Abstract - The Discriminant Random Forest is a novel extension of the Random Forest classification methodology for learning classification tree ensembles. Our methodology leverages Linear Discriminant Analysis to perform multivariate node splitting during tree construction, producing individual classifiers that are stronger and more diverse than their Random Forest counterparts. Application of this methodology to various two-class signal detection tasks has demonstrated that the Discriminant Random Forest achieves reductions in classification error of up to 79.5% relative to that of the conventional Random Forest. Moreover, empirical tests have suggested that this approach is computationally less costly with respect to both memory and efficiency.

Keywords: Random Forest, Linear Discriminant Analysis

1 Introduction

The Random Forest (RF) is a nonparametric ensemble classification methodology whose class predictions are based upon the aggregation of multiple decision tree classifiers. In this paper, we introduce the Discriminant Random Forest (DRF), a novel classifier that extends the conventional RF via a multivariate node splitting technique based upon a linear discriminant function.

Application of the DRF to various two-class signal detection tasks has demonstrated that this approach achieves reductions in classification error of up to 79.5% relative to the RF. Empirical tests suggest that this performance improvement can be largely attributed to the enhanced strength and diversity of its base tree classifiers, which, as demonstrated in [1], lead to lower bounds on the generalization error of ensemble classifiers. Moreover, experiments suggest that the DRF is computationally less costly with respect to both memory and efficiency.

This paper is organized as follows: Section 2 summarizes the motivation and theory behind the Random Forest methodology. We present the Discriminant Random Forest approach in detail in Section 3, and discuss the performance of both the RF and DRF models for two signal detection applications in Section 4. Our conclusions are summarized in Section 5.
approach its optimal performance. In fact, large values of $m$, though they may increase the strength of the individual classification trees, induce higher correlation among them, potentially reducing the overall effectiveness of the forest.

Each tree is grown without pruning until the data at its leaf nodes are homogeneous, or until some other predefined stopping criterion is satisfied. Class predictions are then performed by propagating a test sample through each tree and assigning a class label, or vote, based upon the leaf node that receives the sample. Typically, the sample is assigned to the class receiving the majority vote. Note, however, that the resulting votes can be viewed as approximately i.i.d. random variables, and thus, the Laws of Large Numbers imply that their empirical frequency will approach their true frequency as the number of trees increases. Moreover, the empirical distribution function from which they are drawn will converge to the true underlying distribution function [8]. Ultimately, we can treat the resulting vote frequencies as class-specific probabilities and threshold upon this distribution to make a classification decision.

3 Discriminant Random Forests

3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), pioneered by R.A. Fisher in 1936, is a discrimination technique that utilizes dimensionality reduction to classify items into distinct groups [9]-[11]. The LDA is an intuitively appealing methodology that makes class assignments by determining the linear transformation of the data in feature space that maximizes the ratio of their between-class variance to their within-class variance, achieving the greatest class separation, as illustrated in Fig. 1. The result is a linear decision boundary, identical to that determined by maximum likelihood discrimination, which is optimal (in a Bayesian sense) when the underlying assumptions of multivariate normality and equal covariance matrices are satisfied [12]. It can be shown that, in the two-class case, the maximum class separation occurs when the vector of coefficients, $w$, used to define the linear transformation is as follows

$$w = \Sigma^{-1}(\mu_i - \mu_j),$$  \hspace{1cm} (3)

where $\Sigma^{-1}$ is the common covariance matrix, and $\mu_j$ is the mean vector for class $k$. Typically, when data are limited, we estimate $\Sigma^{-1}$ with the pooled covariance estimate, $S_w$, given by

$$S_w = \sum_{k=1}^{N} S_k, \hspace{1cm} (4)$$

$$S_k = \sum_{i=1}^{N_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^T. \hspace{1cm} (5)$$

In the above equations, $x_{ki}$ and $\bar{x}_k$ denote the $i$th training sample of class $k$ and the corresponding class mean, respectively.

3.2 The Discriminant Random Forest Methodology

Numerous variations of the Random Forest methodology have been proposed and documented in the literature, most of which address node-splitting techniques [13], [14]. Many of these are based upon an assessment of node impurity (i.e., heterogeneity) and include entropy-based methods, minimization of the Gini impurity index, or minimization of misclassification errors. Additional forest-based methods that focus upon alternative aspects of the algorithm include supplementing small feature spaces with linear combinations of available features [1], variations on early stopping criteria, selecting the split at random from the $n$ best splits [14], and PCA transformation of random feature subsets [15].

Our Discriminant Random Forest (DRF) is a novel approach to the construction of a classification tree ensemble in which LDA is employed to split the feature data. Bagging and random feature selection are preserved in this approach, but unlike other forest algorithms, we apply LDA to the data at each node to determine an “optimal” linear decision boundary. By doing so, we allow decision hyperplanes of any orientation in multidimensional feature space to separate the data, in contrast to the conventional forest algorithm, whose boundaries are limited to hyperplanes orthogonal to the axis corresponding to the feature yielding the best split. We have illustrated this effect in Fig. 2, which depicts decision lines in two-dimensional space for the RF (left) and the DRF (right).

These two approaches to node splitting give rise to highly distinctive decision regions. Fig. 3 gives an example of a two-dimensional decision region created by each forest, in which bright blue and bright gold areas represent regions of high posterior probability for the positive and negative classes, respectively. Darker areas indicate regions of greater uncertainty. The decision region produced by the DRF is notably more complex, and its boundaries are fluid and highly intricate, fitting more closely to the training data.

![Fig. 1. LDA transformation and the optimal linear decision boundary.](image)
4 Empirical Results

In the following suite of experiments, we compare the classification performance of the RF (utilizing the misclassification minimization node-splitting criterion) and DRF for two real-world signal detection applications: (1) detecting hidden signals of varying strength in the presence of background noise, and (2) detecting sources of radiation. Both tasks represent two-class problems; thus, we can evaluate the performance of the RF and DRF in terms of false positives, also known as false alarms or type I errors, and false negatives, also known as misses or type II errors. The false positive rate (FPR), false negative rate (FNR), and true positive rate (TPR) are defined as follows:

\[
FPR = \frac{\# \text{ negative samples misclassified}}{\# \text{ negative samples}}
\]

\[
FNR = \frac{\# \text{ positive samples misclassified}}{\# \text{ positive samples}}
\]

\[
TPR = 1 - FNR
\]

4.1 Hidden Signal Detection

The goal in the hidden signal detection application is to detect the presence of an embedded signal. In this application, it is assumed that each detection event requires a considerable amount of costly analysis, making false positives highly undesirable. Hence, we have computed the Area Under the Receiver Operating Characteristic (ROC) Curve, or AUC, integrated over the FPR interval [0, 0.001] and scaled so that a value of 100% represents a perfect detection rate over this low FPR interval. The resulting quantity provides us with a single value that can be used to compare the prediction performance of the classifiers.

The data for this problem are composed of two separate sets. The training data set, \( T_1 \), consists of 7931 negative class samples (i.e., no embedded signal) along with two sets of positive class samples having 40% and 100% embedded signal strength (7598 and 7869 samples, respectively). The \( J_2 \) data set contains 9978 negative class samples and five positive classes having signal strengths of 20%, 40%, 60%, 80% and 100% (7760, 9143, 9327, 9387, and 9425 samples, respectively). The training and testing data sets for each of the following experiments consist of the negative class combined with one of the available positive classes, as indicated in each case. All data samples consist of eight features useful for detecting the presence of embedded signals. We have applied both the RF and DRF forest methodologies at each split dimension \( m \in \{1, 2, \ldots, 8\} \) in an effort to assess the impact of this parameter on their performance.

4.1.1 Training on \( T_1 \), Testing on \( J_2 \)

Fig. 4 shows the plots of the AUC generated by training the forests on \( T_1 \) and testing on \( J_2 \) at signal strengths of 40% and 100%. Each RF or DRF was composed of 500 trees, a sufficient forest size to ensure convergence in AUC. In 14 of the 16 possible combinations of signal strength and split dimension, the DRF performance clearly exceeded that of
the RF. In the remaining two cases, the difference in the detection rate was negligible. Moreover, these results suggest that the DRF is more successful than the RF algorithm in detecting weaker signals and better utilizes more input features. As the split dimension increases, we would expect the trees to become more correlated for both methodologies, resulting in poorer prediction performance. Fig. 4 suggests this trend, but the effect appears to be noticeably less severe for the DRF. The tradeoff between tree strength and correlation with respect to $m$ is discussed in greater detail in Section 4.2. ROC curves for both RF and DRF for the hidden signal detection application are plotted in Fig. 5, indicating that the DRF exhibits superior performance across the low FPR region of interest.

4.1.2 Prediction Performance for J2 with Cross Validation

To more thoroughly explore the impact of signal strength on prediction performance, we trained and tested the RF and DRF on all signal strengths of the J2 data set. We used 5-fold cross-validation (CV) to evaluate the performance of both classifiers. In k-fold cross-validation, the data set is randomly partitioned into $k$ equal-sized and equally-proportioned subsets. For each run $i$, we set aside data subset $i$ for testing, and we train the classifier on the remaining $k-1$ subsets. We use the average of these $k$ estimates to compute our performance estimate. Fig. 6 shows the percentage increase in the AUC achieved by the DRF for split dimensionalities $m \in \{1,2\}$ as compared to the best-performing RF (i.e., $m=1$). As we observed when training on $T1$, the DRF yields substantially better detection performance on weaker signals.

4.2 Radiation Detection

The objective of the Radiation Detection effort is to detect the presence of a radiation source in vehicles traveling through a radiation portal monitoring system that measures the gamma ray spectrum of each vehicle quantized into 128 energy bins. The 128-dimensional normalized gamma ray spectra serve as the input features for the RF and DRF. The negative class is composed of radiation measurements from real vehicles containing no radiation source. The data for the positive class were created by injecting a separate set of negative samples with spectra derived from two isotopic compositions of both Uranium and Plutonium in IAEA Category 1 quantities [16]. These sets of positive and negative samples were then partitioned into non-overlapping, equally-proportioned training and testing sets containing 17,000 and 75,000 samples, respectively.

The ROC curves in Fig. 7 show the TPR (i.e., detection rate) versus the FPR (i.e., false alarm rate) for RF and DRF.

---

**Fig. 4.** AUC for RF and DRF: (top) trained on T1, tested on J2 with 40% signal strength; (bottom) trained on T1, tested on J2 with 100% signal strength. Each classifier is composed of 500 trees.

**Fig. 5.** ROC curves for RF and DRF trained on T1, tested on J2 with 100% signal strength and 170K additional negative samples.

**Fig. 6.** Percentage increase in AUC on J2 for DRF $m=1,2$ over RF $m=1$ at each signal strength. Each forest is composed of 500 trees.
Over most of the low FPR range, the DRF maintains higher detection rates than the RF.

The total number of features available for this application presents an ideal opportunity to thoroughly explore the behavior of the RF and DRF methodologies for high split dimensions. In particular, we wish to investigate their relative computational efficiency, memory considerations, bounds on their generalization error (1), the interplay between tree strength and correlation, and the impact of each of these characteristics on overall algorithm performance. We have utilized the OOB data to compute these characteristics (see [1] for further details) at the minimal forest size required for each algorithm to achieve its peak performance. This can be readily observed in Fig. 8, which shows a plot of the minimum classification error (MCE) achieved by both classifiers with respect to the number of trees in the forest, and indicates that the peak DRF performance was achieved by a forest consisting of approximately 50 trees, far fewer than the 250 trees required by the RF.

In Table 1, performance statistics have been provided for the RF yielding the best MCE and for a DRF whose performance exceeded that of the RF with respect to error, computational requirements and efficiency. Both were trained on a dual-core Intel 6600 2.4 GHz processor with 4GB of RAM. To compute memory usage, we assumed that each RF node must store an integer feature ID and its corresponding floating-point threshold. Each DRF node must store \( m \) integers for its selected feature IDs along with \( m+1 \) floating-point values for its weight vector, \( w \).

Table 1 indicates that the DRF was able to achieve a lower classification error rate than the RF while simultaneously reducing training time and memory usage by 84.8% and 26.5%, respectively. The smaller DRF trees clearly contribute to this improvement in efficiency, but their reduced size also suggests a dramatic increase in tree strength.

From an empirical standpoint, a node splitting strategy that effects a better separation of the data, such as the multivariate LDA technique, naturally generates smaller classification trees as the split dimensionality increases, as
shown in Fig. 9. Though such trees might exhibit superior prediction capabilities (i.e., greater strength), we would generally expect the variation among them to decrease (i.e., increased correlation), potentially leading to a reduction in overall performance.

The key to informative analysis of these two classification methodologies, as introduced in Section 2, lies in our ability to successfully characterize this interplay between the strength and correlation of individual trees. To provide further insight into these behaviors, Fig. 10 compares the OOB estimates of tree strength and correlation for the RF and DRF, along with their classification error and respective generalization error bounds plotted as a function of the split dimensionality, $m$. As expected, the strength of an individual DRF tree is, in general, significantly greater than that of its RF counterpart. Far more remarkable is the reduced correlation among the DRF trees for split dimensions up to $m \approx 90$. The relationship between strength and correlation is typically regarded as a tradeoff [1], in which one is improved at the expense of the other, and the smaller DRF trees might naively be expected to exhibit greater correlation. However, [1] suggests that each base classifier is primarily influenced by the parameter vector representing the series of random feature selections at each node. In the multivariate setting, the number of potential feature subsets at each node increases combinatorially, dramatically enhancing the variability in the parameter vector that characterizes the classifier, which may explain the immediate drop in correlation as $m$ increases. As $m$ approaches the cardinality of the feature set, however, we observe a sudden and severe rise in correlation, behavior that is consistent with the reduced variation in the nodal features used for splitting.

Consistent with the generalization error bound (1), Fig. 10 shows that strength has a greater impact on the classifier performance than the correlation. Even at the highest split dimensionality, the classification error and error bound exhibit only minimal degradation. In contrast, the RF error steadily increases with the split dimensionality. Moreover, the DRF bound is far tighter than that of the RF, even surpassing the RF classification error at $m = 25$.

Interestingly, though the strength and correlation of both methodologies exhibit similar trends, their classification errors exhibit opposing behavior, suggesting that the relationship between strength and correlation is more complex than can be fully explained by our initial experiments.

5 Conclusions

The empirical evidence presented in the previous sections strongly suggests that the Discriminant Random Forest generally produces lower class prediction errors than the Random Forest. In nearly every experiment performed, the DRF achieved lower false positive rates while maintaining higher rates of detection. These improvements are further supported by the superior strength and diversity of the trees produced by the DRF. This methodology also appears to be more successful in the detection of weak signals and may be particularly useful for applications in which low signal-to-noise ratios are typically encountered.

We have found that our methodology achieves far lower prediction errors than the RF when high-dimensional feature vectors are used at each tree node. In general, we expect the performance of any forest to decline as the number of features selected at each node approaches the cardinality of the entire feature set. Under these conditions, the individual base classifiers that compose the forest are nearly identical, negating many of the benefits that arise from an ensemble-based approach. In such cases, the only variation remaining in the forest is due to the bagging of the input data. Though we observed the expected performance degradation for both forest methodologies at extremely high split dimensions, the effect was far less severe for the discriminant-based approach. This result suggests a versatility and robustness in the DRF methodology that may prove valuable for some application domains.

Overall, empirical evidence favors the DRF as a more effective classification methodology than the RF. Although computational efficiency may be adversely impacted by the more complex node-splitting of the DRF at extremely high dimensions, its peak performance is typically achieved at much lower dimensions where it is more efficient than the RF with respect to memory and runtime.

The behavior of the Discriminant Random Forest
methodology is compelling and hints at a complex internal mechanism that invites further investigation. However, we have found significant evidence supporting this technique as a highly robust and successful classification approach across diverse application domains.

6 Acknowledgment

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

7 References