ACTIVE LEARNING FOR SEMI-SUPERVISED MULTI-TASK LEARNING

Hui Li[1], *Xuejun Liao*[2] *and Lawrence Carin* [1,2]

[1]Signal Innovations Group, Inc., Durham, NC, USA
[2]Department of ECE, Duke University, Durham, NC, USA

ABSTRACT

We present an algorithm for active learning (adaptive selection of training data) within the context of semi-supervised multi-task classifier design. The semi-supervised multi-task classifier exploits manifold information provided by the unlabeled data, while also leveraging relevant information across multiple data sets. The active-learning component defines which data would be most informative to classifier design if the associated labels are acquired. The framework is demonstrated through application to a real landmine detection problem.

Index Terms— Active learning, semi-supervised learning, multi-task learning, graph, logistic regression

1. INTRODUCTION

Supervised learning has proven to be an effective technique when sufficient and appropriate labeled data are available. Unfortunately, sufficient labeled data are often not available, particularly when label acquisition is expensive. However, in practice one typically has performed many previous classification "tasks" in the past. If data (labeled and unlabeled) from previous tasks can be shared for a new task, the classification performance for the new task may be improved even in the face of limited labeled data. However, not all of the previous tasks may be related to the new task, and the technical challenge involves inferring the inter-relationships between the multiple data sets, such that the sharing of data across multiple tasks is performed appropriately. This problem is often termed multi-task learning.

Algorithm training based only on labeled data is referred to as supervised learning, while learning based only on unlabeled data is termed unsupervised learning. The concept of integrating all available data, labeled and unlabeled, when training a classifier is typically referred to as semi-supervised learning. Semi-supervised learning [1, 2, 3, 4] and multi-task learning (MTL) [5, 6, 7, 8] have been investigated separately by many researchers. In [9] these two techniques were integrated.

However, the labeled data in [9] were selected randomly. In many sensing applications the acquisition of unlabeled data is relatively inexpensive, while acquiring labels on a subset of the data may be expensive ($e.g.$, requiring a human analyst or near-range sensors). Active learning is a framework that uses information-theoretic measures to define those data that are most informative for labeling [10, 11, 12, 13], allowing one to optimize labeling expenses. In this paper we integrate active learning with semi-supervised MTL, building upon the algorithm in [9].

Below we provide a brief review of the semi-supervised MTL framework, and then demonstrate how it may be extended to perform active acquisition of the labeled data. Example results are provided for a real sensing example.

2. SEMI-SUPERVISED MODEL

Let $G = (\mathcal{X}, \mathbf{W})$ be a weighted graph such that $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is a set of vertices that coincide with the data points in a data manifold, and $\mathbf{W} = [w_{ij}]_{n\times n}$ is the affinity matrix with the (i, j)-th element w_{ij} indicating the immediate affinity between data points $\mathbf{x}_i$ and $\mathbf{x}_j$. Following [2, 3], w_{ij} is defined as $w_{ij} = \exp(-0.5 \|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma_i^2)$, where $\|\cdot\|$ is the Euclidean norm and $\sigma_i > 0$ (techniques for defining σ_i are discussed in [9]).

A Markov random walk on graph $G = (\mathcal{X}, \mathbf{W})$ is characterized by a matrix of one-step transition probabilities $\mathbf{A} = [a_{ij}]_{n\times n}$, where a_{ij} is the probability of transiting from $\mathbf{x}_i$ to $\mathbf{x}_j$ via a single step and is given by $a_{ij} = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}$ [14]. Let $\mathbf{B} = [b_{ij}]_{n\times n} = \mathbf{A}^t$; then (i, j)-th element b_{ij} represents the probability of transiting from $\mathbf{x}_i$ to $\mathbf{x}_j$ in t steps. Note that both the labeled and unlabeled data are used to define the random-walk matrix (no labels are used above), and therefore the model is semi-supervised.

Let $p^*(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ be a base classifier parameterized by $\boldsymbol{\theta}$, which gives the probability of class label y_i of data point $\mathbf{x}_i$. The base classifier can be implemented by any parameterized probabilistic classifier. For binary classification with $y \in \{-1, 1\}$, the base classifier can be chosen as a logistic regression

$$p^*(y_i = 1|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + \exp[-y_i\boldsymbol{\theta}^T\mathbf{x}_i]} \qquad (1)$$

where a constant element 1 is assumed to be prefixed to each $\mathbf{x}$(the prefixed $\mathbf{x}$ is still denoted as $\mathbf{x}$ for notational simplicity), and thus the first element in $\boldsymbol{\theta}$ is a bias term.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

Let $p(y_i|\mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta})$ denote a neighborhood-based classifier parameterized by $\boldsymbol{\theta}$, representing the probability of class label y_i for $\mathbf{x}_i$, given the neighborhood of $\mathbf{x}_i$. The proposed semi-supervised classifier is defined as a mixture

$$p(y_i|\mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta}) = \textstyle\sum_{j=1}^{n} b_{ij}\, p^*(y_i|\mathbf{x}_j, \boldsymbol{\theta}) \tag{2}$$

Let $\mathcal{L} \subseteq \{1, 2, \cdots, n\}$ denote the index set of labeled data in $\mathcal{X}$. The neighborhood-conditioned likelihood function is written as

$$\begin{aligned} & p\big(\{y_i, i \in \mathcal{L}\}|\{\mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\}, \boldsymbol{\theta}\big) \\ & = \prod_{i\in\mathcal{L}} p(y_i|\mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta}) = \prod_{i\in\mathcal{L}} \sum_{j=1}^{n} b_{ij}\, p^*(y_i|\mathbf{x}_j, \boldsymbol{\theta}) \end{aligned} \tag{3}$$

3. THE SEMI-SUPERVISED MTL FRAMEWORK

Suppose we are given M tasks, defined by M partially labeled data sets

$$\mathcal{D}_m = \{\mathbf{x}_i^m : i = 1, 2, \cdots, n_m\} \cup \{y_i^m : i \in \mathcal{L}_m\}$$

for $m = 1, \cdots, M$, where y_i^m is the class label of $\mathbf{x}_i^m$ and $\mathcal{L}_m \subset \{1, 2, \cdots, n_m\}$ is the index set of labeled data in task m. We consider M semi-supervised classifiers, parameterized by $\boldsymbol{\theta}_m$, $m = 1, \cdots, M$, with $\boldsymbol{\theta}_m$ responsible for task m. The M classifiers are coupled by a prior joint distribution over their parameters

$$p(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M) = \textstyle\prod_{m=1}^{M} p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{m-1}) \tag{4}$$

with the conditional distributions in the product defined by

$$\begin{aligned} & p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{m-1}) \\ & = \tfrac{1}{\alpha+m-1}\big[\alpha p(\boldsymbol{\theta}_m|\Upsilon) + \textstyle\sum_{l=1}^{m-1} N(\boldsymbol{\theta}_m; \boldsymbol{\theta}_l, \eta^2\mathbf{I})\big] \end{aligned} \tag{5}$$

where $\alpha > 0$, $p(\boldsymbol{\theta}_m|\Upsilon)$ is a base distribution parameterized by Υ, $N(\,\cdot\,; \boldsymbol{\theta}_l, \eta^2\mathbf{I})$ is a normal distribution with mean $\boldsymbol{\theta}_l$ and covariance matrix $\eta^2\mathbf{I}$.

Each normal distribution represents the prior transferred from a previous task; it is the meta-knowledge indicating how the present task should be learned, based on the experience from a previous task. It is through these normal distributions that information sharing between tasks is enforced.

The base distribution represents the baseline prior, which is exclusively used when there are no previous tasks available, as is seen from (5) by setting $m = 1$. When there are $m - 1$ previous tasks, one uses the baseline prior with probability $\frac{\alpha}{\alpha+m-1}$, and uses the prior transferred from each of the $m - 1$ previous tasks with probability $\frac{1}{\alpha+m-1}$. The α balances the baseline prior and the priors imposed by previous tasks. The role of baseline prior decreases as m increases, which is in agreement with our intuition, since the information from previous tasks increase with m. This model is a simplified version of the Dirichlet process [15].

Assuming that, given $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$, the class labels of different tasks are conditionally independent, the joint likelihood function over all tasks can be written as

$$\begin{aligned} & p\big(\{y_i^m, i \in \mathcal{L}_m\}_{m=1}^M|\{\mathcal{N}_t(\mathbf{x}_i^m) : i \in \mathcal{L}_m\}_{m=1}^M, \{\boldsymbol{\theta}_m\}_{m=1}^M\big) \\ & = \textstyle\prod_{m=1}^{M} \prod_{i\in\mathcal{L}_m} \sum_{j=1}^{n_m} b_{ij}^m\, p^*(y_i^m|\mathbf{x}_j^m, \boldsymbol{\theta}_m) \end{aligned} \tag{6}$$

where the m-th term in the product is taken from (3), with the superscript m indicating the task index. Note that the neighborhoods are built for each task independently of other tasks, thus a random walk is always restricted to the same task (the one where the starting data point belongs) and can never traverse multiple tasks. From (4), (5), and (6), the logarithm of the joint *posterior* of $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$ can be written as

$$\begin{aligned} & \ell_{\text{MAP}}(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M) \\ & = \ln p\big(\{\boldsymbol{\theta}_m\}_{m=1}^M|\{y_i^m, i\in\mathcal{L}_m\}_{m=1}^M, \{\mathcal{N}_t(\mathbf{x}_i^m) : i\in\mathcal{L}_m\}_{m=1}^M\big) \\ & = \textstyle\sum_{m=1}^{M} \big\{ \ln\big[\alpha p(\boldsymbol{\theta}_m|\Upsilon) + \sum_{l=1}^{m-1} N(\boldsymbol{\theta}_m; \boldsymbol{\theta}_l, \eta^2\mathbf{I})\big] \\ & \qquad + \textstyle\sum_{i\in\mathcal{L}_m} \ln \sum_{j=1}^{n_m} b_{ij}^m p^*(y_i^m|\mathbf{x}_j^m, \boldsymbol{\theta}_m) \big\} \end{aligned} \tag{7}$$

The parameters $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$ are sought to maximize the log-posterior, which is equivalent to simultaneously maximizing the prior in (4) and the likelihood function in (6).

4. ACTIVE LEARNING

We take an information-theoretic approach to identifying the data locations at which the labels would be most informative to the classifier parameters. Our approach is based on use of Fisher information [12, 16], which is related to previous uses of active learning [11, 10] as applied to purely supervised single task learning models. The Fisher information involves the log-likelihood; as a result the prior is excluded from the calculation. Since the tasks are connected through the prior, this implies that calculation of Fisher information can be performed for each individual task separately (*not* independently though, since the true parameters are replaced by their most recent estimates, as seen below, which are coupled by the prior). Therefore, we drop each variable's dependence on task index m, for notational simplicity. The data log-likelihood is obtained by taking the logarithm of (3),

$$\begin{aligned} \ell(\boldsymbol{\theta}) & \overset{Def.}{=} \ln p\big(\{y_i, i \in \mathcal{L}\}|\{\mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\}, \boldsymbol{\theta}\big) \\ & = \sum_{i\in\mathcal{L}} \ln \sum_{j=1}^{n} b_{ij}\, p^*(y_i|\mathbf{x}_j, \boldsymbol{\theta}) \end{aligned} \tag{8}$$

where the base classifier is assumed as above to be a logistic-regression classifier. By definition [16], the Fisher information matrix (FIM) for the data likelihood is

$$\begin{aligned} & FIM\,\big\{p\big(\{y_i, i \in \mathcal{L}\}|\{\mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\}, \boldsymbol{\theta}\big)\big\} \\ & = \mathbb{E}_{\{y_i\}_{i\in\mathcal{L}}} \left[\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \left[\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in\mathcal{L}} \mathbb{E}_{y_i}\left[\frac{\sum_{j=1}^{n} b_{ij} p^*(y_j = y_i|\mathbf{x}_j,\boldsymbol{\theta}) p^*(y_j = -y_i|\mathbf{x}_j,\boldsymbol{\theta}) y_i \mathbf{x}_j}{\sum_{k=1}^{n} b_{ik} p^*(y_k = y_i|\mathbf{x}_k,\boldsymbol{\theta})}\right] \\
&\quad \times \left[\frac{\sum_{j=1}^{n} b_{ij} p^*(y_j = y_i|\mathbf{x}_j,\boldsymbol{\theta}) p^*(y_j = -y_i|\mathbf{x}_j,\boldsymbol{\theta}) y_i \mathbf{x}_j}{\sum_{k=1}^{n} b_{ik} p^*(y_k = y_i|\mathbf{x}_k,\boldsymbol{\theta})}\right]^T \\
&= \sum_{i\in\mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i} \qquad (9)
\end{aligned}
$$

where

$$
\mathbf{z}_i \overset{Def.}{=} \sum_{j=1}^{n} b_{ij} p^*(y_j = 1|\mathbf{x}_j,\boldsymbol{\theta}) p^*(y_j = -1|\mathbf{x}_j,\boldsymbol{\theta}) \mathbf{x}_j \qquad (10)
$$

$$
\gamma_i \overset{Def.}{=} \sum_{k=1}^{n} b_{ik} p^*(y_j = 1|\mathbf{x}_k,\boldsymbol{\theta}) \sum_{k=1}^{n} b_{ik} p^*(y_j = -1|\mathbf{x}_k,\boldsymbol{\theta}) \qquad (11)
$$

Assume that $\mathbf{x}_* \in \{\mathbf{x}_1,\cdots,\mathbf{x}_n\}\backslash\mathcal{L}$ is a newly labeled sample and added to the labeled set, so $\mathcal{L}$ changes to $\widetilde{\mathcal{L}}$. The Fisher information matrix changes to

$$
\begin{aligned}
&FIM\left\{p\big(\{y_i, i\in\widetilde{\mathcal{L}}\}|\{\mathcal{N}_t(\mathbf{x}_i) : i\in\widetilde{\mathcal{L}}\},\boldsymbol{\theta}\big)\right\} \\
&= \sum_{i\in\mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i} + \frac{\mathbf{z}_* \mathbf{z}_*^T}{\gamma_*} \qquad (12)
\end{aligned}
$$

The ratio of the determinants (one of several possible measures [12], related to the entropy under a Gaussian approximation for the posterior for the model parameters) of the old and new FIMs is

$$
\begin{aligned}
\psi(\mathbf{x}_j) &= \frac{\det\left[\sum_{i\in\mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i} + \frac{\mathbf{z}_* \mathbf{z}_*^T}{\gamma_*}\right]}{\det\left[\sum_{i\in\mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i}\right]} \\
&= \left(1 + \frac{1}{\gamma_*}\mathbf{z}_*^T \left[\sum_{i\in\mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i}\right]^{-1} \mathbf{z}_*\right) \qquad (13)
\end{aligned}
$$

which we employ as our selection criterion in identifying the most informative data sample for labeling. The criterion $\psi(\mathbf{x}_j)$ is calculated for all $\mathbf{x}_j \in \{\mathbf{x}_1,\cdots,\mathbf{x}_n\} \setminus \mathcal{L}$, and the one with the maximum is the most informative data location to obtain a label. Formally, the active semi-supervised MTL procedure is as follows.

- Initially select few label data for each task;
- Compute the information gain $\psi(\mathbf{x}_j)$ through (13);
- Search the next optimal sample;
- Re-train the parameters for all tasks on semi-supervised MTL framework;
- Go to next task if $\psi(\mathbf{x}_j)$ goes to very small.

The true value of $\boldsymbol{\theta}$ required in calculating $\mathbf{z}$ and γ is replaced with the most recent update of the parameters, following the strategy taken in [10, 12]. To the best of our knowledge, this is the first use of active learning in an MTL setting, and we are also considering a semi-supervised model.

5. EXPERIMENTAL RESULTS

We consider a remote sensing problem based on data collected from real landmine fields[1]. In this problem there are a total of 19 sets of data, collected from various landmine fields (with inert landmine simulants). Each data point is represented by a 9-dimensional feature vector extracted from synthetic aperture radar images. Since this is a detection problem, the class labels are binary, with 1 indicating landmine and 0 indicating clutter (false alarm).

As opposed to the setting in [9], where it is assumed that labeled data from the 19 data sets are available simultaneously, we here assume the much more realistic case for which labeled data are acquired sequentially within one data set (task) at one time. Once the process of label acquisition in a given environment is completed, that environment is not revisited to acquire new labeled data.

Each of the 19 data sets defines a task, in which we aim to find landmines with a minimum number of false alarms. Of the 19 data sets, 1-10 are collected at foliated regions and 11-19 are collected at regions that are bare earth or desert. We expect fewer new labeled data when considering a new task for which environmental conditions stay unchanged from previous tasks (but this is *inferred* by the algorithm, and not imposed by the user).

In the experiment both labeled and unlabeled data are used in training the algorithm. After training, the algorithm is tested on the unlabeled data to calculate the area under ROC curve (AUC) for each data set. We compare the active-learning results with AUC results obtained using random selection of labeled data. For the case where the labeled data are randomly selected, we perform 20 independent trials, and compute the mean as well as error bars of AUC from the trials. Since the data sets are acquired sequentially, the results are presented as AUC as a function of the number of tasks from which labeled data are acquired (the ordering of the tasks is arbitrary; the task order considered here was selected as to make a point on the number of labels actively acquired, as discussed further below).

We observe from the results in Figure 1 that active learning performs much better than random selection for a small number of data sets (tasks). As discussed below, the total number of labels used in random selection of labels is the same as that used for active learning. When the number of tasks increases, the benefit of active learning diminishes since the scarcity of labeled data is overcome via multi-task learning.

In Figure 2 we plot the number of labeled data for each task, as a function of task index. For the active-learning algorithm the total number of labeled data is $n = 174$, across all 19 tasks (this is determined adaptively, by the proposed algorithm). For the random selection of labeled data, the data from

[1] The data from the landmine example are available at www.ece.duke.edu/ lcarin/LandmineData.zip

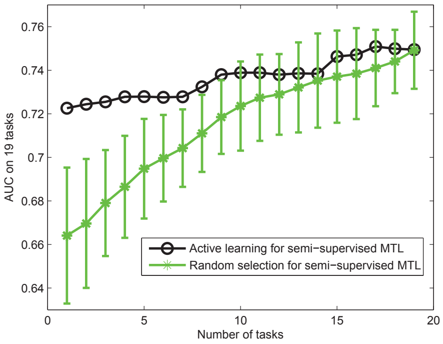

Fig. 1. Performance of active learning for semi-supervised MTL algorithm in comparison to semi-supervised MTL with randomly-selected labeled data. The horizontal axis is the number of tasks from which labeled data are acquired. The vertical axis is the AUC averaged over the 19 tasks.

all 19 tasks are put together, and 174 samples are selected at random for labeling; therefore, the number of labels acquired per task is not constant (the data in Figure 2, for random selection, represents one example). For the active-learning results in Figure 2, note the big jump in the number of labeled data at task $k = 11$. Recall from above that data sets 1-10 are from generally foliated regions and data sets 11-19 are from regions that are generally bare earth or desert. Therefore, the jump in Figure 2 at $k = 11$ is consistent with expectations based on the properties of the environments.

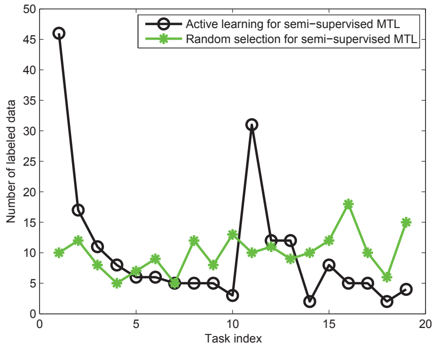

Fig. 2. Number of labeled data using active learning in comparison to number of labeled data with random selection; for the latter, this is one random example.

6. CONCLUSIONS

We have presented an active learning algorithm for semi-supervised MTL. The proposed algorithm actively and sequentially acquires the labels of the most informative data from each task. Experimental results on a real landmine detection problem show that the active acquisition of labeled data yields promising results.

7. REFERENCES

[1] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. 16th International Conf. on Machine Learning (ICML)*. 1999, pp. 200–209, Morgan Kaufmann, San Francisco, CA.

[2] M. Szummer and T. Jaakkola, "Partially labeled classification with Markov random walks," in *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[3] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *The Twentieth International Conference on Machine Learning (ICML)*, 2003, pp. 912–919.

[4] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo, "On semi-supervised classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[5] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[6] B. Bakker and T. Heskes, "Task clustering and gating for Bayesian multitask learning," *Journal of Machine Learning Research*, pp. 83–99, 2003.

[7] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram, "Multi-task learning for classification with Dirichlet process priors," *Journal of Machine Learning Research (JMLR)*, vol. 8, pp. 35–63, 2007.

[8] Gokhan Tur, "Multitask learning for spoken language understanding," in *Proc. IEEE Intl Conf. Acous., Speech, Sig. Proc.*, 2006.

[9] Q. Liu, X. Liao, and L. Carin, "Semi-supervised multi-task learning," in *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2009, MIT Press.

[10] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 589–603, 1992.

[11] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[12] V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, New York, 1972.

[13] G. Tur, R. E. Schapire, and D. Hakkani-Tur, "Active learning for spoken language understanding," in *Proc. IEEE Intl Conf. Acous., Speech, Sig. Proc.*, 2003, pp. 276–279.

[14] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[15] M.D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1995.

[16] T.M. Cover and J.A. Thomas, *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.