Abstract Although Relevance Vector Machine (RVM) is the most popular algorithms in machine learning and computer vision, outliers in the training data make the estimation unreliable. In the paper, a robust RVM model under non-parametric Bayesian framework is proposed. We decompose the noise term in the RVM model into two components, a Gaussian noise term and a spiky noise term. Therefore the observed data is assumed represented as: \( y = Dw + s + e \), where \( Dw \) is the relevance vector component, of which \( D \) is the kernel function matrix and \( w \) is the weight matrix, \( s \) is the spiky term and \( e \) is the Gaussian noise term. A spike-slab sparse prior is imposed on the weight vector \( w \), which gives a more intuitive constraint on the sparsity than the Student’s t-distribution described in the traditional RVM. For the spiky component \( s \), a spike-slab sparse prior is also introduced to recognize outliers in the training data effectively. Several experiments demonstrate the better performance over the RVM regression.

Key words Relevance Vector Machine (RVM); Bayesian nonparametric; Outliers; Spike-slab sparse prior

CLC index TN911.7

DOI 10.1007/s11767-012-0873-0

I. Introduction

Support Vector Machine (SVM) and Relevance Vector Machine (RVM) are the most popular algorithms in machine learning and computer vision such as face recognition, image denoising and image reconstruction. Compared with SVM, RVM has capability in probabilistic predictions and automatic estimation of the error/margin trade-off parameter. It is unnecessary for Kernel functions to meet Mercer’s condition in RVM model. The largest contribution of RVM is that it generates similar performance to SVM with fewest kernel functions. Although RVM is widely used, outliers in the training data (such as salt and pepper noise in the images) make the estimation unreliable.

In the paper, a robust RVM model under non-parametric Bayesian framework is proposed. We decompose the noise term in the RVM model into two components, a Gaussian noise term and an outlier noise term, which we assume to be sparse. Therefore the observed data is assumed represented as: \( y = Dw + s + e \), where \( Dw \) is the relevance vector component, \( D \) is the kernel function matrix and \( w \) is the weight vector; \( s \) represents the sparse term and \( e \) is the Gaussian noise term.

In RVM, the sparsity is often desired when most of the relevance vectors are irrelevant to prediction of the observed data. In the paper, we consider a method for finding sparse solutions to the weight vector \( w \) using the spike-slab sparse prior, which gives a more intuitive constraint on the sparsity than the Student’s t-distribution described in the RVM. Spike-slab prior is also imposed on the sparse component \( s \) to recognize outliers in the training data effectively.

The rest of the paper is organized as follows. We review spike-slab model and present the proposed robust RVM method in the next section. Experiment results on mixture noise denoising and reconstruction of the background and the foreground are showed in Section III. We then conclude in Section IV.
II. Robust RVM Regression

1. Spike-slab model

The spike and slab model first introduced by Mitchell & Beauchamp gives a point with great probability at zero and a bounded uniform distribution elsewhere in the parameter space\(^7\). It specifies mixture prior distributions for each of the coefficient parameters where the spike part helps detect coefficients of zero magnitude by shrinking those coefficients to zero in posterior terms, and the slab part provides the non-zero coefficients with a normal prior thus not requiring extensively informed prior knowledge. Formally, the prior distribution is as follows:

\[
p(\beta_k = 0) = h_k, \quad p(\beta_k < b, \beta_k \neq 0) = (b + f_k)h_k, \\
-\ f_k < b < f_k
\]  

where \(h_k > 0, h_k > 0, \) and \(h_k + 2f_kh_k = 1\). We take \(f_k\) and \(\gamma_k\) the parameter of this distribution, where \(\gamma_k = h_k / h_k\) (the ratio of the heights of the two parts).

George & McCulloch changes the discrete prior specification into a continuous mixture distribution\(^8\). They introduce a new parameter \(\gamma_k\) which is a binomial variable \((\{0, 1\})\) in the following prior distribution:

\[
\beta_k | \gamma_k \sim (1 - \gamma_k)N(0, \sigma_k^2) + \gamma_kN(0, \sigma^2_k) \\
p(\gamma_k = 1) = 1 - p(\gamma_k = 0) = p_k
\]

where we assume that \(\sigma_k^2\) is such a small value and then once \(\beta_k\) falls into the first normal distribution, it is concentrated around 0.

2. Robust RVM model specification

The sparse representation of a signal is modeled by\(^9\):

\[
y = Dw + \varepsilon
\]

where \(y\) is an \(N \times 1\) signal vector, \(w\) is an \((N + 1) \times 1\) sparse coefficient vector, \(D\) is an \(N \times (N + 1)\) matrix called the dictionary and \(\varepsilon\) is an \(N \times 1\) noise vector.

In the RVM formulation, the dictionary \(D\) is generated by the kernel function and remains unchanged, and then the model can be expressed as:

\[
y_i = \sum_{j=1}^{N} w_{ij}K(x_i, x_j) + \varepsilon_i, \quad i = 1, 2, \ldots, N
\]

where, with each \(x_j\), there is an associated kernel function \(K(\cdot, x_j)\).

However, the model is not robust to outliers in the training data with the Gaussian prior of the noise vector and will make the estimation unreliable. In the proposed method, we proposed to split the noise term into two components: a Gaussian noise term \(\varepsilon\) and a spiky term \(s\), which we assume to be sparse, and therefore:

\[
y = Dw + s + \varepsilon
\]

where \(s = [s_{1}, s_{2}, \ldots s_{N}]^T\), and \(\varepsilon = [\varepsilon_{1}, \varepsilon_{2}, \ldots \varepsilon_{N}]^T\).

It is known a prior that most of the features, or columns of \(D\) are irrelevant to the prediction of \(y\), which means most coefficients in \(w\) are exactly zero, while a few coefficients different from zero. Since traditional spike-slab model discussed above can not satisfy this requirement, we modify the model as follow:

\[
w_i \sim (1 - \gamma_i)\delta(w_i) + \gamma_iN(0, \sigma^2_i)
\]

delta distribution introduced here is to enhance the sparsity of the spike term, while once falls into the first term in Eq. (7), \(w_i\) is set to be zero with a high probability.

\(\gamma_i\) is a binary variable and generated as:

\[
r_i \sim \text{Bern}(\pi_i)
\]

The parameter of the Binomial distribution \(\pi_i\) defines the probability weather choosing delta function to govern zero or Gaussian distribution with variance \(\sigma_i^2\) to weight non-zero elements of \(w\). To complete the specification of this hierarchical prior, we must define hyperpriors over \(\pi_i\), suitable prior is beta distribution with parameters \(a\) and \(b\):

\[
\pi_i \sim \text{Beta}(a, b)
\]

The sparse term is modeled as follows:

\[
s_i \sim (1 - p_i)\delta(s_i) + p_iN(0, \sigma_i^2)
\]

\[
p_i \sim \text{Bern}(q_i)
\]

\[
q_i \sim \text{Beta}(c, d)
\]