Generating Fuzzy Term Sets for Software Project Attributes using Fuzzy C-Means and Real Coded Genetic Algorithms

Ali Idri, Ph.D., ENSIAS, Rabat
Alain Abran, Ph.D., ETS, Montreal
Azeddine Zahi, FST, Fes

International Conference on ICT for Muslim World

November, 21-23, 2006, Kuala Lampur, Malaysia

Outline

- Motivations and Objectives
- Fuzzy C-Means for Clustering Software Project Attributes
- Building Membership Functions of Fuzzy sets using Real Coded Genetic Algorithms
- Overview of Empirical Results
- Conclusions and Future Work

Motivations and Objectives

- **Software project attributes are used by estimation models in software engineering to predict some important attributes of future entities such as software development effort, software reliability and programmers productivity**
  - **Software cost estimation models use as inputs software size, software reliability, and experience of the personnel to estimate the required software development effort**

- **Problem :**
  - **Many software project attributes are measured either on Nominal or Ordinal scale type composed of linguistic values such as, *low*, *very low*, *complex*, etc.**
  - **In the COCOMO II software cost estimation model 17 among 23 cost drivers are measured on an Ordinal scale composed of six linguistic values, *very low*, *low*, *nominal*, *high*, *very high*, and *extra-high***

ENSIAS

- **when dealing with linguistic values handling imprecision, uncertainty and partial truth is unavoidable**

⊙ **However, the software engineering community often uses numbers or classical intervals to represent these linguistic values**

- **Such transformation and representation does not mimic the way in which humans interpret linguistic values and consequently cannot deal with imprecision and uncertainty**

⊙ **To overcome this limitation, we have suggested the use of fuzzy sets rather than classical interval (or numbers) to represent linguistic values imitation (Idri, 2000-2006)**

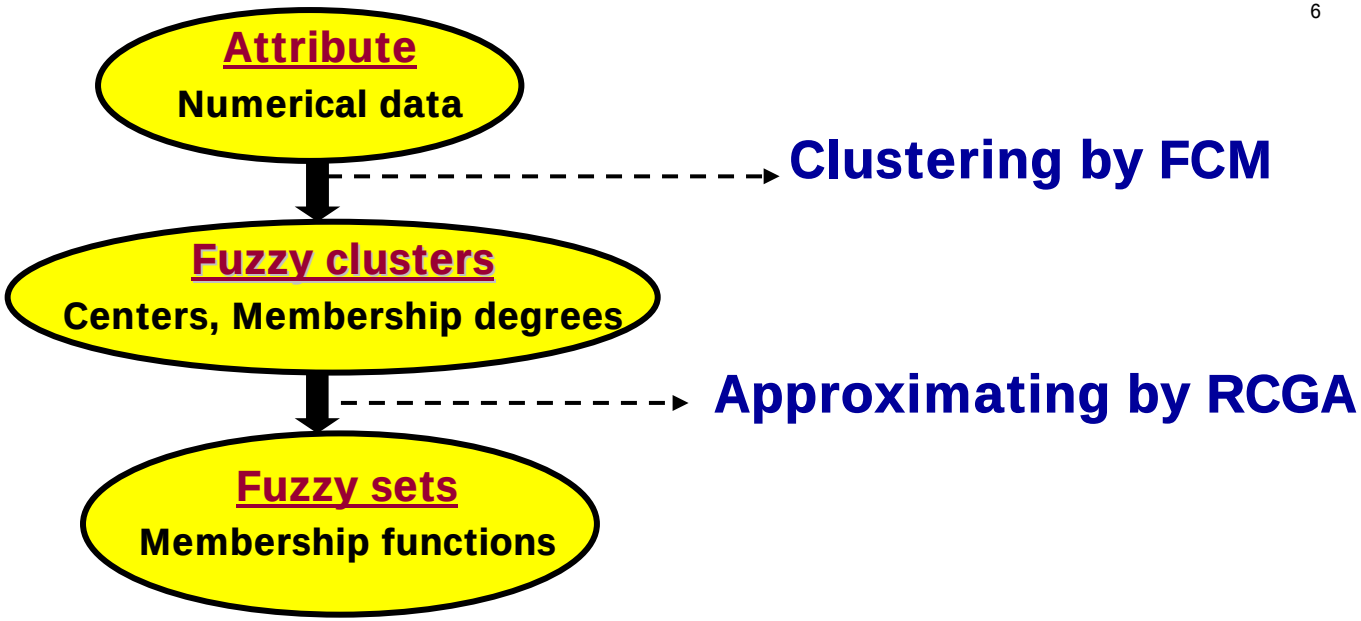
⊙ **The fuzzy sets and their membership functions are defined by using:**

- **Empirical techniques which construct membership functions from expert knowledge; or**
- **Automatic techniques, which construct membership functions from historical data using clustering techniques**

⊙ **Because, in many cases, the descriptions given of the software attributes are insufficient to empirically build their fuzzy representations -> Automatic techniques**

⊙ **Objective**

we suggest the use of the Fuzzy C-Means clustering technique (FCM) and a Real Coded Genetic Algorithm (RCGA) to build the fuzzy representations of the software attributes



- **The validation is done on a dataset that contains 263 historical software projects**
  - **Each project is described by 13 attributes**
    - ✓ **Software size measured in terms of KDSI**
    - ✓ **12 attributes related to the software development environment such as software complexity, the method used in the development and the time and storage constraints imposed on the software**

ENSIAS

Attribues	Designation
SIZE	Software Size
DATA	Database Size
TIME	Execution Time Constraint
STOR	Main Storage Constraint
VIRTMIN, VIRT MAJ	Virtual Machine Volatility
TURN	Computer Turnaround
ACAP	Analyst Capability
AEXP	Applications Experience
PCAP	Programmer Capability
VEXP	Virtual Machine Experience
LEXP	Programming Language Experience
SCED	Required Development

Fuzzy C-Means for Clustering Software Project Attributes

- The FCM algorithm is a fuzzy clustering method used to generate a known number of clusters (c) from a set of numerical data
- The determination of this number is still an open problem in clustering. Often, empirical knowledge or a set of evaluation criteria is used. In this work, we use the Xie-Beni fuzzy cluster validity criterion proposed
- FCM is an iterative algorithm that aims to find cluster centers $(C_j), 1 \le j \le c$ and the matrix $U = (u_{ij}), 1 \le i \le n, 1 \le j \le c$ that minimize the following objective function:

$$Min\ J_m(U,C) = \sum_{i=1}^{i=n}\sum_{j=1}^{j=c}(u_{ij})^m \left\|x_i - c_j\right\|^2 \qquad \sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,\ldots,n$$

- where **m** is the control parameter of fuzziness; $U = (u_{ij})$ is the partition matrix, containing the membership values of all data in all clusters;

⊙ The outline of the FCM algorithm can be stated as follows (Bezdek, 1981):

➢ **Step 1:** Randomly initialize the membership matrix (U) that has the following constraints:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,...,n; \quad 0 \le u_{ij} \le 1$$

➢ **Step 2:** Calculate centroids(c_i) by using the equation:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}$$

➢ **Step 3:** Compute dissimilarity between centroids and data points . Stop if its improvement over previous iteration is below a threshold.

➢ **Step 4:** Compute a new U using the following equation. Go to Step 2.

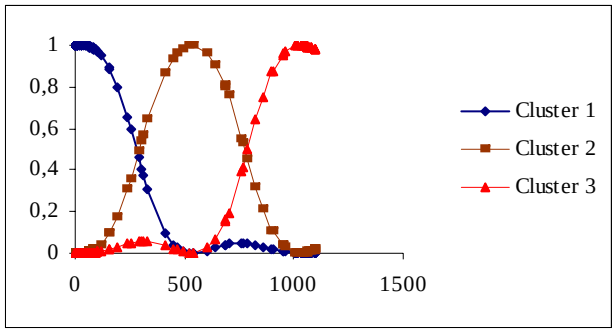
$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{2/(m-1)}}$$

ENSIAS

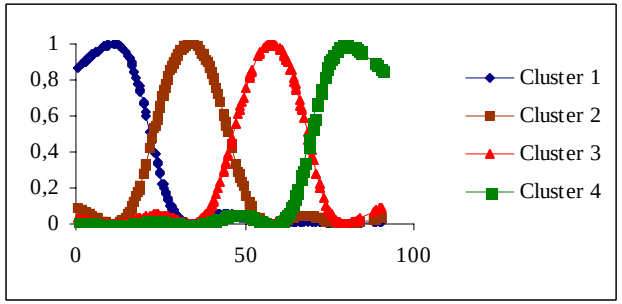
- For each COCOMO'81 software attribute, several experiments were conducted with the FCM algorithm, each time using a different initial matrix U. The desired number of clusters (c) is varied within the interval [3,6]

- The parameter m is fixed to 2 in all experiments.

- We use the Xie-Beni criterion to decide on the number of clusters. For each attribute, we choose the number of clusters that minimizes the value of the Xie-Beni criterion.



Data attribute



Time attribute

Building Membership Functions of fuzzy sets for the COCOMO'81 attributes

- After generating fuzzy sets (clusters) with their partition by means of FCM, we use an RCGA to build membership functions for these clusters;
- Membership functions can be trapezoidal, triangular or Gaussian.
- Our RCGA consists in building a set of membership functions that interpolates and minimizes the mean square error, which is defined as follows:

$$MSE(\mu_1,..,\mu_c) = \frac{1}{n}\sum_{j=1}^{j=n}\left\|(\mu_1(x_j),..,\mu_c(x_j)) - (u_{1j},..,u_{c,j})\right\|^2$$

subject to

$$\sum_{j=1}^{c}\mu_j(x_i) = 1 \qquad \mu_j(x_i)=u_{ij},\ 1\le i\le n;\ 1\le j\le c$$

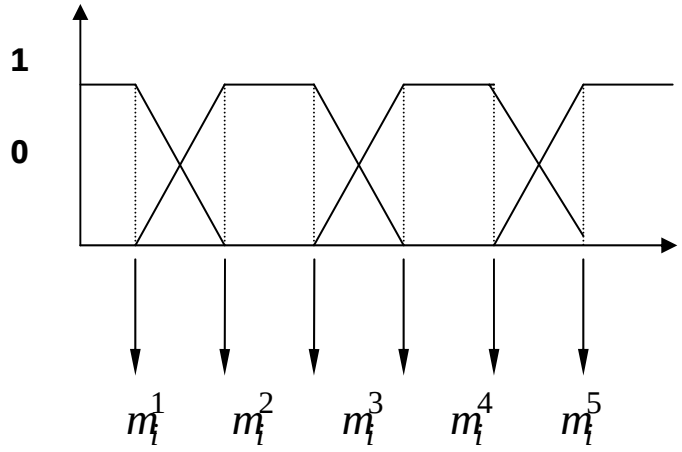
- **The use of an RCGA to find membership functions requires the determination of certain parameters, such as:**
  - **the coding scheme,**
  - **the fitness function, and**
  - **the various genetic operators (selection, crossover and mutation).**

- **Concerning the coding scheme, a chromosome represents the set of the unknown membership functions, $(\mu_j), 1 \le j \le c$, associated with the c fuzzy sets generated by the FCM**

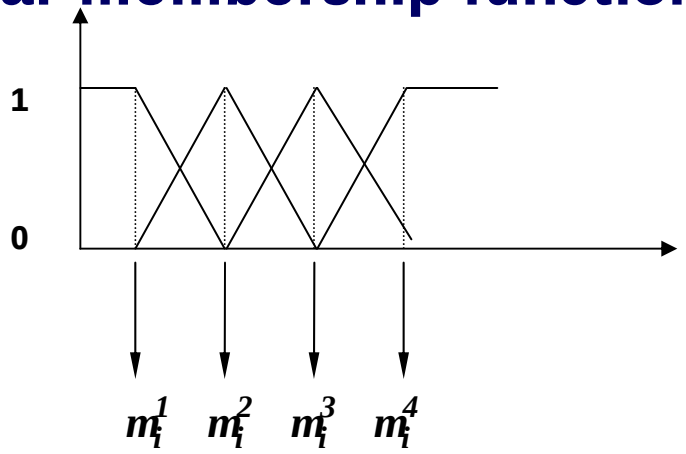
  - **The shape of the membership functions can be trapezoidal, triangular or Gaussian**

  - **Thus, each chromosome encodes a set of membership functions in a real vector $(m_i^1,...,m_i^K)$. The genes m_i^j are obtained from the shape of the membership functions.**

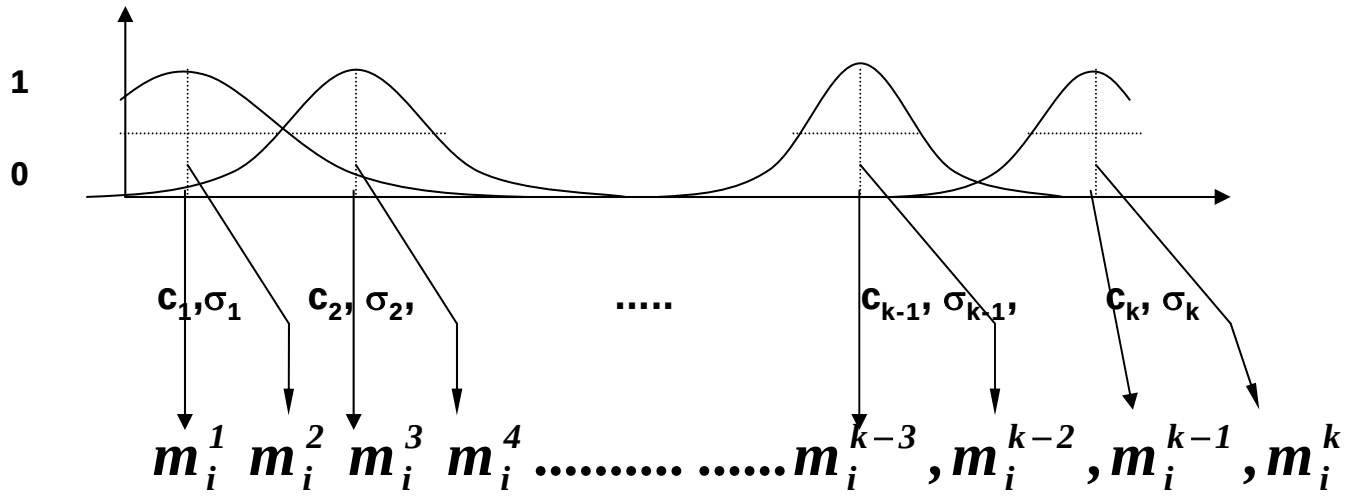
For trapezoidal membership functions



For triangular membership functions



For Gaussian membership functions

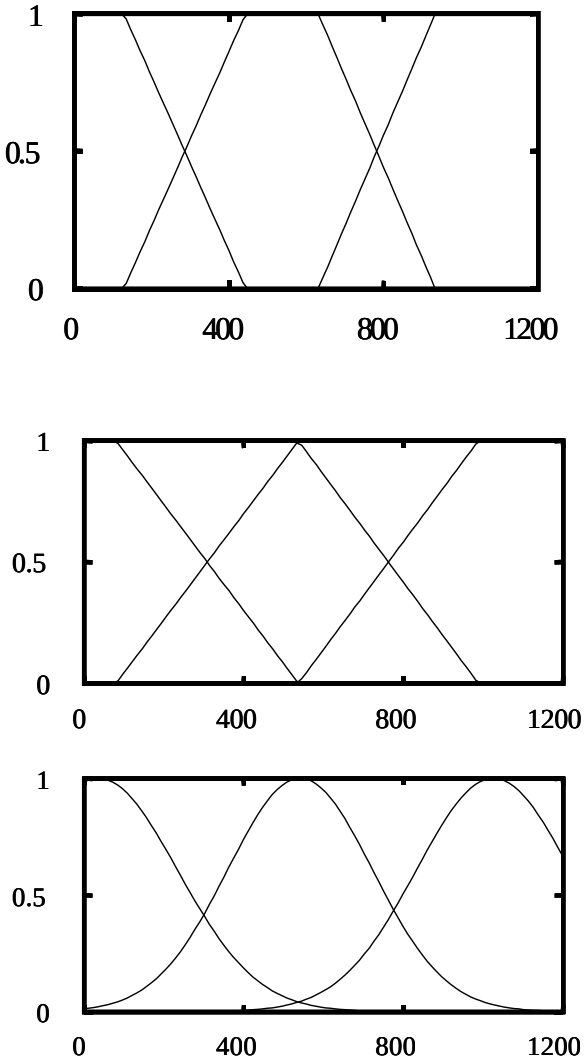


- **The fitness function F is obtained using the following formula**

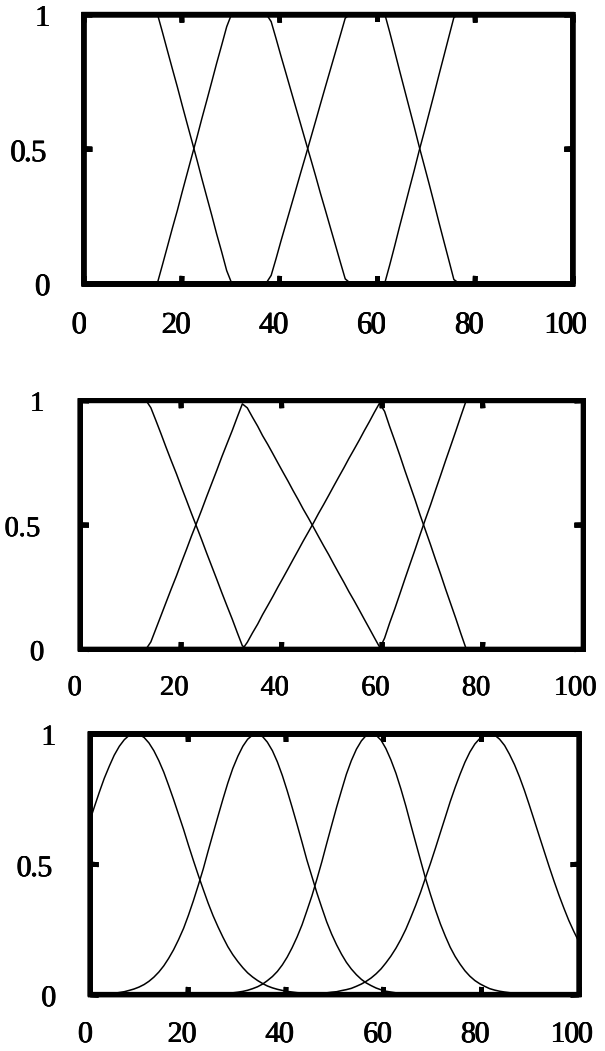
$$MSE(m_i) = \frac{1}{n}\sum_{j=1}^{j=n}\left\|\mu(x_j) - y_j\right\|^2 \qquad F(m_i) = \frac{MSE(m_i)}{\sum_{i=1}^{i=M} MSE(m_i)}$$

- **For the three genetic operators (selection, crossover and mutation), we use those that are specific to RCGAs**

Data attribute



Time attribute



Conclusion and Future Work

- We have used the Fuzzy C-Means clustering technique (FCM) and a Real Coded Genetic Algorithm (RCGA) (the FCM-RCGA process) to build the fuzzy representations for the COCOMO'81 attributes

- The membership functions generated may be trapezoidal, triangular or Gaussian.

- We are currently looking at the accuracy of cost estimation models based on CBR when using the FCM-RCGA process rather than empirical knowledge for building fuzzy sets (SETIT, March, 2007).

ENSIAS

Thank you

idri@ensias.ma
http://lrgl.uqam.ca