Anais VIII Simpósio Brasileiro de Sensoriamento Remoto, Salvador, Brasil, 14-19 abril 1996, INPE, p. 851-856.

Unsupervised Fuzzy C-Means Classification for the Determination of Dynamically Homogeneous Areas

SATURNINO LEGUIZAMÓN [1], HENK PELGRUM [2] and SUSANNA AZZALI [2]

[1] CONAE, Gerencia de Gestión Tecnológica
Casilla Correo 131, 5500 Mendoza, Argentina
Fax: +54-61-288565, e-mail: ntcricyt@arcriba.edu.ar

[2] WINAND STARING CENTRE (ICW),
P.O. Box 35, 6700 AA Wageningen, The Netherlands
Fax: +31-8370-24812, Phone : +31-8370-74323

Abstract. Conventional image classification procedures are often inappropiate for the mapping of continuous phenomena like those appearing in remotely sensed images of the Earth surface. Geographical information, included that remotely sensed, is imprecise in nature. For the allowness of natural fuzziness of such an environment, a fuzzy sets algorithm may be used. This report deals with the development of a fuzzy unsupervised procedure which is applied to the classification of a set of NDVI time series images. The goal of this technique is to locate dynamically homogeneous areas. The results show that fuzzy classification can provide a more detailed information than classical "hard" classification. However, more attention is required at the interpretation stage, due to the continuous nature of the classes. We conclude that the method is most promising and worthy of consideration when a mapping of natural resources is necessary.

Keywords: Remote sensing, fuzzy sets, classification

1 Introduction

Classical image classification techniques were designed for application on phenomena which can be considered to exist in discrete classes. Each pixel is simply allocated to the class with which it displays the greatest level of similarity.

However, full membership of the allocated class is not often the case in natural environment. An alternative approach to classification is required in order to provide a more valuable representation of continuous classes. One way to achieve this is by indicating the relative strength of class membership that a case has relative to all defined classes.

To overcome the potential problem of non-normality, a fuzzy sets procedure may be used. This approach, which indicates the strength of membership, makes no assumptions about the statistical distribution of the data. One approach is based on the fuzzy c-means algorithm (FCM). This paper illustrates the representation of continuous classes and its applicability to locate dynamically homogeneous areas in a continental scale. The application of this method to remotely sensed data has shown interesting results to take it into account in any classification process.

This paper is organised as follows: Section 2 discusses the approach of non supervised clustering in discrete and continuous classes, Section 3 explains the fuzzy c-means classification procedure, Section 4 states the problem of determination of homogeneous regions, Section 5 explains the application of the FCM approach to a given set of data. Section 6 discusses the results and Section 7 presents the conclusions.

2 Fuzzy C-Means Classification

In this Section we will focus our attention on unsupervised classification, also called *clustering*. Clustering refers to a broad spectrum of methods which try to subdivide a data set X into c subsets (clusters), which are pairwise disjoint, all nonempty and reproduce X via union (Besdek et al, 1984). This procedure, in which each point in X is allocated to the class with which it has the greatest level of similarity, is termed a "hard" (i.e. nonfuzzy) c-partition of X.

In the most general case, clusters are defined as groups of points that are "similar" according to some measure of similarity. Usually, "similarity" is defined as proximity of the points according to a distance function. Many algorithms based on minimisation of a mathematical clustering criterium are discussed in the literature (Duda and Hart, 1973)

Classes in which each member has full membership are called discontinuous or discrete classes. On the other hand, classes in which each member belongs in some extent to every cluster or partition, are called continuous classes (McBratney, A.B., and J.J. de Gruijter, 1992).

Continuous classes are a generalisation of discontinuous classes where the indicator function of conventional sets theory, with values 0 or 1, is replaced by the membership function of fuzzy sets theory, with values in the range of 0 to 1.

Let us take the partition of a set of n individuals in *c* discontinuous classes. According to such partition, each individual is a member of exactly one class. This can be numerically represented by a n x *c* membership matrix **M** =(m_{ij}), where m_{ij}=1 if the individual belongs to class j, and m=0 otherwise. In order to ensure that the classes are mutually exclusive, jointly exhaustive and non-empty, the following conditions are applied to **M** :

$$\sum_{j=1}^{c} m_{ij} = 1 \qquad i = 1,...,n \tag{1}$$

$$\sum_{i=1}^{n} m_{ij} > 0 \qquad j = 1,...,c \tag{2}$$

$$m_{ij} \in \{0,1\} \qquad i = 1,...,n;\ j = 1,...,c \tag{3}$$

Condition (3) corresponds to the all-or-nothing status of the memberships in discrete classes. According to the fuzzy sets theory, this condition is relaxed in such a way that partial memberships are allowed, i.e. to take any value between and including 0 and 1. Thus, condition (3) for continuous classes becomes:

$$m_{ij} \in [0,1] \qquad i = 1,...,n;\ j = 1,...,c \tag{3a}$$

Any n x *c* matrix **M** satisfying (1), (2) and (3a), represents a so-called fuzzy partition of the n individuals into *c* classes. Partitions satysfying (1), (2) and (3) are referred to as hard partitions.

As condition (3) implies (3a), hard partitions are special cases of fuzzy partitions.

3 Generation of Continuous Classes

A special case of FCM algorithm was first reported by Dunn in 1973. Dunn's algorithm was later generalised by Bezdek (1980), Bezdek et al. (1984) and Kent et al. (1988). These methods use a n x p data matrix **X** = (x_i) as input, where p denotes the number of variables and x_i denotes the value of individual i for variable .

The most popular fuzzy clustering method to date is the fuzzy c-means, which is a generalisation of the hard c-means clustering.

The hard c-means method minimises the functional *J*(**M**,**C**), which represents the within-class sum-of-square errors between classes, under conditions (1), (2) and (3):

$$J(M,C) = \sum_{i=1}^{n} \sum_{j=1}^{c} m_{ij}\, d^2(x_i, c_j) \tag{4}$$

where **C** = (c_j) is a c x p matrix of centre of classes, c_j denotes the value of the centre of the j class for the variable .

$\mathbf{x}_i = (x_{i1},...,x_{ip})^T$ is the vector which represents the individual i, $\mathbf{c}_j = (c_{j1},...,c_{jp})^T$ is the vector which represents the centre of class j, and $d^2(\mathbf{x}_i,\mathbf{c}_j)$ is the square of the distance between $\mathbf{x}_i$ and $\mathbf{c}_j$ according to a given definition of distance, further denoted by d^2_{ij} to simplify.

J(**M**,**C**) is the sum of the square errors of the distance of each individual to the centre of the given classes.

A fuzzy generalisation of *J*(**M**,**C**) is obtained by a modification of the membership with an exponent . This weighting exponent controls the extent of membership sharing, or the "degree of fuzziness", among the resulting clusters.

Hence, a new J_F(**M**,**C**) function is defined as:

$$J_F(M,C) = \sum_{i=1}^{n} \sum_{j=1}^{c} m_{ij}^{\varphi} d_{ij} \qquad (5)$$

This function is minimised under the conditions (1), (2) and (3a).

The value of φ is chosen from $(0, \infty)$. If $=1$, the solution of (5) is a hard partition, i.e. the result is not fuzzy. As approaches the solution approaches its maximum degree of fuzziness, with $m_{ij}=1/c$ for every pair of i and j. There is no theoretical basis for an optimal selection of φ . It is often chosen on empirical grounds to be equal to 2.

When >1, (5) could be minimised by Picard iteration of the following equations (Bezdek, 1984):

$$m_{ij} = \frac{d_{ij}^{-2(\varphi-1)}}{\sum_{j=1}^{c} d_{ij}^{-2(\varphi-1)}} \quad i=1,...,n; \; j=1,...,c \qquad (6)$$

$$c_j = \frac{\sum_{i=1}^{n} m_{ij}^{\varphi} x_i}{\sum_{i=1}^{n} m_{ij}^{\varphi}} \quad j=1,...,c \qquad (7)$$

According to the previous considerations, the following fuzzy c-means algorithm can be developed:

Fuzzy c-means algorithm:

1) choose the number of classes c, with $1 < c < n$.
2) choose a value for the fuzziness exponent φ, with $\varphi > 1$;
3) choose a definition of distance in the variable space.
4) choose a value for the stopping criterium .
5) initialise $M = M_0$, e.g. with random memberships or with memberships from a hard c-means partition.
6) in the iteration l = 1, 2, 3 ,... re-calculate $C=C_l$ using equation (7) and M_{l-1};
7) re-calculate $M = M_l$ using equation (6) and C_l; compare M_l to M_{l-1} in a convenient matrix norm. if $M_l - M_{l-1}$, then stop; otherwise return to step 6).

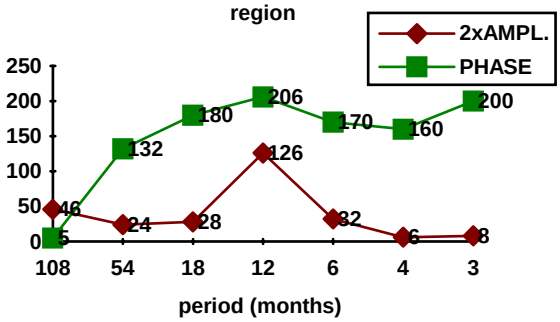
In our investigation we used the exponent φ =2, the Euclidean distance and the initialisation values for the matrix M have been taken from those resulting from a "hard" c-means classification.

4 Dynamically Homogeneous Areas

For many years Menenti et al. (1991, 1993) have been carrying out a study of natural and cultivated areas in South America and Africa using the information provided by Fourier analysis of Normalized Difference Vegetation Index (NDVI) time series. Fourier analysis provides a measure of dynamic behaviour of vegetated regions. The primary objective of that study was to examine and map vegetation-soil-climate units in a continental basis.

Fig. 1 shows Fourier values of amplitude and phase (normalised to 0-255 range) corresponding to an arbitrary region (or point) in South America and considering only 7 harmonics with periods of 3, 4, 6, 12, 18, 54 and 108 months.

Fig. 1.- Amplitude and phase in an arbitrary region



Since Fourier analysis is applied on a pixel by pixel basis, the results of that process are maps of amplitude and time-lag for each period. Some of this maps could be selected to be considered as attributes in a classification analysis.

Geographical regions that exhibit a similar dynamic behaviour can be termed "homogeneous" regions or "iso-growth" zones. The objective of our research is to use Fourier data (images) to locate, by unsupervised fuzzy classification, those geographical areas that exhibit a similar behaviour.

Given the resultant image of a classification analysis, pixels belonging to a given "class" are pixels belonging to a specific homogenoeus area.

Thus, the problem of determination of homogeneous areas becomes a classification problem.

The Fourier transform of the NDVI time series gives a large number of possible choices of atributes in a classification process. Since amplitude and phase values can be combined in different ways, different results can be obtained in that kind of operation. Many comparisons of classification results are necessary to understand which attributes (amplitude, phase, and frequency) are the most appropiate, taking into account the differences between ecosystems.

In our research we consider the result of a discrete FFT (Fast Fourier Transform) applied to 9 years of monthly NDVI time series (years 1982-1991). Therefore, 54 images of amplitude and phase can be obtained. Each harmonic in the Fourier analysis is a component of the "grow cycle" having a precisely defined period. In a given region the amplitude values of different harmonics measure the "variability of vegetation growth", and phase (time-lag) values are an exact measurement of "earliness" or "lateness" of the vegetation growth.

To compare the results of "hard" and FCM classification, 6 arbitrarily selected features were considered: four amplitude components corresponding to 9, 4.5, 1 and 0.5 years, and two phase (time-lag) components corresponding to 1 and 0.5 year.

The geographic region regarded in this experience was a sector of South America. It was determined by data availability, computer memory restrictions and areas already covered by previous studies for comparisons (Menenti et al., 1991, 1993). The number of "classes", or homogeneous areas, was arbitrarily chosen as 6.

Figure 2 shows the result of applying a hard classification method (Isodata) to the set of data mentioned above. Figures 3 and 4 show the results of applying the developed FCM algorithm to the same set of data.

6 Analysis of the Results

Given the Fourier NDVI time series images as an input for an unsupervised classification process, different results are obtained depending on the used approach. Each class define a homogeneous region. In the "hard" classification approach, pixels belonging to a given class are pixels belonging to a specific homogeneous area. In the case of FCM classification process, every pixel belongs to a given class in an extent which is determined by its membership value.

In Fig. 2, which is the result of a hard classification, the 6 classes are clearly identified by their uniform colour on the corresponding area. On the other hand, in Fig. 3 and Fig. 4, which are the resultant images of a FCM classification, classes are not clearly separated and exhibit a more "natural" appearence. This is because each pixel takes a portion of red (R), green (G) and blue (B) colour which is proportional to the degree of membership in one of the 3 classes taken into account.

This result shows that FCM can provide a more detailed information than classical "hard" classification, however this feature entails a problem for the interpretation of the results. One important observation of the resultant image is that the homogeneous regions determined by the FCM approach have a good spatial correlation with the geographical provinces of the South American continent.

Fuzzy classification allows us to obtain a visual and qualitative interpretation of the results as well as a quantitative one, as long as we are able to analyse the membership values of each element of matrix M in the fuzzy process.

The reachness of information provided by fuzzy classification deserves to pay special attention to further interpretation of the biodynamic, climatic, as well as geographical influences.

7 Conclusions

The objective of our study is the determination of dynamically homogeneous vegetated areas, at continental scale, using the Fourier parameters of NDVI time series collected during 9 years.

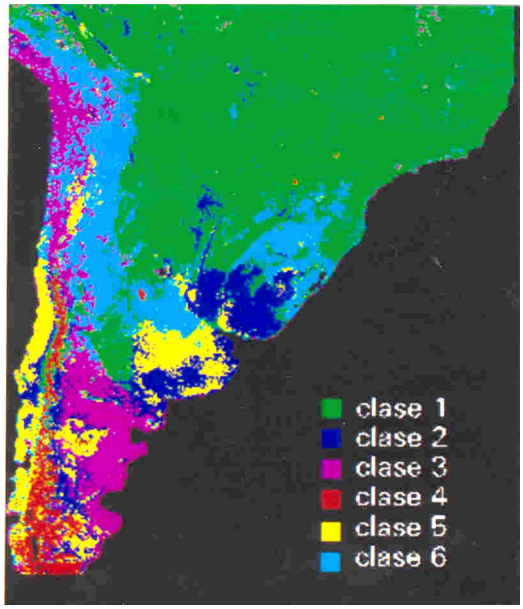


Fig. 2.- Resulting map of a hard classification (Isodata) with 6 "homogeneous areas"in a region of South America

This paper proposes to treat the problem as an unsupervised classification problem by making use of fuzzy set concepts. The unsupervised procedure we have used here is that of the fuzzy c-means algorithm (FCM).

This technique has been applied for the unsupervised classification of NDVI images in a sector of South America.

The results of the process are quantitatively compared to those obtained by applying one of the most popular unsupervised methods: ISODATA. The observation of the results shows that FCM can provide a more detailed information than classical "hard" classification. This property, however, demands more attention at the interpretation stage.

Another important advantage of the FCM approach is that the number of classes to be defined in an unsupervised process can be less than those necessary in hard classification, since "fuzziness" includes many of the hard classes. The excesive number of classes sometimes used in hard classification, is to overcome its limitations.
Fig. 3.- Map of iso-growth regions in South America showing classes 1, 2 and 3 of 6 classes.

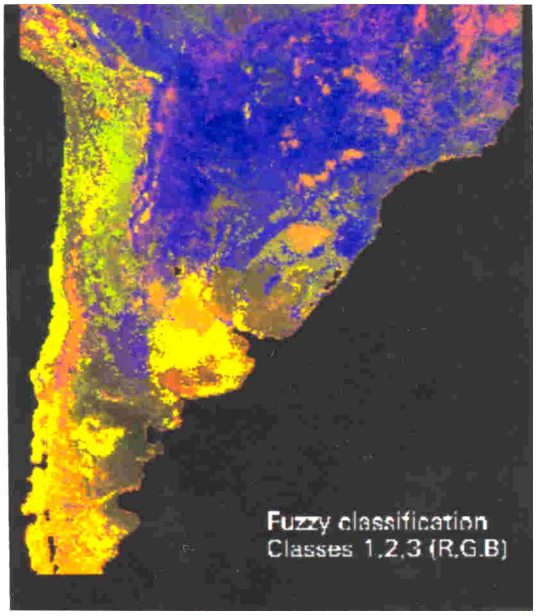


Fig. 3.- Map of iso-growth regions in South America showing classes 1, 2 and 3 of 6 classes

Fig. 4.- Map of iso-growth regions in South America showing classes 4, 5 and 6 of 6 classes.

We conclude that the method is most promising and worthy of consideration when any type of mapping of natural resources is necessary.

We are aware that fuzzy sets contribute to the field of pattern recognition expanding the possibilities of interpretation of the results and giving more chances to a multidisciplinary work. It allows the development of methodological basis for pattern description and classification.

Acknowledgements

This investigation has been sponsored by the Netherlands Remote Sensing Board (BCRS). The authors are indebted to Dr. Massimo Menenti, of the ICW, for his support and valuable suggestions.

References

Bezdek, J.C., 1980. "*A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms*", IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-2, no.1, pp.1-8.

Bezdek, J.C., Ehrlich, R. and Full, W., 1984. "*FCM: the fuzzy c-means clustering algorithm*", Computer & Geosciences, vol 10, 191-203.

Duda, R.O. and P. Hart, 1973. "Pattern Classification and Scene Analysis", John Wiley & Sons.

Dunn, J.C., 1973. "*A fuzzy relative of the ISODATA process and its Use in Detecting Compact, Well Separated Clusters*", J. Cybern, vol. 3, pp. 32-57.

Kent, J.T. and K.V. Mardia, 1988. "*Spatial Classification Using Fuzzy memberships Models*", IEEE Trans. on PAMI, vol. 10, No. 5, September, pp. 659-671.

McBratney, A.B., and J.J. de Gruijter, 1992. "*A Continuoum Approach to Soil Classification by Modified Fuzzy k-Means with Extragrades*", Journal of Soil Science, vol. 43, pp. 159-175.

Menenti, M., S. Azzali, W. Verhoef, R. van Swol, 1991. "*Mapping agroecological zones and time lag in vegetation growth by means of Fourier analysis of time series of NDVI images*", MARS Report 32, DLO The Winand Staring Centre, Wageningen, The Netherlands.

Menenti, M., S. Azzali, A. de Vries, D. Fuller and S. Prince, 1993. "*Vegetation Monitoring in Southern Africa using Temporal Fourier Analysis of AVHHR/NDVI Observations*", Proc. of Intern. Symposium on Remote Sensing in Arid & Semi-Arid Regions, Lanzhou, P R. of China.

About the Authors

Saturnino Leguizamon, received his diploma of Electrical and Electronic Engineer from Universidad de Mendoza, Argentina, in 1967, and obtained his MSc. degree in Engineering Sciences from Universidad de Chile in 1972. His research interest is in the computer analysis of remotely sensed images. For more than 14 years he has been envolved in remote sensing activities by taking part in many projects regarding Earth sciences, its resources and environment. He is also one of the founder members of SELPER in Argentina.

Henk Pelgrum, received his diploma from the Agricultural University, Wageningen, Netherlands in 1993. From 1993 until now he has been working at the Winand Staring Centre for integrated land and water research, Wageningen. His research interest is oriented toward the use of remote sensing for estimation of the surface energy balance. Currently he is doing his PhD research on improving numerical weather prediction models by means of using satellite data as input for the models

Susanna Azzali is an agronomist and remote sensing specialist. Since 1986 she has been working at DLO-Winand Staring Centre (SC-DLO), NL, acquiring a considerable experience in agricultural landuse inventories and monitoring using satellite remote sensing (NOAA-AVHRR, Landsat and TM, SPOT). Thanks to her specialization both in agriculture and irrigation, she carried out several agro-ecological, environmental and irrigation studies using GIS and remote sensing techniques.