Combining Classifiers and Learning Mixture-of-Experts

Lei Xu
Chinese University of Hong Kong, Hong Kong & Peking University, China

Shun-ichi Amari
Brain Science Institute, Japan

INTRODUCTION

Expert combination is a classic strategy that has been widely used in various problem solving tasks. A team of individuals with diverse and complementary skills tackle a task jointly such that a performance better than any single individual can make is achieved via integrating the strengths of individuals. Started from the late 1980' in the handwritten character recognition literature, studies have been made on combining multiple classifiers. Also from the early 1990' in the fields of neural networks and machine learning, efforts have been made under the name of ensemble learning or mixture of experts on how to learn jointly a mixture of experts (parametric models) and a combining strategy for integrating them in an optimal sense.

The article aims at a general sketch of two streams of studies, not only with a re-elaboration of essential tasks, basic ingredients, and typical combining rules, but also with a general combination framework (especially one concise and more useful one-parameter modulated special case, called α-integration) suggested to unify a number of typical classifier combination rules and several mixture based learning models, as well as max rule and min rule used in the literature on fuzzy system.

BACKGROUND

Both streams of studies are featured by two periods of developments. The first period is roughly from the late 1980s to the early 1990s. In the handwritten character recognition literature, various classifiers have been developed from different methodologies and different features, which motivate studies on combining multiple classifiers for a better performance. A systematical effort on the early stage of studies was made in (Xu, Krzyzak & Suen, 1992), with an attempt of setting up a general framework for classifier combination. As re-elaborated in Tab.1, not only two essential tasks were identified and a framework of three level combination was presented for the second task to cope with different types of classifier's output information, but also several rules have been investigated towards two of the three levels, especially with Bayes voting rule, product rule, and Dempster-Shafer rule proposed. Subsequently, the rest one (i.e., rank level) was soon studied in (Ho, Hull, & Srihari, 1994) via Borda count.

Interestingly and complementarily, almost in the same period the first task happens to be the focus of studies in the neural networks learning literature. Encountering the problems that there are different choices for the same type of neural net by varying its scale (e.g., the number of hidden units in a three layer net), different local optimal results on the same neural net due to different initializations, studies have been made on how to train an ensemble of diverse and complementary networks via cross-validation- partitioning, correlation reduction pruning, performance guided re-sampling, etc, such that the resulted combination produces a better generalization performance (Hansen & Salamon, 1990; Xu, Krzyzak, & Suen, 1991; Wolpert, 1992; Baxt, 1992, Breiman, 1992&94; Drucker, et al, 1994). In addition to classification, this stream also handles function regression via integrating individual estimators by a linear combination (Perrone & Cooper, 1993). Furthermore, this stream progresses to consider the performance of two tasks in Tab.1 jointly in help of the mixture-of-expert (ME) models (Jacobs, et al, 1991; Jordan & Jacobs, 1994; Xu & Jordan, 1993; Xu, Jordan & Hinton, 1994), which can learn either or both of the combining mechanism and individual experts in a maximum likelihood sense.

Two stream studies in the first period jointly set up a landscape of this emerging research area, together

Table 1. Essential tasks and their implementations

Two Tasks (a quotation from Xu, Krzyzak and Suen, 1992)

Task 1: "How many and what type of classifiers should be used ~~for a specific application~~?, and for each classifier what type of features should we use?, as well as other problems that are related to the construction of those individual and complementary classifier".

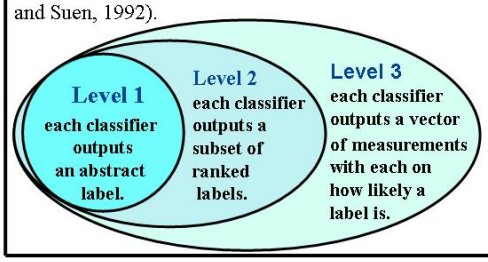
Task 2: "How to combine the results from different existing classifiers so that a better result can be obtained?"

Two Styles of Implementations

Two Stage Implementation

•***Task 1*** **is completed in advance, with the resulted classifiers being diverse and complementary.**

• **Perform *Task 2* in one of three levels** (Xu, Krzyzak and Suen, 1992).



Joint Implementation

Two tasks made jointly or alternatively

under a same criterion	others
• **Mixture of experts (ME)** (Jacobs, et al, 1991; Jordan & Jacobs, 1994); •**Alternative ME** (Xu & Jordan, 1993; Xu, Jordan & Hinton, 1994); • **EM-RBF** (Xu,1998) •**Three layer nets, etc.**	**Stacking , Boosting, …, etc** (Breiman, 1992&94; Wolpert, 1992)

with a number of typical topics or directions. Thereafter, further studies have been further conducted on each of these typical directions. First, theoretical analyses have been made for deep insights and improved performances. For examples, convergence analysis on the EM algorithm for the mixture based learning are conducted in (Jordan & Xu, 1995; Xu & Jordan, 1996). In Tumer & Ghosh (1996), the additive errors of posteriori probabilities by classifiers or experts are considered, with variances and correlations of these errors investigated for improving the performance of a sum based combination. In Kittler, et al (1998), the effect of these errors on the sensitivity of sum rule vs product rule are further investigated, with a conclusion that summation is much preferred. Also, a theoretical framework is suggested for taking several combining rules as special cases (Kittler, 1998) , being unaware of that this framework is actually the mixture-of-experts model that was proposed firstly for combining multiple function regressions in (Jacobs, et al, 1991) and then for combining multiple classifiers in (Xu & Jordan, 1993). In addition, another theoretical study is made on six classifier fusion strategies in (Kuncheva, 2002). Second, there are further studies on Dempster-Shafer rule (Al-Ania, 2002) and other combing methods such as rank based, boosting based, as well as local accuracy estimates (Woods, Kegelmeyer, & Bowyer, 1997). Third, there are a large number of applications. Due to space limit, details are referred to Ranawana & Palade (2006) and Sharkey & Sharkey (1999).

A GENERAL ARCHITECTURE, TWO TASKS, AND THREE INGREDIENTS

We consider a general architecture shown in Fig.1. There are $\{e_j(x)\}_{j=1}^k$ experts with each $e_j(x)$ as either a classifier or an estimator. As shown in Tab.2, a classifier outputs one of three types of information, on which we have three levels of combination. The first two can be regarded as special cases of the third one that outputs a vector of measurements. A typical example is $[p_j(1|x),\ldots,p_j(m|x)]^T$ with each $1 \geq p_j(\ell|x) \geq 0$ expressing a posteriori probability that x is classified to the ℓ-th class. Also, $p_j(\ell|x) = p_j(y=\ell|x)$ can be further extended to $p_j(y|x)$ that describes a distribution for a regression $x \rightarrow y \in R^m$. In Figure 1, there is also a gating net that generates signals $\{\alpha_j(x)\}_{j=1}^k$ to modulate experts by a combining mechanism $M(x)$.