THE VARIOUS METHODS OF CLASSIFICATION FOR REMOTE SENSING DATA

S.R.Ibrahimova

Azerbaijan National Aerospace Agency, Baku, Azerbaijan

ABSTRACT

Data classification is one of the primary tasks in Geocomputation. Traditionally, data classification tasks are based on statistical methodologies such as minimum distance-to-mean (MDM), maximum likelihood classification (MLC) and linear discrimination analysis (LDA). These classifiers have developed over the last century from the mathematical disciplines of set theory and control theory. Over the last 20 years, classification tools have also developed from the emerging fields of connectionism and inference nets within the discipline of Artificial Intelligence (AI); the more notable being the neural network based multi-layered perceptron (MLP), the decision tree and genetic algorithms (GA) such as differential evolution (DE). This paper seeks to compare the advantages and disadvantages of the various classifier types. The results show that, for simple tasks, MDM, LDA and similar classifiers are the best compromise of efficiency and classification ability, whilst for more complex datasets, variants based on the MLP and decision trees are the classifiers of choice.

Keywords: various methods, similar classifiers, neural network

1. INTRODUCTION

Within the Earth Sciences, classification is the process of identifying areas of the Earth's surface, given a particular phenomenological output domain and some different input domain (a set of attributes).

Classification of raw datasets is an important step in the analysis and understanding of geographical features and their relationships. Such data can be remotely sensed, gathered from ground surveys, or even culled from some previous classification.

A classifier's function can be formulated in terms of a mapping of its input variables to its (given) output conditions. We can write:

$$\Re^p \xrightarrow{\Gamma(n)} \Pi^q, \qquad (1)$$

where p is the number of attributes, q is the number of classes and n is the number of samples. The goal of classification is to select an output class from a different phenomenological domain (Π, the classification scheme) to that of the input attributes ($\Re$) for each input vector x^p. These transformation models can be categorized as unsupervised or supervised classifiers. In the case of supervised classification, the user chooses the scheme Π and the classifier learns an approximation $\Gamma'(n, p)$ to the ideal transfer function $\Gamma(n, p)$. This is accomplished by examining a small set (the training set) of the data for which the correct classification has already been determined (by ground survey, or a previous classification). Hence the (common) scheme of ground cover type is often derived from an attribute domain that may comprise several bands of LANDSAT data, as well as ancillary data such as digital elevation models, rainfall, etc.

II. MAIN TEXT

Traditionally, classification of geographic datasets has been based on well-known statistical methods, as implemented in classifiers such as Maximum Likelihood Classification (MLC) or Minimum Distance to Mean [8]. More recently, the discipline of Computer Science has developed classifiers based on machine learning techniques such as decision trees and artificial neural networks. This has lead to a greater choice of classification techniques available to the Earth Scientist. In this paper, we will examine the relative strengths and weaknesses of these various supervised classifiers, giving some comparative results.

Bayesian Classifiers: MLC and MDM. Many statistical classifiers are based on some approximation to the ideal Bayesian classifier, as in most practical applications the optimal Bayesian classifier can never actually be realized. The popular MLC and MDM classifiers are examples of such approximations derived from the following theory of Bayesian probability.

For acceptable classification, a classifier must contain sufficient complexity to enable encoding of the approximated transformation function $\Gamma'(n, p)$ of (1). Bayesian estimation is a process of determining the probable outcome of an event (the *a* posteriori probability) given some new piece of evidence and the original (*a* priori) probability of that outcome. The Bayes Theorem can be restated in terms of classification of data as [3]:

$$\rho(i|x) = \frac{\rho(x|i)\rho(i)}{\rho(x)} \qquad i = 1,...q, \qquad (2)$$

where q - the number of classes, $\rho(i|x)$- the probability of class i given the input vector x, $\rho(x|i)$- the probability of an input vector with characteristics of x given class i, $\rho(i)$ - the probability that class i is present in the dataset, $\rho(x)$- the probability of an input vector with characteristics of x given any class.

Intuitively, to assign a class membership for a given x, we would calculate $\rho(i|x)$ for all classes and assign x to that class i for which $\rho(i|x)$ is a maximum. However for real-world data, it is generally the case that the prerequisite $\rho(x|i)$ is not known and is therefore estimated from the training set as a probability density function (pdf). The specific form of the pdf used for this estimation of $\rho(x|i)$ defines the type of approximation model. The pdf is used as a discriminate rule to identify a given vector x as belonging to a particular class *i*.

If we make the assumption that the cost of misclassifying class i as class j is the same for all i and j , we can rewrite (2) as:

$$x \in i \quad if$$
$$\rho(x|i)\rho(i) > \rho(x|j)\rho(j), \quad for \quad all \quad j \neq i. \qquad (3)$$

This can be alternatively expressed, by taking the logarithm of both sides of the inequality, as:

$$x \in i \quad if$$
$$\ln \rho(x|i) + \ln \rho(i) > \ln \rho(x|j) + \ln \rho(j), \quad for \quad all \quad j \neq i. \qquad (4)$$

and for the normal case where the priors are unknown, or assumed equal, reduces to:

$$x \in i \quad if$$
$$\ln \rho(x|i) > \ln \rho(x|j), \quad for \quad all \quad j \neq i. \qquad (5)$$

The Maximum Likelihood discriminate rule uses (5). The term "maximum likelihood" may be seen as more appropriate if we rewrite (5) as:

$$x \in i \quad if \qquad \ln \rho(x|i) = \max_j (\ln \rho(x|j)). \qquad (6)$$

Consider the case where $\rho(x|i)$ is estimated as a multivariate normal (Gaussian) distribution:

$$\rho(x|i) \cong 2\pi^{-p/2} |\Sigma_i|^{-1/2} e^{\{-f1\}}, \qquad (7)$$

where p is the dimensionality of the input vector x, with Σ_i and u_i ui the sample covariance matrix and sample mean vector, respectively, for class *i*. This assumption of normality underlies the Maximum Likelihood classifier (MLC, more correctly the Gaussian-based Maximum Likelihood Classifier, see [7], the most common statistical classifier for GIS/RS datasets [9]. Substituting (7) into (6) and tidying terms gives the MLC discrimination rule:

$$d_i(x) = -\ln|\Sigma_i| - (x - u_i)\Sigma_i^{-1}(x - u_i) \ , \qquad (8)$$

From (8) it is easier to see how the methodology implicitly sets a lower limit to the sample class size for each *i*. To ensure that the inverse of Σ_i remains non-singular, the number of representative patterns in each class *i*. In practice, it is recommended to maintain size(i) > 10p, for all *i*, so as to provide a minimal set of construction points in each dimension for the Gaussian curves. The Minimum-Distance-to-Mean (MDM) classifier simplifies the discrimination rule of (8) by dropping the covariance term Σ_i and implementing a simpler Euclidean distance-to-mean metric to give a discriminate function:

$$d_i(x) = -(x - u_i)^T (x - u_i) \ . \qquad (9)$$

This produces spheroid decision boundaries in (Euclidean) feature space, rather than the ellipsoid boundaries of the MLC.

Linear Discriminate Analysis. The preceding Bayesian approximation functions are quadratic in nature, as can be seen by examining (8) and (9). They model a volume – that of the particular class distribution. Linear discrimination, as the name suggests, looks for linear combinations of the input variables that can provide an adequate separation for the given classes. Rather than look for a particular parametric form of distribution, LDA uses an empirical approach to define linear decision planes in the attribute space i.e. it models a surface. The discriminate functions used by LDA are built up as a linear combination of the variables that seek to somehow maximize the differences between the classes:

$$z = a_1 x_1 + a_2 x_2 + ... + a_p x_p = a'x \ . \qquad (10)$$

The problem then reduces to finding a suitable vector a. There are several popular variations of this idea, one of the most successful being the Fisher pair wise Linear Discriminate Rule1.

Fisher's Rule is considered a 'sensible' classification, in the sense that it is intuitively appealing. It makes use of the fact that distributions that have a greater variance between their classes than within each class, should be easier to separate. Therefore, it searches for a linear function in the attribute space that maximizes the ratio of the between-group sum-of-squares (B) to the within-group sum-of-squares $(W)^2$. This can be achieved by maximizing the ratio

$$\frac{a'Ba}{a'Wa}, \qquad (11)$$

and it turns out that the vector that maximizes this ratio, a, is the eigenvector corresponding to the largest eigenvalue of $W^{1}B$ i.e. the linear discriminate function z is equivalent to the first canonical variety. Hence the discriminate rule can be written as:

$$x \in i \;\; if \;\; \left|a^T x - a^T u_i\right| < \left|a^T x - a^T u_j\right|, \quad for \;\; all \;\; j \neq i \qquad . \quad (12)$$

The standard LDA can only form linear decision surfaces, although there is no restriction on the orientation of these in the feature space. In the case where the class distributions are unknown, or we have reason to believe they are not normally distributed, we can expect more satisfactory results than the MLC or MDM methods, as it is unconstrained by any prior statistical model.

Decision Trees. Decision trees have evolved from both a statistical consideration [5] and from development in the field of AI [8]. Decision trees are an example of inductive learning and, as such, implement a rule-based classifier. They involve a recursive partitioning of the feature space, based on a set of rules that are learned by an analysis of the training set. A tree structure is developed where, at each branching, a specific decision rule is implemented, which may involve one or more combinations of the attribute inputs. A new input vector then "travels" from the root node down through successive branches until it is placed in a specific class. In essence then, the classification is determined by describing the path from the root node of the tree to a leaf node - each nodal set of rules progressively refining the classification in a hierarchical manner. The tree encodes high levels of complexity where necessary and more simplistic rule combinations when appropriate, so that the tree only becomes complex (deep) where class separation is difficult. Likewise, only attributes that appear to aid the classification problem are considered when rules are defined; other attributes are simply ignored.

The thresholds used for each nodal decision are chosen using minimum entropy or minimum error measures. It is based on using the minimum number of bits to describe each decision at a node in the tree based on the frequency of each class at the node. Alternatively, some minimum error function based on statistics or algebraic distance can be used, although this is not popular in decision trees. This threshold is set by the user, again by experimentation. At some stage the process must be terminated and the criterion used to determine when a class is adequately described has been the subject of much research. With minimum entropy, the stopping criterion is based on the amount of information gained by a rule (the gain ratio).

Artificial Neural Networks – the MLP. MLP's have become increasingly popular as classification tools in a number of fields. They are highly parametric, in the sense that they must be fitted with a large parameter set (their weights and biases) but, similar to decision trees, they do not depend on knowledge of the class distributions, as they use an inductive, data-driven approach to modeling class discrimination. Like the LDA approach, they model decision surfaces, not class distribution volumes. Their main perceived drawback is the complexity and time involved in choosing and setting up the initial network.

MLP's consist of a number of layers of computationally simple units (nodes), that process their input via a non-linear activation function. The layers are attached to each other by a set of plastic weighted connections. The learning phase is devoted to varying the weights in such a way as to produce a classification with minimal error. The error is calculated by implementing some error function E.

This is minimized via some routine or algorithm S. The cost function is generally of the form:

$$E = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(t_j - z_j\right)^2 , \qquad (13)$$

where t_j represents the output at output node j and z_j represents the expected output at that node (given by the training set). S is generally some common numerical minimisation routine, such as gradient descent or conjugate gradient descent.

The use of a cost function such as (13) places a constraint on the MLP classifier that is often overlooked. This "least-squares" form of E assumes a normally distributed noise component within the data, so it is not strictly true to say that the MLP works independently of any distribution trends within the data. This can be alleviated by the implementation of a cost function of a different form, but many of the more effective minimization routines assume such a noise distribution and are hence ineffective unless E is of a similar form to (13). Despite this, MLP's are adept at producing acceptable classification schemas where the class distributions are unknown, sample sizes are small, or there is a high level of noise in the data. Their technique for classification can be viewed in terms of decision surfaces within the attribute space, as can both the LDA and decision tree classifiers. Each decision surface is formed/controlled by each hidden-layer node (and its associated input weights).

However, unlike the decision tree, there is no constraint on the orientation of the surfaces and unlike the LDA classifier, there is no constraint on the number of surfaces or in the way they can be superimposed to produce complex, piecewise linear boundaries.

A Comparison of the Techniques. Comparisons of classification tools are problematic, as there are any number of ways to set-up a given classifier.

Performance of these classifiers can be measured in terms of their: Learning ability; Generalisation ability; Speed.

The first criterion can be determined from measuring the performance on the training set. The second can be determined from measurement on some validation set, which is statistically independent of the training set. The

speed is simply measured as the time taken to converge to the figures given by the second criterion.

The datasets used are described in table 1. Dataset 1 comprises of LANDSAT TM data only. The output classes are well defined crop types and form large, homogeneous regions within the image.

Dataset 2 comprises of LANDSAT TM data plus 7 ancillary layers of data, including digital elevation, geology and flow accumulation. The output classes represent the dominant vegetation cover and, unlike dataset 1, the training sites do not represent contiguous regions in larger target objects, but are instead isolated and 'random' samples (pixels). As such, it provides a much "harder" classification problem and none of the classifiers can be expected to produce a high level of accuracy on this dataset. Table 1 also shows the number of samples in the smallest class within the training sets.

Table 1. Datasets used for these comparisons

Data set	Attributes	Classes	Samples	Min/ class
1	6 – Landsat TM imagery	8 – crop cover	3630	165
2	11 – 4 Landsat TM +ancillary data	9 – floristic classification	1160	53

Table 2 lists the performance on dataset 1, whilst the more complex dataset 2 is used in table 3. Rather than simply adding up the total number of correctly classified samples, all % figures are calculated as the averaged performance over every class in the dataset, thereby removing any bias associated with varying class sample sizes (remembering that these are real-world datasets).

Table 2

Classifier performance on dataset 1

Classifier	Training Set (%)	Validation Set (%)	Time (min:sec)
MDM	53.55	51.25	0:25
MLC	70.65	69.70	1:45
LDA	68.75	65.15	0:35
Decision Tree	73.35	70.05	0:15
MLP	71.60	70.30	3:20

As the tables show, there is quite a difference in performance across the various types of classifiers and datasets. The decision tree produces a useful combination of speed and classification ability. It's ability to generalize is somewhat hampered by the restriction of orthogonal decision surfaces, but the computational complexity of growing the tree is only of the order of O(n). It performs equally well on both the simple and more complex datasets.

Table 3. Classifier performance on dataset 2

Classifier	Training Set (%)	Validation Set (%)	Time (min:sec)
MDM	38.55	37.25	1:30
MLC	46.50	41.00	2:45
LDA	48.05	43.15	0:45
Decision Tree	65.30	52.35	0:35
MLP	70.75	63.10	6:10

Of the statistical methodologies, the MLC gives the best performance on datasets where the assumption of normality can be said to be reasonable. However, in dataset 2, where some of the class samples are quite sparse (remembering that the MLC would require a minimum of 110 samples per class for this dataset) and where much of the ancillary data is either multi-modal or severely skewed, it starts to show it's shortcomings and the empirical approach of statistical classifiers such as LDA can actually outperform it. In these cases, the LDA classifier is more adept at generalization than either of the Bayesian-based classifiers and also gives a useful speed improvement. The LDA is restricted by the number of decision surfaces it can generate (as it can only generate so many covariance matrices) and the positioning of these surfaces is (generally) fixed by the distribution (these arguments also apply, to a lesser extent, to the MDM and MLC).

This is more noticeable when dealing with complex, overlapping distributions such as is present in dataset 2, but it's non-reliance on a known set of class distributions more than compensates, when compared to the Bayesian approximation techniques.

The MLP shows the best learning and generalizing ability, but at a speed sacrifice that may be prohibitive in some instances.

Other Classification Techniques. Although the classifiers presented here are fairly representative of the types of approaches currently available, there are several other methodologies emerging that require some comment.

Logistic Discriminate Analysis is similar to LDA, but with the ability to construct non-linear decision boundaries. Several classifiers based on this technique have been developed and have been shown to give results comparable to the decision tree approach. They are, however, quite complex in implementation, hence rather oblique to analysis .

The self-organising map (SOM) is a variant of the unsupervised Kohonen neural network that is being used with some success as an alternative neural network approach. It essentially does a vector search in attribute space to find a set of "key" vectors that represent each class and then runs a clustering routine to develop decision boundaries. It is a much faster technique than neural networks based on back propagation, giving training times on par with optimized decision trees. It does not provide quite the same generalizing ability as the MLP used here, however, so at this stage, it is hard to see

any distinct advantage that it might have over an optimized decision tree technique, although current research is encouraging .

Genetic algorithms (GA) have recently restirred research interest, particularly a variant known as Differential Evolution . The technique is based on a model of the biological system of splitting and recombining chromosomal sequences. Combinations of attributes are represented as “chromosomes” and those that provide the best class separation per iteration (or “generation”) are selected to “evolve” through to the next iteration. The technique provides a truly global attribute search strategy, rather than the local strategies used in decision trees and neural networks. However, the price is efficiency.

III. CONCLUSIONS

When generalization ability is the dominant criterion for success, unconstrained by efficiency considerations, the MLP is a consistently superior classifier. It can work with sparse, noisy data and does not require any assumptions on the population distribution or the sampling process. By contrast, the popular MLC classifier is not significantly faster, becomes rapidly more inefficient as the number of attribute dimensions increases and gives poorer classification accuracy on real-world problems due to its underlying statistical data requirements. The MDM classifier has a significant benefit in terms of speed, but if this is what is required, the LDA classifier is a better choice; it outperforms the MDM and even the MLC on the more complex dataset.

The decision tree classifier is perhaps the best all-round choice. It is as fast as LDA, approaches the MLP in terms of learning ability and still maintains useful generalizing ability. As long as the noise in the dataset to be classified has been well-modeled in the training set, it is a quite robust classifier that does not require any knowledge of the data distribution.

REFERENCES

1. *1. Dunteman, G. H.* Introduction to Multivariate Analysis. Sage Publications, Beverly Hills, CA. 1984.
2. *Ibrahimova S.R*. Development of new mathematical methods for increase of accuracy rasters transformation. News of NAS Azerbaijan Republic. Institute of Cybernetics, Institute of Information Technologies. Baku, 2003, pp.61-64.
3. *Ibrahimova S.R.* Methods of remote sensing of the ground with application new information technologies. News of NAS of Azerbaijan. 2002, v. XXII, №2-3, pp. 168-173.
4. *Ibrahimova S.R.* The Mathematical Modeling of Sustainable Development of Environmentally Hazardous Area. ISNET Journal of Space Sciences and Technology. Vol.1., No.1, 2004.
5. *Hunt, E. B., Marin, J. and Stone, P. J.* Experiments in Induction. Academic Press, NY, USA, 1966.
6. *Lim, T., Loh, W. Y.* An Empirical Comparison of Decision Trees and Other Classification Methods. Technical Report No. 979, Madison, USA. 1998.
7. *Richards, J. A. Remote* Sensing Digital Image Analysis*: An Introduction.* Springer-Verlag, Berlin. 1986.