HYPERSPECTRAL IMAGE CLASSIFICATION BASED ON A FAST BREGMAN SPARSE MULTINOMIAL LOGISTIC REGRESSION ALGORITHM

J. Li[a, *], J. Bioucas-Dias [a], Antonio Plaza [b]

[a] Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001, Lisbon
Portugal - (jun, bioucas)@lx.it.pt
[b]Department of Technology of Computers and Communications, University of Extremadura
E-10071 Cáceres, Spain - aplaza@unex.es

KEY WORDS: Sparse Multinomial Logistic Regression (SMLR), hyperspectral classification, Bregman iteration

ABSTRACT:

The Sparse Multinomial Logistic Regression (SMLR) method introduced in (Krishnapuram, 2005) is among the state-of-the-art in supervised learning. However its application to large datasets, such as hyperspectral imagery is still a rather challenging task from the computational point of view, sometimes even impossible to perform. In this paper, the Bregman iteration-based SMLR method (Bregman-SMLR) recently introduced in (Bioucas-Dias, 2008) is applied to hyperspectral data classification problems. The Bregman method allows replacing a difficult, non-smooth convex problem with a sequence of quadratic plus diagonal *l2-l1* problems which are very easy to solve (Bioucas-Dias, 2008). Compared with the SMLR algorithm, the reduction of computational complexity is on the order of $d(m-1)^3$ (d is the number of features, and m is the number of classes.) The effectiveness of the proposed method is evaluated with simulated data sets and a real AVIRIS image. Results are presented and compared with others obtained by state-of-the-art supervised algorithms.

1. INTRODUCTION

The *sparse multinomial logistic regression* (SMLR) method introduced in (Krishnapuram, 2005) is among the state-of-the-art in supervised learning. The core of the SMLR is the solution of a two-term optimization problem: one term is the logistic regression and the other is a Laplacian prior which enforces sparseness, thus controlling the machine complexity. However, the SMLR application to large datasets, such as hyperspectral imagery, is still a rather challenging task from the computational point of view, being sometimes even impossible to perform. This is because SMLR has the complexity of the *iterative reweighted least squares* (IRLS) algorithm for maximum likelihood estimation of feature weights. To lighten the SMLR computational burden, a *fast sparse multinomial logistic regression* (FSMLR) was introduced in (Borges, 2006) to implement an iterative scheme (based on the block Gauss-Seidel method) to compute the feature weights of the decision function. The computational gain with respect to the SMLR algorithm is of the order of the number of classes. The FSMLR algorithm is thus well-suited to hyperspectral data sets with a large number of classes.

* Corresponding author. This work was supported by Marie Curie Grant MEST-CT-2005-021175 from the European Commission.

However, when dealing with classification problems with large training sets resulting, for example, from kernel-based regression, the FSMLR method is still very complex in computational terms.

In this paper, the Bregman iteration-based SMLR method (Bregman-SMLR) recently introduced in (Bioucas-Dias, 2008) is applied to hyperspectral data classification problems. The Bregman method allows replacing a difficult, non-smooth convex problem with a sequence of quadratic plus diagonal *l2-l1* problems which are very easy to solve. If *d* is the number of features and *m* is the number of classes, the complexity of the Bregman-SMLR method is $O(d^2)$, which is in contrast with the $O((d(m-1))^3)$ figure of SMLR. As a result, the reduction of computational complexity is on the order of $d(m-1)^3$.

In order to illustrate the effectiveness of the Bregman-SMLR method, we apply it to simulated data sets and real AVIRIS hyperspectral image and compare the obtained results with those provided by the FSMLR, the support vector machines (SVMs), and the linear discriminant analysis (LDA) (Camps-Valls, 2005) in terms of the following aspects: 1) overall accuracy; 2) computational cost; 3) robustness to noise; and 4) number of the training samples required.

2. METHOD

The SMLR used here is, basically, the algorithm introduced in Krishnapuram et Al. (Krishnapuram, 2005). The Bregman-SMLR solves the same optimization problem, but uses the augmented Lagrangian framework. In this section, we briefly review the SMLR the Bregman-SMLR methods.

2.1 Sparse Multinomial Logistic Regression (SMLR)

Let $x = [x_1, ..., x_n]^T \in \Re^{d \times n}, x_i \in \Re^d$ be a vector of observed features, and $y = [y^{(1)}, ..., y^{(m)}]^T$ a 1-of-m encoding of the classes (*n* is the number of samples, *d* is the number of features, and m is the number of classes). The goal of classification is to estimate *y* given *x*. Suppose $\omega^{(i)}$ is the feature-weight vector corresponding to class *i;* then, according to the multinomial logistic regression model, the probability that a given sample *x* belongs to class *i* is given as follows:

$$p(y^{(i)} = 1 \mid x, \omega) = \frac{\exp(\omega^{(i)^T} h(x))}{\sum_{j=1}^{m} \exp(\omega^{(j)^T} h(x))} \quad (1)$$

where, $h(x) = [h_1(x), ..., h_l(x)]^T$ is a vector of l fixed functions of the input. Usual choices for *h(x)* are linear maps $h(x_i) = [1, x_{i,1}, ..., x_{i,n}]^T$, where $x_{i,j}$ means the *j*th component (band) of x_i and kernels $h(x_i) = [1, K(x, x_1), ..., K(x, x_n)]^T$, where $K(\cdot,\cdot)$ is some symmetric kernel function. In this paper, we only consider kernels of the radial basis function (RBF) class.; for the nonlinear mapping guarantees that the transformed samples are more likely to be linearly separable. The SMLR method uses the Maximum A posteriori (MAP) method to estimate the components of ω from the training set:

$$\begin{aligned} \hat{\omega}_{MAP} &= \arg\max_{\omega} L(\omega) \\ &= \arg\max_{\omega} \left[l(\omega) + \log p(\omega) \right] \end{aligned} \quad (2)$$

where $l(\omega)$ is the log-likelihood function,

$$l(\omega) = \sum_{j=1}^{n} \left[\sum_{i=1}^{m} y_j^i \omega^{(i)^T} x_j - \log \sum_{i=1}^{m} \exp(\omega^{(i)^T} x_j) \right] \quad (3)$$

and $p(\omega)$ is a Laplacian prior on ω, which means that $p(\omega) \propto \exp(-\lambda \|\omega\|_1)$, where λ is a regularization parameter controlling the degree of sparseness of $\hat{\omega}_{MAP}$. According to the bound optimization approach (Lange, 2000), the log-likelihood function $l(\omega)$ can be optimized by iteratively maximizing a surrogate function Q, such that:

$$\hat{\omega}_{t+1} = \arg\max_{\omega} Q(\omega \mid \hat{\omega}_t) + \log p(\omega) \quad (4)$$

While the log-likelihood function is concave, the surrogate function $Q(\omega \mid \hat{\omega}')$ can be determined by using a bound on the Hessian H of the log-likelihood. Let B be a negative definite matrix such that $H(\omega) - B$ is positive semi-definite, i.e., $H(\omega) \geq B$ for any ω. A valid surrogate function is (B ö hning, 1992),

$$Q(\omega \mid \hat{\omega}_t) = \omega^T (g(\hat{\omega}_t) - B\hat{\omega}_t) + \frac{1}{2}\omega^T B\omega \quad (5)$$

$$B \equiv -\frac{1}{2}\left[I - 11^T / m\right] \otimes \sum_{j=1}^{n} x_j x_j^T \quad (6)$$

where $\otimes$ is the Kronecher matrix product, $1 = [1,1,...,1]^T$, and

$$g(\omega) = \sum_{j=1}^{n} (y_j^{'} - p_j(\omega)) \otimes x_j \quad (7)$$

where $y_j^{'} = \left[y_j^{(1)}, y_j^{(2)}, ..., y_j^{(m-1)} \right]^T$ and $p_j(\omega) = \left[p_j^{(1)}(\omega), ..., p_j^{(m-1)}(\omega) \right]^T$.

With the inclusion of a Laplacian prior, the objective function becomes

$$L(\omega) = l(\omega) - \lambda \|\omega\|_1 = Q(\omega \mid \hat{\omega}_t) - \lambda \|\omega\|_1 \quad (8)$$

The estimates of ω are then given by:

$$\begin{aligned} \hat{\omega}_{MAP} &= \arg\max_{\omega} L(\omega) \\ &= \arg\max_{\omega} Q(\omega \mid \hat{\omega}_t) - \lambda \|\omega\|_1 \\ &= \arg\max_{\omega} \omega^T (g(\hat{\omega}_t) - B\hat{\omega}_t) + \frac{1}{2}\omega^T B\omega - \lambda \|\omega\|_1 \\ &= \arg\min_{\omega} -\omega^T (g(\hat{\omega}_t) - B\hat{\omega}_t) - \frac{1}{2}\omega^T B\omega + \lambda \|\omega\|_1 \end{aligned} \quad (9)$$

It is not easy to minimize (9) directly in closed form. A line of attack is to replace the l_1 norm with a lower quadratic bound in order to get a surrogate function to iteratively optimize the log-prior. This leads the update function

$$\hat{\omega}_{t+1} = (B - \lambda\Lambda_t)^{-1}(B\hat{\omega}_t - g(\hat{\omega}_t)) \quad (10)$$

Where

$$\Lambda_t = diag\left\{ \left|\hat{\omega}_t^{(1)}\right|^{-1}, ..., \left|\hat{\omega}_t^{d(m-1)}\right|^{-1} \right\} \quad (11)$$

Numerically, (10) is equivalent to solve (Krishnapuram, 2005):

$$\hat{\omega}_{t+1} = \Gamma_t (\Gamma_t B \Gamma_t - \lambda I)^{-1} \Gamma_t (B\hat{\omega}_t - g(\hat{\omega}_t)) \quad (12)$$

where

$$\Gamma_t = diag\left\{ \left|\hat{\omega}_t^{(1)}\right|^{1/2}, ..., \left|\hat{\omega}_t^{d(m-1)}\right|^{1/2} \right\} \quad (13)$$

and $\left|\hat{\omega}_t^{(i)}\right|$ stands for the *i*th value of vector $\left|\hat{\omega}_t\right|$. Now it is possible to estimate the MAP multinomial logistic regression with a Laplacian prior by using the classical IRLS method. And the complexity is the same as the IRLS algorithm for ML estimation.

2.2 Bregman-SMLR

The computational cost involved in solving the linear system presented in (12) is $O((dm)^3)$, which is prohibitive when dealing with large datasets, either with large number of features, or with a very large training dataset. Recently, a Bregman iteration based sparse multinomial logistic regression (Bregman-SMLR) was introduced by J. Bioucas (Bioucas-Dias, 2008), which made possible to deal efficiently with large data sets. In this section, we briefly review the Bregman-SMLR algorithm.

In expression (9), suppose that $\nu = \omega$. Then, we can replace the problem with the following one:

$$\begin{aligned} \left(\hat{\omega}_{t+1}, \hat{\nu}_{t+1}\right) = \arg\min_{\omega,\nu} & -\omega^T (g(\hat{\omega}_t) - B\hat{\omega}_t) \\ & -\frac{1}{2}\omega^T B\omega + \lambda \|\nu\|_1 \\ s.t.: \quad & \omega = \nu \end{aligned} \quad (14)$$

(14) can be iteratively solved (Bioucas-Dias, 2008) as follows:

$$\begin{aligned}(\hat{\omega}_{t+1},\hat{v}_{t+1}) &= \arg\min_{\omega,v} -\omega^T(g(\hat{\omega}_t)-B\hat{\omega}_t) \\ &\quad -\frac{1}{2}\omega^T B\omega+\lambda\|v\|_1+\frac{\beta}{2}\|\omega-v-b_t\|^2 \\ b_{t+1} &= b_t-(\hat{\omega}_{t+1}-\hat{v}_{t+1})\end{aligned} \quad (15)$$

The above minimization is still a difficult problem. However, the minimizations with respect to either ω or v are very easy to compute. Exploiting this fact the Bregman-SMLR iterative scheme proposed in (Bioucas-Dias, 2008) is a follows:

$$\begin{aligned}\hat{\omega}_{t+1} &= \arg\min_{\omega,v} -\omega^T(g(\hat{\omega}_t)-B\hat{\omega}_t)-\frac{1}{2}\omega^T B\omega+\frac{\beta}{2}\|\omega-v-b_t\|^2 \\ \hat{v}_{t+1} &= \arg\min_{\omega,v} -\lambda\|v\|_1+\frac{\beta}{2}\|\omega-v-b_t\|^2 \\ b_{t+1} &= b_t-(\hat{\omega}_{t+1}-\hat{v}_{t+1})\end{aligned} \quad (16)$$

The first step leads to the update function of $\hat{\omega}$:

$$\hat{\omega}_{t+1}=(-B+\beta I)^{-1}\left(-(g(\hat{\omega}_t)-B\hat{\omega}_t)+\beta(\hat{v}_t+b_t)\right) \quad (17)$$

The second step amounts to apply the soft shrinkage function to update e v:

$$\hat{v}_{t+1}=softshrink\left(\omega_{t+1}-b_t,\lambda/\beta\right) \quad (18)$$

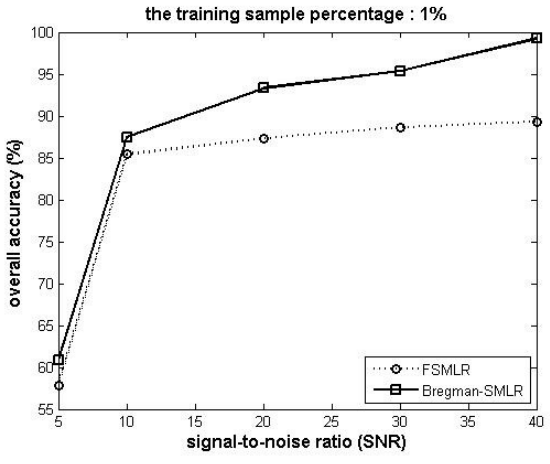
Where B and $g(\hat{\omega}_t)$ are given in (6) and (7), respectively. According to (6), B is fixed, so the part of $(-B+\beta I)^{-1}$ in (17) doesn't need to be updated during the iterations. It can be computed before hand, which greatly lighten the computational complexity, leading to a cost of $O(d^2)$.

3. EXPERIMENT RESULT

In this section, experimental results will be presented. In the first part, simulated data sets are performed to analysis the computational cost, robustness to noise and limited training samples. In the second part shows the results obtained from real hyperspectral imagery.

3.1 Simulated hyperspectral images

Simulated datasets are used to test the proposed method in comparison with the FSMLR algorithm, which demonstrates high quality in supervised hyperspectral classification as shown in (Borges, 2006). The size of the simulated images is $100\times100\times224$ (100×100 is the spatial size of the simulate images, and 224 is the number of spectral bands.), and 10 classes. The featurures are Gausssian vectors with means selected from the USGS (Clark, 2007) library. And covariance matrix $\sigma^2 I$, where I is the identity matrix. The parameter σ determines the signal-to-noise ratio (SNR), as $\text{SNR} \equiv 10\log_{10}\left(E[x^T x]/E[n\sigma^2]\right)$ (n is the number of samples). Figure 1, top, shows the overall accuracy (OA) as a function of SNR using 100 training samples (1% of the whole image).The remaining samples are used for validation. In this case, The Bregman-SMLR algorithm outperforms the FSMLR. Figure 1, middle, shows OA results as a function of the number of training samples with SNR set to 5. Both algorithms obtain similar OA. Figure 1, bottom, shows the computational cost as a function of the number of training samples. As expected, the Bregman-SMLR is much faster than FSMLR.



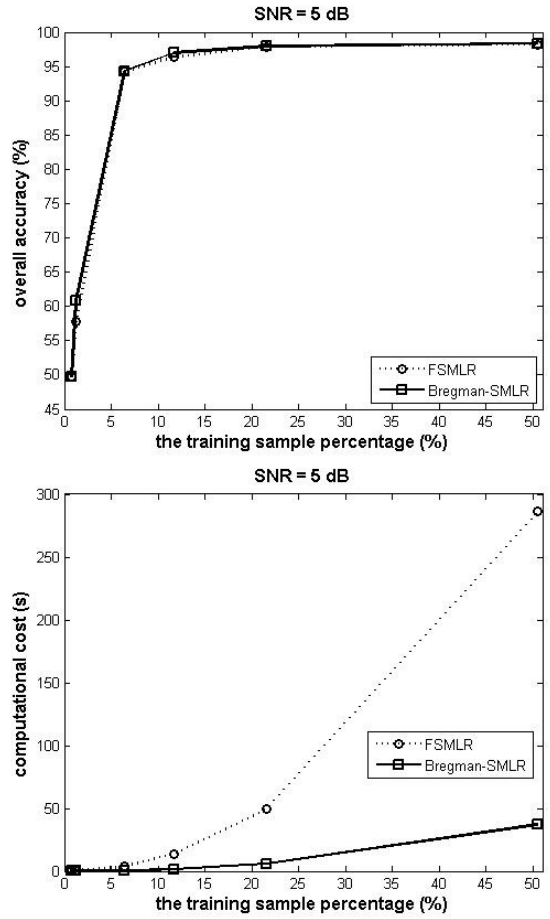


Figure 1. Results on simulated data sets. Each value was obtained from 100 Monte Carlo runs.

3.2 Experiments on real hyperspectral data

Experiments are also carried out using an AVIRIS spectrometer image taken over northwest Indianas Indian Pine test site in June 1992 (Landgrebe, 1992). It contains 145×145 pixels and 220 bands. Noisy bands in number of 20, namely due to water absorption, were discarded during the experiments. The ground truth data image contains 16 classes, 7 of which were discarded for insufficient number of training samples. The remaining 9 classes were used to generate a set of 4757 training samples, with random partition, and 4588 test samples. Table 1 compares the OA results with state-of-the-art supervised algorithms. The Bregman-SMLR obtained much better results than LDA, similar or comparable results to FSMLR and SVMs. Figure 2. shows the computational cost as a function of the number of training samples in comparison with FSMLR algorithm. Considering the computational costs a function of the training set size, the Bregman-SMLR achieves much better performance than the FSMLR method. For 50% of the training set, the FSMLR needs 317.99 seconds while it just needs 33.54 seconds of the Bregman-SMLR.

Bregman-SMLR	FSMLR	SVMs	LDA
91.23%	90.52%	~91%	91.08%

Table 1. Comparison of the proposed method with the results from (Camps-Valls, 2005; Borges, 2006) on a real dataset.

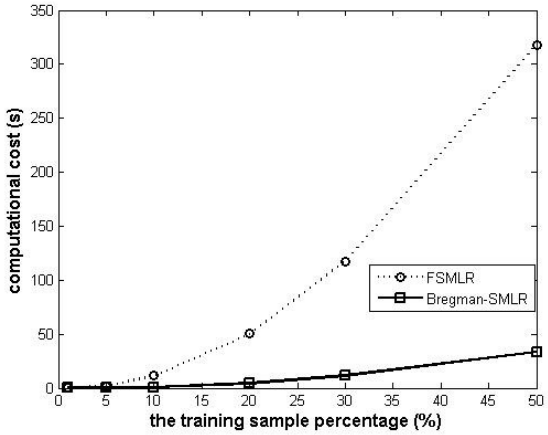


Figure 2. Results on a real dataset

4. CONCLUSIONS

In this paper, a fast Bregman sparse multinomial logistic regression algorithm (Bregman-SMLR) is applied to hyperspectral imagery. Compared with the SMLR algorithm, it is much faster and more efficient.. The performance of the proposed approach was evaluated by using simulated data sets and real AVIRIS hyperspectral imagery. The results obtained show high quality in supervised hyperspectral classification in terms of overall accuracy, robustness to noise, low complexity and limited training samples.

ACKNOWLEDGE

The authors would like to thank Prof. Landgrebe (Purde University, USA) for providing the AVIRIS data.

REFERENCES

Bioucas-Dias, J., 2008. Bregman-SMLR: A fast sparse logistic regression algorithm. Technical Report, Instituto Superior Técnico, TULisbon.

Borges, J. and Bioucas-Dias, J., 2006. Fast Sparse Multinomial Regression Applied to Hyperspectral Data. *International Conference on Image Analysis and – ICIAR*.

Böhning, D., 1992. Multinomial logistic regression algorithm. *Annals of the Institute of Sta-tistical Mathematics,* 44, pp. 197-200.

Camps-Valls, G. and Bruzzone, L., 2005. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6), pp. 1351-1362.

Clark, R. N., Swayze, G. A., Wise, R., Livo, E., Hoefen, T., Kokaly, R. And Sutley S.J., 2007. USGS digital spectral library splib06a. *U.S. Geological Survey, Digital Data Series 231*, 1.

Krishnapuram, B., Carin, L., Figueiredo, M. and Hartemink, A., 2005. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 27(6), pp. 957-968.

Landgrebe, D., 1992. AVIRIS NW Indiana's Indian pine.

Lange, K., Hunter, D. and Yang, I., 2000. Optimizing transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9, pp. 1-59.