Vilnius university
Institute of Data Science and Digital Technologies
L I T H U A N I A

INFORMATICS (N009)

CONTEXTUAL CLASSIFICATION FOR DATA SPECIFIED BY SPATIAL GENERALIZED LINEAR MODELS

Eglė Zikarienė

October 2020

Technical Report DMSTI-DS-N009-19-<no.>

VU Institute of Data Science and Digital Technologies, Akademijos str. 4, Vilnius
LT-08412, Lithuania
www.mii.lt

Abstract

This paper covers empirical part of the doctoral thesis. Empirical research description, algorithm of classification procedure, simulated data description is presented.

Keywords: Bayes discriminant function, Classification algorithm, Conditional Beta distribution.

Contents

1 Introduction 4
2 Conduct Empirical research 4
3 Algorithm 4
3.1 Inputs 5
3.2 Description of functions 6
4 Simulated data 7
5 Real data 8
6 References 9

1 Introduction

An empirical research question: proposed reliability evaluation for contextual classification procedure. This question determines research objectives: construct an algorithm for classification, estimate probability of misclassification and estimate performance of classifiers, examine algorithm performance for a generated data set and apply the algorithm for solving real data problem.

2 Conduct Empirical research

The objective of the research is to empirically estimate classification error rates by simulating different initial situations and perform a comparison. In this section we present a particular and planned design for the research, which depend on the question and we offer ways of answering.

1. Construct the algorithm:
1.1. Statistical analysis for data
1.2. Initial data input description and format description (choosing neighbour schemes).
1.3. Unknown parameter estimation (Unknown model parameter estimation procedure)
1.3. Description of a procedure for BDF value estimation (conditional density function, prior probability evaluation, prior probability estimation evaluation methods)
1.4. Decision making procedure
1.5. Classification error rate probability estimation procedure (evaluating different classification error probability).
2. Research algorithm performance for a generated data set
2.1. Generate data set. Gibbs sampling
2.2. Introduce initial conditions
2.3. Present the results
2.4. Compare the results for different initial conditions
3. Apply the algorithm for real data
3.1. Describe the problem
3.2. Introduce initial conditions
3.3. Apply the proposed algorithm
3.4. Present the results
3.5. Compare the results for different initial conditions

3 Algorithm

This algorithm is designed for spatial data classification using BDF procedure for generalized linear model case when a model belongs to exponential distribution family with beta distribution. Data is analysed utilizing their conditional distributions using MRF property. The algorithm is as shown in Algorithm 1:

Algorithm 1:

Inputs: Data set $\{Z(s): s \in D \subset R^p\}$, Model M for population Ω_l, parameter estimation function *f*, prior probability function *g*, neighborhood system $\partial s \subset D$, $Z(s_0)$ – classification observation.

Algorithm: Choose the model M to the data $\{Z(s): s \in D \subset R^p\}$

Estimate unknown parameters: $\hat{\Psi} = f(Z(s), \Psi)$

Evaluate the prior probability function for class l: $\hat{\pi}_l = g\left(\left\{Z(s)|\mathbf{t}, y = l\right\}, \partial s\right)$

Evaluate the BDF function: $W\left(Z(s_0), \hat{\mathbf{\Psi}}\right)$

Make decision for $Z(s_0)$: $\hat{W}_{lk}^B\left(Z(s_0), \hat{\mathbf{\Psi}}\right) \geq 0,\ l = 1,..,m, k \neq l$

Evaluate probability of misclassification (actual error rate): $P_0^B\left(\hat{\mathbf{\Psi}}\right)$

Outputs: Class label for $Z(s_0)$

Probability of misclassification: $P_0^B\left(\hat{\mathbf{\Psi}}\right)$

Firstly, the data is needed for utilizing the algorithm. The initial data needs to be statistically estimated by finding outliers, correlation relations, proper data model must be chosen. Initial data is introduced together with input information described below.

3.1 Inputs

In this section the user presents initial data that has statistical analysis performed and data model chosen. Beta distribution model must be chosen for the data. In this algorithm the conditional distributions having Markov property in pairwise interactions case are analysed. In this part we choose neighbourhood system.

Probability density function

We focus on auto-beta models, this class of spatial model is constructed under two assumptions: first, the dependence between sites is pairwise and secondly, the full conditionals belong to some exponential family. Assume that for spatial auto-beta model scheme with spatial cooperation (see Hardouin and Yao, 2008), the full conditional density function for feature at location s_i $i = 0,..., n$ in reparametrized form is

$$p\left(Z_0^l = z_0 | T_{-i} = t_{-i}, y_i = l; \mu_i^l \phi_i^l\right) = \exp\left\{\left(\mu_i^l \phi_i^l\right) \ln z_0 + \left(\left(1-\mu_i^l\right)\phi_i^l - 1\right) \ln\left(1 - z_0\right)\right\} \times$$
$$\times \exp\left\{-\ln\left(B\left(\mu_i^l \phi_i^l; \left(1-\mu_i^l\right)\phi_i^l\right)\right)\right\},$$

with conditional mean - μ_i^l and, conditional precision - ϕ_i^l, where

$$\mu_i^l = E\left(Z_i | T_{-i} = t_{-i}, y_i = l\right) = \frac{1 + A_{i1}^l}{2 + A_{i1}^l + A_{i2}^l} \text{ and } \phi_i^l = 2 + A_{i1}^l + A_{i2}^l.$$

Natural parameters:

$$A_{i1}^l = \alpha_{i1}^l - \sum_{i \neq j} \eta_{ij} \ln\left(1 - z_j\right) \quad A_{i2}^l = \alpha_{i2}^l - \sum_{i \neq j} \eta_{ij} \ln\left(z_j\right)$$

Large scale in homogeneities can be modeled via $\alpha_{ik}^l = x'_i \beta_k^l$ where x'_i denotes m vector of explanatory variables at location s_i and β_k^l unknown regression coefficients it is matrix $k \times m$ for class l. Small scale variation $\eta_{ij} = d_k^l \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $d_k^l > 0$, if $d_k^1 = d_k^2$ no differs between class, d^l implies spatial symmetry in class l, d_k^l allow possible anisotropy between the horizontal and vertical directions, $i = 0,1,..., n$, $k, l = 1, 2$. Denote the set of all model parameters by $\Psi = \left(\left\{\beta_k^l\right\}; \left\{\eta_{ij}\right\}\right)$.

Sufficient statistics $T(z_i)=\left[\log(z_i),\log(1-z_i)\right]^T$ and log-partition function $\psi\left(A_{i1}^l,A_{i2}^l\right)=\log B\left(A_{i1}^l+1;A_{i2}^l+1\right)$. We propose the method of maximum pseudo-likelihood to estimate the parameters of the multi-parameter auto- beta model.

Neighbourhood system
First consider the scheme with four nearest neighbours on a two dimension lattice. $S=[1,N]\times[1,N]$: each site $i\in S$ has four neighbours denoted as $\left\{i_e=i+(1,0),\ i_w=i-(1,0),\ i_n=i+(0,1),\ i_s=i-(0,1)\right\}$ with obvious neighbour adjustments near the boundary. In the next step we enlarge the model to a scheme with eight nearest neighbours. Each site then has four more neighbours $\left\{i_{ne}=i+(-1,1),\ i_{nw}=i+(1,1),\ i_{sw}=1+(1,-1),\ i_{se}=1+(1,1)\right\}$ with neighbour adjustments near the boundary. In the next step we enlarge the model to a scheme with twelve nearest neighbours, third order system. Each site then has four more neighbours $\left\{i_{e3}=i+(2,0),\ i_{w3}=i-(2,0),\ i_{n3}=i+(0,2),\ i_{s3}=i-(0,2)\right\}$ with neighbour adjustments near the boundary. Note that in this case, some cliques has three or four elements but we consider pairwise interactions only.

The experiment for data set in the study is performed for every neighbourhood scheme separately. The results are compared.

3.2 Description of functions
In this section functions required for empirical study algorithm implementation are briefly described. The functions with empirical study conditions are given.

Parameter estimation
Parameter estimation for a Markov random field has been studied. The method of maximum likelihood unfortunately needs computer-intensive approximations, since the likelihood function is known only up to a constant that involves the parameters. As a remedy, Besag (1974, 1977) proposed the method of maximum pseudo-likelihood. For the auto-Beta model the normalizing constant is intractable. However, the pseudolikelihood has form:

$$\log\left(L(\Psi)\right)=\sum_{i=1}^{n}\left(A_{i1}^l(\bullet)\cdot\log\left(z(s_i)\right)+A_{i2}^l(\bullet)\log\left(1-z(s_i)\right)-\log B\left(A_{i1}^l+1;A_{i2}^l+1\right)\right),$$

where $A_{i1}^l=x_i\beta_k^l-\sum_{i\neq j}d_k^l\log\left(1-z(s_j)\right)$, $A_{i2}^l=x_i\beta_k^l-\sum_{i\neq j}d_k^l\log\left(z(s_j)\right)$. So using optimization procedure submitted in software R we get estimation:

$$\hat{\Psi}=\arg\max_{\Psi}\left\{\log\left(L(\Psi)\right)\right\}.$$

These estimates we will plug-in into BDF expression.

Prior probability estimation
The prior probabilities depend on the location of focal observation and the number of neighbors (only the closest vs all training sample). The formula for the prior probability for population Ω_1 is

$$\pi_1^0 = \sum_{j\in NN_i^1(k)}\left(\frac{1}{d_{ij}}\right)\Big/\sum_{j\in NN_i(k)}\left(\frac{1}{d_{ij}}\right),$$

where d_{ij} is the distance between sites s_i and s_j, $i,j = 1,\dots,n$, $NN_i(k) = NN_i^1(k) + NN_i^2(k)$, where $NN_i^l(k)$ is the set of sites belonging to the $k-th$ order of neighborhood of s_i in population Ω_l, $l = 1,2$.

BDF estimation

Bayes discriminant function expression for the auto-beta model:

$$W(Z_0,\Psi) = \ln\left(\frac{\pi_0^1 p_{01}}{\pi_0^2 p_{02}}\right) = \left(\alpha_{01}^1 - \alpha_{02}^2\right)\ln(z_0) + \left(\alpha_{01}^1 - \alpha_{02}^2\right)\ln(1-z_0) + \gamma_0(\Psi)$$

where $\gamma_0(\Psi) = \ln(u)$, $u = \dfrac{\pi_0^1 B\left(A_{01}^2+1, A_{02}^2+1\right)}{\pi_0^2 B\left(A_{01}^1+1, A_{02}^1+1\right)}$. By replacing the parameters with their estimators in $W(Z_0,\Psi)$, construct the sample BDF. So when we have the BDF expression, we can take a decision and estimated probability of misclassification.

Decision and probability of misclassification

So BDF allocates the observation in the following way: classifies observation Z_0 given $Z = z$ to the population Ω_1 if $W(Z_0,\Psi) \geq 0$,and to the population Ω_2, otherwise.

The actual error rate for SBDF $W(Z_0,\hat{\Psi})$ is

$$AR(\hat{\Psi}) = \sum_{l=1}^{2}\pi_l^0 \hat{P}_l,$$

where, for $l = 1,2$, $P_l = P_{lz}((-1)^l W_z(Z_0,\Psi) < 0)$,i.e.

$$\hat{P}_1 = \int_W p_{01}(t)dt = \int_0^1 H(-W(Z_0,\hat{\Psi}))p_{01}(t)dt,$$
$$P_2 = \int_{W(Z_0,\Psi)>0} p_{02}(t)dt = \int_0^1 H(W(Z_0,\hat{\Psi}))\ \ p_{02}(t)dt\ ,$$

where $H(\cdot)$ is the Heaviside step function. Now we have descriptions of inputs and using function, so we can go to realization. This algorithm we use for two data set, one is simulated data and the next real data set.

4 Simulated data

Simulated data sets are used for evaluating algorithm reliability estimation. We consider the three different models with neighbours systems as described above. Here we are also interested in the measuring empirically the convergence rate of the probability of misclassification estimators. Therefore, simulations are conducted on increasing lattice sizes. $n = 8\times8,\ 16\times16,\ 32\times32,\ 48\times48,\ 56\times56,\ 64\times64$.

Gibbs sampling

We used the Gibbs sampler (Geman and Geman, 1984), to generate the random field induced by the Markov property, defied over the all lattices sizes. Gibbs sampling is a special case of the Metropolis Hastings algorithm. It is a Markov chain Monte Carlo

(MCMC) algorithm that samples each random variable of a graphical model, one at a time. The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution. In our case, we sample one value of a single lattice point z_i at a time, while keeping everything else fixed. Assume the input lattice has a size of N × N. And we will choose a set of parameter values that satisfy the integrability conditions. The algorithm is as shown in Algorithm 2.

Algorithm 2:
Initialize starting values of $z(s_i)$, for $i = 1,...,N$;
while not at convergence **do**
  Pick an order of the $N \times N$ variables;
  for each variable $z(s_i)$ **do**
    Sample $z(s_i)$ based on $p\left(z(s_i)\middle|N\left(z(s_i)\right)\right)$
    Update $z(s_i)$ to Z
  end
end

One major of iterative simulation pitfall is what is called the "burn-in" or "warm up" problem, which refers to the question of how long to run the chain $Z^{(1)}, Z^{(2)},...$ on grounds that the chain may not yet have reached equilibrium (i.e., the target distribution). Gelman and Rubin (1992) propose a fully quantitative method to monitor the convergence of iterative simulation using several independent sequences, with starting points sampled from an overdispersed distribution. At each step of the iterative simulation, they obtain, for each univariate random variable of interest, an estimate of its distribution and an estimate of the factor by which the scale of this distribution might be reduced if the simulations were continued indefinitely. This potential scale reduction is estimated by the ratio of the current variance estimate using the variance between the several sequence means to the within-sequence variance estimate. When this ratio is near 1, it is considered that the iterative simulation is close enough to convergence and that valid inference for the target distribution can be obtained using data from the next iterations.

5 Real data

Notice that the family of beta distributions offers a large variety of densities on a bounded interval [a, b], which makes the auto-beta models a potentially important class of spatial models, rates, proportions, or concentration indices. For example pant cover, election results.

The modelling approach proposed to tackle spatially sampled proportions will be apply to the real database. It is planned to analyze spatial data with given coordinates. Also, a quantitative variable is monitored in every spatial point that can be modeled using beta distribution. Also, a qualitative variable with two possible values that split the data into two classes. The research objective is to apply proposed classification procedure for the data to evaluate possible class label in a new date point based on measured qualitative variable values.

This empirical study will be expanded, and calculation results will be presented. Resulting data analysis will also be performed: comparison, generalization and conclusions will be presented. Real data analysis and results will also be presented.

6 References

Besag J. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Rotal Statistical Society, vol. 36, p. 192-236.

Besag J. 1977. Efficiency of pseudolikelihood estimation for simple Gaussian fields. Biometrika, vol. 64 (3), p. 616-618.

Hardouin C., Yao J. F. 2008. Multi-parameter auto-models and their application. Biometrika, Oxford University Press (OUP),95 (2), pp.335-349.

Geman S., Geman D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 6, p. 721–741.

Geman S., Rubin D. B. 1992. Inference from iterative simulation using multiple sequences. Statistical science. vol. 7, p. 457–511.