Error rates in spatial classification of Gaussian data with random labeling

Lijana Stabingienė, Kęstutis Dučinskas

Klaipeda University, Department of Statistics
H. Manto 84, LT-92294 Klaipeda
E-mail: lijana.stabingiene@gmail.com; kestutis.ducinskas@ku.lt

Abstract. In spatial classification it is usually assumed that features observations given labels are independently distributed. We have retracted this assumption by proposing stationary Gaussian random field model for features observations. The label are assumed to follow Discrete Random Field (DRF) model. Formula for exact error rate based on Bayes discriminant function (BDF) is derived. In the case of partial parametric uncertainty (mean parameters and variance are unknown), the approximation of the expected error rate associated with plug-in BDF is also derived. The dependence of considered error rates on the values of range and clustering parameters is investigated numerically for training locations being second-order neighbors to location of observation to be classified.

Keywords: supervised classification, Gaussian Random Fields, spatial correlation.

Introduction

Spatial supervised classification is a problem of classifying locations (sites) into several categories by learning the features observation and the adjacency relationships with training sample. Switzer [5] was the first to treat classification of spatial data. It is usually assumed that feature observations are independent conditional on class labels (conditional independence) and normally distributed. This approach is widely used in image classification [3].

In the case of complete parametric certainty, the formula of exact error rate due to Bayes classification rule (BCR) under described assumptions is derived by Nishii and Eguchi [4]. In this paper we have derived the above formula by retracting the assumption of conditional independence. The observation of features to be classified is assumed to be dependent on a training sample.

The stationary Gaussian Random Fields (GRF) model for features and DRF model for class labels are considered. In the case of partial parametric uncertainty, the original approximation of the expected error rate associated with plug-in BDF is proposed. This is the generalization of the similar approximations derived in the case of training sample with fixed sampling design and fixed labels [1]. The numerical analysis of derived exact error rate and proposed approximation of the expected error rate is carried out in the case of isotropic exponential spatial correlation function among features observations. For the second-order neighborhood system, the influence of the some statistical model parameters on the values of considered error rates is numerically evaluated.
1 The main concepts and definitions

The main objective of this paper is to classify the feature observations modeled by stationary Gaussian random field \( \{Z(s): s \in D \subset \mathbb{R}^2 \} \).

The marginal model of observation \( Z(s) \) in class \( \Omega_l \) is

\[
Z(s) = \mu_l + \varepsilon(s),
\]

where \( \mu_l \) is constant mean and the error term is generated by zero-mean stationary Gaussian random field \( \{\varepsilon(s): s \in D \} \) with covariance function defined by the following model for all \( s, u \in D \)

\[
\text{cov} \{ \varepsilon(s), \varepsilon(u) \} = \sigma^2 r(s - u),
\]

where \( r(s - u) \) is the known spatial correlation function and \( \sigma^2 \) is variance as a scale parameter.

Let \( L = \{1, 2\} \) be a label set. A label of location \( s \in D \) associated with \( Z(s) \) is a random variable \( Y(s) \) taking values in \( L \). Let \( S_n = \{s_i \in D: i = 1, \ldots, n\} \) be a set of training locations. Set \( Y = (Y(s_1), \ldots, Y(s_n))^\prime \) and \( Z = (Z(s_1), \ldots, Z(s_n))^\prime \) and call them labels vector and features vector, respectively.

Thus, the vector \( T' = (Z', Y') \) constitutes the training sample.

Suppose that the event \( \{T = t\} \) is equivalent to the event \( \{Z = z\} \cap \{Y = y\} \), where \( t, z, y \) are the realizations of the corresponding random vectors.

Denote by \( R \) the matrix of spatial correlations among components of \( Z \). Suppose that \( S_n \) is fixed, but the labels are distributed randomly on it.

So for \( Y = y \), \( S_n \) is partitioned into the union of two disjoint subsets, i.e., \( S_n = S_y^{(1)} \cup S_y^{(2)} \) where \( S_y^{(l)} \) is the subset of \( S_n \) that contains \( n_l \) locations with labels equal \( l, l = 1, 2 \) \( (n_1 + n_2 = n) \).

Then the model of vector \( Z \) for given \( Y = y \) is

\[
Z = X_y \mu + E_n
\]  

where \( X_y \) is the \( n \times 2 \) design matrix, \( \mu' = (\mu_1, \mu_2) \) and \( E \) is the \( n \)-vector of random errors that has multivariate Gaussian distribution \( N_n(0, \sigma^2 R) \).

Consider the problem of classification (estimation of \( Y(s_0) \)) of the feature observation \( Z_0 = Z(s_0), s_0 \in D, s_0 \notin S_n \) with given training sample \( T \).

Denote by \( r_0 \) the vector of spatial correlations between \( Z_0 \) and \( Z \) given in (1). So we have to deal with conditional Gaussian distribution of \( Z_0 \) given \( T = t \) with means

\[
\mu_{0l}^t = E(Z_0|T = t, Y(s_0) = l) = \mu_l + \alpha_0^t(Z - X_y \mu),
\]

where \( \alpha_0^t \) is spatial correlation between \( Z_0 \) and \( Z \) at \( s_0 \).

\[
\sigma_{0l}^2 = V(Z_0|T = t, Y(s_0) = l) = \sigma^2 R_{0n}
\]

where \( \alpha_0^t = r_0^{-1} R_{0n}, R_{0n} = 1 - r_0^{-1} R_0 \).

2 Error rates of spatial classification

At the beginning we specify the DRF model for class labels.

Denote by \( \{\pi(y) = P(Y = y)\} \) the prior distribution of the labels vector \( Y \).

Liet. mat. rink. LMD darbai, 51:426–430, 2010.
Proposition 1 [Assumption]. The conditional distribution of \( Y(s_0) \) given \( T = t \) depends only on \( Y = y \), i.e., \( \pi_t(y) = P(Y(s_0) = l|T = t), t = 1, 2 \).

Under the assumption that the classes are completely specified the Bayes discriminant function (BDF) \([2]\) minimizing the probability of misclassification is formed by the logarithm of ratio of conditional densities described above. We shall call that situation the case of complete parametric certainty.

Then BDF for classification of \( Z_0 \) given \( T = t \) is

\[
W_t(Z_0) = \left( Z_0 - \frac{1}{2}(\mu_{1t} - \mu_{2t}) \right)'(\mu_{1t} - \mu_{2t})/\sigma_{0t}^2 + \gamma(y)
\]

where \( \gamma(y) = \ln(\pi_1(y)/\pi_2(y)) \).

The conditional Mahalanobis distance given \( T = t \) is

\[
\Delta_{0t} = \left( (\mu_{1t} - \mu_{2t}) \right)/\sigma_{0t} = \Delta_0/\sqrt{R_{0n}},
\]

where \( \Delta_0 = (\mu_{1} - \mu_{2})/\sigma \) is the marginal Mahalanobis distance. It is obvious that \( \Delta_{0t} \) depends on \( S_0 \) but does not depend on \( t \).

Then conditional Bayes error rate (for given \( T = t \)) of classifying \( Z_0 \) by BDF \( W_t(Z_0) \) is

\[
P_t(t) = \sum_{l=1}^{2} \pi_l(y)\Phi(-\Delta_{0t}/2 + (-1)^l\gamma(y)/\Delta_{0t}),
\]

where \( \Phi(\cdot) \) is the standard normal distribution function.

The exact Bayes error rate for \( W_t(Z_0) \) is

\[
E_T(P_t(T)) = \sum_{k=1}^{2} \sum_{l=1}^{2} \pi_l(y)\pi_l(y)\Phi(-\Delta_{0t}/2 + (-1)^l\gamma(y)/\Delta_{0t}),
\]

where \( E_T \) denotes the expectation with respect to \( T \) distribution and \( k \) differences between classes.

Suppose that means \( \{\mu_l\} \) and \( \sigma^2 \) are unknown and need to be estimated from training sample \( T \).

Let \( \hat{\mu} \) and \( \hat{\sigma}^2 \) be the estimates of \( \mu \) and \( \sigma^2 \), based on \( T = t \). Denote the three component vector of parameters by \( \Psi = (\mu, \sigma^2) \) and denote the vector of their estimates by \( \hat{\Psi} = (\hat{\mu}, \hat{\sigma}^2) \).

The plug – in BDF (PBDF) is obtained by replacing the parameters in BDF with their estimates based on \( T = t \). Then PBDF to the classification problem specified above is

\[
W_t(Z_0; \hat{\Psi}) = \left( Z_0 - \frac{1}{2}(\hat{\mu}_{1t} + \hat{\mu}_{2t}) \right)(\hat{\mu}_{1t} - \hat{\mu}_{2t})/\hat{\sigma}_{0t}^2 + \gamma(y)
\]

(2)

where \( \hat{\mu}_{lt} = E(Z_0|T = t; \ Y(s_0) = l) = \mu_l + \alpha_l(z_n - X_n\mu), l = 1, 2 \) and

\[
\hat{\sigma}_{0t}^2 = V(Z_0|T = t; \ Y(s_0) = l) = \hat{\sigma}^2 R_{0n}.
\]

In the considered case the actual error rate \([1]\) for \( W_t(Z_0; \hat{\Psi}) \) is specified by

\[
P_t(\hat{\Psi}) = \sum_{l=1}^{2} \pi_l(y)\Phi(\hat{Q}_l(t)),
\]

(3)
The proof of lemma is based on Taylor series expansion about points ϕ̂ where Z of is actually formed from training sample. Hence the EER for the considered problem is asymptotic scheme for spatial sampling.

Lemma 1. Suppose that observation Z₀ is to be classified by PBDF specified in (2) and let the assumptions of theorem [1] hold. Then the asymptotic approximation of EER defined in (5) is

\[ \text{AEP}_0 = \sum_{y} \sum_{l=1}^{2} \pi(y)\pi_l(y)\phi(-\Delta_{0n}/2 + (-1)^l\gamma(y)/\Delta_{0n}) + \sum_{y} \sum_{l} \pi_l(y)\pi(y)\phi(Q_l(y))(C(y) + 2\gamma^2(y)/(n-2))/\Delta_{0n} \]

where \( \phi(\cdot) \) denotes the standard normal density function and \( C(y) = \Lambda' R_{\mu} \Lambda \Delta_{0n}^2/\rho_0 \), \( \Lambda = X' \rho_0 - H/2 + \gamma(y)G/\Delta_{0n}^2 \).

Proof. The proof of lemma is based on Taylor series expansion about points \( \mu = \hat{\mu} \) and \( \sigma^2 = \hat{\sigma}^2 \) of \( P_T(\hat{\psi}) \) presented in (3), (4). Then taking the expectation of the main term of Taylor described above we complete the proof of lemma. For details see the proof of theorem [1].

3 Numerical example and conclusions

Here we analyze numerically the dependence of exact error rate on some statistical model parameters. Suppose D is 2-dimensional rectangular lattice with unit scaling, \( S_0 = (0,0) \) and \( S_8 \) is the set of second-order neighbors to \( S_0 \). We consider the case of model (3) with constant means and isotropic exponential spatial correlation function given by \( r(h) = \exp\{-|h|/\alpha\} \), where \( \alpha \) is the range parameter. Denote by \( \rho \) the clustering parameter or granularity [4]. The non-negative parameter \( \rho \) gives the degree of spatial dependency of the class labels.

Set \( Y = Y(s_i) \), \( y_i = y(s_i) \), \( i = 1, \ldots, n \). Assume that conditional distribution of \( Y(s_0) \) given \( Y = y \) is

\[ \pi_1(y) = P(Y(s_0) = 1|Y = y) = 1/(1 + \exp(\rho(1 - 2n_1/n))) \].

Liet. mat. rink. LMD darbai, 51:426–430, 2010.
Table 1. Values of $AEP_0$ for $\Delta_0 = 0.2$ in the upper cell and $\Delta_0 = 3$ in the bottom cell, $\pi_4 = 0.5, \pi_3 = \pi_5 = 0.15, \pi_2 = \pi_6 = 0.1$.

| $\rho \setminus \alpha$ | 0    | 0.5  | 1    | 2    | 3    |
|------------------------|------|------|------|------|------|
| 0                      | 0.45957 | 0.45132 | 0.43569 | 0.42269 |
| 0.4                    | 0.07479 | 0.03813 | 0.01009 | 0.00257 |
| 0.8                    | 0.07687 | 0.03933 | 0.01051 | 0.00269 |
| 1.2                    | 0.44256 | 0.43633 | 0.42479 | 0.41542 |
| 1.6                    | 0.08006 | 0.04124 | 0.01113 | 0.00287 |
| 2                      | 0.42771 | 0.42242 | 0.41396 | 0.40820 |
|                        | 0.08450 | 0.04403 | 0.01290 | 0.00311 |
|                        | 0.41229 | 0.40761 | 0.40222 | 0.40078 |
|                        | 0.09066 | 0.04782 | 0.01315 | 0.00341 |
|                        | 0.39718 | 0.39291 | 0.39078 | 0.39447 |
|                        | 0.09846 | 0.05271 | 0.01460 | 0.00380 |

and prior distribution of class labels is $\pi(y) = \pi_{n_1}/C_{n_1}$, where $n_1 = \#(i: y_i = 1, \ i = 1, \ldots, n)$ and $\pi_{n_1} = P(\sum_{i=1}^{n} 1\{Y_i = 1\} = n_1)$, $n_1 = 0, \ldots, n$.

So the results of numerical analysis give us arguments to state that the greater clustering of class labels and stronger spatial correlation between feature observations ensures the smaller spatial classification error.

References

[1] K. Dučinskas. Approximation of the expected error rate in classification of the Gaussian random field observations. Statistics and Probability Letters, 79:138–144, 2009.
[2] K. Fukunaga. Introduction to Statistical Pattern Recognition. Second e. Academic Press, New York, 1990.
[3] R. Nishii. A markov random field-based approach to decision level fusion for remote sensing image classification. IEEE Trans. Geosci. Remote Sensing, 41(10):2316–2319, 2003.
[4] R. Nishii and S. Eguchi. Image classification based on markov random field model with jeffreys divergence. Journal of Multivariate Analysis, 97(9):1997–2008, 2006.
[5] P. Switzer. Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. Math. Geol., 12(4):367–376, 1980.