# Feature Clustering and Ranking

# Abstract

High dimensional feature sets typically contain correlation amongst the features. Traditional feature selections approaches are prone to instability and selection of sub-optimal features in these circumstances. A method for ranking features that is robust to correlation is presented. Average-linkage hierarchical clustering is used to group correlated features into clusters. Clusters are then ranked based on the performance of a naïve Bayes classifier on their constituent features. Individual features can then be selected from the best clusters. Other criteria, such has computation time or measurement cost, can be considered when making this selection. The proposed feature selection method is compared to traditional methods on a number of data sets containing correlation amongst the features. Stability of selected features under different data samplings was shown to be improved with the new method. Selected features produced similar or better accuracies across a variety of classifiers compared to traditional methods.

## Index Terms: Feature selection, hierarchical clustering, high dimensionality, correlation, redundancy

# Introduction

As the number of features increases, the amount of data required to adequately represent class distributions in the increased feature space increases exponentially. This is known as the “curse of dimensionality” (Bishop, 2003). For finite training samples, increasing the features beyond a certain point results in overtraining and a decrease in the classifier accuracy. This is called the “peaking phenomenon” (Jain et al., 2000). The peaking phenomenon makes it necessary to reduce the size of the feature set to a salient minimum in order to achieve an accurate classifier. While support vector machine (SVM) (Burges, 1998) and random forest (Breiman, 2001) classifiers have become popular, partly because of their lack of sensitivity to the peaking phenomenon (Guyon et al., 2002), reducing the number of features is also beneficial from the perspective of measurement costs and feature computation time. The last point is particularly relevant in large scale remote sensing studies of Very High Resolution (VHR) imagery, due to the vast quantities of data requiring processing. A further motivation for reducing features to an informative minimum is the “ugly duckling theorem”, which implies that the more redundant features contained in a data set, the less separable classes become (Jain et al., 2000).

Filter, wrapper and embedded approaches and reason for filter: fast, generic, and proper feature selection i.e. can reduce computation. Instability in traditional and other approaches.

There are two basic approaches to feature set reduction: feature selection and feature extraction. In feature extraction the feature set is mapped into a new feature space of reduced dimensionality (Webb, 2002). Various criteria are used to define the dimensions of the new space such as separability and variance. These may be supervised or unsupervised measures. A disadvantage of the feature extraction approach is that it requires calculation of the full feature set.

Feature selection involves the selection of a subset of features from the original set according to some criterion of subset performance. The number of possible subsets increases combinatorially with the size of the feature set and it is seldom practical to evaluate all possible subsets (Jain et al., 2000). A variety of search methods exist for reducing the portion of feature space searched. Of these, only the branch and bound method is globally optimal, the rest achieve reduced computation at the price of optimality. The complexity of the branch and bound method increases exponentially with the size of the feature set and so it is still computationally impractical for large feature sets (Jain et al., 2000). The fastest and most straightforward search method is simply to rank features based on their individual performance and select the best *N*. Feature ranking approaches are problematic for data sets containing correlation amongst the features. In these situations, correlated features are ranked similarly resulting in sub-optimal and redundant feature sets.

More advanced search methods use greedy sequential approaches, such as forward selection and backward elimination. Greedy search methods are more likely to find the globally optimal feature set than the feature ranking approach as they are exploring more of the search space and are less inclined to select multiple correlated features than the ranking approach.

Feature selection methods can be divided in filter, wrapper and embedded approaches. In the filter approach, generic measures of separability or importance, are used to evaluate feature subsets, while in the wrapper approach, the accuracy of a specific classifier trained on the feature subset is used as the selection criterion (Duin and Tax, 2005). A feature set that is optimal for one classifier is not necessarily optimal for another (Tolosi and Lengauer, 2011), so if using a wrapper approach, one should ideally use the same classifier to label the selected features as the criterion used for feature selection. An embedded approach is one where feature selection is incorporated into the classifier training procedure such as with random forests (Breiman, 2001). Filter approaches have the advantage over wrapper and embedded approaches of making feature selection independent of the classifier, thus allowing for greater flexibility in the choice of classifier (Brown et al., 2012). Filter approaches are also computationally more efficient than wrapper approaches in general . This is an important consideration for large and high dimensional data sets such as those often encountered in remote sensing. For these reasons, our method adopts the filter approach.

High dimensional feature spaces typically contain correlation (i.e. redundancy) amongst the features (Cukur et al., 2015; Tolosi and Lengauer, 2011; Yu and Liu, 2004). The raw bands of aerial multi-spectral imagery often have significant spectral overlap and consequently are correlated with one another. For example, see the overlapping relative spectral responses (RSR’s) of the Intergraph DMC sensor shown in Figure 1. This spectral overlap will exacerbate the correlation amongst features derived from these raw bands (Cukur et al., 2015). In Bioinformatics, gene expression data is also recognised as typically containing a lot of correlation (Sahu and Mishra, 2011; Strobl et al., 2008; Yousef et al., 2007). A number of authors have noted difficulties in selecting features from high dimensional correlated sets. Kononenko et al. (1997), Guyon et al. (2002), Yu & Liu (2004) and Yousef et al. (2007) note that feature redundancy can have a negative impact on the optimality of feature selection. Feature redundancy not only leads to sub-optimal feature selection but also makes selected features unstable and sensitive to small changes in the data used for selection (Guyon and Elisseeff, 2003; Kalousis et al., 2007; Li et al., 2011; Tolosi and Lengauer, 2011).

Correlation can be effectively dealt with using a feature extraction approach, such as principal components analysis (PCA) (Webb, 2002), but this requires computation of the full feature set. This is not practical in computationally demanding applications such as large area remote sensing of aerial imagery. One needs to consider both relevance and redundancy

A number of feature selection approaches have been developed to address the issues of stability and sub-optimality encountered in high-dimensional and correlated data. These methods need to simultaneously consider both feature relevance and redundancy By grouping correlated or redundant features into clusters, one can measure the relevance of each cluster and select single features from relevant clusters . A means of selecting good features in the presence of correlation was devised by Yousef et al. (2007). They use a k-means algorithm to produce a fixed number of clusters of correlated features. A backward elimination type greedy search is then conducted, eliminating whole clusters rather than individual features. The accuracy of a SVM is used as selection criterion. The clustering is then reapplied at each step. A related feature selection method that finds and removes redundancy by clustering features into similar groups is presented by Mitra et al. (2002). They use a novel clustering algorithm to group correlated features based on a similarity measure they call “maximal information compression index”, which is the smallest eigenvalue of the feature covariance. Sahu & Mishra (2011) and Cukur et al. (2015) propose clustering methods based on feature correlation and include an additional step where feature clusters are ranked based on a performance criterion and representative features selected from the best clusters.

Yu & Liu (2004) use a non-linear correlation measure, called symmetrical uncertainty, to measure both feature relevance, by how well correlated features are with class labels, and feature redundancy, by how well correlated features are with each other. An efficient search method, based on the approach of Markov blanket filtering, finds and removes redundant features from a set of relevant ones. This and other methods are incorporated into a common theoretical framework in Brown et al. (2012). A comprehensive empirical study is used to compare the performance of a range of filter approach feature selection algorithms under varying conditions, including correlation amongst the features. They conclude that joint mutual information (JMI) method (Yang and Moody, 1999) provides the best performance overall.

Our method is similar in concept to those of Sahu & Mishra (2011) and Cukur et al. (2015). It clusters correlated features and ranks the clusters using a simple classification accuracy measure. Unlike k-means and backward elimination type approaches, no prior knowledge of the number of clusters is needed, and the clustering is performed only once, making the method computationally efficient. Wrapper and justification. Importance of hand selection for computation.

# Methods

## Formulation

Our method is described as follows:

1. Perform average-linkage hierarchical clustering (Szekely and Rizzo, 2005) of the feature set using the correlation coefficient as the dissimilarity metric.
2. Select a natural number of clusters containing high correlation by visual inspection of the dendrogram.
3. Rank each cluster’s importance by finding the accuracy of a naïve Bayes classifier trained on each individual feature and then finding the median of the feature accuracies in the cluster.
4. Select a single feature from each of the *N* clusters with best importance scores.

Hierarchical clustering provides a simple way of clustering the features that does not require prior knowledge of the number of clusters (Webb, 2002). The method starts with each individual feature in its own cluster and proceeds in a number of steps where the pair of clusters that minimise a criterion are merged at each step. The average linkage criterion was used in our method, and is the average dissimilarity over all pairs of objects in the two clusters. The dendrogram is graphical representation of the clustering process.

Cluster stability and strength of correlation within each cluster are the key factors to consider when choosing the number of clusters and can be visually interpreted from the dendrogram. The naïve Bayes classifier, using a ten bin histogram to model class densities, was chosen to evaluate importance primarily because it makes no assumptions about the form of the class distributions and can thus provide a generic measure of separability. It is simple, fast and recognised as being accurate for a variety of problems (Hand and Yu, 2001). The “naïve” assumption of feature independence is of no consequence when testing individual features. These justifications aside, our reasoning for the choice of dissimilarity metric, clustering algorithm and selection criterion is loose and other options could be also be effective.

The number of clusters, *N*, can be specified by the user based on the size of the training set or chosen using a grid search with the final classifier accuracy as performance measure. To avoid biased accuracy estimates, all classifier accuracy evaluation, for cluster ranking or selection of *N*, is done on unseen test data using a ten-fold cross validation (Bishop, 2003). Major benefits of the cluster ranking method are its speed and that it allows hand-picking of the single features that represent each cluster. The flexibility to choose features allows the user to favour those features that are fastest to compute or perhaps to choose those features that are more readily understood. If these factors are not of importance, the highest individually ranked feature in each cluster can be chosen.

## Data and Ground Truth

A three class, multi-spectral pixel data set was used in this study to evaluate the effectiveness of the proposed method. The pixel data originated from radiometrically corrected VHR aerial images, obtained from Chief Directorate: National Geospatial Information (NGI), a component of the South African Department of Rural Development and Land Reform. The images were captured with an Intergraph DMC camera which provides multi-spectral red, green, blue and near-infrared (NIR) bands at a 0.5m/pixel resolution. The overlapping spectral responses of this sensor are shown in Figure 1. The data set was built as part of a broader vegetation mapping study being conducted in the Little Karoo, an arid region in South Africa.

The three classes of interest are described in Table 1. The data ground truth labels were generated by visual inspection and manual delineation of objects belonging to the various classes in the images. Table 2 shows the total number polygons and pixels ground truthed for each class.

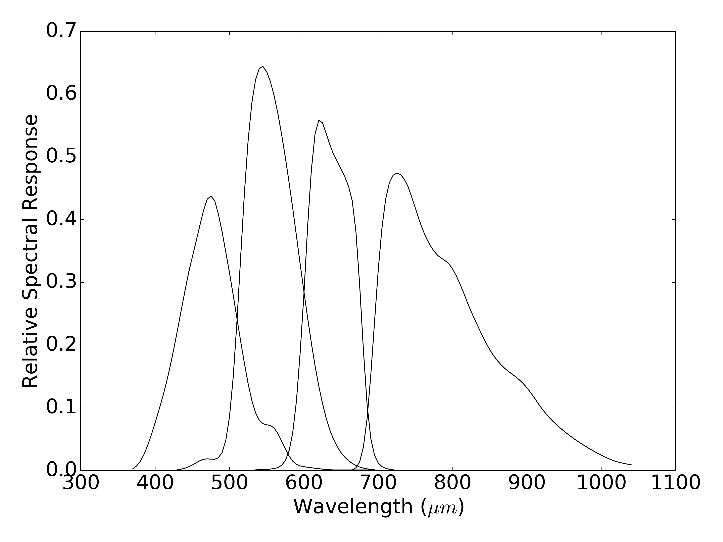


Figure 1 Intergraph DMC relative spectral responses

Table 1 Class descriptions

|  |  |
| --- | --- |
| Class Name | Description |
| Spekboom | *Portulacaria afra* |
| Tree | Any recognisable tree but especially the darker *Euclea* and *Pappea* trees commonly found intermingled in stands of *P. afra*. |
| Background | Bare ground, small shrubs, herbs and anything else not included in the first two classes. |

Table 2 Ground truth size

|  |  |  |
| --- | --- | --- |
| Class Name | Polygons | Pixels |
| Spekboom | 52 | 27260 |
| Tree | 64 | 3357 |
| Background | 44 | 182044 |
| **Total** | **160** | **212661** |

## Features

The full feature set consisted of 46 features and was comprised of a typical combination of spectral features, vegetation indices and texture features. Similar features were used in Li et al. (2010) and Trias-Sanz, Stamon & Louchet (2008). The features can be separated into two broad categories: per-pixel and sliding window features. The per-pixel features are found with the spectral information from only that pixel, while the sliding window features are found from some statistic or function of the pixels inside a small local neighbourhood. The complete feature set and their labels are listed below in Table 3.

Table 3 Features

|  |  |  |
| --- | --- | --- |
| No. | Name | Description |
| 1 | R | Red |
| 2 | G | Green |
| 3 | B | Blue |
| 4 | NIR | Near-infrared |
| 5 | rN | Normalised R |
| 6 | gN | Normalised G |
| 7 | bN | Normalised B |
| 8 | nirN | Normalised NIR |
| 9 | NDVI | Normalised Difference Vegetation Index |
| 10 | RVI | Ratio Vegetation Index |
| 11-14 | tc1-4 | Tasselled cap components |
| 15-18 | pc1-4 | Principal components of raw bands |
| 19-22 | nc1-4 | Principal components of normalised bands |
| 23-26 | Entropy## | Sliding window entropy of pc1, RVI, NDVI and gN |
| 27-30 | Std## | Sliding window standard deviation of pc1, RVI, NDVI and gN |
| 31-34 | Mean## | Sliding window mean of pc1, RVI, NDVI and gN |
| 35-38 | Median## | Sliding window median of pc1, RVI, NDVI and gN |
| 39-42 | Skewness## | Sliding window skewness of pc1, RVI, NDVI and gN |
| 43-46 | Kurtosis## | Sliding window kurtosis of pc1, RVI, NDVI and gN |

Normalised colour features are defined by (Blauensteiner et al., 2006):

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Where are the raw R, G, B and NIR band values and is the band number. The denominator normalises for intensity and thus this space describes the relative contributions of the raw bands to overall intensity. The Ratio Vegetation Index (RVI) is given by

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

It has a range of zero to infinity and increases as the vegetation becomes denser and photosynthetically more active (Myneni et al., 1995). The well-known Normalised Difference Vegetation Index (NDVI) is defined as

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

NDVI is limited to the range -1 to 1 and describes the same relationship as RVI but is easier to visualise and interpret due to its limited range. Both indices are invariant to intensity changes.

The tasselled cap transform of Kauth & Thomas (1976) was approximated for our problem using a principal component transform derived from the variance of the Spekboom class. This way the first component was aligned with Spekboom variation rather than wheat variation as in the original tasselled cap transform.

Entropy is a statistic that describes the amount of randomness in a variable (Webb, 2002). The entropy of the values in the image window is defined by

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | (4) |

Where is the probability in the histogram bin of . 256 bins were used in all cases.

In addition to the entropy, the median and the four central moment features (mean, standard deviation, skewness and kurtosis) of (Li et al., 2010) were included as sliding window features. The first principal component, RVI, NDVI and normalised green channel were all used as inputs to the sliding window feature set. A sliding window size of five pixels was used.

## Evaluation

A data set was produced by extracting the 46 features of Table 3 from the vegetation mapping pixel data described in Section 2.2. The sensitivity of standard feature selection approaches to the following factors was investigated using this data set:

1. Sampling of the data set
2. Selection criterion
3. Subset search method

These results were subsequently compared with the stability of the features selected by the feature clustering and ranking method under the disturbance of factors 1 and 2 above. The accuracy of a diverse selection of classifiers trained on the features selected by the proposed method was also computed as a check on general their validity.

# Results and Discussion

## Standard Feature Selection

We evaluated the sensitivity of standard feature selection approaches to different disturbances using the vegetation mapping data set. The size of the Background class was reduced to be the same as the Spekboom class by taking a random subsample. This was done to expedite computation times.

In the first experiment, a forward selection procedure with naïve Bayes classification accuracy as selection criterion was repeated on bootstrapped samples of the original data. The first six selected features are shown in Table 4 for five different data samplings. The first selected feature remains stable but the rest of the features are sensitive to the specific data set used to evaluate importance. This variation occurs in spite the fact that data contains more than sufficient samples to be representative of the real class distributions.

Table 4 Experiment 1: Forward selection results for different data sets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample Num. | Selected Features (Ordered by ranking) | | | | | |
| 1 | pc2 | tc4 | pc1 | SkewnessNDVI | StdRVI | rc3 |
| 2 | pc2 | rc3 | EntropyPc1 | KurtosisPc1 | SkewnessGn | KurtosisGn |
| 3 | pc2 | rc3 | MeanPc1 | StdGn | MedianRVI | KurtosisNDVI |
| 4 | pc2 | tc4 | EntropyPc1 | StdRVI | pc3 | MeanRVI |
| 5 | pc2 | rc3 | EntropyPc1 | KurtosisGn | SkewnessGn | SkewnessPc1 |

Next, the sensitivity to the selection criterion was evaluated by running a forward selection procedure with following criteria: inter–intra class distance, sum of Mahalanobis distances and the accuracies of naïve Bayes, decision tree and normal Bayes classifiers. Results in Table 5 show significant variation in the first six selected features, although certain features such as pc2, MeanPc1 and MedianPc1 are favoured by differing criteria.

Table 5 Experiment 2: Forward selection results for different selection criteria

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Criterion | Selected Features (Ordered by ranking) | | | | | |
| Inter-intra | rN | NIR | rc3 | R | StdRVI | StdNDVI |
| Mahalanobis | MedianNDVI | rc3 | MeanPc1 | pc2 | pc4 | StdRVI |
| Naïve Bayes | pc2 | rc3 | MedianPc1 | SkewnessPc1 | KurtosisGn | KurtosisPc1 |
| Decision tree | pc2 | MedianGn | MedianPc1 | pc4 | MeanRVI | EntropyGn |
| Normal Bayes | pc2 | gN | MeanPc1 | rc3 | StdRVI | MeanNDVI |

Finally, common subset search methods were tested for their effect on selected features. All except the random forest used a naïve Bayes classification accuracy as selection criterion. It was not practical to evaluate the branch and bound method due to the computation time required. Variation in selected features is again significant and of such a degree that it is difficult to discern any consistency. Results are presented in Table 6.

Table 6 Experiment 3: Feature selection results for different search methods

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Search Method | Selected Features (Ordered by ranking) | | | | | |
| FS | pc2 | rc3 | EntropyPc1 | SkewnessPc1 | KurtosisPc1 | SkewnessGn |
| BE | NIR | tc4 | rc1 | rc3 | EntropyPc1 | MedianNDVI |
| Individual ranking | pc2 | MeanNDVI | MedianNDVI | rN | NDVI | rc1 |
| Random forest | rN | StdRVI | tc4 | MedianPc1 | bN | MedianGn |

## Clustering and Ranking

The seemingly inexplicable variation in results of standard feature selection methods indicates that results are suboptimal. The likely cause of the instability is redundancy in our feature set, which is known to cause this kind of sensitivity (Guyon and Elisseeff, 2003; Li et al., 2011; Tolosi and Lengauer, 2011). The effectiveness of the feature clustering and ranking method was tested on the vegetation mapping data. The dendrogram showing the clustering of the feature set is plotted in Figure 1. The line in red shows the correlation threshold value at which the feature clusters were extracted. This point was selected as a relatively stable point in the hierarchy and one where the correlation amongst features is strong.



Figure 2 Dendrogram showing clustering of correlated features

The clusters were ranked using the naïve Bayes classification accuracy criterion. Table 7 lists the clusters ordered by their importance, with their component features. This table reveals a number of interesting properties of the features. Firstly, it is clear that there is a significant amount of correlation amongst the features in general. The correlation between raw R, G, B and NIR bands is surprisingly strong and these are all grouped into a single cluster. While the definitions of the nirN, NDVI and RVI features are quite different, they are all describing the same spectral property of vegetation, namely high absorption in the red band and high reflectance in the NIR band. This is confirmed by their collection in the same cluster. The mean sliding window feature, median sliding window feature and source feature operated on by those sliding windows are strongly correlated with each other as one would expect.

Table 7 Ranked clusters

|  |  |  |
| --- | --- | --- |
| Num. | Importance (%) | Features |
| 1 | 68.27 | rN, nirN, NDVI, RVI, tc2, pc2, rc1, MeanRVI, MedianRVI, MeanNDVI, MedianNDVI |
| 2 | 61.38 | R, G, B, NIR, tc1, pc1, MeanPc1, MedianPc1 |
| 3 | 60.41 | EntropyPc1 |
| 4 | 55.23 | gN, MeanGn, MedianGn |
| 5 | 54.52 | bN |
| 6 | 53.57 | rc2, rc4 |
| 7 | 50.57 | tc4, rc3 |
| 8 | 49.34 | pc4 |
| 9 | 47.93 | EntropyRVI, StdRVI, EntropyNDVI, StdNDVI |
| 10 | 43.96 | StdPc1 |
| 11 | 43.62 | EntropyGn, StdGn |
| 12 | 42.65 | tc3, pc3 |
| 13 | 41.29 | SkewnessRVI, SkewnessNDVI |
| 14 | 35.27 | SkewnessGn |
| 15 | 35.19 | KurtosisRVI, KurtosisNDVI |
| 16 | 35.03 | SkewnessPc1 |
| 17 | 34.86 | KurtosisGn |
| 18 | 33.86 | KurtosisPc1 |

The first two experiments of Section 3.1 were repeated with the feature clustering and ranking method, to compare its stability to that of standard methods. The effect of different bootstrap samples of the data set on the ranking of the top six clusters in shown in Table 8. The clusters are referred to by their numbers as shown in the first column of Table 7. Clusters 1 to 5 are consistently selected for all data samplings and cluster 6 is selected for all but one of the tests. This is a significant improvement in stability compared to the corresponding results from forward selection in Table 4.

Table 8 Experiment 4: Feature clustering and ranking results for different data sets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample Num. | Selected Clusters (Ordered by ranking) | | | | | |
| 1 | 1 | 2 | 3 | 4 | 5 | 7 |
| 2 | 1 | 2 | 3 | 5 | 4 | 6 |
| 3 | 1 | 2 | 3 | 4 | 5 | 6 |
| 4 | 1 | 2 | 3 | 4 | 5 | 6 |
| 5 | 1 | 2 | 3 | 4 | 5 | 6 |

Table 9 shows the resulting clusters for different selection criteria. Some variation in cluster ranking is seen for different criteria and this is not unexpected as the criteria obviously view cluster importance differently. Results are nevertheless substantially more consistent than the equivalent experiment in Section 3.1 where no single feature was selected for all criteria. Here clusters 1, 2, 3, 5 and 6 are consistently part of the top six for the differing criteria.

Table 9 Experiment 5: Feature clustering and ranking results for different selection criteria

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Criterion | Selected Clusters (Ordered by ranking) | | | | | | |
| Inter-intra | | 1 | 5 | 6 | 3 | 2 | 9 |
| Mahalanobis | | 1 | 5 | 6 | 2 | 3 | 7 |
| Naïve Bayes | | 1 | 2 | 3 | 4 | 5 | 6 |
| Decision tree | | 1 | 2 | 3 | 4 | 5 | 7 |
| Normal Bayes | | 1 | 2 | 6 | 3 | 5 | 4 |

We selected the features NDVI, pc1, EntropyPc1, gN, bN and rc2 from the top six clusters of Table 7. Selection of sliding window features was avoided where possible as they are computationally more demanding than the per-pixel features. There is only one sliding window feature, EntropyPc1 in our final selection. This makes for a very significant improvement in computation time over the full feature set. Decision tree, random forest, Support Vector Machine (SVM) and K-Nearest-Neighbour (KNN) classifiers were trained and tested on the selected features using a ten-fold cross validation. Results are given in Table 10. Similarly good performances are obtained across all classifiers, confirming the value of the selected features.

Table 10 Classifier performance comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | 3 Class Error (%) | CA (Bg / Sb / Tr)a | PA (Bg / Sb / Tr)a | Kappa |
| Decision Tree | 9.46 | 97.64 / 98.10 / 46.86 | 91.22 / 94.49 / 85.91 | 0.87 |
| Random Forest | 9.16 | 98.72 / 97.51 / 47.97 | 89.66 / 96.96 / 85.91 | 0.88 |
| KNN | 10.45 | 97.27 / 97.49 / 80.27 | 97.08 / 98.82 / 72.74 | 0.94 |
| SVM | 10.58 | 97.17 / 95.33 / 79.73 | 94.44 / 98.69 / 75.12 | 0.92 |

a CA = Consumer’s Accuracy (%), PA = Producer’s Accuracy (%), Bg = Background, Sb = Spekboom, Tr = Tree

# Conclusions

Correlation amongst features was identified as a source of instability and inaccuracy in feature selection. A method for ranking correlated clusters of features was presented. Using hierarchical clustering, a natural number of clusters can be selected by observing the stability of correlation relationships in the data using a dendrogram. Clusters are then ranked using an importance measure, calculated as the median of the accuracy of a naïve Bayes classifier trained on each individual feature in the cluster. By selecting individual features from the best clusters, a set of informative features is found while simultaneously removing correlation from the data. The ability to hand-pick features is beneficial as it allows other factors to be included in determining the optimal feature set such as speed of computation and physical interpretability.

Experiments were conducted on a three class vegetation mapping data set consisting of 46 features derived from VHR multi-spectral image pixel data. There was a large amount of correlation amongst the features. It was shown that features selected using a standard forward selection approach were sensitive to different data samplings and selection criteria. Various standard search methods were also found to select vastly different groups of features. Using feature clustering and ranking, selected features were substantially more consistent under different data samplings and different selection criteria. The selected features also produced good accuracies on a variety of classifiers.

A possible weakness in the presented method is the formulation of the importance measure. The clusters are evaluated in isolation which makes selection susceptible to sub-optimality as the best individual clusters are not necessarily the best collectively (Cover, 1974). Ideally the importance of clusters should be evaluated in the context of other clusters so that the effect of informative relationships is incorporated into the ranking. One option would be to follow a backward elimination type approach similar to Yousef et al. (2007) but this is computationally expensive and requires sufficient training data to train on the complete set of features. Alternatively, using the random forest concept of feature importance, we suggest the following as a simple and efficient approach that does not require training on the complete set of features:

1. Find the first principal component of each cluster’s features. Use this as a representative feature for that cluster.
2. Train a naïve Bayes classifier on the full set of representative features and find its accuracy.
3. For each cluster, permute that cluster’s representative feature amongst the samples in the test set and find the accuracy of the naïve Bayes classifier on the disturbed set.
4. The decrease in accuracy between steps 2 and 3 becomes the importance measure for that cluster.

Another aspect of our method that could benefit from further investigation is the correlation coefficient dissimilarity metric. While non-linear relationships between features were captured by this linear measure in our experiments (for example the relationship between RVI and NDVI), this will not always be the case. It would be preferable to use a metric that can describe non-linear correlation, such as the symmetrical uncertainty used in Yu & Liu (2004).

# References

Bishop, C.M., 2003. Neural networks for pattern recognition. Oxford University Press, New York. doi:10.1002/0470854774

Blauensteiner, P., Wildenauer, H., Hanbury, A., Kampel, M., 2006. On colour spaces for change detection and shadow suppression, in: Chum, O., Franc, V. (Eds.), Computer Vision Winter Workshop 2006. Czech Pattern Recognition Society, Telc, Czech Republic, pp. 1–6.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324

Brown, G., Pocock, A., Zhao, M.-J., Lujan, M., 2012. Conditional Likelihood Maximisation: A Unifying Framework for Mutual Information Feature Selection. Journal of Machine Learning Research 13, 27–66. doi:10.1016/j.patcog.2015.11.007

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167. doi:10.1023/A:1009715923555

Cover, T.M., 1974. The best two independent measurements are not the two best. IEEE Transactions on Systems, Man, and Cybernetics SMC-4, 116–117. doi:10.1109/TSMC.1974.5408535

Cukur, H., Binol, H., Uslu, F.S., Kalayci, Y., Bal, A., 2015. Cross correlation based clustering for feature selection in hyperspectral imagery, in: 2015 9th International Conference on Electrical and Electronics Engineering (ELECO). IEEE, Bursa, pp. 232–236. doi:10.1109/ELECO.2015.7394552

Duin, R.P.W., Tax, D.M.J., 2005. Statistical Pattern Recognition, in: Chen, C., Wang, P. (Eds.), Handbook of Pattern Recognition and Computer Vision, 3rd Ed. World Scientific, Singapore, pp. 1–21. doi:10.1142/9789812775320\_0001

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal ofMachine Learning Research 3, 1157–1182. doi:10.1016/j.aca.2011.07.027

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422. doi:10.1023/A:1012487302797

Hand, D.J., Yu, K., 2001. Idiot’s Bayes - not so stupid after all? International Statisitical Review 69, 385–398.

Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 4–37.

Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: A study on high-dimensional spaces. Knowledge and Information Systems 12, 95–116. doi:10.1007/s10115-006-0040-8

Kauth, R.J., Thomas, G.S., 1976. The tasselled cap -- a graphic description of the spectral-temporal development of agricultural crops as seen by LANDSAT, in: Symposium on Machine Processing of Remotely Sensed Data. IEEE, Purdue University of West Lafayette, Indiana, USA, pp. 4B41–4B51.

Kononenko, I., Šimec, E., Robnik-Šikonja, M., 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. Applied Intelligence 7, 39–55. doi:10.1023/A:1008280620621

Li, S., Harner, E.J., Adjeroh, D. a, 2011. Random KNN feature selection - a fast and stable alternative to Random Forests. BMC bioinformatics 12, 450. doi:10.1186/1471-2105-12-450

Li, Z., Hayward, R.F., Zhang, J., Jin, H., Walker, R.A., 2010. Evaluation of spectral and texture features for object-based vegetation species classification using support vector machines, in: Wagner, W., Székely, B. (Eds.), ISPRS TC VII Symposium – 100 Years ISPRS. IAPRS, Vienna, Austria, pp. 122–127.

Mitra, P., Murthy, C. a, Pal, S.K., 2002. Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI 24, 301–312. doi:10.1109/34.990133

Myneni, R.B., Hall, F.G., Sellers, P.J., Marshak, A.L., 1995. The interpretation of spectral vegetation indexes. IEEE Transactions on Geoscience and Remote Sensing 33, 481–486. doi:10.1109/36.377948

Sahu, B., Mishra, D., 2011. A novel approach for selecting informative genes from gene expression data using signal-to-noise ratio and t-statistics, in: 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011). IEEE, Allahabad, India, pp. 5–10. doi:10.1109/ICCCT.2011.6075207

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC bioinformatics 9, 307. doi:10.1186/1471-2105-9-307

Szekely, G.J., Rizzo, M.L., 2005. Hierarchical clustering via joint between-within distances: extending Ward’s minimum variance method. Journal of Classification 22, 151–183. doi:10.1007/s00357-005-0012-9

Tolosi, L., Lengauer, T., 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27, 1986–1994. doi:10.1093/bioinformatics/btr300

Trias-Sanz, R., Stamon, G., Louchet, J., 2008. Using colour, texture, and hierarchial segmentation for high-resolution remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 63, 156–168. doi:10.1016/j.isprsjprs.2007.08.005

Webb, A.R., 2002. Statistical Pattern Recognition. John Wiley & Sons, Ltd, Chichester, UK. doi:10.1002/0470854774

Yang, H.H., Moody, J., 1999. Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. Advances in Neural Information Processing Systems 12, 687–693.

Yousef, M., Jung, S., Showe, L.C., Showe, M.K., 2007. Recursive cluster elimination (RCE) for classification and feature selection from gene expression data. BMC Bioinformatics 8. doi:10.1186/1471-2105-8-144

Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224.