Ensemble Feature Selection for Hyperspectral Imagery

Gidudu, A., Abe, B. and Marwala, T.

School of Electrical and Information Engineering
University of the Witwatersrand, 2050, South Africa
Anthony.Gidudu@wits.ac.za, Bolanle.Abe@students.wits.ac.za, Tshilidzi.Marwala@wits.ac.za

Abstract

The process of relating pixels in a satellite image to known land cover classes is referred to as image classification. As demonstrated in this paper, ensemble feature selection offers a unique approach to land cover mapping, especially for very high dimensional hyperspectral data. Ensemble classification is premised on ensuring diversity among the base classifiers and adopting appropriate means of combining their outputs into a single classification result. This paper explores ensemble feature selection as a means of ensuring diversity for land cover mapping of hyperpectral data. Results show that random selection of features (bands) yielded the best results as compared to building base classifiers depending on search algorithms or as was used in this case, sequentially arranging the features into base classifiers. Of the combination techniques, the single best technique yielded better results than majority vote, however in most cases the difference between the results was not significant.

1. Introduction

The extraction of land cover information from satellite imagery has been one of the major beneficiaries of developments in machine learning. Techniques such as support vector machines, neural networks, fuzzy logic, genetic algorithms etc. which have taken root in remote sensing studies owe their origin to advancements in computational intelligence (and by extension machine learning). Ensemble classification is another such technique that has taken root in machine learning and is making inroads in image classification for land cover mapping. In literature, ensemble classification goes by several names such as multiple classifier systems, committee of classifiers, mixture of experts and ensemble based systems (Polikar [1]). The essence of ensemble classification is to have a final classification result ‘in consultation’ with a group of classifiers. It is akin to getting a second, third (or more) opinion about a financial, medical or social decision one may have to make (Polikar [1]). For ensemble systems to perform effectively, it has been shown that the constituent base classifiers need to have diversity in their predictions (Opitz [2]; Tsymbal et al. [3]; Foody et al. [4]). One way of ensuring diversity in ensemble classifiers is in the use of different feature subsets or so called ensemble feature selection (Tsymbal et al. [3]). In land cover mapping this would entail having an ensemble of different spectral band combinations and having the final classification result based on a pre-stipulated ‘consensus’ (e.g. plurality vote) of the different band combinations (Chen et al. [5]).

Previous work on the application of ensemble classification to land cover mapping has focused mostly on ensuring diversity by using different classifiers. In this paper, the application of ensemble feature classification is explored on hyperspectral data. Hyperspectral data by its nature consists of a very high dimensional feature space (e.g. 200 bands/features) and presents an ideal situation to explore the use of ensemble feature classification for land cover mapping. This paper is organized as follows: section 2.0 gives an overview of ensemble classification, section 3 briefly discusses support vector machines which are the classifiers of choice in this research, section 4 presents the developed methodology in executing
this work, while sections 5 and 6 highlight the results, discussions and conclusions thereof.

2. Ensemble Classification

As alluded to before, ensemble classification is a multiple classifier system in which the aim is to combine the outputs of several classifiers in order to derive an accurate classification (Foody et al. [4]). There is a general consensus that ensemble classifiers yield favorable results compared to those of single systems for a broad range of applications (Bruzzone et al. [6]). However, Foody et al. [4] and Liu et al. [7] argue that whereas the adoption of an ensemble based approach may typically yield a classification with an accuracy that is higher than that of the least accurate classifier used in the ensemble, it may not necessarily be better than each of the base (constituent) classifiers.

Polikar [1] raises a number of reasons justifying the need for an ensemble approach. One reason is that, combining the output of several classifiers by say averaging may reduce the risk of an unfortunate selection of a poor performing classifier (Polikar [1]). Polikar [1] further states that whereas the final output of the ensemble may not beat the performance of the best classifier in the ensemble, it certainly reduces the risk of making a particularly poor choice. In his second justification, Polikar [1] proposes an ensemble approach in the face of large volumes of data, especially in cases where the amount of data may be too large to be handled by a single classifier. Partitioning the data into smaller subsets and training different classifiers with different portions of data and combining the outputs using an intelligent combination rule could potentially prove to an efficient approach (Polikar [1]).

The functionality of ensemble systems involves generating the individual base classifiers and devising a means of combining the outputs of these base classifiers. One way of ensuring improved performance of the ensemble system is to ensure that the individual classifiers make errors differently (Polikar [1]). The premise is that if each classifier makes errors differently, i.e. that there is diversity among the base classifiers, then a strategic combination of these classifiers can reduce the total error (Polikar [1]). Diversity in ensemble systems can be achieved through using different: training datasets, classifiers, features or training parameters (Polikar [1]). Chen et al. [5] categorizes ensemble classification into those based on several different learning algorithms and those based on just one. The former involves using several classifiers on the same dataset. The drawback of this ensemble system is to have to handle different classifiers which increases the complexity of the processing (Chen et al. [5]). In the second categorization, only one classifier is used and the ensemble is created by changing the training set. Two popular examples of this include bagging or bootstrap aggregating (Breiman [8]) and Adaboost or reweighting boosting (Freund et al. [9]).

Under the second categorization is an effective approach for generating an ensemble system by the use of different feature subsets or the so called ensemble feature selection (Opitz [2]). Varying the feature subsets used to generate the base classifier potentially promotes diversity since the classifiers tend to err in different subspaces of the instance space (Oza et al. [10]; Tsymbal et al. [3]). Some of the techniques used to identify features to be used in ensemble systems include genetic algorithms (Opitz [2]), exhaustive search methods and random selection of feature subsets (Ho [11]).

Equally important to the generation of an ensemble is how the base classifier outputs are to be combined. There are two basic approaches in literature which have been suggested as means of integrating ensemble output (Tsymbal et al. [3]): a combination approach and secondly a selection approach. A range of methods are available for the combination of information from multiple classifiers (Giancinto et al. [12]; Valentini et al. [13]; Huang et al. [14]). Some of the methods include majority voting (Chen et al. [5]), weighted majority voting (Polikar [1]) or more sophisticated methods like consensus theory (Benediksson et al. [15] and stacking (Džeroski et al. [16]). A number of selection approaches have also been proposed to solve the integration of ensemble data (Tsymbal et al. [3]). One of the most popular and simplest selection techniques is Cross Validation Majority (CVM) also called single best. In CVM, the cross
validation accuracy for each base classifier is estimated using the training set and then the classifier with the highest accuracy is selected Tsymbal et al. [3].

3. Support Vector Machines

Support Vector Machines (SVMs) are a supervised classification technique having their roots in Statistical Learning Theory. Given a training dataset, the decision boundary between the individual classes is a linear discriminant placed midway between the classes and is expressed as (Foody et al. [4]): \( f(x) = \text{sign} \left( \sum_{i=1}^{r} \alpha_i y_i k(x, x_i) + b \right) \) where \( y \) defines the classes, \( \alpha_i \) for \( i = 1, 2, \ldots, r \) are the Lagrange multipliers, \( b \) is bias and \( k(x, x_i) \) is a kernel function. In many practical cases, a linear discriminant between the training data classes is not feasible and in order to cater for this nature of data, it is nonlinearly projected into a higher dimension space using the kernel \( k(x, x_i) \). Placing a linear discriminant in this high dimension feature space is equivalent to placing a nonlinear discriminant in the previous space. Examples of kernels that can be used for this purpose include: polynomial, sigmoid and Gaussian functions. For each kernel, corresponding parameters are obtained through cross validation before the eventual classification. A more detailed treatise of SVMs can be found in references such as Vapnik [17], Christianini et al. [18] and Campbell [19].

4. Methodology

4.1 Data Description

The hyperspectral data used in this paper was sourced from the AVIRIS sensor [20] and represents Indiana’s Indian Pines in the United States of America. It is a freely accessible online dataset which comes with accompanying ground truth data. Of the 224 bands, 4 were discarded because they contained zeros and of the remaining bands only 180 were used in this research. The rest of the bands were left out because of being affected by atmospheric distortion (Bazi et al. [21]). The classes of interest included; alfalfa, corn-notill, corn-minimum till, corn, grass/pasture, grass/trees, grass/pasture-mowed, hay-windrowed, oats, soybeans-notill, soybeans-minimum till, soybean-clean, wheat, woods, building-grass-tree-drives, stone-steel towers. These classes were selected in reference to the ground truth data.

4.2 Research Design

Based on Chen et al. [5]’s categorization, this paper focused on the ensemble approach dependent on one learning algorithm (In this case Gaussian SVMs), with diversity being enforced through using different feature (band) combinations. Two ensemble feature selection techniques were used namely exhaustive search and random selection of feature subsets. The evaluation function for the exhaustive search was the Bhattacharyya Distance separability index (Bhattacharyya [22]). The results of the base classifiers in each ensemble were combined using two methods; majority voting and an adaptation of Cross Validation majority (CVM) also called single best. In CVM, cross validation data is used as a basis for selecting the best out of the whole ensemble. In this paper, this was modified to consider the final classification results of each base classifier instead. For comparison, another ensemble was derived by sequentially grouping subsequent bands into 10 base classifiers. i.e. bands 1-18 made up the first base classifier, bands 19 – 36 the second base classifier etc, making a total of 10 base classifiers for all the 180 bands.

For each base classifier and corresponding ensemble, classification was carried out in MATLAB with the results being imported into IDRISI Andes for data integration and generation of a land cover map. Classification accuracies were then calculated for each derived land cover map, by making comparisons between the predicted output from the base and ensemble classifiers and the ground truth data. These results were then used as the basis upon which to evaluate ensemble feature classification and its corresponding effect on land cover mapping.

5. Results and Discussions

The Table 1 shows the results of the different ensembles considered. The classification accuracy is given in terms of the Kappa coefficient of agreement (Cohen [23]), which is a measure of
how well the derived map compares with ground truth data. It ranges from 0 to 1 with 0 implying no agreement between predicted land cover and ground truth, and 1 indicating complete agreement.

All the ensembles had ten base classifiers, the figure ten having been arbitrarily chosen. The base classifiers in Ensembles 1, 2, 3 and 4 consisted of 10, 14, 18 and 18 features (bands) respectively, each with different band combinations (feature configurations). Ensembles 1 and 2 were derived from an exhaustive search strategy, with the ten best base classifiers being selected based on their separability indices. Ensemble 3 was constituted by sequentially arranging the 180 bands into ten base classifiers, each with 18 features. On the other hand, all the features constituting the base classifiers in Ensemble 4 were randomly selected.

Table 1: A summary of the classification accuracy of the various ensembles considered

<table>
<thead>
<tr>
<th>Ens.</th>
<th>BC 1</th>
<th>BC 2</th>
<th>BC 3</th>
<th>BC 4</th>
<th>BC 5</th>
<th>BC 6</th>
<th>BC 7</th>
<th>BC 8</th>
<th>BC 9</th>
<th>BC 10</th>
<th>MV</th>
<th>SB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6209</td>
<td>0.6214</td>
<td>0.4591</td>
<td>0.6176</td>
<td>0.6128</td>
<td>0.6149</td>
<td>0.6125</td>
<td>0.6190</td>
<td>0.6202</td>
<td>0.6338</td>
<td>0.6212</td>
<td>0.6338</td>
</tr>
<tr>
<td>2</td>
<td>0.6134</td>
<td>0.6323</td>
<td>0.4737</td>
<td>0.6531</td>
<td>0.6232</td>
<td>0.6128</td>
<td>0.6418</td>
<td>0.6317</td>
<td>0.6112</td>
<td>0.6328</td>
<td>0.6314</td>
<td>0.6418</td>
</tr>
<tr>
<td>3</td>
<td>0.6112</td>
<td>0.6264</td>
<td>0.3383</td>
<td>0.6084</td>
<td>0.6232</td>
<td>0.6128</td>
<td>0.3937</td>
<td>0.4141</td>
<td>0.6112</td>
<td>0.6232</td>
<td>0.4687</td>
<td>0.6314</td>
</tr>
<tr>
<td>4</td>
<td>0.6232</td>
<td>0.6418</td>
<td>0.3937</td>
<td>0.6605</td>
<td>0.6232</td>
<td>0.6128</td>
<td>0.3937</td>
<td>0.4141</td>
<td>0.6112</td>
<td>0.6232</td>
<td>0.4687</td>
<td>0.6314</td>
</tr>
<tr>
<td>5</td>
<td>0.6128</td>
<td>0.6317</td>
<td>0.4141</td>
<td>0.6276</td>
<td>0.6128</td>
<td>0.6128</td>
<td>0.4141</td>
<td>0.4141</td>
<td>0.6112</td>
<td>0.6232</td>
<td>0.4687</td>
<td>0.6314</td>
</tr>
<tr>
<td>6</td>
<td>0.6149</td>
<td>0.6323</td>
<td>0.4885</td>
<td>0.6314</td>
<td>0.6149</td>
<td>0.6128</td>
<td>0.4885</td>
<td>0.5803</td>
<td>0.6112</td>
<td>0.6232</td>
<td>0.4885</td>
<td>0.5803</td>
</tr>
<tr>
<td>7</td>
<td>0.6125</td>
<td>0.6281</td>
<td>0.5803</td>
<td>0.6176</td>
<td>0.6125</td>
<td>0.6125</td>
<td>0.5803</td>
<td>0.5803</td>
<td>0.6112</td>
<td>0.6232</td>
<td>0.5803</td>
<td>0.5803</td>
</tr>
<tr>
<td>8</td>
<td>0.6190</td>
<td>0.6242</td>
<td>0.5288</td>
<td>0.6425</td>
<td>0.6190</td>
<td>0.6190</td>
<td>0.5288</td>
<td>0.6242</td>
<td>0.6176</td>
<td>0.6232</td>
<td>0.6425</td>
<td>0.6242</td>
</tr>
<tr>
<td>9</td>
<td>0.6202</td>
<td>0.6168</td>
<td>0.4687</td>
<td>0.6314</td>
<td>0.6202</td>
<td>0.6202</td>
<td>0.4687</td>
<td>0.4885</td>
<td>0.6125</td>
<td>0.6232</td>
<td>0.4687</td>
<td>0.4885</td>
</tr>
<tr>
<td>10</td>
<td>0.6338</td>
<td>0.6435</td>
<td>0.3593</td>
<td>0.5989</td>
<td>0.6338</td>
<td>0.6338</td>
<td>0.3593</td>
<td>0.3593</td>
<td>0.6125</td>
<td>0.6232</td>
<td>0.3593</td>
<td>0.3593</td>
</tr>
</tbody>
</table>

Table 2: Binomial Test of Significance between Majority Vote and Single Best

| Ensemble | |z| |
|----------|------|
| 1        | 0.99 |
| 2        | 0.82 |
| 3        | 4.43 |
| 4        | 0.99 |

From Table 1, it can be observed that in all cases single best had better results than majority voting. It is also observed that in general, results from ensemble 3 were the poorest, while ensemble 4 yielded the best results. To get a better appreciation of the differences between these results, a binomial test of significance was carried out for each ensemble to ascertain the pairwise difference between majority voting and single best, results of which are illustrated in Table 2.

From Tables 3 and 4, it can be seen that the results of ensemble 4 are significantly better than the results from ensemble 1 and 3. Whereas the results of ensemble 4 are better than ensemble 2, the difference is not significant. The results from
ensemble 3 are significantly worse than all the results of the ensembles 1, 2 and 4.

Of the ensembles considered, evidently the one based on random selection yielded the best classification results. Sequentially selecting bands into base classifiers yielded significantly poorer results. Feature selection resulted in better classification results compared to sequentially selecting the features, however ensemble 2 performed better than ensemble 1. This may have been as a result of using more features in each base classifier. The difference however was not significant.

6. Conclusions

The results show that to effect ensemble classification through feature selection for hyperspectral data, generation of base classifiers can best be done using the random selection of features. This however comes with a disadvantage of not being able to exactly replicate previous results. The other methods used in this research however provided a more ‘controlled environment’ to explore feature selection. Of the said methods, building the base classifiers through sequentially arranging the features resulted in the poorest results. Feature selection through exhaustive search always yielded comparatively better results. Of Ensembles 1 and 2, Ensemble 2 yielded better results. As mentioned before this may be as a result of the base classifiers in Ensemble 2 having more features than Ensemble 1. The significance of the number of features per base classifier pales when it is observed that Ensemble 3 which had 18 features per base classifier performed poorer than Ensembles 1 and 2 which had 10 and 14 features per base classifier. Of the combination methods, apart from Ensemble 3 which proved to be a poorly constituted ensemble, there was no significant difference between majority voting and single best. However, single best always yielded comparatively better results.

7. Acknowledgements

The authors would like to acknowledge the support of the University of the Witwatersrand, Department of Science and Technology and the anonymous reviewers.

8. References


Learning, Bari, Italy, pp 148 – 156. (Morgan Kaufmann)


