Estimating the Natural Number of Classes on Hierarchically Clustered Multi-spectral Images

André R.S. Marçal and Janete S. Borges

Faculdade de Ciências, Universidade do Porto,
DMA, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

Abstract. Image classification is often used to extract information from multi-spectral satellite images. Unsupervised methods can produce results well adjusted to the data, but that are usually difficult to assess. The purpose of this work was to evaluate the Xu internal similarity index ability to estimate the natural number of classes in multi-spectral satellite images. The performance of the index was initially tested with data produced synthetically. Four Landsat TM image sections were then used to evaluate the index. The test images were classified into a large number of classes, using the unsupervised algorithm ISODATA, which were subsequently structured hierarchically. The Xu index was used to identify the optimum partition for each test image. The results were analysed in the context of the land cover types expected for each location.

1 Introduction

Image classification techniques are frequently used to produce land cover maps from multi-spectral satellite images. Usually a supervised classification approach is preferred, making use of training areas to characterise the spectral signature of each class looked for in the image. The results are often disappointing, mainly due to the presence of mixed pixels and an inadequacy between the classes anticipated and the classes actually present in the data. A class identified in training might not be spectrally distinguishable from the other classes. In contrast, there might be some classes in the data, clearly distinguishable from the signal point of view, which were not predicted a-priori. These issues are partly solved when an unsupervised algorithm is applied to the data, but other problems do arise. Unsupervised classification algorithms explore the multi-spectral feature space, looking for densely occupied areas, or clusters, to which classes are assigned. The classes obtained by this process are in principle better suited to the data, but the results can be dependent on the algorithm and the choice of parameters used. This is certainly an important aspect, as the cluster configuration is only considered to be valid if clusters cannot reasonably occur by chance or as a beneficial artefact of a clustering algorithm [1]. Even when this issue is sorted out, there is still a difficulty: labelling the classes produced by the unsupervised classifier. This post-classification labelling is sometimes difficult due to the large number of classes usually created (K). An effective method to assist on this process is to structure the classes hierarchically. A set of K-1 solutions is thus

M. Kamel and A. Campilho (Eds.): ICIAR 2005, LNCS 3656, pp. 447–455, 2005.

made available (classified images with 2, 3, ..., K classes), which brings a new question: which one is the best partition? Or, in an alternative form, what is the "natural" number of classes in the dataset? This is a well-known problem in statistics, but not much explored in image processing, due to the large number of patterns to cluster. This is even more significant in remote sensing, as multi-spectral satellite images are huge data volumes, which are not well manageable for computationally demanding methods.

The validation of a clustering result can be accomplished by carefully applying statistical methods and testing hypotheses [1]. The use of an external index of agreement, such as the Rand index, is appropriate for ascertaining whether the data justify the number of a-priori clusters [2]. Internal examination of validity tries to determine if the structure is intrinsically appropriate for the data [1]. Milligan and Cooper [3] performed a comparative study of 30 similarity indices. However, the large majority of these indices are very demanding computationally, and thus inappropriate for digital images. One criterion that can be applied to large datasets is based on the Minimum of Between Cluster Distance (MBCD) [4]. An improved version of this criterion is proposed by Xu et al. [4]. The purpose of this work was to estimate the usefulness of the Xu similarity index to identify the natural number of clusters in a multi-spectral satellite image.

2 Method

2.1 Similarity Index

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 the class k. The centre of class k is a vector $\mathbf{m}_k$, of dimension d, given by Equation (1), where n_k is the number of patterns assigned to class k.

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

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 centre. 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 [2]. The Ward distance (Equation 3) is used to evaluate the distance between two clusters i and j [4].

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

$$\delta_{ij}^w = \sqrt{\frac{n_i \times n_j}{n_i + n_j}} \times |\mathbf{m}_i - \mathbf{m}_j| \tag{3}$$

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

$$M = min_{i<j}\ \delta_{ij}^{w} \qquad i, j = 1, 2, \ldots, k \tag{4}$$

The initial classification procedure establishes a partition of the data in k classes, which is then hierarchically clustered, producing $k-1$ partitions (with a number of classes $h = k, k-1, \ldots, 2$ classes). The index proposed by Xu, $E(h)$, evaluates the level h of the hierarchical structure by comparing the SSE and MBCD of this level with the proceeding level. The index $E(h)$ is computed using Equation 5, where $J(h)$ is the sum of the J_k for all clusters of partition h.

$$E(h) = \frac{M(h) - M(h+1)}{\sqrt{J(h)} - \sqrt{J(h+1)}} \tag{5}$$

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]. An example of a plot $E(h)$ is presented in Figure 1. The figure shows two partitions, 5 and 8 classes, and the Xu similarity index plot. In this case there is a clear maximum for $h = 5$, indicating that the clustering in 5 classes is the most natural choice for this particular dataset.

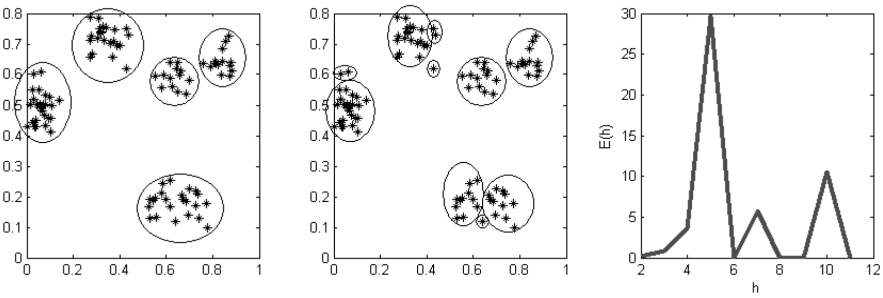

Fig. 1. Example of the Xu index applied to synthetic data. Data classified into 5 classes (left), 8 classes (centre) and Xu index plot (right).

2.2 Hierarchical Classification of Digital Images

Hierarchical clustering methods require the user to specify a measure of dissimilarity between groups of observations. Agglomerative strategies for hierarchical clustering start at the bottom of the hierarchical structure (the level where each cluster contains a single observation) and at each level recursively merge a selected pair of clusters into a single cluster. This produces a hierarchical structure where each level of the hierarchy represents a particular grouping of the data into disjoint clusters of observations. The indices presented in the previous section can be used to decide which level actually represents a ”natural” clustering in the sense that observations within each of its groups are sufficiently more similar to each other than to observations assigned to different groups at that level [5]. Hierarchical clustering algorithms are widely used in some applications as botany and medical diagnosis because they are extremely easy to comprehend [6]. However, the direct application of hierarchical agglomerative methods to digital images is not viable due to the large number of patterns, and, as a result, the enormous computational effort required. An alternative approach is to use an efficient data-clustering algorithm (for example ISODATA) to establish an initial partition of the image data. The tens (or few hundreds) of clusters of this initial partition can then be easily managed to form a hierarchical clustered structure.

The ISODATA (Iterative Self-Organizing Data Analysis Technique) unsupervised classification method is a modification of the k-means algorithm [7]. Both are iterative processes, but the k-means method requires knowledge of the number of classes present in the data. Initially, k centres are seeded along the diagonal (or in other locations) of the feature space. Each pattern (or pixel, for a digital image) is assigned to the class whose centre is closest, according to a given metric (Euclidian distance, for example). Once all patterns are distributed amongst the classes, an updated centre is computed for each class. The process is repeated until all class centres are stable (up to a threshold value), or the iteration limit is reached. The number of classes produced by the ISODATA classifier can vary, within a pre-established range. In each iteration, two or more classes can be merged, a class can be removed or split in two. These decisions are controlled by a set of parameters, which will naturally influence the final results. In the combined methodology, the clusters produced by the ISODATA classifier are used as the initial observations to form the hierarchical clustered structure for the digital images.

3 Index Performance with Synthetic Data

The performance of the Xu index was initially evaluated with synthetic data. Each test was performed on a set of 100 elements, randomly generated with Gaussian distribution curves. The following parameters were considered: data dimensionality (d), number of Gaussians (n), standard deviation of the Gaussians (σ). The number of classes that should be expected is n, although this will be strongly dependent on the random generation process. Each pattern is a d-dimension vector with components between 0 and 1.

The synthetic data generation followed a similar process to the method used by Dubes [2]. It assures that a minimum number of elements are assigned to each cluster, but allows some variability in the number of elements per cluster. The process starts by randomly establishing the n Gaussian centres, assuring that they are at least 2σ apart from each other, and at least at a distance σ from the feature space edges. The number of patterns (100) is divided in $n + 1$ groups.

Each Gaussian curve is assigned a group, and the elements of the remaining group are randomly assigned to any of the Gaussians.

Each dataset generated was classified in k classes (with $k = 2, 3, \ldots, 12$), using MATLAB algorithm ClusterData [8]. The Xu index was computed for each partition, and a plot of the index versus the number of classes created for each dataset. As an illustration, Figure 1 shows the data and the $E(h)$ plot for a test with $d = 2$, $n = 8$, $\sigma = 0.04$. In this case, the number of classes suggested by the index was 5, instead of the 8 expected. However, a visual inspection of the data plot seems to suggest that the choice of 5 classes is actually a reasonable one.

3.1 Evaluation of the Index Performance

A total of 140 sets of parameters were tested: d = 2, 3, 4, 5; n = 4, 5, 6, 7, 8; σ = 0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.10. For each set of parameters, a total of 200 data sets were produced and evaluated, each with 100 patterns. The number of times that the Xu index plot indicated the expected number of clusters was registered, and the success rate computed. Table 1 shows the success rate (in %) for 24 sets of parameters ($n = 6$). For example, for $d = 3$, $n = 6$, $\sigma = 0.02$, the Xu index selected 6 as the natural number of classes 140 out of 200 times, or 70.0 %. The results presented in Table 1 show that the effectiveness of the index decreases with increasing σ and decreasing d. Although not shown in Table 1, the effectiveness of the index also decreases with an increase of the number of Gaussian curves (n), as expected. It is worth mentioning that for high values of σ, the number of classes selected by the index is very often different than the number of Gaussians used to generate the data, but still a reasonable choice. This is illustrated in the example of Figure 1. This helps explaining the low success rate of the index for high values of σ.

Table 1. Success rate of the Xu index with synthetic data (6 Gaussians), for various values of data dimensionality (d) and standard deviation of the Gaussians (σ)

	$\sigma = 0.01$	$\sigma = 0.02$	$\sigma = 0.03$	$\sigma = 0.04$	$\sigma = 0.05$	$\sigma = 0.07$	$\sigma = 0.10$
$d = 2$	86.5%	57.5%	33.5%	18.5%	11.5%	7.5%	6.5%
$d = 3$	94.0%	70.0%	56.5%	31.5%	27.5%	10.0%	4.0%
$d = 4$	87.0%	78.0%	60.5%	52.0%	36.0%	22.0%	12.5%
$d = 5$	83.5%	69.0%	61.5%	54.0%	45.5%	22.5%	8.0%

4 Results with Image Data

Four test images were selected to evaluate the performance of the Xu similarity index. The images selected are small sections (of 512 by 512 pixels) extracted from Landsat TM images of Portugal and Spain, acquired in October 1997.

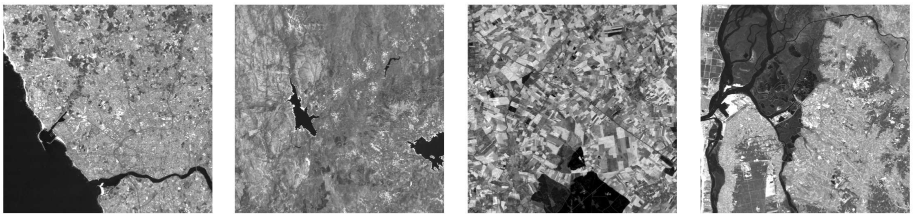

Fig. 2. First principal component of the test images I-Porto, II-Geres, III-Castela, IV-Aveiro (from left to right)

The multi-spectral images have 6 bands, with a 30-meter pixel resolution. The thermal band of Landsat TM was not used due to the lower spatial resolution [9]. The first principal component of each test image is shown in Figure 2, from left to right: I-Porto, II-Geres, III-Castela, IV-Aveiro. The first principal components featured in Figure 2 were only used for displaying purposes. They retained 82.4%, 64.9%, 88.5% and 83.3% of the total variance of the multi-spectral test images I, II, III and IV, respectively.

4.1 Image Classification and Clustering

Each test image was classified using the algorithm ISODATA implemented on the software PCI Geomatics [10]. The same set of parameters was used throughout, including the range of classes allowed (20-40). The classifier converged for a solution with 27 classes for test image III, and with 40 classes for the remaining test images. The classification results were hierarchically structured, using the Euclidian distance metric between the class centres ($\mathbf{m}_k$) as the agglomerative criterion. This produced 39 classified images for test images I, II and IV (with 40, 39, ..., 2 classes), and 26 classified images for test image III.

4.2 Analysis

The Xu similarity index was computed for each classified image, and a plot $E(h)$ produced for each test image. The plots are presented in Figure 3, as a function of the level on the hierarchical structure $-$ the number of classes h. An initial inspection of these plots seems to suggest that the optimum solution, or the natural number of classes, is not always a unique choice.

For test image I, an urban area (the city of Porto), the index has 4 strong maximums for $h = 5$, 9, 11 and 18. In urban areas such as this one, with a pixel size of 30 meters, a great number of mixed pixels should be expected. This can help explaining why there are several possible choices for the "natural" number of classes. The best choice according to the index is for $h = 9$, which was the classified image selected for Figure 4 (left).

Test image II covers a mountainous region (Geres, Portugal), with some bodies of water. In this case the Xu index clearly points to a partition with $h = 4$.

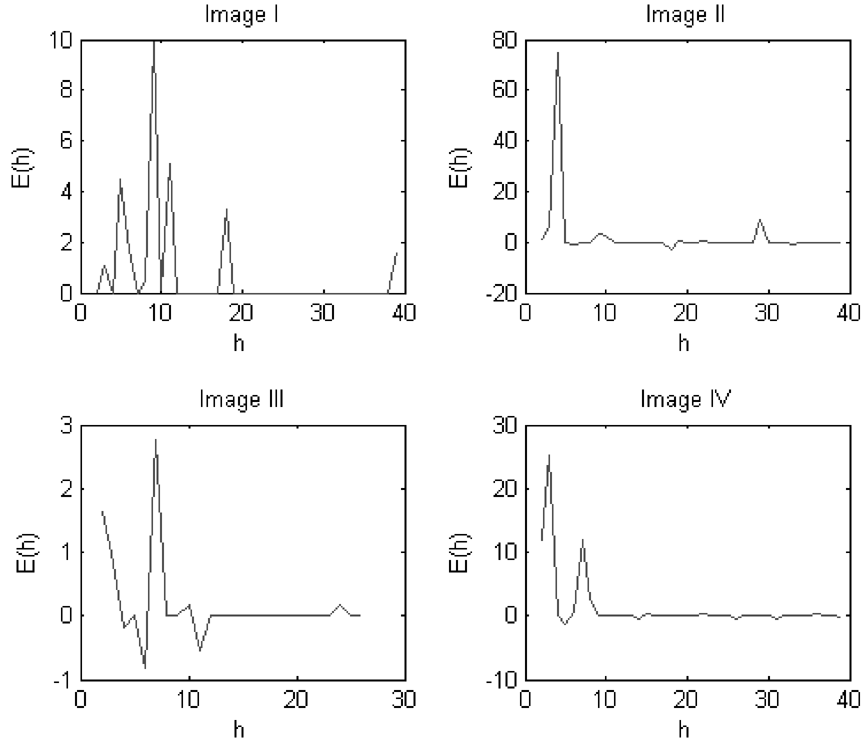

Fig. 3. Xu index plots for test images I(top left), II (top right), III (bottom left), IV (bottom right)

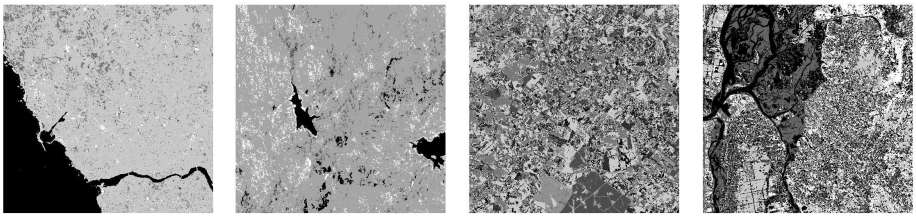

Fig. 4. Classification levels selected for test images I ($h = 9$), II ($h = 4$), III ($h = 7$), IV ($h = 7$) (from left to right)

This is a consistent result, corresponding to four classes with well-distinguished spectral signatures: water, bare soil, sparse and dense vegetation. The magnitude of the index for $h = 4$, compared to the other values of h, suggests that this is the only natural choice for this image, although very subtle local maximum do appear for $h = 9$ and $h = 29$. The result for $h = 4$ is presented in Figure 4.

Test image III covers an agricultural area (in Castela, Spain), with a small urban sector. The index plot seems to indicate a selection of $h = 7$, although the magnitude is in this case rather low. A choice of $h = 3$ could also be done, but the level of discrimination (only 3 classes) is perhaps inappropriate, from a user perspective. The classified image at this level is presented in Figure 4 (3rd from left). There are two well-distinguished classes in a large field in the bottom part of the image. The remaining classes are assigned to smaller fields spread throughout the image.

Test image IV includes a variety of land covers in the Estuary nearby Aveiro, Portugal. There are deep and shallow water, sand, vegetation and urban areas. The plot of $E(h)$ points towards two possible choices: $h = 3$ or $h = 7$. The magnitude of the index is higher for $h = 3$, but from a user perspective, perhaps the partition of the data into 7 classes is a more meaningful one. This later choice is presented in Figure 4 (right).

An additional evaluation of the Xu index adequacy for estimating the number of classes on a multi-spectral satellite image could be done using ground truth data. However, this is a difficult task, as the existing land cover maps (COS90) were produced by air photo interpretation [11]. The land cover maps have much greater spatial detail and diversity of classes than what can be realistically expected from a Landsat TM image. A considerable effort in data generalisation in the existing land cover maps is therefore required in order to make a meaningful comparison between the two datasets.

5 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 often 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 [3], but few are applicable to large data volumes, such as multi-spectral satellite images.

The method tested 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 Xu internal similarity index is used to select the "natural" number of classes from this set of classified images. The final result is a single classified image, although multiple results at multiple levels of the hierarchic structure can also be provided. One aspect that should be taken into account is the fact that the accuracy of the final classified image selected is limited by the initial clustering. Another aspect is that hierarchical methods impose hierarchical structuring whether or not such structure actually exists in the data. The results suggest nevertheless that the method proposed is effective in achieving a coherent result from the data perspective. The results also seemed to be reasonable from an end user point of view, as the number of classes selected were consistent with the diversity of land cover types expected for each test image.

References

1. Jain, A. K., Murty, M. N., Flynn, P. J.: Data clustering: A review. ACM Computing Surveys **31** (1999) 264–323
2. Dubes, R. C.: How many clusters are best - an experiment. Pattern Recognition **20** (1987) 645–663
3. Milligan, G. W., Cooper, M. C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika **50** (1985) 159–179
4. Xu, S., Kamath, M. V., Capson, D. W.: Selection of partitions from a hierarchy. Pattern Recognition Letters **14** (1993) 7–15
5. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: data mining, inference, and prediction. Springer, New York (2001)
6. Ripley, B. D.: Pattern Recognition and Neural Networks. Cambridge University, Cambridge (1996)
7. Tou, J. T., Gonzalez, R. C.: Pattern Recognition Principles. Addison-Wesley (1974)
8. The Math Works: MATLAB The Language of Technical Computing - Using MATLAB : version 6. The Math Works, Inc. (2000)
9. Lillesand, T. M., Kiefer, R. W.: Remote Sensing and Image Interpretation, 4th edition. John Wiley and Sons, New York (2000)
10. PCI Geomatics: X-Pace Reference Manual, Version 8.2. PCI Geomatics, Ontario, Canada (2001)
11. CNIG: Carta de Ocupao do Solo (COS' 90). Centro Nacional de Informao Geogrfica, Lisboa, Portugal (1990)