Analysis of Ensemble Learning Using Simple Perceptrons Based on Online Learning Theory

Seiji MIYOSHI,1,*) Kazuyuki HARA2 and Masato OKADA3,4,5

1Kobe City College of Technology, Kobe 651-2194, Japan
2Tokyo Metropolitan College of Technology, Tokyo 140-0011, Japan
3Graduate School of Frontier Sciences, The University of Tokyo,
Kashiwa 277-8561, Japan
4RIKEN Brain Science Institute, Wako 351-0198, Japan
5Intelligent Cooperation and Control, PRESTO, JST

Ensemble learning of $K$ simple perceptrons, which determine their outputs by sign functions, is discussed within the framework of online learning and statistical mechanics. Hebbian, perceptron and AdaTron learning show different characteristics in their affinity for ensemble learning, that is "maintaining variety among students". Results show that AdaTron learning is superior to the other two rules.

§1. Introduction

Ensemble learning1)–6) means to combine many rules or learning machines (students in the following) that perform poorly. Theoretical studies analyzing the generalization performance by using statistical mechanics7)–8) have been performed vigorously.4)–6) Hebbian learning, perceptron learning and AdaTron learning are well-known as learning rules for a simple perceptron.9)–12) Determining differences among ensemble learnings with these three learning rules is a very attractive problem. We discuss ensemble learning of $K$ simple perceptrons within the framework of online learning and finite $K$.13)–14)

§2. Model

Each student treated in this paper is a perceptron. An ensemble of $K$ students is considered. Connection weights of students are $J_1, J_2, \ldots, J_K$. $J$ and input $x$ are $N$ dimensional vectors. Each component $x_i$ of $x$ is assumed to be an independent random variable that obeys the Gaussian distribution $N(0,1/N)$. Each component of $J_k^0$, that is the initial value of $J_k$, is assumed to be generated according to $N(0,1)$ independently. Each student's output is $\text{sgn}(u_1l_1), \text{sgn}(u_2l_2), \ldots, \text{sgn}(u_Kl_K)$ where $u_kl_k = J_k^Tx$. In this paper, $u_k$ is called a normalized internal potential of a student.

The teacher is also perceptron. The teacher's connection weight is $B$. In this paper, $B$ is assumed to be fixed where $B$ is also an $N$ dimensional vector. Each component $B_i$ is assumed to be generated according to $N(0,1)$ independently. The teacher's output is $\text{sgn}(v)$ where $v = B^Tx$. Here, $v$ represents an internal potential of the teacher.

*) E-mail: miyoshi@kobe-kosen.ac.jp
In this paper the thermodynamic limit $N \to \infty$ is also treated. Therefore, $|x| = 1$, $|B| = |J_k^0| = \sqrt{N}$. Generally, a norm of student $|J_k|$ changes as the time step proceeds. Therefore, the ratio $l_k$ of the norm to $\sqrt{N}$ is considered and is called a length of student $J_k$. That is, $|J_k| = l_k \sqrt{N}$. The common input $x$ is presented to the teacher and all students in the same order. Within the framework of online learning, the update can be expressed as $J_k^{m+1} = J_k^m + f_k^m x^m$, where $m$ denotes time step and $f_k^m = f(\text{sgn}(v^m), u_k^m)$ is a function determined by learning rule.

In this paper, two methods are treated to determine an ensemble output. One is the majority vote of $K$ students. Another method is adopting an output of a new perceptron whose connection weight is the mean of the weights of $K$ students. This method is simply called the weight mean in this paper.

§3. Theory

In this paper, the majority vote and the weight mean are treated to determine an ensemble output. We use $\epsilon = \Theta(-\text{sgn}(v) \sum_{k=1}^{K} \text{sgn}(u_k))$ and $\epsilon = \Theta(-\text{sgn}(v) \cdot \text{sgn}(\sum_{k=1}^{K} u_k))$ as error $\epsilon$ for the majority vote and the weight mean, respectively. Here, $\Theta(\cdot)$ is the step function. Generalization error $\epsilon_g$ is defined as the average of error $\epsilon$ over the probability distribution $p(x)$ of input $x$. Therefore, the generalization error $\epsilon_g$ can be also described as

$$\epsilon_g = \int dx p(x) \epsilon = \int \prod_{k=1}^{K} du_k dv p(u_k, v) \epsilon(u_k, v),$$

by using the probability distribution $p(u_k, v)$ of $u_k$ and $v$. As the thermodynamic limit $N \to \infty$ is also considered in this paper, $u_k$ and $v$ obey the multiple Gaussian distribution based on the central limit theorem. All diagonal components of the covariance matrix $\Sigma$ of $p(u_k, v)$ equal unity. Let us discuss a direction cosine between connection weights as preparation for obtaining non-diagonal components. $R_k \equiv B^T J_k / |B| |J_k|$ is called the similarity between teacher and student. $q_{kk'} \equiv J_k^T J_{k'} / |J_k| |J_{k'}|$ is called the similarity among students. Covariance between an internal potential $v$ of a teacher $B$ and a normalized internal potential $u_k$ of a student $J_k$ equals a similarity $R_k$. Covariance between a normalized internal potential $u_k$ of a student $J_k$ and a normalized internal potential $u_{k'}$ of another student $J_{k'}$ equals a similarity $q_{kk'}$. Therefore, Eq. (3-1) can be rewritten as

$$\epsilon_g = \int \prod_{k=1}^{K} du_k dv p(u_k, v) \epsilon(u_k, v),$$

$$p(u_k, v) = \frac{1}{(2\pi)^{K+1} |\Sigma|^\frac{1}{2}} \exp \left( -\frac{(u_k, v) \Sigma^{-1} (u_k, v)^T}{2} \right),$$

$$\Sigma = \begin{pmatrix}
1 & q_{12} & \cdots & q_{1K} & R_1 \\
q_{21} & 1 & \cdots & \cdots & \\
\vdots & \ddots & \ddots & \ddots & \\
q_{K1} & \cdots & q_{K-1,K} & 1 & R_K \\
R_1 & \cdots & \cdots & R_K & 1
\end{pmatrix}.$$
As a result, a generalization error $\epsilon_g$ can be calculated if all similarities $R_k$ and $q_{k'k}$ are obtained. Differential equations regarding $l_k$ and $R_k$ for general learning rules have been obtained based on self-averaging as follows,\(^9\)

$$\frac{dl_k}{dt} = \langle f_k u_k \rangle + \langle f_k^2 \rangle \frac{2l_k}{l_k}., \quad \frac{dR_k}{dt} = \frac{\langle f_k v \rangle - \langle f_k u_k \rangle R_k}{l_k} - \frac{R_k}{2l_k} \langle f_k^2 \rangle,$$

(3-5)

where $\langle \cdot \rangle$ stands for the sample average. A differential equation regarding $q$ is obtained as follows from self-averaging,\(^4\),\(^13\),\(^14\)

$$\frac{dq_{k'k}}{dt} = \frac{\langle f_{k'k} u_k \rangle - q_{k'k} \langle f_{k'k} u_{k'} \rangle}{l_k} + \frac{\langle f_{k'k} u_{k'} \rangle - q_{k'k} \langle f_k u_k \rangle}{l_k}$$

$$+ \frac{\langle f_k f_{k'} \rangle}{l_k l_{k'}} - \frac{q_{k'k}}{2} \left( \frac{\langle f_k^2 \rangle}{l_k^2} + \frac{\langle f_{k'}^2 \rangle}{l_{k'}^2} \right).$$

(3-6)

\section{Result}

The update procedures $f(\text{sgn}(v), u)$ for Hebbian, perceptron and AdaTron learning are $\text{sgn}(v) \Theta (-u) \text{sgn}(v)$ and $-u \Theta (-u)$, respectively. Using these expressions, $\langle f_k u_k \rangle$, $\langle f_k v \rangle$ and $\langle f_k^2 \rangle$ can be obtained.\(^9\),\(^16\) $\langle f_k u_{k'} \rangle$ and $\langle f_k f_{k'} \rangle$ are also derived analytically.\(^13\),\(^14\) Dynamical behaviors of $R$ and $q$ have been obtained numerically by solving Eqs. (3-5), (3-6) and these sample averages. We have obtained numerical ensemble generalization errors $\epsilon_g$ by using Eqs. (3-2)–(3-4) and the above $R$ and $q$. Figures 1–3 show the relationships between $K$ and the effect of ensemble obtained by the Metropolis method using the values of $R$ and $q$ calculated numerically. In these figures, MV and WM indicate the majority vote and the weight mean, respectively. The ordinates have been normalized by the theoretical ensemble generalization error of $K = 1$ and $t = 50$. Computer simulations were executed with $N = 10^4$.

![Fig. 1. Relationship between $K$ and effect of ensemble in Hebbian learning.](image1)

![Fig. 2. Relationship between $K$ and effect of ensemble in perceptron learning.](image2)
Analysis of Ensemble Learning

Fig. 3. Relationship between $K$ and effect of ensemble in AdaTron learning.

Fig. 4. Relationship between $R$ and $q$.

§5. Discussion

Figures 1 – 3 show that the generalization errors of the three learning rules are all improved by ensemble learning. However, the degree of improvement is small in Hebbian learning and large in AdaTron learning. We discuss the reason for this difference in the following.

Each student moves towards teacher as learning proceeds. Therefore, similarities $R$ and $q$ increase and approach unity, leading to $R$ and $q$ becoming less irrelevant to each other. Then, if $q$ is relatively smaller when compared with $R_k$, variety among students is further maintained and the effect of the ensemble can be considered as large. Therefore, the relationship between $R$ and $q$ is essential in ensemble learning. To illustrate this, Fig. 4 shows the relationship by taking $R$ and $q$ as axes. In this figure, the curve for AdaTron learning is located in the bottom. That is, of the three learning rules, the one offering the smallest $q$ when compared with $R$ is AdaTron learning. In other words, the learning rule in which the rising of $q$ is the slowest and the variety among students is maintained best is AdaTron learning. From the perspective of the difference between the majority vote and the weight mean, Figs. 1 – 3 show that the improvement by weight mean is larger than that by majority vote in all three learning rules.

Acknowledgements

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, with Grant-in-Aid for Scientific Research 13780313, 14084212, 14580438 and 15500151.

References

S. Miyoshi, K. Hara and M. Okada

4) K. Hara and M. Okada, cond-mat/0402069.
13) S. Miyoshi, K. Hara and M. Okada, cond-mat/0403632.