ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 335-340

Sparse Kernel Canonical Correlation Analysis

Lili Tan[1] and Colin Fyfe[2],*

1. Department of Computer Science and Engineering,
The Chinese University of Hong Kong,
Hong Kong.
2. School of Information and Communication Technologies,
The University of Paisley,
Scotland.

Abstract. We review the recently proposed method of Relevance Vector Machines which is a supervised training method related to Support Vector Machines. We also review the statistical technique of Canonical Correlation Analysis and its implementation in a Feature Space. We show how the technique of Relevance Vectors may be applied to the method of Kernel Canonical Correlation Analysis to gain a very sparse representation of a data set and discuss why such a representation may be beneficial to an organism.

1. Introduction

We have previously developed both neural [1] and kernel [2] methods for finding correlations in a data set which are greater than linear correlations. The kernel method was based on the supervised method of Support Vector Machines (SVMs)[7] but lacked the inherent sparseness which SVMs generate - the filters found are functions of the whole data set. In this paper, we use the ideas from Relevance Vector Machines (RVMs) to generate Canonical Correlation vectors which are inherently sparse. We have previously [1] shown how canonical correlation analysis networks can be used to identify stereo correspondence in a data set. While this is still possible with kernel methods, the power of nonlinear kernels is best illustrated on more difficult correspondence problems such as the integration of information across different modalities such as sight and sound.

*Colin Fyfe would like to gratefully acknowledge the support of the Chinese University of Hong Kong for the period when the current work was carried out.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 335-340

2. Kernel Canonical Correlation Analysis

Canonical Correlation Analysis [3] is used when we have two data sets which we believe have some underlying correlation. Consider two sets of input data, from which we draw iid samples to form a pair of input vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$. Then in classical CCA, we attempt to find the linear combination of the variables which gives us maximum correlation between the combinations.

Let Σ_{11} be the covariance matrix of $\mathbf{x}_1$ and similarly with Σ_{12} and Σ_{22}. Then define $K = \Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$. We then perform a Singular Value Decomposition of K to get

$$K = (\alpha_1, \alpha_2, ..., \alpha_k)D(\beta_1, \beta_2, ..., \beta_k)^T \tag{1}$$

where α_i and β_i are the standardised eigenvectors of KK^T and K^TK respectively and D is the diagonal matrix of eigenvalues.

Then the first canonical correlation vectors (those which give greatest correlation) are given by

$$\mathbf{w}_1 = \Sigma_{11}^{-\frac{1}{2}}\alpha_1 \tag{2}$$

$$\mathbf{w}_2 = \Sigma_{22}^{-\frac{1}{2}}\beta_1 \tag{3}$$

with subsequent canonical correlation vectors defined in terms of the subsequent eigenvectors, α_i and β_i.

We have recently extended CCA into the Kernel domain [2]. Consider mapping the input data to a high dimensional (perhaps infinite dimensional) feature space, F. Now, assuming the data has been centred, (we actually will use the same trick as [5] to centre the data later) the covariance matrices in Feature space are defined by

$$\begin{aligned}\Sigma_{11} &= E\{\Phi(\mathbf{x}_1)\Phi(\mathbf{x}_1)^T\}\\ \Sigma_{22} &= E\{\Phi(\mathbf{x}_2)\Phi(\mathbf{x}_2)^T\}\\ \Sigma_{12} &= E\{\Phi(\mathbf{x}_1)\Phi(\mathbf{x}_2)^T\}\end{aligned}$$

and we wish to find those values $\mathbf{w}_1$ and $\mathbf{w}_2$ which will maximise $\mathbf{w}_1^T\Sigma_{12}\mathbf{w}_2$ subject to the constraints $\mathbf{w}_1^T\Sigma_{11}\mathbf{w}_1 = 1$ and $\mathbf{w}_2^T\Sigma_{22}\mathbf{w}_2 = 1$.

In practise we will approximate Σ_{12} with $\frac{1}{M}\sum_i \Phi(\mathbf{x}_{1i})\Phi(\mathbf{x}_{2i})$, the sample average.

We [2] have shown that, using $(K_1)_{ij} = \Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_{1j})$ and $(K_2)_{ij} = \Phi^T(\mathbf{x}_i)$ $\Phi(\mathbf{x}_{2j})$ we require to maximise $\alpha^TK_1K_2^T\beta$ subject to the constraints $\alpha^TK_1K_1^T\alpha$ $= 1$ and $\beta^TK_2K_2^T\beta = 1$. Therefore if we define $\Gamma_{11} = K_1K_1^T$, $\Gamma_{22} = K_2K_2^T$ and $\Gamma_{12} = K_1K_2^T$ we solve the problem in the usual way: by forming matrix $K = \Gamma_{11}^{-\frac{1}{2}}\Gamma_{12}\Gamma_{22}^{-\frac{1}{2}}$ and performing a singular value decomposition on it as before to get

$$K = (\gamma_1, \gamma_2, ..., \gamma_k)D(\theta_1, \theta_2, ..., \theta_k)^T \tag{4}$$

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 335-340

where γ_i and θ_i are again the standardised eigenvectors of KK^T and K^TK respectively and D is the diagonal matrix of eigenvalues [1]

Then the first canonical correlation vectors in feature space are given by

$$\alpha_1 = \Gamma_{11}^{-\frac{1}{2}}\gamma_1 \quad (5)$$

$$\beta_1 = \Gamma_{22}^{-\frac{1}{2}}\theta_1 \quad (6)$$

with subsequent canonical correlation vectors defined in terms of the subsequent eigenvectors, γ_i and θ_i.

Now for any new values $\mathbf{x}_1$, we may calculate

$$\mathbf{w}_1.\Phi(\mathbf{x}_1) = \sum_i \alpha_i \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_1) = \sum_i \alpha_i K_1(\mathbf{x}_i, \mathbf{x}_1) \quad (7)$$

which then requires to be centered as before. We see that we are again performing a dot product in feature space (it is actually calculated in the subspace formed from projections of $\mathbf{x}_i$).

The method requires a matrix inversion and the data sets may be such that one data point may be repeated (or almost) leading to a singularity or badly conditioned matrices. One solution is to add μI, where I is the identity matrix to Γ_{11} and Γ_{22} - a method which was also used in [4]. This gives robust and reliable solutions however it detracts from the precise analytical foundations of the method. The difficulty comes about because the method, like other unsupervised kernel methods, does not have the automatic identification of important data points which the supervised method of Support Vector Machines has. We will use the ideas from Relevance Vector Machines [6] to regain this feature.

3. Relevance Vector Regression

The vectors found by the Relevance Vector Machine method are prototypical vectors of the class types which is a very different concept from the Support Vectors whose positions are always at the edges of clusters, thereby helping to delimit one cluster from another. Relevance Vector Regression uses a dataset of input-target pairs$\{\mathbf{x}_i, t_i\}_{i=1}^N$ It assumes that the machine can form an output y from

$$y(\mathbf{x}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0 \quad (8)$$

and $p(t|\mathbf{x})$ is Gaussian $N(y(\mathbf{x}), \sigma^2)$. The likelihood of the model is given by

$$p(\mathbf{t}|\mathbf{w}, \sigma) = \frac{1}{(2\pi\sigma^2)^{-\frac{N}{2}}} \exp\{-\frac{1}{2\sigma^2}||\mathbf{t} - \Phi\mathbf{w}||^2\} \quad (9)$$

[1] This optimisation is applicable for all symmetric matrices (Theorem A.9.2, [3]).

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 335-340

where $\mathbf{t} = \{t_1, t_2, ..., t_N\}, \mathbf{w} = \{w_0, w_1, ..., w_N\}$, and Φ is the $N*(N+1)$ design (data) matrix. To prevent overfitting, an automatic relevance detection prior is set over the weights

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{N} N(0, \alpha^{-1}) \tag{10}$$

To find the maximum likelihood of the data set with respect to α and σ^2, we iterate between finding the mean and variance of the weight vector and then calculating new values for α and σ^2 using these statistics. We find that many of the α_i tend to infinity which means that the corresponding weights tend to 0. In detail, we have that the posterior of the weights is given by

$$p(\mathbf{w}|\mathbf{t}, \alpha, \sigma^2) \propto |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{w} - \mu)^T \Sigma^{-1} (\mathbf{w} - \mu)\} \tag{11}$$

where

$$\begin{aligned} \Sigma &= (K^T B K + A)^{-1} \\ \mu &= \Sigma K^T B \mathbf{t} \end{aligned} \tag{12}$$

with $A = diag(\alpha_0, \alpha_1, ..., \alpha_N)$ and $B = \sigma^{-2} I_N$.

If we integrate out the weights, we obtain the marginal likelihood

$$p(\mathbf{t}|\alpha, \sigma^{\mathbf{2}}) \propto |B^{-1} + K A^{-1} K^T|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{t}^T (B^{-1} + K A^{-1} K^T)^{-1} \mathbf{t}\} \tag{13}$$

which can be differentiated to give at the optimum,

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} \tag{14}$$

$$(\sigma^2)^{new} = \frac{||\mathbf{t} - \mathbf{K}\mu||^2}{(N - \sum_i \gamma_i)} \tag{15}$$

where $\gamma_i = 1 - \alpha_i \Sigma_{ii}$.

4. Application to CCA

We must first describe CCA in probabilistic terms. Consider two data sets $\Theta_1 = \{\mathbf{x}_1^i, i \in 1...N\}$ and $\Theta_2 = \{\mathbf{x}_2^i, i \in 1...N\}$ defined by probability density functions $p_1(\mathbf{x}_1)$ and $p_2(\mathbf{x}_2)$. Let there be some underlying relationship so that $y_1 = \mathbf{w}_1^T \mathbf{x}_1 + \epsilon_1$ is the canonical correlate corresponding to $y_2 = \mathbf{w}_2^T \mathbf{x}_2 + \epsilon_2$. Then y_1 can be used to predict the value of y_2 and vice versa. Let $y_1 = \rho y_2 + e_1$, where ρ is the correlation coefficient.

Then from the perspective of $\mathbf{x}_1$, the targets, $\mathbf{t}_1$ is given by the other input $\mathbf{x}_2$. i.e.

$$\mathbf{t}_1 = \mathbf{w}_2 K_2 \tag{16}$$

So we are using the current value of $\mathbf{w}_2$ to determine the target for updating the posterior probabilities for $\mathbf{w}_1$. Similarly, we create a target for updating the

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 335-340

probabilities for $\mathbf{w}_2$ using the current value of $\mathbf{w}_1$. We then simply alternate between the two Relevance Vector Machines in a way which is reminiscent of the EM algorithm: from $\mathbf{x}_1$'s perspective, calculation of the new values of $\mathbf{w}_2$ corresponds to the E-step while the calculation of the new values of $\mathbf{w}_1$ corresponds to the M-step and vice-versa from $\mathbf{x}_2$'s perspective.

We have carried out experiments on real and artificial data sets: for example, we generate data according to the prescription:

$$x_{11} = 1 - \sin\theta + \mu_1 \quad (17)$$

$$x_{12} = cos\theta + \mu_2 \quad (18)$$

$$x_{21} = \theta + \mu_3 \quad (19)$$

$$x_{22} = \theta + \mu_4 \quad (20)$$

where θ is drawn from a uniform distribution in $[-\pi, \pi]$ and $\mu_i, i = 1, ..., 4$ are drawn from the zero mean Gaussian distribution N(0, 0.1). Equations (17) and (18) define a circular manifold in the two dimensional input space while equations (19) and (20) define a linear manifold within the input space where each manifold is only approximate due to the presence of noise ($\mu_i, i = 1, ..., 4$).

Thus $\mathbf{x}_1 = \{x_{11}, x_{12}\}$ lies on or near the circular manifold while $\mathbf{x}_2 = \{x_{21}, x_{22}\}$ lies on or near the line. We have previously shown [2] that the kernel method can find greater than linear correlations in such a data set. We now report that the current Relevance Vector Kernel method also finds greater than linear correlations but with a very sparse representation generally. Typically the resulting vectors will have zeros in all but one position with the single non-zero value being very strong. Occasionally, one vector, e.g. $\mathbf{w}_1$ will have a single non-zero value whose correlation with all the other points will be seen in its vector $\mathbf{w}_2$. However in all cases we find a very strong correlation. Figure 1 shows the outputs of a trained CCA-RVM network; the high correlation between y_1 and y_2 is clear. We may also use the CCA-RVM network to find stereo correspondences as before [1], however this task is relatively easy and does not use the full power of the CCA-RVM network.

5. Conclusion

We have previously developed both neural [1] and kernel [2] methods for finding correlations which are greater than linear correlations in a data set. However the previous kernel method found weight vectors which depended on the whole data set. Using the Relevance Vector Machine to determine the Kernel correlations, we still get greater than linear correlations but use only a small number of data points to do so. For an organism which must integrate information from different sensory modalities (or indeed from the same modality but two different organs e.g. two eyes), this is an important saving since the organism need not maintain all its previous memories from the two data streams. The method automatically identifies the most relevant memories which maximise the correlation in feature space.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 335-340

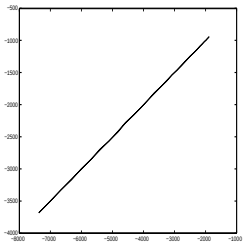

Figure 1: The figure shows a graph of y_1 (horizontal axis) against y_2 (vertical axis) for a Relevance Vector CCA network trained on the negative distance kernel. The high correlation is obvious.

References

[1] P. L. Lai and C. Fyfe. A neural network implementation of canonical correlation analysis. *Neural Networks*, 12(10):1391–1397, Dec. 1999.

[2] P.L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 2001. (Accepted for publication).

[3] K. V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[4] S. Mika, B. Scholkopf, A. Smola, K.-R. Muller, M. Scholz, and G. Ratsch. Kernel pca and de-noising in feature spaces. In *Advances in Neural Processing Systems, 11*, 1999.

[5] B. Scholkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. J. Smola. Input space vs feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10:1000–1017, 1999.

[6] M. Tipping. The relevance vector machine. In Leen T. K. Solla, S. A. and K.-R. Muller, editors, *Advances in Neural Information Processing Systems, 12*. MIT Press, 2000.

[7] V Vapnik. *The nature of statistical learning theory*. Springer Verlag, New York, 1995.