`activelearning`: An `R` Package to Label Unlabeled Data for Classification

John A. Ramey

Fred Hutchinson Cancer Research Center, Baylor University

November 28, 2012

Abstract

Researchers and practitioners are collecting increasingly large and complex data sets for which statistical and machine learning methods are being applied to better understand the data as well as utilizing the correlation structure within the data for predictive and classification purposes. Typically, supervised learning methods are employed as prediction and classification models and require that the response be given for each training observation. The time and resources necessary to collect responses for each training observation is often excessive and infeasible. As we may be able to collect responses on a small subset of the training observations without a response, we often wish to query the response or classification label for those training observations that would yield the most accurate supervised learning model with the least variance. Active learning is a recently established methodology that enables users to select observations optimally that would most improve the prediction or classification performance. In this document, we provide an overview of active learning and several common active-learning methods. Additionally, we discuss our implementation of these methods that are available in our recently developed `activelearning` `R` package. We provide a motivating example where we wish to recognize images of hand-written characters automatically, but with few observations labeled initially, we seek to determine the true label for a subset of the unlabeled observations in order to improve our

automatic recognition of future hand-written characters.

1 Introduction

Perpetual advancements in technologies have enabled researchers and practitioners to acquire complex data sets to better understand data-generating systems. Unfortunately, the size and complexity of these data can be overwhelming to the point that we have large quantities of information at our fingertips, yet our typical first approaches in utilizing the information within the data for predictive and classification purposes may yield anomalous discoveries and poor results. Numerous supervised learning methods have been developed to exploit the structure within the data to classify future, unseen observations based on few labeled training observations, but the accuracy of classification methods (hereafter, *classifiers*) is well-known to degrade as the number of labeled observations is reduced relative to the number of features present in the data.

The usual approach is to train a classifier based on a labeled training data set and then classify unlabeled observations as we obtain them. The implicit assumption here is that we know the true classification of every training observation, but this assumption is often too stringent because obtaining the true labels for each training observation can be costly and time-consuming. Hence, in practice, a small subset of our data may be labeled, whereas the vast majority of the training observations are unlabeled. In the supervised classification paradigm, we use only the labeled data to construct our classifier and ignore the unlabeled observations, but if we only have a handful of labeled observations, our classification performance on future observations may prove ineffective.

Ideally, we would like to obtain the true labels for a sufficient number of training observations from which we construct an efficacious classifier, but as we have discussed previously, the cost and time to procure the labels may hinder any attempts to determine the true classification labels for the majority of the training data. However, the researcher may be able to obtain the true labels for a small number of observations in an effort to improve sufficiently the classification accuracy of the trained classifier. Moreover, the researcher can perhaps achieve a desired accuracy level before the clas-

sifier is employed by obtaining the true labels for a subset of the unlabeled training data. For example, if the researcher's time and budget constraints allow him to hire an expert to provide the true labels for, say 50, training observations, the classification accuracy of these observation perhaps will improve sufficiently. However, from the entire set of unlabeled observations in hand, which observations should the researcher have the expert consider?

The researcher could query the true label from the expert hired for a random subset of training observations. Alternatively, the expert could be presented systematically with a subset of the unlabeled observations, e.g., the observations are presented to the expert in the order in which the data were received. However, these approaches are extremely naive in that they ignore the structure of the data that we already know. Moreover, if there is a region in the sample space for which we have previous knowledge, but there are other regions that we know little about (in terms of the labels), then we may wish to obtain the true classification labels for the training observations in the latter regions.

Active learning is a machine-learning that enables users to determine which observations should be labeled and how many are necessary to achieve a desired accuracy level. In the active learning literature, the researcher queries an **oracle** (i.e., the expert in our example above) with an unlabeled observation to determine its true label. Rather than naively querying the oracle for the true labels via a random or systematic method, several active-learning methods have been proposed in the literature to determine which unlabeled observations should be queried in an optimal manner with the majority of these methods aimed at improving the classification accuracy for the given training data. For an excellent and thorough overview of active learning and its methods, we recommend **?** and **?**.

Within the `activelearning` `R` package available on CRAN (`http://cran.r-project.org/`), we provide an implementation of several of the methods discussed in **?**. We also provide tools to equip researchers to employ active-learning methods in practice. Furthermore, because active learning is closely related to both optimal experimental design (**?**) and sample-size determination (**?**), we have also developed a set of tools to study

the classification accuracy for a data set as we query the true labels for more training observations using any of our implemented active-learning methods. Finally, we demonstrate the benefits of active learning by examining a handwriting recognition example, where our goal is to enhance substantially the character recognition accuracy of future images of handwritten characters.

In this paper we begin with a motivating example in Section 2 based on a large data set with only a few labeled observations. In Section 3 we present several of the proposed methods in the active learning literature to determine which unlabeled observations should be queried. We demonstrate our implementations of these methods in the `activelearning` package, which is available on CRAN (`http://cran.r-project.org/`) in Section 4. Then, in Section 5 we revisit our motivating example with the `activelearning` package and provide concluding remarks in Section 6.

2 Active Learning – Notation and Methods

In this section, we first provide an introduction active learning using notation that we borrow extensively from **?**. Then, we briefly discuss our implementation of several active-learning methods available in the `activelearning` package, including uncertainty sampling, query-by-bagging, and query-by-committee.

2.1 Notation

Supervised learning methods attempt to classify an unlabeled p-dimensional observation $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \mathbb{R}_{p\times 1}$ into one of K groups or classes, where $\mathbb{R}_{m\times n}$ denotes the matrix space of all $m \times n$ matrices over the real field $\mathbb{R}$. Additionally, we assume that we have drawn n_k independently and identically distributed random vectors from the kth class. We construct a supervised classifier from the $N = \sum_{k=1}^{K} n_k$ training observations to predict the class membership of $\boldsymbol{x}$.

2.2 Uncertainty Sampling

Uncertainty sampling is an active-learning method, where an oracle is queried for the true label of the unlabeled training observation that is most *uncertain*, where an unlabeled observation's uncertainty is typically defined by some loss or risk function. We define the uncertainty function about an observation $\boldsymbol{x}$ as $\phi(\boldsymbol{x})$, so that we will query the observation

$$\boldsymbol{x}^* = \arg\max_x \phi(\boldsymbol{x}).$$

? have proposed the **least confidence** function

$$\phi^{LC}(\boldsymbol{x}) = 1 - P_\theta(\widehat{y}|\boldsymbol{x})$$

where $\widehat{y} = \arg\max_y P_\theta(\widehat{y}|\boldsymbol{x})$ is the observation that maximizes the *a posteriori* probability under model θ. Hence, the least confidence observation that is queried is

$$\boldsymbol{x}^*_{LC} = \arg\max_x 1 - P_\theta(\widehat{y}|\boldsymbol{x}) = \arg\min_x P_\theta(\widehat{y}|\boldsymbol{x})$$

which is the observation that minimizes the *a posteriori* probability under the model θ. Although the least confidence function is simple to interpret, it often leads to the selection of outlying feature vectors, which can degrade future classification performance for classifiers that are not robust to outliers. If the majority of the labeled observations are outliers, the trained classifier can still have poor classification accuracy, even if a classifier is designed to be robust to outliers.

Other uncertainty sampling methods include **margin sampling**, which is a correction for the **least confidence** model in the multi-class setting. Perhaps, the most popular proposal for the **uncertainty sampling** function is the Shannon entropy function, so that we query the observation

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} -\sum_{i=1}^{K} P_\theta(y_i|\boldsymbol{x}) \log P_\theta(y_i|\boldsymbol{x}),$$

Notice that for binary classification (i.e., $K = 2$), the Shannon entropy function ap-

proach is equivalent to both the **least confidence** approach and to **margin sampling**.

2.3 Query by Committee

2.4 Query by Bagging

3 Handwriting Recognition Example

3.1 Random Querying: The Naive Approach

4 Concluding Remarks

Active learning is an effective methodology to obtain the true classification labels for large data sets where the vast majority of the observations are unlabeled in order to increase the classification accuracy of future observations. We have implemented several proposed active-learning methods in the `activelearning` `R` package. In addition, we have considered many effective active-learning approaches to query an oracle with unlabeled data to improve the classification performance of future unlabeled observations based on the training data in-hand.

We are excited by the possibilities that active learning can potentially enable. For example, semi-automated systems can be developed to query the true labels of observations about which the system is uncertain. This is especially useful for online systems that must handle large quantities of data, such as text, images, or videos, but the identification of each observation is time-consuming and/or costly. Our implementation of the `activelearning` package is intended to facilitate the advancement of these automated systems in practice.

Practically speaking, we may present an oracle with numerous features in the form of an image in the context of facial recognition, but in other domains, a graphical summary of an individual observation is often difficult to construct and to interpret (e.g. a graph of a linear combination of an observation's features). Also, the presentation of *every* feature of an observation may be unintelligible to the eye of a human oracle. An approach is to present a subset of an observation's features, but caution must be

used if the queried label is applied to the entire observation.

Additionally, the methods that we have considered and implemented have been largely developed for the purpose of text classification, where typically we assume that the feature vectors are probabilistically independent. Although this assumption is widely used in all areas of statistics and machine learning, the approach does not take into account if the data are correlated, especially when the data consists of sequential observations. The active learning methods that we have discussed do not explicitly account for or attempt to model autocorrelations in the data, although some active learning approaches may do this implicitly with ad-hoc reasoning.