Generalized Ambiguity Decompositions for Classification with Applications in Active Learning and Unsupervised Ensemble Pruning

Zhengshen Jiang
Electronics Engineering and
Computer Science, Peking University
Beijing, P.R. China
jiangzhengshen@pku.edu.cn

Hongzhi Liu,* Bin Fu, Zhonghai Wu*
National Engineering Center of
Software Engineering, Peking University
Beijing, P.R. China
{liuhz, fubin1990, wuzh}@pku.edu.cn
*Corresponding author

Abstract

Error decomposition analysis is a key problem for ensemble learning. Two commonly used error decomposition schemes, the classic Ambiguity Decomposition and Bias-Variance-Covariance decomposition, are only suitable for regression tasks with square loss. We generalized the classic Ambiguity Decomposition from regression problems with square loss to classification problems with any loss functions that are twice differentiable, including the logistic loss in Logistic Regression, the exponential loss in Boosting methods, and the 0-1 loss in many other classification tasks. We further proved several important properties of the Ambiguity term, armed with which the Ambiguity terms of logistic loss, exponential loss and 0-1 loss can be explicitly computed and optimized. We further discussed the relationship between margin theory, "good" and "bad" diversity theory and our theoretical results, and provided some new insights for ensemble learning. We demonstrated the applications of our theoretical results in active learning and unsupervised ensemble pruning, and the experimental results confirmed the effectiveness of our methods.

Introduction

Previous work

As a sub-field of machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone (Rokach 2010). To achieve good performance, the base learners should be both accurate and diverse.

It is widely accepted that the generalization error of an ensemble depends on a term related to diversity (Zhou 2012). Thus, error decomposition analysis has long been considered as a key problem in ensemble learning. Two commonly used error decomposition schemes are the classic Ambiguity Decomposition and Bias-Variance-Covariance decomposition. Both the two decompositions are only suitable for regression tasks with square loss. In this work, we generalized the classic Ambiguity Decomposition to classification tasks with a variety of loss functions, including the logistic loss, exponential loss, 0-1 loss, etc.

The classic Ambiguity Decomposition revealed the relationship between the loss of base learners and that of the ensemble for ensemble regression (Krogh and Vedelsby 1995). Its form for a single sample (Brown et al. 2005) is

$$(f_{ens} - y)^2 = \sum_i \omega_i (f_i - y)^2 - \sum_i \omega_i (f_i - f_{ens})^2 \quad (1)$$

where f_i is the output of a base classifier, y is the true value of the considered sample's target, and f_{ens} is a convex combination of the base classifiers, i.e., $f_{ens} = \sum_i \omega_i f_i$, where $\sum_i \omega_i = 1$ and $\omega_i \geq 0$.

Computing the expectation in the sample space yields the classic Ambiguity Decomposition, which is always written as

$$E = \bar{E} - \bar{A}$$

where E is the generalization error of the ensemble, $\bar{E}$ is the average generalization error of the base learners, and $\bar{A}$, which has been considered to be relevant to diversity, is called "ambiguity".

The classic Ambiguity Decomposition assumes that the loss function is square loss which is commonly used in regression but not suitable for classification.

Brown and Kuncheva broke down the Ambiguity term for 0-1 loss into two terms: "good" and "bad" diversity. The good diversity term is taken out of $\bar{E}$ whereas the bad diversity term is added to it (Brown and Kuncheva 2010). Their work only considered classification tasks with 0-1 loss, and the base classifiers were assumed to be combined using majority voting.

Audhkhasi et al. presented a framework to generalize the classic Ambiguity Decomposition to classification problems (Audhkhasi et al. 2013). Their results showed that the ensemble performance approximately decomposed into a difference of the average classifier performance and the diversity of the ensemble. However, their result was a bound of the generalization error. The bound is not tight and the applicability is limited.

Contributions

In this paper, we contributed to the field of machine learning both theoretically and practically.

First, we presented two Generalized Ambiguity Decompositions that can be used not only in regression tasks, but

also in classification tasks with a variety of loss functions, e.g., the logistic loss in Logistic Regression, the exponential loss in Boosting methods, and the 0-1 loss in many classification tasks. These decompositions can be used in many optimization problems in ensemble learning and active learning. All these are not possible by using the classic Ambiguity Decomposition or the "good" and "bad" diversity theory.

Secondly, we proved several important properties of E, $\bar{E}$ and $\bar{A}$. Armed with these properties, the Ambiguity terms can be explicitly computed and optimized. We further discussed the relationship between margin theory, "good" and "bad" diversity theory and our theoretical results, and provided some new insights for ensemble learning.

Lastly, we demonstrated the applications of our theoretical results in active learning and unsupervised ensemble pruning, and the experimental results confirmed the effectiveness of our methods.

The Two Generalized Ambiguity Decompositions

In this section, we will present two Generalized Ambiguity Decompositions which generalize the classic Ambiguity Decomposition to classification problems. We will prove that the classic Ambiguity Decomposition is a special case of both of the generalized decompositions. At the end of this section, we will show that our results provide new insights into ensemble learning both theoretically and practically.

Notations

In this work, we focus on two-class classification problems. But the theoretical results can also be used in regression problems.

We make the usual assumption of PAC learning theory (Valiant 1984) that a task D corresponds to a probability distribution over the input-output space $\mathcal{X} \times \mathcal{Y}$. A sample from D is represented as $(\mathbf{x}, y)$, where $\mathbf{x} \in R^d$ is a vector of the attributes, and $y \in \mathcal{Y}$ is the label. h is a classifier that is trained to predict the label. f is the output of classifier h for input $\mathbf{x}$, i.e. $f = h(\mathbf{x})$. The ensemble of the base classifiers is

$$H(\mathbf{x}) = \sum_i \omega_i h_i(\mathbf{x}) \qquad \text{or} \qquad f_{ens} = \bar{f} = \sum_i \omega_i f_i$$

where ω_i is the weight of classifier h_i. In the following deductions, we assume that $\sum_i \omega_i = 1$ for conciseness reason. However, this assumption is not essential.

We denote the loss function as $l(f, y)$. Commonly used loss functions include the square loss $l(f, y) = (f - y)^2$, the logistic loss $l(f, y) = \log(1 + e^{-yf})$ (Collins, Schapire, and Singer 2002), the exponential loss $l(f, y) = e^{-yf}$ (Collins, Schapire, and Singer 2002), etc.

The generalization error is represented as $E_D\{l(f, y)\}$, where E_D is the expectation in the sample space. Specifically, in the same way as in (Mukherjee et al. 2003), the generalization error is computed as

$$E_D\{l(f, y)\} = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} l(h(\mathbf{x}), y)$$

The Generalized Ambiguity Decompositions

Inspired by the work of Audhkhasi et al., we present the following two Generalized Ambiguity Decompositions.

Theorem 1 (**The First Generalized Ambiguity Decomposition**) *Assume we are dealing with binary classification problems. A set of classifiers $\{h_1, h_2, ..., h_T\}$ have been trained and are combined by weighted averaging $f_{ens} = \sum_i \omega_i f_i$ with $f_i = h_i(\mathbf{x})$ and $\sum_i \omega_i = 1$. Then for any loss function that is twice differentiable, the loss function of the ensemble can be decomposed into*

$$\begin{aligned} l(f_{ens}, y) &= \sum_{i=1}^{T} \omega_i l(f_i, y) \\ &\quad -\frac{1}{2}\sum_{i=1}^{T} \omega_i [l''(f_i^*, y) f_i^2 - l''(f_{ens}^*, y) f_{ens}^2] \end{aligned} \tag{2}$$

Computing the expectation in sample space yields decomposition of the generalization error as following

$$E = \bar{E} - \bar{A}$$

where

$$\begin{aligned} E &= E_D\{l(f_{ens}, y)\} \\ \bar{E} &= \sum_{i=1}^{T} \omega_i E_D\{l(f_i, y)\} \\ \bar{A} &= \frac{1}{2}\sum_{i=1}^{T} \omega_i E_D\{l''(f_i^*, y) f_i^2 - l''(f_{ens}^*, y) f_{ens}^2\} \end{aligned}$$

with f_i^* being some number between zero and f_i, f_{ens}^* being some number between zero and f_{ens}, and $E_D\{\cdot\}$ representing the expectation in sample space.

Proof. For a single sample, the loss of the output f_i given by base classifier h_i can be expanded near zero according to Taylor's theorem:

$$l(f_i, y) = l(0, y) + l'(0, y) f_i + \frac{1}{2} l''(f_i^*, y) f_i^2$$

where f_i^* is an uncertain number between zero and f_i.

For the ensemble, the loss function $l(f_{ens}, y)$ can also be expanded in a similar way, i.e.,

$$l(f_{ens}, y) = l(0, y) + l'(0, y) f_{ens} + \frac{1}{2} l''(f_{ens}^*, y) f_{ens}^2$$

where f_{ens}^* is an uncertain number between zero and f_{ens}.

The weighted average of $l(f_i, y)$ is

$$\sum_{i=1}^{T} \omega_i l(f_i, y) = l(0, y) + l'(0, y) f_{ens} + \frac{1}{2}\sum_{i=1}^{T} \omega_i l''(f_i^*, y) f_i^2$$

Thus

$$\begin{aligned} &\sum_{i=1}^{T} \omega_i l(f_i, y) - l(f_{ens}, y) \\ = \ &\frac{1}{2}\sum_{i=1}^{T} \omega_i [l''(f_i^*, y) f_i^2 - l''(f_{ens}^*, y) f_{ens}^2] \end{aligned}$$

Finally, we get the form of the Generalized Ambiguity Decomposition for a single sample as in Equation 2.

Computing the expectation over the sample space by integrating with respect to the probability density function $p(x)$ yields the form of the Generalized Ambiguity Decomposition for the overall dataset, which is exactly the theorem above. □

Theorem 2 (**The Second Generalized Ambiguity Decomposition**) *With the same assumptions as in Theorem 1, for any loss function that is twice differentiable, the loss function of the ensemble can be decomposed into*

$$l(f_{ens}, y) = \sum_{i=1}^{T} \omega_i l(f_i, y) - \frac{1}{2} \sum_{i=1}^{T} \omega_i l''(f_i^*, y)(f_i - f_{ens})^2$$

and thus with E and $\bar{E}$ remaining the same as in Theorem 1,

$$\bar{A} = \frac{1}{2} \sum_{i=1}^{T} \omega_i E_D\{l''(f_i^*, y)(f_i - f_{ens})^2\}$$

with f_i^* being some value between f_i and f_{ens}, and $E_D\{\cdot\}$ representing the expectation in sample space.

Proof. For a single sample, the loss of the output f_i given by base classifier h_i can be expanded near the ensemble output f_{ens} according to Taylor's theorem:

$$\begin{aligned} l(f_i, y) &= l(f_{ens}, y) + l'(f_{ens}, y)(f_i - f_{ens}) \\ &\quad + \frac{1}{2} l''(f_i^*, y)(f_i - f_{ens})^2 \end{aligned}$$

where f_i^* is a value between f_i and f_{ens}.

Summing over all the loss functions of the base classifiers using weighted averaging yields

$$\sum_{i=1}^{T} \omega_i l(f_i, y) = l(f_{ens}, y) + \frac{1}{2} \sum_{i=1}^{T} \omega_i l''(f_i^*, y)(f_i - f_{ens})^2$$

This is exactly the form of the second Generalized Ambiguity Decomposition for a single sample.

Computing the expectation over the sample space yields the form of the Generalized Ambiguity Decomposition for the overall dataset, which is exactly the theorem above. □

Above theorems are both generalizations of the classic Ambiguity Decomposition, which was derived under the square loss assumption. It can be verified that when we choose $l(f, y) = (f - y)^2$ as the loss function, the Generalized Ambiguity Decompositions become exactly the classic Ambiguity Decomposition, since $l'' = 2$ holds for any f and y.

Property of the Parameter f^*

In above two Generalized Ambiguity Decompositions, there are three uncertain numbers, that is, f_i^* between 0 and f_i, f_{ens}^* between 0 and f_{ens} in Theorem 1, and f_i^* between f_i and f_{ens} in Theorem 2. All the three parameters are in close relationship with Lagrange mean value. In this section, we will prove that in limit situation, the parameters in Theorem 1 can be estimated by

$$\hat{f}_i^* = \frac{f_i}{4} \quad , \quad \hat{f}_{ens}^* = \frac{f_{ens}}{4} \tag{3}$$

and the parameter in Theorem 2 can be estimated by

$$\hat{f}_i^* = \frac{f_i + 3 f_{ens}}{4} \tag{4}$$

Our proof is based on the theorem proved by (Azpeitia 1982) as following.

Lemma 3 *Suppose $f''(x)$ exists in a neighborhood of point a, $f''(x)$ is continuous at a and $f''(a) \neq 0$, ξ is decided by Lagrange mean value theorem:*

$$f(x) = f(a) + f'(\xi)(x - a),$$

then $\lim_{x \to a} \frac{\xi - a}{x - a} = \frac{1}{2}$.

Using this lemma, we can proof Equation 3 and 4 as following.

Theorem 4 *Suppose $f'''(x)$ exists in a neighborhood of point a, $f'''(x)$ is continuous at a and $f'''(a) \neq 0$, the uncertain parameter ξ in the Lagrange remainder of Taylor's theorem ξ is decided by Lagrange mean value theorem:*

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2} f''(\xi)(x - a)^2,$$

then $\lim_{x \to a} \xi = \frac{x + 3a}{4}$.

Proof. According to Lemma 3, we get $f(x) = f(a) + f'(\xi_0)(x - a)$ and $\lim_{x \to a} \xi_0 = \frac{x + a}{2}$.

Expanding $f'(\xi_0)$ according to Lagrange mean value theorem yields

$$\begin{aligned} f'(\xi_0) &= f'(a) + f''(\xi)(\xi_0 - a) \\ &\overset{x \to a}{=} f'(a) + \frac{1}{2} f''(\xi)(x - a) \end{aligned}$$

Substituting above equation into the first one yields

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2} f''(\xi)(x - a)^2,$$

and $\lim_{x \to a} \xi = \frac{\xi_0 + a}{2} = \frac{x + 3a}{4}$. □

Corollaries and Discussions

Optimization Problems in Ensemble Learning The goal of ensemble learning is to minimize the generalization error E. Many tasks in ensemble learning need to optimize an objective function that is related to diversity, such as base classifier generation, ensemble pruning and the optimization of ensemble weights. According to the classic Ambiguity Decomposition, the minimization of E can be decomposed into minimizing $\bar{E}$ and maximizing $\bar{A}$, and $\bar{A}$ is in close relationship with diversity. However, since the classic Ambiguity Decomposition only holds for regression problems, the objective function related to diversity in classification problems is somewhat heuristic. Our newly presented Generalized Ambiguity Decompositions provide theoretical basis to explore the Ambiguity term in classification problems, and enable us to maximize the Ambiguity of many loss functions, such as logistic loss, exponential loss, 0-1 loss, and other less common loss functions. We will further discuss about this later.

Active Learning In the classic Ambiguity Decomposition, $\bar{A}$ can be estimated entirely from unlabelled data, which makes it possible to exploit unsupervised learning methods. Using Theorem 2, we have similar result for logistic loss in classification problems.

Corollary 1 *In classification problems where class label $y \in \{-1, 1\}$, when using logistic loss function, i.e., $l(f, y) = \log(1 + e^{-yf})$, the Ambiguity term $\bar{A}$ is independent with the class label y.*

This is because the second derivative of logistic loss is $l''(f, y) = 1/(e^{yf} + 2 + e^{-yf})$. In the case that $y \in \{-1, 1\}$, $l''(f, y)$ is independent with y, and so do the Ambiguity term according to Theorem 1 or 2.

That is to say, when we use logistic loss in two-class classification problems, unsupervised learning and active learning methods can be used to explore the Ambiguity term. We will demonstrate this in the application section later.

Error Bound and Jensen's Inequality Considering that $l'' \geq 0$ always holds for convex loss functions, we can further prove the following corollary from Theorem 2.

Corollary 2 *For arbitrary convex loss function, the generalization error of the ensemble is always smaller than the weighted average error of the base classifiers, i.e., $E \leq \bar{E}$, since $\bar{A} \geq 0$ always holds for convex loss functions.*

This corollary can also be proved by Jensen's Inequality which has been mentioned many times in the classifier ensemble literature (Krogh and Vedelsby 1995; Audhkhasi et al. 2013). From this corollary, $\bar{E}$ can be viewed as an upper bound of the ensemble error E, which coincides with the fact that the performance of ensemble method is always better than the average performance of the base classifiers.

Properties of E, $\bar{E}$, $\bar{A}$

To further explore the properties of E, $\bar{E}$ and $\bar{A}$, in this section, we analyse the three terms in a special case where the oracle output of base classifiers are combined, i.e., $f_i \in \{-1, 1\}$. In such a case, y should also in $\{-1, 1\}$, although it is not required in our derivations.

The following theorem is about the relationship between the three components E, $\bar{E}$, $\bar{A}$ and $\bar{f}$.

Theorem 5 (E, $\bar{E}$, $\bar{A}$ **only depend on** $\bar{f}$ **and** y) *Assume the outputs of the base classifiers only take the values -1 or 1, which is also called "oracle output", i.e., $f_i \in \{-1, 1\}$. Also assume the base classifiers are combined by weighted voting, i.e., $f_{ens} = \bar{f} = \sum_i \omega_i f_i$ and $\sum_i \omega_i = 1$. Then all the three terms E, $\bar{E}$ and $\bar{A}$ in the Generalized Ambiguity Decomposition only depend on $\bar{f}$(which is also f_{ens}).*

Proof. It is obvious that the term E depends on $\bar{f}$ and y only, since it is defined as the expectation of $l(f_{ens}, y)$ and $f_{ens} = \bar{f}$.

Following we prove that the term $\bar{E}$ only depends on $\bar{f}$ and y.

According to Taylor's theorem, $l(f_i, y)$ can be decomposed as following

$$l(f_i, y) = l(0, y) + l'(0, y) f_i + \sum_{n=2}^{+\infty} \frac{l^{(n)}(0, y)}{n!} f_i^n$$

Considering $f_i \in \{-1, 1\}$, we get

$$f_i^n = \begin{cases} f_i, & \text{for } n = 1, 3, 5, ... \\ 1, & \text{for } n = 2, 4, 6, ... \end{cases}$$

So weighted averaging of $l(f_i, y)$ yields

$$\begin{aligned} \sum_{i=1}^{T} \omega_i l(f_i, y) &= [l(0, y) + \sum_{n=2,4,6,...} \frac{l^{(n)}(0, y)}{n!}] \\ &\quad + [l'(0, y) + \sum_{n=3,5,7,...} \frac{l^{(n)}(0, y)}{n!}] \bar{f} \end{aligned}$$

As can be seen, the only variable is $\bar{f}$ and y. Since $\bar{E}$ is defined as the expectation of $\sum_{i=1}^{M} \omega_i l(f_i, y)$, we can conclude that the term $\bar{E}$ only depends on $\bar{f}$ and y.

Finally, since E and $\bar{E}$ only depend on $\bar{f}$ and y, it is natural that the Ambiguity term, which can be computed by $\bar{A} = \bar{E} - E$, only depend on $\bar{f}$ and y too. □

According to Theorem 5, if the values of $\bar{f}$ are given, all the three terms E, $\bar{E}$ and $\bar{A}$ can be computed, no matter what value the f_i's take.

Following we prove that $\bar{E}$ is always a linear function of $\bar{f}$, and get a concise form of $\bar{E}$ which is easy to compute.

Theorem 6 (**Given class label y, $\bar{e}$ is always a linear function of $\bar{f}$**) *Assume $f_i \in \{-1, 1\}$ and the base classifiers are combined by weighted voting. Denote e, $\bar{e}$ and $\bar{a}$ according to Equation 2 as $e = l(\bar{f}, y)$, $\bar{e} = \sum \omega_i l(f_i, y)$, $\bar{a} = \bar{e} - e$, and E, $\bar{E}$, $\bar{A}$ are the expectations of them, respectively. Then $\bar{e}$ is always a linear function of $\bar{f}$ (note that $l(1, y)$ or $l(-1, y)$ takes different values for different loss)*

$$\bar{e} = \frac{l(1, y) + l(-1, y)}{2} + \frac{l(1, y) - l(-1, y)}{2} \bar{f}$$

and

$$\bar{E} = E_D\{\frac{l(1, y) + l(-1, y)}{2}\} + E_D\{\frac{l(1, y) - l(-1, y)}{2} \bar{f}\}$$

Proof. Let

$$A = \sum_{n=2,4,6,...} \frac{l^{(n)}(0, y)}{n!}, \qquad B = \sum_{n=3,5,7,...} \frac{l^{(n)}(0, y)}{n!}$$

Considering

$$l(1, y) = l(0, y) + l'(0, y) + A + B$$

and

$$l(-1, y) = l(0, y) - l'(0, y) + A - B$$

we have

$$A = \frac{l(1, y) + l(-1, y)}{2} - l(0, y)$$

$$B = \frac{l(1, y) - l(-1, y)}{2} - l'(0, y)$$

Substituting A and B into the last equation in the proof of Theorem 5 yields the theorem. □

Further, we prove the following theorem.

Theorem 7 *Assume $f_i \in \{-1, 1\}$ and the base classifiers are combined by weighted voting, then* ***the straight line represented by $\bar{e}$ always pass through the two endpoints of the curve represented by*** e.

Proof. According to the definition of e, the two endpoints are $l(-1, y)$ and $l(1, y)$.

The two endpoints of the straight line $\bar{e}$ corresponds to the points with $\bar{f} = -1$ and $\bar{f} = 1$. So the endpoints are

$$\frac{l(1,y) + l(-1,y)}{2} - \frac{l(1,y) - l(-1,y)}{2} = l(-1, y)$$
$$\frac{l(1,y) + l(-1,y)}{2} + \frac{l(1,y) - l(-1,y)}{2} = l(1, y)$$

which are exactly the same with that of e. □

Conclusion *Given that $f_i \in \{-1, 1\}$, for all the twice differentiable loss function, we have (1) all the three components E, $\bar{E}$ and $\bar{A}$ only depend on $\bar{f}$; (2) $\bar{e}$ is a linear function of $\bar{f}$, and the function can be computed according to Theorem 6; and (3) the straight line $\bar{e}$ always passes through the endpoints of e.* Using these results, the curves of e, $\bar{e}$ and $\bar{a}$ for different loss functions can be plotted. For space considerations, the plots are not shown in this paper.

It should be noted that the results above also hold for the case of $\sum_i \omega_i \neq 1$, although the theorems were proved in the $\sum_i \omega_i = 1$ case for simplicity reason.

Discussion

Our results provide some new insights in ensemble learning.

Margin Theory has long been used as the explanation of AdaBoost (Freund and Schapire 1995), the widely used ensemble learning algorithm. The margin of a single classifier is defined as yf_i, and the margin of an ensemble $f_{ens} = \bar{f}$, is defined as $y\bar{f} = \sum \omega_i y f_i$. According to Theorem 6, assuming that $y \in \{-1, 1\}$, it can be proved that if $l(1, 1) = l(-1, -1)$ and $l(1, -1) = l(-1, 1)$, then

$$\bar{e} = \frac{l(1,1) + l(-1,1)}{2} - \frac{l(-1,1) - l(1,1)}{2} y\bar{f}$$

Thus, $\bar{e}$ depends only on the margin of the ensemble. Moreover, if the ensemble loss e depends only on $y\bar{f}$, the Ambiguity $\bar{a}$ will depend only on the margin too, which is exactly the case of the logistic loss, exponential loss and 0-1 loss which will be discussed later.

Since $l(-1, 1) \geq l(1, 1)$ always holds for reasonable loss functions, with $margin \stackrel{def}{=} y\bar{f}$, we have

$$\min \bar{E} \Leftrightarrow \max E_D\{margin\} \tag{5}$$

The Ambiguity Terms for Several Loss Functions Following we derive the ambiguity forms for several standard loss under the assumption that $y, f_i \in \{-1, 1\}$.

Logistic loss is usually used in logistic regression, which is a popular technique for classification (Collins, Schapire, and Singer 2002). The loss function is $l(f, y) = \log(1 + e^{-yf})$. After some derivation, we could get that

$$\bar{a} = [-\frac{1}{2} + \log(1 + e)] - \log(e^{y\bar{f}/2} + e^{-y\bar{f}/2})$$

$$\max \bar{A} \Leftrightarrow \min E_D\{\log(e^{\bar{f}/2} + e^{-\bar{f}/2})\} \tag{6}$$

As can be seen, the Ambiguity term for logistic loss is independent from y, which is consistent with the discussion before.

Exponential Loss is widely used in Boosting algorithms (Collins, Schapire, and Singer 2002), e.g. AdaBoost. We take AdaBoost.M1 (Freund and Schapire 1995) into consideration. The loss function used in this algorithm is $l(f, y) = e^{-yf}$. The Ambiguity term is $\bar{a} = (e + e^{-1})/2 - (e - e^{-1})y\bar{f}/2 - e^{-y\bar{f}}$ and

$$\max \bar{A} \Leftrightarrow \min E_D\{\frac{e - e^{-1}}{2} y\bar{f} + e^{-y\bar{f}}\} \tag{7}$$

0-1 Loss is the most commonly used loss functions in classification problems. For two-class problems, $l(f, y) = 1$ if $yf < 0$ otherwise $l(f, y) = 0$. Although it is not differentiable at $f = 0$, it can be approximated to arbitrary precision using some differentiable function, such as the logistic function. Such approximation does not change the results in Theorem 1 to 7. In limit situation, which is exactly the 0-1 loss case, using any of the approximations, $\bar{e}$ can be represented as $\bar{e} = 1/2 - y\bar{f}/2$. So the Ambiguity term is $\bar{a} = \frac{1}{2}(\text{sign}\{y\bar{f}\} - y\bar{f})$ and

$$\max \bar{A} \Leftrightarrow \min E_D\{y\bar{f} - \text{sign}\{y\bar{f}\}\} \tag{8}$$

Since 0-1 loss is not convex, the Ambiguity term can be positive, negative or 0.

Combination of Equation 5 and Equation 6, 7 or 8 forms the optimization problems in ensemble learning with logistic loss, exponential loss and 0-1 loss, respectively. These objective functions can be used in every phases in ensemble learning, i.e., base classifier generation, pruning and combination.

"Good" and "bad" diversity presented by (Brown and Kuncheva 2010) can also be accounted for by our results. The authors broke down the Ambiguity term for 0-1 loss into two terms: "good" and "bad" diversity. Our result is consistent with theirs. In 0-1 loss, "good" and "bad" diversity correspond to the two branches of Ambiguity. In their work, they concluded that the Ambiguity term "can be directly expressed in terms of the average classifier disagreement", which is exactly the margin. We generalized their conclusion to any loss functions that are twice differentiable, and further proved that $\bar{e}$ is a linear function of the margin.

Applications

There are many potential applications of our theoretical results. As the demonstrations, we apply our theoretical results in active learning and unsupervised ensemble pruning.

Active Learning

Active learning would benefit from our theoretical results. According to Theorem 1 and 2, $\bar{e} \geq \bar{a}$ always holds since the generalization error is always non-negative. Therefore we can treat the Ambiguity as a lower bound for the average error of the base classifiers. Thus samples with large Ambiguity are "harder", and the ensemble would benefit the most

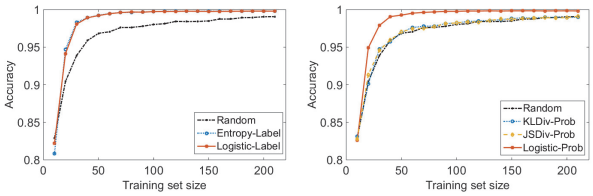

Figure 1: Results for Active Learning on *pendigits* dataset as a representation. "Random" corresponds to the method where the samples were selected randomly. In "Entropy-Label" and "Logistic-Label", class label outputs were used, and new sample was selected according to vote entropy and logistic disagreement respectively. In "KLDiv-Prob", "JSDiv-Prob" and "Logistic-Prob", probability outputs were used, and the disagreement was estimated by KL divergence, JS divergence and logistic disagreement, respectively.

from the samples with the largest Ambiguity. Moreover, diversity is an important property for committee-based active learning methods (Melville and Mooney 2004). Since the Ambiguity term is related to diversity, it could be used as the diversity measure in active learning methods.

We tested the above ideas by a query-by-committee (QBC) like algorithm (Seung, Opper, and Sompolinsky 1992; Settles 2010), where the key role is the *disagreement* measure. Common disagreement measures include *vote entropy* (Dagan and Engelson 1995), *Kullback-Leibler (KL) divergence* (McCallumzy and Nigamy 1998) and *Jensen-Shannon (JS) divergence* (Cover and Thomas 1991).

In our algorithm, samples was selected using the *logistic disagreement* according to Equation 6, i.e.,

$$x^*_{LD} = \underset{x}{\operatorname{argmax}} \ -\log(e^{\bar{f}/2} + e^{-\bar{f}/2}) \tag{9}$$

where $\bar{f}$ is the average output of the base classifiers. This disagreement measure can be used with all kinds of outputs, including class labels, probabilities and scores. **When the base classifiers output class labels, the logistic disagreement is equivalent to vote entropy.**

We used 20 datasets from the UCI Repository (Lichman 2013) in our experiments. An ensemble of 5 CART decision trees (Breiman et al. 1984) was trained using the Bagging (Breiman 1996) algorithm. The initial dataset size was set to be 10. Each time the sample with the largest disagreement was selected out of 200 randomly chosen candidates. The results were averaged over 100 runs, and representative results were shown in Figure 1.

Among the algorithms, random ones performed the worst, which showed the effectiveness of active learning. In the case of class label outputs, Logistic-Label performed equally with Entropy-Label. In the case of probability outputs, with the probabilities scaled to $[-1, 1]$, Logistic-Prob achieved better results than KLDiv-Prob and JSDiv-Prob. All these results confirmed that active learning would benefit from our theoretical results.

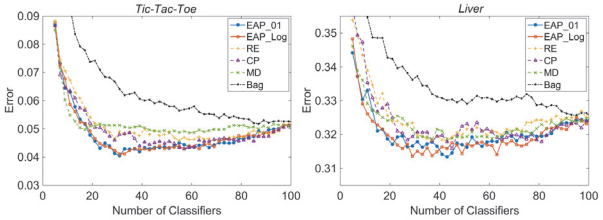

Figure 2: Representative test errors of compared ensemble pruning methods.

Unsupervised Ensemble Pruning

Ensemble pruning is a typical procedure in ensemble learning. Many ensemble pruning algorithms in classification problems need the label y as input, i.e., in a supervised manner. As was stated before, when using logistic loss as the loss function and $y \in \{-1, 1\}$, the Ambiguity term in Theorem 1 or 2 is independent with y. Thus, it is possible to carry out unsupervised ensemble pruning with logistic loss function.

In this section, we experimented on two Error-Ambiguity pruning (EAP) methods: one with 0-1 loss and the other with logistic loss. In our ensemble pruning method, we used a greedy forward procedure, and $\bar{E}/\bar{A}$ was used as the selection criterion. $\bar{E}$ was estimated using the average generalization error of the base classifiers, and can be estimated before the pruning procedure. $\bar{A}$ was computed using Theorem 2 and Equation 4. The pruning method with 0-1 loss is supervised while the one with logistic loss is unsupervised.

We compared our methods on 20 datasets from the UCI Repository to several comparative methods, i.e., Bagging (Bag) (Breiman 1996), Reduce-Error (RE) (Margineantu and Dietterich 1997), Complementarity (CP) (Martinez-Muoz, Hernández-Lobato, and Suarez 2009) and Margin Distance (MD) (Martinez-Muoz, Hernández-Lobato, and Suarez 2009). All the methods were evaluated 30 times on each dataset, and the final performance was obtained by averaging the error rates on test set. Each time we conducted the following steps. First, we randomly split the data set into train/valid/test set. Secondly, we used Bagging (Breiman 1996) to build an ensemble of 101 CART decision trees (Breiman et al. 1984) on the training set. Thirdly the base classifiers were pruned by different ensemble pruning methods. Lastly the performance of the pruned ensemble was evaluated on the testing set. Representative results were shown in Figure 2.

In Figure 2, EAP_01 and EAP_Log methods correspond to our Error-Ambiguity Pruning methods with 0-1 loss and logistic loss, respectively. Interestingly but not surprisingly, although EAP_Log is an unsupervised method, it achieved competitive results compared with other methods. Amongst the pruning methods, our methods decreased faster than the others, which proved the effectiveness of our methods.

Conclusion and Future Work

In this paper, we presented two Generalized Ambiguity Decomposition for classification problems, and discussed sev-

eral important properties of the Ambiguity term. We demonstrated the applications of our theoretical results in active learning and unsupervised ensemble pruning, and the experimental results confirmed the effectiveness of our methods. Interesting directions for future research include more applications of the decompositions, such as base classifier generation, ensemble weight optimization, and so forth.

Acknowledgments

This work is partially sponsored by National Natural Science Fund of China (Grant No. 61232005) and National 863 Program of China (Grant No. 2015AA016009).

References

Audhkhasi, K.; Sethy, A.; Ramabhadran, B.; and Narayanan, S. S. 2013. Generalized ambiguity decomposition for understanding ensemble diversity. *Eprint Arxiv*.

Azpeitia, A. G. 1982. On the lagrange remainder of the taylor formula. *American Mathematical Monthly* 89(5):311–312.

Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. A. 1984. *Classification and regression trees*. CRC press.

Breiman, L. 1996. Bagging predictors. *Machine learning* 24(2):123–140.

Brown, G., and Kuncheva, L. I. 2010. "Good" and "bad" diversity in majority vote ensembles. In *International Conference on Multiple Classifier Systems*, 124–133. Springer.

Brown, G.; Wyatt, J.; Harris, R.; and Yao, X. 2005. Diversity creation methods: a survey and categorisation. *Information Fusion* 6(1):5–20.

Collins, M.; Schapire, R. E.; and Singer, Y. 2002. Logistic regression, adaboost and bregman distances. *Machine Learning* 48(1-3):253–285.

Cover, T. M., and Thomas, J. A. 1991. *Elements of information theory*. Wiley.

Dagan, I., and Engelson, S. P. 1995. Committee-based sampling for training probabilistic classifiers. *Machine Learning Proceedings* 150–157.

Freund, Y., and Schapire, R. E. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, 119–139. Springer.

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems* 7(10):231–238.

Lichman, M. 2013. UCI machine learning repository.

Margineantu, D. D., and Dietterich, T. G. 1997. Pruning adaptive boosting. In *International Conference on Machine Learning*, volume 97, 211–218.

Martinez-Muoz, G.; Hernández-Lobato, D.; and Suarez, A. 2009. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2):245–259.

McCallumzy, A. K., and Nigamy, K. 1998. Employing EM and pool-based active learning for text classification. In *International Conference on Machine Learning*, 359–367.

Melville, P., and Mooney, R. J. 2004. Diverse ensembles for active learning. In *International Conference on Machine Learning*, 584–591.

Mukherjee, S.; Niyogi, P.; Poggio, T.; and Rifkin, R. 2003. Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*.

Rokach, L. 2010. Ensemble-based classifiers. *Artificial Intelligence Review* 33(1-2):1–39.

Settles, B. 2010. Active learning literature survey. *University of Wisconsin, Madison*.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, volume 284, 287–294. ACM.

Valiant, L. G. 1984. A theory of the learnable. *Communications of The ACM* 27(11):1134–1142.

Zhou, Z. H. 2012. *Ensemble methods: foundations and algorithms*. CRC Press.