Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning

Ashish Kapoor, Eric Horvitz and Sumit Basu
Microsoft Research
1 Microsoft Way, Redmond, WA 98052, USA
{akapoor, horvitz, sumitb}@microsoft.com

Abstract

An inescapable bottleneck with learning from large data sets is the high cost of labeling training data. Unsupervised learning methods have promised to lower the cost of tagging by leveraging notions of similarity among data points to assign tags. However, unsupervised and semi-supervised learning techniques often provide poor results due to errors in estimation. We look at methods that guide the allocation of human effort for labeling data so as to get the greatest boosts in discriminatory power with increasing amounts of work. We focus on the application of value of information to Gaussian Process classifiers and explore the effectiveness of the method on the task of classifying voice messages.

1 Introduction

Increased sensing and decreased storage costs are leading to the growing availability of large data sets. However, data resources are often unavailable for machine learning and reasoning because of the high cost of labeling cases. As an example, we may have access to several thousand voice messages stored on a server and wish to build a classification system that could automatically classify voicemail messages into different categories. Unfortunately, performing supervised learning with the data set would require the manual effort of listening to voice messages and applying labels.

Unsupervised learning shows promise for reducing the effort required for tagging, such as its use in preparing data sets for supervised learning. However, pure unsupervised learning, based on notions of clusters and similarity, is often fraught with labeling errors.

We believe that there is a rich space of opportunities within the realm of *complementary computing* [Horvitz and Paek, 2007] for machine learning. We focus on the ideal coupling of human supervision with unsupervised methods and we discuss *selective supervision*, the use of value-of-information to triage human tagging efforts. The active-learning method considers both the cost required to tag data as well as the costs associated with the use of a classifier in a real-world setting. We show how we can minimize the total cost associated with the construction and use of classification systems, where costs are measured in currencies such as monetary quantities or other valuable resources.

Given the cost of labeling previously unlabeled cases and the cost of misclassification—which may be different for different classes—we seek to quantify the expected gain in expected value associated with seeking information on an unlabeled data point. This expected gain, which corresponds to the value of information provided by labeling, is the guiding principle for the active learning framework. We shall show how we can employ this value of information in learning within the Gaussian Process classification framework, and describe the applicability of the approach to both supervised and semi-supervised scenarios.

After a review of some key work in active learning, we describe the active learning framework that uses expected value-of-information (VOI) criterion to triage labeling. Then, we present concepts and computational issues with the use of Gaussian Process classification, and we show how the methods can be extended to handle semi-supervised learning. We highlight the effectiveness of the framework on the task of classifying voice messages and we conclude with experimental results and discussion.

2 Background

Interest has been growing in recent years in *active learning*. Numerous heuristics and schemes have been proposed for choosing unlabeled points for tagging. For example, Freund *et al.* 1997 propose disagreement among the committee of classifiers as a criterion for active learning. Tong and Koller, 2000 explore the selection of unlabeled cases to query based on minimizing the version space within the support vector machines (SVM) formulation. Within the Gaussian Process framework, the method of choice has been to look at the expected informativeness of an unlabeled data point [MacKay, 1992; Lawrence *et al.*, 2002]. Specifically, the idea is to choose to query cases that are expected to maximally influence the posterior distribution over the set of possible classifiers. Additional studies have sought to combine active learning with semi-supervised learning [McCallum and Nigam, 1998; Muslea *et al.*, 2002; Zhu *et al.*, 2003].

All of these methods inherently focus on minimizing the misclassification rate. We focus on the value of moving to a decision-theoretic framework in active learning, where we consider the costs and risks in real-world currencies and employ computations of expected value of information to balance the cost of misdiagnosis with the costs of providing labels.

3 Decision-Theoretic Active Learning

A *linear classifier* parameterized by $\mathbf{w}$ classifies a test point $\mathbf{x}$ according to: sign($f(\mathbf{x})$), where $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. Given a set of training data points $\mathcal{X}_L = \{\mathbf{x}_1, .., \mathbf{x}_n\}$, with class labels $\mathcal{T}_L = \{t_1, .., t_n\}$, where $t_i \in \{1, -1\}$, the goal of a learning algorithm is to learn the parameters $\mathbf{w}$. Most classification techniques focus on minimizing classification error. However, preferences about the relative numbers of false positives and negatives produced by a classification system can vary by person and task. These preferences can be expressed in terms of real-world measures of cost such as a monetary value and we can seek to minimize the expected cost for the use of a classifier over time. Additionally, we have the cost of tagging cases for training, which can vary for cases in different classes or with other problem-specific variables.

We aim to quantify the values of acquiring labels of different data points and to use computations of these values as a guiding principle in active learning. Intuitively, knowing the label of one or more currently unlabeled points may reduce the total risk in the classification task. On the other hand, labels are acquired at a price. The difference in the reduction in the total expected cost of the use of the classifier, which we shall refer to as the risk, and the cost of acquiring a new label is the expected value of information for learning that label. The real-world cost associated with the usage of a classifier is a function of the number of times that a classifier will be used in the real world, so a probability distribution over usage is considered in the computation of expected cost.

For simplicity, we shall focus in the discussion on two-class discrimination problems. The methods discussed in the paper can be generalized in a straightforward manner to handle multiple classes. We note that the work presented here makes a *myopic* assumption, where we only seek to label one data point at a time. This can be generalized by applying lookahead procedures that consider the acquisition of labels for different sets of points.

Let us define the risk matrix $\mathbf{R} = [R_{ij}] \in I\!R^{2\times 2}$, where R_{ij} denote the cost or risk associated with classifying a data point belonging to class i as j. We use the index 2 to denote the class -1. We assume that the diagonal elements of $\mathbf{R}$ are zero, specifying that correct classification incurs no cost. Thus, given the labeled set $\mathcal{X}_L$ with labels $\mathcal{T}_L$ we can train a classifier $f(\mathbf{x})$ and compute the total risk on the labeled data points as:

$$J_L = \sum_{i \in L_+} R_{12}(1 - p_i) + \sum_{i \in L_-} R_{21} p_i \qquad (1)$$

Here, p_i denotes the probability that the point $\mathbf{x}_i$ is classified as class +1, i.e. $p_i = p(\text{sign}(f(\mathbf{x}_i)) = 1|\mathbf{x}_i)$. Further, L_+ and L_- are the indices of positively and negatively labeled points respectively. Note, that p_i is the predictive distribution and depending upon the classification technique may or may not be available. Predictive distributions are available for Gaussian Process classification (Section 4) and other probabilistic classifiers, including probabilistic mappings of outputs of SVMs [Platt, 2000].

Beyond labeled cases, we also have a set of unlabeled data points $\mathcal{X}_U = \{\mathbf{x}_{n+1}, .., \mathbf{x}_{n+m}\}$, that we wish to classify. We seek to include the total risk associated with the unlabeled data points:

$$J_U = \sum_{i \in U} R_{12}(1 - p_i) \cdot p_i^* + R_{21} p_i \cdot (1 - p_i^*) \qquad (2)$$

Here, $p_i^* = p(t_i = 1|\mathbf{x}_i)$ is the *true* conditional density of the class label given the data point. As we do not have the true conditional, we cannot compute this expression exactly. However, we can approximate p_i^* with p_i; thus, we approximate the total risk on the unlabeled data points as:

$$J_U \approx \sum_{i \in U} (R_{12} + R_{21})(1 - p_i) \cdot p_i \qquad (3)$$

Now, let C_i denote the cost of knowing the class label of $\mathbf{x}_i$. We assume that the costs C_i and the risks R_{12} and R_{21} are measured with the same currency. This assumption does not impose significant constraints as different currencies can be transformed into a single utility by using appropriate real-world conversions.

Given the risks (J_L and J_U), we approximate[1] the expected misclassification cost per point as $\bar{J} = \frac{J_L + J_U}{n+m}$. Assuming a *closed* world, where the system only encounters the $n + m$ points in $\mathcal{X}_L \cup \mathcal{X}_U$, the expected cost is the sum of the total risk ($J_{all} = (n + m)\bar{J}$) and the cost of obtaining the labels:

$$U = J_{all} + \sum_{i \in L} C_i = J_L + J_U + \sum_{i \in L} C_i \qquad (4)$$

Upon querying the new point, we may see a reduction in the total risk. However, a cost is incurred when we query a label and computing the difference in these quantities triages the selection of cases to label. Formally, we define the VOI of an unlabeled point $\mathbf{x}_j$ as the difference in the reduction in the total risk and the cost of obtaining the label:

$$VOI(\mathbf{x}_j) = U - U^j = (J_{all} - J_{all}^j) - C_j \qquad (5)$$

Here, U^j and J_{all}^j denote the total expected cost and the total misclassification risk respectively if we consider $\mathbf{x}_j$ as labeled. The VOI quantifies the gain in utilities in terms of the real-world currency that can be obtained by querying a point; hence, our strategy would be to choose next for labeling the point that has the highest value of information. This results in minimization of the total cost U that consists of the total risk in misclassification as well as the labeling cost. We note that this approach differs from the earlier methods in active learning where the focus has been to minimize the classification error.

Now, let us consider $\mathbf{x}_j$ for querying. Note that we need to compute the expression for VOI before we know the label for $\mathbf{x}_j$ and the total risk J_{all}^j cannot be computed before we know the actual label t_j. Similarly, C_j cannot be computed if the costs of labels are different for different classes. We approximate the terms J_{all}^j for the j^{th} data point with an expectation of the empirical risk as: $J_{all}^j \approx p_j J^{j,+} + J^{j,-}(1 - p_j)$. Here $J^{j,+}$ and $J^{j,-}$ denote the total risks when $\mathbf{x}_j$ is labeled as class 1 and class -1 respectively. These risks can be written as sum of risks on labeled and unlabeled data as following:

$$J^{j,+} = J_L^{j,+} + J_U^{j,+} \quad \text{and} \quad J^{j,-} = J_L^{j,-} + J_U^{j,-} \qquad (6)$$

[1]This approximation can be used for any case, whether previously seen or unseen, provided the density $p(\mathbf{x})$ does not change.

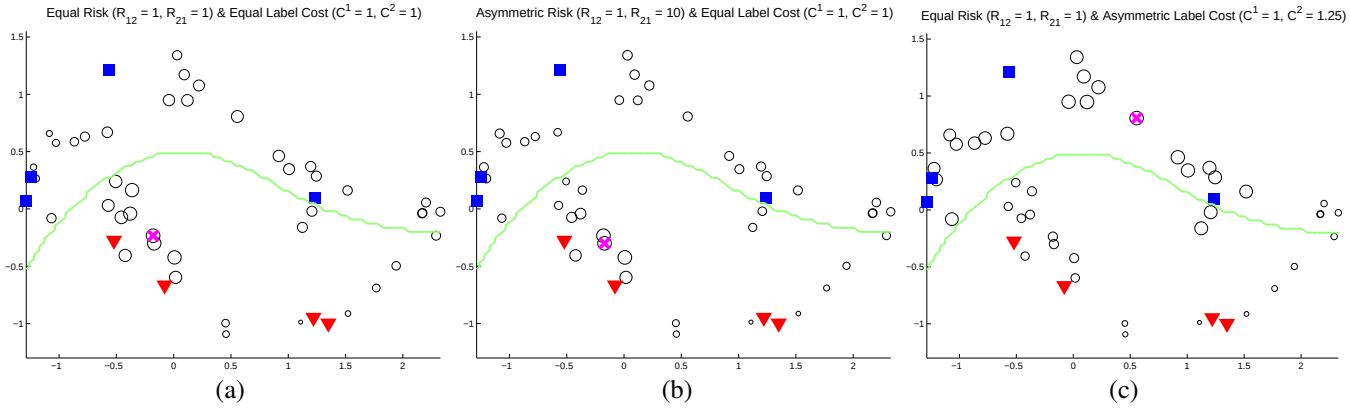

Figure 1: Selection of points to query for label based on the value of information. The circles represent the unlabeled cases and the radii correspond to the VOI of labeling each case. The squares (class 1) and the triangles (class -1) represent previously labeled cases. The different figures correspond to the following situations: (a) symmetry in the costs of both the risks and the labeling, (b) asymmetric risk, and (c) asymmetric label costs. The cross corresponds to the selection of the next query. The curve denotes the decision boundary based on the currently available labels.

To calculate these risks we first compute $p^{j,+}$, the resulting posterior probability upon adding $\mathbf{x}_j$ as a positively labeled example in the active set. Now, using similar expressions to equations 1 and 3 we can compute $J_L^{j,+}$ and $J_U^{j,+}$, the risks on the labeled and the unlabeled data points when $\mathbf{x}_j$ is assumed to be positively labeled. The corresponding computations follow for $J_L^{j,-}$ and $J_U^{j,-}$ as well. Similarly, we can use expectation of C_j if costs of labeling vary by class.

Thus, our strategy is to select cases for labeling that have the highest VOI:

$$j_{sel} = \arg\max_{j \in U} VOI(\mathbf{x}_j) \tag{7}$$

We note that whenever $VOI(\mathbf{x}_{j_{sel}})$ is less than zero, we have a condition where knowing a single label does not reduce the total cost; thus, this situation can be employed as a stopping criterion. Stopping criteria for open-world situations would include the computation of gains in accuracy of the classifier over multiple uses, based on a probability distribution over expected cases and the lifespan of the system. We note that a greedy policy might indicate stopping when there is still potential for further reduction of the overall cost via querying a set of points.

We now demonstrate the approach, starting with illustrations with a toy data set. We shall employ Gaussian Process classification (see section 4) within the active learning framework. Figure 1 shows the selection of unlabeled points to query based on the VOI criterion. The sample data consists of two half moons in a multidimensional space, where the top half belongs to class +1 and the bottom to the class -1. In the simulation, we start with a few cases that are already labeled and are represented as squares for class 1 and triangles for the class -1. The different graphs in the figure correspond to different settings of risks (R_{12} and R_{21}) and labeling costs. We assume that C^1 and C^2 are the costs for querying points that belong to class $+1$ and -1 respectively. The unlabeled points are displayed as circles and the radii correspond to the VOI of labeling these cases. The next case selected to be queried is marked with a cross. Figure 1(a) shows the VOI for all the unlabeled data points and the case selected for the next query when the risks and the cost of labelings are equal for both classes. For this situation, cases that are nearest to the decision boundary are associated with the highest VOI. Choosing cases that minimize the objective for overall cost corresponds to the selection of queries that would minimize the classification error; hence, the points at the decision boundary are the ones that are the most informative. Figure 1(b) illustrates the situation where it is far more expensive to misclassify a point belonging to class -1. Due to this asymmetry in risks, the points that are likely to belong to class -1, but that also lay close to the decision boundary, have the highest VOI. Figure 1(c) depicts the situation where obtaining a label for a point in class -1 is 1.25 times as expensive to obtain the label for a point belonging to class 1. The VOI is highest for those points that are more likely to belong to class 1 and that are close to the decision boundary. The sample data set illustrates how VOI can be used effectively to guide tagging supervision such that it minimizes both the operational and training costs of a classifier.

We point out that we have assumed a *closed system* where both the set of the labeled and the unlabeled data are available beforehand. Prior work on active learning includes studies that assume an *open system*, where the data points arrive one at a time. The methods discussed in this paper can be extended to handle the open system case by using the average empirical risk ($\frac{J_L}{n}$) as a guiding principle as earlier used by Zhu *et al.*, 2003. Note that this is not a transductive learning framework; the final classification boundary depends only on the labeled data. Both the labeled and the unlabeled data points are used only to determine which cases to query. Once trained, the classifier can be applied to novel test points, beyond the original set of labeled and the unlabeled points. We shall describe in Section 4.2 extensions to handle the transductive case by using a semi-supervised learning algorithm to train the classifier. Before that, we will pause to review Gaussian Process classification.

Table 1: Features extracted from voice messages

Prosodic features (* includes max, min, mean and variance)	**Metadata information**
Duration of silence*	Is Weekend?
Duration of voiced segment*	Is AM on a work day?
Absolute pitch*	Is PM on a work day?
Length of productive segment*	Is after hours on a work day?
Length of pause*	Size in bytes
Change in pitch during productive segments*	Size in seconds
Rate features (syllable, silence, productive segments, pauses)	Is external caller?

4 Gaussian Process Classification

In this work, we explore active learning with the use of Gaussian Process (GP) classifiers. One of the advantages of using GP classification is that we directly model the predictive conditional distribution $p(t|\mathbf{x})$, consequently making it easy to compute the actual conditional probabilities without any calibrations or post-processing. GP methods provide a Bayesian interpretation of classification. With the approach, the goal is to infer the posterior distribution over the set of all possible classifiers given a training set:

$$p(\mathbf{w}|\mathcal{X}_L, \mathcal{T}_L) = p(\mathbf{w}) \prod_{i \in L} p(t_i|\mathbf{w}, \mathbf{x}_i) \tag{8}$$

Here, $p(\mathbf{w})$ corresponds to the prior distribution over the classifiers and is selected typically so as to prefer parameters $\mathbf{w}$ that have a small norm. Specifically, we assume a spherical Gaussian prior on the weights: $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$. The prior imposes a smoothness constraint and acts as a regularizer such that it gives higher probability to the labelings that respect the similarity between the data points. The likelihood terms $p(t_i|\mathbf{w}, \mathbf{x}_i)$ incorporate the information from the labeled data and different forms of distributions can be selected. A popular choice is the probit likelihood: $p(t|\mathbf{w}, \mathbf{x}) = \Psi(t \cdot \mathbf{w}^T\mathbf{x})$. Here, $\Psi(\cdot)$ denotes the cumulative density function of the standard normal distribution. The posterior prefers those parameters that have small norm and that are consistent with the training data.

Computing the posterior, $p(\mathbf{w}|\mathcal{X}, \mathcal{T})$, is non-trivial and approximate inference techniques such as Assumed Density Filtering (ADF) or Expectation Propagation (EP) are typically required. The idea behind ADF is to approximate the posterior $p(\mathbf{w}|\mathcal{X}_L, \mathcal{T}_L)$ as a Gaussian distribution, *i.e.* $p(\mathbf{w}|\mathcal{X}_L, \mathcal{T}_L) \approx \mathcal{N}(\bar{\mathbf{w}}, \Sigma_{\mathbf{w}})$. Similarly, EP is another approximate inference technique. EP is a generalization of ADF, where the approximation obtained from ADF is refined using an iterative message passing scheme. We refer readers to Minka, 2001 for the details.

Given the approximate posterior $p(\mathbf{w}|\mathcal{X}, \mathcal{T}) \sim \mathcal{N}(\bar{\mathbf{w}}, \Sigma_{\mathbf{w}})$, a frequent practice is to choose the mean $\bar{\mathbf{w}}$ of the distribution as the point classifier. The mean, which is also called the *Bayes point*, classifies a test point according to: $\text{sign}(\bar{\mathbf{w}}^T\mathbf{x})$. It is relatively straightforward to generalize to the non-linear case by using the kernel trick, where the idea is to first project the data into a higher dimensional space to make it separable [Evgeniou *et al.*, 2000].

One of the byproducts of using the GP classification framework is that we obtain a predictive distribution $p(\text{sign}(f(\mathbf{x}))|\mathbf{x})$:

$$p(\text{sign}(f(\mathbf{x})) = \mathbf{1}|\mathbf{x}) = \Psi\left(\frac{\bar{\mathbf{w}}^T\mathbf{x}}{\sqrt{\mathbf{x}^T\Sigma_{\mathbf{w}}\mathbf{x} + 1}}\right) \tag{9}$$

Unlike other classifiers, the GP classification models the predictive conditional distribution $p(t|\mathbf{x})$, making it easy to compute the actual conditional probabilities without any calibrations or post-processing. Probabilistic interpretations have been made of other kernel classifiers such as SVM [Sollich, 1999] and other attempts that map the output of the classifiers directly to the probability [Platt, 2000]. Our approach is to use this predictive distribution in the selective-supervision framework to compute expected risks and to quantify the value of information.

4.1 Computational Issues

As mentioned earlier, ADF or EP can be used for approximate inference in GP classification. However, the proposed scheme for selecting unlabeled points is computationally expensive. Note, that computational complexity for EP is $O(n^3)$, where n is the size of labeled training set. In the proposed method, we have to compute VOI for every unlabeled data point, requiring us to perform EP twice for every point under consideration.

A faster alternative is to use ADF for approximating the new posterior over the classifier rather than computing EP. Specifically, to compute the new posterior $p^{j,+}(\mathbf{w}|\mathcal{X}_{L\cup j}, \{\mathcal{T}_L \cup +1\})$ we can compute the Gaussian projection of the old posterior multiplied by the likelihood term for the j^{th} data point. That is: $p^{j,+}(\mathbf{w}|\mathcal{X}_{L\cup j}, \{\mathcal{T}_L \cup +1\}) \approx \mathcal{N}(\bar{\mathbf{w}}^{j,+}, \Sigma_{\mathbf{w}}^{j,+})$, where $\bar{\mathbf{w}}^{j,+}$ and $\Sigma_{\mathbf{w}}^{j,+}$ are respectively the mean and the covariance of $p(\mathbf{w}|\mathcal{X}_L, \mathcal{T}_L) \cdot \Psi(1 \cdot \mathbf{w}^T\mathbf{x}_j)$. This is equivalent to performing ADF starting with the old posterior $p(\mathbf{w}|\mathcal{X}_L, \mathcal{T}_L)$ and incorporating the likelihood term $\Psi(1 \cdot \mathbf{w}^T\mathbf{x}_j)$ and does not require $O(n^3)$ operations to compute VOI for every unlabeled data point. We can use similar computations to approximate $p^{j,-}(\mathbf{w}|\mathcal{X}_{L\cup j}, \{\mathcal{T}_L \cup -1\})$.

4.2 From Supervised to Semi-Supervised Learning

The underlying classifier in the proposed framework is based on Gaussian Processes and it can be easily extended for the semi-supervised case [Kapoor, 2006; Sindhwani *et al.*, 2005]. Specifically, at the core in the GP classification is the kernel matrix $\mathbf{K}$, where entry K_{ij} encodes the similarity between the i^{jth} and the j^{th} data points. Rather than using $\mathbf{K}$ as the

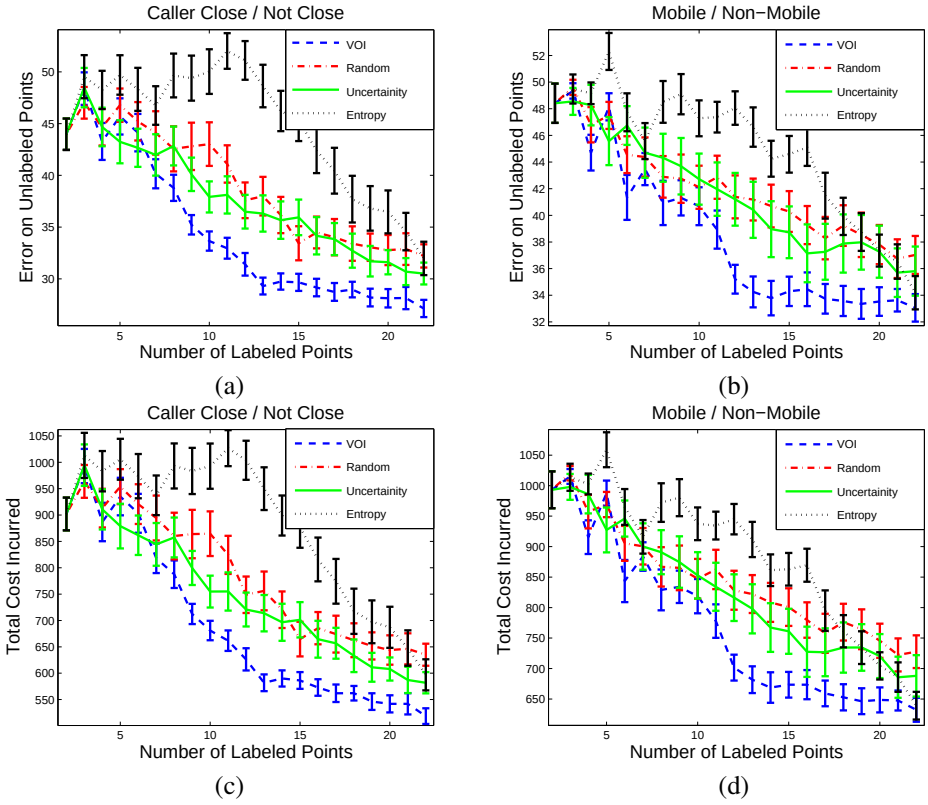

Figure 2: Comparison of different active learning schemes. Graphs (a) and (b) show the error on unlabeled points versus the number of labels for classifying voicemails. Graphs (c) and (d) show the total cost incurred versus the number of labels. VOI criteria can provide good classification performance with significantly lower costs. The results are averaged over 20 runs and the error bars represent the standard error.

similarity matrix for GP classification, we can use the *inverse* of the transformed Laplacian:

$$r(\Delta) = \Delta + \sigma \mathbf{I} \quad \text{where} \quad \Delta = \mathbf{D} - \mathbf{K}$$

Here, $\mathbf{D}$ is the diagonal matrix where the diagonal elements are: $D_{ii} = \sum_j K_{ij}$ and $\sigma > 0$ is added to remove the zero eigenvalue from the spectrum of $r(\Delta)$. Intuitively, rather than computing similarity directly via the kernel K_{ij}, the inverse of the transformed Laplacian computes the similarity over a manifold. Thus, the unlabeled data points help in classification by populating the manifold and using the similarity over the manifold to guide the decision boundary. The extension of GP classification to handle semi-supervised learning has been recently studied [Kapoor, 2006; Sindhwani *et al.*, 2005] and is related to the graph-based methods for semi-supervised learning. The rest of the active learning framework can be used as it is on top of this semi-supervised GP classification framework.

5 Sample Challenge: Classifying Voicemail

We shall now move to a challenging classification task that highlights the value of employing the selective supervision methods. Specifically, our goal is to build a system that can classify voice messages in several ways, including whether the messages are urgent vs. non-urgent, the caller is personally close vs. not close to the person being called, and to detect if the caller is calling from a mobile phone. The classification task is related to prior work on the prioritization and routing of email messages with statistical classifiers [Horvitz *et al.*, 1999].

Given a set of voice messages, we first extract features that promise to be of value in discriminating among the target classes. Specifically, we look at the prosody and metadata that accompanies the messages.

Prosodic features: We consider multiple prosodic features including syllable rate, pause structure, and pitch dynamics. We employ a pitch tracker and then extract the prosodic features summarized in table 1.

Message metadata: We also extract metadata from the voice messages. Specifically, we extract features that indicate the day and the time of the call. Additionally, we consider the size of the voicemail in bytes as well as the length of the message in seconds. We also extract features that indicate whether the caller is calling from outside the recipient's organization. Several metadata features are shown in table 1.

We now explore the selective supervision concepts applied to the voicemail classification challenge. The data set consists of 207 labeled voice messages received by a single user over a period of 8 months. We explore the use of the methods to guide the labeling efforts for *supervised* classification. Annotating a voicemail is tedious and shorter voicemails can be labeled more quickly than the longer ones. Thus, we use the asymmetric cost criterion where the cost of a label scales with the length of the voicemail. Specifically, we assume that the cost of labeling voicemail is 0.01 US dollars per second of message length. Further, for all of the experiments we assume

Table 2: Average accuracy (standard error) on the unlabeled points and on the total cost when starting with one labeled point per class and choosing 20 other points. The bold figures indicate significant performance difference with 95% confidence.

Task	Accuracy		Cost	
	VOI	Random	VOI	Random
Close?	**72.9±0.8**	67.8±1.1	**518.5±14.9**	635.0±20.8
Mobile?	**66.9±1.0**	63.0±1.4	**632.2±19.5**	727.6±26.8
Urgent?	53.2±0.5	53.2±0.6	899.5±10.1	913.0±12.7

that misclassification costs $R_{12} = R_{21} = 10$ US dollars. For Gaussian Process classification, the polynomial kernel of degree 2 is used.

We compare the selective-supervision strategy in supervised learning with three other schemes: 1) selecting points randomly, 2) choosing the point where the classification is most uncertain (*i.e.,* $j_{sel} = \arg\min_{j \in U} |p_j - 0.5|$) and 3) choosing the point that is likely to change the posterior over $\mathbf{w}$ the most [MacKay, 1992; Lawrence *et al.*, 2002], *i.e.,* $j_{sel} = \arg\max_{j \in U} p_j \Delta^{j,+} + (1 - p_j)\Delta^{j,-}$. Here, $\Delta^{j,+}$ and $\Delta^{j,-}$ are the differential entropy scores [Lawrence *et al.*, 2002] when j^{th} point is added as a positively and a negatively labeled point in the training set respectively. The term being maximized quantifies the expected change in the posterior over $\mathbf{w}$ when point $\mathbf{x}_j$ is added to the training set.

Graphs (a) and (c) in Figure 2 compare the different active learning schemes on the task of detecting if the caller is personally close to the person being called. Similarly, graphs (b) and (d) in Figure 2 show the plots for the task of classifying whether the voice messages originated from a mobile phone. For the studies, results are averaged over 20 runs, where, for each run, we select one case per class as labeled and the rest of the points are selected according to the different active learning policies. Note, that, in the beginning, when there are a very few labeled data points, it is difficult to estimate the misclassification risk ($R_L + R_U$). However, with increasing numbers of labeled cases, the Gaussian Process classification will typically provide a better estimate of the posterior $p(\text{sign}(f(\mathbf{x}) = 1|\mathbf{w}, \mathbf{x})$.

We found that the VOI policy results in significant gains over the other methods, both in terms of the accuracy as well as the cost. Table 2 shows the average classification accuracy and the standard error on the unlabeled data points together with the total cost (the sum of the misclassification and training costs) incurred after querying 20 points guided by the VOI criterion and random selection policy. As indicated by the results, the VOI criterion provides significant gains over random sampling in terms of cost and accuracy for classifying messages as personally close versus not close and mobile versus non-mobile. The VOI criterion for selective supervision provides valuable guidance on labeling efforts under budget constraints. We found that the gain with the use of VOI was not significant for detecting urgency. We hypothesize that this is due to the poor separability of the data.

6 Conclusion

We have investigated a decision-theoretic approach to classification, focusing on the use of value of information as the basis for guiding supervision. We showed how the risks of misclassification and cost of obtaining labels can be used to quantify the value of querying for labels of unknown cases. We applied and tested the ideas within the Gaussian Process classification framework. Finally, we reviewed results obtained from applying the methods to the task of building predictive models for classifying voice messages, where the cost of labeling a voicemail scales with the length of messages. We are continuing to explore challenges with the efficient computation of VOI and with studying the value of using the methods in triaging labeling effort in selective supervision.

References

[Evgeniou *et al.*, 2000] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[Freund *et al.*, 1997] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3), 1997.

[Horvitz and Paek, 2007] E. Horvitz and T. Paek. Complementary computing, User Modeling and User-Adapted Interaction 17. *Special Issue on Statistical and Probabilistic Methods for User Modeling*, 2007.

[Horvitz *et al.*, 1999] E. Horvitz, A. Jacobs, and D. Hovel. Attention-sensitive alerting. In *Uncertainty in Artificial Intelligence*, 1999.

[Kapoor, 2006] A. Kapoor. *Learning Discriminative Models with Incomplete Data*. PhD thesis, Massachusetts Institute of Technology, 2006.

[Lawrence *et al.*, 2002] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process method: Informative vector machines. *Neural Information Processing Systems*, 15, 2002.

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

[McCallum and Nigam, 1998] A. K. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *International Conference on Machine Learning*, 1998.

[Minka, 2001] T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institue of Technology, 2001.

[Muslea *et al.*, 2002] I. Muslea, S. Minton, and C. A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *International Conference on Machine Learning*, 2002.

[Platt, 2000] J. C. Platt. Probabilities for support vector machines. *Advances in Large Margin Classifiers*, 2000.

[Sindhwani *et al.*, 2005] V. Sindhwani, W. Chu, and S. S. Keerthi. Semi-supervised Gaussian processes. Technical Report YRL-2005-60, Yahoo! Research Labs, 2005.

[Sollich, 1999] P. Sollich. Probabilistic methods for support vector machines. *Neural Information Processing Systems*, 12, 1999.

[Tong and Koller, 2000] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *International Conference on Machine Learning*, 2000.

[Zhu *et al.*, 2003] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining at ICML*, 2003.