A Generic Active Learning Framework for Class Imbalance Applications

Aditya R. Bhattacharya[1]
arb17b@my.fsu.edu
Ji Liu[2]
jliu@cs.rochester.edu
Shayok Chakraborty[1]
shayok@cs.fsu.edu

[1] Department of Computer Science
Florida State University
Florida, USA

[2] Department of Computer Science
University of Rochester
New York, USA

Abstract

Active learning algorithms automatically identify the most informative samples from large amounts of unlabeled data and tremendously reduce human annotation effort in inducing a robust machine learning model. Real-world data often exhibit significantly skewed class distributions, where samples from one class dominate over the other. While active learning has been extensively studied, there have been limited research efforts to develop active learning algorithms specifically for class imbalance applications. In this paper, we propose a novel framework to address this research challenge. We pose the active sample selection as a constrained optimization problem and derive a linear programming relaxation to select a batch of samples. Contrary to existing algorithms, our framework is generic and is applicable to both binary and multi-class problems, where the imbalance may exist across multiple classes. Our extensive empirical studies on four vision datasets spanning three different application domains (face, facial expression and handwritten digits recognition) with varied degrees of class imbalance demonstrate the promise and potential of the method for real-world imbalanced data applications.

1 Introduction

Due to the widespread deployment of inexpensive sensors, modern era is witnessing an unprecedented growth in the amount of digital data in varied forms (images, videos, text etc.). This has expanded the possibilities of solving real-world problems using computational learning frameworks. However, while gathering large amounts of unlabeled data is cheap and easy, annotating them with class labels (to induce a model) is a time-consuming and labor-intensive process. This has set the stage for research in the field of *active learning* (AL). Active learning algorithms automatically identify the salient and informative samples from large amounts of unlabeled data; this not only reduces the human annotation effort in training a model, but also exposes the model to the most salient exemplars from the underlying data population [31]. Active learning has demonstrated promising performance in a variety of applications including computer vision [5], text mining [35], spam filtering [30] and bio-informatics [25] among others.

One of the fundamental problems in automated data analysis and prediction is the problem of imbalanced class distributions, where some classes (majority classes) have a significantly higher number of examples in the training set than other classes (minority classes) [1, 2]. Class imbalance is commonly prevalent in a variety of domains such as medical diagnosis [14, 24], fraud detection [6] and others [3, 27]. For instance, in medical diagnosis, the frequency of one class (e.g. cancer) can be much less than the other class (e.g. healthy patient). This can have a serious detrimental effect in training machine learning models [18]. Learning algorithms for class imbalance problems include oversampling methods [17, 23], undersampling methods [21], thresholding [28], cost-sensitive learning [10] and one-class classification [19]. A popular oversampling technique called Synthetic Minority Oversampling TEchnique (SMOTE) [7], which augments artificial samples created by interpolating neighboring data points, has depicted impressive performance in several applications.

Although both active learning and imbalanced data learning are well-researched problems, there have been limited research efforts in developing active learning algorithms specifically for imbalanced data applications. In this paper, we propose a novel framework to address this important challenge. We propose a sample selection criterion based on informativeness and data geometry; the active selection is then posed as a constrained optimization problem and a linear programming relaxation is derived to select a batch of samples. Our framework can handle binary as well as multi-class settings, where more than one classes can be minority classes. While active learning algorithms have been studied to discover anomalies and rare categories [26] [16], our focus in this research is to develop an AL framework for imbalanced data, which are common in real-world applications. The rest of the paper is organized as follows: we present a survey of related techniques in Section 2; the details of our framework are presented in Section 3; Section 4 depicts the results of our empirical studies; and we conclude with discussions in Section 5.

2 Related Work

In this section, we present a brief survey of active learning techniques for imbalanced data applications. Active learning (AL) has received significant research attention in the machine vision community [31]. In a typical pool-based setup, the learner is exposed to a pool of unlabeled samples and it iteratively queries informative samples for manual annotation. The most common query strategy in active learning is uncertainty sampling, where unlabeled samples with the highest classification uncertainties are queried for annotation. The uncertainty of an unlabeled sample can be quantified by its Shannon's entropy [15], its distance from the decision boundary in the feature space for SVM classifiers [35], the disagreement among a committee of classifiers about the label of the sample [13] and also by combining multiple criteria such as uncertainty, representativeness and diversity [32]. Although active learning, in general, has been extensively studied, relatively fewer research efforts have focused on the problem of class imbalanced active learning. Existing methods mostly use data balancing techniques (such as over/under sampling), together with an active learning algorithm (such as uncertainty sampling) to address this problem.

The SVM-based AL algorithm proposed by Ertekin *et al.* [11, 12] is based on the observation that the imbalance ratio of the classes within the margin is much smaller than that of the entire dataset. The proposed method focuses only on a random subset of samples within the SVM margin (thereby addressing the imbalance issue) and queries samples closest to the decision hyperplane from this subset. Along similar lines, Zieba and Tomczak [37] proposed a boosted SVM algorithm for class imbalance active learning, based on sampling near the decision boundary and misclassification cost estimation. Yang and Ma proposed an ensemble-based framework for class imbalanced AL, where artificial data samples were created to address the class-imbalance and margin-based uncertainty sampling was used for active sample selection [36]. Chairi *et al.* [4] approached the problem by undersampling the dataset to reduce the class-imbalance, followed by margin-based sampling for active learning. Tomanek and Hahn [34] addressed this problem in the context of named entity recognition (NER), where data re-sampling was applied to reduce the imbalance ratio and a disagreement based scheme was used for active sample selection. The Uncertainty Sampling with Biasing Consensus (USBC) algorithm proposed by Chen and Mani [8] used a multi-model committee, together with uncertainty sampling using least confidence (with higher weight on the minority class) to address class imbalanced active learning.

However, these AL algorithms assume a binary classification setting, where one class is the majority class and the other is the minority class; they cannot be applied to multi-class problems, where the imbalance may exist across more than one classes. For instance, consider an object recognition problem, where the goal is to recognize objects in image. This is a multi-class problem in which multiple classes can be potentially imbalanced, since common objects (such as trees, buildings etc.) tend to occur in a larger number of images than uncommon objects (such as butterflies, frisbees etc.). In this paper, we propose a generic active learning framework for imbalanced data to address this research challenge. We now describe our framework.

3 Proposed Framework

Consider an active learning problem, where we are given a labeled training set L_t and an unlabeled set U_t at time t. Let w_t be the learning model trained on L_t and C be the set of classes in the problem. We are further given that there is a class imbalance in the data, and a set $C_{min} \subset C$, which contains the set of minority classes in the problem. Note that we do not make any assumptions about the cardinalities of the two sets C and C_{min}, that is, our method is applicable to multi-class problems with an arbitrary number of minority classes. Our objective is to select a batch B containing k unlabeled samples such that the model w_{t+1} trained on $L_t \cup B$ depicts good generalization capability. We quantify the utility score of a batch of unlabeled samples and attempt to select a batch furnishing the maximal utility score. In this research, the utility of a batch of samples was quantified using their informativeness and diversity scores. This ensures that the selected samples are individually informative and they have high diversity (minimal redundancy) among them. Such selection criteria has been used in previous active learning research [32].

Computing informativeness: We used two criteria to compute the information content of an unlabeled sample: (i) *Classification entropy*, which computes a confidence score of an unlabeled sample. We use the weighted Shannon's entropy for this purpose, which has been used in previous research on learning in the presence of skewed class distributions [20]. The weighted entropy of an unlabeled sample x_i is computed as:

$$E(x_i) = -\sum_{j=1}^{C} g_j p_j \log(g_j p_j) \tag{1}$$

where p_j is the posterior probability of x_i with respect to class j, computed by the current model w_t and g_j is the importance weight assigned to class j and is given by $g_j = \frac{1}{C.b_j}$, where b_j is the proportion of samples in class j in the training set [20]. A high value of $E(x_i)$ denotes that the model has low confidence of prediction on the sample and that the sample is informative.

(ii) *Data geometry*, which gives an estimate of confidence of the sample with respect to every class. This condition enables us to specifically focus on the classes that are under-represented in the data. For an unlabeled sample x_i and a class label j, let d_i^j denote the distance between x_i and its m nearest neighbors in the training set, with the same class label j. Similarly, let d_i^{-j} denote the distance between x_i and its m nearest neighbors in the training set, with class label other than j. The ratio of these distances is then computed $\alpha_{ij} = \frac{d_i^j}{d_i^{-j}}$. A value of α_{ij} close to 1 implies that the unlabeled sample x_i is at the border between class j and other classes. It should thus have a low confidence with respect to class j. If the value of α_{ij} is very low (close to 0), it signifies that the distance of x_i to its closest neighbors in class j is much lower than that in other classes, which means that there is a high chance it belongs to class j; the confidence of x_i with respect to class j should thus be high. If α_{ij} has a very high value (much greater than 1) it implies that the distance of x_i to its closest neighbors in class j is much higher than that in other classes, which means that there is a high chance it belongs to one of the classes other than j. In this case also, it should have a high confidence with respect to class j. We define a term q_{ij} as the absolute difference between 1 and α_{ij}; a low value of q_{ij} denotes that the unlabeled sample x_i has a low confidence (and hence informative) with respect to class j:

$$q_{ij} = |1 - \alpha_{ij}| \tag{2}$$

This is illustrated in Figure 1, which shows an example of a 3-class classification problem. From a data geometry perspective, the unlabeled sample should have a value of α_{ij} very close to 1 for both classes 1 and 2. Thus, q_{ij} will be low for these classes and the sample will have low confidence with respect to these classes. On the other hand, the value of α_{ij} will be high (much greater than 1) for class 3 and consequently, q_{ij} will be high for class 3. The unlabeled sample will thus have a high confidence with respect to class 3. This supports our intuition, as given the data geometry, it is very unlikely that the unlabeled sample will belong to class 3.

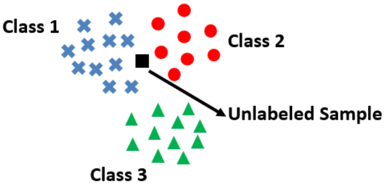


Figure 1: Illustration of confidence computation using data geometry. The unlabeled sample has low confidence with respect to classes 1 and 2, but high confidence with respect to class 3. Best viewed in color.

Given $E(x_i)$ and q_{ij}, we compute a confidence matrix $P \in \Re^{C \times |U_t|}$, where each column represents a sample and the values in the rows denote the confidence of the corresponding unlabeled sample with respect to all classes:

$$P(j,i) = \frac{q_{ij}}{E(x_i)} \tag{3}$$

Computing redundancy: To avoid selecting samples that are individually informative but mutually redundant, we compute a redundancy matrix $R \in \Re^{|U_t| \times |U_t|}$ where R_{ij} denotes the redundancy between samples x_i and x_j in the unlabeled set. We used the cosine similarity to quantify the redundancy between a pair of samples, where a low value of the similarity denotes low redundancy. We thresholded the similarity values at 0, so that R contains only non-negative entries. The matrix R is computed as follows:

$$R(i,j) = \min(0, cos(x_i, x_j)) \tag{4}$$

3.1 Active Sample Selection

Given the confidence matrix P and the redundancy matrix R, our objective is to select k unlabeled samples which furnish minimal confidence and minimum redundancy among them. To consider the class imbalance, we attempt to select a specific number of samples which furnish low confidence with respect to the minority classes. We define a binary matrix $M \in \{0,1\}^{|U_t| \times C}$ where each row corresponds to an unlabeled sample and each column corresponds to a class. A value of 1 in a row denotes that the sample should be selected for annotation and the position of 1 in a particular row of M denotes the class producing the minimum confidence for that sample. Such a formulation enables us to consider the confidence of an unlabeled sample with respect to the individual classes, which is critical in applications involving class imbalance. The active batch selection is thus posed as the following optimization problem:

$$\begin{aligned} \min_{M} \quad & \text{trace}(MP) + \lambda (Me)^\top R(Me) \\ \text{s.t.} \quad & M_{ij} \in \{0,1\}, \forall i, j \\ & M_i.e \leq 1, \forall i \\ & \sum_{i,j} M_{ij} = k \\ & \langle M, E \rangle = \rho \end{aligned} \tag{5}$$

where λ is a weight factor governing the relative importance of the uncertainty and redundancy terms, e is a vector of length C with all entries 1, M_i denotes row i of matrix M, $\langle \cdot, \cdot \rangle$ denotes the matrix inner product operator, E is a matrix of the same dimension as M with all 1s in the columns corresponding to the minority classes and 0s elsewhere and $\rho (< k)$ is a constant. The first constraint denotes that M is a binary matrix; the second constraint signifies the each row of M can have at most one entry as 1; the third constraint denotes that M will have k entries as 1, which is the pre-specified batch size; and the fourth constraint denotes that M should have a specific number (ρ) of 1s in the columns corresponding to the minority classes. This enforces the algorithm to select ρ out of the k samples by considering the confidence in the minority classes. Thus, by specifically focusing on the confidence of a sample with respect to the minority classes, we attempt to guide the active sample selection process to address the imbalance issue in the data. We now discuss an efficient strategy to solve this optimization problem, as detailed in the following theorem.

Theorem 1. *The optimization problem defined in Equation (5) can be expressed as an equivalent linear programming (LP) problem.*

Proof. The first term in the objective function can be expressed as a linear term: $trace(MP) = \sum_{i,j} P_{ij}.M_{ji}$. The second term is simplified as follows:

$$\begin{aligned}
(Me)^\top R(Me) &= \sum_{i,j} R_{ij}(Me)_i(Me)_j \\
&= \sum_{i,j} R_{ij}\langle M_i.e, M_j.e\rangle \\
&= \sum_{i,j} R_{ij}\langle M_i, M_j.ee^\top\rangle \\
&= \sum_{i,j} R_{ij}\langle M_j^\top M_i, ee^\top\rangle \quad \text{(by laws of inner product)} \\
&= \sum_{i,j} R_{ij} \sum_{a,b} M_{ia}.M_{jb} \quad \text{(since } ee^\top \text{ is a matrix of all 1's)} \\
&= \sum_{i,j}\sum_{a,b} R_{ij} M_{ia}.M_{jb} \\
&= \sum_{i,j}\sum_{a,b} R_{ij} V_{ijab},
\end{aligned}$$

where $V_{ijab} = M_{ia}.M_{jb}$ (the derivation uses the algebra of inner product operation and the fact that $ee^\top$ is a matrix of all 1's). Since M is a binary matrix with only 0 and 1 entries, V_{ijab} will equal 1 when both M_{ia} and M_{jb} are 1 and will equal 0 otherwise. Considering the binary constraints on M, this quadratic equality can be expressed as an equivalent linear inequality as follows:

$$V_{ijab} = M_{ia}.M_{jb} \Leftrightarrow M_{ia} + M_{jb} \geq 2V_{ijab} \tag{6}$$

A simple observation reveals that the linear inequality also produces a value of 1 for V_{ijab} when both M_{ia} and M_{jb} are 1 and 0 otherwise. The optimization problem in Equation (5) can thus be expressed as follows:

$$\begin{aligned}
\min_{M,V} \quad & \sum_{i,j} P_{ij}.M_{ji} + \lambda \sum_{i,j}\sum_{a,b} R_{ij} V_{ijab} \\
\text{s.t.} \quad & M_{ij}, V_{ijab} \in \{0,1\}, \forall i,j,a,b \\
& M_i.e \leq 1, \forall i \\
& \sum_{i,j} M_{ij} = k \\
& \langle M, E\rangle = \rho \\
& M_{ia} + M_{jb} \geq 2V_{ijab}
\end{aligned} \tag{7}$$

In this optimization problem, both the objective function and the constraints are linear in the variables M and V. It is thus a linear programming (LP) problem. □

We vectorize the variables, append them one below the other and express the objective function and the constraints in terms of the new variable. The integer constraints on M and V are then relaxed into continuous constraints and the problem is solved using an off-the-shelf LP solver. After obtaining the continuous solution, we recover the integer solution of our variable of interest M, using a greedy approach where the highest entries in each row of M are reconstructed as 1 and the other entries as 0, observing the constraints.

4 Experiments and Results

We conducted an extensive set of experiments to study the performance of our framework against competing baselines, the effect of batch size and its performance in multi-class settings where more than one classes can be the minority classes. These are detailed below.

4.1 Datasets and Experimental Setup

Our search revealed that most of the publicly available datasets where there is a natural class imbalance, are not from the vision domain. We therefore selected 4 challenging vision datasets (from different application domains) and used the *step imbalance* technique proposed in previous research [2] to impart class imbalance in the data. **Face Recognition**: We used two datasets in our experiments: the VidTIMIT [29] and the NIST Multiple Biometric Grand Challenge (MBGC) [33], both of which contain recordings of subjects under unconstrained natural conditions. 25 subjects were selected at random for our experiments. **Facial Expression Recognition**: We also studied the performance of our algorithm on the MindReading dataset, which was collected to help individuals with autism spectrum disorder (ASD) recognize facial expressions [9]. It contains images of the 6 basic emotions from a number of subjects. **Handwritten Digits Recognition**: We further validated the performance of our framework on the MNIST dataset [22] (containing images of handwritten digits from 10 classes), which is extensively used in computer vision research.

We first studied the performance of our algorithm on binary classification problems with a single minority class. This was done to facilitate a fair comparison against the baseline methods, which work only on binary problems. Each dataset was binarized in the following way: one class was selected at random from each dataset and was considered as the minority class; all the other classes were coalesced and was considered the majority class. This resulted in the following imbalance ratios for each of the datasets: VidTIMIT $(1:24)$, MBGC $(1:24)$, MindReading $(1:5)$ and MNIST $(1:9)$. This strategy of introducing imbalance in the data is inspired from previous research on imbalanced learning with vision data [2].

Each dataset was divided into an initial training set (1%), an unlabeled set (89%) and a test set (10%) (the number of images used for each dataset are depicted in Table 1). For a given batch size k, each algorithm queried k samples from the unlabeled set in each iteration. The selected samples were then labeled and appended to the training set; the model was updated and tested on the test set. The process was continued iteratively until a stopping condition was satisfied (taken as 25 iterations in this work). The objective was to study the improvement in performance on the test set with increasing sizes of the training set. We used the F1-score as the evaluation metric in this research, as it is commonly used in class imbalanced learning applications [1]. All the results were averaged over 3 runs (with different initial training, unlabeled and test sets) to rule out the effects of randomness. The weight parameter λ was selected as 1 based on preliminary experiments, the batch size k was taken as 5 (we also studied the effect of this parameter on learning performance) and the parameter ρ was taken as 2.

Dataset	Training	Unlabeled	Test
VidTIMIT	87	7788	875
MBGC	85	7610	855
MindReading	25	2279	257
MNIST	91	8099	910

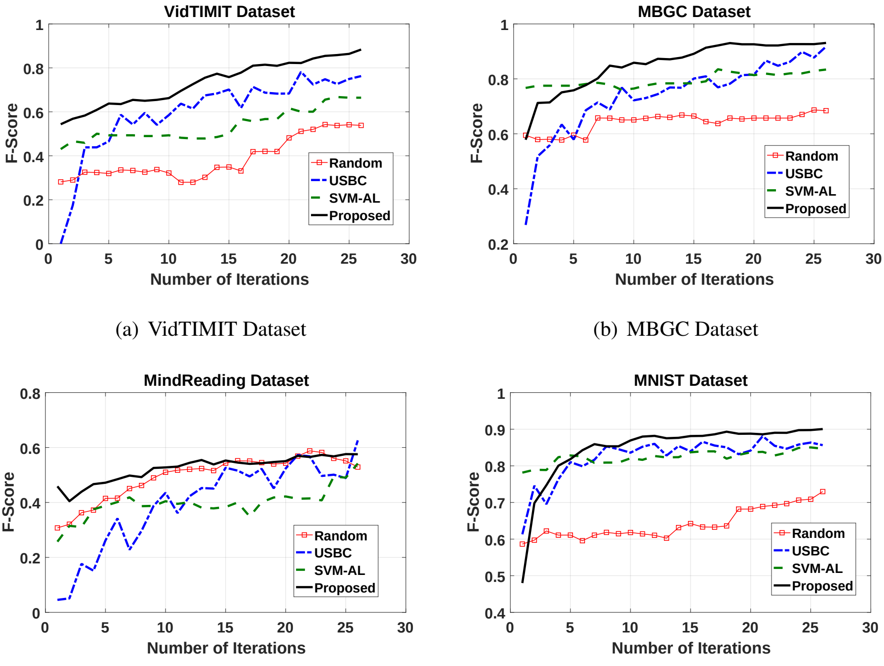
Table 1: Dataset Details

4.2 Comparison Baselines

We used three comparison baselines in our work: (*i*) *Random Sampling*, where a batch of unlabeled samples is queried at random; (*ii*) *SVM-AL* [11]; and (*iii*) *USBC*, which has demonstrated impressive performance in an active learning challenge [8]. These baseline methods were selected to capture the two most common active learning techniques for imbalanced data classification (margin-based uncertainty sampling [11], query-by-committee together with ensemble learning [8]).

4.3 AL Results: Binary Problems with One Minority Class

The AL performance results are depicted in Figure 2. In each graph, the x-axis denotes the number of active learning iterations and the y-axis denotes the F1-score on the test set. We note that *Random Sampling* sometimes depicts good performance (as in the MindReading dataset), but is not consistent across datasets in its performance. The *SVM-AL* and *USBC* algorithms perform better than *Random Sampling*, but is not as good as the proposed method. Our framework consistently depicts impressive performance across all the datasets; at any given iteration, it produces the highest F1-score most of the times. The improvement is particularly evident for the face recognition datasets VidTIMIT and MBGC. This shows that our algorithm is efficiently identifying the most informative unlabeled samples and is capable of inducing a robust model with minimal human effort. The results unanimously corroborate the promise and potential of our method for real-world applications with class imbalance.



(a) VidTIMIT Dataset

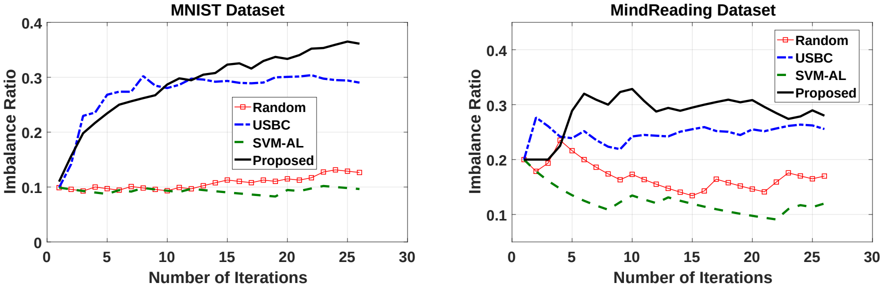
(b) MBGC Dataset

(c) MindReading Dataset

(d) MNIST Dataset

Figure 2: Performance of the proposed framework on binary datasets with one minority class. Best viewed in color.

Figure 3(a) shows a plot of the class imbalance ratio (ratio of the number of samples in the minority and majority classes in the training set) against the number of active learning iterations for the MNIST dataset. Since this is a binary classification problem, an imbalance ratio of 0.5 implies that the data is perfectly balanced between the two classes. The training set has an initial imbalance ratio of about 0.1. The figure depicts that our method achieves an imbalance ratio of about 0.36 after 25 AL iterations. *USBC* attains a ratio of about 0.29 while *Random Sampling* and *SVM-AL* do not depict a considerable growth in the imbalance ratio. The constraint $\langle M,E \rangle = \rho$ enforces our method to specifically focus on the samples furnishing low confidence with respect to the minority classes and query them accordingly. Figure 3(b) shows the results for the MindReading dataset, where the starting imbalance ratio is approximately 0.2. The proposed method attains a ratio of about $0.28 - 0.3$ after the AL iterations, whereas the baselines achieve a much lower ratio (except *USBC*). Thus, besides querying the salient and exemplar samples, our method also attempts to balance the class distribution in the training set. This combined criteria accounts for its superior performance, as evident from Figure 2.



(a) MNIST Dataset

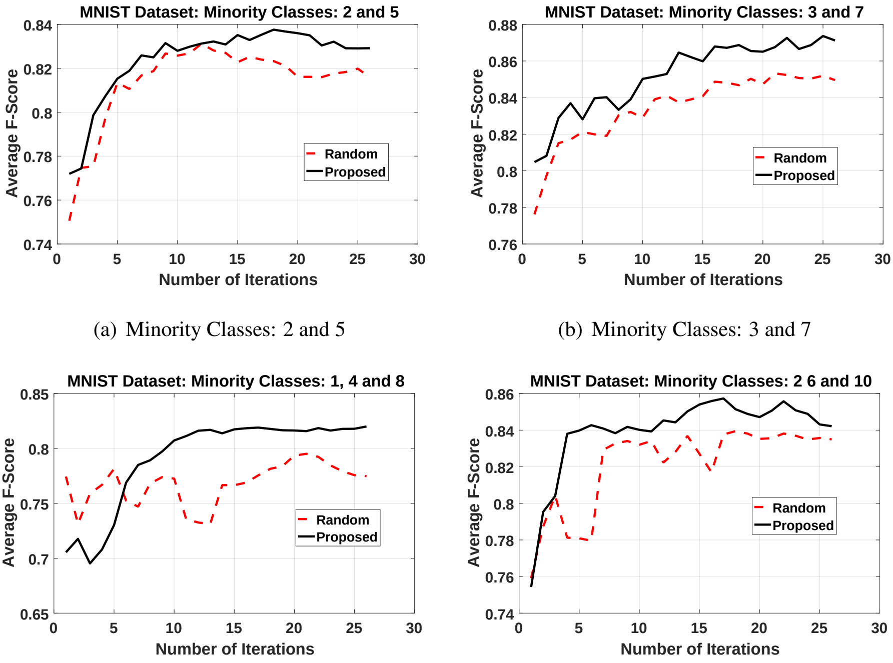
(b) MindReading Dataset

Figure 3: Study of imbalance ratio. The proposed method attempts to balance the class distribution in the training set and achieves the highest imbalance ratio. Best viewed in color.

We also conducted an experiment to study the effect of batch size on the learning performance. The MNIST dataset was used with batch size 3, 5, and 10. The results are presented in the Supplemental File, due to space constraints. The results depict a similar pattern as in Figure 2; our algorithm outperforms the baselines across all batch sizes.

4.4 AL Results: Multi-class Problems with Multiple Minority Classes

As evident from the formulation detailed in Section 3, our method can also be applied to multi-class problems, where there are multiple minority classes. We conducted experiments on the MNIST dataset (with all 10 classes) to study the multi-class performance of our method. We studied two different settings, with 2 and 3 minority classes (out of 10). We randomly deleted 70% of the samples from each of the classes designated as a minority class, to impart class imbalance in the data. We used *Random Sampling* as the comparison baseline for this experiment, as the other AL methods are specifically designed for a binary classification problem with a single minority class and do not generalize to multiple minority classes. Since this is a multi-class problem, we used the average F1-score (averaged across all classes) as the performance metric. Figure 4(a) presents the results with two minority classes $(2,5)$. We once again note that our algorithm outperforms *Random Sampling* and depicts a faster growth in the average F1-score with increasing number of iterations. Figure 4(b) also shows a similar pattern, where $(3,7)$ were taken as the minority classes, instead of 2 and 5. Figures 4(c) and 4(d) show the performance with 3 minority classes $(1,4,8)$ and $(2,6,10)$ respectively. Our framework depicts better performance than *Random Sampling* in both experiments.



(a) Minority Classes: 2 and 5

(b) Minority Classes: 3 and 7

(c) Minority Classes: 1, 4 and 8

(d) Minority Classes: 2, 6 and 10

Figure 4: Performance of the proposed framework on the multi-class MNIST dataset with multiple minority classes. Best viewed in color.

5 Conclusion and Future Work

In this paper, we proposed a novel active learning framework for class-imbalance problems. The active sample selection was posed as a constrained optimization problem, based on the confidence and diversity criteria, which was shown to be equivalent to a linear programming problem. Our extensive empirical studies on a variety of applications demonstrated the potential of our algorithm over competing baselines. More importantly, our method generalizes to a multi-class setting, with an arbitrary number of minority classes. This corroborates the flexibility and the usefulness of the algorithm for real-world classification applications. As part of future research, we plan to study the performance of our framework on other problem domains, such as multi-label learning. We also plan to compute the value of ρ adaptively in each iteration, based on the class distribution in the training set.

References

[1] P. Branco, L. Torgo, and R. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49, 2016.

[2] M. Buda, A. Maki, and M. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018.

[3] C. Cardie and N. Howe. Improving minority class prediction using case-specific feature weights. In *International Conference on Machine Learning (ICML)*, 1997.

[4] I. Chairi, S. Alaoui, and A. Lyhyaoui. Sample selection based active learning for imbalanced data. In *International Conference on Signal-Image Technology and Internet-Based Systems*, 2014.

[5] S. Chakraborty, V. Balasubramanian, and S. Panchanathan. Dynamic batch mode active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[6] P. Chan and S. Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 1998.

[7] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research (JAIR)*, 16, 2002.

[8] Y. Chen and S. Mani. Active learning for unbalanced data in the challenge with multiple models and biasing. In *JMLR Workshop on Active Learning and Experimental Design*, 2011.

[9] R. El-Kaliouby and P. Robinson. Mind reading machines: Automated inference of cognitive mental states from video. In *IEEE International Conference on System, man and Cybernetics*, 2004.

[10] C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

[11] S. Ertekin, J. Huang, L. Bottou, and C. Giles. Learning on the border: Active learning in imbalanced data classification. In *Conference on Information and Knowledge Management (CIKM)*, 2007.

[12] S. Ertekin, J. Huang, and C. Giles. Active learning for class imbalance problem. In *ACM Conference on Information Retrieval (SIGIR)*, 2007.

[13] Yoav Freund, Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997. ISSN 0885-6125.

[14] J. Grzymala-Busse, L. Goodwin, W. Grzymala-Busse, and X. Zheng. An approach to imbalanced data sets based on changing rule strength. In *Rough-Neural Computing*, 2004.

[15] A. Holub, P. Perona, and M. Burl. Entropy-based active learning for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2008.

[16] T. Hospedales, S. Gong, and T. Xiang. Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 25, 2013.

[17] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[18] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6, 2002.

[19] N. Japkowicz, S. Hanson, and M. Gluck. Nonlinear autoassociation is not equivalent to pca. *Neural Computation*, 12, 2000.

[20] A. Kirshners, S. Parshutin, and H. Gorskis. Entropy-based classifier enhancement to handle imbalanced class problem. *Procedia Computer Science*, 104, 2017.

[21] M. Kubat and S. Matwin et al. Addressing the curse of imbalanced training sets: one-sided selection. In *International Conference on Machine Learning (ICML)*, 1997.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[23] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2015.

[24] B. Namee, P. Cunningham, S. Byrne, and O. Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial intelligence in medicine*, 24, 2002.

[25] H. Osmanbeyoglu, J. Wehner, J. Carbonell, and M. Ganapathiraju. Active machine learning for transmembrane helix prediction. *BMC Bioinformatics*, 11(1), 2010.

[26] D. Pelleg and A. Moore. Active learning for anomaly and rare-category detection. In *Neural Information Processing Systems (NIPS)*, 2004.

[27] P. Radivojac, N. Chawla, A. Dunker, and Z. Obradovic. Classification and knowledge discovery in protein databases. *Journal of Biomedical Informatics*, 37, 2004.

[28] M. Richard and R. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3, 1991.

[29] Conrad Sanderson. *Biometric Person Recognition: Face, Speech and Fusion*. VDM Verlag, June 2008. ISBN 3639027698.

[30] D. Sculley. Online active learning methods for fast Label-Efficient spam filtering. In *Fourth Conference on Email and AntiSpam*, 2007.

[31] B. Settles. Active learning literature survey. In *Technical Report 1648, University of Wisconsin-Madison*, 2010.

[32] D. Shen, J. Zhang, J. Su, G. Zhou, and C. Tan. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2004.

[33] M. Tistarelli and M. Nixon. Advances in biometrics: Icb. *SpringerLink*, 2009.

[34] K. Tomanek and U. Hahn. Reducing class imbalance during active learning for named entity annotation. In *International Conference on Knowledge Capture*, 2009.

[35] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research (JMLR)*, 2:45–66, 2001.

[36] Y. Yang and G. Ma. Ensemble-based active learning for class imbalance problem. *Journal of Biomedical Science and Engineering*, 3, 2010.

[37] M. Zieba and J. Tomczak. Boosted svm with active learning strategy for imbalanced data. *Soft Computing*, 19, 2015.