Knowledge Base Data Mining and Machine Learning in a Parallel Computing Environment

William J. Campbell, Samir Chettri

Abstract

The expectation of this research is to greatly broaden the use of remotely sensed imagery by providing a novitiate user, access to embedded information and knowledge without embarking upon a full-scale research project to complete the content extraction, storage and retrieval process. The intent of our approach is to develop an intelligent system that can adapt to changes or new information and learn from these changes. This will drastically alter the approach researchers take in using any digital imagery by opening the scientific discovery process, particularly to disciplines that have not traditionally used imagery due to the complexity of the image processing techniques. We hope to accomplish this by the judicious use of declarative and procedural knowledge, engineering, and automatic feature or image object labeling using recent classification techniques on BEOWULF parallel computing architectures

Introduction

The transformation of raw imagery data into information and eventually into knowledge for distribution into decision-making processes has not changed significantly since Landsat was launched in 1972. Image processing, analysis, product generation and delivery tasks are performed on a case-by-case basis using mostly manual methods and basic classification tools. This was not a problem when data availability was low and a skilled researcher could spend a considerable amount of time working with just one Landsat scene. Today's data volumes and availability have increased exponentially since then, yet the methods and tools to extract and distribute information have remained at approximately the same capacity level. A majority of funding has been allocated to the development of spacecraft platforms and sensors with much less spent developing techniques for analysis. A large portion of the science resources available to researchers is spent on extracting the content from satellite imagery even before the knowledge discovery begins.

Accessibility of data, derived products, and processing algorithms must be assured through the (next generation) Internet and Intranets". We are developing methods to provide users of all levels of expertise access to information and knowledge derived from remotely sensed data without the need to embark upon a research project to complete the typical content extraction, storage, and retrieval process. We employ *naïve Bayes* image classification algorithms for automated region probability determination and labeling. Subsequently (and perhaps simultaneously) we explore the use of *Bayesian multinets* for machine learning and automated content definition. The resulting image content products are updated using spatial data from various sources to build informative priors to update the resulting image products from either the *naïve Bayes* or *Bayesian multinet* approaches.

One of the important exercises in the geospatial sciences is bridging the gap between what is sensed from above and what is reality on the ground. This is typically accomplished via mapping campaigns that include field visits, or at the very least substantial a priori knowledge of ground characteristics for the region of interest.

The ability to acquire, store, and process newly acquired imagery and collected field data in a timely manner is therefore important to all users of remotely sensed data, regardless of application. However, all users suffer similar technological barriers such as processing capabilities, digital data transfer capabilities, and data storage limitations such as photographs (with GPS inscribed in the images), GIS files, Hand-held Spectrometer readings, Sun Photometer data, audio, and video data.

Classification of imagery has a long and varied history [**Jan 2000**]. A standard taxonomy of classification first divides classifiers into *supervised* and *unsupervised*. In each category, the classifiers may be *parametric* or *non-parametric*. We are using all varieties of supervised and unsupervised classifier (whether parametric or non-parametric). Typically we will first use the unsupervised methods to either reduce the dimensionality or as a way of visualizing the high dimensional data. Below we briefly describe the algorithms to be studied. We indicate with an asterisk (*) whether we have previously applied the method to remotely sensed data. Note that even if we have applied the methods to remotely sensed data, we have not done extensive testing of these methods – which we would do on various data sets. With the exception of the SVM method (below), none of the techniques have been applied to Hyperspectral data, which is one focus area of interest.

A. Unsupervised classification:

- *Kohonen Feature Map (*)*: This is a way to reduce high dimensional data to lower dimensional data and is considered to be a non-linear version of principal components analysis [**Kohonen 1989**] [**Haykin 1999**].

- *Generative Topographic Mapping (GTM)*: Was to be a principled (i.e., Bayesian) alternative to the Kohonen Feature Map and is used for visualizing high dimensional data in lower dimensions. [**Bishop et al. 1998**].

- *Probabilistic Principal Components Analysis*: This is similar to the GTM in that it uses Bayesian theory as a starting point to develop a principled alternative to principal components analysis (PCA). Additionally, mixtures of probabilistic principal component analyzers are developed in [**Bishop et al. 1999**].

- *Kernel Based Principal Component Analysis*: This is another non-linear principal component analysis method. It uses techniques developed in the theory of support vector machines (discussed under the supervised classification methods) [**Scholkopf et al. 1999**].

- *Artificial Neural Network/Independent Component Analysis (ANN/ICA):* Typically, one assumes a single class of land coverage for each pixel at the location (x, y), even if several types of land canopy might exist within each pixel. We show how Artificial Neural Network/Independent Component Analysis (ANN/ICA) can provide sub-pixel class percentage composition. **[Szu 1999-1] [Szu 1999-2].**

All of the above unsupervised classification methods are applicable to both multi-spectral and hyperspectral data without exception. All four would be used for dimensionality reduction or high-dimensional data viewing/abstraction.

B. Supervised classification:

- *Maximum Entropy Spectral Unmixing (ME) (*)*: Uses PCA (or the non-linear variants, see item a) above) to determine end-members and obtains the fractions of a land cover class in a pixel using methods from quantum statistical mechanics & Bayesian statistical inference [**Chettri et al. 1996**], [**Chettri et al. 1997-1**].

- *Gaussian Maximum Likelihood Classification (GMLC) (*)*: Assumes a Gaussian distribution for each class [**Richards-Jan 2000**]. Provided in many toolboxes.

- *Support Vector Machines (SVM) (*)*: Recently, powerful new statistical techniques known as the SVM, using the fundamental concept of the *Vapnik-Chervonenkis dimension* have been developed that helps bypass the *curse of dimensionality*. The application of SVM's to hyperspectral data can be seen in [**Gualt-Chettri 1999**] [**Gualt-Chettri 2000**].

- *Mixture Model Neural Networks (MMNN)(*)*: These use the Expectation Maximum (EM) algorithm to model the underlying density as a mixture of Gaussians. Under slightly restrictive assumptions they can be applied directly to hyperspectral data . The EM algorithm is a general purpose iterative method that is applied in the calculation of Maximum Likelihood (ML) estimates. Since image classification in remote sensing can be stated as an ML problem, the EM algorithm has wide applications. [**Chettri et al. 1997-2**].

- *Back-propagation Neural Network (BPNN) (*)*: [**Campbell et al. 1989-1**] presents what is probably the first application of BPNN's to remotely sensed data. They are remarkably robust & powerful classification methods & remain a standard in the remote sensing classification literature.

Not all of the above supervised classification methods may be applied directly to hyperspectral data – for example only Maximum Entropy (ME) and SVM require no modification at all to be so applicable. All the others usually require modification of the data (i.e., reduction to lower dimensions) or strong assumptions on the data itself. However, given the great volumes of data and the greater number of channels becoming available (i.e., hyperspectral data) we are modifying the above methods to work on parallel architectures as well as combine the classifiers in a clever manner

Parallelize Selected Algorithms On A BEOWULF Cluster

We parallelize selected algorithms on the Highly-parallel Integrated Virtual Environment (theHIVE), TheHIVE is a Beowulf-class parallel machine and provides inexpensive, re-programmable, high performance computing. We will leverage this resource to accomplish our computing requirements.

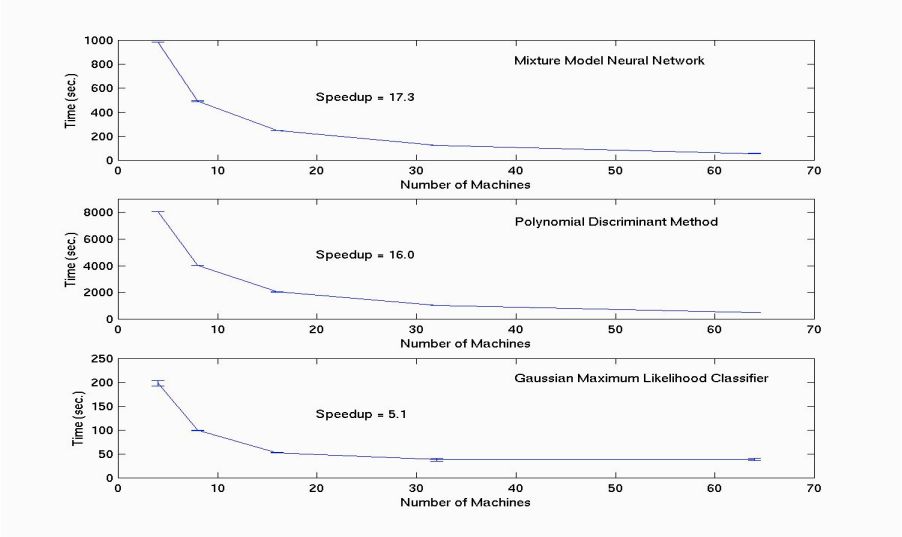
Table 1 shows which classifiers are currently parallelized on theHIVE and which are not. We are investigating the applicability (speed, classification accuracy, etc.) of those algorithms that have not been applied for classification of remotely sensed data. After this process, we will choose to parallelize some or all of the set below. All of the algorithms have been chosen because they show promise as unsupervised/supervised image classifiers.

CLASSIFICATION METHOD	STATUS	APPLICABILITY
Gaussian Maximum Likelihood Classifier (GMLC)	*S, P*	*M*
Mixture Model Neural Network	*S, P*	*M*
Support Vector Machines	*S*	*M, H*
Back Propagation Neural Network	*S*	*M, H*
Mixed pixel analysis using Maximum Entropy Methods	*S*	*M, H*
Hierarchical Image Segmentation	*S, P*	*M, H*
Kohonen feature maps	*S, P*	*M, H*
Generative Topographic Mapping	*S*	*M, H*
Probabilistic PCA	*S*	*M, H*
Kernel based PCA	*S*	*M, H*

Table 1: Status (*S* = Serial, *P* = Parallel code available) and applicability (*M* = applicable to Multispectral, *H* = applicable to Hyperspectral) of classification algorithms both supervised and unsupervised.

Of the supervised classification algorithms that have been parallelized, common features of their performance on theHIVE stand out. These are illustrated in Figure 2 [**Smit et al. 2000**]. Here we see that there is no super-linear speedup – rather we reach a point of diminishing returns, which are about 16 machines for the given architecture.

Figure 1: Classification time vs. number of machines on theHIVE for several algorithms.



It is of interest to note whether we can get additional classification accuracy by running several classifiers in parallel (for the same image or multi-source images of the same region) and then combining the classifier outputs suitably. We are investigating the "Behavior-Knowledge Space Method," described in [**Huang-Suen 1995**]. Thus we can truly see the benefits of parallel processing – running classifier fusion on a serial machine would be overly time consuming – especially when the dimensionality of the data is high and the size of the image large or a large number of images need to be processed.

Perform initial classification using Bayesian Networks

Thus far in our processing approach we have been using *naïve Bayesian* classifiers. These methods work in the following manner – Given a set A_i of attributes and a class label c, the naïve Bayesian classifier learns the conditional probability of A_i from training set. The strong assumption made in the naïve Bayesian classifier is that the A_i's are independent given the value of the class C. The structure of the naïve Bayesian classifier is given in Figure 2, below.

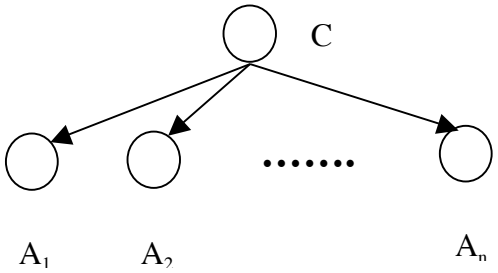


Figure 2: Naïve Bayesian classifier.

A Bayesian Network (BN) [**Frey 1998**] is a Directed Acyclic Graph (DAG) in which the nodes represent variables and the (directed) arcs between nodes imply causation. For example if an arc goes from node A to node B, this implies that A causes B. Additionally the Bayesian networks specify a conditional probability distribution between P (A| P_1, P_2, … P_n) where A is a node whose parents are P_1, P_2, … P_n. Thus a BN consists of two parts: A topology and estimation of the parameters of the conditional distribution. Both are difficult problems and require different skill sets. Someone may specify the network topology with an expert knowledge of the problem domain and who is an expert on cause and effect relationships.

Inference in singly connected BN's can be shown to reduce to *local computations* with message passing between nodes. This form of probability estimation is provably non-linear Polynomial (NP) hard even though it gives exact answers. An alternative to probability propagation is Markov Chain Monte Carlo (MCMC) **[Gilks et al. 1996]**. This method uses either the Gibb's sampling or the Metropolis-Hastings algorithm for conditional-distribution estimation but is computationally intense. However parallelization efforts on a Beowulf cluster may mitigate the calculation times. For classification of multi-spectral or hyper-spectral data we will create different BN's to model the probability of each class separately. This allows for a more flexible model. Such groups of networks are called *Bayesian Multinets* [**Friedman et al. 1997**].

Building A Database Of Informative Priors For Bayesian Classification Of Images.

Once a classification is performed using *Bayesian Multinets* or the *naïve Bayesian classifiers* (either singly or as a group) we will combine this classification with deterministic topographical and temporal information. This deterministic topographic information could be for example the presence/absence of roads, drainage networks, soil type elevation, situation of other transportation networks (railroads, rivers, canals), slope, aspect etc.

The mathematical steps of generating priors are as follows [**Frigessi-Stander 1994**]:

- Build first and second order spatial neighbors (known as cliques).
- Model spatial homogeneity. This reflects the idea that every pixel tends to have the same classification as its neighbors. Rotation invariance needs to be taken into account.
- Model the deterministic information mathematically. This is the difficult part and considerable research must go into this. An unpublished document [**Campbell 1988**] combined with [**Frigessi-Stander 1994**] will be used as a starting point. For example we will consider the following types of rules:
  - Adjacency rules
  - Temporal variation rules

Here are examples of rules we can use and quantify mathematically:

1. The higher the density of roads around a pixel, the more likely the pixel is to be urban.
2. Change from agriculture to urban allowed but urban to agriculture is not allowed.

For example a spatial rule is as follows: z_i is the deterministic information we can use for each pixel *i:* z_i = (road present or absent, z_{i1}, z_{i4}). In order to quantify rule 1 above we would write $W(x_i|z_i) = B_5\, z_{i4} - B_6\, z_{i1}$. Here x_i is the cover type (say water, urban etc.) and z_i is the available deterministic information (in this case z_{i1} is the distance from pixel *i to* the nearest road and z_{i4} is the proportion of roads in a window around pixel *i*. Thus, $W(x_i|z_i)$ represents the deterministic information that a pixel *i* is urban (x_i = road) given the deterministic information z_i. B_5 and B_6 are positive parameters.

The spatial homogeneity and the deterministic rules help form the prior information we have of an image that can be represented as a Markov Random Field. The likelihood is constructed from the *naïve Bayesian* (or the *Bayesian Multinets*) classification and the contingency matrix for the given image [**Richards-Jan 2000**]. Using the usual version of Bayes' theorem we can construct the posterior probability of a classified image = prior x likelihood. Of course, Bayes theorem permits updates to the posterior probability based on new information. A key problem in our approach as described above is the need to build a set of rules that we can use to build reliable priors. To this end we will build a knowledge base as described below.

The classic steps to building a knowledge base of knowledge acquisition, knowledge analysis, knowledge-system design, and knowledge base implementation will be followed. The knowledge base system design will be based on well-known principles described in [**Stefik 1995**].

The knowledge base will have to be iteratively improved as classification is performed on remotely sensed scenes and the classification accuracy is assessed. Classification accuracy will be measured using standard statistical techniques (*k-hat statistic*, *kappa statistic*) used in the remote-sensing literature **[Richards-Jan 1999]**.

The expectation of this research is to drastically broaden the use of imagery by providing a novitiate user, access to the embedded information and knowledge without embarking upon a research project to accomplish the content extraction, storage and retrieval process. Our state-of-the-art approach intent is to develop an intelligent system that can adapt to changes or new information and learn from these changes. This will drastically alter the approach researchers take in using any digital imagery by opening the scientific discovery process, particularly to disciplines that have not traditionally used imagery due to the complexity of the image processing techniques.

References

[**Bishop et al. 1998**]. Bishop, C. M., Svensen, M. and Williams, C. K. I. "GTM: the Generative Topographic Mapping." *Neural Computation*. Vol. 10, pp. 215-315, 1998.

[**Bishop et al. 1999**]. Bishop, C. M., Tipping, M. E. "Latent Variable Models and Data Visualization." In Kay, J. W. and Titterington, D. M. eds. *Statistics and Neural Networks,* pp. 147-164, Oxford University Press, 1999.

[**Campbell 1988**]. Campbell, W. J. "Rules for expert system used to process neural network output." *Unpublished manuscript circa 1988.*

[**Campbell et al. 1989-1**]. Campbell, W. J., Hill, S., Cromp, R. "Automatic Labeling and Characterization of Objects Using Artificial Neural Networks", *Telematics and Informatics*, Vol. 6, Nos. 3/4, pp. 259-271, 1989.

[**Chettri et al. 1996**]. Chettri, S. R. and Netanyahu, N. "Spectral unmixing of remotely sensed imagery using Maximum Entropy." In David Shaefer et al. Eds. *25th AIPR Workshop – Emerging Applications of Computer Vision*, SPIE Volume 2962, 16-18 October 1996, Washington, DC, pp. 55-62.

[**Chettri et al. 1997-1**]. Chettri, S. R., Garegnani, J., Robinson, J., Coronado, P., Cromp, R. F., Netanyahu, N., Campbell, W. J. "Multi-resolution maximum entropy spectral unmixing." *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 1997, pp. 347-352.

[**Chettri et al. 1997-2**]. Chettri, S. R., Murakami, Y., Isamu, N., and Garegnani, J. "Comparing the computational complexity of the PNN, the P/dm and the MMNN." In Selander, J. M. ed. *26th AIPR Workshop – Exploiting New image Sources and Sensors*. Vol. 3240, pp. 1260-132, 1997.

[**Friedman et al. 1997**]**.** Friedman, N., Geiger, D. and Goldszmidt, M., "Bayesian Network Classifiers", *Machine Learning*, 29:131-163, 1997.

[**Frigessi-Stander 1994**]. Frigessi, A. and Stander, J. "Informative priors for the Bayesian classification of Satellite Images." *Journal of the American Statistical Association*, Vol. 89, No. 426, pp. 703-709, 1994.

[**Frey 1998**]. Frey, B. J., *Graphical Models for Machine Learning and Digital Communication*, MIT Press, 1998.

[**Gilks et al. 1999**]. Gilks, W.R. et al., *Markov Chain Monte Carlo In Practice*, Chapman and Hall, 1996.

[**Gualt-Chettri 1999**]. Gualtieri, J. A., Chettri, S. R., Cromp, R. F. et al. "Support Vector Machine Classifiers as applied to AVIRIS data." Published in the proceedings of the 1999 Airborne Geoscience Workshop, JPL, February 8-11, 1999.

[**Gualt-Chettri 2000**]. Gualtieri, J. A. and Chettri, S. R. "Support Vector Machines for classification of hyperspectral data." Published in the International Geoscience and Remote Sensing Symposium, Hawaii, August 2000.

[**Haykin 1999**]. Haykin, S. *Neural Networks: A comprehensive foundation*. Prentice-Hall, 1999.

[**Huang-Suen 1995**]. Huang, Y. S. and Suen, C. Y. "A method of combining multiple experts for the recognition unconstrained handwritten numerals." *IEEE PAMI*, Vol. 17, No. 1, pp. 90-94, 1995.

[**Jan 2000**]. Jain, A. K., Duin, R. P. W and Mao, J. "Statistical pattern recognition: A Review." *IEEE PAMI*, V. 22, No. 1, pp. 4-37, 2000.

[**Kohonen 1989**]. Kohonen. T. *Self-Organization and Associative Memory*. Springer-Verlag, Third ed., 1989.

[**Richards-Jan 2000**]. Richards, J. and Jia X., *Remote Sensing Digital Image Analysis – An Introduction,* Springer, 3rd Ed., 2000.

[**Scholkopf et al. 1999**]. Scholkopf, B. Smola, A. J., and Muller, K. R. "Kernel Principal Component Analysis." In Scholkopf, B., Burges, C. J. C. and Smola. A. J. eds. *Advances in Kernel Methods,* pp. 327-352, MIT Press, 1999.

[**Smit et al. 2000**]. Smit, M., Garegnani, J., Bechdol, M., Chettri, S. R. "Parallel image classification on TheHIVE." Accepted for publication and presentation at *AIPR 2000 Imagery in the New Millennium*, October 16-18 2000, Washington, D.C.

[**Stefik 1995**]. Stefik, Mark, *Introduction to Knowledge Systems*, Morgan Kaufmann, 1995.

[Szu 1999-1]. H. Szu, "Progresses in unsupervised artificial neural networks of blind image demixing, " New tech of IEEE Ind. Elec. Soc. Newsletter, pp. 7-12, June 1999.

[Szu 1999-2]. H. Szu, "ICA-an enabling tech for intelligent Sensory processing", IEEE Trans. Circuits and Systems Newsletters pp. 14-41, Dec. 1999.