Comments on the Skidmore and Turner Supervised Nonparametric Classifier

Peng Gong
Earth-Observations Laboratory, Institute for Space and Terrestrial Science, North York, Ontario M3J 3K1, Canada
J. Douglas Dunlop
Earth-Observations Laboratory, Institute for Space and Terrestrial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

ABSTRACT: In this paper, the supervised nonparametric classifier proposed by Skidmore and Turner (1988) is discussed. Because the notation in the Skidmore and Turner paper is enigmatic with a published erratum (Skidmore, 1989), the algorithm has been restated in a consolidated form. The computational efficiency of the original algorithm is improved.

INTRODUCTION

PARAMETRIC CLASSIFIERS assume that the form of each class probability distribution is known. To obtain the exact class probability distribution, only certain parameters need to be estimated from training samples (Swain and Davis, 1978). For instance, the commonly used maximum-likelihood classifier (MLC) is a parametric classifier. It assumes that the class probability distribution is multivariate normal (i.e., unimodal). The probability distribution for a particular class is modeled completely by the means vector and the covariance matrix. Estimates of these parameters are obtained from training samples. However, for certain types of classifications, researchers find that the parametric assumption of a class used by the MLC does not hold (Ince, 1987; Kliparchuk and Eyton, 1987). To circumvent this problem, researchers have attempted to develop and/or evaluate nonparametric classifiers, because they do not assume any form for the probability distributions of the classes. All the information about the class probability distribution comes from the analysis of training samples.

Recently, Skidmore and Turner (1988) proposed a new supervised nonparametric classifier. A modification to the original algorithm was suggested by Lowell (1989) and rebutted by Skidmore and Turner (1989). A 14 to 16 percent overall accuracy improvement in the classification of forest cover types was obtained in comparison with the results obtained by using the MLC with SPOT High Resolution Visible (HRV) (Skidmore and Turner, 1988) and Landsat Thematic Mapper (TM) data (Skidmore, 1989).

In this paper, the supervised nonparametric classifier proposed by Skidmore and Turner (1988) is examined. Its computational efficiency is improved. Because the notation in the Skidmore and Turner paper is enigmatic with a published erratum (Skidmore, 1989), the algorithm will first be restated in a consolidated form. The confusion arises in the Skidmore and Turner notation because of the ambiguous use of summation symbols and subscripts. None of the summation notation shows which variable is being summed over. In some equations, it is meant to be summation over **X** (the random variable) while in the others what is meant is summation over i or j (classes). The symbol F_j is used inside a summation over j which leaves the reader wondering because F_j is the total number of training area pixels, which is constant. This notion was not clarified, even in the response article (Skidmore and Turner, 1989).

SKIDMORE AND TURNER'S ALGORITHM

Each pixel in the m (N in Skidmore and Turner (1988)) channel multichannel image has m gray-level values which are considered as a measurement vector $\mathbf{X} = (x_1, x_2, \ldots, x_m)$ in the m-dimensional gray-level space. Provided the image is quantized into 8 bits, it is possible for each axis in the gray-level space to have gray levels ranging from 0 to 255. Labeled training samples are obtained for each of k classes (Skidmore and Turner (1988) used j in three ways: to denote the number of classes, subscript variable in their Equations 1 and 2, and subscript for the total number of samples used. In this paper, j is only used as a subscript variable). For class i, a probability density function in the m-dimensional gray-level space, $F_i(\mathbf{X})$, can be obtained from the training samples

$$f_i(\mathbf{X}) = n_i(\mathbf{X})/N_i \quad i = 1,2, \ldots, k \tag{1}$$

where $n_i(\mathbf{X})$ ($F_i(\mathbf{X})$ in Skidmore and Turner (1988)) is the number of pixels with a gray-level vector **X** and N_i (F_i in Skidmore and Turner, (1988) is the total number of sample pixels for training class i:

$$N_i = \sum_{\mathbf{X}} n_i(\mathbf{X}) \quad \text{for all } \mathbf{X}. \tag{2}$$

According to Skidmore and Turner (1988), class densities $f_i(\mathbf{X})$ are used to approximate the class conditional probability distribution

$$P(\mathbf{X}|i) = f_i(\mathbf{X}) \quad i = 1,2, \ldots, k. \tag{3}$$

To classify a pixel with its gray-level vector **X** into one of the k classes, the *a posteriori* probability $P(i|\mathbf{X})$, for **X** belonging to class i, is estimated through

$$P(i|\mathbf{X}) \doteq \frac{P(\mathbf{X}|i)P(i)}{\sum_{j=1}^{k} P(\mathbf{X}|j)P(j)}. \tag{4}$$

In the above equation $P(i)$ is the *a priori* probability of class i. Skidmore and Turner (1988) used class areal proporations resulting from unsupervised clustering of the image to estimate these *a priori* probabilities. Substituting Equations 1 and 3 into Equation 4, Skidmore and Turner (1988) have

$$P(i|\mathbf{X}) = \frac{(n_i(\mathbf{X})/N_i)\ P(i)}{\sum_{j=1}^{k} (n_j(\mathbf{X})/N_j)\ P(j)} \tag{5}$$

0099-1112/91/5710–1311$03.00/0

(Equation 1 in Skidmore and Turner (1988)). This is exactly what we want from Bayes' theorem in deriving the *a posteriori* probabilities (or empirical probabilities). Skidmore and Turner (1988, p.1416) further proposed "a weighting factor to normalize training area fields of different size." The normalization, however, caused some confusion. A correction was made on p.900 in the June 1989 issue of the *PE&RS* which, unfortunately, is still misleading. Based on our interpretation of the context in Skidmore and Turner (1988), the correct expression with normalization factors is

$$P(i|\mathbf{X}) = \frac{(N/N_i)n_i(\mathbf{X})\ P(i)}{\sum_{j=1}^{k}(N/N_j)n_j(\mathbf{X})\ P(j)} \tag{6}$$

(Equation 2 in Skidmore and Turner (1988)) where the normalization factor N is the sum of all training area pixels,

$$N = \sum_{i=1}^{k} N_i. \tag{7}$$

For each vector $\mathbf{X}$ in the training sample of class i, Skidmore and Turner (1988) estimated its *a posteriori* probability using Equation 6. Subsequently, they constructed a look-up table for each class, with its entries being every possible vector $\mathbf{X}$ and output the conditional *a posteriori* probabilities. For a pixel with gray-level vector $\mathbf{X}$, they compared all the *a posteriori* probabilities taken from look-up tables and assigned the pixel to the class with the greatest probability of being correct.

DISCUSSION OF THE ALGORITHM

From Equation 6 it is obvious that N is a common factor in every term of the summation and therefore It can move in front of the summation symbol Σ. The N in the denominator then cancels with the N in the numerator, making Equation 6 mathematically equivalent to Equation 5. This means that "the weighting factor to normalize the training area fields of different size" (Skidmore and Turner, 1988, p.1416) is a redundant term which has no beneficial effect on the classification. Because Equation 6 is made more complex by including the multiplicative term, it is therefore computationally less efficient.

From Equation 5, it can be seen that the denominator is a constant in all the *a posteriori* class probabilities for a given gray-level vector $\mathbf{X}$. This implies that, when the *a posteriori* probabilities are compared, only the numerator contributes to the discrimination. Therefore, in construction of the look-up table for class i, only the numerator $[n_i(\mathbf{X})/N_i]P(i)$ needs to be calculated. The decision rule is thus simplified to

$\mathbf{X} \rightarrow$class i iff for $j = 1,2, ..., k$

$$(n_i\ (\mathbf{X})/N_i)P(i) \geq (n_j\ (\mathbf{X})/N_j)P(j). \tag{8}$$

For classification purposes, the decision rule in Equation 8 is equivalent to comparing empirical probabilities as Skidmore and Turner (1988) did. But Equation 8 is computationally more efficient. If empirical probabilities are needed for purposes other than classification, Equation 5 should be used. If $P(i)$ $i = 1,2,...,k$ are not known beforehand, as is frequently the case in practice, an equal *a priori* probability for each classis assumed. This simplifies Equation 8 to the following:

$\mathbf{X} \rightarrow$ class i iff for $j = 1,2, ..., k$

$$n_i(\mathbf{X})/N_i \geq n_j(\mathbf{X})/N_j \tag{9}$$

In fact, only one look-up table is needed for the nonparametric classifier. For each gray-level vector $\mathbf{X}$ to be classified, the corresponding entries in the look-up table are extracted and compared using Equation 8 or Equation 9. The resultant class label is then assigned. For those gray-level vectors in the m-dimensional space which were not sampled during training of any class, a label "unclassified" is assigned.

The advantage of this algorithm lies in the fact that it does not require any assumption of the probability density function for a class. The class probability density function is exactly the density of the class training sample. Because the density can take any form, the algorithm is not limited to the Gaussian probability density as is the case with the MLC.

Similar to histogram-based clustering (Letts, 1978), the Skidmore and Turner (1988) classifier requires large look-up tables when the number of image channels is large. For a two-channel image, both channels quantized to 8 bits, there are 256 by 256 gray-level combinations between the two channels. The look-up table therefore requires 256 by 256 entries in order to handle all the possible gray-level vectors that an image pixel may have. The number of entries in the look-up table increases exponentially as more channels are added to the image. In practice, it is difficult to classify an image with more than three channels at an 8-bit quantization level. To employ this algorithm, one has to either reduce the quantization level or use fewer image channels. To reduce the number of image channels, one may apply principal component analysis or other types of data transformation to the original images. Judicious channel selection is also useful.

Skidmore and Turner (1988) suggested using a "collapsing factor" to merge similar gray-levels. However, the collapsing factor is discussed only in terms of empirical results without its statistical basis or its historical development being addressed. In fact, Parzen (1962) formalized the compromise between resolution and statistical significance inherent in the estimation of probability density estimation. He established the conditions for convergence of the estimated density $p_n(\mathbf{X})$ to the true density $p(\mathbf{x})$ and he also developed a method for ensuring convergence known as *Parzen Windowing*.

The practical significance of this, as applied to the Skidmore and Turner algorithm, is that the number of samples must be large enough that every possible vector $\mathbf{X}$ has a statistically significant chance of being sampled. If there are 256 possible measurement vectors, then, based on random chance, 256 samples give vector a 1/256 probability of being sampled. If this seems low, then consider that for a three channel image 16,777,216 samples are required to get the same probability of populating any arbitray $\mathbf{X}$. This is known as "the curse of dimensionality" (Duda and Hart, 1973, p.95) because typically the image will not contain this many samples. By increasing the collapsing factor, each vector has a greater chance of being sampled and thus there are fewer gaps or holes in the density estimate which can lead to unclassified measurement vectors. In practice, there is always a compromise that must be made between the collapsing factor and the total sample size. The optimum classification accuracy can only be achieved by a careful balance of the two. Although very densely clustered classes may be estimated accurately with a small number of samples, confidence that the density estimate is accurate can only be achieved by adhering to Parzen's convergence criteria. In the nonparametric method, if a gray-level vector is not included in the training samples, pixels having that gray-level vector cannot be labeled during classification. In this respect, the MLC is more flexible because, since MLC uses the continuous Gaussian normal distribution to model the class probability distribution, it can assign a class probability to any gray-level vector.

CONCLUSION

By analyzing Skidmore and Turner's supervised nonparametric classifier, some computational considerations have been identified to simplify the classifier and to enhance its efficiency. One of the most important aspects of probability density estimation, the relationship between the total number of samples and the size of the sampling window, was seemingly overlooked in their analysis.

ACKNOWLEDGMENT

The authors are grateful to Dr. P.J. Howarth in Geography, Dr. E. Jernigan in System Design Engineering, and Mr. D. Barber in the Earth-Observations Laboratory at the University of Waterloo for their comments on a draft of this manuscript. Peng Gong's studies were supported by the International Development Research Centre (IDRC), Ottawa, Canada. J. Douglas Dunlop is supported by a Centre of Excellence grant from the Province of Ontario to the Institute for Space and Terrestrial Science.

REFERENCES

Duda, R. O., and P. E. Hart, 1973. *Pattern Classification and Scene Analysis*. Wiley and Sons, New York, 482p.

Ince, F., 1987. Maximum likelihood classification, optimal or problematic? a comparison with the nearest neighbour classification. *International Journal of Remote Sensing*, 8(12): 1829–1838.

Letts, P., 1978. Unsupervised classification in the ARIES image analysis system. *Proceedings of the 5th Canadian Symposium on Remote Sensing*, pp. 61–71.

Lowell, K. E., 1989. A probabilistic modification of the decision rule in the Skidmore/Turner supervised nonparametric classifier. *Photogrammetic Engineering & Remote Sensing*, 55(6): 897–899.

Kliparchuk, K., and J. R. Eyton, 1987. Understanding feature spaces. *Proceedings of the 11th Canadian Symposium of Remote Sensing*, Waterloo, Ontario, Canada, pp.579–589.

Parzen, E., 1962. On estimation of a probability density function and mode, *Ann. Math. Stat.*, 33,1065–1076.

Skidmore, A. K., 1989. An expert system classifies Eucalypt forest types using Thematic Mapper data and a digital terrain model. *Photogrametric Engineering & Remote Sensing*, 55(10): 1449–1464.

Skidmore, A. K., and B. J. Turner, 1988. Forest mapping accuracies are improved using a supervised nonparametric classifer with SPOT data. *Photogrammetric Engineering & Remote Sensing*, 54(10): 1415–1421.

———, 1989. Why areal accuracy is not correctly estimated using Lowell's modification of the supervise nonparametric classifier. *Photogrammetric Engineering & Remote Sensing*, 55(6): 899–900.

Swain, P. H., and S. M. Davis, 1978. *Remote Sensing: The Quantitative Approach*. McGraw-Hill, New York, 396p.

(Received 7 December 1989; revised and accepted 4 December 1990)

Response

WE WOULD LIKE to thank P. Gong and J. D. Douglas for re-expressing the supervised nonparametric classifier algorithm in a form which assisted their understanding of the algorithm, and which may help other readers in interpreting the algorithm.

However, we wish to comment on a number of points that Gong and Douglas raised.

(1) The denominator in Equation 2 (Skidmore and Turner, 1988) is needed if the empirical probabilities are to be calculated. If the empirical probabilities are not required, then the denominator is not needed for the discrimination of the classes at vector space **X**.

(2) The questions posed in the last paragraph of the section titled "Discussion of the Algorithm" may be simply answered:

(i) Five Euclidean distance units were the rejection criteria used for the Euclidean distance classifier.

(ii) As explained on page 1419 of the original paper, only those pixels with a 75 percent empirical probability of correct classification were chosen (from the 316 test pixels). This probability was chosen to indicate that as the empirical probability increases, so does the mapping accuracy. Additional empirical probabilities could be taken to further investigate the relationship between empirical probability and mapping accuracy.

(iii) As stated in the original paper, only those vector spaces with the denominator in Equation 2 (Skidmore and Turner, 1988) equal to 0 will remain unclassified (i.e., those vector spaces with no training area pixels and therefore an empirical probability of 0 percent).

Referring specifically to the conclusions by Gong and Douglas:

(3) We claimed that the classifier is new for the analysis of remotely sensed data, not that the underlying mathematics was original (see reference to Geisser [1982] in the original paper). Neither Parzen (1962) nor Duda and Hart (1973) developed their concepts into a classification algorithm.

(4) The modification to the supervised nonparametric classifier proposed by Gong and Douglas, and discussed in point (1) above, would appear to offer some improvements in processing efficiency, *provided* the empirical probabilities are not required.

(5) Parzen (1962) and Duda and Hart (1973) refer to probability density functions, while we consider the probability at individual vector spaces. Gong and Douglas do not establish (or test) the relationship between the collapsing factor approach we used and Parzen windows.

We welcome the opportunity to address the issues raised by Gong and Douglas, and find it gratifying that the algorithm is generating this interest.

A. K. Skidmore and B. J. Turner
Forestry Commission of N.S.W.
GPO Box 2667
Sidney, NSW 2000
Australia