A metamorphosis of Canonical Correlation Analysis into Multivariate Maximum Margin Learning

Sandor Szedmak[1], Tijl De Bie[2] and David R. Hardoon[3] *

1 - Electronics and Computer Science, ISIS Group
University of Southampton, SO17 1BJ, United Kingdom

2 - K.U. Leuven, OKP, Tiensestraat 102, 3000 Leuven, Belgium, and,
University of Bristol, Engineering Mathematics, BS8 1TR, United Kingdom
3 - Department of Computer Science
University College London, WC1E 6BT, United Kingdom

Abstract. Canonical Correlation Analysis(CCA) is a useful tool to discover relationship between different sources of information represented by vectors. The solution of the underlying optimisation problem involves a generalised eigenproblem and is nonconvex. We will show a sequence of transformations which turn CCA into a convex maximum margin problem. The new formulation can be applied for the same class of problems at a significantly lower computational cost and with a better numerical stability.

1 Introduction

In machine learning one of the most important questions is how one can guarantee satisfactory performance of a learner on earlier unseen cases. To be more precise, let us define the (supervised) learning problem in the following way. The learner receives a subset of ordered pairs (a sample) $\mathcal{S} = \{(x_i, y_i),\ i = 1, \ldots, m\}$ from the direct product $(\mathcal{S} \subseteq)\mathcal{X} \times \mathcal{Y}$. We will refer to the domain as the input and to the set of values as the output in the sequel. The task is to find a function $f : \mathcal{X} \to \mathcal{Y}$ for which $f(x) \approx y$. To this end, one specifies a function class from which f can be chosen, and within this function class the f which minimises an approximation error on $\mathcal{S}$ is sought for. This search can generally be accomplished by formulating and solving an optimisation problem. The main building blocks of such an optimisation problem usually are: a loss function which measures the approximation error f makes on $\mathcal{S}$, and a regularisation term which restricts the functions space f is chosen from. Both these ingredients can be either terms in the objective function or constraints in the optimisation task.

How to find a proper function f in a given learning problem depends on the type of input and output space. If both are vector spaces, then we need learning methods that search for vector valued functions. For this purpose in classical statistics some regression based approaches have been developed. One of those tools is Canonical Correlation Analysis (CCA), which can be regarded as an extension of the Multivariate Linear Regression (MLR), where not only

*Supported by the European Community Projects SMART and PASCAL

the input but also the output may have several components, i.e. belongs to a higher dimensional vector space. MLR tries to find a linear combination of the input variables that maximises the correlation between the output variable and this linear combination. CCA looks for linear combinations of the input variables and the output variables that maximise the correlation between these linear combinations. The reader can consult to [1] and [5]. Both references give the kernelization of CCA referred to as KCCA.

In this paper, we will elucidate a connection between CCA and maximum margin learning. We will assume that $\mathcal{X}$ and $\mathcal{Y}$ are vector spaces. Our goal is to start from CCA and, by gradually transforming it, to arrive at a maximum margin learning problem that searches for a linear function f that maps inputs to outputs as given by $\mathbf{y} = f(\mathbf{x}) = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is a matrix or more generally a linear operator. The accuracy of a function f will be measured by the covariance between the real output $\mathbf{y}$ and its prediction $\mathbf{W}\mathbf{x}$. And lastly, the result will be invariant with respect to scalings of the input and the output vectors with the same factor.

To arrive at this maximum margin formulation, we will borrow some inspiration from a game theoretic scenario. In this scenario the learner (the first player) competes with nature (the other player), in trying to find the best function. Nature tries to present the learner with cases where the predictor function behaves poorly, thus forcing the learner to build up the most generally reliable function. A very recent book following this approach is [3].

The game behind the learning task accumulating all the aforementioned building blocks can be formalised as a minimax optimisation problem

$$\begin{array}{ll} \min_{\omega} \max_{\alpha} & L(\boldsymbol{\omega}, \boldsymbol{\alpha}) \\ \text{s.t.} & \boldsymbol{\omega} \in \mathcal{D}_{\omega},\ \boldsymbol{\alpha} \in \mathcal{D}_{\alpha}, \end{array} \tag{1}$$

where $\mathcal{D}_{\omega}$ and $\mathcal{D}_{\alpha}$ cover the allowed strategies of the learner and nature. Often the function L may be regarded as a Lagrangian functional, with $\boldsymbol{\omega}$ the primal and $\boldsymbol{\alpha}$ the dual variables. Hence, the minimax formulation can express a very wide range of reasonable optimisation problems. This minimax problem is convex if L is convex in ω and concave in α. Some of the most recent algorithms to solve the convex minimax problem are built around the extragradient method for solving variational inequalities, a further generalisation of the minimax schema, see for example [9] and [8]. These algorithms can tackle with very large scale problems in reasonable time.

2 From CCA to Maximum Margin Learning

Consider the supervised learning problem where we are given a sample $\{(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, \ldots, m\}$ with $\mathbf{x}_i \in \mathcal{X}$, $\mathbf{y}_i \in \mathcal{Y}$ and $\mathcal{X}, \mathcal{Y}$ linear vector spaces, both equipped with a bilinear inner product. With $\mathbf{X}$ and $\mathbf{Y}$ we will denote matrices containing the output and the input vectors in their rows respectively. Otherwise all vectors are column vectors. $\mathbf{1}$ is a vector of ones and $\mathbf{0}$ is a vectors of zeros. $\|\cdot\|$ denotes the

ℓ_2 norm of the vectors in its argument. In what follows we assume that the output and input vectors are centred: $\mathbf{1}^T\mathbf{x}_i = 0$ and $\mathbf{1}^T\mathbf{y}_i = 0$ for any $i = 1, \ldots, m$. $\|\mathbf{A}\|_F$ denotes the Frobenius norm of matrix $\mathbf{A}$ and $\langle \mathbf{A}, \mathbf{B}\rangle_F = \mathbf{trace}(\mathbf{A}^T\mathbf{B})$ is the Frobenius inner product. The notations for matrices are capitalised, and matrices and vectors are bold Latin or Greek characters.

CCA is formalised by the following optimisation problem:

$$\max_{w_x, w_y} \mathbf{corr}(\mathbf{Y}\mathbf{w}_y, \mathbf{X}\mathbf{w}_x) = \max_{w_x, w_y} \frac{\langle \mathbf{Y}\mathbf{w}_y, \mathbf{X}\mathbf{w}_x\rangle}{\|\mathbf{Y}\mathbf{w}_y\|\|\mathbf{X}\mathbf{w}_x\|}. \tag{2}$$

We will start from this formulation, and gradually transform it into a maximum margin problem. The first step in the metamorphosis is a simple reformulation to follow the linear prediction schema we put forward in the introduction:

$$\frac{\langle \mathbf{Y}\mathbf{w}_y, \mathbf{X}\mathbf{w}_x\rangle}{\|\mathbf{Y}\mathbf{w}_y\|\|\mathbf{X}\mathbf{w}_x\|} \quad \Rightarrow \quad \frac{\langle \mathbf{Y}, \mathbf{X}\mathbf{w}_x\mathbf{w}_y^T\rangle_F}{\|\mathbf{Y}\|_F\|\mathbf{X}\mathbf{w}_x\mathbf{w}_y^T\|_F}.$$

With this step the correlation between vectors is transformed into correlation between matrices. The numerators of both expression are equal to each other. Thus, the covariances, the unnormalised correlations, are equal. By substituting $\hat{\mathbf{W}}$ for $\mathbf{w}_x\mathbf{w}_y^T$. we arrive at what we call RegressiveCCA(RCCA). Now note that

$$\arg\max_{\hat{\mathbf{W}}} \frac{\langle \mathbf{Y}, \mathbf{X}\hat{\mathbf{W}}\rangle_F}{\|\mathbf{Y}\|_F\|\mathbf{X}\hat{\mathbf{W}}\|_F} \quad = \quad \arg\min_{\hat{\mathbf{W}}} \frac{\|\mathbf{Y}\|_F\|\mathbf{X}\hat{\mathbf{W}}\|_F}{\langle \mathbf{Y}, \mathbf{X}\hat{\mathbf{W}}\rangle_F}.$$

This is a fractional programming problem. Assuming that $\langle \mathbf{Y}, \mathbf{X}\hat{\mathbf{W}}\rangle_F \geq \lambda > 0$, this can be reformulated into a conditional optimisation problem (see e.g. [2]):

$$\begin{array}{ll} \min & \frac{1}{\lambda}\|\mathbf{Y}\|_F\|\mathbf{X}\hat{\mathbf{W}}\|_F \\ \text{s.t.} & \langle \mathbf{Y}, \mathbf{X}\hat{\mathbf{W}}\rangle_F \geq \lambda. \end{array} \tag{3}$$

Let $\mathbf{W} = \hat{\mathbf{W}}/\lambda$ and let the constant factor $\|\mathbf{Y}\|_F$ be dropped, then we get

$$\begin{array}{ll} \min & \|\mathbf{X}\mathbf{W}\|_F \\ \text{s.t.} & \langle \mathbf{Y}, \mathbf{X}\mathbf{W}\rangle_F \geq 1, \end{array} \tag{4}$$

a linearly constrained convex optimisation problem since the Frobenius norm is a convex function. It contains a data dependent regularisation term in the objective and a maximum margin constraint to force a high covariance whilst reducing the capacity of the learner. This task can be solved efficiently as a second order cone programming problem (see [6, 2]). Remember that when we unfolded the fractional problem in (3), we assumed that there is a λ that is strictly positive. When this is not the case, the constraint in (4) is infeasible. To solve this problem, we can introduce a slack variable in a similar way as in the soft margin formulation of Support Vector Machines (SVM's). This leads to

$$\begin{array}{ll} \min & \|\mathbf{X}\mathbf{W}\|_F + C\xi \\ \text{s.t.} & \langle \mathbf{Y}, \mathbf{X}\mathbf{W}\rangle_F \geq 1 - \xi, \end{array} \tag{5}$$

where C is a penalty constant balancing between the regularisation and the error expressed by ξ. This is a first efficiently solvable convex variant of CCA.

2.1 An SVM style solution

In this subsection, we further reformulate the optimisation problem, guided by game theoretic arguments as outlined in the introduction. In particular, we will formulate the problem as a minimax optimisation problem of the cost function, with the maximum taken over all strategies of nature, and the minimum over all strategies of the learner. To this end we first change the data dependent objective function $\|\mathbf{XW}\|_F$ into a data independent one $\frac{1}{2}\|\mathbf{W}\|_F^2$. Now, note that the constraint in (4) can be expressed as sum of terms corresponding to the sample items

$$\langle \mathbf{Y}, \mathbf{XW} \rangle_F = \textstyle\sum_{i=1}^m \langle \mathbf{y}_i, \mathbf{W}\mathbf{x}_i \rangle.$$

Then we write the following cost function as a function of a strategy $\boldsymbol{\alpha}$ of nature:

$$\begin{aligned} L(\mathbf{W}, \boldsymbol{\alpha}) &= \tfrac{1}{2}\|\mathbf{W}\|_F^2 \boxed{- \textstyle\sum_{i=1}^m \alpha_i \langle \mathbf{y}_i, \mathbf{W}\mathbf{x}_i \rangle + \sum_{i=1}^m \alpha_i}, \\ & \alpha_i \geq 0, \ i = 1, \dots, m. \end{aligned}$$

Regarding $L(\mathbf{W}, \boldsymbol{\alpha})$ as a Lagrangian, we can write the primal problem as

$$\begin{array}{ll} \min & \frac{1}{2}\|\mathbf{W}\|_F^2 \\ \text{s.t.} & \langle \mathbf{y}_i, \mathbf{W}\mathbf{x}_i \rangle \geq 1, \ i = 1, \dots, m. \end{array}$$

Hence, we have a minimax based maximum margin learner for arbitrary output vectors taken from any abstract linear space. Note that we can opt to incorporate slack variables, to allow some items to violate the general rule. Below we present the resulting problem parallel to the SVM learning problem, to highlight the similarities. In this general formulation, we allow that the input and output items are not taken directly from a linear vector space: they may be embedded by the functions ψ and ϕ into proper vector spaces denoted by $\mathcal{H}_\psi$ and $\mathcal{H}_\phi$.

	Binary class learning	Vector label learning
	Support Vector Machine(SVM)	Maximum Margin Robot(MMR)
min	$\frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}^T\boldsymbol{\xi}$	$\frac{1}{2}\|\mathbf{W}\|_F^2 + C\mathbf{1}^T\boldsymbol{\xi}$
w.r.t.	$\mathbf{w} : \mathcal{H}_\phi \rightarrow \mathbb{R}$, normal vector	$\mathbf{W} : \mathcal{H}_\phi \rightarrow \mathcal{H}_\psi$, linear operator
	$\boldsymbol{\xi} \in \mathbb{R}^m$, error vector	$\boldsymbol{\xi} \in \mathbb{R}^m$, error vector
s.t.	$y_i \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) \geq 1 - \xi_i$	$\langle \boldsymbol{\psi}(\mathbf{y}_i), \mathbf{W}\boldsymbol{\phi}(\mathbf{x}_i) \rangle_{\mathcal{H}_\psi} \geq 1 - \xi_i$
	$\boldsymbol{\xi} \geq \mathbf{0}, \ i = 1, \dots, m$	$\boldsymbol{\xi} \geq \mathbf{0}, \ i = 1, \dots, m$

Both problems imply duals with the same structure and computational complexity.

$$\begin{array}{ll} \max & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j=1}^m \alpha_i \alpha_j \overbrace{\langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle}^{\kappa_{ij}^\phi} \overbrace{\langle \boldsymbol{\psi}(\mathbf{y}_i), \boldsymbol{\psi}(\mathbf{y}_j))\rangle}^{\kappa_{ij}^\psi} \\ \text{w.r.t.} & \alpha_i \in \mathbb{R}, \\ \text{s.t.} & 0 \leq \alpha_i \leq C, \ i = 1, \dots, m. \end{array}$$

The predictor function in the vector learning case is given by

$$y \Leftarrow \tilde{\mathbf{y}} = \arg \boldsymbol{\psi}(\tilde{\mathbf{y}}) = \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x}) = \textstyle\sum_{i=1}^m \alpha_i \boldsymbol{\psi}(\mathbf{y}_i) \underbrace{\langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}) \rangle}_{\kappa_i^\phi(\mathbf{x})},$$

as linear combination of the known label vectors. If the outputs are not explicitly given then we need to solve pre-image problem. It can be solved by

$$\tilde{\mathbf{y}} = \arg\max_{y_t \in \hat{\mathcal{Y}}} \langle \boldsymbol{\psi}(\mathbf{y}_t), \boldsymbol{\psi}(\tilde{\mathbf{y}}) \rangle, \tag{6}$$

where $\hat{\mathcal{Y}}$ is a set of possible outputs. This inversion makes sense if $\hat{\mathcal{Y}}$ is a finite set with reasonably small size or $\hat{\mathcal{Y}}$ satisfies some simple and convex constraints, e.g. it is a hyperplane or a ball. We should also assume that the vectors in the inner product are normalised to the same length.

3 Experiment

We present an image-text retrieval experiment. The source of the data is an annotated set of images, 695 items, from the University of Washington.[1] To each image a set of words is attached expressing the most characteristic properties of the images, e.g. if the image shows a garden then the words are: flowers, trees, grass etc. The number of words describing the images varies over the images.

The information conveyed by the images was transformed by the following procedure. For each image the so called SIFT features [7] were computed and those features were classified by k-means clustering. Based on this a dictionary of “visual words” was built up. The feature vector of an image in the experiment contained a histogram of the visual words detected on the image, as in [4]. On the other hand, the textual annotation was preprocessed as well, e.g. words were changed from plural into singular. Then, the word indicators give the labels to the images. As a result, we had 132 textual words and 3000 visual words to characterise the images.

In the test we considered the image features (the visual words) as inputs, and the descriptor words were predicted. The accuracy was measured by the proportion of the correctly predicted words for all images. In the test a ten fold cross validation procedure was applied. To estimate the best parameters for the polynomial and the Gaussian kernels the training set was split into validation training and validation test sets. The parameters providing the highest result in the validation had been chosen and applied on the test set at the end.

Table 1 summarises the result. The convex variants are faster and generally significantly more accurate than the KCCA, furthermore they are much less sensitive on ill-conditioned kernel matrices, which may not be exactly positive definite because of numerical reason.

4 Conclusion

We give convex alternatives to the CCA method to the problems where both the inputs and outputs are vectors. These methods could be seen as regression approaches and they can be extended to solve problems arising in statistics, e.g. General or Generalized Linear Model type of problems. Our assumptions

[1] www.cs.washington.edu/research/imagedatabase/groundtruth

Method	Properties	Precision	Recall	F1	Comp. Time(s)
KCCA	Linear, 10 factor	44.9(1.4)	36.4(1.5)	40.2(1.3)	0.59
KCCA	Poly(2,0), 10 factor	43.2(2.5)	34.9(2.3)	38.6(2.3)	0.59
KCCA	Gauss(0.8), 10 factor	25.3(10.4)	20.2(8.2)	21.8(7.7)	0.58
MMR	Linear, unnorm.	24.3(1.8)	32.1(3.2)	27.6(2.2)	**0.03**
MMR	Linear, norm.	45.7(1.1)	37.9(2.1)	41.4(1.4)	0.39
MMR	Poly(5,0), norm.	**51.1**(1.8)	**50.1**(2.2)	**50.6**(1.9)	0.18
MMR	Gauss(0.25), norm.	50.0(2.6)	49.4(1.6)	49.7(1.4)	0.11
Random baseline		3.63			

Table 1: Precision, Recall and F1 measures provided by the different methods are in percentages and the computational times in seconds. () contains the corresponding Standard Deviation. Properties show the types of the kernels, the number of the factors considered in KCCA, and when the data vectors were or were not normalised to the length of one. The time data is received by matlab code on a Pentium machine of 3.5 GHz.

involve similarities measured in angles instead of distances. They move the regression type problems from Euclidian space into Projective geometry and the deep understanding of this movement is an important task for future research.

References

[1] T. De Bie, N. Cristianini, and R. Rosipal. Eigenproblems in pattern recognition. In E. Bayro-Corrochano, editor, *Handbook of Geometric Computing : Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics*, pages 129–170. Springer-Verlag, Heidelberg, 2005.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[3] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games.* Cambridge University Press, 2006.

[4] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *XRCE Research Reports, XEROX.* The 8th European Conference on Computer Vision - ECCV, Prague, 2004.

[5] David R. Hardoon and John Shawe-Taylor. Kcca for different level precision in content-based image retrieval. In *Submitted to Third International Workshop on Content-Based Multimedia Indexing*, IRISA, Rennes, France, 2003.

[6] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. In *Linear Algebra and its Applications, Special Issue on Linear Algebra in Control, Signals and Image Processing*, volume 284, pages 193–228. 1998.

[7] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision, Corfu*, 1999.

[8] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 2004.

[9] Y. Nesterov. Dual extrapolation and its application for solving variational inequalities and related problems. In *CORE Discussion Paper/68, September 2003.* 2003.