SENS'2006

Second Scientific Conference with International Participation

SPACE, ECOLOGY, NANOTECHNOLOGY, SAFETY

14 – 16 June 2006, Varna, Bulgaria

MULTISPECTRAL DATA CLASSIFICATION BY NON-LINEAR METHODS

Hristo S. Nikolov[a], Nina Jeliazkova[b]

[a] *STIL-BAS, Acad. G. Bonchev bl.3, Sofia 1113, Bulgaria*

[b] *IPP – BAS, , Acad. G. Bonchev bl.25a, Sofia 1113, Bulgaria*

Key words: non-linear methods, classification, texture

Abstract.

The last remote instruments provided the scientific community two important improvements – hyperspectral sensors with increased spatial resolution. This opened new trends in classification of small areas of the land cover and anthropogenic objects. Together with this more precise spectral and spatial new challenges are posed to the algorithms for data processing, namely the exponentially increased volume of data. One promising method to overcome this problem is to focus the research only on that features which best describe the object of interest. These features may include not only these mentioned above but also textural and geographical ones. The idea of this research was to establish a framework for remotely sensed data classification based on non-linear methods developed recently, such as data mining and kernel based "kd-trees". The results obtained confirmed some improvements and simplicity from computational point, which means robustness, and slight increase of the map accuracy.

This study was partially supported by NSFB under Contracts NZ-1410/04 and MUNZ-1502/05.

INTRODUCTION

The objective of classification methods is to determine to which class a given sample belongs. The observation vector is usually obtained through some measurement process and serves as the input to a decision system by which we assign the sample to one of the given classes. Probabilistic methods, discriminant analysis, nearest-neighbour classifiers, neural networks and decision trees are representative classification techniques. In the context of remote sensing, the observation vector consists of the spectral responses of land cover objects that form either image pixels or regions. In the paper we demonstrate the advantages of using Bayesian classifier, based on a recent very fast algorithm for nonparametric density estimation, to the problem of land cover classification. First we compare the performance of the proposed algorithm with several different classification algorithms from STATLOG project using a small benchmark dataset. Then a larger data set taken from Corine Land Cover project for Bulgaria is used to compare the proposed algorithm with the k-nearest neighbours classifier and Naïve Bayes classification.

As an alternative to the crisp classification methods, where each pixel is classified to exactly one class, the soft classification methods assign multiple class memberships to a pixel. Soft classification methods provide more realistic interpretation of the real world, where land cover intergrades gradually and boundaries between classes are sometimes vague. The accuracy of a crisp classification is usually summarized in an error (confusion) matrix, where rows represent classes as observed (ground truth) and columns represent predicted classes (Congalton and Green, 1993). The cell *(i,j)* contains the number of pixels from *class i*, predicted as *class j*.

Bayes' decision theory is a fundamental statistical approach to the problem of classification (Duda et al., 2000; Fukunaga, 1990). This approach is based on the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known beforehand or estimated from data. We select to use the classification approach based on Bayes' decision rule, because it is known to be theoretically optimal in the sense of minimal classification error. Bayes' rule provides only a general framework for the supervised classification and several algorithms can be derived if different approaches to estimate probability densities are employed. We favor the nonparametric density estimation in order to obtain the necessary probability values, instead of assuming that the data is distributed according to one of the standard distributions. According to the statistical tests applied to the data used, the hypothesis of data normality was rejected. Moreover, it was noted (Landgrebe 2000, Landgrebe 1998) that in the context of remote sensing image classification, higher order moments of probability distributions are more important for the classification. According to statistical tests applied to our data (Jarque-Bera test using the corresponding Matlab function), the hypothesis of data normality is rejected. Therefore, we could expect more precise classification results if the data distribution is reflected more accurately by a nonparametric technique. In addition, we took the advantage of a recently proposed Very Fast Algorithm for Multivariate Kernel Density Estimation (Gray, 2003) .We compare the algorithm performance with Naïve Bayes approach, which is also based on Bayesian decision theory, but relies on a strong assumption of features independence. Another restriction of the particular Naïve Bayes implementation used is that the probability densities of the features are assumed normal. We also compare the algorithm performance with k-nearest neighbor classifier, which classifies a query point based on the class membership of the majority of the closest *k* neighbors of the query point. "Closeness" relies on a predefined distance measure, usually Euclidean distance. Obviously the decision is affected by the user specified parameter *k* and the distance measure.

METHOD

The nonparametric technique of choice in this paper is the kernel density estimation. It is known to approximate the true density of the data if enough data points are observed (Silverman 1986.). The idea of kernel density estimation is to model the density as a sum of the influences of the data points. The influence of a data point is given by a kernel function, which is symmetric and has maximum at the data point. Examples for kernel functions are Gaussian bell curve, square wave function, etc. The density function takes higher values in regions, where some kernel functions have a significant overlap. Given random sample $\mathbf{x}_1$, $\mathbf{x}_2$,…$\mathbf{x}_n$ from an unknown true density **f(x)**, the kernel density estimate $\hat{f}(x)$ of **f(x)** at the point $x \in R^d$ is

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where *h* is the *smoothing parameter (bandwidth)*, *n* is the number of data, K() is a kernel function, which satisfies $\int_{-\infty}^{\infty} K(x,h)dx = 1$.

One of the recent advances reduces the complexity (Gray, 2000, Gray, 2003 - Very

Fast Multivariate Kernel Density Estimation) and provides fast and accurate computation of kernel density estimate by using computational geometry to organize the data. This is achieved by a hierarchical representation of the data, which divides the points into hierarchy of subsets and caches sufficient statistics for each subset. The implementation relies on adaptive algorithms to build k-dimensional tree –*kd*-tree (Moore 2000). The cost of constructing such a tree is O(NlogN), assuming uniform data distribution. An example of the cached statistics is the bounding box statistics, which represent a box, containing all points at given node. The hierarchy and cached statistics are used to approximately estimate the kernel density using dual-tree algorithm (Gray,2003). Two *kd*-trees are build, one for the training set and one for the query data set (e.g. test data set in classification scenario). The heart of algorithm is the idea that the interaction between points in a bounding box in the first tree and the points in a bounding box in the second tree can be approximated by a constant (making use of the cached statistics). If the approximation is not sufficiently accurate, then the procedure repeats recursively by comparing children nodes of the tree. The required accuracy can be defined by the user, for exact estimation the algorithm has to reach the leaf nodes of the tree. Thus the approximation level allows tradeoffs between evaluation quality and computation speed. The implementation of this algorithm is available as a Matlab toolbox, distributed by GNU LGPL license (Ihler, 2004). The kernel bandwidth in each dimension is selected by crossvalidation.

For the land cover type recognition we developed a Matlab script, making use of this toolbox for the probability density estimation. The classification procedure proposed is as follows:

1. The Principal Component Analysis (Kahrunen-Loeve Transform) is performed over all training data (all classes) in order to obtain orthogonal variables.
2. Multivariate conditional probability densities for each of the *C* classes are estimated via dual tree algorithm (Gray, 2003);
3. Posterior probability for each pixel of training data set is calculated using the estimated densities and Eq.1, assuming equal prior probabilities for all classes. Thus for a pixel we obtain *C* probability values representing the degree of class membership. $p\ (w_i \mid x),\ i=1..C$.
4. Pixel-by-pixel classification. A pixel is assigned to the class with maximum posterior probability. The accuracy of classification is summarized in an error matrix
5. Region classification. In land cover type recognition we are more interested in recognizing continuous regions than single points. Here we propose two approaches in order to make decision for each continuous region of specific cover type, which take advantage of the “soft classification”:

We assess the classification quality by test data set. The test data is also projected into principal components space, obtained in the step 1 above, and the posterior probability for each pixel of test data set is calculated using densities estimated from the training data (step 2). The same statistics as for the training data (confusion matrix, mean class probability, KL distance) are calculated and presented in corresponding tables and figures.

DATA

In this research we considered 2 scenarios:

1. A satellite image (667 x 663 pixels), obtained by Landsat Thematic Mapper is used to assess

the behavior of the proposed algorithm. The spectral information from 7 bands is used as 7 input features; all 12 identified classes are predicted;

2. The same data set as in the second scenario, but each pixel is represented by 7 values for each spectral band and additional 7 values per each of its eight immediate neighbors, thus giving (1+8)x7=63 features per pixel. Only six most important classes are predicted.

The results can be summarized as follows:

- Test set pixel by pixel classification is not good by means of error rate for all but few classes. Similar results are obtained by kNN and Naïve Bayes algorithms (Table 1).
- Region classification (by using mean probability over pixels in a region, or KL distance between training and test distribution) is correct, except for class 212. A visual examination of the class 212 reveals that the two regions used as test and training set respectively are very different. Perhaps more data for this class could help to improve the classification;
- By using only 7 spectral features, the texture information of the image is ignored. This makes classification of regions with similar spectral response (for example forests and fruit trees) very hard;

It is known (Fukunaga, 1990) that to discriminate N classes, at least N-1 features are necessary. Therefore, in this scenario it is unrealistic to expect perfect classification. To overcome these shortcomings, the next scenario was considered.

Class\Accuracy %	1NN , 7 spectral bands		5NN, 7 spectral bands		Naïve Bayes, 7 spectral bands		This paper, 7 spectral bands	
	Training	Test	Training	Test	Training	Test	Training	Test
112	100	51.10	78.90	61.60	27.7	30.8	49	47.11
121	100	11.90	20.80	8.7	21.4	37.6	28	27.27
142	100	2.40	9.50	2.10	0	0	34	22.31
211	100	7.30	69.70	8.90	14.6	4.6	39	21.09
212	100	0.10	55.10	0	64.0	1.2	100	0.88
22	100	26.20	59.30	29.20	28.8	43.4	66	35.31
231	100	64.60	83.90	75.60	63.8	69.9	72	51.28
24	100	35.10	40.30	24.10	1.60	1.6	41	53.72
311	100	92.20	83.60	95.6	84.3	96.6	91	91.72
32	100	23.50	32.10	17.10	7.00	6.8	34	21.62
411	100	4.60	17.90	0.40	3.90	1.1	44	22.82
512	100	94.3	81.40	95.50	82.4	98.6	94	96.96
All	100	47.24	70.44	48.72	48.52	43.09	57.67	41.01

Table 1.Training and test accuracy (percent of correctly classified pixels) of kNN, Naïve Bayes and nonparametric classification method for the data set in scenario 1 (12 classes, 7 features).

Class	1NN, 63 principal components		Naïve Bayes, all 63 principal componentsC		Naïve Bayes, 12 principal components		This paper, 12 principal components	
	Training	Test	Training	Test	Training	Test	Training	Test
112	100	58	52	49	69	64	92.05	70.62
142	100	3.2	16	13	2.40	2.30	90.67	46.43
221	100	60.2	67	81	45	74	99.42	81.80
231	100	83.9	54	58	79	82	97.81	74.65
311	100	95.8	35	98	86	97	99.99	87.83
512	100	97.7	100	00	85	99.90	100	90.88

Table 2.Training and test accuracy (percent of correctly classified pixels) of kNN, Naïve Bayes and nonparametric classification method for the data set in scenario 2 (6 classes, 63 features).

CONCLUSIONS

The proposed method for non-linear classification of multispectral data provides better accuracy than conventional ones. The assigned class-membership of a pixel is a “soft classification”, i.e. a probability of a pixel to belong to the class is provided, instead of “yes”/“no” answer. This could be very helpful in the context of classification of remote sensing imagery with moderate spatial resolution. In addition to pixel-by-pixel classification of an image, it allows classification of predefined regions of the image as a whole. The classifications of regions as integral objects are accurate even in difficult scenario as using only spectral features and discriminating among 12 classes. The high error rate for some of the classes is because of the insufficient data and not using the relevant features (texture), which will be in a focus of further research. The focus of a future research we put on developing better methods for dimensionality reduction of the feature space.

REFERENCES

Congalton R. & Green K., A practical look at the sources of confusion in error matrix generation. Photogrammetric Engineering and Remote Sensing 59, 641-644, 1993

Devroye L., L. Gyorfi, G. Lugosi. A probabilistic Theory of Pattern Recognition, Springer, 1996.

Duda R., P. Hart, D. Stork. Pattern Classification, 2nd ed., John Wiley & Sons, 2000.

Foody, G.M., 1999. The continuum of classification fuzziness in thematic mapping, Photogrammetric Engineering and Remote Sensing, 65, 443-451.

Fukunaga, K., Introduction to Statistical Pattern Recognition, San Diego, California, Academic Press Inc., 1990.

Gray A. and A W Moore, `N-Body' problems in statistical learning, NIPS, 2000, pp. 521-527.

Gray A. and A. Moore, Nonparametric Density Estimation: Toward Computational Tractability, Proc. SIAM International Conference on Data Mining , San Francisco, USA, 2003.

Gray A. and A. Moore, Very Fast Multivariate Kernel Density Estimation using via Computational Geometry, in Proceedings, Joint Stat. Meeting 2003, August 3 - 7, 2003 · San Francisco, California

Hand, D. (1982), Kernel Discriminant Analyis, Research Studies, New York.

K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press,1972.

King, R.D., Feng, C., & Sutherland, A. (1995) StatLog: Comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence.* 9, 289-333

David Landgrebe, Information Extraction Principles and Methods for Multispectral and Hyperspectral Image Data, Chapter 1 of *Information Processing for Remote Sensing,* edited by C. H. Chen, published by the World Scientific Publishing Co., Inc., 1060 Main Street, River Edge, NJ 07661, USA, 2000

MATLAB KDE class, http://ssg.mit.edu/~ihler/code/kde.shtml.

Mitchie, D., Spiegelhalter, D. and Taylor C. (eds) (1994). Machine learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence, Ellis Horwood.

Moore A., The Anchors Hierarchy: Using the Triangle Inequality to Survive High Dimensional Data. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, USA p: 397 – 405, 2000

Silverman B. W.. Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986.

Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

http://www.cs.waikato.ac.nz/ml/weka/