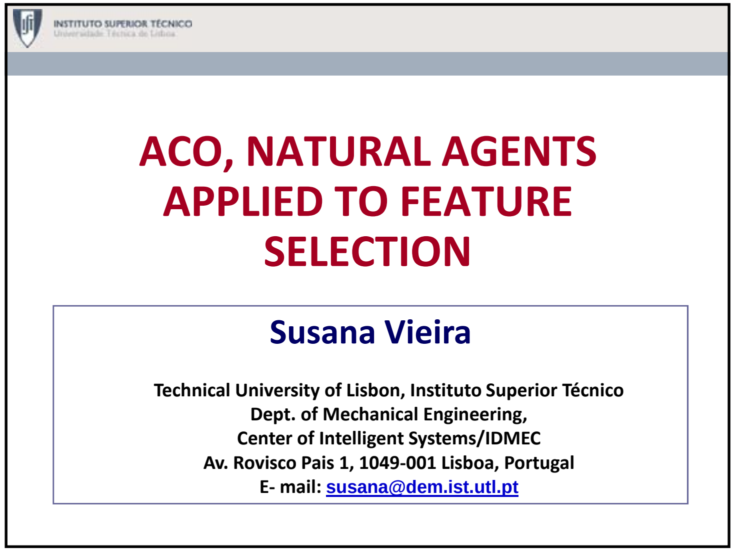
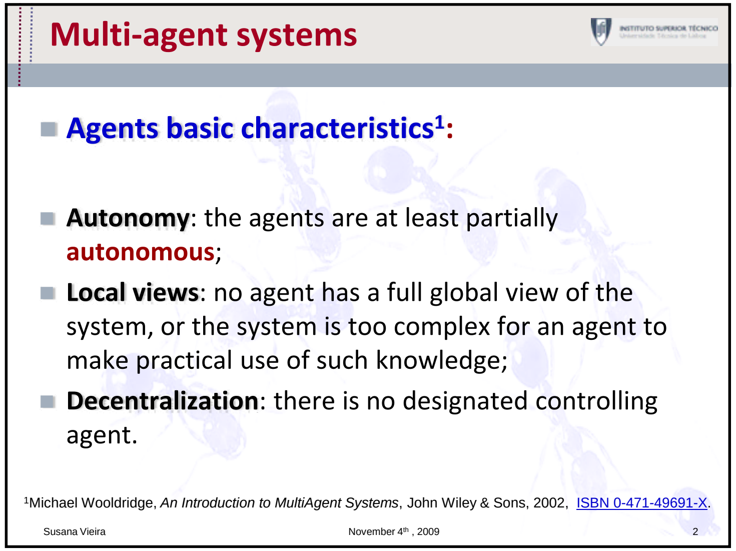


ACO, NATURAL AGENTS APPLIED TO FEATURE SELECTION

Susana Vieira

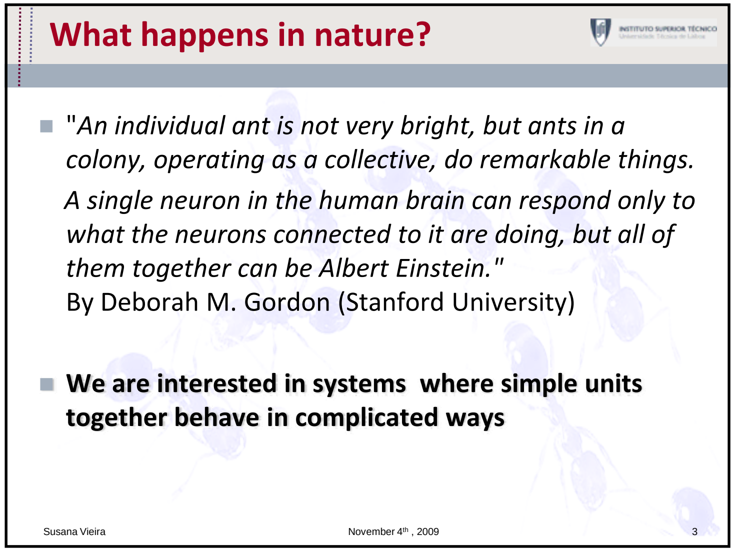
Technical University of Lisbon, Instituto Superior Técnico
Dept. of Mechanical Engineering,
Center of Intelligent Systems/IDMEC
Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
E- mail: susana@dem.ist.utl.pt



Multi-agent systems

- **Agents basic characteristics[1]:**
- **Autonomy**: the agents are at least partially **autonomous**;
- **Local views**: no agent has a full global view of the system, or the system is too complex for an agent to make practical use of such knowledge;
- **Decentralization**: there is no designated controlling agent.

[1] Michael Wooldridge, *An Introduction to MultiAgent Systems*, John Wiley & Sons, 2002, ISBN 0-471-49691-X.

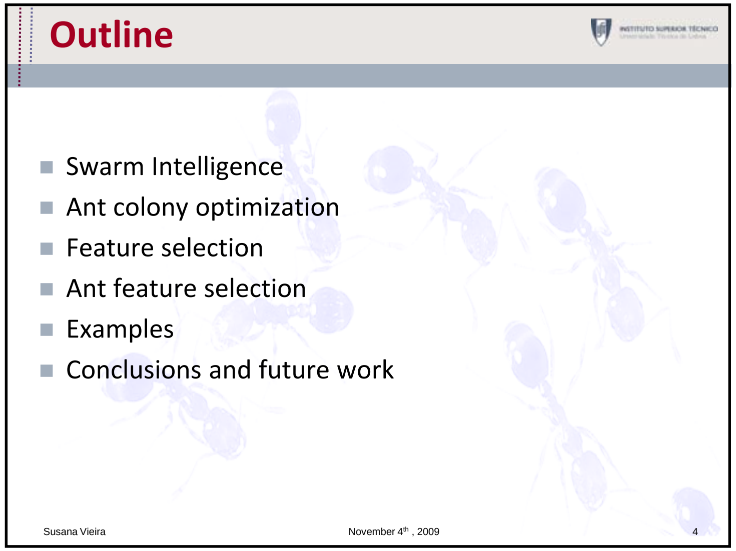


What happens in nature?

- "*An individual ant is not very bright, but ants in a colony, operating as a collective, do remarkable things.*

  A single neuron in the human brain can respond only to what the neurons connected to it are doing, but all of them together can be Albert Einstein."
  By Deborah M. Gordon (Stanford University)

- **We are interested in systems where simple units together behave in complicated ways**



Outline

- Swarm Intelligence
- Ant colony optimization
- Feature selection
- Ant feature selection
- Examples
- Conclusions and future work

Learning from Nature

- Nature has inspired researchers in many different ways.
  - **Airplanes** have been designed based on the structures of **birds**' wings.
  - **Robots** have been designed in order to imitate the movements of **insects**.
  - **Resistant materials** have been synthesized based on **spider webs**.
- After millions of years of evolution all these species developed solutions for a wide range of problems. Some ideas can be developed by **taking advantage of the examples that Nature offers**.

Learning from Nature

- Some **social systems** in Nature can present an **intelligent collective behavior** although they are composed by **simple individuals.**
- The intelligent **solutions** to problems naturally **emerge** from the **self-organization** and **communication** of these individuals.
- These systems provide important **techniques** that can be **used in the development of** distributed **artificial intelligent systems**.

Swarm Intelligence

- Based on the study of emergent collective intelligence of groups of simple agents

Ant Colony

Bird Flock

Fish School

Animal Herd

Swarm Intelligence

- Swarm Intelligence is an **artificial intelligence technique** based on the study of collective behavior in self-organized systems.
  - Swarm Intelligence systems are typically made up of a population of **simple agents interacting locally** with one another and with their **environment**. This interaction often lead to the **emergence of global behavior**.
- The main bio-inspired algorithms that have been developed are:
  - **Ant Colony Optimisation** (ACO)
  - **Particle Swarm Optimisation** (PSO)

Natural ants

- Individual ants are simple insects with limited memory and capable of performing simple actions.

- However, an ant colony expresses a complex collective behavior providing **intelligent solutions to problems** such as:
  - carrying large items
  - forming bridges
  - **finding the shortest routes from the nest to a food source**, prioritizing food sources based on their distance and ease of access.

What is special about ants?

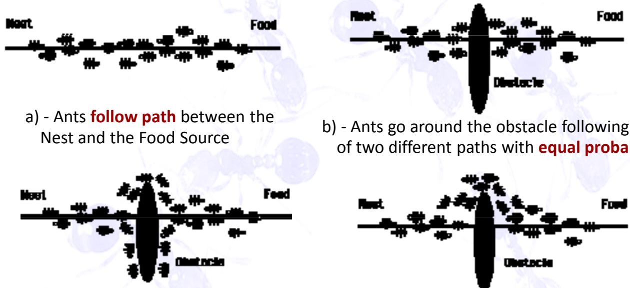
- Ants can perform complex tasks:
  - nest building, food storage
  - garbage collection, war
  - ***foraging*** *(to wander in search of food)*
- There is no management in an ant colony
  - collective intelligence
- They communicate using:
  - ***pheromones*** *(chemical substances)*, sound, touch
- Curiosities:
  - Ant colonies exist for more than 100 million years
  - Myrmercologists estimate that there are around 20 000 species of ants

Ant colony optimization

- **Ant Colony Optimization** is one of the most used method of the *Artificial Life* algorithms.
  - **Introduced by:** Marco Dorigo (1992), and is starting to be used in industrial applications.
  - **Applications:** Travelling salesman problem, vehicle routing, quadratic assignment problem, internet routing, logistics scheduling.
- There are also some applications of ACO in clustering and data mining problems, including **feature selection**.

The foraging behaviour of ants

- How can almost blind animals manage to learn the shortest route paths from their nests to the food source and back?



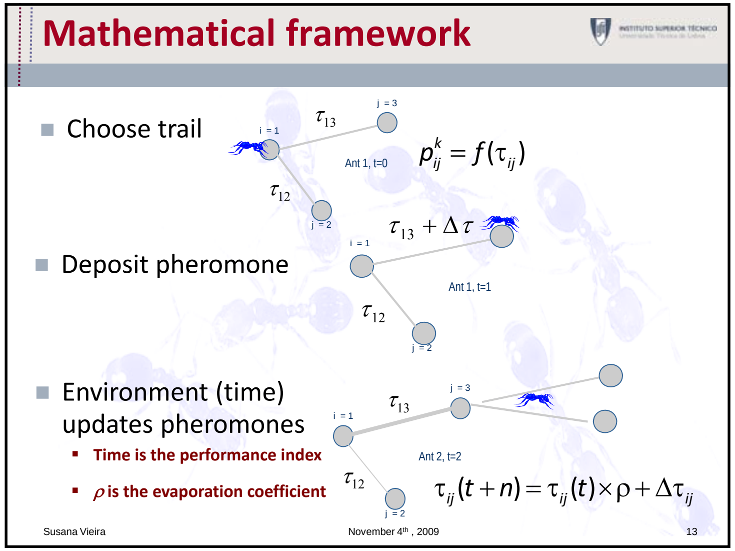
a) - Ants **follow path** between the Nest and the Food Source

b) - Ants go around the obstacle following one of two different paths with **equal probability**

c) - On the shorter path, more **pheromones** are laid down

d) – At the end, **all ants follow** the shortest path.

Fotos: http://iridia.ulb.ac.be/~mdorigo/ACO/RealAnts.html



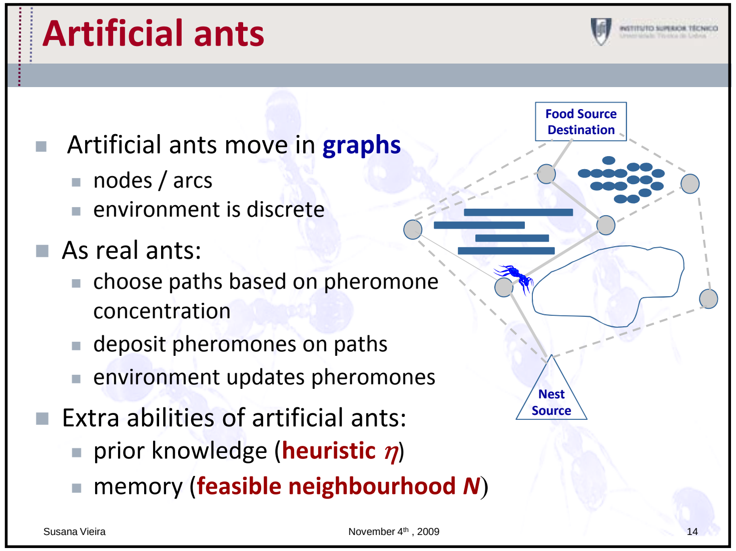
Mathematical framework

- Choose trail

$$p_{ij}^k = f(\tau_{ij})$$

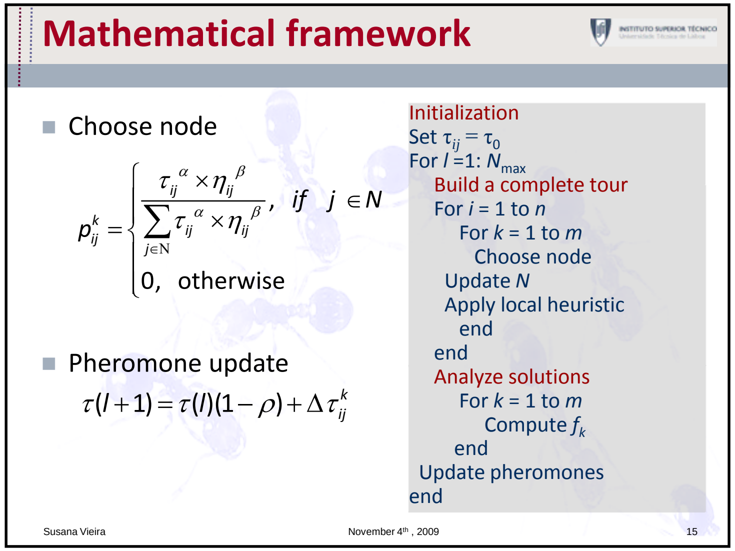
- Deposit pheromone
- Environment (time) updates pheromones
  - **Time is the performance index**
  - **ρ is the evaporation coefficient**

$$\tau_{ij}(t+n)=\tau_{ij}(t)\times\rho+\Delta\tau_{ij}$$



Artificial ants

- Artificial ants move in **graphs**
  - nodes / arcs
  - environment is discrete
- As real ants:
  - choose paths based on pheromone concentration
  - deposit pheromones on paths
  - environment updates pheromones
- Extra abilities of artificial ants:
  - prior knowledge (**heuristic η**)
  - memory (**feasible neighbourhood *N***)



Mathematical framework

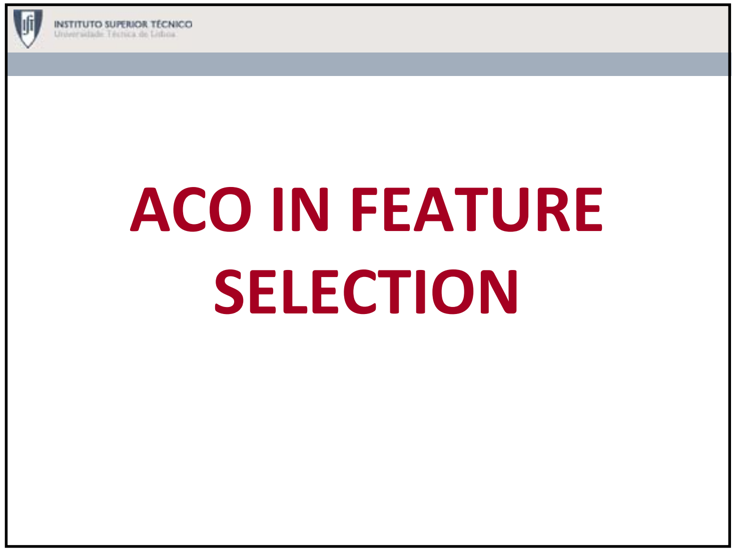
- Choose node

$$p_{ij}^k=\begin{cases}\dfrac{\tau_{ij}{}^{\alpha}\times\eta_{ij}{}^{\beta}}{\sum_{j\in N}\tau_{ij}{}^{\alpha}\times\eta_{ij}{}^{\beta}}, & if \quad j\in N\\ 0, & \text{otherwise}\end{cases}$$

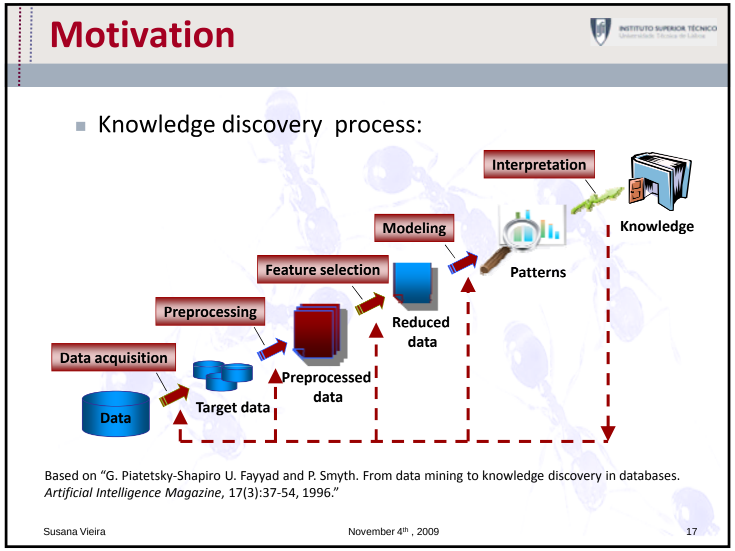
- Pheromone update

$$\tau(l+1)=\tau(l)(1-\rho)+\Delta\tau_{ij}^k$$

```
Initialization
Set $\tau_{ij} = \tau_0$
For $l$ =1: $N_{max}$
    Build a complete tour
    For $i$ = 1 to $n$
        For $k$ = 1 to $m$
            Choose node
        Update $N$
        Apply local heuristic
        end
    end
    Analyze solutions
        For $k$ = 1 to $m$
            Compute $f_k$
        end
    Update pheromones
end
```



ACO IN FEATURE SELECTION



Motivation

- Knowledge discovery process:

Interpretation

Knowledge

Modeling

Patterns

Feature selection

Reduced data

Preprocessing

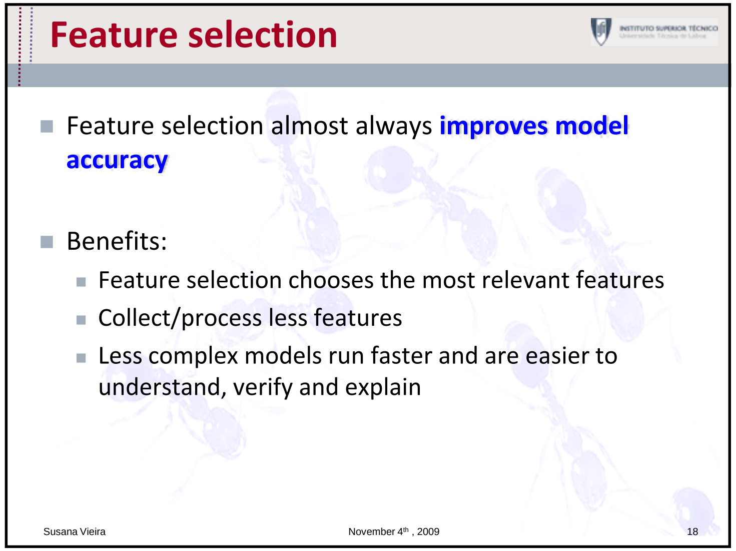
Preprocessed data

Data acquisition

Target data

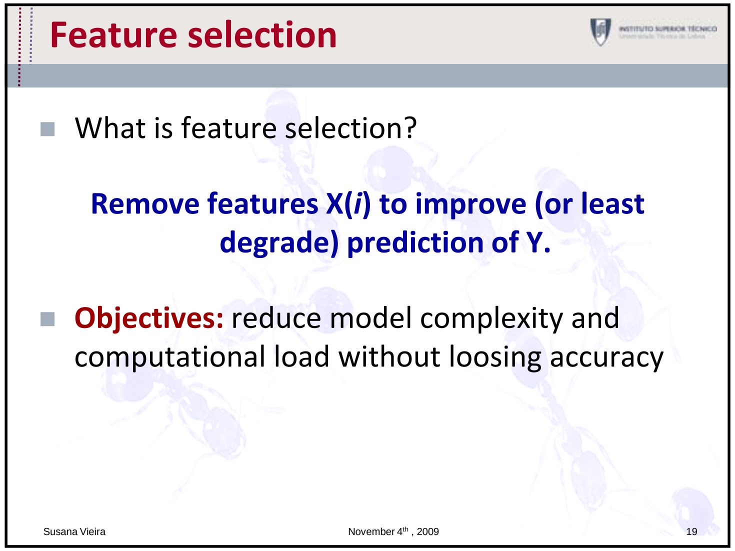
Data

Based on "G. Piatetsky-Shapiro U. Fayyad and P. Smyth. From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine*, 17(3):37-54, 1996."



Feature selection

- Feature selection almost always **improves model accuracy**
- Benefits:
  - Feature selection chooses the most relevant features
  - Collect/process less features
  - Less complex models run faster and are easier to understand, verify and explain

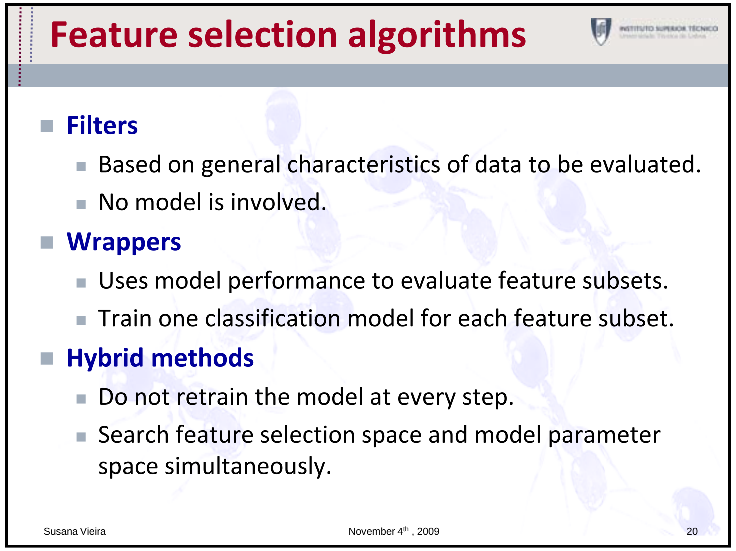


Feature selection

- What is feature selection?

Remove features X(*i*) to improve (or least degrade) prediction of Y.

- **Objectives:** reduce model complexity and computational load without loosing accuracy



Feature selection algorithms

- **Filters**
  - Based on general characteristics of data to be evaluated.
  - No model is involved.
- **Wrappers**
  - Uses model performance to evaluate feature subsets.
  - Train one classification model for each feature subset.
- **Hybrid methods**
  - Do not retrain the model at every step.
  - Search feature selection space and model parameter space simultaneously.

Objective function

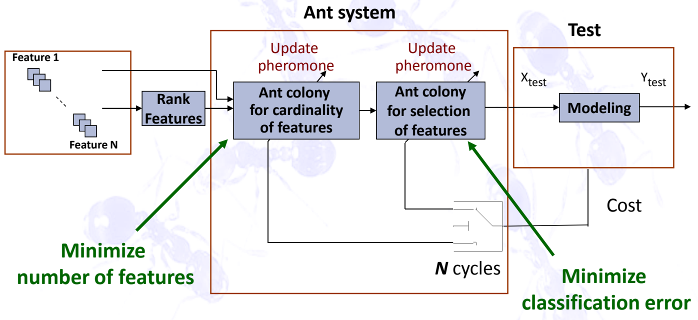
- **Main objectives:**
  - Minimize the number of misclassifications, or the **classification error**
  - Reduce the number of features, or the **features cardinality**

$$\text{minimize } f = w_1 e + w_2 N$$

➢ Tradeoff **precision** vs. **accuracy**.

Optimization algorithm

- Multicriteria algorithm:


Ant system
Test
Feature 1
Feature N
Rank Features
Update pheromone
Ant colony for cardinality of features
Update pheromone
Ant colony for selection of features
X_{test}
Y_{test}
Modeling
Cost
N cycles
Minimize number of features
Minimize classification error

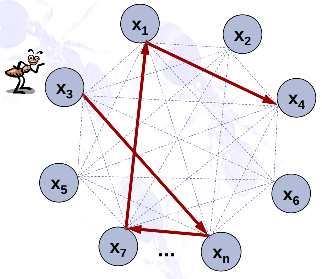
ACO for feature selection

- **Second colony:**
  - Choose node

$$p_{ij}^k = \begin{cases} \dfrac{\tau_{ij}{}^{\alpha} \times \eta_{ij}{}^{\beta}}{\sum\limits_{j \in N} \tau_{ij}{}^{\alpha} \times \eta_{ij}{}^{\beta}}, & if \quad j \in N \\ 0, & \text{otherwise} \end{cases}$$

  - Pheromone update

$$\tau(l+1) = \tau(l)(1-\rho) + \Delta\tau_{ij}^k$$



Subset: $\{x_3, x_n, x_7, x_1, x_4\}$

Choosing node in graph

- Probability of an ant choosing node *i* (cardinality of features):

$$p_i^k(t) = \frac{\left[\tau_{n_i}(t)\right]^{\alpha_n} \cdot \left[\eta_{n_i}\right]^{\beta_n}}{\sum\limits_{l \in J_i^k} \left[\tau_{n_{il}}(t)\right]^{\alpha_n} \cdot \left[\eta_{n_{il}}\right]^{\beta_n}}$$

- Probability of an ant choosing node *j* (selection of features):

$$p_j^k(t) = \frac{\left[\tau_{f_j}(t)\right]^{\alpha_f} \cdot \left[\eta_{f_j}\right]^{\beta_f}}{\sum\limits_{l \in J_j^k} \left[\tau_{f_{jl}}(t)\right]^{\alpha_f} \cdot \left[\eta_{f_{jl}}\right]^{\beta_f}}$$

Heuristics of ant systems

- **Heuristic for feature cardinality**: Fisher's score for the features

$$F(i)=\frac{\left|\mu_{c_1}(i)-\mu_{c_2}(i)\right|^2}{\sigma^2_{c_1}(i)+\sigma^2_{c_2}(i)}$$

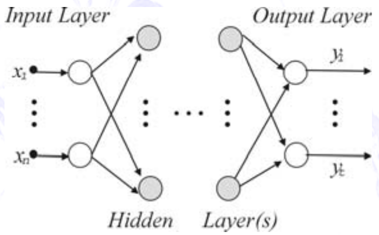
mean and variance values of feature i for the samples in class c_1 and c_2

- **Heuristic for selection of features**: classification error $e(i)$ for the individual features

$$\eta_f(i)=\frac{1}{e(i)}$$

Modeling

- ***Takagi-Sugeno*** fuzzy models are used;
  - ***Antecedents*** A^i are fuzzy sets obtained using fuzzy clustering – membership functions.
  - ***Consequents*** y_i are estimated using least squares estimation.
- ***Feedforward*** Neural Network are used;



Data sets

- Examples:

Data sets	Number of features	Number of classes	Size of data set
Wine	13	3	178
Breast cancer	9	2	699

Results (Wine)

Methods	Reduced Subsets	Classification accuracy (%)		
		Best	Mean	Worst
AFS Approach	4-8	100	99.8	98.9
Corcoran and Sen (1994)	13	100	99.5	98.3
Ishibuchi et al. (1999)	13	99.4	98.5	97.8
Roubos et al. (2003)	4-7	99.4	-	98.3
Mendonça et al. (T-D) (2007)	11	100	99.9	99.4
Mendonça et al. (B-U) (2007)	4	100	98.5	92.7

Results (Breast Cancer)

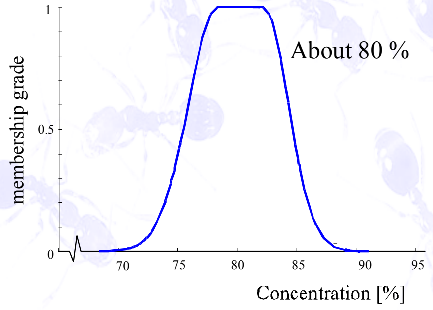
Methods	Reduced Subsets	Classification accuracy (%)		
		Best	Mean	Worst
AFS Approach	2-5	100	96.4	91.3
Wang et al. (POSAR) (2004)	4	95.94	-	-
Wang et al. (CEAR) (2004)	4	94.20	-	-
Wang et al. (DISMAR) (2003)	5	95.94	-	-
Wang et al. (GAAR) (2000)	4	95.65	-	-
Wang et al. (PSORSFS) (2007)	4	95.80	-	-
Abony et al. (GG: R = 2) (2003)	8-9	95.71	90.99	84.28
Abony et al. (Sup: R = 2) (2003)	7-9	98.57	92.56	84.28
Abony et al. (GG: R = 4) (2003)	9	98.57	95.14	88.57
Abony et al. (Sup: R = 4) (2003)	8-9	98.57	95.57	90.0

Fuzzy goals and constraints

- Let *A* be a given set of possible alternatives which contains a solution to a decision making problem under consideration.
- A **fuzzy goal** *G* is a fuzzy set on *A*, characterized by μ_G: *A* → [0,1], represents the degree to which the alternatives satisfy the specified decision goal.
- A **fuzzy constraint** *C* is a fuzzy set on *A* characterized by μ_C: *A* → [0,1], constrains the solution to a fuzzy region within the set of possible solutions..

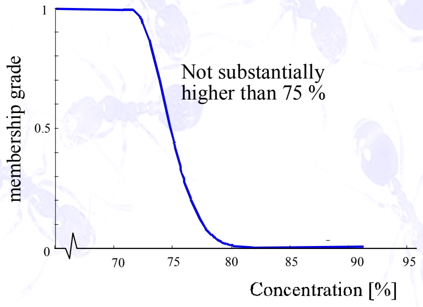
Fuzzy goal

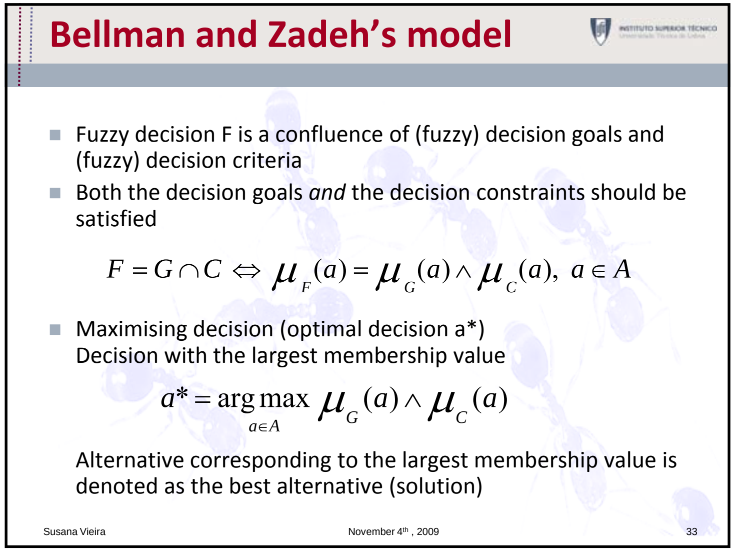
- **Goal**: "Product concentration should be *about* 80%".



Fuzzy constraint

- **Constraint**: "Product concentration should be *not substantially higher* than 75%".





Bellman and Zadeh's model

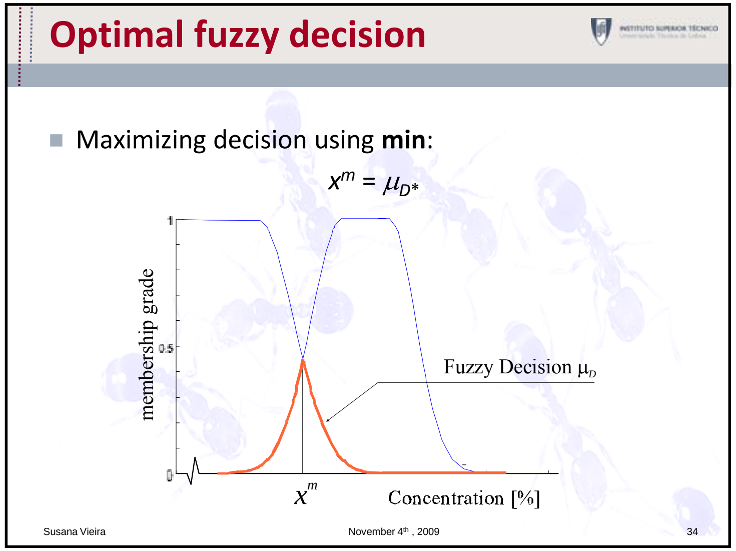
- Fuzzy decision F is a confluence of (fuzzy) decision goals and (fuzzy) decision criteria
- Both the decision goals *and* the decision constraints should be satisfied

$$F = G \cap C \Leftrightarrow \mu_F(a) = \mu_G(a) \wedge \mu_C(a),\ a \in A$$

- Maximising decision (optimal decision a*)
  Decision with the largest membership value

$$a^* = \underset{a \in A}{\arg\max}\ \mu_G(a) \wedge \mu_C(a)$$

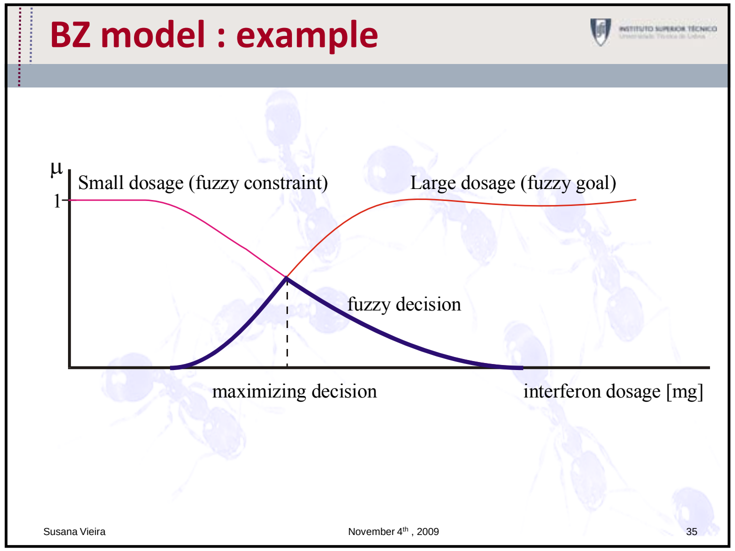
Alternative corresponding to the largest membership value is denoted as the best alternative (solution)



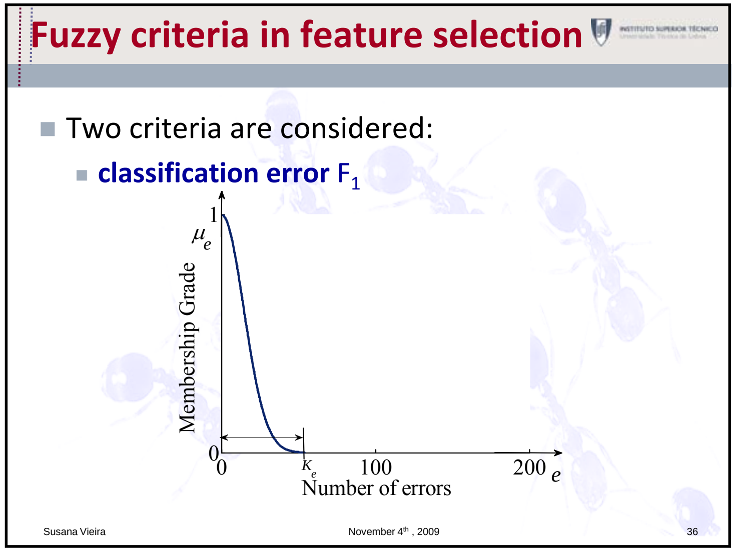
Optimal fuzzy decision

- Maximizing decision using **min**:

$$x^m = \mu_{D^*}$$

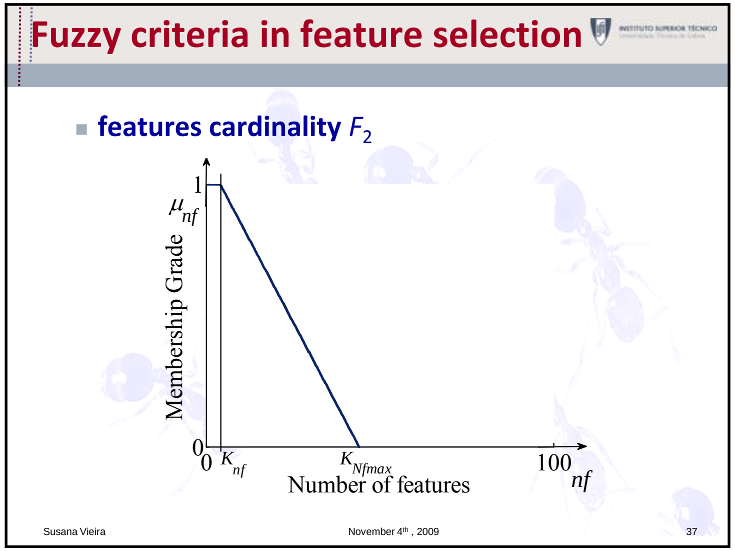


BZ model : example



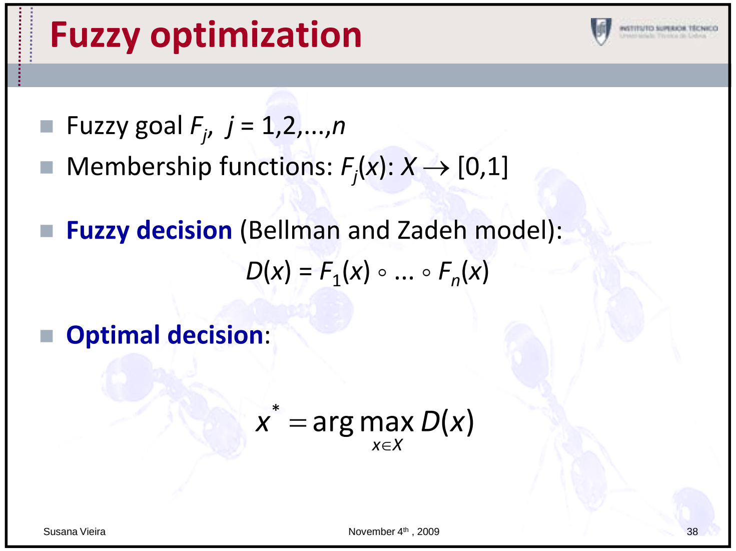
Fuzzy criteria in feature selection

- Two criteria are considered:
  - **classification error** F_1



Fuzzy criteria in feature selection

- **features cardinality** F_2



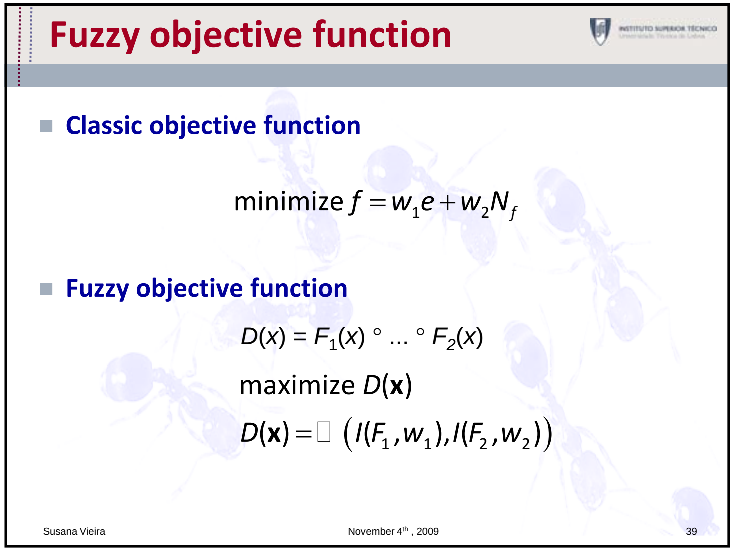
Fuzzy optimization

- Fuzzy goal F_j, $j = 1,2,\ldots,n$
- Membership functions: $F_j(x)\colon X \rightarrow [0,1]$
- **Fuzzy decision** (Bellman and Zadeh model):

$$D(x) = F_1(x) \circ \ldots \circ F_n(x)$$

- **Optimal decision**:

$$x^* = \arg\max_{x \in X} D(x)$$



Fuzzy objective function

- **Classic objective function**

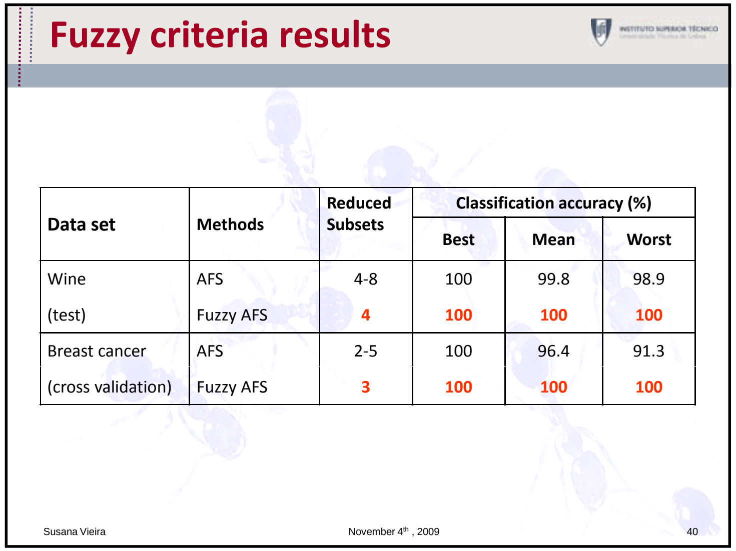
$$\text{minimize } f = w_1 e + w_2 N_f$$

- **Fuzzy objective function**

$$D(x) = F_1(x) \circ \ldots \circ F_2(x)$$

$$\text{maximize } D(\mathbf{x})$$

$$D(\mathbf{x}) = □\left(I(F_1, w_1), I(F_2, w_2)\right)$$



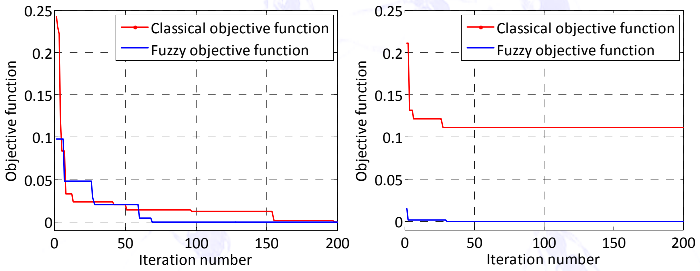
Fuzzy criteria results

Data set	Methods	Reduced Subsets	Classification accuracy (%)		
			Best	Mean	Worst
Wine	AFS	4-8	100	99.8	98.9
(test)	Fuzzy AFS	4	100	100	100
Breast cancer	AFS	2-5	100	96.4	91.3
(cross validation)	Fuzzy AFS	3	100	100	100

Results: classical vs fuzzy criteria

- Classical versus fuzzy objective function convergence:

Wine **Breast Cancer**



Data sets

- Examples:

	Data sets	Number of features	Number of classes	Size of data set
1	Breast cancer original	9	2	699
2	Wine	13	3	178
3	Vote	16	2	300
4	Diagnostic breast cancer	32	2	569
5	Prognostic breast cancer	34	2	198
6	Sonar	60	2	208
7	Musk	166	2	476

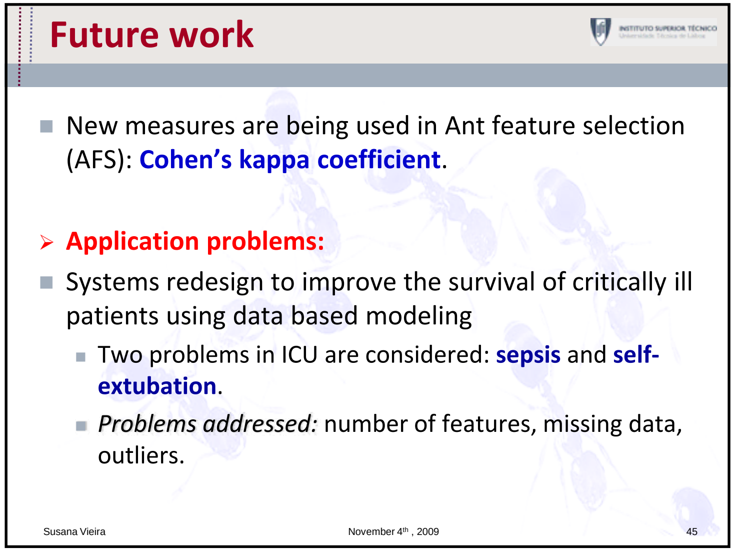
Results: classical vs fuzzy criteria

- Classification rates with 10-fold cross validation:

data set	Fuzzy Models					Neural Networks				
	No FS	AFS	NF	FOF	NF	No FS	AFS	NF	FOF	NF
1	82.6	99.5	4-8	100	4	79.5	99.4	2	100	4
2	84.5	97.7	2-5	98.7	3	84.8	99	2-4	99.3	3-4
3	80.0	99.7	2-5	100	2-3	73.3	98.7	2-3	99.0	2
4	77.2	99.5	2-3	99.5	3	74.0	96.3	2-3	98.6	4
5	78.9	85.6	2	87.3	3-4	77.8	78.9	2-3	78.9	2-4
6	60.2	86.6	2-3	86.7	2	55.4	83.6	2-4	84.2	3-15
7	77.7	78.3	2-20	85.0	6-22	74.7	79.8	2-6	83.5	9-107
Ave.	77.3	92.4	-	93.9	-	74.2	90.8	-	91.9	-
WTL	0/0/7	0/1/6	-	6/1/0	-	0/0/7	0/1/6	-	6/1/0	-

Conclusions

- Ants are natural multi-agents systems.
- A feature selection algorithm based on two cooperative ant colonies was presented.
- The problem is divided into two contradictory objectives: choosing the **features cardinality** and selecting the **most relevant features**.
- Fuzzy objective functions for feature selection are used.
- **Fuzzy objective functions** help the convergence of ACO.

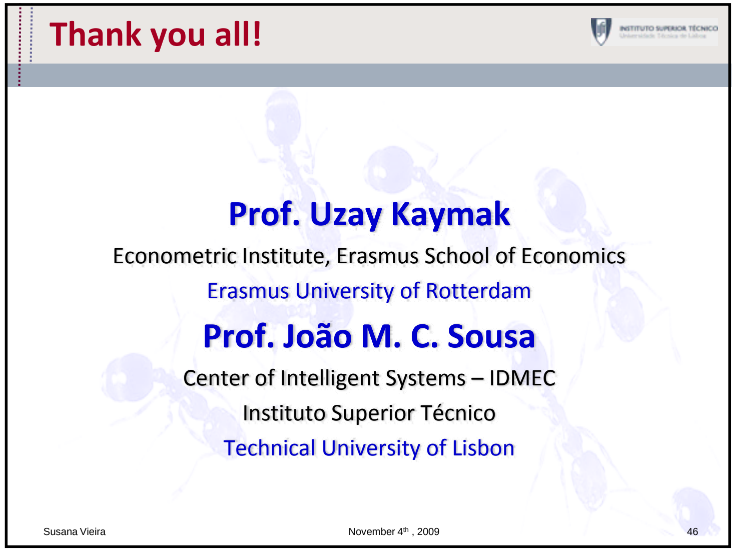


Future work

- New measures are being used in Ant feature selection (AFS): **Cohen's kappa coefficient**.

➢ **Application problems:**

- Systems redesign to improve the survival of critically ill patients using data based modeling
  - Two problems in ICU are considered: **sepsis** and **self-extubation**.
  - *Problems addressed:* number of features, missing data, outliers.



Thank you all!

Prof. Uzay Kaymak

Econometric Institute, Erasmus School of Economics

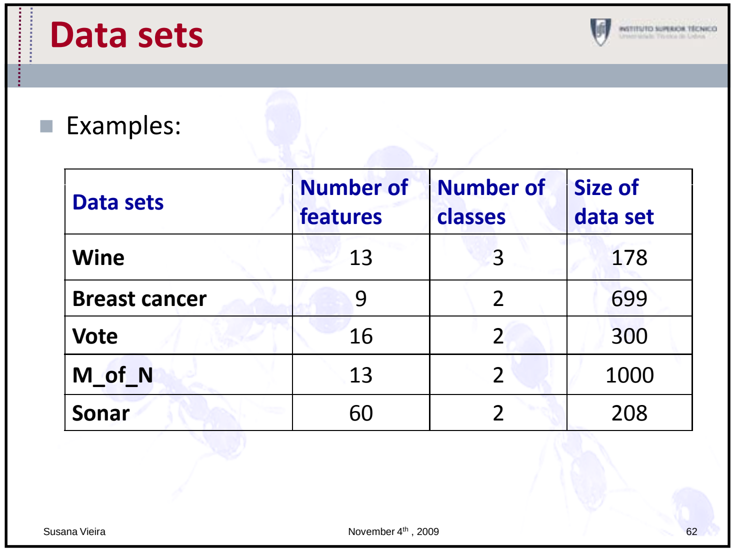
Erasmus University of Rotterdam

Prof. João M. C. Sousa

Center of Intelligent Systems – IDMEC

Instituto Superior Técnico

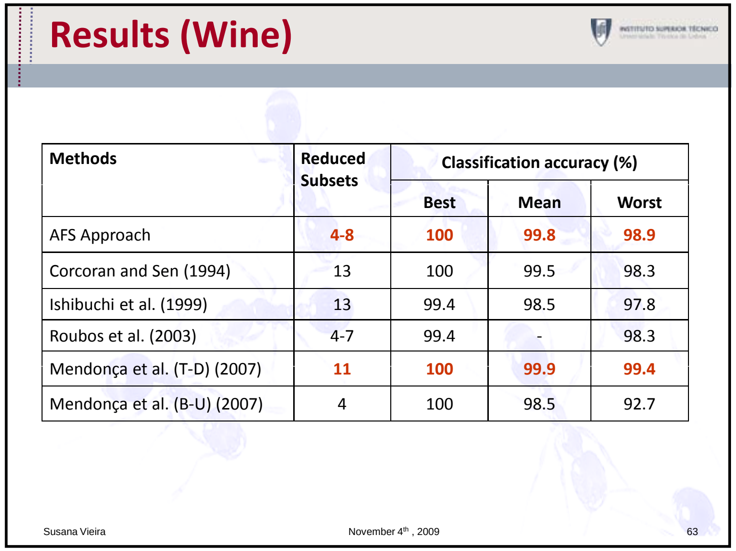
Technical University of Lisbon



Data sets

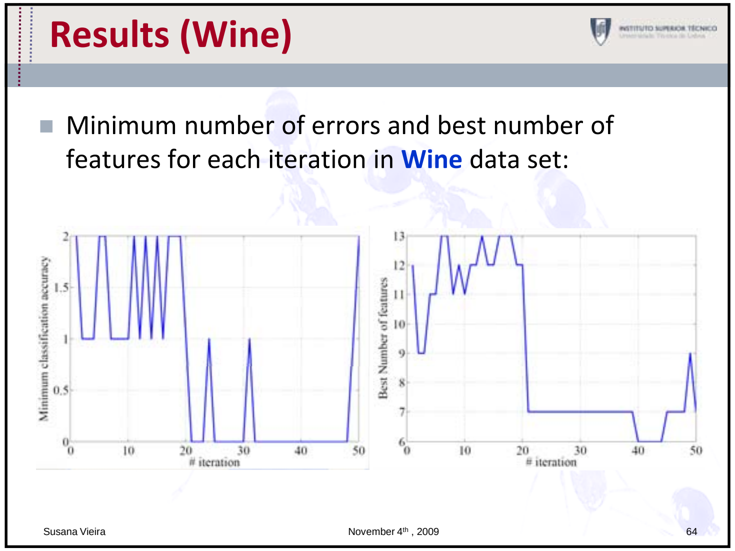
- Examples:

Data sets	**Number of features**	**Number of classes**	**Size of data set**
Wine	13	3	178
Breast cancer	9	2	699
Vote	16	2	300
M_of_N	13	2	1000
Sonar	60	2	208



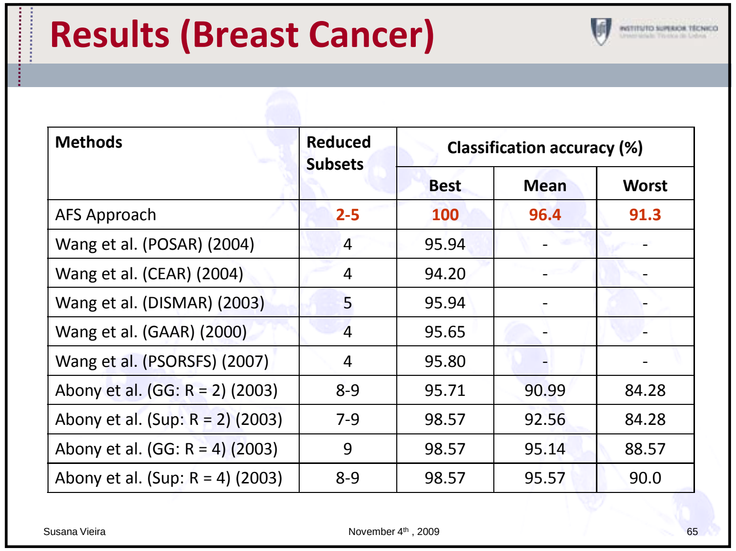
Results (Wine)

Methods	**Reduced Subsets**	**Classification accuracy (%)**		
		Best	**Mean**	**Worst**
AFS Approach	**4-8**	**100**	**99.8**	**98.9**
Corcoran and Sen (1994)	13	100	99.5	98.3
Ishibuchi et al. (1999)	13	99.4	98.5	97.8
Roubos et al. (2003)	4-7	99.4	-	98.3
Mendonça et al. (T-D) (2007)	**11**	**100**	**99.9**	**99.4**
Mendonça et al. (B-U) (2007)	4	100	98.5	92.7



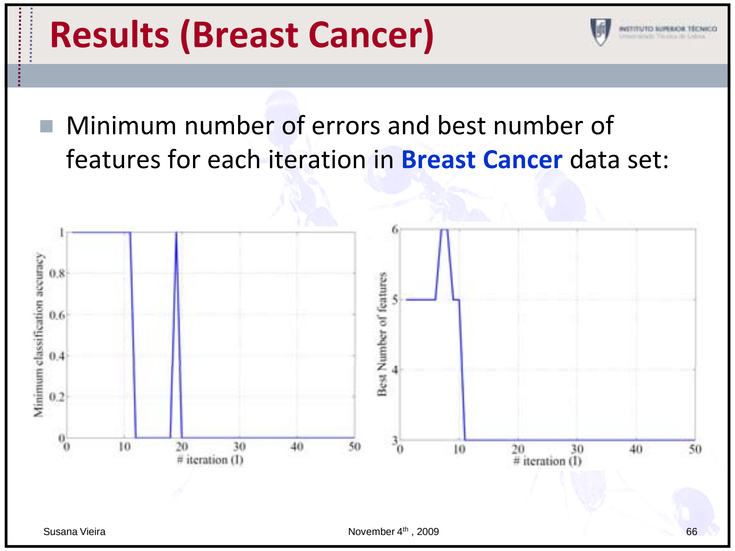
Results (Wine)

- Minimum number of errors and best number of features for each iteration in **Wine** data set:



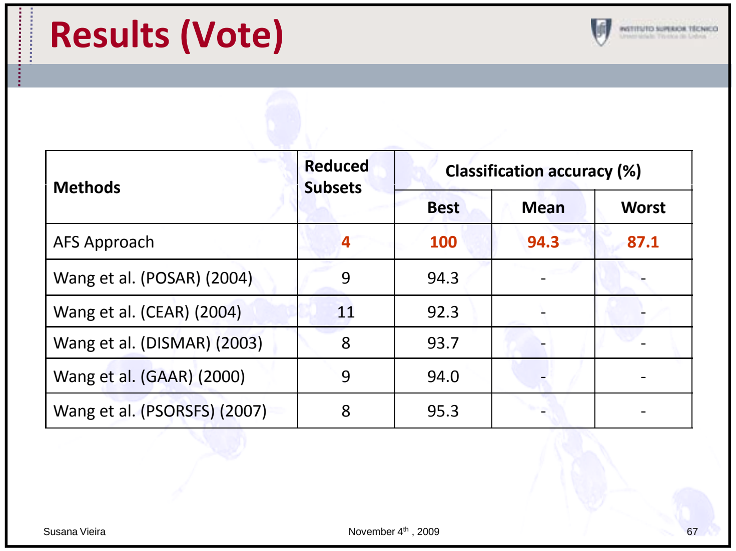
Results (Breast Cancer)

Methods	Reduced Subsets	Classification accuracy (%)		
		Best	Mean	Worst
AFS Approach	2-5	100	96.4	91.3
Wang et al. (POSAR) (2004)	4	95.94	-	-
Wang et al. (CEAR) (2004)	4	94.20	-	-
Wang et al. (DISMAR) (2003)	5	95.94	-	-
Wang et al. (GAAR) (2000)	4	95.65	-	-
Wang et al. (PSORSFS) (2007)	4	95.80	-	-
Abony et al. (GG: R = 2) (2003)	8-9	95.71	90.99	84.28
Abony et al. (Sup: R = 2) (2003)	7-9	98.57	92.56	84.28
Abony et al. (GG: R = 4) (2003)	9	98.57	95.14	88.57
Abony et al. (Sup: R = 4) (2003)	8-9	98.57	95.57	90.0



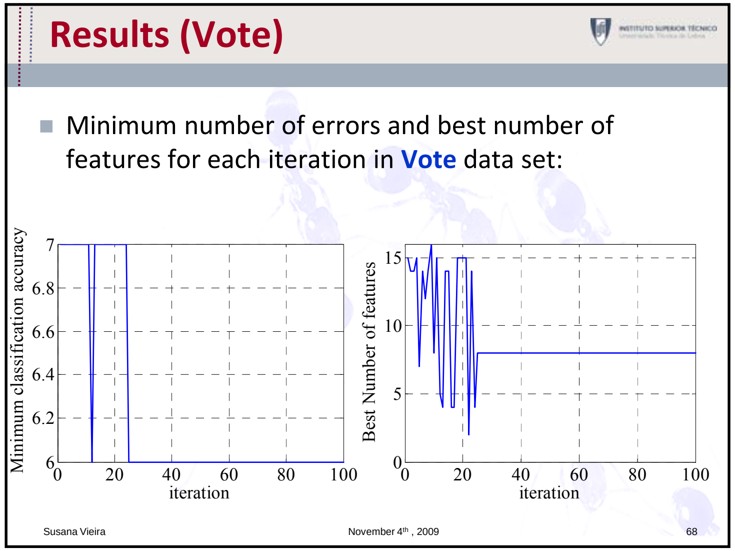
Results (Breast Cancer)

- Minimum number of errors and best number of features for each iteration in **Breast Cancer** data set:



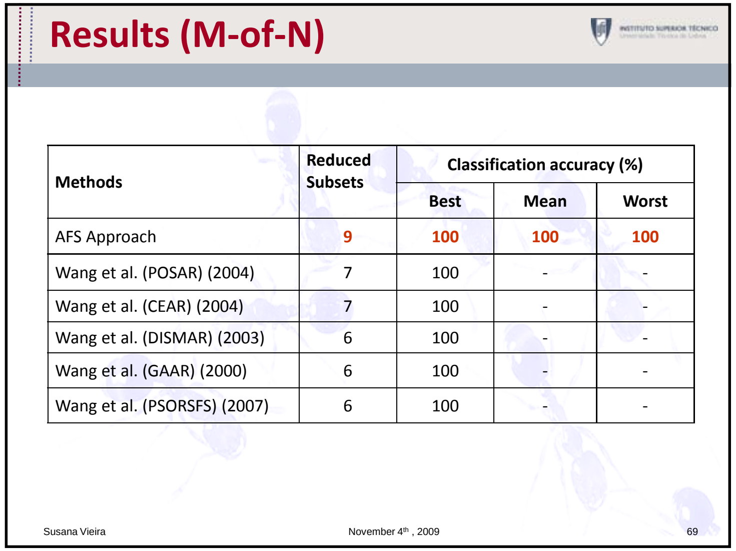
Results (Vote)

Methods	Reduced Subsets	Classification accuracy (%)		
		Best	Mean	Worst
AFS Approach	4	100	94.3	87.1
Wang et al. (POSAR) (2004)	9	94.3	-	-
Wang et al. (CEAR) (2004)	11	92.3	-	-
Wang et al. (DISMAR) (2003)	8	93.7	-	-
Wang et al. (GAAR) (2000)	9	94.0	-	-
Wang et al. (PSORSFS) (2007)	8	95.3	-	-



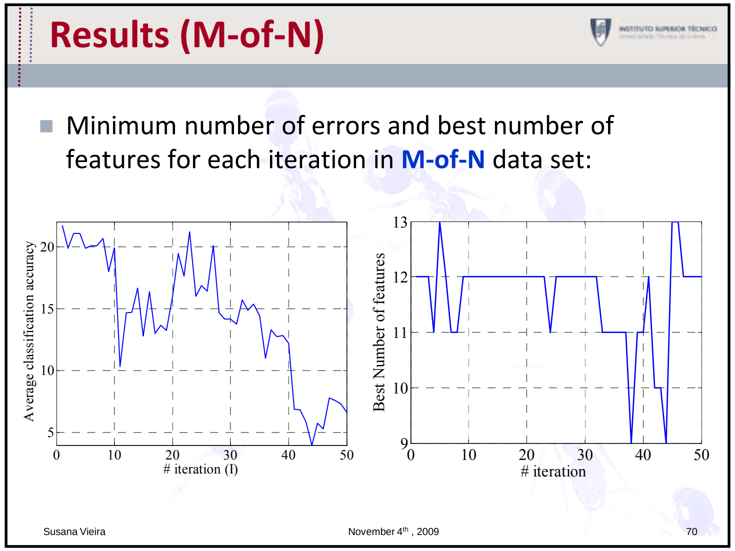
Results (Vote)

- Minimum number of errors and best number of features for each iteration in **Vote** data set:



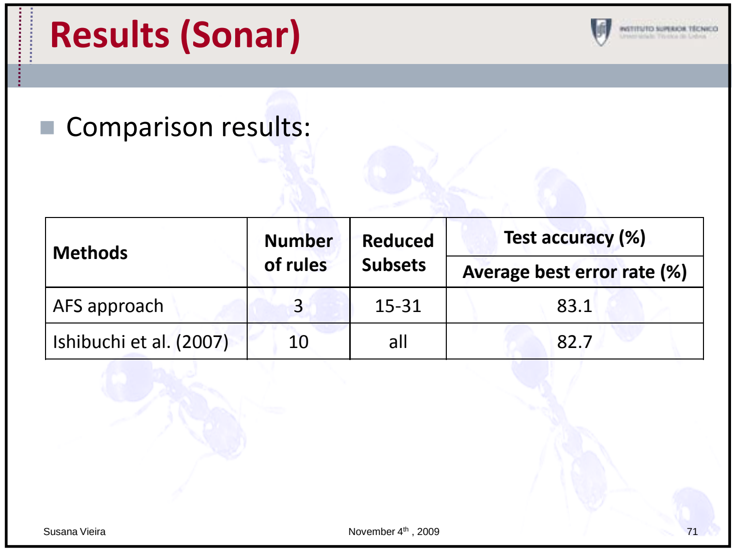
Results (M-of-N)

Methods	Reduced Subsets	Classification accuracy (%)		
		Best	Mean	Worst
AFS Approach	9	100	100	100
Wang et al. (POSAR) (2004)	7	100	-	-
Wang et al. (CEAR) (2004)	7	100	-	-
Wang et al. (DISMAR) (2003)	6	100	-	-
Wang et al. (GAAR) (2000)	6	100	-	-
Wang et al. (PSORSFS) (2007)	6	100	-	-



Results (M-of-N)

- Minimum number of errors and best number of features for each iteration in **M-of-N** data set:



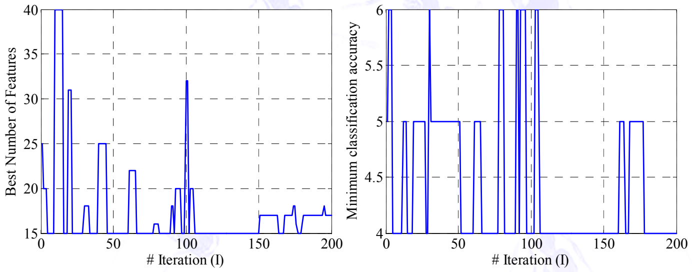
Results (Sonar)

- Comparison results:

Methods	Number of rules	Reduced Subsets	Test accuracy (%)
			Average best error rate (%)
AFS approach	3	15-31	83.1
Ishibuchi et al. (2007)	10	all	82.7

Results (Sonar)

- Minimum number of errors and best number of features for each iteration in **Sonar** data set:



Results

- Computational time and number of rules:

Dataset	#Samples	#Features	#Rules	Time (s)
Wine	178	13	9	830
Breast Cancer	699	9	4	567
Vote	300	16	6	434
M-of-N	1000	13	6	125
Sonar	208	60	6	1343