Information-Theoretic Feature Selection for Classification: Applications to Fusion of Hyperspectral and LiDAR Data

Qiang Zhang, Victor P. Pauca, Robert J. Plemmons, Robert Rand, and Todd C. Torgersen,

Abstract

In an era of massive data gathering, analysts are facing an increasingly difficult decision regarding the amount of data features to consider for the purpose of target detection and classification. While more data may offer additional information at the cost of potentially prohibitive increase in computation time, it is unclear whether classification accuracy increases accordingly. In this study, we address this question from an information-theoretic standpoint, providing upper and lower bounds of misclassification error, using Chernoff Information (CI) as an upper bound and the resistor-average distance as a lower bound. Our results are applied to the problem of classification from the fusion of real hyperspectral and LiDAR data, using numerical experiments and analysis. The results show that, in principle, adding extra data features does not decrease the CI, or equivalently the classification accuracy. However, empirically this is not the case, and our results show that in fact the indiscriminate inclusion of more data can lead to reduced classification accuracy. We offer a hill-climbing algorithm to incrementally rank and reorder features in such a way that adding an extra feature to any first-$k$ features does not lead to information loss or, equivalently, reduction in classification accuracy.

Q. Zhang is with the Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina.
Victor P. Pauca is with the Department of Computer Science, Wake Forest University, Winston-Salem, North Carolina.
Robert J. Plemmons is with the Departments of Computer Science and Mathematics, Wake Forest University, Winston-Salem, North Carolina.
Robert Rand is with the National Geospatial-Intelligence Agency.
Todd C. Torgersen is with the Department of Computer Science, Wake Forest University, Winston-Salem, North Carolina.
Index Terms

Chernoff information, feature selection, high-dimensional data classification, fusion of hyperspectral and LiDAR data, remote sensing.

I. INTRODUCTION

The science of remote sensing for the environment and other applications in its broadest sense includes general aerial images, satellite, and spacecraft observations of parts of the earth. Examples of sensing modalities include, e.g., hyperspectral imaging (HSI) and Light Detection and Ranging (LiDAR) data. HSI remote sensing technology allows one to capture images using a range of spectra from ultraviolet to visible to infrared. Hundreds of images of a scene or object are created using light from different parts of the spectrum. These hyperspectral images can be used, for example, to detect objects at a distance, to identify surface minerals, objects, and buildings from space, and space objects from the ground, leading to useful classification of data features. LiDAR is a remote sensing technology that measures distance by illuminating a target with a laser and analyzing the reflected light. Each LiDAR point has the position and color in a (x,y,z,R,G,B) format, along with range information. Its wide applications include geography, archaeology, geology, forestry, remote sensing, and atmospheric physics [1].

A tremendously rich literature exists on material identification, clustering, or spectrally unmixing HSI data, such as the match filter and related methods [2], [3], principal component analysis (PCA) [4], nonnegative matrix factorization (NMF) [5], [6], [7], independent component analysis (ICA) [8], [9], support vector machine (SVM) [10], projection pursuit [11] and hierarchical clustering models [12]. In addition, recent advances in sensor development have allowed for both HSI and LiDAR data to be captured simultaneously [13], [14], thus providing both geometrical as well as spectral information for a given scene. A typical dataset collected in this way can contain millions of LiDAR points and hundreds of gigabyte HSI images, see, e.g., [15], [16], [17], [18]. The fusion of multi-modality data, e.g., HSI and LiDAR, is an important, but challenging, aspect of remote sensing. The huge amounts of data, provided by conventional and multi-modal sensing platforms, present a significant challenge regarding developments of effective computational techniques for processing, storage, transmission, and analysis of such data.
Facing these large amounts of data, analysts often have to make a first decision on whether to use the entire dataset or a subset for a given classification algorithm. Quite often the dataset becomes so large that it is computationally prohibitive to use the whole dataset, and thus the question becomes which subset to choose. Another similar question arises that whether hundreds of HSI images plus millions of LiDAR points would guarantee a better identification or classification of objects. The curse of dimensionality is a well-known fact, that is better described in the following [19],

$$\lim_{p \to \infty} \frac{d_{\text{max}} - d_{\text{min}}}{d_{\text{min}}} = 0,$$

where $p$ is the number of dimensions, and $d_{\text{max}}$ and $d_{\text{min}}$ are respectively the maximum and the minimum distances to differentiate two objects. In other words, it becomes harder and harder to differentiate two objects when data dimension increases. Clearly, more is not necessarily better, but could it be worse?

Using the misclassification error as the criterion for evaluating whether more data would be better or worse for classification, we do not rely on any specific classification algorithms, such as PCA or SVM, but establish both an upper bound of the misclassification error, namely the Chernoff information [20], and a lower bound, namely the resistor-average distance (RAD) [21], a symmetrized Kullback-Leibler distance. This approach avoids empirically comparing different match filters or clustering algorithms, some of which often excel in one dataset but fall behind in others, while the bounds allow us evaluate which subspaces would be more appropriate in classifying objects.

In this study we demonstrate that, theoretically, adding more data will not decrease the classification accuracy, but empirically this is not necessarily true. For a real hyperspectral dataset, we have shown that incrementally adding bands from shorter to longer wavelengths can at some point reduce the Chernoff information, meaning that adding more data could instead reduce the classification accuracy, as also observed in [22]. This raises a caution on approaches arbitrarily adding extra dimensions of data, e.g., concatenating data from different modalities such as HSI and LiDAR. On the other hand, the Chernoff information may also provide us an understanding of the most information-rich dimensions or bands, e.g., a subset of wavelengths that would best differentiate one material from others.
In the literature of variable selection for high-dimensional clustering, or equivalently the band selection in hyperspectral data analysis, various importance measures have been proposed, e.g., density-based methods such as the excess mass test for multimodality, [23], [24], [25], criterion-based methods such as sparse $k$-means [26] using Between-Clusters Sum of Square (BCSS), or model-based methods such as Bayesian Information Criterion [27]. The band selection is also an important research topic in the hyperspectral imaging, where we see information-theoretic approaches such as [28], [29], and for more comprehensive comparisons, see, e.g., [30]. In the following discussions, by a feature we mean a variable either as a spectral band or a LiDAR variable for each spatial pixel.

The differences between our approach and variable-selection, also known as band-selection, methods in the hyperspectral literature lie in several aspects. First, rather than evaluating each feature separately, such as by the information entropy and the contrast measure [31], or in a small neighborhood of features, such as by the first and second spectral derivatives [32], the spectral ratio measure [33], and the correlation measure [34], the upper and lower bounds of the misclassification error are computed using a subset of features. A hill-climbing algorithm using the upper bound ranks and adds features incrementally from the most information-rich to the least, and each new feature’s information is evaluated conditional upon the set of already chosen features.

Second, the information bounds are independent of any particular classification algorithm or dataset, while guaranteeing that no classification algorithm will perform better or worse than these bounds. Our proposed method is much easier to compute and can be applied to not only hyperspectral data but also LiDAR, while some of the band-selection methods are devised specifically for hyperspectral data. For example, the spectral-derivative methods depend on the smoothness of neighboring-band variations that may not be applicable to LiDAR, which often shows abrupt changes.

Third, unlike the fixed measures such as the information entropy or the spectral derivatives, the ranks of the features can vary depending upon the underlying statistical models assumed by classifiers. This third aspect needs to raise enough attention to analysts that rather than feeding a chosen subset of data to a classifier without knowing its underlying statistical assumptions, we should be more aware of these assumptions to an extent that information-theoretic bounds can
be established to validate the empirical results which can vary significantly from one dataset to another.

The structure of the paper is as follows. In Sec. II, we introduce the definition of Chernoff information and its fast computation, specifically for the exponential family of distributions. Some empirical results are presented here to illustrate further developments. In Sec. III and IV, we discuss the feature selection method based on Chernoff information that employs all the bands while guaranteeing a nondecreasing Chernoff information sequence. In Sec. V, the method is illustrated through the problem of fusing real HSI and LiDAR datasets. We then conclude in Sec. VI with discussions.

II. CHERNOFF INFORMATION

The expected cost of misclassification \( E \), also called the Bayesian Error (Risk), is formulated as,

\[
E = \sum_{i} p_i \sum_{k \neq i} P(k|i),
\]

where \( p_i \) is the prior probability of class (object/target) \( i \), \( P(k|i) \) is the probability of classifying class \( i \) as \( k \). \( E \) is minimized if an observed vector \( x \) in population \( k \) satisfies

\[
p_k f_k(x) > p_i f_i(x), i \neq k,
\]

where \( f_i(x) \) is the distribution of class \( i \). If we assume a uniform prior distribution, then the classification rule can be stated as: find \( k \), such that

\[
f_k(x) > f_i(x), i \neq k.
\]

Consider a univariate two-class case in which each class has a normal distribution, as shown in Fig. 1, where \( f_i, i = 1, 2 \), is a normal distribution with mean \( \mu_i \), and the decision region is \((-\infty, \frac{\mu_1 + \mu_2}{2})\) for class 1 and \((\frac{\mu_1 + \mu_2}{2}, +\infty)\) for class 2. Moving the cut point to the left or right of \((\mu_1 + \mu_2)/2\) increases \( E \), shown as the shaded area in Fig. 1.

To establish an upper bound on the classification error, we use the two-class case as an example, as

\[
E = \int_x \min(p_1 f_1, p_2 f_2) dx.
\]
Using the following relationship,
\[ \min(a, b) \leq a^\alpha b^{1-\alpha}, \forall \alpha \in [0, 1], \]  
we obtain a first bound on \( E \) as
\[ E \leq p_1^\alpha p_2^{1-\alpha} \int_x f_1^\alpha(x)f_2^{1-\alpha}(x)dx. \]  
Define Chernoff \( \alpha \)-divergence as,
\[ C_\alpha(f_1 : f_2) = -\log \int_x f_1^\alpha(x)f_2^{1-\alpha}(x)dx, \]
and Chernoff information as,
\[ C^*(f_1 : f_2) = \max_{\alpha \in [0,1]} C_\alpha(f_1 : f_2). \]
Then we have
\[ E \leq p_1^\alpha p_2^{1-\alpha} \exp \left[-C^*(f_1 : f_2)\right], \]
and for the uniform prior distribution case, we simply have
\[ E \leq \exp \left[-C^*(f_1 : f_2)\right]. \]
Clearly, higher \( C^* \) means lower misclassification error and hence will be more desirable. From the definitions of Chernoff \( \alpha \)-divergence and Chernoff information, we know they are both nonnegative and asymmetric, i.e., \( C_\alpha(f_1 : f_2) \neq C_\alpha(f_2 : f_1) \), but \( C_\alpha(f_2 : f_1) = C_{1-\alpha}(f_1 : f_2) \).
Applications of Chernoff information include, e.g., sensor networks [35], image segmentation and registration [36], [37], face recognition [38] and edge segmentation [39]. Next, we look at some basic properties of Chernoff information.

In multivariate distributions, if we have one variable independent from the rest, we know Chernoff $\alpha$-divergence or Chernoff information of all variables would be decomposable. This is stated in the following proposition.

**Proposition 1.** Let $f_1$ and $f_2$ be two joint distributions of a $d$-variate random vector $\mathbf{x} = (x_1, \ldots, x_d)$ and a random variable $y$. If $y$ is independent from $\mathbf{x}$, then

$$C_\alpha(f_1, f_2) = C_\alpha(f_{1x}, f_{2x}) + C_\alpha(f_{1y}, f_{2y}),$$

where $f_{1x}$ and $f_{2x}$ are the marginal distributions of $\mathbf{x}$, and $f_{1y}$ and $f_{2y}$ are the marginal distributions of $y$.

The proof is simply from the decomposition of the distribution functions, $f_1$ and $f_2$, due to the independence assumption, and hence ignored here. Proposition 1 tells us that adding independent variables for classification would never reduce Chernoff information, or to increase the upper bound of misclassification error. If variable $y$ is not independent from $\mathbf{x}$, the Chernoff $\alpha$-divergence between $f_1$ and $f_2$ would still be no less than that between $f_{1x}$ and $f_{2x}$.

**Proposition 2.** If $y$ is not independent, then

$$C_\alpha(f_1, f_2) \geq C_\alpha(f_{1x}, f_{2x}).$$

**Proof:** By the conditional distribution, $f_i(\mathbf{x}, y) = f_i(y|\mathbf{x})f_{ix}(\mathbf{x})$, $i = 1, 2$, we have

$$\int_{\Omega_x} \int_{\Omega_y} f_1^\alpha(\mathbf{x}, y) f_2^{1-\alpha}(\mathbf{x}, y) d\mathbf{x} dy$$

$$= \int_{\Omega_x} \int_{\Omega_y} f_{1x}^\alpha(\mathbf{x}) f_1^\alpha(y|x)$$

$$f_{2x}^{1-\alpha}(\mathbf{x}) f_2^{1-\alpha}(y|x) dy d\mathbf{x}.$$  (14)

Denote $g_\alpha(\mathbf{x}) = \int_{\Omega_y} f_1^\alpha(y|x) f_2^{1-\alpha}(y|x) dy$, and notice that $\forall \mathbf{x}, \forall \alpha, 0 < g_\alpha(\mathbf{x}) < 1$. Hence, the
integral above becomes
\[
\int_{\Omega} f_{1x}^\alpha(x) f_{2x}^{1-\alpha}(x) g_\alpha(x) dx \leq \int_{\Omega} f_{1x}^\alpha(x) f_{2x}^{1-\alpha}(x) dx.
\] (15)

Hence, \( C_\alpha(f_1, f_2) \geq C_\alpha(f_{1x}, f_{2x}) \) by definition.

Propositions 1 and 2 confirm that adding a variable, regardless of being independent or not, should not reduce Chernoff information, or increase the misclassification error.

There is generally no closed form of Chernoff information, except special cases such as the 1-parameter exponential family. For example, the Chernoff information between two univariate-Gaussian distributions with a common standard deviation is
\[
C^*(f_1 : f_2) = \frac{1}{8\sigma^2} (\mu_1 - \mu_2)^2.
\] (16)
However, if we assume the class distributions belong to the exponential family, the relationships between Chernoff \( \alpha \)-divergence and Jensen \( \alpha \)-divergence and Bregman divergence will lead to a simple algorithm for computing CI [20]. The exponential family of distributions can be written as the following,
\[
p(x) = h(x)e^{\langle \theta, f(x) \rangle - F(\theta)},
\] (17)
where \( \theta \) is the natural parameter vector, \( f(x) \) is the sufficient statistics vector, \( F(\theta) \) is the cumulative generating function, and \( h \) is simply a normalizing function. For example, the multivariate normal distribution, \( \mathcal{N}(\mu, \Sigma) \), has \( \theta = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}) \), \( f(x) = (x, xx^T) \), and \( F(\theta) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2} \log \det(2\pi\Sigma) \). Jensen \( \alpha \)-divergence on the natural parameters is defined as
\[
J_\alpha(\theta_1, \theta_2) = \alpha F(\theta_1) + (1 - \alpha) F(\theta_2)
+ F(\alpha\theta_1 + (1 - \alpha)\theta_2),
\] (18)
and Bregman divergence on the natural parameters is defined as
\[
B(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - [\theta_1 - \theta_2]F'(\theta_1),
\] (19)
where \( F \) is the cumulative generating function in the definition of the exponential family.

In the exponential family of distributions, we have the following relationships,
• Chernoff $\alpha$-divergence on members of the same exponential family is equivalent to a Jensen $\alpha$-divergence on the corresponding natural parameters;
• the maximal Jensen $\alpha$-divergence can be computed as an equivalent Bregman divergence;
• Bregman divergence is also the same as the Kullback-Leibler divergence. The former is a distance acting on natural parameters, while the latter is a distance on canonical parameters.

These relationships help us avoid the integration and optimization steps in the computation of Chernoff information, but to use a simple geodesic-bisection algorithm to find the optimal $\alpha^*$. The algorithm is summarized in pseudo-codes in Algorithm 1.

**Input**: natural parameters of two distributions belonging to the same exponential family, $\theta_1$ and $\theta_2$.

**Output**: $\alpha^*$ and $\theta^* = (1 - \alpha^*)\theta_1 + \alpha^*\theta_2$.

1. Initialize $\alpha_m = 0, \alpha_M = 1$.
2. while $\alpha_M - \alpha_m > tol$ do
3. Compute midpoint, $\alpha' = (\alpha_m + \alpha_M)/2$.
4. if $B(\theta_1 : \theta) < B(\theta_2 : \theta)$ then
5. $\alpha_m = \alpha'$;
6. else
7. $\alpha_M = \alpha'$;
8. end
9. end

**Algorithm 1**: A geodesic-bisection algorithm for computing the Chernoff information.

For a size-$n$ sample, Chernoff information asymptotically upper-bounds the misclassification error [40], i.e.,

$$\lim_{n \to \infty} E = e^{-nC(f_1, f_2)}. \quad (20)$$

The error also has an arbitrarily tight lower bound by an information-theoretic distance [41],

$$\lim_{n \to \infty} E = e^{-nC(f_1, f_2)} \geq e^{-nR(f_1, f_2)}, \quad (21)$$

where $R(f_1, f_2)$ is the resistor-average distance (RAD) [21] defined as

$$R(f_1, f_2) = (KL(f_1, f_2)^{-1} + KL(f_2, f_1)^{-1})^{-1}, \quad (22)$$
and $KL(f_1, f_2)$ is the Kullback-Liebler (KL) distance from $f_1$ to $f_2$. As stated before, the exponential-family distribution allows computing KL distance with Bregman divergence on the natural parameters, from which we can derive RADs between pairs of distributions, and hence, a lower bound of the misclassification error.

III. CHERNOFF INFORMATION AND HIGH-DIMENSIONAL CLUSTERING

The multivariate normal (MVN) distribution is the most popular distribution assumed in clustering HSI and LiDAR data, e.g., the matched filter [2], [3], the orthogonal subspace projection [42], the adaptive match filter (AMF) [43], and the adaptive cosine estimator (ACE) [44]. Hence, in this section we focus on this particular member of the exponential family, even though Algorithm 1 can be applied to any member in this family.

Let $f_1 \sim N(\mu_1, \Sigma_1)$ and $f_2 \sim N(\mu_2, \Sigma_2)$ represent two $d$-variate normal distributions. Then the Chernoff $\alpha$-divergence between these two distributions is,

$$C_\alpha(f_1, f_2) = \frac{1}{2} \log \frac{\alpha \Sigma_1 + (1 - \alpha) \Sigma_2}{|\Sigma_1|^\alpha |\Sigma_2|^{1-\alpha}} + \frac{\alpha(1-\alpha)}{2} (\mu_1 - \mu_2)^T [\alpha \Sigma_1 + (1 - \alpha) \Sigma_2] (\mu_1 - \mu_2).$$

(23)

Now we add an extra dimension to both $p$ and $q$, and denote $\tilde{p} \sim N(\tilde{\mu}_1, \tilde{\Sigma}_1)$ and $\tilde{q} \sim N(\tilde{\mu}_2, \tilde{\Sigma}_2)$ as two $d+1$-dimensional normal distributions, where $\tilde{\mu}_1 = (\mu_1; m_1)$, $\tilde{\mu}_2 = (\mu_2; m_2)$, and

$$\tilde{\Sigma}_1 = \begin{pmatrix} \Sigma_1 & s_1 \\ s_1^T & \sigma_1^2 \end{pmatrix}, \tilde{\Sigma}_2 = \begin{pmatrix} \Sigma_2 & s_2 \\ s_2^T & \sigma_2^2 \end{pmatrix}.$$

(24)

Proposition 2 tells us that $C_\alpha(\tilde{p}, \tilde{q}) \geq C_\alpha(p, q)$, and hence $C^*(\tilde{p}, \tilde{q}) \geq C^*(p, q)$. To validate this empirically, we select a small HSI dataset of size $60 \times 100 \times 58$, with the first two numbers being sizes of two spatial dimensions and the last number being the size of the spectral dimension. For the details on the dataset, please see Sec. V. We transform the dataset into a matrix of size $6,000 \times 58$, and incrementally increase the number of bands, $d$, from 1 to 58 following the wavelength order. At each $d$, we feed the data to the $k$-means clustering algorithm, and then use the resulting centroids and pixel memberships to calculate the cluster-conditional means and covariance matrices, from which we can compute CIs between pairs of cluster-conditional...
distributions. We then take the minimum CI between all pairs which represents the largest possible misclassification error and plot it against the number of bands used in clustering, shown in Fig. 2. Here we set the number of clusters, $k$, as 5. From Fig. 2, we observe a general increasing trend, but also ups-and-downs in the middle, which clearly violates Proposition 2.

Investigating this problem, we find that the empirical covariances of subspaces are changing when a new band is added, i.e., $\tilde{\Sigma}_{1:d,1:d} \neq \Sigma$, or the assumption in Eq. (24) is violated. Here, $\tilde{\Sigma}_{1:d,1:d}$ indicates the first $d$-columns and first $d$-rows of $\tilde{\Sigma}$. The only possible reason for this is that the cluster memberships of some pixels switched after a new band is added. Figure 3 shows the number of pixels that switched their class membership after a new band is added. In the worst case, we could have close to 30% pixels changing memberships, as shown by the peak. Clearly, though correct classification could be done in lower dimensions, adding more data can potentially alter these classifications leading to wrong conclusions, and the membership-switching problem presents a challenge in clustering high dimensional data that is worth further investigation.

As claimed by Proposition 1, adding an independent variable would not decrease Chernoff information, and because PCA would provide us with the space spanned by orthogonal principal components, namely the principal component (PC) space, we can test the trend of CI after incrementally adding PCs. Another concern in high dimensional data clustering is that the data
variances at different bands can also alter clustering results, see, e.g., [45], [46]. For example, adding a large-variance variable could easily alter sample memberships as illustrated in Fig. 4, where we see point A in the one-dimensional case, on the horizontal axis, would be classified as the red class, but in the two-dimensional case, it would be classified as the blue class. The two centroids, $C_1$ and $C_2$, are more separated on the vertical axis, indicating a larger overall variance on this second dimension. For this reason, it might be advisable to order features by their variances, which is also tested here. Similarly, one could argue that normalizing each band by its variance could help avoid the membership switching. All three cases are tested and shown in Figure 5. The PCA approach clearly enjoys a monotonically increasing trend after incrementally adding PCs, while ordering by variance also results in a monotonic increasing trend, although it rises slower than the CI sequence by PCA in the beginning, but we observe a sharp jump at the 35th ordered variable. Disappointingly, normalizing bands does not help in removing the ups-and-downs, as seen on the black CI curve.

IV. FEATURE SELECTION WITH CHERNOFF INFORMATION

Though PCA enjoys the clear advantage given its orthogonality property in the PC space, many times researchers would like to classify in the original feature space, which would be more physically or chemically meaningful. For example, it is hard to physically interpret PC
Fig. 4. An example of membership switching after adding an extra dimension.

Fig. 5. Comparison of the minimum CIs with different orders of data.
vectors, which must contain negative entries, in the HSI or LiDAR data. In the original feature space, besides ranking by variances, we can also rank them by various importance measures, as suggested in the introduction section, e.g., the excess-mass measure \cite{23}, \cite{24}, \cite{25} or the BCSS used in $k$-means. For example, given a HSI dataset, rather than following the wavelength order, we can rank and choose the most important bands before applying clustering algorithms. Here, we will introduce a hill-climbing scheme that would incrementally search and add the most information-rich set of features, while guaranteeing no loss of information in terms of classification. The end result would be a set of chosen features, which can then be reordered by, for example, the wavelength order for algorithms such as spectral unmixing.

Starting from an empty set $\hat{X}$, at each iteration the algorithm searches for a variable, $X_i$, that provides us the largest minimum CI between all pairs of clusters, i.e.,

$$\hat{i} = \max_i \min_{j,k,j \neq k} CI(\hat{f}_i, \hat{f}_k),$$

(25)

where $\hat{f}_k$ is the empirical distribution of the $k^{th}$ cluster estimated from the data, $\hat{X}$ concatenated with $X_i$. The algorithm then adds $X_i$ to $\hat{X}$. Denote the maximum CI at the $j^{th}$ iteration as $c_j$ and the sample membership set as $M_j$. To guarantee a nondecreasing sequence of CI, if $c_j < c_{j-1}$, we use Proposition 2 by fixing the membership as the last iteration, i.e., $M_j = M_{j-1}$, and recompute CIs between all pairs before finding $\hat{i}$. The pseudo-codes are included in Algorithm 2.

Compared to the excess-mass measure or the BCSS, Chernoff information is more flexible with model assumptions, because it allows a wide variety of exponential-family distributions, while the excess-mass measure usually requires continuous data and the BCSS assumes the homoscedasticity, or the equal variance. It also helps on choosing the number of clusters, since the minimum CI between all pairs of clusters would dwindle with an increasing number of clusters. The fixing-membership idea also ensures no loss of information or features, especially when noisy or highly correlated data is added that could potentially alter the correct classification.

V. Experiments

The HSI and LiDAR datasets were simultaneously acquired by sensors on the same aircraft over Gulfport, Mississippi \cite{9} in 2010. The ground sampling distance is 1 meter. The original
**Input**: d-dimensional data, $X$.

**Output**: Reordered d-dimensional data, $\hat{X}$, and a CI sequence.

1. Initialize an index set as $S = \{1, 2, \ldots, d\}$.
2. $\hat{X} = \emptyset$, $j = 1$.
3. while $S \neq \emptyset$ do
   4. for $i \in S$ do
      5. $X_i = concatenate(\hat{X}, X_i)$.
      6. Cluster $X_i$ using, e.g., k-means, and assign membership to $M_j$.
      7. Estimate $\Sigma_k$ and $\mu_k$.
      8. Compute $CI_{ijk}$ between all pairs of clusters.
   9. ($\hat{i}, \hat{j}, \hat{k}$) = max $i$ min $j,k,j \neq k$ $CI_{ijk}$.
10. $c_j = CI_{ijk}$.
11. if $c_j < c_{j-1}$ then
12.    $M_j = M_{j-1}$.
13.    Repeat steps 4 to 11.
14. end
15. $\hat{X} = concatenate(\hat{X}, X_{\hat{i}})$.
16. $S = S \setminus \hat{i}$.
17. $j = j + 1$.
18. end

**Algorithm 2**: A hill-climbing algorithm for reordering features.

HSI dataset has 72 bands ranging from $0.4\mu m$ to $1.0\mu m$. After removing noisy bands, we use 58 bands on a $60 \times 100$ scene, and the coregistered LiDAR point cloud is preprocessed as the implicit geometry (IG) functions, including the population function, the validity function and the distance function, each containing 12 features. For details of IG functions, see [47]. The scene, as shown in Fig. 6, contains parking lots, buildings, vehicles, grass and trees.

The fusion of HSI and LiDAR data considered next is on the feature level, meaning we simply
Fig. 6. The top figure shows a false-color image of an HSI image at 0.72\(\mu\)m. The middle figure shows the height profile of the scene from LiDAR. The bottom figure shows a 3D aerial view of the scene by Google Map (Imagery@2014 Google, Map data@2014 Google).
concatenate two datasets together, with a total of 70 features. The purposes of the following experiments include: first, whether concatenating HSI data with LiDAR would increase the classification accuracy; second, compare the hill-climbing approach with other three ordering schemes, namely the natural wavelength order, the PC order and the variance order; third, among the three IG functions which would perform the best.

The experiments are set up in such a way that for each combination of HSI and LiDAR IG, i.e., HSI only, HSI + Validity, HSI + Distance, and HSI + Population, we will compute the CI sequences of four different orders: the wavelength order, the PC order, the variance order and the hill-climbing order. We use the $k$-means clustering algorithm and set the number of clusters at 5. We derive the empirical cluster-conditional means and covariances from the pixel memberships given by $k$-means, before transforming them into natural parameters and computing the statistical distances.

Each subplot in Fig. 7 shows four curves of CIs which correspond to four different combinations of HSI and LiDAR features. From the up-and-downs in all four HSI-only curves, we can see that adding features following the wavelength order could reduce CIs. Adding LiDAR features does increase CIs, or reduce the misclassification errors. The PC order mostly enjoys an increasing sequences of CIs as more PCs are added, though we observe occasionally small fluctuations, possibly due to the sensitivity of $k$-means algorithm to initial conditions. Ordering features by their variances would not guarantee nondecreasing sequences of CIs, as we observe some large drops in the green curves. At last, the hill-climbing approach not only enjoys nondecreasing sequences of CIs, it also has the largest CIs in the end when all features are included. The largest increase in CI is observed by HSI and the Distance function.

With the same experimental setup, we also computed the RADs, shown in Fig. 8, the exponential function of which provides us lower bounds of misclassification error. As expected, they have greater values than CIs, and mostly follow the same trends, though in the hill-climbing approach they could have occasional drops. Note that the hill-climbing algorithm can also be based on the RADs, rather than CIs, which would provide a nondecreasing sequence of RADs. However, we cannot guarantee nondecreasing sequences for both RAD and CI. For misclassification errors, we transform the CIs and RADs into the upper and lower bounds of $E$, by $e^{-R(f_1, f_2)} \geq E \leq e^{-C(f_1, f_2)}$, and plot them in Fig. 9. More directly, we can see the combination of HSI with the Distance
function has the quickest drop in error when more features are added, and for a given tolerance level this combination would need the smallest number of features. For example, with only 10 features, HSI+Distance would guarantee a misclassification error below 0.01.

VI. CONCLUSIONS AND DISCUSSIONS

In this study we have introduced both upper and lower bounds of misclassification errors by information-theoretic statistical measures, that would help us identify and collect features for the best classification accuracy, without reliance on any particular classification algorithm. Recent studies of Chernoff information allow us to compute it quickly with a geodesic-bisection algorithm on the exponential family of distributions, which is often widely assumed in analyzing hyperspectral and LiDAR data. We have shown that, theoretically, adding more features should not reduce Chernoff information, or the classification accuracy, but empirically on sample data,
Fig. 8. The RADs of increasing numbers of features in four different orders.

adding more features does not guarantee Chernoff information will not be lost. Facing this problem, we have devised a hill-climbing algorithm to reorder the features based on their contribution to Chernoff information, and hence to guarantee a nondecreasing sequence of CI, as more features are added. The proposed approach does not remove any features, contrary to band-selection methods, while it is also flexible in the use of underlying distributions of clustering models.

We have learned several things from this study. First, there are not necessarily “bad bands” or “good bands” in all hyperspectral images, which often reminds us of data quality measures such as the signal-to-noise ratio (SNR). For example, in Fig. 4, data in both the x-axis and the y-axis, appear to show a fairly good two-cluster pattern, but which can still cause the membership switching of certain points. Only in the more extreme cases, when the SNRs of certain bands
are so low or the data are so noisy that they are almost random, we should consider calling these “bad bands” and remove them.

Second, because of the dependence of the CI on the assumption of statistical models, the determination of information-rich bands is model-dependent, meaning that one “good” band for a certain model might be “bad” for another model. Hence, analysts might be encouraged to have a good understanding of underlying statistical-model assumptions before applying certain band-selection methods.

Third, as the model complexity increases, e.g., when the number of clusters in $k$-means increases, the membership switching can be more severe and hence we might expect more ups- and-downs in the CIs after incrementally adding features, see, e.g., a comparison of CIs of the 3-cluster model with the CIs of the 5-cluster model in Fig. 10. In these cases, we need to be
more careful in choosing the features to use. Also, the membership switching is not necessarily only happening to the more rudimentary clustering algorithms, such as $k$-means, because as clusters are less separated and as the number of clusters increases, we can still observe similar behaviors by more sophisticated algorithms.

Fourth, it might not be a good practice to incrementally add spectral bands following the natural wavelength order as shown in Fig. 2. In our experiments, the wavelength order appears to show poor CI sequences. We suggest that analysts consider using feature-selection techniques, the CI approach being one of them, to rank and reorder the bands before applying their data analysis methods. However, for physical interpretation of clustered data, the information-rich bands selected by the hill-climbing algorithm should be reordered back in the wavelength order. For example, the indices of the top-five ranked spectral bands by the hill-climbing algorithm could be, \{30, 27, 3, 50, 10\}, and analysts should reorder them as, \{3, 10, 27, 30, 50\}, before applying clustering algorithms for better interpretation of clustering results.

**ACKNOWLEDGMENT**

This work is sponsored by National Geospatial-Intelligence Agency (NGA) contract HM1582-10-C-0011 and by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-
REFERENCES


