Kernel Canonical Correlation Analysis With Applications

Blaž Fortuna
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: blaz.fortuna@ijs.si

Abstract

This paper provides an overview of Kernel Canonical Correlation Analysis. KCCA is a technique for finding common semantic features between different views of data. Applications on text retrieval, categorization and image retrieval based on text queries are presented.

1 Introduction

Canonical Correlation Analysis (CCA) is a method of correlating two multidimensional variables. It makes use of two different views of the same semantic object (eg. the same text document written in two different languages) to extract representation of the semantic. Input to CCA is a paired dataset $S = \{(u_i, v_i); u_i \in U, v_i \in V\}$, where U and V are two different views on the data – each pair contains two views of the same document. The goal of CCA is to find the common semantic space W and the mappings from each U and V into W space. All documents from U and V can be mapped into W to obtain a view independent representation.

Example Let space V be vectors-space model for English and U vector-space model for French text documents. Paired dataset is than a set with pairs made of English documents, together with their French translation. The output of CCA on this dataset is a semantic space where each dimension shares similar English and French meaning. By mapping English or French documents into this space, language unexpanded representations are obtained. In this way standard machine learning algorithms can be used on multi-lingual datasets.

2 Theoretical Foundations

Canonical Correlation Analysis ([1], [2]) can be seen as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximized. Canonical Correlation Analysis seeks a pair of linear transformations, one for each of the sets of variables, such that, when the set of variables are transformed, the corresponding co-ordinates are maximally correlated.

Let $S = \{(u_i, v_i); u_i \in U, v_i \in V\}$ be a paired dataset. By using the CCA, we can find directions $f_u \in U$ and $f_v \in V$ in the two spaces so that the projections $\{(f_u; u_i)\}_{i=1}^N$ and $\{(f_v; v_i)\}_{i=1}^N$ of the feature vectors of documents from the two views would be maximally correlated. Formally, the CCA is to maximize canonical correlation ρ in space $U \times V$ which is defined as

$$\rho = \max_{(f_u, f_v) \in U \times V} \frac{\sum_{i=1}^N < f_u, u_i >< f_v, v_i >}{\sqrt{\sum_i < f_u, u_i >^2 \sum_i < f_v, v_i >^2}}$$

In an attempt to increase the flexibility of the feature selection, kernelisation of CCA (KCCA) can be applied to map the hypothesis to a higher-dimensional feature space. There we search for f_u and f_v in the space spanned by the corresponding feature vectors, i.e. $f_u = \sum_l \alpha_l u_l$ and $f_v = \sum_m \alpha_m v_m$. The upper equation can be rewritten as

$$\sum_i < f_u, u_i >< f_v, v_i >=$$

$$\sum_i \sum_{lm} \alpha_l \beta_m < u_l, u_i >< v_m, v_i >=$$

$$\alpha^T K_u K_v \beta,$$

where $\alpha = (\alpha_1, \ldots, \alpha_N)$, $\beta = (\beta_1, \ldots, \beta_N)$ and K_u and K_v are Gram matrixes of $\{u_i\}_{i=1}^N$ and $\{v_i\}_{i=1}^N$. In order to force non-trivial learning on the correlation, we introduce a regularization parameter to penalize the norms of the associated weights. The problem becomes

$$\rho = \max_{\alpha,\beta} \frac{\alpha^T K_u K_v \beta}{\sqrt{(\alpha^T K_u^2 \alpha + \tau \alpha^T \alpha)(\beta^T K_v^2 \beta + \tau \beta^T \beta)}}.$$

Because regularized equation is not affected by rescaling of α or β, optimization problem is subject to the two constraints

$$\alpha^T K_u^2 \alpha + \tau \alpha^T \alpha = 1,$$

$$\beta^T K_v^2 \beta + \tau \beta^T \beta = 1.$$

By using corresponding Lagrangian and *Kuhn-Tucker* conditions we can rewrite the upper optimization problem as a generalized eigenvalue problem

$$\begin{pmatrix} 0 & K_u K_v \\ K_v K_u & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} =$$

$$\lambda \begin{pmatrix} K_u^2 + \tau I & 0 \\ 0 & K_v^2 + \tau I \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Note that the size of generalized eigen problem is $2N$,where N is the size of the paired dataset. This can be reduced by using *incomplete Cholesky decomposition* to N or less when seeking only approximate solution. In the upper derivation we assumed that we have two different views of documents. (U and V). CCA can be generalized to more views, but than the trick to reduce the size of eigen problem can not be used.

3 Applications of KCCA

3.1 Labels

A similar problem to CCA is to select features of highest correlation between documents and their labels. The method for finding these features is called Partial Least Squares (PLS) [1]. PLS could also be thought as a method which looks for directions that are good at distinguishing the different labels. Similarity between this problem and CCA can be noticed when viewing labels as another "different view of documents".

3.2 Cross-Language Text Mining

With KCCA we can construct a semantic space into which text documents, written in different languages, can be mapped to obtain language independent representation. This highly reduces the complexity of dealing with different languages since we can apply standard machine learning algorithms to the data mapped into the semantic space. Another method for dealing with multi-lingual datasets is CL-LSI [4].

Text document retrieval The semantic space for languages can be used at searching databases with documents in different languages. First, all documents from the database are mapped into the semantic space. Than, queries can be viewed as documents and can be mapped into the semantic space. The result of a query is a set of documents from the database that are the closest to the mapped query in the semantic space. The advantage of this approach is that the results are independent of the language in which the query was issued.

This approach was shown and tested in [3] on 'house debates' part of 36^{th} Canadian Parliament proceedings corpus. Text chunks were split into paragraphs and paragraphs were treated as separate documents. Part of this dataset was used for generating the semantic space with KCCA and the rest of the documents were used for testing. Short queries were generated from the five most probable words from each test document. The relevant documents were the test documents themselves in monolinguistic retrieval (English query - English document, table 1) and their mates in cross-linguistic (English query - French document, table 2) test. Each test was done for different dimensions d of the generated semantic space.

Text categorization Another application of the semantic space is categorization of multi-lingual

d	100	200	300	400	full
cl-lsi	53	60	64	66	70
cl-kcca	60	63	70	71	73
cl-lsi	82	86	88	89	91
cl-kcca	90	93	94	95	95

Table 1: English → English top-ranked (left) and top-ten (right) retrieval accuracy

d	100	200	300	400	full
cl-lsi	30	38	42	45	49
cl-kcca	68	75	78	79	81
cl-lsi	67	75	79	81	84
cl-kcca	94	96	97	98	98

Table 2: English → French top-ranked (left) and top-ten (right) retrieval accuracy

documents. First, the semantic space is generated from the paired dataset with KCCA. Than, the labled training set for categorization is mapped into the semantic space. Note that these labled documents do not need to be paired anymore. Even more, they can even come from only one language. Once training set is mapped into semantic space standard classification algorithms can be used, eg. SVM. Another way of using SVM is to learn classifier on labled documents from one language and than transfer it trough semantic space into other language's vector-space model.

This approach was shown and tested in [5] on NTCIR-3 patent retrieval test collection, with paired documents in English and Japanese. The classifier was learned on documents in one language and was used to classify documents in another language. The training set for Topic 01 had 827 annotated documents with 26 relevant, Topic 07 had 366 annotated documents with 102 relevant documents. The classifier was trained on English training set. Results are in table 3.

d	50	100	150	full
Eng-tr	78.1	97.7	99.2	100.0
Eng-ts	36.0	41.0	44.4	46.9
Jp-tr	79.4	92.5	98.4	99.2
Jp-ts	41.1	42.4	48.9	49.1
Eng-tr	87.6	93.9	95.8	97.1
Eng-ts	85.1	87.4	87.0	87.9
Jp-tr	87.4	92.9	95.4	96.8
Jp-ts	77.2	77.7	77.3	78.4

Table 3: Average precision [%]: the classifier learned on English training set was used on English training and test sets and on Japanese training and test sets. On left are results for Topic 01 and on right for Topic 07.

3.3 Machine Translation and KCCA

The goal of KCCA is to generate language independent semantic space. But, in order to use KCCA, paired dataset is needed. This can be tricked by using machine translation tools, for example *Google Language Tools* [1], to artificially generate paired dataset from monolinguistic dataset. Semantic space obtained from this kind of paired dataset can than be used for text as described upper.

So far, this approach was only tested on documents written in the same language as original documents used for generating paired dataset for KCCA. Generated semantic space was compared to normal vector-space model (BOW + TFIDF) and to Latent Semantic Indexing (LSI). Documents from Reuters-24578 dataset were used with Mod-Apte split. First 400 documents were translated into German using Google. Than this documents were used for generating semantic space with KCCA and with LSI (only English copies are necessary for LSI). Learning was done on very small sets. In one experiment only 5 documents were relevant out of 25 and, in other, 10 documents were relevant out of 50. Documents for learning were randomly chosen and averaged over 10 runs. The classifier was than tested on whole testing set of Mod-Apte split (around 3000 documents). Results are in table 4 and 5.

3.4 Image-Text Retrieval

The goal here is the retrieval of images based on a text query, but without any labeling associated with the image. The database used for generating the semantic space contains images retrieved from

[1] `http://www.google.com/language_tools`

Cat	BOW	KCCA	LSI
Earn	97.2	96.9	96.7
Acq	75.4	90.2	84.4
Corn	55.3	27.0	49.2
Grain	69.3	65.1	69.2
Trade	47.6	43.2	36.8
Earn	98.2	95.7	97.6
Acq	74.4	90.2	84.4
Corn	60.5	27.9	55.6
Grain	76.7	66.1	76.0
Trade	67.3	59.7	57.0

Table 4: Average precision [%] for classifier trained on 25 documents (top) and on 50 documents (bottom).

Cat	BOW	KCCA	LSI
Earn	92.9	93.5	91.9
Acq	72.0	84.7	78.5
Corn	51.3	32.7	45.4
Grain	62.0	59.5	61.3
Trade	46.2	42.0	36.2
Earn	94.0	93.6	92.9
Acq	79.4	86.6	82.1
Corn	55.2	32.3	53.2
Grain	69.4	59.9	66.8
Trade	62.3	56.5	53.5

Table 5: Break Even Point [%] for classifier trained on 25 documents (top) and on 50 documents (bottom).

the Internet with attached text – dataset for KCCA contains pairs of image and attached text.

This approach was shown and tested in [2]. Images in database were split into three classes. For each query a set of 10 or 30 images was chosen that best match the query text. Success is considered if the images are of the same label as query text (first part of Table 6). At second test successful match was considered if the image that actually matched with the chosen text is contained in the set (second part of Table 6).

Image Set	KCCA (30 dim)	KCCA (5 dim)
10	85 %	91 %
30	83 %	91 %
10	17 %	60 %
30	32 %	69 %

Table 6: Results for Image-Text Retrieval

References

[1] J. Shawe-Taylor, N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004

[2] D. R. Hardon, S. Szedmark, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. Technical Report CSD-TR-03-02, Department of Computer Science, Royal Holloway, University of London, 2003.

[3] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances of Neural Information Processing Systems 15*, 2002.

[4] M. L. Littman, S. T. Dumais, and T. K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette, editor, *Cross language information retrieval.* Kluwer, 1998.

[5] Yaoyong Li and John Shawe-Taylor. Using KCCA for Japanese-English cross-language information retrieval and classification