Regression Canonical Correlation Analysis

Jan Rupnik
Jan.Rupnik@ijs.si

Blaz Fortuna
Blaz.Fortuna@ijs.si

Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana

Abstract

In this paper we present Regression Canonical Correlation Analysis, an extension of Canonical Correlation Analysis, where one of the dimensions is fixed and demonstrate how it can be solved efficiently. We applied the extension to the task of query translation in the context of Cross-Lingual Information Retrieval.

1 Introduction

Canonical Correlation Analysis (CCA) technique is often used to aid Cross-Lingual Information Retrieval (CLIR) systems. It provides mappings from each of the languages into a "language independent" semantic space where documents and queries from various languages can be matched against each other. It also helps at extending query to other semantically similar words.

Training data for CCA consists of aligned corpus, a collection of documents from two or more languages where each document is aligned with its translation in other languages. CCA is an unsupervised learning method and focuses on the most frequent semantic dimensions in the corpus. This results in good performance on queries related to these dimensions. However, if an obscure query is used, which does not appear in one of the frequent semantic dimensions, it can result in a poor performance.

In this paper we propose a method, Regression Canonical Correlation Analysis (rCCA), derived from CCA, which assumes that the input in one language is already fixed and optimizes only over the remaining languages. This provides a way of supervising CCA with the actual query. We also demonstrate how it can be solved efficiently, making it usable in real-world CLIR systems and evaluate its performance on the data from CLEF competition.

2 Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) is a dimensionality reduction technique similar to Principal Component Analysis (PCA), with an additional assumption that the data consists of feature vectors that arose from two sources (two views) that share some information (a set of feature vectors computed from audio information and a set of feature vectors computed from the frames in a video recording). Instead of looking for linear combinations of features that maximize the variance (PCA) we look for a linear combination of feature vectors from the first view and a linear combination for the second view, that are maximally correlated.

Formally, let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the set of n sample points (pairs of observation vectors) where $x_i \in \boldsymbol{R}^p$ and $y_i \in \boldsymbol{R}^q$ represent feature vectors from p (or q)-dimensional vector spaces. Let $\boldsymbol{X} = [x_1, \ldots, x_n]$ and let $\boldsymbol{Y} = [x_1, \ldots, x_n]$ be the matrices with observation

vectors as columns, which are viewed as two samples of observations of two random vectors (X and Y). The idea is to find two linear functional (row vectors) $\alpha \in \boldsymbol{R}^p$ and $\beta \in \boldsymbol{R}^q$ so that the random variables αX and βY are maximally correlated (α and β map the random vectors to random variables, by computing weighted sums of vector components). By using the sample matrix notation $\boldsymbol{X}$ and $\boldsymbol{Y}$ this problem can be formulated as the following optimization problem:

$$\max_{\alpha \in R^p, \beta \in R^q} \alpha \boldsymbol{X}\ \boldsymbol{Y}'\ \beta'$$
$$s.t.$$
$$\alpha \boldsymbol{X}\ \boldsymbol{X}'\alpha' = 1$$
$$\beta \boldsymbol{Y}\ \boldsymbol{Y}'\beta' = 1$$

The optimization problem can be reduced to an eigenvalue problem and includes inverting the variance matrices $\boldsymbol{XX'}$ and $\boldsymbol{YY'}$. If they are not invertible one uses a regularization technique by replacing them with $(1-\kappa)\ \boldsymbol{XX'} + \kappa\ \boldsymbol{I}$, where $\kappa\ \epsilon\ \boldsymbol{R}$ and $\boldsymbol{I}$ is the identity matrix.

A single canonical variable is usually inadequate in representing the original random vector, that is why one looks for $k-1$ other projection pairs $(\alpha_2, \beta_2), \ldots, (\alpha_k, \beta_k)$, so that α_i, and β_i are highly correlated and each α_i is uncorrelated to α_j for $j \neq i$.

2.1 Information retrieval

Aggregating the projection vectors into matrices gives us two linear mappings, one for each view. One way of looking at them is that they map feature vectors into a space of mutual information. Given a query in the first view and a corpus of target documents in the second view, the task is to find which documents are relevant for that query. We can use the CCA projections to project the query and the target corpus into the same space and use cosine-similarity measure to compare the query and the corpus. Computing all the similarities enables us to rank the target corpus according to its similarity to the query and select a few of the documents with the highest rank as relevant. The projection vectors computed are pairs of general concepts in the two views that are strongly related (they have similar co-occurrence patterns). If the query is very specific it could be unrelated to the concepts discovered by CCA, which means that it cannot be represented well in the common space (loss of information).

3 Regression Canonical Correlation Analysis (rCCA) and IR

The main idea behind our proposed approach to customizing CCA for IR is to avoid looking for a large set of concept vectors that is general enough to represent the documents and rather solve an optimization problem for each query separately. The method requires a set of pairs of observations (training sets) and a source query vector in the first view. We then find the vector in the second view that is maximally correlated to the query across the training sets. The optimization problem that is solved is the CCA optimization problem with the first projection vector α fixed and set as the query. As a consequence the unit variance constraint for the query vector $q\ \epsilon\ \boldsymbol{R}^p$ (column vector) can be omitted since the variance of the query does not affect the solution of the problem.

$$\max_{\beta \in R^q} q'\ \boldsymbol{X}\ \boldsymbol{Y}'\ \beta$$
$$s.t.$$
$$\beta' \boldsymbol{Q} \beta = 1,$$

Where $\boldsymbol{Q} = \boldsymbol{YY'}$ *or* $(1-\kappa)\boldsymbol{YY'} + k\boldsymbol{I}$ if we use regularization. This problem can be reduced to a system of linear equations by writing the Lagrangian of the optimization problem (let λ denote the Lagrange multiplier) and setting its gradient to zero. This results in:

$$\boldsymbol{YX}'q + 2\lambda \boldsymbol{Q}\beta = 0.$$

Multiplying with β and using the variance constraint results in a second equation:

$$\beta' \boldsymbol{YX}'q + 2\lambda = 0.$$

Solving the first equation for β and substituting it in the second equation solves λ:

$$\lambda = -\frac{q'\boldsymbol{XY}'\beta}{2} = -\frac{q'\boldsymbol{XY}'}{2}\left(-\frac{1}{2\lambda}\boldsymbol{Q^{-1}YX}'q\right)$$

$$\lambda = \sqrt{\frac{q'\boldsymbol{XY}'\boldsymbol{Q^{-1}YX}'q}{4}}.$$

The eigenvalue problem is transformed to a system of linear equations. The solution involves computing $g := \boldsymbol{Q^{-1}(YX'}q)$ and setting the solution β to:

$$\beta = -\frac{g}{\sqrt{q'\boldsymbol{XY}'g}}.$$

If computing the inverse of $\boldsymbol{YY'}$ or $(1-\kappa)\boldsymbol{YY'} + k\boldsymbol{I}$ is not feasible, we can use the fact that we need to compute the inverse evaluated on a single vector only, which means that we can use iterative methods like the conjugate gradient method. The method only involves matrix-vector computations which means that we can exploit special structural properties of the matrix $\boldsymbol{Y}$ (such as sparsity) to speed up computations.

This section assumed that the data was centered. Centering the matrices before optimization is one way, but sometimes that is not desirable (centering can turn a sparse matrix into a full matrix). It is possible to implement centering on the fly with almost no extra computational cost.

4 Experiments: CLEF domain specific information retrieval task

The comparison between CCA and rCCA for IR was conducted on a text IR task, where we used the data from the CLEF IR competition (Domain-specific Track). The two methods were evaluated on the German and English parts of the data set. As part of the task we were given bilingual parallel resources for German and English, pseudo-aligned corpus GIRT for German and English of approximately 150,000 documents and the Cambridge Scientific Abstracts (CSA) corpus for English. We were also provided a set of 100 queries for each language, where 75 of them were used in tracks before 2007 and 25 were used in 2007. Note that the CSA corpus was used only in the 2007 track, this is why we give separate results for previous years and 2007.

The documents in each corpus were transformed into feature vectors using the bag-of-words vector space model. English vector space was roughly 70,000 dimensional and the German language resulted in 250,000 dimensions. The higher dimensionality of the later vector space stems from the fact that phrases are commonly concatenated into single words in German language.

Table 1: Comparison of CCA and rCCA (Mean Average Precision)

QUERY SET	CORPUS	CCA (GIRT)	RCCA (GIRT)
German 2006	English GIRT	20.53	26.49
English 2006	German GIRT	21.23	28.92
German 2007	English GIRT, CSA	17.84	27.25
English 2007	German GIRT	24.72	31.78

For each of the query we were provided with a set of relevant documents. Experiment was conducted as follows. We selected a random subset of 30,000 training documents and computed a 2000 dimensional semantic space for standard CCA. We then mapped the source queries and the target corpus into their common 2000-dimensional semantic space, and computed rankings of target documents for each query. In rCCA we mapped the queries directly to the space of the target corpus and computed the rankings. We could compare the rankings of both methods by using the lists of relevant documents for each query. The final

mean average precision (see Table 1) was obtained by averaging mean average precisions over queries.

5 Conclusions

We demonstrated that rCCA, the proposed extension to the CCA, is useful and leads to better performance in the context of CLIR. We also demonstrated that it can be solved efficiently which enables its application in real-world CLIR systems.

Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under NeOn (IST-4-027595-IP), SMART (IST-033917) and PASCAL2 (IST-NoE-216886).

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System*. New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hiippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.