STRUCTURAL ANALYSIS OF GRAY MATTER VIA PRUNED CLASSIFIER

Q. V. Tran[1], *J. Kukal*[1], *J. Horáček*[2]

[1]KSE, FJFI, CVUT v Praze

[2]PCP, CNS, 3LF UK v Praze

Abstract

Traditional logistic regression can be easily generalized for multi-class partition task. When the number of input variables is fixed and the patterns are non-separable, the task can be solved via likelihood maximization. Using Bayesian approach we formulate alternative but regularized optimization task. The adequate objective function for minimization is smooth and convex, then unimodal, its minimum is not in infinity and thus easily obtainable. The second extension consists of pruning the structure of multi-classifier to obtain the best model. The implementation of pruning process leads to the binary optimization task, which was solved via Fast Simulated Annealing (FSA) heuristics. The classifier and its implementation in the Matlab environment were used for the structural analysis of anatomically labeled domains of gray matter which were obtained from human brain MRI scans. The biomedical data include atlased brain maps of schizophrenic and control normal patients. Various geometrical characteristics of 116 domains are used as complete model of schizophrenia and the optimum sub-model was found.

1 Introduction

Traditional *probit regression* according to Bliss [1] is oriented to classification into *two classes* strongly connected with Gaussian normal distribution. Lately, Berkson [2] introduced *logit regression* with logistic function inside. The logistic model is widely used [4, 5, 8] and its generalization to *multi-class model* [4, 8] is known. There are also favorable multi-classification tasks beginning with *iris flower classification* according to Fisher [3]. Kukal and Vyšata [3] recommended to built up a kind of *soft multi-classifier* with *constrained gain* together with maximum *sensitivity and specificity* and design its learning as multi-criteria optimization task. Tran et all [7] discussed the possibility of *ANN pruning*, which is based on logistic regression and likelihood ratio testing. It was discussed only for two classes and non-separable patterns. The difficulties arise when the number of classes is higher then two and when the classes are separable due to dimensionality of input space. But there is a useful and not prohibited idea to transform the original data into another space with the higher dimensionality and thus better separability of patterns. The next section introduces generalized logistic model for multi-classification, first.

2 Multi-Class Logistic Model

Let $n, N \in \mathbf{N}$, $N > 1$ be number of *properties* and number of *classes*. In this case the *classifier* f has n inputs and N outputs. The classifier realizes a *partition* among N classes just when it is described as a function

$$\mathrm{f} : \mathbf{R}^n \to \mathbf{Q}_N \text{ where } \mathbf{Q}_N = \{ \boldsymbol{y} \in \mathbf{P}_N \mid \| \boldsymbol{y} \|_1 = 1\} \subset \mathbf{P}_N = [0, 1]^N \tag{1}$$

According to [4, 8] we can generalize logit model of logistic regression for N classes as

$$y_i = \frac{\exp(s_i)}{\sum_{k=1}^{N} \exp(s_k)} \text{ for } i = 1,\ldots, N \tag{2}$$

where

$$s_i = \sum_{j=0}^{n} v_{i,j} x_j \text{ for } i = 1,\ldots, N \tag{3}$$

$\boldsymbol{x} \in \mathbf{R}^{n+1}, \boldsymbol{V} \in \mathbf{R}^{N\times(n+1)}$ and $x_0 \equiv 1$.

The model (2, 3) has redundant parameters. Expanding ratio in (2) by a factor exp(–s_1) we obtain the better form

$$y_i = \frac{\exp(s_i - s_1)}{\sum_{k=1}^{N}\exp(s_k - s_1)} = \frac{\exp(h_i)}{\sum_{k=1}^{N}\exp(h_k)} \text{ for } i = 1,\ldots, N \tag{4}$$

where

$$h_i = \sum_{j=0}^{n}(v_{i,j} - v_{1,j})x_j = \sum_{j=0}^{n} w_{i,j}x_j \text{ for } i = 1,\ldots, N \tag{5}$$

$\boldsymbol{W} \in \mathbf{R}^{N\times(n+1)}$ and $w_{1,j} = 0$ for $j = 0,\ldots, n$.

The model (4, 5) with unknown matrix $\boldsymbol{W}$ has only $(N – 1)(n + 1)$ free parameters. The sensitivity of this model to input variables depends only on reduced matrix $\boldsymbol{W}_{\text{red}}$, which is defined via relationship

$$\boldsymbol{W} = \begin{pmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{b} & \boldsymbol{W}_{\text{red}} \end{pmatrix} \tag{6}$$

where $\boldsymbol{W}_{\text{red}} \in \mathbf{R}^{(N-1)\times n}, \boldsymbol{b} \in \mathbf{R}^{N-1}$ is a bias vector for N classes.

3 Maximum Likelihood Estimate and Its Regularization

The estimation of model parameters is frequently performed via maximization of likelihood function or equivalently as a minimization of its negative logarithm. Resulting optimum point (if exists) is called *maximum likelihood estimate*. Let m be number of patterns, $(\boldsymbol{x}_k, c_k)$ be a *pattern* where

$c_k \in \{1,\ldots, N\}$ be *class index* for $k = 1, \ldots, N$. Adequate optimization task for maximum likelihood estimation is

$$\Phi(\boldsymbol{W}) = -\ln \mathrm{L}(\boldsymbol{W}) = \min \tag{7}$$

where L is likelihood function for given pattern set. Recognizing that $\boldsymbol{y}_k = \mathrm{f}(\boldsymbol{x}_k) > \boldsymbol{0}$ is a vector of class membership probabilities and denoting its i th component as $(\boldsymbol{y}_k)_i$, we obtain explicitly

$$\Phi(\boldsymbol{W}) = -\sum_{k=1}^{m}\ln(\boldsymbol{y}_k)_{c_k} \tag{8}$$

which is smooth and convex, thus unimodal function. But in special cases (including separable patterns) the minimum of (8) does not exist and the norm of $\boldsymbol{W}$ approaches infinity during search process. A kind of task regularization is necessary in this case. When the regularization is based on statistical theory, we can convert the task to M-estimate finding and testing.

Bayesian approach is used in our paper to perform statistically correct regularization. It is based on *a priori knowledge* of distribution of estimated matrix $\boldsymbol{W}_{\text{red}}$. Let $w_{i,j} \sim \mathrm{N}(0,\sigma^2)$ be independent stochastic variables with Gaussian normal distribution for $i = 2,\ldots, N, j = 1,\ldots, n$. Here the standard deviation

$\sigma > 0$ is supposed to be a priori known. But there is not a priori knowledge of bias vector $\boldsymbol{b}$. Replacing likelihood by conditional probability in (7) we obtained new optimization task

$$\Psi(\boldsymbol{W}) = \frac{1}{2\sigma^2}\|\boldsymbol{W}_{\text{red}}\|_F^2 - \sum_{k=1}^{m}\ln(\boldsymbol{y}_k)_{c_k} = \min \tag{9}$$

Here, $\|\ldots\|_{\mathrm{F}}$ is Frobenius norm and resulting function in (9) is smooth and convex, again. But in this case, the optimum of (9) exists in all cases. After the formula rearrangement

$$\Psi(\boldsymbol{W}) = \sum_{k=1}^{m}\left(-\ln(\boldsymbol{y}_k)_{c_k} + \frac{1}{2m\sigma^2}\|\boldsymbol{W}_{\text{red}}\|_F^2\right) = \sum_{k=1}^{m}\psi_k(\boldsymbol{W}) = \min \tag{10}$$

we recognize, that it is just M-estimate finding task. Thus we can use the theory of M-estimates for model and sub-model testing. The minimization of (9) is easy to perform in the Matlab environment using `fminunc` or `fminsearch` function.

4 Model and Sub-Model Testing

There are many approaches to model testing [8]. One of them is *likelihood ratio test* [8]. It is based on the comparison of given model and its sub-models. There are many models, which can be derived from the full model (4, 5). They can be generated by a *control matrix* $\boldsymbol{B} \in \{0,1\}^{(N-1)\times n}$, where $b_{i,j} = 1$ means that input variable x_j is used for the calculation of h_i and thus $w_{i,j}$ is unknown parameter. When $b_{i,j} = 0$, then $w_{i,j} = 0$ and thus fixed. The indexing system in $\boldsymbol{B}$ is the same as in $\boldsymbol{W}$. The number of free parameters for the estimation is then

$$Q = \sum_{i=2}^{N} \sum_{j=1}^{n} b_{i,j} \text{ satisfying } 0 \le Q \le (N-1)n \tag{11}$$

The *full model* is a special case of the model with $Q = (N - 1)n$. *Constant model* with $Q = 0$ is an opposite extreme case. Let $\boldsymbol{B}$, $\boldsymbol{B}_{\text{sub}}$ be control matrices of given model and its sub-model and Q, Q_{sub} adequate number of free parameters. Then the relationship between their structures is

$$\boldsymbol{B}_{\text{sub}} \le \boldsymbol{B} \wedge Q_{\text{sub}} < Q \tag{12}$$

The traditional *likelihood ratio test* (LR-test) of model and its sub-model difference is based on the testing criterion

$$LR = 2(\ln L - \ln L_{\text{sub}}) \tag{13}$$

where L, L_{sub} are maximized likelihood functions for the model and its sub-model. The stochastic variable LR has limiting distribution $\chi^2_{Q-Q_{\text{sub}}}$ with $Q - Q_{\text{sub}}$ degrees of freedom. The adequate p_{value} is directly obtained as

$$p_{\text{value}} = 1 - \mathrm{F}_{Q-Q_{\text{sub}}}(LR) \tag{14}$$

where F is cumulative distribution function of chi-squared distribution. The comparison of the model and its sub-model via LR-test brings a view on the significance of the model, which is compared with constant model. Let Ψ, Ψ_0 be optimum values of (10) for the model with control matrix $\boldsymbol{B}$ and for the constant model. Comparing the model with constant one, we can measure the model significance via probability

$$p_0 = 1 - \mathrm{F}_Q(2(\Psi_0 - \Psi)) \tag{15}$$

Lower value of p_0 indicates the higher significance of given model and various models can be ordered according to p_0 to obtain the best one.

Let m_k be number of patterns in k^{th} class. Then the optimal output of constant model is $y_k = m_k/m$

for $k = 1,\dots, N$ and the optimum value of (10) is

$$\Psi_0 = -m \sum_{k=1}^{N} y_k \ln y_k \tag{16}$$

5 Model Pruning as Binary Optimization Task

The minimization of (10) is easy task of convex programming but it is only a subject of inner loop. Statistical meaning of model pruning is in finding of the best model or its control matrix $\boldsymbol{B}$, respectively.. There are $2^{(N-1)n}$ possibilities how to design the control matrix $\boldsymbol{B}$. We find the best pruning as a global minimum of objective function

$$\mathrm{q}(\boldsymbol{B}) = \log_{10} p_0 \tag{17}$$

which is a *binary optimization task* with many local minima. The task was solved by the application of Fast Simulated Annealing (FSA). The probability of mutation p_{mut} was fixed for each element of matrix $\boldsymbol{B}$ to obtain $\boldsymbol{B}_{\mathrm{new}}$. The difference between q($\boldsymbol{B}_{\mathrm{new}}$) and q($\boldsymbol{B}$) was modulated by Cauchy distribution with amplitude $T_k > 0$. Resulting probability of state changing from $\boldsymbol{B}$ to $\boldsymbol{B}_{\mathrm{new}}$ is thus enumerated as

$$p_{\mathrm{new}} = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{q(\boldsymbol{B}) - q(\boldsymbol{B}_{\mathrm{new}})}{T_k} \qquad (18)$$

The cooling strategy was set to

$$T_k = \frac{T_0}{1 + \dfrac{k}{n_0}} \qquad (19)$$

where T_0, n_0, k are initial temperature, index scale and index of state change. The FSA method was supposed to be a heuristics, which increases the probability of global optimum reaching in the task (17). The numeric difficulties with the evaluation of p_0 in (15) for (17) are solvable by the asymptotic expansion of cumulative distribution function complement. The FSA algorithm together with objective function (17) were also created in the Matlab environment.

6 Biomedical Application

The new Matlab library for regularized pruning of multi-logidtic model was used for optimum sub-model finding in the case of schizophrenia classification. The complete set of 98 patients was split into schizophrenic (54 patients) and control normal (44 patients) groups. Their 3D MRI brain scans were labeled and atlased via Statistical Parametric Mapping (SPM5) technique and then used as morphological data source. Resulting labeled 3D scan (of every patient) consists of 116 anatomical domains with respect to gray matter filling. Various geometric characteristics were used to obtain the vector description of labeled 3D scan: number of separated domains, total volume, radius of maximum internal sphere, diameter, number of watershed domains, maximum volume of watershed domain, maximum internal sphere radius over watershed domains and maximum diameter over watershed domains. The classification into two clases (schizophrenia, control normal) was performed over 166×8 morphological properties of gray matter. The optimum sub-model prefers total volumes in selected anatomical regions in general which means that traditional volumetric approach plays the significant role.

7 Conclusion

Extended logistic regression model was used for the realization of regularized multi-classifier. Parameter estimation was converted to convex optimization task for free minimization. Optimum sub-model selection was inspired by likelihood ratio test and then performed using Fast Simulated Annealing. The properties of regularized multi-classifier were studied on schizophrenia classification task together with model structure pruning.

Acknowledgement

The paper was created under the support of grant OHK4 - 165/11 CTU in Prague.

References

[1] Bliss C.I., The method of probits, *Science*, Vol.79, No.2037, 1934, pp. 38–39.

[2] Berkson J., Application of the logistic function to bio-assay, *J Am Stat Assoc*, Vol.39, 1944, pp. 357–365.

[3] Fisher, R.A., The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7: 179–188, 1936

[4] Hilbe J. M., *Logistic Regression Models*, Chapman & Hall/CRC Press, 2009

[5] Hosmer, D. W., Stanley, L., *Applied Logistic Regression*, New York; Chichester, Wiley, 2000

[6] Kukal J., Vyšata O., Learning of Soft Classifier via Differential Evolution, *Proc. 14th Int. Conf. on Soft Computing, Mendel 2008*, 181-185, VUT Brno, Brno, 2008

[7] Tran Q.V., Kukal J., Kalčevová J., Boštík J., Logistic Regression as Bridge Between Statistics and Artificiall Inteligence, *Proc. 16th Int. Conf. on Soft Computing, Mendel 2010*, 491-494, VUT Brno, Brno,2010

[8] Wooldridge J. M., *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2002

tran@vse.cz

jaromir.kukal@fjfi.cvut.cz

horacek@pcp.lf3.cuni.cz