In Proc. 7th International Conference on Information Processing and Management of Uncertainty (IPMU'98), Paris, France, 1998, 1890-1891.

On Combining Multiple Classifiers

Ludmila I. Kuncheva
School of Mathematics, University of Wales, Bangor
Bangor, Gwynedd LL57 1UT, United Kingdom
e-mail: L.I.Kuncheva@bangor.ac.uk

Abstract

We consider the classification accuracy of a combination of multiple classifiers. Based on a probabilistic framework, we suggest a multiplicative aggregation connective which we call *probabilistic product*. A case study with the `satimage` data from ELENA database shows that the probabilistic product aggregation is superior to some aggregation rules of the same type: majority vote, maximum, minimum, simple average, and product.

1 Introduction

Combination of multiple classifiers (CMC) is a two-level paradigm where the individual classifier outputs (first level) are combined to infer the final classification decision (second level), thereby aiming at a higher classification accuracy. In this paper we "borrow" an aggregation connective which has been originally developed for aggregation of expert probability assessments [1]. Putting it into a pattern recognition framework we show that this connective provides the (Bayes-) optimal classification decision when the individual (Bayes-optimal) classifiers use independent subsets of features.

Let $C = \{C_1, \ldots, C_L\}$ be the set of L classifiers and let $\Omega = \{1, 2, \ldots, M\}$ be the labels of a full group of M mutually exclusive classes. Let $X = \{X_1, \ldots, X_n\}$ be the set of features and let $\mathcal{S}^X$ denote the respective feature space. We search for a *classifier* $D : \mathcal{S}^X \to \Omega$. using some *aggregation rule* $\mathcal{F}$ that operates on the outputs of C_i, $i = 1, \ldots, L$. It is argued in [2] that the exact Bayesian approach to find $\mathcal{F}$ is very difficult to apply because dependencies between the experts (first-level classifiers here) have to be modeled. Therefore many authors consider simple heuristical aggregation connectives. In cases where favorable assumptions may hold more rigorous approaches can be used. Our experiments show that even when the assumptions may not hold the rigorous approach still gives good results.

2 Probabilistic product

We study an aggregation rule that comes from expert opinion fusion [1]. This rule performs classifier fusion using continuous-valued classifier outputs. Let $d_{i,j}(x)$ be the degree of "support" given by classifier C_i to the hypothesis that x comes from class j. Typically, $d_{i,j}(x) \in [0, 1]$. Besides the usual interpretation as an estimate of the posterior probability $P(j|x)$ by classifier C_i, $d_{i,j}(x)$ can be viewed as typicalness, belief, certainty, possibility, etc., not necessarily coming from statistical classifiers.

Let us assume that the set of features can be partitioned into L nonoverlapping subsets $\{X^{(1)}, \ldots, X^{(L)}\}$, which are conditionally independent, i.e.,

$$p(x|j) = p([\mathbf{x}_1 \ldots \mathbf{x}_L]^T|j) = \prod_{i=1}^{L} p(\mathbf{x}_i|j),$$

$$\mathbf{x}_i \in \mathcal{S}^{X^{(i)}}, \quad j = 1, \ldots, M.$$

where $p(.|j)$ are the class-conditional probability density functions and $\mathcal{S}^{X^{(i)}}$ is the feature space generated by the feature set $X^{(i)}$.

Let $P(j)$ denote the prior probability for class j. If all classifiers produce $d_{i,j}(x) = P(\mathbf{x}_i|j)$ (the true posterior probability for class j on the feature set $X^{(i)}$), the true overall probability is

$$P(j|x) = P(j|[\mathbf{x}_1 \ldots \mathbf{x}_L]^T) = \frac{P(j)\, p(x|j)}{p(x)} =$$

$$= \frac{\prod_{i=1}^{L} p(\mathbf{x}_i)}{p(x)} \; \frac{\prod_{i=1}^{L} P(j|\mathbf{x}_i)}{P(j)^{L-1}}.$$

Noticing that the first multiplier does not depend on the class label, we can design a set of discriminant functions proportional to the true posterior probabilities

$$\mathcal{G}_j(x) = \frac{\prod_{i=1}^{L} d_{i,j}(x)}{P(j)^{L-1}}, \quad j = 1, \ldots, M. \qquad (1)$$

The the (Bayes-) optimal class label assigned to x is that of the highest $\mathcal{G}_j(x)$, $j = 1, \ldots, M$.

In the experiments we used the estimates of $P(j)$ on the training sample as

$\hat{P}(j) = \frac{N_j}{N}$, $j = 1, \ldots, M$, where N_j is the number of elements of the training set from class j and N is the total training sample size.

3 Experimental illustration

To illustrate the proposed aggregation rule we used the `satimage` data from ELENA database (anonymous ftp at `ftp.dice.ucl.ac.be`, directory `pub/neural-nets/ELENA/databases`). The `satimage` data consists of 6435 patterns (pixels) with 36 attributes in 6 classes. We applied linear and quadratic classifiers as the first-level classifiers setting up the following four designs: 18 2-feature classifiers (features were grouped as $\{(1,2), (3,4), \ldots, (35,36)\}$, we conducted experiments with all 18 classifiers, and with the best 3,5,7, and 9 classifiers *ranked individually*); 12 3-feature classifiers; 9 4-feature classifiers; and 6 6-feature classifiers. With each design we performed one set of experiments with LDC and one set with QDC. The dataset was split into two equal parts used in turn for training and testing. We compared the *probabilistic product* ((1) and (2)) with the majority vote, maximum, minimum, (unweighted) average, and (unweighted) product. The results are shown in Table 1 for LDC and Table 2 for QDC. The entries in the table are the estimates of classification accuracy (in %) averaged over the two parts used as testing sets. We chose to show the test accuracy corresponding to the *highest training* result. The "winning" CMC are typed in boldface for each experimental setting.

4 Conclusions

The results lead us to the following conclusions:

1. The increase in classification accuracy of CMC compared to the best single classifier is not high. The major reason is that the individual classifiers are not independent.
2. Because of the large data sets (both for training and test) and the favorable class structure, the correlation between training and test accuracy is high. The benefit is twofold: (i) the CMC chosen on the training results performed well on the test set; (ii) small differences in classification accuracy can be found significant.
3. In 5 out of 8 experiments the *probabilistic product* aggregation was superior to the other aggregations. In the other 3 experiments it was the second best after the simple average which is traditionally good and robust.
4. The *probabilistic product* worked well even though the feature sets might have been dependent.

Table 1.: Classification accuracy (in %) with LDC

Number of features	2	3	4	6
Single best	79.60	81.76	81.98	82.33
Majority	80.54	82.38	83.39	83.31
Maximum	79.84	81.78	83.48	83.17
Minimum	79.28	80.74	82.33	82.11
Average	80.66	**83.24**	83.71	**84.06**
Product	79.96	82.10	82.98	82.92
Probabilistic product	**81.94**	82.81	**83.88**	83.40

Table 2.: Classification accuracy (in %) with QDC

Number of features	2	3	4	6
Single best	81.10	84.30	84.76	84.94
Majority	81.96	84.91	85.18	85.58
Maximum	80.88	84.56	85.18	85.58
Minimum	81.63	84.30	85.00	85.51
Average	82.56	85.13	**85.96**	86.18
Product	83.88	84.94	85.60	86.17
Probabilistic product	**84.92**	**85.78**	85.92	**86.70**

References

[1] R.F. Bordley. A multiplicative formula for aggregating probability assessments. *Management Science*, 28:1137–1148, 1982.

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