Available online at www.sciencedirect.com
ScienceDirect
Journal of Interactive Marketing 24 (2010) 42–54
JOURNAL OF INTERACTIVE MARKETING
www.elsevier.com/locate/intmar

Semi-Supervised Response Modeling

Hyoung-joo Lee [a], Hyunjung Shin [b], Seong-seob Hwang [c], Sungzoon Cho [c,*] & Douglas MacLachlan [d]

[a] Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, UK
[b] Department of Industrial and Systems Engineering, Ajou University, Suwon, 443-749, Republic of Korea
[c] Department of Industrial Engineering, Seoul National University, Seoul, 151-744, Republic of Korea
[d] Department of Marketing and International Business, University of Washington, Seattle, WA, 98195-3200, USA

Abstract

Response modeling is concerned with identifying potential customers who are likely to purchase a promoted product, based on customers' demographic and behavioral data. Constructing a response model requires a preliminary campaign result database. Customers who responded to the campaign are labeled as *respondents* while those who did not are labeled as *non-respondents*. Those customers who were not chosen for the preliminary campaign do not have labels, and thus are called *unlabeled*. Then, using only those labeled customer data, a classification model is built in the *supervised* learning framework to predict all existing customers. However, often in response modeling, only a small part of customers are labeled, and thus available for model building, while a large number of unlabeled data may give valuable information. As a method to exploit the unlabeled data, we introduce *semi-supervised* learning to the interactive marketing community. A case study on the CoIL Challenge 2000 and the Direct Marketing Educational Foundation data sets shows that the transductive support vector machine, one of widely used semi-supervised models, can identify more respondents than conventional supervised models, especially when a small number of data are labeled. Semi-supervised learning is a viable alternative and merits further investigation.

Keywords: Scoring model; Response modeling; Classification; Semi-supervised learning

Introduction

Response modeling has been an effective strategy for customer relationship management (CRM) campaigns. Its aim is to construct a scoring model that predicts whether or not each customer will respond to a given mailing campaign (Deichmann et al. 2002; Ha, Cho, and MacLachlan 2005; Levin and Zahavi 2001; Zahavi and Levin 1997a). Based on the model, marketers give promotions or offers to customers who are predicted to respond. A well-targeted mailing campaign will increase revenue while a mistargeted or unwanted mailing will not only increase the marketing cost without adding any value but also may worsen the customers' relationship with the firm (Gönül, Kim, and Shi 2000).

A typical modeling procedure can be outlined as follows (Zahavi and Levin 1997a,b). First, a preliminary campaign is launched to a small subset of customers and the responses are observed. We then assign class labels to the customers, that is, *respondents* to those who responded and *non-respondents* to those who did not. Customers who have not been targeted in the preliminary campaign, on the other hand, do not have class labels, and thus are called *unlabeled*. Second, using those labeled customer data with the known class labels, a prediction model is built to classify the customers into respondents and non-respondents. Note that unlabeled customer data are not directly involved in model building. Finally, for each customer, a score is estimated as to whether he or she will respond to a future campaign. An actual mailing campaign is targeted to a subset of customers with high scores. In the machine learning literature, this type of approach is called *supervised* learning[1] in the sense that modeling is guided by the class labels. A model constructed by supervised learning can exploit information only

* Corresponding author.
E-mail address: zoon@snu.ac.kr (S. Cho).

[1] The term, "learning," stems from the analogy that a model is constructed by observing data. Although it is often used in the machine learning literature essentially interchangeably with "model building" we will refrain from using it throughout this paper except for established terminology.

1094-9968/$ - see front matter © 2009 Direct Marketing Educational Foundation, Inc. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.intmar.2009.10.004

from those labeled customers who have been targeted in the preliminary campaign.

Some researchers view this as a suboptimal use of available data since we have the unlabeled customer data as well, which may also give valuable information for building a prediction model. Recently, the so-called *semi-supervised* learning framework has been proposed to exploit unlabeled data in a principled manner (Chapelle, Schölkopf, and Zien 2009; Seeger 2000; Zhu 2005). It has been proposed to deal with situations where labeled data are difficult to collect while unlabeled data are readily available or relatively easy to collect. Previous research has shown that semi-supervised learning using unlabeled data can improve upon typical supervised learning in terms of classification accuracy (Chapelle, Schölkopf, and Zien 2009; Chawla and Karakoulas 2005). We believe that semi-supervised learning is well-suited for response modeling tasks where there is a small amount of labeled data but a large amount of unlabeled data. This is the case when, for example, a direct marketer is testing a new list of customers. However, to the best of our knowledge, no research has been reported that attempts to apply semi-supervised learning to response modeling.

The goal of this paper is to introduce semi-supervised learning to the interactive marketing community. We present a method applicable to response modeling: transductive support vector machines (TSVMs)[2] (Joachims 1999; Sindhwani and Keerthi 2007), a semi-supervised extension of the supervised support vector machines (SVMs) (Cristianini and Shawe-Taylor 2000). A case study is conducted on two well-known data sets: the *CoIL Challenge 2000* data set (CoIL2000) (van der Putten and van Someren 2000) and the *Direct Marketing Educational Foundation* data set 4 (DMEF4). The results show that the TSVM outperforms supervised models when there is a small amount of labeled data.

This paper is organized as follows. In the next section, semi-supervised learning is introduced and discussed in the light of response modeling. The Response models section describes response models that are applied to the case study presented in the Case study section. Finally, the Conclusion section provides conclusions and discusses some issues and future research directions.

Semi-supervised learning

Suppose that we have a data set of n customers. Among them, n_ℓ customers have been targeted in a preliminary campaign and have known class labels. For those *labeled* customers, a class label of +1 or *positive* is assigned to a respondent in the previous campaign, or −1 or *negative* to a non-respondent. The remaining $n_u(=n-n_\ell)$ customers who do not have labels are called *unlabeled*. Without loss of generality, we can denote the labeled set by $\mathcal{L}=\{(X_i,y_i)\}_{i=1}^{n_\ell}$ and the unlabeled set by $\mathcal{U}=\{X_i\}_{i=n_\ell+1}^{n}$. A d-dimensional feature vector, $\mathbf{x}_i\in R^d$, describes demographic information and purchase history of the ith customer, and a class label, $y_i\in\{-1,+1\}$, indicates his or her response. Note that class labels for the unlabeled customers, $\{y_i\}_{i=n_\ell+1}^{n}$, are not available.

The objective of response modeling is to identify customers who are likely to respond. In particular, a model estimates a score, usually in the form of a probability, that each customer will respond, that is, $f(\mathbf{x}_i)=p(y_i=+1|\mathbf{x}_i)\in[0,1]$ for $i=1,..n$. Then, for marketing decisions, a predicted label is assigned to each customer:

$$\hat{y}_i=\begin{cases}+1, & \text{if } f(\mathbf{x}_i)\geq\theta,\\ -1, & \text{if } f(\mathbf{x}_i)<\theta,\end{cases} \tag{1}$$

where a cut-off point, θ, is determined by considering various conditions such as strategic goals, marketing costs, inventory levels, and so on. A marketing campaign would be targeted to a subset of customers whose predicted labels are +1, that is, $\mathcal{S}=\{i|\hat{y}_i=+1\}$.

Response modeling has been considered exclusively in the supervised learning setting, which is illustrated in Fig. 1a. A model is built by estimating the functional relationship between the feature vectors and the class labels in the labeled set, $\mathcal{L}$. Then, the model is applied to all existing customers to obtain their predicted scores. Supervised learning approaches have generally worked well for response modeling or other prediction tasks. However, labeled data are often difficult, expensive, or time-consuming to obtain, as they require actual mailing campaigns. In such a situation, only a small number of labeled data may be available, which may not be sufficient for building an accurate model. One might wonder whether the unlabeled data can be utilized in model building, and if so, how.

Semi-supervised learning can provide an answer for the question (see Fig. 1b). While the true labels of the unlabeled customers are unknown, their feature vectors are readily available. A semi-supervised model is built using the feature vectors of the unlabeled customers, $\mathcal{U}$ as well as the labeled data, $\mathcal{L}$. Of course, supervised learning models can also use the unlabeled customer data for preprocessing such as feature selection/construction and customer segmentation. After that, however, the unlabeled data are not involved in model building at all. On the other hand, semi-supervised learning can exploit the unlabeled data in both preprocessing and model building stages.

While the term "semi-supervised" was first used by Merz, St. Clair, and Bond (1992), inclusion of a small number of unlabeled data in classification has been studied since the 1960s by several researchers (Agrawala 1970; Fralick 1967; Hartley and Rao 1968; McLachlan 1975, 1977; Scudder 1965), who demonstrated the benefit of doing so either analytically or empirically. The first argument for semi-supervised learning with a large number of unlabeled data was made by O'Neill (1978) who stated:

> "Even when there is no control over the ratio of available unclassified and classified data, unclassified observations should certainly not be discarded."

[2] "Transductive" means that one can only deal with data at hand, but cannot handle unseen data. In other words, a modeling procedure should be performed all over again every time new data arrive, such as hierarchical clustering for example. It contrasts with inductive modeling where one obtains a tangible model in a functional form that can handle unseen data. Although most of recent TSVMs are in fact inductive, they are called transductive by convention.

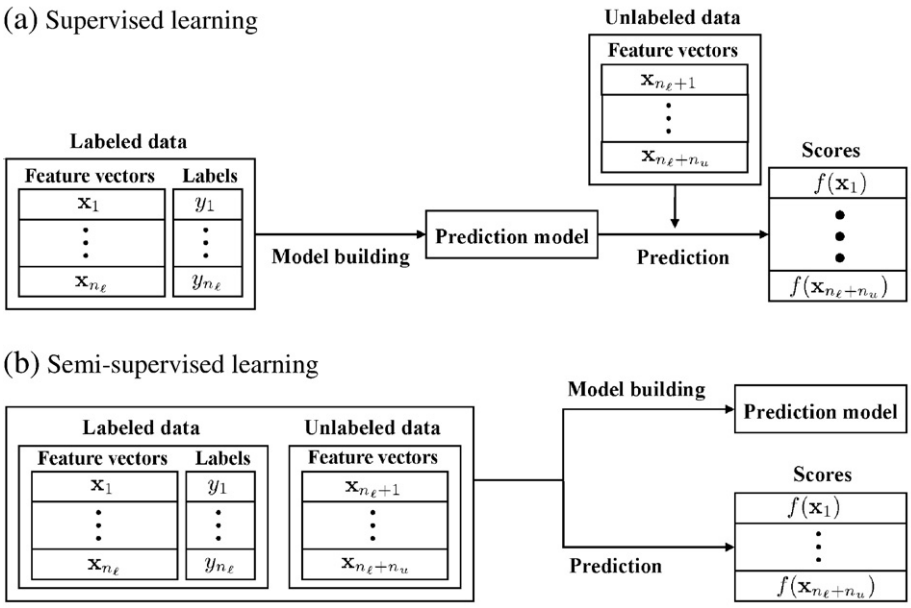

Fig. 1. Outlines of supervised and semi-supervised learning for response modeling.

By the mid-1990s, it had been proven analytically that unlabeled data can reduce misclassification error although they are much less valuable than labeled ones (Castelli and Cover 1995, 1996; Ratsaby and Venkatesh 1995). Though based on different assumptions, they arrived at equivalent conclusions; the classification error is expected to converge to the Bayes optimal error, polynomially with respect to the number of unlabeled data and exponentially with respect to the number of labeled data. Since the late 1990s, semi-supervised learning has gained popularity in the machine learning community, mostly due to the emergence of applications such as text classification (Blum and Mitchell 1998; Joachims 1999; Nigam et al. 2000; Sindhwani and Keerthi 2007) and bioinformatics (Lanckriet et al. 2004; Shin and Tsuda 2006; Shin, Lisewski, and Lichtarge 2007; Weston et al. 2005) where labeled data are far more difficult to collect than unlabeled data. Semi-supervised learning approaches have been shown empirically to outperform supervised ones in various classification tasks.

The difference between supervised and semi-supervised learning can be interpreted from a probabilistic point of view (Seeger 2000; Zhu 2005). The joint probability of a feature vector-label pair, $(\mathbf{x},y)$, can be decomposed as $p(\mathbf{x},y)=p(\mathbf{x})p(y|\mathbf{x})$. Supervised learning focuses only on the term $p(y|\mathbf{x})$. For example, logistic regression estimates $p(y=+1|\mathbf{x})$, but does not take $p(\mathbf{x})$ into account at all. In other words, it is not concerned with how the feature vectors are distributed. This contrasts with unsupervised learning, for example clustering for customer segmentation, in which the objective is essentially to estimate $p(\mathbf{x})$ with no predetermined labels available. On the other hand, a semi-supervised learning model estimates $p(y|\mathbf{x})$ while taking $p(\mathbf{x})$ into account. Thus, semi-supervised learning can be considered to complement supervised learning with additional information from the distribution of the feature vectors.

The benefit of semi-supervised learning can be attributed to incorporating $p(\mathbf{x})$ for *regularization* (Cozman and Cohen 2006; Seeger 2000), which indicates restricting the flexibility of a model to prevent overfitting (Tikhonov and Arsenin 1977). In supervised learning, a small number of labeled data points may lead to overfitting and high variance for estimated model parameters, and in turn lead to high rates misclassification error (Friedman 1997). In semi-supervised learning, it is assumed that similar feature vectors lead to similar labels, which is called the *smoothness* assumption. According to this assumption, a model will avoid putting a decision boundary in regions with high $p(\mathbf{x})$ so that a sudden change of class labels should not arise between similar feature vectors. As a result, one can avoid overfitting since unlabeled data discourage the model from being overly sensitive to irregularities in a small number of labeled cases. In this sense, semi-supervised learning can be considered a regularization technique, like for example, ridge regression (Malthouse 1999) that incorporates a penalty term to constrain flexibility of a model. The difference is that in semi-supervised learning, regularization is done by unlabeled data, while in ridge regression a model is regularized, in a Bayesian sense, by a prior belief on model parameters.

In particular, to improve performance using semi-supervised learning, the following conditions are prerequisite (Chapelle, Schölkopf, and Zien 2009):

- The smoothness assumption: Although the degree of appropriateness of this assumption depends on the data at hand and should be assessed based on domain-specific knowledge, it is reasonable to believe that the assumption holds in most classification problems including response modeling. In fact, without the assumption, most supervised learning models would also fail. A stronger form of

the assumption is the *cluster* assumption that data in a cluster are likely to have the same labels. Therefore, if a data set is believed to consist of well-separated clusters, a semi-supervised learning model is highly likely to improve upon a supervised model. On the contrary, if two classes of data severely overlap or a data set is known to have some discontinuity of labels over the feature space, semi-supervised learning might not be helpful.

- A small amount of labeled data and a large number of unlabeled data: Obviously, a large amount of unlabeled data is needed to make an influence on the resulting model. This is not a critical issue, since unlabeled data are abundant in most response modeling problems. On the other hand, as stated above, with a small amount of labeled data, a supervised learning model is likely to overfit and perform poorly, especially in a high-dimensional space. In this case, unlabeled data would be particularly helpful to avoid overfitting. Conversely, if there is a huge amount of labeled data available, there should be very little room for improvement since supervised learning would perform very well. From the proofs of Castelli and Cover (1995, 1996) and Ratsaby and Venkatesh (1995) that labeled data are far more valuable than unlabeled data, it is clear that improvement of semi-supervised learning diminishes as the amount of labeled data increases. However, it is nearly impossible to determine analytically for a particular problem how much labeled data are needed for unlabeled data to be unnecessary. In the Case study section, we investigate empirically the effect of the volume of the labeled data in response modeling.

One simple example of a semi-supervised model is a so-called "cluster-and-label" approach (Nigam et al. 2000). Consider a finite mixture model (McLachlan and Peel 2000) that estimates the distribution of labeled and unlabeled data by a mixture of K individual distributions, for example a Gaussian mixture model. A probability density of a feature vector $\mathbf{x}$ can be written as follows:

$$p(x) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k), \tag{2}$$

where, for each component k, $p(k)$ is the mixing probability and $p(\mathbf{x}|k)$ is the component-conditional probability density function. This is a fully unsupervised model as the class labels, $_{yi}$'s, are not considered at all. Also consider a supervised model for $p(y|k)$ using the labeled data, for example by simply counting the number of cases from each class assigned to each cluster. If we assume that $\mathbf{x}$ and y are conditionally independent given k, that is, $p(\mathbf{x},y|k)=p(\mathbf{x}|k)p(y|k)$, we have the following model:

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x},y)}{p(\mathbf{x})} = \frac{\sum_k p(k)p(\mathbf{x}|k)p(y|k)}{\sum_k p(k)p(\mathbf{x}|k)} \tag{3}$$

We have already estimated $p(k)$ and $p(\mathbf{x}|k)$ by the unsupervised model in Eq. (2) using both the labeled and unlabeled data, while $p(y|k)$ has been estimated by the supervised model using the labeled data. Thus, the model in Eq. (3) can be considered semi-supervised with supervised and unsupervised concepts incorporated. The model would be accurate when the cluster assumption is satisfied so that the mixture model in Eq. (2) matches the data structure. Moreover, to estimate accurately the parameters of the mixture models, many unlabeled data points are needed. Note, however, that model identification and estimation of Eq. (2) are usually very difficult, especially for a high-dimensional data set. Many semi-supervised learning methods in practice, including one employed in this paper, do not estimate $p(\mathbf{x})$ explicitly.

Fig. 2 illustrates an example of decision boundaries obtained by supervised and semi-supervised learning. Two classes of data were generated by two Gaussian distributions. Unlabeled data are represented by small dots while labeled

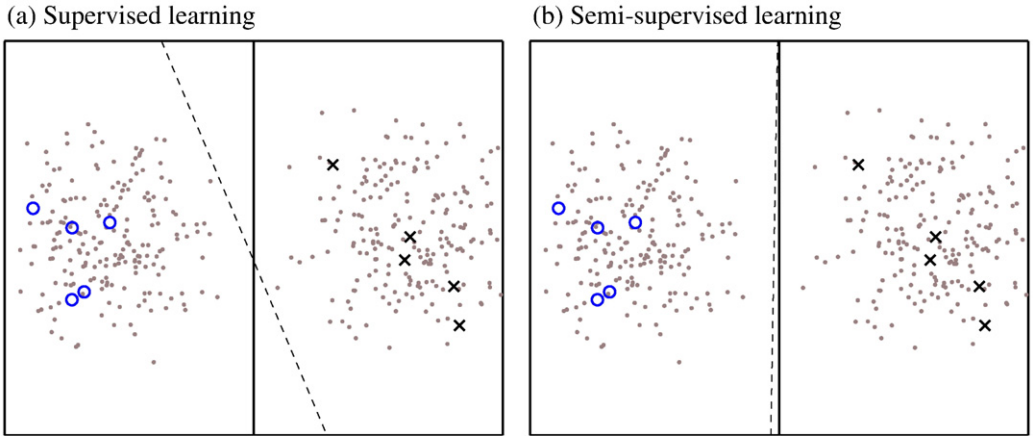

Fig. 2. Comparison of decision boundaries derived from supervised and semi-supervised learning frameworks: 'o' denotes a positively labeled data point, 'x' a negatively labeled one, and '.' an unlabeled one.

data from the two classes are represented by o's and x's, respectively. The thick solid line is the true, that is, Bayes optimal, decision boundary and the thin dashed lines are the decision boundaries obtained using supervised and semi-supervised models. Looking at both the labeled and unlabeled data, it is obvious that the data consist of two clusters and that the decision boundary should lie between the two clusters. A supervised model, unable to exploit the unlabeled data, may result in a decision boundary that differs significantly from the true boundary. A semi-supervised model, on the other hand, by incorporating the unlabeled data, gives a decision boundary much closer to the true boundary.

It is also possible, though very difficult, to apply semi-supervised learning to regression, that is, with continuous output variables. However, semi-supervised learning for regression has not yet been as well established and widely used in practice for classification due to its obvious difficulty. Namely, unlabeled data can have a very large, infinite in theory, possible number of output values while only a small number of discrete labels are considered in classification. For more details, see Lafferty and Wasserman (2007) and the references therein.

Response models

This section briefly describes supervised and semi-supervised response models in our case study. Three supervised models are considered: logistic regression, and linear and radial basis function (RBF) SVMs. As a semi-supervised model, the TSVM is considered.

Logistic regression

Logistic regression is one of the most popular response models and has been used in many papers as a "baseline" model for benchmark purposes, for example Deichmann et al. (2002), Ha, Cho, and MacLachlan(2005), and Zahavi and Levin (1997a). Logistic regression estimates a posterior probability that a customer will respond as follows,

$$f(\mathbf{x}_i) = p(y_i = +1 \,|\, \mathbf{x}_i) = \frac{1}{1 + exp\left[-\left(\beta_0 + \beta^T \mathbf{x}_i\right)\right]}, \tag{4}$$

where β_0 and $\beta=[\beta_1, \ldots\beta_d]^T$ are $d+1$ parameters to be estimated.

Support vector machines (SVMs)

SVMs (Cristianini and Shawe-Taylor 2000) have recently been widely used for various classification tasks including response modeling (Shin and Cho 2006; Viaene et al. 2001). SVMs are concerned with finding a hyperplane that separates two classes in the labeled set with a maximum margin, shown in Fig. 3a as the distance between the two hyperplanes. The following optimization problem is considered:

$$\min \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n_l} \xi_i, \tag{5}$$

subject to $y_i(\boldsymbol{w}^{\mathrm{T}}\Phi(\mathbf{x}_i)+b)\geq 1-\xi_i$, $\xi_i\geq 0, i=1,\ldots,n_l$, where ξ_i is a margin error on $\mathbf{x}_i$ and $\mathbf{w}$ is a vector orthogonal to the hyperplanes. The objective is to not only minimize the error but also maximize the margin between the two classes. By margin maximization, which plays a role of the shrinkage term in ridge regression (Malthouse 1999), SVMs are known to reduce overfitting. A user-defined parameter C controls the trade-off between the margin and the error. A mapping function $\Phi(\cdot)$ projects vectors from the feature space into a kernel space. We do not need to know the mapping explicitly since the data appear only in the form of inner products in the dual form, to which the kernel trick is applied, that is, $\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)=k(\mathbf{x}_i, \mathbf{x}_j)$. In our case study, two different kernel functions are employed:

$$\text{linear kernel} : k\left(X_i, X_j\right) = X_i^T X_j \tag{6}$$

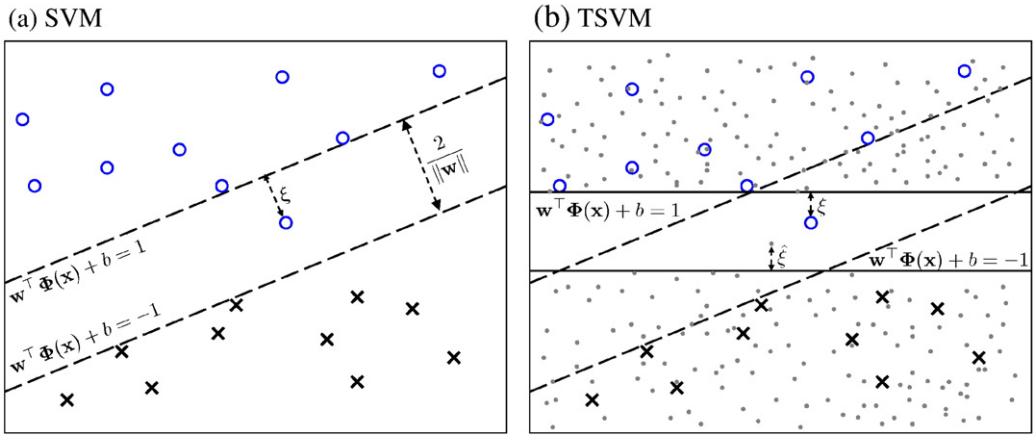

Fig. 3. Margin maximization for an SVM and a TSVM: A margin is the distance between the two separating hyperplanes, that is, $\frac{2}{||w||}$. Margin errors are denoted by ξ for a labeled data point and $\hat{\xi}$ for an unlabeled data point.

$$\text{RBF kernel}: k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right), \quad (7)$$

where σ is a pre-determined kernel width. We will call SVMs with those kernels as linear and RBF SVMs, respectively. A linear SVM generates a linear decision boundary while an RBF SVM generates a more flexible nonlinear boundary.

In the optimal solution, the decision function of a feature vector $\mathbf{x}_i$ is given by

$$g(\mathbf{x}_i) = \mathbf{w}^T \Phi(\mathbf{x}_i) + b = \sum_{\mathbf{x}_j \in L_{SV}} \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i) + b, \quad (8)$$

where $\mathcal{L}_{SV}$ is the set of support vectors and α_j's are the Lagrangian multipliers related to the support vectors. From the decision function, we compute the score of the customer $\mathbf{x}_i$ in a form of a probability as proposed in Platt (1999),

$$f(\mathbf{x}_i) = p(y_i = +1 \,|\, \mathbf{x}_i) = \frac{1}{1 + \exp(-g(\mathbf{x}_i))}. \quad (9)$$

The solution to Eq. (5) can be obtained by quadratic programming techniques. The computational complexity is known to be proportional to the cubic of the number of cases, that is, $O(n_\ell^3)$, which might be prohibitive for very large data sets. Among many efficient approximation methods to overcome this issue, we employ a method proposed in Fan, Chen, and Lin (2005)[3], which solves the problem in $O(n_\ell)$ by decomposing it into many small problems.

Transductive support vector machines (TSVMs)

There are several requirements for a response model, two of which are specifically relevant to semi-supervised learning. First is scalability. A customer database usually contains hundreds of thousands or even millions of customers. While supervised models deal with only labeled data, semi-supervised models deal with a possibly larger amount of unlabeled data as well. Second, a model should be inductive rather than transductive. Suppose that we have built a model, $\mathcal{M}$, based on a data set, $\mathcal{D} = \{\mathcal{L}, \mathcal{U}\}$. We should be able to estimate the scores of an additional set of unlabeled data, $\mathcal{U}'$, without building a new model, $\mathcal{M}'$, from scratch. Many popular semi-supervised methods, for example graph-based methods (Shin and Tsuda 2006), fail to satisfy those two requirements. We present a method that is both scalable and inductive: the TSVM. Later in this section, we will clarify how the TSVM is inductive.

The TSVM (Joachims 1999) was proposed to exploit unlabeled data, as an extension of the standard supervised SVMs. As mentioned in the section on Semi-supervised learning, the smoothness assumption entails the condition that the decision boundary should avoid high-density regions. In the case of SVMs, it is equivalent to avoiding putting data, labeled or unlabeled, within the margin. In Fig. 3b, the dashed lines are the maximum margin boundary resulting from an SVM that ignores the unlabeled data. With the unlabeled data, a TSVM puts the decision boundary (the solid lines) at the maximum margin for both the labeled and unlabeled data.

We describe a formulation given in Sindhwani and Keerthi (2007) as follows:

$$\min \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{2n_\ell}\sum_{i=1}^{n_\ell} \xi_i^2 + \frac{\lambda^*}{2n_u}\sum_{i=n_\ell+1}^{n} \hat{\xi}_i^2, \quad (10)$$

$$\text{subject to } y_i(\mathbf{w}^T\Phi(\mathbf{x}_i)+b) \geq 1-\xi_i, \xi_i \geq 0, i = 1, \ldots, n_\ell,$$
$$\hat{y}_i(\mathbf{w}^T\Phi(\mathbf{x}_i)+b) \geq 1-\hat{\xi}_i, \hat{\xi}_i \geq 0, i = n_\ell, \ldots n,$$
$$\frac{1}{n_u}\sum_{i=n_\ell+1}^{n} \max\left[0, \text{sign}\left(\mathbf{w}^T\Phi(\mathbf{x}_i) + b\right)\right] = r,$$

where $\hat{y}_i$ and $\hat{\xi}_i$ are the predicted label and the predicted margin error for an unlabeled sample, $\mathbf{x}_i$, respectively. Aside from a few notational differences, this formulation is equivalent to incorporating Eq. (5) with the additional terms relating to the unlabeled data. Both the known labels, y_i's, and the predicted labels, $\hat{y}_i$'s, appear in this optimization problem since we need to consider both the labeled and the unlabeled data. In particular, the predicted labels are included in the second set of constraints to discourage a model from having the unlabeled data within the margin. The last term in the objective function plays a role of minimizing the predicted margin error on the unlabeled data. The two user-defined parameters, λ and λ^*, determine the relative importance of the margin and the errors on the labeled and unlabeled data, respectively. The last constraint determines a fraction, r, of the unlabeled data to be classified positive, which is to alleviate the class imbalance problem. Usually, r is set to the response rate in the labeled data.

As the linear kernel is employed in this paper, the TSVM can be considered a direct extension of the linear SVM. The decision function is essentially the same as Eq. (8), except that the set of support vectors now contains some unlabeled data and their predicted labels as well:

$$g(\mathbf{x}_i) = \mathbf{w}^T\Phi(\mathbf{x}_i) + b = \sum_{\mathbf{x}_j \in L_{SV}} \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i) + \sum_{\mathbf{x}_j \in U_{SV}} \alpha_j \hat{y}_j k(\mathbf{x}_j, \mathbf{x}_i) + b, \quad (11)$$

where $\mathcal{L}_{sv}$ and $\mathcal{U}_{sv}$ are the sets of support vectors for the labeled and the unlabeled data, respectively. The scores are computed identically to Eq. (9). Afterward, if we are given an additional set of unlabeled data, $\mathcal{U}'$, we have two choices: induction or transduction. Firstly, we can apply the decision function, Eq. (11), directly to those new data. Alternatively, we can go through another procedure of semi-supervised modeling by formulating and solving a new optimization problem involving $\mathcal{D}' = \{\mathcal{L}, \mathcal{U}, \mathcal{U}'\}$. The decision on which procedure to choose depends on the data at hand, time and resources available, and so on.

The most critical concern for the TSVM is its computational complexity. Finding the exact solution to Eq. (10) is NP-hard. Much attention has been paid to developing efficient

[3] Its software implementation, LIBSVM, is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

approximation methods. This paper employs the linear TSVM based on the modified Newton method proposed in Sindhwani and Keerthi (2007)[4], which has been shown to be both accurate and efficient. For example, it was successfully applied to problems with more than 800,000 data.

Case study

This section presents a case study conducted on two data sets: the CoIL2000 and the DMEF4, for which four types of response models are built. We briefly describe the data sets. After introducing a few criteria to evaluate and compare their performance, we show that the semi-supervised models perform well on the data sets.

Data sets

The CoIL2000 data set[5] was used in the CoIL Challenge competition in 2000 (van der Putten and van Someren 2000) and has been analyzed by a number of researchers (Elkan 2001; Kim et al. 2005). It is concerned with a real world business problem of an insurance company, which would like to predict which customers are potentially interested in a caravan insurance policy. Each customer is represented by an 85-dimensional feature vector including demographic and product usage information. Originally, the labeled set contains 5822 customers whose class labels are known while the unlabeled set contains 4000 customers. The response rates for both sets are roughly 6% with 348 and 238 respondents, respectively.

In the competition, the goal was to identify 800 most-likely-to-purchase customers out of the 4000 unlabeled customers. In this way, however, we can evaluate modeling performance on only one realization of data split. As performance of a model may show a large variation with regard to a specific data split (Malthouse 2001), 30 different splits were generated and used for evaluation. All results reported in this paper were averaged over the 30 realizations of data splits. For each split, we randomly sampled 4,000 customers who would be treated as unlabeled. To investigate the effect of the amount of labeled data, n_l, labeled data were further sampled from the remaining 5,822 customers so that they account for 5, 10, 20, 30, 40, and 50% of total data. In other words, with the number of unlabeled data fixed at 4,000, we varied the numbers of labeled data to 211, 444, 1,000, 1,714, 2,667, and 4,000. Feature selection/construction is not our main concern although it is very important. We used a set of features with which the best performance in the competition was reported (Elkan 2001). These features, as listed in Table 1, are related to car ownership, wealthiness, and propensity to spend on insurance coverage. We also applied a log transformation to features with highly skewed distributions.

The DMEF4 data set[6] involving an up-scale catalog mailing task has been analyzed by various researchers (Ha, Cho, and MacLachlan 2005; Kim, Lee, and Cho 2008; Lee and Cho 2007; Malthouse 2001, 2002; Shin and Cho 2006). The original problem is to predict the amount that each customer will spend during the test period, from September 1992 to December 1992. It is formulated into a classification problem where a class label is +1 for a respondent who spent a non-zero amount and −1 for a non-respondent who did not spend at all. The data set contains 101,532 customers, each of whom is described by 91 features. The response rate is 9.4% with 9571 respondents and 91,961 non-respondents.

Table 1
Eight features used for the CoIL2000 data set.

Name	Description
MKOOPKLA	Purchasing power class
PPERSAUT	Contribution car policies
PBRAND	Contribution fire policies
AWAPART	Number of private third party insurance
APERSAUT	Number of car policies
ABRAND	Number of fire policies
APLEZIER	Number of boat policies
ABYSTAND	Number of social security insurance policies

Preprocessing was performed similarly as with the CoIL2000 data set. For each split, we sampled 50,000 unlabeled data, and labeled data corresponding to the portions, p_l, of 0.5, 1, 5, 10, 20, 30, 40, and 50%, which correspond to the numbers of labeled data, n_l, of 251, 505, 2,632, 5,556, 12,500, 21,429, 33,333, and 50,000, respectively. We used 17 features, some original and some derived, reported in Malthouse (2002), as listed in Table 2. The function $I(\cdot)$ for Tran51–Tran55 is an indicator function that returns 1 if the condition is satisfied or 0 otherwise. A log transformation was also applied to features with highly skewed distributions.

Experimental settings

We employed three supervised and one semi-supervised models. The former includes logistic regression, and linear and RBF SVMs, while the latter is the TSVM. In an actual response modeling situation, scores are estimated for all customers, labeled and unlabeled. In our case study, however, predictions are made only on unlabeled customers for performance evaluation purposes. The supervised models were constructed using the feature vector-label pairs only of the labeled customers and then the models were applied to the feature vectors of the unlabeled customers to estimate their scores. On the other hand, the TSVM estimated the scores of the unlabeled customers using the feature vector-label pairs of the labeled customers and the feature vectors of the unlabeled customers. Note that, for both types of models, the class labels of the unlabeled customers had been assumed to be unknown until they were evaluated.

The models were evaluated in terms of how well they performed on the unlabeled data. For the CoIL2000 data set, we used, as a performance measure, the number of actual respondents among the subset of the top 800 customers for

[4] Its software implementation, SVMlin, is available at http://people.cs.uchicago.edu/~vikass/svmlin.html.
[5] Available at http://www.liacs.nl/~putten/library/cc2000/.
[6] Available at http://www.directworks.org/Educators/Default.aspx?id=632.

Table 2
17 features used for the DMEF4 data set.

Name	Description
Original features	
Purseas	Number of seasons with a purchase
Falord	LTD fall orders
Ordtyr	Number of orders this year
Puryear	Number of years with a purchase
Sprord	LTD spring orders

Name	Formulation	Description
Derived variables		
Recency		Order days since 10/1992
Comb2	$\Sigma_{m=1}^{14}\text{ProdGrp}_m$	Number of product groups purchased from this year
Tran25	1/(1+lorditm)	Inverse of latest-season items
Tran38	1/recency	
Tran42	log(1+ordtyr × falord)	Interaction between the number of orders
Tran44	$\sqrt{\text{ordhist} \times \text{sprord}}$	Interaction between LTD orders and LTD spring orders
Tran46	$\sqrt{\text{comb2}}$	
Tran51	$I(0 \le \text{recency} < 90)$	
Tran52	$I(90 \le \text{recency} < 180)$	
Tran53	$I(180 \le \text{recency} < 270)$	
Tran54	$I(270 \le \text{recency} < 366)$	
Tran55	$I(366 \le \text{recency} < 730)$	

the sake of consistency with the original task. A more accurate model would contain more respondents in the subset. On the other hand, such a canonical performance measure does not exist for the DMEF4 data set. Moreover, to compare the response models fairly and avoid effects of a particular cut-off point, we used the area under receiver operating characteristics (AUROC) (Bradley 1997). An AUROC is computed by integrating true positive rates over false positive rates on an ROC curve, as indicated by the shaded area in Fig. 4. It is equivalent to an expected true positive rate when randomly selecting a false positive rate. Naturally, the more accurate a model, the larger its AUROC. Random guessing, represented by the straight line between (0, 0) and (100, 100), would produce an AUROC value of 50%. Analysis in terms of AUROC gives us a robust estimate of performance since various cut-off points are "integrated out" into the value. Also, we compared the response models using lift charts, generated by plotting a series of lift values, that is, response rates for selected subsets divided by a response rate for the whole population.

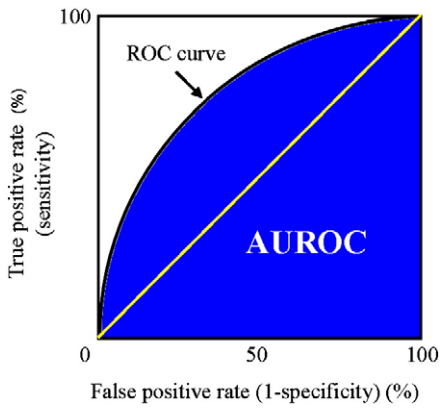

Fig. 4. An example of an ROC curve and its AUROC.

One of the most obvious difficulties in response modeling is that the class distribution is severely unbalanced, that is, $n_+ << n_-$ where n_+ and n_- are the numbers of respondents and non-respondents in the labeled set, respectively. Among the models employed, only the TSVM can handle class imbalance by adopting the last constraint in Eq. (10). For logistic regression, we balanced the labeled data by subsampling n_+ customers out of n_- non-respondents so that the response rate in the newly sampled labeled set should be 50%. For SVMs, we assigned a larger cost to the respondent class to ensure that the class is not neglected. The objective function in Eq. (5) is now modified as follows:

$$\min \frac{1}{2}||\mathbf{w}||^2 + \frac{n_-}{n_+} C \sum_{\{i|y_i=+1\}} \xi_i + C \sum_{\{i|y_i=-1\}} \xi_i \qquad (12)$$

All the models except logistic regression require a set of user-defined parameters to be optimized. One should pre-specify the trade-off parameter, C in Eq. (5), for both SVMs while additionally for the RBF SVM, the kernel width, σ in Eq. (7), should be determined in advance. For the TSVM, the two parameters, λ and λ^* in Eq. (10), should be specified. All of these parameters were chosen for each split of labeled data by five-fold cross-validation (Efron and Gong 1983), as illustrated in Fig. 5. At first, the labeled data are partitioned into five groups of approximately equal sizes. A model is constructed using four groups and then used to score the remaining group, which is for a moment assumed to be unlabeled. This procedure is repeated for all five possible choices of the hold-out groups, indicated by the filled blocks. In this manner, we can estimate performance using only the labeled data. For each model, this cross-validation procedure goes through various possible sets of parameters, among which the one with the highest AUROC is selected. Table 3 lists the optimal sets of parameters obtained by cross-validation. Note that slightly different sets of parameters were selected for different data splits and that the most frequently selected parameters are listed as a guideline.

Response modeling performance

Fig. 6a shows the numbers of actual respondents among top 800 customers identified by the models, averaged over 30

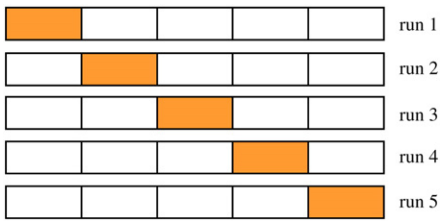

Fig. 5. Five-fold cross-validation.

Table 3
Sets of parameters selected by cross-validation.

p_ℓ (%)	CoIL2000 data set			DMEF4 data set		
	Linear SVM	RBF SVM	TSVM	Linear SVM	RBF SVM	TSVM
	C	C,σ	λ,λ^*	C	C,σ	λ, λ^*
0.5				10^{-2}	1, 2^3	10^{-2}, 10^{-4}
1				10^{-2}	1, 2^3	10^{-2}, 10^{-1}
5	1	10^{-1}, 2^2	10, 10	10^{-1}	1, 2^3	10^{-4}, 10^{-6}
10	1	10^{-2}, 2^2	10, 10	10^{-1}	1, 2^3	10^{-4}, 10^{-6}
20	10^{-1}	10^{-2}, 2^2	1, 1	10^{-2}	10^2, 2^3	10^{-4}, 10^{-6}
30	1	10^{-1}, 2	10^{-2}, 10^{-2}	10^{-1}	10^2, 2^2	10^{-4}, 10^{-6}
40	1	10^{-1}, 2	10^{-2}, 10^{-2}	10^{-2}	10^2, 2^2	10^{-4}, 10^{-6}
50	1	10^{-1}, 2^2	10^{-2}, 10^{-2}	10^{-1}	10^2, 2^2	10^{-4}, 10^{-6}

different data splits of the CoIL2000 data set. The supervised models, as expected, performed better in general as the number of labeled cases increased. On the other hand, TSVM improved until p_ℓ was 20% and then more or less stabilized. It was better than any of the supervised models for all numbers of labeled data. In particular, their differences were larger for smaller numbers of labeled data cases. For example, with p_ℓ=5%, there were only 198 non-respondents and 13 respondents among the 211 labeled data cases. The supervised models suffered from the lack of labeled data, especially respondents. On the other hand, TSVM could overcome it by exploiting 4000 feature vectors of unlabeled data. Furthermore, only with p_ℓ=5%, TSVM captured as many respondents as logistic regression did with p_ℓ=50%. The Kolmogorov–Smirnov test was carried out since the distributions of the results were clearly non-Gaussian. TSVM identified more respondents than any supervised model for all numbers of labeled data with a significance level of 5%.

The cut-off point of 800 may be arbitrary. To evaluate their robustness regardless of a particular choice of a cut-off point, we compared them in terms of the AUROC, as shown in Fig. 6b. The general trends here are similar to those in Fig. 6a. The supervised models performed poorly for small amounts of labeled data but improved as the number increased. TSVM was clearly the best model again in terms of the AUROC. The differences were statistically significant for all cases except p_ℓ=50% where TSVM and logistic regression resulted in essentially equivalent AUROCs. Furthermore, the AUROC of TSVM with p_ℓ=10% was no worse than that of any other model with any portion of labeled data. It should be also noted that TSVM can be considered a direct extension of linear SVM since our TSVM model employed the linear kernel. TSVM outperformed linear SVM by an almost constant difference, suggesting that the unlabeled data contributed to the improvement consistently. In addition, a linear model might be sufficient for this data set since RBF SVM, a nonlinear model, was not substantially better than the linear models.

Fig. 7 shows the lift charts for the numbers of labeled data with 5≤p_ℓ≤30%. In the top row, the advantages of TSVM are even more noticeable. TSVM identified much more of the respondents than any of the supervised models for mailing depths of 50% or less. Performance at lower mailing depths is particularly important since a marketer would rather focus on those customers with high scores. For example, for p_ℓ=5%, TSVM identified about 50% more respondents than the best supervised model, RBF SVM, for the 1% mailing depth. As TSVM, as shown in Fig. 6, did not improve much as the number of labeled data increased, its advantage over the other models gradually diminished although it still continued to be the best model.

The AUROCs for the DMEF4 data set are depicted in Fig. 8. TSVM was clearly the best with statistical significance, for p_ℓ=20%, which is equivalent to n_ℓ≤12,500. Similarly to the CoIL2000 data set, the supervised models struggled when there was only a small amount of labeled data. The supervised models managed to perform as well as TSVM for p_ℓ=30%, where the labeled set contained roughly 21,500 customers. In comparison, TSVM with p_ℓ=5%, with only about 2,600 customers, resulted in high AUROCs, which was comparable to or better than any supervised model with any number of labeled data. Also for this data set, improvement of TSVM over linear SVM was very consistent. Similarly to the CoIL2000 data set, RBF SVM was not much better than the

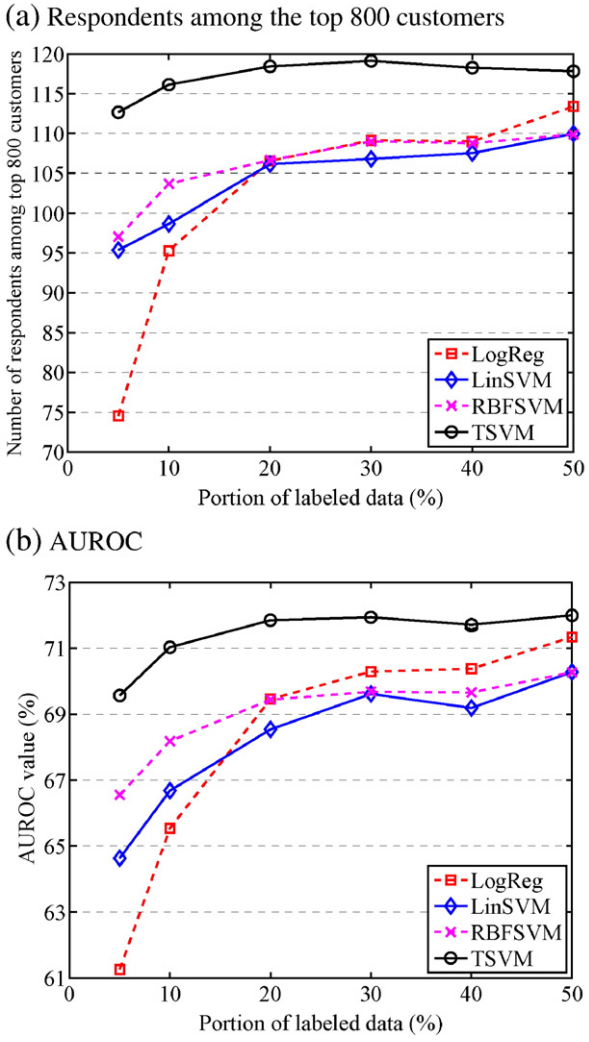

Fig. 6. Performance for the CoIL2000 data set.

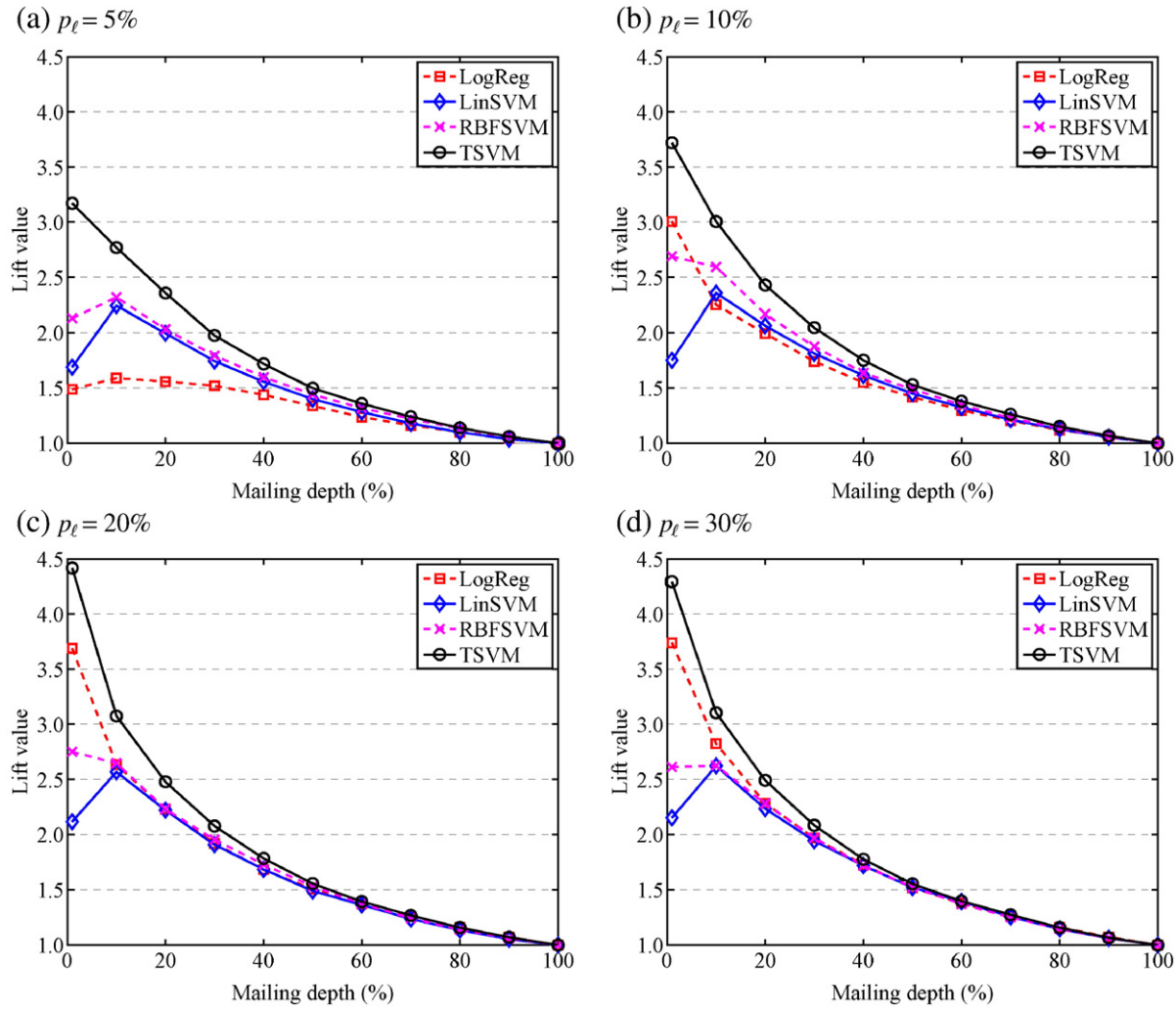

Fig. 7. Lift charts for the CoIL2000 data set with respect to the number of labeled data.

linear models, implying that the data set may have a linear structure.

Fig. 9 shows the lift charts for $0.5 \leq p_l \leq 10\%$. Again, TSVM identified more respondents than the supervised models for lower mailing depths. As the proportion of labeled data increased, however, the supervised models gradually caught up with TSVM. Still, its performance was respectable. For instance, with p_l of 5 and 10%, it identified about 100 and 50 more respondents than linear SVM for the mailing depth of 10%, respectively.

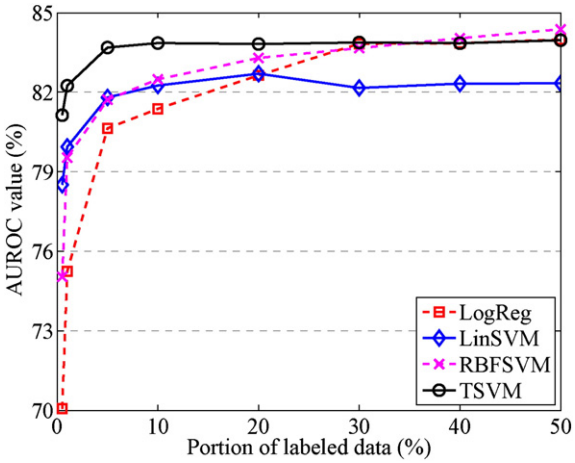

Fig. 8. Average AUROC values for the DMEF4 data set.

In summary, our semi-supervised model, TSVM, can improve upon the supervised models, especially when only a small number of labeled cases are available. For the CoIL2000 data set, it was better than the supervised models for any amount of labeled data, from 200 to 4000. For the DMEF4 data set, it outperformed the supervised models until roughly 20,000 labeled cases became available. The comparison between TSVM and linear SVM suggested that the improvement was causal, not just coincidental. Moreover, TSVM built with a small number of labeled cases can produce results comparable to the supervised models with a much larger number of labeled data. Even when there were plenty of labeled data, $n_l = 50{,}000$ for example, TSVM was no worse than other models. We would argue that using TSVM in any situation would be harmless. The lift charts showed that TSVM identified many more respondents, in particular when a small number of customers were

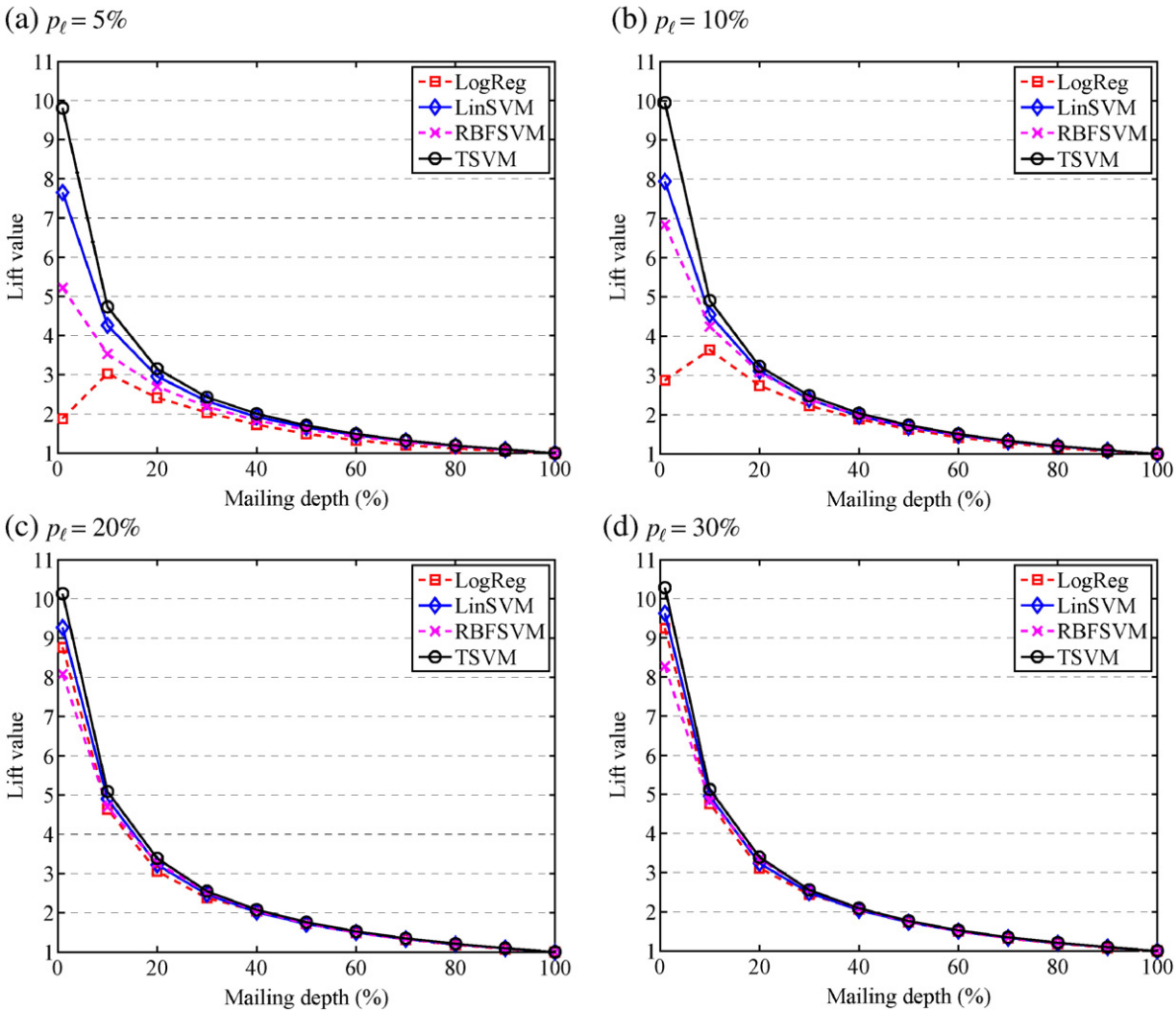

Fig. 9. Lift charts for the DMEF4 data set with respect to the number of labeled data.

assumed to be contacted, which is important from a marketing point of view.

Conclusion

This paper introduced semi-supervised learning and presented a method applicable to response modeling. The TSVM, an efficient and inductive semi-supervised model, exploits unlabeled data as well as labeled data by maximizing the margin between two classes for both types of data. Our case study on the CoIL2000 and the DMEF4 data sets showed that semi-supervised learning should be considered as a viable option for response modeling, especially when a small number of labeled cases are available. From the results on the two data sets, we would argue that the TSVM can provide performance gain over the supervised learning models when there are less than 20,000 labeled cases available. In addition, TSVMs resulted in performance competitive to the supervised models that were built using a much larger number of labeled cases. One might save cost and time for a preliminary mailing campaign to collect labeled data by employing TSVMs instead of logistic regression for example.

Several future research directions need to be addressed. First of all, more empirical studies need to be conducted on various response modeling and other scoring tasks where there might also be a lack of labeled data. Scoring models may have other applications in marketing such as churn modeling (Jamal and Bucklin 2006; Lemmens and Croux 2006; Neslin et al. 2006), personalization (Bucklin and Sismeiro 2009; Montgomery and Smith 2009; Murray and Häubl, 2009), estimation of customer lifetime value and customer equity (Blattberg, Malthouse, and Neslin 2009; Fader and Hardie 2009; Gupta 2009; Jamal and Zhang 2009; Kumar et al. 2009), and managing communications and promotions for multichannel retailers (Neslin and Shankar 2009). Second, a comparative study involving various semi-supervised learning methods (Chapelle, Schölkopf, and Zien 2009; Zhu 2005) can be considered. Some might be more suited for response modeling than the TSVM. One obstacle is that many of them are computationally inefficient. Much research has been concentrated on reducing time and memory requirements. Third, we assumed in this paper that labeled and unlabeled data are drawn from the same distribution: the labels missing completely at random (MCAR) case. In practice, unlabeled cases may be systematically different from labeled

ones: the labels missing not at random (MNAR) case. If labeled data have been obtained from a previous campaign targeted to customers who were more likely to respond, the labeled data can be systematically different from the rest of the data. Semi-supervised learning algorithms need to be expanded to deal with MNAR cases (Chawla and Karakoulas 2005). Fourth, in this paper, and in many other papers dealing with response modeling, the underlying relationships between feature vector-label pairs are assumed to remain static. Hence, results from a past campaign are used for building a model for a future campaign. However, that is not always the case (Malthouse and Derenthal 2008; Zahavi and Levin 1997b). Many events could happen between the two campaigns that could change the relationships. Moreover, a different marketing strategy may be adopted for the future campaign. Thus the past results may not necessarily account for the situations at the time of the future campaign. Finally, we did not consider feature selection/construction. However, feature selection schemes can improve performance of response models as different approaches may require different sets of features.

Acknowledgments

H. Shin would like to gratefully acknowledge support from Post Brain Korea 21 and the research grant from Korean Government (MOEHRD, KRF-2008-531-D00032). S. Hwang and S. Cho were supported by the Basic Research Program of the Korea Science and Engineering Foundation (R01-2005-000-103900-0), the Brain Korea 21 program in 2007, and Engineering Research Institute of Seoul National University. We are especially grateful to the editor and an anonymous reviewer for their valuable comments and suggestions.

References

Agrawala, Ashok K. (1970), "Learning with a Probabilistic Teacher," *IEEE Transactions on Information Theory*, 16, 4, 373–9.

Blattberg, Robert, Edward Malthouse, and Scott Neslin (2009), "Lifetime Value: Empirical Generalizations and Some Conceptual Questions," *Journal of Interactive Marketing*, 23, 2, 157–68.

Blum, Avrim and Tom Mitchell (1998), "Combining Labeled and Unlabeled Data with Co-Training," Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT 1998). p. 92–100.

Bradley, Andrew P. (1997), "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition*, 30, 7, 1145–59.

Bucklin, Randolph E. and Catarina Sismeiro (2009), "Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing," *Journal of Interactive Marketing*, 23, 1, 35–48.

Castelli, Vittorio and Thomas M. Cover (1995), "On the Exponential Value of Labeled Samples," *Pattern Recognition Letters*, 16, 1, 105–11.

——— and Thomas M. Cover (1996), "The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter," *IEEE Transactions on Information Theory*, 42, 6, 2102–17.

Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien (2009), *Semi-supervised Learning*. MIT Press.

Chawla, Nitesh V. and Grigoris Karakoulas (2005), "Learning from Labeled and Unlabeled Data: An Empirical Study Across Techniques and Domains," *Journal of Artificial Intelligence Research*, 23, 331–66.

Cozman, Fabio and Ira Cohen (2006), "Risks of Semi-Supervised Learning," *Semi-supervised Learning*. chapter 4. MIT Press. p. 57–71.

Cristianini, Nello and John Shawe-Taylor (2000), *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

Deichmann, Joel, Abdolreza Eshghi, Dominique Haughton, Selin Sayek, and Nicholas Teebagy (2002), "Application of Multiple Adaptive Regression Splines (MARS) in Direct Response Modeling," *Journal of Interactive Marketing*, 16, 4, 15–27.

Efron, Bradley and Gail Gong (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, 37, 1, 36–48.

Elkan, Charles (2001), "Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000," Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001). p. 426–31.

Fader, Peter S. and Bruce G.S. Hardie (2009), "Probability Models for Customer-Base Analysis," *Journal of Interactive Marketing*, 23, 1, 61–9.

Fan, Rong-En, Pai-Hsuen Chen, and Chih-Jen Lin (2005), "Working Set Selection Using Second Order Information for Training Support Vector Machines," *Journal of Machine Learning Research*, 6, 1889–918.

Fralick, Stanley C. (1967), "Learning to Recognize Patterns Without a Teacher," *IEEE Transactions on Information Theory*, 13, 1, 57–64.

Friedman, Jerome H. (1997), "On Bias, Variance, 0/1 Loss, and the Curse of Dimensionality," *Data Mining and Knowledge Discovery*, 1, 1, 55–77.

Gönül, Füsun F., Byung-Do Kim, and Mengze Shi (2000), "Mailing Smarter to Catalog Customers," *Journal of Interactive Marketing*, 14, 2, 2–16.

Gupta, Sunil (2009), "Customer-Based Valuation," *Journal of Interactive Marketing*, 23, 2, 169–78.

Ha, Kyoungnam, Sungzoon Cho, and Douglas MacLachlan (2005), "Response Models Based on Bagging Neural Networks," *Journal of Interactive Marketing*, 19, 1, 17–30.

Hartley, Herman Otto and Jon N.K. Rao (1968), "Classification and Estimation in Analysis of Variance Problems," *Review of the International Statistical Institute*, 36, 2, 141–7.

Jamal, Zainab and Randolph E. Bucklin (2006), "Improving the Diagnosis and Prediction of Customer Churn: a Heterogeneous Hazard Modeling Approach," *Journal of Interactive Marketing*, 20, 3-4, 16–29.

——— and Alex Zhang (2009), "DMEF Customer Lifetime Value Modeling (Task 2)," *Journal of Interactive Marketing*, 23, 3, 279–83.

Joachims, Thorsten (1999), "Transductive Inference for Text Classification Using Support Vector Machines," Proceedings of the 16th International Conference on Machine Learning (ICML 1999). p. 200–9.

Kim, Dongil, Hyoung-joo Lee, and Sungzoon Cho (2008), "Response Modeling with Support Vector Regression," *Expert Systems with Applications*, 34, 2, 1102–8.

Kim, YongSeog, W. Nick Street, Gary J. Russell, and Filippo Menczer (2005), "Customer Targeting: a Neural Network Approach Guided by Genetic Algorithms," *Management Science*, 51, 2, 264–76.

Kumar, V., Ilaria Dalla Pozza, J. Andrew Petersen, and Denish Shah (2009), "Reversing the Logic: the Path to Profitability Through Relationship Marketing," *Journal of Interactive Marketing*, 23, 2, 147–56.

Lafferty, John and Larry Wasserman (2007), "Statistical Analysis of Semi-Supervised Regression," *Advances in Neural Information Processing Systems*, 20, 801–8.

Lanckriet, Gert R.G., Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William Stafford Noble (2004), "A Statistical Framework for Genomic Data Fusion," *Bioinformatics*, 20, 16, 2626–35.

Lee, Hyoung-joo and Sungzoon Cho (2007), "Focusing on Non-Respondents: Response Modeling with Novelty Detectors," *Expert Systems with Applications*, 33, 2, 522–30.

Lemmens, Aurélie and Christophe Croux (2006), "Bagging and Boosting Classification Trees to Predict Churn," *Journal of Marketing Research*, 43, 2, 276–86.

Levin, Nissan and Jacob Zahavi (2001), "Predictive Modeling Using Segmentation," *Journal of Interactive Marketing*, 15, 2, 2–22.

Malthouse, Edward C. (1999), "Ridge Regression and Direct Marketing Scoring Models," *Journal of Interactive Marketing*, 13, 4, 10–23.

——— (2001), "Assessing the Performance of Direct Marketing Scoring Models," *Journal of Interactive Marketing*, 15, 1, 49–62.

——— (2002), "Performance-Based Variable Selection for Scoring Models," *Journal of Interactive Marketing*, 16, 4, 37–50.

——— and Kirstin M. Derenthal (2008), "Improving Predictive Scoring Models Through Model Aggregation," *Journal of Interactive Marketing*, 22, 3, 51–68.

McLachlan, Geoffrey J. (1975), "Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis," *Journal of the American Statistical Association*, 70, 350, 365–9.

——— (1977), "Estimating the Linear Discriminant Function from Initial Samples Containing a Small Number of Unclassified Observations," *Journal of the American Statistical Association*, 72, 358, 403–6.

——— and David Peel (2000), *Finite Mixture Models*. Wiley-Interscience.

Merz, Christopher J., Daniel C. St. Clair, and William E. Bond (1992), "SeMi-Supervised Adaptive Resonance Theory (SMART2)," *Proceedings of International Joint Conference on Neural Networks (IJCNN*, 1992, 3, 851–6.

Montgomery, Alan and Michael Smith (2009), "Prospects for Personalization on the Internet," *Journal of Interactive Marketing*, 23, 2, 130–7.

Murray, Kyle B. and Gerald Häubl (2009), "Personalization without Interrogation: Towards More Effective Interactions between Consumers and Feature-Based Recommendation Agents," *Journal of Interactive Marketing*, 23, 2, 138–46.

Neslin, Scott A. and Venkatesh Shankar (2009), "Key Issues in Multichannel Management: Current Knowledge and Future Directions," *Journal of Interactive Marketing*, 23, 1, 70–81.

———, Sunil Gupta, Wagner Kamakura, Lu Junxiang, and Charlotte H. Mason (2006), "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research*, 43, 2, 204–11.

Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell (2000), "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, 39, 2-3, 103–34.

O'Neill, Terence J. (1978), "Normal Discrimination with Unclassified Observations," *Journal of the American Statistical Association*, 73, 364, 821–6.

Platt, John C. (1999), "Probabilities for SV Machines," Advances in Large Margin Classifiers. chapter 5. MIT Press. p. 61–74.

Ratsaby, Joel and Santosh S. Venkatesh (1995), "Learning from a Mixture of Labeled and Unlabeled Examples with Parametric Side Information," Proceedings of the 8th Annual Conference On Computational Learning Theory (COLT 1995). p. 412–7.

Scudder, Henry J. (1965), "Probability of Error of Some Adaptive Pattern Recognition Machines," *IEEE Transactions on Information Theory*, 11, 3, 363–71.

Seeger, Matthias (2000), Learning with Labeled and Unlabeled Data. Technical Report, University of Edinburgh. Available at http://www.kyb.mpg.de/bs/people/seeger/papers/review.pdf.

Shin, Hyunjung and Sungzoon Cho (2006), "Response Modeling with Support Vector Machines," *Expert Systems with Applications*, 30, 4, 746–60.

——— and Koji Tsuda (2006), "Prediction of Protein Function from Networks," Semi-supervised learning. chapter 20. MIT Press. p. 361–75.

———, Andreas Martin Lisewski, and Olivier Lichtarge (2007), "Graph Sharpening Plus Graph Integration: a Synergy that Improves Protein Functional Classification," *Bioinformatics*, 23, 23, 3217–24.

Sindhwani, Vikas and S. Sathiya Keerthi (2007), "Newton Methods for Fast Solution of Semi-Supervised Linear SVMs," Large Scale Kernel Machines. chapter 7. MIT Press. p. 155–74.

Tikhonov, Andrey Nikolayevich and Vasilii Yakovlevich Arsenin (1977), *Solutions of Ill Posed Problems*. Wiley.

van der Putten, Peter and Maarten van Someren (2000), Coil Challenge 2000: the Insurance Company Case. TeReport, chnical Leiden Institute of Advanced Computer Science. Available at http://www.liacs.nl/~putten/library/cc2000/PUTTEN~1.pdf.

Viaene, Stijn, Bart Baesens, Tony Van Gestel, Johan A.K. Suykens, Dirk Van den Poel, Jan Vanthienen, Bart De Moor, and Guido Dedene (2001), "Knowledge Discovery in a Direct Marketing Case using Least Squares Support Vector Machines," *International Journal of Intelligent Systems*, 16, 9, 1023–36.

Weston, Jason, Christina Leslie, Eugene Ie, Dengyong Zhou, Andre Elisseeff, and William Stafford Noble (2005), "Semi-Supervised Protein Classification using Cluster Kernels," *Bioinformatics*, 21, 15, 3241–7.

Zahavi, Jacob and Nissan Levin (1997a), "Issues and Problems in Applying Neural Computing to Target Marketing," *Journal of Direct Marketing*, 11, 4, 63–75.

——— and Nissan Levin (1997b), "Applying Neural Computing to Target Marketing," *Journal of Direct Marketing*, 11, 4, 76–93.

Zhu, Xiaojin (2005), Semi-supervised Learning with Graphs. PhD thesis, Carnegie Mellon University. Available at http://pages.cs.wisc.edu/~jerryzhu/pub/thesis.pdf.