Kernelized Sorting

Authors

Novi Quadrianto[1] | Le Song[2] | Alex J. Smola[3]

1: RSISE, ANU & SML, NICTA | 2: SCS, CMU | 3: Yahoo! Research

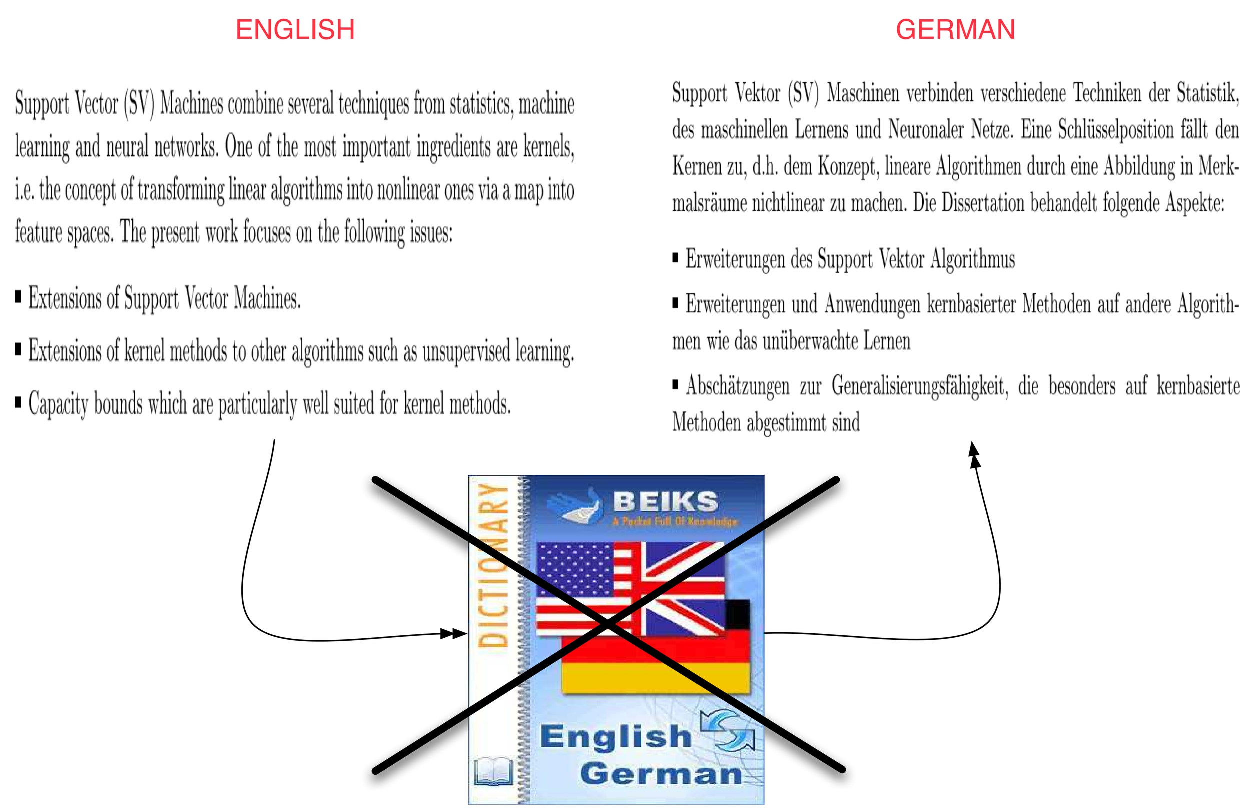
Abstract

Matching pairs of objects from different domains is a fundamental operation in data analysis. It typically requires the definition of a similarity measure between the classes of objects to be matched. For many cases, we may be able to design a cross domain similarity measure based on prior knowledge or to observe one based on the co-occurence of such objects. In some cases, however, such a measure may not exist or it may not be given to us beforehand.

We develop an approach which is able to perform matching by requiring a similarity measure only within each of the classes. This is achieved by maximizing the dependency between matched pairs of observations by means of the Hilbert-Schmidt Independence Criterion. This problem can be cast as one of maximizing a quadratic assignment problem with special structure and we present a simple algorithm for finding a locally optimal solution.

Problem Statement

Assume we are given two collections of documents purportedly covering the same content, written in two different languages. Can we determine the correspondence between these two sets of documents without using a dictionary?



(Formal) problem formulation:

Given

- two sets of observations $X = \{x_1, \dots, x_m\}$ and $Y = \{y_1, \dots, y_m\}$

Find

- a permutation matrix $\pi \in \Pi_m$,

$$\Pi_m := \left\{\pi | \pi \in \{0,1\}^{m\times m} \text{ where } \pi 1_m = 1_m, \pi^\top 1_m = 1_m\right\}$$

such that $\left\{(x_i, y_{\pi(i)}) \text{ for } 1 \le i \le m\right\}$ is maximally dependent.

Hilbert-Schmidt Independence Criterion

- HSIC is the square of the Hilbert-Schmidt norm of the cross covariance operator

$$\mathcal{C}_{xy} = \mathbf{E}_{xy}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)],$$
$$\text{where } \mu_x = \mathbf{E}[\phi(x)], \mu_y = \mathbf{E}[\psi(y)].$$

- In term of kernels, HSIC can be expressed as

$$\begin{aligned}\text{HSIC} = \|\mathcal{C}_{xy}\|^2_{\text{HS}} = &\mathbf{E}_{xx'yy'}[k(x,x')l(y,y')] + \\ &\mathbf{E}_{xx'}[k(x,x')]\mathbf{E}_{yy'}[l(y,y')] - \\ &2\mathbf{E}_{xy}[\mathbf{E}_{x'}[k(x,x')]\mathbf{E}_{y'}[l(y,y')]].\end{aligned}$$

- A biased estimator of HSIC given finite sample $Z = \{(x_i, y_i)\}_{i=1}^m$ drawn from $\Pr_{xy}$ is

$$\widehat{\text{HSIC}} = (m-1)^{-2} \operatorname{tr} HKHL = (m-1)^{-2} \operatorname{tr} \bar{K}\bar{L},$$
$$\text{where } K \in \mathbb{R}^{m\times m}, K_{ij} = k(x_i, x_j)$$
$$\bar{K} := HKH.$$

Advantages of HSIC are

- Computing HSIC is simple: only the kernel matrices K and L are needed;
- HSIC satisfies concentration of measure conditions, i.e. for random draws of observation from $\Pr_{xy}$, HSIC provides values which are very similar;
- Incorporating prior knowledge into the dependence estimation can be done via kernels.

Kernelized Sorting

Optimization problem

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi_m} \left[\operatorname{tr} K\pi^\top L\pi\right]$$

Sorting as a special case

For scalar x_i and y_i and a linear kernel on both sets, we can rewrite the optimization problem

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi_m} \left[X^\top \pi Y\right]^2$$

This is maximized by sorting X and Y.

Optimization

Convex Objective and Convex Domain

- Define π as a doubly stochastic matrix,

$$P_m := \left\{ \begin{array}{l} \pi \in \mathbb{R}^{m\times m} \text{ where } \pi_{ij} \ge 0 \text{ and} \\ \sum_i \pi_{ij} = 1 \text{ and } \sum_j \pi_{ij} = 1 \end{array} \right\}$$

- The objective function $\operatorname{tr} K\pi^\top L\pi$ is convex in π.

Convex-Concave Procedure

- Compute successive linear lower bounds and maximize

$$\pi_{i+1} \leftarrow \operatorname{argmax}_{\pi \in P_m} \left[\operatorname{tr} K\pi^\top L\pi_i\right]$$

This will converge to a local maximum.

- Initialization is done via sorted principal eigenvector.

Related Work

Instead of HSIC, we can use Mutual Information to measure the dependence between random variables x_i and $y_{\pi(i)}$. MI is defined as, $I(X,Y) = h(X) + h(Y) - h(X,Y)$. We can approximate MI maximization by maximizing its lower bound. This then corresponds to minimizing an upper bound on the joint entropy $h(X,Y)$.

Optimization problem

$$\pi^* = \operatorname{argmin}_{\pi \in \Pi_m} \left[\log |HJ(\pi)H|\right],$$
$$\text{where } J_{ij} = K_{ij} L_{\pi(i),\pi(j)}.$$

This is related to the optimization criterion proposed by Jebara (2004) in the context of aligning bags of observations by sorting via minimum volume PCA.

Applications

Image Layout

Layout of 284 images into a 'NIPS 2008' letter grid using Kernelized Sorting. Gaussian RBF kernel is used for the image objects and also for the positions of the grid. One can see that images are laid out in the letter grid according to their color grading.

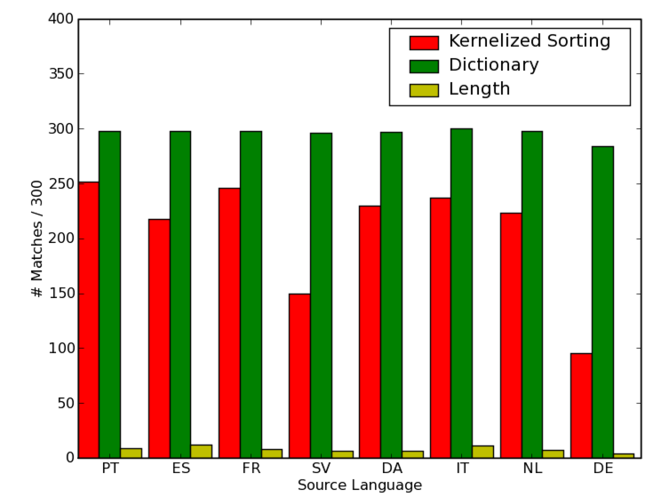
Layout of 320 images into a 16×20 grid using Kernelized Sorting.

Layout of 320 images into a 16×20 grid using Generative Topographic Mapping. A compressed representation of images is shown. GTM does not guarantee unique assignments of images to nodes.

Image Matching

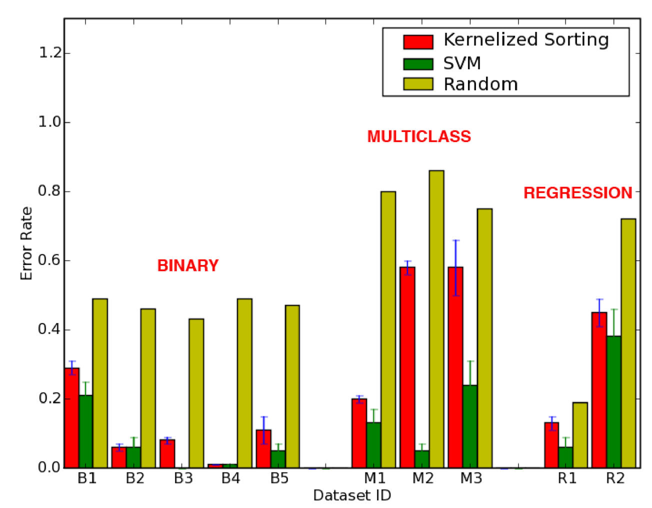
Image matching as obtained by Kernelized Sorting. The images are cut vertically into two equal halves and Kernelized Sorting is used to pair up image halves that originate form the same images. 140 pairs out of 320 are correctly matched. This is quite respectable given that chance level would be 1 correct pair.

Multilingual Document Matching



- Matching non-English documents (source languages) to its English translation (target language) of the Europarl Parallel Corpus.
- Standard TF-IDF features of a-bag-of-words kernel is used for Kernelized Sorting.
- Result comparisons with length based and dictionary based document matching.

Data Attribute Matching



- Matching data attributes (or dimensions) of binary, multiclass, and regression data sets from UCI repository and LibSVM site.
- Homogenous misclassification loss and squared loss are used to measure performance.
- Gaussian RBF kernel is used for Kernelized Sorting.
- Results comparisons with random permutations and SVM classification/regression on original unsplitted data sets.

Summary

- We generalize sorting by maximizing dependency between matched pairs of observations via HSIC.
- Applications of our proposed sorting algorithm range from data visualization to image, data attribute and multilingual document matching.