6. Classifier Ensembles and Feature Selection

6.1. Feature selection for classifier ensembles

6.2. Feature extraction for classifier ensembles: Rotation Forests.

6.3. Feature selection from classifier ensembles

6.4. Feature extraction from classifier ensembles: Boosted linear projections

6.5. Another look at Random Subspace
1. Optimise the combiner

2. Use error-correcting output codes (ECOC)

3. Try different classifier models
e.g., decision trees, neural networks,
nearest neighbour, etc.

4. Train each classifier on a different subset of features

5. Alter the training data: sample from the whole
data set, inject noise (bagging, boosting,
random forests)
1. Random subspace method: choose the subsets of features randomly
2. Select a feature subset for each class: \( L = c \); use this subset to build a classifier for the multi-class problem.
3. Use natural groups of features
4. Select subsets of features by a genetic algorithm

Select/extract features for the ensemble (to make the ensemble better)
Select/extract features from the ensemble (to use in another classifier)

Feature Selection for the ensemble (systematic)

- Favourite class model (Oza and Tumer) Select a feature subset for each class: \( L = c \) (one versus all); use this subset to build a classifier for the multi-class problem.

- Incremental model (Gunter and Bunke) Select feature subsets sequentially. The subsets for the classifiers further in the ensemble must not be the same as the previously selected subsets.
  
The merit of a feature subset for the current classifier is evaluated on the basis of the ensemble performance, not the individual classifiers (this enforces diversity)

  Any feature selection method can be used. The authors propose the floating search (Pudil) as robust and successful.
**Feature Selection for the ensemble (systematic)**

- **Iterative model** (Puuronen et al.)
  1. Start by building an ensemble by, say, the favourite-class model.
  2. Identify the median classifier, i.e., the classifier with the smallest total diversity with each member of the ensemble.
  3. Change the absent-present status of all the features, one at a time, thereby building \( n \) new classifiers, where \( n \) is the number of features.
  4. Calculate the ensemble accuracy when replacing the median classifier by each one of the new classifiers, one at a time. Keep the classifier (feature subset) that gives the best ensemble performance.
  5. If there is no improvement at step 4, then stop. Else, continue from 2.

*Example:* Suppose that the median classifier uses features 1, 3 and 4 (out of 5 features). Then the mask set is \([1 \ 0 \ 1 \ 1 \ 0]\). The 5 new classifiers will use \([0 \ 0 \ 1 \ 1 \ 0]\), \([1 \ 1 \ 1 \ 1 \ 0]\), \([1 \ 0 \ 0 \ 1 \ 0]\), \([1 \ 0 \ 1 \ 0 \ 0]\), and \([1 \ 0 \ 1 \ 1 \ 1]\).

---

**Feature Selection for the ensemble (random)**

- **Random Subspace [Ho 1998]**
  - Sample random subsets of features

- **Genetic algorithms**
  - The ensemble is represented by the whole population, i.e., we evolve feature subsets
  - The ensemble is represented by a chromosome i.e., we evolve a population of ensembles
Feature Extraction for the ensemble


Feature Extraction for the ensemble – Rotation Forest

“This altogether gives a very bad impression of ill-conceived experiments and confusing and unreliable conclusions. ... The current spotty conclusions are incomprehensible, and are of no generalization or reference value.”

“This is a potentially great new method and any experimental analysis would be very useful for understanding its potential. Good study, with very useful information in the Conclusions.”
Random Forests (Breiman)

For $k = 1, \ldots, L$
- Take a bootstrap sample $S_k$ from $Z$ of size $N$.
- Build a tree-classifier $D_k$ using $S_k$ as the training set.
- To split a node of the tree, select the best feature among $M$ randomly selected candidate-features (e.g., $M=5$).

End $k$

Majority voting: for an unlabeled $x$, take the votes of the $L$ classifiers and calculate $g_k(x) = \sum$ votes for $\omega_k$, $k = 1, \ldots, c$.

Pick the class with the largest support.

Rotation Forests (Rodriguez et al., IEEE TPAMI, 2006)

For $k = 1, \ldots, L$
- Take a bootstrap sample $S_k$ from $Z$ of size $N$.
- Build a tree-classifier $D_k$ using $S_k$ as the training set.
- Form a new set of extracted features and use this set to build the classifier.

End $k$

Majority voting: for an unlabeled $x$, take the votes of the $L$ classifiers and calculate $g_k(x) = \sum$ votes for $\omega_k$, $k = 1, \ldots, c$.

Pick the class with the largest support.
Rotation Forests - the feature extraction part

Step 1. Split randomly the feature set into K subsets (Assume K is a factor of the number of features)

Step 2. For each feature subset, apply PCA on the data using only these features and a random sub-sample of the classes

Step 3. Pool all principal components to form a new set of extracted features.

Note 1: No principal components are discarded.

Note 2: Applying a PCA is equivalent to rotating the feature axes. K different rotations are carried out to obtain a new set with extracted features.

---

Feature Extraction for the ensemble – Rotation Forest

Rotation Forests - the feature extraction part

- initial features
- random split into 3 subsets
- PCA on each subset using a bootstrap sample from a random subset of classes
- pool all PCs
- train a classifier (decision tree) on the new feature set and a bootstrap sample from Z
Rotation Forests - why do they work?

**Diversity.** Each decision tree uses different set of axes which are not obtained through a simple rotation of the original space. Trees are sensitive to rotation of the axes because they build classification regions with lines parallel to the axes.

**Accuracy.** No principal components are discarded, hence all the information is preserved. The whole data set is used to train each classifier (with different extracted features), therefore we do not expect degradation of accuracy due to sub-sampling the data.

Kappa-error diagrams for the UCI vowel data
Rotation Forests - EXTRACTION OF FEATURES

Step 1. Split randomly the feature set into K subsets (Assume K is a factor of the number of features)

Step 2. For each feature subset, apply PCA on the data using only these features and a random sub-sample of the classes

Step 3. Pool all principal components to form a the new set of extracted features.

Classification Step: Majority voting: for an unlabeled x, take the votes of the \( L \) classifiers and calculate \( \mu_k(x) = \sum \text{votes for } \omega_k, \quad k = 1, \ldots, c \). Pick the class with the largest support.
Rotation Forests - EXTRAECTION OF FEATURES

1. Is splitting essential?

Step 1. Split randomly the feature set into K subsets (Assume K is a factor of the number of features)

Step 2. For each feature subset, apply PCA on the data using only these features and a random sub-sample of the classes

Step 3. Pool all principal components to form a new set of extracted features.

Classification Step: Majority voting: for an unlabeled x, take the votes of the L classifiers and calculate \( \mu_k(x) = \sum \omega_k \), \( k = 1, \ldots, c \). Pick 3. What happens for different L?

Feature Extraction for the ensemble – Rotation Forest – an experiment

Data (32 data sets from UCI)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Extrainction</th>
<th>labor</th>
<th>lymphography</th>
<th>pendigits</th>
<th>prima-diabetes</th>
<th>primary-tumor</th>
<th>segment</th>
<th>sonar</th>
<th>soybean</th>
<th>splice</th>
<th>vehicle</th>
<th>vote</th>
<th>vowel-c</th>
<th>vowel-n</th>
<th>waveform</th>
<th>wisconsin-breast-cancer</th>
<th>zoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>anneal</td>
<td>6</td>
<td>898</td>
<td>32</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>audiology</td>
<td>23</td>
<td>226</td>
<td>69</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>autos</td>
<td>7</td>
<td>205</td>
<td>10</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>balance-scale</td>
<td>3</td>
<td>625</td>
<td>0</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breast-cancer</td>
<td>2</td>
<td>286</td>
<td>10</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cleveland-14-heart</td>
<td>5</td>
<td>307</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>credit-rating</td>
<td>2</td>
<td>690</td>
<td>9</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>german-credit</td>
<td>2</td>
<td>1000</td>
<td>13</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>glass</td>
<td>7</td>
<td>214</td>
<td>0</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heart-statlog</td>
<td>2</td>
<td>270</td>
<td>0</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hepatitis</td>
<td>2</td>
<td>155</td>
<td>13</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horse-colic</td>
<td>2</td>
<td>368</td>
<td>16</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hungarian-14-heart</td>
<td>5</td>
<td>294</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hyperthyroid</td>
<td>4</td>
<td>3772</td>
<td>22</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ionosphere</td>
<td>2</td>
<td>351</td>
<td>0</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iris</td>
<td>3</td>
<td>150</td>
<td>0</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Feature Extraction for the ensemble – Rotation Forest – an experiment

% of data sets (out of 32) where the respective ensemble method is best

Rotation Forest
Random Forest
Bagging
Boosting

Ensemble size

Feature Extraction for the ensemble – Rotation Forest – an experiment

Average rank

Random Forest
Bagging
AdaBoost
Rotation Forest

Ensemble size
Feature Extraction for the ensemble – Rotation Forest – an experiment

1. Is splitting essential?
YES. Non-sparse methods do not work well

2. K = ?
No consistent pattern was found. M = 3 features works well (recommended)

3. What happens for different L?
Better than the other methods for small L but also better for large L!

4. Is PCA the best “rotation” method?
YES. Slight edge over COMPLETELY random choice of the sparse matrix.

Feature Selection from the ensemble


For very large feature sets \( n >> N \).

Sample random subsets (as in RS) → assign weights → sum up the weights → select the top M features
**Feature Selection from the ensemble**


For very large feature sets ($n >> N$).

Sample random subsets (as in RS)

assign weights

sum up the weights

select the top M features

---

**Feature Extraction from the ensemble**


Recall: AdaBoost

re-sample training data set based on previous error

extract linear projections which minimise the weighted error

extracted (linear) features
Another look at the Random Subspace method

$n$: total number of features
$Q$: number of “important” features
$M$: sample size
$L$: ensemble size

Definition 1. A classifier is called usable iff its feature set contains one or more of the important features.

Definition 2. The usability of an ensemble is the proportion of usable classifiers out of $L$.

Definition 3. The degree of coverage of an ensemble is the proportion of important features (out of $Q$) selected in at least one feature subset.

Definition 4. Let $X$ and $Y$ be the feature subsets for two classifiers in the ensemble, and $X_q \subseteq X$ and $Y_q \subseteq Y$ be the respective subsets containing the important features. The feature set diversity (FSD) between the two classifiers is defined as

$$|X_q \cup Y_q| - |X_q \cap Y_q|,$$

where $|.|$ denotes cardinality.

The research hypothesis here is that ensembles would have better accuracy if they have

- high usability
- high degree of coverage
- high feature set diversity

Thus we set out to calculate the relationship between $(L,M)$, on the one hand, and the probability of full coverage, complete usability and guaranteed diversity between all pairs of classifiers, on the other hand. In addition, we will try to support the research hypothesis.

The expected benefit will be an automatic method for selection of $L$ and $M$ for a given $n$ and a hypothesised $Q$ (application to fMRI classification).
Another look at the Random Subspace method

The research hypothesis here is that ensembles would have better accuracy if they have

- high usability
- high degree of coverage
- high feature set diversity

Thus we set out to calculate the relationship between \((L,M)\), on the one hand, and the probability of complete usability and guaranteed diversity between pairs of classifiers, on the other hand. In addition, we will try to support the research hypothesis.

The expected benefit will be an automatic method for selection of \(L\) and \(M\) for a given \(n\) and a hypothesised \(Q\) (application to fMRI classification).

---

PhD Reviewer...

The Joke

Giorgio Valentini’s PhD Thesis
**Random Subspace ensembles**

| \( n = 12 \) | number of features |
| \( Q = 4 \) | important features |
| \( M = 6 \) | sample size |
| \( L = 4 \) | ensemble size |

**PROBLEM:** The sample of a feature subset for each classifier is random, uniform and without replacement. The \( L \) samples (each of size \( M \)) are taken independently.

Calculate the probability that all the classifiers in the ensemble will be usable (i.e., the feature subset of every classifier contains at least one important feature).

**Answer behind this shape**

0.74