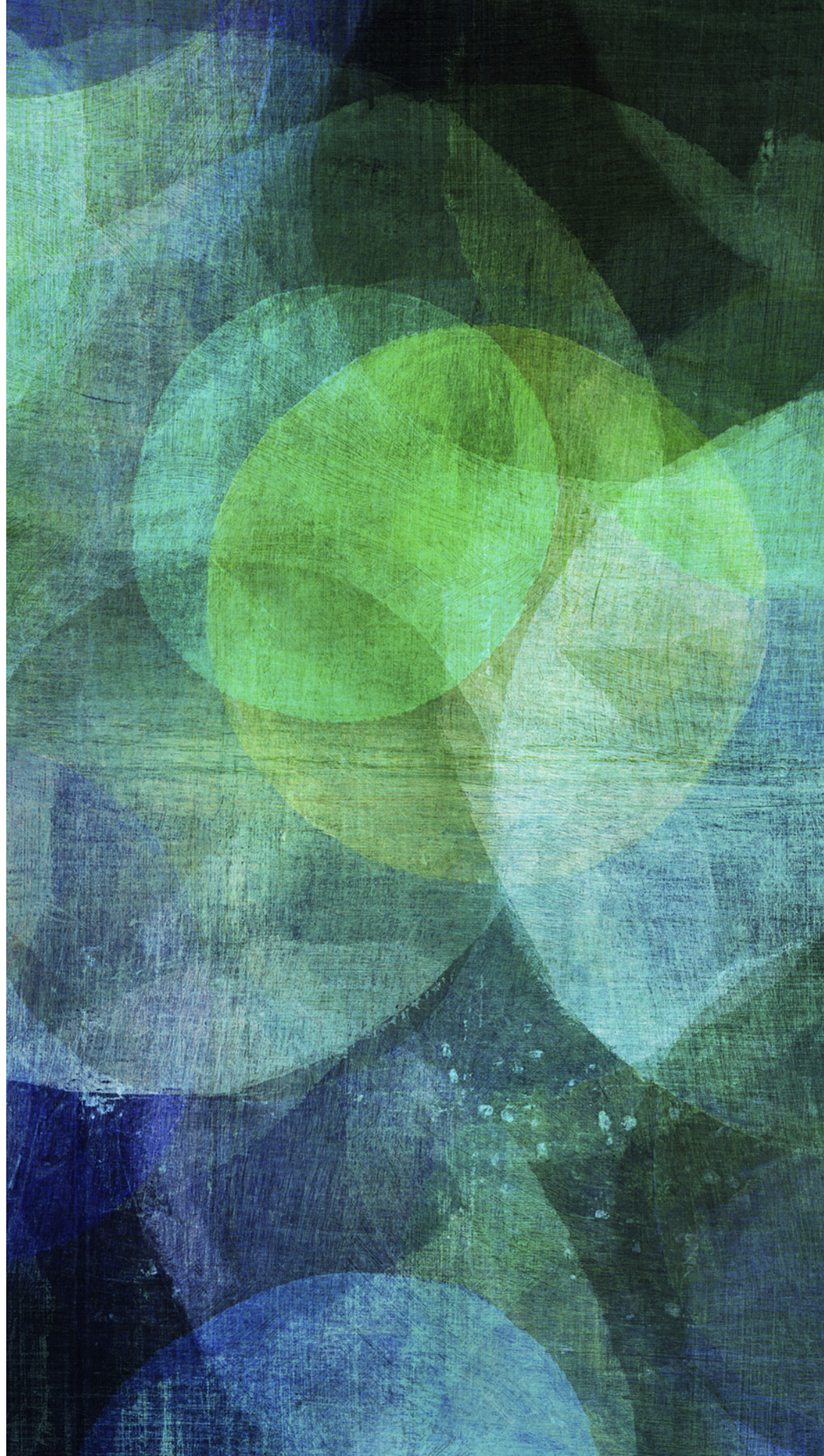
A HYBRID METHOD FOR IMPUTATION OF MISSING VALUES USING OPTIMIZED FUZZY C-MEANS WITH SUPPORT VECTOR REGRESSION AND A GENETIC ALGORITHM

Saman Khamesian

Spring 2024 - EMIL Lab Presentation

CONTENTS

INTRODUCTION



INTRODUCTION

- Missing values in datasets should be extracted from the datasets or should be estimated before they are used for classification, association rules or clustering in the preprocessing stage of data mining.
- Missing values are highly undesirable in data mining, machine learning and other information systems.
- Missing values typically occur because of:
  - ✓ Sensor faults
  - ✓ Lack response in scientific experiments
  - ✓ Data transfer problems in digital systems
  - ✓ …
- To deal with missing values in datasets:
  - ✓ Ignoring
  - ✓ Deleting
  - ✓ Zero or mean estimation methods
- The primary disadvantages of these estimation methods are the loss of efficiency.

INTRODUCTION

- There are three types of missing data:
  - ✓ Missing data completely at random (MCAR)
    The missing value has no dependency on any other variable.
  - ✓ Missing at random (MAR)
    The missing value depends on other variables. The missing value can be estimated using other variables.
  - ✓ Missing not random (MNAT)
    The missing value depends on other missing values, and thus missing data cannot estimated from existing variables.
- A typical dataset containing missing values can be divided into two sections:
  - ✓ Complete rows
    A complete record is a row of the dataset in which no attributes have missing value(s).
  - ✓ Incomplete rows
    An incomplete record is a row of the dataset in which one or more columns have missing value(s).

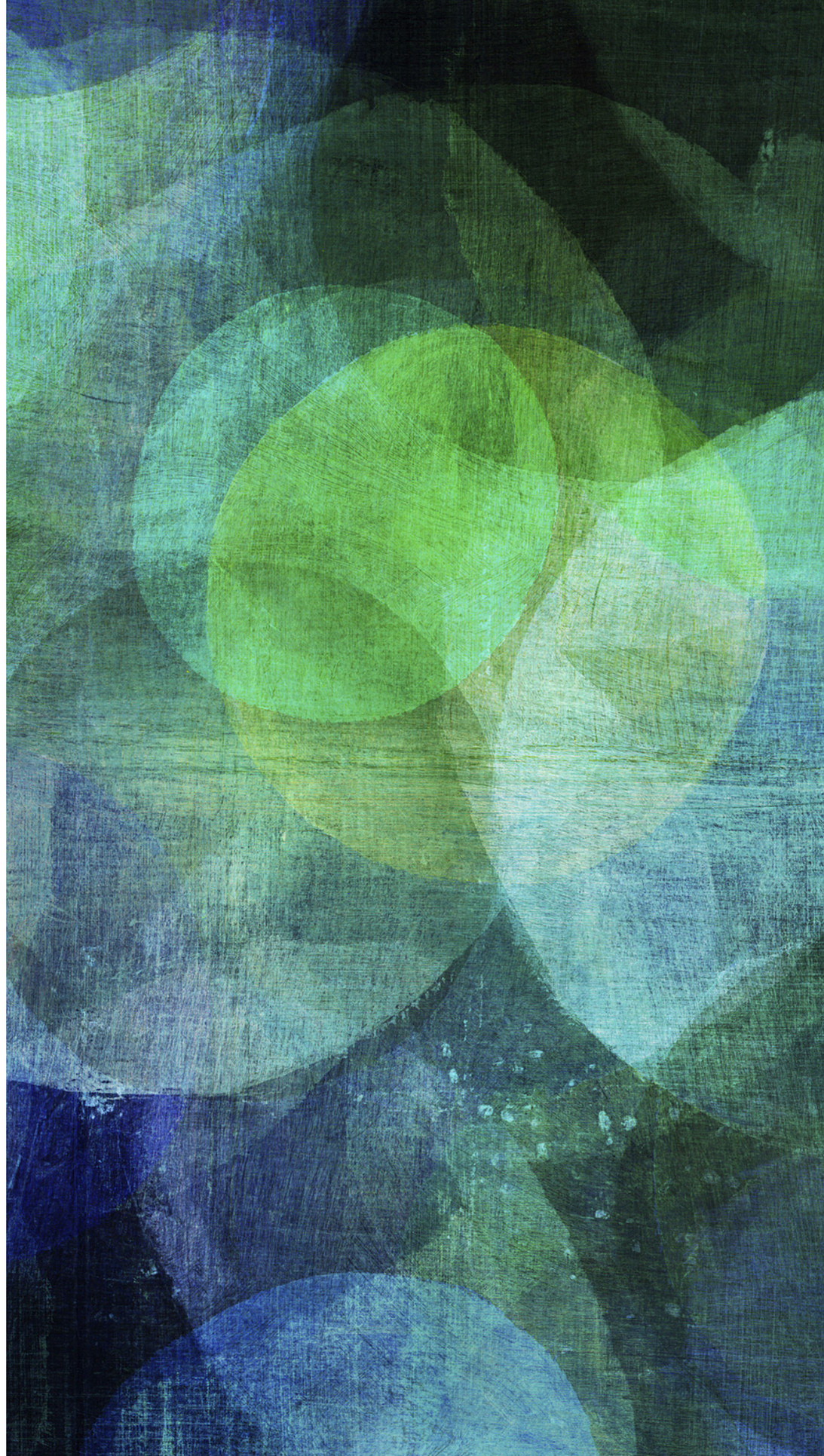
INTRODUCTION

- Y1, Y2, Y3, Y4, Y5 and Y6 are records (rows).
- X1, X2, X3, X4 and X5 are attributes (columns).
- Y2, Y5 and Y6, which do not have any missing values, are complete rows, and Y1, Y3 and Y4, which have missing values, are called incomplete rows.
- X1 and X4 which are available in all records, are reference attributes, and X2, X3 and X5, which are non-reference attributes.

	X1	X2	X3	X4	X5
Y1	0.113524	0.084785	?	0.625473	0.06385
Y2	0.112537	0.138211	0.15942	0.625473	0.068545
Y3	0.110563	?	0.144928	0.624212	0.083568
Y4	0.110563	0.170732	0.146998	0.623581	?
Y5	0.108588	0.129501	0.144928	0.624212	0.076056
Y6	0.108588	0.082462	0.112836	0.626103	0.015023

A section of a dataset with missing values

FCM MODEL



FUZZY C-MEANS INTRODUCTION

- Given a set of objects, the overall objective of clustering is to divide the dataset into groups based on the similarity of the objects and to minimize the intra-cluster dissimilarity.
- Fuzzy c-means (FCM) (Also called soft K-means or Fuzzy K-means) is a method of clustering that allows one datum to belong to two or more clusters.
- In fuzzy clustering, each data object x_i has a membership function, which describes the degree to which the data object belongs to a certain cluster c_j:

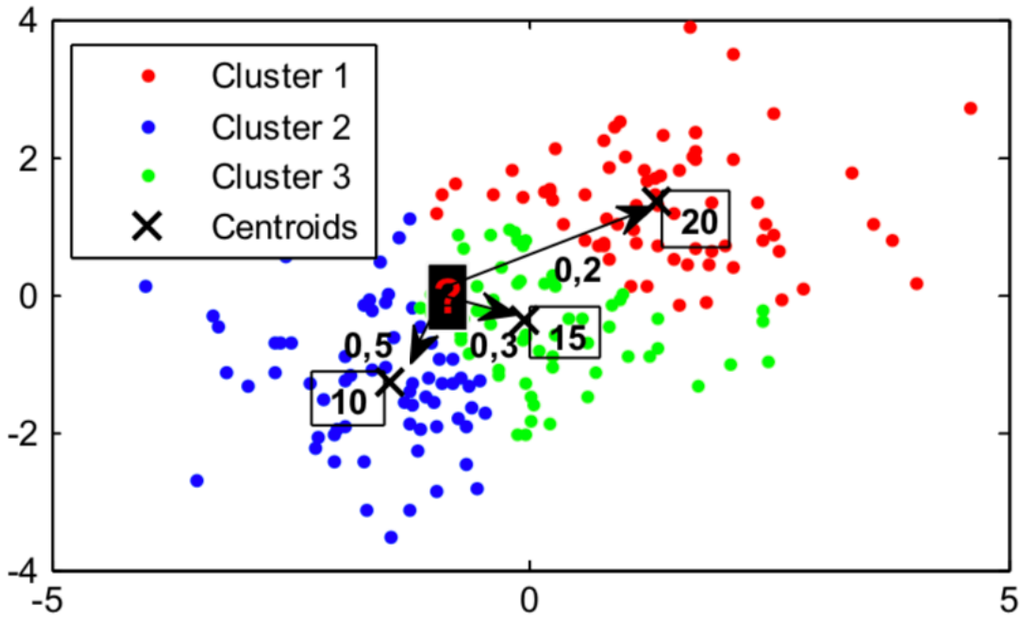
$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{||x_i - c_j||}{||x_i - c_k||}\right)^{\frac{2}{m-1}}} \text{ (membership function)} \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \text{ (centroid cluster j)}$$

- $u_{ij} \in [0,1]$
- C: number of clusters $2 \leq C \leq N$
- m: is a parameter called the weighting factor $1 < m < \infty$ (This parameter controls the amount of fuzziness in the clustering process)
- There is no theoretical optimal choice of c and m. The parameters can be changed depending on the characteristics of the dataset and relation of attributes with each other.

FCM MODEL

- In fuzzy c-means imputation, the missing value of the incomplete data object x_i is estimated using the information about membership degrees and the values of the cluster centroids:

  1. Randomly select C complete data objects as C centroids (use only reference attributes to compute the cluster centroids)
  2. Iteratively update membership functions and centroids until the overall distance meets the user-specified distance threshold ε.
  3. Impute non-reference attributes for each incomplete object: $x_{i,j} = \sum_{k=1}^{C} U(x_i, c_k) \,.\, c_{k,j}$



Fuzzy c-means (FCM) imputation

Assume C=3 and m=2

Membership values are estimated as 0.5, 0.3 and 0.2

Centroids are estimated as 10, 15 and 20

The missing value (?) is calculated as:

? = 0.5*10 + 0.3*15 + 0.2*20 = 13.5

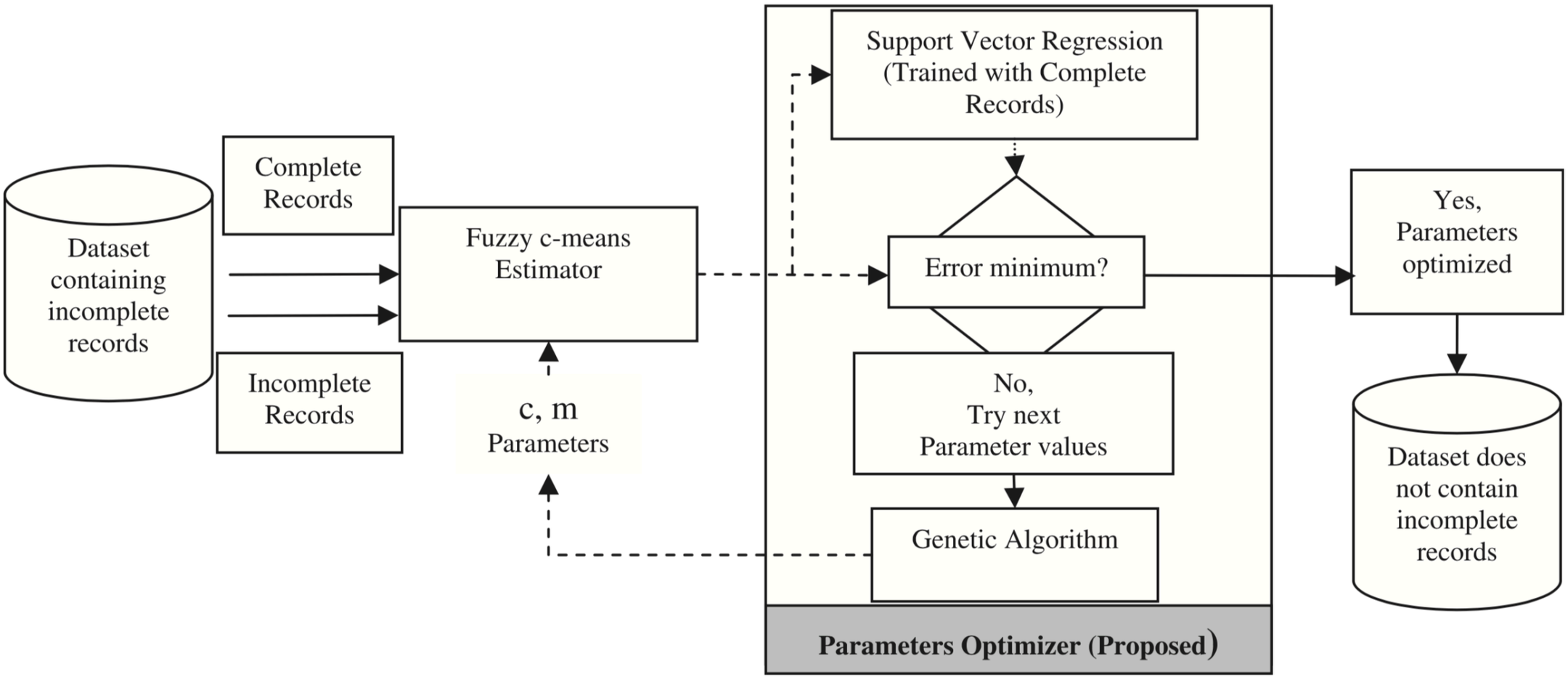
FCM AND SVR+GA

- We can estimate missing values using fuzzy c-means, where C is the number of clusters and m is the parameter of weighting factor.
- The optimal cluster number (C) and weighting factor (m) should be determined to obtain the best predictive accuracy.
- The purpose of the genetic algorithm, in cooperation with support vector regression, is to minimize error.
- The pseudocode of the proposed method, fuzzy c-means imputation optimized with SVR-GA, is as follows:

1. Train the support vector regression algorithm with the dataset (complete) rows, for which $\text{Input}(X) \approx \text{Output}(Y)$.
2. Estimate the dataset (incomplete) rows using fuzzy c-means, and compare the fuzzy c-means output with the SVR output vector.
3. Obtain the optimized c and m parameters by using the genetic algorithm to minimize the difference between the SVR output and the fuzzy c-means output.
4. Estimate the missing values using fuzzy c-means with optimized parameters.

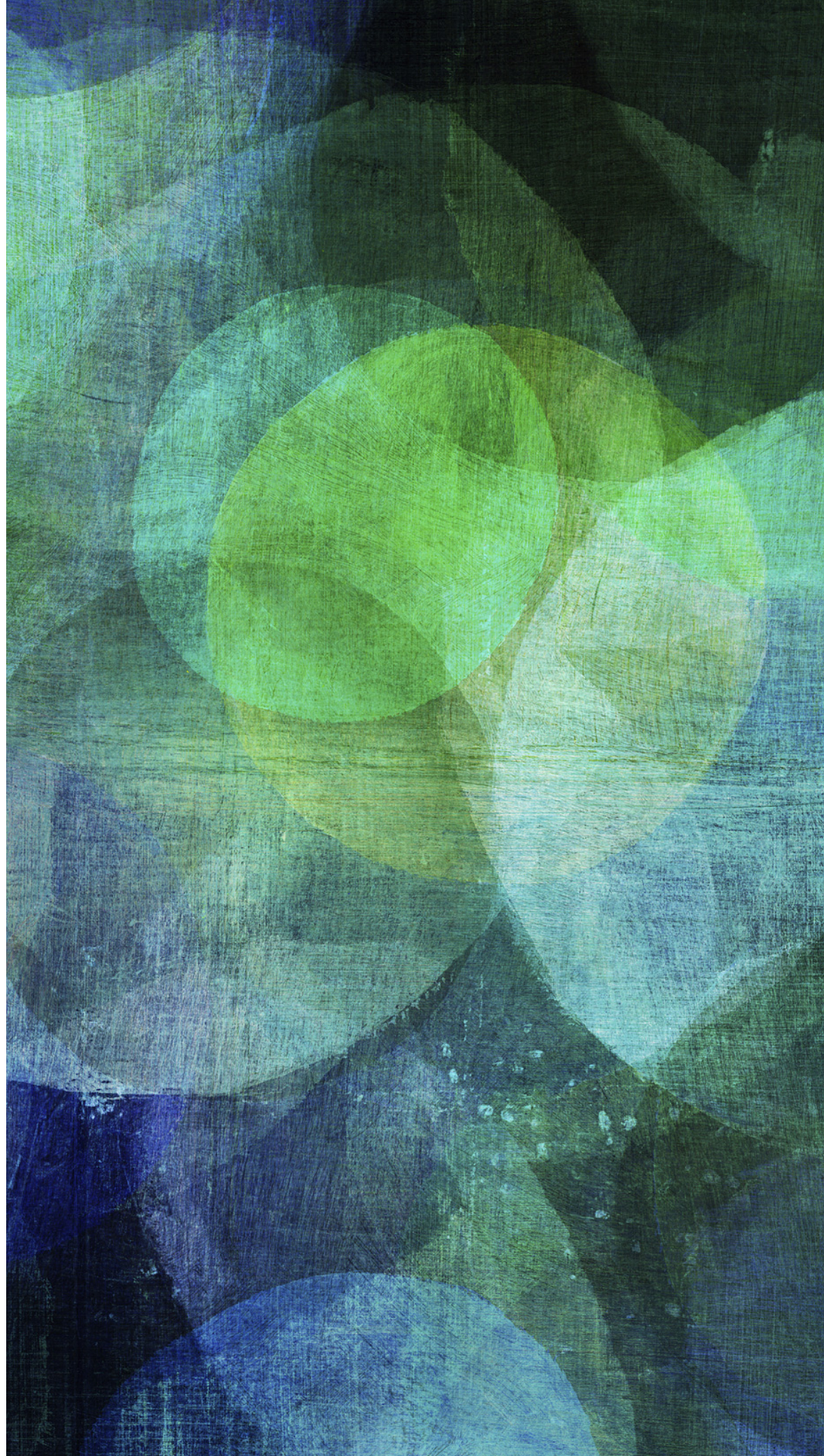
FCM AND SVR+GA

- The minimized error function is, $error = (X - Y)^2$ where X is the output of the support vector regression (SVR) prediction and Y is the output of fuzzy c-means algorithm prediction.



The proposed fuzzy c-means (FCM) SVR-GA imputation method

EXPERIMENTAL RESULTS



EXPERIMENTAL RESULTS

- System information:
  - ✓ Intel Core 2 quad-core CPU
  - ✓ 4.00 GB RAM
  - ✓ Microsoft Windows XP SP2 operating system

- Implementation information:
  - ✓ MATLAB R2009b version 7.9
  - ✓ Using LS-SVM toolbox
    - ✓ Using a radial basis kernel → $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$
  - ✓ Using genetic algorithm toolbox
    - ✓ Population size = 20
    - ✓ Generations = 40
    - ✓ Crossover fraction = 60%
    - ✓ Mutation fraction = 3%

EXPERIMENTAL RESULTS

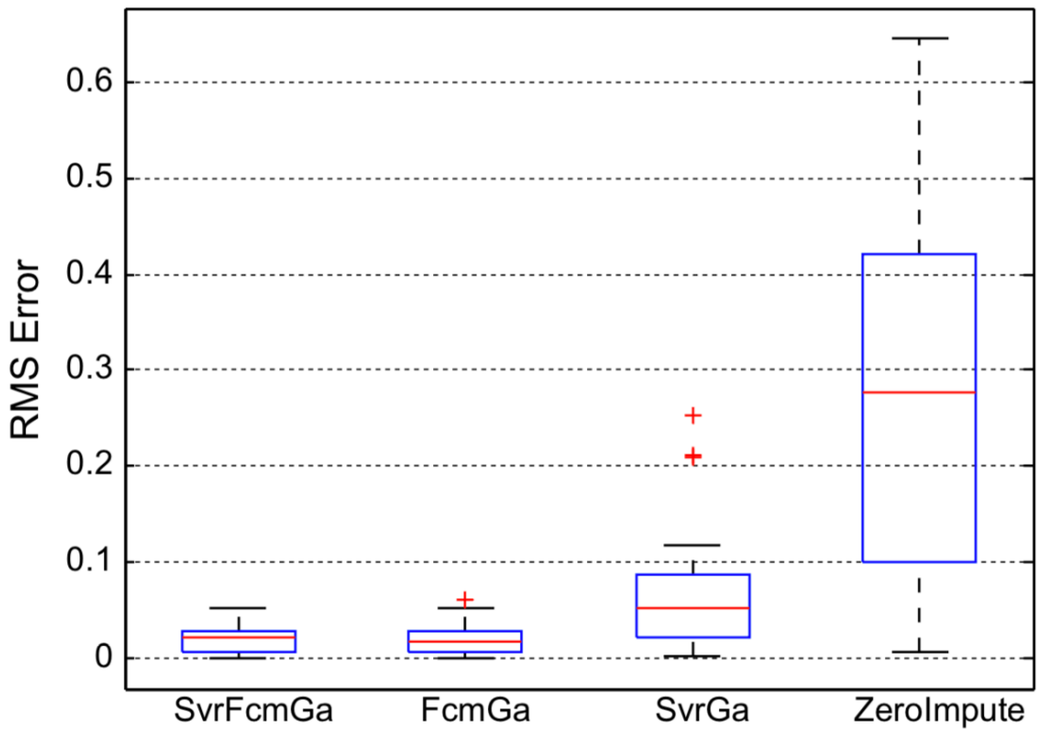
- Datasets information:
  - ✓ All datasets are available in UCI Repository of Machine Learning.
  - ✓ All the datasets are transformed using a min–max normalization to {0,1} before use, Because of Testing the algorithms under equal conditions.
  - ✓ All datasets are artificially regenerated such that they have 1%, 5%, 10%, 15%, 20% and 25% missing value ratios.

Dataset name	Records	Attributes
Glass	214	11
Haberman	306	4
Iris	150	5
Musk1	476	167
Wine	178	14
Yeast	1489	9

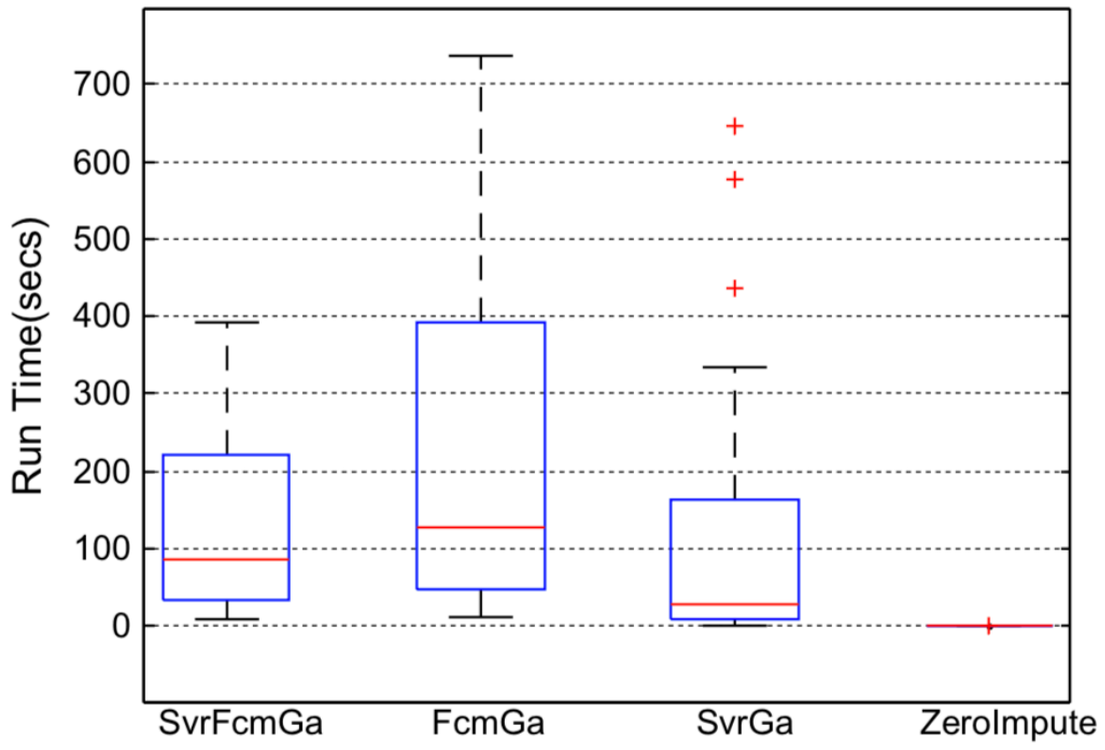
The datasets used

EXPERIMENTAL RESULTS

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x_i})^2}{n}}$$



RMS error in overall datasets for datasets missing 1–25% of the values



Runtime (s) for datasets in which 1–25% of the data are missing

REFERENCES

- I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Information Sciences*, vol. 233, pp. 25–35, Jun. 2013.
- V. Vapnik, S.E. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, Adv. Neur. Inform. 9 (1997) 281–287.
- H.H. Feng, G.S. Chen, C. Yin, B.R. Yang, Y.M. Chen, A SVM regression based approach to filling in missing values, Proc. Knowled.-Based Intell. Inform. Eng. Syst. 3683 (Pt 3) (2005) 581–587.
- T. Marwala, Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques, Information Science Reference, Hershey PA, 2009.

“

Thank you for your attention