Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 2, pp. 403–412
DOI : 10.47974/JIOS-1557

A conditionally positive definite kernel function for clustering of incomplete data

Sonia Goel *
Meena Tushir †
Department of Electrical and Electronics Engineering
Maharaja Surajmal Institute of Technology
Guru Gobind Singh Indraprastha University
New Delhi
India

Abstract

Clustering of incomplete data sets that contains missing features is one of the most widely studied problems in the literature, and several imputation and non-imputation techniques are used to solve this problem. A weighted sum of the Euclidean distance from the datum to the corresponding clusters is used in Fuzzy c-means clustering. It has been observed that the kernel-based clustering techniques outperform the conventional algorithms in terms of accuracy. This is due to their ability to handle non-linear data and map it to higher dimensional space while preserving its internal structure. Kernel functions are really important when it comes to the performance of kernel-based clustering methods. Choosing the right kernel function isn't simple. Among the various clustering algorithms that have been examined in the literature, the, Gaussian kernel function has been found to be more useful.. This paper suggests a conditionally positive definite kernel function that can be used in the unsupervised clustering of incomplete data. Numerical analysis shows that the conditionally positive definite kernel function also performs well on datasets with incomplete features.

Subject Classification: *58-XX.*

Keywords: *Clustering, Incomplete data, Imputation, Non-imputation techniques. Kernel function, Positive definite & conditionally positive definite kernel function.*

* *E-mail:* **soniagoel@msit.in (Corresponding Author)**
† *E-mail:* **meenatushir@yahoo.com**

© Taru Publications

1. Introduction

There's a lot of focus in the literature on how to group data sets with missing features. Data sets with missing features are called incomplete data sets. To make them fit for clustering, some researchers use statistical methods to pre-process the data, while others use traditional algorithms to group the incomplete data. Jerez et al proposed machine learning and statistical imputation techniques to predict patients' chances of recurrence of breast cancer in a large real-world dataset [1]. Four different strategies for the clustering of incomplete data were suggested in which three strategies employ modified versions of the FCM algorithm [2]. A KNN imputation procedure which uses mutual information as a feature weighted distance metric for missing data imputation is explored in [3]. The incomplete data can be filled using various statistical methods [4-5]. A unified framework supporting multiple imputation methods was proposed in literature [6]. Three methods of missing data imputation: mean time of day (ToD), mean ToD imputation, and linear interpolation are proposed in literature. These methods are used to evaluate the continuous stability (IS) and continuous variability (IV) of the data [7]. Zhang et al. [8] proposed nearest neighbor selection for iteratively KNN imputation. Interpolation involves three different techniques: linear interpolation, quadratic interpolation, and cubic interpolation [9]. Goel & Tushir [10] explored an extensive review of different algorithms for missing data handling in fuzzy clustering. Expectation maximization clustering for blood detection in wireless capsule endoscopy [11] is explored in literature. Kernel methods in which the natural grouping of the data points within the input space is preserved, and the related structure is simplified have been integrated with fuzzy methods is explored in literature [12]. The KFCM (Gaussian kernel functions) clustering technique was first proposed by Girlomi [13] to classify unsupervised data. Hybrid c means clustering utilizing kernel function is presented in literature [14]. In this work, linear interpolation is used to impute missing attributes of the data, and conditionally positive definite kernel function [15] is further explored for the kernel-based clustering of incomplete data.

The structure of the paper is as follows: in Section 2, the positive definite kernel function, as well as the conditionally positive definite kernel function and Kernel fuzzy c means clustering (KFCM) is discussed. Section 3 shows the methodology and simulation results with the proposed strategy for different data sets that have different missing proportions. Section 4 concludes the paper.

2. Overview of Kernel Functions and KFCM

A kernel function is a function that represents the dot product of two feature vectors in the high dimension of a data space. The technique employs a mapping function, which defines a non-linear transformation from $x \to \emptyset(x)$ to a linearly separable data structure in the original data space which is illustrated in Figure 1. A kernel function $K(o_i, o_j)$ corresponds to a dot product of two feature vectors in a high-dimensional feature space for a set of vectors given as follows: $O = \{o_1, o_2 \dots o_k \dots \dots, o_N\}$ in an s-dimensional feature space R^s.

Therefore, $K(o_i, o_j) = \emptyset(o_i).\emptyset(o_j)$. The mapping function $\emptyset$: $R^2 \rightarrow H = R^3$ and $(o_{i1}, o_{i2}) \rightarrow (o_{i1}^2, o_{i2}^2, \sqrt{2}o_{i1}o_{i2})$. Then the dot product in the feature space H is calculated as

$$\emptyset(o_i).\emptyset(o_j) = (o_{i1}^2, o_{i2}^2, \sqrt{2}o_{i1}o_{i2}).(o_{j1}^2, o_{j2}^2, \sqrt{2}o_{j1}o_{j2}) = K(o_i, o_j) \quad (1)$$

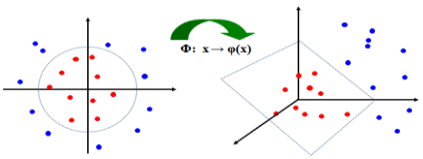


Figure 1
Non-linearly separable data mapped into the higher dimensional feature space

2.1. *Positive-definite and conditionally definite positive kernel functions*

i. A function K: $\aleph \times \aleph \rightarrow$ Ř is called a positive definite ($\aleph$ be a non-empty set) if and only if it is
   a) Symmetric, that is, $K(o, o') = K(o', o)$ for any two objects $o, o' \in \aleph$
   b) Positive-definite, i.e., $c^T Kc = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(o_i, o_j) \geq 0$ (2)
   holds for any $o_1, \ldots\ldots\ldots o_n \in \aleph$ and $c_1, \ldots\ldots\ldots c_n \in$ Ř

ii. A symmetric function K: $\aleph \times \aleph \rightarrow$ Ř for $n \geq 2$, satisfying

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(o_i, o_j) \geq 0 \text{ for all } c_i \in \text{Ř, with } \sum_{i=1}^{n} c_i = 0 \quad (3)$$

is called conditionally positive definite(CPD).

We have explored the Log kernel function which is CPD presented in Table 1.

Table 1
Positive-definite and conditionally definite positive kernel functions

Name of Kernel	Kernel Function	Attribute	Type
Gaussian	$exp\left(-\frac{\|o_i - o_j\|^2}{2\sigma^2}\right)$	$K(o, o) = 1$	Positive-definite
Log	$\log\left(\frac{1}{1 + \beta\|o_i - o_j\|^2}\right)$	$K(o, o) = 0$	Conditionally Positive definite

2.2 *Kernel Fuzzy C- Means (KFCM) and its Variants*

KFCM [15] utilizes a distance metric that's different from the standard Euclidean norm. The objective function in the KFCM model is calculated as:

$$J_{KFCM}(O: U, V) = \sum_{k=1}^{N} \sum_{i=1}^{p} (u_{ik})^m \|\emptyset(o_k) - \emptyset(v_i)\|^2 \quad (4)$$

Where $\|\emptyset(\mathrm{o_k}) - \emptyset(\mathrm{v_i})\|^2$ is the square of the distance between $\emptyset(\mathrm{o_k})$ and $\emptyset(\mathrm{v_i})$.

The distance between two points in the feature space is determined as follows:

$$\|\emptyset(o_k) - \emptyset(v_i)\|^2 = K(o_k, o_k) - 2K(o_k, v_i) + K(v_i, v_i) \quad (5)$$

$K(\mathrm{o_k}, \mathrm{o_k}) = 1$ and $K(\mathrm{v_i}, \mathrm{v_i}) = 1$ for Gaussian kernel function. Therefore,

$$\|\emptyset(o_k) - \emptyset(v_i)\|^2 = 2(1 - K(o_k, v_i)) \quad (6)$$

Objective function for G-KFCM with Gaussian kernel function is as follows:

$$J_{G-KFCM}(O: U, V) = 2 \sum_{k=1}^{N} \sum_{i=1}^{p} (u_{ik})^m \big(1 - K(o_k, v_i)\big) \quad (7)$$

$$\nabla_{\mathrm{u_{ik}}} \mathrm{J} = 0$$

Partition matrix $\mathrm{u_{ik}}$ is updated as

$$u_{ik} = \frac{\left(1 / \left(1 - K\left(o_k, v_i\right)\right)\right)^{1/(m-1)}}{\sum_{i=1}^{p} \left(1 / \left(1 - K\left(o_k, v_i\right)\right)\right)^{1/(m-1)}} \quad (8)$$

Using the Gaussian kernel function, we get

$$J_{G-KFCM}(O: U, V) = 2 \sum_{k=1}^{N} \sum_{i=1}^{p} (u_{ik})^m \left(1 - exp\left(-\frac{\|o_k - v_i\|^2}{2\sigma^2}\right)\right) \quad (9)$$

where σ is defined as the Kernel width

$$J_{G-KFCM}(O: U, V) =$$

$$2\left[\sum_{k=1}^{N} \sum_{i=1}^{p} (u_{ik})^m - \sum_{k=1}^{N} \sum_{i=1}^{p} (u_{ik})^m \left(exp\left(-\frac{\|o_k - v_i\|^2}{2\sigma^2}\right)\right)\right] \quad (10)$$

Centroids $\mathrm{v_i}$ for G-KFCM algorithm is calculated as follows:

$$v_i = \frac{\sum_{k=1}^{N} (u_{ik})^m K(o_k, v_i)\, o_k}{\sum_{k=1}^{N} (u_{ik})^m K(o_k, v_i)} \quad (11)$$

2.2.1 KFCM with Log kernel function (L-KFCM)

The objective function for L-KFCM is expressed as follows:

$$J_{L-KFCM}(O: U, V) = 2 \sum_{k=1}^{N} \sum_{i=1}^{p} (u_{ik})^m \left(1 - log\left(\frac{1}{1 + \beta\|o_k - v_i\|^2}\right)\right) \quad (12)$$

Centroids $\mathrm{v_i}$ and partition matrix u_{ik} are updated as:

$$v_i = \frac{\sum_{k=1}^{N}(u_{ik})^m\left(\frac{1}{1+\beta\|o_k - v_i\|^2}\right)o_k}{\sum_{k=1}^{N}(u_{ik})^m\left(\frac{1}{1+\beta\|o_k - v_i\|^2}\right)} \tag{13}$$

$$u_{ik} = \frac{\left(log\left(1+\beta\|o_k - v_i\|^2\right)\right)^{-\frac{1}{m-1}}}{\sum_{i=1}^{p}\left(log\left(1+\beta\|o_k - v_i\|^2\right)\right)^{-\frac{1}{m-1}}} \tag{14}$$

2.3 *Methodology for conditionally positive definite kernel function for clustering of incomplete data*

Given a set of points O which contains some missing values, we minimize $J_{FCM}(O: U, V)$ to determine U, V, and O.

The given data set is represented as $O = [O_W, O_M]$

where $O_W = \{o_W \in O | o_W$ is a complete datum $\}$ and

$O_M = \{o_M \in O | o_M$ is a incomplete datum$\}$

Partition matrix values of missing features are required for the updation of missing features.

Therefore, $u_{(ik)M}$ is separated from the partition matrix u_{ik}.

Missing features $o_{(k)M}$ in O_M for G-KFCM algorithm are updated as

$$o_{(k)M} = \frac{\sum_{k=1}^{N}(u_{(ik)M})^m K(o_{(k)M}, v_i)\, v_i}{\sum_{k=1}^{N}(u_{(ik)M})^m K(o_{(k)M}, v_i)} \tag{15}$$

Missing features $o_{(k)M}$ in O_M for L-KFCM are updated as

$$o_{(k)M} = \frac{\sum_{k=1}^{N}(u_{(ik)M})^m\left(\frac{1}{1+\beta\|o_{(k)M} - v_i\|^2}\right)v_i}{\sum_{k=1}^{N}(u_{(ik)M})^m\left(\frac{1}{1+\beta\|o_{(k)M} - v_i\|^2}\right)} \tag{16}$$

The algorithm used in our work can be summarized as follows:

Algorithm

1. A predefined number of clusters, a termination criterion, for an incomplete dataset is defined.
2. Use linear interpolation imputation to estimate all missing features.
3. FCM is used to initialize the cluster centers and the fuzzy partition matrix.
4. KFCM is used with positive definite (Gaussian) kernels and conditionally positive (Log) kernels.
5. v1 = v

6. Membership degrees for the partition matrix for positive definite (Gaussian) kernels and conditionally positive (Log) kernels are updated with Eq. 8 and 14, respectively.
7. Centroids are updated with Equation 11 and Equation 13, respectively.
8. Update all missing features for G-KFCM and L-KFCM using Eqs 15 &16.
9. Until $\|v_1 - v\| < \in$

3. Simulation Results

The effectiveness of the two kernel functions integrated into KFCM clustering that manage incomplete data is evaluated using four complete data sets (SEED, IRIS, WINE, and GLASS) and one incomplete data set [19]. In order to assess the performance of a proposed algorithm, three assessment indices are employed i.e. average prototype error, percentage accuracy, and misclassification rate. Misclassification rate is calculated by aggregating all incorrectly assigned data points (A) within each cluster, and the lower the misclassification rate, the better the outcome of the cluster.

$$Misclassification\ rate(\ Mis) = \sum_{s=1}^{N} A_s \tag{17}$$

The formula for calculating percentage accuracy (% accuracy) is as follows:

$$\%\ Accuracy = \frac{N - \sum_{S=1}^{N} A_S}{N} \times 100 \tag{18}$$

To assess the performance of the clustering algorithms, the average prototype error between the cluster centers computed using various algorithms and the actual cluster centers is found and is given by: .

$$error = \|v_{actual} - v_{calculated}\|^2 \tag{19}$$

Where, v_{actual} is the actual cluster centre and $v_{calculated}$ is the cluster centre calculated by different algorithms respectively.

For every algorithm, various percentages of missing data—10, 20, 30, 40,are introduced. All experiments are performed in MATLAB 2023a environment. The first simulation is performed on a random artificial data set as shown in Figure 2. The artificial data set is 2-dimensional, with two unequal-sized clusters with 35 and 81 data points and centres $\begin{pmatrix} 4.65 & 0.01 \\ 16.17 & -0.04 \end{pmatrix}$. Numerical analysis has been conducted to compare the performance of KFCM with Gaussian and Log kernel functions on artificial data sets. The log kernel function performs comparable to the Gaussian kernel function when applied to the artificial dataset. The results of the final centroids and the error rate recognized by their respective algorithms are tabulated in Table 2 and 3 respectively. In order to confirm the results, real data sets, including the IRIS, SEED, WINE, and GLASS datasets having varying degrees of missing rate, was used in the experiments for calculating misclassification (M), accuracy (%), and error (E).

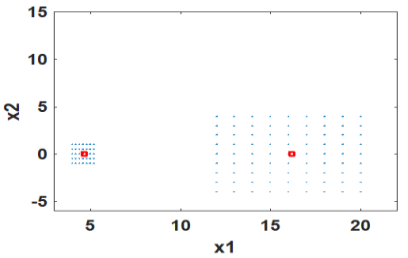


Figure 2
Artificial dataset (Cluster centres are presented by 'Ɣ')

When applied to IRIS data with varying degrees of missing rate (10-40%), Table 4 compares the performance of the Gaussian and Log kernel functions with a step size of 10.

At 10% of missing rate, Gaussian functions outperform Log functions, and at 20%, 30%, and 40%, G-KFCMs and LFCMs are almost identical. Table 5 shows that Gaussian kernel functions perform better at lower rates of missing, while L-KFCMs are the most successful when the missing rate is higher. Table 6 and 7 presents the simulation results conducted on WINE and GLASS data sets. In both datasets performance of L-KFCM is better in terms of all performance criterions.

Table 2
Centroids produced by G-KFCM and L-KFCM for artificial data

Missing Ratio Clustering Algorithm	10%		20%		30%		40%	
G-KFCM	4.59	-0.01	4.61	-0.01	4.61	-0.01	4.61	-0.02
	16.00019	-0.05	16.12	-0.07	16.07	-0.10	15.99	-0.002
L-KFCM	4.63	0.003	4.62	-0.07	4.60	-0.03	4.63	-0.08
	16.0	-0.10	16.0	0.011	16.0	-0.10	15.8	-0.04

Table 3
Simulation results for %age Error for the artificial data

Missing Ratio Clustering Algorithm	10%			20%			30%			40%		
	Cluster I	ClusterII	Average Error	ClusterI	ClusterII	Average Error	Cluster I	ClusterII	Average Error	ClusterI	ClusterII	Average Error
G-KFCM	0.16	0.03	0.09	0.14	0.05	0.10	0.14	0.09	0.12	0.15	0.12	0.13
L-KFCM	0.13	0.12	0.13	0.14	0.12	0.13	0.15	0.11	0.13	0.13	0.27	0.20

Table 4
Simulation Results (Iris dataset) for G-KFCM, and L-KFCM algorithms

Algorithm/Missing Ratio Performance Indices	G_KFCM				L_KFCM			
	10%	20%	30%	40%	10%	20%	30%	40%
Misclassification	9	12	13	14	11	13	14	14
Accuracy	94	92	91.3	90.7	92.7	91.3	90.7	90.7
Error	0.001	0.001	0.002	0.002	0.001	0.002	0.002	0.002

Table 5
Simulation Results (Seed dataset) for G-KFCM, and L-KFCM algorithms

Algorithm/Missing Ratio Performance Indices	G_KFCM				L_KFCM			
	10%	20%	30%	40%	10%	20%	30%	40%
Misclassifications	17	19	21	22	19	20	22	22
Accuracy	91.9	90.9	90	89.5	90.9	90.5	89.5	89.5
Error	0.043	0.044	0.045	0.048	0.041	0.041	0.043	0.045

Table 6
Simulation Results (Wine dataset) for G-KFCM, and L-KFCM algorithms

Algorithm/Missing Ratio Performance Indices	G_KFCM				L_KFCM			
	10%	20%	30%	40%	10%	20%	30%	40%
Misclassifications	10	12	16	19	10	12	15	16
Accuracy	94.3	93.2	91.0	89.3	94.3	93.2	91.6	91.0
Error	0.05	0.06	0.095	0.07	0.048	0.06	0.095	0.11

Table 7
Simulation Results (Glass dataset) for G-KFCM, and L-KFCM algorithms

Algorithm/Missing Ratio Performance Indices	G_KFCM				L_KFCM			
	10%	20%	30%	40%	10%	20%	30%	40%
Misclassifications	18	18	20	21	17	18	**20**	**20**
Accuracy	91.6	91.6	90.7	90.2	92.1	91.6	90.7	90.7
Error	0.16	0.28	0.37	0.56	0.018	0.15	0.156	0.16

4. Conclusions

The effects of conditionally positive definite (Log) or positive definite (Gaussian) kernel functions on KFCM clustering of incomplete data have been thoroughly analyzed in this work. According to the literature, the best kernel function for KFCM algorithms is Gaussian. We have simulated log and Gaussian kernel functions in this study and embedded them in KFCM for incomplete data. We have performed in-depth analyses on four real-world datasets with different missing rates, as well as artificial data. This leads us to the conclusion that the conditionally positive definite log kernel function performs well across a variety of datasets.

References

[1] Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2), 105-115 (2010).

[2] Hathaway, R. J., & Bezdek, J. C., Fuzzy c-means clustering of incomplete data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31(5), 735-744 (2001).

[3] García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., & Verleysen, M., K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72(7-9), 1483-1493 (2009).

[4] Little, R.J. and Rubin, D.B., Statistical Analysis with Missing Data. 793, John Wiley & Sons, Hoboken (2019).

[5] Enders,C. K.,& Baraldi, A. N., Missing data handling methods. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, 139-185 (2018).

[6] Farhangfar,A.,Kurgan, L.A.,& Pedrycz, W., A novel framework for imputation of missing values in databases. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(5), 692-709 (2007).

[7] Weed, L., Lok, R., Chawra, D.,& Zeitzer, J., The impact of missing data and imputation methods on the analysis of 24-hour activity patterns. Clocks & Sleep, 4(4), 497-507 (2022).

[8] Zhang, S., Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, 85(11), 2541-2552 (2012).

[9] Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A., Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. In Materials Science Forum, 803, 278-281 (2015, January) Trans Tech Publications Ltd.

[10] Goel, S., & Tushir, M., Different approaches for missing data handling in fuzzy clustering: a review. *Recent Advances in Electrical & Electronic Engineering*, 13(6), 833-846 (2020).

[11] Hwang, S., Oh, J., Cox, J., Tang, S. J., & Tibbals, H. F., Blood detection in wireless capsule endoscopy using expectation maximization clustering. In Medical Imaging 2006: Image Processing, 6144, p. 61441P. International Society for Optics and Photonics (2016).

[12] K. R. Muller, S. Mika, G. Ratsch and K. Tsuda, An introduction to kernel-based learning algorithms, *IEEE transactions on neural networks*, 12(2), 181-201 (2001).

[13] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE transactions on Neural Networks*,13(3), 780-784 (2002).

[14] M. Tushir and S. Srivastava, A new kernel based hybrid c-means clustering model, In 2007 IEEE International Fuzzy Systems Conference, 1-5 (2007).

[15] Nigam, J., Tushir, M., & Rai, D., A conditionally positive definite kernel functions for possibilistic clustering. *International Journal of Artificial Intelligence and Soft Computing*, 6(1), 75-91 (2017).

[16] K. Bache and M. Lichman, UCI Machine Learning Repository [http://archive. ics. uci.edu/ml]. Irvine, CA: University of California, "School of information and computer science (2013).