ON EXPONENTIALLY CONSISTENCY OF LINKAGE-BASED HIERARCHICAL CLUSTERING ALGORITHM USING KOLMOGROV-SMIRNOV DISTANCE

Tiexing Wang, Yang Liu, and Biao Chen

Department of EECS, Syracuse University, Syracuse, NY, 13244, USA
email:{twang17,yliu106,bichen}@syr.edu

ABSTRACT

This paper focuses on performance analysis of linkage-based hierarchical agglomerative clustering algorithms for sequence clustering using the Kolmogrov-Smirnov distance. Data sequences are assumed to be generated from *unknown* continuous distributions. The goal is to group the data sequences whose underlying generative distributions belong to one cluster without *a priori* knowledge of both the underlying distributions as well as the number of clusters. Upper bounds on the clustering error probability are derived. The upper bounds help establish the fact that the error probability decays exponentially fast as the sequence length goes to infinity and the obtained error exponent bound has a simple form. Tighter upper bounds on the error probability of single-linkage and complete-linkage algorithms are derived by taking advantage of the simplified metric updating for these two special cases. Simulation results are provided to validate the analysis.

***Index Terms*—** Kolmogorov-Smirnov distance, clustering, exponential consistency, probability of error, hierarchical clustering algorithm.

1. INTRODUCTION

Sequence clustering is of interest to a broad range of applications. Examples include market segmentation [1], image clustering [2, 3], and meteorological parameters characterization [4–6]. This paper considers clustering of sequences generated by *unknown* continuous distributions. Each data sequence may represent a sequence of observations or features in temporal, spatial, or other dimensions. Each class (or cluster) is defined by distributions that are close to each other using a suitably defined metric. Different clusters are assumed to be separated from each other, again, with respect to a given metric. While Euclidean distance and other vector norms have been used for sequence clustering, metrics that characterize distribution distances (e.g., the KS distance) are more relevant for the clustering at hand as sequences are generated according to some underlying distributions. We choose in the present work the Kolmogrov-Smirnov (KS) metric as probability metric; the KS distance is a true probability metric that metrizes weak convergence and has the desired concentration property [7].

This work was supported by the National Science Foundation under Grant CNS-1731237.

The above clustering problem where each cluster consists of distributions that are close to each other belongs to the general problem of unsupervised learning and has been well studied in the literature (see, e.g., [8, 9]). There are generally two classes of approaches: partitional and hierarchical. The partitional clustering algorithms include k-means clustering [10–12] and k-medoids clustering [13–15]; they usually start with some initial cluster centers, often randomly chosen, and then iteratively assign each data sequence to a cluster followed by cluster center updating. The knowledge of the number of clusters is usually required for partitional clustering algorithms. Hierarchical clustering algorithms include both hierarchical agglomerative clustering (HAC) algorithms and hierarchical divisive clustering algorithms. The HAC algorithms start with singletons and proceed to merge clusters according to pairwise distances. Hierarchical divisive clustering algorithms, on the other hand, start with one cluster consisting of all the data sequences and then proceed to split the clusters [16, 17]. The knowledge of the number of clusters is not necessary for hierarchical clustering algorithms. However, the threshold for merging and splitting may be required for hierarchical clustering algorithms.

The HAC algorithms can be further divided into two groups - linkage-based algorithms and centroid-based algorithms. Examples for linkage-based algorithms include single-linkage (SLINK) [18], complete-linkage (CLINK) [19], weighted pair group method with arithmetic mean (WPGMA), and unweighted pair group method with arithmetic mean (UPGMA) [20]. Centroid-based clustering algorithms include unweighted pair-group centroid method and weighted pair-group centroid method [21]. The above HAC algorithms can be unified using the Lance-Williams dissimilarity (LWD) update formula (see Eq. (4)), which computes the distance between clusters in a recursive manner [22]. We note that LWD can include an infinite set of HAC algorithms by changing the weights for the terms in the updating formula.

The primary focus of this paper is to study asymptotic clustering performance of the linkage-based HAC algorithms under the KS distance. We establish that a large class of linkage-based HAC algorithms are exponentially consistent, that is, the clustering error probability varnishes exponentially fast as the sequence length goes to infinity. While exponential consistency holds for k-medoids clustering algorithms under the KS distance and MMD distribution metrics [7, 23, 24],

there have been no parallel results in the literature for the HAC algorithms.

The rest of the paper is organized as follows. Section 2 introduces the clustering problem and review preliminaries on the KS distance. The general HAC algorithm is introduced in Section 3 followed by the analysis of linkage-based clustering algorithms in Section 4. Tighter upper bounds on the error probabilities of SLINK and CLINK are provided in Section 4.2. Section 5 contains simulation results.

2. SYSTEM MODEL AND PRELIMINARIES

2.1. Preliminaries of KS distance

Denote by F_p the cumulative distribution function (c.d.f.) of distribution p. The KS distance between distributions p and q is defined as

$$d_{KS}(p,q) = \sup_{a\in\mathbb{R}} |F_p(a) - F_q(a)|. \tag{1}$$

Let $\mathbf{x}$ be an independent and identically distributed (i.i.d.) sequence generated by distribution p. The empirical c.d.f. of $\mathbf{x}$ is given by

$$F_{\mathbf{x}}(a) = \frac{1}{n}\sum_{i=1}^{n} 1_{[-\infty,a]}(\mathbf{x}[i]),$$

where $1_A(\cdot)$ is the usual indicator function. The empirical KS distance between two sequences $\mathbf{x}$ and $\mathbf{y}$ is the KS distance between the corresponding empirical c.d.f., and denoted by $d_{KS}(\mathbf{x},\mathbf{y})$ for notational convenience.

2.2. Clustering Problem

Suppose there are K distribution clusters denoted by $\mathcal{P}_k$ for $k = 1,\ldots,K$, where K is fixed but unknown. Define respectively the intra-cluster distance of $\mathcal{P}_k$ and the inter-cluster distance between $\mathcal{P}_k$ and $\mathcal{P}_{k'}$ for $k \neq k'$ as

$$\begin{aligned} d_{KS}(\mathcal{P}_k) &= \sup_{p_i,p_{i'}\in\mathcal{P}_k} d_{KS}(p_i,p_{i'}), \\ d_{KS}(\mathcal{P}_k,\mathcal{P}_{k'}) &= \inf_{p_i\in\mathcal{P}_k,p_{i'}\in\mathcal{P}_{k'}} d_{KS}(p_i,p_{i'}), \end{aligned} \tag{2}$$

where $d_{KS}(\cdot,\cdot)$ is the KS distance defined in (1). Thus $d_{KS}(\mathcal{P}_k)$ and $d_{KS}(\mathcal{P}_k,\mathcal{P}_{k'})$ are respectively the diameter of $\mathcal{P}_k$ and the distance between $\mathcal{P}_k$ and $\mathcal{P}_{k'}$. Define

$$\begin{aligned} d_L &= \max_{k=1,\ldots,K} d_{KS}(\mathcal{P}_k), \\ d_H &= \min_{k\neq k'} d_{KS}(\mathcal{P}_k,\mathcal{P}_{k'}), \\ \Sigma &= d_H + d_L, \\ \Delta &= d_H - d_L. \end{aligned} \tag{3}$$

We further assume that $d_L < d_H$.

Suppose M_k data sequences are generated from the distributions in $\mathcal{P}_k$, hence a total of $\sum_{k=1}^{K} M_k = M$ sequences are to be clustered. Without loss of generality, assume that each sequence $\mathbf{x}_{k,j_k} = [\mathbf{x}_{k,j_k}[1],\ldots,\mathbf{x}_{k,j_k}[n]]$ consists of n i.i.d. samples generated from $p_{k,j_k} \in \mathcal{P}_k$ for $k = 1,\ldots,K$ and $j_k \in \{1,\ldots,M_k\}$. Note that for any k, p_{k,j_k}'s are not necessarily distinct. Thus $\mathbf{x}_{k,j_k}$'s can be generated from the same distribution for the same k. Additionally, all the data sequences are assumed to have the same length; our analysis can be easily extended to the case with different sequence lengths by replacing n with the minimum sequence length.

A clustering algorithm is said to be *consistent* if for any $0 \le d_L < d_H$,

$$\lim_{n\to\infty} P_e = 0,$$

where P_e is the probability of clustering errors and n is the sequence length. The algorithm is said to be *exponentially consistent* if for any $0 \le d_L < d_H$,

$$B = \lim_{n\to\infty} -\frac{1}{n}\log P_e > 0.$$

For the case where a clustering algorithm is exponentially consistent, we are also interested in characterizing (the bound for) the error exponent B.

2.3. Additional Notations

Denote by $\mathcal{C}_l \sim \mathcal{P}_k$ the sequences in $\mathcal{C}_l$ that are generated from $\mathcal{P}_k$. When specific reference to a cluster after the t-th iteration is needed, $\mathcal{C}_l^t$ will be used instead for sequences in the l-th cluster.

3. HAC ALGORITHMS WITH LWD UPDATE

Define the dissimilarity matrix of K clusters to be

$$\mathbf{D} = \begin{bmatrix} 0 & d(\mathcal{C}_1,\mathcal{C}_2) & \cdots & d(\mathcal{C}_1,\mathcal{C}_K) \\ d(\mathcal{C}_2,\mathcal{C}_1) & 0 & \cdots & d(\mathcal{C}_2,\mathcal{C}_K) \\ \vdots & \vdots & \vdots & \vdots \\ d(\mathcal{C}_K,\mathcal{C}_1) & d(\mathcal{C}_K,\mathcal{C}_2) & \cdots & 0 \end{bmatrix},$$

where $d(\mathcal{C}_l,\mathcal{C}_{l'})$ is the dissimilarity (i.e., the distance metric for distribution clustering) between clusters $\mathcal{C}_l$ and $\mathcal{C}_{l'}$ and satisfies 1) $d(\mathcal{C}_l,\mathcal{C}_{l'}) \ge 0$, 2) $d(\mathcal{C}_l,\mathcal{C}_l) = 0$, and 3) $d(\mathcal{C}_l,\mathcal{C}_{l'}) = d(\mathcal{C}_{l'},\mathcal{C}_l)$. In each iteration, HAC algorithms try to merge two clusters $\mathcal{C}_{l_1}$ and $\mathcal{C}_{l_2}$ if

$$d(\mathcal{C}_{l_1},\mathcal{C}_{l_2}) = \min_{l\neq l'} d(\mathcal{C}_l,\mathcal{C}_{l'}) \le d_{th},$$

with d_{th} a pre-determined threshold. The algorithm stops if

$$\min_{l\neq l'} d(\mathcal{C}_l,\mathcal{C}_{l'}) > d_{th}.$$

The general HAC algorithm is summarized in Algorithm 1. Note that any HAC algorithm converges within M steps. The LWD update formula provides a unified view for dissimilarity updating after each merge step [22]. Suppose $\mathcal{C}_{l_1}$ and $\mathcal{C}_{l_2}$ are merged. Then the LWD between $\mathcal{C}_{l_1}\cup\mathcal{C}_{l_2}$ and $\mathcal{C}_{l_3}$ is given by

$$\begin{aligned} d(\mathcal{C}_{l_1}\cup\mathcal{C}_{l_2},\mathcal{C}_{l_3}) &= \alpha_1 d(\mathcal{C}_{l_1},\mathcal{C}_{l_3}) + \alpha_2 d(\mathcal{C}_{l_2},\mathcal{C}_{l_3}) \\ &\quad + \beta d(\mathcal{C}_{l_1},\mathcal{C}_{l_2}) + \gamma\,|d(\mathcal{C}_{l_1},\mathcal{C}_{l_3}) - d(\mathcal{C}_{l_2},\mathcal{C}_{l_3})|. \end{aligned} \tag{4}$$

Algorithm 1 HAC Algorithm

Input: Data sequences $\{\mathbf{y}_i\}_{i=1}^{M}$ and threshold d_{th}.
Output: Partition set $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$.
$\mathcal{C}_i = \{\mathbf{y}_i\}$ for $i = 1, \ldots, M$, and construct the corresponding $\mathbf{D}$.
while $\min_{\mathcal{C}_l, \mathcal{C}_{l'} \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots\}} d(\mathcal{C}_l, \mathcal{C}_{l'}) \leq d_{th}$ **do**
    Merge $\mathcal{C}_{l_1}$ and $\mathcal{C}_{l_2}$ if $d(\mathcal{C}_{l_1}, \mathcal{C}_{l_2}) = \min_{\mathcal{C}_l, \mathcal{C}_{l'} \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots\}} (\mathcal{C}_l, \mathcal{C}_{l'})$,
    Update the distance matrix $\mathbf{D}$.
end while
Return $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$

Table 1: Coefficients of linkage-based HAC algorithms

SLINK	$\alpha_1 = \alpha_2 = 0.5$, $\beta = 0, \gamma = -0.5$.
CLINK	$\alpha_1 = \alpha_2 = 0.5$, $\beta = 0, \gamma = 0.5$.
UPGMA	$\alpha_1 = \frac{\lvert\mathcal{C}_{l_1}\rvert}{\lvert\mathcal{C}_{l_1}\rvert + \lvert\mathcal{C}_{l_2}\rvert}$, $\alpha_2 = 1 - \alpha_1$, $\beta = 0, \gamma = 0$.
WPGMA	$\alpha_1 = \alpha_2 = 0.5$, $\beta = 0, \gamma = 0$.

The choices of coefficients in (4) for typical linkage-based HAC algorithms are given in Table 1, where $|\mathcal{C}|$ denotes the cardinality of $\mathcal{C}$ [25]. For the rest of the paper, linkage-based clustering algorithms with LWD update are assumed to satisfy

$$\alpha_i \geq 0 \text{ for } i = 1, 2, \tag{5a}$$
$$\alpha_1 + \alpha_2 = 1, \tag{5b}$$
$$|\gamma| \leq \min\{\alpha_1, \alpha_2\}, \tag{5c}$$
$$\beta = 0. \tag{5d}$$

Thus $d(\mathcal{C}_{l_1} \cup \mathcal{C}_{l_2}, \mathcal{C}_{l_3})$ in (4) is always non-negative and

$$d(\mathcal{C}_{l_1} \cup \mathcal{C}_{l_2}, \mathcal{C}_{l_3}) \geq \min\{d(\mathcal{C}_{l_1}\mathcal{C}_{l_3}), d(\mathcal{C}_{l_2}\mathcal{C}_{l_3})\},$$
$$d(\mathcal{C}_{l_1} \cup \mathcal{C}_{l_2}, \mathcal{C}_{l_3}) \leq \max\{d(\mathcal{C}_{l_1}\mathcal{C}_{l_3}), d(\mathcal{C}_{l_2}\mathcal{C}_{l_3})\}.$$

Equation (5d) is a necessary condition for linkage-based clustering algorithms, which implies that $d(\mathcal{C}_l, \mathcal{C}_{l'})$ is only a function of $d(\mathbf{y}_i, \mathbf{y}_{i'})$, where $\mathbf{y}_i \in \mathcal{C}_l$ and $\mathbf{y}_{i'} \in \mathcal{C}_{l'}$.

4. LINKAGE-BASED ALGORITHMS

This section presents an upper bound on the error probability of linkage-based HAC algorithms generated from the LWD update formula with coefficients satisfying (5). The complete proof of the results omitted due to the space limit.

4.1. General Case

Proposition 1. *If the linkage-based HAC algorithm updates* $\mathbf{D}$ *by* (4)*, then for* $t \geq 0$ *and* $l \neq l'$*,*

$$d(\mathcal{C}_l^t, \mathcal{C}_{l'}^t) = \sum_{i:\, \mathbf{y}_i \in \mathcal{C}_l^t} \sum_{i':\, \mathbf{y}_{i'} \in \mathcal{C}_{l'}^t} \theta_{ii'}^t(\mathbf{y}_i, \mathbf{y}_{i'})\, d(\mathbf{y}_i, \mathbf{y}_{i'}). \tag{6}$$

Moreover, if the LWD update satisfies (5)*, then* $\theta_{ii'}^t(\mathbf{y}_i, \mathbf{y}_{i'}) \geq 0$ *and for any* $l \neq l'$*,*

$$\sum_{i:\, \mathbf{y}_i \in \mathcal{C}_l^t} \sum_{i':\, \mathbf{y}_{i'} \in \mathcal{C}_{l'}^t} \theta_{ii'}^t(\mathbf{y}_i, \mathbf{y}_{i'}) = 1. \tag{7}$$

Outline of the Proof. Equation (6) can be proved by induction while (7) results from (6) and (5). □

Intuitively, with (5), the updated metric in (4) can be rewritten as a convex combination of $d(\mathcal{C}_{l_1}, \mathcal{C}_{l_3})$ and $d(\mathcal{C}_{l_2}, \mathcal{C}_{l_3})$, leading to (7).

Proposition 2. *Suppose a linkage-based HAC algorithm uses update in* (4) *and the KS distance metric is used. If data sequences are generated from distributions satisfying* $d_L < d_H$*, then for* $d_{th} \in (d_L, d_H)$ *and sufficiently large* n*,*

$$P\left(d\left(\mathcal{C}_{l_1}^t, \mathcal{C}_{l_3}^t\right) \leq d_{th}\right) \leq 4M^2 e^{-nb_1}, \tag{8a}$$
$$P\left(d\left(\mathcal{C}_{l_1}^t, \mathcal{C}_{l_2}^t\right) > d_{th}\right) \leq 4M^2 e^{-nb_2}, \tag{8b}$$

where $\mathcal{C}_{l_1}^t, \mathcal{C}_{l_2}^t \sim \mathcal{P}_k$*,* $\mathcal{C}_{l_3}^t \sim \mathcal{P}_{k'}$ *for* $k \neq k'$*,* $b_1 = \frac{(d_H - d_{th})^2}{2}$ *and* $b_2 = \frac{(d_{th} - d_L)^2}{2}$.

Outline of the Proof. Equation (8) is proved by the union bound, Proposition 1 and lemmas in [7]. □

Therefore, for sequence clusters obtained after the t-th iteration by any linkage-based algorithm, any cluster pair generated from the same distribution is close to each other whereas any cluster pair generated from different distribution clusters is sufficiently separated.

Theorem 4.1. *Suppose a linkage-based clustering algorithm uses update in* (4) *and the KS distance is used. If data sequences are generated from distributions satisfying* $d_L < d_H$*, then for* $d_{th} \in (d_L, d_H)$ *and sufficiently large* n*, the error probability upon convergence is upper bounded by*

$$P_e \leq 4M^5 e^{-nb_1} + 4M^4 e^{-nb_2}.$$

Outline of the Proof. The idea of proving the upper bound on the error probability is as follows. We first show the exponential decay of the probability of the event, denoted by $\hat{E}^t$, that after the t-th iteration there exists one cluster that contains sequences generated from two distribution clusters while the clustering result after the $(t-1)$-th iteration is correct. Suppose the clustering algorithm converges after T iterations with T finite. We then show the exponential decay of the probability of the event, denote by $\hat{H}^T$, that there exist two clusters in clustering output such that all the sequences in the two cluster are generated form the same distribution clusters. The error event is the union of $\hat{E}^t$ for $t = 1, \ldots, T$ and $\hat{H}^T$. Since the algorithm converges after at most M iterations, the exponential consistency is established using the union bound. □

4.2. Tighter Error Bounds for SLINK and CLINK

Tighter upper bounds on the error probability for SLINK and CLINK can be derived by taking advantage of the fact the inter-cluster distance is computed using a single pair of sequences. The entry $d(\mathcal{C}_l, \mathcal{C}_{l'})$ in $\mathbf{D}$ for SLINK is given by

$$d_S(\mathcal{C}_l, \mathcal{C}_{l'}) = \min_{\mathbf{y}_1 \in \mathcal{C}_l, \mathbf{y}_2 \in \mathcal{C}_{l'}} d(\mathbf{y}_1, \mathbf{y}_2). \tag{9}$$

The distance between two clusters for CLINK is given by

$$d_C(\mathcal{C}_k, \mathcal{C}_{k'}) = \max_{\mathbf{y}_1 \in \mathcal{C}_k, \mathbf{y}_2 \in \mathcal{C}_{k'}} d(\mathbf{y}_1, \mathbf{y}_2). \tag{10}$$

The following theorem provides a tighter upper bound on the error probability of SLINK and CLINK.

Theorem 4.2. *Given the KS distance, the error probability of SLINK and CLINK for $d_{th} \in (d_L, d_H)$ and sufficiently large n is upper bounded by*

$$P_{e,S} \leq 4M^3 e^{-nb_1} + 4M^2 e^{-nb_2}.$$

Outline of the Proof. The idea of proving the upper bound on the error probability is the same as the proof of Theorem 4.1. The only difference is that the distance between two clusters for both SLINK and CLINK only depends on a pair of sequences from the two clusters. □

The bound on error probability in Theorem 4.2 is tighter than the general bound in Theorem 4.1 by a factor of $\frac{1}{M^2}$.

5. EXPERIMENTAL RESULTS

This section provides some experimental results for both linkage and centroid based algorithms. Set $K = 5$, $M_k = 5$ for $k = 1, \ldots, 5$, and $\mathbf{x}_{k,j_k}[i] \in \mathbb{R}$. Gaussian $\mathcal{N}(\mu_{k,j_k}, \sigma^2)$ and Gamma $\Gamma(a_{k,j_k}, b)$ distributions are used in the simulation. The probability density function (p.d.f.) of a $\Gamma(\alpha, \beta)$ is given as

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) \quad (x > 0),$$

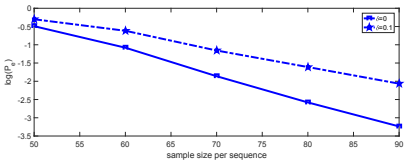
where $\alpha > 0$, $\beta > 0$ and $\Gamma(\cdot)$ is the Gamma function. For this experiment, we set $\sigma = 1$, $\beta = 1$, and

$$\mu_{k,j_k} = (k-1) + \left(j_k - \frac{M_k + 1}{2}\right)\frac{\delta}{2},$$
$$\alpha_{k,j_k} = 2.5(k-1) + \left(j_k - \frac{M_k + 1}{2}\right)\frac{\delta}{2},$$

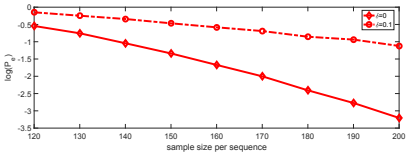
where $j_k = 1, \ldots, 5$, $\delta = 0$ and 0.1. Note that when $\delta = 0$, sequences belonging to the same distribution cluster are generated from a single distribution.

The Monte Carlo experiment for a given sample size continues until following two conditions are both satisfied:

1. the number of trials that provides incorrect clustering output reaches 1000,
2. the total number of trials reaches 5×10^4.

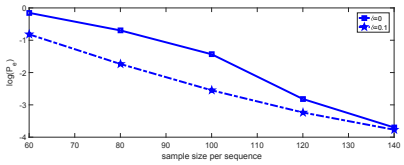


(a) SLINK

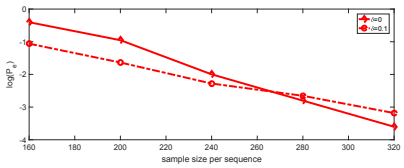


(b) CLINK

Fig. 1: Gaussian distributions under the KS distance



(a) SLINK



(b) CLINK

Fig. 2: Gamma distributions under the KS distance

The error probabilities of SLINK, and CLINK under the KS distance are given in Figs. 1 and 2. One can observe that $\log P_e$ is a linear function of the sample size, validating the exponential consistency of these algorithms. Furthermore, SLINK outperforms CLINK under the KS distance in terms of the error probability. One possible reason is that the distance between two clusters estimated by (10) tends to underestimate the number of clusters. Thus, a larger d_{th} may help improve the performance of CLINK. Moreover, the slope of $\log P_e$ with respect to n, i.e., the quantity $-\frac{\log P_e}{n}$, is non-decreasing as δ becomes smaller. In the current simulation setting, this implies a larger Δ under the KS distance.

However, with Gamma distributions, $\log P_e$ given $\delta = 0$ can be larger than $\log P_e$ given $\delta = 0.1$. A possible reason is that the KS distance between two sequences is always lower bounded by $\frac{1}{n}$, which has an amplified effect on the clustering result when all sequences in the same cluster are generated from a single distribution.

6. REFERENCES

[1] Y. Sakurai, L. Li, R. Chong, and C. Faloutsos, "Efficient distribution mining and classification," in *Proc. SIAM int. conf. data mining*, Atlanta, Georgia, USA, Apr. 2008, pp. 632–643.

[2] E. Spellman, B. C. Vemuri, and M. Rao, "Using the KL-center for efficient and accurate retrieval of distributions arising from texture images," in *Proc. IEEE Conf. Comput. Vision, Pattern Recognition (CVPR)*, San Diego, CA, USA, June 2005, vol. 1, pp. 111–116.

[3] C. Lin, C. Chen, H. Lee, and J. Liao, "Fast k-means algorithm based on a level histogram for image retrieval," *Expert Syst. Applicat.*, vol. 41, no. 7, pp. 3276 – 3283, 2014.

[4] M. Vrac, L. Billard, E. Diday, and A. Chédin, "Copula analysis of mixture models," *Computational Stat.*, vol. 27, no. 3, pp. 427–457, 2012.

[5] R. Moreno-Sáez, M. Sidrach de Cardona, and L. Mora-López, "Data mining and statistical techniques for characterizing the performance of thin-film photovoltaic modules," *Expert Syst. Applicat.*, vol. 40, no. 17, pp. 7141 – 7150, 2013.

[6] R. Moreno-Sáez and L. Mora-López, "Modelling the distribution of solar spectral irradiance using data mining techniques," *Environmental Modelling, Software*, vol. 53, pp. 163 – 172, 2014.

[7] T. Wang, Q. Li, D. Bucci, Y. Liang, B. Chen, and P. Varshney, "K-medoids clustering of data sequences with composite distributions," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2093–2106, 2019.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[9] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, Cambridge, 2012.

[10] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, pp. 129–137, Mar. 1982.

[11] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *KDD Workshop Text Mining*, 2000.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Trans. Pattern Anal., Mach. Intell.*, vol. 24, pp. 881–892, Jul. 2002.

[13] L Kaufman and P Rousseeuw, "Clustering by means of medoids," *Statistical data anal. based on the L1-norm and related methods*, pp. 405–416, 1987.

[14] M. Laan, K. Pollard, and J. Bryan, "A new partitioning around medoids algorithm," *J. Statistical Computation and Simulation*, vol. 73, no. 8, pp. 575–584, 2003.

[15] Hae-Sang Park and Chi-Hyuck Jun, "A simple and fast algorithm for k-medoids clustering," *Expert syst. applicat.*, vol. 36, no. 2, pp. 3336–3341, 2009.

[16] P. Macnaughton-Smith, W. T. Williams, M. B. Dale, and L.G. Mockett, "Dissimilarity analysis: a new technique of hierarchical sub-division," *Nature*, vol. 202, no. 4936, pp. 1034, 1964.

[17] M. Chavent, Y. Lechevallier, and O. Briant, "Divclus-t: A monothetic divisive hierarchical clustering method," *Computational Stat. & Data Anal.*, vol. 52, no. 2, pp. 687–701, 2007.

[18] J. C. Gower and G. J. S. Ross, "Minimum spanning trees and single linkage cluster analysis," *J. Royal Stat. Society. Series C (Applied Stat.)*, vol. 18, no. 1, pp. 54–64, 1969.

[19] S.C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[20] R.R. Sokal, "A statistical method for evaluating systematic relationship," *University of Kansas science bulletin*, vol. 28, pp. 1409–1438, 1958.

[21] Peter HA Sneath and Robert R Sokal, *Numerical taxonomy. The principles and practice of numerical classification.*, W. H. Freeman, San Francisco, 1973.

[22] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies: 1. hierarchical systems," *Comput. J*, vol. 9, no. 4, pp. 373–380, 1967.

[23] T. Wang, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney, "Clustering under composite generative models," in *Proc. Annu. Conf. Inform. Sci. Syst. (CISS)*, Princeton, NJ, USA, Mar. 2018, pp. 338–343.

[24] T. Wang, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney, "Exponentially consistent k-means clustering algorithm based on Kolmogrov-Smirnov test," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 2296–2300.

[25] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.