Is Independence Good For Combining Classifiers?

L.I. Kuncheva, C.J. Whitaker, C.A. Shipp
School of Informatics,
University of Wales, Bangor
Gwynedd LL57 1UT, UK
E-mail: {l.i.kuncheva,c.j.whitaker}@bangor.ac.uk

R.P.W. Duin
Faculty of Applied Sciences
Delft University of Technology
P.O. Box 5046, 2600 GA Delft, The Netherlands
E-mail: duin@ph.tn.tudelft.nl

Abstract

Independence between individual classifiers is typically viewed as an asset in classifier fusion. We study the limits on the majority vote accuracy when combining **dependent** *classifiers. Q statistics are used to measure the dependence between classifiers. We show that dependent classifiers could offer a dramatic improvement over the individual accuracy. However, the relationship between dependency and accuracy of the pool is ambivalent. A synthetic experiment demonstrates the intuitive result that, in general, negative dependence is preferable.*

Table 1. Tabulated values of the majority vote accuracy of L independent classifiers with individual accuracy p

	$L = 3$	$L = 5$	$L = 7$	$L = 9$
$p = 0.6$	0.6480	0.6826	0.7102	0.7334
$p = 0.7$	0.7840	0.8369	0.8740	0.9012
$p = 0.8$	0.8960	0.9421	0.9667	0.9804
$p = 0.9$	0.9720	0.9914	0.9973	0.9991

1. Introduction

Let $\mathcal{D} = \{D_1, \ldots, D_L\}$ be a set (pool) of classifiers such that $D_i : \Re^n \rightarrow \Omega$, where $\Omega = \{\omega_1, \ldots, \omega_c\}$, assigns $\mathbf{x} \in \Re^n$ a class label $\omega_j \in \Omega$.

The majority vote method of combining classifier decisions, one of many methods in this important research area [2, 3, 4, 5, 6, 7, 8, 9], is to assign the class label ω_j to $\mathbf{x}$ that is supported by the majority of the classifiers D_i.

Finding independent classifiers is one aim of classifier fusion methods for the following reason. Let L be odd, $\Omega = \{\omega_1, \omega_2\}$, and all classifiers have the same classification accuracy p. The majority vote method with independent classifier decisions gives an overall correct classification accuracy calculated by the binomial formula

$$P_{maj} = \sum_{m=0}^{\lfloor L/2 \rfloor} \left(\begin{array}{c} L \\ m \end{array} \right) p^{L-m}(1-p)^m, \qquad (1)$$

where $\lfloor a \rfloor$ denotes the largest integer smaller than a. The probability of a correct classification for $p = 0.6, 0.7, 0.8, 0.9$ and $L = 3, 5, 7, 9$ is shown in Table 1. Then the majority vote method with independent classifiers is guaranteed to give a higher accuracy than individual classifiers when $p > 0.5$.

In this study we are interested in combining **dependent** classifiers and establishing a relationship between the dependence and the accuracy of the pool. If all classifiers are totally positively dependent (i.e., they are identical) there will be no improvement over p. However, if there are negatively dependent, i.e., commit mistakes on strongly different objects, we could expect improvement over the predicted value for independent classifiers.

2. Dependency between classifiers

Let $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ be a labeled data set, $\mathbf{z}_j \in \Re^n$ coming from the classification problem in question. For each classifier D_i we design an N-dimensional output vector $\mathbf{y}_i = [y_{1,i}, \ldots, y_{N,i}]^T$ of *correct classification*, such that $y_{j,i} = 1$, if D_i recognizes correctly $\mathbf{z}_j$, and 0, otherwise. There are various statistics to assess the similarity of D_i and D_k [1]. The Q statistic for two classifiers is

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \qquad (2)$$

where N^{ab} is the number of elements $\mathbf{z}_j$ of $\mathbf{Z}$ for which $y_{j,i} = a$ and $y_{j,k} = b$ (see Table 2).

For statistically **independent** classifiers, $Q_{i,k} = 0$. Q varies between -1 and 1. Classifiers that tend to recognize *the same* objects correctly will have positive values of Q, and those which commit errors on different objects will render Q negative.

Table 2. A 2×2 table of the relationship between a pair of classifiers

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	N^{11}	N^{10}
D_i wrong (0)	N^{01}	N^{00}

Total, $N = N^{00} + N^{01} + N^{10} + N^{11}$.

Shown below are four 2×2 tables and the respective Q's $(N = 100)$.

$Q = -1$

0	50
50	0

$Q = -0.5$

30	30
30	10

$Q = 0$

36	24
24	16

$Q = 1$

50	0
0	50

3. A synthetic example

Let $\mathcal{D} = \{D_1, D_2, D_3\}$ and $N = |\mathbf{Z}| = 10$. We assume that all three classifiers have the same individual accuracy of correct classification, $p = 0.6$. This is manifested by each classifier labeling correctly 6 of the 10 elements of **Z**. Given these requirements, *all* possible combinations of distributing 10 elements into the 8 combinations of outputs of the three classifiers are shown in Table 3. For a correct overall decision by the majority vote for some $\mathbf{z}_j \in \mathbf{Z}$, at least two of the three outputs $\mathbf{y}_i$ should be 1. The last column of Table 3 shows the majority vote accuracy of each of the 28 possible combinations. It is obtained as the proportion (out of 10 elements) of the sum of the entries in columns '111', '101', '011' and '110' (two or more correct votes). The best and the worst cases are highlighted in the table.

The table offers at least two interesting facts

- We can gain up to 30 % increase in the classification accuracy over the individual rate (best case in Table 3). This is a substantial improvement, especially noticing that the accuracy of the majority vote of 3 *independent* classifiers, each one of accuracy 0.6, is 0.648 (Table 1).

- Combining classifiers using the majority vote is beneficial or "neutral" in a great deal of the cases. In this example, in 12 of the 28 cases (42.9 %) the combined accuracy is greater than the limit for independent classifiers ($P_{maj} \geq 0.7$). For another 11 cases (39.3 %), the accuracy did not improve on the individual rate ($P_{maj} = p = 0.6$). In the remaining 5 cases (17.8 %) the overall accuracy was below the individual error rate ($P_{maj} < 0.6$). It is unknown which of these 28 distributions is most likely to occur in a real-life experiment. Therefore, even though most of the cases are no worse than the individual classifiers, improvement over p is *not guaranteed*.

Table 3. All possible combination of correct/incorrect classification of 10 objects by three classifiers so that each classifier recognizes exactly 6 objects. The entries in the table are the number of occurrences of the specific binary output of the three classifiers in the particular combination. The majority vote accuracy P_{maj} is shown in the last column.

No	1 1 1	1 0 1	0 1 1	0 0 1	1 1 0	1 0 0	0 1 0	0 0 0	P_{maj}
1	0	2	2	2	4	0	0	0	0.8
2	0	2	3	1	3	1	0	0	0.8
3	**0**	**3**	**3**	**0**	**3**	**0**	**0**	**1**	**0.9**
4	1	1	1	3	4	0	0	0	0.7
5	1	1	2	2	3	1	0	0	0.7
6	1	2	2	1	2	1	1	0	0.7
7	1	2	2	1	3	0	0	1	0.8
8	2	0	0	4	4	0	0	0	0.6
9	2	0	1	3	3	1	0	0	0.6
10	2	0	2	2	2	2	0	0	0.6
11	2	1	1	2	2	1	1	0	0.6
12	2	1	1	2	3	0	0	1	0.7
13	2	1	2	1	2	1	0	1	0.7
14	2	2	2	0	2	0	0	2	0.8
15	3	0	0	3	2	1	1	0	0.5
16	3	0	0	3	3	0	0	1	0.6
17	3	0	1	2	1	2	1	0	0.5
18	3	0	1	2	2	1	0	1	0.6
19	3	1	1	1	1	1	1	1	0.6
20	3	1	1	1	2	0	0	2	0.7
21	**4**	**0**	**0**	**2**	**0**	**2**	**2**	**0**	**0.4**
22	4	0	0	2	1	1	1	1	0.5
23	4	0	0	2	2	0	0	2	0.6
24	4	0	1	1	1	1	0	2	0.6
25	4	1	1	0	1	0	0	3	0.7
26	5	0	0	1	0	1	1	2	0.5
27	5	0	0	1	1	0	0	3	0.6
28	6	0	0	0	0	0	0	4	0.6

For each pool $\mathcal{D}$, there are $L(L-1)/2$ pairs of classifiers. Denote by $Q_{i,j}$ the Q value for classifiers D_i and D_j. The Q statistic was calculated for each pair of classifiers for each of the 28 combinations. For the winning combination ($P_{maj} = 0.9$), $Q_{1,2} = Q_{2,3} = Q_{1,3} = -0.5$. For the worst case ($P_{maj} = 0.4$), $Q_{1,2} = Q_{2,3} = Q_{1,3} = 0.333$. Although supporting the intuition that negative dependence is beneficial, this result appears to be not very indicative. Table 4 shows the sorted P_{maj} and the corresponding $Q_{1,2}, Q_{2,3}$ and $Q_{1,3}$. As seen in the table, there is no clear pattern of relationship between P_{maj} and the Q's. For a general observation, we averaged separately the

Q's for all 12 combinations for which $P_{maj} > 0.648$ (favorable) and the 16 combination for which $P_{maj} \leq 0.648$ (unfavorable). The averaged Q of the favorable combinations is -0.1227, and that of the unfavorable combinations is 0.2873. However, the values of the Q's for both groups: favorable and unfavorable, are scattered in the whole range from -1 to 1, and extracting a consistent relationship seems impossible. Triple dependence between classifiers for the 28 combinations was also calculated by

$$Q_{123} = \frac{N^{111}N^{001}N^{010}N^{100} - N^{011}N^{101}N^{110}N^{000}}{N^{111}N^{001}N^{010}N^{100} + N^{011}N^{101}N^{110}N^{000}}, \tag{3}$$

and is shown as the last column in Table 4.

Table 4. Sorted by P_{maj} combination from Table 3, the corresponding pairwise and triple Q's. The long dash means that the value could not be calculated (division by zero).

No	P_{maj}	$Q_{1,2}$	$Q_{1,3}$	$Q_{2,3}$	Q_{123}
21	0.4	0.33	0.33	0.33	1.0
15	0.5	0.88	-0.50	-0.50	1.0
17	0.5	0.33	-0.50	0.33	1.0
22	0.5	0.88	0.33	0.33	1.0
26	0.5	0.88	0.88	0.88	1.0
8	0.6	1.00	-1.00	-1.00	—
9	0.6	0.88	-1.00	-0.50	—
10	0.6	0.33	-1.00	0.33	—
11	0.6	0.33	-0.50	-0.50	1.0
16	0.6	1.00	-0.50	-0.50	—
18	0.6	0.88	-0.50	0.33	—
19	0.6	0.33	0.33	0.33	0.5
23	0.6	1.00	0.33	0.33	—
24	0.6	0.88	0.33	0.88	—
27	0.6	1.00	0.88	0.88	—
28	0.6	1.00	1.00	1.00	—
4	0.7	0.88	-1.00	-1.00	—
5	0.7	0.33	-1.00	-0.50	—
6	0.7	-0.50	-0.50	-0.50	1.0
12	0.7	0.88	-0.50	-0.50	-1.0
13	0.7	0.33	-0.50	0.33	-1.0
20	0.7	0.88	0.33	0.33	-1.0
25	0.7	0.88	0.88	0.88	-1.0
1	0.8	0.33	-1.00	-1.00	—
2	0.8	-0.50	-1.00	-0.50	—
7	0.8	0.33	-0.50	-0.50	-1.0
14	0.8	0.33	0.33	0.33	-1.0
3	0.9	-0.50	-0.50	-0.50	-1.0

The same type of synthetic experiment was carried out for $N = 100$. From the total of 36151 possible combinations, 14941 (41.3 %) have $P_{maj} > 0.648$ (favorable group). The worst part of the unfavorable group, i.e., with $P_{maj} < 0.6$, consisted of 11270 (31.2 %) combinations. The averaged values of Q for the two groups are similar to the values in our previous example, -0.1109 for the favorable group and 0.2320 for the unfavorable one. Figure 1 represents the histograms of all Q's for the favorable and unfavorable groups of combinations. Generally, the favorable Q's tend to be more on the negative side. Figure 2 shows the relationship between P_{maj} and the triple dependence Q_{123}.

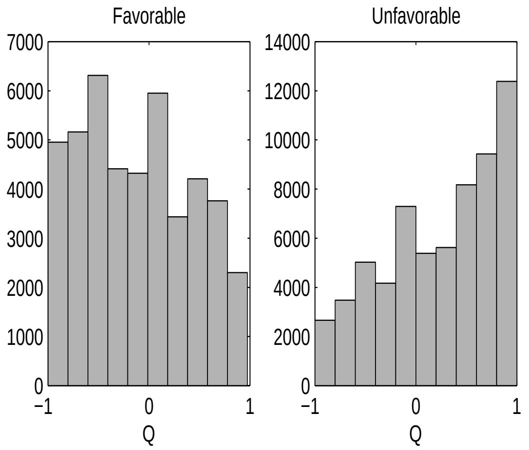

Figure 1. Histograms of the Q statistic for the "favorable" and "unfavorable" combinations of classifier outputs, $N = 100$.

The simulation was run for $L = 3$ classifiers (any number of classes c) with $N = 10$, 20, and 30 and with individual accuracy $p = 0.6$, 0.7, 0.8 and 0.9. Table 5 shows the minimal and the maximal accuracy P_{maj}.

Table 5. The minimal and the maximal accuracy of the majority vote P_{maj} of $L = 3$ classifiers of accuracy p with N objects

p	$N = 10$		$N = 20$		$N = 30$	
	P_{max}	P_{min}	P_{max}	P_{min}	P_{max}	P_{min}
0.6	0.9	0.40	0.9	0.40	0.9	0.40
0.7	1.0	0.60	1.0	0.55	1.0	0.56
0.8	1.0	0.75	1.0	0.70	1.0	0.70
0.9	1.0	0.90	1.0	0.85	1.0	0.86

To represent the overall dependence of a pool of classifiers we took the average and the maximum of the three Q's. Shown in Table 6 are dependence thresholds for $N = 10$, 20, and 30, and $p = 0.6$, 0.7, 0.8 and 0.9. The average and the maximal Q were retrieved for each pool $\mathcal{D}$ and the *maximal* value of Q was identified. All pools of classifiers whose pairwise dependences Q (the average or the maxi-

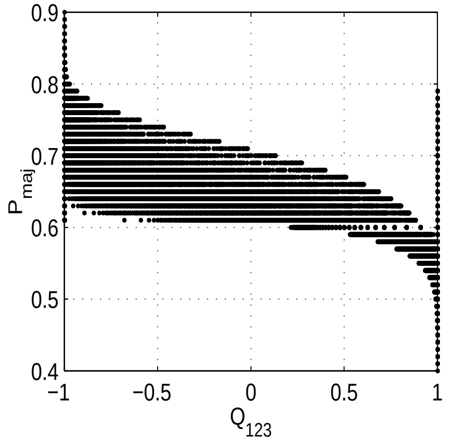

Figure 2. Relationship between the majority vote accuracy P_{maj} and triple dependence Q_{123} for 3 classifiers

mum) have been "more negative" than the threshold belong in the favorable group, i.e., they are **better than a pool of independent classifiers**.

The triple dependence Q_{123} did not appear to be as useful as might have been expected. Neither Table 4 nor Figure 2 demonstrate an unequivocal relationship between P_{maj} and Q_{123}.

Table 6. Threshold dependence values guaranteeing that the combination is "favorable"

p	$N = 10$		$N = 20$		$N = 30$	
	Q_{av}	Q_{max}	Q_{av}	Q_{max}	Q_{av}	Q_{max}
0.6	-0.375	-0.5	-0.375	-0.5	-0.43	-0.5
0.7	-0.63	-1.0	-0.45	-0.47	-0.52	-0.66
0.8	-1.0	-1.0	-0.6	-1.0	-0.7	-1.0
0.9	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0

4. Conclusions

By a synthetic example we explore the gain when combining dependent classifiers instead of independent ones. Our results support the intuition that negatively related classifiers are better than independent classifiers, and we also show that this relationship is ambivalent. In other words, if we want to be *guaranteed* a result better than the predicted accuracy for independent classifiers, we have to ensure that Q's are "sufficiently" negative (Table 6).

References

[1] A. Afifi and S. Azen. *Statistical Analysis. A Computer Oriented Approach.* Academic Press, N.Y., 1979.

[2] J. Benediktsson and P. Swain. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:688–704, 1992.

[3] H. Drucker, C. Cortes, L. Jackel, Y. LeCun, and V. Vapnik. Boosting and other ensemble methods. *Neural Computation*, 6:1289–1301, 1994.

[4] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[5] L. Kuncheva, J. Bezdek, and R. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 1999. (accepted).

[6] L. Lam and C. Suen. Optimal combination of pattern classifiers. *Pattern Recognition Letters*, 16:945–954, 1995.

[7] K.-C. Ng and B. Abramson. Consensus diagnosis: A simulation study. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:916–928, 1992.

[8] K. Woods, W. Kegelmeyer, and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410, 1997.

[9] L. Xu, A. Krzyzak, and C. Suen. Methods of combining multiple classifiers and their application to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:418–435, 1992.