5-th international scientific conference
Applied information and communication technologies
26-27 April 2012
Latvia University of Agriculture

# POSSIBILITIES OF APPLYING CLUSTERING ALGORITHMS IN DATA ANALYSIS

Pēteris Grabusts
Rezekne Higher Educational Institution

April 26, 2012

# RULE EXTRACTION METHODS

**The motivation for studying rule extraction methods through clustering is the following:**

- **the amount of multidimensional data to be analysed becomes too large for the potentialities of statistical analysis;**
- **popular neural network methods operate by the black box principle that complicates interpretation of the results for the user;**
- **previously unknown regularities are present in the data;**
- **the regularities found can be represented in a way that is easy to perceive and understand for the user.**

# RULE EXTRACTION METHODS

The main requirement which is put forward to the results of data analysis is that the results must always be interpreted as correctly as possible. The rules that represent the regularities found have to be stated as simple and easy to understand logical expressions. Namely, they must look as these logical rules:

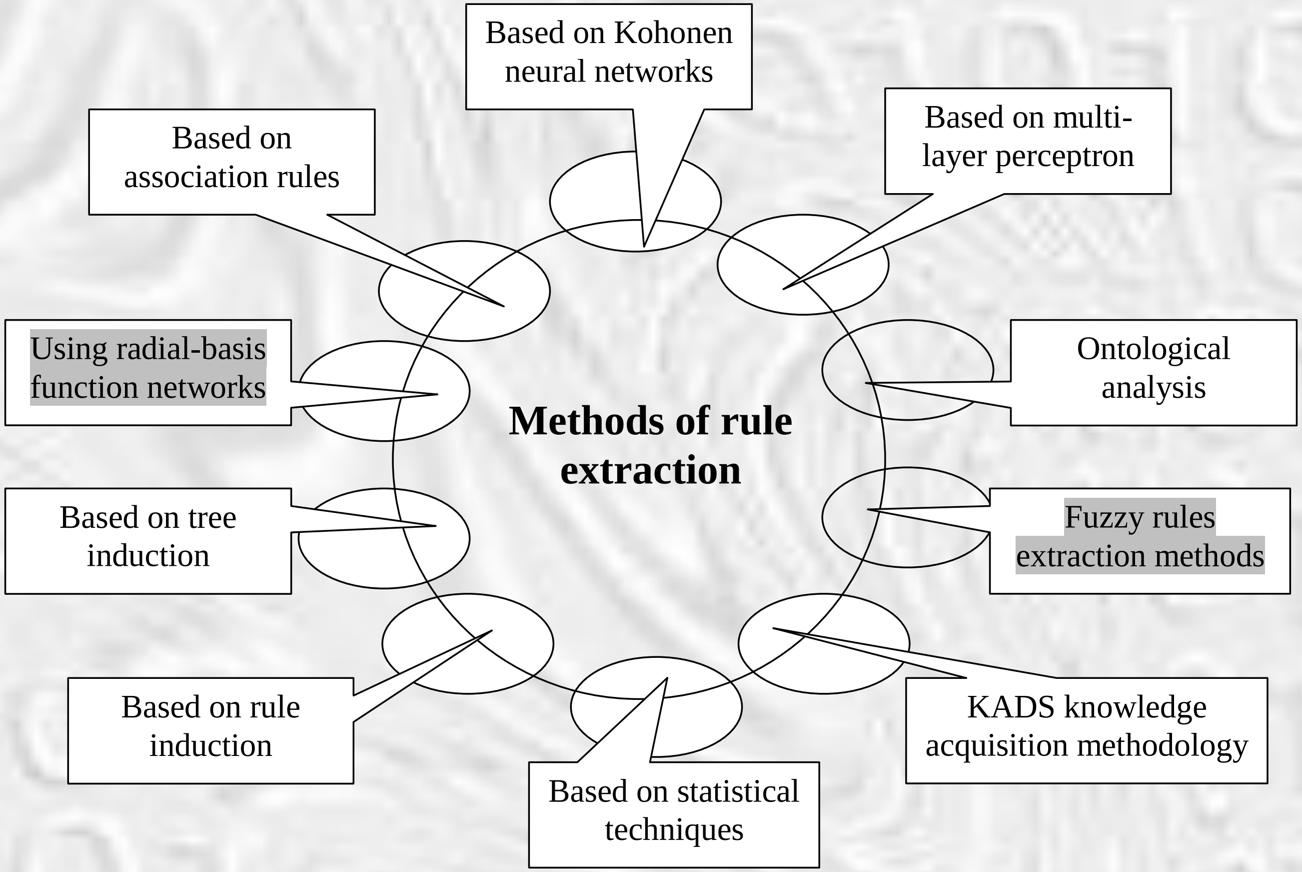
IF {(Event 1) AND (Event 2) AND ... (Event N)} THEN ...

In what follows, the author will employ logical conditional rules (production rules) of this kind:

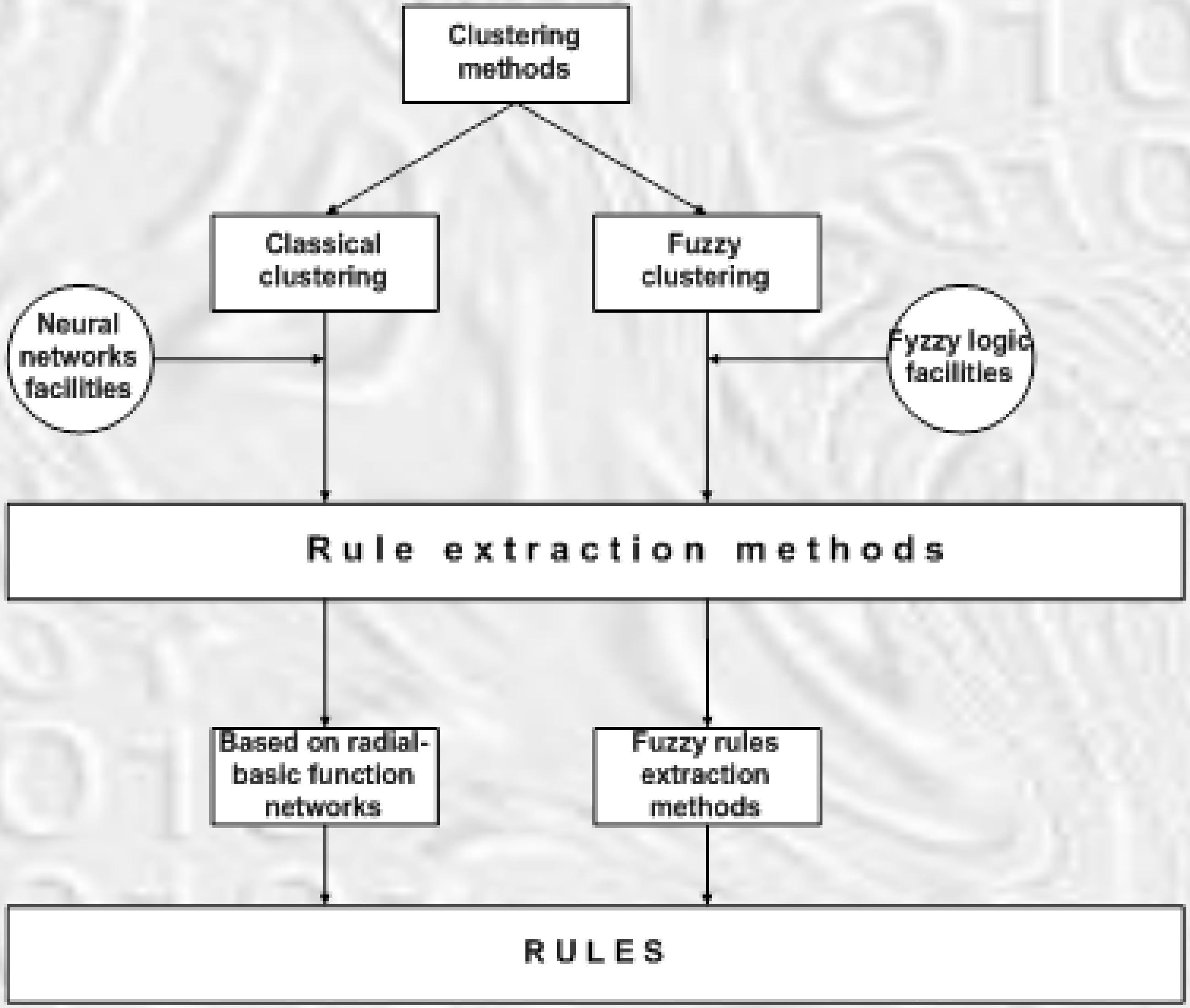
$$IF\ \underbrace{(Antecedent\,1)\ and\ (Antecedent\ 2)\ and ... (Antecedent\ N)}_{A}\ THEN\ \underbrace{(Consequent)}_{B}\ .$$

# RULE EXTRACTION METHODS

- Expert Systems
- Rule-based Classifier
- Classifier based on fuzzy rules
- Knowledge-based Agent Control
- Pattern Recognition
- Rule-based Forecasting
- Rule-based Prediction
- Tool to detect Patterns in databases
- Tool to detect Regularities in databases
- Tool for extraction of useful knowledge from raw data
- Genetic-based Machine Learning



# RULE EXTRACTION METHODS

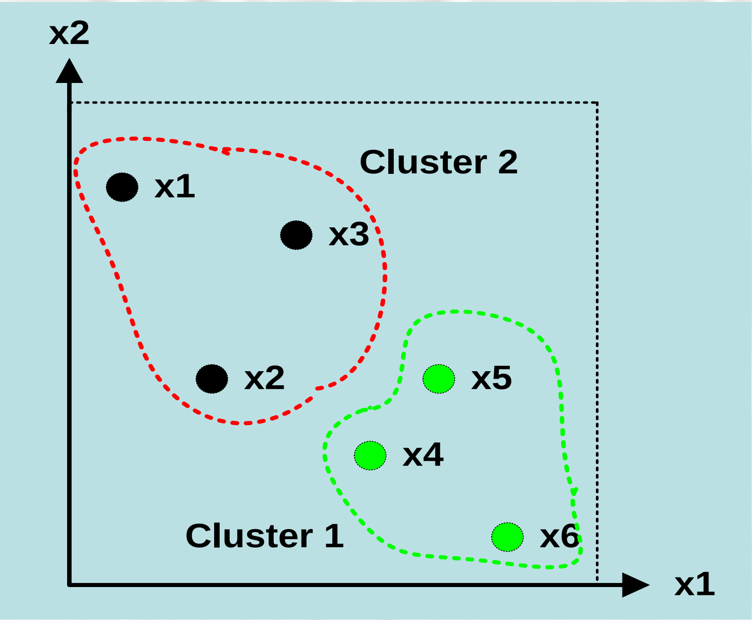


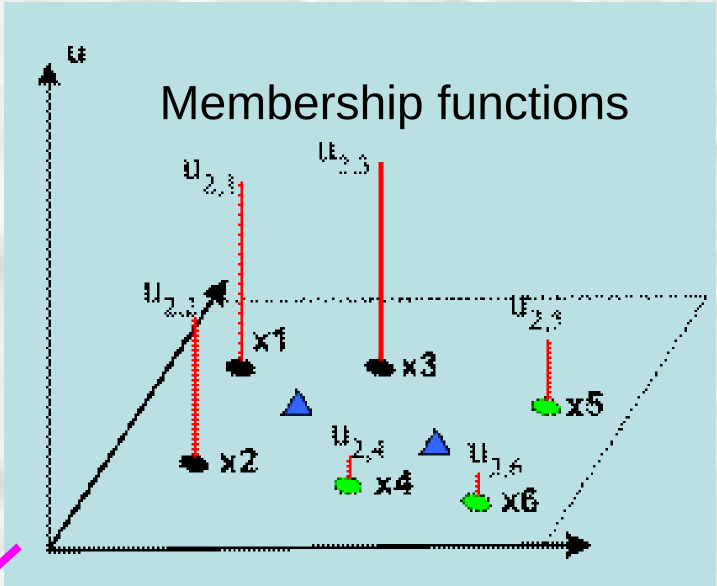
- **FUZZY RULE BASE DESIGN METHOD**
- **EXTRACTING RULES FROM TRAINED RBF NEURAL NETWORK**

(Using clustering !)

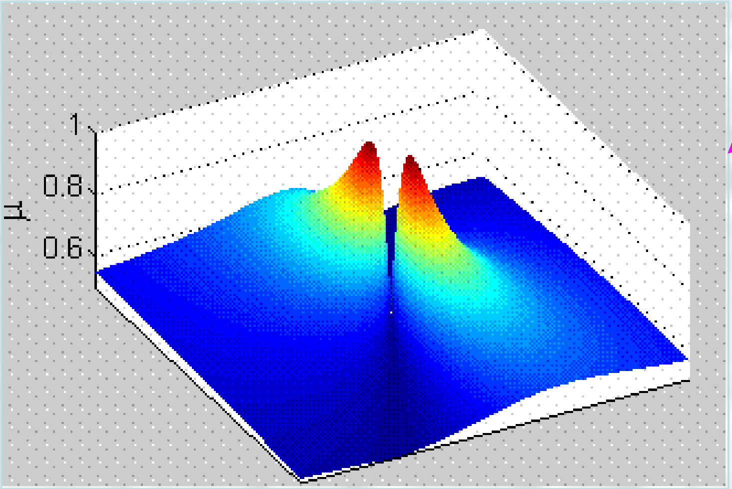
# I. FUZZY RULE BASE DESIGN

# FUZZY CLUSTERING




Distribution of membership functions



# FCM

The fuzzy c-means algorithm allows each data point to belong to a cluster to a degree specified by a membership grade, and thus each point may belong to several clusters.

- Membership matrix $m_{ik}$ - [0,1] :

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

- Objective function is :

$$J(M, c_1, c_2, \ldots c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{k=1}^{K} m_{ik}^q d_{ik}^2$$

- Optimal center:

$$c_i = \frac{\sum_{k=1}^{K} m_{ik}^q u_k}{\sum_{k=1}^{K} m_{ik}^q}$$

# FCM ALGORITHM

**FCM algorithm**

(1) Initialise the membership matrix M with random values between 0 and 1.

(2) Calculate cluster centres $c_i$ (i=1,2,..., c).

(3) Compute the objective function. Stop if either it is below a certain threshold level or its improvement over the previous iteration is below a certain tolerance.

(4) Compute a new M.

(5) Go to step 2.

⇓

## Fuzzy Rule Base

# Fuzzy classifier

**<u>The fuzzy classifier</u>** is based on the set of final rules R for which the following holds:

**R: If $x_1$ is $\mu^{(1)}_R$ and … and $x_p$ is $\mu^{(p)}_R$ Then class is $C_R$**

## <u>Distribution of Rules:</u>

$$R(x_1,\ldots x_p) = \begin{cases} C, & if\ \ \mu_C^{(R)}(x_1,\ldots,x_p) > \mu_D^{(R)}(x_1,\ldots,x_p) \quad for\, all\ \ D \in C,\ D \neq C \\ \notin C, & otherwise \end{cases}$$

## <u>Two dimension space:</u>

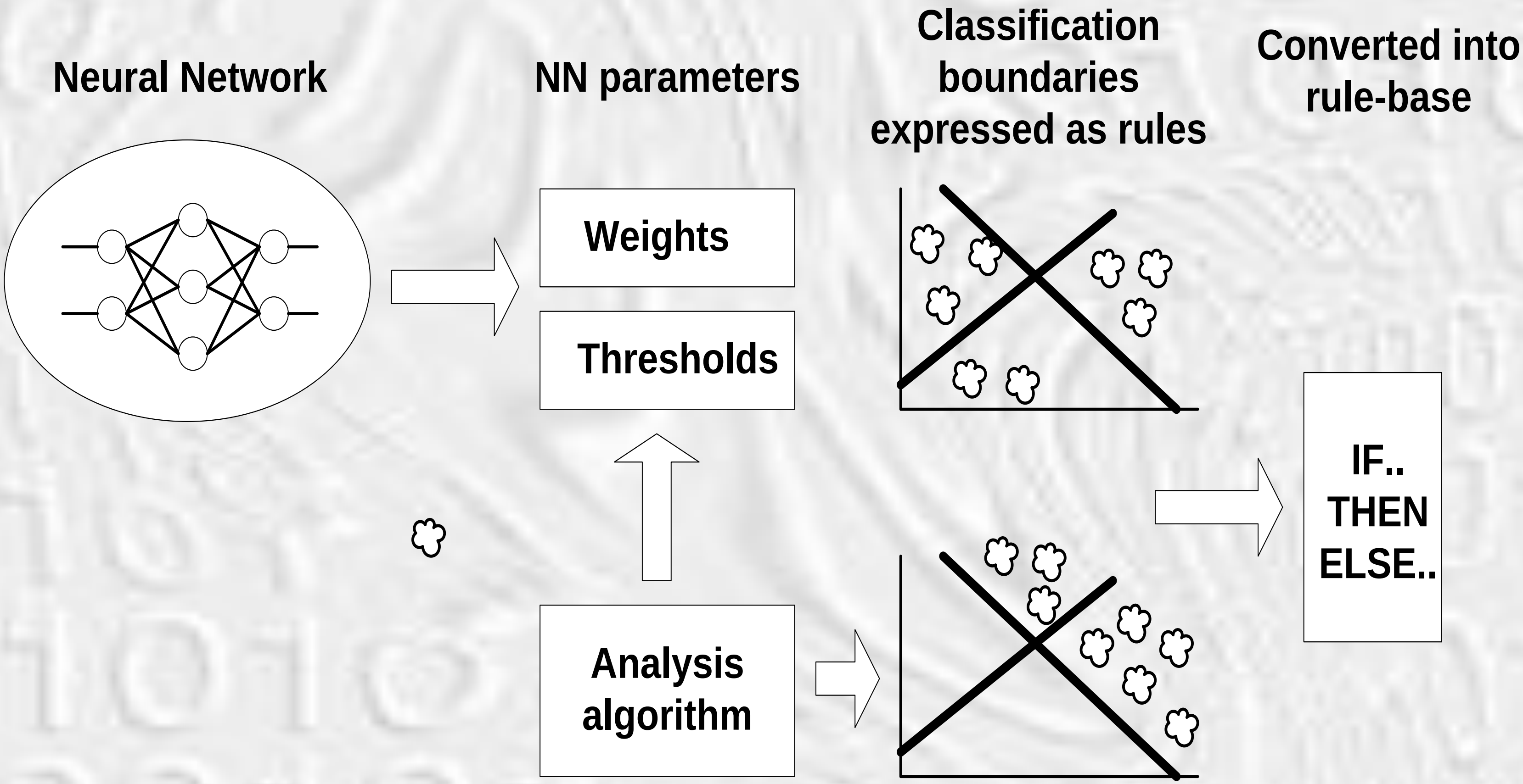
IF x is μ1 and y is v1 THEN class is A

5 stages

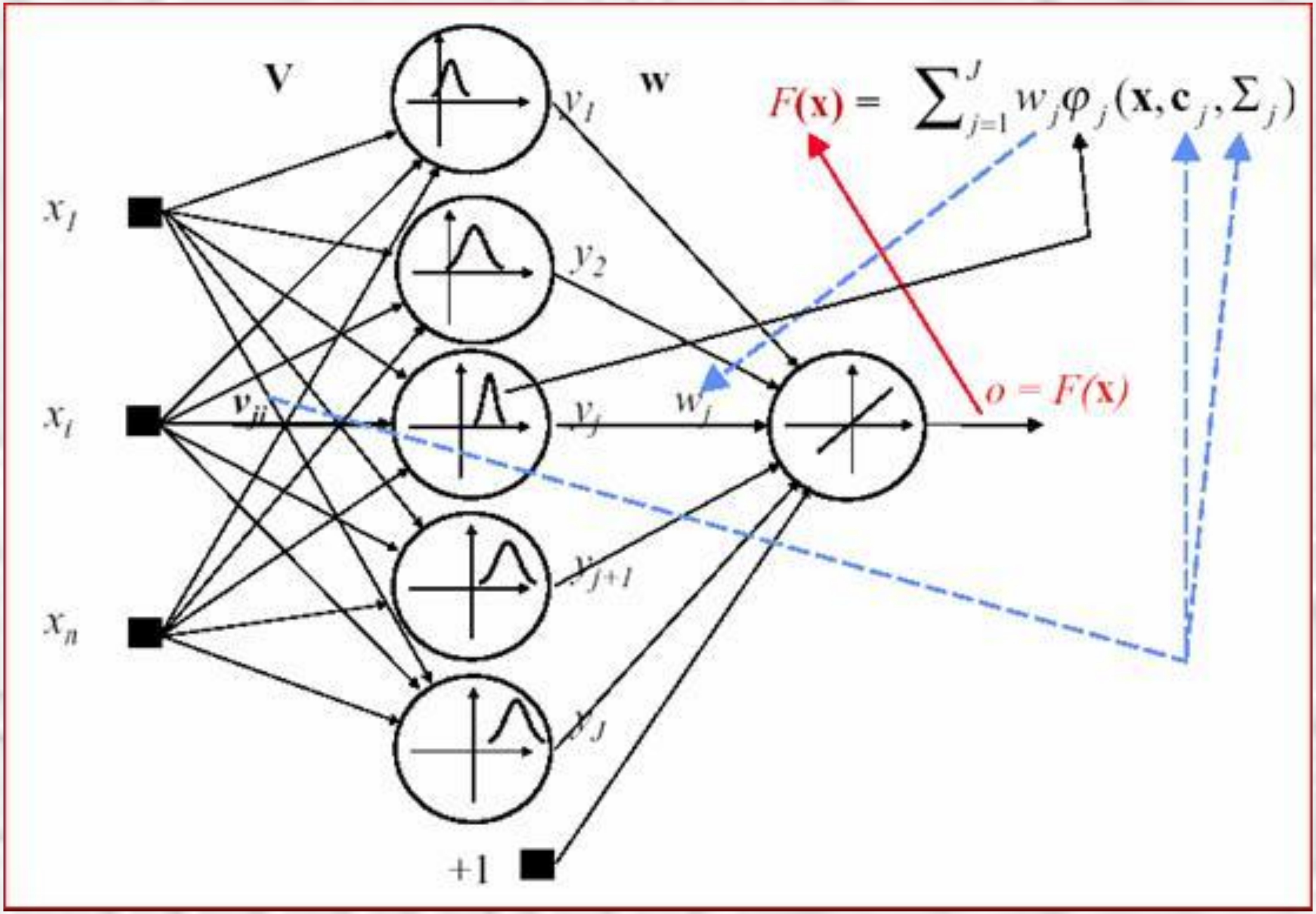
IF x is μ2 and y is v2 THEN class is B

# II. EXTRACTING RULES FROM TRAINED RBF NEURAL NETWORK

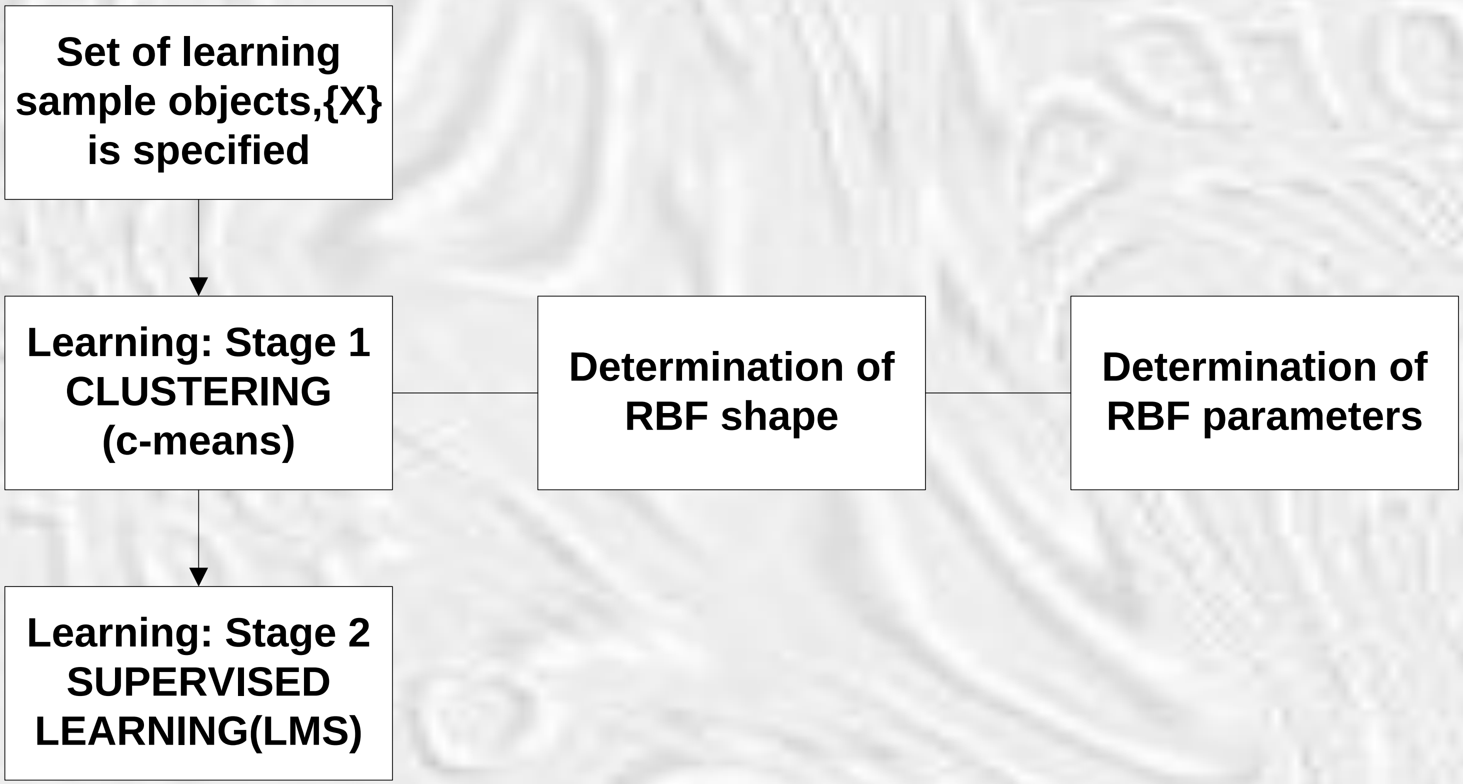
# Rule extraction process from NN



# RBF Network Architecture



# Clustering – the first stage of the RBF network learning



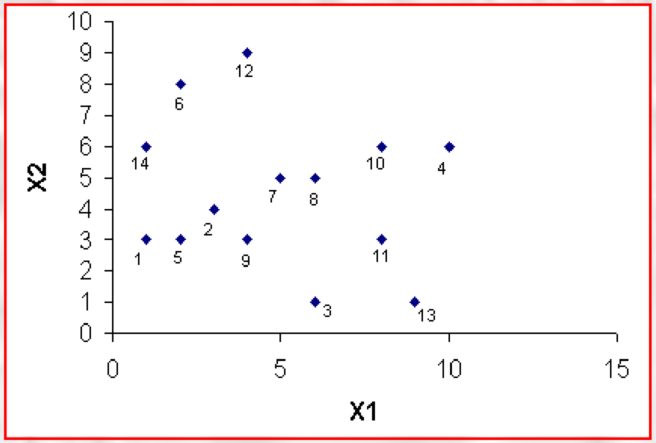
# RULEX algorithm

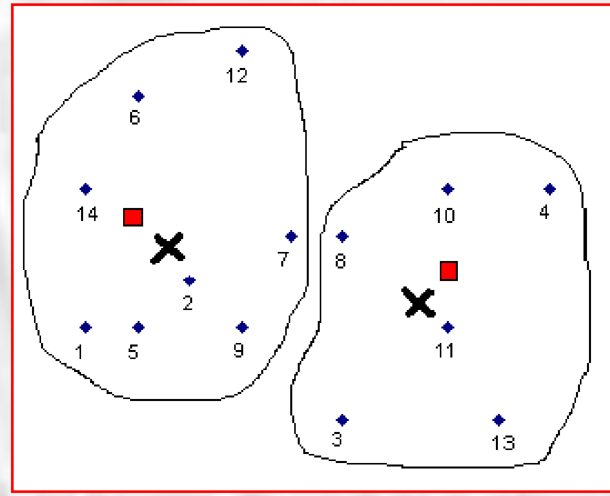
The local nature of each RBF hidden unit enables a simple translation into a single rule:

IF $\text{Feature}_1$ is TRUE AND IF $\text{Feature}_2$ is TRUE AND IF $\text{Feature}_n$ is TRUE THEN $\text{Class}_x$,

| | |
|---|---|
| **Input:** | **Hidden weights $\mu$ (centre positions)**<br>**Gaussian radius spread $\sigma$**<br>**Steepness S** |
| **Output:** | **One rule per hidden unit** |
| **Procedure:** | **Train RBF network on data set**<br>**For each hidden unit:**<br>    **For each $\mu_i$**<br>        $X_{lower} = \mu_i - \sigma_i + S$<br>        $X_{upper} = \mu_i + \sigma_i - S$<br>**Build rule by:**<br>    **antecedent=[ $X_{lower}$, $X_{upper}$]**<br>    **Join antecedents with AND**<br>    **Add class label**<br>**Write rule** |

# Simple example




**Centri : $\mu_1$=(-0.73; 0.26) un $\mu_2$=(0.97;-0.35).**

**Rādiusi: $\sigma_1^2$ = 1.07 un $\sigma_2^2$ = 1.04**

**Cluster 1.** $X_{1\_lower}$= -0.73 - 1.03 + 0.6= - 1.16; $X_{2\_lower}$=0.26 - 1.03 +0.6= -0.17;
$X_{2\_upper}$= -0.73+ 1.03 - 0.6 = -0.3; $X_{2\_upper}$=0.26 +1.03 - 0.6= 0.69.

**Cluster 2.** $X_{1\_lower}$= 0.97 - 1.01 + 0.6= 0.56; $X_{2\_lower}$= -0.35 - 1.01 +0.6= -0.76;
$X_{2\_upper}$= 0.97+ 1.01 - 0.6 = 1.38; $X_{2\_upper}$= -0.35 +1.01 - 0.6= 0.06.

**Steepness=0.6**

**IF ($x_1 \geq$ -1.16 AND $\leq$ -0.3) AND IF ($x_2$ $\geq$ - 0.17 AND $\leq$ 0.69) THEN CLASS 1**
**IF ($x_1 \geq$ 0.56 AND $\leq$ 1.38) AND IF ($x_2$ $\geq$ - 0.76 AND $\leq$ 0.06) THEN CLASS 2**

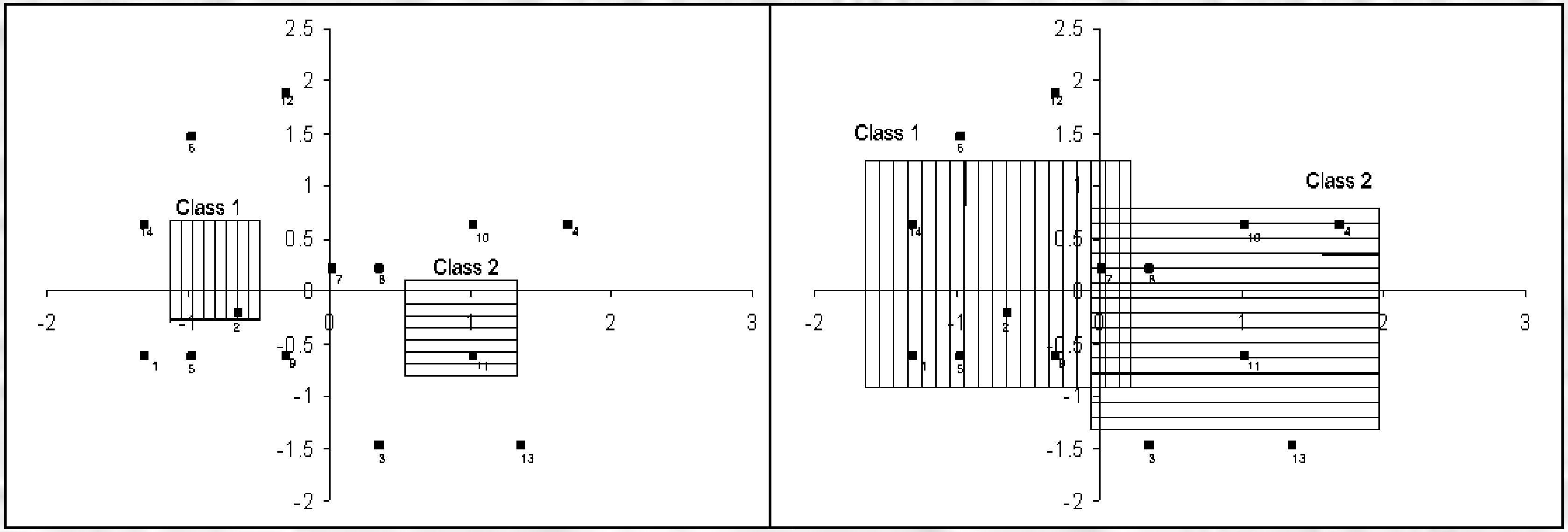
---

**Steepness=0**

**IF ($x_1 \geq$ -1.76 AND $\leq$ 0.3) AND IF ($x_2$ $\geq$ - 0.77 AND $\leq$ 1.29) THEN CLASS 1**
**IF ($x_1 \geq$ -0.04 AND $\leq$ 1.98) AND IF ($x_2$ $\geq$ - 1.36 AND $\leq$ 0.66) THEN CLASS 2**

# Results

Regions of rules are represented in Figure.

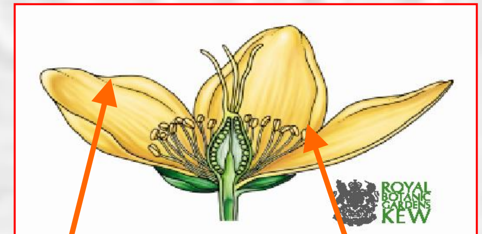



a) b)

Region of rules: steepness=0.6 (a) and steepness=0 (b)

# Application: IRIS data set

setosa, versicolor un virginica.


Petal Sepal

| Setosa | | | |
|---|---|---|---|
| SL | SW | PL | PW |
| 5.1 | 3.5 | 1.4 | 0.2 |
| 4.9 | 3.0 | 1.4 | 0.2 |
| 4.7 | 3.2 | 1.3 | 0.2 |
| 4.6 | 3.1 | 1.5 | 0.2 |
| 5.0 | 3.6 | 1.4 | 0.2 |
| ....... | ....... | ....... | ....... |

| Versicolor | | | |
|---|---|---|---|
| SL | SW | PL | PW |
| 7.0 | 3.2 | 4.7 | 1.4 |
| 6.4 | 3.2 | 4.5 | 1.5 |
| 6.9 | 3.1 | 4.9 | 1.5 |
| 5.5 | 2.3 | 4.0 | 1.3 |
| 6.5 | 2.8 | 4.6 | 1.5 |
| ....... | ....... | ....... | ....... |

| Virginica | | | |
|---|---|---|---|
| SL | SW | PL | PW |
| 6.3 | 3.3 | 6.0 | 2.5 |
| 5.8 | 2.7 | 5.1 | 1.9 |
| 7.1 | 3.0 | 5.9 | 2.1 |
| 6.3 | 2.9 | 5.6 | 1.8 |
| 6.5 | 3.0 | 5.8 | 2.2 |
| ....... | ....... | ....... | ....... |

4 parameters:

SL – *sepal length*
SW – *sepal width*
PL - *petal length*
PW – *petal width*

# Results

| | Parameter S=-0.9 | Parameter S=0 |
|---|---|---|
| Values of centers and radii | Class 1 = 5.01 3.42 1.46 0.24<br>Class 2 = 5.94 2.77 4.26 1.33<br>Class 3 = 6.59 2.97 5.55 2.03<br>Values of radii = 0.30 0.61 0.87 | Class 1 = 5.01 3.42 1.46 0.24<br>Class 2 = 5.94 2.77 4.26 1.33<br>Class 3 = 6.59 2.97 5.55 2.03<br>Values of radii = 0.30 0.61 0.87 |
| Rules correctly describe elements of classes (%) | 100 | 58.7 |
| Rule of Class 1 | IF (X1>= 3.80 AND < 6.21 ) AND<br>IF (X2>= 2.21 AND < 4.62) AND<br>IF (X3>= 0.26 AND < 2.67) AND<br>IF (X4>= -0.96 AND < 1.45)<br>THEN SETOSA | IF (X1>= 4.70 AND < 5.31 ) AND<br>IF (X2>= 3.11 AND < 3.72) AND<br>IF (X3>= 1.16 AND < 1.77) AND<br>IF (X4>= -0.06 AND < 0.55)<br>THEN SETOSA |
| Rule of Class 2 | IF (X1>= 4.42 AND < 7.45 ) AND<br>IF (X2>= 1.26 AND < 4.28) AND<br>IF (X3>= 2.75 AND < 5.77) AND<br>IF (X4>= -0.19 AND < 2.84)<br>THEN VERSICOLOR | IF (X1>= 5.32 AND < 6.55 ) AND<br>IF (X2>= 2.16 AND < 3.38) AND<br>IF (X3>= 3.65 AND < 4.87) AND<br>IF (X4>= 0.71 AND < 1.94)<br>THEN VERSICOLOR |
| Rule of Class 3 | IF (X1>= 4.82 AND < 8.36 ) AND<br>IF (X2>= 1.20 AND < 4.74) AND<br>IF (X3>= 3.78 AND < 7.32) AND<br>IF (X4>= 0.26 AND < 3.80)<br>THEN VIRGINICA | IF (X1>= 5.72 AND < 7.46 ) AND<br>IF (X2>= 2.10 AND < 3.84) AND<br>IF (X3>= 4.68 AND < 6.42) AND<br>IF (X4>= 1.16 AND < 2.90)<br>THEN VIRGINICA |

# Results of training set B (arbitrary 20 elements of every class)

| Correct | Values of parameter S | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | -0.9 | -0.8 | -0.7 | -0.6 | -0.5 | -0.4 | -0.3 | -0.2 | -0.1 | 0 | 0.1 | 0.2 |
| Class 1 | 49 | 49 | 48 | 48 | 45 | 40 | 39 | 27 | 14 | 9 | 2 | 0 |
| Class 2 | 50 | 49 | 49 | 48 | 45 | 44 | 40 | 36 | 28 | 20 | 10 | 3 |
| Class 3 | 49 | 49 | 48 | 47 | 45 | 43 | 43 | 42 | 39 | 35 | 29 | 23 |
| % | 98.7 | 98 | 96.7 | 95.3 | 90 | 84.7 | 81.3 | 70 | 54 | 42.7 | 27.3 | 17.3 |

# Conclusions

- Such rule extraction technique is shown through IRIS data set experimental results.
- The extracted rules can help discover and analyze the hidden knowledge in data sets further.
- The experiments have shown that these methods can be viewed as alternatives to traditional data analysis methods.
- The correct adjustment of parameters in both methods proposed will allow minimizing data processing risks in the analysis of data.

Thanks !