# Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data

# Abstract

High dimensional remote sensing data sets typically contain redundancy amongst the features. Traditional approaches to feature selection are prone to instability and selection of sub-optimal features in these circumstances. They can also be computationally expensive, especially when dealing with very large remote sensing datasets. This article presents an efficient, deterministic feature ranking method that is robust to redundancy. Average-linkage hierarchical clustering is used to group correlated features into clusters. A relevance criterion is evaluated on each feature. Clusters are then ranked based on the median of the relevance values of 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 filter approach methods on a number of remote sensing data sets containing feature redundancy. Mutual information and naïve Bayes relevance criteria were evaluated in conjunction with the feature selection methods. It was shown that, using the new method, the stability of selected features improved under different data samplings, while similar or better classification accuracies were achieved 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 involving 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).

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. Principal Components Analysis (PCA) (Webb, 2002) is an example of a popular feature extraction method.

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 exists 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 is as such 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*. However, feature ranking approaches are problematic for data sets containing feature redundancy. 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. The forward selection (FS) approach starts with an empty feature set and proceeds in a number of steps where at each step the feature from the remaining unselected features, that improves an accuracy criterion the most, is added to the model (Bishop, 2003). The backward elimination (BE) method starts with the full set of features and removes one at each step. The feature whose removal produces the best accuracy according to some criterion is eliminated from the set at each step (Bishop, 2003). BE is computationally costly and requires sufficient reference data to be able to train and assess the classifier on the full feature set. 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 redundant features than the ranking approach (Webb, 2002).

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). 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). In general, filter approaches are also computationally more efficient than wrapper approaches. This is an important consideration for large and high dimensional data sets such as those often encountered in remote sensing.

High dimensional feature spaces typically contain feature redundancy (Cukur et al., 2015; Tolosi and Lengauer, 2011; Yu and Liu, 2004). Although feature correlation and redundancy are related, they are not strictly the same thing (Brown et al., 2012; Guyon and Elisseeff, 2003). Features can help improve separability when the within class correlation is stronger than the between class correlation. We use the term “redundancy” to refer to correlation of features between classes. The raw bands of aerial multi-spectral imagery often have significant spectral overlap and consequently are correlated with one another. This spectral overlap will exacerbate the redundancy amongst features derived from these raw bands (Cukur et al., 2015).

A number of authors have noted difficulties in selecting features from high dimensional sets. Kononenko et al. (1997), Guyon et al. (2002), Yu & Liu (2004) and Yousef et al. (2007) noted 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).

Redundancy can be effectively dealt with using a feature extraction approach, such as PCA, but this requires computation of the full feature set. This is not practical in computationally demanding applications such as analysing very high resolution (VHR) imagery over large areas. A number of feature selection approaches have been developed to address the issues of stability and sub-optimality encountered in high dimensional and redundant data. These methods consider the trade-off between feature relevance (i.e. how much information the feature contains about the class labels) and redundancy.

A means of selecting good features from redundant spaces was devised by Yousef et al. (2007). They used a k-means algorithm to produce a fixed number of clusters of correlated features. The accuracy of a SVM classifier is found for all the features of each cluster. The lowest performing clusters are eliminated, the remaining features combined, and the process is repeated until a desired number of clusters is reached. A related feature selection method that finds and removes redundancy by clustering features into similar groups was presented by Mitra et al. (2002). They used 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) also used k-means clustering to group redundant features. The best feature, according to an importance measure, is then selected from each cluster. Cukur et al. (2015) proposed a similar method, where redundant features are clustered and top ranked features selected from each cluster using an importance measure called “minimum redundancy maximum relevance” (mRMR).

A two-step procedure called the “Fast Correlation Based Filter” (FCBF), was developed by Yu & Liu (2004). It first creates a reduced set of relevant features, and then removes redundant features from this set using a search procedure based on Markov blanket filtering. A non-linear correlation measure, called symmetrical uncertainty, is used to measure both feature relevance and redundancy. Relevance is measured by how well features are correlated with class labels and redundancy is measured by how well features are correlated with each other.

A number of feature importance measures (including the FCBF) were incorporated by Brown et al. (2012) into a common theoretical framework. These measures all consider both relevance and redundancy in some way. A comprehensive empirical study was used to compare the performance (in terms of stability and classifier accuracy) of these measures. The study tested the criteria in a FS approach, under varying conditions, including redundancy in high dimensional feature spaces. They concluded that joint mutual information (JMI) (Yang and Moody, 1999) provides the best feature selection performance overall.

With the exception of FCBF, the above feature selection procedures can be grouped into two categories:

1. Approaches that use some form of clustering of similar features to identify and remove redundancy, usually followed by a ranking of cluster relevance to choose features from the most informative clusters.
2. Approaches that use a single measure of feature importance that incorporates the trade-off between feature relevance and redundancy, after which a FS or simple ranking procedure is used to select the best features.

In gene expression studies, such as DNA microarray studies, all features have similar measurement costs and computation times (Inza et al., 2004), but this is not the case for remote sensing problems, where some features might carry significantly larger computation time or measurement cost burdens than others (Blaschke, 2010). Where a group of correlated features are similarly relevant, it is beneficial to select the feature with the lowest computation time or measurement cost. Consideration of these costs can potentially result in a substantial reduction in classification computation time for this form of remote sensing problem.

Feature selection in redundant feature spaces has received attention in the field of bioinformatics, but is still largely unexplored in remote sensing applications. Of the methods reviewed, only those of Cukur et al. (2015) and Brown et al. (2012) were applied to remotely sensed data. Cukur et al. (2015) worked with two hyperspectral data sets and Brown et al. (2012) included one remote sensing data set in the fifteen they used for comparing criteria. In this paper we propose a filter approach feature selection method for addressing the problem of high dimensional, redundant feature spaces of remotely sensed data. We adopt the filter approach due its relative speed and separation of feature selection and classification tasks. The method follows the clustering and ranking approach of category 1 above. While the methods of Sahu & Mishra (2011) and Yousef et al. (2007) use k-means clustering, the proposed method uses hierarchical clustering. This has the advantage of producing deterministic results, not requiring prior knowledge of the number of clusters and allowing user selection of the final partitioning. We compare the performance of the proposed method to popular feature selection approaches, on a number of remote sensing data sets. The proposed method also allows consideration of computation time and measurement cost in selecting features from correlated clusters of similarly relevant features. While many feature selection evaluations only consider classification accuracy (Cukur et al., 2015; Guyon et al., 2002; Mitra et al., 2002; Sahu and Mishra, 2011; Strobl et al., 2008; Yousef et al., 2007; Yu and Liu, 2004), we present measures of computation time, classification accuracy and stability of selected features under different data samplings.

# Methods

## Formulation

The proposed 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 value of a relevance criterion for each individual feature and then finding the median of the feature relevance values 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 grouping 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 dissimilarity criterion are merged at each step. The average-linkage criterion, which is the average dissimilarity over all pairs of objects in the two clusters, was used in the proposed method, with the correlation coefficient as the dissimilarity measure. The dendrogram is a graphical representation of the clustering process (MathWorks, 2016). 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. An example dendrogram is shown in Figure 1. The inverted U-shaped lines show which clusters are combined into new clusters. The height of the horizontal line indicates the magnitude of the dissimilarity (i.e. average linkage correlation distance) between clusters. The dotted line shows an example dissimilarity threshold at which to extract clusters from the hierarchy. At this threshold, the clusters are highly correlated (i.e. the dissimilarity is small) and the cluster contents are stable (i.e. the next level in the hierarchy only occurs at a substantially larger dissimilarity).

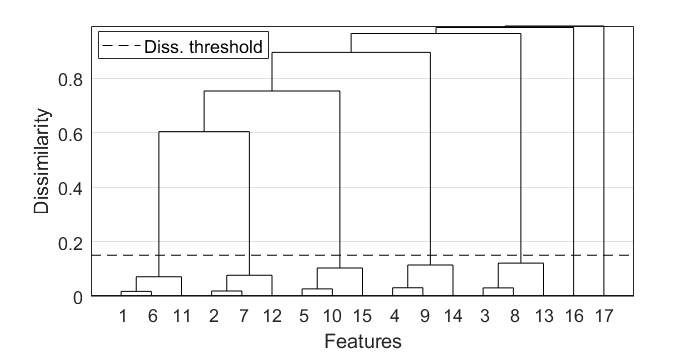


Figure 1 Example dendrogram showing chosen threshold at which to extract clusters

We investigated the performance of two different feature relevance measures: the accuracy of a naïve Bayes classifier and the mutual information (MI) between the feature and the class labels. The naïve Bayes classifier, using a histogram to model class densities, was chosen 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.

MI is a measure of the dependence between two random variables (Brown et al., 2012). Given two random variables and , with probability distributions and and joint probability distribution , the MI between and is defined as:

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

The MI between a feature and the class labels gives a useful indication of that feature’s relevance or importance (Brown et al., 2012). The probability distributions in Equation (1) are not known and are estimated using histograms.

The number of clusters to select, *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 evaluations, for cluster ranking or selection of *N*, are done on unseen test data using a five-fold cross validation (Bishop, 2003).

## Data Sets

Five remote sensing and one synthetic data set were used for comparing the proposed method against popular existing feature selection methods. The data sets are detailed in Table 1. The “difficulty” in the last column is calculated as , as in Brown et al. (2012), where is the number of objects, the number of features and the number of classes. Smaller values indicate that the data is less representative of the underlying class distributions which results in more challenging feature selection and classification tasks. The Spekboom set consists of 46 spectral and textural features derived from four band multi-spectral, 0.5 m spatial resolution aerial imagery. The classes represent three types of vegetation found in the Little Karoo, a semi-arid region in South Africa. It was created as part of a vegetation mapping project being conducted by the authors.

The two class synthetic data set was generated to have redundancy amongst the features. The first five features for class were generated from a normal distribution, , with mean mean and standard deviation of one (). The mean, , of each distribution, was generated from the standard normal distribution, . The same number of objects were generated for each class. To introduce redundancy, an additional five features were generated by adding normally distributed noise, , to the original five features. A further five redundant features were similarly generated, but by adding normally distributed noise, , to the original features. Finally, two spurious features, sampled from a standard normal distribution, , were added to the data set.

The Statlog Landsat and Urban Land Cover data sets were obtained from the UCI Machine Learning Repository (Lichman, 2013). The Statlog Landsat features are generated from six band multi-spectral pixel values in three by three neighbourhoods. The data set consists of six land cover classes. The features of the Urban Land Cover data set are comprised of multi-scale spectral, size, shape and textural measures, derived from high resolution aerial imagery (Johnson and Xie, 2013).

Kennedy Space Centre (KSC) and Botswana are public hyperspectral data sets with vegetation and land cover classes (GIC, 2014). The Botswana data were acquired by the Hyperion sensor on board the NASA EO-1 satellite and consist of 145 bands in the 400-2500 nm portion of the spectrum, at a 30 m pixel resolution. The KSC data were acquired by the NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor and consist of 176 bands in the 400-2500 nm range, acquired at a spatial resolution of 18 m.

Table 1 Data sets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Abbrev. | Features | Objects | Classes | Difficulty |
| Spekboom | Spekboom | 46 | 57877 | 3 | 419.40 |
| Synthetic | Synthetic | 17 | 10000 | 2 | 294.12 |
| Statlog Landsat | Landsat | 36 | 3756 | 6 | 17.39 |
| Urban Land Cover | Urban | 147 | 261 | 9 | 0.20 |
| Botswana | Botswana | 145 | 1330 | 14 | 0.66 |
| Kennedy Space Centre | KSC | 176 | 1365 | 13 | 0.60 |

## Evaluation

The proposed Feature Clustering and Ranking (FCR) method was compared to a number of other established feature selection methods. We adopted a similar, although reduced, evaluation approach to that of Brown et al. (2012). The compared methods included the standard selection approaches of ranking, FS and BE. These standard approaches and FCR, were each evaluated with two different feature relevance criteria: MI and the naïve Bayes classification accuracy. The MI relevance criterion for FCR and ranking approaches finds the MI between individual features and the class labels. To integrate the MI relevance criterion into FS and BE, it is necessary to compute the MI of a set of multiple candidate features with the class labels. In this situation, the candidate features are first merged into a joint variable and then the MI of the class labels with this joint variable is computed (Brown et al., 2012). We used histograms with ten bins along each dimension to approximate probability densities for both the MI and naïve Bayes criteria (Webb, 2002). This approximation is made to avoid difficulties and inefficiencies associated with estimating probability densities for continuous variables (Brown et al., 2012).

Brown et al. (2012) compared the performance of a several feature selection criteria in redundant high dimensional spaces, and found the JMI criterion gave the best overall performance in terms of classification accuracy and stability. Based on these results, FS with the JMI criterion was included in our study to represent the “state of the art” performance. The JMI measure for feature is

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

where are the class labels and is the set of previously selected features. JMI considers the MI between the class labels and the joint variables , which are the pairwise combinations of the candidate feature with each feature already selected. It measures how well the candidate feature complements selected features in describing the class labels. The evaluated methods and their criteria are detailed in Table 2.

Table 2 Method and criteria combination

|  |  |  |
| --- | --- | --- |
| Method | Criterion | Abbreviation |
| Forward selection | Joint Mutual Information | FS-JMI |
| Feature Clustering and Ranking | naïve Bayes | FCR-NaïveBC |
| Ranking | naïve Bayes | Rank-NaïveBC |
| Forward selection | naïve Bayes | FS-NaïveBC |
| Backward elimination | naïve Bayes | BE-NaïveBC |
| Feature Clustering and Ranking | Mutual information | FCR-MI |
| Ranking | Mutual information | Rank-MI |
| Forward selection | Mutual information | FS-MI |
| Backward elimination | Mutual information | BE-MI |

To quantify the stability of the selected features, we used the consistency index developed by Kuncheva (2007). If and and are subsets of the full feature set , with , and , the consistency index is

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

It’s value lies in the range , where positive values indicate similar sets, zero indicates a random relation and negative values indicate an anti-correlation between the feature sets(Kuncheva, 2007). To evaluate stability for a particular method, we select features from bootstrap samples of the data. The consistency index is found for each pairwise combination of selected features over ten bootstraps of the data. These pairwise consistency indices are averaged to give a measure of overall stability.

A k-nearest-neighbour (k-NN) classifier (with ) was used to evaluate the accuracy of the features selected by each method. k-NN is a generic classifier that makes no assumptions about the data and requires no tuning. While other classifiers may be more accurate in particular situations, k-NN allows a relative comparison of the feature selection methods, independent of the influence of classifier tuning for specific data. For each of the feature sets found from the bootstrap samples, the k-NN accuracy was found as the average per-class accuracy from a ten-fold cross validation. For each method and data set combination, an overall accuracy was computed as the average of the bootstrap accuracies.

The number of features to select for each data set was fixed across methods. This parameter was selected by using the accuracy of a k-NN classifier (), trained on the first *N* features selected by FS-NaïveBC, as the criterion in a grid search. A low value of *N* that achieved good accuracy was selected for each data set.

The FCR methods (FCR-MI and FCR-NaïveBC) required some specific treatment to integrate them into the evaluation. The dissimilarity threshold at which to extract the feature clusters from the hierarchy was determined by visual inspection of the dendrogram for each dataset. Thresholds at which the clusters were both stable and strongly correlated were favoured (see section 2.1 for a description of these concepts). The chosen threshold was used across all bootstraps of the data set. After bootstrapping, clusters were assigned unique indices, ensuring identical clusters had the same index. The consistency index was then found using the selected cluster indices rather than feature indices. This was done to simulate hand-selection of preferred features from the best clusters for each bootstrap, while, in practice, allowing the FCR algorithm to automatically choose the top ranked feature from each cluster (for the sake of simplicity and speed). In other words, the cluster index was used to represent the index of the preferred feature that could otherwise have been selected by hand from the cluster contents. The clusters represent a reduced set of possible indices compared to the features. This reduced set in turn reduces the possibility of stability errors. It is acknowledged that this formulation of the consistency index may favour FCR, but it was regarded as a necessary modification to allow comparison of the different methods.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

We followed a similar approach to that of Brown et al. (2012), for computing a single “non-dominated” ranking of the methods that considers stability and accuracy performance simultaneously. The concept of “Pareto optimality”, is used to find a single optimal solution in terms of multiple criteria. In the context of our evaluation, the “Pareto front” is the set of methods on which no other method can improve without degrading either the accuracy or stability. The methods in this set are called “non-dominated” (Mishra and Harit, 2010). Successive Pareto fronts can be formed iteratively by finding the current Pareto front of the set of methods that excludes members of the previous fronts. A method was thus given a non-dominated rank of *N* if it was a member of the *N*th Pareto front. The average of the non-dominated ranks for each method over the six data sets was used to produce an overall ranking. Using the approach of (Demšar, 2006), a Friedman test and post-hoc Nemenyi tests were conducted on the method ranks to look for statistically significant differences between individual methods.

The bulk of the software implementation was done in MatlabTM, making use of the PRTools toolbox (TU Delft, 2015). The MI and JMI criteria were however computed using the FEAST (FEAture Selection Toolbox) C++ implementation (Brown et al., 2012).

# Results and Discussion

The chosen FCR correlation dissimilarity thresholds and corresponding number of features selected for each data set are detailed in Table 3.

|  |  |  |
| --- | --- | --- |
|  |  | Dissimilarity Threshold |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

The results of the stability and accuracy evaluations for each method and data set combination are shown in Figure 2 and Figure 3 respectively. The methods appear along the x axis in order of their mean stability in Figure 2, and mean accuracy in Figure 3, over the six data sets. FS-MI was the most stable overall, but had one of the poorest accuracies. Similarly, FS-NaïveBC is the most accurate overall, but is the least stable. While neither FCR-NaïveBC nor FCR-MI achieve the best overall accuracy or stability, they are amongst the top three methods for both performance measures. The wide range of stabilities confirms the sensitivity of some methods to variations in the data. The method accuracies span a smaller range than the method stabilities. Nonetheless, there are substantial differences in accuracy between the best and worst methods. Compared to the other data sets, the stability of the Spekboom, Synthetic and Landsat data is noticeably superior. As reflected in the “difficulty” values in Table 1, these data sets are more representative of the underlying distributions and are thus less sensitive to disturbances.

The ranking methods, Rank-MI and Rank-NaïveBC both have poor accuracy performance. This is expected as these methods do not consider feature complementarity and only measure relevance of features in isolation. The relatively poor accuracy and stability of FS-JMI was surprising in the context of the results of Brown et al. (2012), where it produced the best overall performance. Perhaps FS-JMI is more competitive when applied to higher dimensional data sets, containing hundreds or thousands of features, such as those that were used in Brown et al. (2012). Another contributing factor to this anomaly may be that, while we fixed the number of selected features to a small number, Brown et al. (2012) evaluated performance for a variable number of features and found that for some data, the superiority of the FS-JMI method only became apparent after many (more than ten) features had been selected. The FS-JMI results nevertheless provide a benchmark that helps confirm the usefulness of the FCR method for the type of data investigated in our study.

As with classifier design, there is a “curse of dimensionality” problem with computing the MI of joint variables. As the number of features increases, the number of objects needed to adequately represent the feature distribution increases exponentially (Brown et al., 2012). For this reason, the MI criterion is not well suited for evaluating the BE method, which requires computation of the relevance criterion for the full feature set. This likely explains the poor performance of BE-MI in terms of both accuracy and stability.



Figure 2 Method stability per data set (methods along the x axis are ordered by their mean stability over the data sets)



Figure 3 Method accuracy per data set (methods along the x axis are ordered by their mean accuracy over the data sets)

The method execution times, summed over the six data sets, are provided for reference in Table 4. The execution time of FCR competes well with the other methods. The NaiveBC criterion is slower to compute than the MI criterion as it uses a five-fold cross-validation, implemented in MatlabTM, to evaluate the classification accuracy. MI is computed once-off using the efficient FEAST C++ implementation. Methods using the NaiveBC criterion are consequently slower than their MI counterparts. JMI is faster than the related FS-MI method, as the criterion only requires MI computations between pairwise combinations of features and the class labels, while the MI criterion is evaluated on the combination all selected features. BE is known to be less efficient than FS (Guyon and Elisseeff, 2003), and is slowest of the tested methods.

Table 4 Method cumulative execution time over all data

|  |  |
| --- | --- |
| Method | Time (s) |
| Rank-MI | 5.77 |
| JMI | 6.25 |
| FCR-MI | 6.46 |
| FS-MI | 24.71 |
| FCR-NaïveBC | 122.23 |
| Rank-NaïveBC | 145.55 |
| BE-MI | 815.23 |
| FS-NaïveBC | 831.99 |
| BE-NaïveBC | 5726.67 |

Table 5 presents the non-dominant ranking of the methods, in terms of both accuracy and stability. The best ranked method overall is FCR-MI, with FCR-NaïveBC, FS-MI and FS-NaïveBC occupying the second rank position. While the FS-MI and FS-NaïveBC produced the best performance for stability and accuracy respectively, FCR-MI and FCR-NaïveBC achieved a better compromise between these two measures. The Rank-NaïveBC, Rank-MI and BE-MI methods are ranked lowest due to the known limitations of these methods. The Friedman and Nemenyi tests did not reveal any significant differences between the method rankings over the data sets. Further data sets and or bootstraps would need to be included in the evaluation to show significance.

If the clustering step were omitted, FCR-MI and FCR-NaïveBC would simplify to Rank-MI and Rank-NaïveBC respectively. FCR-MI and FCR-NaïveBC show a substantial improvement in performance compared to Rank-MI and Rank-NaïveBC which lends support to the effectiveness of the clustering step. Considering the combination of the MI and naïve Bayes criteria with each method in isolation, there is a general trend for MI to produce better stability and naïve Bayes to produce better accuracy. While FCR works well with either criterion, the results favour the use of MI as it is faster and produces a better non-dominant ranking than naïve Bayes. On the whole the evaluation study shows the FCR method to be effective at selecting accurate and stable features from high dimensional remote sensing data containing redundancy.

Table 5 Non-dominated ranking of methods by accuracy and stability

|  |  |
| --- | --- |
| Method | Rank |
| FCR-MI | 1.50 |
| FCR-NaïveBC | 1.83 |
| FS-NaïveBC | 1.83 |
| FS-MI | 1.83 |
| BE-NaïveBC | 2.83 |
| JMI | 3.00 |
| Rank-NaïveBC | 3.33 |
| Rank-MI | 3.50 |
| BE-MI | 3.50 |

# Conclusions

Feature redundancy 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 a relevance criterion evaluated 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 redundancy from the data. The ability to hand-pick features is beneficial as it allows other factors, such as speed of computation and physical interpretability, to be considered when determining an effective feature set.

The effectiveness of the proposed FCR method was evaluated by comparing its accuracy, stability and execution time to a set of popular feature selection methods. A set of five remote sensing and one synthetic data set, all containing redundancy, were used to compare the methods. The feature selection methods were each tested in combination with two criteria for feature relevance: the MI between the candidate feature(s) and the class labels, and the accuracy of a naïve Bayes classifier trained on the candidate feature(s). The FS method in combination with the JMI criterion is recognised as providing competitive performance on high dimensional problems (Brown et al., 2012), and consequently was also included in the tested methods. FS-NaïveBC provided the best accuracy performance but the worst stability performance. In a similar vein, FS-MI provided the best stability performance but the second worst accuracy performance. The FCR method performed well overall, with both naïve Bayes and MI criteria. Although FCR did not quite achieve the best performance in accuracy or stability alone, it was the highest ranked method when considering the accuracy and stability measures in combination. Another benefit of FCR is its relative speed compared to greedy search FS and BE type methods.

While the MI criterion is attractive in its ability to describe non-linear correlations and complementarity between features (Brown et al., 2012), the curse of dimensionality limits its application in practice, particularly for high dimensional problems. Part of the motivation for the JMI and other criteria presented in Brown et al. (2012), is to circumvent this problem by formulating low dimensional approximations to MI. The naïve Bayes criterion also avoids the curse of dimensionality as it treats the features as independent. Intuitively, one may suspect that this assumption is overly crude to be useful in practice, but naïve Bayes does perform well for a variety of real world problems (Hand and Yu, 2001), including the data sets used in our study. In FCR, the features are evaluated in isolation, so dimensionality and representivity issues are less of a concern than for the FS and BE methods which evaluate combinations of features. Although FCR worked well with either criterion, the results favour the use of MI as it was faster and produced a better non-dominant ranking than naïve Bayes.

The need for user specification of dissimilarity threshold is a weakness of the FCR method. This is a subjective choice and different thresholds can lead to different sets of selected features. No user intervention is required for the other methods considered in this study. A possible way to automate this choice would be to extract clusters from all levels in the hierarchy, select a set of features from the best clusters at each level, and then use a performance measure such as the accuracy of a k-NN classifier to choose the best set of features overall. The need for visual inspection of the dendrogram to make the choice of dissimilarity threshold also limits the dimensionality of data that the FCR method can practically be applied to. For data sets of hundreds or thousands of features, the dendrogram would likely be too cluttered to make a sensible choice and a feature selection algorithm other than FCR may be more appropriate. It is worth noting that for problems where feature stability and user specification of preferred variables are not required, FS-NaïveBC may be a more sensible choice of feature selection method. It achieved the best accuracy results and does not require any user intervention as with FCR.

Our reasoning for the choice of dissimilarity metric for clustering, clustering algorithm and feature relevance measure was loose and other options could also prove effective. A possible limitation of the proposed method is the use of ranking for selecting the best clusters. Ranking is known to be sub-optimal in the presence of redundancy (Cover, 1974). While much of the feature redundancy will be eliminated by selecting individual features from correlated clusters, it seems likely that remnant redundancy will be present in many problems. 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 selection. This could conceivably be done by adopting a FS or BE type approach to clusters rather than features. Another aspect of our method that could benefit from further investigation is the correlation coefficient dissimilarity metric. It would be preferable to use a metric that can describe non-linear correlation, such MI or 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

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65, 2–16. doi:10.1016/j.isprsjprs.2009.06.004

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

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

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

GIC, 2014. Hyperspectral Remote Sensing Scenes [WWW Document]. URL http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral\_Remote\_Sensing\_Scenes

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.

Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J., 2004. Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence in Medicine 31, 91–103. doi:10.1016/j.artmed.2004.01.007

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.

Johnson, B., Xie, Z., 2013. Classifying a high resolution image of an urban area using super-object information. ISPRS Journal of Photogrammetry and Remote Sensing 83, 40–49. doi:10.1016/j.isprsjprs.2013.05.008

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

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

Kuncheva, L.I., 2007. A stability index for feature selection, in: International Multi-Conference: Artificial Intelligence and Applications. IASTED, Innsbruck, Austria, pp. 390–395.

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

Lichman, M., 2013. UCI Machine learning repository [WWW Document]. URL http://archive.ics.uci.edu/ml

MathWorks, 2016. Hierarchical Clustering - Statistics and Machine Learning Toolbox [WWW Document]. Matlab Documentation. URL https://www.mathworks.com/help/stats/hierarchical-clustering.html (accessed 3.1.17).

Mishra, K.K., Harit, S., 2010. A fast algorithm for finding the non dominated set in multi objective optimization. Multi-Objective Optimization using Evolutionary Algorithms 1, 35–39.

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

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

TU Delft, 2015. PRTools [WWW Document]. URL http://prtools.org/prtools/ (accessed 10.20.16).

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.