8.2 Methods for Constructing Ensembles of Classifiers

The methods generally vary the input given to each classifier. These include sub-sampling the training set, manipulating the features of the training set, manipulating the output target, injecting randomness and some methods which are specific to particular algorithms.

8.2.1 Sub-sampling the Training Examples

The learning algorithm is run several times, each time with a reduced data set obtained from the training set. This method runs particularly well on algorithms where the output classifier undergoes major changes in response to small changes in the training set. The data set can be reduced by choosing a sub-set of the training set or by using a different sub-set of the features in each case.

Bagging

On each run, this technique presents the classifier with a training set that consists of m training examples drawn randomly from the original training set of n items. Such a training set is called a bootstrap replicate of the original training set and the technique is called *bootstrap aggregation*. Each bootstrap replicate contains about two-thirds of the original training set with some patterns possibly appearing several times.

EXAMPLE 2

Consider the following training set :

$$X_1 = (1, 1, X); \quad X_2 = (2, 1, X); \quad X_3 = (3.3, 1, X); \quad X_4 = (1, 2, X);$$

$$X_5 = (2, 2, X); \quad X_6 = (5, 1, O); \quad X_7 = (6, 1, O); \quad X_8 = (5, 2, O);$$

$$X_9 = (6, 2, O); \quad X_{10} = (5, 3, O)$$

Here each triplet gives the first feature, the second feature and the class label. This is shown in Figure 8.2.

If there is a test pattern at (4, 2), using the given training set, it is closer to Class O and will be classified in this case as belonging to Class O if the nearest neighbour algorithm is used. If bagging is used and two patterns are drawn at random from each class to be used for classification, the class of the test pattern will vary depending on the patterns drawn. If patterns 3 and 5 are drawn from Class X and patterns 7 and 9 are drawn from Class O, then the test pattern will be classified as belonging to Class X. If patterns 1 and 4 are drawn from Class X and patterns 6 and 7 from Class O, then the test pattern will be classified as belonging to Class O. If the classification is

done a number of times with different patterns from the two classes at random, these results have to be combined to find the class of the test pattern.

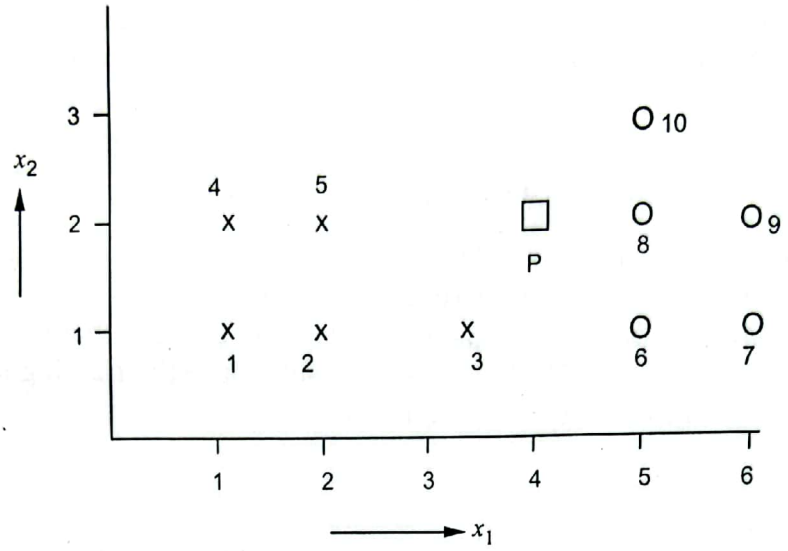

Figure 8.2 Training data set having two classes

Leaving Out Disjoint Sub-sets

The original training set is divided into a number of disjoint sub-sets. Then different overlapping training sets are constructed by dropping one of these sub-sets.

EXAMPLE 3

Consider the data set given in Figure 8.2. Let us take the disjoint sub-sets as $S_1 = \{1, 2\}$, $S_2 = \{4, 5\}$, $S_3 = \{3\}$, $S_4 = \{6, 7\}$, $S_5 = \{8, 10\}$, $S_6 = \{9\}$.

If sub-sets S_1 and S_4 are left out, then the data set will consist of the points {3, 4, 5, 8, 9, 10} and P will be classified as belonging to Class O. In this way, by leaving out one sub-set from each class, we can get different sets of points. If S_1 and S_5 are left out, the data set will consist of the points {3, 4, 5, 6, 7, 9} and P will be classified as belonging to Class X. If S_1 and S_6 are left out, the data set will consist of {3, 4, 5, 6, 7, 8, 10} and P will be classified as belonging to Class O. If S_2 and S_4 are left out, the data set will consist of {1, 2, 3, 8, 9, 10} and P will be classified as belonging to Class O. If S_2 and S_5 are left out, the data set will consist of {1, 2, 3, 6, 7, 9} and P will be classified as belonging to Class X. If S_2 and S_6 are left out, the data set will be {1, 2, 3, 6, 7, 8, 10} and P will be classified as belonging to Class O. If S_3 and S_4 are left out, the data set will consist of {1,2,4,5,8,9,10} and P will be classified as belonging to Class O. If S_3 and S_5 are left out, the data set will consist of {1, 2, 4, 5,

6, 7, 9} and P will be classified as belonging to Class X. If S_3 and S_6 are left out, the data set will consist of {1, 2, 4, 5, 6, 7, 8, 10}, and P will be classified as belonging to Class O. Thus, if this is done a number of times, the classification of P will depend on which sub-sets are left out. The combination of these decisions will decide on the classification of P.

ADABOOST Algorithm

The general problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules-of-thumb is referred to as *boosting*. The booster is provided with a set of labelled training examples $(x_1, \theta_1), ..., (x_N, \theta_N)$, where θ_i is the label associated with instance x_i. On each round $t = 1, ..., T$, the booster devices a distribution D_t over the set of examples and uses a weak hypothesis h_t having low error ϵ_t with respect to D_t. Any classifier can be used at this point. The weak classifier can be a decision stump—a decision tree with depth one. In other words, the classification is based on a single decision node. Thus, distribution D_t specifies the relative importance of each example for the current round. After T rounds, the booster must combine the weak hypothesis into a single prediction rule.

One boosting algorithm is the ADABOOST which maintains a probability distribution $p_t(x)$ over the training set. In each iteration t, it draws a training set of size m by sampling with replacement according to the probability distribution $p_t(x)$. The classifier is used on this training set. The error rate e_t of this classifier is computed and used to adjust the probability distribution on the training set. The probability distribution is obtained by normalising a set of weights, $w_t(i), i = 1, .., n$ over the training set. The effect of the change in weights is to place more importance on training examples that were misclassified by the classifier and less weight on the examples that were correctly classified. The final classification is constructed by a weighted vote of the individual classifiers. Each classifier is weighted according to its accuracy for the distribution p_t that it was trained on.

If the input is a training set S of size n, the inducer is $\mathcal{I}$ and the number of trials T, the algorithm is as follows :

STEP 1: $S' = S$ with weights assigned to be 1; $m = n$;

STEP 2: Consider $i = 1$ to T

STEP 3: $C_i = \mathcal{I}(S')$

STEP 4: $\epsilon_i = \dfrac{1}{m} \displaystyle\sum_{x_j \in S' : C_i(x_j) \neq y_j} \text{weight}(x)$

STEP 5: If $\epsilon_i > \frac{1}{2}$, set S' to a bootstrap sample from S with weight 1 for every instance and go to Step 3.

STEP 6: $\beta_i = \frac{\epsilon_i}{(1-\epsilon_i)}$

STEP 7: For each $x_j \in S'$, if $C_i(x_j) = y_j$ then weight(x_j) = weight(x_j) . β_i

STEP 8: Normalise the weights of instances so that the total weight of S' is m.

STEP 9:

$$C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} \log \frac{1}{\beta_i}$$

In the above algorithm, every weight is taken as 1 and we start with the whole training data. A classifier is chosen. The ϵ error of using this classifier is obtained by adding the weights of the patterns which are misclassified and dividing the sum by the total number of patterns m. This gives the equation in Step 4. β is calculated as in Step 6 which represents the error in classification. The weights of the samples which are classified correctly are updated(reduced) by multiplying their weight with β. They are then normalised so that they sum to m. This is carried out for different classifiers. The inducer (I) decides the classifier to be used. Step 5 is used to see that the error does not become too high for the classifier. Step 9 gives the equation to be used to combine the classifiers for a test pattern. The term $\log \frac{1}{\beta}$ represents the classification accuracy of the particular classifier. A test pattern P is classified according to different hypotheses. For each class, the summation of $\log \frac{1}{\beta}$ is carried out only for the hypothesis for which it is classified as belonging to that class. The class which has the highest term for the summation is chosen as the class of the test pattern.

EXAMPLE 4

Consider Figure 8.2. Let a weight of 1 be assigned to all the samples, i.e., weight$(i)=1$, $i = 1,...,10$. Consider three hypotheses where Hypothesis 1 and Hypothesis 2 are decision stumps.

Hypothesis 1

Let the first hypothesis be that if $x_1 \leq 3$, the pattern belongs to Class X and Class O otherwise. This hypothesis misclassifies pattern 3. Which means

$$\epsilon_1 = \frac{1}{10} = 0.1$$

$$\beta_1 = \frac{0.1}{0.9} = 0.1111$$

$$\text{weight}(1) = 1 \times 0.1111 = 0.1111$$

Similarly the weights of the other patterns except pattern 3 will be 0.1111. Only the weight of pattern 3 remains as 1. Normalising,

$$\text{weight}(1) = \frac{0.1111}{1.9999} \times 10 = 0.5555$$

$$\text{weight}(2) = 0.5555,\ \text{weight}(4) = 0.5555,\ \text{weight}(5) = 0.5555,$$

$$\text{weight}(6) = 0.5555,\ \text{weight}(7) = 0.5555,\ \text{weight}(8) = 0.5555,$$

$$\text{weight}(9) = 0.5555 \text{ and } \text{weight}(10) = 0.5555$$

$$\text{weight}(3) = \frac{1}{1.9999} \times 10 = 5.0002$$

Hypothesis 2

Let the second hypothesis be that if $x_1 \leq 5$, the pattern belongs to Class X and Class O otherwise.

$$\epsilon_2 = \frac{1}{10} \times (0.5555 + 0.5555 + 0.5555) = 0.16665$$

$$\beta_2 = \frac{0.16665}{1 - 0.16665} = 0.2000$$

$$\text{weight}(1) = 0.5555 \times 0.2 = 0.1111;\ \text{weight}(2) = 0.1111;$$

$$\text{weight}(3) = 5.0002 \times 0.2 = 1.00004;\ \text{weight}(4) = 0.1111;$$

$$\text{weight}(5) = 0.1111;\ \text{weight}(6) = 0.5555;$$

$$\text{weight}(7) = 0.1111;\ \text{weight}(8) = 0.5555;$$

$$\text{weight}(9) = 0.1111;\ \text{weight}(10) = 0.5555;$$

Normalising,

$$\text{weight}(1) = \frac{0.1111}{3.33314} \times 10 = 0.333319;\ \text{weight}(2) = 0.333319;$$

$$\text{weight}(3) = \frac{1.00004}{3.33314} \times 10 = 3.0003;\ \text{weight}(4) = 0.333319;$$

$$\text{weight}(5) = 0.333319;\ \text{weight}(6) = \frac{0.5555}{3.33314} \times 10 = 1.6666;$$

$$\text{weight}(7) = 0.333319;\ \text{weight}(8) = 1.6666;$$

weight(9) = 0.333319; weight(10) = 1.6666;

Hypothesis 3

Let the third hypothesis be that if $x_1 + x_2 \leq 3.5$, the pattern belongs to Class X and Class O otherwise. This hypothesis misclassifies pattern 3 and 5.

$$\epsilon_3 = \frac{1}{10} \times (3.0003 + 0.333319) = 0.3334$$

$$\beta_3 = \frac{0.3334}{1 - 0.3334} = 0.5002$$

weight(1)= 0.333319 × 0.5002 = 0.16673; weight(2)= 0.16673;

weight(3)= 3.0003; weight(4)= 0.16673;

weight(5)= 0.333319; weight(6)= 1.6666 × 0.5002 = 0.8336;

weight(7)= 0.16673; weight(8)= 0.8336; weight(9)= 0.16673; weight(10)= 0.8336;

Normalising,

$$\text{weight}(1) = \frac{0.16673}{6.668069} \times 10 = 0.2502; \text{ weight}(2) = 0.2502;$$

$$\text{weight}(3) = \frac{3.0003}{6.668069} \times 10 = 4.4995; \text{ weight}(4) = 0.2502;$$

$$\text{weight}(5) = \frac{0.333319}{6.668069} \times 10 = 0.4999; \text{ weight}(6) = \frac{0.8336}{6.668069} \times 10 = 1.2501;$$

weight(7)= 0.2502; weight(8)= 1.2501;

weight(9)= 0.2502; weight(10)= 1.2501;

If we take a test pattern (4, 2), according to the first hypothesis, it belongs to class O; according to the second hypothesis, it belongs to class X; and according to the third hypothesis, it belongs to class O. For Class X,

$$\Sigma \log \frac{1}{\beta_i} = \log \frac{1}{0.2} = 0.699$$

For Class O,

$$\Sigma \log \frac{1}{\beta_i} = \log \frac{1}{0.1111} + \log \frac{1}{0.5002} = 1.2551$$

P will be classified as belonging to Class O.