Supervised Image Classification based on AdaBoost with Contextual Weak Classifiers

Ryuei NISHII
Kyushu University
Graduate School of Mathematics, Hakozaki, Higashiku, Fukuoka 812-8581, Japan
Email: nishii@math.kyushu-u.ac.jp

Shinto EGUCHI
Institute of Statistical Mathematics
Minami-azabu, Minatoku, Tokyo 106-8569, Japan
Email: eguchi@ism.ac.jp

Abstract — AdaBoost, one of machine learning techniques, is employed for supervised classification of land-cover categories of geostatistical data. We introduce contextual classifiers based on neighboring pixels. First, posterior probabilities are calculated at all pixels. Then, averages of the posteriors in various neighborhoods are calculated, and the averages are used as contextual classifiers. Weights for the classifiers can be determined by minimizing the empirical risk with multiclass. Finally, a linear combination of classifier is obtained. The proposed method is applied to artificial multispectral images and shows an excellent performance similar to the MRF-based classifier with much less computation time.

I. Introduction

AdaBoost proposed by [1] is one of machine learning techniques and has been progressed for pattern recognition rapidly and widely. AdaBoost combines weak classifiers into a weighted voting machine, and it shows high performance in various fields, see [2–4].

In this paper, we introduce AdaBoost for contextual image classification for geostatistical data. First, we prepare a posterior probability given feature vector at each pixel. The posteriors can be defined by machine learning techniques as well as statistical methods. Then, averages of the posteriors are calculated in various types of neighborhoods, and we positively use them as classifiers. The weights for the classifiers are obtained by minimizing the empirical risk with multiclass. Finally, a linearly combined classifier gives a contextual voting machine.

The proposed method is shown to be a very fast algorithm. It needs only 1/100 of time required by MRF-based classifiers implemented by [5]. The performance of our method keeps equivalent to that of MRF-based classifiers. Also, the proposed method is very flexible since the posteriors can be derived from various techniques: support vector machines (SVM), AdaBoost, and artificial neural networks (ANN), see e.g. [6–8].

In Section II, Real AdaBoost is explained. The determination of the coefficient of classifiers are discussed in a multiclass case. Section III defines contextual classifiers. Neighborhoods of a pixel and the averages of the posteriors in the neighborhoods are defined, which will be viewed as classifiers. The proposed method is applied to a numerical example in Section IV. It shows an excellent performance similar to MRF-based classifiers. Section V concludes the paper.

II. Real AdaBoost with multiclass

Let \( D = \{1, \ldots, n\} \) be a training area consisting of \( n \) pixels, and each pixel is belonging to one of \( g \) possible categories \( C_1, \ldots, C_g \). Suppose that an \( m \)-dimensional feature vector \( x_i \in \mathbb{R}^m \) is observed at each pixel \( i \). A label of the category covering the pixel \( i \) is denoted by \( y_i \in \{1, \ldots, g\} \).

Let a function \( f(x, y) \) be a (weak) classifier. We allocate a feature vector \( x \) into a category label:

\[
 y_f(x) = \arg \max_{y \in \{1, \ldots, g\}} f(x, y). \tag{1}
\]

In the ordinary setting of AdaBoost, the function \( f \) is restricted to take only zero or one as follows.

\[
 f(x, y) = \begin{cases} 
 1 & \text{if } y = y_f(x) \\
 0 & \text{otherwise.} 
\end{cases} \tag{2}
\]

Typical weak classifiers are decision stumps given by the functions \( \delta \text{ sign}(x^j - t) \) in binary class with label set \( \{-1, 1\} \), where \( \delta = \pm 1 \), \( t \in \mathbb{R} \), \( x^j \) is the \( j \)-th variate of the feature vector \( x \) and \( \text{sign}(z) \) is the sign of the argument \( z \).

We also consider the case such that the classifier \( f \) takes real values. For example, a posterior probability

\[
 p(y | x_i) = p(x_i | y) / \sum_{y' = 1}^{g} p(x_i | y') \tag{3}
\]

takes positive values, where \( p(\cdot | y) \) is a class-conditional probability density of the category \( C_y \).
AdaBoost aims to combine weak classifiers into a strong classifier. Let \( F = \{ f(x, y) \} \) be a set of weak classifiers. The loss of misclassification due to the classifier \( f \) is assessed by the following exponential loss function:

\[
L(x, y | y_0) = \exp \left[ f(x, y) - f(x, y_0) \right],
\]

for \( y \neq y_0, y = 1, \ldots, g \), where \( y_0 \) be the true label of the feature vector \( x \). The loss function (4) is an extension of the exponential loss with binary class. The empirical risk is defined by

\[
R_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} \sum_{y=1}^{g} \exp \left[ f(x_i, y) - f(x_i, y_i) \right],
\]

where \((x_i, y_i)\) are training data on the area \( D \).

Let \( f \) and \( F \) be classifiers (fixed). Then, the optimal coefficient \( c \) which gives the minimum value of the empirical risk \( R_{emp}(F + cf) \) is denoted by \( c^* \):

\[
c^* = \arg \min_c \{ R_{emp}(F + cf) \}.
\]

If the function \( f \) takes 0-1 values like (2), the optimal coefficient \( c^* \) can be expressed in the closed form. If \( f \) takes real values, there is no closed form to the optimal coefficient \( c^* \). We will take an iterative procedure for the estimation. The convergence of the procedure is, however, very fast.

Real AdaBoost procedure is as follows.

1. Find a weak classifier \( f \in F \) and the coefficient \( c \) which minimize the empirical risk \( R_{emp}(cf) \) defined by the formula (5), say \( f_1 \) and \( c_1 \).
2. Consider the empirical risk \( R_{emp}(c_1 f_1 + cf) \) with given \( c_1 f_1 \) in the previous step. Then, find the optimal classifier \( f \) and the coefficient \( c \) which minimize the empirical risk, say \( f_2 \) and \( c_2 \).
3. This is repeated \( T \)-times and obtain the final classifier \( F_T = c_1 f_1 + \cdots + c_T f_T \).

Then, a test vector \( x \) is classified into the label \( \arg\max_y F_T(x, y) \) for \( y \in \{1, \ldots, g\} \). We note that AdaBoost is a classifier based on a weighted majority vote.

III. Neighborhoods and Contextual Classifiers

As we have seen in Section II, AdaBoost requires multiple classifiers defined in the feature space. We will add contextual classifiers to the set of noncontextual classifiers so as to give contextual classification.

Let \( d(i, j) \) be a distance between centers of two pixels \( i \) and \( j \) in the domain \( D \). Then, define a subset \( U_r(i) \) of \( D \) with center \( i \) and radius \( r \) by

\[
U_r(i) = \{ j \in D \mid d(i, j) = r \}, \quad r = 0, 1, \sqrt{2}, 2, \ldots
\]

Note that the subsets \( U_r(i) \) are different to ordinary neighborhoods. See Fig. 1 for the subsets \( U_r(i) \) for \( r = 0, 1, \sqrt{2}, 2, \ldots \).

![Figure 1. Isotropic subsets \( U_r(i) \) with a center pixel \( i \) and radius \( r \) = 0, 1, \( \sqrt{2} \), 2.](image)

It is seen that \( U_0(i) = \{ i \} \), \( U_1(i) \) is the first-order neighborhood of the pixel \( i \), and \( U_1(i) \cup U_{\sqrt{2}}(i) \) forms the second-order neighborhood.

Suppose that a statistical model is applied to the distribution in the feature space. Then, the posterior probability \( p(y | x_i) \) defined by (3) is a measure of confidence of the current classifier. Such a measure is defined by machine learning techniques as well as statistical techniques. See [6-8] for posteriors introduced in SVM and AdaBoost.

Define the average of posterior probabilities in the subset \( U_r(i) \) by

\[
\bar{p}_r(y | x_i) = \left\{ \sum_{j \in U_r(i)} p(y | x_j) / |U_r(i)|, \right. \quad \text{if } |U_r(i)| > 0
\]

\[
0, \quad \text{otherwise}
\]

for \( r = 0, 1, \sqrt{2}, \ldots \), where \( |S| \) denotes the cardinality of set \( S \). Obviously, it holds that \( \bar{p}_0(y | x_i) = p(y | x_i) \). Hence, noncontextual classification is done by the posterior \( \bar{p}_0(y | x_i) \), which is a strong classifier. If the spatial dependency of the categories exists, the averaged posteriors \( \bar{p}_1(y | x_i) \) in the adjacent pixels to the pixel \( i \) also have information for classification. If the spatial dependency is strong, \( \bar{p}_r(y | x_i) \) with large \( r \) is also useful. Thus, we adopt the average of the posteriors \( \bar{p}_r(y | x_i) \) as a classifier of the center pixel \( i \).

Obviously, the importance of the posteriors for classifying the center pixel \( i \) will be in the following order:

\[
\bar{p}_0(y | x_i), \quad \bar{p}_1(y | x_i), \quad \bar{p}_{\sqrt{2}}(y | x_i), \ldots
\]

The coefficients to the classifiers \( \bar{p}_r(\cdot | \cdot) \) can be tuned by minimizing the empirical risk (5).

Thus, candidates of contextual classifiers are as follows.

(1a) The averaged posteriors \( \bar{p}_r(y | x_i) \) (6) in the subset \( U_r(i) \).
IV. NUMERICAL EXPERIMENTS

Our method is examined through multispectral images generated over the image (a) of Fig. 2 with three categories ($g = 3$). The labels 1, 2 and 3 correspond to the colors black, white and grey. The numbers of pixels from the categories are respectively given by 3330, 1371 and 3580. We simulate four-dimensional spectral images ($m = 4$) at each pixel of the true image (a) following multivariate normal distributions independently with mean vectors $\mu(1) = (0 0 0 0)^T$, $\mu(2) = (1 1 0 0)^T/\sqrt{2}$, $\mu(3) = (1.0498 - 0.6379 0 0)^T$ and with common variance-covariance matrix $\sigma^2 E_4$, where $E_4$ denotes the identity matrix. Test data are similarly generated over the same image (a), and the normal distributions are used for driving the posteriors.

Considering the importance order (7) of the averages, we applied the following classifiers: $c_0 \tilde{p}_0$, $c_0 \tilde{p}_0 + c_1 \tilde{p}_1$, ..., $c_0 \tilde{p}_0 + c_1 \tilde{p}_1 + \cdots + c_r \tilde{p}_r$ to test data, where the coefficient $c_r$ is sequentially tuned by minimizing the empirical risk $R_{\text{emp}}(c_0 \tilde{p}_0 + c_1 \tilde{p}_1 + \cdots + c_{r-1} \tilde{p}_{r-1} + c_r \tilde{p}_r)$ with respect to constant $c$.

Fig. 3 plots the coefficients $c_0, c_1, ..., c_\sqrt{\sigma^2}$ against radius $r$ in the case $\sigma^2 = 1$. Magnitude of the coefficients give a reliability of the corresponding classifiers. The coefficient $c_0$ is not large, and this is ascertained by the fact that the error rate due to $\tilde{p}_0$ (the noncontextual classifier) is 41.75%.

Fig. 4 shows test error rates due to the classifiers $c_0 \tilde{p}_0 + c_1 \tilde{p}_1 + \cdots + c_r \tilde{p}_r$ for $r = 0, 1, ..., \sqrt{50}$ with variance $\sigma^2 = 1$. The curve shows that the neighborhood information reduces the error rate rapidly, and it is stable even if the radius $r$ is too large.

Fig. 2 (b) is the image obtained by the posterior $\tilde{p}_0$, which is equivalent to the linear discriminant function. The classified images are getting closer to the true label (a) as the radius $r$ is getting larger.

Table I compares error rates due to Gaussian MRF-based (GMRF) and the proposed classifiers in two cases: $\sigma^2 = 1, 4$. Each row is corresponding to GMRF with the neighborhood system $U_1(i) \cup U_{\sqrt{2}}(i) \cup \cdots \cup U_r(i)$ and the classifier $c_0 \tilde{p}_0 + c_1 \tilde{p}_1 + \cdots + c_r \tilde{p}_r$. The third and fifth columns show CPU time (seconds) required to GMRF and the proposed method respectively. GMRF needs much time for tuning the dependency parameter [5], whereas the the proposed method determines the coefficient sequentially with several iterations. It is shown that the proposed method is very fast procedure compared with the ordinary MRF-based classifier, and that both the classifiers show similar performance.

V. CONCLUSION

Real AdaBoost is introduced to provide contextual image classification. Weak classifiers based on posteriors on
Figure 3. Estimated coefficients for the posteriors in ring regions

Figure 4. Error rates for training data due to the classifiers $c_0\bar{p}_0 + c_1\bar{p}_1 + \cdots + c_r\bar{p}_r$

Table I

<table>
<thead>
<tr>
<th>radius</th>
<th>$\sigma^2 = 1$</th>
<th>$\sigma^2 = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>41.75</td>
<td>54.56</td>
</tr>
<tr>
<td>$\sqrt{2}$</td>
<td>5.01</td>
<td>4.61</td>
</tr>
<tr>
<td>$\sqrt{2}$</td>
<td>3.57</td>
<td>5.23</td>
</tr>
<tr>
<td>$\sqrt{5}$</td>
<td>3.23</td>
<td>4.61</td>
</tr>
<tr>
<td>2</td>
<td>7.75</td>
<td>10.18</td>
</tr>
<tr>
<td>$\sqrt{10}$</td>
<td>6.10</td>
<td>5.12</td>
</tr>
<tr>
<td>$\sqrt{13}$</td>
<td>9.52</td>
<td>13.56</td>
</tr>
<tr>
<td>4</td>
<td>11.70</td>
<td>15.58</td>
</tr>
<tr>
<td>$\sqrt{20}$</td>
<td>21.78</td>
<td>20.58</td>
</tr>
<tr>
<td>$\sqrt{25}$</td>
<td>30.58</td>
<td>22.97</td>
</tr>
</tbody>
</table>

Table I shows the error rates (%) of classification results based on GMRF and Spatial Real AdaBoost.

Features of the proposed method are as follows.

- Various types of posteriors can be implemented.
- Various types of neighborhoods can be implemented.
- The method requires much less computation time and shows similar performance to MRF-based classifiers.

Selection of types of posteriors and neighborhoods is a new problem. Selection of the number of classifiers is also an old and new problem.

References


