RC24740 (W0902-041) February 6, 2009
Computer Science

IBM Research Report

Classification Via Compressed Random Fields

**Avishy Carmi, Guillermo Cecchi, Dimitri Kanevsky,
Bhuvana Ramabhadran, Irina Rish**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Contents

Abstract

A classification method based on random field models is derived. A few innovative concepts that are incorporated into the algorithm such as efficient training via Kalman filtering and adaptation via extended Baum-Welch (EBW) demonstrate improvement both in computational complexity and classification accuracy. The most significant contribution of this work, however, is the derivation of an online compression and training mechanism that is capable of representing the sparsity patterns which arise naturally in the high dimensional data sets. Thus, the improved classifier uses compressed sensing techniques for learning the statistical relations within the random field models. The resulting representations are compressed in the sense that only few connections are considered for each node. The performance of the new algorithm is demonstrated in fMRI classification.

Chapter 1

Random Field-Based Classification

1.1 Problem Formulation

Let the data set $X \in \mathbb{R}^m$ (of which m is very large) be associated with some response $Y \in \mathscr{Y}$ where $\mathscr{Y}$ is some real-valued discrete space (i.e., classes). We are aimed at predicting Y, the class associated with X. For simplicity, we assume here only two classes $\mathscr{Y} = \{0, 1\}$. The assumption underlying the following derivations is that both X and Y are real-valued random variables of which the joint probability density function (pdf) $p(X, Y)$ exists.

In what follows we denote $\bar{X}$ the reduced feature space obtained by some feature extraction algorithm. The Appendix section details the two simple feature ranking methods (cross correlation and mutual information) used for obtaining the results presented inhere.

1.2 Random Field Models

Let $G(X, e)$ be an undirected graph, where e is an edge representing statistical dependency between vertices in $\bar{X}$. Let us denote G_i the neighborhood of $X_i \in \bar{X}$. For every class $Y = \theta$ we define a parametric functional relation of the form

$$\varphi_\theta(X_i, G_i, W_i) = 0 \tag{1.1}$$

where $W_i \sim p_{W_i}(\cdot)$ is a noise random variable representing uncertainty. Now, suppose that we can express the following relation

$$W_i = \varphi_\theta^{-1}(X_i, G_i) \tag{1.2}$$

then it easily follows that

$$p(X_i \mid G_i, \theta) = p_{W_i}(\varphi_\theta^{-1}(X_i, G_i)) \det \left(\nabla_{X_i} \varphi_\theta^{-1}(X_i, G_i)\right) \tag{1.3}$$

is the pdf describing statistical relation between X_i and G_i for a given class $Y = \theta$. The above formulation describes a random field (RF) of which the joint pdf can be approximated by the pseudo-likelihood

$$p(\bar{X} \mid Y = \theta) = \frac{1}{Z_\theta} \prod_{i=1}^{n} p(X_i \mid G_i, \theta), \qquad X_i \in \bar{X} \tag{1.4}$$

where n denotes the total number of nodes in $\bar{X}$. The normalizing constant Z_θ is given as

$$Z_\theta = \sum_{\bar{X} \in \Omega} \prod_{i=1 (X_i \in \bar{X})}^{n} p(X_i \mid G_i, \theta) \tag{1.5}$$

1.2.1 The Predicted Class

The predicted class is taken as the one with the highest probability $p(\bar{X}_{test} \mid Y = \theta)$, where X_{test} denotes the test data set. In the binary case, one has to compare

$$p(\bar{X}_{test} \mid Y = 0) \gtrless p(\bar{X}_{test} \mid Y = 1) \tag{1.6}$$

or equivalently

$$\frac{p(\bar{X}_{test} \mid Y=0)}{p(\bar{X}_{test} \mid Y=1)} \gtrless 1 \tag{1.7}$$

Further defining

$$l := \prod_{i=1}^{n} \frac{p(X_i \mid G_i, \theta=0)}{p(X_i \mid G_i, \theta=1)} \tag{1.8}$$

and

$$c := \log \frac{Z_0}{Z_1} \tag{1.9}$$

yields an equivalent test to (1.6)

$$\log l = \sum_{i=1}^{n} \log \frac{p(X_i \mid G_i, \theta=0)}{p(X_i \mid G_i, \theta=1)} \gtrless c \tag{1.10}$$

where the constant c (which rarely can be computed straightforwardly) can be tuned using either training or development data sets (see appendix). Using Eq. (1.10), the predicted class is obtained as

$$\hat{Y} = \begin{cases} 0, & \log l > c \\ 1, & \log l < c \end{cases} \tag{1.11}$$

A few aspects concerning the converges of this likelihood ratio test are given in the Appendix.

1.2.2 Gaussian Random Fields

Let us assume linear functionals of the form

$$X_i = \left(\beta_i^\theta\right)^T G_i + W_i, \qquad X_i \in \bar{X}, \quad W_i \sim \mathcal{N}(0, \sigma_w^2) \tag{1.12}$$

where $G_i \in R^{n-1}$ and $\beta_i^\theta \in R^{n-1}$, $i \in [1, n]$. Following this, the conditional pdf $p(X_i \mid G_i, \theta)$ can be expressed by means of the pdf of W_i as

$$p(X_i \mid G_i, \theta) \propto \exp\left\{-\frac{1}{2}\sigma_w^{-1} \parallel X_i - \left(\beta_i^\theta\right)^T G_i \parallel_2^2\right\} \tag{1.13}$$

In practice, the random parameter vector associated with the class θ, β_i^θ, is estimated using the training data set. Let $\hat{\beta}_i^\theta$ be an estimator of β_i^θ, then

$$\beta_i^\theta = \hat{\beta}_i^\theta + \tilde{\beta}_i^\theta \tag{1.14}$$

where $\tilde{\beta}_i^\theta$ is the estimation error. Substituting (1.14) into (1.12) gives

$$X_i = \left(\hat{\beta}_i^\theta + \tilde{\beta}_i^\theta\right)^T G_i + W_i \tag{1.15}$$

Further defining $\zeta_i^\theta := \left(\tilde{\beta}_i^\theta\right)^T G_i + W_i$ yields

$$X_i = \left(\hat{\beta}_i^\theta\right)^T G_i + \zeta_i^\theta \tag{1.16}$$

which is similar to (1.12) with the only difference of β_i^θ replaced by its estimate. The conditional pdf $p(X_i \mid G_i, \theta)$ can now be expressed in terms of $\hat{\beta}_i^\theta$ instead of the unknown β_i^θ. Thus, assuming $\zeta_i \sim \mathcal{N}(0, \sigma_i^\theta)$ yields

$$p(X_i \mid G_i, \theta) \propto \exp\left\{-\frac{1}{2\sigma_i^\theta} \parallel X_i - \left(\hat{\beta}_i^\theta\right)^T G_i \parallel_2^2\right\} \tag{1.17}$$

In what follows we shall see that ζ_i^θ represents the innovation noise in the Kalman filtering formulation. This sequence has some well-known statistical properties [1].

1.3 Efficient Training via Kalman Filtering

We use the Kalman filter (KF) algorithm for training the RF models of every class in a computationally efficient manner. The KF estimates the parameters β_i^θ, $i \in [1, n]$ sequentially using the training samples thereby allowing significant reduction of computational load.

Suppose that there are k_θ training samples for class $Y = \theta$, and let $X_{train}^\theta := \left\{\bar{X}(1), \ldots, \bar{X}(k_\theta)\right\}$ be the set of these samples. The KF is the best linear estimator in the minimum mean square error (MMSE) sense [1], that is

$$\hat{\beta}_i^\theta = \arg\min_{\hat{\beta}_i^\theta} E\left\{\| \beta_i^\theta - \hat{\beta}_i^\theta \|_2^2\right\} \tag{1.18}$$

Taking (1.12) as the measurement equation while assuming $\sigma_w = 1$ yields the following KF recursion

Initialization:

$$P_0 = \gamma^{-1} I, \ \left(\hat{\beta}_i^\theta\right)_0 = 0, \quad \gamma << 1 \tag{1.19}$$

Measurement update:

$$K_k = P_k G_i(k) \left[G_i(k) P_k G_i(k)^T + 1\right]^{-1} \tag{1.20a}$$

$$\left(\hat{\beta}_i^\theta\right)_{k+1} = \left(\hat{\beta}_i^\theta\right)_k + K_k \left[X_i(k) - G_i(k)^T \left(\hat{\beta}_i^\theta\right)_k\right] \tag{1.20b}$$

$$P_{k+1} = \left(I - K_k G_i(k)^T\right) P_k \tag{1.20c}$$

It should be noted that the KF is used here for parameter estimation rather than state estimation. However, if the training samples are obtained from time-series then the conventional KF algorithm, which includes a time-propagation stage, may be more adequate.

The next stage consists of computing the conditionals $p(X_i \mid G_i, \theta)$ in (1.17). For that purpose we need to know the statistics of ζ_i^θ, the innovation. It is well known from KF theory that $(\zeta_i^\theta)_k$ is a zero-mean white Gaussian sequence [1]

$$(\zeta_i^\theta)_k \sim \mathcal{N}\left(0, G_i(k)^T P_k G_i(k) + 1\right) \tag{1.21}$$

In this work we compute the sample covariance of ζ_i^θ as

$$\sigma_i^\theta = \frac{1}{k_\theta - 1} \sum_{j=1}^{k_\theta} \left[X_i(j) - \left(\hat{\beta}_i^\theta\right)_{k_\theta}^T G_i(j)\right] \left[X_i(j) - \left(\hat{\beta}_i^\theta\right)_{k_\theta}^T G_i(j)\right]^T \tag{1.22}$$

1.3.1 Generalization of The KF Formulation

The linear connections (1.12) can be generalized as follows. Consider two sets of nodes $G_i \in R^r$ and $G_j \in R^m$ satisfying the relation

$$G_i = \beta_{ij}^\theta G_j + W_{ij} \tag{1.23}$$

where $\beta_{ij}^\theta \in R^{r \times m}$.

In order to implement the previously described KF scheme for estimating the matrices β_{ij}^θ we rewrite the above equation as follows

$$G_i = (G_j^T \otimes I_{r \times r}) \bar{\beta}_{ij}^\theta + W_{ij} \tag{1.24}$$

where

$$\bar{\beta}_{ij}^\theta := \text{Vec}\left(\beta_{ij}^\theta\right) \tag{1.25}$$

is the vectorized form of β_{ij}^θ and $\otimes$ is Kronecker product. The KF can now be applied for estimating $\bar{\beta}_{ij}^\theta$ using (1.24).

[1] The innovations process is non-stationary.

1.3.2 Adaptation

Given the test data set $\bar{X}_{test} = \{\bar{X}(1), \ldots, \bar{X}(k_{test})\}$, we adapt the MRF models of every class using extended Baum-Welch (EBW) iterations as follows (see [2–4])

$$\left(\hat{\beta}_i^\theta\right)_{j+1} = \left[I - D_j G_i(j)^T\right] \left(\hat{\beta}_i^\theta\right)_j + D_j X_i(j) \tag{1.26}$$

where D_j is some tuning matrix which can be set as the Kalman gain matrix, K_j (where j denotes the test sample index), to ensure convergence [4]. Finally, the sample covariance of ζ_i^θ is updated as

$$\begin{aligned}\left(\sigma_i^\theta\right)_{new} = & \frac{k_\theta - 1}{k_\theta + k_{test} - 1} \left(\sigma_i^\theta\right)_{old} \\ & + \frac{1}{k_\theta + k_{test} - 1} \sum_{j=1}^{k_{test}} \left[X_i(j) - \left(\hat{\beta}_i^\theta\right)_{k_{test}+1}^T G_i(j)\right] \left[X_i(j) - \left(\hat{\beta}_i^\theta\right)_{k_{test}+1}^T G_i(j)\right]^T\end{aligned} \tag{1.27}$$

1.3.3 Adding A Bias Term

In some cases adding a bias term to (1.12) may improve the training model. Following this, the linear functional takes the form

$$X_i = \left(\beta_i^\theta\right)^T G_i + b_i + W_i \tag{1.28}$$

where b_i is some unknown bias. Alternatively, the bias can be estimated as part of β_i^θ by simply using

$$X_i = \left(\beta_i^\theta\right)^T \bar{G}_i + W_i \tag{1.29}$$

where $\bar{G}_i = [G_i, 1]$.

Chapter 2

Compressed Random Fields

Using notions from the theory of compressed sensing we devise an online training and compression algorithm that replaces the ordinary KF described previously. The compression, as it is demonstrated in the ensuing, significantly improves the classification accuracy. The new compression algorithm is extensively detailed in [5].

2.1 Sparse Signal Recovery

Consider an $\mathbb{R}^n$-valued random signal x that is sparse in some known orthonormal sparsity basis $\psi \in \mathbb{R}^{n \times n}$, that is

$$z = \psi^T x, \quad |\text{supp}(z)| << n \tag{2.1}$$

where $\text{supp}(z)$ denotes the support of z. The signal x is measured using a sequence of noisy observations given by

$$y_k = Hx + \zeta_k = H' z + \zeta_k \tag{2.2}$$

where ζ_k is a zero-mean white Gaussian sequence with covariance R_k, and $H := H' \psi^T \in \mathbb{R}^{m \times n}$ with $m < n$.

Letting $y^k := [y_1, \dots, y_k]$, our problem is defined as follows. We are interested in a y^k-measurable estimator $\hat{x}$ such that the minimum mean square error (MMSE) $E\left[\| x - \hat{x} \|_2^2\right]$ is minimized.

2.2 The Combinatorial Problem and Compressed Sensing

It has already been shown that in the deterministic case (i.e., when z is a parameter vector) one can recover z (and therefore also x, i.e., $x = \psi z$) with high accuracy by solving the optimization problem [6, 7]

$$\min \| \hat{z} \|_0 \ \text{ s.t. } \sum_{i=1}^{k} \| y_i - H' \hat{z} \|_2^2 \leq \epsilon \tag{2.3}$$

for sufficiently small ϵ. Following a similar approach, in the stochastic case it can be shown that the sought-after optimal estimator satisfies

$$\min \| \hat{z} \|_0 \ \text{ s.t. } E_{z|y^k}\left[\| z - \hat{z} \|_2^2\right] \leq \epsilon \tag{2.4}$$

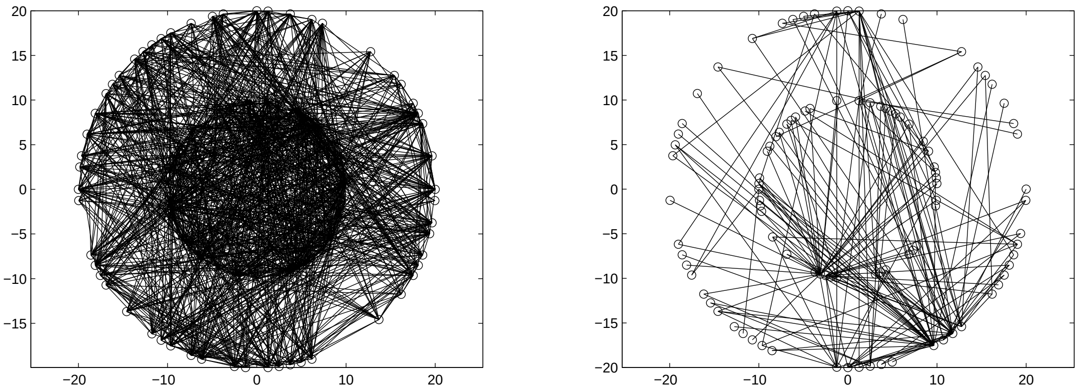
Unfortunately, the above optimization problems are NP-hard and cannot be solved efficiently. Recently, it has been shown that if the sensing matrix H' obeys a so-called *restricted isometry hypothesis* (RIH) then the solution of the combinatorial problem (2.3) can almost always be obtained by solving the convex optimization [6, 8]

$$\min \| \hat{z} \|_1 \ \text{ s.t. } \sum_{i=1}^{k} \| y_i - H' \hat{z} \|_2^2 \leq \epsilon \tag{2.5}$$

This is a fundamental result in the new emerging theory of compressed sensing (CS) [6, 8]. The essential idea here is that the convex l_1 minimization problem can be efficiently solved. Another insights provided by CS are related to the construction of sensitivity matrices that satisfy the RIH. These underlying matrices are random by nature which in turn has shed a new light on the way observations should be sampled. For an extensive review of CS the reader is referred to [6, 8].

2.3 Online Compression and Training

It is more likely that compressed random field models will better represent the patterns that arise naturally in the high dimensional data. This may be thought of as a model reduction approach for eliminating insignificant statistical relations. Following this approach we have substituted the ordinary KF training algorithm with the Compressed sensing-embedded KF (CSKF) of [5]. The obtained random field model in this case is compressed in the sense that only few connections are used for every node (see Fig. 2.1).



(a) No Compression **(b) With Compression**

Figure 2.1. Visualization of the random field models constructed by the ordinary KF (left panel) and the CS-embedded KF (right panel). All nodes are mapped onto two circles.

Chapter 3

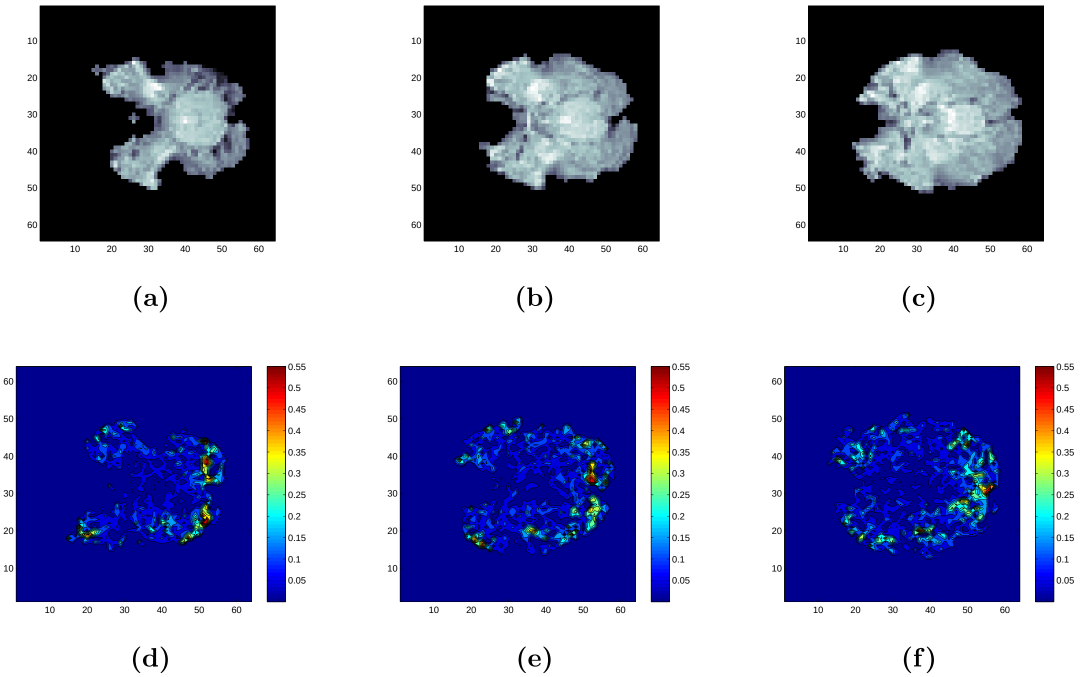
fMRI Classification

The proposed classification algorithm is applied to fMRI analysis. The performance evaluation consists of both the compressed and uncompressed random field models. The following study involves three data sets. The first two data sets consists of several subjects performing a task, such as reading a sentence or looking at a picture. The third data set (neurospin) consists of normal and mentally ill subjects (schizophrenia).

3.1 Princeton Data Set

The data X is a vector consisting of $14,043$ elements (voxels). The testing scenario and the fMRI datasets are the ones used in [9]. The total number of samples is 84. In this case Y represents the stimuli response which can take either of the two classes -1 or $+1$ (there are exactly 42 samples of each class). The training and testing data sets are obtained using two-out cross validation, that is, at every run two testing samples (one of each class) is taken out of the original set, leaving 82 training samples. This procedure is repeated 84 times. The classification algorithm is tested using Monte Carlo runs in which the original data set, consisting of 84 samples, is randomly permuted.

The upper panel in Fig. 3.1 shows various fMRI scans of different brain sections. The corresponding cross correlation maps of these sections are shown in the lower panel in the same figure.

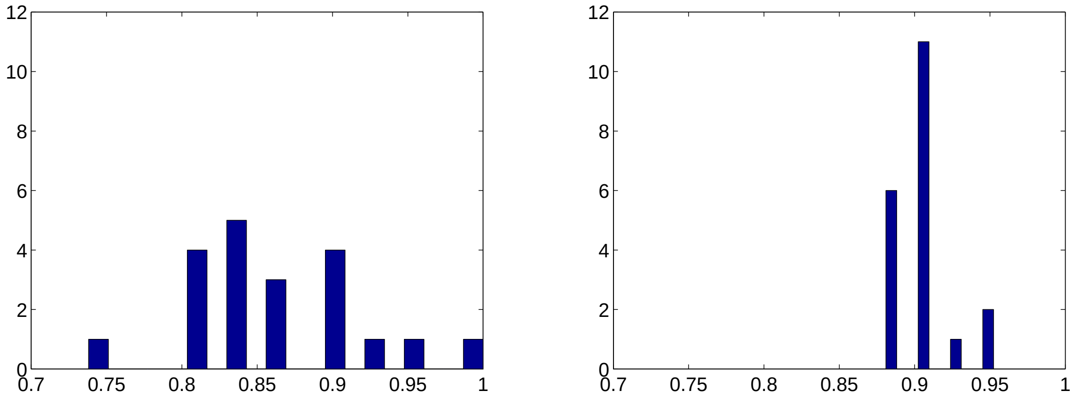


(a) (b) (c)

(d) (e) (f)

Figure 3.1. fMRI scans of different brain sections and their corresponding cross correlation maps.

The performance of the algorithm with 100 nodes based on 20 Monte Carlo runs is shown in figures 3.2. From this figure it can be clearly seen that the compression scheme significantly improves the mean classification accuracy (86% without compression and 91% with compression).

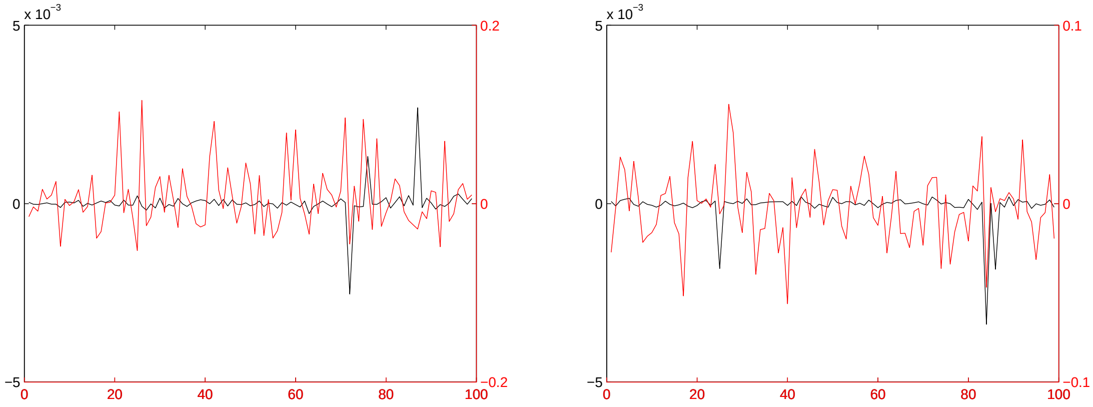


(a) Without compression

(b) With compression

Figure 3.2. Distribution of classification accuracy based on 20 Monte Carlo runs. The mean accuracy when using compression increases to 91%, approximately 5% more than in the uncompressed case. Random field consists of 100 nodes.

The connectivity of 2 distinct nodes of both the compressed and the uncompressed random field models is depicted in Fig. 3.3. This figure clearly shows the effect of compression on the estimated parameters β_i^θ.



(a) Node 1

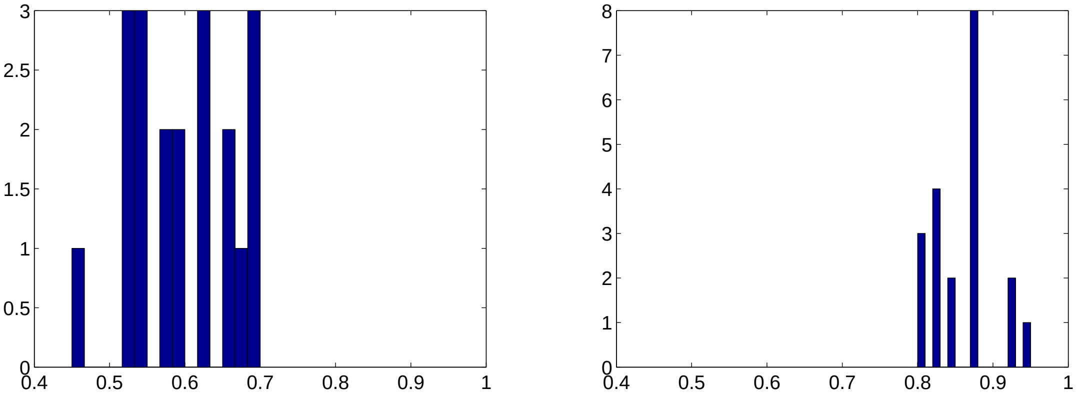
(b) Node 2

Figure 3.3. Connectivity (β_i^θ) of 2 distinct nodes of the ordinary (red line) and the compressed (black line) random field models. The connectivity of the compressed random field clearly shows a sparsity pattern.

3.2 CMU Data Set

This data set is the one that was used in [10]. The detailed description of this data set can be found at the StarPlus web-site at http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/. In this example we have used the first image only of each trial in which the subject is being shown either a picture or a sentence. Thus, only 40 trials are used out of the 54 that are included in this data set (in all others the subject is in rest). As suggested by the provided documentation in the StarPlus web-site we have used the following region of interest: 'CALC', 'LIPL', 'LT', 'LTRIA', 'LOPER', 'LIPS', 'LDLPFC'. The results are shown for the subject numbered 04847.

The classification performance of the algorithm over 20 Monte Carlo runs is shown in Fig. 3.4. With a total of 80 nodes the mean accuracy with and without compression is 87% and 59%, respectively.

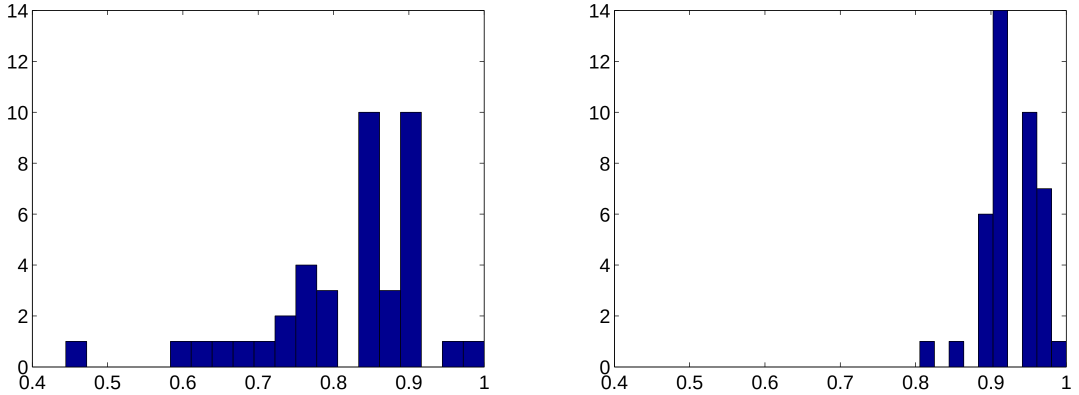


(a) Without compression **(b) With compression**

Figure 3.4. Distribution of classification accuracy based on 20 Monte Carlo runs. The mean accuracy when using compression increases to 87%, approximately 28% more than in the uncompressed case. Random field consists of 80 nodes.

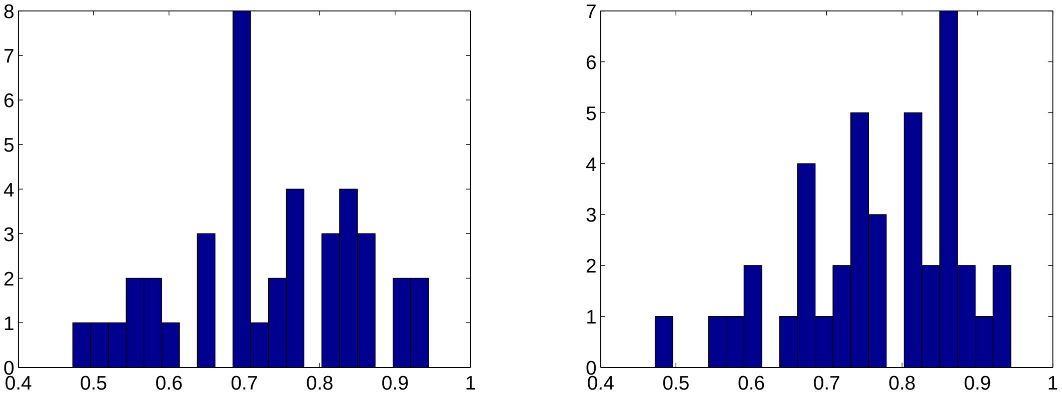
3.3 Neurospin Data Set

This data set consists of 9 normal and 9 mentally ill subjects. There are two samples associated with each subject. As before, the algorithm is tested using two-out cross validation. However, in this case every sample consists of two trails performed by the same subject. The performance of the algorithm with and without compression is shown in Fig. 3.5. In this case compression improves the mean accuracy by more than 10% (82% without compression and 93% with compression). The last figure for this data set, Fig. 3.6, demonstrates the effect of adaptation on the classification accuracy of the uncompressed variant.



(a) Without compression **(b) With compression**

Figure 3.5. Distribution of classification accuracy based on 40 Monte Carlo runs. The mean accuracy when using compression increases to 93%, approximately 11% more than in the uncompressed case. Random field consists of 100 nodes.



(a) Without adaptation

(b) With adaptation

Figure 3.6. Distribution of classification accuracy based on 40 Monte Carlo runs. The mean accuracy when using adaptation increases to 77%, approximately 4% more than in the non adaptive case. Random field consists of 50 nodes. No compression is used.

Appendix A

A.1 Optimal Tuning of c

The constant c in (1.10) can be taken as the one maximizing the accuracy of prediction based on some development dataset. Let $X_{dev}^{\theta} = \{X_{\theta}(1), \dots, X_{\theta}(k_{\theta})\}$ be a dataset associated with the class $Y = \theta$, and let also

$$d_{\theta}(j) := \log p(\bar{X}_{\theta}(j) \mid Y = 0) - \log p(\bar{X}_{\theta}(j) \mid Y = 1) \tag{A.1}$$

We are aimed at minimizing the following objective function

$$c^{*} = \arg\max_{c} \left[\eta_1 \sum_{j=1}^{k_1} \mathbf{1}\left(d_{\theta=1}(j) \le c\right) + \eta_0 \sum_{j=1}^{k_0} \mathbf{1}\left(d_{\theta=0}(j) \ge c\right) \right] \tag{A.2}$$

where $\mathbf{1}(a \in A)$ is the indicator function of the event $a \in A$ (i.e., a function which takes the value 1 if $a \in A$, and takes the value 0 otherwise). The constants η_0 and η_1 are the relative counts of both classes, that is, $\eta_0 := k_0/(k_0 + k_1)$ and $\eta_1 := k_1/(k_0 + k_1)$.

A.2 Feature Ranking

The preliminary results obtained in this work are based on a reduced feature space obtained using either of the following feature selection methods.

A.2.1 Cross Correlation Ranking

The cross correlation of the ith element $X_i \in X$ and Y is given by

$$\rho_{X_i,Y} = \frac{E\{(X_i - \mu_{X_i})(Y - \mu_Y)\}}{(E\{X_i\}^2 - \mu_{X_i}^2)^{1/2}(E\{Y\}^2 - \mu_Y^2)^{1/2}} \tag{A.3}$$

Given k training samples $X_{train} = \{X(1), ...X(k)\}$ with known responses $Y_{train} = \{Y(1), ...Y(k)\}$, Eq. (A.3) is approximated by

$$\hat{\rho}_{X_i,Y} = \frac{k \sum_{j=1}^{k} X_i(j)Y(j) - \sum_{j=1}^{k} X_i(j) \sum_{j=1}^{k} Y(j)}{(k \sum_j X_i(j)^2 - (\sum_j X_i(j))^2)^{1/2}(k \sum_j Y(j)^2 - (\sum_j Y(j))^2)^{1/2}} \tag{A.4}$$

The reduced set $\bar{X}$ is then obtained as

$$\bar{X} = \{X_i \mid \hat{\rho}_{X_i,Y} \ge \rho_{Th}\} \tag{A.5}$$

where $\rho_{Th} > 0$ is some predetermined threshold value.

A.2.2 Mutual Information Ranking

The mutual information (MI) of X_i and Y is given as

$$\begin{aligned} I(X_i, Y) &= \sum_{Y} \sum_{X_i \in X_{train}} p(X_i, Y) \log \left[\frac{p(X_i, Y)}{p(X_i)p(Y)} \right] \\ &= \sum_{Y} \sum_{X_i \in X_{train}} p(X_i \mid Y)p(Y) \log \left[\frac{p(X_i \mid Y)}{p(X_i)} \right] = \sum_{Y=0,1} p(Y) \sum_{X_i \in X_{train}} p(X_i \mid Y) \log \left[\frac{p(X_i \mid Y)}{p(X_i)} \right] \end{aligned} \tag{A.6}$$

Let us assume that $p(Y=0)=p(Y=1)=1/2$ (i.e., balanced training set), and

$$p(X_i \mid Y) = \mathcal{N}\left(X_i - \mu_{X_i|Y}, \sigma^2_{X_i|Y}\right) \tag{A.7a}$$

$$p(X_i) = \mathcal{N}\left(X_i - \mu_{X_i}, \sigma^2_{X_i}\right) \tag{A.7b}$$

The statistics of the Gaussian pdfs above can be approximated using the training data samples as

$$\mu_{X_i|Y} = \frac{1}{k_Y}\sum_{j=1}^{k_Y} X_i^Y(j) \tag{A.8a}$$

$$\sigma^2_{X_i|Y} = \frac{1}{k_Y - 1}\sum_{j=1}^{k_Y}(X_i^Y(j) - \mu_{X_i|Y})^2 \tag{A.8b}$$

$$\mu_{X_i} = \frac{1}{2}\mu_{X_i|Y=0} + \frac{1}{2}\mu_{X_i|Y=1} \tag{A.8c}$$

$$\sigma^2_{X_i} = \frac{k_0 - 1}{k-1}\sigma^2_{X_i|Y=0} + \frac{k_1 - 1}{k-1}\sigma^2_{X_i|Y=1} \tag{A.8d}$$

where

$$X_i^a(j) = \{X_i \in X(j) \cap Y(j) = a\} \tag{A.9}$$

and k_Y denotes the number of training samples of class Y. Substituting the above in Eq. (A.6) yields

$$\begin{aligned} I(X_i, Y) = \frac{1}{2}\sum_{\theta=0,1}\sum_{j=1}^{k} C_\theta^i \exp\left\{-\frac{1}{2}\frac{(X_i(j) - \mu_{X_i|Y=\theta})^2}{\sigma^2_{X_i|Y=\theta}}\right\} \\ \times\left[\log C_\theta^i - \frac{1}{2}\frac{(X_i(j) - \mu_{X_i|Y=\theta})^2}{\sigma^2_{X_i|Y=\theta}} - \log \bar{C}^i + \frac{1}{2}\frac{(X_i(j) - \mu_{X_i})^2}{\sigma^2_{X_i}}\right] \end{aligned} \tag{A.10}$$

where C_θ^i and $\bar{C}^i$ are normalization constants

$$C_\theta^i = \left[\sum_{j=1}^{k}\exp\left\{-1/2\frac{(X_i(j) - \mu_{X_i|Y=\theta})^2}{\sigma^2_{X_i|Y=\theta}}\right\}\right]^{-1} \tag{A.11a}$$

$$\bar{C}^i = \left[\sum_{j=1}^{k}\exp\left\{-1/2\frac{(X_i(j) - \mu_{X_i})^2}{\sigma^2_{X_i}}\right\}\right]^{-1} \tag{A.11b}$$

The reduced set $\bar{X}$ is then obtained as

$$\bar{X} = \{X_i \mid I(X_i, Y) \geq I_{Th}\} \tag{A.12}$$

where $I_{Th} > 0$ is some predetermined threshold value.

A.3 Convergence Aspects

A.3.1 Ergodic Sums

If the following conditions hold

- The uncertainty random variables W_i, $\forall i$ are independent and identically distributed (iid).
- The Jacobian $\nabla_{X_i}\varphi_\theta^{-1}(X_i, G_i)$ is independent of θ.

then

$$l = \prod_{i=1}^{n}\frac{p_W(\varphi^{-1}_{\theta=0}(X_i, G_i))}{p_W(\varphi^{-1}_{\theta=1}(X_i, G_i))} \underset{<}{\overset{>}{}} \exp\{c\} \tag{A.13}$$

can be interpreted as a likelihood ratio test where nodes act as samples. It can be shown (using the strong ergodic theorem or the strong law of large numbers) that in this case

$$\lim_{n\to\infty} l = \begin{cases} +\infty, & \text{if } \theta = 0 \text{ is the true class} \\ 0, & \text{if } \theta = 1 \text{ is the true class} \end{cases} \tag{A.14}$$

The above argumentation implies that regardless of the value of c the test yields the correct class for some $n > n'$, the number of nodes in the RF model.

A.3.2 Convergence to a True Class

It has been pointed out that the accuracy (i.e., convergence to the correct class) depends on the value of c and the number of nodes n. Under the conditions previously mentioned the strong law of large numbers (SLLN) yields

$$\lim_{n\to\infty} \frac{1}{n} \log l = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i \mid G_i, \theta = 0)}{p(X_i \mid G_i, \theta = 1)} = \alpha \tag{A.15}$$

where

$$\alpha := \begin{cases} -KL\left\{p(\cdot \mid \cdot, \theta = 0) \ \| \ (p(\cdot \mid \cdot, \theta = 1)\right\}, & \text{if } \theta = 0 \text{ is the true class} \\ \ \ KL\left\{p(\cdot \mid \cdot, \theta = 1) \ \| \ (p(\cdot \mid \cdot, \theta = 0)\right\}, & \text{if } \theta = 1 \text{ is the true class} \end{cases} \tag{A.16}$$

and $KL\{p_1 \parallel p_2\}$ denotes the Kullback-Leibler divergence between the pdfs p_1 and p_2. Note that the definition (A.16) implies $\alpha < 0$ if the true class is $\theta = 0$ and $\alpha > 0$ if the true class is $\theta = 1$. According to the central limit theorem

$$\zeta = (\frac{1}{n} \log l - \alpha) \sim \mathcal{N}\left(0, O(1/n)\right) \tag{A.17}$$

assuming large enough n. Thus,

$$\frac{1}{n} \log l = \alpha + \zeta, \qquad \zeta = O(1/n^{1/2}) \tag{A.18}$$

or, equivalently

$$l = \exp\{n\alpha\} \exp\{O(1/n^{1/2})\} \tag{A.19}$$

Eqs. (A.13) and (A.19) imply

$$\exp(n\alpha) \exp(O(1/n^{1/2})) \gtrless \exp(c) \tag{A.20}$$

yielding

$$\exp(n\alpha) \gtrless \exp(c - O(1/n^{1/2}) \tag{A.21}$$

and

$$\alpha \gtrless \frac{c}{n} - O(1/n^{3/2}) \tag{A.22}$$

The above clearly shows that the effect of c diminishes as $n \to \infty$. Moreover, the accuracy depends on c, n and α, the expected discrimination information of one class over the other.

Bibliography

[1] Mendel, J. M., *Lessons in Estimation Theory for Signal Processing, Communications, and Control*, Prentice Hall, 1995.

[2] Gopalakrishnan, P., Kanevsky, D., Nahamoo, D., and Nadas, A., "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," *IEEE Trans. Information Theory*, Vol. 37, No. 1, January 1991.

[3] Kanevsky, D., "Extended Baum Transformations For General Functions, II," Tech. Rep. RC23645(W0506-120), Human Language Technologies, IBM, 2005.

[4] Carmi, A. and Kanevsky, D., "Matrix form of Extended Baum Transformations," Tech. Rep. RC, Human Language Technologies, IBM, 2008.

[5] Carmi, A., Gurfil, P., and Kanevsky, D., "A Simple Method for Sparse Signal Recovery from Noisy Observations Using Kalman Filtering," Tech. Rep. RC24709, Human Language Technologies, IBM, 2008.

[6] Candes, E. J., Romberg, J., and Tao, T., "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information," *IEEE Transactions on Information Theory*, Vol. 52, 2006.

[7] Chartrand, R., "Exact Reconstruction of Sparse Signals via Nonconvex Minimization," *IEEE Signal Processing Letters*, Vol. 14, 2007, pp. 707–710.

[8] Candes, E. J., "Compressive Sampling," European Mathematical Society, Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006.

[9] Rish, I., Grabarnik, G., Cecchi, G., Periera, F., and Gordon, G. J., "Closed-Form Supervised Dimensionality Reduction with Generalized Linear Models," The 25th International Conference on Machine Learning, Helsinki, Finland, 2008.

[10] Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., and Newman, S., "Learning to Decode Cognitive States from Brain Images," *Machine Learning*, Vol. 57, 2004, pp. 145–175.