Anomaly Detection Challenges - Challenge I

Hamza Tahir (03670002) and Muhammad Hamza Usmani (03669506)

Technical University of Munich

1 Introduction

This brief report serves as a purpose to present and explain the methodologies applied to tackle the first challenge in the Practical: Anomaly Detection Challenges. Section 2 discusses the challenge task and the data set for the machine learning/anomaly detection task. Section 3 explains the approaches adopted for the task.

2 The Challenge

Machine Learning Task The machine learning task for this challenge is to determine if given soil sample is that of normal soil or a cotton soil. The task is a supervised learning, binary classification with two decision classes:

Table 1. Decision Classes

Class	Numeric Representation
Normal Soil	0
Cotton Soil	1

The Data set The dataset of the challenge consisted of a subset of the Satellite Imagery dataset originally published by Ashwin Srinivasan. The training set consists of 4435 samples while the test set consists of 2000 samples. Samples have 36 features, that characterize spatial neighborhood and spectral band characteristics. 70% of training samples have missing values for randomly chosen feature values.

3 Methodology

This section explains the data analysis and the machine learning process to build the model for classifying the given samples as normal or cotton soil samples.

3.1 Data Preprocessing

The given dataset has missing values some features in many samples, to build a generic classifier it is therefore important to treat the missing values to build a robust classifier that is relatively generic. The authors used the following techniques to deal with the missing values in the dataset:

Feature Mean: The missing values were imputed by average of values. Different averaging schemes were adopted, under first scheme as proposed by Runkler [1], the missing value of a particular feature is replaced by average value of of the feature. This averaging technique is generic to replace missing values for all kinds of datasets.

Feature Maximum: In this technique, missing values can be treated as outliers and thus they can be imputed with maximum possible values [1]. Under this scheme, missing values are replaced by maximum value of the feature. This is exactly the same as the Feature Mean technique, only difference is replacing average with the maximum feature value.

Nine-Neighborhood Mean: Another spatial average based technique used is 9-neighborhood averages, considering the given dataset comprises of four different observations in a nine-neighborhood, missing values were also replaced by averages in a particular nine-neighborhood.

Immediate Neighborhood Mean: This approach is similar to the Nine-Neighborhood one but here image locality is leveraged to better replace missing values. The underlying assumption is that closely related pixels have similar values. Therefore, rather than using the entire 9 neighborhood, only 'immediate' neighbors are used for the average value. 'Immediate' neighbors are defined as the pixels that are to the left, right, top and bottom of the 9-neighborhood. Figure 1. summarizes all imputation techniques:

3.2 Machine Learning Techniques

Following machine learning techniques were used to classify given samples as "Cotton Soil" or "Normal Soil":

1. Naive Bayes
2. Support Vector Classifier (SVC)
3. K-Nearest Neighbors (KNN)
4. Decision Trees
5. Neural Network - Multi Layer Perceptron
6. Random Forest

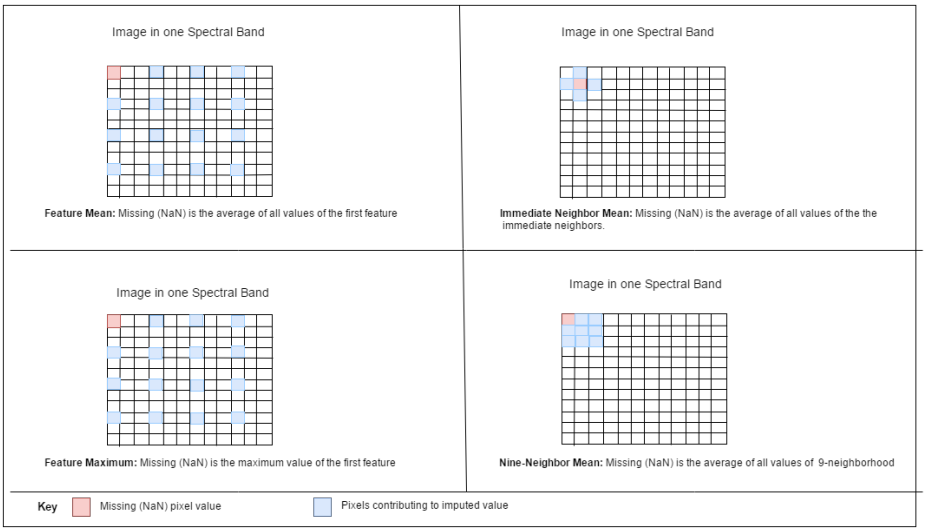

Fig. 1. Missing Value Schemes

3.3 Generalizing - Developing a Robust Classifier

To build a robust classifier, that is relatively general and is not restricted only to the given training set following techniques were used:

Three-fold Cross Validation: Three-fold cross validation was used to overcome the problem of over-fitting, and to build a model to that will generalize to an independent dataset, Hawkins et. al. (see [2]).

Feature Weighing: Another technique used to improve accuracy and to generalize the classification model was to weigh the features. Under the assumption that not all features contribute equally to determine the class of a sample, different features can have different weights that represent their relative importance to the classification model. In particular, we could not be certain that a certain spectral band was more or less important for the classification of cotton soil. It might be the fact that the green spectrum was more important, or maybe the infra-red spectrum.
Feature weighing is a complex technique. In our model we used basic techniques based upon correlation to determine weights of the features. First, we extracted each feature out individually and trained our classifiers on each separately. The feature that had the lowest accuracy after 3-fold validation was then completely removed from the training of the final model. This way, we fine-tuned our model to better estimate the classification.

4 Results

Results of the challenge are summarized in this section. We present two basic results. One is with out classifiers trained on all features, without weighing. Second with our classifiers trained with only three features (first, second and fourth), therefore giving 0 weight to the third feature.

Table 2. Best Average Cross-Validation Accuracies (without feature weighing)

Training Accuracies				
Technique	Feature Mean	Feature Max	Nine-Neighbor	Immediate-Neighbor
Decision Trees	97.519	98.528	96.550	97.723
Multi-Layer Perceptron	96.280	97.768	97.768	98.016
Naive Bayes	98.422	96.979	98.535	98.512
KNN	98.858	96.731	98.557	98.783
SVC	97.813	94.295	97.858	98.377
Random Forest	98.647	97.610	97.970	98.715
Average	**97.796**	**96.985**	**97.873**	**98.354**

Table 3. Best Average Cross-Validation Accuracies (with feature weighing)

Training Accuracies				
Technique	Feature Mean	Feature Max	Nine-Neighbor	Immediate-Neighbor
Decision Trees	97.994	97.002	97.610	97.566
Multi-Layer Perceptron	97.250	95.942	97.565	97.632
Naive Bayes	98.377	96.397	98.557	98.512
KNN	97.881	96.619	98.693	98.85
SVC	97.565	94.656	98.197	98.377
Random Forest	98.625	98.197	98.715	98.625
Average	**97.948**	**96.468**	**98.222**	**98.260**

As one might expect, the average methods that leveraged the spatial locality of the image pixels proved to have the highest training accuracy in all the classifiers. Feature weighing on the other hand, had negligible effect on the result of the training accuracies.
The best classifier overall proved to be the KNN model, followed closely by the Random Forest approach. This also makes sense, as these classifiers are overall robust to high variance and high dimensional data.

References

1. Runkler, T: Data Analysis Models and Algorithms for Intelligent Data Analysis. Springer, 24–25 (2011)
2. Hawkins D. , Basak S. , and Denise M. Assessing Model Fit by Cross-Validation J. Chem. Inf. Comput. Sci., 2003, 43 (2), pp 579586 (2003)