CLASSIFICATION AND FUZZY SETS

Foody, G. M.
Department of Geography, University of Southampton, UK

Keywords: Classification, fuzzy and crisp sets, uncertainty

Contents

Summary

This contribution focuses on issues connected with classification and fuzzy sets in geoinformatics. It reviews the use of classification as a means of summarising information, particularly with regard to supervised classifications that are commonly used in geoscience and remote sensing applications. The procedure for class allocation in widely used approaches to classification, such as those based on maximum likelihood, evidential reasoning and feedforward neural networks, is discussed. Some of the key limitations to the conventional approaches to classification are discussed with particular regard to issues connected with the use of hard or crisp sets, especially when uncertainty prevails. The concept of fuzzy sets is briefly discussed before illustrating the potential to derive fuzzy classifications that may sometimes be more appropriate than conventional hard classifications in geoinformatics.

1. Introduction

Classification is an attractive and useful means of describing and summarising information. It is a natural response to large volumes of information as the simplification achieved by the classification aids communication and analysis. It is not surprising, therefore, that classification has an important role to play in everyday life as well as in science and technology, particularly as a means of summarising or generalising information.

In essence, classification may be considered to be the process by which cases of some phenomenon are grouped together into a set of classes on the basis of their relative similarity with regard to their properties. Thus, for example, a soil may be allocated to a class (e.g., podsol, brown earth *etc*.) on the basis of properties such as its texture, colour, organic matter content and structure. There are many approaches to, and methods of, classification. The classification may, for instance, be hierarchical with groups of differing size encountered at different levels.

Thus, for example, within a hierarchical classification of vegetation, a group representing forests would reside at a level above that comprising deciduous, coniferous and mixed forests. This type of classification is commonly encountered and may be achieved by agglomerative methods in which the analysis begins by grouping together similar individuals before ultimately combining similar groups into larger groups until, potentially, one (super) group contains all the individuals, or by divisive methods, in which one large group is successively sub-divided into smaller groups and ultimately the individuals separated. In this chapter, however, the focus is on classification methods widely used in the theme of geoinformatics and particularly in geoscience and remote sensing applications.

Classification methods are widely used in geoinformatics and are a topic of considerable current research. In particular, the last decade has seen a considerable increase in interest in classifications that may be considered to be fuzzy as this can provide a more useful and meaningful summary of a data set than traditional (crisp) classification. Further details on this issue will be provided after a brief review of traditional classification.

2. Major approaches to classification

The classifications that are generally, but by no means only, encountered in geoscience and remote sensing applications are non-hierarchical and are either supervised or unsupervised. This section aims to provide a brief overview of these methods and focuses, in particular, on those of considerable current interest.

2.1 Unsupervised classification

An unsupervised classification groups together cases with similar attributes. With this approach to classification, cases of the phenomenon of interest are acquired and grouped together according to their relative similarity. The key issue to note, however, is that the classes have not been defined in advance. The classification will simply group together cases that are similar and there is no guarantee that the groupings derived will relate to classes of interest. This type of classification is, therefore, most attractive as an exploratory analysis, particularly when there is little prior knowledge available. An unsupervised classification may be achieved using a variety of clustering algorithms. Commonly, conventional statistical algorithms are used. The *k*-means clustering algorithm and (its relative) the ISODATA algorithm are, for example, popular in remote sensing applications.

When the underlying assumptions of such techniques have not been satisfied, attention has often turned to alternative approaches with techniques such as Kohonen's self organising feature map (SOFM) neural network attracting increasing attention. Neural computing methods have become popular in geoinformatics as, in general, they offer a powerful yet assumption-free means of data analysis. A major attraction of the SOFM for unsupervised classification, in addition to the general advantages of neural computing over traditional methods, is that the network's output, typically presented in a two-dimensional space, not only clusters cases by similarity but also highlights the relative similarity of the cases and clusters. Consequently, the SOFM can not only classify a data set but can also be used to arrange the cases in a manner that reflects their similarity analogous to an ordination analysis that is widely used in community ecology.

Although unsupervised classifications are useful and widely used in the geosciences, particularly as part of exploratory data analysis, there are significant problems with their use. Of particular concern is that the clusters formed may not relate to informative or useful classes. Thus, for example, a data set containing a wealth of variables on soil properties acquired in the field may be input to an unsupervised classification and a set of classes identified in each of which the soils have similar properties in terms of say B horizon colour, depth and pH. The derived classes may be of little interest, however, if the classification was to aid an afforestation programme in which soil drainage was the major concern. Although the classes defined may have some relation to soil drainage (e.g., the colour may indicate waterlogging *etc*.), a classification in which the soils had been grouped according to variables that have a marked and direct influence on soil hydrology such as texture, structure and stoniness may be more useful. Consequently, it is often more appropriate to attempt to group the cases into predefined classes of interest. This may be achieved through the application of a supervised classification.

2.2 Supervised classification

A major difference between an unsupervised and supervised classification is that with a supervised classification the classes are defined at the outset. Thus, at the outset of a supervised classification the properties of each class are defined in terms of a set of variables. The measured values of these variables for a case of unknown class membership are then compared against those of the classes and each case allocated to the class with which it has the greatest similarity. This type of classification is very common, especially in remote sensing applications as a tool for thematic mapping applications. This application of a classification aims essentially to convert the remotely sensed image, depicting typically the spatial distribution of Earth surface reflectivity in a number of spectral wavebands, into a thematic map such as one depicting the spatial distribution of land cover classes. In this context, the classification is generally applied on a per-pixel basis and has three distinct stages. First, the training stage, in which pixels of known class membership in the remotely sensed image are characterised and class 'signatures' derived. These training statistics describe the typical remotely sensed response of each class. In the second stage, these training statistics are used to allocate pixels of unknown class membership in the image to a class in accordance to some decision rule. This is achieved by comparing the remotely sensed response observed for that pixel with the signature of each class defined in the training stage. Third, the quality of the classification is evaluated. This is generally based on the accuracy of the classification that is assessed by comparing the actual and predicted class of membership for a set of pixels not used in training the classification.

-
-
-

Bibliography

Belward, A. S. (this Theme). Developments in global land cover mapping. [Provides a discussion of classification techniques for thematic mapping applications].

Bezdek, J. C. (1993). Fuzzy models - what are they, and why? *IEEE Transactions on Fuzzy Systems* **1**, 1-6 [Provides an overview of fuzzy and crisp sets, and explains the distinction between probabilities and fuzzy membership values].

Bezdek, J. C., Ehrlich, R. and Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers and Geosciences* **10**, 191-203 [Provides a discussion on fuzzy classification and presents the FCM algorithm].

Bosserman, R. W. and Ragade, R. K. (1982). Ecosystem analysis using fuzzy set theory. *Ecological Modelling* **16**, 191-208 [Provides an introduction to fuzzy sets for the study of complex ecological systems].

Fisher, P. (2000). Sorites paradox and vague geographies. *Fuzzy Sets and Systems* **113**, 7-18 [Discusses the vagueness that is an inherent property of many geographical phenomena].

Foody, G. M. (1999). The continuum of classification fuzziness in thematic mapping. *Photogrammetric Engineering and Remote Sensing* **65**, 443-451. [This discusses the supervised classification of remotely sensed data and illustrates classifications of varying degrees of fuzziness derived with a neural network].

Grunsky, E. C. (this Topic). The use of multivariate statistical methods for the evaluation of geochemical survey data. [Provides a review of multivariate statistical analyses, including cluster analysis, used in a geoscience context].

Klir, G. and Folger, T. A. (1988). *Fuzzy Sets, Uncertainty, and Information*, London: Prentice-Hall, International. [This provides a comprehensive introduction to fuzzy sets and uncertainty as well as the fundamentals of evidential reasoning].

Krishnapuram, R. and Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems* **1**, 98-93 [Provides a discussion of fuzzy clustering and introduces the PCM algorithm].
Openshaw, S. (1996). Fuzzy logic as a new scientific paradigm for doing geography. *Environment and Planning A* **28**, 761-768 [Provides a simple overview of the argument for adopting of fuzzy sets and fuzzy logic in the study of complex systems].

Schalkoff, R. J. (1992). Pattern Recognition: Statistical, Structural and Neural Approaches, New York: Wiley. [Provides an introduction to pattern recognition that encompasses supervised and unsupervised classification by various approaches including neural networks].

Schowengerdt, R. A. (1997). *Remote Sensing: Models and Methods for Image Processing*, second edition, San Diego: Academic Press. [Provides a discussion of supervised and unsupervised classification using both crisp and fuzzy sets in a remote sensing context].

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control* **8**, 338-353 [The highly influential article that introduced fuzzy sets].

Biographical Sketch

Giles Foody is Professor of Geography in the School of Geography at the University of Southampton, UK. His main research interests are in environmental remote sensing.