D.Middle$	1.6
$T_1.Low$	2.8
$T_2.High$	3.6
$T_3.Middle$	2.8

Minimum Support = 1.5

The fuzzy generalized mining algorithm — *Example* (cont.)

The candidate set C2 is generated from L1 and
the items of C2 must not have ancestor or descendant relation in the taxonomy



The new representation of the given taxonomy in this example.

→

(B.Low , C.Middle)
(B.Low , D.Middle)
(B.Low , T3.Middle)
(C.Middle , D.Middle)
(C.Middle , T1.Low)
(C.Middle , T3.Middle)
(D.Middle , T1.Low)
(D.Middle , T2.High)
(T1.Low , T3. Middle)
(T2. High , T3. Middle)

The fuzzy generalized mining algorithm — *Example* (cont.)

The fuzzy membership values of each transaction data for the candidate 2-itemsets are calculated.

1	$\left(\frac{0.6}{A.Low}+\frac{0.4}{A.Middle}\right)\left(\frac{0.4}{C.Low}+\frac{0.6}{C.Middle}\right)\left(\frac{0.8}{E.Low}+\frac{0.2}{E.Middle}\right)\left(\frac{0.6}{T_1.Low}+\frac{0.4}{T_1.Middle}\right)\left(\frac{0.8}{T_2.Middle}+\frac{0.2}{T_2.High}\right)\left(\frac{0.8}{T_3.Low}+\frac{0.2}{T_3.Middle}\right)$
2	$\left(\frac{0.6}{B.Low}+\frac{0.4}{B.Middle}\right)\left(\frac{0.8}{C.Middle}+\frac{0.2}{C.High}\right)\left(\frac{0.8}{D.Middle}+\frac{0.2}{D.High}\right)\left(\frac{0.6}{T_1.Low}+\frac{0.4}{T_1.Middle}\right)\left(\frac{0.2}{T_2.Middle}+\frac{0.8}{T_2.High}\right)\left(\frac{0.8}{T_3.Middle}+\frac{0.2}{T_3.High}\right)$

The membership values for $B.Low \cap C.Middle$

TID	$B.Low$	$C.Middle$	$B.Low \cap C.Middle$
T1	0.0	0.6	0.0
T2	0.6	0.8	0.6
T3	0.8	0.2	0.2
T4	0.0	0.4	0.0
T5	0.0	0.0	0.0
T6	0.8	0.6	0.6

0.0+0.6+0.2+0.0+
0.0+0.6 = 1.4

The fuzzy generalized mining algorithm — *Example* (cont.)

The scalar cardinality (count) of each candidate 2-itemset in C2

The fuzzy counts of the itemsets in C_2

Itemset	Count
$(B.Low, C.Middle)$	1.4
$(B.Low, D.Middle)$	0.8
$(B.Low, T_3.Middle)$	1.6
$(C.Middle, D.Middle)$	1.0
$(C.Middle, T_1.Low)$	1.2
$(C.Middle, T_3.Middle)$	1.6
$(D.Middle, T_1.Low)$	1.2
$(D.Middle, T_2.High)$	1.2
$(T_1.Low, T_3.Middle)$	1.6
$(T_2.High, T_3.Middle)$	2.4

Minimum Support = 1.5

The fuzzy generalized mining algorithm — *Example* (cont.)

The confidence value for the association rule is calculated

- (B.Low , T3.Middle)
  If B=Low then T3=Middle
  If T3=Middle then B=Low
- (C.Middle , T3.Middle)
  If C=Middle then T3=Middle
  If T3=Middle then C=Middle
- (T1.Low , T3.Middle)
  If T1=Low then T3=Middle
  If T3=Middle then T1=Low
- (T2.High , T3.Middle)
  If T2=Low then T3=Middle
  If T3=Middle then T2=Low

⇨ Confidence :

$$\frac{\sum_{i=1}^{6}(B.Low \cap T_3.Middle)}{\sum_{i=1}^{6}(B.Low)} = \frac{1.6}{2.2} = 0.73$$

The fuzzy generalized mining algorithm — *Example* (cont.)

Given confidence threshold is set at 0.7. The following three rules are kept

If B=Low then T3=Middle (confidence value = 0.73)
If T3=Middle then B=Low (confidence value = 0.53)
If C=Middle then T3=Middle (confidence value = 0.62)
If T3=Middle then C=Middle (confidence value = 0.57)
If T1=Low then T3=Middle (confidence value = 0.8)
If T3=Middle then T1=Low (confidence value = 0.57)
If T2=Low then T3=Middle (confidence value = 0.67)
If T3=Middle then T2=Low (confidence value = 0.86)

Minimum Confidence = 0.7

The fuzzy generalized mining algorithm — *Example* (cont.)

Its close ancestor rule mined out is “If T1 =Low, then T3 =Middle”.
The support interest measure of the rule “If B=Low, then T3 =Middle” is alculated

If B=Low then T3=Middle

If T1=Low then T3=Middle

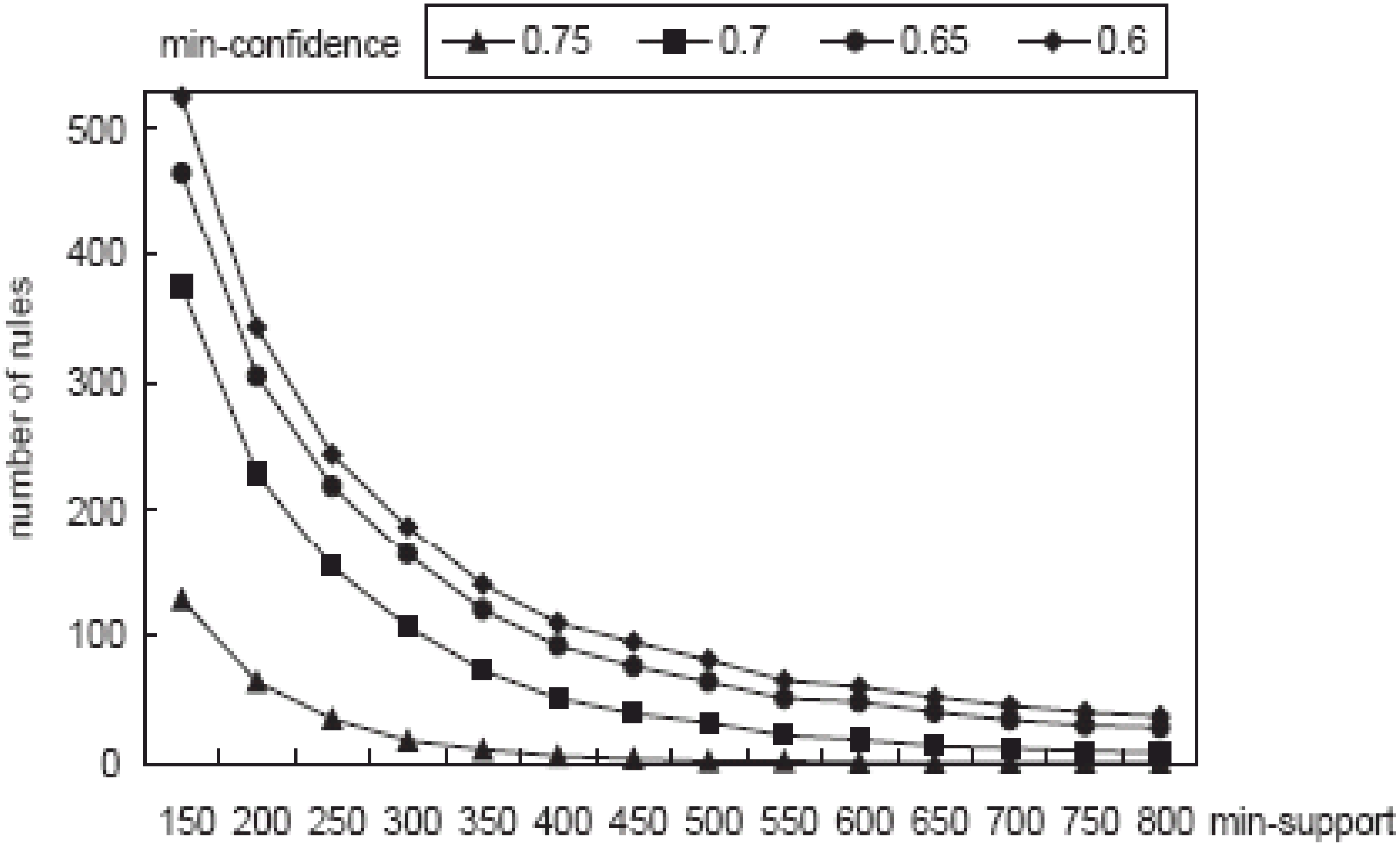
If T3=Middle then T2=Low

$$I_{support}(B.Low \rightarrow T_3.Middle) = \frac{count_{B.Low \rightarrow T_3.Middle}}{\frac{count_{B.Low}}{count_{T_1.Low}} \times \frac{count_{T_3.Midde}}{count_{T_3.Middle}} \times count_{T_1.Low \rightarrow T_3.Midde}} = \frac{1.6}{\frac{2.2}{2.8} \times \frac{2.8}{2.8} \times 1.6} = 1.27.$$

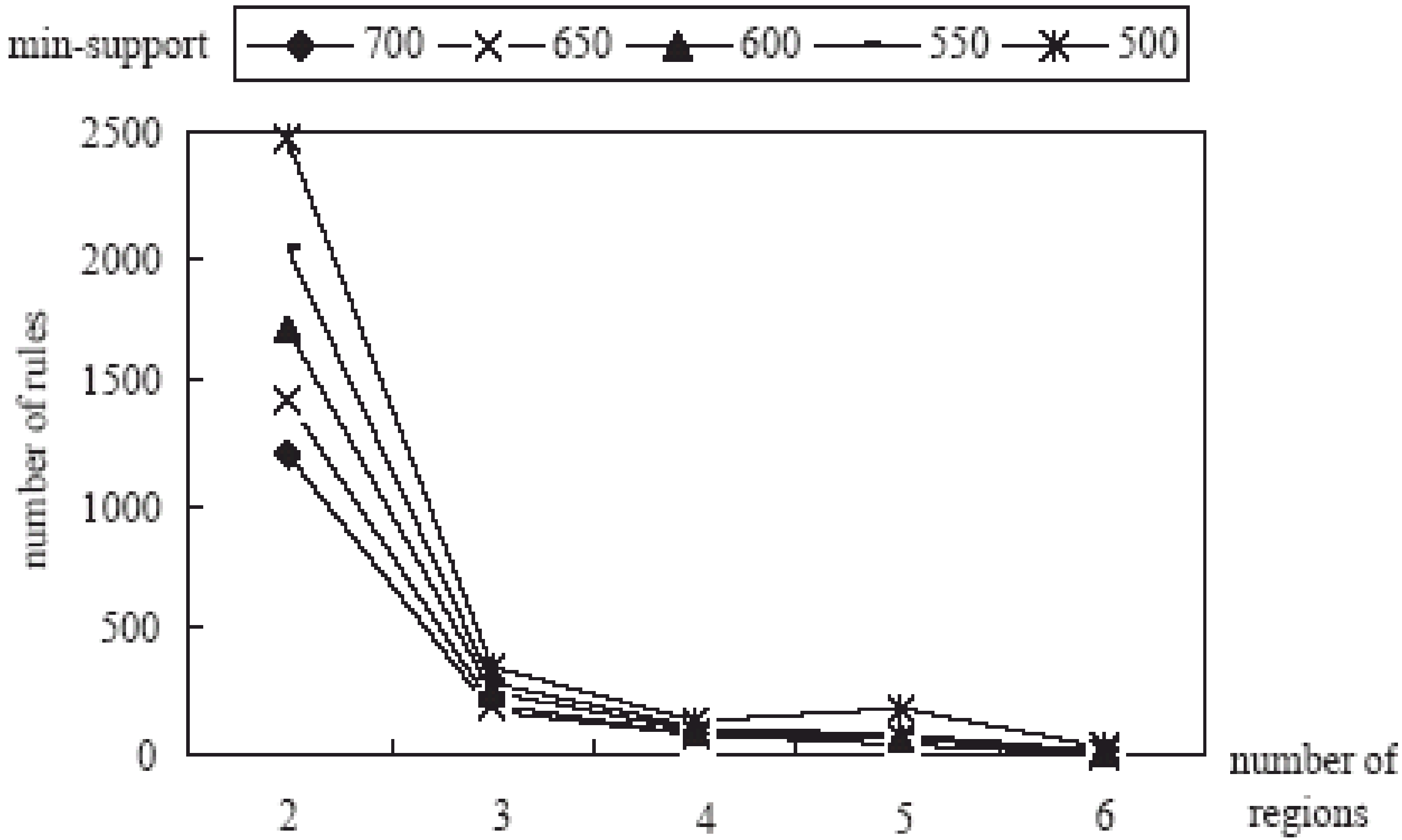
$$I_{confidence}(B.Low \rightarrow T_3.Middle) = \frac{confidence_{B.Low \rightarrow T_3.Middle}}{\frac{count_{T_3.Middle}}{count_{T_3.Middle}} \times confidence_{T_1.Low \rightarrow T_3.Middle}} = \frac{0.73}{\frac{2.8}{2.8} \times 0.8} = 0.9.$$

interest threshold R =1.5

Experimental results



Experimental results



Fuzzy Association Rules: Comments

Gets smoother mining results due to its fuzzy membership characteristics.

Advanced approach to be presented later:

J. Alcalá-Fdez, R. Alcalá, M.J. Gacto, F. Herrera, **Learning the Membership Function Contexts for Mining Fuzzy Association Rules by Using Genetic Algorithms**. *Fuzzy Sets and Systems,* *doi:10.1016/j.fss.2008.05.012* ***, in press (2008)***.

Fuzzy Data Mining: Final Comments

- Fuzzy sets/logic is a useful form of knowledge representation, allowing for approximate but natural expression of some types of knowledge.
- An alternative way is to include uncertainty of input data while using crisp logic rules.
- Results may sometimes be better than with other systems since it is easier to include a priori knowledge in fuzzy systems.
- Adaptation of fuzzy rule parameters leads to neurofuzzy systems, genetic fuzzy systems, …

Outline

- ✓Introduction: Soft Computing Techniques in Data Mining
- ✓Fuzzy Data Mining
- ✓Evolutionary Data Mining
- ✓Concluding Remarks

Evolutionary algorithms and Data Mining

- EAs were not specifically designed as data mining techniques, as other approaches like neural networks.

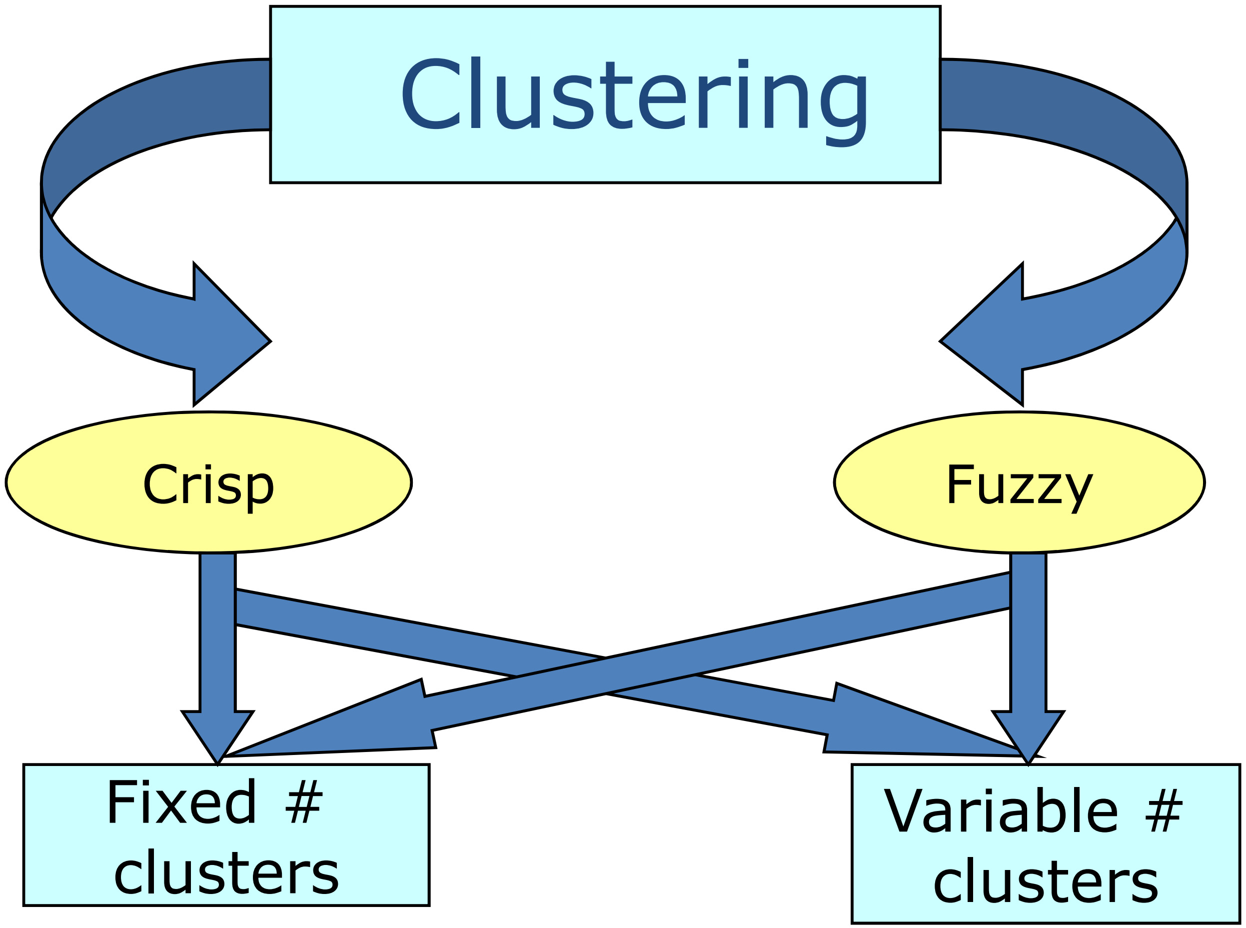
  However, it is well known that a learning task can be modelled as an optimization problem, and thus solved through evolution.

- Their powerful search in complex, ill-defined problem spaces have permitted applying EAs successfully to a huge variety of machine learning and knowledge discovery tasks.

- Their flexibility and capability to incorporate existing knowledge are also very interesting characteristics for the problem solving.

Evolutionary algorithms and Data Mining

- **Evolutionary clustering**
- **Evolutionary feature/instance selection**
- **Evolutionary rule learning**
- **Genetic programming – tree based coding**
- **Evolutionary neural networks**
- **Evolutionary optimization based data mining processes, ...**



"Genetic Algorithm Based Clustering Technique'', Patt. Recog., 33, 1455-1465, 2000.

"Clustering using Simulated Annealing with Probabilistic Redistribution'' Int. Journal of Pattern Recognition and Artificial Intelligence, vol 15, no. 2, pp. 269-285, 2001.

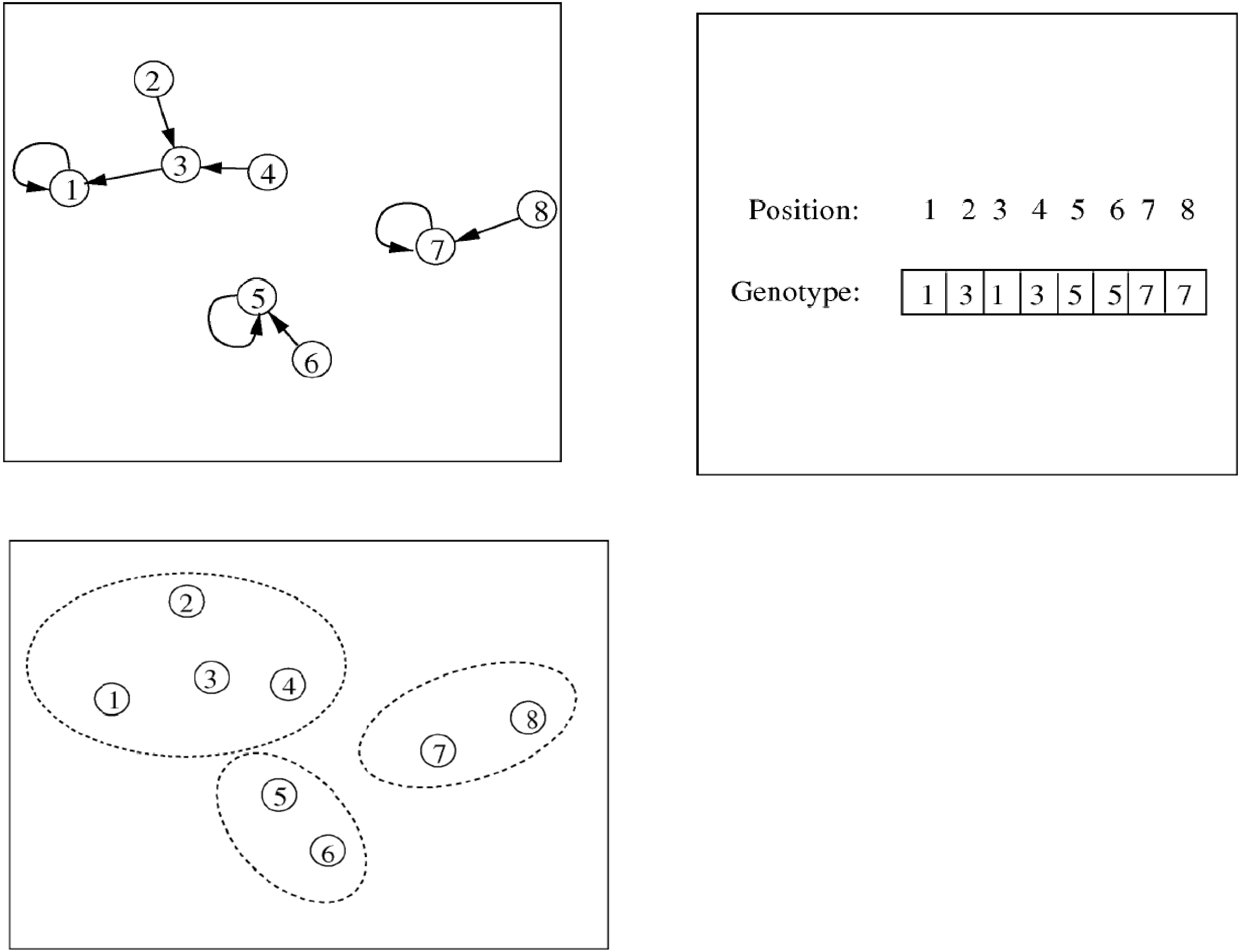
"Non-parametric Genetic Clustering : Comparison of Validity Indices'', IEEE Trans. on Systems, Man and Cybernetics Part-C, vol. 31, no. 1, pp. 120-125, 2001.

''Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification'', Pattern Recognition.

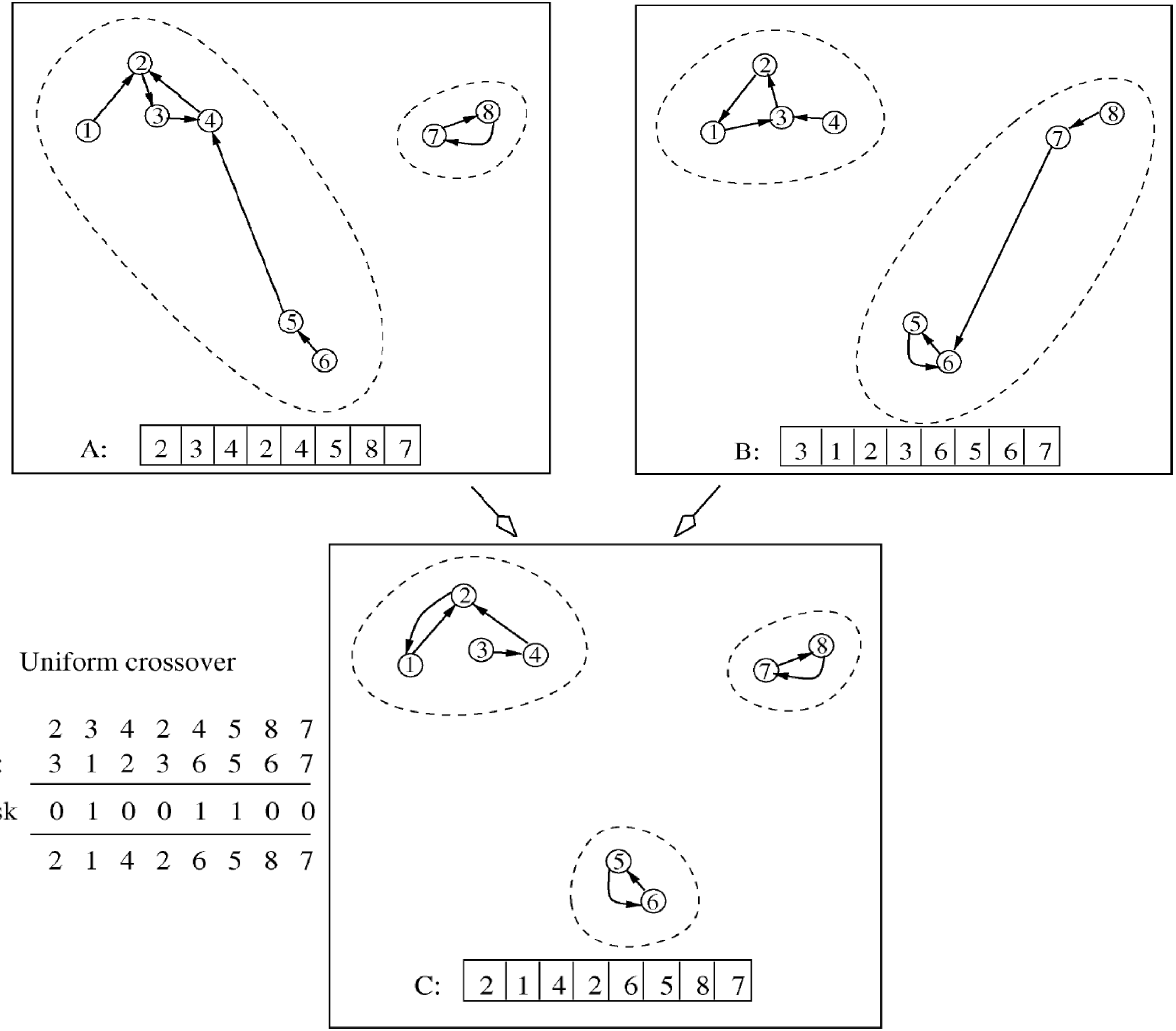
Evolutionary Clustering

Example: Locus-based adjacency representation

- Genetic representation and operators
  - Locus-based adjacency representation
    - No need to fix the number of clusters
    - Well-suited for standard crossover operators
  - Uniform crossover
    - One-point or two point
  - Neighborhood-biased mutation operator
    - Quickly discard unfavorable links
    - Explore feasible solutions



Position: 1 2 3 4 5 6 7 8
Genotype: 1 3 1 3 5 5 7 7

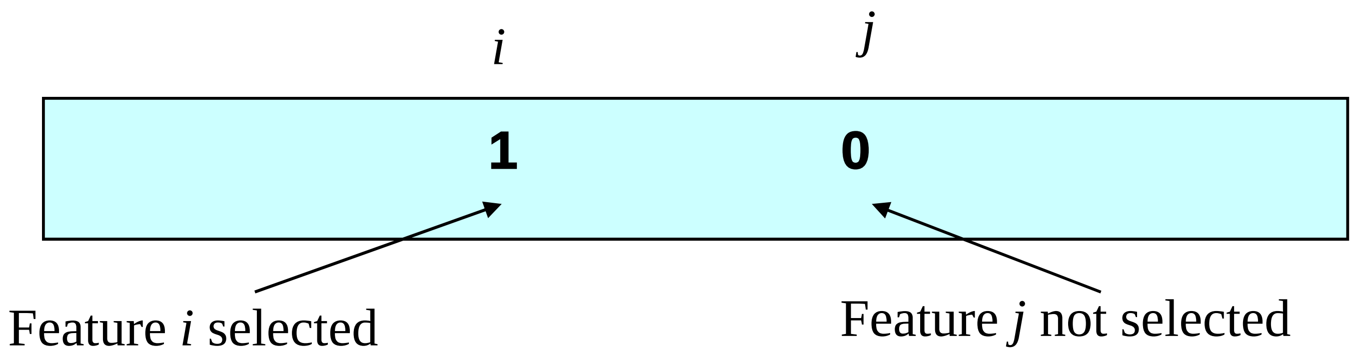


Uniform crossover

A:	2	3	4	2	4	5	8	7
B:	3	1	2	3	6	5	6	7
mask	0	1	0	0	1	1	0	0
C:	2	1	4	2	6	5	8	7

Feature Selection

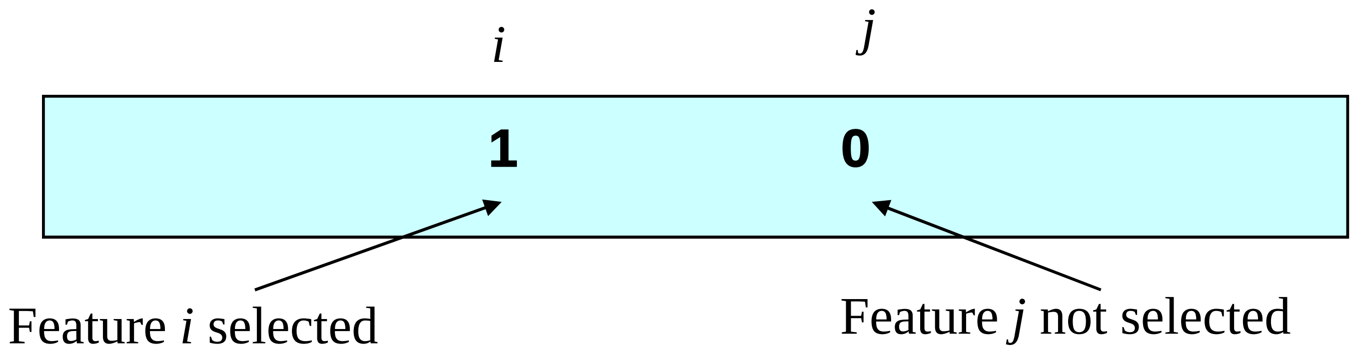
- GA encodes the features



- The two objectives computed and kept separate
- Non dominated sorting for ranking the individuals
- Genetic operations
- Outputs a set of solutions

Evolutionary Instance Selection

- GA encodes the features



- The two objectives computed and kept separate
- Non dominated sorting for ranking the individuals
- Genetic operations
- Outputs a set of solutions

Evolutionary Prototype Selection

- Application of Evolutionary Algorithms to PS Problem (EPS).
- Representation of solutions:
  - A chromosome consists on *n* genes with two posible states: 0 and 1 (*n* is the number of instances in training data).
  - Example:

0	1	1	1	0	0	1	0	0	1

Evolutionary Prototype Selection

- Fitness Function:
  - It combines two values: classification rate (*clas_rat*) and percentage f reduction (*perc_red*)

$$Fitness(s) = \alpha \cdot clas\_rat + (1 - \alpha) \cdot perc\_red$$

- Cano et al. (2003) studied four models:
  - Generational Genetic Algorithm (GGA)
  - Steady-State Genetic Algorithm (SGA)
  - Heterogeneous Recombination and Cataclysmic Mutation (CHC)
  - Population Based Incremental Learning (PBIL)

Evolutionary Rule Learning

A chromosome codes a rule or a set of rules.

Each gen codes a rule or a rule part:

IF $cond_1 \wedge ... \wedge cond_n$

THEM Class = C_i

Evolutionary Rule Learning

Some kind of rules

Interval rules

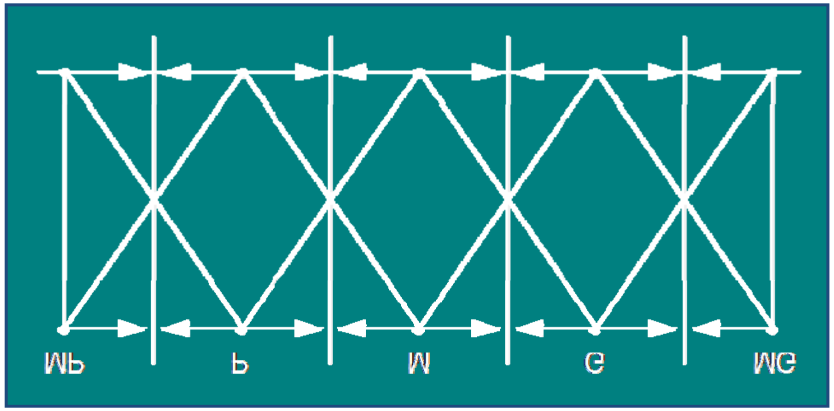
IF $x_1 \in [a_1,b_1] \wedge \ldots \wedge x_n \in [a_n,b_n]$
Then Class = C_i

Fuzzy rules

IF x_1 es M $\wedge \ldots \wedge$ x_n es G
Then Clase = C_i with certain degree r_i

Reference

GENETIC FUZZY SYSTEMS.
Evolutionary Tuning and Learning of Fuzzy Knowledge Bases.
O. Cordón, F. Herrera, F. Hoffmann, L. Magdalena
World Scientific, Julio 2001. ISBN 981-02-4016-3



Evolutionary Rule Learning

There exists four models according to the way to code the rules for learning rule bases (they are known as Genetics-based machine learning)

Chromosome = Rule Base

Model Pittsburgh: GASSIST, CORCORAN, GIL

Chromosome = Rule

Model Michigan (XCS, UCS)

LCS - Learning Classifier Systems

IRL Model – Iterative Rule Learning (SIA, HIDER)

GCCL Model – Genetic Cooperative-Competitive Learning (REGAL, LOGEMPRO)

Evolutionary Rule Learning

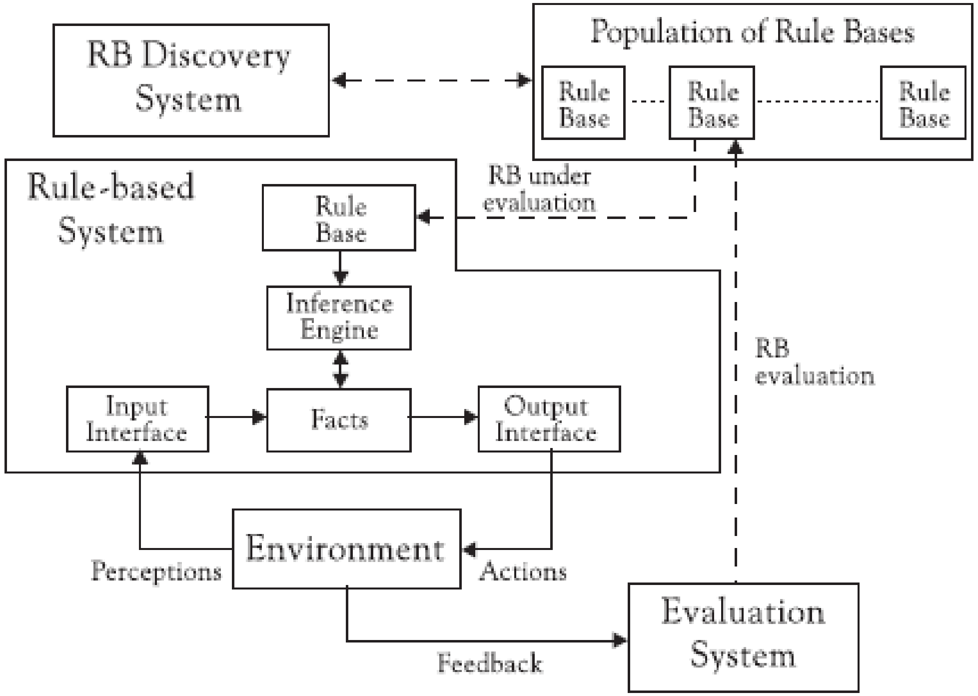
Pittsburgh Learning Approach:

Each chromosome encodes a whole rule set and the derived RB is the best individual of the last population.

The fitness function evaluates the performance at the complete RB level. However, the search space is huge, thus making difficult the problem solving and requiring sophisticated designs.

Mainly used in

off-line learning.



Evolutionary Rule Learning

Michigan Learning Approach (Learning Classifier Systems):

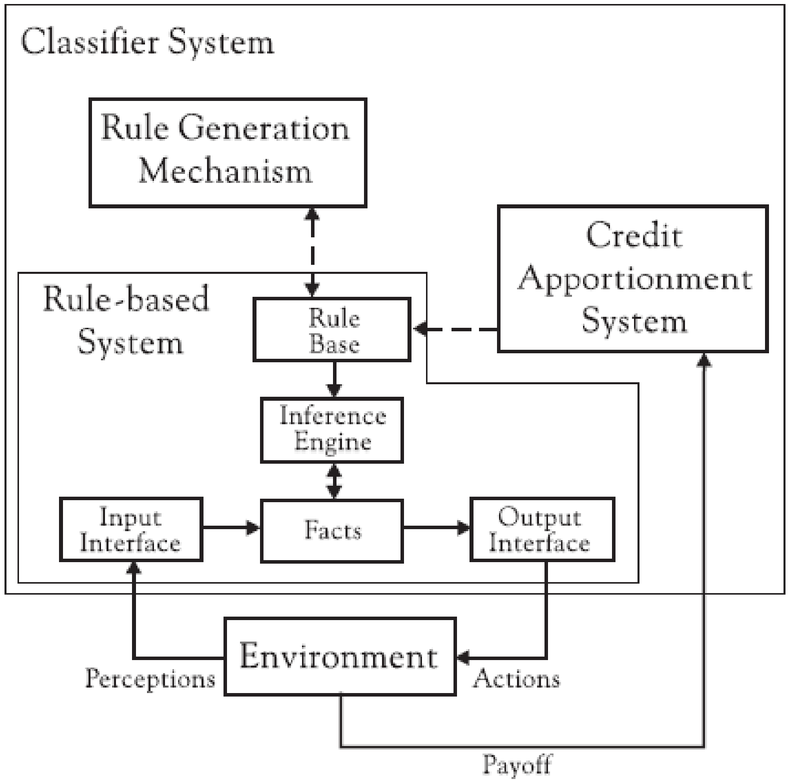
Each chromosome encodes a single rule.

Reinforcement mechanisms (reward (credit apportion) and weight penalization) are considered to adapt rules through a GA

Low weight (bad performing) rules are substituted by new rules generated by the GA.

The key question is to induce collaboration in the derived RB as the evaluation procedure is at single rule level (cooperation vs. competition problem (CCP)).

Mainly used in on-line learning.



Evolutionary Rule Learning

Iterative Rule Learning Approach:

Intermediate approach between the Michigan and Pittsburgh ones, based on partitioning the learning problem into several stages and leading to the design of multi-stage learning.

As in the Michigan approach, each chromosome encodes a single rule, but a new rule is learnt by an iterative rule generation stage and added to the derived RB, in an iterative fashion, in independent and successive runs of the GA.

The evolution is guided by data covering criteria (rule competition). Some of them are considered to penalize the generation of rules covering examples already covered by the previously generated fuzzy rules (soft cooperation).

Evolutionary Rule Learning

Iterative Rule Learning Approach:

A second post-processing stage is considered to refine the derived RB by selecting the most cooperative rule set and/or tuning the membership functions (cooperation induction)

Hence, the CCP is solved taking the advantages of both the Michigan and Pittsburgh approaches (small search space and good chances to induce cooperation)

Mainly used in off-line learning (modeling and classification applications).

Evolutionary Rule Learning

Iterative Rule Learning Approach:

A second post-processing stage is considered to refine the derived RB by selecting the most cooperative rule set and/or tuning the membership functions (cooperation induction)

Hence, the CCP is solved taking the advantages of both the Michigan and Pittsburgh approaches (small search space and good chances to induce cooperation)

Mainly used in off-line learning (modeling and classification applications).

Evolutionary Rule Learning

Cooperative-competitive learning Approach:

The *GCCL* approach, in which the complete population or a subset of it encodes the RB.

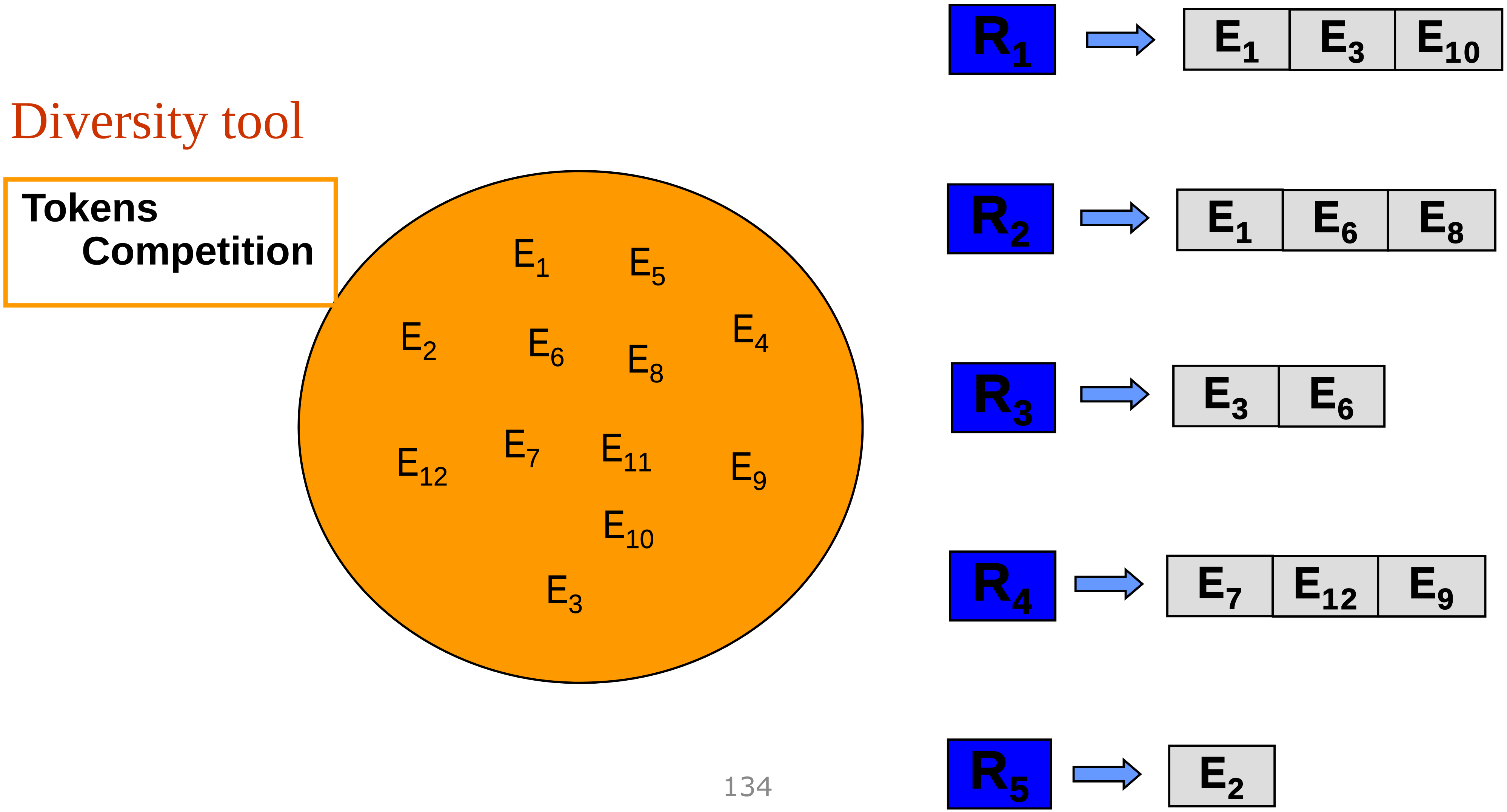
In this model the chromosomes compete and cooperate simultaneously.

COGIN (Greene and Smith 1993), REGAL (Giordana and Neri 1995) and LOGENPRO (Wong and Leung 2000) are examples with this kind of representation.

Evolutionary Rule Learning

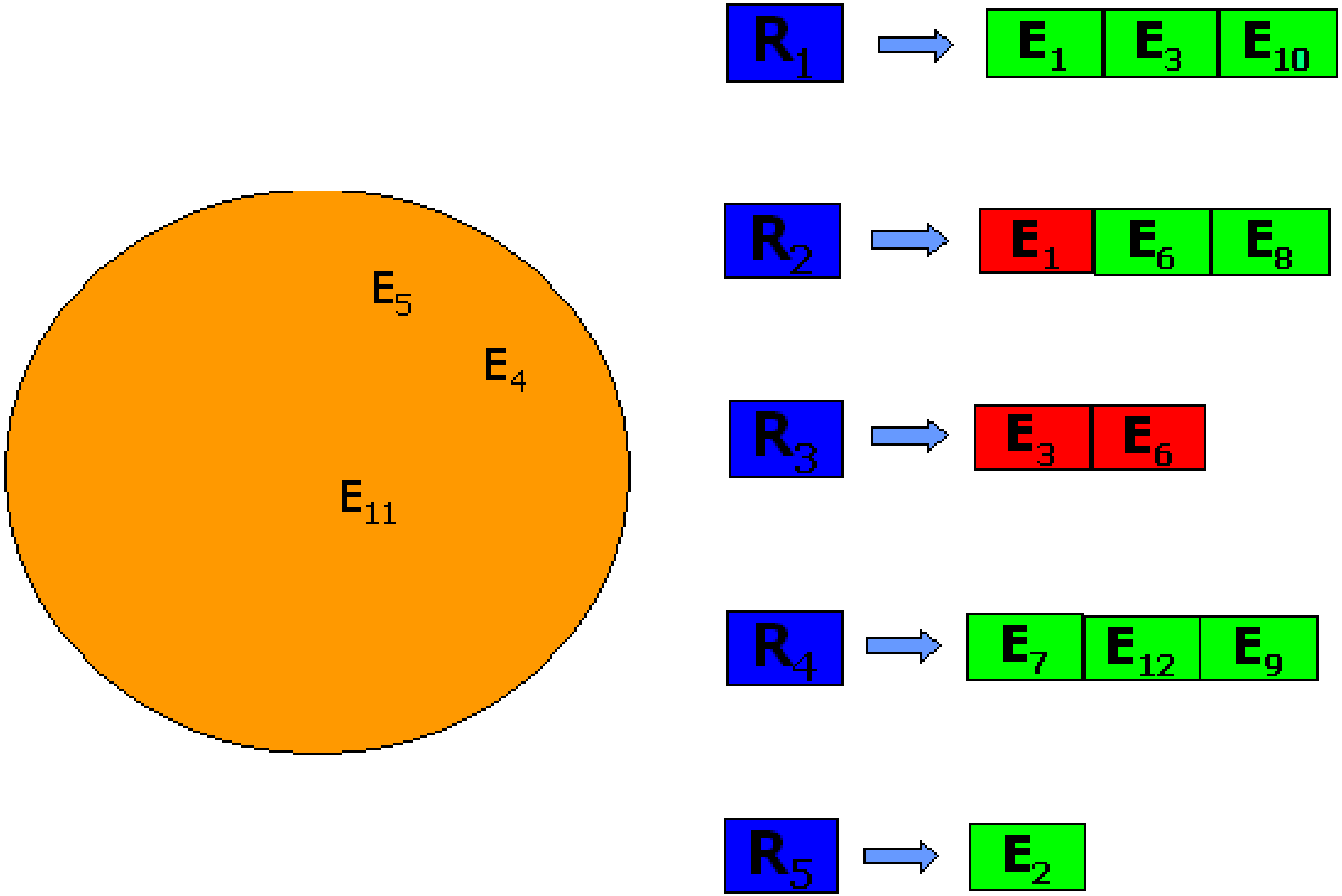
Cooperative-competitive learning Approach:

Diversity tool



Evolutionary Rule Learning

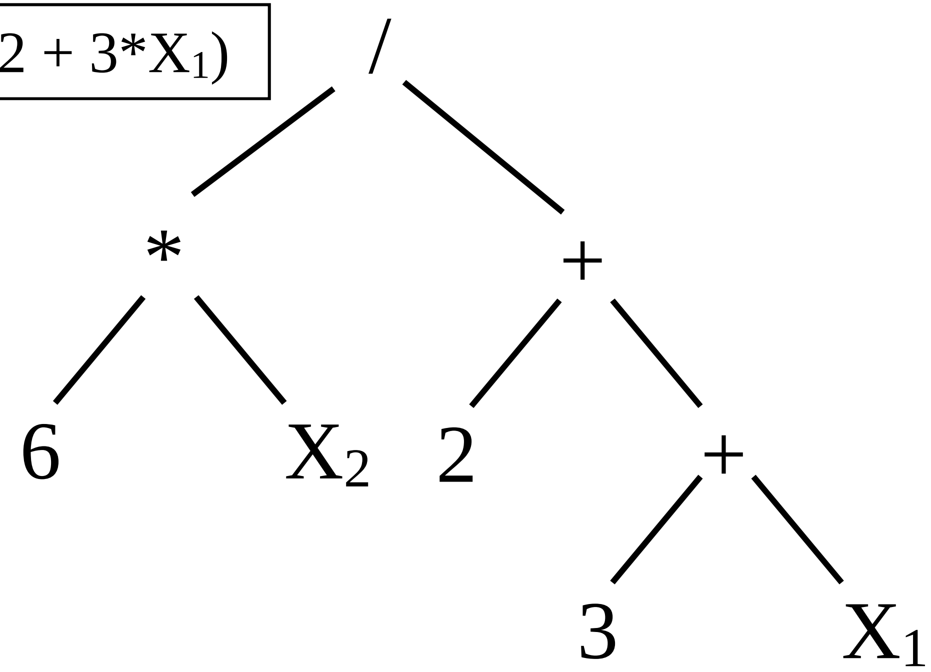
Cooperative-competitive learning Approach:



Genetic programming – tree based coding

Genetic Programming

$g = 6*X_2 / (2 + 3*X_1)$



- Symbols: {X_i, constants}
- Functions: { +, -, *, /, $\sqrt{\ }$, ...}

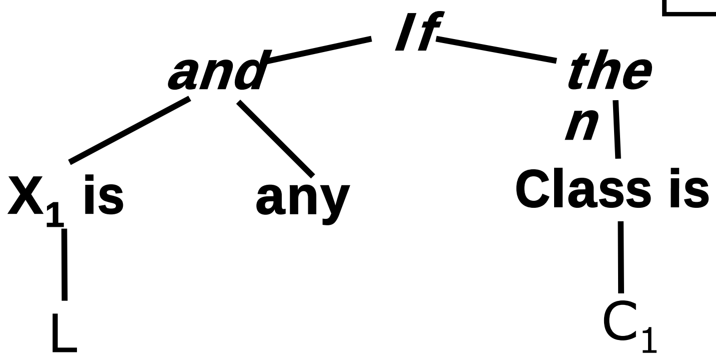
We can use Genetic Programming for learning regression functions, discriminant functions, ...

Genetic programming – tree based coding

Genetic Programming

Free grammar for designing rules

- start → [If], antec, [then], conseq, [.]
- antec → descriptor1, [and], descriptor2.
- descriptor1 → [any].
- descriptor1 → [X_1 is] label.
- descriptor2 → [any].
- descriptor2 → [X_2 is] label.
- label → {member(?a, [L, M, H, L or M, L or H, M or H, L or M or H])}, [?a].
- conseq → [Class is] descriptorClass
- descriptorClass → {member(?a, [C_1, C_2, C_3])}, [?a].



If X_1 is L *then* Class is C_1

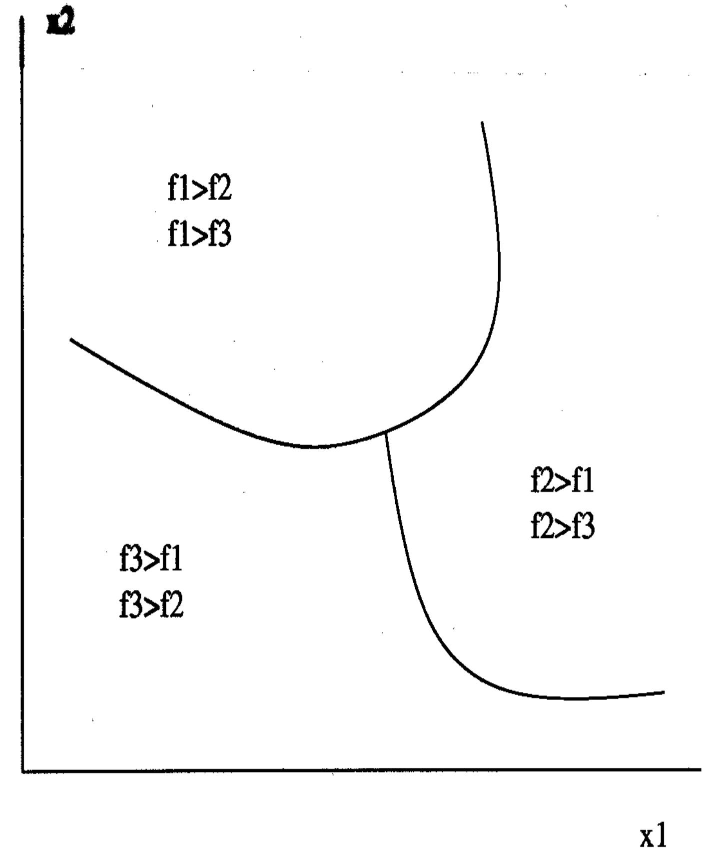
Genetic programming – tree based coding

Classification based on discriminant functions

Model 1:

To get M functions

$f_1, \ldots, f_M$ such that $f_k(x) > f_i(x),\ i \neq k$ when x belongs to class k.



Genetic programming – tree based coding

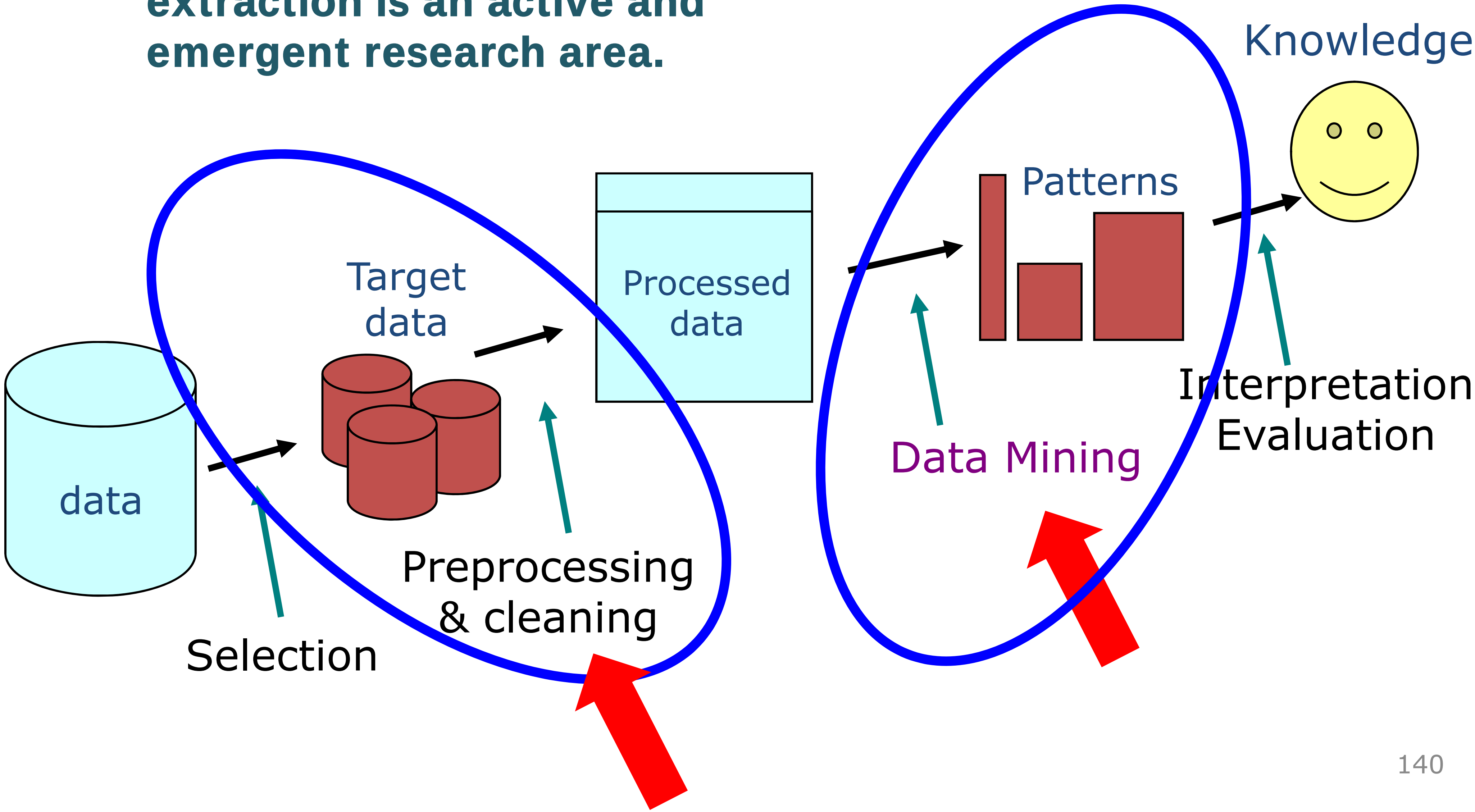
Classification based on discriminant functions

Model 2:

IF $f_1(a) > 0$ Them class 1
else
    IF $f_2(a) > 0$ Them clase 2
    else
        ....
        IF $f_M(a) > 0$ Them class M
        else
            the pattern can not be classified

Evolutionary Data Mining: Final Comments

The use of Evolutionary Algorithms as toof for knowledge extraction is an active and emergent research area.



Outline

- ✓ Introduction: Soft Computing Techniques in Data Mining
- ✓ Fuzzy Data Mining
- ✓ Evolutionary Data Mining
- ✓ Concluding Remarks

Concluding Remarks

Soft Computing based techniques provide useful tools for data mining, making use of their main features:

- Commonsense knowledge may sometimes be captured in an natural way using fuzzy rules.
- ANN: Machinery for learning and curve fitting (Learns from examples)
- GAs are *Appropriate* and *Natural Choice* for problems which need – Optimizing Computation Requirements, and Robust, Fast and Close Approximate Solutions

Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.