Special Session on Random Forest and Ensembles

Maite Termenón[1]

[1]Computational Intelligence Group

2012 January 27

Article to Present

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 2, FEBRUARY 2002

A Theoretical Study on Six Classifier Fusion Strategies

Ludmila I. Kuncheva, *Member, IEEE*

Summary

- Two things to demostrate:
    - Selection of fusion methods is very important, as important as selection of classifiers on the ensemble.
    - Assumptions and decisions of a previous work are not correct.
- They propose six fusion methods and calculate the theoretical probability of error for each, depending on:
    - followed distribution (normal and uniform),
    - true posterior probability,
    - number of classifiers, L.
- They reproduce the experiment of the previous work and compare the results.

Outline

Outline

Introduction

Frame subtitles are optional. Use upper- or lowercase letters.

- Let $D = \{D_1, \ldots, D_L\}$ be a set of classifiers.
- By combining the individual output, we aim at a higher accuracy than that of the best classifiers.
- This study is inspired by a publication by Alkoot and Kittler [1], where classifier fusion methods are experimentally compared.

Assumptions

1. All classifiers produce soft class labels. We assume that $d_{j,i}(\mathbf{x}) \in [0,1]$ is an estimate of the posterior probability $P(\omega_i|\mathbf{x})$ offered by classifier D_j for an input $\mathbf{x} \in \Re^n$, $i = 1,2, \;\; j = 1,\ldots,L$.
2. There are two possible classes $\Omega = \{\omega_1, \omega_2\}$. We consider the case where, for any $\mathbf{x}$, $d_{j,1}(\mathbf{x}) + d_{j,2}(\mathbf{x}) = 1$, $j = 1,\ldots,L$.
3. A single point $\mathbf{x} \in \Re^n$ is considered and the true posterior probability is $P(\omega_1|\mathbf{x}) = p > 0.5$. Thus, the Bayes-optimal class label for $\mathbf{x}$ is ω_1 and a classification error occurs if label ω_2 is assigned.
4. The classifiers commit independent and identically distributed errors in estimating $P(\omega_1|\mathbf{x})$.

Two distribution

- Two distributions of $d_{j,1}(\mathbf{x})$ are discussed:
  - Normal distribution: $N\left(p,\sigma^2\right)$, $\sigma \in [0.1,1]$
  - Uniform distribution spanning the interval $[p-b,p+b]$, $b \in [0.1,1]$

Fusion methods

- The support for class w_i, $d_i(x)$, yielded by the team is:

$$d_i(\mathbf{x}) = \mathcal{F}(d_{1,i}(\mathbf{x}), \ldots, d_{L,i}(\mathbf{x})), \quad i = 1, 2, \qquad (1)$$

where $\mathscr{F}$ is the chosen fusion method.

- Fusion Methods: minimum, maximum, average, median, mayority vote and oracle.

Mayority Vote

- We first harden individual decisions by assigning class labels:
  - $D_j(\mathbf{x}) = w_1$ if $d_{j,1}(\mathbf{x}) > 0.5$
  - $D_j(\mathbf{x}) = w_2$ if $d_{j,1}(\mathbf{x}) \leq 0.5$
  - $j = 1, \ldots, L$
- Class label most represented among the L (*label*) outputs is chosen.

Oracle

- It is an abstract fusion model.
- If at least one of the classifiers produces the correct class label, then the team produces the correct class label too.
- Usually used in comparative experiments.

To demostrate

- Consensous among researchers:
  - The major factor for a better accuracy is the diversity in the classifier team.
  - So, fusion method is of a secondary importance.
- However, a choice of an appropiate fusion method can improve further on the performance of the classifier.

Outline

Definitions

- Denote P_j the output classifier D_j for class w_1 and let

$$\hat{P}_1 = \mathcal{F}(P_1, \ldots, P_L) \tag{2}$$

be the fused estimate of $P(w_1 \mid \mathbf{x})$.

- And so,

$$\hat{P}_2 = \mathcal{F}(1 - P_1, \ldots, 1 - P_L). \tag{3}$$

Definitions

- Individual estimates P_j are i.i.d random variables, such $P_j = p + \varepsilon_j$, with:
  - Probability Density Function (pdf): $f(y), y \in \Re$
  - Cumulative Distribution Function (cdf): $F(t), t \in \Re$
- Then $\hat{P}_1$ is a random variable too with pdf $f_{\hat{P}_1}(y)$ and cdf $F_{\hat{P}_1}(t)$.

Probability of Error (I)

- For single classifier, average and median: $\hat{P}_1 + \hat{P}_2 = 1$
- For oracle and majority vote:
  - $\hat{P}_1 = 1, \hat{P}_2 = 0$ if class w_1 is assigned and viceversa.
- Probability of error:

$$P_e = P(\text{error}|\mathbf{x}) = P(\hat{P}_1 \leq 0.5) = F_{\hat{P}_1}(0.5) = \int_0^{0.5} f_{\hat{P}_1}(y)dy \qquad (4)$$

for the single best classifier, average, median, majority vote and oracle.

Probability of Error (II)

- For the minimun and maximum rules, class label is decided by the maximum of $\hat{P}_1$ and $\hat{P}_2$.
- An error will occur if $\hat{P}_1 \leq \hat{P}_2$:

$$P_e = P(\text{error}|\mathbf{x}) = P(\hat{P}_1 \leq \hat{P}_2) \tag{5}$$

for the minimum and maximum.

Normal Distribution

- $N\left(p,\sigma^2\right)$. We denote by $\Phi\left(z\right)$ the cdf of $N\left(0,1\right)$.
- Thus:

$$F(t) = \Phi\left(\frac{t-p}{\sigma}\right). \tag{6}$$

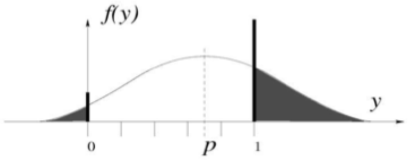
Uniform Distribution

- Uniform distribution within $[p-b, p+b]$:

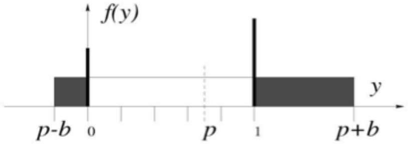
$$f(y) = \begin{cases} \frac{1}{2b}, & y \in [p-b, p+b]; \\ 0, & \text{elsewhere}, \end{cases}$$
$$F(t) = \begin{cases} 0, & t \in (-\infty, p-b); \\ \frac{t-p+b}{2b}, & t \in [p-b, p+b]; \\ 1, & t > p+b. \end{cases} \tag{7}$$

Considerations

- In [1], distributtions are clipped, so all P_js were in [0,1].



(a)



(b)

Considerations

- A theoretical analysis with clipped distribution is not straightforward.
- The clipped distributions are actually mixtures of a continuous random variable in the interval $(0, 1)$ and a discrete one taking values 0 or 1.
- In this theoretical analysis, distributions are not clipped.

Error for Single Classifier

- Normal distribution:

$$P_e = \Phi\left(\frac{0.5 - p}{\sigma}\right), \tag{8}$$

- Uniform distribution:

$$P_e = \frac{0.5 - p + b}{2b}. \tag{9}$$

Introduction

- They are identical for $c = 2$ and any number of classifiers L.
- Substituting $\mathscr{F} = max$ in (2):
  - Team's support for w_1 is $\hat{P}_1 = max_j \{P_j\}$
  - support for w_2 is $\hat{P}_2 = max_j \{1 - P_j\}$

Classification error

- If:

$$\max_j\{P_j\} < \max_j\{1-P_j\}, \quad (10)$$

$$p + \max_j\{\epsilon_j\} < 1 - p - \min_j\{\epsilon_j\}, \quad (11)$$

$$\epsilon_{\text{max}} + \epsilon_{\text{min}} < 1 - 2p. \quad (12)$$

- For the minimum fusion method:

$$\min_j\{P_j\} < \min_j\{1-P_j\}, \quad (13)$$

$$p + \epsilon_{\text{min}} < 1 - p - \epsilon_{\text{max}}, \quad (14)$$

$$\epsilon_{\text{max}} + \epsilon_{\text{min}} < 1 - 2p, \quad (15)$$

Probability of error

- The probability of error for minimum and maximum is:

$$P_e = P(\epsilon_{\max} + \epsilon_{\min} < 1 - 2p) \quad (16)$$
$$= F_{\epsilon_s}(1 - 2p), \quad (17)$$

Normally distributed Pjs

- ε_j are also normally distributed with mean 0 and variance σ^2.
- We cannot:
    - assume that ε_{max} and ε_{min} are independent.
    - analyze their sum as a distributed variable.
- There are *order statistics* and $\varepsilon_{min} \leq \varepsilon_{max}$.
- So, we have not attempted a solution for the normal distribution case.

Uniform Distributions Pjs

- Taken from [8], where the pdf of midrange $(\varepsilon_{min} + \varepsilon_{max})/2$ is calculated for L observations.
- We derived $F_{\varepsilon_s}(t)$ to be:

$$F_{\epsilon_s}(t) = \begin{cases} \frac{1}{2}\left(\frac{t}{2b}+1\right)^L, & t \in [-2b, 0]; \\ 1 - \frac{1}{2}\left(1-\frac{t}{2b}\right)^L, & t \in [0, 2b]. \end{cases} \quad (18)$$

Noting that $t = 1 - 2p$ is always negative,

$$P_e = F_{\epsilon_s}(1-2p) = \frac{1}{2}\left(\frac{1-2p}{2b}+1\right)^L. \quad (19)$$

Normal probability error for Average

- Average fusion method gives $\hat{P}_1 = \frac{1}{L}\sum_{j=1}^{L} P_j$.
- If $P_1, \ldots, P_L$ are normally distributed and independent then $\hat{P} \sim N\left(p, \frac{\sigma}{L}\right)$
- Probability of error is:

$$P_e = P(\hat{P}_1 < 0.5) = \Phi\left(\frac{\sqrt{L}(0.5 - p)}{\sigma}\right). \qquad (20)$$

Uniform probability error for Average

- Assumption: the sum of L independent variables is a variable of approximately normal distribution.
- The higher the L, the more accurate the approximation.
- Knowing the varaince of uniform distribution for $P_j = \frac{b^2}{3}$, we can assume $\hat{P} \sim N\left(p, \frac{b^2}{3L}\right)$.
- Probability of error is:

$$P_e = P(\hat{P}_1 < 0.5) = \Phi\left(\frac{\sqrt{3L}(0.5 - p)}{b}\right). \qquad (21)$$

Median and Majority Vote

- We restrict our choice of L to odd numbers only.
- For the median fusion method:

$$\hat{P}_1 = \text{med}\{P_1, \ldots, P_L\} = p + \text{med}\{\epsilon_1, \ldots, \epsilon_L\} = p + \epsilon_m. \qquad (22)$$

- Then, the probability of error is:

$$P_e = P(p + \epsilon_m < 0.5) = P(\epsilon_m < 0.5 - p) = F_{\epsilon_m}(0.5 - p), \qquad (23)$$

where F_{ϵ_m} is the cdf of ϵ_m.

Median and Majority Vote

- From the order stadistics theory [8]:

$$F_{\epsilon_m}(t) = \sum_{j=\frac{L+1}{2}}^{L} \binom{L}{j} F_\epsilon(t)^j [1 - F_\epsilon(t)]^{L-j}, \qquad (24)$$

where $F_\epsilon(t)$ is the distribution of ϵ_j, i.e., $N(0, \sigma^2)$ or uniform in $[-b, b]$.

Error for Median and Majority Vote

- Normal distribution:

$$P_e = \sum_{j=\frac{L+1}{2}}^{L} \binom{L}{j} \Phi\left(\frac{0.5-p}{\sigma}\right)^j \left[1-\Phi\left(\frac{0.5-p}{\sigma}\right)\right]^{L-j}. \quad (25)$$

- Uniform distribution:

$$P_e = \begin{cases} 0, & p-b > 0.5; \\ \sum_{j=\frac{L+1}{2}}^{L} \binom{L}{j} \left(\frac{0.5-p+b}{2b}\right)^j \left[1-\frac{0.5-p+b}{2b}\right]^{L-j}, & \text{otherwise.} \end{cases} \quad (26)$$

Probability Error for Oracle

- The probability of error for the oracle is:

$$P_e = P(\text{all incorrect}) = F(0.5)^L \qquad (28)$$

- Normal distribution:

$$P_e = \Phi\left(\frac{0.5 - p}{\sigma}\right)^L, \qquad (29)$$

- Uniform distribution:

$$P_e = \begin{cases} 0, & p - b > 0.5; \\ \left(\frac{0.5-p+b}{2b}\right)^L, & \text{otherwise.} \end{cases} \qquad (30)$$

Outline

Description

- Reproduction of part of the experiments from [1].
- Two figures:
  - Normally distributed P_js.
  - Uniformly distributed P_js.

Results Normal Distribution

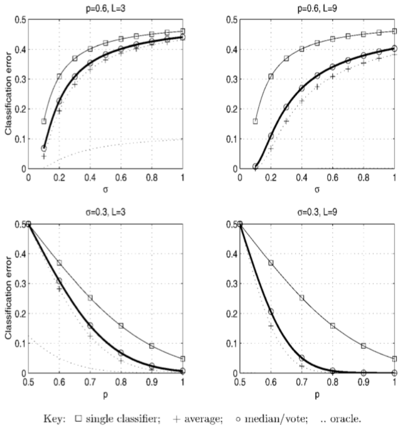


Figure: P_e for normally distributed P_js.

Results Uniform Distribution

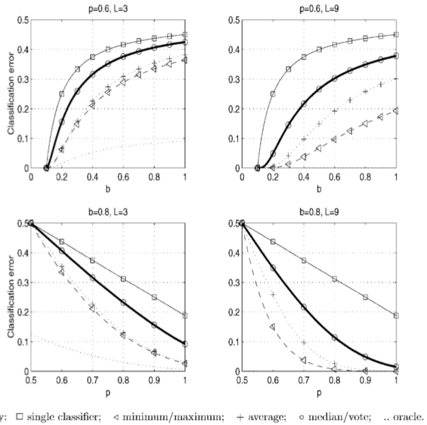


Figure: P_e for uniformly distributed P_js.

Findings in results

- Individual error is higher that the error of any fusion methods.
- Oracle model is the best of all.
- The more classifiers, the lower the error.

More Interesting Findings

- Average and median/vote have same performance for normally distributed (aprox), but different for uniform distribution (average is better).
- Average method is outperformed by minimum/maximum method, contrary to findings in literature.

Comparing to [1]

- Results are different.
- They found a threshold for b where min, max and product change from the best to worst fusion methods.
- Discrepancy can be attributed to clipped-distribution effect.
- This study uses $L = 9$ classifiers instead of $L = 8$ to avoid ties.
- For small values of b and σ, the sets of results are similar.

Outline

Summary

- Six simple classifier methods have been studied theoretically.
- We give formulas for classification error at a single point in the feature space, $\mathbf{x} \in \Re^n$.
- Conditions:
  - Two classes $\{w_1, w_2\}$.
  - Each classifier gives an output P_j as an estimate of the posterior probability $P(w_1 \mid \mathbf{x}) = p > 0.5$.
  - P_j are i.i.d coming from a fixed distribution with mean p.

Multiclass

- For c classes:
  - It is not enough that $P(w_1 \mid \mathbf{x}) > \frac{1}{c}$ for a correct classification
  - Only $P_1, \ldots, P_L$ are not enough, we also need to specify conditions for the support for other classes.
  - $P(w_1 \mid \mathbf{x}) > 0.5$ is sufficient but not necessary for a correct classification.
  - True classification error can only be smaller than P_e.

Conclusions

- It is claimed in the literature that combination methods are less important than the diversity of the team.
  - Normally distributed errors: fusion methods gave very similar performance, but,
  - Uniformly distributed error: methods differed significantly, especially for higher L.
- So, **combination methods are also relevant in combining classifiers**.

Limitations

- The most restrictive and admittedly unrealistic assumption is the independence of the estimates.
- It is recognized that "independently built" classifiers exhibit positive correlation, due to that difficult parts of the feature space are difficult for all classifiers.
- Ensemble design methods (ADAboost), try to overcome this unfavorable dependency by enforcing diversity.
- However, it is difficult to measure or express this diversity in a mathematically tractable way.

My opinion

- It is an interesting theorical paper.
- It remarks the idea of when using ensembles, combination methods selection is very important.
- Approach is based on a previous work and classification experiments improve previous results.
  - If some developments are too complicated, they argument why is complicated and do not calculate it.
  - One fusioon method does not fin with the teoretical framework, and they change it for other one.
  - They change the number of classifiers on the experiment to avoid complications.

References I

- [1] F. Alkoot and J. Kittler, “Experimental Evaluation of Expert Fusion Strategies,” Pattern Recognition Letters, vol. 20, pp. 1361–1369, 1999.
- [8] A. Mood, F. Graybill, and D. Boes, Introduction to the Theory of Statistics, third ed. McGraw-Hill, 1974.