PROJECT 2

ECE759 Pattern Recognition-Spring 2016

Laura Gonzalez
Charles West
Selene Schmittling

April 23, 2016

1 Introduction

In this project we utilize the dataset used in Project 1. This data set includes two groups of images: one set are labeled "Water" images and the other is labeled "Non-water" images. In project 1, the images were manually cropped and water pixels labeled for both sets of images.

In this project we are tasked with

1. defining an architecture for and running a Neural Network for classification;
2. defining appropriate parameters for and implementing a support vector machine.
3. identifying feature vectors that can describe water vs. non-water and using techniques to determine if the feature vectors are "good".
4. proposing an unsupervised classification technique to classify regions into two clusters.

In this section, we briefly describe the techniques we explored in general. In the Methodology section, we describe our specific implementations and in Results, we describe the results of our systems and compare them to the Bayes classifier we used in the first Project.

1.1 Neural Network

1.2 Support Vector Machine ("SVM")

Support Vector Machines are another technique used to classify data. When classifying linearly separable data, SVM maximizes the distance between the two classes which improves generalizability for new data. In essence, it creates the maximum buffer between the classes so that testing points are more accurately assigned to their appropriate class. While SVM can be

1.3 Feature Selection

Feature selection is a key process in classifying data. Some features will be better suited to a classification problem then others. Ideal features will linearly separate the classes such that data points in the same cluster are close together and data points in different clusters are far apart. Even if the features are not linearly separable, it will be best if they are able to separate the classes in space.
One of the issues involved in trying to identify whether features are well chosen is that the vectors are frequently larger than 3 in dimension which means there is no way to visualize the data points in relation to each other. One of the measures that can be used to identify how well spaced data points are is to use scatter matrices and calculations which use them. Scatter matrices quantify the between cluster and within cluster distance. Feature vectors for which the within cluster distance is very small (low variance) and for which between cluster distance is large (i.e., the data is very separated) is desirable.
Additional information can be gained from calculations with utilize information in these matrices. [DESCRIBE J1-J3].
Another tool that can be used to deal with feature vectors is Principal Components Analysis (PCA). PCA is a tool which quantifies the axis along which maximal variance in the data points is found. It is possible to identify these axes and then identify a basis for the data which more separates the data.

1.4 Unsupervised classification

Unsupervised classification is used in cases where labels are not available. Unsupervised methods include: [list]. Each of these methods requires the choice of certain parameters. It is not always clear what the values of the parameters should be. Another issue with unsupervised classification is in identifying what the generated clusters mean. It is frequently difficult to describe the significance of the clustering. This significance will rely heavily on the data that is used for clustering. For instance. clusters of gene expression data has been associated with genes which are involved in the same process (co-expressed). It is unclear what the clusters mean for color spaces. However, when we clustered water pixels exclusively in Project 1, it seemed as if the clusters corresponded to "shadows".

2 Methodology

2.1 Neural Network

We implemented the Neural Network using C++ and relevant libraries including OpenCV and [Neural Network library].

2.1.1 Architecture Choice

2.1.2 Implementation

2.2 SVM

We implemented the SVM using OpenCV's SVM.

2.2.1 Parameter Choice

2.2.2 Implementation

2.3 Detecting and Classifying "Water" vs. "Non-water" Regions

2.3.1 Features Chosen

All work on features was performed in MATLAB, in some cases in the IBM Virtual Computing Lab. We had only the Image Processing toolbox available. For this reason, we focused on color spaces and gradient. In Project 1 we focused exclusively on the RGB ("Red/Green/Blue") color space. We decided to see if other color spaces could provide additional useful information. We also decided to include the magitude and direction of the gradient. Since areas of water generally don't have a lot of edges, we hoped that this would help identify water. Color spaces used were RGB, HSV (Hue/Saturation/Value) and YCbCr which provides luminance (Y) and chrominance (Cb and Cr) information.
In order to see if we could find dimensions which most explained the variance in the data, we applied Principal Components Analysis using MATLAB's pca() function.
We calculated the scatter matrices on the data to determine whether the data points were separated by "water" vs. "non-water".

2.3.2 Principal Components Analysis

2.3.3 Scatter Matrices

3 Results

3.1 Neural Networks

3.2 SVM

3.3 Detecting and Classifying "Water" vs. "Non-water" Regions

4 Discussion