LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-644160

Ensemble Feature Selection in Scientific Data Analysis

A. Sisto, C. Kamath

September 23, 2013

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Ensemble Feature Selection in Scientific Data Analysis

[1]Aaron Sisto, [2]Chandrika Kamath
Department of Materials Science and Engineering, Stanford University
Computation Directorate, Lawrence Livermore National Laboratory

Feature selection has emerged as an important method of both examining and predicting results from scientific experiments. However, current feature selection methods cannot consistently provide meaningful descriptions of feature importance in the complex, high-dimensional, natural datasets encountered in scientific data analysis. Here, an ensemble feature selection method is demonstrated to collectively reduce the bias associated with individual feature selection algorithms and provide uncertainty estimates on feature importance ranks for both synthetic and natural datasets.

Introduction

Interpretation of scientific data, generated both by experimental measurements and simulation, often necessitates identification of underlying mechanisms and correlation between input parameters and quantities of interest. This is particularly difficult if the system of interest is not well understood, or the high dimensionality of the data precludes analysis by direct observation [1]. In such cases, it is desirable to identify input parameters that are particularly influential in determining the behavior of measurable quantities so that experiments can be designed and executed more efficiently. For this purpose, data mining and machine learning algorithms have emerged in recent years as extremely powerful tools in analysis of massive, complex datasets [2, 3, 4]. The utility of such approaches is due, in part, to the category of algorithms that accomplish dimensionality reduction of the feature, or input, space of the dataset. Dimensionality reduction refers to either a subset or alternative representation of the input parameters, or features, which are thought to influence the output quantities [5, 6]. Advantages of such approaches include improved interpretability of the data by domain experts, visualization, reduced data storage requirements and improved predictive capabilities. The objective of dimensionality reduction algorithms is to attain a lower dimensional feature space which represents the output quantities equally well compared to the original features, and in doing so, describes the inherent dimensionality of the system. In many cases, the true objective in scientific data analysis is to discover causal relations between input and output quantities. However, rigorously describing causal relations generally remains elusive due to incomplete information available about the physical system as well as impractical computational cost associated with subspace searching [7, 8].

The category of dimensionality reduction in which a subset of features are chosen or ranked is referred to as feature selection (FS) [9]. Feature selection algorithms reduce the dimensionality of the feature space by either choosing an optimal subset of the original features or by ranking the individual features using a certain measure of importance.In contrast to FS methods, linear and nonlinear dimensionality reduction algorithms [10, 11] such as Principal Component Analysis (PCA) and Isomap transform the input parameters by

mapping the feature space onto a space of lower dimensionality. The primary difference between the two approaches is the representation of the reduced dimension of the feature space. In scientific data analysis, and specifically, in the context of this study, it is advantageous to retain the original representation of the data using feature selection, allowing domain experts to gain insight into the relations between experimental input/output parameters.

Filter Methods

A class of FS algorithms known as filter methods selects features based on correlations between each feature and the outcome class (c-correlations) [12]. Filter methods are simple in the sense that the feature selection is only biased by the choice of correlation measure in the filter method. However, a general expression for the importance measure in subset or feature ranking has not been established, although many specific formulae have been suggested. This is not surprising given that the "importance" of a feature set may change in meaning depending on the goal of the analysis. Nonetheless, the main objective in filter-based FS methods is to identify important features based on the principles of maximized relevance and minimized redundancy. Relevance refers to the degree of correlation between the feature and the outcome class, while redundancy refers to the degree of correlation between features (f-correlation). The latter type of correlation is generally not described within simple filter methods as additional complexity in the ranking algorithm is required to determine how the ranks change from those reflecting purely c-correlations [13, 14]. The discussion of correlation is kept abstract at this point, as many different descriptions are available, each identifying slightly different relationships between variables. Several quantities indirectly described by correlation, which are useful in feature ranking, are **separability, information, consistency**, **dependency**. Furthermore, these quantities are appropriate for choosing methods with fundamentally different biases, since each quantity describes a different representation of the data. **Separability** is a well-known feature ranking metric that shows the degree to which changes in class labels separate or divide the feature values. Features that exhibit high separability are capable of describing the class label independently, with specific feature values corresponding explicitly to specific class labels. **Information** is a general measure used to quantify the relative disorder in a dataset in the presence of an additional feature. If a feature adds structure to the data, thereby reducing the disorder of the initial dataset, it is considered relatively important. A reference for minimal information content is a purely random variable as the entropy associated with this variable is maximized and it can only add disorder to the data. **Consistency** in a dataset containing multiple data points, or instances, refers to the degree of differentiation between class labels and identical instances [15]. Low consistency indicates many recurring class labels for varying feature values. This makes determination of feature importance or predictability difficult as the correlation between all feature and class labels becomes unclear. Conversely, high consistency data corresponds to a minimal degree of uncertainty in the class labels and is ideal for analysis. However, recognition of either high or low consistency in a given dataset is extremely important and can even provide physical understanding of the dataset and experimental conditions. For example, low consistency could arise from high degrees of error introduced by the measurement device or this could evidence the presence of uncontrolled or unmonitored experimental parameters influencing

the outcome. The **dependency** of features in a dataset refers to f-correlation, defined above. Certain filter methods, such as information gain and Pearson coefficient, can quantify f-correlation, although this does not yield any inherent information about the overall feature importance in itself.

An additional consideration that must be taken into account when applying feature ranking methods is the presence of features that exhibit either partial correlation to the outcome, or seemingly zero relevance. Such features may be incorrectly highly ranked in the former case, and lowly ranked in the latter. To elucidate, high correlation to the class outcome does not necessarily preclude even higher correlation to other features or subsets. In fact, this situation is encountered frequently and consideration of only c-correlation results in multiple highly ranked feature which may each describe the outcome equally well. In a physical sense, this is perfectly acceptable and does not distinguish between features which may themselves be causally related. However, this also does not help to reduce the dimensionality of the system and thus, simplify the experimental conditions, and in this sense, the FS has not achieved its goal. Clearly, some measure of f-correlation is necessary in this case to prevent redundant feature ranking. Conversely, a variable or variable subset which is completely irrelevant with respect to the outcome may become extremely relevant when considered in conjunction with other features. Without searching through all likely feature subsets, this quality of the seemingly irrelevant variable may remain undetected by a simple filter method.

Finally, intelligent sampling of data points to use in feature selection is potentially necessary when experimental data is extremely noisy or inconsistent [16]. The method of active sampling can be extremely effective and identify truly important features, which have been misranked due to conditions of the data.

Wrapper Methods

A more complex set of feature selection algorithms known as wrapper algorithms seeks to provide a definition of "goodness" of the feature subset by using the accuracy of a classification method to evaluate a chosen feature subset [17]. Thus, each subset is chosen independently of the classifier, but the importance is gauged by the accuracy of an arbitrary induction algorithm. Wrapper methods often achieve higher classification accuracy while incurring a greater computational cost [18]. Thus, for datasets with a large number of features, such algorithms often become unfeasible. Additionally, the choice of feature subset is highly sensitive to both the filter method and the induction algorithm. One method to evaluate the bias associated with a filter method is to perform a wrapper feature selection using a common induction algorithm.

Hybrid methods involving classifier-FS algorithms have also been proposed in the context of wrapper FS. These methods seek to determine an optimal ensemble of feature subsets or feature weights using boosting [19]. Often, subsets or feature weights are optimized based on how well specific data points are predicted by an induction algorithm, with misclassified examples weighted either higher or lower. The flexibility of using multiple feature subsets or dynamic weights in this way can facilitate much higher classification accuracy.

Embedded Methods

Another common methodology employed to evaluate feature subsets is a combination of filter and classification algorithms in what are known as embedded methods. In embedded methods, a learning algorithm is used primarily to distinguish between different feature subsets, similar to wrapper methods. The filter methods discussed above are incorporated directly into the learning algorithm, with the objective function guiding the search through feature subsets. Thus, the filter is not used independently of the classifier, but both work self consistently to obtain an optimal feature subset for classification. Common embedded methods include decision trees and artificial neural networks. These methods have been found to achieve much greater accuracy in classification; the computational cost is typically greater than filter methods but less than wrapper methods. Furthermore, in contrast to filter methods, both the bias of the filter method and the inference algorithm simultaneously impact the resulting feature selection and a high degree of understanding about both types of algorithms is necessary for proper use.

The latter two FS methods (wrapper and embedded) denote a wide range of algorithms which perform extremely well for classification with complex, natural datasets. However, due to the sensitivity of the resulting subset or feature ranking on the classification procedure, in general, the selected feature subset cannot be considered to have any physical interpretability. Furthermore, the initial correlation criteria describing feature relevance and redundancy is obscured by the objective of minimized misclassification error, completely removing the utility of these methods as feature ranking algorithms except in the sense that features influence the classifier. In fact, using inference as a feature selection metric can often further mystify the truly important input quantities because, as shown elsewhere, multiple feature subsets can yield excellent classification results [20, 21]. Thus, the classification algorithm is often unable to resolve the true importance ranking of features or subsets, corresponding closely to a fictitious causal network.

The need for both feature ranking and classification approaches arises from the confluence of two main goals: physical insight and predictive ability, both related to selection of feature subsets. The first, identification of features important to observable quantities, is often much less deterministic, given the results of filter methods described above. To illustrate, various sources of uncertainty arise from both the condition of the dataset as well as the data analysis methods. Often, experimental datasets do not include a large enough number of measurements to adequately sample the entire input/output space. Additionally, scientists are limited by which input and output quantities can be controlled and measured accurately. Furthermore, the measurement itself contributes varying degrees of noise into the physical parameters being measured, further obscuring direct observation of the quantities of interest. These conditions alone introduce a large amount of ambiguity into the data being analyzed and require careful choices in the selection of FS criteria. The bias associated with FS methods is also not easily quantifiable and estimates of the variability of arbitrary FS algorithms when applied to natural datasets are not well defined.

Validation Methods

Various methods have been proposed to account for the variability of feature selection results with respect to conditions of the data or data mining algorithms. Variance in feature ranks between different feature selection methods is important to identify, and is often undesirable. In some cases, high variance in the feature ranks is due to inadequacy in the correlation or FS method or inability of the chosen feature subset to accurately describe the entire range of outcome values. Additionally, varying degrees of variance can also provide more subtle insight into the underlying structure of the data. Knowledge of the variability of quantities related to individual features can guide experimentalists to explore narrower ranges in the feature values to better understand the physical system. Finally, overall uncertainty estimates in the feature rankings can provide at least a heuristic means of choosing future experimental approaches.

A simple validation approach is to introduce a "fake" variable, randomly from a Gaussian distribution and compare this feature on the same basis as the features in the dataset [22]. The random variable has absolutely no correlation to the outcome or other features and should be ranked lowest as a result. Coincidently, extremely noisy features or uncorrelated variables can achieve the same level of importance in FS algorithms, or lower. Thus, the random variable can be used to eliminate uncorrelated features by providing a lower bound on the correlation factor, intrinsic to each method.

Many of the validation methods often involve bootstrapping approaches to gauge the variance in the feature rankings or classification results with respect to variability in the data [23, 24]. The approach of bootstrapping involves sampling from a subset of the instance array, either with replacement or without. The frequency with which individual variables appear in important feature subsets or are highly ranked in the individual bootstrap iterations can be used to judge to overall influence of a feature. Additionally, random subsets of the instance array taken from specific ranges of a particular variable can yield more detailed insight into the structure of the measured quantities. Predictive accuracy can be quantified from these results as well, offering a quantitative description of small sample domain effects.

Methods that attempt to quantify the uncertainty associated with FS algorithm bias involve similar bootstrapping approaches, but also necessarily include multiple methods operating on the same random subset of the data. FS is conducted over a population of models and then important feature subsets are extracted based on the frequency with which they appear in the various models. However, a well-defined ensemble FS method has not yet been proposed and a means to quantify uncertainty due to FS model bias remains unresolved. The differing descriptions of correlation given by various feature selection algorithms is worrisome due to the fact that most scientific data mining studies involve only a small number of feature selection algorithms. In such cases, the variability in the feature subset rankings is not well described and the results therefore have a low degree of certainty. Furthermore, in the absence of a classification result to determine the performance of the feature subset, it is not correct to consider any feature subset incorrect. Instead, different descriptions of feature importance are achieved by different methods and the results can be contradictory when analyzing natural datasets. These datasets are often noisy, incomplete and the available input and output parameters can exhibit varying types of correlation in different ranges. Analogous to the No Free Lunch Theorem [25], no individual

feature selection algorithm can adequately capture all correlations in complex, high-dimensional data and at least some degree of variability between different algorithms is expected.

Finally, associated with uncertainty quantification estimates from FS model bias, consensus-based selection methods [26] have been widely used, particularly in datasets with a large number of features. Such methods typically involve comparing feature rankings across multiple FS algorithms and selecting features that occur more frequently, with higher rankings. The procedure for making these types of selections remain primarily heuristic given the high sensitivity of ranking distributions on the type and number of features in the dataset and the choice of FS algorithms.

Feature Selection Methods

The common filter selection methods described above often involve computing the correlation measure between features or classes. Basic filter methods rely on the description of c-correlation to determine the degree of relevance of individual feature or feature subsets. This general quantity of correlation can be further divided into classical linear correlation and information theory. Nomenclature used in the following algorithms is as follows: A capital letter (e.g. X, Y) represents a random variable, a subscript (e.g. X_i, Y_i) represents the i^{th} value (instance) of the corresponding random variable. Data points (row vectors) and tensor quantities are written in bold, with a superscript (e.g. $\mathbf{X}^i$, $\mathbf{Y}^i$) representing the i^{th} feature or element in the vector. Classical linear correlation methods such as Pearson correlation coefficient and Chi-squared statistic are often useful in quickly determining the degree of linear correlation between input and output quantities or between inputs.

Pearson Correlation Coefficient

The Pearson correlation coefficient is an extremely simple and fast description of correlation, given by the expression:

$$\rho_{Y,X} = \frac{COV(Y,X)}{\sigma_X \sigma_Y} \tag{1}$$

here, COV is the covariance matrix and σ is the variance of each continuous feature. Nominal features can be used as well, with an appropriate discretization scheme. Nonlinear correlation can also be determined by transforming the input or output quantities first by a nonlinear mapping and subsequently applying the linear correlation method. However, these method are limited in usefulness given complex functional relationships and noisy, high-dimensional data. Despite this, Pearson correlation can be an extremely useful and efficient to use, especially in the present ensemble approach as a means to contrast the information-based methods.

A second category of correlation measure is that associated with information theory. Information theory relies on the description of Shannon entropy, expressed identically to the concept of entropy in thermodynamics:

$$H(X) = -\sum_i P(X_i)\log\big(P(X_i)\big) \tag{2}$$

In general, broad distributions, such as those found in noisy data, result in higher entropy, while highly localized distributions result in lower entropy. Although the expression is straightforward to compute, the distribution functions used to compute the entropy give rise to various descriptions of correlation between input or output quantities. These various correlation algorithms are summarized below.

Information Gain

The Shannon entropy is used to define the degree of importance of random variable Y given information about X. This method requires nominal, normalized variables, but can be extended to numerical values if an appropriate discretization scheme is used. Additionally, the lack of a normalization factor in the IG expression below results in a bias toward features that take on a higher number of values.

$$IG(Y|X) = H(Y) - H(Y|X) \tag{3}$$

This method can also be used to calculate f-correlations since the operator, H, acts on one or more arbitrary random variables.

Symmetrical Uncertainty

The normalization requirement and bias toward higher numbers of values in the IG expression is removed in this expression, which gives the normalized IG.

$$SU(Y,X) = 2\left[\frac{IG(Y|X)}{H(Y)+H(X)}\right] \tag{4}$$

Similar to information gain, this method describes the information content in a given feature, while including descriptions of the entropic structure of both feature and class, in addition to the quantity calculated in information gain. This method can also be used to calculate f-correlations.

Kullback-Leibler Distance

The KL distance is an entropy based measure, computed for each feature j, as a summation over the distributions of feature values X_i given class labels $Y_{m,n}$. This quantity measures the degree to which features are split by class labels and is related to the concept of separability discussed above.

$$\Delta_j = \sum_m^c \sum_n^c \sum_i P(\boldsymbol{X}_i^j|Y_m) log\left(\frac{P(\boldsymbol{X}_i^j|Y_m)}{P(\boldsymbol{X}_i^j|Y_n)}\right) \tag{5}$$

where the number of class labels in Y is denoted by the summation over c values.

Cross-Entropy

The f-correlation quantities are computed from the conditional probability distributions in the class labels, given two feature vectors.

$$H(\boldsymbol{X}^i, \boldsymbol{X}^j) = -\sum_i P(Y_i|\boldsymbol{X}^j) log P(Y_i|\boldsymbol{X}^j) \tag{6}$$

RANK algorithm

An instance-based similarity measure is used to compute the information gain for each feature, proposed by Dash and Liu [27]. The definition of distance **D** is taken to be Euclidean in the case of numerical features and the Hamming distance is used with nominal features.

Numerical features

$$\boldsymbol{S_{i,j}} = e^{-\alpha \boldsymbol{D_{i,j}}} \tag{7}$$

Euclidean distance

$$\boldsymbol{D_{i,j}} = \left[\sum_{\boldsymbol{k}} \left(\frac{(\boldsymbol{X_i^k} - \boldsymbol{X_j^k})}{\max(\boldsymbol{X^k}) - \min(\boldsymbol{X^k})}\right)\right]^{\frac{1}{2}} \tag{8}$$

Nominal features using Hamming distance

$$\boldsymbol{S_{i,j}} = \frac{\sum_k |\boldsymbol{X_i^k} = \boldsymbol{X_j^k}|}{N} \tag{9}$$

$$H = -\sum_i \sum_j -\boldsymbol{S}_{i,j} log \boldsymbol{S}_{i,j} - (1 - \boldsymbol{S}_{i,j}) log(1 - \boldsymbol{S}_{i,j}) \tag{10}$$

α is calculated from the average distance between features:

$$\alpha = \frac{-\ln(0.5)}{\bar{\boldsymbol{D}}} \tag{11}$$

Similarity between data points arises from changes in individual feature values. Thus, this method uses the concept of consistency, described above, to relate the structure of the instances in the dataset in the presence of a specific feature to information content via the similarity matrix. Increased similarity-based entropy associated with the removal of a certain feature indicates a higher level of importance than a feature that causes a decrease in entropy.

Many other FS methods have been proposed in the literature and these 5 were chosen in an effort to sample dissimilar descriptions of correlation, as described above. For example, a method such as Gain Ratio, which is very similar to Information Gain and Symmetrical Uncertainty, was found to give almost identical results to one or both of these other methods, in all test cases. The choice of the similarity-based RANK method is also not unique or necessarily optimal. The nearest-neighbor FS method [28] is also a similarity-based metric, involving the calculation of distance between datapoints. Instead of weighting feature importance using an expression of the entropy, distances from the outcome are weighted and compared for each feature to arrive at an importance factor. Many methods have also been proposed as similarity preserving [29]. These methods, including Laplacian and Fisher score, could also provide meaningful ranks on limited data types where other methods fail to distinguish between features. The choice of Pearson coefficient is similarly not unique although this method has been used extensively in both feature selection and in calculating f-correlations in conjunction with higher accuracy methods. Other statistical methods such as t-test, Chi-squared statistic, ANOVA and Wilcoxon are widely used.

Feature-feature correlations

As described above, an extension of simple filter methods to determine the degree of redundancy between features requires computation of f-correlation. As described above, many of the included methods can describe f-correlations as well as c-correlations. However, the incorporation of these quantities into feature ranking requires a modification of the simple filter algorithms. Specifically, filter algorithms typically output an importance factor associated with the internal degree of c-correlation for each feature. These are not comparable between filter methods without some kind of normalization scheme, which has not been established globally. A comparable output quantity, however, is the feature ranking which is computed by sorting the feature importance factors. The ranking reflects only the degree of relevance of each feature to the class outcome. Incorporation of f-correlation in feature ranking has been proposed in algorithms such as the Fast Correlation-Based Filter (FCBS) [30] and Markov-Blanket [31], as well as indirectly in full feature subset search algorithms such as FOCUS[32] and RELIEF[33]. The first two methods, FCBS and MB allow for systematic inclusion of f-correlation in feature ranking and approach the concept of causal relations much more closely [34]. However, these methods still rely on a specific description of both f- and c-correlation, and thus, incur the bias from the corresponding correlation algorithm.

The recurring issue of bias in feature ranking has led some researchers to propose methods of quantifying uncertainty in the output or feature selection methods[35]. This seems a necessary step in the development of a robust, accurate feature selection approach, with potential to improve both accuracy and performance of classification while yielding important domain information. In this report, an ensemble algorithm is proposed to calculate feature rankings based on filter methods. The ensemble approach achieves a high degree of computational efficiency and scalability due to the simplicity of the involved c-correlation algorithms, while providing robust feature selection criteria and uncertainty quantification reflecting the ranking distributions from each individual filter. As test cases, both natural and synthetic datasets are employed to evince the influence of bias in the filter method as well as robustness of the ensemble algorithm. Finally, a natural dataset obtained from

measurements taken from additive manufacturing simulations is examined both with the ensemble method, as well as more complex methods such as FCBS and MB. These results were obtained using a preliminary version of the data and a more complete analysis of the data will be presented elsewhere.

Ensemble feature rank algorithm

The ensemble rank of each feature in a dataset is computed based on a weighted average over the distributions of ranks from each FS algorithm. The individual FS algorithm rank distributions are constructed by bootstrapping samples of the instance array from the original dataset and computing the ranks for each feature. After iterating in the bootstrapping section, the ranks from each iteration are binned and normalized, resulting in a distribution of ranks for each feature in the original dataset, for each FS algorithm. The individual distributions are then used as weights in the average of feature ranks over each FS algorithm. Within the individual rank distributions for each FS algorithm, the average and variance can be used to measure the certainty in the rank output by the specific FS algorithm, since the mean rank for each feature corresponds to the rank computed for the full instance array. The mean and variance of the ranks for each feature after the ensemble procedure is used as an indication of the uncertainty in the overall feature rank. Because FS algorithms with fundamentally different descriptions of correlation are used in the ensemble rank, robustness is assumed to the extent that the FS algorithms collectively are sufficient to identify the major c-correlations in the dataset. The algorithm is delineated as follows:

<u>Ensemble Rank Algorithm</u>

1. Iterate over N FS methods

2. Iterate over M bootstrap samples of the original instance array; Compute feature ranks at each iteration

3. Construct the FS method rank distributions for each feature from ranks over all bootstrap iterations

4. Compute aggregate ranks for each feature from the (M) rank distributions in each (N) FS methods from the weighted average:

$$\mathrm{Rank(i)} = \frac{\sum_{\mathrm{method}} \sum_{\mathrm{rank}} \mathrm{P(rank)} \cdot \mathrm{rank}}{\mathrm{N} \cdot \mathrm{M}}$$

5. Compute error bounds from the standard deviation in the weighted average computed in 4, with respect to all FS methods.

Description of synthetic test datasets

Synthetic datasets are extremely valuable as diagnostic test cases for feature selection algorithms because they provide a means to explicitly probe the bias associated with a

specific algorithm. With precise control over the types of datasets and correlations in synthetic data, extreme circumstances such as high-dimensionality, noise, nonlinear correlation and partial relevance or redundancy can be explored. In this study, synthetic datasets were generated by embedding low dimensional, nonlinear functions of random variables in a higher-dimensional manifold.

Swiss

Two dimensional data is generated from functions of the form:

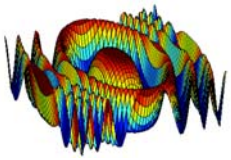
$$[x \cdot \sin(x) \quad x \cdot \cos(x)]$$

where x is a random variable varying from [-1:1] with 1000 values. These data are then embedded in a third dimension of noise. The resulting 3-dimensional data is illustrated below:

Cosinus-Hills

An output dataset is generated from a random variable taking 1000 values in the range [-1:1]:

$$[\cos(2\pi({x_1}^2 + 3 \cdot {x_2}^2)]$$



Difficult

A 5-dimensional dataset of random variables is embedded in a 10-dimensional nonlinear manifold. The dataset is named “difficult” because none of the original generating variables remain in the final 10-dimensional manifold. Instead, they are each transformed by different nonlinear functions. This dataset is well-suited for nonlinear dimensionality reduction methods, but most likely problematic for feature selection algorithms due to the absence of the inherent dimensional representations in the final dataset.

CORRAL

The CORRAL dataset was originally proposed as a means to test feature selection in the presence of highly correlated feature, a nonlinear concept and irrelevant features. The first four features completely determine the target concept:

$$(A \text{^} B) \vee (C \text{^} D)$$

The 5th feature is irrelevant and the 6th feature is highly correlated with a 25% error rate.

Description of natural test datasets

Iris

This dataset is extremely well-known in the machine learning community as it has proved numerous studies of feature selection and dimensionality reduction since its inception. The dataset contains 4 features and 1 nominal class taking on 3 values. It has been determined that only two features are relevant to the class and even one independently can accurately predict the class outcome.

Table 1. Summary of datasets

Dataset	# Features	# Class Labels	# Instances	Important Features
IRIS	4	3	150	**{3,4}**
CORRAL	6	2	256	**{1,2,3,4}**
Cosinus-Hills	2	continuous	1024	**{1,2}**
Swiss	2	2	1000	**{1,2}**
Difficult	10	continuous	1024	**{ }**

Results

The calculated feature ranks are shown below both for each FS method, as well as the ensemble ranks. To understand the volatility in the ranks between feature methods, the governing correlation metric associated with each method must be considered. Additionally, knowledge of the feature ranks from individual methods does not provide a clear level of uncertainty in the outcome. Furthermore, features can be scored very closely, but given very different ranks since the rank is simply a sorted representation of the calculated importance values. Thus, both the variability between methods as well as the uncertainty associated with a simple rank representation in each method must be assessed within the ensemble method. The latter issue is also resolved by the bootstrapping feature ranking step, which is carried out for each FS method individually. Averaged feature ranks are calculated from 1000 bootstrap iterations with a subset size of 70% of the original number of instances. Cross-validation was performed to determine the threshold for convergence in the ensemble ranks for both number of iterations and size of the random subset. By calculating a distribution in the feature ranks, features that are very close in importance factor in a single calculation involving the training data exhibit broadened, overlapping feature rank distributions. This indicates that to some extent, the ranks of either feature can be reversed or even become interchanged with other features ranked much differently in the calculation with the entire training set. The former issue is resolved more easily from the general concept of the ensemble ranking method, as the fundamental objective in such methods is to evince the overall estimate and uncertainty in the ranks with respect to individual sampling methods.

The rank distributions calculated from bootstrap sampling for each individual FS method are useful in understanding certain characteristics of the data or usefulness of the method. For example, in Figure 1 below, within each method, the features are ranked from 1

to the total number of features, including any additional noise features, with the highest rank corresponding to the highest degree of importance. The bootstrap algorithm results in the rank distribution shown, where the normalized frequencies of the ranks corresponding to each feature are shown along the vertical axis.

The discrepancy between methods is immediately evident from the IRIS dataset tests. Although features 3 and 4 are consistently ranked above 1, 2 and the random index, 5, the individual rankings are highly sensitive to the correlation method (Figure 1). The distribution of ranks within each method provides insight into both the condition of the data, as well as the performance of the method. The average ranks of features 3 and vary depending on the method and even within a single method, there is some degree of variation with respect to sampled data. These results yield much more information than single feature ranks with the training data and the ensemble ranks reflect these characteristics. It can be seen in Figure 2 that the standard error estimates of the 3rd and 4th features overlap, corresponding to finite probability that these features are weighted identically. Examination of the f-correlations between each of these features and a third random variable confirm this interpretation. The results for Symmetrical Uncertainty and Pearson coefficient are shown in Tables 2 and 3, respectively. These methods clearly illustrate the conclusions drawn from the ensemble rank: both feature are strongly correlated to each other, while being slightly less correlated to the reference random variable. Therefore, the high degree of relevance and high degree of redundancy indicated by this level of f-correlation lead to the well-known conclusion that both features are predictive of the outcome, while only one may be necessary.

Table 2. f-correlation values calculated with Symmetrical Uncertainty for the Iris dataset

Feature Index	3	4	5
3	0.00	5.47	4.59
4	5.47	0.00	2.55
5	4.59	2.55	0.00

Table 3. f-correlation values calculated with Pearson coefficient for the Iris dataset

Feature Index	3	4	5
3	0.00	0.95	0.04
4	0.95	0.00	0.07
5	0.04	0.07	0.00

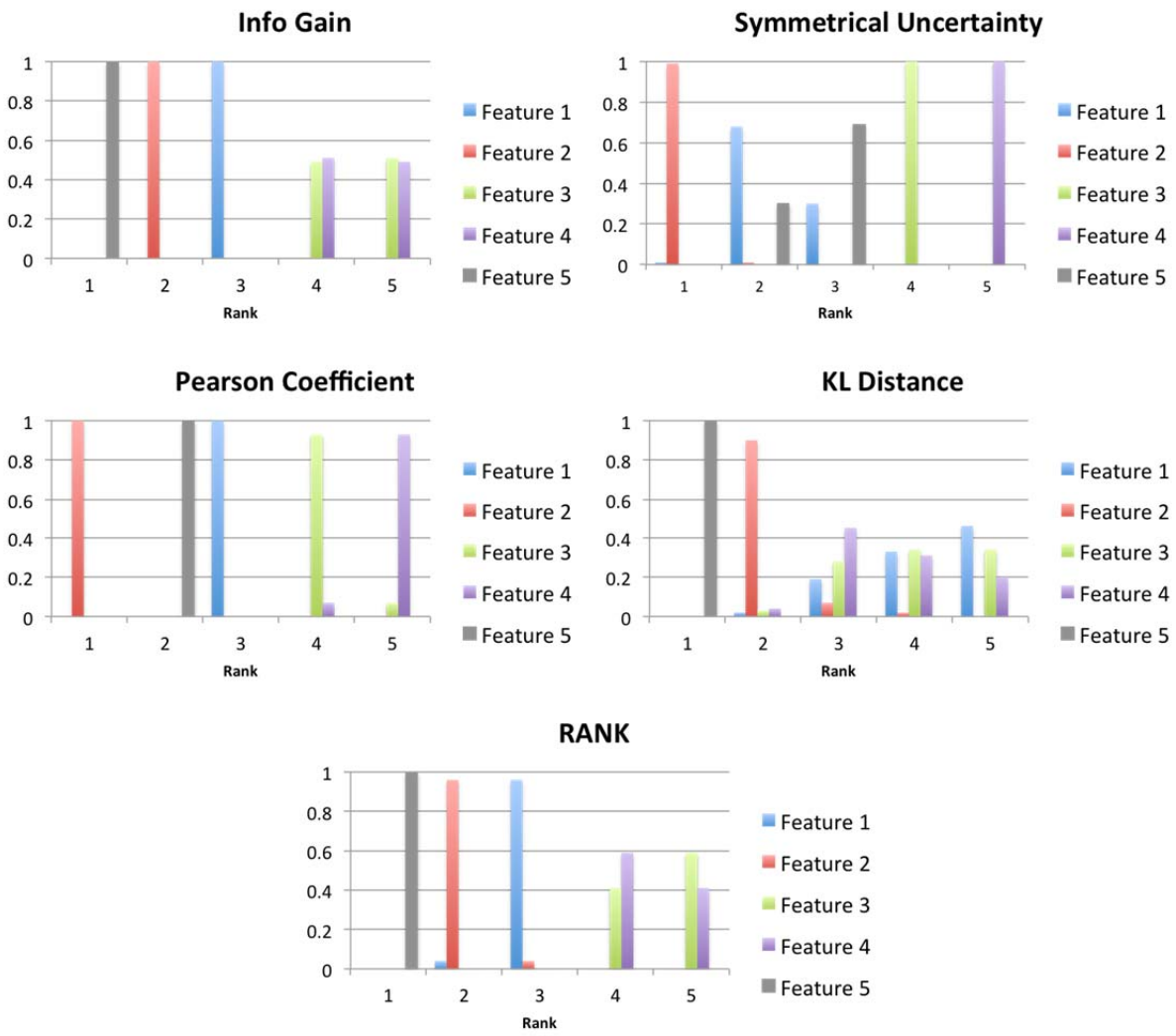


Figure 1. Individual bootstrapping feature ranks for IRIS dataset

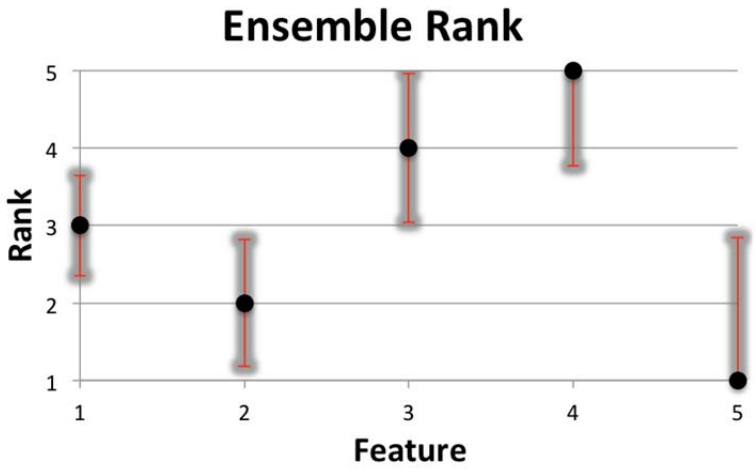


Figure 2. Ensemble feature ranks for IRIS dataset with standard error bars

The CORRAL dataset is examined using the same ensemble rank algorithm and rank distribution methodology. As expected, none of the methods are able to distinguish between the actual predictive variables (1,2,3,4) and the partially correlated variable, 6, although all methods correctly rank the random (7) and uncorrelated variable (5) lowest (Figure 3). Similarly, the uncorrelated variable 5, is ranked lowest with the random variable. This is an important example that illustrates the failure of individual FS methods to differentiate between features. In this case, all relevant features (1-4) and feature 6, are ranked approximately evenly, above the random variable, yielding limited insight into the available quantities. In this case, it is necessary to incorporate the f-correlation quantities in the analysis. Tables 4 and 5 show the f-correlations calculated with symmetrical uncertainty and Pearson coefficient, respectively. The first characteristic observed in these calculations is the identical structure of the f-correlation array between the two methods. Both methods show that the relevant features (1-4) are completely uncorrelated, as well as feature 5, which is purposely uncorrelated to the class. The random feature has a low, uniform degree of correlation to all other features while the purposely highly class-correlated feature is also strongly correlated to the other features. Combined with the ensemble rank results (Figure 4), the f-correlation measure provides a definite evaluation of feature importance. Here, features 1-4 are both highly correlated to the class while uncorrelated to other features. These are the conditions for high relevance and low redundancy, used as indications of importance. Contrastingly, features 5 and 7 have a low degree of relevance, while feature 6 has a high degree of relevance and high degree of redundancy.

Table 4. f-correlation values calculated with Symmetrical Uncertainty for the CORRAL dataset

Feature Index	1	2	3	4	5	6	7
1	0.00	0.00	0.00	0.00	0.00	**18.17**	**0.29**
2	0.00	0.00	0.00	0.00	0.00	**18.17**	**0.29**
3	0.00	0.00	0.00	0.00	0.00	**18.17**	**0.29**
4	0.00	0.00	0.00	0.00	0.00	**18.17**	**0.29**
5	0.00	0.00	0.00	0.00	0.00	0.00	**0.29**
6	**18.17**	**0.18**	**18.17**	**18.17**	0.00	0.00	**0.28**
7	**0.29**	**0.09**	**0.29**	**0.29**	**0.29**	**0.28**	0.00

Table 5. f-correlation values calculated with Pearson coefficient for the CORRAL dataset.

Feature Index	1	2	3	4	5	6	7
1	0.00	0.00	0.00	0.00	0.00	**0.18**	**0.10**
2	0.00	0.00	0.00	0.00	0.00	**0.18**	**0.09**
3	0.00	0.00	0.00	0.00	0.00	**0.18**	**0.13**
4	0.00	0.00	0.00	0.00	0.00	**0.18**	**0.10**
5	0.00	0.00	0.00	0.00	0.00	0.00	**0.10**
6	**0.18**	**0.18**	**0.18**	**0.18**	0.00	0.00	0.00
7	**0.10**	**0.09**	**0.13**	**0.10**	**0.10**	0.00	0.00

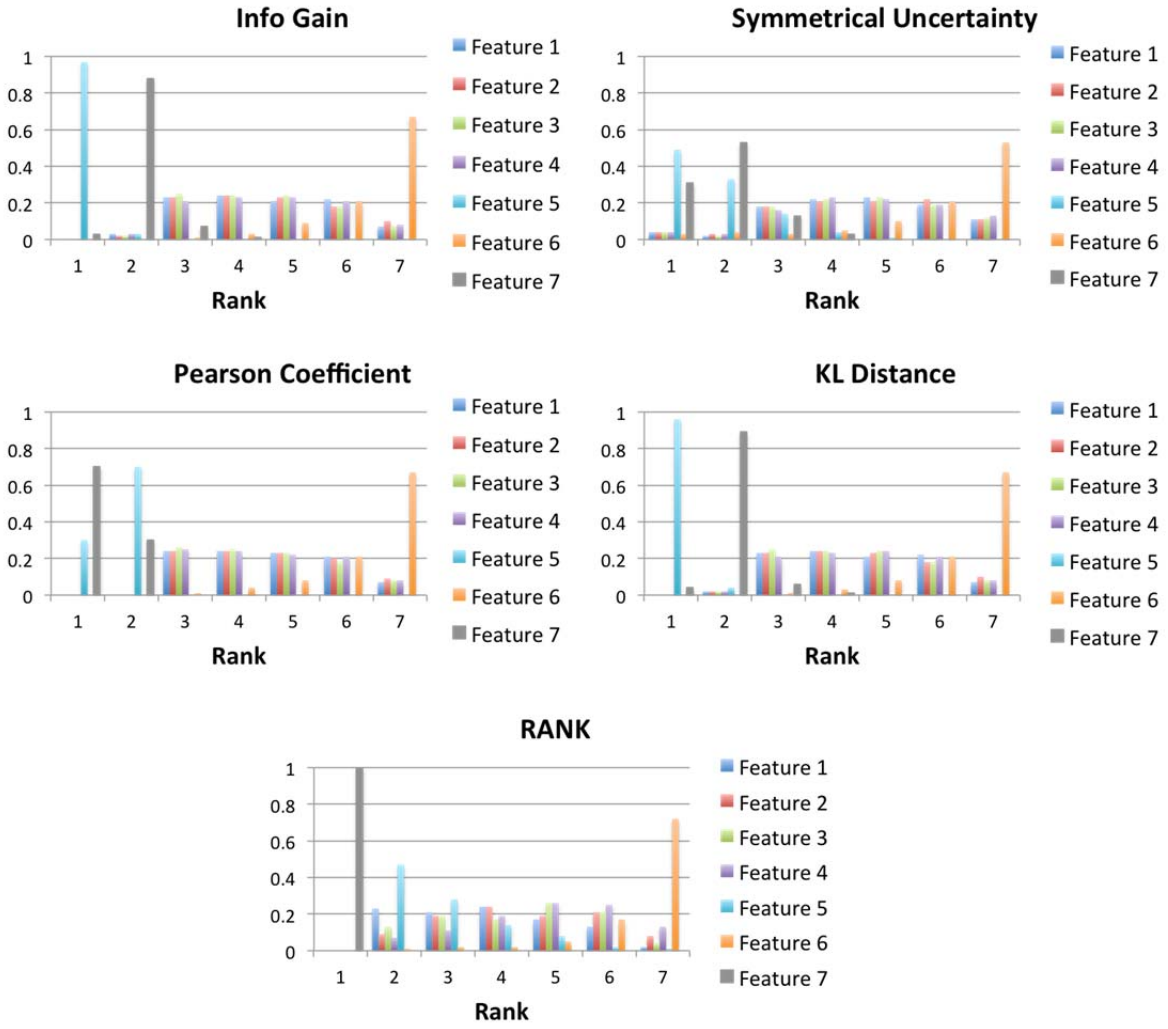


Figure 3. Individual bootstrapping ranks for CORRAL dataset

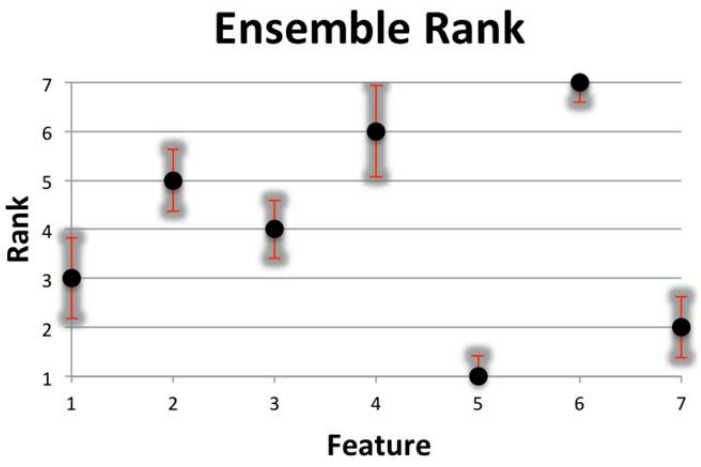


Figure 4. Ensemble feature ranks for CORRAL dataset with standard error bars

Ensemble rank results for the two synthetic datasets, Cosinus-Hills and Swiss (Figures 6, 8), show a correct, high ranking of the first two features and a minimum rank given to the 3rd, random feature. Furthermore, the first two features exhibit overlapping error bounds, indicating, as expected, that they are both given similar feature weights in the individual methods.

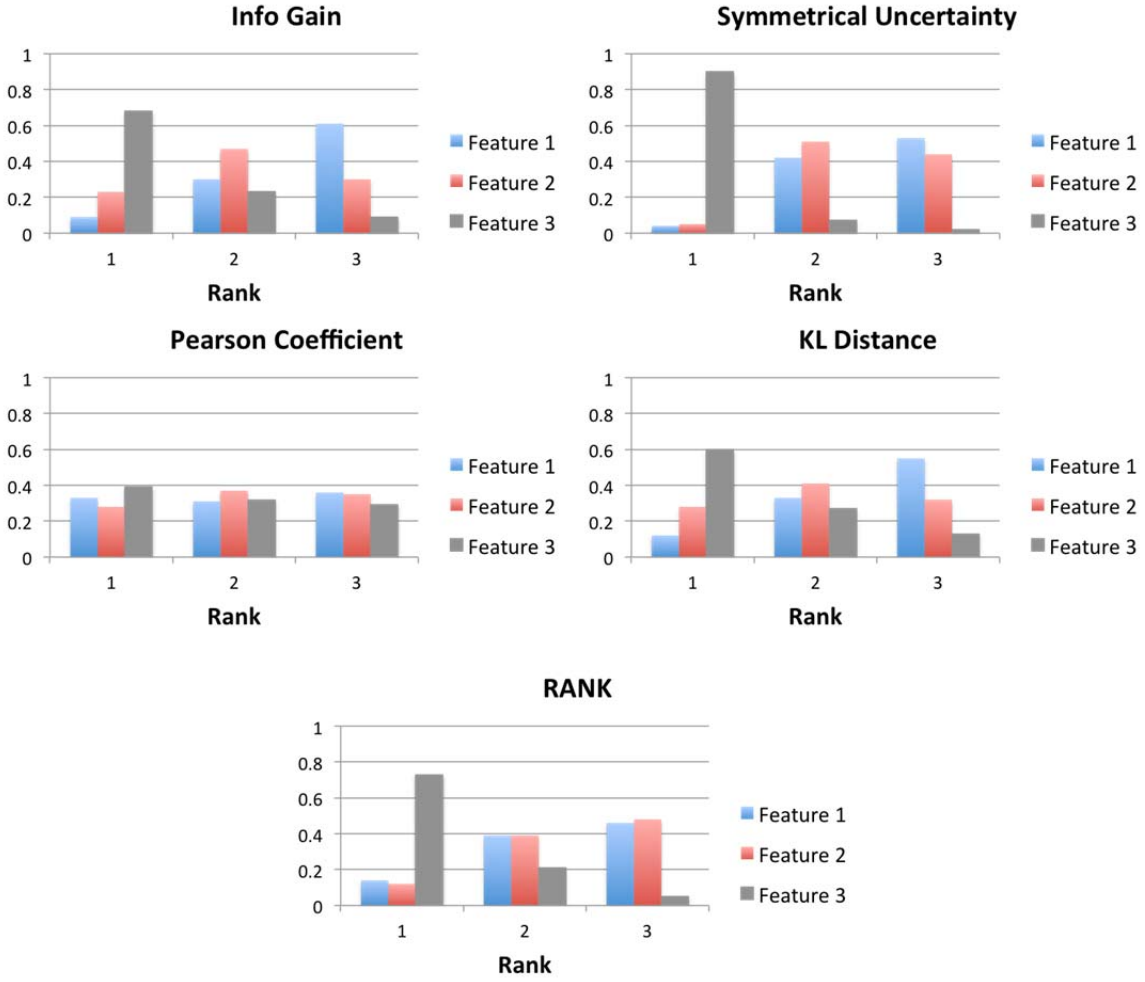


Figure 5. Individual bootstrapping feature ranks for Cosinum-Hills dataset

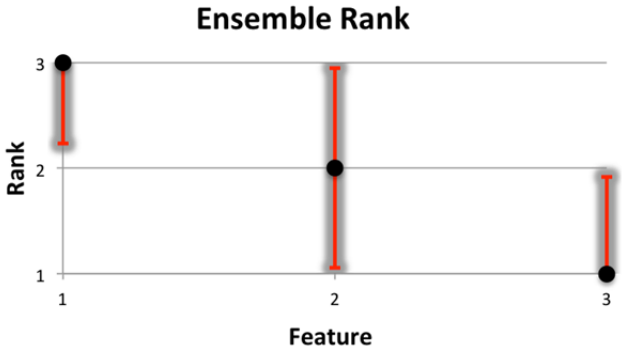


Figure 6. Ensemble feature ranks for Cosinus-Hills dataset with standard error bars

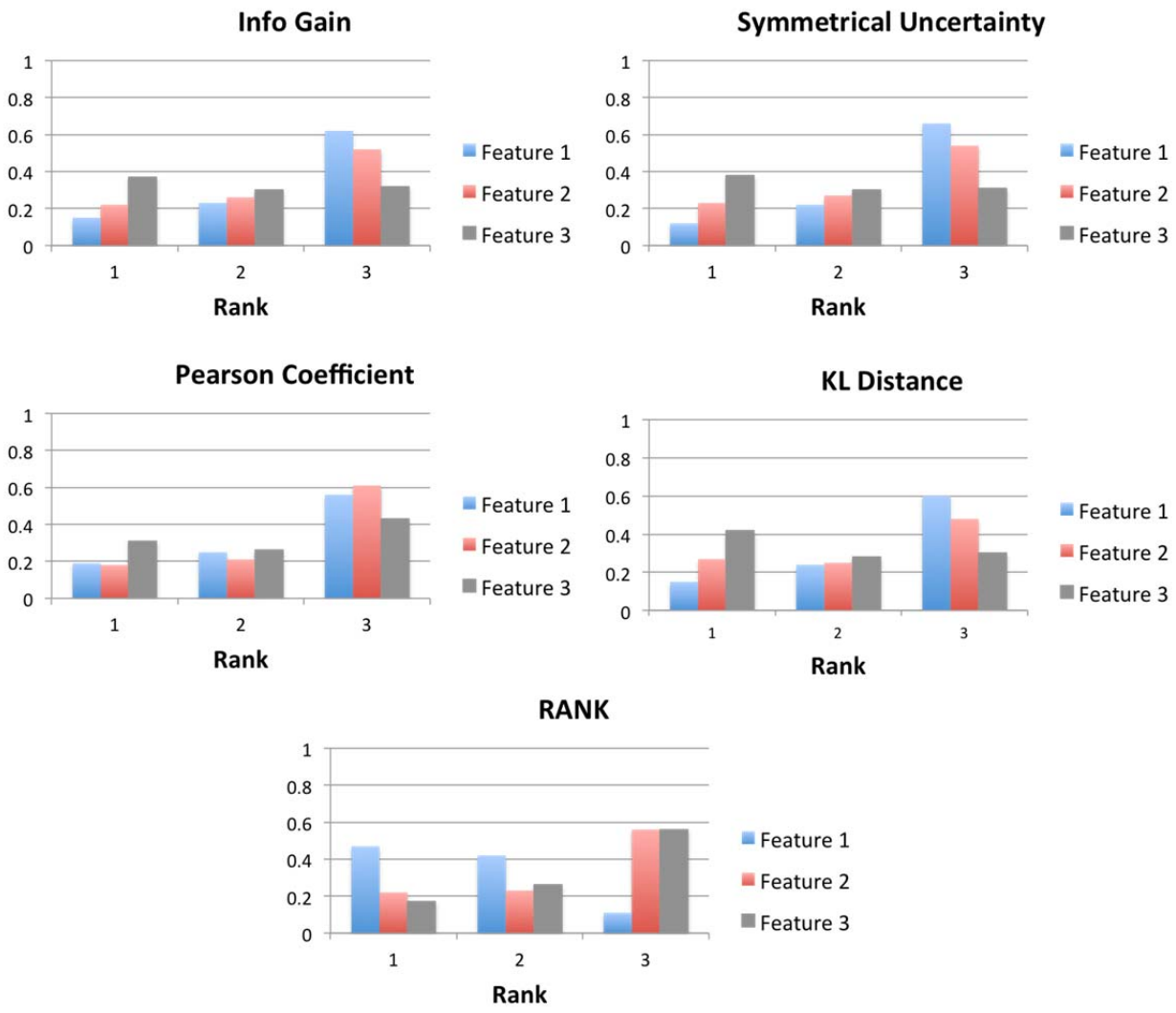


Figure 7. Individual bootstrapping feature ranks for Swiss dataset

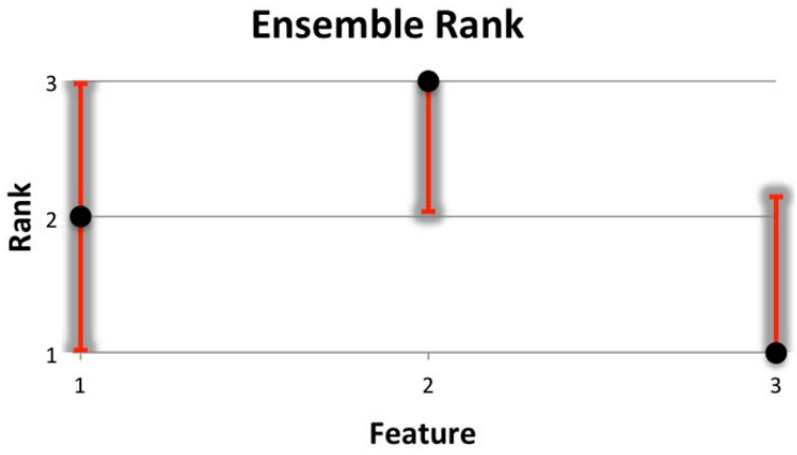


Figure 8. Ensemble feature ranks for Swiss dataset with standard error bars

Finally, the Difficult dataset ensemble rank results are shown below in Figure 10. The results from filter methods such as those used here are not expected to perform well on this dataset since other dimensionality reduction techniques are required when transformation of the original features is necessary. However, this dataset is potentially related to a natural dataset in which all quantities are transformed by some experimental measurement network and none of the intrinsic variables are included. Here, as well as in most datasets with many complex features, ensemble rank is an extremely efficient, advantageous procedure for gaining a conceptual understanding of the data. Due to the higher dimensionality of this dataset, relative to the others studied here, as well as the lack of any strong correlations to the outcome, no clear trends are evident from only examining the results for individual methods in Figure 9. The various methods exhibit a limited amount of structure in their own rank distributions, with very few being distinguishable from the rest. Pearson coefficient and KL distance are the least descriptive, weighting all features almost identically. Information gain is similarly non-descriptive, besides a high ranking for feature 3. However, Symmetrical Uncertainty results show relatively peaked rank distributions and RANK results are even more so. This situation is characterized by multiple FS methods providing little insight into the intrinsic dimensionality of the data. The ensemble rank results confirm this statement with the random feature ranked equivalently with the highest ranked feature of the original dataset. This is an important result, nonetheless, because it is an indication that none of the FS methods should be trusted to describe the importance of the features in this dataset and instead, other methods should be considered. This is not unlike the motivating observations for the No Free Lunch Theorems and the exact situation for which the ensemble rank method is necessary. In this case, the ensemble ranks reflect the structure in the RANK and Symmetrical Uncertainty rank distributions, while also describing the level of uncertainty arising from the inability of other methods to sufficiently discriminate between the same features. Furthermore, inclusion of the multiple non-descriptive methods is not completely detrimental to the ensemble averages since they each weight all features almost identically. Instead, these methods help to identify the complex correlation that is inherent to the dataset. However, it is also apparent that the features included in the ensemble rank must exhibit very different bias toward conditions of the data. If many features are included, all yielding very similar results, the ensemble rank will be incorrectly weighted by these methods and the average, as well as the uncertainty, will be skewed.

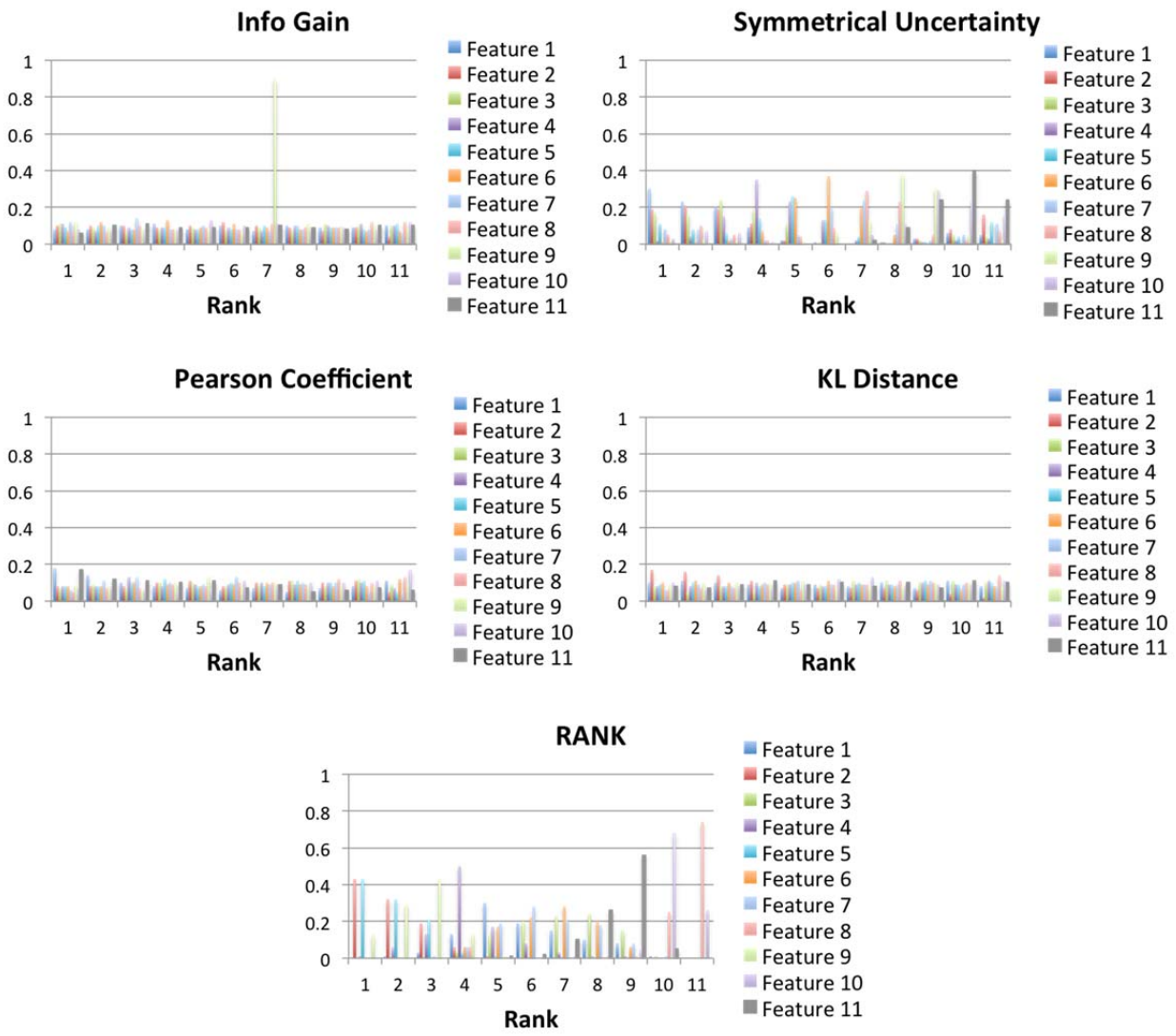


Figure 9. Individual bootstrapping feature ranks for Difficult dataset

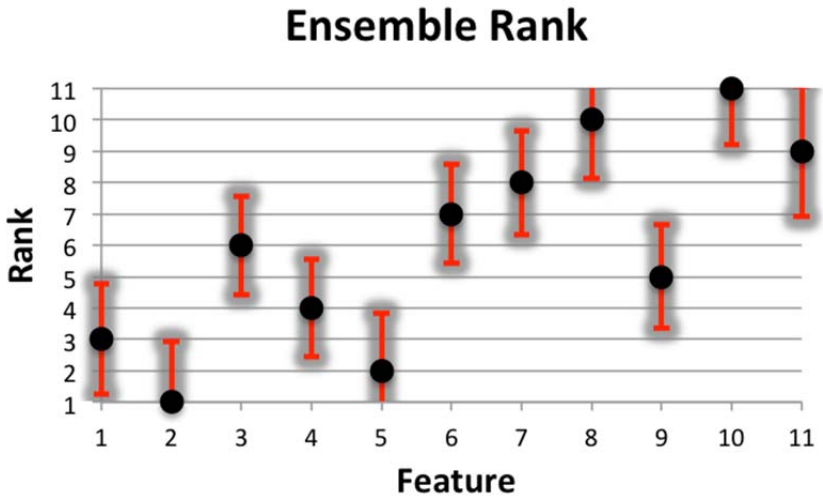


Figure 10. Ensemble feature ranks for Difficult dataset with standard error bars

Conclusions

The choice of feature rank algorithms remains an unresolved issue and the only sense of validation for choosing one method over another, given similar computational complexity, is a deep understanding of how each algorithm describes the quantity of interest. Earlier, it was mentioned that prediction and feature importance are two relevant objectives with slightly different interpretations in terms of feature ranking. However, this statement must be appended to include the possibility of coincidence between features that are considered important (using the principles of maximum relevance and minimum redundancy) as well as yielding the maximum predictive capability. This is certainly not the most common situation since the feature rank corresponding to greatest classification accuracy is not necessarily unique. Nonetheless, the method of ensemble ranking is a general approach which, given the appropriate choice of FS methods, can yield a significant amount of information about the structure of the data as well as correlation-related quantities. It has yet to be determined if the same feature that are highly ranked with this algorithm are coincidently those that achieve the highest classification accuracy, but such a finding would also provide a solution to the seemingly arbitrary selection of features, in the sense of physical interpretability.

As evinced in the results above, FS methods with fundamentally different sources of bias describe feature importance drastically differently. Such difference should not be avoided, but actually have relevance to the physical interpretation of the feature variables. Furthermore, combined with the ensemble approach, feature ranks can be calculated robustly for an arbitrary dataset, while yielding a quantifiable level of certainty in the overall average. These are the necessary prerequisites for feature selection in natural scientific datasets.

Acknowledgements

The work of Aaron Sisto was supported by the DOE Computational Science Graduate Fellowship.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] E. Cantu-Paz. S. Newsam and C. Kamath. Feature selection in scientific applications. *International Conference on Knowledge Discovery and Data Mining.* 2004.

[2]Y. Saeys, I. Inza and P. Larranaga, A review of feature selection techniques in bioinformatics. *Bioinformatics*. pp. 2507-2517. 2007.

[3] E. P. Xing, M. I. Jordan and R. M. Karp. Feature selection for high-dimensional genomic microarray data. *Proceedings of the 18th International Conference on Machine Learning.* pp. 601-608. 2001.

4] Y. Yang and J. O. Pedersen, A comparative study on feature selection in text categorization. *Proceedings of the 14th International Conference on Machine Learning*. pp. 412-420. 1997.

[5] L. J. P. van der Maaten, E. O. Postma and H. J. van den Herik. Dimensionality reduction: A comparative review. Journal of Machine Learning Research. 2008.

[6] L. O. Jimenez and D. A. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical and asymptotic properties of multivariate data. *IEEE Transactions on Systems, Man and Cybernetics.* pp. 39-54. 1997.

[7] C. Glymour and G. F. Couper. *Computation, causation, and discovery.* AAAI Press/ The MIT Press, Menlo Park, California, Cambridge Massachusetts, London, England. 1999.

[8] J. Pearl. Causality: models, reasoning and inference. Cambridge University Press. 2000.

[9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Machine Learning Research*. pp. 1157-1182. 2003.

[10] C. J. C. Burges. *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers.* Kluwer Academic Publishers. 2005.

[11] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha and D. D. Lee. Spectral methods for dimensionality reduction. In *Semisupervised Learning.* Cambridge, MA, USA. 2006.

[12] G. John, R. Kohavi and K. Pfleger. Irrelevant features and the subset selection problem. *Proc. ML-94*. pp. 121-129. 1994.

[13] M. Hall. Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the 17th International conference on machine learning.* pp. 359-366. 2000.

[14] J. Ye and Q. Li, Feature Reduction via Generalized Uncorrelated Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering.* pp. 1312-1322. 2006.

[15] M. Dash and H Liu. Consistency-based search in feature selection. *Artificial Intelligence.* pp. 155-176. 2003.

[16] H. Liu, H. Lu and L. Yu. Active sampling: An effective approach to feature selection. *Proceedings of the 3rd SIAM International Conference on Data Mining.* pp. 244-248. 2003.

[17] R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence.* pp. 273-324. 1997.

[18] P. Langley. Selection of relevant features in machine learning. *Proc. of the AIII, Fall symposium on relevance.* AAAI Press. 1994.

[19] S. Das. Filters, wrappers and a boosting-based bybrid for feature selection. *Proceedings of the 18th International Conference on Machine Learning. pp. 74-81. 2001.*

[20] Y. Yang, Y. Xiao and M. R. Segal. Identifying differentially expressed genes from microarray experiments via statistical synthesis. *Bioinformatics.* pp. 1084-1093. 2005.

[21] K. Y. Yeung, R. E. Bumgarner and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics.* pp. 2394-2402. 2005.

[22] V. G. Tusher, R. Tibshirani and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*. pp. 5116-5121. 2001.

[23] J. Bi, K. Bennett, M. Embrechts, C. Breneman and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research.* pp. 1229-1243. 2003.

[24] J. Yu and X. Chen. Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data. *Bioinformatics*. pp. 487–494. 2005.

[25] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on evolutionary computation.* pp. 67-82. 1997.

[26] R. Teramoto and H. Fukunishi. Consensus scoring with feature selection for structure-based virtual screening. *J. Chem. Info. Model.* pp. 288-295. 2008.

[27] M. Dash and H. Liu. Feature selection for clustering. *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. pp. 110-121. 2000.

[28] N. A. Chuzhanova, A. J. Jones and S. Margetts. Feature selection for genetic sequence classification. *Bioinformatics*. pp. 139-143. 1998.

[29] Z. Zhao, L. Wang, H. Liu and J. Ye. On similarity preserving feature selection. *IEEE Transactions on Knowledge and Data Engineering.* pp. 619-632. 2013.

[30] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proceedings of the 20th International Conference on Machine Learning.* pages 856-863. 2003.

[31] D. Koller and M. Sahami. Toward optimal feature selection. *Proceedings of the 13th International Conference on Machine Learning*. pages 284-292, 1996.

[32] H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. *Proc. AAAI-91*. pp. 547-552. 1991.

[33] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning Journal*. pp. 23-69. 2003.

[34] I. Guyon. C. Aliferis, A. Elisseeff. Causal feature selection. Unpublished manuscript; downloaded from the Web in June2013.

[35] C. Sima, U. Braga-Neto and E. R. Dougherty. Superior feature-set ranking for small samples using bolstered error estimation. *Bioinformatics.* pp. 1046-1054. 2005.