Comparison of spatial classification rules with different conditional distributions of class label

Giedrius Stabingis, Kęstutis Dučinskas, Lijana Stabingienė

Department of Mathematics and Statistics, Klaipėda University
H. Manto str. 84, LT-92294 Klaipėda, Lithuania
gstabingis@gmail.com; kestutis.ducinskas@ku.lt; lijana.stabingiene@gmail.com

Received: 21 January 2013 / **Revised:** 15 July 2013 / **Published online:** 25 November 2013

Abstract. In this paper spatial classification rules based on Bayes discriminant functions are considered. The novelty of this work is that the statistical supervised classification method is improved by extending the influence of spatial correlation between observation to be classified and training sample. Such methods are used for data containing spatially correlated noise. Method accuracy is tested experimentally on artificially corrupted images. This classification rule with distance based conditional distribution for class label shows advantage against other classification rule ignoring such influence and against other commonly used supervised classification methods.

Keywords: Bayes discriminant functions, supervised classification, spatial dependency.

1 Introduction

The incorporation of the spatial information (image texture, direction, closeness and other) into image classification is highly potential [1]. In the series of papers (see e.g., [2–4]) the incorporation of geostatistical information of features into plug-in versions of classifiers is based on the marginal distribution of the observation to be classified. Thus the geostatistical Bayes classifiers based on conditional feature distribution of the observation to be classified were investigated.

Author Dučinskas [5] discusses the problems of classification for Gaussian random field observations. In his paper he offer to take into account the spatial closeness of the points, i.e., the correlation of their feature values, which is typical to the spatial data. According to the Tobler's first law of geography: everything is related to everything else, but near things are more related then distant things [6]. The formal property that describes this is spatial autocorrelation. Spatial autocorrelation represents the degree to which that correlation changes with distance [7].

In this paper the statistical supervised classification method [8] is extended by incorporating more influence from the spatial dependency into classification problem and achieving higher accuracy. This method can be used in image analysis and in other fields, where data used for classification is corrupted by spatially correlated noise.

In the earlier papers of authors [9] conditional independence assumption is changed. Also the observation of features to be classified is assumed to be dependent to the features observations in training sample. The stationary Gaussian Random Field (GRF) model for features and discrete Markov Random Field (MRF) model for class label are considered.

The earlier method was applied in papers [8, 10] and the defended doctoral thesis [11] was based on it. This method was applied in the real situation for the remotely sensed image, covered with clouds, classification [12]. The error rates for the earlier method were presented and investigated in [13]. The error rates of the multivariate case was presented and investigated in [14].

The main idea of the method proposed in this paper is the following. In the classification methods proposed by the authors, the prior class probabilities are calculated only from the neighbor observations of the observation to be classified. These neighbor observation are from the training sample and based on some neighboring scheme. Usually such prior class probabilities are calculated only according to the amount of the observations belonging to concrete class. But these observations are far from the observation to be classified in different distances and this must be also added into account while calculating prior probabilities. In such way the prior information is evaluated more accurately when the observations of some class are closer to the observation to be classified then the observations from the other classes.

In this paper when assigning the object into one of the classes the classification rule with distance based posterior distribution for class label is used. Advantage of classification rule with distance based posterior distribution for class label against one ignoring spatial proximity between locations is shown visually and confirmed numerically.

In order to verify the reliability of the method the large experiment is performed. During this experiment artificially corrupted images of different symbols are classified with the method proposed in this paper, with the earlier method of the authors with less influence of the spatial dependency and with other commonly used supervised classification methods. These commonly used methods are not investigated in this paper, they are used only for the general orientation of the supervised classification methods. The comparison of these methods lets other researchers who are using supervised classification methods to evaluate better the methods proposed by the authors. The classification with these methods is done on the same data and on the same training sample as for the methods proposed by the authors. These commonly used supervised classification methods are: Support Vector Machines (SVM), Neural Networks (NNet), Random Forests (RF) and Multinomial Logistic Regression (Logit). The advantage of SVMs classifiers is their capabilities to learn from small number of samples [15]. The class (label) of a new sample is determined by a linear combination of the kernel functions evaluated on a certain subset of the examples the support vectors and the input. The coefficients of the combination are obtained as a solution to a convex optimization problem occurring at the learning stage [16]. Neural networks rely on the iterative derivation of weights which effectively define hyper-planes and hyper-regions in the pattern feature space [17]. Although artificial neural network methods are frequently found to give a higher total classification accuracy when compared to other methods, they do not always perform universally well [18]. Random Forests method grows many classification trees and then every tree gives the class label for the observation to be classified. Then the class is assigned according to the most classes given by all trees in the forest [19].

Numerical and visual analysis of proposed discriminant function in the case of isotropic exponential spatial correlation for the nearest neighbor neighborhood system using eight nearest neighbors *NN*(8) is done. All calculations are done in R system [20]. For the commonly used supervised classification methods the rassclass package from the R system is used.

2 Method description

In this paper extended statistical supervised classification method is presented. Spatial classification rule based on the plug-in Bayes discriminant function with posterior distribution of class label ignoring the distances among the locations is denoted by SCR. The extended method depending on distances among unclassified locations and training sample locations is called SCRD (Spatial Classification Rule with Distance).

In this paper features are modeled by stationary Gaussian random field (GRF) $\{Z(s)\colon s \in D \subset \mathbb{R}^2\}$, and class labels are modeled by discrete Markov random field (MRF). Such modeling is common in image analysis. In the context of image analysis index s means pixel.

The marginal model of observation $Z(s)$ in class Ω_l is $Z(s) = \mu_l + \varepsilon(s)$, where μ_l is the mean, and the error term $\varepsilon(s)$ is generated by zero-mean stationary Gaussian random field $\{\varepsilon(s)\colon s \in D\}$ with covariance function defined by model $\operatorname{cov}\{\varepsilon(s), \varepsilon(u)\} = \sigma^2 r(s-u)$ for all $s, u \in D$, where $r(s-u)$ is the spatial correlation function and σ^2 is variance as a scale parameter. During the experiments, the exponential covariance function is used:

$$C(h) = \sigma^2 \exp\left\{-\frac{|h|}{\alpha}\right\}, \tag{1}$$

where α is the correlation range parameter. $r(s-u) = r(h) = \exp\{-|h|/\alpha\}$, where h is the Euclidean distance between s and u locations.

Let $L = \{1, 2\}$ be a label set. A label of pixel $s \in D$ associated with $Z(s)$ is a random variable $Y(s)$ taking values in L. Let $S_n = \{s_i \in D;\ i = 1, \dots, n\}$ be a set of training pixels. Set $Y = (Y(s_1), \dots, Y(s_n))'$ and $Z = (Z(s_1), \dots, Z(s_n))'$ 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.

Assume that the model of Z for given $Y = y$ is $Z = X_y \mu + E$, where X_y is a design matrix, $\mu' = (\mu_1, \mu_2)$ and E is the n-vector of random errors that has multivariate Gaussian distribution $\mathcal{N}_n(0, \sigma^2 R)$. 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 $T = t$. Here s_0 is the location of the observation to be classified. $Z_n = (Z(s_i \mid s_i \in S_n))$ is the feature vector from the neighbor observations of the observation to be classified. Denote by r_0 the vector of spatial correlations between Z_0 and Z_n. Also denote by R the matrix of spatial correlations among components of Z_n. R and r_0 components are calculated according to the Eq. (1). Since Z_0 is correlated with training sample, we have to deal with conditional Gaussian distribution of Z_0 given $T = t$ $(Z = z, Y = y)$ with means μ_{lt}^0 and variance σ_{0t}^2 that are defined by

$$\begin{aligned} \mu_{lt}^0 &= \mathbf{E}\big(Z_0 \bigm| T = t;\ Y(s_0) = l\big) = \mu_l(s_0) + \alpha_0'(z_0 - X_y\mu), \\ \sigma_{0t}^2 &= \mathbf{V}\big(Z_0 \bigm| T = t;\ Y(s_0) = l\big) = \sigma^2 R_{0n}, \end{aligned} \tag{2}$$

where $\alpha_0' = r_0' R^{-1}$, $R_{0n} = 1 - r_0' R^{-1} r_0$ and $l = 1, 2$.

In this methodology assumption that the posterior distribution of $Y(s_0)$ given $T = t$ depends only on $Y = y$ and N_0 is made. The posterior distribution of $Y(s_0)$ is

$$\pi_l(y) = \mathbf{P}\big(Y(s_0) = l \bigm| T = t\big), \quad l = 1, 2.$$

Suppose that means $\{\mu_l(s)\}$ and σ^2 are unknown and need to be estimated from training sample T. Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the estimates of μ and σ^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 Bayes discriminant function (PBDF) is obtained by replacing the parameters in Bayes discriminant function (BDF) with their estimates based on $T = t$. Then PBDF to the classification problem specified above is

$$W_t(Z_0; \hat{\Psi}) = \left(Z_0 - \frac{\hat{\mu}_{1t}^0 + \hat{\mu}_{2t}^0}{2}\right) \frac{\hat{\mu}_{1t}^0 - \hat{\mu}_{2t}^0}{\hat{\sigma}_{0t}^2} + \gamma(y), \tag{3}$$

where $\hat{\mu}_{lt}^0 = \hat{\mu}_l + \alpha_0'(z_n - X_y\hat{\mu})$ and $\hat{\sigma}_{0t}^2 = \hat{\sigma}^2 R_{on}$, $\gamma(y) = \ln(\pi_1(y)/\pi_2(y))$.

SCR is denoted the classification rule based on the posterior distribution of $Y(s_0)$ specified by

$$\pi_1(y) = \frac{1}{1 + \exp\{\rho(1 - \frac{2n_1}{n})\}}, \quad n_1 = 0, 1, \ldots, n, \tag{4}$$

when $I_0 = \{i\colon\ s_i \in N_0 = \mathit{NN}(8),\ i = 1, \ldots, n\}$ and where ρ is non negative constant called a clustering parameter, and n_1 is the number of locations from N_0 with label equal 1. Here $\mathit{NN}(8)$ is the nearest neighbor scheme with eight nearest neighbors.

In this paper proposed spatial classification rule SCRD is based on the following posterior distribution

$$\pi_1(y) = \frac{\sum_{i \in I_0} \frac{\delta(y_i = 1)}{d(s_i, s_0)}}{\sum_{i \in I_0} \sum_{j=1}^2 \frac{\delta(y_i = j)}{d(s_i, s_0)}} = \frac{\sum_{i \in I_0} \frac{\delta(y_i = 1)}{d(s_i, s_0)}}{\sum_{i \in I_0} \frac{1}{d(s_i, s_0)}}, \tag{5}$$

where $\delta(\cdot)$ is the 0-1 indicator function and $d(\cdot\,, \cdot)$ denotes the Euclidean distance function between locations. For the case of two classes $\pi_2 = 1 - \pi_1$.

According to the situation illustrated in Fig. 1 using SCR method the probabilities of class labels are $\pi_1 = 1/2$ and $\pi_2 = 1/2$. While using SCRD method the prior probability

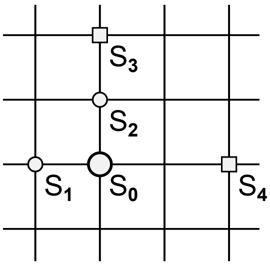

Fig. 1. Illustration of situation.

of the first class increases because s_1 and s_2 locations are closer to the observation to be classified s_0 then the locations s_3 and s_4 from the second class. According to the SCRD method for the situation illustrated in Fig. 1 π_1 prior probability is calculated:

$$\begin{aligned}\pi_1(y) &= \frac{\left(\frac{\delta(y_1=1)}{1} + \frac{\delta(y_2=1)}{1} + \frac{\delta(y_3=1)}{2} + \frac{\delta(y_4=1)}{2}\right)}{\sum_{i=I_0}\left(\frac{\delta(y_i=1)}{d(s_i,s_0)} + \frac{\delta(y_i=2)}{d(s_i,s_0)}\right)} = \cdots \\ &= \frac{2}{\left(\frac{1}{1}+\frac{0}{1}\right)+\left(\frac{1}{1}+\frac{0}{1}\right)+\left(\frac{0}{2}+\frac{1}{2}\right)+\left(\frac{0}{2}+\frac{1}{2}\right)} = \frac{2}{3}.\end{aligned}$$

3 Method illustration

The aim of this experiment is to test the overall accuracy of proposed method. The artificially corrupted images of different symbols are used in this experiment for classification. All calculations are done with statistical computing software R [20].

3.1 Preparation for the experiment

In this experiment 100 different images are used. The dimensions of every image are 200×200 pixels. Every initial image consists of pure white and pure black color pixels. All these images are prepared outside the R software and read using rtiff package. This package reads .tiff type image and produces the number matrix corresponding image pixels. With this package black pixels become 0 and white pixels become 1 inside of the number matrix. All other gray level pixel are from the interval $(0, 1)$. Several of initial images are shown in figure Fig. 2.

According to every initial image the training sample is randomly generated for every of the classes. 0.8% of image points are used in training sample. That is only about 320 from 40000 points are taken for the training sample. The first class is sampled from white pixels and the second from black pixels. Sampling is done proportionally for each class.

Further initial images are corrupted by spatially correlated Gaussian random fields. Random fields are generated with geoR package inside R software using isotropic exponential covariance function, and variance equal to 1. Also, during this experiment, the influence of correlation range parameter α is investigated. So each of the initial images

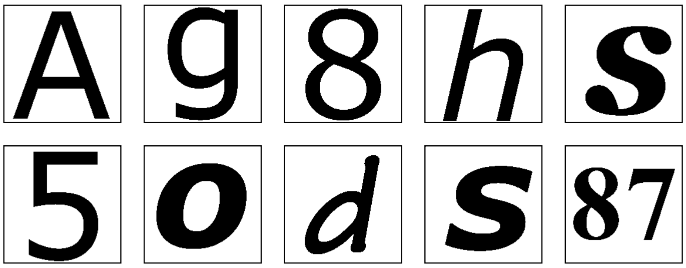

Fig. 2. Several of initial images used in experiment. 100 such different initial images are used in the experiment.

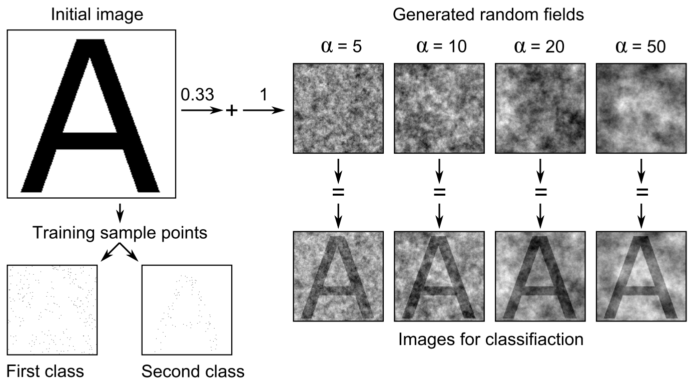

Fig. 3. Experiment preparation scheme. 100 different images, for 4 different α values, are prepared according to this scheme for the experiment.

is corrupted by four different Gaussian random fields where spatial correlation range parameter α is equal to 5, 10, 20 and 50. Four different Gaussian random fields generated with different α parameter values are shown in Fig. 3. These four Gaussian random fields are generated separately for every initial image.

Every generated random field is a number matrix and this matrix is normalized in order to gain values between 0 and 1. Then such field is combined with initial image by summing their matrix values with proportion $1 : 0.33$. It means that during this summation the random field matrix is taken as is – with values between 0 and 1, but initial image matrix is multiplied by 0.33 before summation. This is done in order to get the corrupted image which is corrupted hard enough for such classification problems. After this summation the resulting matrix is once more normalized in order to gain values between 0 and 1. These data normalizations are done in order to get the situation similar to the situation when real corrupted image is read. The whole experiment preparation scheme is shown in Fig. 3.

According to the preparation scheme presented in Fig. 3, 100 different images were prepared for classification. This gives us 400 different images because of 4 different α values which were investigated. All these 400 images are used for the classification.

3.2 Results of the experiment

During the experiment all 400 different corrupted images are classified by 6 different classification methods. As described above the classification process is done using 6 different classification methods. Two of them are methods of this letter authors. These are SCR – the older method of the authors and SCRD – the method proposed in this letter. Other four methods are supervised classification method commonly used for image per pixel classification. These methods are Logit, RF, NNet and SVM. After the classification lots of resulting images were obtained. One of the classification sets with letter "B" is presented in Fig. 4.

According to the visual classification results presented in Fig. 4 it can be shown that older method of authors (SCR) performs better when the other commonly used classification methods. The new classification method SCRD – proposed in this letter is even better. The classification errors appear at the same places for both methods but for the new method these error places are smaller. Also from the visual results it ca be seen that results become better for the authors methods when spatial correlation range parameter α increases. For the other commonly used methods this situation is opposite.

In this letter only one set off classified images is presented, but other sets show very similar results. After the classification of all the images all the resulting images are analyzed numerically. The average classification accuracy, standard deviation of classification accuracy, the minimum and the maximum classification accuracies are calculated for every method and presented in the Table 1.

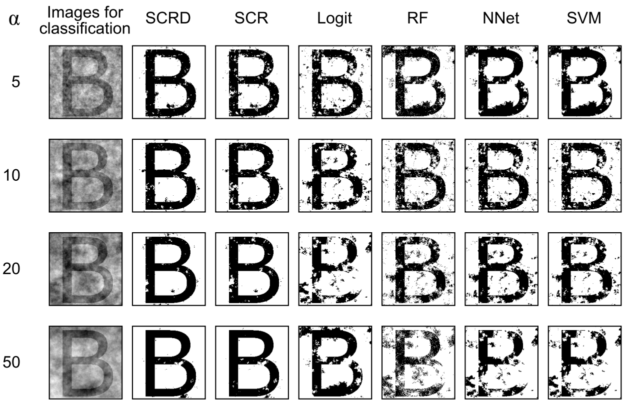

Fig. 4. Visual results of the experiment. Such classification sets are obtained for 100 different initial images.

Table 1. Overall classification accuracy results.

Method	SCRD	SCR	Logit	RF	NNet	SVM
Average	0.984	0.977	0.908	0.888	0.907	0.908
σ	0.009	0.014	0.030	0.037	0.030	0.030
Min	0.946	0.925	0.759	0.731	0.761	0.756
Max	0.999	0.999	0.978	0.965	0.977	0.978

Table 2. Average and minimum classification accuracy according to parameter α.

Methods:	SCRD	SCR	Logit	RF	NNet	SVM
α	Average classification accuracy					
5	0.973	0.962	0.927	0.910	0.926	0.926
10	0.980	0.971	0.915	0.896	0.914	0.915
20	0.988	0.984	0.905	0.882	0.904	0.904
50	0.994	0.993	0.886	0.863	0.886	0.886
α	Minimum classification accuracy					
5	0.946	0.925	0.879	0.858	0.878	0.881
10	0.958	0.943	0.863	0.841	0.857	0.859
20	0.972	0.963	0.817	0.753	0.814	0.815
50	0.976	0.972	0.759	0.731	0.761	0.756

As it was mentioned before, the influence of the parameter α is tested in this paper. The results of the overall classification accuracy for increasing α parameter is presented in Table 2.

According to the numerical results presented in Tables 1, 2 it can be stated that the method SCRD proposed in this paper is better then the older method of the authors (SCR). Also the classification accuracy for SCRD and SCR methods increases for the larger α values while other commonly used classification methods looses their accuracy.

4 Conclusions

The quite large classification experiment is performed, so the results can be considered as representative enough for such classification problems.

Visual and numerical results show that the incorporation off more spatial dependencies into the classification problem increases the classification accuracy.

The SCRD method, proposed in this letter, is more accurate when the other method SCR, which is also proposed by the authors.

According to the visual analysis the new method (SCRD) gets errors at the same places like the older authors method (SCR) but the area off these errors is significantly decreased.

Other commonly used supervised classification methods are influenced by the spatial correlation. The experiment results show that the accuracy of these methods decreases then the spatial correlation grows.

This new method can also be used for all image per pixel and other spatial supervised univariate classification problems especially when data consists spatially correlated noise (variation).

References

1. R.M. Haralick, Statistical and structural approaches to texture, *Proc. IEEE*, **67**:786–804, 1979.

2. M.A. Oliver, R. Webster, A geostatistical basis for spatial weighting in multivariate classification, *Math. Geol.*, **21**:15–35, 1989.

3. P.M. Atkinson, Spatially weighted supervised classification for remote sensing, *Int. J. Appl. Earth. Obs.*, **5**:277–291, 2004.
4. P.M. Atkinson, D.K. Naser, A geostatistically weighted k-NN classifier for remotely sensed imagery, *Geogr. Anal.*, **42**:204–225, 2010.
5. K. Dučinskas, Approximation of the expected error rate in classification of the Gaussian random field observations, *Stat. Probab. Lett.*, **79**:138–144, 2009.
6. W. Tobler, A computer movie simulating urban growth in Detroit region, *Econ. Geogr.*, **46**:234–240, 1970.
7. J.G. Liu, P.J. Mason, *Essential Image Processing and GIS for Remote Sensing*, Wiley-Blackwell, UK, 2009.
8. K. Dučinskas, L. Stabingienė, G. Stabingis, Image classification based on Bayes discriminant functions, *Procedia Environ. Sci., Spatial Statistics 2011 – Mapping Global Change*, **7**:218–223, 2011.
9. L. Stabingienė, G. Stabingis, K. Dučinskas, Comparison of the classification methods for the images modeled by Gaussian random fields, *Liet. mat. rink. LMD darbai*, **52**:200–204, 2011.
10. L. Stabingienė, G. Stabingis, K. Dučinskas, Comparison of linear discriminant functions in image classification, *Liet. mat. rink. LMD darbai*, **51**:227-231, 2010.
11. L. Stabingienė, *Image Analysis Using Bayes Discriminant Functions*, PhD Dissertation, Vilnius, 2012.
12. L. Stabingienė, Classification of the real remotely sensed image covered with clouds, *Liet. mat. rink. LMD darbai*, **53**:117–122, 2012.
13. K. Dučinskas, L. Stabingienė, Expected bayes error rate in supervised classification of spatial gaussian data, *Informatica*, **22**(3):371–381, 2011.
14. K. Dučinskas, Error rates in classification of multivariate Gaussian random field observation, *Lith. Math. J.*, **51**(4):477–485, 2011.
15. A.K. Jain, P.W. Duin, J. Mao, Statistical pattern recognition: A review, *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**(1):4–37, 2000.
16. W.W.Y. Ng, A. Dorado, D.S. Yeung, W. Pedrycz, E. Izquierdo, Image classification with the use of radial basis function neural networks and the minimization of the localized generalization error, *Pattern Recognition*, **40**:19–32, 2007.
17. R.P. Lippmann, An introduction to computing with neural nets, *IEEE ASSP Magazine*, **2**:4–22, 1987.
18. I. Kanellopoulos, G.G. Wilkinson, J. Megier, Integration of neural network and statistical image classification for land cover mapping, in: *Proceedings of International Geoscience and Remote Sensing Symposium, Tokyo, 18–21 August 1993*, Vol. 2, IEEE, 1993, pp. 511–513.
19. L. Breiman, Random forests, *Mach. Learn.*, **45**(1):5-32, 2001.
20. D. Wiesmann, D. Quinn, rasclass: Supervised Raster Image Classification, The Comprehensive R Archive Network, `http://cran.r-project.org/web/packages/rasclass/index.html`, 2012.