

Some questions on ranking of classifiers in the presence of label noise

A. E. Koudou, B. Lamiroy, N. Villa-Vialaneix

Workshop Labelnoise
Nancy, November 29, 2017

Outline

Problem setting

Related work

Some directions of research

1. Problem setting

- ▶ Suppose we are given n objects that belong to one of the two groups $\{0, 1\}$.
- ▶ The group membership of object i is further denoted by y_i but is not actually observed.
- ▶ Instead of that, we are given observations $(\tilde{y}_i)_{i=1, \dots, n}$, for which we can suppose that the probability that \tilde{y}_i is equal to y_i is known and equal to p_0 .

- ▶ Further, we are given m algorithms, $(A_j)_{j=1,\dots,m}$ that produce classifications of the objects: $A_j(i) = y_i^j$.
- ▶ These algorithms are ranked according to the misclassification rate with respect to the observations \tilde{y}_i :

$$\text{rank}(A_j) < \text{rank}(A_{j'}) \Leftrightarrow \sum_{i=1}^n \mathbf{1}_{\{y_i^j \neq \tilde{y}_i\}} < \sum_{i=1}^n \mathbf{1}_{\{y_i^{j'} \neq \tilde{y}_i\}}.$$

- ▶ These ranks are further denoted by $(\tilde{r}_j)_{j=1,\dots,m}$.
- ▶ However, they do not correspond to the ranks $(r_j)_{j=1,\dots,m}$ that could have been obtained using the true probabilities that the algorithms A_j produce a label y_i^j equal to y_i .

- ▶ Denote by p_j the probability that the classification given by A_j is equal to the true class label.
- ▶ The general question is to know what is the influence of p_0 on the differences between $(r_j)_{j=1,\dots,m}$ and $(\tilde{r}_j)_{j=1,\dots,m}$.

Specific questions could be:

- ▶ are we able to know that if p_0 is small enough, the ranking is (mostly) unchanged?
- ▶ are we able to propose a test that would say that, given p_0 , which algorithms are not significantly different?

2. Related work

The following conclusions can be derived from Raj *et al* (2011)

- ▶ the probability that A_j agrees with the observed class label, \tilde{y}_i , is, under independence assumption between classifiers errors and labels errors,

$$\begin{aligned}t_j &= P(y_i^j = \tilde{y}_i) \\&= P(y_i^j = \tilde{y}_i; \tilde{y}_i = y_i) + P(y_i^j = \tilde{y}_i; \tilde{y}_i \neq y_i) \\&= P(y_i^j = y_i; \tilde{y}_i = y_i) + P(y_i^j \neq y_i; \tilde{y}_i \neq y_i) \\&= p_0 p_j + (1 - p_0)(1 - p_j) \\&= (2p_0 - 1)p_j + 1 - p_0.\end{aligned}$$

- ▶ Thus, if $p_0 > 0.5$, then $t_j > t_{j'} \Leftrightarrow p_j > p_{j'}$ meaning that an algorithm has a better performance than another one with respect to the observed labels if and only if it is the case with respect to the true labels.
- ▶ t_j can be estimated by $\tilde{t}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{y_i^j = \tilde{y}_i\}}$.
- ▶ A test is derived for H_0 : “ $p_j = p_{j'}$ ” against H_1 “ $p_j \neq p_{j'}$ ”.
- ▶ It is simply based on the binomial distribution variance and the fact that, asymptotically, $\frac{\tilde{t}_j - \tilde{t}_{j'}}{\sqrt{\text{Var}(\tilde{t}_j) + \text{Var}(\tilde{t}_{j'})}}$ follows a normal distribution.
- ▶ p_0 is not even needed to perform this test and can be worse than p_j .

- ▶ Question : comparing algorithms (A_j) by using the values of $(t_j)_j$, can one compute the probability of change of ranking, i.e. the probability that $\tilde{t}_j < \tilde{t}_{j'}$ when $p_j > p_{j'}$?
- ▶ This question is very close to the one addressed by Lamiroy and Pierrot, where they used an approach based on combinatorics to derive the probability of change in ranking and validated their result by simulations.

The Lam and Stork (2003) approach

Under the assumption of independence of the error of the classifier and the error on the class label, Lam and Stork expressed the "true error rate" of the algorithm A_j as

$$P(y_i \neq y_i^j) = \frac{P(y_i^j \neq \tilde{y}_i) - P(y \neq \tilde{y}_i)}{1 - 2P(y \neq \tilde{y}_i)}.$$

They gave some bounds on this expression in a special case of non independence.

Some directions of research

- ▶ The approach of Raj *et al* (2011) is based on the independence of the error of the classifier and of the error on the class label.
- ▶ However, this setting does not seem very realistic: it is expected that classifiers are more often mistaken on the objects that have error on class labels more often. This setting can be formalized using the conditional probabilities:

$$\mathbb{P}(y_i^j = y_i | \tilde{y}_i = y_i) = p_j^T \quad \text{and} \quad \mathbb{P}(y_i^j = y_i | \tilde{y}_i \neq y_i) = p_j^F$$

with $p_j^T \geq p_j$ and $p_j^F \leq p_j$
(the independence case is $p_j = p_j^F = p_j^T$).

- ▶ Note also that

$$p_j = p_j^T p_0 + p_j^F (1 - p_0)$$

(which, in particular, implies that $p_j^T = p_j \Leftrightarrow p_j^F = p_j$).

- ▶ We have

$$\begin{aligned}
 t_j &= p_0 p_j^T + (1 - p_0)(1 - p_j^F) \\
 &= p_j p_0 + (1 - p_j)(1 - p_0) + \underbrace{r_j p_0 + s_j (1 - p_0)}_{\text{divergence to the independence case}}
 \end{aligned}$$

with $p_j^T = p_j + r_j$ and $p_j^F = p_j - s_j$.

- ▶ Let us now suppose that A_j and $A_{j'}$ compare such that $\text{rank}(A_j) \leq \text{rank}(A_{j'})$, which is equivalent to $p_j \geq p_{j'}$.
- ▶ Then, we would like to know when the ranks as given by the observed labels are indicative of this relative performance. Similarly as in Raj, we suppose that the observed ranks are reasonable, meaning that $p_0 \geq 0.5$. With this condition,

$$p_j p_0 + (1 - p_j)(1 - p_0) \geq p_{j'} p_0 + (1 - p_{j'})(1 - p_0)$$

and the ranks are preserved on the observed labels if and only if

$$\begin{aligned} t_j \geq t_{j'} &\Leftrightarrow r_j p_0 + s_j(1 - p_0) \geq r_{j'} p_0 + s_{j'}(1 - p_0) \\ &\Leftrightarrow r_j - r_{j'} \geq \frac{1 - p_0}{p_0} (s_j - s_{j'}) \end{aligned}$$

or, equivalently, if and only if

$$p_j^T - p_{j'}^T \geq \frac{1 - p_0}{p_0} (p_j^F - p_{j'}^F).$$

Estimation

t_j and $t_{j'}$ are not really observed. Only,




$$\tilde{t}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{y_i^j = \tilde{y}_i\}}$$

can be observed. However, if the observations are i.i.d., $n\tilde{t}_j$ follows a binomial distribution $\mathcal{B}(n, t_j)$.

Some questions related to the Lamiroy-Pierrot approach :

- ▶ Other measures of dissimilarity than the Kullback-Leibler divergence, such as total variation distance?
- ▶ Extension to more than two algorithms ?

References

-  C. P. Lam and D. G. Stork (2003). Evaluating classifiers by means of test data with noisy labels. In *Proceedings of IJCAI*.
-  B. Lamiroy and P. Pierrot (2016). Statistical performance metrics for use with imprecise ground truth. *11th International Workshop on Graphics recognition*.
-  B. Raj, R. Singh and J. Baker (2011). A paired test for recognizer selection with untranscribed data. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 5676-5679. IEEE Signal Processing Society.