# Leachable Component Clustering

**Miao Cheng[a], Xinge You[b]**

[a]School of Computer Science, Guangxi Normal University, Guilin, China

[b]School of Electronics and Information Communications, Huazhong University of Science and Technology, Wuhan, China

August 21-25 • Montréal, Québec
ICPR 2022
26TH International Conference on Pattern Recognition

## Abstract

Clustering attempts to partition data instances into several distinctive groups, while the similarities among data belonging to the common partition can be principally reserved. Furthermore, **incomplete data** frequently occurs in many real-world applications, and brings perverse influence on pattern analysis. As a consequence, the specific solutions to data imputation and handling are developed to conduct the missing values of data, and independent stage of knowledge exploitation is absorbed for information understanding. In this work, a novel approach to clustering of incomplete data, termed **leachable component clustering**, is proposed. Rather than existing methods, the proposed method handles data imputation with **Bayes alignment**, and collects the lost patterns in theory. Due to the simple numeric computation of equations, the proposed method can learn optimized partitions while the **calculation efficiency** is held. Experiments on several artificial incomplete data sets demonstrate that, the proposed method is able to present superior performance compared with other state-of-the-art algorithms.

## Contributions

- The proposed method seeks for optimized imputation, by exploiting the imputation models with respect to preservation of intrinsic distributions.
- The proposed imputation models aim to fulfill the lost elements with the defined objectives, while fixed solutions can be obtained with calculation of the equations.
- The proposed framework can be generalization of analysis of incomplete data, and leads to the possible extensions of further advances.

## Contact

Miao Cheng
Email: mcheng@mailbox.gxnu.edu.cn
miao_cheng@outlook.com

## Self-Expressive Clustering

Given a set of data instances $X = [x_1, x_2, \cdots, x_n] \in R^{d\times n}$, and the amount of clustering $c > 0$. The clustering can be explained a self-expressive problem and defined as

$$J_{SEC} = arg\min_{Z} \frac{1}{2}\|X - XZ\|^2 + \gamma\|Z\|_F^2$$
$$s.t. \quad Z^T 1 = 1, \quad 0 \le Z \le 1.$$

## Latent Component Models

**● Principle Component Imputation Model**

According to the self-expressive property, the orthogonal components of $X$ and $\tilde{X}$ nearly share the common bases, e.g.,

$$J_P = arg\min_{P}\|X - PP^T X\|^2$$
$$s.t. \quad P^T P = I$$

**● Self-Expressive Representation Learning**

Representation learning attempts to learn the representative patterns of the original data with respect to specific objectives, e.g.,

$f: x_i \to y_i, \quad i = 1,2,\cdots,n.$

The similarity objectives of instances are normally referred, and defined as

$$J_S = arg\min_{y} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|Sim\left(\widehat{x_{i,j}}, \widehat{x_{j,i}}\right) - Sim\left(y_i, y_j\right)\right\|^2$$

With the similarity function $Sim(\cdot,\cdot)$, $\widehat{x_{i,j}}$ and $\widehat{x_{j,i}}$ indicate the labeled instances corresponding to the valid features between two instances, which are respectively defined as

$$\widehat{x_{i,j}} = x_i \odot l_i \odot l_j$$

and

$$\widehat{x_{j,i}} = x_j \odot l_j \odot l_i$$

The $l_i$ denotes the labels that represent the validation and missing values of instance $x_i$, e.g.,

$$l_{iq} = \begin{cases} 1, & if \quad x_{iq} \neq 0 \\ 0, & otherwise. \end{cases}$$
$$q = 1,2,\cdots,d.$$

## Leachable Component Learning

**● Distribution Alignment Model**

The distribution of samples with respect to a specific data $x_i$ can be approximately calculated as

$$G_{m,\sigma^2}(x_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x_i - m\odot l_i)^T(x_i - m\odot l_i)}{2\sigma^2}\right)$$

As a consequence, it attempts to learn the ideal approximation $\widetilde{x_i}$ that can reserve the statistical distributions of each incomplete data, such as

$$J_{BA} = arg\min_{\widetilde{x_i}}\left|G_{\widetilde{m},\widetilde{\sigma^2}}(\widetilde{x_i}) - G_{m,\sigma^2}(x_i)\right|,$$
$$i = 1,2,\cdots,n.$$

**● Solution of Latent Leachable Learning**

For the proposed model, it is to solve the following equation for each incomplete data,

$$\widetilde{x_i}\cdot\widetilde{x_i} - 2\widetilde{m}\cdot\widetilde{x_i} + \widetilde{m}\cdot\widetilde{m} + 2\tilde{\sigma}^2 log\left(\sqrt{2\pi}\tilde{\sigma}G_{m,\sigma^2}(x_i)\right) = 0$$

As a consequence, the ideal solution $\widetilde{x_i}$ to the proposed imputation model can be calculated as

$$\widetilde{x_i} = \widetilde{m} \pm \sqrt{2\tilde{\sigma}^2\tau\left(log\left(\sqrt{2\pi}\tilde{\sigma}G_{m,\sigma^2}(x_i)\right)\right)}.$$

Here, $\tau(\cdot)$ denotes the absolute function, such as

$$\tau(x) = \begin{cases} |x|, & if \quad x \neq 0 \\ 0, & othereise. \end{cases}$$

## Extensions

a. The proposed method is able to be recycled, due to its intrinsic relationship with data means.
b. Similarly, the original IVP imputation model aims to solve the equation associated with the reservation of information volumes,

c. $\sum_{j=1}^{k}\left(\widetilde{x_i}\cdot\widetilde{x_i} - 2\cdot x_j\cdot\widetilde{x_i} + x_j\cdot x_j\right) - \sum_{j=1}^{k} S_f(x_i, x_j)$
$= k\cdot\widetilde{x_i}\cdot\widetilde{x_i} - 2\cdot\widetilde{x_i}\cdot\sum_{j=1}^{k} x_j + \sum_{j=1}^{k} x_j\cdot x_j - \sum_{j=1}^{k} S_f(x_i, x_j)$

and

$$S_f(x_i, x_j) = Sim_f\left(\widehat{x_i}, \widehat{x_j}\right)$$

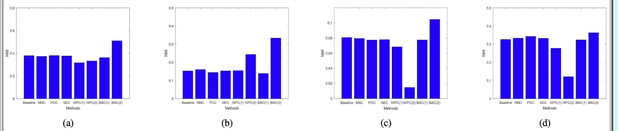
## Experiments

**● Data Sets**

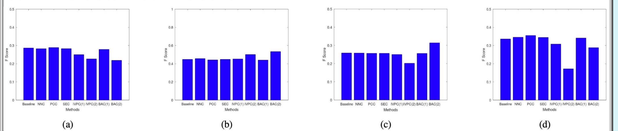
| Data | Partition | Dimensionality | Instance |
|---|---|---|---|
| Birds | 20 | 50 | 3625 |
| Firewall | 3 | 11 | 3000 |
| Flower | 5 | 50 | 4323 |
| Monkey Species | 10 | 50 | 1098 |

**● Results**

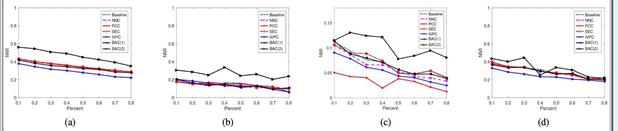
a. The obtained Normalized Mutual Information



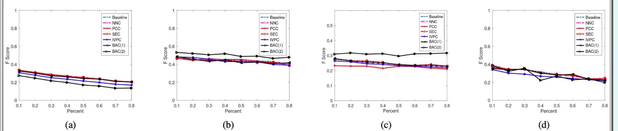
b. The obtained F Score



c. The obtained NMI associated with different percent of null elements



d. The obtained Rand Index associated with different percent of null elements



e. The time complexities (seconds) associated with different percent of null elements