Active Learning via Transductive Experimental Design

Kai Yu KAI.YU@SIEMENS.COM
Siemens, Corporate Technology, Otto-Hahn-Ring 6, Munich 81739, Germany

Jinbo Bi JINBO.BI@SIEMENS.COM
Siemens, Medical Solutions, 51 Valley Stream Parkway, Malvern PA 19355, USA

Volker Tresp VOLKER.TRESP@SIEMENS.COM
Siemens, Corporate Technology, Otto-Hahn-Ring 6, Munich 81739, Germany

Abstract

This paper considers the problem of selecting the most informative experiments $\mathbf{x}$ to get measurements y for learning a regression model $y = f(\mathbf{x})$. We propose a novel and simple concept for active learning, *transductive experimental design*, that explores available unmeasured experiments (i.e.,unlabeled data) and has a better scalability in comparison with classic experimental design methods. Our in-depth analysis shows that the new method tends to favor experiments that are on the one side *hard-to-predict* and on the other side *representative* for the rest of the experiments. Efficient optimization of the new design problem is achieved through alternating optimization and sequential greedy search. Extensive experimental results on synthetic problems and three real-world tasks, including questionnaire design for preference learning, active learning for text categorization, and spatial sensor placement, highlight the advantages of the proposed approaches.

1. Introduction

Recent years have seen considerable interests in learning with labeled and unlabeled data (Seeger, 2000), since labels are often expensive to obtain whereas vast amount of unlabeled data are easily available. Semi-supervised learning (Zhou et al., 2004; Zhu, 2005) solves the problem by exploring additional information given by unlabeled data. *Active learning* reduces the labeling costs in a different but complementary way, which chooses the most informative data to label.

There has been a long tradition of research on active learning in the machine learning community. Typically discriminant models prefer to choosing uncertain or hard-to-predict data, and generative models tend to select typical data. Uncertain data can be atypical and even outliers. It is thus essential to unify these two different views. Active learning is also referred to as *experimental design* in statistics (Atkinson & Donev, 1992). In order to learn a predictive function from *experiment-measurements* pairs, experimental design selects the most informative experiments to measure, given that conducting an experiment is expensive.

This paper studies active learning for regression problems in the context of experimental design. We briefly review classic methods, such as A-optimal, D-optimal and E-optimal design methods, and point out their shortcomings, such as insufficient exploration of available unmeasured data and the poor scalability. These drawbacks motivate us to propose a novel and simple concept, *transductive experimental design*. The key idea is to select data points that are the most contributive to predictions on unlabeled test data that are *given beforehand*. We provide insights into the suggested method: it seeks for data points that are hard to predict and meanwhile representative to unexplored test data. By deriving equivalent formulations of the transductive design, we are able to devise tractable optimizing procedures that produce desired performance, and a better scalability than those classical methods.

The paper is organized as follows. We briefly review active learning in Sec. 2.1 and experimental design in Sec. 2.2. In Sec. 3 we introduce the concept of transductive experimental design, and derive solutions in Sec. 4. Finally we empirically evaluate the suggested methods in Sec. 5 and conclude in Sec. 6.

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

2. Related Work

2.1. Active Learning

In machine learning community there has been extensive research on active learning. Existing approaches either select the most uncertain data given previously trained models (Freund et al., 1997), or choose the most informative data that optimize some expected gain (Cohn & Ghahramani, 1996; MacKay, 1992; Chapelle, 2005). The latter typically requires expensive retraining of models when evaluating each candidate. Some other approaches assume generative models and explore the dependency between inputs and outputs (Nigam et al., 2000; Seeger, 2000). Active learning methods for support vector machines (Tong, 2001) and Gaussian processes (Guestrin et al., 2005) have also been suggested.

2.2. Experimental design

Classic experiment design considers learning a linear function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, $\mathbf{w} \in \mathbb{R}^d$, from *measurements* $y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i, i = 1, \dots, m$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is measurement noise, and $\mathbf{x}_1, \dots, \mathbf{x}_m$ are *experiments* chosen from n candidates $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^d$, $n > m$. The goal of experimental design is to find a set of experiments $\mathbf{x}_i$ that together are maximally informative. Following the convention in the machine learning literature, we call experiments $\mathbf{x}$ as *data*, and measurements y as *labels*.

In the rest of this paper, we use $\mathbf{X}$ to represent both the matrix $[\mathbf{x}_1, \dots, \mathbf{x}_m]^\top \in \mathbb{R}^{m \times d}$ and the set $\{\mathbf{x}_i\}$, and $\mathbf{V}$ to represent both $[\mathbf{v}_1, \dots, \mathbf{v}_n]^\top \in \mathbb{R}^{n \times d}$ and the set $\{\mathbf{v}_i\}$. The meanings will be clear in the contexts. $|\mathbf{X}| = m$ and $|\mathbf{V}| = n$ respectively denote the sizes of two sets.

The maximum-likelihood estimate of $\mathbf{w}$ is obtained by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 \right\}, \qquad (1)$$

It is known that the estimation error $\mathbf{e} = \mathbf{w} - \hat{\mathbf{w}}$ has zero mean and a covariance matrix given by $\sigma^2 \mathbf{C}_\mathbf{w}$, where $\mathbf{C}_\mathbf{w}$ is the inverted Hessian of $J(\mathbf{w})$

$$\mathbf{C}_\mathbf{w} = \left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1}, \qquad (2)$$

and σ is a constant. The matrix $\mathbf{C}_\mathbf{w}$ characterizes the confidence of the estimation, or the *informativeness of the selected data*. Let m_j denote the number of times for which $\mathbf{v}_j$ is chosen in $\mathbf{X}$, so we have $m_1 + \cdots + m_n = m$. Then an optimization problem can be formulated as minimization of some measurement of estimation error derived from $\mathbf{C}_\mathbf{w}$. For example, the so-called *A-optimal design* minimizes the trace of $\mathbf{C}_\mathbf{w}$

$$\begin{aligned} \min_{m_1, \dots, m_n} \quad & \mathrm{Tr}\Big[\big(\sum_{j=1}^{n} m_j \mathbf{v}_j \mathbf{v}_j^\top \big)^{-1} \Big] \qquad (3) \\ \text{subject to} \quad & m_j \geq 0, m_1 + \cdots + m_n = m, m_i \in \mathbb{Z} \end{aligned}$$

where $\mathrm{Tr}(\cdot)$ is the trace. To relax the integer constraint $m_j \in \mathbb{Z}$, we set $\tau_j = m_j / m$ and ignore $m\tau_j \in \mathbb{Z}$, then A-optimal design becomes

$$\begin{aligned} \min_{\tau_1, \dots, \tau_n} \quad & \mathrm{Tr}\Big[\big(\sum_{j=1}^{n} \tau_j \mathbf{v}_j \mathbf{v}_j^\top \big)^{-1} \Big] \qquad (4) \\ \text{subject to} \quad & \boldsymbol{\tau} \succeq 0, \mathbf{1}^\top \boldsymbol{\tau} = 1 \end{aligned}$$

where $\boldsymbol{\tau}$ is the vector of τ_j's, and $\mathbf{1}$ a column vector of ones. This has been shown to be a convex semidefinite programming (SDP) problem (Boyd & Vandenberghe, 2004). There exist other two common variants: *D-optimal design* minimizes the logarithm determinant of $\mathbf{C}_\mathbf{w}$ and *E-optimal design* minimizes the 2-norm of $\mathbf{C}_\mathbf{w}$. The selected data are the m data points $\mathbf{v}_i$ associated with the largest weights τ_i. Very recently in the machine learning community a robust E-optimal design method (Flaherty et al., 2006) was applied to biological experiments.

3. Transductive Experimental Design

3.1. Motivations

The classic experimental design methods described in Sec. 2.2 have the following shortcomings.

- The optimization criteria based on $\mathbf{C}_\mathbf{w}$ does not directly characterize the quality of predictions on test data. If the test data are given beforehand, it is more sensible to directly assess the quality of predictions $y = f(\mathbf{x})$ on the test data.

- Standard experimental design only considers linear functions and is thus restrictive in applications.

- Very importantly, classic experimental design has to solve a SDP problem, which is often very slow when dealing with hundreds of data points.

To overcome these problems, this paper proposes experimental design in a *transductive* setting, where the focus is on the predictive performance on known test data, as well as the development of efficient solutions.

3.2. Formulations

A general setting may consider a different set $\mathbf{T}$ of test data points besides candidates in $\mathbf{V}$. Here for simplification we assume that the two sets are the same. In this section we will first focus on linear functions and then generalize it to the nonlinear case by applying *reproducing kernels*. The scalability issue will be addressed in Sec. 4.

Let us consider a regularized linear regression problem

$$\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 + \mu \|\mathbf{w}\|^2 \right\} \tag{5}$$

where $\mu > 0$ and $\|\cdot\|$ is the vector 2-norm. Similar as before, the inverted Hessian is computed as

$$\mathbf{C}_\mathbf{w} = \left(\frac{\partial J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)^{-1} = (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I})^{-1} \tag{6}$$

Compared with Eq. (2), the newly introduced regularization improves numerical stability since $\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I}$ is full-rank. Let $\mathbf{f} = [f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)]^\top$ be the function values on all the available data $\mathbf{V}$, then the predictive error $\mathbf{f} - \hat{\mathbf{f}}$ has the covariance matrix $\sigma^2 \mathbf{C}_\mathbf{f}$ with

$$\begin{aligned}\mathbf{C}_\mathbf{f} &= \mathbf{V}\mathbf{C}_\mathbf{w}\mathbf{V}^\top = \mathbf{V}(\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{V}^\top \\ &= \frac{1}{\mu}\left[\mathbf{V}\mathbf{V}^\top - \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right],\end{aligned} \tag{7}$$

where the Woodbury inversion identity is applied. In contrast to $\mathbf{C}_\mathbf{w}$, $\mathbf{C}_\mathbf{f}$ directly characterizes the quality of predictions on the target data $\mathbf{V}$. The average predictive variance on $\mathbf{V}$ is given by $\frac{\sigma^2}{n}\mathrm{Tr}(\mathbf{C}_\mathbf{f})$. A sensible design objective is to select m data points $\mathbf{X}$ from $\mathbf{V}$ such that a high confidence of predictions on the available test data $\mathbf{V}$ is ensured. Therefore we formulate the transductive experimental design problem as a minimization of the predictive variance on test data $\mathbf{V}$. Since n, μ, σ and $\mathrm{Tr}(\mathbf{V}\mathbf{V}^\top)$ are constants, we define the problem as

Definition 3.1. ***Transductive experimental design:***

$$\begin{aligned}\max_{\mathbf{X}} \quad & \mathit{Tr}\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right] \\ \textit{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m\end{aligned} \tag{8}$$

Since $\mathrm{Tr}(\mathbf{C}_\mathbf{f}) = \mathrm{Tr}(\mathbf{C}_\mathbf{w}\mathbf{V}^\top\mathbf{V})$, the classical A-optimal design can be seen as a subcase of transductive design, however with a restrictive assumption $\mathbf{V}^\top\mathbf{V} \propto \mathbf{I}$.

3.3. Interpretations

The following theorem helps to understand the behaviors of the proposed transductive experimental design.

Theorem 3.2. *Transductive experimental design is equivalent to*

$$\begin{aligned}\min_{\mathbf{X},\mathbf{A}} \quad & \sum_{i=1}^{n} \|\mathbf{v}_i - \mathbf{X}^\top \mathbf{a}_i\|^2 + \mu\|\mathbf{a}_i\|^2 \\ \textit{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m, \\ & \mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times m}\end{aligned} \tag{9}$$

Proof. We rewrite the cost function as $L(\mathbf{X}, \mathbf{A}) = \|\mathbf{V} - \mathbf{A}\mathbf{X}\|_F^2 + \mu\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)$, where $\|\cdot\|_F$ is the Frobenius norm for matrices. Then

$$\begin{aligned}L(\mathbf{X},\mathbf{A}) &= \mathrm{Tr}\left[(\mathbf{V} - \mathbf{A}\mathbf{X})(\mathbf{V} - \mathbf{A}\mathbf{X})^\top\right] + \mu\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top) \\ &= \mathrm{Tr}\left[\mathbf{V}\mathbf{V}^\top - \mathbf{A}\mathbf{X}\mathbf{V}^\top - \mathbf{V}\mathbf{X}^\top\mathbf{A}^\top + \mathbf{A}\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top + \mu\mathbf{A}\mathbf{A}^\top\right] \\ &= \mathrm{Tr}(\mathbf{V}\mathbf{V}^\top) - \mathrm{Tr}\left[\mathbf{A}\mathbf{X}\mathbf{V}^\top + \mathbf{V}\mathbf{X}^\top\mathbf{A}^\top - \mathbf{A}(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})\mathbf{A}^\top\right]\end{aligned}$$

By taking the partial derivatives of $L(\mathbf{X}, \mathbf{A})$ with respect to $\mathbf{A}$, it is easy to see that given $\mathbf{X}$, the optimum of $\mathbf{A}$ to minimize $L(\mathbf{X}, \mathbf{A})$ has the form

$$\mathbf{A}^* = \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}$$

Plugging this result into the loss function, we can get

$$\begin{aligned}&\min \|\mathbf{V} - \mathbf{A}\mathbf{X}\|_F^2 + \mu\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top) \\ &= \mathrm{Tr}(\mathbf{V}\mathbf{V}^\top) - \mathrm{Tr}\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right]\end{aligned}$$

Given $\mathrm{Tr}(\mathbf{V}\mathbf{V}^\top)$ is a constant, the minimization problem with respect to $\mathbf{X}$ becomes the maximization of $\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}$, which completes the proof. □

Theorem 3.2 transforms the problem into a regularized least squares formalism. Interestingly, it demonstrates an equivalence to finding the optimal set of basis vectors $\mathbf{X}$ to approximate the whole set of vectors $\mathbf{V} \equiv \{\mathbf{v}_i\}$ by $\hat{\mathbf{v}}_i = \mathbf{X}^\top \mathbf{a}_i$. Based on the projection theorem of least squares estimator, the approximations can be seen as (regularized) projections of $\mathbf{V}$ onto the linear subspace spanned by $\mathbf{X}$. Therefore, transductive experimental design has a clear geometric interpretation: it tends to find *representative* data samples $\mathbf{X}$ that span a linear space to retain most of the information of $\mathbf{V}$. In contrast, standard experimental design methods do not pursue this property.

On the other hand, the minimization in (9) encourages to particularly "focus on" those $\mathbf{v}_i$ with large norms, or even to directly include them into $\mathbf{X}$. Intuitively, it is hard to obtain stable predictions for those $\mathbf{v}_i$ with big norms, because a small disturbance to $\mathbf{w}$ can cause a big variation of $f(\mathbf{v}_i) = \mathbf{w}^\top \mathbf{v}_i$. Therefore, Theorem 3.2 indicates that the selected $\mathbf{X}$ tends to well represent those *hard-to-predict* test cases in $\mathbf{V}$. Furthermore, in the context of sequential design (see Sec. 4.1), $\mathbf{v}_i$ are actually residuals of data after being approximated by previously selected data, which means that $\mathbf{v}_i$ with a larger norm correspond to data that are under-represented by previously chosen data. Therefore, transductive experimental design tends to select data representative to those yet unexplored data in a sequential design.

Like other experimental design methods, despite the fact that we consider a supervised learning problem, the data selection itself is independent of measurements $\mathbf{y} \equiv \{y_i\}$. The reason is that the least squares cost has only a linear dependency between $\mathbf{w}$ and $\mathbf{y}$, which makes the Hessian of $J(\mathbf{w})$ independent of $\mathbf{y}$. Note that for classification, not the focus of this paper, the situation is different.

3.4. Kernel Transductive Experimental Design

Now we are ready to handle nonlinear functions. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) with a kernel function

$$k(\mathbf{x}, \mathbf{v}) = \langle \phi(\mathbf{x}), \phi(\mathbf{v}) \rangle, \ \mathbf{x}, \mathbf{v} \in \mathbb{R}^d \tag{10}$$

where $\phi : \mathbb{R}^d \rightarrow \mathcal{H}$ is a feature mapping, then $f \in \mathcal{H}$ has the form $f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x})$. Plugging it into (5), we obtain a regularized linear regression in the feature space. It is well-known that the solution has the form $f(\mathbf{x}) = \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x})$, with coefficients $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m]^\top$ estimated via a kernel regression

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^m \Big[\sum_{j=1}^m \alpha_j k(\mathbf{x}_j, \mathbf{x}_i) - y_i \Big]^2 + \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

Let's denote the data in the transformed feature space by $\mathbf{X} = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_m)]^\top$ and $\mathbf{V} = [\phi(\mathbf{v}_1), \ldots, \phi(\mathbf{v}_n)]^\top$, and plug them into (8), we directly obtain the kernelized transductive experimental design

$$\begin{aligned} \max_{\mathbf{X}} \quad & \mathrm{Tr}\left[\mathbf{K}_{\mathbf{v}\mathbf{x}}(\mathbf{K}_{\mathbf{x}\mathbf{x}} + \mu \mathbf{I})^{-1}\mathbf{K}_{\mathbf{x}\mathbf{v}}\right] \\ \text{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m \end{aligned} \tag{11}$$

where $(\mathbf{K})_{ij} = k(\mathbf{v}_i, \mathbf{v}_j)$, $(\mathbf{K}_{\mathbf{v}\mathbf{x}})_{ij} = k(\mathbf{v}_i, \mathbf{x}_j)$ and $(\mathbf{K}_{\mathbf{x}\mathbf{x}})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. In the new kernelized version we can directly work with a kernel function, like RBF kernel, without explicitly referring to the feature mapping $\phi(\cdot)$. $f(\mathbf{x})$ can be nonlinear if a nonlinear kernel is adopted. In the case of linear kernels, the kernel regression and kernel transductive experimental design are equivalent to their counterparts introduced in Sec. 3.2. In the rest of this paper, we will mainly consider the kernel version.

4. Optimization Approaches

Although the transductive design has a simple interpretation, the involved optimization problem is a difficult combinatorial optimization problem, as indicated by the following theorem. We have to resort to tractable approximations.

Theorem 4.1. *Transductive experimental design is NP-hard.*

Proof. Based on theorem 3.2, a special case of the problem is to select $m < n$ basis vectors from n candidates to approximate a *single* vector in the least squares criterion. The case is known as a sparse linear regression problem with a cardinality constraint, which has been proven to be NP-hard in (Natarajan, 1995). The transductive design is NP-hard since it approximates multiple vectors using sparse basis. □

4.1. Sequential Optimization

In this subsection, we develop a very simple sequential greedy optimization approach. We first formulate transductive experimental design as a sequential optimization problem. Given previously selected data $\mathbf{X}_1$, a *sequential* transductive design seeks m new data $\mathbf{X}_2 \subset \mathbf{V}$ in the following way

$$\begin{aligned} \max_{\mathbf{X}_2} \quad & \mathrm{Tr}\left[\mathbf{K}_{\mathbf{v}\mathbf{x}}(\mathbf{K}_{\mathbf{x}\mathbf{x}} + \mu \mathbf{I})^{-1}\mathbf{K}_{\mathbf{x}\mathbf{v}}\right] \\ \text{subject to} \quad & \mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2, \mathbf{X}_2 \subset \mathbf{V}, |\mathbf{X}_2| = m \end{aligned} \tag{12}$$

Problem (12) can be written as a canonical form of transductive experimental design

$$\begin{aligned} \max_{\mathbf{X}_2} \quad & \mathrm{Tr}\left[\tilde{\mathbf{K}}_{\mathbf{v}\mathbf{x}_2}(\tilde{\mathbf{K}}_{\mathbf{x}_2\mathbf{x}_2} + \mu \mathbf{I})^{-1}\tilde{\mathbf{K}}_{\mathbf{x}_2\mathbf{v}}\right] \\ \text{subject to} \quad & \mathbf{X}_2 \subset \mathbf{V}, |\mathbf{X}_2| = m \end{aligned} \tag{13}$$

where the kernel matrix $\tilde{\mathbf{K}}$ is obtained by deflating the original kernel matrix $\mathbf{K}$ by $\mathbf{X}_1$:

$$\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{K}_{\mathbf{v}\mathbf{x}_1}(\mathbf{K}_{\mathbf{x}_1\mathbf{x}_1} + \mu \mathbf{I})^{-1}\mathbf{K}_{\mathbf{x}_1\mathbf{v}} \tag{14}$$

Problem (13) can be understood as a kernel version of the following procedure: after approximating $\mathbf{V}$ by $\mathbf{X}_1$, the approximation residuals $\tilde{\mathbf{V}}$ form a new kernel matrix $\tilde{\mathbf{K}} = \tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top$, and a set of m vectors from $\tilde{\mathbf{V}}$ are selected to further approximate $\tilde{\mathbf{V}}$. As pointed out in Sec. 3.3, the algorithm tends to select data that are typical among those under-represented by $\mathbf{X}_1$.

Since it is very simple to select just one data point, we propose an *easy-to-implement* algorithm that iteratively performs the following two steps until m data points have been selected. Note that there is no need for matrix inverse.

Algorithm 1: Sequential Design

- Select $\mathbf{x} \in \mathbf{V}$ with the highest $\|\mathbf{K_x}\|^2/(k(\mathbf{x},\mathbf{x}) + \mu)$, and add $\mathbf{x}$ into $\mathbf{X}$, where $\mathbf{K_x}$ and $k(\mathbf{x},\mathbf{x})$ are $\mathbf{x}$'s corresponding column and diagonal entry in current $\mathbf{K}$;
- Update $\mathbf{K} \leftarrow \mathbf{K} - \frac{\mathbf{K_x}\mathbf{K_x}^\top}{(k(\mathbf{x},\mathbf{x})+\mu)}$;

4.2. Alternating Optimization

Sequential optimization is a greedy process that can be suboptimal. In this subsection we first transform the problem into an equivalent regression-like formalism, which makes it possible to relax the discrete nature of the problem and then design non-greedy mathematical programming solutions.

Theorem 4.2. *Let* $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_n]^\top$ *and* $\pi_1 \geq \ldots \geq \pi_n$ *be the eigenvectors and eigenvalues of* $\mathbf{K} = \mathbf{V}\mathbf{V}^\top$. *Then transductive experimental design is equivalent to*

$$\min_{\mathbf{X},\mathbf{C}} \quad \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K_{vx}}\mathbf{c}_i\|^2 + \mu\pi_i\|\mathbf{c}_i\|^2 \quad (15)$$

$$\textit{subject to} \quad \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m, \quad \mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_n]^\top \in \mathbb{R}^{n\times m}$$

Proof. Let $\mathbf{V}$ has the singular value decomposition $\mathbf{V} = \mathbf{Q}\mathbf{\Pi}^{1/2}\mathbf{P}^\top$. Then based on Theorem 3.2, given $\mathbf{X}$, at the minimum of $\|\mathbf{V} - \mathbf{A}\mathbf{X}\|_F^2 + \mu\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)$ there are

$$\begin{aligned}\|\mathbf{V} - \mathbf{A}^*\mathbf{X}\|_F^2 &= \|\mathbf{V} - \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\|_F^2 \\ &= \|\mathbf{V}\mathbf{P} - \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{P}\|_F^2 \\ &= \|\mathbf{Q}\mathbf{\Pi}^{1/2} - \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{P}\|_F^2\end{aligned}$$

and

$$\begin{aligned}&\mu\mathrm{Tr}(\mathbf{A}^*\mathbf{A}^{*\top}) \\ &= \mu\mathrm{Tr}\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-2}\mathbf{X}\mathbf{V}^\top\right] \\ &= \mu\mathrm{Tr}\left[\mathbf{Q}\mathbf{\Pi}^{1/2}\mathbf{P}^\top\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-2}\mathbf{X}\mathbf{P}\mathbf{\Pi}^{1/2}\mathbf{Q}^\top\right] \\ &= \mu\mathrm{Tr}\left[\mathbf{\Pi}^{1/2}\mathbf{P}^\top\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-2}\mathbf{X}\mathbf{P}\mathbf{\Pi}^{1/2}\right].\end{aligned}$$

Let $\mathbf{C} = \mathbf{P}^\top\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}$, then the minimum of $\|\mathbf{V} - \mathbf{A}\mathbf{X}\|_F^2 + \mu\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)$ must have the form

$$\|\mathbf{Q}\mathbf{\Pi}^{1/2} - \mathbf{K}_{\mathbf{vx}}^\top\mathbf{C}^\top\|_F^2 + \mu\mathrm{Tr}[\mathbf{C}^\top\mathbf{\Pi}\mathbf{C}].$$

where we have applied $\mathrm{Tr}[\mathbf{\Pi}^{1/2}\mathbf{C}\mathbf{C}^\top\mathbf{\Pi}^{1/2}] = \mathrm{Tr}[\mathbf{C}^\top\mathbf{\Pi}\mathbf{C}]$. Obviously, minimizing the above new cost function with respect to $\mathbf{C} \in \mathbb{R}^{n\times m}$ is a variational formalism of minimization in problem (15) with respect to $\mathbf{w}$. The proof is finished. □

Theorem 4.2 shows that the transductive design is equivalent to choosing m columns in K that can be used to best approximate eigenvectors of K. Due to the weighting by eigenvalues π_i in (15), a better efficiency can be achieved by considering only those leading eigenvectors $\mathbf{q}$ of $\mathbf{K}$. Note that $\mathbf{X}$ is a subset of $\mathbf{V}$ assuming that all available data are given in $\mathbf{V}$ as its rows. Denote a matrix $\mathbf{B}$ as an $n \times n$ diagonal matrix with its j-th diagonal element equal to $\beta_j \in \{0, 1\}$. We call $\mathbf{B}$ an indicator matrix indicating whether or not an according data point will appear in $\mathbf{X}$. If $\beta_j = 1$, $\mathbf{v}_j$ is included in $\mathbf{X}$. Then $\mathbf{K_{vx}}\mathbf{c}_i = \mathbf{K}\mathbf{B}\boldsymbol{\alpha}_i$ where $\boldsymbol{\alpha}_i$ is an n vector with its subset of m components (indicated in $\mathbf{B}$) equal to $\mathbf{c}_i$ correspondingly. Then the transductive design problem is equivalent to the following integer program:

$$\min_{\mathbf{B},\boldsymbol{\alpha}_i} \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \mu\pi_i\|\mathbf{B}\boldsymbol{\alpha}_i\|^2$$

$$\text{subject to } \mathbf{B} = \mathit{diag}(\boldsymbol{\beta}), \quad \mathrm{Card}(\boldsymbol{\beta}) = m, \quad \beta_j \in \{0, 1\}, \quad j = 1, \cdots, n. \quad (16)$$

Problem (16) is often computationally intractable since it requires branch-and-bound procedure to optimize integer variables $\boldsymbol{\beta}$. We relax constraints on integer variables $\boldsymbol{\beta}$ to allow them to take real numbers. Then β_j corresponds to a scaling factor indicating how significantly the corresponding data in $\mathbf{V}$ contributes to the minimization of (16). We then enforce the sparsity of $\boldsymbol{\beta}$. Sparsity can be enforced by employing regularization conditions on $\boldsymbol{\beta}$, such as the 0-norm penalty which controls the cardinality of $\boldsymbol{\beta}$, (notice 0-norm is not really a vector norm, see (Weston et al., 2003)), or the 1-norm penalty which is less stringent than the 0-norm penalty. To derive computationally efficient and scalable formulations, we relax the problem to use 1-norm penalty on $\boldsymbol{\beta}$ instead of restricting its cardinality. Problem (16) becomes

$$\min_{\mathbf{B},\alpha_i} \quad \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \mu\pi_i\|\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \gamma\|\boldsymbol{\beta}\|_1$$

$$\text{subject to} \quad \mathbf{B} = \mathit{diag}(\boldsymbol{\beta}), \quad \beta_j \geq 0, \quad j = 1, \cdots, n. \quad (17)$$

The residual term $\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}\mathbf{B}\boldsymbol{\alpha}_i$ in problem (17) is bilinear with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_i$. Taking the 2-norm of the residual introduces polynomial terms of high order in terms of its variables and thus the problem is still arduous to solve. We propose an alternating optimization approach (Bezdek & Hathaway, 2003) to problem (17) by repeating steps depicted in Algorithm 2, which is similar, in spirit, to the principle of Expectation-Maximization (EM) algorithms. Moreover, note that $\|\boldsymbol{\beta}\|_1 = \sum \beta_j$ due to the nonnegativity of β_j.

Algorithm 2: Alternating Design

- Fix $\mathbf{B}$ to the current solution (initially to the identity matrix $\mathbf{I}$), convert $\tilde{\mathbf{K}} \leftarrow \mathbf{KB}$, solve the following problem for optimal $\boldsymbol{\alpha}_i$,

$$\min_{\boldsymbol{\alpha}_i} \quad \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \tilde{\mathbf{K}}\boldsymbol{\alpha}_i\|^2 + \mu\pi_i\|\mathbf{B}\boldsymbol{\alpha}_i\|^2 \tag{18}$$

- Fix $\boldsymbol{\alpha}_i$ to the solution obtained at the above step, convert $\mathbf{K}_i \leftarrow \mathbf{K} \cdot diag(\boldsymbol{\alpha}_i)$, solve the following problem for optimal $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned}\min_{\boldsymbol{\beta}\geq 0} \quad & \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}_i\boldsymbol{\beta}\|^2 + \mu\pi_i\|\boldsymbol{\beta} \otimes \boldsymbol{\alpha}_i\|^2 \\ & +\gamma\|\boldsymbol{\beta}\|_1\end{aligned} \tag{19}$$

- $\mathbf{B} \leftarrow \mathbf{B} \otimes diag(\hat{\boldsymbol{\beta}})$

where $\otimes$ denotes the component-wise multiplication between two matrices. The algorithm takes a greedy scheme in the third step of the iterations, assuring data samples receiving small scaling factors in early iterations will continue receiving small weights.

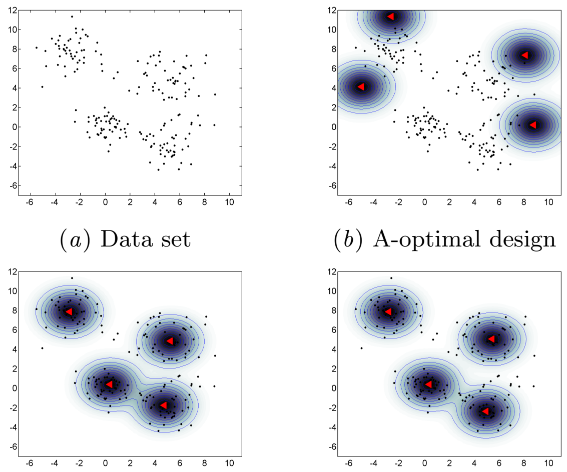
The first step of Algorithm 2 solves a simple ridge regression problem which can be de-coupled to minimize $\|\sqrt{\pi_i}\mathbf{q}_i - \tilde{\mathbf{K}}\boldsymbol{\alpha}_i\|^2 + \mu\pi_i\|\mathbf{B}\boldsymbol{\alpha}_i\|^2$ for each individual $\boldsymbol{\alpha}_i$. Thus, problem (18) actually has a closed-form solution, which is to solve $\mathbf{B}\left(\mathbf{KK} + \mu\pi_i\mathbf{I}\right)\mathbf{B}\boldsymbol{\alpha}_i = \sqrt{\pi_i}\mathbf{BKq}_i$ where the diagonal matrix $\mathbf{B}$ may not be full rank. The solution $\hat{\boldsymbol{\alpha}}_i = \mathbf{B}^{-1}\left(\mathbf{KK} + \mu\pi_i\mathbf{I}\right)^{-1}\mathbf{Kq}_i$ where $\mathbf{B}^{-1}$ denotes the diagonal matrix whose nonzero diagonal elements equal the inverse of nonzero diagonal components of $\mathbf{B}$. Note that the matrix inversion $\left(\mathbf{KK} + \mu\pi_i\mathbf{I}\right)^{-1}$ only needs to be calculated in the first iteration and can then be reused in later iterations, thus gaining computational efficiency.

The second step of Algorithm 2 solves a quadratic programming problem. Denote $\Lambda_i = diag(\boldsymbol{\alpha}_i)$. The problem (19) can be rewritten in the following canonical form of a quadratic program:

$$\begin{aligned}\min_{\boldsymbol{\beta}\geq 0} \quad & \boldsymbol{\beta}^\top \sum_i \left(\Lambda_i(\mathbf{KK} + \mu\pi_i\mathbf{I})\Lambda_i\right)\boldsymbol{\beta} \\ & + \left(\gamma\mathbf{e}^\top - 2\sum_i \sqrt{\pi_i}\mathbf{q}_i^\top\mathbf{K}\Lambda_i\right)\boldsymbol{\beta}.\end{aligned} \tag{20}$$

5. Experiments

In this section we test the kernel transductive experimental design in a number of settings. To the best of our knowledge, there is no kernel A-optimal design in the literature. For a fair comparison to our approach, we first use kernel PCA to map data points into an n-dimensional linear space, and then apply the standard A-optimal design solved by the SeDuMi optimization package[1]. For those methods that compute nonnegative coefficients for each candidate data points, like alternating transductive design and A-optimal design, we choose those m data points having the biggest coefficients. In all the investigated problems, μ is fixed as 0.1.



(a) Data set (b) A-optimal design

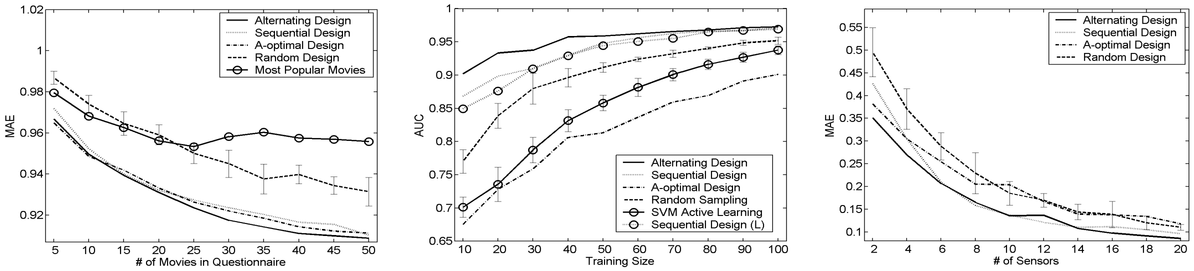
(c) Sequential design (d) Alternating design

Figure 1. Experimental design ($m = 4$) on synthetic I. Selected data are marked by red triangles, gray levels and contours indicate the predictive variance of the learned function in the input space (darker means lower variance).

Synthetic problem I: We generate a mixture of four Gaussian components in a 2-D space, as shown in Fig. 1-(a). An RBF kernel with length scale 1.8 is used. Classical experimental design, such as A-optimal design, attempts to choose data on the border of data set as shown Fig. 1-(b), where the low predictive variance area covers a space without many data samples present. In contrast, as shown in Fig. 1-(c) and (d), the two variants of transductive design both select representative data regarding the whole distribution.

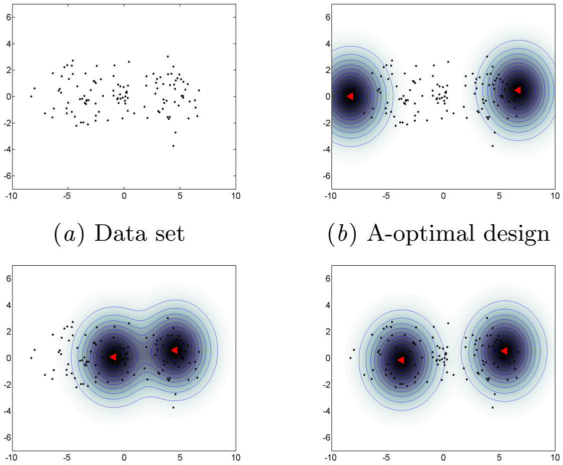
Synthetic problem II: In this case we show that sequential design can sometimes obtain suboptimal solutions while the alternating approach is superior. As shown in Fig. 2-(a), the data set consisted of two major Gaussians in the left and right sides and a minor Gaussian in the middle. An RBF kernel with the length scale 2.5 is applied. The sequential transductive design first picked up a point near the center of the data and then picked another point close to the right border, as shown in Fig. 2-(c). This result is less optimal than that of non-sequential solutions shown in Fig. 2-(d).

[1] http://sedumi.mcmaster.ca/



(a) Questionnaire Design (b) Text Categorization (c) Sensor Placement

Figure 3. Experimental design in various applications



(a) Data set (b) A-optimal design

(c) Sequential design (d) Alternating design

Figure 2. Experimental design ($m = 2$) on synthetic problem II. The sequential transductive design is suboptimal compared with the non-sequential solution.

Questionnaire Design for Recommender Systems: The so-called "cold-start problem" of recommender systems refers to the difficulty of providing accurate recommendations if the system does not know a user's preferences on any products. To solve the problem, the system usually requires the user to rate a set of products. In this experiment we consider questionnaire design to select informative products. Our study is based on the well-known Eachmovie data set, which contains 74,424 users' numerical ratings $\{1, 2, 3, 4, 5, 6\}$ on 1648 movies. We follow the same setting as (Breese et al., 1998), which chose 5000 users' ratings as training data and a different set of 5000 users for test. Then each movie is seen as being represented by a 5000-dimensional feature vector formed by 5000 training users' ratings on it. In forming feature vectors, we estimate each training user's mean rating and use it to centralize this user's ratings. Given a test user's ratings on a set of movies, we apply regularized linear regression (or equivalently, kernel regression with linear kernels) to predict this user's ratings on other unrated movies. Mean absolute error (MAE) is the most common accuracy metric in the literature. We use experimental design methods to choose $m = 5, 10, 15, \ldots, 45, 50$ movies. Since each test user only watched a subset of movies, given a particular questionnaire, some test users may have no ratings on the chosen movies. In this case we use each movie's mean ratings as predictions. To alleviate the sparsity problem, we restrict the question candidates to those 100 most popular movies, regarding to their numbers of received ratings in the training set. MAE is evaluated on test users' ratings on movies outside of the questionnaire. The results are shown in Fig. 3-(a). As a baseline, "Random Design" chooses m movies randomly, repeated by 10 times. The mean and standard deviation are plotted. The second baseline, "Most Popular Movies", chooses the m most frequently rated movies. A bit surprisingly, choosing the most popular movies does not bring advantages over random guessing. This is because that the most popular movies are rated highly by nearly all the users and thus give no information about user tastes. MAE is even increased as m going over 25, because less popular movies and relatively more *hard* movies are left for accuracy evaluation. Very positively, all the experimental design methods outperform the two baselines. In this task transductive design shows results comparable to that of A-optimality.

Text Categorization: We validate experimental design methods on text categorization based on a subset of Newsgroup corpus, which contains 8014 dimensional TFIDF features and 3970 documents, covering four categories *autos*, *motorcycles*, *baseball*, *hockey*. We conduct one-against-all scheme for each category and thus treat the problem as binary classification ($y = \{-1, 1\}$). Due to the unbalance of two classes, AUC score is used to evaluate the accuracy, averaged over the 4 topics. We apply kernel regression with linear kernels, which has shown the state-of-art for text categorization compared to SVMs (Zhang & Yang, 2003). A SVM active learning algorithm described in (Tong, 2001) is also examined, which chooses data closest to the classification boundary. For each run of this method we initialize with a SVM trained on a pair of randomly chosen positive and negative examples. With 10 random initializations, mean and errorbar are computed. As the baseline, random design is also repeated 10 times to produce the errorbars. The results are shown in Fig. 3-(b). Transductive design methods significantly outperform the competitors. For example, AUC based on just 10 selected training examples achieves 90.2%, in contrast to 77.0% with random sampling. Interestingly, SVM active learning and A-optimal design perform much worse than random sampling. This is because that Newsgroup data has a very clear clustering structure (like synthetic problem I). As illustrated in the synthetic problem I, A-optimal design does not explore this structure. SVM active learning tends to select untypical data and thus does not either. Since SDP by SeDuMi affords A-optimal design with up to 400 candidates, we have to restrict the candidates of A-optimality to a random set of 397 documents. To make the comparison fair, we also apply sequential transductive design based on the same candidate set (shown as "Sequential Design (L)") and still produce much better results.

Sensor Placement: The application is to measure indoor temperature based on optimal placement of sensors. The data was previously applied in (Guestrin et al., 2005), consist of snapshots of measurements from 54 sensors in a hall within two days. We apply experimental design to select sensors such that remaining sensors' measurements can be optimally predicted. In this case we employ nonlinear kernels between locations, offered by the authors of (Guestrin et al., 2005). Fig. 3-(c) shows the results measured by MAE, averaged over 10,000 snapshots. Tansductive design generally outperforms random selection. A-optimal design does not show much advantages, largely because all the sensors are not uniformed distributed in the space.

6. Conclusions

In this paper we proposed *transductive experimental design* for active learning of regression models. As a key advantage over classical methods, it fully explores the available unlabeled data and demonstrates sensible data selection properties. Efficient solutions were developed. The achieved experimental results suggest its wide applicability to real-world applications. In the near future it would be interesting to develop a similar idea for classification models.

References

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experiment designs.* Oxford Statistical Science Series. Oxford University Press.

Bezdek, J., & Hathaway, R. (2003). Convergence of alternating optimization. *Neural, Parallel Sci. Comput.*, *11*, 351–368.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization.* Cambridge University Press.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence* (pp. 43–52).

Chapelle, O. (2005). Active learning for Parzen window classifier. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics* (pp. 49–56).

Cohn, D., & Ghahramani, Z. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Flaherty, P., Jordan, M. I., & Arkin, A. P. (2006). Robust design of biological experiments. *NIPS 18.*

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, *28*, 133–168.

Guestrin, C., Krause, A., & Singh, A. (2005). Near-optimal sensor placements in gaussian processes. *Proc. of the International Conference on Machine Learning (ICML).*

MacKay, D. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590–604.

Natarajan, B. (1995). Sparse approximate solutions to linear systems. *SIAM Journal of Computing*, *25*, 227–234.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.

Seeger, M. (2000). *Learning with labeled and unlabeled data* (Technical Report). Edinburgh University.

Tong, S. (2001). *Active learning: Theory and applicaitons.* Doctoral dissertation, Stanford University.

Weston, J., Elisseeff, A., Schölkopf, B., & Tipping, M. (2003). Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, *3*, 1439–1461.

Zhang, J., & Yang, Y. (2003). Robustness of regularized linear classifcation methods in text categorization. *The 26th Annual International SIGIR Conference (SIGIR'99).*

Zhou, D., Bousquet, O., Lal, T., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems 16.* MIT Press.

Zhu, X. (2005). Semi-supervised learning literature survey.