Challenges and Opportunities in Applied Machine Learning

Carla E. Brodley, Umaa Rebbapragada, Kevin Small, Byron C. Wallace

■ *Machine-learning research is often conducted in vitro, divorced from motivating practical applications. A researcher might develop a new method for the general task of classification, then assess its utility by comparing its performance (for example, accuracy or AUC) to that of existing classification models on publicly available data sets. In terms of advancing machine learning as an academic discipline, this approach has thus far proven quite fruitful. However, it is our view that the most interesting open problems in machine learning are those that arise during its application to real-world problems. We illustrate this point by reviewing two of our interdisciplinary collaborations, both of which have posed unique machine-learning problems, providing fertile ground for novel research.*

What do citation screening for evidence-based medicine and generating land-cover maps of the Earth have in common? Both are real-world problems for which we have applied machine-learning techniques to assist human experts, and in each case doing so has motivated the development of novel machine-learning methods. Our research group works closely with domain experts from other disciplines to solve practical problems. For many tasks, off-the-shelf methods work wonderfully, and when asked for a collaboration we simply point the domain experts to the myriad available open-source machine-learning and data-mining tools. In other cases, however, we discover that the task presents unique challenges that render traditional machine-learning methods inadequate, necessitating the development of novel techniques.

In this article we describe two of our collaborative efforts, both of which have addressed the same question: if you do not initially obtain the classification performance that you are looking for, what is the reason and what can you do about it? The point we would like to make is that much of the research in machine learning focuses on improvements to existing techniques, as measured over benchmark data sets. However, in our experience, when applying machine learning, we have found that the choice of learning algorithm (for example, logistic regression, SVM, decision tree, and so on) usually has only a small effect on performance, and for most real-world problems it's easy to try some large subset of methods (for example, using cross-validation) until one finds the best method for the task at hand. In our view, the truly interesting questions arise when none of the available machine-learning algorithms performs adequately.

In such cases, one needs to address three questions: (1) why is the performance poor? (2) can it be improved? and (3) if so, how? Over the past 15 years, our research group has worked on over 14 domains (Draper, Brodley, and Utgoff 1994; Moss et al. 1997; Brodley and Smyth 1997; Brodley et al. 1999; Shyu et al. 1999; Lane and Brodley 1999; Kapadia, Fortes, and Brodley 1999; Friedl, Brodley, and Stahler 1999; Stough and Brodley 2001; MacArthur, Brodley, and Broderick 2002; Dy et al. 2003; Early, Brodley, and Rosenberg 2003; Aisen et al. 2003; Pusara and Brodley 2004; Fern, Brodley, and Friedl 2005; Lomasky et al. 2007; Preston et al. 2010; Rebbapragada et al. 2009; 2008b; 2008a), with an emphasis on effectively collaborating with domain experts. In the majority of these, poor performance was due not to the choice of learning algorithm, but rather was a problem with the training data. In particular, we find that poor classifier performance is usually attributable to one or more of the following three problems. First, insufficient training data was collected — either the training data set is too small to learn a generalizable model or the data are a skewed sample that does not reflect the true underlying population distribution. Second, the data are noisy; either the feature values have random or systematic noise, or the labels provided by the domain expert are noisy. Third, the features describing the data are not sufficient for making the desired discriminations. For example, a person's shoe size will not help in diagnosing what type of pulmonary disease he or she has.[1]

The unifying theme of this work is that application-driven research begets novel machine-learning methods. The reason for this is twofold: first, in the process of fielding machine learning, one discovers the boundaries and limitations of existing methods, spurring new research. Second, when one is actively working with domain experts, it is natural to ask how one might better exploit their time and expertise to improve the performance of the machine-learning system. This has motivated our research into interactive protocols for acquiring more training data (for example, active learning and variants), cleaning existing labeled data, and exploiting additional expert annotation beyond instance labels (that is, alternative forms of supervision).

Active Learning without Simplifying Assumptions

When confronted with inadequate classifier performance, perhaps the most natural strategy for improving performance is to acquire more labeled training data. Unfortunately, this requires that the domain expert spend valuable time manually categorizing instances (for example, biomedical citations or land-surface images) into their respective classes. It is critical to judiciously use expert time and minimize the expert's overall effort.

In an ongoing collaboration that investigates methods to semiautomate biomedical citation screening for systematic reviews, we have developed novel strategies that better exploit experts by using their annotation time wisely and providing a framework that incorporates their domain knowledge into the underlying machine-learning model. In this section we review this collaboration, illustrating how working closely with domain experts can motivate new directions in machine-learning research by exposing inadequacies in conventional techniques.

Systematic Reviews

Over the past few decades, evidence-based medicine (EBM) (Sackett et al. 1996) has become increasingly important in guiding health-care best practices. Systematic reviews, in which a specific clinical question is addressed by an exhaustive assessment of pertinent published scientific research, are a core component of EBM. Citation screening is the process in which expert reviewers (clinical researchers, who are typically doctors) evaluate biomedical document abstracts to assess whether they are germane to the review at hand (that is, whether they meet the prespecified clinical criteria). This is an expensive and laborious, but critical, step in conducting systematic reviews. A systematic review typically begins with a PubMed[2] query to retrieve candidate abstracts. The team of clinicians must then identify the relevant subset of this retrieved set. The next step is to review the full text of all "relevant" articles to select those that ultimately will be included in the systematic review. Figure 1 illustrates the systematic review process. Shown in log scale are three piles containing the total number of papers in PubMed in 2011 (left), potentially relevant papers (middle), and papers deemed relevant and requiring full-text screening (right). Here we focus on the process between the middle and right piles, known as citation screening.

The torrential volume of published biomedical literature means reviewers must evaluate many thousands of abstracts for a given systematic review. Indeed, for one such review concerning disability in infants and children, researchers at the Tufts Evidence-Based Practice Center screened more than 30,000 citations. This process is performed for every systematic review. The number of citations examined per review is increasing as EBM becomes more integral to the identification of best practices. Figure 2 illustrates the magnitude of this task. In general, an experienced reviewer can screen (that is, label) an abstract in approximately 30 seconds. For the aforementioned review in which more than 30,000 citations were screened,

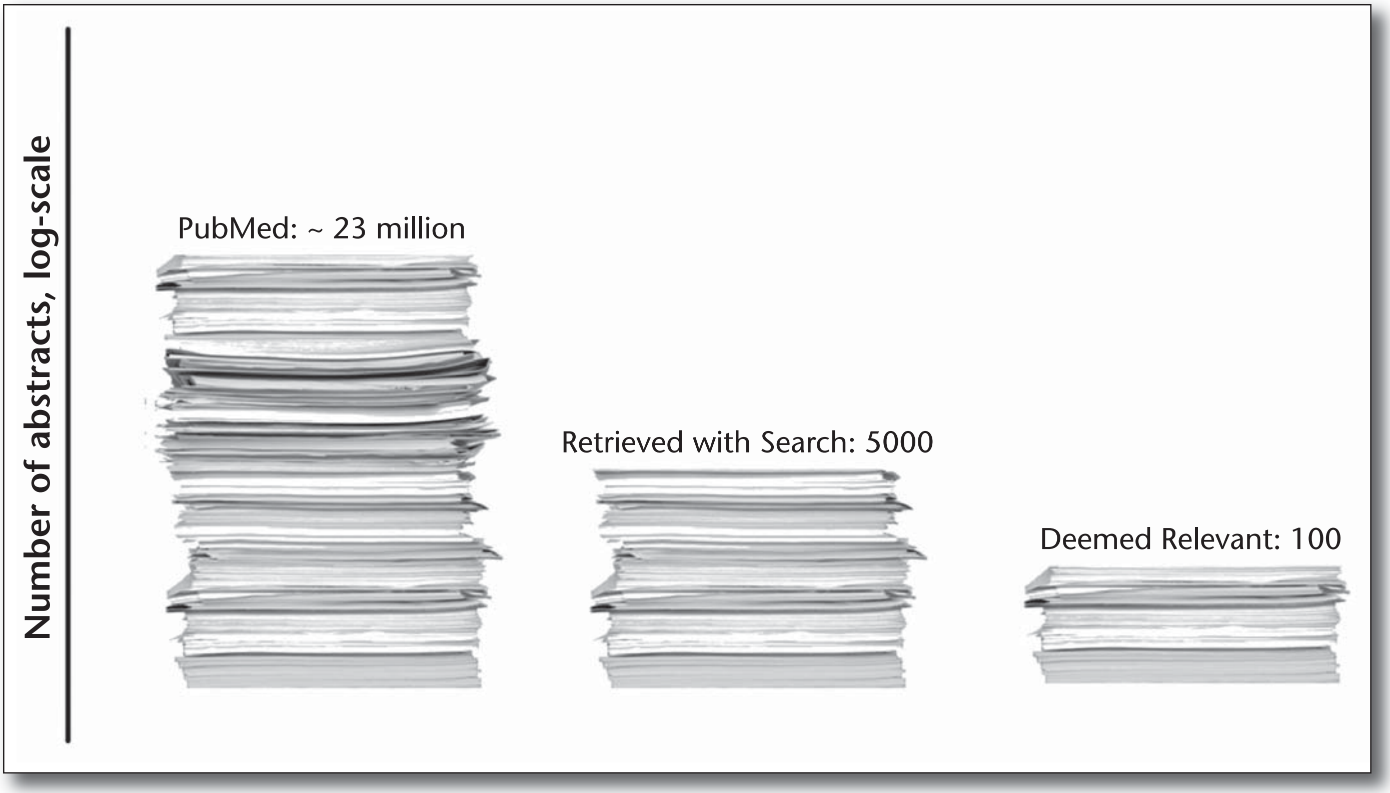

Figure 1. The Systematic Review Screening Process.

this translates to more than 250 (tedious) hours of uninterrupted work. Moreover, because doctors usually perform this task, citation screening is incredibly expensive.

Addressing the Limitations of Off-the-Shelf Machine Learning

In collaboration with the evidence-based practice center, our group is developing machine-learning methods to semiautomate citation screening. We begin by framing the task as a binary text classification problem; each citation is either relevant or irrelevant to the current systematic review. The aim then is to induce a classifier capable of making this distinction automatically, allowing the reviewers to safely ignore citations screened out (that is, labeled irrelevant) by the classifier. The combination of expensive annotation time and the requirement that a new classifier must be learned for each systematic review has motivated our interest in the active learning framework (Settles 2009), which attempts to mitigate the amount of labeled data required to induce a sufficiently performing classifier by training the model interactively. The intuition is that by selecting examples cleverly, rather than at random, less labeled data will be required to induce a good model, thereby saving the expert valuable time.

Active learning has been shown to work quite well empirically for in vitro settings (for example, Mccallum and Nigam [1998], Tong and Koller [2000]). In particular, uncertainty sampling, in which at each step in active learning the unlabeled instance about whose predicted label the model is least confident is selected for labeling, has become an immensely popular technique. We were therefore disappointed when uncertainty sampling failed to outperform random sampling for our systematic review data sets with respect to the recall-focused evaluation metric of interest for this application.

This last statement raises an important point: If evaluating uncertainty sampling as if these were "benchmark" data sets about which we knew nothing aside from the feature vectors and their labels, the results using active learning were quite good in terms of accuracy. Figure 3 illustrates this point. Shown are two comparisons of querying strategies: random sampling, in which the instances to be labeled and used as training data are selected randomly, and the active learning strategy known as SIMPLE (Tong and Koller 2000), an uncertainty-sampling technique. The results shown are averages taken over 10 runs. As a base learner for both querying strategies, we used support vector machines (SVMs). On the left side of the figure, we have plotted overall accuracy (assessed over a hold-

Figure 2. Abstracts to Be Screened.

These stacks of paper (printed abstracts) represent only a portion of the roughly 33,000 abstracts screened for a systematic review concerning disability in infants and children. This task required approximately 250 hours of effort by researchers at Tufts Medical Center.

out set) against the number of labels provided by the domain expert. Note that, as one might hope, the active learning curve dominates that of random sampling. One might conclude that active learning ought then to be used for this task, as it produces a higher accuracy classifier using fewer labels (that is, less human effort). However, consider the right side of the figure, which substitutes sensitivity for accuracy as the metric of performance. Sensitivity is the proportion of positive instances (in our case, relevant citations) that were correctly classified by the model.

In the citation screening task, we care more about sensitivity than we do about the standard metric of accuracy for two reasons. First, of the thousands of citations retrieved through an initial search, only about 5 percent, on average, are relevant. This is a phenomenon known as class imbalance. Second, for a given systematic review, it is imperative that all relevant published literature be included, or else the scientific validity of the review may be compromised. In the nomenclature of classification, this means we have highly asymmetric costs. In particular, false negatives (relevant citations the classifier designates as irrelevant) are much costlier than false positives (irrelevant citations the classifier designates as relevant); false negatives may jeopardize the integrity of the entire review while false positives incur additional labor for the reviewers (in the form of reviewing the article in full text to determine its relevance). Although it has previously been recognized that class imbalance and cost-sensitive classification are important for real-world problems (Drummond, Holte, et al. 2005; Provost 2000; He and Garcia 2009), many new studies still use accuracy or AUC to evaluate new learning methods. It is thus important to reiterate our point: without understanding

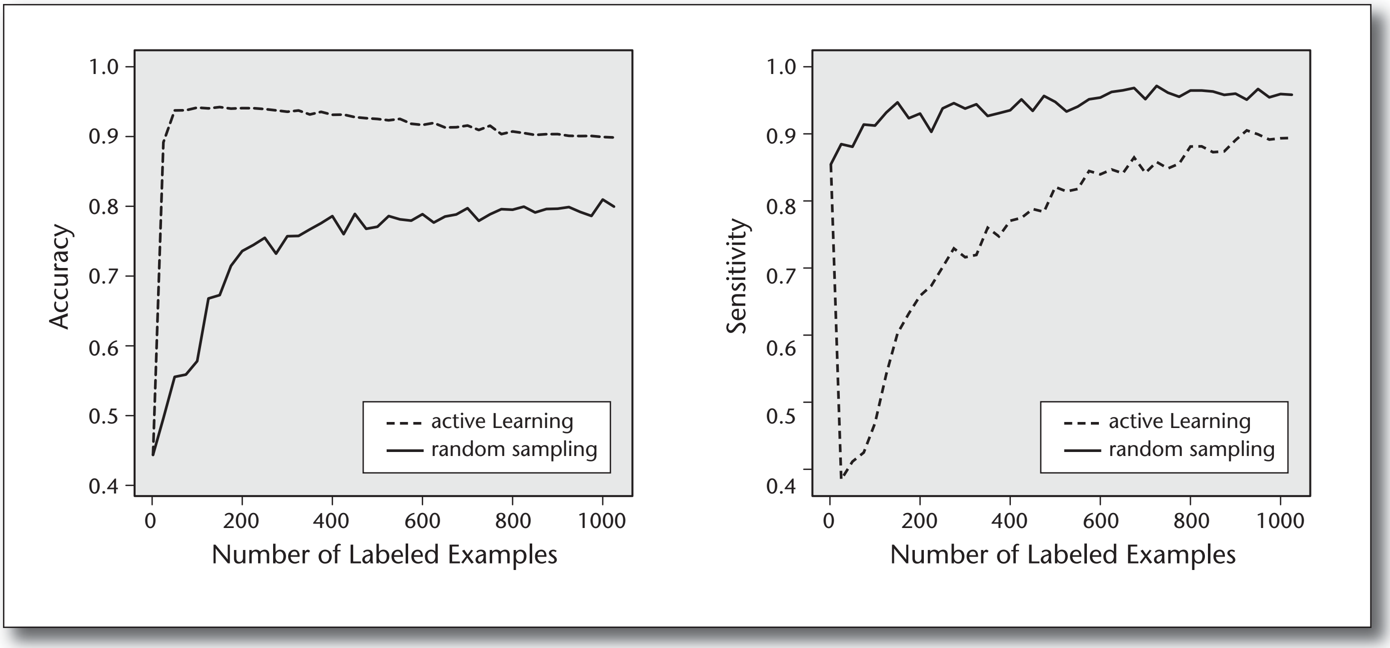

Figure 3. The Performance of Two Different Querying Strategies on a Systematic Review Data Set.

(1) Random sampling, in which labels are requested for instances at random, and (2) the popular active learning technique SIMPLE (Tong and Koller 2000), wherein the learner requests labels for instances about whose label it is most uncertain. On the left, learners are compared as to accuracy (that is, total proportion of correctly classified instances); on the right, sensitivity (accuracy with respect to the minority class of relevant citations). Due to the imbalance in the data set, the two metrics tell very different stories about the relative performance of the respective querying strategies.

the data one is working with, it is easy to draw misleading conclusions by considering the wrong metric. It is impossible to conclude that one learning algorithm is better than another without reference to a classification task with associated misclassification costs.

Fortunately, researchers are beginning to acknowledge the shortcomings of existing learning techniques when they are deployed for real-world tasks. For example, Attenberg and Provost's (2010) recent call-to-arms paper addresses class imbalance, nonuniform misclassification costs, and several other factors that can lead to poor performance of fielded active learning. In our own work, the observation that traditional uncertainty sampling performed poorly in our application led to two developments. First, in light of the above discussion regarding appropriate evaluation metrics, we adapted a method from the field of medical decision making that has been used for diagnostic test assessment (Vickers and Elkin 2006) to elicit a metric from the domain expert reflecting the implicit relative costs they assign to false negatives and false positives. This metric can then be used to compare classifiers in a task-specific way. Second, we developed a novel active learning algorithm that works well under class imbalance by exploiting domain knowledge in the form of labeled terms to guide the active learning process (Wallace et al. 2010).

The strategy of exploiting labeled terms was motivated by our interaction with the reviewers during an ongoing systematic review. When initially experimenting with uncertainty sampling, we would show reviewers citations for which the model was least certain. The reviewers would often remark that the model was being confused by certain things. For example, in one case the model was consistently confounded by published trials that were otherwise relevant to the review, but in which the subjects were mice rather than humans. Unfortunately, there was no way for them to impart this information directly to the model. We therefore developed an algorithm that exploits labeled terms (for example, words like *mice* that indicate a document containing them is likely to be irrelevant) to guide the active learning process. Figure 4 shows the labeled terms for a systematic review of proton beam radiation therapy.

In particular, we build a simple, intuitive model over the labeled phrases in tandem with a linear-kernel support vector machine (Vapnik 1995) over a standard bag-of-words (BOW) representation of the corpus.[3] For the former, we use an odds-ratio based on term counts (that is, the ratio of positive to negative terms in a document). Specifically, suppose we have a set of positive features (that is, *n*-grams indicative of relevance), P^F, and a set of negative features N^F. Then, given a document d to classify, we can generate a crude prediction model for d being relevant, stated as:

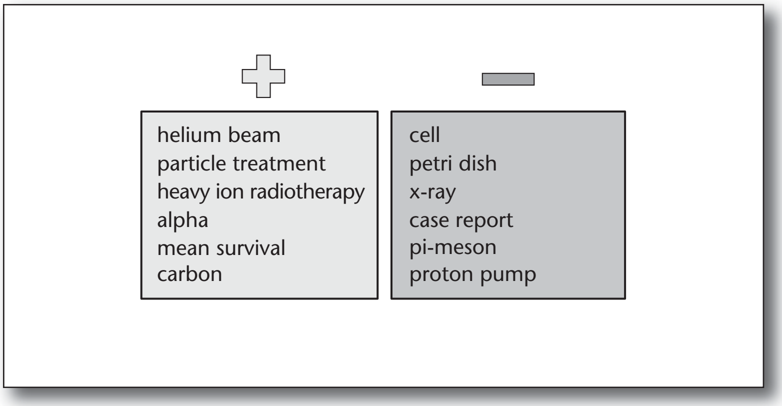

Figure 4. An Example of Labeled Terms for the Proton Beam Data.

The left side shows the terms that if present indicate that the document is relevant, and the right side shows terms likely to be present in irrelevant abstracts.

$$\frac{\sum_{w^+ \in \mathcal{P}^F} I_d(w^+)+1}{\sum_{w^- \in \mathcal{N}^F} I_d(w^-)+1}$$

where $I_d(w)$ is an indicator function that is 1 if w is in d and 0 otherwise. Note that we add pseudocounts to both the negative and positive sums, to avoid division by zero. The value of this ratio relative to 1 gives a class prediction and the magnitude of the ratio gives a confidence.[4] For example, if d contains 10 times as many positive terms as it does negative terms, the class prediction is + and a proxy for our confidence is 10 (note that because we're using an odds-ratio, the confidence can be an arbitrarily large scalar).

Our active learning strategy then makes use of both this simple, semantic or white-box labeled terms model and the black-box SVM model during active learning using a framework known as "co-testing" (Muslea, Minton, and Knoblock 2006), which works as follows. Suppose we have two models, M_1 and M_2. Define contention points as those unlabeled examples about whose labels M_1 and M_2 disagree. Now request the label for one of these points. Because at least one of these models is incorrect, such points are likely highly informative. In our particular case, these points are those about whose labels the black-box SVM model disagrees with what the domain expert has explicitly communicated to us. Intuitively, it is desirable to teach the SVM to make classifications that tend to agree with the information the expert has provided. This approach improves the performance of the system substantially; experts have to label fewer citations to identify the relevant ones than they do using the uncertainty sampling approach. The key point here is that actually working with the domain experts motivated a new method that ultimately reduced workload. Had we been working with only benchmark data sets, we would not have had the labeled terms, and thus could not have developed this approach. This methodology is now used for each new systematic review, with the potential to dramatically reduce workload.

Note that recently, several other researchers in machine learning also recognized the potential benefit of exploiting labeled features in addition to labeled instances for classifier induction. This type of learning scenario is often referred to as dual-supervision. For recent developments, see Attenberg, Melville, and Provost (2010).

Addressing Other Simplifying Assumptions

More recently, we have addressed several other simplifying assumptions that are not valid when applying active learning in a real-world setting. One such assumption is that there is a single, infallible expert who can provide labels at a fixed cost. In our case, however, there is often a group of annotators (for example, medical experts participating in a systematic review) who can provide labels of varying quality and cost: some are experienced reviewers whose time is costly, others are novices whose time is comparatively cheap.

We call this *multiple expert active learning* (Wallace et al. 2011). The task is, given a panel of experts, a set of unlabeled examples, and a budget, who should label which examples? Although recent work has explored somewhat similar scenarios, such as crowd sourcing with Mechanical Turk (Sheng, Provost, and Ipeirotis 2008), our domain differs in an important way — we have a setting in which all annotators must posses a requisite minimum aptitude for annotating instances, precluding the use of low-cost, untrained annotators through crowd sourcing.

A similar problem was explored by Donmez and Carbonell (2008). They proposed a decision-theoretic approach to the task, wherein the instance to be labeled and the expert to do the labeling is picked at each step in active learning in order to maximize local utility (that is, the estimated value of acquiring a label for the selected instance normalized by the cost of the chosen expert). This strategy is intuitively appealing, but has a few drawbacks for our domain. In particular, we found it quite difficult to produce a reasonable value of information for each unlabeled instance. More problematically, we found that, on our data, this greedy strategy tended repeatedly to select cheaper labelers, accruing a noisy training set. This in turn gave rise to comparatively inaccurate models. We had no way to take into account the negative utility of incorrectly labeled instances, particularly because predicting which instances novices would mislabel proved a difficult task (see Wallace et al. [2011] for a more detailed discussion).

While conducting our experiments with real experts, we noticed that novice labelers (reviewers) appeared capable of gauging the reliability of the labels they provided (that is, they knew when they knew). Motivated by this observation, we set out to exploit the metacognitive abilities of the "novice" experts. Specifically, in addition to simply labeling an example, we ask experts to indicate if they are uncertain about their classification. If they are, we pass the example on to a more experienced expert (in our domain, as presumably in most, cost monotonically increases with expertise). In Wallace et al. (2011) we demonstrated that this approach outperforms all other strategies proposed, in terms of cost versus classifier performance. Moreover, we presented empirical evidence that novice labelers are indeed conscious of which examples they are likely to mislabel, and we also argued that automatically identifying difficult instances before labeling is not feasible. Thus we are relying on relatively cheap human computation rather than automated methods, as described in Rebbapragada (2010).

Summary

In summary, while developing a fielded machine-learning system for semiautomating citation screening for systematic reviews we ran into problems when applying off-the-shelf machine-learning technologies. These issues spurred new research questions, that we addressed by working closely with domain experts. Thus far, the Tufts Evidence-Based Practice Center has used our machine-learning methods prospectively to prioritize work when conducting two systematic reviews (that is, we used the model to select the documents that the experts should screen, and also to inform them when they had likely identified all of the relevant citations). At this point in both of the aforementioned reviews, the remaining citations (which the model labeled as irrelevant) were screened by a less costly human (that is, not a doctor). In both cases, our model was correct in its designation of the remaining citations as irrelevant. As we further validate our model empirically, our goal is for experts to trust its decisions enough to eliminate the human screening of irrelevant citations.

We are working on an online tool for citation screening that integrates our methods (active learning, labeling task allocation, and the automatic exclusion of irrelevant citations). We are currently conducting a larger scale empirical evaluation and plan to make our system available to evidence-based practice centers outside of Tufts in Summer 2012.

Cleaning Up after Noisy Labelers

In the preceding section we considered the strategy of collecting additional training data to improve classifier performance. We now turn our attention to a second common cause of poor model generalization accuracy: noisy data. In particular, labeling errors in training data negatively affect classifier accuracy. Experts make errors because of inadequate time spent per example, annotator fatigue, inconsistent labeling criteria, poor data quality, ambiguity among classes, and shifts in decision criteria. For example, when using multispectral satellite imagery data to create maps of global land cover, there are two potential sources of labeling errors: poor image quality and ambiguity between multimodal classes (Friedl et al. 2010). The "Open Shrubland" class in the International Geosphere-Biosphere Programme (IGBP) is bimodal. Both high-latitude tundra and desert cacti qualify for this label. However, high-latitude tundra is spectrally similar to the "Grassland" class and thus "Grassland" and "Open Shrubland" are often confused. In this domain, the experts can consult auxiliary data sources where possible to make a high confidence decision, but this incurs significant additional cost; many of the original labels were assigned prior to the availability of high-quality data sources (for example, Google Earth), whose review could resolve previously ambiguous cases.

In the case of labeling medical abstracts for systematic reviews, as mentioned earlier, experts spend an average of 30 seconds per abstract, often skimming the text to make a decision. To understand why labeling errors are made in this domain, we examined the error rate of a particular expert on a systematic review investigating proton beam cancer therapy. When we asked the expert to perform a second review of certain examples, he discovered that had he read some abstracts more thoroughly, he would have assigned them a different label. This expert admitted feeling fatigue and boredom during the original labeling process, which caused him to relax his criteria on assigning articles to the relevant class. He also noted that because his decision criteria evolved over time, he would have assigned different labels to earlier examples had he gone back and relabeled them. Our physician in this case was an expert, thus demonstrating that even experienced labelers are not immune from making errors in this domain.

Involving Experts to Correct Label Errors

Existing fully automated approaches for mitigating label noise generally look to improve the resulting classifier by avoiding overfitting to the noise in the data. Operationally, such methods often have the goal of removing label noise as a preprocessing step for the training data prior to applying the supervised machine-learning method. Preprocessing methods attempt to detect noisy examples in a single pass of the training data set. Methods differ in how they handle suspected mislabelings. Some discard suspected examples (Zeng and Martinez 2003; Valizadegan and Tan 2007; Gamberger, Lavrac, and Dzeroski 1996; Gamberger, Lavraselj, and Groselj 1999; Brodley and Friedl 1996; 1999; Venkataraman et al. 2004; Verbaeten 2002; Verbaeten and Assche 2003; Zhu, Wu, and Chen 2003), a few attempt to correct them (Zeng and Martinez 2001; Wagstaff et al. 2010), and a few perform a hybrid of the two actions (Lallich, Muhlenbach, and Zighed 2002; Muhlenbach, Lallich, and Zighed 2004; Rebbapragada and Brodley 2007). The primary drawback of "single pass" approaches is that they introduce Type I and II errors. They may eliminate or flip the labels on clean examples, while leaving truly mislabeled examples unchanged.

For both the land-cover and the citation-screening domains, after recognizing that we have labeling errors, we asked the question: how can we involve our expert in this process? Our idea was to leverage the strength of fully automated techniques but outsource the task of relabeling to a domain expert. Our approach, called *active label correction* (ALC), assumes the availability of additional data sources that the original annotator or a second annotator can consult to improve their decision. Fortunately, this is a realistic scenario, even in domains when expert time comes at a premium. For example, in the domain of citation screening, one can retrieve the full text if the abstract has not provided sufficient information to choose a label, and for land-cover classification, there are different procedures for labeling data, some of which take considerably more time and effort than others. For such domains, if the expert is presented a small, focused set of examples that are likely mislabeled, he or she can justify spending the additional time necessary to ensure those examples are properly labeled.

By highlighting the difficult cases for rereview, ALC ensures quality labels for examples that are ambiguous, and potentially more informative for classification. Note that this scenario is different than the one described above in multiple-expert active learning, in which one has several experts and a single pool of unlabeled data. Here we have one expert, either the original or a more qualified alternate, who is prepared to spend the necessary time analyzing select instances that were already labeled using a cheaper procedure suspected of introducing errors.

Cleaning Up the Land-Cover Training Data

In Rebbapragada (2010) we explore many instantiations of ALC, run several experiments on real-world data both within scenarios for which we do and do not know the amount of noise, and apply it to a database of known land-cover types used to train a global land-cover classifier provided by the Department of Geography and Environment at Boston University. The source of the data is the system for terrestrial ecosystem parameterization (STEP) database (Muchoney et al. 1999), which is a collection of polygons drawn over known land-cover types. The database was created from data collected by the moderate resolution imaging spectroradiometer (MODIS). Each observation in this database measures surface reflectance from seven spectral bands, one land-surface temperature band, and one vegetation index (a transformation of the spectral bands that is sensitive to the amount of vegetation present in the pixel). The data has been aggregated by month and includes the mean, maximum, and minimum surface reflectance measurements for one year. The number of MODIS pixels in a site ranges from 1 to 70; each site is given a single class label. The size of the site depends on the landscape homogeneity and is constrained to have an area between 2 and 25 square kilometers.

The labeling scheme is determined by the International Geosphere-Biosphere Programme (Muchoney et al. 1999), which is the consensus standard for the Earth-science modeling community. The IGBP legend contains 17 land-cover classes. Ideally, the land-cover classes should be mutually exclusive and distinct with respect to both space and time. By these criteria, the IGBP scheme is not ideal because it tries to characterize land cover, land use, and surface hydrology, which may overlap in certain areas (in the next section we discuss efforts in how to rethink this classification scheme for the geography community using machine-learning methods). The version of STEP used to produce the MODIS Collection 5 Land Cover product was initially labeled at Boston University during 1998–2007 using various information sources such as Landsat imagery. Since then new information sources have become available such as high-resolution Google Earth imagery, independently produced maps of agriculture, MODIS-derived vegetation indices, and ecoregion maps (Friedl et al. 2010). These new sources of information make it an ideal candidate for ALC.

We partnered with our domain experts to evaluate ALC's ability to find all instances that might be mislabeled. We ran six rounds of ALC; for each round we asked the domain expert to examine the 10 sites for which our method had the lowest confidence on their assigned class labels as computed from using the calibrated probabilities from an SVM.[5] At each round, a new SVM was computed including the (possibly) corrected training data of the last round. Thus our expert examined 60 unique sites in the evaluation.

Our expert analyzed each of the selected sites and provided his personal primary and secondary labeling for each site, the confidence on the composition of the site itself, and some detailed comments. Prior to our involvement, the domain experts had revisited some of the labels on this data set. Thus, some sites had two labels: the "original" label and a "corrected" label. Out of the 60 that our method chose for rereview (using the original labels), 19 were sites that the geographers had already identified as potentially mislabeled. Of those 19, our expert agreed with 12 labels in the "corrected" data. Of the remaining 7, he decided to relabel 4 of them. For 2, although he disagreed with the "corrected" label, he was not confident of the true label. On the last he determined that the "orginal" label was more accurate. Thus in 18 of 19 cases, the expert confirmed that those examples were truly mislabeled. Of the 41 sites that had consistent labels between original and corrected, the expert verified that 16 of these sites were truly mislabeled. Of the other 25 sites, he did make adjustments to 3 in which he found other errors and wanted to revisit later, either for deletion or to regroup the site with other observations. Thus, for this data set ALC was successful in isolating the difficult cases that require more human attention. Informally, the expert told us that "[ALC] is both useful and important to the maintenance of the database. (1) It identifies sites that are very heterogeneous and mixtures of more than one land-cover class. (2) It identifies sites that are mislabeled or of poor quality. I would like to see this analysis on some of the other STEP labels besides the IGBP class since the IGBP classification has a number of problems with class ambiguity."

Summary

In summary, when trying automatically to classify regions of Earth's surface, we needed to address the issue that the available training data had labeling errors. Through discussions with our domain experts we ascertained that more costly, labor-intensive processes were available and that they were willing to reexamine some labels to correct for errors that if left untouched would decrease the ultimate performance of an automated classification scheme. Indeed, experiments illustrated that this was a better use of their time than labeling new data with the less costly procedure (Rebbapragada 2010). Because we were working closely with our domain experts, we developed a new way in which to use human expertise, which we named active label correction. We plan to apply these ideas more extensively to both the citation screening and the land-cover classification domain.

Rethinking Class Definitions

In working with the geographers on trying to correct label noise in their training data, we sat back and brainstormed about the causes of the errors. Even with the ability to use auxiliary information sources, for some sites it is impossible to come up with a single correct label despite significant additional effort. Our conjecture was that the features were not sufficient to make the class discriminations required in the IGBP classification scheme. For this domain, our ability to gather more data is limited by the physical properties of MODIS, thus the avenue of collecting additional data in the form of more features is closed. Indeed, for any real-world domain, before applying a supervised learning method, we must identify the set of classes whose distinction might be useful for the domain expert, and then determine whether these classifications can actually be distinguished by the available data.

Where do class definitions originate? For many data sets, they are provided by human experts as the categories or concepts that people find useful. For others, one can apply clustering methods to find automatically the homogeneous groups in the data. Both approaches have drawbacks. The categories and concepts that people find useful may not be supported by the features (that is, the features may be inadequate for making the class distinctions of interest). Applying clustering to find the homogeneous groups in the data may find a class structure that is not of use to the human. For example, applying clustering to high-resolution CT scans of the lung (Aisen et al. 2003) in order to find the classes of pulmonary diseases may group together in one cluster two pulmonary diseases that have radically different treatments or, conversely, it may split one disease class into multiple clusters that have no meaning with respect to diagnosis or treatment of the disease.

Combining Clustering and Human Expertise

To address this issue, we developed a method for redefining class definitions that leverages both the class definitions found useful by the human experts and the structure found in the data through clustering. Our approach is based on the observation that for supervised training data sets, significant effort went into defining the class labels of the training data, but that these distinctions may not be supported by the features. Our method, named *class-level penalized probabilistic clustering* (CPPC) (Preston et al. 2010), is designed to find the natural number of clusters in a data set, when given constraints in the form of class labels. We require a domain expert to specify an $\mathcal{L} \times \mathcal{L}$ matrix of pairwise probabilistic constraints, where $\mathcal{L}$ is the number of classes in the data. This matrix makes use of the idea that the expert will have different levels of confidence for each class definition, and preferences as to which class labels may be similar. Thus, each element of the matrix defines the expert's belief that a pair of classes should be kept separate or should be merged. Elements on the diagonal reflect the expert's opinion of whether a particular class is multimodal. Using the instances' class labels and the probabilities in the matrix provides us with pairwise probabilistic instance-level constraints.

Our framework for discovering the natural clustering of the data given this prior knowledge uses a method based primarily on the penalized probability clustering (PPC) algorithm (Lu and Leen 2007). Penalized probability clustering is a constraint-based clustering algorithm, in which Lu and Leen defined a generative model to capture the constraints of the domain and the clustering objective. The model allows the user to provide probablistic must- and cannot-link constraints; a must-link constraint between a pair of instances tells PPC to try to put the instances in the same cluster and a cannot-link contraint indicates they should not be in the same cluster. The probability helps the algorithm determine how much weight to place on the constraint. To make the constraint-based clustering tractable they use a mean field approximation (Lu and Leen 2007).[6] In Preston et al. (2010) we provide the details of this work and we provide a framework to evaluate a class clustering model using a modification of the Bayesian information criterion that measures the fit of a clustering solution to both the data and to the specified constraints while adding a term to penalize for model complexity.

Redefining the Land-Cover Classification Scheme

Before we applied CPPC to the land-cover data set, the geographers speculated that (1) large, complex classes such as agriculture may contain several distinguishable subclasses and (2) geographic regions classified as one of the mixed classes, such as "mixed forest" should most likely be merged with either of its two subclasses: "deciduous broadleaf forest" or "evergreen broadleaf forest." In the latter case, some sensors cannot distinguish well between these types of forest. To evaluate CPPC's ability to uncover meaningful class structure, we presented the geographers with two maps built using a land-cover data set for North America. One map was generated using CPPC, the other with expectation maximization (EM) clustering. They were not told which was the new solution. They determined that the map generated using the CPPC clusters defined far more meaningful class structure than the map generated from the EM clusters, and further that the map was better than the map generated by applying supervised learning to the original labeled data. The results showed that CPPC finds areas of confusion where classes should be merged or where separation within a class provides additional useful information, such as corn and wheat clusters within the agriculture class (see figure 5 for an illustrative example of the results).

Based on these results, our geographers are interested in using this tool to redesign the land-cover hierarchy. By observing the results of the constraints, one can better assess what types of actual divisions exist in the data and whether or not our desired class descriptions are feasible distinctions (that is, can actually be distinguished by the features).

Summary

Our work in redefining the labeling scheme for land-cover classification arises from an observation that even when the experts tried to correct for labeling errors in their data set, it was not possible to create an accurate classifier from the data. The reason is that the features do not support the classes defined by the IGBP classification scheme. This led to research on how to redefine this scheme using the best of both worlds: the work already done on examining each training data point to come up with a land-cover label and the ability of unsupervised learning to uncover structure in the data. Thus our ideas resulted from working on an application for which off-the-shelf machine-learning methods failed to give the acceptable performance and by considering novel ways in which to use our domain expertise. We have begun to work with two groups of doctors to apply some of these same ideas to develop a classification scheme for grouping clinical trials and for developing a classification scheme for chronic obstructive pulmonary disease (COPD).

Conclusions

A primary objective in our research is to consider how we can best utilize domain expertise to improve the performance of machine-learning and data-mining algorithms. This article describes two such collaborations, which led to methods for utilizing human input in the form of imparting additional domain knowledge (that is, terms indicative of class membership), correction of labeling errors, and the use of training labels and expert knowledge to guide the redefinition of class boundaries. Some of our previous research efforts led to the ideas of human-guided feature selection for unsupervised learning (Dy and Brodley 2000) as applied to content-based image retrieval for diagnosing high-resolution CT images of the lung; determining what experiments should be run next when the process of generating the data cannot be separated from the process of labeling the data (Lomasky et al. 2007) for the generation of training data for an artificial nose; and user-feedback meth-

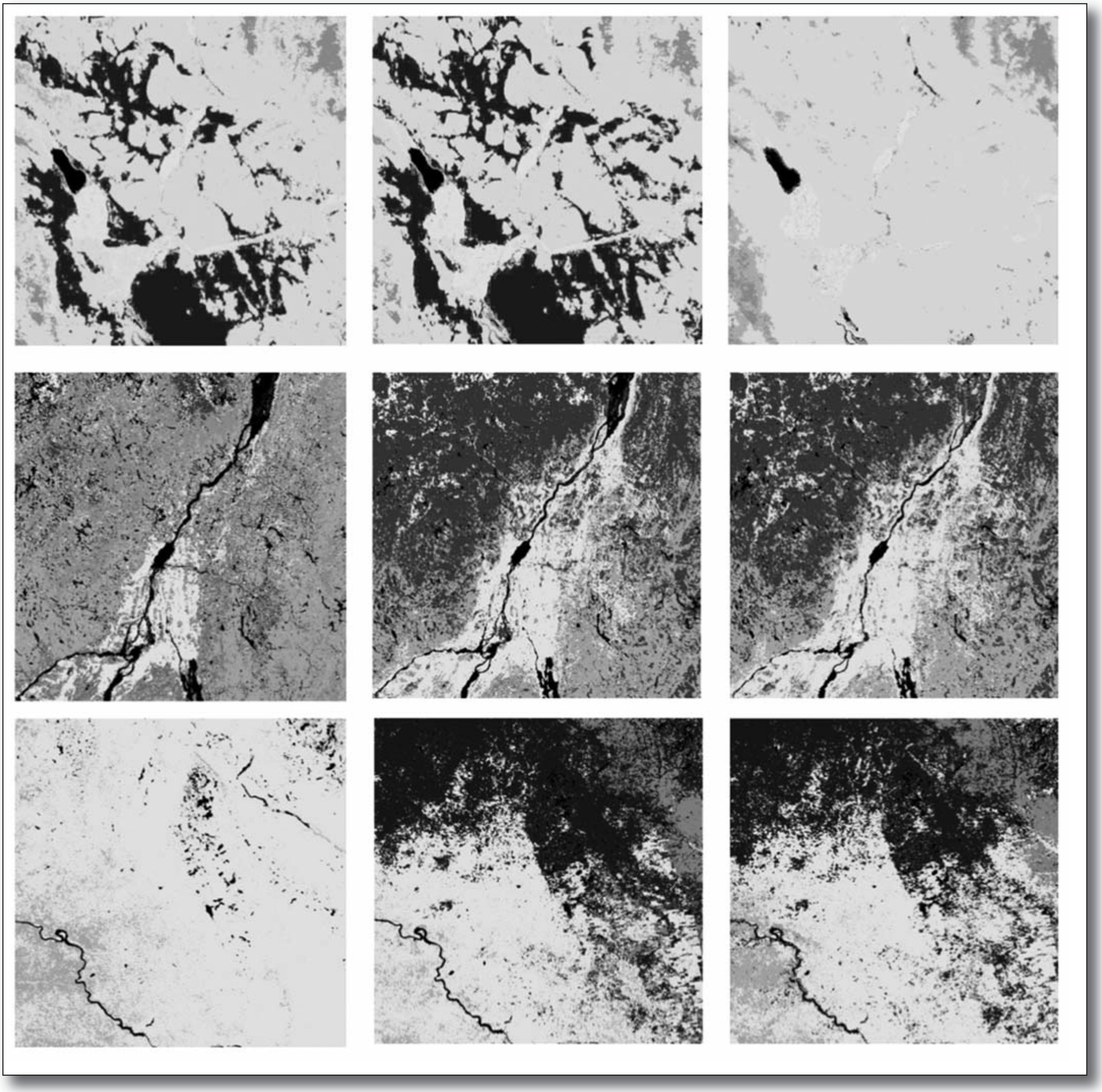

Figure 5. Maps Using Original IGBP Classification (left), CPPC Clustering (middle), and EM Clustering (right).

The top row shows a map tile centered on the the agricultural oasis just south of the Salton Sea in southern California. It represents a region with very sparse desert shrub cover (darker gray) and some agriculture (light gray) and denser forest/grassland (medium gray). The original IGBP classification contains a barren or sparsely vegetated class that is similar (spectrally) to the open shrubland class and is preserved in the clustering with constraints but not the EM clustering. In this case, CPPC preserves the class structure of interest. The middle row shows an area centered on Montreal. It represents a transition between southern hardwoods (deciduous broadleaf) and boreal conifers (evergreen needleleaf) forests. It is also an area of transition between forest and the agricultural region along the river. The original IGBP classification has two "mixture" classes that mostly disappear after clustering: mixed forest and agriculture/natural mosaic (mixture of trees and agriculture). In this case CPPC correctly decides to merge classes. The bottom row shows an area on the border of North and South Dakota in the west with Minnesota to the east. It is almost completely covered with agriculture with a small patch of natural prairie to the southwest of the image. The original IGBP classification has one single agriculture class (light gray). As we see with the two clustering results there are actually three distinct agricultural signals within this area, which indicate different types of crops, different harvesting times, or the presence/absence of irrigation. In this case, CPPC is able to find subclasses withing a single class long thought to be multimodal.

ods for anomaly detection of astrophysics data to find anomalous sky objects.

Clearly theory and algorithm development are both crucial for the advancement of machine learning, but we would argue strongly that being involved in an application can highlight shortcomings of existing methodologies and lead to new insights into previously unaddressed research issues. To conclude, we are motivated by the belief that contact with real problems, real data, real experts, and real users can generate the creative friction that leads to new directions in machine learning.

Acknowledgments

We would like to thank our domain experts Mark Friedl and Damien Sulla-Menashe from the Department of Geography and Environment at Boston University, and Tom Trikalinos from Tufts Medical Center at Tufts University for their willingness to spend countless hours with us. We would like to thank Roni Khardon and Dan Preston for their work in helping to develop the CPPC algorithm. We would like to thank Roni Khardon, Foster Provost, and Kiri Wagstaff for reading and commenting on this work. Our research was supported by NSF IIS-0803409, IIS-0713259, and R01HS0 18494. The research was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. US Government sponsorship is acknowledged.

Notes

[1] The exception is cystic fibrosis, which affects mostly the young.

[2] PubMed is a repository of published biomedical papers. Sometimes other databases are used in addition to, or instead of, PubMed.

[3] Bag-of-Words is a featurespace encoding for textual data in which documents are mapped to vectors whose entries are functions over word counts. The simplest such representation is binary BOW, in which each column represents a particular word and the value of this column for a given document is set to 1 or 0, indicating that the word is present or not, respectively.

[4] In order to ensure that the magnitude is symmetric in the respective directions, one may either flip the ratio so that the numerator is always larger than the denominator, or one may take the log of the ratio.

[5] The SVM used a polynomial kernel (order 2) where the C parameter was set to 1.0.

[6] A straightforward implementation of the PPC algorithm using mean field approximation (Jaakkola 2000) results in $O(N^2)$ computational complexity due to the number of constraints (that is, one constraint for each pair of instances), where N is the total number of instances. In CPPC, we reduce this time complexity to $O(NL)$ by taking advantage of the repetitions in the set of instance pairwise constraints that are induced from the class pairwise constraints. In general, $L << N$, as the number of classes is generally much smaller than the number of instances.

References

Aisen, A. M.; Broderick, L. S.; Winer-Muram, H.; Brodley, C. E.; Kak, A. C.; Pavlopoulou, C.; Dy, J.; and Marchiori, A. 2003. Automated Storage and Retrieval of Medical Images to Assist Diagnosis: Implementation and Preliminary Assessment. *Radiology* 228(1): 265–270.

Attenberg, J., and Provost, F. 2010. Why Label When You Can Search?: Alternatives to Active Learning for Applying Human Resources to Build Classification Models Under Extreme Class Imbalance. In *Proceedings of the 16th ACM SIGKDD International cConference on Knowledge Discovery and Data Mining,* 423–432. New York: Association for Computing Machinery.

Attenberg, J.; Melville, P.; and Provost, F. 2010. A Unified Approach to Active Dual Supervision for Labeling Features and Examples. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases,* 40–55. Berlin: Springer.

Brodley, C. E., and Friedl, M. A. 1996. Identifying and Eliminating Mislabeled Training Instances. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence,* 799–805. Palo Alto, CA: AAAI Press.

Brodley, C. E., and Friedl, M. A. 1999. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research* 11: 131–167.

Brodley, C. E., and Smyth, P. J. 1997. Applying Classification Algorithms in Practice. *Statistics and Computing* 7(1): 45–56.

Brodley, C. E.; Kak, A. C.; Dy, J. G.; Shyu, C. R.; Aisen, A.; and Broderick, L. 1999. Content-Based Retrieval from Medical Image Databases: A Synergy of Human Interaction, Machine Learning, and Computer Vision. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence,* 760–767. Palo Alto, CA: AAAI Press.

Donmez, P., and Carbonell, J. G. 2008. Proactive Learning: Cost-Sensitive Active Learning with Multiple Imperfect Oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management,* 619–628. New York: Association for Computing Machinery.

Draper, B. A.; Brodley, C. E.; and Utgoff, P. E. 1994. Goal-Directed Classification Using Linear Machine Decision Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(9): 888–893.

Drummond, C.; Holte, R.; et al. 2005. Severe Class Imbalance: Why Better Algorithms Aren't the Answer. In *Proceedings of the 16th European Conference of Machine Learning and Knowledge Discovery in Databases,* 539–546. Berlin: Springer.

Dy, J., and Brodley, C. E. 2000. Visualization and Interactive Feature Selection for Unsupervised Data. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 360–364. New York: Association for Computing Machinery.

Dy, J.; Brodley, C.; Kak, A.; Pavlopoulo, C.; Aisen, A.; and Broderick, L. 2003. The Customized-Queries Approach to Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(3): 373–378.

Early, J. P.; Brodley, C. E.; and Rosenberg, K. 2003. Behavioral Authentication of Server Flows. In *Proceedings of the 19th Annual IEEE Computer Security Applications Confer-*

ence, 49–55. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Fern, X. Z.; Brodley, C. E.; and Friedl, M. A. 2005. Correlation Clustering for Learning Mixtures of Canonical Correlation Models. In *Proceedings of the Fifth SIAM International Conference on Data Mining,* 439–448. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Friedl, M. A.; Brodley, C. E.; and Stahler, A. 1999. Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales. *IEEE Transactions on Geoscience and Remote Sensing* 37(2): 969–977.

Friedl, M. A.; Sulla-Menashe, D.; Tan, B.; Schneider, A.; Ramankutty, N.; Sibley, A.; and Huang, X. 2010. MODIS Collection 5 Global Land Cover: Algorithm Refinements and Characterization of New Datasets. *Remote Sensing of Environment* 114(1): 168–182.

Gamberger, D.; Lavrac, N.; and Dzeroski, S. 1996. Noise Elimination in Inductive Concept Learning: A Case Study in Medical Diagnosis. In *Algorithmic Learning Theory: Seventh International Workshop,* Lecture Notes in Computer Science 1160, 199–212. Berlin: Springer.

Gamberger, D.; Lavrac, C.; Groselj, C. 1999. Experiments with Noise Filtering in a Medical Domain. In *Proceedings of the Sixteenth International Conference on Machine Learning,* 143–151. San Francisco: Morgan Kaufmann Publishers.

He, H., and Garcia, E. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21(9): 1263–1284.

Jaakkola, T. S. 2000. Tutorial on Variational Approximation Methods. In *Advanced Mean Field Methods: Theory and Practice,* 129–159. Cambridge, MA: The MIT Press.

Kapadia, N.; Fortes, J.; and Brodley, C. 1999. Predictive Application-Performance Modeling in a Computational Grid Environment. In *Proceedings of the 8th IEEE International Symposium on High Performance Distributed Computing* (HPDC 99), 47–54. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Lallich, S.; Muhlenbach, F.; and Zighed, D. A. 2002. Improving Classification by Removing or Relabeling Mislabeled Instances. In *Proceedings of the Thirteenth International Symposium on the Foundations of Intelligent Systems,* 5–15. Berlin: Springer-Verlag.

Lane, T., and Brodley, C. E. 1999. Temporal Sequence Learning and Data Reduction for Anomaly Detection. *ACM Transactions on Computer Security* 2(3): 295–331.

Lomasky, R.; Brodley, C.; Aerneke, M.; Walt, D.; and Friedl, M. 2007. Active Class Selection. In *Proceedings of the Eighteenth European Conference on Machine Learning and Knowledge Discovery in Databases,* 640–647. Berlin: Springer.

Lu, Z., and Leen, T. K. 2007. Penalized Probabilistic Clustering. *Neural Computation* 19(6): 1528–1567.

MacArthur, S.; Brodley, C. E.; and Broderick, L. 2002. Interactive Content-Based Image Retrieval Using Relevance Feedback. *Computer Vision and Image Understanding* 88(3): 55–75.

McCallum, A., and Nigam, K. 1998. Employing EM and Pool-Based Active Learning for Text Classification. In *Proceedings of the Fifteenth International Conference on Machine Learning* (ICML), 350–358. San Francisco: Morgan Kaufmann.

Moss, E.; Utgoff, P.; Cavazos, J.; Precup, D.; Stefanovic, D.; Brodley, C.; and Scheeff, D. T. 1997. Learning to Schedule Straight-Line Code. In *Advances in Neural Information Processing Systems 9,* 929–935. Cambridge, MA: The MIT Press.

Muchoney, D.; Strahler, A.; Hodges, J.; and LoCastro, J. 1999. The IGBP Discover Confidence Sites and the System for Terrestrial Ecosystem Parameterization: Tools for Validating Global Land Cover Data. *Photogrammetric Engineering and Remote Sensing* 65(9): 1061–1067.

Muhlenbach, F.; Lallich, S.; and Zighed, D. A. 2004. Identifying and Handling Mislabelled Instances. *Journal of Intelligent Information Systems* 22(1): 89–109.

Muslea, I.; Minton, S.; and Knoblock, C. A. 2006. Active Learning with Multiple Views. *Journal of Artificial Intelligence Research* 27: 203–233.

Preston, D.; Brodley, C. E.; Khardon, R.; Sulla-Menashe, D.; and Friedl, M. 2010. Redefining Class Definitions Using Constraint-Based Clustering. In *Proceedings of the Sixteenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 823–832. New York: Association for Computing Machinery.

Provost, F. 2000. Machine Learning from Imbalanced Data Sets 101. In *Learning from Imbalanced Data Sets: Papers from the AAAI Workshop*. AAAI Technical Report WS-00-05. Palo Alto, CA: AAAI Press.

Pusara, M., and Brodley, C. E. 2004. User ReAuthentication Via Mouse Movements. In *Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security.* New York: Association for Computing Machinery.

Rebbapragada, U. 2010. Strategic Targeting of Outliers for Expert Review. Ph.D. Dissertation, Tufts University, Boston, MA.

Rebbapragada, U., and Brodley, C. E. 2007. Class Noise Mitigation Through Instance Weighting. In *Proceedings of the 18th European Conference on Machine Learning (ECML) and the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases* (PKDD), Lecture Notes in Computer Science Volume 4701, 708–715. Berlin: Springer.

Rebbapragada, U.; Lomasky, R.; Brodley, C. E.; and Friedl, M. 2008a. Generating High-Quality Training Data for Automated Land-Cover Mapping. In *Proceedings of the 2008 IEEE International Geoscience and Remote Science Symposium,* 546–548. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Rebbapragada, U.; Mandrake, L.; Wagstaff, K.; Gleeson, D.; Castao, R.; Chien, S.; and Brodley, C. E. 2009. Improving Onboard Analysis of Hyperion Images by Filtering Mislabeled Training Data Examples. In *Proceedings of the 2009 IEEE Aerospace Conference*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Rebbapragada, U.; Protopapas, P.; Brodley, C. E.; and Alcock, C. 2008b. Finding Anomalous Periodic Time Series: An Application to Catalogs of Periodic Variable Stars. *Machine Learning* 74(3): 281.

Sackett, D.; Rosenberg, W.; Gray, J.; Haynes, R.; and Richardson, W. 1996. Evidence Based Medicine: What It Is and What It Isn't. *BMJ Journal* 312(7023): 71.

Settles, B. 2009. Active Learning Literature Survey. Technical Report 1648, Department of Computer Science,University of Wisconsin–Madison, Madison, WI.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), 614–622. New York: Association for Computing Machinery.

Shyu, C.; Brodley, C. E.; Kak, A.; Kosaka, A.; Aisen, A.; and Broderick, L. 1999. Assert: A Physician-in-the-Loop Content-Based Image Retrieval System for HRCT Image Databases. *Computer Vision and Image Understanding* 75(1–2): 111–132.

Stough, T., and Brodley, C. E. 2001. Focusing Attention on Objects of Interest Using Multiple Matched Filters. *IEEE Transactions on Image Processing* 10(3): 419–426.

Tong, S., and Koller, D. 2000. Support Vector Machine Active Learning with Applications to Text Classification. In *Journal of Machine Learning Research* 2: 999–1006.

Valizadegan, H., and Tan, P.-N. 2007. Kernel-Based Detection of Mislabeled Training Examples. In *Proceedings of the Seventh SIAM International Conference on Data Mining,* 309–319. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory.* Berlin: Springer.

Venkataraman, S.; Metaxas, D.; Fradkin, D.; Kulikowski, C.; and Muchnik, I. 2004. Distinguishing Mislabeled Data from Correctly Labeled Data in Classifier Design. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence,* 668–672. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Verbaeten, S. 2002. Identifying Mislabeled Training Examples in ILP Classification Problems. In *Proceedings of the Twelfth Belgian-Dutch Conference on Machine Learning.* Utrecht, The Netherlands: University of Utrecht, Faculty of Mathematics and Computer Science.

Verbaeten, S., and Assche, A. V. 2003. Ensemble Methods for Noise Elimination in Classification Problems. In *Multiple Classifier Systems, 4th International Workshop*, Volume 4. Berlin: Springer.

Vickers, A. J., and Elkin, E. B. 2006. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making* 26(6): 565–574.

Wagstaff, K.; Kocurek, M.; Mazzoni, D.; and Tang, B. 2010. Progressive Refinement for Support Vector Machines. *Data Mining and Knowledge Discovery* 20(1): 53–69.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2010. Active Learning for Biomedical Citation Screening. In *Proceedings of the Seventeenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 173–182. New York: Association for Computing Machinery.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2011. Who Should Label What? Instance Allocation in Multiple Expert Active Learning. In *Proceedings of the Eleventh SIAM International Conference on Data Mining* (SDM), 176–187. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Zeng, X., and Martinez, T. 2001. An Algorithm for Correcting Mislabeled Data. *Intelligent Data Analysis* 5(1): 491–502.

Zeng, X., and Martinez, T. R. 2003. A Noise Filtering Method Using Neural Networks. In *Proceedings of the 2003 IEEE International Workshop of Soft Computing Techniques in Instrumentation, Measurement, and Related Applications.*Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Zhu, X.; Wu, X.; and Chen, S. 2003. Eliminating Class Noise in Large Datasets. In *Proceedings of the TwentiethInternational Conference on Machine Learning,* 920–927. Palo Alto, CA: AAAI Press.

Carla E. Brodley is a professor and chair of the Department of Computer Science at Tufts University. She received her PhD in computer science from the University of Massachusetts at Amherst in 1994. From 1994–2004, she was on the faculty of the School of Electrical Engineering at Purdue University. She joined the faculty at Tufts in 2004 and became chair in September 2010. Professor Brodley's research interests include machine learning, knowledge discovery in databases, health IT, and personalized medicine. She has worked in the areas of intrusion detection, anomaly detection, classifier formation, unsupervised learning, and applications of machine learning to remote sensing, computer security, neuroscience, digital libraries, astrophysics, content-based image retrieval of medical images, computational biology, chemistry, evidence-based medicine, and personalized medicine. In 2001 she served as program cochair for ICML and in 2004 as general chair. She is on the editorial boards of *JMLR, Machine Learning*, and *DKMD*. She is a member of the AAAI Council, and is a board member of IMLS and of the CRA Board of Directors.

Umaa Rebbapragada received her B.A. in mathematics from the University of California, Berkeley, and her M.S. and Ph.D. in computer science from Tufts University. She currently works in the Machine Learning and Instrument Autonomy group at the Jet Propulsion Laboratory of the California Institute of Technology. Her research interests are machine learning and data mining, with specific focus on anomaly detection and supervised learning in the presence of noisy labels and imperfect experts. She has applied her research to problems in astrophysics, land-cover classification, and earthquake damage assessment.

Kevin Small is a research scientist within the Institute for Clinical Research and Health Policy Studies at Tufts Medical Center. He received his Ph.D. in computer science from the University of Illinois at Urbana-Champaign in 2009. His research interests are in machine learning with an emphasis on scenarios where there is an interaction with a domain expert during the learning procedure and methods that exploit known interdependencies between multiple learned classifiers (such as structured learning). He has applied his research primarily to information extraction and related natural language tasks, with a recent focus on health informatics domains.

Byron Wallace received his B.S. in computer science from the University of Massachusetts, Amherst, and his M.S. in computer science from Tufts University. He is currently a Ph.D. candidate in computer science at Tufts. He is also research staff in the Institute for Clinical Research and Health Policy Studies at Tufts Medical Center. His research interests are in machine learning with a focus on its applications in health informatics.