ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse72852019.pdf
https://doi.org/10.30534/ijatcse/2019/72852019

Band Selection by Divergence Distance Based on Gaussian Mixture Model for Hyperspectral Image Classification

Mohammed LAHLIMI[1], Mounir AIT KERROUM[2], Youssef FAKHRI[3]
[1, 2 & 3]Laboratory of Research in Computer Science and Telecommunications,
Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco
[1] lahlimi.mohammed@gmail.com
[2] maitkerroum@gmail.com
[3] yousseffakhri@yahoo.fr

ABSTRACT

In this work, we investigate a new band selection approach by Divergence distance based on the Gaussian Mixture Model (GMM) for Hyperspectral image classification. The main motivation in modeling the Divergence distance with GMM is due to the fact that GMM is well known to be less sensitive to estimation error problem than non-parametric models and can capture non Gaussian statistics of multivariate data. To estimate the parameters of GMM, the Expectation Maximization (EM) with the Bayesian Information Criterion (BIC) and a Robust Expectation Maximization (REM) algorithm are used. This investigation is inspired by our previous work on the Bhattacharyya distance hence we are particularly interested in using the Divergence distance to find out which one gives better results. The performances of the proposed approach are compared to those of the Bhattacharyya distance in terms of global classification accuracy and numbers of retained bands through two Classifiers; Extreme Learning Machine (ELM) and Support Vector Machine (SVM). The experiments are carried out on three hyperspectral images, the Indiana Pines (92AV3C), the Botswana and the Kennedy Space Center dataset (KSC).

Key words : Band Selection, BIC, Bhattacharyya distance, Divergence distance, Hyperspectral Imaging, GMM, REM.

1. INTRODUCTION

The Hughes phenomenon [1] is one of the main challenges in remote sensing [2]. With the increase of data dimensionality and due to small sample size problem (SSSP) [2], a good estimate of the class parameters can't be found, as result the classifier will not be properly trained [3]. Hence, reducing the data dimensionality before the classification process is essential.

Dimensionality reduction can be accomplished in two different approaches, **band selection** [3] [4] [5] and **band extraction** [6] [7] [8]. Band extraction consists on finding a linear/nonlinear transformation to a lower dimensional feature space [6]. In remote sensing, band extraction tries to separate classes based on their spectral characteristics [9]. The Principal Component Analysis (PCA), Segmented Principal Component Analysis (SPCA), Independent Component analysis (ICA), Orthogonal Subspace Projection (OSP) and others [10] [11] [12] have been used to reduce the data volume. Because of the transformation, the original data are replaces by new set of variables with no actual physical meaning [6], which can be a disadvantage in some cases. In the other hand, bands selection attempts to identify a subset from the original pool by selecting the bands that contribute to the classification task by means of maximizing a class separability criterion [6]. Between this two dimensionality reduction methods, band selection is the preferred one in this study, as the physical meaning of the data remains unchanged [13]. Its main goal is to identify and choose only those bands that improve the classification task based on the chosen criterion [6]. Existing Approaches for band selection can be classified in two groups: the wrapper approach [6] which consist on using the error of the classifier itself as criterion for the band selection, it produce a subset with the high classification score, but the drawback of this technique is that the results is biased toward the classifier[6]. Unlike the wrapper approach, the filter approach [6] deploy metrics and distances to evaluate the bands without involving the classifier.

In [14] [15], the authors used the Divergence distance criterion to evaluate the bands effectiveness. We noticed in those studies, the required probability estimation to model this distance is often done under the assumption of the Normal distribution. However, in remote sensing, many factors [16] can affect the spectral response of an hyperspectral image. As consequence, using the single normal distribution assumption to describe the data is not flexible enough to capture the complex data structures of the real world [17]. GMM, in the other hand, is known to be less sensitive to estimation error problem than non-parametric models [18] and captures non Gaussian statistic of multivariate data [16] through modeling the data with more than one weighted Gaussian component.

This study investigate a new band selection method using the divergence distance based on GMM. The main challenge in GMM is the estimation of its parameters. In literature, the Expectation-Maximization (EM) algorithm [19] is often used,

however the EM algorithm for GMM is quite sensitive to the initial values [20] and the number of its components K is user defined. A good choice of the parameter K is important as it can directly affect the estimation of the covariance matrix. As when the ratio of the number of training samples to the number of bands is small, we can easily end up with the "Hughes phenomenon" and the classification results may not be satisfactory [2]. To overcome this shortcoming, two approach are proposed; a Robust Expectation Maximization (REM) algorithm as defined in [20] since it can automatically obtain an optimal number of clusters K and a GMM based on the Bayes Information Criterion (BIC).

Our main contributions in this study is a new technique for hyperspectral band selection based on the Divergence distance using GMM-REM and GMM-BIC. To assess the efficiency of the proposed approach, experiments were carried out on three hyperspectral Images: The Indiana Pines (92AV3C) scene firstly used by David Landgrebe and his students [21] [22] [23]. The Initial experiment were done on a four class subset [23] [24] of the Indiana Pines scene to lower the computation time and to have enough samples for a good probability estimation. The other hyperspectral scene are the Botswana dataset and the Kennedy space center KSC also used in several of studies such as [25] [26]. The selected bands with the proposed approach are compared in terms of number of retained band and in terms of classification accuracy.

The remaining of this paper is organized as follows: section 2 and 3 present the proposed band selection algorithm; the experimental results and comments are presented in section 4 and finally the conclusion in section 5.

2. BACKGROUND

2.1 Divergence Distance

Given two classes ω_1, ω_2 and a band vector x, between two distributions, $p(x|\omega_1)$ and $p(x|\omega_2)$, the Divergence-distance is the sum of the two Kullback Leibler divergences [27] and it is interpreted as the amount of information necessary to change the prior probability distribution into posterior probability distribution [28]. The divergence-distance is a similarity measurement used in information theory defined as [6]:

$$J_d(\omega_1,\omega_2) = \int [p(x|\omega_1) - p(x|\omega_2)] \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \quad (1)$$

The highest the value of J_d, the most dissimilar the band pair are. For a multi-class problem, it can be computed as the average divergence between each pairwise classes (ω_i, ω_j) :

$$J = \sum_i \sum_j P(\omega_i)P(\omega_j)J_d(\omega_i,\omega_j) \quad (2)$$

In previous works [14] [15] the authors estimated the equation (1) under the assumption of a normal distributions with means μ_1, μ_2 and covariance matrices Σ_1, Σ_2. Hence, the equation (1) can be simplified to:

$$J_d = \frac{1}{2}(\mu_i - \mu_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) + \frac{1}{2} tr(\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I) \quad (3)$$

In remote sensing, the spectral response of hyperspectral image can be affected by many factors [16]. As result, the above-simplified equation (3) is not flexible enough to capture the complex data structures met in real world settings [17].

2.2 Gaussian Mixture Model

The Gaussian Mixture Model (GMM) captures non-Gaussian statistic of multivariate data [16]. GMM models the density as the sum of one or more weighted Gaussian components [22], and usually less sensitive to estimation error problem than purely non-parametric models [18]. For a GMM, a probability density function is written as the sum of K gaussian components:

$$p(x|\omega) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k) \quad (4)$$

where K the number of mixture component, π_k the mixing weight $(0 \le \pi_k \le 1$ and $\sum_{k=1}^{K} \pi_k = 1)$ and $p(x|\mu_k, \Sigma_k)$ a d-dimensional gaussian distribution

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} exp[-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)] \quad (5)$$

μ_k and Σ_k, are respectively the mean and the covariance matrix of the k^{th} component. While the parameters $\{\pi c, \mu_c, \Sigma_c\}$ are estimated by the EM algorithm [19].

3. BAND SELECTION BASED ON DIVERGENCE DISTANCE

3.1 Problem formulation

Let $F = \{B_i\}_{i=1}^{d}$ be a set of d-dimensional band space. The goal is to find an optimal subset $S = \{B'_i\}_{i=1}^{d'}, S \subset F, d' \le d$ that keeps the maximum amount of discriminant information as possible while discarding any redundancy or irrelevant bands according to a cost-function (Divergence, Bhattacharyya ...). Given a band set $F = \{B_i\}_{i=1\dots N}$ and the output class ω, find a subset S_M, with $M < N$ that optimizes the objective function. The Sequential forward selection (SFS) is the simplest greedy search algorithm [29]. With an empty set of bands S at the beginning, we start to add sequentially the band that maximizes the cost function when combined with the bands that have already been selected. The main advantage of SFS algorithms is that is relatively low computational burden [30]. The ideal greedy selection algorithm to solve our problem can be described by the following procedures similar to previous work under [29]:

(1)		Initialization: Set $F \leftarrow$ "initial set of $\|F\|$ input Bands" and set $S \leftarrow$ "empty set".
(2)		Computation of the cost-function J of the divergence distance on equation (2).
(3)		Choice of the first Band: find the Band that maximizes J on step (2), set $F \leftarrow F \backslash \{B_i\}$ and set $S \leftarrow \{B_i\}$.
(4)		Greedy selection: repeat until $\|S\| = d'$.
	(a)	Computation of the cost function Divergence distance: $\forall B_i \in F$, compute J.
	(b)	Selection of the next Band: chose the Band $B_i \in F$ that maximizes J, set $F \leftarrow F \backslash \{B_i\}$ and set $S \leftarrow S \cup \{B_i\}$.
(5)		Output the set S containing the selected Bands.

This algorithm is the same as [29] [31] [32] except for the fourth step. Instead of calculating Mutual Information as a cost function between multiple variables, we propose in this work the use of the Divergence Distance criterion between multiple bands in order to select the salient ones for hyperspectral image classification. In the following section, we show how to estimate Divergence distance using GMM.

3.2 Divergence Distances based on GMM

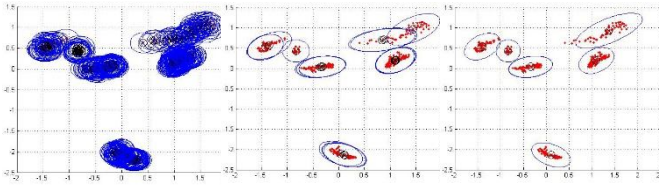
Since we are using data from a real world setup, the cost-function J_d will be calculated using GMM. For each class pair ω_i and ω_j, the equation (1) can be expressed as:

$$J_d(\omega_i, \omega_j) = \sum [p(x|\omega_i) - p(x|\omega_j)] \ln \frac{p(x|\omega_i)}{p(x|\omega_j)} \qquad (6)$$

Now if we replace $p(x|\omega_i)$ and $p(x|\omega_j)$ by its expression from equation (4) and (5), the Divergence distance based on GMM can be expressed as follow:

$$J_d(\omega_i, \omega_j) = \sum [\sum_{k=1}^{K_i} \pi_{i,k} p(x|\mu_{i,k}, \Sigma_{i,k}) - \sum_{k=1}^{K_j} \pi_{j,k} p(x|\mu_{j,k}, \Sigma_{j,k})] \ln \frac{\sum_{k=1}^{K_i} \pi_{i,k} p(x|\mu_{i,k}, \Sigma_{i,k})}{\sum_{k=1}^{K_j} \pi_{j,k} p(x|\mu_{j,k}, \Sigma_{j,k})} \qquad (7)$$

In order to compute the cost-function based Divergence distance by GMM equation (7), a number of parameters must be estimated: π the mixing coefficient, μ the mean, Σ the covariance matrix and K the number of clusters. The main the challenge when using the GMM is to estimate its parameters, with the Expectation-Maximization (EM) algorithm [33], three of those parameters π, μ, Σ can be estimated however the number of component K is user defined, hence it needs to be given a priori, usually after observing the nature of the data. The choice of parameter K is quite important in this study as it can directly affect the estimation of the covariance matrix, since we can easily end up with the "Hughes phenomenon". Next, we present approaches to optimally choose the number of component K:



(a) Iteration 1 (b) Iteration 10 (c) Iteration 30

Figure 1: Example of the REM implementation (a)-(b), all data points are used for Initialization, then discarding clusters that do not met required criteria; (c) the processes convergent to an optimum number of clusters $k = 6$ after 30 iterations.

GMM with Bayesian Information Criterion: To choose the right number of components K for GMM to represent the true distribution of data, a model selection technique will be used. This will provide a method to maximize the likelihood of the training data while attempting to avoid over fitting [34]. One model selection is the Bayes Information Criterion (BIC) [16], introduced by [35], and is defined as:

$$BIC = -2 \times \ln(likelihood) + \ln(N) \times k \qquad (8)$$

where k and N respectively are the number of parameters estimated and the number of observations. The model that minimize the BIC criterion is considered better [36] [37].

GMM with Robust Expectation-Maximization: The second approach we propose to estimate/compute the Divergence distance based on GMM is the Robust Expectation Maximization (REM) algorithm [20]. With the EM algorithm, the number of clusters K is user defined and it has to be defined before hand. REM was developed to automatically obtain an optimal number of clusters K, thus the number of component will no longer have to be defined a priori. The REM algorithm uses all data points as seeds to solve the problem of choosing cluster centers and when a cluster doesn't met the required criteria, it is discarded and the number of component K is decreased until achieving automatically an optimal number of clusters (as seeing in figure 1). For more detail about the algorithm, see [20].

3.3 Regularization problem

For the estimation of the covariance matrix, it is well known that small sample size datasets usually cause "Hughes phenomenon" and singularity problems [22] and by partitioning the already small set of data into multiple clusters and then estimating their statistics, we can end up with an ill-conditioned mixture model [38]. Since the covariance matrix of each component should be invertible in order to compute equation (5), the sample size of each component should not be less than the dimensionality of the data [22]. For GMM, the "Hughes phenomenon" is mostly related to the estimation of the covariance matrix [39]. One way around this problem is the regularization techniques of sample covariance matrix:

Leave One Out Covariance (LOOC): To reduce the estimation error and to avoid the singularity of covariance matrix in equation (5), we use the regularization process [3]

[22] [39] [40]. Let S the covariance matrix and $diag(S)$ its diagonal version, the following covariance estimators are commonly used for the regularization process:

$$S_i^{looc}(\alpha_i) = \begin{cases} (1-\alpha_i)diag(S_i) + \alpha_i S_i & \text{if } 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & \text{if } 1 \le \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & \text{if } 2 \le \alpha_i \le 3 \end{cases} \quad (9)$$

The optimization strategy consists of evaluating several values of α_i through maximizing the average log likelihood of the Gaussian density [40]. Since in our case we are using an iterative approach to select bands, using the regularization techniques as described by equation (9), can add to the complexity of the algorithm and to the computation time.

Maximum Entropy Covariance Selection (MECS): The MECS method deals directly with singular and unstable covariance matrices; it uses the principle of maximizing the information under an incomplete and consequently uncertainty context rather than optimizing classification accuracy or group likelihood [41]. It is on combining the sample group covariance matrices and the pooled covariance matrix [41]. We are particularly interested in this method because according to [41], MECS: - does not require an optimization procedure, - can be used whenever the sample group covariance matrices are poorly estimated or ill posed, - perform at least as well as any other method and at a much lower computational cost.

4. EXPERIMENTAL STUDY

4.1 Dataset

4.1.1 Indian Pines dataset

This hyperspectral image was gathered by AVIRIS sensor site in North-western Indiana on June 12, 1992 over the Indian Pines test. First used by David Landgrebe and his students [21] [22] [23] [40] and it has become a benchmark for testing hyperspectral supervised classification algorithms. The data is an image of 145×145 pixels with a spatial resolution of 18m by 224 bands in the wavelength range of $0.4 - 2.510^{(-6)}\ m$.

4.1.2 Botswana dataset

This scene was acquired on May 31, 2001 by Hyperion sensor over a strip of $7.7km$ on the Okavango Delta, Botswana. 242 bands are collected in the wavelength range of $400 - 2500nm$. The UT Center preprocessed the data for Space Research, removed noisy bands, and identified 14 classes / land cover types.

4.1.3 Kennedy Space Center

The data was gathered by AVIRIS sensor on March 23, 1996 over the Kennedy Space Center (KSC), Florida. The data consist of 176 bands - water absorption band removed - collected in the wavelength range of $400 - 2500nm$ of $10nm$ width and a spatial resolution of 18m. Due to certain vegetation types with similar spectral signatures, the land cover for this scene is difficult to define. KSC personnel developed the classification scheme for the data and 13 classes for the site were identified.

4.2 Experimental setup

The proposed band selection approach with the divergence distance presented in this work was tested in Matlab (R2014a), on a 64-bit PC with an i7 microprocessor (2.20GHz) and 6 GB of RAM. We first run the experiment using the proposed approach on the benchmark dataset Indian Pine (92AV3C), and then we conducted the same experiment on Botswana and Kennedy Space Center datasets.

For classification purposes, the dataset is divided into two halves. We choose a chessboard selection of pixels with ground truth to yield a training/testing split of 50%. The selected bands then are fed to SVM and ELM classifiers in order de show their classification performances. The first used classifier is SVM through the LIBSVM library with RBF as kernel function and the grid search technique to find the C and γ parameters [42]. The second used classifier is the ELM [43] [44], which is extremely fast with good generalization [45]. In this classifier, we used RBF as activation function for ELM and the grid search technique to choose the number of hidden neurons.

4.3 Results and discussions

The first experiment to take place is the assessment of the divergence distances. To measure its effectiveness, we run the test on the Indian Pine benchmark dataset. We point out that this scene has been often used in various studies such as [21] [22] [23] [40].

The purpose of this experiment is to evaluate each band independently from the rest and see how it ranks in terms of class separability according to the cost-function of the probabilistic distances Bhattacharyya and Divergence. The higher the value we get the more the classes are separable on that band. In figure 2 we can notice that band region $170 \sim 190$ have the highest value and indeed the first selected band is 168 using divergence.

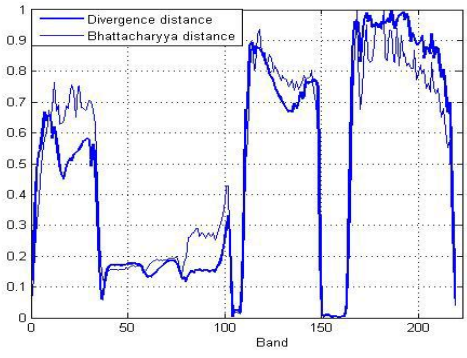
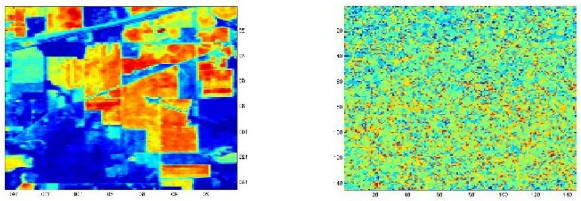
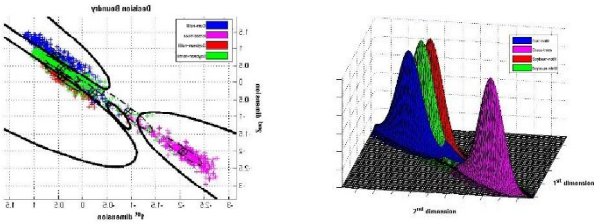


Figure 2: Divergence VS Bhattacharyya score for each band for 92AV3C dataset.



(a) **(b)**

Figure 3: (a) is the first selected band 168 for 92AV3C dataset using divergence, **(b)** is the band 153 one of the noisy band that been discarded by the selection algorithm.



(a) **(b)**

Figure 4: (a) The decision boundary with the first two selected bands with divergence distance based GMM-REM for the 4-class 92AV3C subset scene, **(b)** the probability estimation of each class.

It has been reported in previous studies [23] [24] that the bands 104-108, 150-163 and 220 in Indian Pine are the region of water absorption therefore they contain only noise and no useful information which can be seen clearly in figure 2 as they got the lowest value and present on figure 3b a noisy image. Hence, the bands selection approach with the divergence distance did not choose any of those noise bands in the selection process, as they need to be discarded automatically. Hence, the Divergence distance based on GMM is capable on measuring the pertinence of a band.

The second experiment is to evaluate the performances of the first two selected bands by drawing a decision boundary between the spectral classes using the 92AV3C subset scene. In order to get an easier visual inspection the test is done on a portion of the Indian Pine dataset containing only the four class with the highest number of samples instead of using all the 16 classes similar to [23] [24]. The first two selected bands with the divergence distance are 168 and 142. In figure 4, we can see that data is highly correlated yet with just these two bands we were able to draw a decision boundary between the four classes of the data set and separate one class from the rest.

The overall classification score with SVM for the first two selected bands is 81.74%. The other classes in the other hand are still mixed up and will need to go on a higher dimension with more bands before achieving the desired separability between the classes. For the Indian Pine sub scene, a classification score accuracy of 93.44% with SVM is achieved with only five bands, and a classification score of 97.31% at the dimension thirty.

The final experiment in this study is to evaluate the Divergence distance in contrast of our previous work on the Bhattacharyya distance in order to find out which criterion gives better results for band selection using the three datasets: 92AV3C, KSC and Botswana.

For the Indian Pine dataset, we do notice from figure 5a and 5b that the Divergence distance with GMM-REM estimation gives a slightly better classification score than its competitors. According to figure 5a, the Bhattacharyya approach with the GMM-BIC estimation has head start with the first five selected bands. The Divergence distance with GMM-REM then catch up to the Bhattacharyya, from then on it stays on top. Meanwhile, in figure 5b with the ELM classifier, the divergence with GMM-REM almost stays on top through the entire selected band but with a very small difference from the Bhattacharyya with GMM-BIC.

In the KSC dataset, with the first two selected bands, we can see from figure 6a and 6b that there is a big gap between the Bhattacharyya and the Divergence distance in terms of classification score with SVM and ELM in favor of the divergence distance. Between the two distances with the same GMM estimation a gap around 7% is found and a gap around 14% between the divergence with BIC and Bhattacharyya with REM. After the first seven selected band, there is almost no difference between the two distances in terms of an overall classification score with SVM and ELM.

According to figure 7a of the Botswana dataset, with the first two selected bands, a margin of 10% is found in favor for the Bhattacharyya with the GMM-REM estimation in the SVM classifier. From figure 7a and 7b, the Bhattacharyya distance performs better with first five selected bands, and almost stays on top of the other distance through the whole selected band pool but with a very small margin.

By looking at the curves in figure 5, 6 and 7, with GMM-REM estimation not only we can get a pretty good classification accuracy and but also it can resist and delay the Hugh phenomenon, which means that we can get a robust probability estimation for our probabilistic distance and the results will be more reliable. Unlike the estimation with GMM-BIC, the Hughes Phenomenon started to manifest itself clearly around the band fifteen in figure 7 and twenty-five in figure 5b. In fact, we did notice in our experiment that the more bands we add to the pool the more GMM-BIC is forced to lower its number of component K since the number of observation is already small until it starts to model the data using the normal distribution, the thing that can explain the early manifestation on curse of dimensionality.

The experimental results of the Divergence distance based on GMM BIC and REM, compared to the Bhattacharyya distance in the same setup, shows that the classification curves

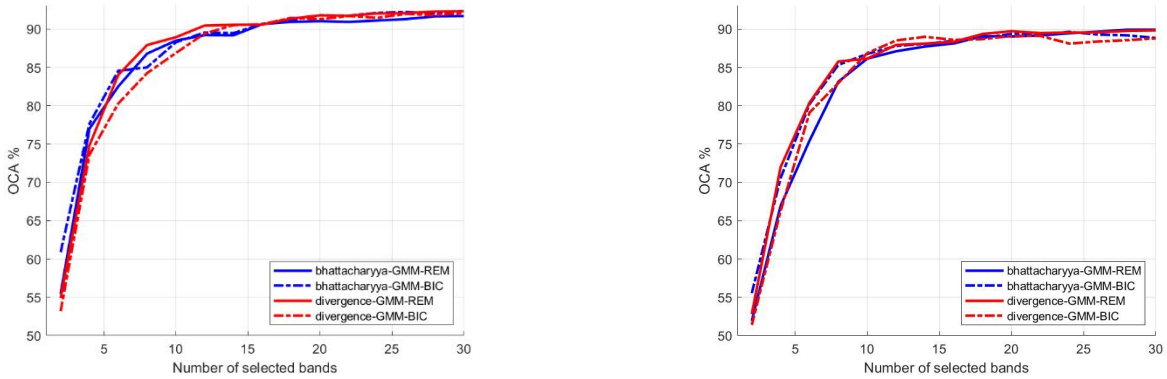


Figure 5: Overall classification Accuracy of the selected bands for dataset 92AV3C using **(a)** SVM Classifier, **(b)** ELM Classifier.

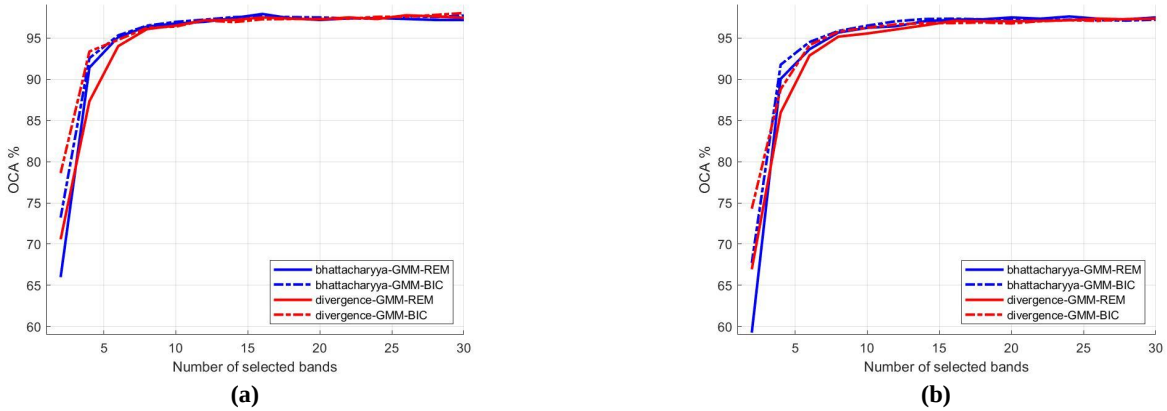


Figure 6: Overall classification Accuracy of the selected bands for dataset KSC using **(a)** SVM Classifier, **(b)** ELM Classifier.

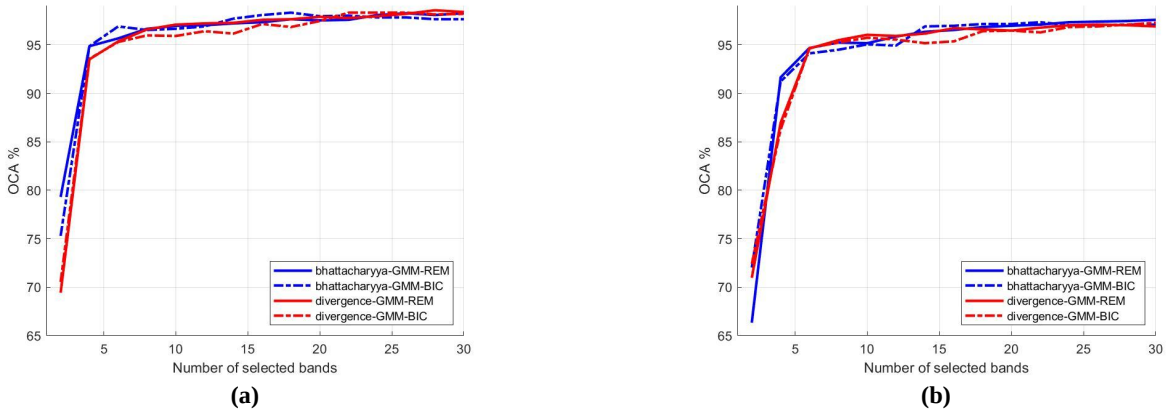


Figure 7: Overall classification Accuracy of the selected bands for dataset Botswana using **(a)** SVM Classifier, **(b)** ELM Classifier.

Between the two distance overlaps with each other's. Depending on the number of the selected bands, on how well the GMM was estimated, on how well the classifier parameter were chosen and on the data set itself how correlated it is and how its post treatment was to deal with the outliers. We do notice that Divergence distance performs the best at times and others times the Bhattacharyya distance performs better. According to figure 5, 6 and 7, the results are close to each other and the margin between the classification curves of the selected bands with both distances is not wide enough to concur on the superiority of one on the others. Therefore, it is hard to decide which one of the distances is the best. Thus, we can conclude that in our setup, the Divergence distance performs as well as the Bhattacharyya distance.

5. CONCLUSION

This paper presented a new band selection algorithm with a GMM based Divergence distances for hyperspectral image classification, using the sequential forward selection technique to reduce the data dimension. Since the EM algorithm is sensitive to the initial values and its number of components needs to be user defined, a GMM BIC and GMM REM were introduced in this study to model the Divergence distance. Our main contribution in this work is a new approach to give a robust estimation of Divergence distance using GMM-BIC and GMM-REM algorithm for hyperspectral band selection.

The initial performed experiment with divergence distance on the Indian Pine dataset has shown the reliability of the criterion as a similarity measure. It has the ability to evaluate the pertinence of a band and thus choose the best bands from a given hyperspectral image dataset and discard the ones with no relevant information.

The experimental results, have demonstrated the effectiveness of our proposed method in terms of classification accuracy with fewer bands. In fact, with the GMM-REM estimation, not only we can get a good classification accuracy but also it can resist and delay the Hugh phenomenon, which means that we can get a robust probability estimation for our probabilistic distance and the results will be more reliable.

This investigation was inspired by our previous work on the Bhattacharyya distance, thus we were particularly interested in using the Divergence distance to find out which one gives better results. On the three used datasets 92AV3C, KSC and Botswana, The experimental study showed that between the two distances, the results are close to each other; therefore, it is hard to decide which one is the best in our current setup. Thus, we can conclude that the Divergence distance performs as well as the Bhattacharyya distance.

REFERENCES

1. G. Hughes, **On the mean accuracy of statistical pattern recognizers**,IEEE transactions on information theory, vol. 14, no. 1, pp. 55–63, 1968. https://doi.org/10.1109/TIT.1968.1054102
2. B. M. Shahshahani and D. A. Landgrebe, **The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon**, IEEE Transactions on Geoscience and Remote Sensing, vol. 32, no. 5, pp. 1087–1095, Sep 1994. https://doi.org/10.1109/36.312897
3. J. Richards, **Remote Sensing Digital Image Analysis: An Introduction**. Springer Berlin Heidelberg, 2012.
4. M. Lahlimi, M. Ait Kerroum, and Y. Fakhri, **Band selection with bhattacharyya distance based on the gaussian mixture model for hyperspectral image classification**, in Recent Advances in Electrical and Information Technologies for Sustainable Development, S. El Hani and M. Essaaidi, Eds. Cham: Springer International Publishing, 2019, pp.87–94. https://doi.org/10.1007/978-3-030-05276-8_10
5. J. Wang, X. Wang, K. Zhang, K. Madani, and C. Sabourin, **Morphological band selection for hyperspectral imagery**, IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 8, pp. 1259–1263, Aug 2018. https://doi.org/10.1109/LGRS.2018.2830795
6. A. Webb and K. Copsey, **Statistical Pattern Recognition**. Wiley, 2011. https://doi.org/10.1002/9781119952954
7. A. Datta, S. Ghosh, and A. Ghosh, **Unsupervised band extraction for hyperspectral images using clustering and kernel principal component analysis**, International Journal of Remote Sensing, vol. 38, no. 3, pp. 850–873, 2017.
8. M. P. Uddin, M. A. Mamun, and M. A. Hossain, **Feature extraction for hyperspectral image classification**, in 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dec 2017, pp. 379–382.
9. J. A. Richards and J. Richards, **digital image analysis. Springer**, 1999, vol. 3.
10. P. K. Varshney and M. K. Arora, **Advanced image processing techniques for remotely sensed hyperspectral data**. Springer Science & Business Media, 2004. https://doi.org/10.1007/978-3-662-05605-9
11. K. Burgers, Y. Fessehatsion, S. Rahmani, J. Seo, and T. Wittman, **A comparative analysis of dimension reduction algorithms on hyperspectral data**, LAMDA Research Group, pp. 1–23, 2009.
12. M. Mazumder, **Feature extraction techniques for speech processing: A review**, International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, pp. 285–292, 08 2019. https://doi.org/10.30534/ijatcse/2019/5481.32019
13. C. Lee, D. Landgrebe et al., **Feature extraction based on decision boundaries**, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 15, no. 4, pp. 388–400, 1993.
14. R. Huang and M. He, **Band selection based on feature weighting for classification of hyperspectral data**, IEEE Geoscience and Remote Sensing Letters, vol. 2, no. 2, pp. 156–159, 2005. https://doi.org/10.1109/LGRS.2005.844658
15. Z. Du, M. K. Jeong, and S. G. Kong, **Band selection of hyperspectral images for automatic detection of poultry skin tumors**, IEEE Transactions on Automation Science and Engineering, vol. 4, no. 3, pp. 332– 339, 2007.
16. W. Li, S. Prasad, and J. E. Fowler, **Hyperspectral image classification using gaussian mixture models and markov random fields**, Geoscience and Remote Sensing Letters, IEEE, vol. 11, no. 1, pp. 153–157, 2014.

17. M. M. Dundar and D. A. Landgrebe, **A cost-effective semisupervised classifier approach with kernels**, IEEE Transactions on Geoscience and Remote Sensing, vol. 42, no. 1, pp. 264–270, Jan 2004. https://doi.org/10.1109/TGRS.2003.817815
18. M. M. Dundar and D. A. Landgrebe, **Toward an optimal supervised classifier for the analysis of hyperspectral data**, Geoscience and Remote Sensing, IEEE Transactions on, vol. 42, no. 1, pp. 271–277, 2004.
19. W. L. Martinez and A. R. Martinez, **Computational statistics handbook with MATLAB**. CRC press, 2007, vol. 22.
20. M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, **A robust em clustering algorithm for gaussian mixture models**, Pattern Recognition, vol. 45, no. 11, pp. 3950–3961, 2012. https://doi.org/10.1016/j.patcog.2012.04.031
21. L. O. Jimenez and D. A. Landgrebe, **Supervised classification in highdimensional space: geometrical, statistical, and asymptotical properties of multivariate data**, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 28, no. 1, pp. 39–54, Feb 1998.
22. B.-C. Kuo and D. A. Landgrebe, **A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction**, Geoscience and Remote Sensing, IEEE Transactions on, vol. 40, no. 11, pp. 2486–2494, 2002.
23. S. Tadjudin and D. A. Landgrebe, **Robust parameter estimation for mixture model**, IEEE Transactions on Geoscience and Remote Sensing, vol. 38, no. 1, pp. 439–445, 2000.
24. G. Camps-Valls and L. Bruzzone, **Kernel methods for remote sensing data analysis**. John Wiley & Sons, 2009. https://doi.org/10.1002/9780470748992
25. S. Wang and C. Wang, **Research on dimension reduction method for hyperspectral remote sensing image based on global mixture coordination factor analysis**, The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 40, no. 7, p. 159, 2015.
26. A. Datta, S. Ghosh, and A. Ghosh, **Band elimination of hyperspectral imagery using partitioned band image correlation and capacitory discrimination**, International Journal of Remote Sensing, vol. 35, no. 2, pp. 554–577, 2014.
27. S. Theodoridis and K. Koutroumbas, **Pattern Recognition, Fourth Edition**. Academic Press, 2008.
28. L. I. Kuncheva, **Combining Pattern Classifiers: Methods and Algorithms**. Wiley-Inderscience, 2004. https://doi.org/10.1002/0471660264
29. M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, **Textural feature selection by joint mutual information based on gaussian mixture model for multispectral image classification**, Pattern Recogn. Lett., vol. 31, no. 10, pp. 1168–1174, Jul. 2010.
30. L. Burrell, O. Smart, G. K. Georgoulas, E. Marsh, and G. J. Vachtsevanos, **Evaluation of feature selection techniques for analysis of functional mri and eeg**," in DMIN, 2007, pp. 256–262.
31. R. Battiti, **Using mutual information for selecting features in supervised neural net learning**, IEEE Transactions on Neural Networks, vol. 5, no. 4, pp. 537–550, July 1994. https://doi.org/10.1109/72.298224
32. N. Kwak and Chong-Ho Choi, **Input feature selection for classification problems**, IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 143–159, Jan 2002. https://doi.org/10.1109/72.977291
33. R. O. Duda, P. E. Hart, and D. G. Stork, **Pattern Classification (2nd Edition)**. Wiley-Interscience, 2000.
34. K. Z. Yu, **Generating gaussian mixture models by model selection for speech recognition**, 2006.
35. G. Schwarz et al., **Estimating the dimension of a model**, The annals of statistics, vol. 6, no. 2, pp. 461–464, 1978.
36. J. Chen and Z. Chen, **Extended bayesian information criteria for model selection with large model spaces**, Biometrika, vol. 95, no. 3, pp. 759–771, 2008. https://doi.org/10.1093/biomet/asn034
37. H. D.-G. Acquah, **Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of an asymmetric price relationship**, Journal of Development and Agricultural Economics, vol. 2, no. 1, pp. 001–006, 2010.
38. M. M. Dundar and D. Landgrebe, **A model-based mixture-supervised classification approach in hyperspectral data analysis**, Geoscience and Remote Sensing, IEEE Transactions on, vol. 40, no. 12, pp. 2692–2699,2002. https://doi.org/10.1109/TGRS.2002.807010
39. M. Fauvel, C. Dechesne, A. Zullo, and F. Ferraty, **Fast forward feature selection for the nonlinear classification of hyperspectral images**, arXiv preprint arXiv:1501.00857, 2015.
40. S. Tadjudin and D. A. Landgrebe, **Covariance estimation with limited training samples**, IEEE Transactions on Geoscience and Remote Sensing, vol. 37, no. 4, pp. 2113–2118, July 1999.
41. C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa, **A new covariance estimate for bayesian classifiers in biometric recognition**, IEEE Transactions on circuits and systems for video technology, vol. 14, no. 2, pp.214–223, 2004.
42. S. Dhariwal, **An efficient approach for semantic image classification using normalization method**, International Journal of Advanced Trends in Computer Science and Engineering, pp. 1268–1274, 08 2019.

https://doi.org/10.30534/ijatcse/2019/37842019

43. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, **Extreme learning machine:a new learning scheme of feedforward neural networks**, in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, vol. 2. IEEE, 2004, pp. 985–990.
44. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, **Extreme learning machine: theory and applications**, Neurocomputing, vol. 70, no. 1, pp. 489–501, 2006. https://doi.org/10.1016/j.neucom.2005.12.126
45. A. S. Kumar, **Ensemble online sequential extreme learning machine and swarm intelligent based feature selection for cleveland heart disease prediction system**, International Journal of Advanced Trends in Computer Science and Engineering, vol. 6, no. 5, 2017.