Multi-task Learning with Weak Class Labels: Leveraging iEEG to Detect Cortical Lesions in Cryptogenic Epilepsy

Bilal Ahmed bahmed01@cs.tufts.edu
Department of Computer Science, Tufts University
Medford, MA, USA

Thomas Thesen thomas.thesen@med.nyu.edu
Karen E. Blackmon karen.blackmon@nyumc.edu
Ruben Kuzniecky ruben.kuzniecky@nyumc.edu
Orrin Devinsky od4@med.nyu.edu
Comprehensive Epilepsy Center, Department of Neurology
School Of Medicine, New York University
New York, NY, USA

Jennifer G. Dy jdy@ece.neu.edu
Department of Electrical and Computer Engineering, Northeastern University
Boston, MA, USA

Carla E. Brodley c.brodley@neu.edu
College of Computer and Information Science, Northeastern University
Boston, MA, USA

Abstract

Multi-task learning (MTL) is useful for domains in which data originates from multiple sources that are individually under-sampled. MTL methods are able to learn classification models that have higher performance as compared to learning a single model by aggregating all the data together or learning a separate model for each data source. The performance of these methods relies on label accuracy. We address the problem of simultaneously learning multiple classifiers in the MTL framework when the training data has imprecise labels. We assume that there is an additional source of information that provides a score for each instance which reflects the certainty about its label. Modeling this score as being generated by an underlying ranking function, we augment the MTL framework with an added layer of supervision. This results in new MTL methods that are able to learn accurate classifiers while preserving the domain structure provided through the rank information. We apply these methods to the task of detecting abnormal cortical regions in the MRIs of patients suffering from focal epilepsy whose MRI were read as normal by expert neuroradiologists. In addition to the noisy labels provided by the results of surgical resection, we employ the results of an invasive intracranial-EEG exam as an additional source of label information. Our proposed methods are able to successfully detect abnormal regions for all patients in our dataset and achieve a higher performance as compared to baseline methods.

1. Introduction

Multi-task learning (MTL) simultaneously learns multiple related prediction tasks which can be represented using a shared common structure (Caruana, 1997). MTL is ideally suited for domains in which the data is collected from different sources, each of which when considered individually does not have enough data to facilitate learning a reliable prediction model. However, because the data are collected from disparate sources, the underlying distribution for each source has its own distinct characteristics. This causes a co-variate shift that negatively impacts the performance of a classifier learned by simply pooling the data together. MTL exploits the "relatedness" among tasks by sharing the common information, through joint representation and regularization.

MTL requires accurately annotated training data. However, in many domains there is significant label noise that arises due to human subjectivity, imprecise measurements, etc. Because MTL allows information to be shared among tasks during the learning phase, the impact of label noise can be compounded, undermining model performance. In this paper we formulate an MTL model that learns from imprecise labels, given access to an additional source of information which provides a score for an instance quantifying the confidence associated with its label.

Our research is motivated by the task of identifying cortical malformations in the MRIs of patients suffering from treatment-resistant epilepsy (TRE) caused by focal cortical dysplasia (FCD) (Bernasconi et al., 2011; Thesen et al., 2011; Hong et al., 2014). The machine learning task is to develop a classifier that can distinguish between normal and abnormal cortical tissue using MRI data. For these patients the only treatment to lead a normal seizure-free life is surgical removal of the affected cortical tissue. However, in patients suffering from FCD which is one of the leading causes of TRE, 45-60% have a normal MRI (Wang et al., 2013). In the absence of a visually detected lesion (*MRI-negative*), the success of surgical resection drops from 66% to 29% (Bell et al., 2009). From a machine learning perspective detecting FCD lesions for MRI-negative patients has two major challenges: *1) Label noise:* The goal of resective surgery is to remove the entire lesion. If any part of the lesion is left behind, the outcome will not be successful. The margin(s) around the lesion is marked in a "generous" manner to maximize the likelihood of removing the entire lesion. This introduces false positives in training data.*2) Inter-patient variability:* The morphology of the human brain is affected by different demographic factors such as age and gender (Salat et al., 2004). Therefore, treating the data from different patients in an identical manner leads to poor classification accuracy.

To address inter-patient variability we treat each patient as a separate classification task. To this end, we use the patient's MRI to isolate the ressected region (positive instances) and extract the same region from an age and gender matched healthy control (negative instances). We then use MTL to learn a common classifier (across all tasks), using the datasets gathered from all the subjects. The common classifier is then used to detect FCD lesions in new patients.

Before undergoing resective brain surgery, all patients are subjected to an invasive intracranial EEG (iEEG) exam. A board of certified epileptologists reviews recorded electrical activity to determine the region that is responsible for generating the seizure i.e., the seizure onset zone. To isolate the abnormal region, each electrode is labeled as being part of the seizure onset zone or not. However, for MRI-negative patients there is no visible lesion to guide precise electrode implantation, which results in sampling errors in about 40% of the cases (Hong et al., 2014). Nevertheless, iEEG provides valuable information for locating the seizure onset zone. In this work, we augment the MRI labels i.e., the ressected regions for MRI-negative patients with the results of iEEG analysis to mitigate the effects of label noise. Using this combined supervision, our proposed approach detects lesions in the resection zones for all MRI-negative patients in our sample, as compared to baseline methods that achieve lower detection rates and/or have higher false positive rates. It should be noted here, that during classification we only have access to patient's MRI data, and the extra supervision from iEEG is used only for training. Thus, our proposed method serves as a pre-surgical patient evaluation tool that detects candidate lesional regions on a patient's MRI data; the candidate lesions are further evaluated using invasive iEEG and video monitoring to locate the final target for resective surgery.

The contributions of this work are: 1) we extend the regularized multitask learning framework (Evgeniou and Pontil, 2004) to incorporate auxiliary label information when the training data has weak labels; 2) we model the case when the auxiliary information has similar semantics across tasks, and the case when its semantics differ among tasks; and 3) we cast the task of detecting FCD lesions in MRI-negative patients, as an MTL problem and incorporate the results of their iEEG exams to provide additional supervision to ameliorate the problem of label noise that arises when resection zones are used as ground truth.

2. Related Work

MTL has been successfully applied in various application domains, such as object detection (Torralba et al., 2004) and conjoint analysis (Lenk et al., 1996). A number of different approaches have been taken to develop robust MTL frameworks, including hierarchical Bayes (Thrun and Pratt, 1996), regularization (Evgeniou and Pontil, 2004), and Gaussian processes (Wang and Khardon, 2012). In this work we extend the regularized MTL framework (Evgeniou and Pontil, 2004) to incorporate auxiliary supervision when the training labels are uncertain. We have chosen this framework as it admits a solution similar to support vector learning (Vapnik, 1995), which along with learning large-margin classifiers allows the use of kernel functions (Schölkopf and Smola, 2001).

Turning now to prior efforts in dealing with label noise, there are two main approaches. The first is to identify the noisy labels (Brodley and Friedl, 1999), and either discard them or assign them lower weights (Rebbapragada and Brodley, 2007). The second approach assumes that we are provided with scores quantifying the uncertainty of each training label. For example, in learning from probabilistic labels (Smyth et al., 1994; López-Cruz et al., 2013) the class of each instance is specified by a probability distribution over the possible class labels. Our work falls within this second approach but instead of assuming that the training labels are "soft" we assume that we have access to a secondary source of information, providing the means to infer the probability for a particular instance as belonging to either the positive or negative class. In this regard, the work that is closest to ours is Nguyen et al., (Nguyen et al., 2011), in which the authors assume the availability of an additional source of label information. They model this side information as inducing a pair-wise ranking in the context of learning a binary classifier. The main difference between their approach and our work is that we consider the inclusion of additional label information in the context of MTL rather than for single task learning. We go beyond simply incorporating the additional label information into the regularized MTL framework, by taking two different modeling approaches. In the first approach we assume that the additional source behaves uniformly across tasks, while in the second approach we allow its underlying semantics to be different for different tasks.

3. MTL with Auxiliary Label Information

We first provide the details of our notation and review the regularized MTL framework (Evgeniou and Pontil, 2004). For clarity and ease of comparison we follow the notation of (Evgeniou and Pontil, 2004).

Notation: We consider that we have data from T related classification tasks given as (x_i^t, y_i^t), where $x_i^t \in \mathbb{R}^d$ and $y_i^t \in \{-1, 1\}$ for all $i \in \{1, 2, \ldots, m\}$ and $t \in \{1, 2, \ldots, T\}$. All tasks share the same feature space and, without loss of generality we assume that all tasks have an equal number (m) of training instances and the underlying data distributions for all tasks are different but related. The goal then is to simultaneously learn T classifiers one per task, such that $f_t(x_i^t) = y_i^t$; $\forall t \in \{1, 2, \ldots, T\}$.

In addition to labeled data we also have a label score r_i^t for each training instance. This score is considered relative to either the positive or the negative class and represents the degree of "positive-ness" or "negative-ness" of the instance. For the sake of clarity in the rest of the paper we assume that this score is generated relative to the positive class. This score induces a pairwise ranking of instances for each task: $(i, j) \in \Pi_t \ : \ r_i^t \geq r_j^t$. The ranking function Π_t is adapted from rank-SVM (Joachims, 2002). The score assigned by the ranking function i.e., r_i^t reflects the degree of belief about an instance belonging to the positive class.

3.1 Regularized Multi-task Learning (MTL)

The regularized MTL approach (Evgeniou and Pontil, 2004) learns a separate classification function, $f_t(x) = w_t \cdot x$ for each individual task t, defined as: $w_t = w_0 + v_t$, where $w_0 \in \mathbb{R}^d$ is a vector that represents the parameters common to all tasks and $v_t \in \mathbb{R}^d$ are the task-specific parameters. If the tasks are highly similar, then the v_t are small relative to w_0, and vice versa. All the classifiers f_t can be learned simultaneously by solving the following optimization problem:

$$\begin{aligned} &\min_{w_0, v_t, \xi_i^t} \quad \sum_{t=1}^{T} \sum_{i=1}^{m} \xi_i^t + \frac{\lambda_1}{T} \sum_{t=1}^{T} \|v_t\|^2 + \lambda_2 \|w_0\|^2 \\ &\text{subject to:} \\ &\forall i, \ \ \forall t: \quad y_i^t \, (w_0 + v_t) \cdot x_i^t \ \geq \ 1 - \xi_i^t \ , \ \ \xi_i^t \ \geq \ 0 \end{aligned} \tag{1}$$

where, λ_1 and λ_2 are regularization parameters that control the relatedness of the tasks.The structure of MTL is captured by the kernel function $K_{st}(.)$:

$$K_{st}(x_i^t, x_j^s) = \left(\frac{1}{\mu} + \delta_{st}\right) x_i^t \cdot x_j^s \tag{2}$$

that arises from the dual of Equation 1. This multi-task kernel couples the inner-product of the instances across tasks based on $\mu = \frac{T\lambda_2}{\lambda_1}$ which represents the degree of relatedness among the tasks. The standard inner product in Equation 2 can be replaced with any kernel function (Evgeniou and Pontil, 2004).

The MTL framework outlined above requires accurate class labels during the learning phase, and the presence of label noise can seriously undermine its performance. We consider the case when the labels are noisy but there is another source of obtaining supplementary label information that can accurately grade the "positive-ness" of an instance. This additional source may represent the subjective judgment of a domain expert(s) about a particular instance, based on either the same set of features that are available to the learner or some other view of the data.

In this work we interpret the auxiliary label scores as the output of a pairwise ranking function. We can model the behavior of the ranking function as being *globally-consistent* i.e., it does not not vary across tasks. In other words, if we consider the ranking function as representing an expert's judgment, then this assumption would require that his/her evaluation criteria for ascertaining the rank of an instance does not change from one task to another. This is a strong assumption, because most real-life experts will make varying judgments based on the nature of the task at hand. We can model this variation by taking a *task-specific* approach that allows each task to have its own ranking function.

From the perspective of MTL, the difference between the two approaches is whether or not the ranking information is shared across tasks. In the *task-specific* approach, we cannot compare the ranks of two instances belonging to different tasks, because the underlying ranking functions are different. In the *globally-consistent* case, rank information can be shared among tasks without introducing any discrepancies. Similar to rank-SVM (Joachims, 2002), we take the rank of each instance as being proportional to its distance from the separating hyperplane:

$$(i, j) \in \Pi_t \ : \ w_t \cdot x_i^t \geq w_t \cdot x_j^t$$

where, Π_t is the ranking function for task t. For each pair of instances belonging to task t we augment the original MTL problem (Equation 1) with pairwise rank constraints (Nguyen et al., 2011).

3.1.1 Globally-Consistent Label Ranking (GC)

Here we consider all pairwise ranking functions Π_t, $\forall t$ to be identical. In this case we modify the original MTL framework (Equation 1) by adding rank constraints:

$$\begin{aligned}
&\min_{w_0, v_t, \xi_i^t, \eta_{pq}^t} && \frac{1}{2}\sum_{t=1}^{T} \|v_t\|^2 + \frac{\mu}{2}\|w_0\|^2 + C\sum_{t=1}^{T}\sum_{i=1}^{m} \xi_i^t + C' \sum_{t=1}^{T} \sum_{(p,q)\in\Pi_t} \eta_{pq}^t \\
&\text{subject to:} && \\
&\forall i, \ \ \forall t: && y_i^t \left(w_0 + v_t\right) \cdot x_i^t \ \geq \ 1 - \xi_i^t \ , \quad \xi_i^t \ \geq \ 0 \\
&\forall t, \forall \left(p, q\right) \in \Pi_t: && v_t \cdot \left(x_p^t - x_q^t\right) \geq 1 - \eta_{pq}^t \ , \quad \eta_{pq}^t \ \geq \ 0
\end{aligned} \tag{3}$$

where η_{pq}^t are slack variables. C' is a positive scalar and represents the relative cost of violating a rank constraint. It is defined as: $C' = aC, a \in \mathbb{R}^+$. The rank constraints can be re-written as:

$$z_{pq}^t \left(w_0 + v_t\right) \cdot \Delta_{pq}^t \ \geq \ 1 - \eta_{pq}^t$$

where, $z_{pq}^t = 1$ are the labels for each pseudo-example: $\Delta_{pq}^t = x_p^t - x_q^t$. By augmenting the data of each task with the pseudo-examples Δ_{ij} and their corresponding labels $z_{ij} = 1$ we combine the two sets of constraints to solve a single classification problem. However, the number of pseudo-examples is quadratic in terms of the number of instances in the original dataset for each task. This will cause the number of positive instances to be substantially higher than the number of negative instances, which in the worst case scenario can result in a degenerate solution in which the resulting hyperplanes only respect the rank constraints. The trade-off between preserving the ranking and accurate classification is controlled by the cost parameters C' and C, respectively. The cost parameters are analogous to the cost parameter of the traditional support vector machine (SVM) (Schölkopf and Smola, 2001). We set these parameters based on a grid search over a hold-out dataset, which is the standard procedure for training SVMs. The optimization problem in Equation 3 can be efficiently solved by formulating its dual using the kernel function from Equation 2 (Supplemental)[1].

It is worth mentioning that no pseudo-examples are created by comparing the ranks of two instances from different tasks. The assumption of global consistency is exploited in the construction of the rank constraints (c.f., Equation 3), in which both w_0 and v_t are required to preserve the ranking.

3.1.2 Task-Specific Label Ranking (TS)

To model the peculiarities that may exist in the auxiliary information as we move from one task to another, we can limit the influence that the rank constraints have on the overall solution, by limiting them to affect only the task-specific components. This is formulated similar to Equation 3, with modified rank constraints:

$$\forall t, \forall\, (p,q) \in \Pi_t : \; z_{pq}^t \; v_t \cdot \Delta_{pq}^t \geq 1 - \eta_{pq}^t \tag{4}$$

By not allowing the rank information to directly influence the shared component w_0, the ranking function Π_t is no longer coupled across tasks, and can behave differently for different tasks. It should be noted that although the shared weight vector w_0 is not required to preserve the rankings, it is still indirectly affected by the rank constraints through v_t ($\because w_t = w_0 + v_t$). This problem can be solved efficiently by formulating its dual (Supplemental), which gives rise to a new kernel function. Let $\tilde{X}^t \in \mathbb{R}^d$ be the augmented data for task t obtained by combining all the original data instances (x_i^t) and the pseudo-examples (Δ_{pq}^t), and let u_i^t be an indicator variable which is 0 when the instance is a pseudo-example and 1 otherwise. Then the new kernel is given by:

$$K_{st}(\tilde{x}_k^t, \tilde{x}_l^s) = \left(\frac{u_k^t u_l^s}{\mu} + \delta_{st} \right) \tilde{x}_k^t \cdot \tilde{x}_l^s \tag{5}$$

This is a multitask kernel that does not allow the ranking function Π_t to directly impact w_0, restricting the auxiliary label information from being directly shared among tasks.

In both the globally-consistent and task-specific cases, the final optimization takes the form of a standard quadratic program (QP) with box constraints (Boyd and Vandenberghe, 2004) which can be easily solved using any off-the-shelf QP-solver.[2]

4. Detecting Cortical Malformations in Epilepsy

For patients suffering from treatment-resistant epilepsy (TRE) caused by FCD, early detection and subsequent surgical removal of the lesion area is the only treatment. A com-

1. Please refer to the supplementary material accompanying this paper.
2. We used the QP solver included in the optimization toolbox for Matlab (http://www.mathworks.com).

prehensive clinical evaluation for surgery candidates involves a neurological exam, scalp electroencephalography (EEG), neuropsychological exam, and a visual inspection of the patient's MRI. In 45-60% of such cases, no lesion is detected on the MRI by expert neuroradiologists (Wang et al., 2013). Our research focuses on developing machine learning algorithms that can detect these subtle FCD lesions using the patient's MRI data. To this end we use training data that includes patients who underwent surgery and their ressected tissue was histopathologically verified to contain FCD.

4.1 Surface-Based Morphometry and Lesion Detection

Surface based morphometry (SBM) models the cortical surface as a folded two-dimensional surface. It is extracted by delineating the boundary between the gray and white matter using T1-weighted MRI images (Dale et al., 1999). After extraction the cortical surface can be registered to a standard surface also known as a group-atlas. Registration allows for a more precise comparison of individual cortical structures across subjects (Fischl and Dale, 2000). SBM has been used successfully for analyzing and detecting neurological abnormalities in Schizophrenia (Rimol et al., 2012), Autism (Nordahl et al., 2007), and Epilepsy (Thesen et al., 2011; Hong et al., 2014). SBM has been used in conjunction with different machine learning and statistical techniques to identify lesions in MRI-positive FCD patients by classifying each vertex on the cortical surface. These approaches include neural networks (Besson et al., 2008), uni-variate z-score based thresholding (Thesen et al., 2011), stratified ensemble of logistic regression classifiers (Ahmed et al., 2015), semi-supervised conditional random fields (Ahmed et al., 2014) and linear discriminant analysis (LDA) (Hong et al., 2014).

Recall, that one of the major confounding factors in this data is the label noise that arises from using resection zones as ground truth. In our approach to minimize the effects of label noise we do not pre-process the resection zone prior to classifier training based on ad-hoc heuristics (Ahmed et al., 2015) or subjective expert opinions (Hong et al., 2014; Besson et al., 2008). Instead, we augment the resection zones with the results of the iEEG analysis as positive labels for the training data.

None of the detection schemes address the co-variate shift arising from inter-patient variability in brain morphology and the differences in patterns of seizure onset and intensity. To overcome this co-variate shift we treat each patient as a separate classification task. To this end, we isolate a patient's resection zone and the same region from an age and gender matched control. Both the lesional and normal images are segmented to obtain super-pixels which are used as data instances. We learn all patient specific classifiers simultaneously using our proposed MTL approaches. In regularized MTL framework (c.f. section-3.1), the trained model is tested on previously left-out data from the same tasks that generated the training data. However, in our case test data comes from new patients that were not part of the training data. For classifying data from out-of-sample tasks, we discard task-specific parameter vectors v_t and use the mean component w_0 (Van Esbroeck et al., 2016).

4.2 Data Description and Pre-processing

The dataset consists of 18 MRI-negative FCD patients collected over a three year period from New York University's comprehensive epilepsy center which is a leading level-4 epilepsy treatment center. All 18 patients underwent successful resective surgery and histopathological examination showed evidence of FCD. This may seem a small set but note, that only a few MRI-negative patients proceed to surgery, and out of those only a third have successful outcomes (Bell et al., 2009). The controls were matched from a cohort of 115 neurotypical controls. All patients and controls were scanned using the same imaging protocol (Supplemental).

Our training and test data comprises of image patches taken from the resected regions of the patients, and corresponding regions from matched controls. We focus only on the resected regions because all the patients included in our experiments were seizure-free after surgery, which shows that their resected regions contained the primary FCD lesion(s). Furthermore, for new MRI-negative patients, *seizure-semiology* i.e., signs and symptoms of a seizure provides credible but crude estimation about the location of the epileptogenic zone (Rosenow and Lüders, 2001; Noachtar and Peters, 2009; Tufenkjian and Lüders, 2012). In these cases our methods can be applied to the suspected cortical region.

Segmentation: All the patient and control surfaces were registered to an average surface. After registration the resected region(s) for each patients and the corresponding region(s) from his/her matched control was isolated, and flattened to a standard 2-d image. We use cortical thickness to represent the intensity values.It should be noted that even though we use cortical thickness to obtain the super-pixels, a larger set of morphological features (including cortical thickness) is used to characterize each super-pixel.

We use Quickshift (Vedaldi and Soatto, 2008) for unsupervised segmentation. Quickshift has two parameters: size of the Gaussian kernel (σ_{QS}) used by a Parzen window density estimator, and the maximum distance (δ_{QS}) between two pixels permitted while remaining part of the same segment. All the patients and controls were segmented using the same set of Quickshift parameters ($\sigma_{QS} = 8, \delta_{QS} = 32$) that were optimized using a parameter estimation set of three patients, which are distinct from the fifteen patients whose results are reported here. After segmentation, each super-pixel is treated as an independent instance. We used first and second-order statistics of cortical thickness, gray-white contrast (GWC), curvature, sulcal depth, Jacobian distance and local gyrification index (LGI) to represent each super-pixel (Thesen et al., 2011). Additionally, we also included the average surface area measured on both the pial and white matter surfaces.

Creating Electrode Maps: The spatial resolution of the iEEG and MRI are not identical: the MRI voxels (and corresponding surface vertices) are smaller than the point spread function that defines the iEEG generators. More importantly, localizing the source for a given electrode is subject to the ill-posed inverse problem (Yuan et al., 2012). These are well-known limitations of iEEG. Solving this problem is beyond the scope of this work. However, a sphere with half the diameter of the inter-electrode distance (approximately ten mm), presents a reasonable criterion for matching iEEG and MRI. Therefore, all surface vertices within a radius of five mm (Euclidean distance) from the electrode's location were considered within range of the electrode. We only selected electrodes that were labeled by the experts as either being part of the seizure-onset zone or recorded abnormal electrical activity during seizure onset.

	Recall						False Postive Rate					
Id.	LDA	SLR	SVMR	MTL	GC	TS	LDA	SLR	SVMR	MTL	GC	TS
P01	0.07	0.04	0.48	-	-	0.09	-	0.05	0.22	-	-	0.21
P02	0.11	0.19	0.39	0.28	0.09	0.25	0.18	0.14	0.56	0.32	0.15	0.29
P03	0.17	-	0.52	0.22	0.02	0.11	0.07	0.03	0.26	0.08	-	0.03
P04	-	0.12	0.30	0.20	-	0.20	0.1	0.10	0.35	0.17	-	0.13
P05	0.15	0.03	0.60	0.03	-	0.03	-	0.02	0.26	0.03	-	0.03
P06	0.12	-	0.62	-	-	0.09	-	0.05	0.39	0.04	-	0.09
P07	-	-	0.30	-	-	0.09	0.14	0.03	0.26	0.02	-	-
P08	0.16	0.13	0.67	0.12	0.06	0.12	0.07	0.04	0.30	0.06	0.03	0.03
P09	0.15	0.06	0.49	0.08	0.46	0.22	-	-	0.48	-	0.16	0.05
P10	0.04	-	0.42	0.04	-	0.04	0.05	0.01	0.23	0.09	-	0.13
P11	0.08	-	0.37	0.07	-	0.05	-	0.05	0.25	0.11	-	0.15
P12	-	0.05	0.57	0.05	-	0.05	0.07	0.07	0.23	0.05	-	0.05
P13	0.13	0.21	0.95	0.16	-	0.44	0.13	0.11	0.70	0.13	-	0.14
P14	-	-	0.28	0.14	-	0.21	0.02	0.16	0.65	0.16	-	0.15
P15	0.09	0.15	0.53	0.04	-	0.1	0.15	0.05	0.39	-	-	0.12
Mean	0.09	0.07	0.50	0.10	0.04	0.14	0.07	0.06	0.37	0.08	0.02	0.11

Table 1: Detailed results for MRI-negative subjects. LDA is the Fisher linear discriminant analysis based method (Hong et al., 2014), SLR represents the stratified classification scheme (Ahmed et al., 2015), SVMR is a single task SVM with rank constraints (Nguyen et al., 2011), MTL represents regularized MTL (Evgeniou and Pontil, 2004) without rank constraints, GC and TS are the globally-consistent and the task-specific approaches, respectively ('-' represents a value of zero).

5. Results

The lesions of all 18 MRI-negative patients used in our experimental evaluation were located in the temporal region, which is one of the most prevalent localization of FCD in adults (Krsek et al., 2008). All patients had undergone successful resective surgery and were histopathologically verified to have FCD. We make two sets of comparisons: one designed to evaluate the effectiveness of the MTL framework and the other to evaluate whether including auxiliary supervision in the MTL framework boosts the lesion detection rate over MTL alone.

Baseline Selection: To show that the regularized MTL framework (Evgeniou and Pontil, 2004) can be used effectively to detect FCD lesions, we adapt the LDA based classification scheme (Hong et al., 2014) as one of the baselines. To this end we use LDA to classify superpixels using the same set of features as used by our proposed MTL methods. Additionaly, we also compare the performance of our proposed methods to a stratified vertex-based classification approach (Ahmed et al., 2015) (SLR). In all of our experiments we have not used any post-processing for any of the proposed methods or baselines.

Experimental Setup: A leave-one-patient-out cross-validation strategy was used, in which we left out one patient's data and trained on the remaining patients. Hyperparameters for the proposed methods and baselines were set during each round, using the data of three MRI-negative patients and their matched controls (Supplemental) whose iEEG data was not available. We will refer to this set of three patients as the model parameter set (MPS). Data in MPS are distinct from the patients and controls used in our experiments.

Performance Analysis: We have developed the proposed methods keeping in view their final use as focus-of-attention tools for neuroradiologists for detecting visually elusive FCD lesions. Therefore, the detection rate i.e., the number of patients whose lesions are correctly detected constitute the main result. For a more detailed performance analysis, we calculate the recall and false positive rate (FPR) based on surface area. Recall represents the percentage of the surface area on the patient's resection zone correctly identified as lesional, and FPR is the percentage of the surface area incorrectly labeled as lesional on the matched control's selected region. Note that the estimates for recall should be considered as *lower bounds* because they are calculated using the noisy resection zones as ground truth.

Table-1 compares the detection rate (number of patients with positive recall), recall and FPR of the proposed *task-specific* (TS) and the *globally-consistent* (GC) approaches with baseline methods. Among the baselines that do not rely on iEEG data, MTL outperforms both LDA and SLR by correctly detecting the lesions in twelve patients, whereas LDA and SLR detect the lesion in eleven and nine patients, respectively. Not only does MTL achieve a higher detection rate, it also has higher average recall than both LDA and SLR. On the other hand, TS detects the lesion in all fifteen patients and achieves higher average recall as compared to LDA, SLR and MTL. Both TS and SVMR have the same detection rate, but TS clearly outperforms SVMR as far as average FPR is concerned, the average FPR of TS is significantly lower than that of SVMR which is unacceptably high (37%), when interpreted in the context of its use as a focus-of-attention tool. The lower FPR of TS shows that an MTL based formulation coupled with auxiliary supervision leads to superior detection rate and lower FPR. Turning now to GC, we see that it has the worst performance in terms of detection rate among all the methods. The low detection rate of GC when compared to TS, substantiates our assumption that the information obtained from the iEEG analysis has task-specific semantics and the criteria used for determining the seizure-onset zone differ on a patient by patient basis. When this information was shared freely among tasks, the label noise was further compounded leading to low detection rate of the GC method.

6. Conclusion

In this work we addressed the problem of MTL in the presence of uncertain labels, assuming an additional source of supervision, that we modeled as a pairwise ranking function. To this end, we extended the regularized MTL framework (Evgeniou and Pontil, 2004) by incorporating additional rank constraints. We modeled the case when there is a single ranking function for all tasks, and the case where each task has its own ranking function. In the latter *task-specific* case we developed a new operator-valued kernel (Evgeniou et al., 2005) for ensuring that ranks are not directly shared between tasks. In all cases, the model parameters were found by solving a quadratic optimization problem (QP) with box-constraints (Boyd and Vandenberghe, 2004), which is solvable with any standard QP solver. We demonstrated the efficacy of the proposed method on the challenging problem of detecting epileptogenic lesions. The task-specific approach detected lesions within the resections of all patients, whereas the baselines either achieved lower detection rates or had an unacceptably high false positive rate. Identifying the abnormal region in cryptogenic epilepsies is based on a confluence of evidence from multiple sources such as MRI, PET, etc. In this regard, our proposed method will have a deeper impact in the application domain by enhancing the sensitivity of the patient evaluation methodology. Furthermore, the proposed methods can be broadly applied to other domains in which decisions are made based on converging evidence from disparate sources.

References

B. Ahmed, C. E. Brodley, K. E. Blackmon, R. Kuzniecky, G. Barash, C. Carlson, B. T. Quinn, W. Doyle, J. French, O. Devinsky, and T. Thesen. Cortical feature analysis and machine learning improves detection of mri-negative focal cortical dysplasia. *Epilepsy & Behavior*, 48:21 – 28, 2015.

Bilal Ahmed, Thomas Thesen, Karen Blackmon, Yijun Zhao, Orrin Devinsky, Ruben Kuzniecky, and Carla Brodley. Hierarchical conditional random fields for outlier detection: An application to detecting epileptogenic cortical malformations. *Proceedings of the 31st International Conference on Machine Learning*, pages 1080–1088, 2014.

M.L. Bell, S. Rao, E.L. So, et al. Epilepsy surgery outcomes in temporal lobe epilepsy with a normal MRI. *Epilepsia*, 50(9):2053–2060, 2009.

A. Bernasconi, N. Bernasconi, B.C. Bernhardt, and D. Schrader. Advances in mri for 'cryptogenic' epilepsies. *Nat Rev Neurol.*, 7(2):99–108, 2011.

P. Besson, N. Bernasconi, O. Colliot, et al. Surface-based texture and morphological analysis detects subtle cortical dysplasia. In *MICCAI '08*, pages 645–652, 2008.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 0521833787.

Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. *J. A.I. Res.*, 11:131–167, 1999.

Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997.

A.M. Dale, B. Fischl, M.I. Sereno, et al. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *ACM SIGKDD*, KDD '04, pages 109–117, 2004. ISBN 1-58113-888-1.

Theodoros Evgeniou, Charles A. Micchelli, Massimiliano Pontil, et al. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, 6:615–637, 2005. ISSN 1532-4435.

B. Fischl and A.M. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS*, 97(20):11050–11055, 2000.

Seok-Jun Hong, Hosung Kim, Dewi Schrader, Neda Bernasconi, Boris C. Bernhardt, and Andrea Bernasconi. Automated detection of cortical dysplasia type II in MRI-negative epilepsy. *Neurology*, 83(1):48–55, 2014.

T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, KDD '02, pages 133–142, 2002.

Pavel Krsek, Bruno Maton, Brandon Korman, Esperanza Pacheco-Jacome, Prasanna Jayakar, Catalina Dunoyer, et al. Different features of histopathological subtypes of pediatric focal cortical dysplasia. *Annals of Neurology*, 63(6):758–769, 2008.

Peter J. Lenk, Wayne S. DeSarbo, Paul E. Green, and Martin R. Young. Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.

Pedro L. López-Cruz, Concha Bielza, Pedro Larrañaga, et al. Learning conditional linear gaussian classifiers with probabilistic class labels. In *CAEPIA '13*, pages 139–148, 2013.

Quang Nguyen, Hamed Valizadegan, Milos Hauskrecht, et al. Learning classification with auxiliary probabilistic information. In *IEEE ICDM '11*, pages 477–486, 2011.

S. Noachtar and A.S. Peters. Semiology of epileptic seizures: A critical review. *Epilepsy Behav.*, 15(1):2–9, 2009.

C.W. Nordahl, D. Dierker, and I., et al. Mostafavi. Cortical folding abnormalities in autism revealed by surface-based morphometry. *J Neurosci.*, 27(43):11725–11735, 2007.

Umaa Rebbapragada and Carla E. Brodley. Class noise mitigation through instance weighting. In *ECML '07*, pages 708–715, 2007.

L.M. Rimol, R. Nesvg, and D.J., et al. Hagler Jr. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biological Psychiatry*, 71(6):552–560, 2012.

Felix Rosenow and Hans Lüders. Presurgical evaluation of epilepsy. *Brain*, 124(9):1683–1700, 2001.

D.H. Salat, R.L. Buckner, A.Z. Snyder, et al. Thinning of the cerebral cortex in aging. *Cerebral Cortex*, 14(7):721–730, 2004.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001. ISBN 0262194759.

Padhraic Smyth, Usama M. Fayyad, Michael C. Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. In *NIPS '94*, pages 1085–1092, 1994.

Supplemental. Supplementary material.

T. Thesen, B.T. Quinn, C. Carlson, et al. Detection of epileptogenic cortical malformations with surface-based MRI morphometry. *PLoS ONE*, 6(2):e16430, 2011.

S. Thrun and L. Pratt. *Learning to Learn*. Kluwer Academic Publishers, 1996.

A Torralba, K.P. Murphy, W.T. Freeman, et al. Sharing features: efficient boosting procedures for multiclass object detection. In *IEEE CVPR '04*, pages 762–769, 2004.

K. Tufenkjian and H.O. Lüders. Seizure semiology: its value and limitations in localizing the epileptogenic zone. *J Clin Neurol.*, 8(4):243–250, 2012.

Alex Van Esbroeck, Landon Smith, Zeeshan Syed, Satinder Singh, and Zahi Karam. Multi-task seizure detection: addressing intra-patient variation in seizure morphologies. *Machine Learning*, 102(3):309–321, 2016.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. pages 705–718, 2008.

Yuyang Wang and Roni Khardon. Sparse gaussian processes for multi-task learning. In *ECML/PKDD'12*, pages 711–727, 2012.

Z. I. Wang, A. V. Alexopoulos, S. E. Jones, Z. Jaisani, I. M. Najm, and R. A. Prayson. The pathology of magnetic-resonance-imaging-negative epilepsy. *Mod Pathol*, 26(8):1051–1058, 2013.

Jinxian Yuan, Yangmei Chen, and Edouard Hirsch. Intracranial electrodes in the presurgical evaluation of epilepsy. *Neurological Sciences*, 33(4):723–729, 2012.

Supplementary Material

1. MTL with Rank Constraints

Here we provide the primal and dual problems for both the globally consistent and task-specific label ranking appproaches.

1.1 MTL with globally-consistent label ranking:

The primal optimization problem is given as:

$$\min_{w_0,v_t,\xi_i^t,\eta_{pq}^t} \left\{ \frac{1}{2}\sum_{t=1}^{T}\|v_t\|^2 + \frac{\mu}{2}\|w_0\|^2 + \right.$$
$$\left. +C\sum_{t=1}^{T}\sum_{i=1}^{m}\xi_i^t + C'\sum_{t=1}^{T}\sum_{(p,q)\in\Pi_t}\eta_{pq}^t \right\} \tag{1}$$

subject to:

$$\begin{aligned} \forall i, \quad \forall t: \; y_i^t\,(w_0+v_t)\cdot x_i^t \;&\geq\; 1-\xi_i^t \\ \forall t, \forall\,(p,q)\in\Pi_t: z_{pq}^t\,(w_0+v_t)\cdot\Delta_{pq}^t \;&\geq\; 1-\eta_{pq}^t \\ \xi_i^t \;\geq\; 0 \;\;, \quad \eta_{pq}^t \;&\geq\; 0 \end{aligned}$$

where η_{pq}^t and ξ_i^t are slack variables and $z_{pq}^t = 1$ are the labels for each pseudo-example: $\Delta_{pq}^t = x_p^t - x_q^t$. Note that, pseudo-examples are created on a per task basis i.e., there are no pseudo-examples resulting from comparing the ranks of two instances from different tasks. The dual of above problem can be formulated as:

$$\begin{aligned} \max_{\alpha_i^t,\beta_{pq}^t} \; & \left\{ \sum_{t=1}^{T}\sum_{i=1}^{m}\alpha_i^t + \sum_{t=1}^{T}\sum_{(p,q)\in\Pi_t}\beta_{pq}^t - \right. \\ & \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{s=1}^{T}\sum_{j=1}^{m}\alpha_i^t y_i^t \alpha_j^s y_j^s K_{st}\left(x_i^t, x_j^s\right) - \\ & -\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{s=1}^{T}\sum_{(p,q)\in\Pi_t}\alpha_i^t y_i^t \beta_{pq}^s z_{pq}^s K_{st}\left(x_i^t, \Delta_{pq}^s\right) - \\ & \left. -\frac{1}{2}\sum_{t=1}^{T}\sum_{(p,q)\in\Pi_t}\sum_{s=1}^{T}\sum_{(k,l)\in\Pi_s}\beta_{pq}^t z_{pq}^t \beta_{kl}^s z_{kl}^s K_{st}\left(\Delta_{pq}^t, \Delta_{kl}^s\right) \right\} \end{aligned} \tag{2}$$

subject to:

$$\forall i, \quad \forall t: \; 0 \leq \alpha_i^t \leq C \;,\; 0 \leq \beta_{pq}^t \leq C'$$

where, α_i^t and β_{pq}^t are Lagrange multipliers corresponding to the classification and rank constraints, respectively. $K_{ij}(.,.)$ is defined as (Evgeniou and Pontil, 2004):

$$K_{st}(x_i^t, x_j^s) = \left(\frac{1}{\mu} + \delta_{st}\right) x_i^t \cdot x_j^s \tag{3}$$

The optimal solutions for both w_0 and v_t can be obtained by formulating and solving the Lagrangian function for Equation 1.

$$w_0^* = \frac{1}{\mu}\left[\sum_{t=1}^{T}\sum_{i=1}^{m}\alpha_i^t y_i^t x_i^t + \sum_{s=1}^{T}\sum_{(p,q)\in\Pi_s}\beta_{pq}^s z_{pq}^s \Delta_{pq}^s\right] \tag{4}$$

$$v_t^* = \sum_{i=1}^{m}\alpha_i^t y_i^t x_i^t + \sum_{(p,q)\in\Pi_t}\beta_{pq}^t z_{pq}^t \Delta_{pq}^t \tag{5}$$

1.2 MTL with task-specific label ranking:

In this formulation we limit the influence of rank constraints on the overall solution, by restricting them to affect only their own task-specific components. This is formulated as:

$$\begin{aligned}\min_{w_0, v_t, \xi_i^t, \eta_{pq}^t} & \left\{\frac{1}{2}\sum_{t=1}^{T}\|v_t\|^2 + \frac{\mu}{2}\|w_0\|^2 + \right. \\ & \left. C\sum_{t=1}^{T}\sum_{i=1}^{m}\xi_i^t + C'\sum_{t=1}^{T}\sum_{(p,q)\in\Pi_t}\eta_{pq}^t\right\}\end{aligned} \tag{6}$$

subject to:

$$\begin{aligned}\forall i,\ \ \forall t: &\ \ y_i^t\,(w_0 + v_t)\cdot x_i^t \ \geq\ 1 - \xi_i^t \\ \forall t, \forall\,(p,q)\in\Pi_t: &\ \ z_{pq}^t\ v_t\cdot\Delta_{pq}^t \geq 1 - \eta_{pq}^t \\ &\ \ \xi_i^t \ \ \geq\ 0\ \ ,\ \ \ \eta_{pq}^t\ \ \geq\ 0\end{aligned}$$

By not allowing the rank information to directly influence the shared component w_0, the ranking function Π_t is no longer coupled across tasks, and can behave differently for different tasks. It should be noted that although the shared weight vector w_0 is not required to preserve the rankings, it is still indirectly affected by the rank constraints through v_t (since, $w_t = w_0 + v_t$). The dual for Equation 6 can be formulated in a form identical to Equation 2, by using a new kernel function. Let $\tilde{X}^t \in \mathbb{R}^d$ be the augmented data for task t obtained by combining all the original data instances (x_i^t) and the pseudo-examples (Δ_{pq}^t), and let u_i^t be an indicator variable defined as:

$$u_k^t = \begin{cases} 1 & \text{if } \tilde{x}_k^t = x_i^t, \\ 0 & \text{if } \tilde{x}_k^t = \Delta_{pq}^t. \end{cases}$$

where, $k \in \{1, \ldots, |\tilde{X}^t|\}$. Let the kernel function be defined as:

$$K_{st}(\tilde{x}_k^t, \tilde{x}_l^s) = \left(\frac{u_k^t u_l^s}{\mu} + \delta_{st}\right)\tilde{x}_k^t \cdot \tilde{x}_l^s \tag{7}$$

This is an operator-valued kernel (Evgeniou et al., 2005) that does not allow the ranking function Π_t to directly impact w_0, thus, restricting the auxiliary label information from being directly shared among tasks.

In this formulation, the optimal solution for v_t does not change and remains identical to Equation 5. The difference lies in the optimal solution for the shared parameter vector w_0, which in this case is given as:

$$w_0^* = \frac{1}{\mu}\sum_{t=1}^{T}\sum_{i=1}^{m}\alpha_i^t y_i^t x_i^t \tag{8}$$

As expected, it can be seen that w_0 is no longer affected by the rank constraints.

2. Imaging

Participants were selected from a large registry of patients with epilepsy. Criteria for inclusion in this study included: (1) completion of a high resolution T1-weighted MRI scan; (2) surgical resection to treat focal epilepsy; (3) diagnosis of FCD on neuropathological examination of the resected tissue.

2.1 Imaging Protocol

Imaging was performed on a Siemens Allegra 3T scanner. Image acquisitions included a conventional 3-plane localizer and a T1-weighted volume pulse sequence (TE=3.25 ms, TR =2530 ms, TI =1100 ms, flip angle =7 deg field of view (FOV) = 256 mm, matrix = 256x256, vertex size =1x1x1.3 mm, scan time: 8:07 min). Acquisition parameters were optimized for increased gray/white matter image contrast. The T1-weighted image was reoriented into a common space, roughly similar to alignment based on the AC-PC line. Images were corrected for nonlinear warping caused by no-uniform fields created by the gradient coils.

2.2 Surface Extraction

The MRI sequences were processed using the FreeSurfer software package[1], which performs automated tissue segmentation to recreate 3D representations of the cortical surfaces from structural MRI scans (Dale et al., 1999). Briefly, after skull striping, the method (Dale et al., 1999) involves: (i) segmentation of the white matter, (ii) tessellation of the gray/white matter boundary, (iii) inflation of the folded surface, and (iv)correction of topological defects. Once the surface was reconstructed it was further refined by classifying all white matter vertices in the MRI volume to create the gray/white matter boundary. The gray/white matter junction was delineated up to submillimeter accuracy by further refining the white matter surface. After refining the gray/white matter junction the pial surface was located by deforming the surface outward. Each segmentation and reconstruction underwent manual inspection and editing, when necessary. However, the high image quality and gray-white contrast in the initial images resulted in minimal editing requirements for both patient and control scans. Surface reconstruction was followed by a registration process that involved morphing the reconstructed surface to an average spherical representation that accurately matched sulcal and gyral features across individual subjects while minimizing metric distortion (Fischl et al., 1999).

3. Model Hyperparameters

In this section we explain the different model hyperparamaters for the proposed methods and the baselines. We also provide the details of how their values were set in our experiments. All model parameters for the proposed approaches and the baselines were set using the data of three MRI-negative patients and their matched controls. No iEEG data was available for these three patients, and hence we used them for setting the model hyperparameters. We

1. Available at http://surfer.nmr.mgh.harvard.edu/

Parameter	Range
μ	$10^{-7}, 5^{-6}, 10^{-6}, 5^{-5}, \ldots, 10^{3}$
C	$2^{-10}, 2^{-9}, \ldots, 2^{10}$
γ	$2^{-10}, 2^{-9}, \ldots, 2^{10}$
a	$10^{-6}, 5^{-5}, 10^{-5}, \ldots, 10^{3}$

Table 1: Range of values for the model hyper-parameters used in the grid search. The grid search optimized the area under the curve (AUC) over the model parameter set (MPS) consisting of three patients whose data is distinct from the fifteen patients used for performance analysis.

will refer to this set of three patients as the model parameter set (MPS). The data for these three patients are distinct from the fifteen patients and controls used in our experiments and whose results are reported in the paper.

3.1 Segmentation

The standard quick shift algorithm is a fast mode seeking algorithm similar to mean shift (Vedaldi and Soatto, 2008). It performs a hierarchical segmentation of the image, where the sub-trees represent image segments. It has two parameters namely the size of the Gaussian kernel (σ_{QS}) used by a Parzen window density estimator, and the maximum distance (δ_{QS}) between two pixels permitted while remaining part of the same segment. The scale parameter σ_{QS} is varied to change the average size of segments, and δ_{QS} is set to be a multiple of σ_{QS} (Vedaldi and Soatto, 2008). Thus, higher values of σ_{QS} produce larger segments. All the patients and controls were segmented using the same set of Quickshift parameters ($\sigma_{QS} = 8, \delta_{QS} = 32$) that were optimized using the MPS.

3.2 Setting the Hyper-Parameters

It should be noted that the hyper-parameters were set individually for each test subject using the MPS. Below we provide the details about how the parameters were set for the different baselines and our proposed methods:

- *LDA:* The detection threshold ($\tau \in [0, 1]$) for LDA was optimized by maximizing the area under the curve (AUC) over the MPS.
- *SLR:* The detection threshold ($\tau \in [0, 1]$) for logistic regression was optimized by maximizing the area under the curve (AUC) over the MPS.
- *SVMR:* We also used a single-task SVM using the RBF kernel that incorporates ranking constraints based on the model in (Nguyen et al., 2011) as a baseline. In addition to the cost parameter (identical to the cost parameter of tradtional SVM) and the scale parameter of the RBF kernel, there is a third parameter a that defines the relative cost of violating a ranking constraint ($C' = aC$). All three parameters were set by optimizing the AUC using the MPS.

- *MTL:* This corresponds to the regularized MTL framework that does not incorporate any auxiliary supervision, and uses the resection zone as class labels. The model parameters were set using a grid-search (Evgeniou and Pontil, 2004) and include misclassification cost (C), task-relatedness parameter (μ) and the scale (γ) of the RBF kernel. To find suitable values for the parameters we used a three-level grid and optimized the area under the curve (AUC) over the MPS.

- *GC & TS:* These are the proposed methods that incorporate auxiliary supervision derived from iEEG data. In addition to the three model parameters for MTL, there is a fourth parameter a that defines the relative cost $C' = aC$ of violating a rank constraint (c.f., Equations 1 and 6). To find suitable values for the parameters we designed a four-level grid and optimized the AUC over the MPS.

Table 1 lists the ranges for the parameters used in the grid search for SVMR, MTL, GC and TS.

References

A.M. Dale, B. Fischl, M.I. Sereno, et al. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *ACM SIGKDD*, KDD '04, pages 109–117, 2004. ISBN 1-58113-888-1.

Theodoros Evgeniou, Charles A. Micchelli, Massimiliano Pontil, et al. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, 6:615–637, 2005. ISSN 1532-4435.

B. Fischl, M.I. Sereno, and A.M. Dale. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999.

Quang Nguyen, Hamed Valizadegan, Milos Hauskrecht, et al. Learning classification with auxiliary probabilistic information. In *IEEE ICDM '11*, pages 477–486, 2011.

A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. pages 705–718, 2008.