Discussion on “Random-projection ensemble classification”

Meimei Liu and Guang Cheng

We congratulate Cannings and Samworth for an inspiring piece of work. Accuracy and stability are two main principles in designing classification algorithms; see [3]. This short note empirically examines how these two measures are affected by the choice of \((B_1, B_2, d)\), which in turn determines the computational cost of the proposed method in the paper.

According to [2], one (statistically meaningful) way to define instability for a classification procedure \(\Phi\) is

\[
\text{CIS}(\Phi) = \mathbb{E}_{D_1, D_2}[d(\hat{\phi}_{n1}, \hat{\phi}_{n2})]
\]

where \(d(\hat{\phi}_{n1}, \hat{\phi}_{n2}) = \mathbb{P}_X(\hat{\phi}_{n1}(X) \neq \hat{\phi}_{n2}(X))\) and \(\hat{\phi}_{ni} = \Phi(D_i)\) is the classifier trained based on the sample \(D_i\) for \(i = 1, 2\), which is drawn from the same population.

In Figure 1, we fix \(d = 5\) and study how misclassification rate and CIS are affected by different combinations of \((B_1, B_2)\). Figure 1 (a) shows that once \(B_1\) is large enough the misclassification rate will not change too much as \(B_2\) grows. However, the pattern of misclassification rates is roughly the same as \(B_1\) grows under different choices of \(B_2\). This might indicate that \(B_1\) plays a more prominent role than \(B_2\) in determining the misclassification rate. By examining Figure 1 (b) on CIS in a similar way, we find that the roles of \(B_1\) and \(B_2\) are more comparable, though. We further investigate the least computational cost needed to achieve the best accuracy and stability. In Figure 1 (a), three dots, denoted as \((B_1^*, B_2^*)\), are found to have the smallest value of \(B_1 \times B_2\), i.e., 275, among all combinations of \((B_1, B_2)\) leading to the smallest misclassification rate. In contrast in Figure 1 (b) for CIS, we need a higher computational budget, i.e., \(B_1^{**} \times B_2^{**} = 300\), to obtain the best stability.

However, the projection dimension \(d\) is another factor in determining the computational cost. Hence, in Figure 2, we fix the total number of random projections, i.e., \(B_1 \times B_2\), while varying \(d\). Figure 2 (a) shows that an increase of \(B_1\) leads to a smaller misclassification rate, but this improvement is no longer obvious when \(B_1\) is sufficiently large. However, the pattern for CIS in Figure 2 (b) is not that clear. Another

---

1Ph.D student, Department of Statistics, Purdue University, West Lafayette, IN 47906. E-mail: liu1197@purdue.edu. Research Sponsored by NSF CAREER Award DMS-1151692

2Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906. E-mail: chengg@purdue.edu. Research Sponsored by NSF CAREER Award DMS-1151692, DMS-1418042, and ONR N00014-15-1-2331.
phenomenon in Figure 2 is that misclassification rate and CIS in all curves cannot be further improved as $d$ grows beyond some critical point $d^*$ and $d^{**}$ respectively. Since $B_1 \times B_2$ are fixed in all curves, a sharp lower bound of $d$ might be viewed as the computational limit of the proposed algorithm from a statistical perspective.

All these empirical observations require new theoretical understanding on high dimensional classification problems from the perspective of computational cost.

References


Figure 1: Heat map of (a) misclassification rate \((d = 5)\) and (b) CIS \((d = 5)\) under various values of \((B_1, B_2)\): the training data set of size 200 and the testing data set of size 1000 were generated following Section 6.1.2 of the paper; the \(k\)NN classifier is considered here; CIS is calculated by averaging the disagreement of two classifiers on the testing data with 100 replications.
Figure 2: Effect of projection dimension on (a) the misclassification rate and (b) CIS ($d$ ranges from 2 to 20, $n = 200$ and $B_1 \times B_2 = 200$; the $k$NN classifier is considered here): $\circ, B_1 = 20, B_2 = 10$; $\triangle, B_1 = 40, B_2 = 5$; $+, B_1 = 50, B_2 = 4$; $\times, B_1 = 100, B_2 = 2$. 

4