Applying ISOMAP to the Learning of Hyperspectral Image *

X. Rosalind Wang and **Suresh Kumar** and **Tobias Kaupp**
and **Ben Upcroft** and **Hugh Durrant-Whyte**
ARC Centre of Excellence for Autonomous Systems
Australian Centre for Field Robotics (J04)
The University of Sydney
{ r.wang, s.kumar, t.kaupp, b.upcroft, hugh } @cas.edu.au

Abstract

In this paper, we present the application of a non-linear dimensionality reduction technique for the learning and probabilistic classification of hyperspectral image. Hyperspectral image spectroscopy is an emerging technique for geological investigations from airborne or orbital sensors. It gives much greater information content per pixel on the image than a normal colour image. This should greatly help with the autonomous identification of natural and man-made objects in unfamiliar terrains for robotic vehicles. However, the large information content of such data makes interpretation of hyperspectral images time-consuming and user-intensive.

We propose the use of Isomap, a non-linear manifold learning technique combined with Expectation Maximisation in graphical probabilistic models for learning and classification. Isomap is used to find the underlying manifold of the training data. This low dimensional representation of the hyperspectral data facilitates the learning of a Gaussian Mixture Model representation, whose joint probability distributions can be calculated offline. The learnt model is then applied to the hyperspectral image at runtime and data classification can be performed.

1 Introduction

Hyperspectral image spectroscopy is an emerging technique that obtains data over a large number of wavelengths per image pixel over an area. Spectral analysis is the extraction of quantitative or qualitative information from reflectance spectra based on the wavelength-dependent reflectance properties of materials [Rencz, 1999]. Hyperspectral sensors are characterised by the very high spectral resolution that usually results in hundreds of observed wavelength channels per pixel of the image. These channels permit very high discrimination capabilities in the spectral domain including material quantification and target detection. The application of infrared reflectance spectroscopy with Short-Wave Infrared (SWIR, light from 1300 to 2500 nanometre in wavelength) also allows recognition of subtle mineralogic and compositional variation [Thompson *et al.*, 1999].

Currently, the process of hyperspectral analysis is user intensive, requiring a large amount of data analysis, and expert input. The hyperspectral data is often represented as a "cube" of information where the layers of the cube are the images at the spectral bands. Systems are available for the simultaneous viewing of this information in the spatial and spectral domains [Rencz, 1999]. The user is able to select points on the image and the program displays the spectrum at the specified location and the closest spectral match to it. The user then applies his/her own knowledge and other methods to interpret the data.

While most methods of spectral analysis require a large amount of knowledge and understanding in spectroscopy and field sites, there are some that do not and can be applied in a relatively straightforward manner. The most widely used of these methods is Principal Component Analysis (PCA) [Rencz, 1999]. PCA first calculates from the full data set a covariance matrix from which eigenvalues and eigenvectors are extracted and sorted according to decreasing eigenvalue. The k eigenvectors having the largest eigenvalues are chosen, giving the inherent dimensionality of the subspace of the original data [Duda *et al.*, 2001]. The amount of spectral variability contained in each component is given by the eigenvalue, and the relative proportion or contribution of each band to that component is given by the eigenvector. The method, however is linear and scene specific, and thus hard to port to different regions and environmental conditions.

Another method of analysing hyperspectral data is to use graphical models such as Bayesian Networks

*This work is supported by the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC) and the New South Wales State Government.

(BN) [Jensen, 2001] [Friedman *et al.*, 1997]. BNs allow the learning of a model that captures the relationship between the channels of hyperspectral data [Wang and Ramos, 2005] [Wang *et al.*, 2005]. The classification of new data can then be inferred from the model. It is, however, computationally intensive with a time complexity of $O(n^2 \cdot N)$, where n is the number of vertices and N is the number of nodes in the network, for learning a Tree Augmented Naive Bayes Network [Friedman *et al.*, 1997].

In this paper, we present a non-linear manifold mapping algorithm, Isomap, as described by Tenenbaum *et al.* [Tenenbaum *et al.*, 2000] in combination with statistical learning as a method for analysing hyperspectral data. The Isomap algorithm is able to reduce the high dimensional data into a low dimensional manifold efficiently. We further apply the Expectation Maximisation algorithm [Dempster *et al.*, 1977] to learn a BN representing the joint probability distribution of the relationships between the data, manifold and the labels, thus allowing us to classify the data.

This paper is organised as follows. In Section 2, we present the data collected for the study. In Section 3, we first describe the Isomap algorithm in detail. The probabilistic methods used for learning the relationship between the manifold learnt from Isomap and the hyperspectral data as well as how the labels are fit within the model is then discussed. Finally in Section 4, we present the experimental results and discuss the shortcomings of the algorithm, and any improvements that can be achieved in the future.

2 Data collection

A hyperspectral visible near-infrared and short wave infrared (0.4 - 2.5 μm) dataset was collected in Australia near Marulan, New South Wales, roughly 200km south west from Sydney. The dataset (Fig. 1) was collected using the 125 band HyMap instrument [Cocks *et al.*, 1998] at an altitude of approximately 1.5km covering approximately 150km^2 with an average resolution of 3.3m per pixel. The dataset was collected on 24th February, 2005 between 1200-1300 hours for maximum sunlight exposure.

The test area allows access to a wide range of rural conditions such as grass land, river, trees, farm animals, man-made objects like buildings, as well as wildlife areas. This gives us a good place to test for learning representations of natural objects in a rural environment.

The dataset was collected over the area because it is used extensively as a testing ground for both the air and ground robotic vehicles by our group. It is hoped that having air surveyed hyperspectral image data would give ground vehicles more information for navigation before deployment in an unknown natural environment where there is no geometrical shapes for easy identification.

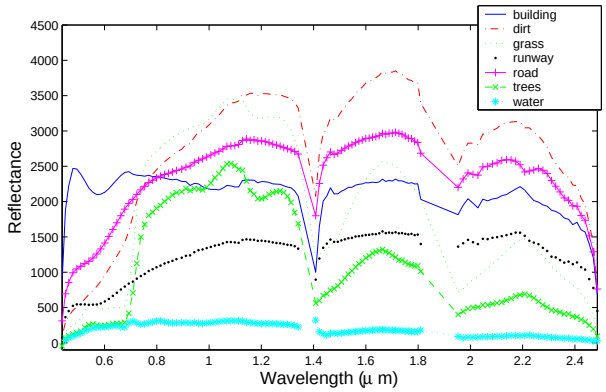

Figure 2: Spectra of the training data.

Training data is necessary for an autonomous system to learn from the typical data it might encounter during operation. For our purposes, we require a set of training data that takes into account trees, open grassland, rivers and dams, and man-made structures. This will, in turn, help plan and determine the traversability for a robotic vehicle.

We pick our training data from the swath shown highlighted by the red rectangle in Fig. 1. Pixels of the following objects were gathered from the image and manually labelled for training data:

1. buildings
2. dirt
3. grass
4. road (dirt surface)
5. runway (bitumen surface)
6. trees
7. water (from the dams and the river)

Approximately 1300 data points for training were collected, and Fig. 2 shows the different spectra of the training data for a typical pixel representative of each class.

3 Approach to the Problem

Our approach to the problem of classifying the hyperspectral data into their respective classes is to use a combination of non-linear dimensionality reduction with a probabilistic graphical model. The procedure for analysing the hyperspectral image is:

1. Select pixels from the image for training and testing data.
2. Apply Isomap on the training data to find the embedded manifold.

Figure 1: (left) Hyperspectral image data of the Marulan area, north is up, image is approx. 16 km across. (right) Study region as enclosed by the rectangle on the right.

3. Build a BN to learn the mapping of this data to the manifold, and the classes of the manifold.
4. Test the classification on the testing data.

3.1 Nonlinear Dimensionality Reduction

On a hyperspectral image, each pixel contains data gathered on hundreds of spectral channels across the electromagnetic spectrum. Each pixel thus can be treated as a point in high dimensional space, to classify this data using a BN would require a node for each channel [Wang and Ramos, 2005] [Wang *et al.*, 2005], resulting in a large and complex network. However, by applying appropriate dimensionality reduction techniques, we can reduce this high dimensional hyperspectral data into a much lower dimensional space, requiring only a few BN nodes. This in turn reduces computational time for both learning and inference.

Nonlinear dimensionality reduction (NLDR) techniques find an intrinsic low dimensional structure embedded in a high dimensional observation space. We choose to use a technique, *isometric feature mapping*, or Isomap, as described by Tenenbaum *et. al.* [Tenenbaum *et al.*, 2000]. It reliably recover low dimensional nonlinear structure in realistic perceptual data sets.

The Isomap method formulates the NLDR as the problem of finding a Euclidean feature space embedding of a set of observations that explicitly preserves their intrinsic metric structure; the metric structure is quantified as the distance between the points along the manifold.

The sensor data $\mathbf{Z}$ is assumed to lie on a smooth nonlin-

ear manifold embedded in the high dimensional observation space. It constructs an implicit mapping $f : \mathbf{Z} \rightarrow \mathbf{X}$ that transforms the data $\mathbf{Z}$ to a low dimensional Euclidean feature (state) space $\mathbf{X}$, which preserves the distances between observations as measured along geodesic paths on the manifold. In summary, the steps of the algorithm are [Tenenbaum *et al.*, 2000]:

1. Find the K nearest neighbours of all data;
2. Compute the distance matrix that contains the shortest path between all pairs of neighbours;
3. Construct d-dimensional embedding by finding the eigenvalues and eigenvectors of the inner product of the distance matrix.

Once a manifold is found, we can use graphical models to learn the mapping from high dimensional hyperspectral data to the low dimensional manifold.

3.2 Probabilistic Methods

The Isomap algorithm and indeed most NLDR algorithms are inherently *deterministic*, i.e. they do not provide a measure of *uncertainty* of the underlying states of the high dimensional observations. Furthermore, when applying any dimensionality reduction algorithm, the resulting manifold will be scene specific.

Therefore, to capture the uncertainties of the data, another method needs to be employed to learn the mapping from the high dimensional data to the low dimensional manifold. This learnt relationship can then be used for inference of the manifold of data from different areas, thus classify these new data. A probabilistic method can not only perform the classification, but can also encapsulate the uncertainties inherent in the low dimensional state inferenced from noisy high dimensional observations.

Bayesian network

We chose to use a graphical model, specifically a Bayesian Network [Jensen, 2001], to model the relationship between the high dimensional data and the low dimensional manifold. Bayesian Networks are graphical representations of multivariate joint probability distributions that exploit the dependency structure between distributions, describing them in a compact and natural manner [Friedman and Koller, 2003].

Figure 3 shows the model used to learn the mapping between the observation, $\mathbf{Z}$, and the manifold, $\mathbf{X}$ as employed by Kaupp *et. al.* [Kaupp *et al.*, 2005]. Both variables are represented as multi-dimensional Gaussian nodes. The edge from $\mathbf{X}$ to $\mathbf{Z}$ defines the relationship of $P(\mathbf{Z}|\mathbf{X})$, i.e. the sensor model, as assumed by Isomap: that the observed data lies on the manifold.

Two other nodes are present in the model: S and L, both are represented by discrete probability distributions. Therefore, the joint distribution of $P(\mathbf{X}, S)$, for example,

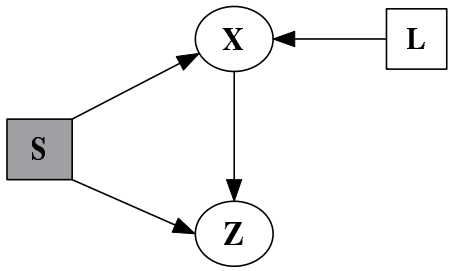

Figure 3: Bayesian Network used to learn the mapping and the classification of the hyperspectral data.

is a Gaussian mixture model, and the size of these nodes represent the number of components in the mixture.

The node S corresponds to a spatial region on the manifold over which a mixture component is representative. This representation conveniently handles highly nonlinear manifolds through the ability to model the local covariance structure of the data in different areas of the manifold. This node is shaded to show that it is a hidden variable, since this node is not observed during the learning process.

Lastly, the node L is added to show the label or the class of the data. The label node points to the manifold data to show that $\mathbf{X}$ is dependant on L. This node is added because the number of components in S can be different from the number of classes represented by the data; some classes can consist of several Gaussian clusters on the manifold, or vice versa.

The joint probability distribution of all the random variables in the model shown in Fig. 3 is expressed as:

$$P(\mathbf{z}, \mathbf{x}, s, l) = P(\mathbf{z}|\mathbf{x}, s)P(\mathbf{x}|s, l)P(s)P(l) \qquad (1)$$

where: $\mathbf{z} \in \mathbf{Z}$, $\mathbf{x} \in \mathbf{X}$ and the dependencies are given by:

$$P(\mathbf{z}|\mathbf{x}, s) = \frac{1}{(2\pi)^{D/2}|\Psi_s|^{1/2}} \times \exp\left\{-\frac{1}{2}[\mathbf{z} - \Lambda_s\mathbf{x} - \mu_s]^T\Psi_s^{-1}[\mathbf{z} - \Lambda_s\mathbf{x} - \mu_s]\right\} \qquad (2)$$

and

$$P(\mathbf{x}|s, l) = \frac{1}{(2\pi)^{d/2}|\Sigma_{s,l}|^{1/2}} \times \exp\left\{-\frac{1}{2}[\mathbf{x} - \nu_{s,l}]^T\Sigma_{s,l}^{-1}[\mathbf{x} - \nu_{s,l}]\right\} \qquad (3)$$

For the classification problem, we need to learn the parameters of these distributions. The various parameters are: The prior probabilities $P(s)$, which follows a multinomial distribution; The probabilities $P(l)$, which is a vector of $|L|$ in size; For the conditional probabilities

described above, we have the mean vectors $\nu_{s,l}$ and μ_s, the full covariance matrix $\Sigma_{s,l}$, the diagonal covariance matrix Ψ_s and the loading matrix Λ_s.

EM Algorithm

When all the variables in a model are observed during the learning process, then we can use the Maximum Likelihood to estimate the parameters. In this model, however, S is not observed, thus *Expectation Maximisation* (EM) [Dempster *et al.*, 1977] is used to learn the parameters of the network. The EM algorithm provides a general approach to maximum-likelihood parameter estimation when the observation has incomplete data. Ghahramani and Hinton [Ghahramani and Hinton, 1996] showed the general solution of EM for such mixture of factor analysers.

Inference

Once all the parameters are estimated off-line, we can apply the model to new data online to find the classifications of the pixels on the new image. The classification of the new pixel can be found by the marginalising the joint distribution of Eqn. 1 to give $P(l)$ [Jensen, 2001], where each value of $P(l_j)$ gives the probability of the pixel being in class j:

$$P(l) = \sum_{\mathbf{z},\mathbf{x},s} P(\mathbf{z} = \mathbf{z}_i|\mathbf{x}, s)P(\mathbf{x}|s, l)P(s)P(l) \qquad (4)$$

4 Results and discussion

4.1 Mapping the Training Data

We take the training data as described in Section 2 and apply the Isomap algorithm to find the embedding manifold. The dimensionality of this manifold can be determined by the residual variance, as shown in Fig. 4, at each dimension of the mapped data. From this plot, we can deduce that the first three dimensions of Isomap result would be a good description the hyperspectral data.

Figure 5 and 6 show the first three dimensions of the manifold. We can see that most of the classes of the training data are separated on the manifold. This manifold is calculated by setting $k = 14$ in step 1 of the Isomap algorithm as outlined in Section 3.1. We found that below this value the resulting manifolds vary.

There are a couple of overlapping classes in the manifold: Firstly, a small proportion of the class 3 (grass) and class 6 (trees), which can be seen clearly in the top of Fig. 6. Secondly, we can see that some of the data for class 2 (dirt) and class 4 (road) share similar spaces. This is due to the fact that the road presented here is made of dirt. These cases could cause potential problems of misclassification later.

Therefore, from these two manifolds, we can see that despite the similarities between the IR spectra of the

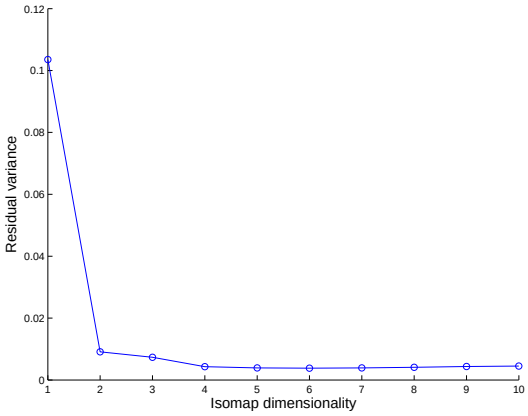

Figure 4: The residual variance of the hyperspectral training data as presented.

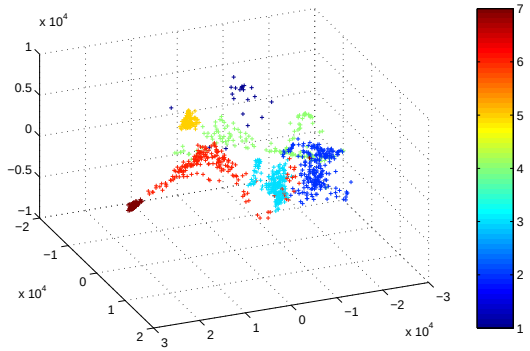

Figure 5: The mapping of the training data in 3D, the axes are the first three eigenvectors of the manifold.

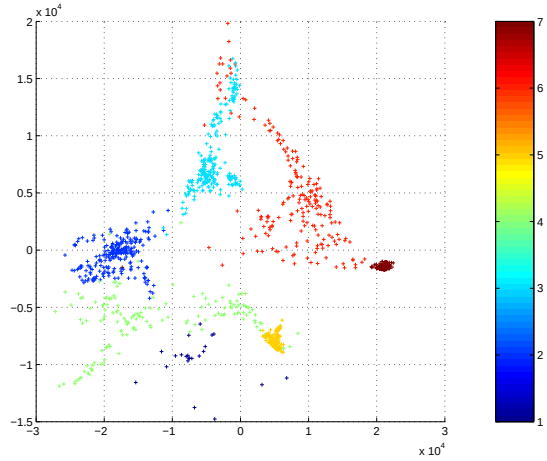

Figure 6: Top view of the 3D mapping, i.e. showing here the first two eigenvectors. As with Fig. 5, each colour represent a distinct training class as listed in Section 2.

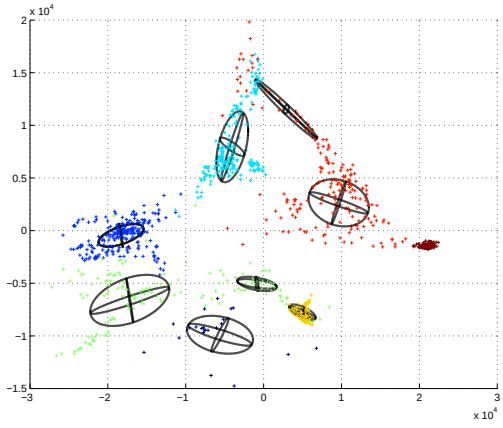

Figure 7: Top view of the clustering of the mapping 3d data.

data, and most importantly, the appearance of the individual pixels for various class objects under a visual sensor, Isomap can separate most of the classes in an unsupervised manner.

4.2 Classification of the Testing Data

We set the dimensionality of the nodes in the network as shown on Fig. 3 as: $|\mathbf{X}| = 3$ as discussed above, $|\mathbf{Z}| = 125$ for the observed hyperspectral data, and $|L| = 7$ for the number of classes listed.

The choice of the best number of mixture components for the node $\mathbf{S}$ is crucial to the off-line computation for the model. We chose the following approach to address the over-fitting concerns: We chose a threshold for the minimum value any $P(s)$ should take. If the computed probabilities $P(s)$ are all significantly greater than the threshold, the model is refined. If any resulting $P(s)$ is smaller than the threshold, the number of components will then be reduced. We found that $|S| = 10$ best described the state space, Fig. 7 shows the 10 components of the embedding using k-means clustering [Duda *et al.*, 2001].

Using the training data gathered, we can learn the parameters of the Bayesian network shown in Fig. 3 by applying the EM algorithm. From the data set in Fig. 1(b), we gather approximately another 2700 data points data for testing. The test data are: The test data are shown in Table 1. These are the only classes used due to their sufficient sample size in the test image.

Table 1 shows the result from running the graphical model on the test data. Here, a pixel is classified as class 'x' if it has the maximum probability among all the classes. From this table, we can see that for the test data of water and trees, this classifier has very good results with correct classification in the high 90's. The classification of the road in the image is still fairly good at almost 89%. Most of the rest of the pixels are classified as 'dirt', which is what the road is composed of.

The test data for grass has the worst classification results here, where less than half of the test data are correctly classified while most of the data are labelled as 'trees' by the classifier. From Fig. 2 of the spectra samples of the data we can see that the respective spectra of the two data are very similar. Figures 5 and 6 confirm these two classes have some of their mapped data very close on the manifold. Therefore, the classification of hyperspectral 'grass' data using only pixels do not yield results with high correction rates. It could be possible to improve this using a small patch of the image data thus giving the classifier more information at one time.

4.3 Classification of the Image Data

Figure 8 shows the classification result of the entire test image data. In comparison with the colour image of the test data in Fig. 1, we can see that most of the area has been classified correctly.

In the last section, we found that the results for the test data for grass have a very low correct classification rate, however, sub-figure (a) of the image result shows otherwise. In comparison with the colour image of the test data on Fig. 1, we can see that nearly all of the grass area in the image have been correctly classified. The only major discrepancy in the result is the patch on the top right corner, between rows 200 and 400. In sub-figure (c), we see that these pixels have been classified as trees instead. We believe this is because the spectra for these particular pixels are very different from the rest of the grass data and are more similar to the tree data. Furthermore, as they were not originally included in the training data, the classifier's result grouped the pixels into the 'trees' label. An incremental learning method or using small image patches, which take advantages of correlations between the neighbouring pixels, might correct this error.

Of the other results, sub-figure (b) shows the roads clearly through the image. There are patches of light gray areas, especially on the lower half of the image. These are the background dirt signatures that are a part of the bush areas. Since the road is composed of dirt, some of these pixels do result in approximately 20% probability of being classified as 'road'. Only very few individual pixels scattered around the area have high probabilities of being 'road', again we believe this can be rectified by learning and inferencing small image patches instead.

Sub-figure (c) shows the result for 'trees'. The large patch in the bottom half of the image is the forest/bush as can be seen in Fig. 1. On the top half, we can clearly see the trees that grows on the bank of the river. We can also see the small patches of trees around (250, 400) that is a part of the farmer's house.

Table 1: The resulting classification of the test data.

Test Data			Resultant Class Label Number of data classed as (% of total data)						
Class	Name	Total	1	2	3	4	5	6	7
3	grass	812	-	2 (0.246)	**380 (46.8)**	1 (0.123)	-	429 (52.8)	-
4	road	440	1 (0.227)	32 (7.27)	3 (0.682)	**391 (88.9)**	1 (0.227)	-	12 (2.72)
6	trees	647	2 (0.309)	1 (0.155)	16 (2.473)	2 (0.309)	-	**626 (96.8)**	-
7	water	784	-	-	-	-	-	16 (2.04)	**768 (98.0)**

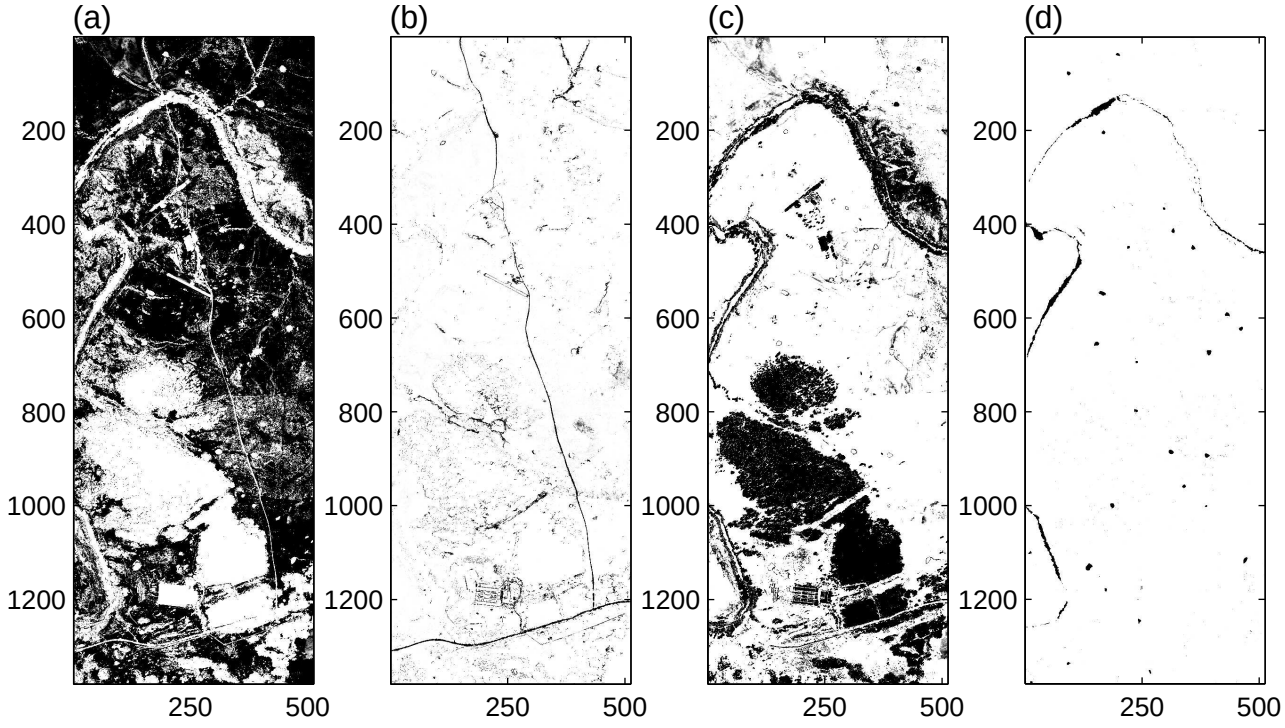

Figure 8: Result of the inference for the label node of the graphical model on the image data as shown on Fig. 1(b). The sub-figures are: (a) Grass, (b) Road, (c) Trees, and (d) water. Black indicates a 100% certainty of classified data and white 0%. The axes are the pixel numbers of the image.

The water result in sub-figure (d) shows clearly the river through the image. Furthermore, it shows scattered around the image small areas of water. These are the dams on the farm, most of which can be seen clearly on the colour image.

5 Conclusion

In this paper we have investigated the application of a dimensionality reduction technique, specifically that of Isomap, in combination with probabilistic statistical learning to analyse hyperspectral data. We found that the manifold embedding is fairly consistent and able to group the classes together. By applying probabilistic methods, we can then learn a model of the embedding, and thus use the learnt model to classify the new data. We found in most cases when applied to single pixels, the classification rate has high accuracy. However, when two classes are very similar in spectral appearance, the classification can be poor.

References

[Cocks *et al.*, 1998] T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields. The HyMap airborne hyperspectral sensor: the system, calibration and performance. In *1st EARSeL Converence*, pages 37–42, 1998.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–39, 1977.

[Duda *et al.*, 2001] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., New York, 2001.

[Friedman and Koller, 2003] Nir Friedman and D. Koller. Being Bayesian about network structure: A bayesian approach to structure discovery in Bayesian Networks. *Machine Learning*, 50:95–126, 2003.

[Friedman *et al.*, 1997] Nir Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[Ghahramani and Hinton, 1996] Zoubin Ghahramani and Geoffrey E. Hinton. The EM algorithm for mixture of factor analyzers. Technical report, Department of Computer Science, University of Toronto, May 1996.

[Jensen, 2001] F. V. Jensen. *Bayesian Networks and Decision Graphs.* Springer-Verlag, New York, 2001.

[Kaupp *et al.*, 2005] Tobias Kaupp, Alexei Makarenko, Suresh Kumar, Ben Upcroft, and Stephan Williams. Humans as information sources in sensor networks. In *Proceedings of the 2005 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Edmonton, Alberta, Canada, August 2005.

[Rencz, 1999] A. N. Rencz, editor. *Remote Sensing for the Earth Sciences.* John Wiley & Sons, Inc., third edition, 1999.

[Tenenbaum *et al.*, 2000] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–23, December 2000.

[Thompson *et al.*, 1999] A. J. B. Thompson, P. L. Hauff, and A. J. Robitaille. Alteration mapping in exploration: application of short-wave infrered (SWIR) spectroscopy. *Socity of Economic Geologists Newsletter*, 39, October 1999.

[Wang and Ramos, 2005] X. Rosalind Wang and Fabio T. Ramos. Applying structural EM in autonomous planetary exploration missions using hyperspectral image spectroscopy. In *Proceedings 2005 International Conference on Robotics and Automation*, Barcelona, Spain, April 2005.

[Wang *et al.*, 2005] X. Rosalind Wang, Adrian J. Brown, and Ben Upcroft. Applying incremental EM to bayesian classifiers in the learning of hyperspectral remote sensing data. In *Proceedings of The Eighth International Conference on Information Fusion*, Philadelphia, PA, USA, July 2005.