Ensembles of diverse classifiers often outperform single classifiers as has been well-demonstrated across several applications. Existing training algorithms either learn a classifier ensemble on pre-defined feature sets or independently perform classifier training and feature selection. Neither of these schemes is optimal. We pose feature subset selection and training of diverse classifiers on selected subsets as a joint optimization problem. We propose a novel greedy algorithm to solve this problem. We sequentially learn an ensemble of classifiers where each subsequent classifier is encouraged to learn data instances misclassified by previous classifiers on a concurrently selected feature set. Our experiments on synthetic and real-world data sets show the effectiveness of our algorithm. We observe that ensembles trained by our algorithm performs better than both a single classifier and an ensemble of classifiers learnt on pre-defined feature sets. We also test our algorithm as a feature selector on a synthetic dataset to filter out irrelevant features.

**Index Terms** — Classifier ensemble, diversity, simulated annealing, loss function optimization

1. **INTRODUCTION**

Ensembles of multiple classifiers have outperformed single classifier systems in several applications (e.g. Netflix challenge [1], DARPA GALE [2], Intoxication sub-challenge of Interspeech 2011 [3]). Dietterich [4] noted that ensemble of classifiers can have potentially lower generalization error and can capture rather complex class boundaries using simple classifiers. We propose a novel algorithm to train an ensemble of classifiers, jointly performing class boundary optimization and feature selection for each classifier. Consider an ensemble of K classifiers \( \{ f_1, \ldots, f_K \} \) to model a dataset \( S = \{(x_1, y_1), \ldots, (x_M, y_M)\} \) of size \( M \) from a probability distribution \( P \). The data-samples \( x_m \) lie in a \( D \) dimensional feature space such that \( x_m = \{x_{m1}, \ldots, x_mD\} \). We aim to learn the target labels \( y_m \) using the ensemble, where \( y_m \) is drawn from a set of \( J \) classes \( \{c_1, \ldots, c_J\} \). Ensemble learning optimizes a set of loss functions \( \{L_1, \ldots, L_K\} \) over \( S \), where \( L_k \) corresponds to \( f_k(k = 1, K) \). The learning scheme optimizes the loss functions over a set of parameters \( \{\Theta_1, \Lambda_1\}, \ldots, \{\Theta_K, \Lambda_K\} \), where \( \Theta_k \) and \( \Lambda_k \) determine the class boundaries and features subset used by the \( k \)th classifier, respectively. We represent \( \Lambda_k \) as a \( D \)-dimensional binary indicator vector \( \{\lambda_{k1}, \lambda_{k2}, \ldots, \lambda_{kD}\} \) where \( \lambda_{kd} = 1 \) indicates the use of \( d \)th feature while training the \( k \)th classifier and vice versa. Existing methods enforce diversity in the ensemble by using empirically-chosen diverse loss functions for each classifier or subjecting various constraints over \( \Lambda_k \) and/or \( \Theta_k \) while training \( f_k \). We instead pose ensemble learning as a joint optimization problem over the parameters \( \{\Theta_1, \Lambda_1\}, \ldots, \{\Theta_K, \Lambda_K\} \) in (1).
compare the performance of our algorithm against a single classifier and an ensemble learnt by optimizing only $\Theta_k$ with a preset $\Lambda_k = 1_D$.

Apart from classification experiments, we also test our algorithm as a feature selector. Prominent wrapper feature selection schemes (such as forward feature selection) [14] are usually computationally expensive when as filter schemes (such as feature selection through mutual information) use a selection criterion which is usually different from the loss function. As our algorithm performs coupled feature selection and class boundary optimization, we avoid any extra computations as introduced by shuttling optimization between $\Lambda_k$ and $\Theta_k$ in the filter schemes. Furthermore, as our algorithm optimizes $\Lambda_k$ over $L_k$, the selected features are directly relevant to the loss function. We present an application to our algorithm as a feature selector and present results on a synthetic dataset with dummy features.

2. ENSEMBLE CREATION AND FUSION

Training an ensemble can be broken down into training a diverse set of classifiers followed by fusing their decisions. We present a novel algorithm for training diverse classifiers in section 2.1 and describe the fusion strategies in section 2.2.

2.1. Training diverse classifiers

We first propose an algorithm to sequentially optimize each loss function $L_k$ over $\{A_k, \Theta_k\}$, thereby jointly determining the feature subset and the class boundaries. We focus on designing an ensemble of discriminative probabilistic classifiers of the form shown in (4). Such classification models are popular across several applications [15, 16].

$$f_k(x_m) = f_k(x_m / \Lambda_k, \Theta_k) = \arg \max_{j \in \{1, .., J\}} p_k(c_j / x_m, \Lambda_k, \Theta_k) \tag{4}$$

We estimate the posterior probability of each $c_j$ ($j = 1, .., J$) given $x_m$, based on a soft-max function (5) and assign the class with the maximum probability. The class assignment parameter $\Theta_k$ determines linear class boundaries based on $J$ $D$-dimensional vectors $\Theta_{kj} = \{ \theta_{kj1}, .., \theta_{kdj} \}$, $j = 1, .., J$.

$$p_k(c_j / x_m) = p_k(c_j / x_m, \Lambda_k, \Theta_k) = \frac{\exp(\alpha_{mkj})}{\sum_{j=1}^J \exp(\alpha_{mkj})} \tag{5}$$

where $\alpha_{mkj} = x_m \times \text{diag}(\Lambda_k) \times \theta_{kj}^T$.

We define the loss function $L_k$ as negative of weighted accuracy function $A_k$ defined in (7).

$$A_k(\Lambda_k, \Theta_k, W_k) = \frac{1}{M} \sum_{m=1}^M w_{km} \times \delta(f_k(x_m) - y_m) \tag{7}$$

where $\delta(f_k(x_m) - y_m) = \begin{cases} 1 & \text{if } f_k(x_m) = y_m \\ 0 & \text{otherwise} \end{cases}$

We introduce diversity by assigning different weights to the data sample while learning each classifier. The sample weights $W_k = \{w_{k1}, .., w_{kM}\}$ while training the $k^{th}$ classifier are heuristically set as shown in equation (8). This scheme assigns a higher weight $w_{km}$ to the instance $x_m$ if the probability of the true class $y_m$ in the previously trained classifiers is low. Hence the return is higher if the $k^{th}$ classifier correctly classifies $x_m$. This encourages each subsequent classifier to learn data instances misclassified by previously trained classifiers.

Algorithm 1

<table>
<thead>
<tr>
<th>Line</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\eta$</td>
<td>Gradient ascent scaling constant</td>
</tr>
<tr>
<td>2</td>
<td>$\theta_{thr}$</td>
<td>Threshold for convergence</td>
</tr>
<tr>
<td>3</td>
<td>$\Lambda_k = 1_D$</td>
<td>Initial training using all dimensions</td>
</tr>
<tr>
<td>4</td>
<td>$\Theta_k$</td>
<td>Random initialization for $\Theta_k$</td>
</tr>
<tr>
<td>5</td>
<td>${U_1, .., U_D}$</td>
<td>$D$ uniform random variables; $U_d \in [0, 1]$</td>
</tr>
<tr>
<td>6</td>
<td>${u_1, .., u_D}$</td>
<td>A realization of ${U_1, .., U_D}$</td>
</tr>
<tr>
<td>7</td>
<td>$U_{bin}$</td>
<td>Binary variable where the $d^{th}$ element $U_{bin}(d) = 1$ if $u_d &lt; p_\Lambda$</td>
</tr>
<tr>
<td>8</td>
<td>$\Lambda_k$</td>
<td>An alternate dimension subset where $\Lambda_{kd} = NOT(\lambda_{kd})$ with probability $p_\Lambda$</td>
</tr>
<tr>
<td>9, 10</td>
<td>$\Theta_k^{new}, \Theta_k^{rew}$</td>
<td>$\Theta_k$ after gradient ascent on $\Lambda_m, \Lambda_m$ respectively</td>
</tr>
<tr>
<td>11, 12</td>
<td>$A^{old}, A^{new}, A^{new'}$</td>
<td>Weighted accuracies using: $(\Lambda_k, \Theta_k), (\Lambda_k, \Theta_k^{rew}), (\Lambda_k, \Theta_k^{new'})$ respectively</td>
</tr>
</tbody>
</table>

Table 1. Intermediate variables in Algorithm 1 and their description. We state the optimization procedure in Algorithm 1. We evaluate the binary indicator vector $A_k$ by performing simulated annealing [17] based binary integer programming to maximize $A_k$. As $A_k$ is not differentiable, we evaluate $\Theta_k$ by performing gradient ascent on the weighted data log-likelihood reward function $O_k$ as a proxy for $A_k$ (Note we maximize $A_k, O_k$ as it is negative of loss function). We list the set of intermediate variables and their description in Table 1.
1. The inner loop performs optimization on \(A_k, O_k\) over two candidate feature subsets \(\Lambda_k\) and \(\Lambda_k'\). The candidate feature subset \(\Lambda_k\) is derived from \(\Lambda_k\) where we flip \(\lambda_{k,i}\) with a flipping probability \(p_k\). We retain the dimension subset with a higher \(A_k\) after performing gradient ascent to determine \(\Theta_k\).

2.2. Fusing diverse classifiers

Given an ensemble of classifiers, we define a fusion function over the outputs of the classifiers \(\{f_1(x_m), \ldots, f_K(x_m)\}\) and an importance variable \(I = \{i_1(x_m), \ldots, i_K(x_m)\}\) as shown in (10). We state two methods to determine \(I\) after (10).

\[
g(x_m) = \arg \max_{j \in (1, \ldots, J)} \prod_{k=1}^{K} p_k(c_j/x_m)^{i_k(x_m)} \tag{10}
\]

- **Oracle fusion function** \((g_{orc})\): This function uses a 1-in-K encoding scheme over the importance variable \(I\). Given there exists at least one classifier that correctly classifies \(x_m\), \(g_{orc}\) deterministically sets an \(i_k(x_m)\) to 1 if \(f_k(x_m) = y_m\). In case none of the classifiers output the correct class, this function assigns an incorrect class. Note that this scheme is not practically feasible as we are using the true labels in determining \(I\).

- **Equal weighting scheme** \((g_{EW})\): This is a proxy for above scheme where each \(i_k(x_m)\) is assigned a value \(1/K\). \(g_{EW}\) assumes that each classifier in the ensemble is equally important.

In the next section, we test our ensemble learning algorithm and the fusion schemes on several datasets.

3. EXPERIMENTAL RESULTS AND DISCUSSION

We test our algorithm on a synthetic dataset followed by several real world datasets. We also present an application to our algorithm as a feature selector and test it on the synthetic dataset.

3.1. Synthetic dataset

We design a two-dimensional dataset with \(x_m = (x_{m1}, x_{m2}); (0 < x_{m1} < 1, 0 < x_{m2} < 1)\) and classes \(\{c_1, c_2\}\). As our algorithm trains linear classifiers, we chose the class assignment boundary to be quadratic (11) to mimic a practical scenario where the true class boundary may not be perfectly modeled by the classifier. We conduct two sets of experiments on this synthetic dataset. The first experiment shows that the trained ensemble is a better classifier whereas the second experiment illustrates our algorithm as a feature selector.

\[
y_i = \begin{cases} c_1 & \text{if } (x_{i1})^2 + (x_{i2})^2 < 1 \\ c_2 & \text{otherwise} \end{cases} \tag{11}
\]

3.1.1. As a better classifier

We train the ensemble with an equal number of samples from both the classes, with varying dataset size. A large test set size of 50,000 data-samples per class ensures small sampling bias while calculating the classification accuracy. We choose a single classifier trained on all the features \(K = 1; \Lambda_1 = I_D = I_2\) as our baseline. We train two kinds of ensembles. In the first case, we solely optimize over \(\Theta_k\) and set \(\Lambda_k = I_2\). This is analogous to several previously described ensemble learning schemes which fix \(\Lambda_k\) and optimize over \(\Theta_k\). This is achieved by setting the flipping probability \(p_k = 0\) in the algorithm. In the second case, we optimize over both set of parameters \(\Lambda_k, \Theta_k\) and \(p_k\) is empirically set to 0.01. We compare the accuracies obtained by these ensembles to study the effect of including \(\Lambda_k\) in optimization. \(K\) is tuned on the train set itself for maximum accuracy. \(\eta, \text{thr}\) are empirically set. Figure 1 reports mean relative improvements from models trained on 100 different datasets for each dataset size.

We observe that while an ensemble of classifier beats the baseline classifier every time, the relative improvement in accuracy varies with the dataset size. \(g_{orc}\) performs substantially better than \(g_{EW}\). The ensembles provide a higher relative improvement for a smaller dataset size for both \(g_{orc}, g_{EW}\). This suggests that our algorithm works better particularly in the case of sparse distribution of data-points. We also observe that optimization over both \(\{\Lambda_k, \Theta_k\}\) performs better than optimization on just \(\Theta_k\) with a pre-assigned value to \(\Lambda_k\). This supports our case on optimization over both the parameters \(\{\Lambda_k, \Theta_k\}\).

3.1.2. As a feature selector

We run our algorithm on an extended synthetic dataset obtained after appending 6 dummy features \((x_{m3}, \ldots, x_{m8}); (0 < x_{md} < 1, d = 3..8)\) to \((x_{m1}, x_{m2})\) in the above synthetic dataset. The dummy features play no role in determining the class. We define a symmetric feature co-selection matrix \(F\) on \(\Lambda_k\) obtained after running the algorithm. The \((p, q)^{th}\) entry \(F(p,q)\) in the matrix is defined as shown in (12). \(F(p,q)\) for \(p \neq q\) determines the proportion of times the \(p^{th}\) and the \(q^{th}\) features are used together in the \(K\) classifiers. \(F(p,q)\) for \(p = q\) gives an individual feature’s selection frequency. We binarize \(F\) based on two thresholds \(T_{cross}, T_{indiv}\) to obtain the most frequently selected pair of features and most frequently selected features. This matrix \(F^{bin}\) is shown in (13).

\[
F(p,q) = \begin{cases} F(p,q) & \text{if } p \neq q \\ T_{cross} & \text{if } p = q \end{cases} \tag{12}
\]

\[
F^{bin} = \begin{cases} F(p,q) & \text{if } p \neq q \\ F(p,q) > T_{cross} \text{ and } p \neq q \end{cases} \tag{13}
\]
3.2. Real-world datasets

We test our algorithm on three real world datasets listed in Table 2. We randomly split the data into training (80%), development (10%) and test (10%) sets. We set our baseline as a single classifier trained on all the features. We train two sets of ensembles. In the first ensemble training, we optimize on \( \Theta \) and set \( \Lambda_k = 1_D \) and in the second case we optimize on both the parameters, with \( p_\lambda = .01 \). The development set is used to tune \( K \). We report the mean accuracies on the baseline classifier and the mean relative improvement over the baseline accuracy using the ensembles over 100 iterations of random splitting in Table 2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( D, J, M )</th>
<th>Baseline accuracy</th>
<th>Relative Improvement</th>
<th>( g_{orc} )</th>
<th>( g_{EV} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone [18]</td>
<td>8.3, 4177</td>
<td>54.2</td>
<td>36.5</td>
<td>35.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Image Segmentation [18]</td>
<td>19.7, 2310</td>
<td>91.3</td>
<td>3.9</td>
<td>5.4</td>
<td>1.0</td>
</tr>
<tr>
<td>IEMOCAP [19]</td>
<td>200.4, 5498</td>
<td>54.9</td>
<td>43.2</td>
<td>50.3</td>
<td>2.9</td>
</tr>
</tbody>
</table>

\( ^1 \Theta \) indicates optimization over \( \{ \Theta \_k \} \) with \( \Lambda_k = 1_D \) (\( p_\lambda = 0 \)), \( \{ \Theta, \Lambda \} \) indicates optimization over \( \Theta, \Lambda_k \) (\( p_\lambda = 0.01 \)).

Table 2. Mean accuracies on the test set over 100 iterations of cross validation on various datasets (\( D \): Number of features, \( J \): Number of classes, \( M \): Number of data-samples in the dataset).

3.3. Discussion

We observe that the ensemble learning techniques consistently outperform the single classifier case. \( g_{orc} \) achieves substantial gains over the baseline result. We summarize this in Theorem 1 which proves that given an ensemble of classifier and the oracle fusion function \( g_{orc} \), we will always achieve a lower generalization error than all classifiers in the ensemble. We define generalization error as the probability (\( P_r \)) of a classifier \( f \) assigning an incorrect class to an instance \( x \) drawn from the data distribution \( P \) in (14). \( y \) is the correct class corresponding to \( x \).

\[
e_{err}(f) = P_r(\{x \sim P, y \neq f(x)\})
\]

(14)

**Lemma 1.** Probability of intersection of events \( E_1, \ldots, E_K \) is less than probability of any single event \( E_k \); \( k = 1..K \) (Frechet inequality [20]).

\[
Pr[E_1 \land E_2 \land \ldots \land E_K] \leq Pr[E_k]; \forall k = 1..K
\]

(15)

**Theorem 1.** \( err_P(g_{orc}) \leq err_P(f_k); \forall k = 1..K \)

*Proof.* \( err_P(g_{orc}) = Pr_{\{x \sim P\}}[g_{orc}(x) \neq y] \)

(14)

(Since \( g_{orc}(x) \) is incorrect if all the classifiers are incorrect.)

\[
\leq Pr_{\{x \sim P\}}[f_k(x) \neq y] = err_P(f_k); \forall k = 1..K \text{ (from Lemma 1)}
\]

We observe that even though the proxy fusion scheme \( g_{EV} \) beats the baseline classifier it performs worse that \( g_{orc} \). This indicates a severe shortcoming in our fusion proxy. We also observe that ensemble of classifiers optimized on both \( \Theta \), \( \Lambda_k \) outperforms those optimized over \( \Theta \), with \( \Lambda_k = 1_D \) in every case except for the oracle fusion for Abalone dataset. This supports our previous premise that joint optimization should provide us with a better ensemble over pre-assigning \( \Lambda_k \) and then optimizing over \( \Theta \). We also observe that our algorithm gives the best relative improvement in case of IEMOCAP dataset, where we have relatively smaller number of data-samples given the feature dimensionality. This is consistent with our observation on the synthetic dataset where we obtained a higher relative improvement when the data is sparsely distributed.

4. CONCLUSION

We present a novel algorithm for learning an ensemble of classifiers that performs joint optimization to determine class boundaries and the feature set for each classifier in the ensemble. We introduce a loss function that introduces data-driven diversity. We sequentially optimize the loss function for each classifier in the ensemble. We use an oracle fusion function and an equal weighting function to obtain the final decision from the ensemble. We present our results on several datasets and observe that not only does our algorithm trains better classifier ensembles, it can also filter out features unrelated to class assignments.

Even though we achieve better classification, there is a room for several improvements. In particular, one can suggest other fusion proxies to better approximate the oracle fusion scheme. We also aim at providing theoretical bounds for ensemble performance given the fusion function and the set of classifiers. As a further experimentation, one can also test the joint parameter optimization on other ensemble learning techniques with different loss functions. Further, we need better investigation over the performance of our algorithm and relate it to data sparsity and the number of features. One can also work on the feature selection scheme and suggest further improvements to the selection procedure.

5. ACKNOWLEDGMENT

This research is supported by NSF, NIH and DARPA.
6. REFERENCES


