Conformal Multi-Instance Kernels

Matthew B. Blaschko
Max Planck Institute for Biological Cybernetics
Spemannstr. 38
72076 Tübingen, Germany
matthew.blaschko@tuebingen.mpg.de

Thomas Hofmann
Google
Freigutstr. 12
8002 Zürich, Switzerland
thofmann@google.com

Abstract

In the multiple instance learning setting, each observation is a bag of feature vectors of which one or more vectors indicates membership in a class. The primary task is to identify if any vectors in the bag indicate class membership while ignoring vectors that do not. We describe here a kernel-based technique that defines a parametric family of kernels via *conformal transformations* and jointly learns a discriminant function over bags together with the optimal parameter settings of the kernel. Learning a conformal transformation effectively amounts to weighting regions in the feature space according to their contribution to classification accuracy; regions that are discriminative will be weighted higher than regions that are not. This allows the classifier to focus on regions contributing to classification accuracy while ignoring regions that correspond to vectors found both in positive and in negative bags. We show how parameters of this transformation can be learned for support vector machines by posing the problem as a multiple kernel learning problem. The resulting multiple instance classifier gives competitive accuracy for several multi-instance benchmark datasets from different domains.

1 Introduction

Multiple-instance learning (MIL) as introduced in [8, 5] is a generalization of supervised classification in which class labels are associated with sets of feature vectors, called *bags*, instead of individual patterns. However, compared to generic classification problems involving sets of instances, the MIL setting has more specific semantics, namely that a bag belonging to a class (i.e. a positive example in binary classification) contains class-characteristic feature vectors that are not contained in negative bags. It is the presence or absence of such feature vectors that determines the classification outcome for the bag as a whole.

This setting has numerous interesting applications, ranging from drug design [5] to text categorization [2] and tasks in computer vision [11, 13] or image retrieval [11, 15]. In these applications, the feature vectors in a bag may correspond to different aspects, alternate data representations, or local patches, parts, or fragments. For instance, in visual object recognition, each feature vector may encode local image properties, some of which might come from the actual object of interest (e.g. a face in the context of face recognition), but some of which might come from an arbitrary background.

Learning a classifier from multiple instance data involves two aspects: (i) dealing with the intrinsic variability of the characteristic and non-characteristic feature vectors and (ii) identifying, either explicitly or implicitly, the truly characteristic feature vectors in (positive) bags. The first problem corresponds to standard supervised classification, while the second problem deals with ambiguity that is specific to the MIL scenario.

Because of the intrinsic variablility of the feature vectors, nearby vectors can be assumed to be similarly characteristic. We therefore wish to modulate the contribution of an individual vector to a classification decision based on where in the feature space it is located. Conformal transformations

directly address the ambiguity introduced by the MIL setting by weighting individual feature vectors by a discriminatively learned function over the feature space. Regions of the feature space that are indicative of class membership are weighted highly, while uninformative regions will have a weight close to zero. We therefore use conformal transformations of a base kernel to define generalized set kernels; such transformations have been investigated by [14] in a different context. In some sense, conformal kernels allow us to recast the diverse density [10] idea in the kernel setting: the geometry of the input space is locally magnified or shrunken based on the discriminative power of feature vectors in the region. As shown by [14] this corresponds in the metric case to a local adaptation of the (Riemannian) metric tensor in a way that locally preserves angles.

2 Multi-instance Kernels

The standard set kernel [7] over sets $p = \{x_1, \ldots, x_N\}$ and $p' = \{x'_1, \ldots, x'_{N'}\}$ of patterns is defined as follows,

$$k(p, p') = \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \kappa(x_i, x'_j)\,. \tag{1}$$

where κ is some pattern-level base kernel. It is straightforward to verify that this is equivalent to the inner product between the centroids of the sets of vectors in the feature space implicitly defined by κ.

In the multi-instance kernel of [6] the set kernel is modified to better address the multi-instance learning problem by exponentiating $\kappa(x_i, x'_j)$ by a power $q \geq 1$. Theoretically, this ensures separability of any training set of bags as $q \to \infty$, however at the cost of a diagonally-dominant kernel matrix with very poor generalization capabilities. In the special case of Gaussian kernels for which $0 \leq \kappa(x, x') \leq 1$, the overall summation is dominated by the pairs of vectors $(x, x') \in p \times p'$ with the smallest angle between them, while the influence of feature vectors that do not closely match vanishes. This has to be expected to perform well on a multiple instance problem only under the assumption that vectors indicating class membership are *tightly clustered*, while non-indicative vectors have a low probability of being close together (and consequently will typically have small values $\kappa(x_i, x'_j)$). This obviously does not address the issues of non-discriminative features as each vector in the bag is weighted equally. If only a small fraction of the vectors indicates class membership, the other vectors will overwhelm the result. Multi-instance kernels hence do not actually solve the multi-instance learning problem. Rather, they solve the problem of defining a kernel between sets of vectors that makes *no distinction* between vectors that are indicative of class membership and those that are not.

3 Conformal Multi-Instance Kernels

3.1 Conformal Kernel Transformations

A conformal kernel modifies a kernel function in a way that preserves angles between vectors in the mapped space [1, 14] and that can be used to locally stretch or magnify the feature space. In particular, a conformal kernel is one that has the form

$$k_\theta(x, x') = c_\theta(x) c_\theta(x') \kappa(x, x') \tag{2}$$

where $\kappa(x, x')$ is a base kernel between patterns, e.g. a Gaussian kernel, and $c_\theta > 0$. k_θ is a valid kernel with feature map $\Phi_\theta(x) = c_\theta(x)\Phi(x)$, with $\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$, as can be easily checked.

Although angles are preserved, the conformally transformed mapping, Φ_θ expands regions of the space where c_θ is large, and contracts regions of the embedding space where it is small. To see this note that for the input space $\mathbb{R}^n$ the $n \times n$ positive-definite matrix $\mathbf{G}(x) = (g_{ij}(x))_{i,j}$, with

$$g_{ij}(x) \equiv \left(\frac{\partial}{\partial x_i} \Phi(x) \right) \cdot \left(\frac{\partial}{\partial x_j} \Phi(x) \right) \tag{3}$$

is the Riemannian metric tensor induced on the manifold in the embedding space. The volume form in a Riemannian space is defined as

$$dV = \sqrt{g(x)} dx_1 \ldots dx_n \tag{4}$$

where $g(x) = \det \mathbf{G}(x)$. The factor $\sqrt{g(x)}$ represents how a local area is magnified in the embedding space under the mapping Φ. In general, the conformal transformation of equation 2 yields

$$\tilde{g}_{ij}(x) = \frac{\partial c_\theta(x)}{\partial x_i}\frac{\partial c_\theta(x)}{\partial x_j} + c_\theta(x)^2 g_{ij}(x) \tag{5}$$

The first term can be made small by specifying relatively slow varying conformal functions. Focusing on the second term, we see that large values of $c(x)$ serve to increase the volume element of the embedding space while small values will decrease the volume element. In a maximum margin classifier, this means that regions of the space corresponding to large values of $c_\theta(x)$ will be given greater emphasis than those with small values.

3.2 Conformal Multi-instance Kernels

We propose that a conformal multi-instance kernel be defined as a modification of a standard set kernel in which the base kernel between individual patterns is modified conformally.

$$k(p, p') = \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} c_\theta(x_i) c_\theta(x'_j) \kappa(x_i, x'_j) \tag{6}$$

The conformal transformation $c_\theta > 0$ allows altering the geometry of the feature space in a way that puts more emphasis on *relevant* input space regions.

Conformal multi-instance kernels, therefore, are set kernels over conformally modified base-kernels, where the conformal transformation is parametrized by θ, which is estimated from labeled training data as part of the learning process.

In this paper, we specify $c_\theta(x)$ to have the specific form

$$c_\theta(x) = \sum_{r=1}^{q} \theta_r \tilde{\kappa}(x, \mu_r) \tag{7}$$

that is, for a given set of expansion points μ_r in the input space, we assume the positive function c_θ can be written as a radial basis function network with kernel $\tilde{\kappa}$ (one may or may not set $\tilde{\kappa} = \kappa$). In the experiments performed here, we have set $\tilde{\kappa}$ to be a Gaussian. Intuitively, a large value of θ_r indicates that the neighborhood of μ_r is a discriminative region of the feature space. In the next section, we discuss two techniques for learning the parameters, θ, for a support vector machine.

In order to find suitable expansion points μ_r, we perform a highly-scalable unsupervised clustering stage prior to discriminative learning. Specifically, we have followed the buckshot heuristic proposed by [4] which finds R clusters from n data vectors, based on a randomly generated sample of size $\sqrt{Rn}$ using k-means. We have found the solution not to be too sensitive to the random sample and initialization of k-means and suggest optimizing R via cross-validation.

4 Learning Algorithms

We propose to simultaneously optimize a support vector machine classifier while optimizing the parameters θ in Eq. (7). To do this, we make use of generalization bounds to define functions to minimize with respect to SVM parameters α and θ. Two bounds that have been used in this context previously are the well-studied radius-margin bound [3] and the trace-margin bound [9].

4.1 Optimizing the Trace-Margin Bound

Lanckriet et al. make use of the trace-margin bound and semi-definite programming to simultaneously learn the kernel parameters and optimize the margin. As this is one convex optimization procedure, this has advantages both in speed and optimality. When $\mathbf{K}_\theta$ can be defined as a positively weighted linear combination of kernel matrices

$$\mathbf{K}_\theta = \sum_{l=1}^{q} \theta_l \mathbf{K}_l, \quad \theta > 0\,, \tag{8}$$

the optimization can be formulated as a quadratically constrained quadratic program (QCQP, c.f. [9], Theorem 17). More recently, multiple kernel learning has been formulated as a semi-infinite linear program which is generally faster than the QCQP approach [12].

If Eq. (6) can be written as a linear combination of positive definite functions, we can optimize θ using multiple-kernel learning. As we expand Eq. (6)

$$k_\theta(p, p') = \frac{1}{NN'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \left(\sum_{l=1}^{q} \theta_l \tilde{\kappa}(x_i, \mu_l) \right) \left(\sum_{m=1}^{q} \theta_m \tilde{\kappa}(x'_j, \mu_m) \right) \kappa(x_i, x'_j) \tag{9}$$

we see that this function is quadratic in parameters θ. However, we notice that for localized kernels such as Gaussian kernels we often will have a situation where

$$\tilde{\kappa}(x_i, \mu_l)\tilde{\kappa}(x'_j, \mu_m)\kappa(x_i, x'_j) \approx 0 \tag{10}$$

for $l \neq m$. When $\|x_i - x'_j\|$ is small, $\kappa(x_i, x'_j)$ will be large, but at least one of $\tilde{\kappa}(x_i, \mu_k)$ or $\tilde{\kappa}(x'_j, \mu_l)$ will be close to zero. When $\|x_i - x'_j\|$ is large, $\kappa(x_i, x'_j)$ will be close to zero and the product will be close to zero. Therefore we may attempt to simplify the conformal MI-kernel as follows

$$\begin{aligned} k(p, p') &\approx \frac{1}{NN'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \left(\sum_{l=1}^{q} \theta_l^2 \tilde{\kappa}(x_i, \mu_l)\tilde{\kappa}(x'_j, \mu_l) \right) \kappa(x_i, x'_j) && (11) \\ &= \sum_{l=1}^{q} \theta_l^2 \left(\frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \tilde{\kappa}(x_i, \mu_l)\tilde{\kappa}(x'_j, \mu_l)\kappa(x_i, x'_j) \right) && (12) \end{aligned}$$

which is linear in the (transformed) parameters $\rho_l \equiv \theta_l^2$. As we will show in the experimental section, this simplified kernel is still able to capture the MI-kernel idea, while leading to learning algorithms with much improved scalability.

5 Experimental Results

In our experiments, we implemented the linear combination of kernels in equation 12 as well as an additional matrix computed using the standard set kernel. This gives a guarantee that the trace-margin bound will be no worse than a standard set kernel alone. We used cross-validation to select the bandwidth associated with the conformal transformation, and the bandwidth for the base Gaussian kernel between patterns.

We begin with an illustrative toy example in which the data lie in a one-dimensional space, figure 1. The positive and negative bags are generated by sampling five points from mixture of Gaussians distributions, each with two equally-weighted centers at $\{5, 0\}$ and $\{0, -5\}$ respectively. As expected, the clustering step closely approximates these centers. Interestingly, the multiple-kernel learning step places nearly all the mass at the RBF located at 5. This is due to that a few instances from positive bags drawn from the zero mean Gaussian have relatively large negative values. The data are therefore more separable as viewed through the "lens" of the RBF centered at 5 than the RBF centered at -5. Due to the one-norm regularization properties of multiple kernel learning [12], the more informative kernel is accorded the majority of the weight. In repeated experiments, the RBF centered at 0 was given a weight of 0, as was the standard set kernel. Qualitatively similar results were obtained when optimizing the radius-margin bound, indicating that the multiple kernel learning approach to optimization is a reasonable approximation in this case.

The empirical results reported here are from three main domains: pharmaceuticals, object recognition in computer vision, and text categorization. Specifically, we report results for the benchmark MUSK data sets, detection of specific kinds of animals in images, and categorization of TREC document data sets. Details of the multi-instance learning test suite can be found in [2]. Table 1 shows comparative results for our technique and the best result previously published in [2]. In every case the kernel was able to achieve results that were comparable to the previous best results. Note that due to space restrictions, the comparative results are taken from the best of several variants of their algorithm, and the best results for the MUSK datasets are reported for the IAPR algorithm, which is highly optimized to the MUSK problem. We refer the interested reader to [2] for further details and

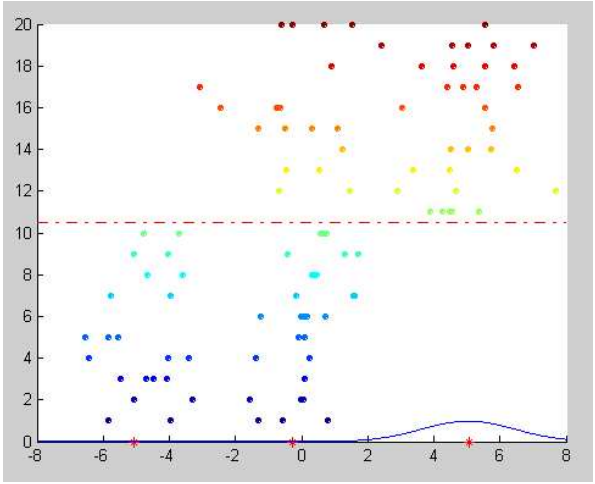

Figure 1: A toy example. The horizontal axis represents the feature values, while the vertical axis represents which bag a feature belongs to. The red asterisks are the RBF centers selected using k-means clustering. Bags below the dotted horizontal line belong to class -1, while bags above the dotted line belong to class 1. The learned conformal function is plotted in blue at the bottom.

Table 1: Classification accuracy on benchmark datasets

	MUSK 1	**MUSK 2**	**Elephant**	**Fox**	**Tiger**
Conformal Kernels	90.22	86.96	83.5	61.5	84.5
Multi-instance SVM	92.4 (IAPR)	89.2 (IAPR)	82.2	59.4	84
EM Discriminative Density	84.8	84.9	78.3	56.1	72.1

	TREC 1	**TREC 2**	**TREC 3**
Conformal Kernels	94	76.25	86
Multi-instance SVM	93.9	84.5	87
EM Discriminative Density	85.8	84.0	69

results. It should also be noted that we did not re-implement the techniques reported in [2] so we are not able to report statistical significance. These initial results suggest that this approach may be well suited to computer vision, where multiple regions of the input feature space need to be considered for classification.

In the results reported here, the number of cluster centers was fixed at 300. The approximate range of bandwidths for the RBF network was selected based on the singular values of the clusters obtained during the buckshot clustering step. The overlap of the support between RBFs was relatively low, and the approximation in equation 12 holds well in the experiments performed here. The result of cross-validation was typically that the bandwidth associated with the cluster centers was two to three times as large as the bandwidth for the base kernel. Set kernels using the same bandwidths as the conformal multi-instance kernel performed at approximately chance.

6 Conclusions and Future Work

We have presented here a novel form of kernel for solving the multi-instance learning problem that performs competitively with previous results. Previous maximum margin approaches have either treated every pattern within a bag equally (e.q. [6]) or have modified the SVM formulation (e.g. [2])), while our approach models the semantics of the multi-instance problem by a specific choice of a family of kernels.

It is interesting to note that in our formulation, regions corresponding to patterns only occurring in negative bags may be emphasized by the conformal transformation. In content based image retrieval, a classifier looking for images of tigers may determine that tigers and automobiles are very unlikely to co-occur. Any image in which a car appears will therefore be considered to be less likely to also have a tiger present. In previous MIL approaches, background patterns are explicitly assumed to have no effect on classification.

The clustering approach used here may be able to be improved upon by placing the radial basis function centers in a way that makes use of bag labels. As the conformal transformation will emphasize areas that have a high concentration of positive or negative features, we may be able to increase performance given a fixed number of RBF centers by placing them in these regions to begin with. We are also interested in different forms of the conformal function, such as spectral decompositions.

Acknowledgments

The first author is supported by a Marie Curie fellowship under the EU funded project PerAct, EST 504321. This work is funded in part by the CLASS project, IST 027978.

References

[1] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[3] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, March 2001.

[4] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *SIGIR*, 1992.

[5] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[6] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proc. 19th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2002.

[7] D. Haussler. Convolutional kernels on discrete structures. Technical Report UCSCCRL-99-10, Computer Science Department, University of California at Santa Cruz, 1999.

[8] J. D. Keeler, D. E. Rumelhart, and W.-K. Leow. Integrated segmentation and recognition of hand-printed numerals. In *Advances in Neural Information Processing Systems (NIPS*3)*, pages 557–563, 1990.

[9] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004.

[10] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998.

[11] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proc. 15th International Conf. on Machine Learning*, pages 341–349. Morgan Kaufmann, San Francisco, CA, 1998.

[12] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006.

[13] Paul Viola, John Platt, and Cha Zhang. Multiple instance boosting for object detection. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006.

[14] Si Wu and Shun-Ichi Amari. Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers. *Neural Process. Lett.*, 15(1):59–67, 2002.

[15] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts. Content-based image retrieval using multiple-instance learning. In *Proc. 19th International Conference on Machine Learning*, pages 682–689.