ASKING FOR A SECOND OPINION: RE-QUERYING OF NOISY MULTI-CLASS LABELS

Jack W. Stokes[⋆] *Ashish Kapoor*[⋆] *Debajyoti Ray*[†]

[⋆] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
[†] VideoAmp Inc., 4500 Via Marina, Marina del Rey, CA 90292, USA

In this paper, we propose a new maximum margin-based, active learning algorithm for identifying incorrectly labeled training data. The algorithm combines a round-robin approach for investigating each class with a simple, yet effective ranking metric called maximum negative margin (MNM). Samples are given to an expert for re-evaluation to determine if they are indeed mislabeled. We also propose using five active learning metrics, including uncertainty sampling with margin sampling (USMS) and minimum margin, for the noisy label task which have previously been used in the standard active learning setting for identifying new samples to label. USMS is very competitive with maximum negative margin. In addition, we consider other information theoretic objective criteria for this new task including uncertainty sampling with entropy, query-by-committee with voting entropy, and K-nearest neighbor with voting entropy, but these consistently perform worse than MNM and USMS. The MNM noisy label active learning algorithm can be useful in several different scenarios including data cleansing as a preprocessing step before training and identifying mislabeled examples in the test set.

Index Terms— Active Learning, Data Cleansing

1. INTRODUCTION

In real-world, large-scale applications of supervised learning, one or more experts, who provide labels, interact with the classifier model. Typically, in a system that is deployed and used over many years, the experts also learn over time. As an expert investigates new failure cases, she may change her mind as to which label to give to an item. In the case of multiple experts, the first may assign one label to an item while another provides a completely different label to an identical but separate item. Thus, the algorithm has to cope with label noise due to the fact that the experts may not always be right, may change their mind over time, or may not agree. Instead of relying solely on experts to provide labels, a separate system may assign labels using automated classifiers or rule-based approaches which can serve as an additional source of label noise. Under these scenarios, the mislabeled samples can reduce the efficacy of the final trained classification system. As we demonstrate in Section 3, mislabeled samples can significantly reduce a classifier's accuracy compared to a similar classifier trained using correctly labeled data.

The primary goal of this work is to improve the classification accuracy of a multiclass linear classifier. To do so, we indicate potentially mislabeled samples in the training set to the human analyst for correction, although the same algorithm can also be applied to the test set. In this paper, we take an active learning approach by assigning a metric score to each sample to rank for investigation and possible relabeling. Samples are ranked for an expert which are most likely to be mislabeled and the system seeks confirmation that the sample's label is indeed noisy. In active learning, the algorithm pays a cost to obtain labels (in terms of the expert's time, for example), and the goal is to pay the minimum cost to obtain *informative* labels in order to achieve a high classification accuracy [1]. Similar to the standard active learning task of asking experts to label new samples for training, we show that active learning reduces the number of candidate samples which need to be re-investigated for labeling errors.

As noted in Section 4, there have been two previous efforts which utilized active learning for the noisy label problem [2, 3]. In [2], the information criteria is a function of the optimum margin classifier used by Guyon *et al.* and the results are presented for binary classification. In [3], the approach is based on ranking samples according to the misclassification cost, $1 - P(y|\mathbf{x})$, again in the binary setting. In this paper, we present several solutions to the multiclass problem which represents most large-scale supervised learning tasks. We also provide a solution for any linear classifier instead of requiring a specific type of classification algorithm which is the case for [2]. It is important to note that while the proposed algorithm uses a linear classifier for identifying mislabeled samples, it can also be to used to first clean a training set which can then be used to train a nonlinear classifier such as a neural network or decision tree.

In contrast to the misclassification cost for binary classification, we introduce a simple metric called maximum negative margin (MNM) in Section 2 as the active learning objective for identifying label noise in multiclass data. A hyperplane corresponding to an individual class (i.e. label) in a multiclass linear classifier is directional; examples with a positive score are predicted to belong to that particular class. The best performing MNM metric uses the intuitive observation that a mislabeled sample is likely to be located on the opposite side of the hyperplane corresponding to the sample's label. In addition, potentially mislabeled examples located the furthest distance from the hyperplane may contribute more to improving the resulting classifier trained with the corrected data. In addition to the minimum negative margin algorithm, we also *propose* and evaluate using several other common active learning metrics, including uncertainty sampling with margin sampling (USMS), uncertainty sampling with label entropy (USE), minimum margin (MM), Query-By-Committee with voting entropy (QBC), and K-Nearest Neighbor with label entropy (KNN), in the *new* setting of noisy label identification. Instead of ranking items with the highest active learning score which can cause many samples to be returned for a single class, we instead propose a round-robin approach which identifies the next most anomalous sample for each class for all of the metrics proposed in this paper.

2. ACTIVE LEARNING WITH NOISY LABELS

A classifier trained with noisy labels can provide dramatically different results than a classifier that is trained with correctly labeled data. Figure 1 illustrates the effect of the mislabeled example given

by the maroon "+" on the left side of the θ_+ decision boundary (i.e. indicated as "Mislabeled Example") which corresponds to the "+" class. The θ_+ decision boundary shifts significantly after the mislabeled point is assigned its true label. Thus, an algorithm, such as generalized binary search (GBS) [4], that assumes all labels are correct, can generalize poorly to unseen test data. We summarize one

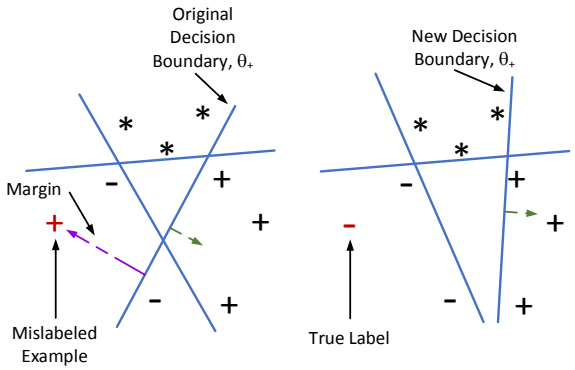


Fig. 1. (Left) One of the points in the $(-)$ category is mislabeled as $(+)$, leading to the learned θ_+ decision boundary. (Right) The new θ_+ decision boundary shifts significantly after obtaining the true label $(-)$.

iteration of the round-robin, active learning algorithm for identifying mislabeled samples in Figure 2. In step 1, several of the algorithms (MNM, MM, USE, USMS) first retrain a linear classifier on the previously updated training set including all corrected labels. Consider a set of samples $\mathbf{x} \in \mathcal{R}^N$ with multiclass labels $y \in \{1, \cdots, Y\}$ where $|Y|$ is the maximum number of classes in the dataset. For the experiments in the following section we use a logistic regression classifier trained with one-versus-all minimizing the cross-entropy loss function, but other linear classifiers such as the support vector machine (SVM) can also be used [5]. For QBC [6], an ensemble of classifiers is trained with random subsets of the training data. In our work, we consider subsets containing 50% of the samples in the training set. For KNN, the method does not require a new classifier to be trained on all of the corrected training data for each round, but one can be trained in order to evaluate how well this improved classifier performs on a holdout test set.

In step 2, we evaluate the posterior probability $P(y|\mathbf{x})$ for all samples in the training set, which were not investigated in earlier rounds, using the classifier trained in step 1 for USE or USMS and the individual posterior probabilities for each classifier in the ensemble for QBC. In the next step, these posterior probabilities are then used to evaluate the objective functions for these algorithms' metrics. In step 3, compute each algorithm's metric, $U(\mathbf{x})$, for all algorithms and rank the largest metrics in step 4 for all remaining labeled samples which have not been re-examined. Next in step 5, one training example is selected in a round-robin fashion from each class according to the largest metric of examples labeled with that class. In step 6, the total number of examples selected in this round is updated from the results in step 5. If more examples are needed to meet the desired number of examples in this iteration (e.g. 100), step 5 is repeated. After the desired number of samples in the iteration have been selected, the samples are given to the expert for investigation.

Maximum Negative Margin: Inspired by [16], we propose the maximum negative margin (MNM) active learning metric for investigating potentially mislabeled samples. By definition, the margin is the distance from an example to the decision boundary (i.e. hyperplane) with parameters $\theta_{y'}$ where $\theta_{y'}$ is chosen to correspond to each distinct label y'. We conjecture for the MNM algorithm that samples which are mispredicted (i.e. have a negative margin) and located the furthest distance away from the hyperplane are good candidates for investigation. Results in Section 3 provide empirical evidence to support this hypothesis. For the MNM active learning metric, we rank the samples according to

$$U(\mathbf{x}_j) = -(\theta \cdot \mathbf{x}_j)/||\theta||_2. \quad (1)$$

As with each of the remaining active learning metrics, we rank the samples with the largest scores, $U(x_j)$, for the analyst to investigate.

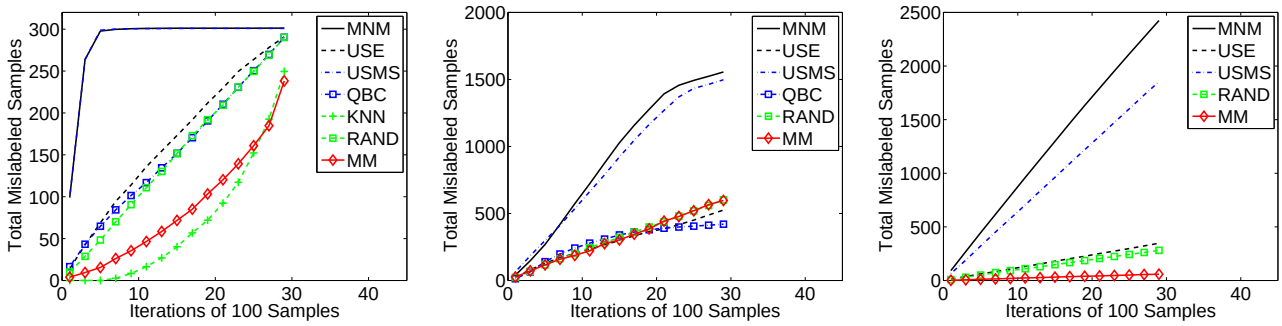
New Algorithms with Previous Metrics: In Table 1, we discuss several metrics which have been previously discussed in the standard active learning setting but are now proposed for the task of selecting potentially mislabeled samples. The minimum margin (MM) metric was originally suggested in [7], and the goal is to identify mislabeled samples near the decision boundary. The uncertainty

1. For MNM, MM, USE or USMS, train a classifier on the labeled samples. For QBC, compute the ensemble of classifiers on random subsets of the training data. For KNN, an optional classifier can be learned to evaluate the performance on a holdout test set.
2. For USE or USMS, evaluate the classifier for the labeled samples which have not been investigated in earlier iterations. For QBC, evaluate each classifier in the ensemble on the remaining labeled samples, j.
3. Compute the objective function $U(\mathbf{x}_j)$ ((1), Table 1) for the labeled training example j.
4. Rank the metric for each sample in the remaining training set according to the largest $U(\mathbf{x}_j)$.
5. Select the next samples to be re-investigated chosen as follows. For each class y',
   (a) Choose one sample with the largest $U(\mathbf{x}_j)$ among the examples labeled as y'.
   (b) For the USMS metric, if not enough samples for a class are found from 5(a), select the labeled sample with the next highest output probability $P(y_i|\mathbf{x}_j)$ corresponding to the desired class y_i.
6. Repeat step 5 until the desired number of samples have been selected for review in this iteration.

Fig. 2. One iteration of the proposed mislabeled sample active learning algorithm as pseudo-code.

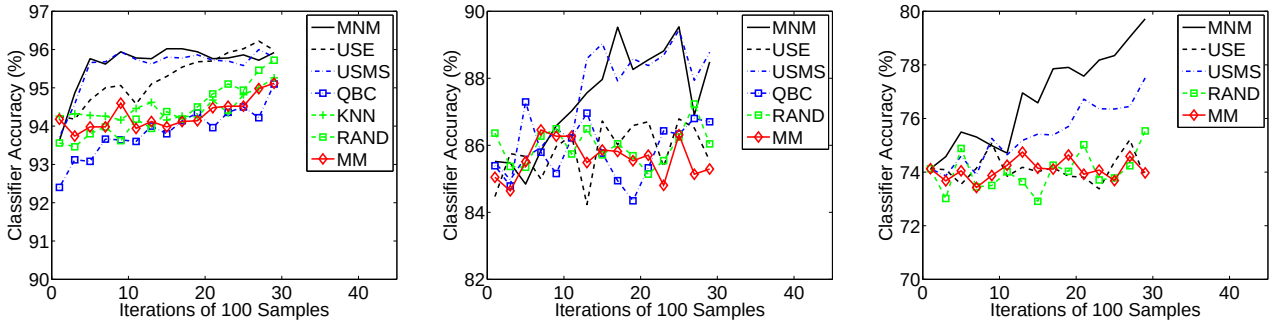
Alg	Metric
MM	$U(\mathbf{x}_j) = -\lvert\theta \cdot \mathbf{x}_j\rvert/\lvert\lvert\theta\rvert\rvert_2$
USE	$U(\mathbf{x}_j) = -\sum_{y' \in Y} P_\theta(y'\lvert\mathbf{x}_j) \log P_\theta(y'\lvert\mathbf{x}_j)$
USMS	$U(\mathbf{x}_j) = \min_{m,n \neq j} \lvert P(y_m\lvert\mathbf{x}_j) - P(y_n\lvert\mathbf{x}_j)\rvert$
QBC	$U(\mathbf{x}_j) = -\sum_{y' \in Y} \frac{V(y',\mathbf{x}_j)}{C} \log \frac{V(y',\mathbf{x}_j)}{C}$
KNN	$U(\mathbf{x}_j) = -\sum_{y' \in Y} \frac{V(y',N_K)}{K} \log \frac{V(y',N_K)}{K}$

Table 1. Different active learning metrics proposed for the noisy label sample detection task.



(a) Digits (b) Shuttle (c) MNIST

Fig. 3. Cumulative number of correctly identified mislabeled samples for each iteration of candidate samples.



(a) Digits (b) Shuttle (c) MNIST

Fig. 4. Classifier accuracy on a holdout test set for each iteration of candidate samples.

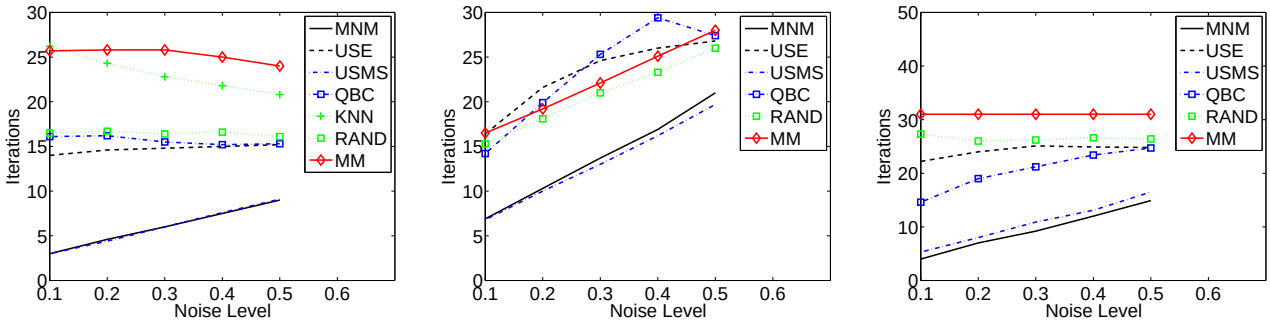
sampling with label entropy (USE) algorithm [8] assigns a value based on label uncertainty, i.e. where the classifier is most uncertain about the label that it assigns to a point. A variation of the USE metric is uncertainty sampling with margin sampling (USMS) [9]. Instead of considering the entropy with respect to all classes, USMS only considers the two classes with the highest posterior probabilities. Query-By-Committee with voting entropy (QBC), originally proposed by Dagan *et al.* [6] for the standard active learning task, computes the entropy of the percentage of votes belonging to each class. For QBC, C is the total number of committee members and $V(y', \mathbf{x}_j)$ is the number of committee members which predict the sample has label y'. In the K-nearest neighbor with label entropy (KNN) metric [10], we instead compute the entropy of the previously assigned labels of nearby samples. For KNN, N_K are the K-nearest neighbors of $\mathbf{x}_j$, and $V(y', N_K)$ is the number of samples with label y'. The distance measure used to select the K-nearest neighbors is the Euclidean distance for all datasets used in Section 3.

3. EXPERIMENTAL RESULTS

In this section, we compare the performance of the different metrics proposed above for detecting incorrectly labeled samples. We begin by describing the datasets used in this study. We then compare MNM with algorithms previously proposed for standard active learning for this new task of discovering noisy sample labels.

For our analysis, we consider the two UCI datasets [11] including Digits (3000/500) and Shuttle (30000/10000), and the MNIST dataset (50000/10000) [12] where (N_{train}/N_{test}) are the size of the randomly selected training and holdout test sets, respectively. All of the datasets are multiclass. We randomly and uniformly reassign the label to another class for a given noise level to construct the noisy labeled dataset. The noise level for the training datasets is set to 10% for all experiments in Figures 3 and 4. We do not add additional noise to the test set in order to estimate the true change in accuracy for the different metrics. The results are obtained by averaging 10 trials for each dataset. Running each algorithm for 30 iterations with 100 samples per iteration, the first 3000 samples are considered. For all QBC experiments, we set the number of classifiers in the ensemble to $C = 15$. Similarly, we set $K = 10$ for the KNN experiments. We use a logistic regression classifier for our study since it is widely applied, efficient for large-scale datasets, and does not require rely on setting specific hyper-parameters other than the step size. The step size used to train each logistic regression classifier is 1e-2 which is selected using hyper-parameter tuning. In addition, no regularization is included since we employ early stopping.

Figure 3(a) indicates how the cumulative number of correctly detected mislabeled samples varies over each iteration for the Digits dataset. We include random sampling, denote as RAND, where an expert randomly chooses labeled samples to investigate as a baseline for comparison. For the initial iterations, the MNM and USMS algorithms identify the most number of mislabeled samples. The total number of labeled samples in the Digits dataset is less than 3000. Thus, all mislabeled samples will be detected by all algorithms because the oracle analyst is asked to re-investigate all labeled samples in the training set. For the Shuttle dataset shown in Figure 3(b), the MNM algorithm clearly detects more mislabeled samples early in



(a) Digits (b) Shuttle (c) MNIST

Fig. 5. Number of iterations to discover a percentage of mislabeled samples for a range of noise levels.

the active learning re-query process after the first few iterations. The USMS algorithm is the second best algorithm for quickly identifying mislabeled samples. Figure 3(c) depicts that MNM continues to identify large numbers of mislabeled samples over the first 30 iterations in MNIST. In fact, approximately 80% of all of the proposed candidate samples are indeed confirmed to be mislabeled.

While being able to accurately detect mislabeled samples is indeed critical, the most important metric is how quickly does each algorithm reduce the error rate of a classifier trained on the updated training set. If an algorithm consistently detects mislabeled samples which do not improve the results of the retrained classifier, the expert will become frustrated and refuse to re-investigate previously labeled samples. In Figure 4, we evaluate the logistic regression classifier accuracy retrained, for each iteration, with the updated (i.e. corrected) training set and evaluated on a holdout test set. From the figure, the classifier trained with the MNM objective function performs best for most of the ranges in the Digits and MNIST datasets. After the first ten iterations for the Shuttle dataset, the classifier trained with MNM and USMS each perform well, depending on the iteration number. While not always outperforming the other methods for all datasets, MNM offers the best overall improvement in classifier accuracy in this figure. By being able to observe a concrete performance improvement using either the MNM or USMS algorithms, experts will remain engaged and continue to re-evaluate potentially mislabeled samples.

In the previous experiments, the noise level was set to a fixed level (10% for the datasets). In the final set of experiments, we investigate the performance of this set of algorithms over a range of label noise. We determine the average number of iterations required to identify a fixed percentage (e.g. 50% for Digits, 10% for Shuttle, 5% for MNIST) of mislabeled samples for a noise level ranging from 10% to 50%. Figure 5 indicates that the MNM algorithm requires a lower number of iterations to find the mislabeled samples for each noise level for the MNIST dataset as well as the higher noise ranges of the Digits dataset. MNM is competitive with USMS for the Shuttle dataset and the lower noise range of the Digits dataset.

4. RELATED WORK

Using active learning to help identify mislabeled samples is not new. Guyon *et al.* first suggested an information theoretic measure for ranking potentially mislabeled samples which is similar to active learning [2]. However, the metric was derived from the coefficients of their optimum margin classifier. Thus their algorithm is not generic, and its current form, is also specified for binary labels. Nallapati *et al.* propose using active learning to discover mislabeled samples in the context of text classification [3]. Their active learning algorithm uses misclassification cost to rank samples with binary labels for re-inspection. Our work differs from that of Nallapati's approach in that we employ several different active learning metrics and show that these new algorithms can be used with multiclass data.

Settles provides an excellent tutorial on active learning [1]. In text classification, unlabeled data is plentiful and active learning methods provide a way to pick the most informative point to label. The method developed in [14] selects the unlabeled point with the highest label uncertainty, i.e. where the classifier is most uncertain. The objective function also includes a Naive Bayes density model to prevent sampling from low-density regions. An important early paper was written by Brodley *et al.* [15] on identifying mislabeled samples by creating an ensemble of classifiers that are used to filter potentially mislabeled samples. Only samples which are predicted by the ensemble members are used to train the final classifier. Other samples are predicted to be mislabeled and held out of the final training set. The notion of an intrinsic margin is used to group samples into three categories, e.g. typical, critical, and noisy, based on three methods including a support vector machine [16]. Abe *et al.* propose a method for detecting outliers in general using active learning [17]. Valizadegan *et al.* [18] propose a kernel-based method for identifying mislabeled samples. This algorithm requires solving an optimization problem for a binary set of labels. The round robin approach has some similarity to active class selection [19]. A system which first partitions the data into subsets and then identifies mislabeled samples based on rules was proposed by Zhu *et al.* [20].

5. CONCLUSIONS

Labeling errors occur due to human error, a difference in opinion as to the true nature of the example between experts for groups of similar items, or labeling using mistake prone automation. In this paper we present several new active learning algorithms for detecting mislabeled samples using discriminative and information theoretic criteria. The new maximum negative margin metric accomplishes the two main goals for a detecting mislabeled samples at large-scale, namely, the metric is fast to compute, and it often improves the accuracy of a classifier trained with the correctly labeled dataset more quickly than other metrics. As with most machine learning algorithms, the performance of MNM is dataset dependent.

6. REFERENCES

[1] B. Settles, "Active learning literature survey," Tech. Rep., Univ. of Wisconsin - Madison, January 26 2010.

[2] I. Guyon, N. Matic, and V. Vapnik, "Discovering informative patterns and data cleaning," in *Proceedings of AAAI-94 Workshop on Knowledge Discovery in Databases*, 1994.

[3] R. Nallapati, M. Surdeanu, and C.D. Manning, "Corractive learning: Learning from noisy data through human interaction," in *Proceedings of IJCAI Workshop on Intelligence and Interaction*, 2009.

[4] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Advances in Neural Information Processing Systems*, 2005.

[5] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[6] I. Dagan and S. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proceedings of the International Conference on Machine Learning (ICML)*, 1995.

[7] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," in *Proceedings of Journal of Machine Learning Research*, 2001.

[8] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[9] T. Scheffer, C. Decomain, and S.Wrobel, "Active hidden markov models for information extraction," in *In Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA)*, 2001.

[10] Prateek Jain and Ashish Kapoor, "Active learning for large multi-class problems," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] K. Bache and M. Lichman, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`, 2013.

[12] "Mnist database," `http://yann.lecun.com/exdb/mnist/`, 1998.

[13] B. Settles, *Active Learning*, Morgan & Claypool, 2012.

[14] A.K. McCallum and K. Nigam, "Employing em and pool-based active learning for text classification," in *Proceedings of International Conference on Machine Learning*, 1998.

[15] C. Brodley and M.A. Friedl, "Identifying and eliminating mislabeled training instances," in *Proceedings AAAI Conference on Artificial Intelligence*, 1996.

[16] L. Li, A. Pratap, H. Lin, and Y. S. Abu-Mostafa, "Improving generalization by data categorization," in *In Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2005.

[17] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in *In Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[18] H. Valizadegan and P-N. Tan, "Kernel based detection of mislabeled training examples," in *Proceedings SIAM International Conference on Data Mining*, 2007.

[19] R. Lomasky, C.E. Brodley, M. Aernecke, D. Walt, and M. Friedl, "Active class selection," in *In Proceedings of the European Conference on Machine Learning (ECML)*, 2007.

[20] T. Zhu, X. Wu, and Q. Chen, "Eliminating class noise in large datasets," in *In Proceedings of the Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, 2013.