Estimation of the "natural" number of classes of a multispectral image

André R. S. Marçal , Janete S. Borges
Faculdade de Ciências
Universidade do Porto
Porto, Portugal
andre.marcal@fc.up.pt, jsborges@fc.up.pt

Abstract — **The availability of an effective internal similarity index to determine the "natural" number of classes in a multi-spectral satellite image would benefit the process of unsupervised image classification. Two similarity indices (DB and Xu) were tested in sections of multi-spectral satellite images from Landsat TM, SPOT HRVIR, ASTER and IKONOS. The images were initially clustered into a manageable number of classes using an unsupervised classification algorithm. These results were then structured hierarchically, and the internal similarity indices computed for each level. The inspection of the DB and Xu index plots were used to select the "natural" number of classes for each test image.**

Similarity index; Hierarchical clustering; Image Classification

I. INTRODUCTION

Multi-spectral satellite images are often used to produce thematic maps through image classification. Supervised classification methods are most commonly used. These methods require a prior identification of training areas, which are used to characterize the spectral signature of each class. The results from supervised classification are often unsatisfactory, both due to the presence of mixed pixels and to the lack of knowledge of the actual classes present in the dataset. Classes that are not spectrally distinguishable might be looked for, and others that have a clear signature in the feature space might not have been initially predicted.

Unsupervised classification algorithms explore the multi-spectral feature space. The classes are assigned to densely occupied areas, or clusters. The classes obtained by this process are in principle better suited to the data, but the results vary with the choice of algorithm and the associated parameters. The cluster configuration is valid if clusters cannot reasonably occur by chance or as a beneficial artifact of a clustering algorithm [1]. Even if a valid configuration is achieved there will still be a problem: labeling the classes produced by the unsupervised classifier. One of the reasons for the difficulty of this a posteriori labeling is the large number of classes usually produced by the classifier. One possible strategy to assist in this process is the hierarchically structuring of the resulting K classes. This allows the user to label the classes hierarchically, and also to have classified images at various levels. A set of solutions is thus made available (classified images with 2, 3, ..., K classes), which brings a new question: what is the best partition? Or, in an alternative form, what is the "natural" number of classes in the dataset?

A number of similarity indices can be used to provide an answer to that question in a classification process [2]. However, most of these indices are not applicable to multi-spectral images, due to the huge computation effort required. This paper presents the results of the application of two internal similarity indices on multi-spectral satellite images: Davies-Bouldin [3], and a combined sum of squared error and minimum of between-cluster distance proposed by Xu [4].

II. SIMILARITY INDICES

Let $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be the patterns to classify, and $\mathbf{x}_i$ a vector of the d-dimension feature space. For digital images, the patterns are the image pixels. The classification of the image corresponds to the establishment of a partition $\{C_1, C_2,..., C_K\}$ for the n patterns, so that $i \in C_k$ if $\mathbf{x}_i$ belongs to class k, k=1,2,…,K. The center of class k is a vector, of dimension d where n_k is the number of patterns assigned to class k

$$\mathbf{m}_k = \frac{1}{n_k} \sum_{i \in C_k} \mathbf{x}_i \cdot \tag{1}$$

The Sum-of-Squared Error (SSE) for class k (J_k) is the sum of the quadratic distances between all its elements and the class center (2). The distance $\delta(\mathbf{x},\mathbf{y})$ between two vectors $\mathbf{x}$ and $\mathbf{y}$ is computed using a metric, such as the Minkowski distance or the Euclidian distance [5].

$$J_k = \sum_{i \in C_k} \delta^2(\mathbf{x}_i, \mathbf{m}_k) \tag{2}$$

A. Davies and Bouldin index

The classification established a partition of the data in K classes, which are then hierarchically clustered, producing K-1 partitions (with a number of classes $h = K$, K-1, …, 2 classes). Once a partition h is established, $R_{i,j}$ provides a "within to between" separability for the pair of classes i, j (3). The values of $R_{i,j}$ will be low when both classes are "compact" and well separated from each other.

This work was done with the support of "Centro de Investigação em Ciências Geo-Espaciais, Faculdade de Ciências da Universidade do Porto", financed by "Fundação de Ciência e Tecnologia" through POCTI/FEDER.

$$R_{i,j} = \frac{J_j/\sqrt{n_j} + J_i/\sqrt{n_i}}{\delta(\mathbf{m}_j, \mathbf{m}_i)} \quad (3)$$

The Davies-Bouldin (DB) index of a partition h is the average of the highest $R_{i,j}$ of each class (4). The lowest the value of $DB(h)$ the better is the separability between classes, and therefore the better the partition h.

$$DB(h) = \frac{1}{K-1}\sum_{h=2}^{K} R_h \ , \quad R_h = \text{Max}\{ R_{i,j} , \forall i \neq j \} \quad (4)$$

The DB index was used with the Minkowski distance [3] and the Euclidian distance [5]. A modification of the DB index is to use the Mahalanobis distance [6] as the metric in (2). However, the results were not significantly different that those using the Euclidian distance.

B. *Xu index*

The Ward distance (5) is used to evaluate the distance between two clusters i and j [4].

$$\delta_{ij}^{w} = \sqrt{\frac{n_i \times n_j}{n_i + n_j}} \times \left| \mathbf{m}_i - \mathbf{m}_j \right| \quad (5)$$

A dissimilarity measure, in terms of the Minimum of Between-Cluster Distances (MBCD), can be defined for a partition with k classes (6). Both SSE and MBCD alone are insufficient to establish a criterion for the best partition. However, the two can be used combined to form an index [4].

$$M = \min_{i<j} \delta_{ij}^{w} \ , \quad i,j = 1,2,\ldots, k \quad (6)$$

The index proposed by Xu, *E(h)*, evaluates the level h of the hierarchical structure, comparing the SSE and MBCD of this level with the proceeding. The index *E(h)* is computed using (7), where *J(h)* is the sum of the J_k for all clusters of partition h.

$$E(h) = \frac{M(h) - M(h+1)}{\sqrt{J(h)} - \sqrt{J(h+1)}} \quad (7)$$

When plotting the index E as a function of h, a significant maximum of *E(h)* should be expected to appear at level h^*, where lie h^* natural groupings or clusters [4].

C. *Example with synthetic data*

The performance of the DB and Xu indices was verified using synthetic data. An example of application of both indices to a synthetic data set is presented in this section. A total of 100 two-dimensional patterns were generated using 8 Gaussian curves, with a standard deviation of 0.04. The data was classified in k classes (with k=2, 3, ..., 12), using MATLAB algorithm ClusterData [7]. The synthetic data is presented in Fig.1, as well as the data clustered in 5, 7 and 10 classes.

The DB and the Xu indices were computed for each partition. A plot of both indices is presented in Fig. 2. The Xu index clearly points to a solution for k=7, with a second choice for k=5 but clearly less favored. On the contrary, the DB index does not distinguish significantly between the solutions with 5 and 7 classes, with only a very slight advantage towards k=5. The combined analysis of both indices would suggest a choice of 7 classes as the most reasonable in this case.

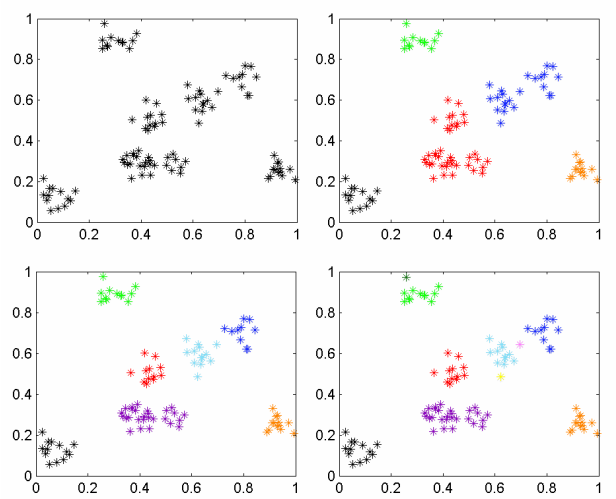

Figure 1. Synthetic data (top left) clustered in 5 classes (top right), 7 classes (bottom left) and 10 classes (bottom right).

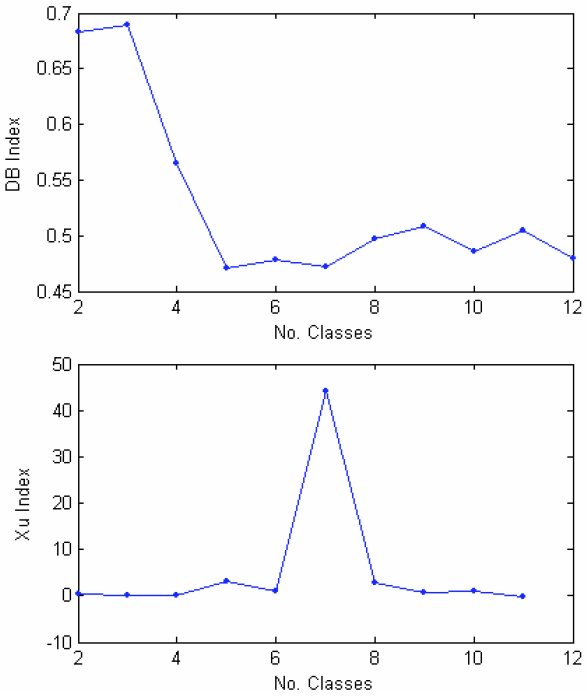

Figure 2. Davies-Bouldin index (top) and Xu index (bottom) plots for the synthetic data of Fig.1.

III. RESULTS

Four satellite multi-spectral images were selected to evaluate the performance of the DB and Xu similarity indices. These images are small sections of 512 by 512 pixels, and are presented in Fig. 3 as false color composites, with the near infrared in the green channel. Image I is from Landsat TM and covers an agricultural region. Image II is from SPOT HRVIR, of an estuary and the surrounding wetlands. Image III is from the ASTER sensor and covers an area mainly with forest, with a river and small urban parts. Image IV is an IKONOS multi-spectral image of an urban area.

Figure 3. Test images: I – Landsat TM (top left), II – SPOT HRVIR (top right), III – ASTER (bottom left), IV – IKONOS (bottom right)

Each test image was classified using the algorithm ISODATA implemented on the software PCI Geomatics [8]. The same set of parameters was used throughout, including the range of classes allowed (20-40). In all for cases the classifier converged with a solution of more than 30 classes. The classification results were hierarchically structured, using the Euclidian distance metric between the class centers ($\mathbf{m}_k$) as the agglomerative criterion. A total of 29 classified images (with 30, 29, ..., 2 classes) were available from each test image, and were used to apply the DB index and the Xu index.

A. *Test image I – Landsat TM*

Test image I has 6 spectral bands with a pixel size of 30m. The thermal band of TM was not used due to its lower spatial resolution. The values computed for the DB and Xu indices are presented in Fig. 4, plotted versus the number of classes in each partition. The Xu index has three clear maxima, pointing to possible solutions for h=4, h=13 or h=16. The DB index is not as convincing, but it seems to indicate that a choice of h=16 might be a wise one, as there is a local minimum at this level. In this case there is a single value to select – a partition of the image into 16 classes.

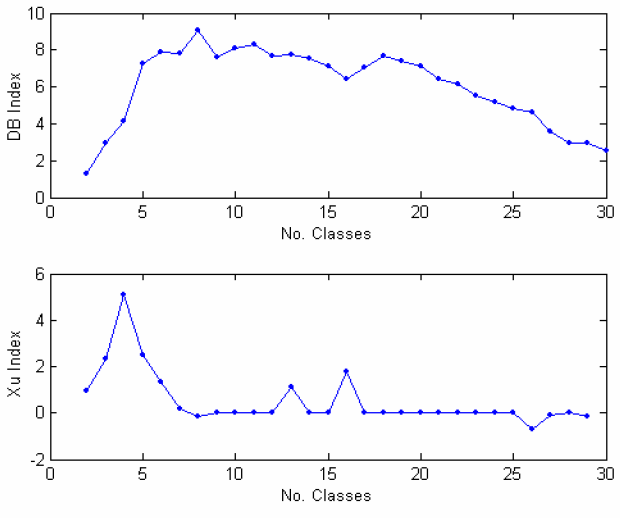

Figure 4. DB index (top) and Xu index (top) applied to test image I.

B. *Test image II – SPOT HRVIR*

Test image II has 4 spectral bands with a pixel size of 20m. The plots for the DB index and the Xu index for this image are presented in Fig. 5. The Xu index first favors a partition in 27 classes, followed by a partition in 4 classes. There are two other local maxima, but not very prominent. The DB index is consistent with both choices (for h=4 or h=27), although it also favors h=13 or h=17. But combining the information from both indices, the best solutions are for h=4 and h=27.

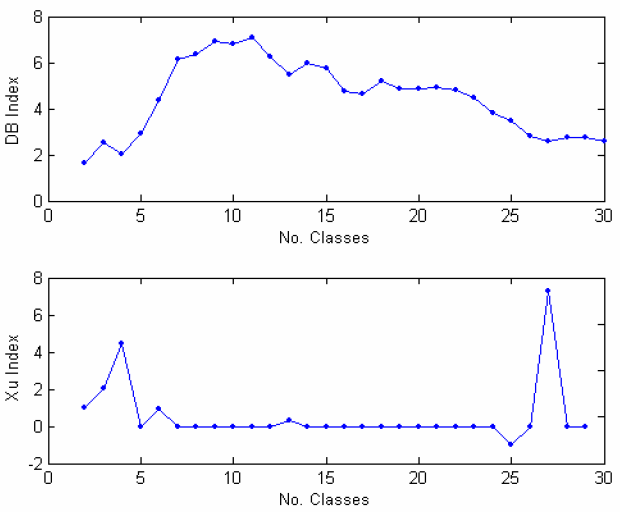

Figure 5. DB index (top) and Xu index (top) applied to test image II.

C. *Test image III – ASTER*

Test image III is from ASTER, a sensor with 14 spectral bands. However, at the best spatial resolution (15 m) only 3 bands are available. The classifier only used the first 3 bands of ASTER. The plots for the internal similarity indices DB and Xu are presented in Fig. 5. The Xu index has a very strong

maximum for h=2, which is also the minimum for the DB index. This is clearly a classification very "natural" from the data perspective, as the 2 classes are water and land, which have a very different spectral signature in the visible and near-infrared parts of the spectrum. However, as this might not a very useful classification from the user point of view, other possible choices for h should be considered. Unfortunately, in this case the 2 indices do not provide consistent indications. The Xu index has slight maxima at h=13 and h=23. The DB index points to h=8, h=19 and h=26. The value h=23 might be selected by the DB index plot in a second group of candidates.

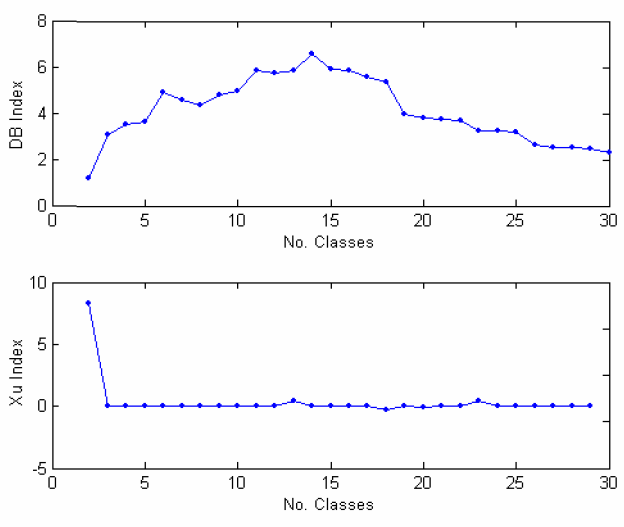

Figure 6. DB index (top) and Xu index (top) applied to test image III.

D. *Test image IV – IKONOS*

Test image IV is a section of an IKONOS multi-spectral image, with 4 bands and a pixel size of 4m. The DB index and Xu index plots are presented in Fig. 7. The Xu index clearly selects two partitions as good ones: h=3 and h=8. The DB index indication of the best number of classes is not so obvious. The most favored seem to be h=3, h=7, h=8 and h=19. These values include the two selected by the Xu index, which are therefore the suitable values for this image.

IV. CONCLUSIONS

Unsupervised classification methods have great potential for the classification of multi-spectral satellite images, as they permit the identification of the classes that are naturally distinguishable in the data. One of the reasons that justify the fact that these methods are somehow neglected for satellite image classification is the difficulty in assessing the results produced. A number of statistical indices have been developed and used to assess the classification results [2], but few are applicable to large data volumes, such as multi-spectral satellite images. The availability of an effective internal similarity index to determine the "natural" number of classes in a multi-spectral satellite image would benefit the process of unsupervised image classification.

The method proposed here starts by clustering the multi-spectral image, using an unsupervised classification algorithm, into a manageable number of classes. These are then structured hierarchically, and the internal similarity indices proposed by Davies and Bouldin [3] and Xu [4] are computed. Both indices are applicable to large multi-spectral images. The DB and Xu index plots are used to select the "natural" number of classes from the set of classified images. The plots of the Xu index are generally easier to interpret, as the maxima are very distinguishable. The choice of the best number of classes from the DB index plot is usually not so obvious. However, the combined use of both indices can provide a choice for the "natural" number of classes, or perhaps 2 or 3 good choices.

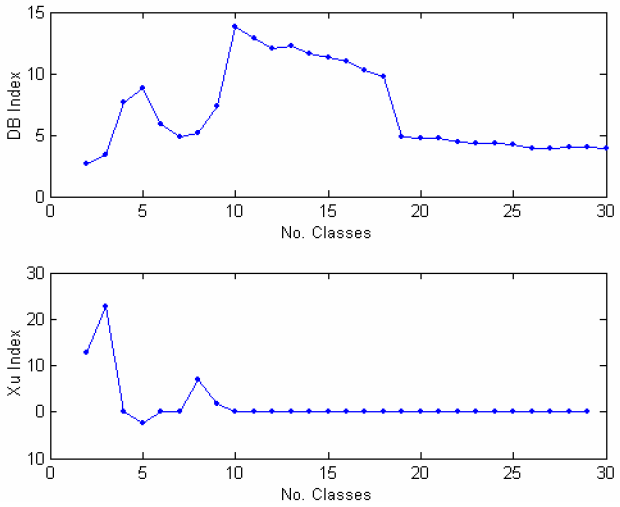

Figure 7. DB index (top) and Xu index (top) applied to test image IV.

REFERENCES

[1] Jain, A. K., Murty, M. N., Flynn, P. J., "Data clustering: A review". ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999.

[2] Milligan, G. W. and Cooper, M. C., "An examination of procedures for determining the number of clusters in a data set". Psychometrika, Vol. 50, No. 2, pp. 159-179, 1985.

[3] Davies, D. L., Bouldin, D. W., "A cluster separation measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, pp. 224-227, 1979.

[4] Xu, S., Kamath, M. V., Capson, D. W., "Selection of partitions from a hierarchy", Pattern Recognition Letters, Vol. 14, No. 1, pp. 7-15, 1993.

[5] Dubes, R. C., "How many clusters are best - an experiment". Pattern Recognition, Vol. 20, No. 6, pp. 645-663, 1987.

[6] J.A. Richards and X. Jia, Remote Sensing Digital Image Analysis, 3rd ed., Berlin, Germany, Springer-Verlag, 1999.

[7] The MathWorks: MATLAB The Language of Technical Computing – Using MATLAB: version 6. The MathWorks, Inc., 2000.

[8] PCI GEOMATICS, X-Pace Reference Manual, Version 8.2, PCI Geomatics, Ontario, Canada, 2001.