

# Software Project Similarity Measurement Based on Fuzzy c-Means

1

**MOHAMMAD AZZEH** [M.Y.A.AZZEH@BRAD.AC.UK](mailto:M.Y.A.AZZEH@BRAD.AC.UK)

**DR. DANIEL NEAGU** [D.NEAGU@BRAD.AC.UK](mailto:D.NEAGU@BRAD.AC.UK)

**PROF. PETER COWLING** [P.I.COWLING@BRAD.AC.UK](mailto:P.I.COWLING@BRAD.AC.UK)

**DEPARTMENT OF COMPUTING,  
SCHOOL OF INFORMATICS,  
UNIVERSITY OF BRADFORD,  
BRADFORD, U.K.,  
BD7 1DP**

# Agenda

2

- **Motivation**
- **The Problem.**
- **Proposed Solution.**
- **Results.**
- **Conclusions.**

# Motivation

3

- **Analogy software estimation is based on assumption:**  
“similar software projects have similar software effort”
- **But...**
  - Incomplete and noisy data.
  - Measurement and similarity assessment uncertainty.
  - Complex interaction between attributes.
  - Data type (Ordinal and Nominal scale).

# The Problem

4

- Two projects that may seem similar may indeed be different in a critical way.
- the uncertainty in assessing similarities and differences means that two different estimators could develop significantly different views and effort estimates.
- This uncertainty stem from:
  - Data Collection tool.
  - The type of information available (Nominal and ordinal scale attributes),
  - Attribute measurement, and
  - skill of estimator.

# Existing techniques

5

- **Nearest neighborhood based techniques such as Euclidean distance, weighted Euclidean, maximum distance, But:**
  - it is difficult to handle categorical variables.
  - They are sensitive to the irrelevant features and missing values.
- **Fuzzy distance between two linguistic variables.**
  - is compatible only with dataset that described as linguistic quantifiers (e.g. Low, high) such as COCOMO dataset.

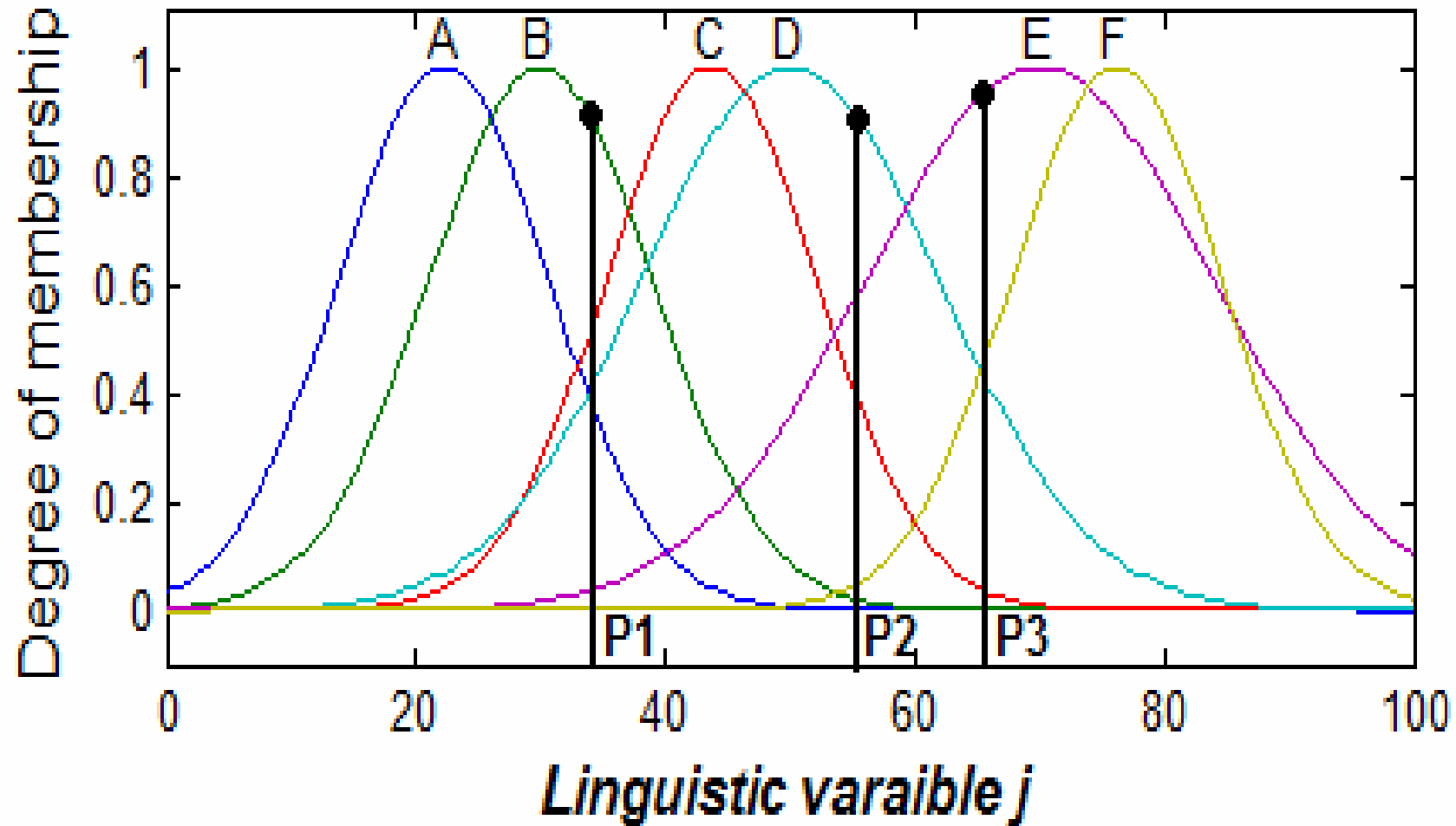
# The Proposed Model

6

- The proposed model is to assess similarity between software projects based on Fuzzy c-Means and Fuzzy Logic.
- The model assumes that the two software projects that most belong to the same fuzzy cluster at  $k^{\text{th}}$  feature have one similarity degree.

# The First approach

7



# The First Approach

8

- **Definition 1.** Let  $p_x$  and  $p_y$  be two software projects described by  $M$  features. Let  $F_j(p_x)$  mostly belong to fuzzy set and  $F_j(p_y)$  mostly belong to fuzzy set . The similarity between two software projects in terms of features  $j$  is given as:

$$SM_j(F_j(p_x), F_j(p_y)) = \min\left(e^{\frac{-(x-y)^2}{(\sigma_x + \sigma_y)^2}}, 1\right)$$

- where  $x, y$  are the mean values and  $\sigma_x, \sigma_y$  are the standard deviation for and respectively.

# The First Approach

9

- **Definition 2.** The Overall similarity between two software projects  $p_x$  and  $p_y$  is given by equation 8:

$$SM_1(p_x, p_y) = \text{avg}_{j=1}^M (SM_j(F_j(p_x), F_j(p_y)))$$

# Second Approach—based on Partition matrix of Fuzzy c-Means

Project	Cluster $A_1$	Cluster $A_2$	...	Cluster $A_k$
$P_1$	$\mu_{A_1}(p_1)$	$\mu_{A_2}(p_1)$	...	$\mu_{A_k}(p_1)$
$P_2$	$\mu_{A_1}(p_2)$	$\mu_{A_2}(p_2)$	...	$\mu_{A_k}(p_2)$
...	...	...	...	...
$P_n$	$\mu_{A_1}(p_n)$	$\mu_{A_2}(p_n)$	...	$\mu_{A_k}(p_n)$

$$SM_2(p_x, p_y) = \sum_{i=1}^k \min(\mu_{A_i}(p_x), \mu_{A_i}(p_y))$$

# Property of similarity measurement

11

- **Property 1.**  $SM(p_x, p_x)=1$
- **Property 2.**  $SM(p_x, p_y) < SM(p_x, p_x)$ .
- **Property 3.**  $SM(p_x, p_y) = SM(p_y, p_x)$
- **Property 4.** Let  $p_x, p_y, p_z$  be three projects where  $p_x$  mostly belong to fuzzy set  $A$ , and  $p_y$  mostly belong to fuzzy set  $B$ ,  $p_z$  mostly belong to fuzzy set  $C$ , and  $F_j(p_x) < F_j(p_y) < F_j(p_z)$  for all features then  $SM(p_x, p_z) < \min(SM(p_x, p_y), SM(p_y, p_z))$ .

# Advantages of proposed models

12

- **Advantages:**
  - handling categorical in the same way numerical attributes are handled where they are represented by fuzzy sets.
  - using Fuzzy C-means has the advantage to group close projects together in the same cluster and then represent them in the same fuzzy set.
  - Reducing uncertainty in similarity assessment.

# Results

13

	P1	P2	P3	P4	P5
P1	1	0.97	0.86	0.93	0.656
P2	0.97	1	0.684	0.959	0.635
P3	0.87	0.684	1	0.841	0.568
P4	0.93	0.956	0.845	1	0.58
P5	0.65	0.626	0.569	0.59	1

# Results

14

Evaluation Criteria	First proposed approach		
	<i>One Analogy</i>	<i>Mean of two analogies</i>	<i>Mean of three analogies</i>
MMRE	18.23%	16.06%	13.55%
BMMRE	22.4%	17.7%	17.0%
PRED(25%)	72%	78%	84%
Second proposed approach			
	<i>One Analogy</i>	<i>Mean of two analogies</i>	<i>Mean of three analogies</i>
MMRE	20.3%	16.6%	16.8%
BMMRE	26.6%	23.4%	22%
PRED(25%)	72%	74%	74%
Euclidean Distance			
	<i>One Analogy</i>	<i>Mean of two analogies</i>	<i>Mean of three analogies</i>
MMRE	59.4%	62.8%	65.6%
BMMRE	62.2%	60.1%	64.7%
PRED(25%)	44.4%	44.6%	45.4%
Weighted Euclidean Distance			
	<i>One Analogy</i>	<i>Mean of two analogies</i>	<i>Mean of three analogies</i>
MMRE	59.4%	55.2%	56.7%
BMMRE	62.2%	59%	59.3%
PRED(25%)	44.4%	47.9%	50.2%

# Outstanding issues

15

- when two different projects feature values fall in the fuzzy set and is given unity similarity degree as similarity between a project to itself.
- our approach is restricted to Gaussian membership function.
- the project must be defined by more than one feature!!

# Questions?