Combining Classifiers
Sections 4.1 - 4.4

Nicolette Nicolosi
Ishwarryah S Ramanathan

October 17, 2008

4.1 - Types of Classifier Outputs

1. Abstract Level: Each classifier D_i returns a label $s_i \in \Omega$ for $i = 1$ to L. A vector $s = [s_i, ..., s_L]^T \in \Omega^L$ is defined for each object to be classified, using all L classifier outputs. This is the most universal level, so any classifier is capable of giving a label. However, there is no additional information about the label, such as probability of correctness or alternative labels.

2. Rank Level: The output of each classifier $D_i \in \Omega$, and alternatives are ranked in order of probability of being correct. This type is frequently used for systems with many classes.

3. Measurement Level: D_i returns a c-dimensional vector $[d_{i,1}, ..., d_{i,c}]^T$, where $d_{i,j}$ is a value between 0 and 1 that represents the probability that the object to be classified is in the class ω_j.

4. Oracle Level: Output of D_i is only known to be correct or incorrect, and information about the actual assigned label is ignored. This can only be applied to a labeled data set. For a data set Z, D_i produces the output vector
$y_{ij} = \{1$ if z_j is correctly classified by D_i; 0 otherwise$\}$

4.2 - Majority Vote

Consensus Patterns

1. Unanimity - 100% agree on choice to be returned

2. Simple Majority - 50% + 1 agree on choice to be returned

3. Plurality - Choice with the most votes is returned

Majority Vote

Classifiers output a c-dimensional binary vector $[d_{i,1}, ..., d_{i,c}]^T \in \{0,1\}^c$, where $i = 1, ..., L$ and $d_{i,j} = 1$ if D_i labels x in ω_i, and $d_{i,j} = 0$ otherwise. In this case, plurality will result in a decision for ω_k if

$$\sum_{i=1}^{L} d_{i,k} = \max_{i=1}^{c} \sum_{i=1}^{L} d_{i,j},$$

and ties are resolved in an arbitrary manner.

The plurality vote is often called the majority vote, and it is the same as the simple majority when there are two classes ($c = 2$).

Threshold Plurality

A variant called threshold plurality vote adds a class ω_{c+1}, to which an object is assigned when the ensemble cannot decide on a label, or in the case of a tie. The decision then becomes:

ω_k, if $\sum_{i=1}^{L} d_{i,k} >= \alpha * L$
ω_{c+1}, otherwise

where $0 < \alpha <= 1$

Using the threshold plurality, we can express the simple majority by setting $\alpha = \frac{1}{2} + \epsilon$, where $0 < \epsilon < \frac{1}{L}$, and the unanimity vote by setting $\alpha = 1$.

Properties of Majority Vote

Some assumptions for the following discussion:

1. The number of classifiers, L, is odd (makes it simple to break ties).

2. The probability that a classifier will return the correct value is denoted by p.

3. Classifier outputs are independent of each other. This makes the joint probability:
$P(D_{i_1} = s_{i_1}, ..., D_{i_K} = s_{i_K})$
$= P(D_{i_1} = s_{i_1}) * ... * P(D_{i_K} = s_{i_K})$, where s_{i_i} is the label give by classifier D_{i_i}.

The majority vote gives an accurate label if at least $\lfloor \frac{L}{2} \rfloor + 1$ classifiers return correct values. So the accuracy of the ensemble is:

$$P_{maj} = \sum_{m=\lfloor \frac{L}{2} \rfloor + 1}^{L} \binom{L}{m} p^m (1-p)^{L-m}$$

Condorcet Jury Theorem

The Condorcet Jury Theorem supports the intuitive expectation that improvements over the individual accuracy p will only occur when p is larger than 0.5.

1. If $p > 0.5$, P_{maj} is monotonically increasing (strictly increasing) and $P_{maj} \to 1$ as $L \to \infty$.

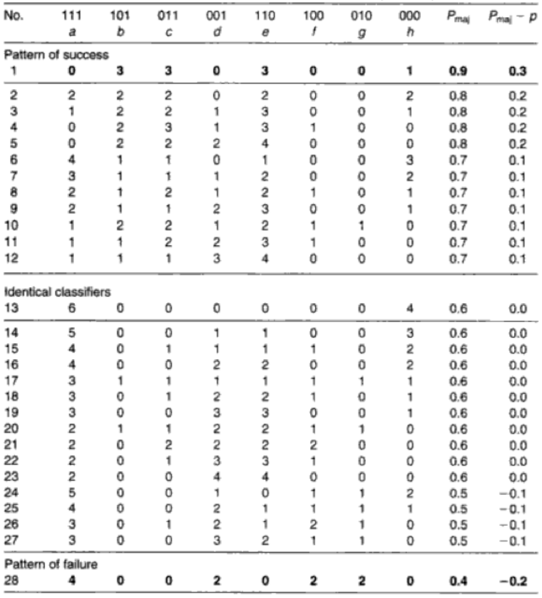
2. If $p < 0.5$, P_{maj} is monotonically decreasing and $P_{maj} \to 0$ as $L \to \infty$

3. If $p = 0.5$, $P_{maj} = 0.5$ for any L.

Limits on Majority Vote

$D = \{D_1, D_2, D_3\}$ is an ensemble of three classifiers, each of which has the same probability of correctly classifying a sample ($p = 0.6$). All combinations distributing 10 elements into the 8 combinations of outputs can be represented if we represent each classifier output as either a 0 or a 1. For example, 101 would represent the case where the first and third, but not the second, classifiers correctly labeled a certain number of samples X.

TABLE 4.3 All Possible Combinations of Correct/Incorrect Classification of 10 Objects by Three Classifiers so that Each Classifier Recognizes Exactly Six Objects.



No.	111	101	011	001	110	100	010	000	P_{maj}	$P_{maj} - p$
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*		
Pattern of success										
1	**0**	**3**	**3**	**0**	**3**	**0**	**0**	**1**	**0.9**	**0.3**
2	2	2	2	0	2	0	0	2	0.8	0.2
3	1	2	2	1	3	0	0	1	0.8	0.2
4	0	2	3	1	3	1	0	0	0.8	0.2
5	0	2	2	2	4	0	0	0	0.8	0.2
6	4	1	1	0	1	0	0	3	0.7	0.1
7	3	1	1	1	2	0	0	2	0.7	0.1
8	2	1	2	1	2	1	0	1	0.7	0.1
9	2	1	1	2	3	0	0	1	0.7	0.1
10	1	2	2	1	2	1	1	0	0.7	0.1
11	1	1	2	2	3	1	0	0	0.7	0.1
12	1	1	1	3	4	0	0	0	0.7	0.1
Identical classifiers										
13	6	0	0	0	0	0	0	4	0.6	0.0
14	5	0	0	1	1	0	0	3	0.6	0.0
15	4	0	1	1	1	1	0	2	0.6	0.0
16	4	0	0	2	2	0	0	2	0.6	0.0
17	3	1	1	1	1	1	1	1	0.6	0.0
18	3	0	1	2	2	1	0	1	0.6	0.0
19	3	0	0	3	3	0	0	1	0.6	0.0
20	2	1	1	2	2	1	1	0	0.6	0.0
21	2	0	2	2	2	2	0	0	0.6	0.0
22	2	0	1	3	3	1	0	0	0.6	0.0
23	2	0	0	4	4	0	0	0	0.6	0.0
24	5	0	0	1	0	1	1	2	0.5	−0.1
25	4	0	0	2	1	1	1	1	0.5	−0.1
26	3	0	1	2	1	2	1	0	0.5	−0.1
27	3	0	0	3	2	1	1	0	0.5	−0.1
Pattern of failure										
28	**4**	**0**	**0**	**2**	**0**	**2**	**2**	**0**	**0.4**	**−0.2**

In the table, there is a case where the majority vote is correct 90 percent of the time. This is unlikely, but it is an improvement over the individual rate $p = 0.6$. There is also a case in which the majority vote is correct only 40 percent of the time, which is worse than the individual rate. These best and worst possible cases are "the pattern of success" and "the pattern of failure," respectively.

Patterns of Success and Failure

p_i is an individual accuracy for classifier D_i. Let $l = \lfloor \frac{L}{2} \rfloor$.

A pattern is a success pattern if the:

1. Probability of any combination of $\lfloor \frac{L}{2} + 1 \rfloor$ correct and $\lfloor \frac{L}{2} \rfloor$ incorrect votes is α

2. Probability of all L votes being incorrect is γ

3. Probability of any other combination is 0

The pattern of success occurs when exactly $\lfloor \frac{L}{2} \rfloor + 1$ votes are correct. This results in using the minimum number of votes required, without wasting votes. In this case,

$$P_{maj} = \binom{L}{l+1} \alpha,$$

and the pattern of success is possible when $P_{maj} \leq 1$, so $\alpha \leq \frac{1}{\binom{L}{l+1}}$. Using these definitions, we can rewrite the accuracy $p = \binom{L-1}{1} \alpha$. Substituting this rewritten definition for p, we obtain:

$$P_{maj} = \frac{pL}{l+1} = \frac{2pL}{L+1}$$

If $P_{maj} \leq 1$ and $p \leq \frac{L+1}{2L}$, then:

$$P_{maj} = \min \left\{ 1, \frac{2pL}{L+1} \right\}$$

A pattern is a failure pattern if the:

1. Probability of any combination of $\lfloor \frac{L}{2} \rfloor$ correct and $\lfloor \frac{L}{2} \rfloor + 1$ incorrect votes is β

2. Probability of all L votes being incorrect is δ

3. Probability of any other combination is 0

The pattern of failure occurs when exactly l out of L classifiers are correct. In this case,

$$P_{maj} = \delta = 1 - \binom{L}{l} \beta$$

The accuracy p can be rewritten using P_{maj} and α:

$$p = \delta + \binom{L-1}{l-1} \beta$$

These equations can be combined to give:

$$P_{maj} = \frac{pL-1}{l+1} = \frac{(2p-1)(L+1)}{L+1}$$

Matan's Upper and Lower Bounds

A classifier D_i has accuracy p_i, and L classifiers are ordered so that $p_1 \leq p_2 \leq p_3, \ldots, \leq p_L$. Let $k = l+1 = \frac{(L+1)}{2}$ and $m = 1, 2, 3, \ldots, k$.

The upper bound is the same as the pattern of success:

$$\max P_{maj} = \min \left\{ 1, \sum k, \sum k-1, \ldots, \sum 1 \right\}$$

where

$$\sum m = \left(\frac{1}{m}\right) \sum_{i=1}^{L-k+m} p_i$$

The lower bound is the same as the pattern of failure:

$$\min P_{maj} = \max \left\{ 0, \xi(k), \xi(k-1), \ldots, \xi(1) \right\}$$

where

$$\xi(m) = \left(\frac{1}{m}\right) \sum_{i=k-m+1}^{L} p_i - \frac{(L-k)}{m}$$

4.3 - Weighted Majority Vote

Adding weights to the majority vote is an attempt to favor the more accurate classifiers in making the final decision. Representing label outputs in the following way uses them as "degrees of support" for the classes:

$$d_{i,j} = \begin{cases} 1 & \text{if } D_i \text{ labels x in } \omega_j, \\ 0 & \text{otherwise.} \end{cases}$$

The discriminant function for class ω_j is:

$$g_j(x) = \sum_{i=1}^{L} b_i d_{i,j}$$

where b_i is a coefficient for D_i. The discriminant function is the sum of coefficients for classifiers in the ensemble for which the output on x is ω_j.

4.4 - Naive Bayes Combination

Naive Bayes combination assumes that classifiers are mutually independent given a class label. In practice, the classifiers are frequently dependent upon each other in spite of this assumption. Interestingly, the Bayes classifier is still often fairly accurate and efficient in these situations. The probability that D_j labels x in class $s_j \in \Omega$ is $P(s_j)$. The conditional independence is then:

$$P(s|\omega_k) = P(s_1, s_2, ..., s_L|\omega_k) = \prod_{i=1}^{L} P(s_i|\omega_k)$$

From this equation, it follows that the posterior probability necessary to label x is:

$$P(\omega_k|s) = \frac{P(\omega_k)P(s|\omega_k)}{P(s)}$$

$$P(\omega_k|s) = \frac{P(\omega_k)\prod_{i=1}^{L} P(s_i|\omega_k)}{P(s)},$$

for $k = 1, ..., c$

The denominator is ignored because it is irrelevant for ω_k, so the support for ω_k is:

$$\mu_k(x) \propto P(\omega_k) \prod_{i=1}^{L} P(s_i|\omega_k)$$

One way to select weights

Consider an ensemble of L independent classifiers. D_i denotes a classifier and p_i denotes its associated individual accuracy. The accuracy of the ensemble is maximized by assigning weights:

$$b_i \propto \log \frac{p_i}{1-p_i}$$

For each classifier D_i, a c by c confusion matrix CM^i defined by applying D_i to the training set. The (k,s)th entry of the matrix $cm^i_{k,s}$ represents the number of elements that belong to ω_k that were assigned the ω_s by D_i. This confusion matrix can be used to estimate the probability $P(s_i|\omega_k)$. Specifically,

$$P(s_i|\omega_k) = \frac{cm^i_{k,s}}{N_k}$$

The estimated posterior probability for ω_s is $\frac{N_k}{N}$. With this, we can rewrite the support equation for

ω_k:

$$\mu_k(x) \propto \left(\frac{1}{N_k^{L-1}}\right) \prod_{i=1}^{L} cm_{k,s_i}^{i}$$

If the estimate for $P(s_i|\omega_k)$ is zero, $\mu_k(x)$ is nullified.