Robust Learning with Highly Skewed Category Distributions

Selen Uguroglu

CMU-LTI-13-020

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
Jaime Carbonell (Chair)
Anatole Gershman
Jeff Schneider
Charles Elkan (UCSD)

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies*

Keywords: Active learning, supervised learning, kernel methods, cost-sensitive learning

To my mother...

Abstract

Highly skewed category distributions are abundant in many real-world tasks in data mining, such as medical diagnosis (rare diseases), text categorization (rare topics), and fraud detection (when most transactions are legitimate). Under extreme class skew, most supervised learning algorithms tend to minimize loss by labeling every instance with the majority class(es), leading to poor recall on the minority class(es). However, true misclassification costs may be much greater when minority class instances are missed, e.g. a massive but rare fraud missed, or an uncommon life-threatening condition misclassified as benign. Hence, a means of detecting rare but consequential classes is required, and that is the topic of this dissertation.

Whereas learning under extreme class skew has been previously investigated, many challenges remain: e.g. disjunctive majority classes and minority-majority class overlap. Prior research did not consider incorporating the structure of minority class into the learning process. In this dissertation, we address class imbalance under the compactness hypothesis, i.e. minority class forms one or more compact clusters in the feature space. Furthermore, we introduce several learning algorithms to address class imbalance under two other assumptions: disjunctive majority class and overlapping classes. We also propose new active learning strategies in cases when there are insufficient labeled minority class instances to learn accurate concept descriptions under highly-skewed settings. Our algorithms are based on a variety of methods/paradigms, including multiple kernel learning, maximum mean discrepancy, and cost-sensitive learning.

We evaluate the new and baseline methods on several real-world datasets with a particular focus on the Womens' Ischemic Syndrome Evaluation (WISE) dataset, to demonstrate a practical application in medical diagnosis. We show that when the assumptions are satisfied, leveraging the structure of classes, such as compact minority class, disjunctive majority class, leads to better prediction performance, quantified by the improvement in F-1 and AUC measures. Our empirical results reveal an improvement in F-1 as much as 28%.

Acknowledgments

First and foremost, I would like to thank my advisor, Jaime Carbonell, for his mentorship during this PhD. His unconditional support allowed me to freely pursuit my research interests, and his close guidance made sure that I was always on the right track. He taught me how to analyze a problem from different perspectives, and helped me become the independent researcher that I am. I always admired his endless knowledge and wisdom; he will always remain a role model of mine.

I was extremely fortunate to have Anatole Gershman, Charles Elkan, and Jeff Schneider on my thesis committee. Their invaluable feedback and suggestions helped me greatly improve this thesis. I would like to thank Cédric Archambeau for his mentorship during my internship at Xerox Research Centre Europe. The insightful discussions we had helped me broaden my vision as a researcher. I would like to thank my former co-advisor Eric Xing for sparking my interest in bioinformatics, and for his guidance through the early bumps in PhD. Thanks to Judith Klein-Seetharaman for introducing me to challenging problems in data mining, and for her deep confidence that I will overcome them.

During my time at Carnegie Mellon, I have had the opportunity to learn from various faculty, collaborators, post-docs, and fellow graduate students. I am grateful to everyone of them, they greatly influenced my development as a researcher.

I have had a splendid time in Pittsburgh, I have been fortunate to have amazing friends, amazing housemates. This thesis would not be possible without their support.

Last but not least, I would like to thank my family, especially my mother. She has always stressed the importance of education, “the gold bracelet” that I must have. This dissertation is dedicated to her...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

A skewed class distribution occurs when at least one class has many fewer instances relative to other classes. For instance, in topic classification, articles written on specific, obscure topics are fewer in a given dataset, compared to news articles on broad topics such as science or politics. In medical diagnosis, the number patients with rare diseases is much smaller compared to the general population. In astronomy, image datasets collected by astronomical surveys have a high degree of imbalance: known objects (stars, galaxies) account for 99.9% of the data, and unknown objects (those with the greatest interest) constituting the rest. Many other real-world tasks deal with highly skewed datasets; examples can be extended to credit-card fraud detection [72], intrusion detection [49], text categorization [85], [100], detection of oil spills [55], web page classification [40], and sentence boundary detection [65].

The main challenge in datasets with high-category skew arises from inseparable classes. One or more classes can be composed of small, possibly overlapping disjuncts. Recent studies have shown that small disjuncts and class overlaps in the dataset aggravates class imbalance problem; in the presence of enough labeled data class imbalance alone do not hinder the classification performance significantly as long as classes are separable [30], [18], [15], [42], [92], [93], [94], [47].

Under extreme class skew, most standard machine learning algorithms perform poorly on minority classes. The general goal of these algorithms is to maximize the classification accuracy by minimizing 0/1 loss. However, this strategy operates under the assumption that the training set has close to a balanced class distribution or instances are associated with equal misclassification costs. Therefore under high-class skew, standard algorithms minimize overall error by ignoring under-represented classes: On a skewed dataset, where the majority class constitutes 99.9% of the data, an ineffective learner can maximize classification accuracy by assigning every instance to the majority class. In expense of incorrectly classifying all minority class instances, this leads to 99.9% accuracy.

The real problem stems from the unequal misclassification costs of the classes: incorrect classification of minority class instances can be highly costly, i.e. labeling fraudulent transactions as normal transactions may result in deep financial losses, misclassifying sick patients as healthy may be detrimental to patients' health. On the other hand, an error in the opposite direction, e.g. suggesting a credit card transaction needs to be verified, may only cause an inconvenience. Therefore, methods that detect rare but consequential classes are required.

Prior work includes re-sampling strategies, that aim to balance the class distribution, or cost-sensitive methods that modify learning algorithm to incorporate unequal misclassification costs. However, none of these strategies put adequate emphasis on small disjuncts or class overlaps, furthermore the structure of classes is often ignored by prior work. This explains why strategies such as cost-sensitive learning or re-sampling can be effective in certain imbalanced datasets, but fail in others.

In this dissertation, we address the class imbalance problem by leveraging shared properties of the minority class instances. We can state the thesis hypothesis as follows:

Hypothesis 1. Compactness of the minority class
In many real world applications, minority class instances tend to exhibit similar patterns. In medical diagnosis, high risk patients or patients with rare diseases may share similar physiological symptoms or genetic markers. In credit-card fraud detection, even if fraudulent transactions may resemble legitimate transactions, they may share unique features or patterns. With the compactness hypothesis, we assume that the minority class instances for each class form compact clusters with each other in the feature space or subspace [39].

In this thesis, we examine the validity of the compactness hypothesis, and propose several novel algorithms for efficient learning under high class skew. We especially focus on the performance of the proposed and baseline algorithms on the Womens' Ischemic Syndrome Evaluation (WISE) dataset, as this dataset is highly skewed with potential class overlaps. We give detailed information on the WISE dataset in the next section.

In certain problem settings, sufficient labeled data from the minority class may not be available for training. In this dissertation, we propose algorithms for such related problem settings as well. More specifically, we propose new active learning algorithms to utilize when there is not enough labeled data, unlabeled data is abundant, skewed class distribution is anticipated.

1.2 Thesis Statement

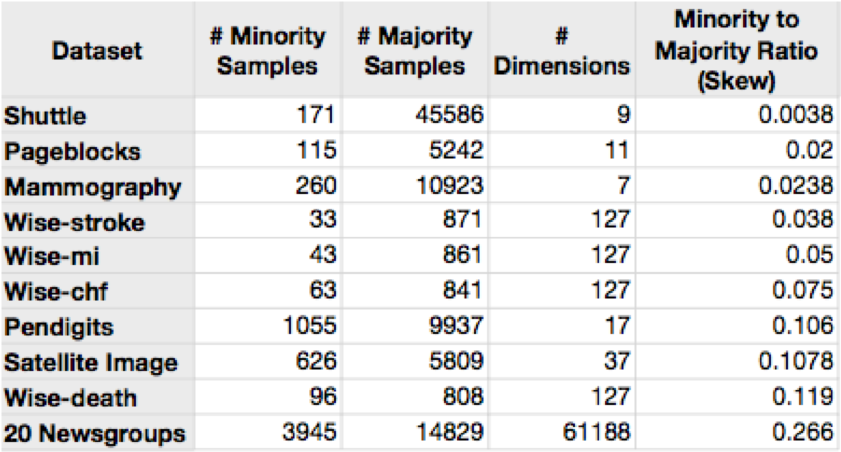
This thesis demonstrates that leveraging the structure of the classes (e.g. compact minority class, disjunctive majority class) in the learning process leads to better prediction performance on the minority class than state of the art baselines under high-class skew.

1.3 Thesis Outline

The structure of this dissertation is as follows: Chapter 2 reviews prior work. Chapter 3 presents initial data analysis on the WISE dataset. Chapter 4 examines the validity of the compactness hypothesis and applies it on the WISE dataset. Chapter 5 describes two algorithms to combat class imbalance when small disjuncts or class overlaps are present. Chapter 6 proposes active learning algorithms to utilize when the presence of class imbalance in the dataset is known a priori. Chapter 7 provides concluding remarks.

1.4 Data

In this section, we describe the data used for evaluation. The applications of our approach to different tasks such as medical diagnosis, or text classification are shown using a variety of skewed datasets. Dataset statistics such as the number of minority class instances, the number of majority class instances, the number of dimensions in each dataset and the skew level are shown in Table 1.1. Skew level is defined as the ratio of the minority class instances to the majority class instances in a dataset. There are 4 different WISE datasets each corresponding to different event prediction tasks: mortality, cardiovascular heart failure (CHF), myocardial infraction (MI) and stroke prediction.



Dataset	# Minority Samples	# Majority Samples	# Dimensions	Minority to Majority Ratio (Skew)
Shuttle	171	45586	9	0.0038
Pageblocks	115	5242	11	0.02
Mammography	260	10923	7	0.0238
Wise-stroke	33	871	127	0.038
Wise-mi	43	861	127	0.05
Wise-chf	63	841	127	0.075
Pendigits	1055	9937	17	0.106
Satellite Image	626	5809	37	0.1078
Wise-death	96	808	127	0.119
20 Newsgroups	3945	14829	61188	0.266

Table 1.1: Description of the datasets used in the experiments: the number of minority and majority class instances, the number of dimensions, and class skew are shown for each dataset

1.4.1 Womens' Ischemic Syndrome Evaluation (WISE) Dataset

WISE study is a National Heart, Lung, and Blood Institute sponsored clinical study conducted to understand the clinical presentation of coronary artery disease in women [5]. During the study, female patients who had been assigned coronary angiogram had underwent several diagnostic tests to understand the causes of their chest pain or myocardial ischemia [5].

Before the administration of diagnostic tests, baseline evaluation data is collected from each patient. Baseline evaluation data includes demographic, clinical, angiographic, activity level information about patients, as well as physical symptoms such as the location and severity of the pain. After baseline evaluation, patients underwent several invasive and non-invasive diagnostic tests. Non-invasive tests are procedures performed without the insertion of needle, instruments or fluids into the body, and invasive tests are procedures that can range from blood tests (as it involves needles) to surgeries. Diagnostic tests that were administered in the WISE study include electrocardiogram, Dobutamine stress tests, pharmacologic stress tests without Dobutamine, angiogram, exercise stress test, radionuclide perfusion, brachial artery ultrasound [89]. Among these procedures, angiogram is seen as the gold standard test, however it is very invasive and costly.

Baseline evaluation data is available from the most of the patients, however not all patients had underwent all of the diagnostic tests. To limit additional problems due to missing data and focus only on the class imbalance problem, experiments in Chapter 4 and 5 uses only baseline evaluation data.

Events we try to predict in the WISE dataset are death, cardiovascular heart failure, myocardial infraction or stroke. The dataset is extremely skewed, in addition to Table 1.1, Figure 1.1 shows the number of patients who had one or more adverse events through out the study versus the patients who had not experienced the cardiac event or events.

In this thesis, we hypothesize that the WISE dataset has overlapping classes due to its time dependence: Features used to predict the events are collected in the beginning of the study. Over time, patients can get better, get worse, or stay the same. Hence initially sick patients may have no adverse events, or healthier patients may have complications due to heart disease. Therefore, we believe that this is an interesting dataset to study for the class imbalance problem.

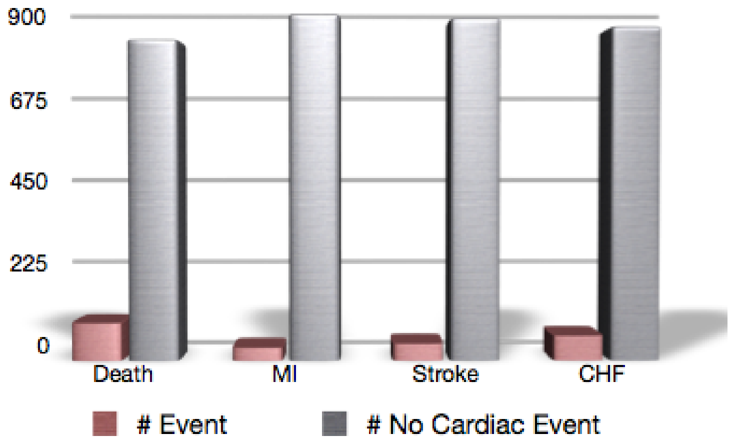


Figure 1.1: The distribution of cardiac events in the WISE dataset

1.4.2 20 Newsgroups

20 Newsgroups is text categorization dataset that contains around 20,000 documents categorized into 20 different topics. Following the literature [4], we picked the science category (sci.crypt, sci.electronics, sci.med, sci.space) as the minority class, and left the rest of the categories as the majority class. Disjunctive nature of the majority class, and class overlaps (i.e. potential overlaps between sci.crypt category and comp category) may aggravate class imbalance problem.

1.4.3 UCI Datasets

We compared the performances of the algorithms using several real-world datasets from the UCI Machine Learning repository [6]. These are the benchmark datasets that are commonly used by previous research in class imbalance. Selected UCI datasets are listed as follows:

- Satimage: Following the literature [90], we chose the smallest class as minority class, and collapsed the rest of the 5 the classes to one majority class.
- Mammography: There are two classes in this dataset, and the task is to distinguish between benign and malignant tumors. Only 260 out of 11183 tumors are malignant.
- Pendigits: Digit 0 is randomly selected as the minority class and the rest of the digits are collapsed into one big majority class.
- Pageblocks: Class 5 is chosen as the minority class, and the largest two classes (1 and 2) are collapsed into one big majority class.

1.5 Evaluation Metrics

Accuracy is often not used as a reliable evaluation metric on imbalanced datasets: by assigning every minority class to the majority class, any classifier can achieve high accuracy rates if the dataset is highly skewed. Therefore, throughout this thesis we relied on other (more robust) metrics such as the area under the ROC curve (AUROC), sensitivity, specificity and F-1 score. Specificity measures the proportion of true negatives in the dataset, whereas, sensitivity, (also known as recall) measures the proportion of the true positives. A varying decision threshold can be applied on the class probabilities given by a classifier to get a plot of sensitivity against specificity. This plot is defined as the ROC (Receiver Operating Curve), and commonly used to evaluate machine learning algorithms as well as statistical models in medicine. Since it is hard to compare two ROC plots over a range of thresholds, the area under the ROC curve (AUROC) is used for comparison. An AUROC score of 0.5 means that for every true positive, classifier generates a false positive. An AUROC score below 0.5 indicates a poor classifier. Higher AUROC score indicates the classifier is good at generating more true positives with fewer false positives [12].

F-1 score is another commonly used metric in imbalanced classification; it is the harmonic mean of precision and recall. Formulations of these metrics are given in Equation 1.1.

$$\begin{aligned}
\text{Accuracy} &= \frac{\#\text{true positives + true negatives}}{(\#\text{total number of test instances})} \\
\text{Specificity} &= \frac{\#\text{true negatives}}{(\#\text{true negatives} + \#\text{false positives})} \\
\text{Sensitivity (recall)} &= \frac{\#\text{true positives}}{(\#\text{true positives} + \#\text{false negatives})} \\
\text{Precision} &= \frac{\#\text{true positives}}{(\#\text{true positives} + \#\text{false positives})} \\
\text{F-1 score} &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}
\end{aligned} \tag{1.1}$$

Chapter 2

Literature Review

In this chapter, we review relevant work in imbalanced classification, active learning, anomaly detection and rare class discovery.

2.1 Imbalanced Classification

Research in imbalanced classification follows two main directions, data centric approaches and algorithm centric approaches. Data centric methods aim to modify the class distribution in the training set prior to learning. The simplest strategies for balancing the class distribution is randomly removing majority class samples from training data (random under-sampling) or replicating minority class samples (random over-sampling).

Despite its simplicity, random under-sampling has been shown to be effective against class imbalance [31]. However, under-sampling reduces information in the dataset, and increases variance in estimating model parameters [27]. Therefore, under-sampling should be performed in an informed manner, since randomly discarding training examples can potentially lead to the removal of informative samples; moreover, noisy samples might remain in the dataset.

Several informed under-sampling methods have been proposed to overcome the weaknesses of random under-sampling. However, many of these methods have not tackled class overlaps or small disjuncts extensively; yet alone incorporating the structure of minority class into the learning process. Some of these strategies treat minority class samples overlapping with the majority class as noise. For instance, one-sided selection (OSS) removes majority class instances that are distant from the k-Nearest Neighbor decision boundary, along with the minority class instances that overlap with the majority class [54]. Yen et al. clusters training data prior to sampling in order to provide representative samples from each “disjunct” [95]. Then they label every instance within a cluster of mostly majority class members as majority class [95]. Similarly, if a cluster has more minority class samples, all samples from that cluster are treated as if they are from the minority class. Unfortunately, these strategies are not well suited to datasets with many small overlapping disjuncts: examples from different classes that lie in the overlap region are either treated as if they belong to the same class, or removed from the training set. In this thesis, we do not treat overlapping examples as noise, on the contrary, we divert the focus of the classifier on the overlap regions. We believe that this strategy is better for learning adequate concept descriptions, by putting more emphasis on *hard to learn* examples.

Sampling from the overlap regions is not a new concept in imbalanced classification. Mani et al. proposed an informed under-sampling method that selects majority class instances based on their distance to the surrounding minority class samples [98]. They described 4 different variations of the proposed method. One of the variations, NearMiss-2, selects majority class instances whose average distance is smallest to k-farthest minority class examples. The goal is to focus on the overlap region by selecting negative samples that are close to all positive samples [98]. However, their method is prone to sampling mainly from only one region, and it can lead to reduced variability in the dataset. With our proposed method, robust under-sampling, we focus on the overlap regions, with the aim to keep intra-class variability in the majority class as well.

Random over-sampling duplicates minority class instances to balance the class distribution. Classifier trained on the final dataset may appear to be more robust (if the replicates are classified correctly), however over-sampling can potentially lead to over fitting [73], [52]. Moreover, over-sampling increases the computational burden, therefore it is not well suited to very large datasets. Finally, it has been shown that under-sampling is more effective against class imbalance than over-sampling [22]. Therefore, in this thesis we primarily focus on under-sampling, yet we still provide an overview of the over-sampling methods in this chapter.

Among the popular over-sampling methods, Synthetic Minority Over-Sampling (SMOTE) [90] can be listed. It creates synthetic minority class examples based on their distance to each other in the feature space. Each minority sample is connected with a line to any or all of its k-nearest neighbors; the number of nearest neighbors chosen depends on the amount of over-sampling needed [90]. Since SMOTE does not directly replicate minority examples, but synthesize new ones, it avoids over-fitting, as opposed to random over-sampling [35]. However, it gives each minority class instance equal importance and it does not put emphasis on hard to classify samples, i.e. examples that lie on the overlap region. In this thesis, we show that our new methods outperform SMOTE, demonstrating that methods that are robust to disjunctive/overlapping classes lead to higher prediction performances than this benchmark informative over-sampling method. More recent work in informed over-sampling addresses the question of which examples should be selected to replicate. Borderline-SMOTE assumes instances closest to the decision boundary are more important, since they may be prone to misclassification, and it uses them to generate artificial instances. Similarly, AdaSYN produces more synthetic examples from minority instances that are harder to learn [36]. These methods assume that having more instances around the decision boundary can improve prediction performance. However, if the decision boundary is misplaced, this hypothesis may be invalid. Furthermore, none of these approaches operate under disjunctive majority class assumption.

With Cluster Based over-sampling Jo et al. addresses small disjunct problem [47]. Their method clusters data prior to classification and random over-sampling is performed cluster by cluster [47]. Majority and minority class are clustered separately, and they are both oversampled based on the size of the cluster and maximum class size [47]. Their method is not well suited to very large datasets (as with other over-sampling techniques) and clustering should be reliable since over-sampling depends on the size of the cluster.

Algorithm centric approaches include cost-sensitive learning and ensemble learning [34], [11]. There are three main directions in cost sensitive learning:

- Making a cost-sensitive learner from a cost-insensitive learner, such as Adaboost [26], SVMs [29], decision trees [21], [85], [44], random forests [13], kNN [84], [89], [63].
- Assigning instances to the lowest risk class (class with lowest expected misclassification cost) [75]. Methods of this type usually rely on accurate class-conditional probabilities [75]. Examples include MetaCost [19], one of the earliest and the most well-known technique. MetaCost can be applied to any classifier: it wraps a meta-learning stage around the classifier such that the classifier minimizes the new cost rather than 0/1 loss [19]. Zadrozny et al. proposed a more general technique to apply when probabilities are unknown, and instances can have different costs [97].
- Converting a cost-insensitive classifier to cost-sensitive classifier by changing the distribution of training samples [96], [2], [75].

Adjusting the decision threshold or posterior probabilities can be listed as alternative methods [52]. Decision threshold of traditional cost-insensitive classifiers is set to 0.5. Provost et al. suggests modifying this threshold as a way to tackle class imbalance [73]. It has been shown that moving the decision threshold has the same effect as adjusting the cost matrix [67]. An example cost matrix is given in Table 2.1:

	Actual Positive	Actual Negative
Predicted Positive	C_{11}	C_{10}
Predicted Negative	C_{01}	C_{00}

Table 2.1: Cost matrix for binary classification

Elkan shows that in a special case, where the decision threshold of the cost-insensitive classifier is 0.5, and $C_{00} = C_{11} = 0$, the number of majority class instances should be multiplied by $\dfrac{C_{10}}{C_{01}}$ [23]. When costs are unknown, the suggested way is to multiply it by ratio of the number of minority class instances to the number of majority class instances in the dataset [46]. It has been also empirically demonstrated that random under/over-sampling has the same effect of adjusting the cost matrix, or moving the decision threshold [67], [9]. In this thesis, we compare our proposed methods to cost-sensitive learners, and empirically show that, if the assumption of the learners are satisfied, our proposed methods outperform baseline cost-sensitive methods.

2.2 Active Learning

Active learning is a branch of machine learning, where the goal is to achieve greater learning performance with fewer training instances. This is achieved by giving the learning algorithm the control over the data from which it learns. Active learning is well suited to problems where unlabeled data is abundant, initial training (labeled) data is scarce, and labels are difficult to obtain. As opposed to passive learning, rather than requiring labels for the entire unlabeled dataset, an active learner can choose training instances with a querying strategy and require labels for only those that are selected. An optimal querying strategy can minimize labeling costs, while achieving greater learning performance than passive learning with less training data. It has been shown by Dasgupta et al. [16] that under some constraints, active learning approaches can reduce the number of queries needed to learn a concept exponentially, compared to passive learning.

Based on where queried instances are sampled from, there are three main problem settings in active learning: membership query synthesis, pool based active learning and stream based active learning [76]. In membership query synthesis, queried instances may be artificially synthesized; in pool based active learning, queried instances are selected from a readily available pool of unlabeled instances; and in stream based active learning, the learner has an option to query or discard instances coming in a stream [76]. In this review, we focus on pool based active learning, since it is the most relevant setting for learning under high-class distribution.

Among the most commonly used query strategies, uncertainty sampling [60], query-by-commitee [77], [28], and expected error reduction [74] can be listed. Uncertainty sampling queries the unlabeled instances whose labels the learner is the the least certain about [76]. For probabilistic models, this translates to having a posterior probability close to 0.5 [60], [59]. For support vector machines, uncertain instances are the ones that are closest to the separating hyperplane. Query-by-committee utilizes an ensemble of learners (committee) and select instances with the most disagreement in labels; disagreement can be measured by vote entropy or the Kullback-Leiber divergence [77]. This can be thought as a generalization of entropy based uncertainty sampling. Expected error reduction strategy calculates the expected future risk (such as 0/1 loss) of every unlabeled instance over all possible labels, and select the one with the minimal risk [76]. There are also ensemble-based active learning strategies, which incorporate ensemble-methods such as boosting or bagging into active learning [68], or aim to increase diversity among committee members [69]. For a detailed information on these strategies, readers can refer to the extensive review by Settles [76].

There are two main motivations of using active learning on skewed datasets: Minimizing the number of required labels to learn concept descriptions, and balancing the skewed class distribution. Previous approaches in this field either balance the initial training dataset first and then use a standard query strategy; or develop a new query strategy that is robust for datasets with high class skew. Among the first group of approaches, Haines et al. synthesized artificial minority class instances to balance the dataset with random subspace resampling and applied uncertainty based active learning to the modified balanced dataset [78]. Zhu et al. proposed BootOS algorithm, which eliminates within-class imbalance problem with bootstrap based over-sampling then uses uncertainty sampling as an active learning strategy [101].

Researchers in the second group leaned towards using a cost-sensitive learner, rather than first balancing the training set. Ertekin et al. observed that class imbalance is less severe around the decision boundary thus they sampled instances that are closest to the SVM hyperplane [24]. Subsequently, Bloodgood and Shanker observed that, under high class skew, SVM hyperplane can be too close to the positive examples, which could lead to low recall [7]. To alleviate this problem, they used cost-sensitive SVMs by setting the cost ratio either to the level of imbalance in the initial training data, or to the estimate of overall corpus imbalance [7]. Ratio estimates come from a random subsample of the data, which under extreme class skew, may not reveal an accurate class distribution [4]. Li et al. proposed an active learning strategy that is based on both *uncertainty* and *certainty* measures to ensure that the active learner samples a balanced set, and sampled instances are informative [61]. Moreover, these strategies assume that the decision boundary learnt from the initial labeled set is reliable enough to guide querying strategy. If the labeled set is small, and if majority class consists of many small disjuncts, the initial learner may not be reliable enough to give an accurate certainty or uncertainty measure. Attenberg et al. has shown that under extreme class skew with overlapping disjuncts, classifier may not be able to produce reliable posterior probabilities especially during the initial stages of active learner [4]. In this thesis, to account for the unreliable class probabilities, we propose a new strategy: active learning with maximum probability. The location of the decision boundary does not hold significant importance for the maximum probability strategy, the classifier is rather used as a "ranker". The goal is to keep the labeled set as balanced as possible by querying unlabeled instances with the highest probability belonging to the minority class. If the classifier mistakes them being in the minority class, then these *majority* samples are highly informative for learning. Tomanek and Hahn applied query by committee based active learning strategies for named entity recognition. They propose multiple algorithms where either minority class instances are oversampled during active learning to balance the dataset or labeled set is kept as balanced as possible by incorporating class specific costs while modifying the active learning algorithm [88]. This method is closely related to the "cost-sensitive" active learning mentioned above.

To avoid sampling before learning, and to minimize computational costs, in this dissertation, we propose a strategy that resembles balancing the dataset prior to active learning. This strategy, active learning with threshold selection, utilizes a varying decision threshold to account for varying skew levels in the dataset. Finally, we propose an "unsupervised" active learning strategy to use when the initial training set does not have enough labels to learn the concept descriptions. This method measures the similarity between unlabeled instances with the minority class instances, and queries those with the greatest overlap.

2.3 Unsupervised Anomaly Detection

Learning under skewed class distributions is closely related to anomaly detection problem in machine learning, hence we provide a brief overview of the recent work in anomaly detection. Unlike imbalanced classification, anomaly detection is often unsupervised (or labels are known only for the "normal" data). Anomalies are patterns that do not resemble the majority (normal) patterns and are assumed to be isolated examples, with minimal shared properties. They can be categorized under point anomalies (an instance is anomalous with respect to the data), contextual anomalies (an instance is anomalous within a context) or collective anomalies (instances are anomalous together but not by themselves). Much work in anomaly detection focus on individual outliers rather than small classes, or self-similar items. An anomaly detection algorithm either classifies an instance as normal or anomaly, or it assigns an anomaly score to it. Unsupervised anomaly detection methods can be clustering based, information theoretic based or spectral based [10]. Clustering methods [25], [83], [58], assume majority ("normal") instances belong to large clusters, and anomalies either form small clusters or do not form any cluster. This assumption may not be valid in many applications where majority class is disjunctive with many small clusters. In such cases, distinguishing anomalous clusters from normal clusters can be extremely hard. Moreover, point anomalies that lie in the normal cluster, but away from the cluster center, are almost impossible to detect. Information theory based methods [57], [51], [91], [71] calculate a score based on a metric. The idea is that anomalies add information to the dataset, causing significant changes to the overall score. Finding the right information metric that is sensitive enough to detect alterations is often a challenge. Spectral Analysis methods [62], [56], [87] assume that, unlike anomalous data, normal data can be explained in the reduced dimension. Using eigen decomposition, anomalies can be detected by observing the lowest eigenvalues, which may have a lot of variability for anomalous instances.

There are also classifier-based anomaly detection methods that do not necessarily depend on labeled data: Jakkula et al. utilizes one class support vector machines to detect anomalous sensor events [45]. One class support vector machine is also used for feature selection for anomaly detection [53], [99], [66]. Kloft et al. develops a method based on SVDD that automatically selects different sets of features, rather than a single feature set [53]. This approach allows to obtain several detectors, as opposed to single one, for various type of network intrusion attacks.

Anomaly detection has applications in various domains including fraud detection, insurance risk modeling and spam detection. There is a greater literature for unsupervised anomaly detection, interested readers can refer to the extensive review by Chandola et al. [10].

2.4 Rare Class Detection

Both imbalanced classification and rare class detection deal with highly skewed datasets, where the latter aims to detect the rare classes de-novo from a few examples. Pelleg et al. fits a mixture model to the training data, and uses active learning to sample points with different criteria such as low likelihood, uncertainty, a combination of both, and interleaving [14]. He et al. assumes minority classes form a compact cluster in the feature space, and the distribution of the majority class is sufficiently smooth [37], [38], [39]. Stokes et al. proposes the system ALAADIN where anomaly score is computed by the sum of negative log likelihood, and items with the highest anomaly score is presented for labeling [82]. Dasgupta et al. presents an active learning scheme that exploits the cluster structure in the data, which was proven to be effective in rare category detection [17]. Hospedales et al. observes that generative and discriminative classifier performances vary with the training sample size so they change the classifier (SVM or logistic regression) and active query strategy (uncertainty or low likelihood) in different stages during the learning process [43].

The main difference between rare class detection and imbalanced classification is that the former focuses on the detection of rare categories when no labeled samples are available apriori and the latter assumes that there are labeled examples from the minority (rare) class, and it tries to improve prediction performance especially on the minority class. Compact minority class hypothesis has been investigated under rare class detection, and in this thesis, we demonstrate that it is closely tied to imbalanced classification as well.

Chapter 3

Cost Sensitive Learning for Heart Disease

3.1 Introduction

Many tasks in medical diagnosis deal with learning under high class skew; clinical datasets have only a few examples from the minority class, and yet, correct identification of positive samples is extremely important. For instance, in the risk stratification of heart disease, the number of high-risk patients tend to be much lower than the number of low-risk patients, and mis-categorizing a high-risk patient with severe stenosis into low-risk can have fatal consequences.

In this chapter, we present our preliminary analysis of the Womens' Ischemic Syndrome Evaluation (WISE) dataset, to provide an example of a medical diagnosis task that deals with high class skew. Our contributions in this work are as follows: Firstly, we evaluate the prediction performances of several cost-sensitive algorithms on the WISE dataset. Secondly, we compare cost-sensitive algorithms with a widely used clinical risk stratification metric, American Health Association risk guidelines. Our findings in this chapter revealed that state of the art cost-sensitive algorithms provide a better risk-stratification than widely adopted guidelines.

3.2 Cost-sensitive supervised learning algorithms

In this section, we give formulations of the cost-sensitive algorithms used in the experiments: cost-sensitive K-NN, cost-sensitive SVM, and cost-sensitive logistic regression. The binary classification problem is as follows: Given labeled training dataset $\mathcal{D}$ of n tuples, $\mathcal{D}$ = $\{(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, $(\mathbf{x}_3, y_3)$, ..., $(\mathbf{x}_n, y_n)\}$ and the target space, $\mathcal{T}$ = $\{t_1, t_2\}$, where $\mathbf{x}_i \in \mathcal{R}^p$ are the feature vectors, the goal is to infer binary class labels $y_i \in \{t_1, t_2\}$. Since the WISE dataset has high class skew, i.e., $P(t_1) << P(t_2)$, we refer t_1 as the minority class, and t_2 as the majority class.

3.2.1 Cost-sensitive KNN

k-NN is a non-linear supervised learning algorithm that classifies an instance based on the votes (labels) of its k-closest neighbors in the feature space. k-NN has been successfully utilized for clinical diagnosis by prior research [79]. It is particularly applicable to datasets where data forms natural clusters in the feature space, and there are partial class overlaps.

Given a distance measure d, k-NN first finds the set of k-nearest neighbors, N_i of a test instance x_i. Then, for each label t in the target space, it computes the number of neighbors, $V_i(t)$, with the label t. Formally, under the majority voting scheme, $V_i(t)$ is computed based on the following formula:

$$V_i(t) = \sum k \in N_i(I(t, y_k)) \tag{3.1}$$

where I is an indicator function, that is I(t, y_k) = 1 if t = y_k, 0 otherwise. Predicted target variable of $\mathbf{x}_i$ can then found using Equation 3.2:

$$\hat{y}_i = \arg\max_{t \in T} C_t V_i(t) \tag{3.2}$$

Cost-insensitive k-NN may perform poorly on datasets with class imbalance. Under high class skew, samples from the majority class dominate the neighborhood of any test instance. This leads to higher votes for the majority class than for the minority class. As a result, k-NN tends to label each example with the majority class label.

Cost-sensitive k-NN (C-KNN), addresses this problem by assigning class-based weights to each instance, thus weighing the votes of the neighbors. With the new weighted voting scheme, having n number minority class instances in the vicinity is more important than having n number majority class instances (given that the minority class has higher weights then the majority class). This can also penalize uninteresting classes, whether majority or not.

With the weighted voting scheme, the domination of majority class in the feature space can be alleviated. Denoting the weight vector corresponding to class labels t_1 , t_2 with $\mathbf{w} = [w_{t_1}, w_{t_2}]$, weighted voting scheme can be formalized as follows:

$$\hat{y}_i = \arg\max_{t \in T} w_t V_i(t) \tag{3.3}$$

3.2.2 Cost-sensitive SVM

Cost-sensitive SVM is one of the most commonly used methods in imbalanced classification. Cost-sensitive SVMs address class imbalance by altering the location of the decision boundary based on the cost vector C, as opposed to placing it in the middle of two classes. The objective function of the cost-sensitive SVM under the cost vector $\mathbf{C} = [C_{t_1}, C_{t_2}]$, is given with the Equation 3.4.

$$\begin{aligned}
&\min_{\mathbf{w},e,b} \frac{1}{2}\|\mathbf{w}\|^2 + C_{t_1} \sum_{i|y_i=t_1} \epsilon_i + C_{t_2} \sum_{i|y_i=t_2} \epsilon_i \\
&\quad \text{subject to:} \\
&\quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \epsilon_i, \qquad \forall i = 1 \ldots n \\
&\quad \epsilon_i \geq 0 \qquad \forall i = 1 \ldots m
\end{aligned} \tag{3.4}$$

where class labels $\mathbf{y}_i$ can be {-1,1}.

3.2.3 Cost-sensitive Logistic Regression

Cost-sensitive logistic regression is essentially equivalent to up or down sampling the training instances based on the class costs. In this section, we will explain this by augmenting the log likelihood function of logistic regression with class costs.

Under the logistic regression model, posterior probability a sample belonging to the minority class is given in Equation 3.5:

$$P_\theta(\hat{y} = 1|\mathbf{x}) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}} \tag{3.5}$$

Assuming that the training instances are identically and independently distributed, and there are m instances in the training set, the log likelihood can be written as Equation 3.6:

$$l(\theta) = \sum_{i=1}^{m} y_i \log(P_\theta(y = 1|\mathbf{x}_i)) + (1 - y_i) \log(P_\theta(y_i = 0|\mathbf{x}_i)) \tag{3.6}$$

As can be seen in Equation 3.7, integrating the cost vector $\mathbf{C} = [C_{t_1}, C_{t_2}]$ to the log likelihood, where C_{t_1} is the cost associated with the minority class and C_{t_2} is the cost associated the majority class, is equivalent to providing more/less training instances to the model (depending on the C).

$$l(\theta) = \sum_{i=1}^{m} C_{t_1} y_i \log(P_\theta(y = 1|\mathbf{x}_i)) + C_{t_2}(1 - y_i) \log(P_\theta(y_i = 0|\mathbf{x}_i)) \tag{3.7}$$

3.3 Empirical Results

3.3.1 Performance comparison of the cost-sensitive classifiers

In this section, we compared the performances of cost-sensitive k-NN, cost-sensitive SVM, and cost-sensitive logistic regression on mortality, MI, stroke, CHF and any cardiac event prediction using the WISE dataset. We converted the problem to a binary classification problem: for each event type (such as MI), the goal is to classify patients having that event to the minority class, and classify patients who did not have that event to the majority class. For each algorithm, the costs are set to the ratio of minority to majority class instances in the dataset. Python scikit-learn [1] is used for the implementation. For k-NN, the number of neighbors is set to 10. RBF kernel is selected for SVM. l_2 penalty is used to regularize logistic regression.

The AUROC and F-1 comparison of each algorithm on all 5 prediction tasks are shown in Figures 3.2 and 3.1 respectively. Results are averaged over 10-fold cross validation, and mean scores are reported. As can be seen in the Figures, cost-sensitive SVMs outperforms cost-sensitive logistic regression and cost-sensitive K-NN on all 5 prediction tasks. The highest performance in AUROC is achieved in CHF prediction task: SVM achieved an AUROC score of 0.877. The highest

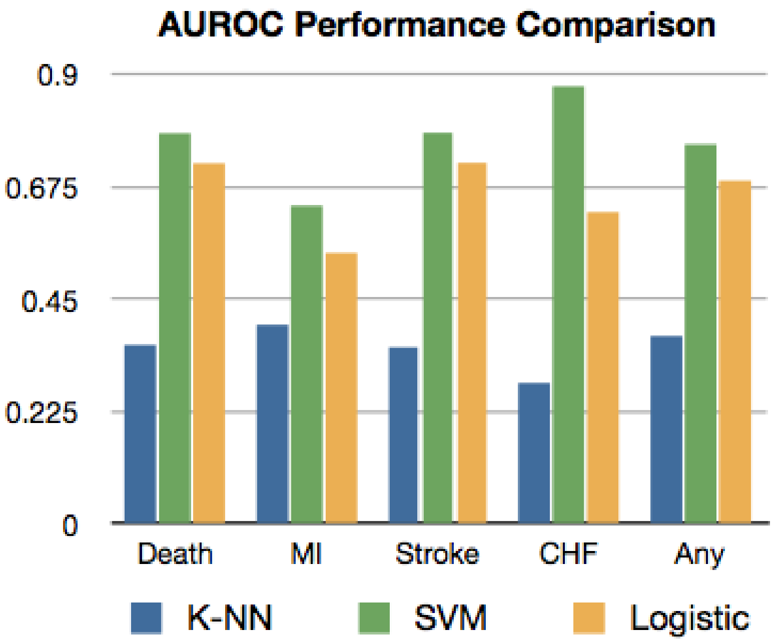


Figure 3.1: AUROC comparison of cost-sensitive KNN, cost-sensitive SVM and cost-sensitive logistic regression on mortality, CHF, MI, stroke, and any cardiac event prediction tasks

performance in F-1 is achieved in mortality prediction, highest score being 0.399. Even though the AUROC scores of SVM on all 5 prediction tasks are higher than 0.5, the F-1 performances of all 3 algorithms are unsatisfactory. This motivated our thesis study: tackling class imbalance under different assumptions in order to achieve high prediction performance on the minority class.

3.3.2 Comparison with the baseline method

American Heart Association (AHA) provides a set of widely-used guidelines to identify patients who are at high or low-risk of having heart disease. It calculates the risk score based on the factors that affect the likelihood of heart disease, such as diabetes, excess weight, high blood pressure, smoking etc. Based on the presence of all or some of these factors, patients are categorized into different risk groups.

In this section, we compared each of the cost-sensitive methods with the AHA guidelines. To make this comparison, Framingham score and metabolic syndrome indicator is added to the baseline evaluation data, since they are used by AHA to calculate the risk score [70]. Using cost-sensitive classifiers, we partitioned patients into high and low risk groups, based on whether they had any heart-disease related adverse events (death, stroke, MI or CHF).

After classification with the respective methods, we employed Kaplan-Meier survival analysis [50] to compare the rates of having an adverse event. Rather than removing patients who dropped

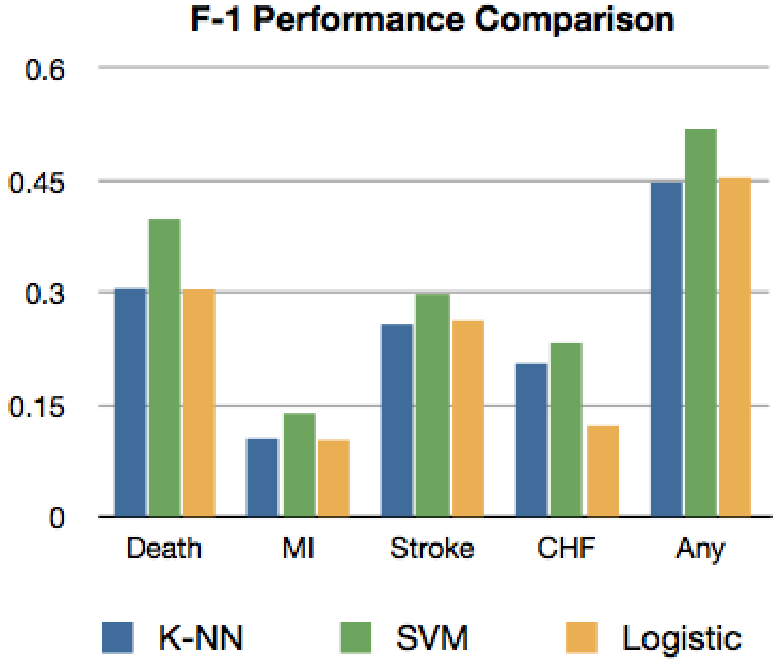


Figure 3.2: F-1 comparison of cost-sensitive KNN, cost-sensitive SVM and cost-sensitive logistic regression on mortality, CHF, MI, stroke, and any cardiac event prediction tasks

Time t	**Event**	**No-Event**	Total
Group 1 (High Risk)	d_t^1	$n_t^1 - d_t^1$	n_t^1
Group 2 (Low Risk)	d_t^2	$n_t^2 - d_t^2$	n_t^2
Total	d_t	$N_t - d_t$	N_t

Table 3.1: 2×2 table used to compute the hazard ratio

out of the WISE trial or who passed away, we included them in the analysis since Kaplan-Meier is suitable for patient censoring. The survival curves for each of the methods are shown in Figures 3.3 and 3.4. Patients who left the trial prior to completion (censored patients) are represented by ticks. Red line corresponds to the predicted lower-risk patients, and black line corresponds to high-risk patients.

In order to be able to compare the survival curves, we calculated the hazard ratios (HR) and p-values for each method using the MStat package [20]. Hazard ratio is a measure that is frequently used in clinical trials. To calculate it, we followed MStat guidelines, and formed the Table 3.1 for the two risk groups: where d_t^i stands for the number of events in group i at time t and n_t^i stands for the number of patients in group i at time t. N_t is the total number of patients and d_t stands for the total number of events at time t. Based on 3.1, the number of observed

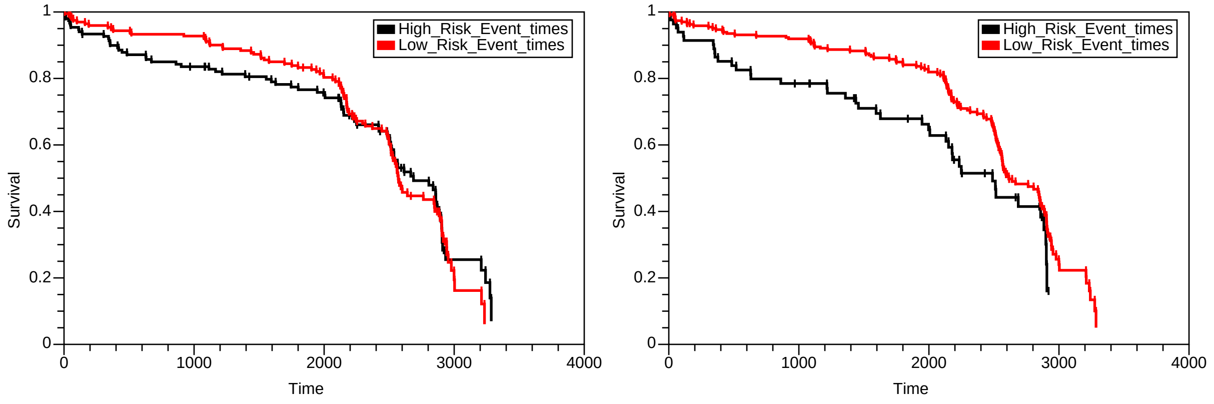


Figure 3.3: Survival curves after classification with AHA guidelines (left), and after classification with cost-sensitive SVM (right)

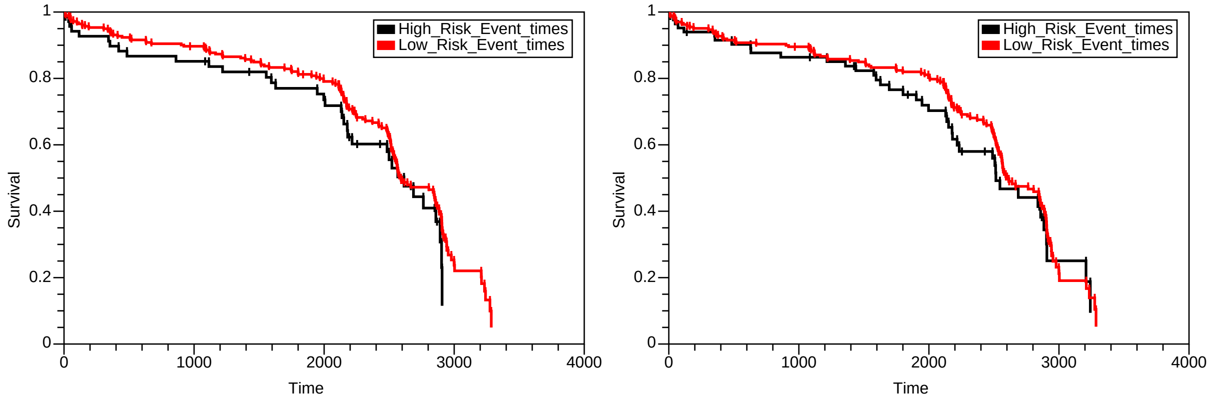


Figure 3.4: Survival curves after classification with cost-sensitive KNN (left), and cost-sensitive logistic regression (right)

events, and the total number of expected events for each group i can be written as:

$$\begin{aligned} O^i &= \sum_{t=1}^{T} d_t^i \\ E^i &= \sum_{t=1}^{T} \frac{n_t^i d_t^i}{N_t} \end{aligned} \tag{3.8}$$

Following [20], hazard ratio of the algorithms on the two risk groups is computed with 3.9. First group is the high-risk group and the second group is the low-risk group.

$$\text{Hazard Ratio} = \frac{O^1/E^1}{O^2/E^2} \tag{3.9}$$

If hazard ratio = 1, then there is no difference between the two groups, if it is > 1, then events are more frequent in the high risk group than in the control group, if it is < 1 events are less frequent in the high risk group than in the control group. In our case, control group is the low-risk patients, and a higher hazard ratio signifies that the corresponding classification algorithm is better in identifying high risk patients.

The hazard ratios and p-values for each method is shown in Table 3.2. As can be seen in Table 3.2, prediction with AHA guidelines results in a hazard ratio that is lower than 1, meaning that patients who are classified as high-risk are less likely to have any adverse event than patients in the low-risk group. Evidently, this is incorrect, the hazard ratio is expected to be higher than 1.

Among the machine learning algorithms, SVM is better than classifying patients to the correct risk groups: it achieves the highest hazard ratio (hazard ratio $= 1.565$) with the lowest p-value ($p = 0.01$). Having a p-value lower than 0.05 demonstrates statistical significance, hence it is safe to conclude that cost-sensitive SVM should be preferred over AHA guidelines in the risk stratification of heart disease.

Method	**HR**	**P Value**
AHA	0.9813	0.9009
K-NN	1.264	0.2132
SVM	1.565	0.01055
Logistic Regression	1.197	0.3074

Table 3.2: Hazard ratios and p-values of the predictions by AHA and the cost-sensitive algorithms

3.4 Chapter Conclusions

In this chapter, we utilized various cost-sensitive algorithms for risk stratification in heart disease. This work is among the first experiments we conducted on the WISE dataset, which to our knowledge, had not been analyzed using machine learning techniques before. Our empirical results revealed that cost-sensitive SVM outperforms other cost-sensitive methods such as cost-sensitive KNN, and cost-sensitive logistic regression on all 5 prediction tasks, mortality, stroke, MI, CHF and any cardiac event prediction. Using cost-sensitive SVM, we achieved over 0.783 AUROC score for mortality prediction. This is promising for using data mining in the diagnosis of heart disease in women.

Our biggest contribution in this chapter is that our approach outperforms state-of-the-art, conventional risk guidelines for cardiovascular heart disease. This suggests that using machine learning methods rather than rule or guideline based systems should be in the roadmap for segmenting patients into correct risk groups. However, even though we achieved better performance than the conventional guidelines using cost-sensitive learning methods, we obtained unsatisfactory F-1 scores. This motivated the later works in this dissertation: there is a need to develop more advanced methods than the current supervised learning algorithms to achieve high prediction performance on the minority class under high-class skew. For this purpose, in the next chapter, we study learning under high-class skew under the compactness assumption.

Chapter 4

Multiple Kernel Learning for Imbalanced Classification

4.1 Introduction

Learning under skewed class distributions is related to anomaly detection: both tasks focus on the correct identification of the minority class patterns. Unlike imbalanced classification, anomaly detection is often unsupervised (or labels are known only for "normal" data). Furthermore, in most unsupervised anomaly detection problems, anomalies are assumed to be isolated examples, with minimal shared properties, and there is a single "normal" pattern. This assumption may not hold in most highly-skewed classification problems: for instance, in the diagnosis of heart disease, patients who are under high risk of having a myocardial infarction may present similar symptoms, such as high blood pressure, or severe angina. Yet, healthier patients may not necessarily conform to a single "normal" pattern. Similarly, in mammography, malignant masses may be similar in terms of shape and density, but benign masses may differ. This is consistent with the minority class compactness hypothesis.

One-class SVMs have been utilized successfully in unsupervised anomaly detection to identify patterns that are unusual or different from the normal pattern. The assumption is that the majority class samples can be contained within a hypersphere. If the data does not naturally satisfy this assumption, it can be mapped to a (possibly) higher dimension with a kernel to fit the assumptions of the learner. However, there is one caveat to this approach: kernel can be chosen arbitrarily or using cross validation, if the right kernel is not used, the assumption may remain unsatisfied, hence the final prediction may not be optimal.

In this section, we first hypothesize that the minority class forms a compact cluster in the feature space, and majority class examples lie outside of this cluster. Compactness assumption has been shown useful in previous work in rare category characterization [39]. However, previous work failed to show that there are datasets that the compactness assumption may not hold true with the chosen kernel. In this chapter, we show that on the WISE dataset, automatically learning kernel combination may be necessary; as one-class SVMs are not as effective as they are in anomaly detection. We conclude the chapter by stating that on the WISE dataset, even with multiple kernel learning, minority class itself cannot be contained in a hypersphere, suggesting that the clinical presentation of acute heart disease may be in several different forms.

4.2 Motivation

In this section, we describe the motivation behind this chapter. As we mentioned before, in highly skewed classification, we can assume that the minority class forms compact clusters in the feature space, and this structure can be leveraged towards better prediction. To test this hypothesis, we conducted three separate classification experiments: first one assumes that the minority class is compact and learns only from the minority class (with *nNegative* instances from the majority class added in the training set), the second experiment assumes that the majority class is compact and learns only from the majority class (with *nPositive* instances from the minority class added in the training set). Third experiment uses all data, regardless of their labels. We refer to first and second experiments as Positive Compact and Negative Compact respectively, and refer third experiment as one-class SVM. Rather than utilizing a portion of the labeled data for Positive and Negative Compact methods, one could also try a transductive approach, similar to transductive SVMs [48]. However, our goal is to test the compactness hypothesis, so we have not explored that option in this section.

All experiments use Gaussian RBF kernel, and experiments are done using Python Scikit-learn [1] one-class SVM classifier. *nNegative* and *nPositive* are found by taking %0.01 of the total number of minority and majority class instances in the training set respectively.

We measured the F-1 score of Positive Compact, Negative Compact, one-class SVM and cost-sensitive logistic regression on highly skewed datasets from the UCI database, and then we measured their performance on the WISE dataset for mortality, myocardial infarction (MI), congestive heart failure (CHF), and stroke prediction. For cost-sensitive logistic regression, we multiplied the number of majority class instances with the class skew in the training set $\left(\frac{\#minoritysamples}{\#majoritysamples}\right)$. Figure 4.1 shows the results on the Satimage and Pendigits datasets. As evident from the graphs, Positive Compact approach outperforms all of the baseline approaches, Negative Compact, one-class SVM and cost-sensitive logistic regression. This leads to two major findings: 1. Data supports the major hypothesis of this thesis: positive (minority) class forms a compact cluster in the feature space 2. A method that leverages the cluster structure of minority class outperforms state of the art baseline approaches.
Figures 4.2 and 4.3 show the F-1 performances on the WISE dataset. On the WISE dataset, Positive Compact underperforms compared to Negative Compact, One-Class SVM, and cost-sensitive logistic regression. This indicates that the compactness assumption may not hold on certain datasets. As a next step, rather than invalidating the hypothesis completely, we investigated whether this is a result of choosing the wrong kernel. In the following section, we apply Multiple Kernel Learning approach with compactness assumption on the WISE dataset to see whether the right kernel combinations improve prediction performance.

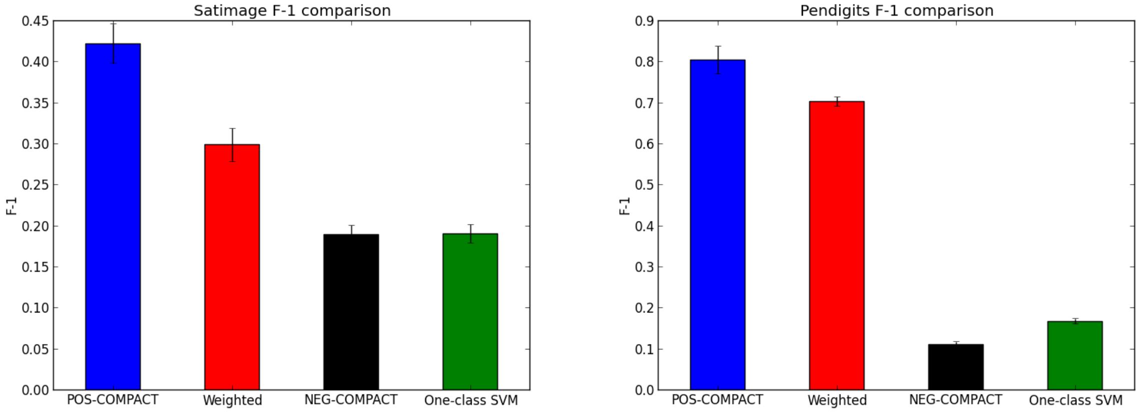


Figure 4.1: Evaluating whether the minority class is compact on Satimage and Pendigits datasets

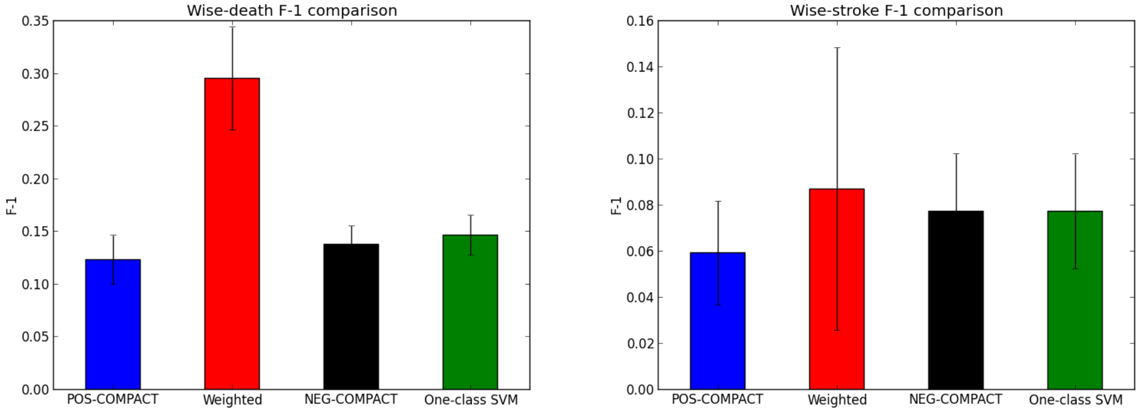


Figure 4.2: Evaluating whether the minority class is compact on the WISE dataset (mortality and stroke prediction)

4.3 Problem Formulation

In this section, we provide the formulations for one-class SVMs and multiple kernel learning for one class learning.

4.3.1 One-class SVM

One-class SVM tries to model the data by placing majority of the instances in a hypersphere with a center $\mathbf{c}$ and radius $R > 0$ [86]. The goal is to place all (or as many) instances within the hypersphere while minimizing the volume by minimizing R^2. The instances that cannot be placed inside the hypersphere are penalized with slack variables, $\xi \geq 0$. Given n instances $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n \in \mathcal{X}$, one-class SVM can be formulated with Equation 4.1. C is the parameter that controls the tradeoff between volume and point-wise violations. As $C \to \infty$, hypersphere

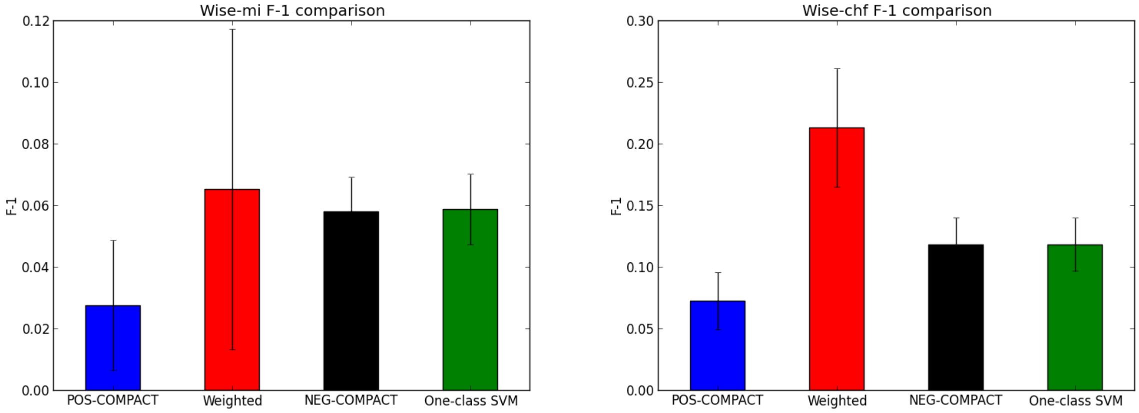


Figure 4.3: Evaluating whether the minority class is compact on the WISE dataset (MI and CHF prediction)

includes all points, and as $C \to 0$, hypersphere reduces to the centroid.

$$\begin{aligned} \underset{\mathbf{c},R,\xi,\beta}{\arg\min} \quad & R^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \forall_{i=1}^{n} : ||\mathbf{x}_i - \mathbf{c}||^2 \leq R^2 + \xi_i \\ & \forall_{i=1}^{n} : \xi_i \geq 0 \end{aligned} \tag{4.1}$$

If the data is not naturally clustered in an Euclidean space, mapping it to a higher dimensional space via a kernel function may be necessary. In the kernelized formulation, the dot product is replaced by the kernel, k, i.e. $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$, where ϕ is a feature map to a D dimensional feature space, $\phi(\mathbf{x}) \to R^D$. Using kernels, one-class SVM can now be written as 4.2:

$$\begin{aligned} \underset{\mathbf{c},R,\xi,\beta}{\arg\min} \quad & R^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \forall_{i=1}^{n} : ||\phi(\mathbf{x}_i) - \mathbf{c}||^2 \leq R^2 + \xi_i \\ & \forall_{i=1}^{n} : \xi_i \geq 0 \end{aligned} \tag{4.2}$$

As an example of a kernel function, Gaussian or Polynomial kernels can be given (formulations are given in Equation 4.3):

$$\begin{aligned} k(\mathbf{x}_i, \mathbf{x}_j) &= e^{\frac{-||\mathbf{x}_i - \mathbf{x}_j||}{\sigma}} \\ k(\mathbf{x}_i, \mathbf{x}_j) &= \langle \mathbf{x}_i, \mathbf{x}_j + 1 \rangle^q, q \in \mathcal{N} \end{aligned} \tag{4.3}$$

One-class SVM defines instances that lie inside the hypersphere, i.e. $||\phi(\mathbf{x}_i) - \mathbf{c}||^2 - R^2 < 0$ as normal data points. The instances that lie outside of the hypersphere, i.e. $||\phi(\mathbf{x}_i) - \mathbf{c}||^2 - R^2 > 0$ are classified as anomalous points.

4.3.2 One-class Multiple Kernel Learning

Rather than using a single, fixed kernel that is determined outside of learning, multiple kernel learning (MKL) allows us to combine different kernel combinations and pick the one that works the best automatically [32]. In MKL, there are d different feature mappings, $\phi_1, \phi_2 \ldots \phi_d$ where $\phi_i(\mathbf{x}) \to R^{D_i}, i = 1 \ldots d$ where D_i is the dimensionality of the i^{th} feature space. Each feature mapping corresponds to a different kernel. These kernels can then be combined linearly, or non-linearly. In this chapter, we focus on weighted linear combination of kernels.

Applying MKL paradigm to one-class SVM, we can re-write Equation 4.1 as Equation 4.4:

$$\begin{array}{ll} \underset{\mathbf{c},R,\xi,\beta}{\arg\min} & R^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \forall_{i=1}^{n} : ||\varphi_\beta(x_i) - \mathbf{c}||^2 \leq R^2 + \xi_i \\ & \forall_{i=1}^{n} : \xi_i \geq 0 \\ & \sum_{i=1}^{d} \beta_i = 1 \\ \text{where} & \varphi(x_i) = (\sqrt{\beta_1}\phi_1(x_i), \sqrt{\beta_2}\phi_2(x_i), ..., \sqrt{\beta_d}\phi_d(x_i))^T \end{array} \tag{4.4}$$

$\sum_{i=1}^{d} \beta_i = 1$ imposes an L_1 constraint on the kernel coefficients, thus allows us to select the combination that best describes the data. This optimization problem can be solved with semi-infinite linear programming (SILP). To do that, the non-convex objective function of Equation 4.4 should be re-arranged to solve for the Lagrangian. This can be done through variable substitution by introducing the variable $\mathbf{v}$ and setting $\mathbf{c}_j = \frac{\mathbf{v}_j}{\sqrt{\beta_j}}$ [53]. This will be clear if the first constraint is expanded:

$$\begin{array}{ll} \underset{\mathbf{c},R,\xi,\beta}{\arg\min} & R^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \forall_{i=1}^{n} : \sum_{j=1}^{k} \beta_j \langle \phi_j(\mathbf{x}_i), \phi_j(\mathbf{x}_i) \rangle - 2\sum_{j=1}^{k} \langle \sqrt{\beta_j}\phi_j(\mathbf{x}_i), \mathbf{c} \rangle + \langle \mathbf{c}, \mathbf{c} \rangle \leq R^2 + \xi_i \\ & \forall_{i=1}^{n} : \xi_i \geq 0 \\ & \sum_{i=1}^{d} \beta_i = 1 \\ \text{where} & \varphi(x_i) = (\sqrt{\beta_1}\phi_1(x_i), \sqrt{\beta_2}\phi_2(x_i), ..., \sqrt{\beta_d}\phi_d(x_i))^T \end{array}$$

Setting $\varphi(\mathbf{x}_i) = 0$ to solve for $\forall i = 1 \dots n_1$ $R^2 = \langle \mathbf{c}, \mathbf{c} \rangle + \gamma^2$, we can remove the non-linear dependency between β and $\mathbf{c}$, and obtain Equation 4.5 [53].

$$\begin{aligned} &\underset{\mathbf{v},\gamma,\xi,\beta}{\arg\min} && \sum_{j=1}^{k} \frac{1}{\beta_j} \langle \mathbf{v}_j, \mathbf{v}_j \rangle + \gamma^2 + C \sum_{i=1}^{n_1} \xi_i \\ &\text{subject to} && \forall_{i=1}^{n} : \sum_{j=1}^{k} \beta_j \langle \phi_j(\mathbf{x}_i), \phi_j(\mathbf{x}_i) \rangle - 2 \sum_{j=1}^{k} \langle \phi_j(\mathbf{x}_i), \mathbf{v}_j \rangle \leq \gamma^2 + \xi_i \\ &&& \forall_{i=1}^{n} : \xi_i \geq 0 \\ &&& \sum_{i=1}^{d} \beta_i = 1 \\ &&& \beta \geq 0 \end{aligned} \tag{4.5}$$

Introducing Lagrange multipliers $C \geq \alpha \geq 0$ we can obtain the Lagrangian in Equation 4.6.

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{k} \beta_j K_j(x_i, x_i) - \alpha_i \alpha_j \sum_{j=1}^{k} \beta_j K_j(x_i, x_l) \tag{4.6}$$

4.3.3 Optimization Problem

The Lagrangian should be maximized with respect to α while minimizing with respect to β, hence, the objective function is $\min_{\beta} \max L(\alpha)$. This problem can be cast as a semi-infinite linear programs of the form Equation 4.7, where θ is an upper bound on the objective function $\min_\beta \max L(\alpha)$. If we denote the objective function $\min_\beta \max L(\alpha)$ as $\Theta(\alpha, \beta)$, and the optimal value of the objective function as α^*, then $\forall \alpha$ and β, $\Theta(\alpha^*, \beta) \geq \Theta(\alpha, \beta)$ [53]. Therefore, minimizing θ is equivalent to maximizing $L(\alpha)$. SILP programs of this form can be used via column generation method. Interested readers can refer to Sonnenburg et al.'s paper [80].

$$\begin{aligned} &\min_{\beta,\theta} && \theta \\ &\text{such that} && \theta \geq \sum_{j=1}^{k} \beta_j \left(\sum_{i=1}^{n} \alpha_i K_j(\mathbf{x}_i, \mathbf{x}_i) \quad - \sum_{i=1}^{n} \sum_{t=1}^{n} \alpha_i \alpha_t K_j(\mathbf{x}_i, \mathbf{x}_t) \right) \\ &&& \forall_{i=1}^{n} 0 \leq \alpha_i \leq C \sum_{i=1}^{n} \alpha_i = 1 \\ &&& \forall_{j=1}^{d} \beta_j \geq 0 \sum_{j=1}^{d} \beta_j = 1 \end{aligned} \tag{4.7}$$

4.4 Empirical Results

In this section, we explain the experimental settings, and present the results on the WISE dataset to predict mortality, stroke, CHF and MI. For multiple kernel learning, we use 9 base kernels:

- Gaussian Kernel: For the σ parameter in Equation 4.3, six different values were used: 0.5, 1, 2, 5, 7, 10
- Polynomial Kernel: Three different degrees were used for the degree parameter in Equation 4.3: 1, 2, 3

We used the Shogun toolkit [81] for MKL implementation. As explained in the previous sections, Shogun toolkit solves semi-infinite linear program formulation of one-class MKL problem using constraint-generation. Experiments are repeated 10 times; mean and standard deviation of F-1 scores are shown in Figures 4.4 and 4.5. We compared the MKL approach to random under sampling (RUS), one-class SVM, and cost-sensitive logistic regression (denoted as weighted in the Figures). For stroke prediction, MKL slightly outperforms the baseline approaches, however for CHF, MI and mortality prediction, it slightly underperforms. On average, MKL's performance is comparable to the most commonly used techniques to deal with high class skew. On the other hand, one-class SVM approach is not well suited to this type of problem, as demonstrated by its poor empirical performance.

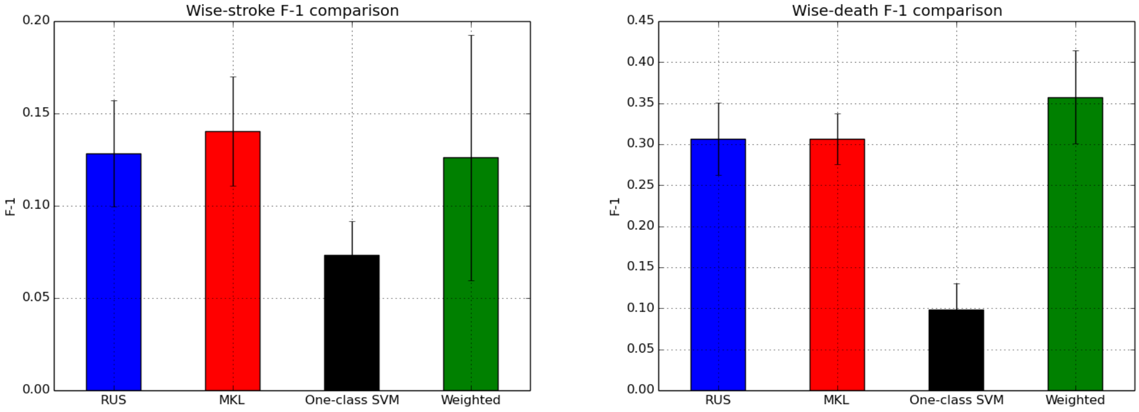


Figure 4.4: F-1 comparison on the WISE dataset (stroke prediction on the left, mortality prediction on the right)

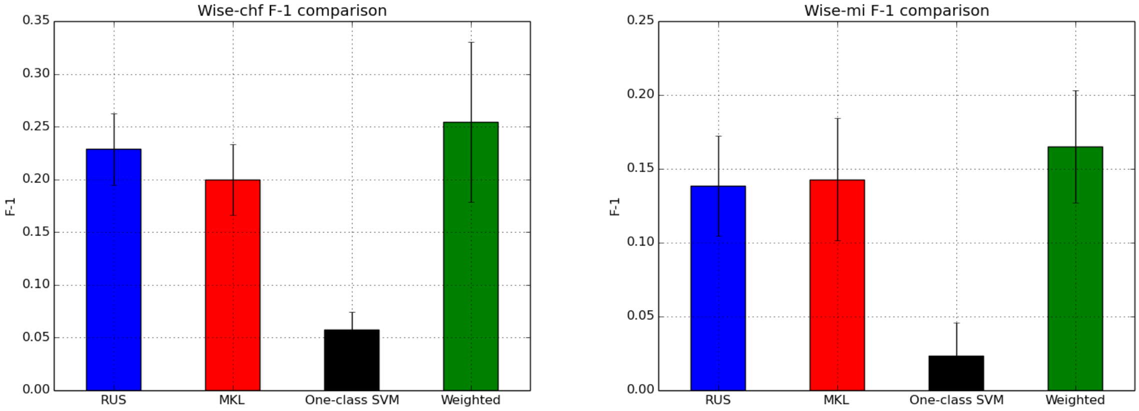


Figure 4.5: F-1 comparison on the WISE dataset (CHF prediction on the left, MI prediction on the right)

4.5 Chapter Conclusions

In this chapter, we shifted our focus to deal with highly skewed class distribution on the WISE dataset and introduced a technique based on multiple kernel learning. Our assumption was that minority class forms compact clusters in the feature space. Under this assumption, we demonstrated that a method that leverages this structure to the learning process, outperforms baseline approaches. This assumption is satisfied through the use of a single kernel, such as Gaussian kernel on selected UCI datasets. However, this assumption was not verified in 4 different prediction tasks of the WISE study. To test whether this is due to usage of the wrong kernel, we experimented with multiple kernel learning to select the right kernel combination.

Under compactness assumption, our experiments revealed that MKL does not outperform baseline approaches, especially cost-sensitive logistic regression, indicating that the minority class of the WISE dataset may not be compact. This may be due to the following: 1. The presentation of heart disease in patients may differ significantly 2. Adverse events (CHF, MI, stroke, or death) are recorded at the end of the trial; however, features that are used in prediction are the measurements taken in the beginning of the trial. Sick patients may get sicker (or healthier) towards the end of the trial. Consequently, a “healthy” patient may end up suffering from an adverse event, or a “sick” patient may not have any complications at all (or have it after the end of the trial). Class overlaps of these sort may have invalidated the compactness assumption.

As evident by the poor performances of one-class SVMs demonstrated in this section, on the WISE dataset, neither majority nor minority class conform to a single “normal” pattern. This suggests the possibility of having small disjuncts in the dataset. In the next chapter, we introduce algorithms that deal with class overlaps and small disjuncts.

Chapter 5

Robust Sampling with Distribution Separation for Highly Skewed Datasets

5.1 Introduction

Poor performances of the learning algorithms on datasets with highly skewed class distribution are not always caused by class imbalance alone, small disjuncts and overlapping classes aggravate the problem [93], [47]. Commonly used methods to deal with class skew, such as random under sampling of the majority class and cost-sensitive learning do not tackle either of these issues; they merely solve the class imbalance problem by rebalancing the dataset or by assigning different misclassification costs to classes. They can be effective strategies on datasets where poor performance is caused only by class imbalance; however they are not optimal strategies on every dataset.

In this chapter, we propose two algorithms that deal with small disjuncts and class overlaps respectively. Our first algorithm, Cluster-Classify, clusters the majority class prior to training, and trains different classifiers on each cluster. The idea of training classifiers on subsets of majority class, rather than throwing away useful data is not a new concept in imbalanced learning. Previously, ensemble based classifier strategies have been proposed to overcome the weaknesses of under sampling. EasyEnsemble [64] randomly subsamples a subset N from the majority class, which has equal number of instances with the minority class. A classifier is trained on this subset and the minority class and added to the ensemble of classifiers. The process is repeated T times and the ensemble classifier is used for prediction [64]. Unlike Cluster-Classify, EasyEnsemble does not address the small disjunct problem, as the subsets are sampled randomly, whereas Cluster-Classify decides on which classifier to be used by leveraging the disjunctive nature of majority class. Similarly, in Roughly Balanced Bagging, the dataset is divided into K equal sized subsets, and a base learner is trained on each subset [41]. Predicted label is obtained by averaging the posterior probability of each base learner [41]. As EasyEnsemble, this algorithm does not address the small disjunct problem.

Our second algorithm, Robust Under Sampling, under-samples from the overlap region, removing redundant majority class data simultaneously. Informed under-sampling methods with a focus on class overlaps have been proposed by previous research on imbalanced learning. For instance, Mani and Zhang [98] proposed an informed under-sampling method with 4 different variations: NearMiss-1, NearMiss-2, NearMiss-3 and “most distant”. NearMiss-1 selects majority class instances whose average distance to k-nearest minority class examples is smallest. NearMiss-2 selects majority class instances whose average distance is smallest to k-farthest minority class examples. This ensures selecting negative examples that are close to all minority class instances [98]. NearMiss-3 selects a fixed number of majority class per minority class instances to ensure every minority class example is surrounded by a majority class member [98]. “most distant” selects examples whose average distance to k-nearest minority class members are farthest. They found NearMiss-2 to be the best performing variation; however, even though NearMiss-2 samples majority class mainly from the overlap region, it doesn’t try to increase the variability in the majority class. With the robust under-sampling algorithm, we also try to maximize intra-class similarity. In the following sections, we describe our algorithms in full detail, and present empirical results.

5.2 Cluster-Classify

In this section, we hypothesize that the majority class consists of small non-overlapping disjuncts. Under this hypothesis, we propose an algorithm to deal with class imbalance. The basic idea is to cluster the majority class prior to classification, and learn a different classifier on each of the clusters. Rather than building an ensemble of classifiers, one classifier is chosen during testing based on the cluster assignment of the test instance. We refer this algorithm as Cluster-Classify throughout the chapter.

Algorithm 1 gives a detailed description of the Cluster-Classify algorithm. First, majority class instances are grouped into different clusters with a given clustering algorithm. The number of clusters is given to the algorithm as a parameter (denoted as K in 1). Each cluster has less majority class instances than the total number of majority class instances in the dataset (given that $K > 1$); hence a less skewed class distribution can be achieved per cluster. For each cluster, a separate classifier is trained on the instances from only that cluster and from the entire minority class.

After training K classifiers (and obtaining K models), one final multi-class classifier is trained only on the majority class instances using cluster assignments as labels. During testing, this classifier is used to determine which cluster a test instance belongs to. If the current instance is a majority class instance, then its “disjunct” can be predicted using the final classifier. If it is a minority class instance, then it can be assigned to any cluster, since the final classifier is trained only on the majority class. In this case, the problem reduces to random under sampling, as the cluster assignment could be random, and the cluster-classifier is trained using a subset of the majority class. After finding the cluster assignment of the test instance, the classifier of that cluster is then used to predict its label.

Algorithm 1 Algorithm: Cluster-Classify

```
K: # of clusters, $\mathcal{L}$: Training set, $\mathcal{T}$: Test set
$\mathbf{X}_{\text{MAJ}}$: Majority class instances in $\mathcal{L}$
$\mathbf{X}_{\text{MIN}}$: Minority class instances in $\mathcal{L}$
Cluster $\mathbf{X}_{\text{MAJ}}$ to K clusters
models: K x 1 empty array
for k = 1: K do
    $\mathbf{X}^{k}_{\text{MAJ}}$: Majority class instances in the k$^{\text{th}}$ cluster
    $\mathcal{L}^{k} = \mathbf{X}^{k}_{\text{MAJ}} \cup \mathbf{X}_{\text{MIN}}$
    models[k]: Classification model trained on $\mathcal{L}^{k}$
    predicted-labels$^{k}$: Test model$^{k}$ on $\mathcal{T}$
end for
model$^{\text{cluster}}$: Train classifier only on $\mathbf{X}_{\text{MAJ}}$ using cluster assignments as labels
predicted-labels$^{\text{cluster}}$: Test model$^{\text{cluster}}$ on $\mathcal{T}$
final-predictions: $N_{\text{test}}$ x 1 empty array, where $N_{\text{test}}$ is the number of test instances in $\mathcal{T}$
for each $\mathbf{x}_i \in \mathcal{T}$ do
    predicted-cluster: predicted-labels[i]$^{\text{cluster}}$
    cluster-vote: predicted-labels$^{\text{predicted-cluster}}$
    final-predictions[i] = cluster-vote
end for
return final-predictions
```

5.3 Robust Under-sampling

In this section, we introduce robust under-sampling which is an informative under-sampling method with the premise to deal with class overlaps as well as class imbalance. This strategy tries to keep as many instances as possible from the overlap region while increasing variation. By keeping only majority class instances that closely resemble the minority class in the training set, we shift the classifier's focus primarily on the overlap region. By maximizing the distance between majority class samples, we remove instances that resemble those that are already in the dataset; hence eliminating redundancy and balancing the class distribution. Algorithm 2 describes this strategy in more detail.

In the beginning of the algorithm, training instances are split into minority and majority class instances. The distance (denoted as $Dist_1$) between randomly selected b majority class samples and the minority class is computed. As the next step, the distance between the selected b majority class samples and the rest of the majority class example is computed ($Dist_2$ in 2). As a distance metric, we picked maximum mean discrepancy, which is explained in the next section. The goal is to minimize the inter class distance by selecting majority class instances with maximal similarity to the minority class (overlap region) and maximize intra class distance by selecting those with minimal similarity to the ones that are already sampled. Therefore, the distance we are trying to minimize is $Dist = Dist_1 - Dist_2$. This process is repeated $maximum - repeats$ times. Both $maximum - repeats$ and b are given as parameters to the algorithm. At the end of the algorithm, instances with the smallest $Dist$ are kept, and the rest is removed from the training set to balance out the class distribution.

Algorithm 2 Algorithm: Robust Under Sampling

```
maximum-repeats, b: number of samples, $\mathcal{L}$: Training set, $\mathcal{T}$: Test set
i = 0
$\mathbf{X}_{\text{MAJ}}$: Majority class instances in $\mathcal{L}$
$\mathbf{X}_{\text{MIN}}$: Minority class instances in $\mathcal{L}$
dists: maximum-repeats x 1 empty array
while i $\leq$ maximum-repeats do
    $\mathbf{X}_S^i$ = randomly choose b from $\mathbf{X}_{\text{MAJ}}$
    $\mathbf{X}_{US}$ = $\mathbf{X}_{\text{MAJ}} \setminus \mathbf{X}_S$
    $D_1$ = Dist($\mathbf{X}_{\text{MIN}}$, $\mathbf{X}_S^i$)
    $D_2$ = Dist($\mathbf{X}_{\text{US}}$, $\mathbf{X}_S^i$)
    $D_f$ = $D_1 - D_2$
    dists[i]= $D_f$
    i = i + 1
end while
min-ind: index with the minimum distance; dists[min-ind] = min(dists)
$\mathcal{L}^{new}$ = $\mathbf{X}_S^{\text{min-ind}} \cup \mathbf{X}_{\text{MIN}}$
model: Train classifier on $\mathcal{L}^{new}$
prediction-labels = Test model on $\mathcal{T}$
return prediction-labels
```

5.3.1 MMD

Maximum Mean Discrepancy is a non-parametric method that addresses the two sample problem: Checking whether the two distributions are equal by comparing samples from the two distributions [8], [33]. The null hypothesis is that the distributions are equal, and this hypothesis is rejected, if the MMD distance between the two distributions are non-zero. The distance is measured by the difference between the empirical means of the samples mapped in a Reproducing Kernel Hilbert Space (RKHS) [8], [33]. The formal definition can be given as follows:
Definition: Let $\mathbf{X}^S$ and $\mathbf{X}^T$ be two samples drawn from probability distributions p and q respectively. n_S and n_T denote the number of instances $\mathbf{X}^S$ and $\mathbf{X}^T$. Let $\phi(\mathbf{x})$ is a feature space map such that $\phi(\mathbf{x})$: $\mathbf{x} \rightarrow \mathcal{H}$, where $\mathcal{H}$ is a RKHS. Then, the empirical estimate of MMD is defined with Equation 5.1:

$$\text{MMD}(\phi, \mathbf{X}^S, \mathbf{X}^T) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(\mathbf{x}_i^S) - \frac{1}{n_T} \sum_{i=1}^{n_T} \phi(\mathbf{x}_i^T) \right\|_{\mathcal{H}}^2 \tag{5.1}$$

Based on the MMD theory, MMD[$\mathcal{H}, p, q$] = 0 if and only if $p = q$, hence the distance is 0 if and only if the two samples come from the same distribution [8]. The advantages of MMD is that it is a non-biased test that can be easily applicable to high-dimensional data. For more information, interested readers can refer to [8] or [33].

5.4 Empirical Results

We compared Cluster-Classify and Robust Under Sampling (denoted as MMD) with SMOTE, cost-sensitive logistic regression (denoted as weighted), regular logistic regression, random under sampling and random oversampling. SMOTE [90] is an oversampling approach that creates synthetic minority class examples based on their distance to each other in the feature space. For SMOTE, we set the number of nearest neighbors to 10. 3 times the number of minority class instances are synthetically generated and added to the training set.

Random under-sampling removes majority class instances in the dataset in order to balance class distribution (or make it less skewed). In our experiments, after random under-sampling the class distribution is balanced. For random over-sampling, the number of duplicates is set to 3 times the number of minority class instances and added to the training set. For cost-sensitive logistic regression, we multiplied the number of majority class instances with the class skew in the training set $\left(\frac{\#minoritysamples}{\#majoritysamples}\right)$. We conducted experiments on the datasets from the UCI repository and on the WISE dataset. The following subsection provides the empirical results.

5.4.1 Experiments on the UCI dataset

We compared the performance of the algorithms on Satimage, Pendigits, Pageblocks and Mammography datasets. Each experiment is repeated 10 times, and mean F1 scores are reported. For each of the repeats, 70% of the dataset is randomly selected for training, and the remaining 30% is used for testing. Logistic regression is used as a final classifier to evaluate the algorithms. For Cluster-Classify implementation, k-means is used for clustering, the number of clusters is set to 10. On every cluster, an SVM classifier with an RBF kernel is trained, and an SVM classifier is used to decide the cluster a test instance belongs to. Python Scikit-learn toolkit [1] is used for implementation.

As can be seen in Figures 5.1 and 5.2, Cluster-Classify performs better than the rest of the methods, including SMOTE, on Satimage, Pendigits, and Mammography datasets. As mentioned in the introduction, for Satimage and Pendigits, one class is picked as minority class and the rest of the classes are collapsed into one big majority class. Consequently, the majority class in Satimage and Pendigits datasets do contain small disjuncts. Therefore, we can state that Cluster-Class is a promising approach to use on skewed datasets with small disjuncts. For the Pageblocks dataset only the largest two classes are collapsed into a big majority class, therefore it is not as disjunctive as Satimage or Pendigits datasets. Hence, even though Cluster-Classify performs better than random under sampling or cost-sensitive logistic regression on Pageblocks, it underperforms compared to a benchmark method such as SMOTE.

The performance of robust under-sampling is comparable to random under-sampling on all 4 UCI datasets. As evident by the good performance of Cluster-Classify, these datasets consist of small disjuncts. In cases where majority class is disjunctive, sampling from the overlap region is almost similar to random under-sampling, as there can be many intra-class or inter-class overlaps. This suggests that robust under-sampling may be better suited to datasets, where degradation of prediction performance under high-class skew, is caused by class overlaps alone, but not by small disjuncts.

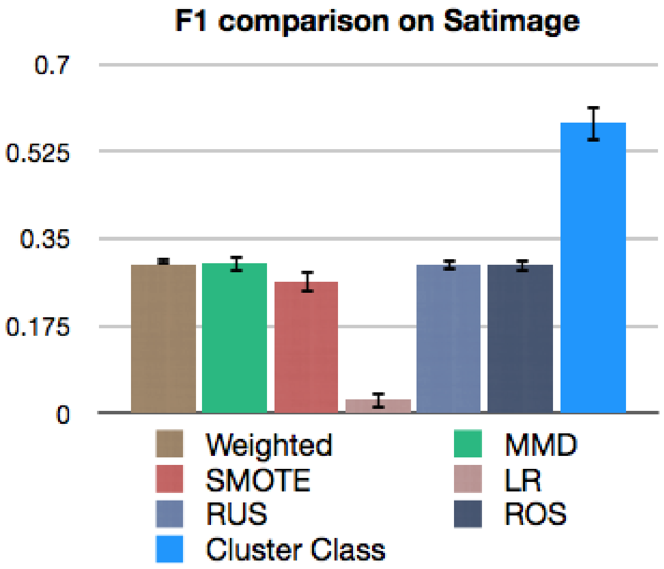


Figure 5.1: F-1 performance comparison of the algorithms on the Satimage dataset

5.4.2 Experiments on the WISE dataset

In this section, we present empirical performance of the algorithms on the WISE dataset. Experimental settings are similar to those from the previous section. On mortality prediction, Cluster-Classify significantly outperforms all the baseline methods measured by AUROC, demonstrating that the majority class does have small disjuncts. For the rest of the prediction tasks, the performance of Cluster-Classify falls behind the rest of the benchmark methods. However, we believe that further experiments with varying number of clusters can reverse this result, and improve Cluster-Classify's performance.

Experiments on the WISE dataset reveal that robust under-sampling performs *worse* than random under-sampling. This can be explained as follows: with robust under sampling, our goal is to sample mainly from the overlap region, and increase intra-class variation in the dataset. However, if overlaps are caused by noise (i.e., samples in the overlap region are noisy examples), this strategy can actually lead to poor prediction performance, since it keeps mainly noisy samples in the dataset.

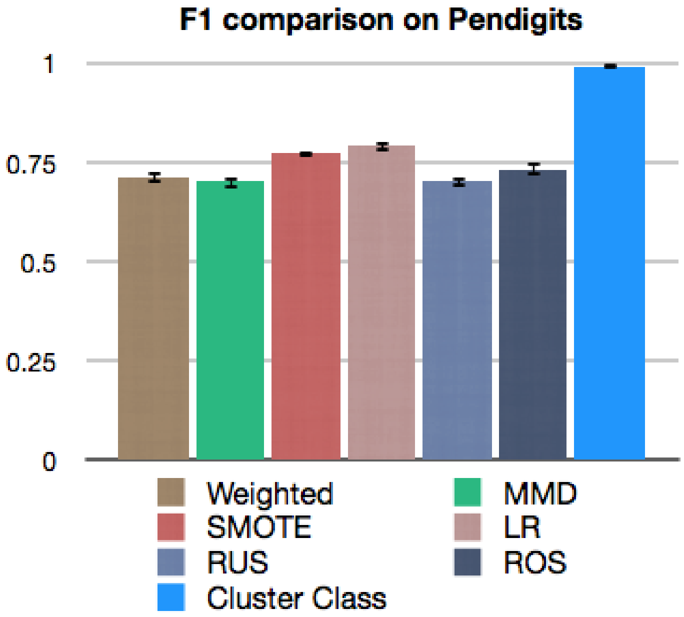


Figure 5.2: F-1 performance comparison of the algorithms on the Pendigits dataset

On other 3 prediction tasks on the WISE dataset, random under sampling outperforms SMOTE, cost-sensitive learning and both of our proposed algorithms. This supports the noisy class hypothesis that we have just speculated. By removing some members of the majority class, random under-sampling could be reducing noise.

5.5 Chapter Conclusions

In this chapter, we proposed two novel algorithms to improve the learning performance on imbalanced datasets in the presence of class overlaps and small disjuncts. Cluster-Classify's performance on the datasets with known small disjuncts is better than all the other algorithms; however, its performance depends on the number of clusters. As our experiments revealed, when the number of clusters is set close to the "actual" number of disjuncts, the performance of Cluster-Classify is impressive compared to rest of the algorithms, but when the number of clusters is much higher than the actual number, its performance gain is relatively smaller. In the future, it would be interesting to explore strategies to set the cluster size, as the number of disjuncts may not known apriori.

Informative under-sampling from the overlap region has been shown effective by previous research in imbalanced classification [98]. In this chapter, we introduced a novel algorithm that increases variation within the majority class, while sampling from the overlap region. However, in this chapter, we have not been able to achieve substantial improvement with robust under-sampling over random under-sampling on the UCI datasets. As evident by the competitive performance of Cluster-Classify, UCI datasets that we used in the experiments potentially con-

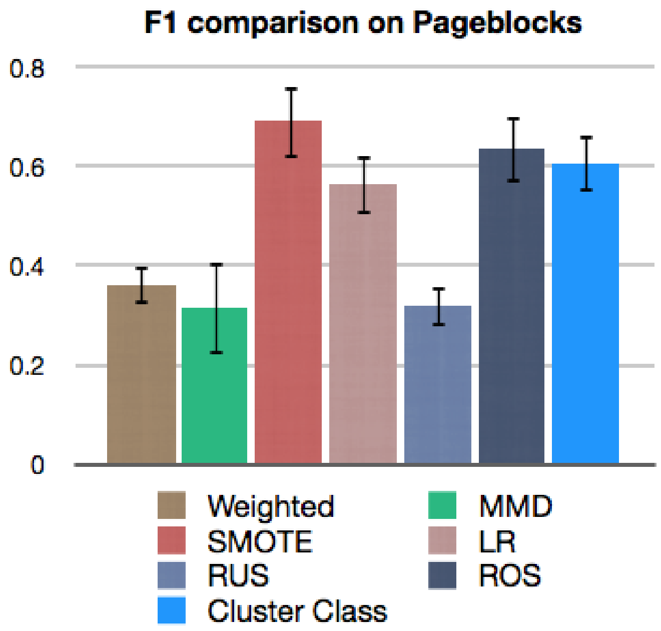


Figure 5.3: F-1 performance comparison of the algorithms on the Pageblocks dataset

sists of small disjuncts. Therefore, focusing on a single overlap region does not provide enough information to learn all the concept descriptions in the dataset. To conclude, informative under-sampling from the overlap region is better suited for datasets where classes are not disjunctive, and degradation in prediction performance is caused by class overlaps and extreme class skew alone.

Some prior research in imbalanced classification assumed that examples lie in the overlap region are noisy [54], [95]. This assumption may hold on the WISE dataset: sampling from the overlap region degrades the prediction performance compared to random under-sampling. In the future, new techniques to understand the reasons behind overlapping classes should be developed. Finally, in the future, one may try to down-weight the majority class instances that are not selected, rather than removing them completely.

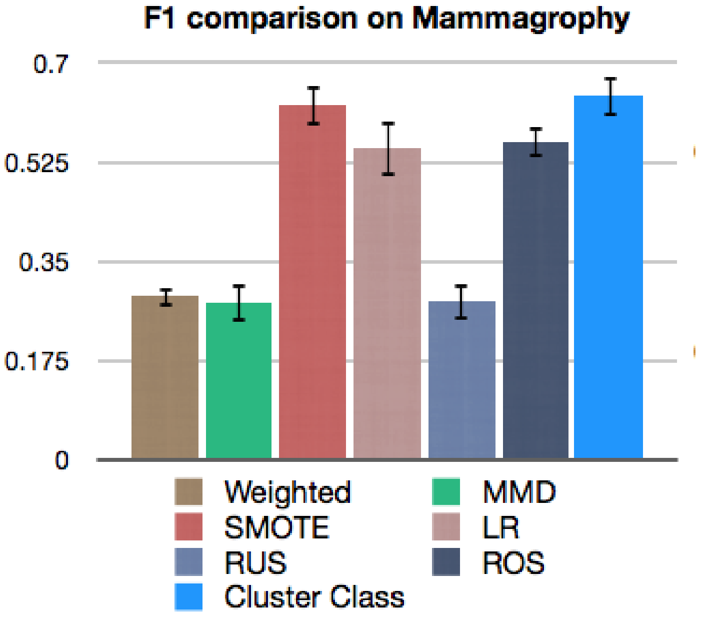


Figure 5.4: F-1 performance comparison of the algorithms on the Mammography dataset

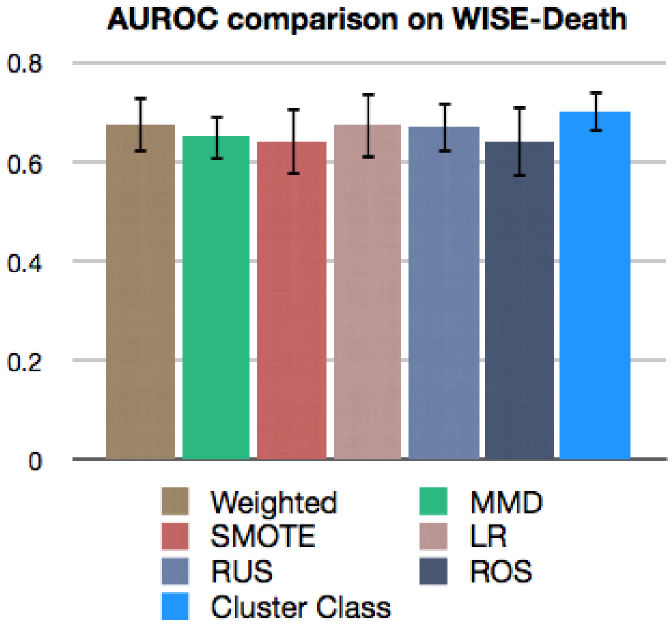


Figure 5.5: AUROC performance comparison of the algorithms on the WISE dataset (mortality prediction)

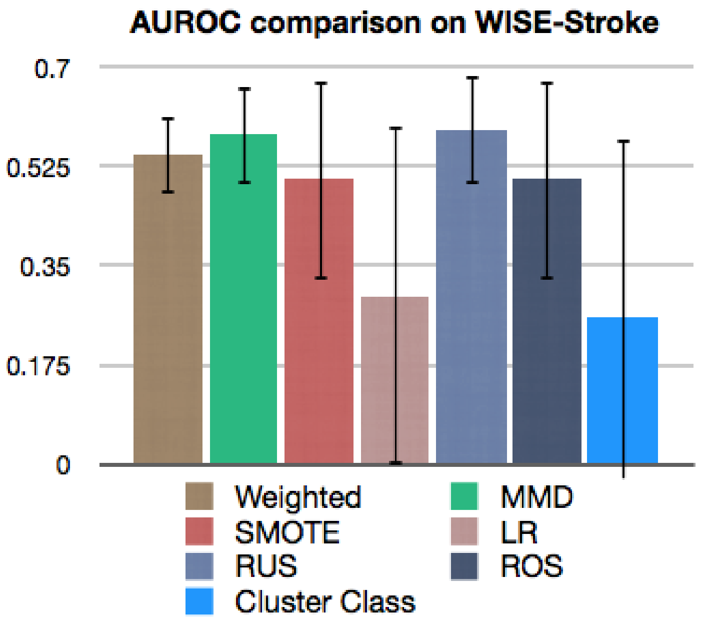


Figure 5.6: AUROC performance comparison of the algorithms on the WISE dataset (stroke prediction)

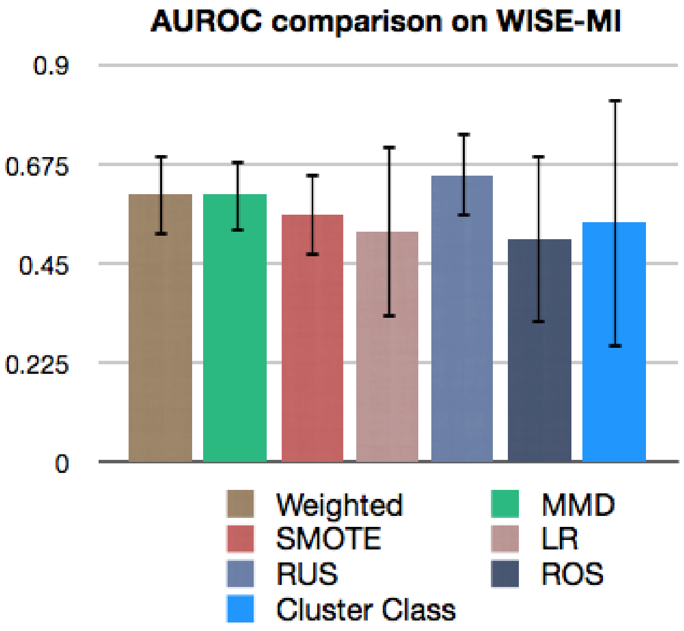


Figure 5.7: AUROC performance comparison of the algorithms on the WISE dataset (MI prediction)

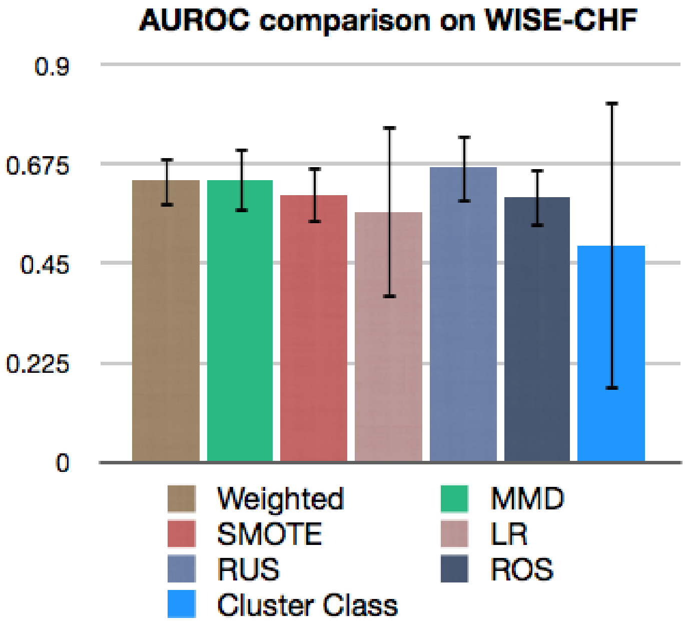


Figure 5.8: AUROC performance comparison of the algorithms on the WISE dataset (CHF prediction)

Chapter 6

Active Learning for Highly Skewed Datasets

6.1 Introduction

Despite having abundant unlabeled data, obtaining sufficient labels for supervised machine learning may be hard in some real-world problems. In topic classification, compared to all available news articles, the number of categorized (labeled) articles is much smaller. Similarly, in automatic speech recognition (ASR), untranslated speech data is abundant, however labeled (translated) corpus is limited.

Active learning can reduce the label complexity by guiding the labeling process, through requesting labels of more informative instances [76], [16]. One basic idea is to train a model on the initial available labeled set, use the trained model to select instances from the unlabeled pool, request labels from the selected instances and re-iterate the process. Even though the standard active learning strategies, such as uncertainty sampling, can greatly reduce the label complexity, they may fail in the presence of class imbalance. In highly skewed settings, they may request labels mainly from the majority class, and the training set may lack representative instances from the minority class.

In this chapter, we introduce three new active learning strategies tailored for highly skewed datasets: 1. Maximum probability sampling 2. Threshold sampling 3. Sampling with maximum mean discrepancy. The goal of the maximum probability sampling is to request labels from instances that the current model identifies as minority with high probability. Threshold sampling requests labels from all instances having posterior probabilities that are above a certain probability threshold, which is determined by the level of skewness of the current training set. Finally, sampling with maximum mean discrepancy selects instances for labeling that are similar to the minority class instances. Similarity is measured by the maximum mean discrepancy metric.

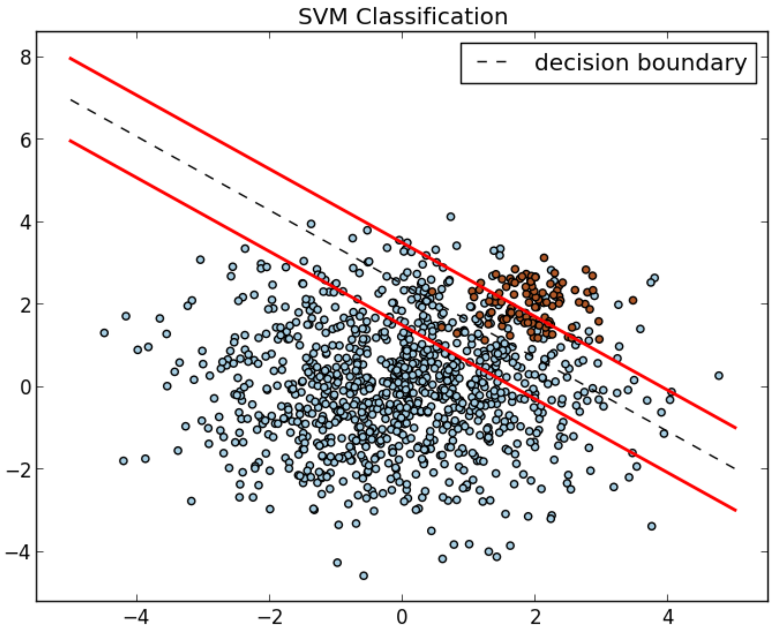
We apply these active learning methods to various tasks, including adverse event prediction on the WISE clinical dataset, and the text categorization on the 20 Newsgroup dataset. We compare the proposed and baseline approaches based on the AUROC performance on the test set. We conclude the chapter by providing guidelines on when to use each strategy, as experiments reveal that the performance of each strategy varies based on the characteristics of the dataset, such as the dimensionality, and the number of labeled minority class instances.

6.2 Uncertainty Sampling

In this section, we review the baseline method that was used in the experiments: uncertainty sampling. The basic idea of uncertainty sampling is to label instances that the classifier is prone to classify incorrectly, i.e., those that lie close to the decision boundary. Under a probabilistic model, this strategy can be formulated as follows:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} |0.5 - P_{\tilde{\theta}}(\hat{y} = 1|\mathbf{x})| \tag{6.1}$$

where $\tilde{\theta}$ is the current model. For support vector machines, uncertainty sampling selects unlabeled examples that lie in between margins (between the red lines in Figure 6.2). Uncertainty sampling is shown to be ineffective when the dataset is highly skewed [4]. It has been conjectured that this is especially true if the majority class has many disjuncts that overlap with the minority class. In such cases, uncertainty sampling may request labels mainly from the majority class, aggravating the class imbalance problem further. In this chapter, we evaluate the performance of uncertainty sampling on a variety of datasets with varying skew levels. We also examine the validity of the hypothesis that uncertainty sampling strategy performs less than ideal on datasets with skewed class distribution.



6.3 Active Learning with Maximum Probability

In this section, we describe our proposed strategy: active learning with maximum probability. The motivation behind this strategy is as follows: by requesting labels from the unlabeled instances that the current model classifies as minority with high probability, we can achieve a more balanced class distribution in the (final) labeled set. This strategy does not rely on the discriminative power of the initial classifier as much as uncertainty sampling: the exact placement of the decision boundary does not matter as long as the initial classifier can learn the majority class pattern even when trained on a small dataset. Hence, avoiding repeated queries to the majority class is possible, if the initial classifier can roughly separate between the minority and majority class. We will explain this intuition in detail in this section.

Under a probabilistic model of binary classification, active learning with maximum probability can be formulated as follows:

$$\mathbf{x}^{*} = \arg\max_{\mathbf{x}} P_{\tilde{\theta}}(\hat{y} = 1|\mathbf{x}) \tag{6.2}$$

where $\tilde{\theta}$ is the current model. Algorithm 3 outlines this strategy in detail.

Algorithm 3 Algorithm: Active Learning with Maximum Probability

```
maximum-iteration, nselect: # instances that will be selected for labeling, $\mathcal{L}$: Labeled data, $\mathcal{U}$: Unlabeled data
i = 0
while i $\leq$ maximum-iteration do
    model: train classifier on $\mathcal{L}$
    prediction-probabilities: prediction probabilities of model on $\mathcal{U}$
    $\mathbf{X}s$: nselect instances in $\mathcal{U}$ with the highest prediction probabilities
    $\mathbf{y}s$: request labels for the instances in $\mathbf{X}s$
    $\mathcal{U}$ = $\mathcal{U} \setminus (\mathbf{X}s, \mathbf{y}s)$
    $\mathcal{L}$ = $\mathcal{L} \cup (\mathbf{X}s, \mathbf{y}s)$
    i = i + 1
end while
```

At each active learning epoch, a probabilistic model is trained on the labeled set, and then used to obtain prediction probabilities of the unlabeled set. Prediction probabilities are then sorted, and the top $nselect$ examples are chosen for labeling. In contrast to uncertainty sampling, here the goal is to achieve a balanced class distribution in the labeled set, rather than trying to improve classification performance at each epoch. During the initial stages of active learning, labeled data may be too scarce; hence the placement of the decision boundary may not be optimal. Or, if the labeled data is imbalanced, the decision boundary may be biased towards the minority class (if class imbalance is not corrected). Active learning with maximum probability selects instances that are likely from the minority class regardless of the placement of the decision boundary. To illustrate, we trained a probabilistic classifier, logistic regression, on a randomly generated 2-dimensional dataset twice: once with correcting the class imbalance by assigning

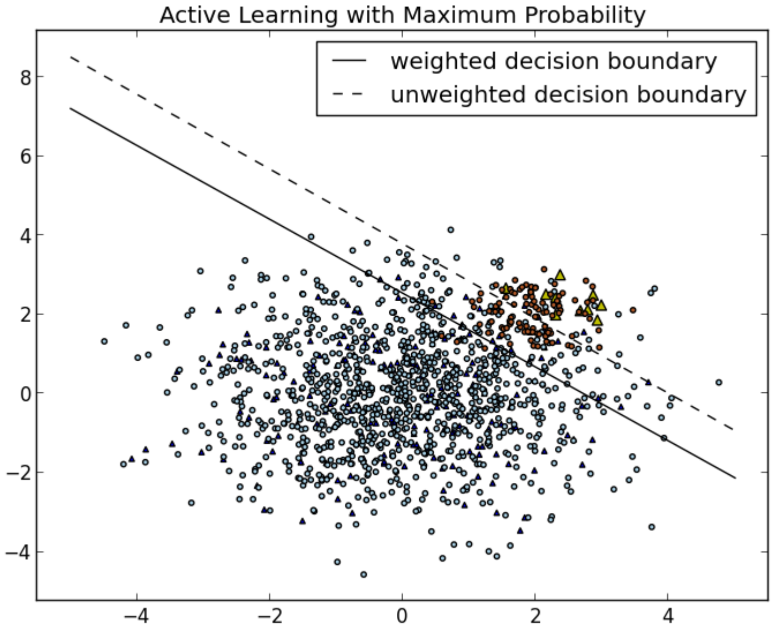


Figure 6.1: Instances that are selected by maximum probability, with respect to the decision boundary learnt by cost-sensitive and regular logistic regression

class specific costs inversely proportional to the class frequency, and once using cost-insensitive logistic regression. We plotted the decision boundaries, along with the labeled, unlabeled, and selected examples in Figure 6.3. The decision boundary of the weighted classifier is shown with the black solid line, and the decision boundary of the unweighted classifier is shown with dashed black line. 2-dimensional training examples are shown in blue (negative/majority class), and orange (positive/minority class) circles. Unlabeled examples are shown in triangles, 5 instances with maximum probability are shown in green, and unselected instances are shown in blue. In both cases, the instances that the maximum probability active learning selects are the same and lies within the minority class region, regardless of the placement of the decision boundary.

6.4 Active Learning with Threshold Selection

In maximum probability based active learning, top n unlabeled instances are selected for labeling with highest posterior probabilities belonging to the minority class. In threshold selection based active learning, rather than fixing the number of instances that will be chosen for labeling, we determine a threshold based on the skewness of the training set. All instances with probabilities that are above that threshold are then chosen for labeling (up to a maximum number determined by a parameter). As we will explain later, this algorithm is specifically designed for datasets where the severity of class imbalance is not known apriori. Algorithm 4 explains this strategy in full detail.

Algorithm 4 Algorithm: Active Learning with Threshold Selection

```
minimum-select, maximum-select, maximum-iteration, $\mathcal{L}$: Labeled data, $\mathcal{U}$: Unlabeled data

i = 0
while i $\leq$ maximum-iteration do
    *model*: train classifier on $\mathcal{L}$
    prediction-probabilities: prediction probabilities of *model* on $\mathcal{U}$
    skew = compute skew with 6.3
    $\mathbf{X}s$: find instances in $\mathcal{U}$ such that their prediction probabilities are greater than $1 -$ skew
    if $|\mathbf{X}s| == 0$ then
        $\mathbf{X}s$: choose min-select instances with the highest prediction probabilities
    end if
    if $|\mathbf{X}s| >$ max-select then
        $\mathbf{X}s$ = choose only max-select instances with the highest prediction probabilities
    end if
    $\mathbf{y}s$ = request labels for the instances in $\mathbf{X}s$
    $\mathcal{U} = \mathcal{U} \setminus (\mathbf{X}s, \mathbf{y}s)$
    $\mathcal{L} = \mathcal{L} \cup (\mathbf{X}s, \mathbf{y}s)$
    i = i + 1
end while
```

At each active learning epoch, logistic regression classifier is trained on the available labeled data and then used to predict probabilities of the instances in the unlabeled dataset. Additionally, skewness of the training set is computed using Equation 6.3. In the beginning, the training set is small but balanced ($skew <= 0.5$), therefore the threshold is set to 0.5. If the skew increases (the ratio of labeled minority class instances to the total number of training instances decreases), then the threshold automatically increases to include examples with higher probability of being in the minority class. This is to prevent requesting more labels from majority class and it allows flexibility to the active learning strategy: if the dataset is balanced, then the threshold remains close to 0.5. If the dataset is extremely skewed ($skew < 0.1$) then the active learner needs to have high confidence to be able to request the label of an instance.

Maximum probability fixes the number of selected instances for labeling at each epoch. Threshold selection is more flexible in terms of the number of instances that are being labeled at each epoch. If there is not any instance with posterior probabilities higher than the threshold, then the instance with the highest probability (of belonging to the minority class) is chosen for labeling. $maximum - select$ parameters determines the maximum number of instances to be labeled at each epoch. If there are too many instances with posterior probabilities higher than threshold, only $maximum - select$ instances will be queried. Otherwise, all instances with posterior probability higher than the threshold with be labeled.

$$\text{skew} = \frac{\text{\# of minority class instances in the dataset}}{\text{Total number of instances in the dataset}} \tag{6.3}$$

6.5 Active Learning with Maximum Mean Discrepancy

Active learning strategies described in the previous sections all rely on a probabilistic classifier. However, especially during the early stages of active learning when the labeled data is scarce, the classifier may not be reliable. Therefore, in this section we describe an unsupervised active learning method. Starting with an initial set of minority class instances, this method repeatedly compares a subset of instances in the unlabeled pool to the minority class, and selects instances from the subset with the lowest maximum mean discrepancy distance [8]. Minority class instances and candidate unlabeled instances are both mapped to a high dimensional Reproducing Kernel Hilbert Space (RKHS) with a kernel. In our experiments, we picked the radial basis kernel. Algorithm 5 explains this strategy in detail.

Algorithm 5 Algorithm: Active Learning with Maximum Mean Discrepancy

```
maximum-iteration, maximum-repeat: partition repeats, ns: number of selected examples,
  $\mathcal{L}$: labeled data, $\mathcal{U}$: unlabeled data
i = 0
while i $\leq$ maximum-iteration do
    $\mathbf{X}_{\text{MIN}}$: Minority class instances in $\mathcal{L}$
    nmin = $|\mathbf{X}_m|$, # of minority instances in the training data
    dists: maximum-repeat x 1 empty array
    $\mathcal{P}$: maximum-repeat x 1 collection of sets
    for $r = 1 \ldots$ maximum-repeat do
        $P^{\text{tmp}}$: select nmin random instances from $\mathcal{U}$
        dists[r] = compute MMD$(P^{\text{tmp}}, \mathbf{X}_{\text{MIN}})$ using Equation ??
        $\mathcal{P}[r]$ = $P^{\text{tmp}}$
    end for
    min-ind: index with the minimum distance; dists[min-ind] = min(dists)
    $\mathbf{X}s$: randomly select ns examples from $\mathcal{P}$[min-ind]
    $\mathbf{y}s$: request labels for the instances in $\mathbf{X}s$
    $\mathcal{U}$ = $\mathcal{U} \setminus (\mathbf{X}s, \mathbf{y}s)$
    $\mathcal{L}$ = $\mathcal{L} \cup (\mathbf{X}s, \mathbf{y}s)$
    i = i + 1
end while
```

6.6 Empirical Results

We evaluated the performance of the algorithms after each iteration with cost-sensitive logistic regression and tested the active learning strategies on 20 Newsgroups, Pageblocks, Pendigits, Mammography and Satimage datasets. We performed 10 random repeats, and the results are averaged over all repeats. At each repeat, %30 percent of the dataset is reserved as the test set, and 20 samples (10 from the minority class, and 10 from the majority class) are randomly chosen as the initial labeled training set. Active learning strategies pick the best samples from the remainder of the dataset (the rest of the dataset is treated as "unlabeled").

At each active learning iteration, 3 samples are selected by the current active learning strategy and added to the training set with their labels. There are a total of 300 active learning iterations (maximum iteration is 300). In every 3 iteration, a cost-sensitive logistic regression classifier is trained on the current labeled training set and evaluated on the test set (therefore there are 100 evaluations). Costs are inversely proportional to the class skew in the labeled set. We reported the performance of the active learning algorithms for each dataset with respect to iterations in the Figures below. Figure 6.2 shows the AUROC results on the 20 Newsgroups dataset. On this dataset, active learning based on maximum probability outperforms the rest of the algorithms especially in the beginning, when the labeled dataset is small. Under insufficient labeled data, the classifier estimates may not be reliable, hence selecting examples based on the decision boundary is not guaranteed to improve the prediction performance. The reliability of the classifier improves as more labeled samples are collected. At this point, it is advisable to switch to uncertainty sampling from maximum probability, since uncertainty sampling selects examples that could improve the performance of the classifier by fine-tuning the decision boundary. The left figure in Figure 6.2 shows the number of minority class samples with respect to the active learning iterations. This figure demonstrates interesting results on the effect of the chosen samples to the prediction performance. As evident from the Figure, threshold sampling queries more minority class samples than the rest of the methods. However, its performance falls behind uncertainty sampling, which does not query nearly as many examples as active learning with maximum probability. This shows that the number of minority class samples in the training set is not as significant, informativeness of each sample is more important for the prediction performance.

UCI datasets are smaller than the 20 Newsgroups dataset, both in terms of dimensionality and total number of instances. Figures 6.3, 6.4, 6.5, and 6.6 shows the AUROC comparison on Mammography, Pageblocks, Pendigits, and Satimage datasets respectively. To be able to see the effect of the class skew and the number of minority class instances in the unlabeled pool, for each of the UCI dataset, we altered the class skew by randomly removing the minority class instances, thus increasing the skew.

As it can be seen from the Figure 6.4, active learning based on uncertainty outperforms the rest of the active learning strategies on the original dataset (skew is not altered). However, when the skew is increased (the left Figure), active learning based on maximum mean discrepancy has a domination over the rest of the methods. When we removed %75 percent of the minority class instances from the training set, the performance of the active learning algorithms that rely on a classifier suffer tremendously. Since active learning based on mmd is an unsupervised method that measures the similarity between unlabeled instances to the labeled minority class instances, it does not suffer from the lack of minority class data. Figure 6.3 also confirms this hypothesis: active learning based on mmd outperforms the rest of the algorithms, under both lower-skew and higher-skew (to a certain extent) settings. Satimage and Pendigits have more minority class instances than Pageblocks and Mammagrophy, and the skew is less severe. The initial AUROC score on the Pendigits dataset is high, suggesting that the classifier can learn the concept descriptions even with the initial small labeled set (Figure 6.5). By adding samples that are closest to the decision boundary, uncertainty sampling can fine-tune the decision boundary, thus it performs better than the rest of the approaches. On the original Satimage dataset, active learning with mmd significantly outperforms the rest of the methods, however, when the skew is increased active learning with threshold selection performs better than active learning with mmd (Figure 6.6). Removing minority class instances does not effect the performance of active learning with threshold selection, but the performance of active learning with maximum mean discrepancy drops. Hence, under moderately skewed settings, relying on a classifier based strategy is a better approach.

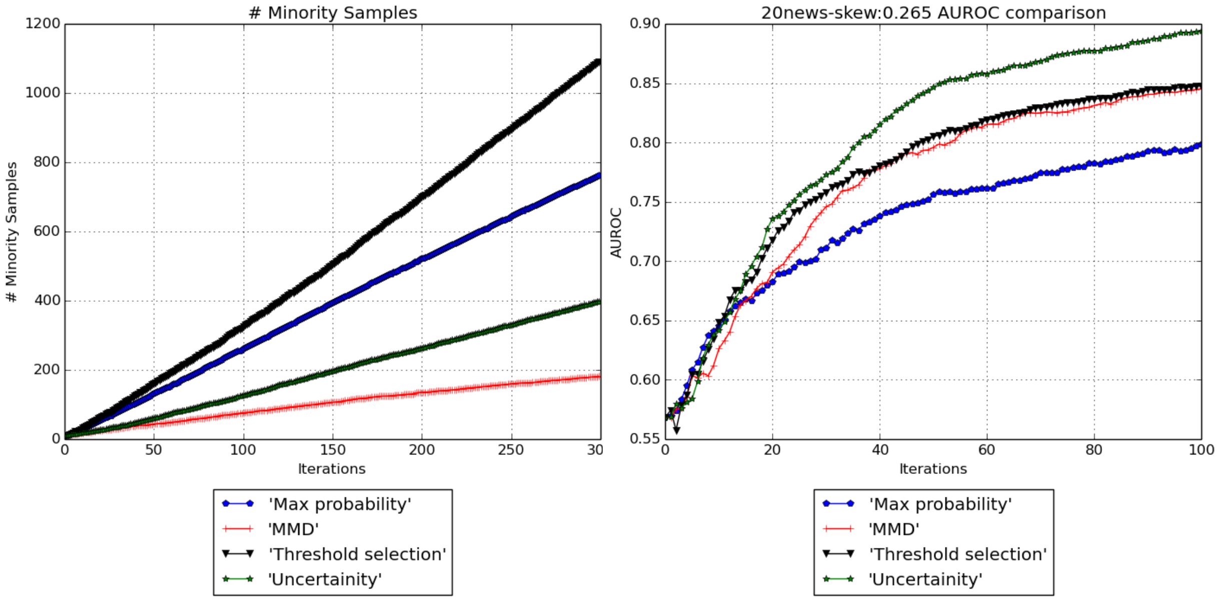


Figure 6.2: The left Figure shows the number of minority class samples with respect to active learning iterations. The right Figure compares the performance of the algorithms as measured by AUROC score on the test set.

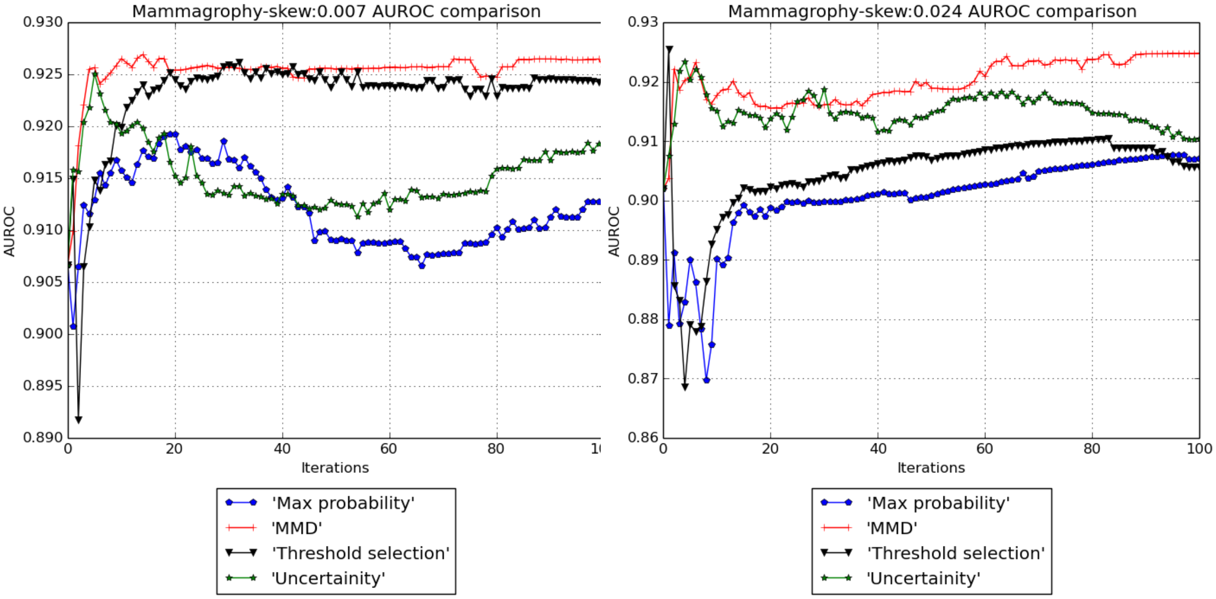


Figure 6.3: AUROC comparison on the Mammography dataset. The left Figure shows the AUROC results when the skew is increased, the right Figure shows the AUROC results on the unmodified dataset

6.7 Chapter Conclusions

In this chapter, we proposed three novel active learning strategies: active learning with maximum probability, threshold selection and maximum mean discrepancy. We evaluated these strategies on a variety of tasks, including text classification and medical diagnosis. Our proposed algorithms outperformed one of the most commonly used baseline active learning algorithm, uncertainty sampling, in most of the datasets. However, we found out that, which algorithm performs the best depends on the characteristics of the dataset, such as its dimensionality, and the number of available minority class instances.

To provide a guideline of which active learning algorithm to choose, we propose the flow chart in Figure 6.7. If the dataset is high dimensional, and if there are insufficient labeled samples in the training set, then it is better to use maximum probability especially in the beginning stages of active learning. After enough labeled data is collected, probability estimates of the classifier becomes more reliable. At this point, uncertainty sampling is better in sampling instances that could better improve the prediction performance.

In cases where the dataset is not high dimensional, the optimal algorithm depends on the skew level, and the size of labeled pool. If extreme skew is anticipated, and a few minority examples are expected in the unlabeled set, then using an unsupervised similarity based algorithm, such as active learning with maximum mean discrepancy would be a better choice. Under moderate skew levels, a classifier based approach such as threshold selection, or uncertainty sampling would be more effective. If there are sufficient labeled minority class instances to learn a good classi-

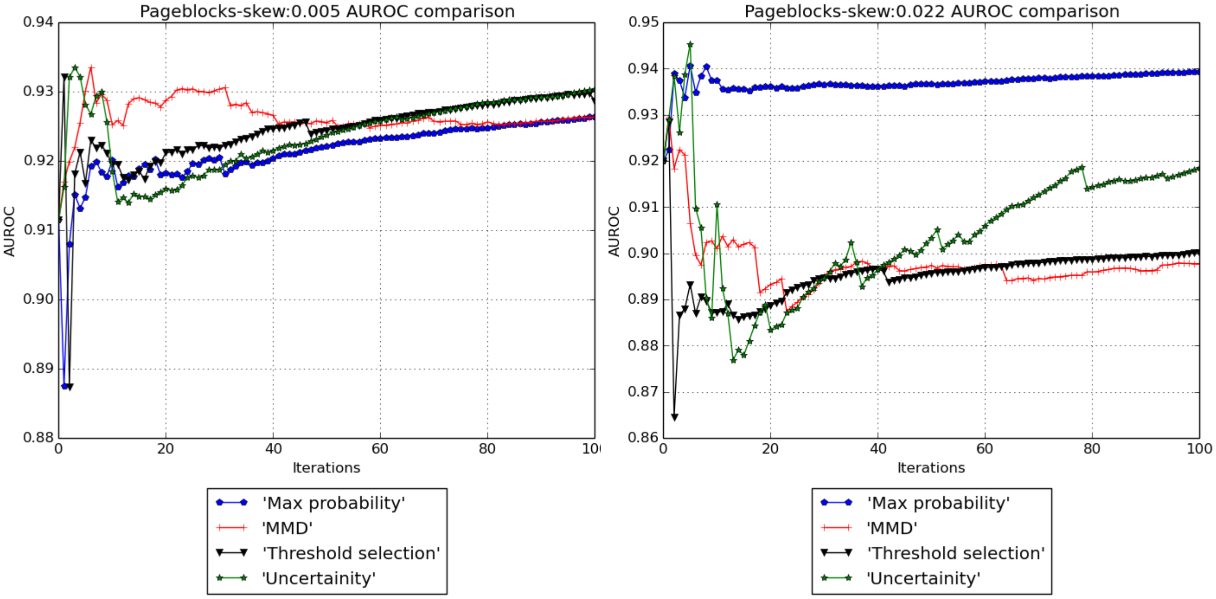


Figure 6.4: AUROC comparison on the Pageblocks dataset. The left Figure shows the AUROC results when the skew is increased, the right Figure shows the AUROC results on the unmodified dataset

fier, than uncertainty sampling outperforms other strategies, since it can fine-tune the decision boundary.

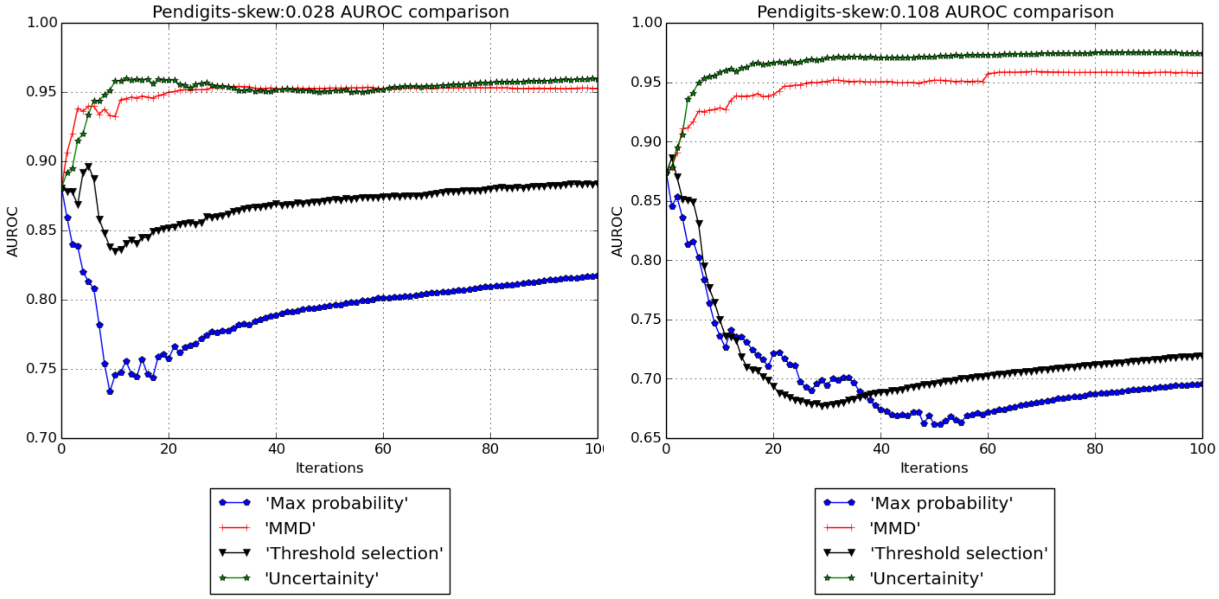


Figure 6.5: AUROC comparison on the Pendigits dataset. The left Figure shows the AUROC results when the skew is increased, the right Figure shows the AUROC results on the unmodified dataset

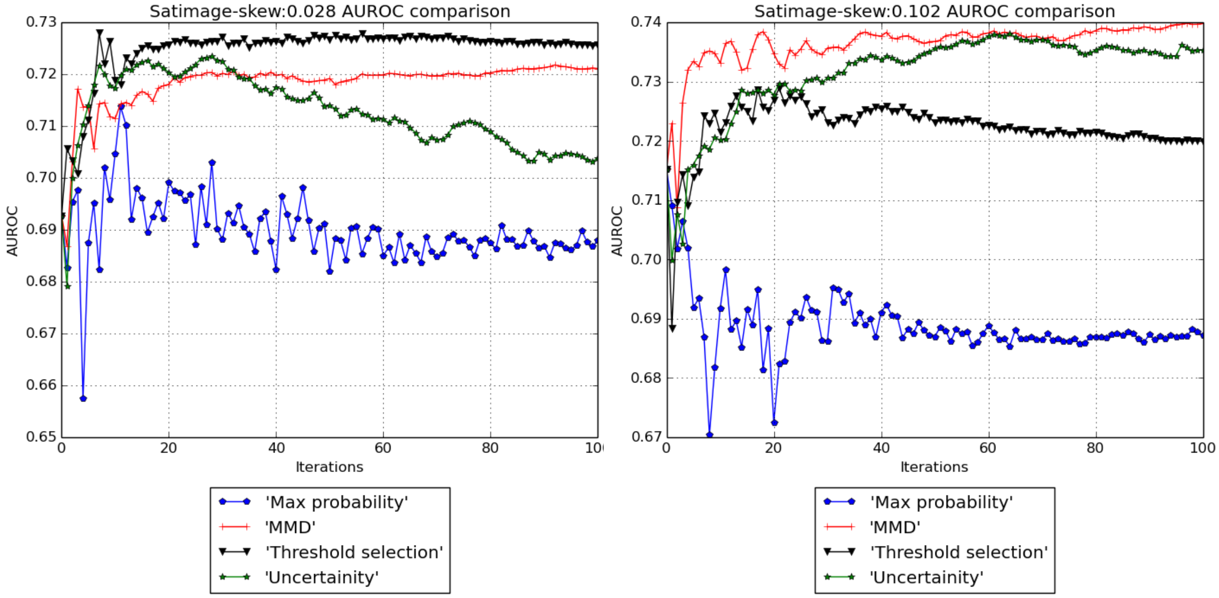


Figure 6.6: AUROC comparison on the Satimage dataset. The left Figure shows the AUROC results when the skew is increased, the right Figure shows the AUROC results on the unmodified dataset

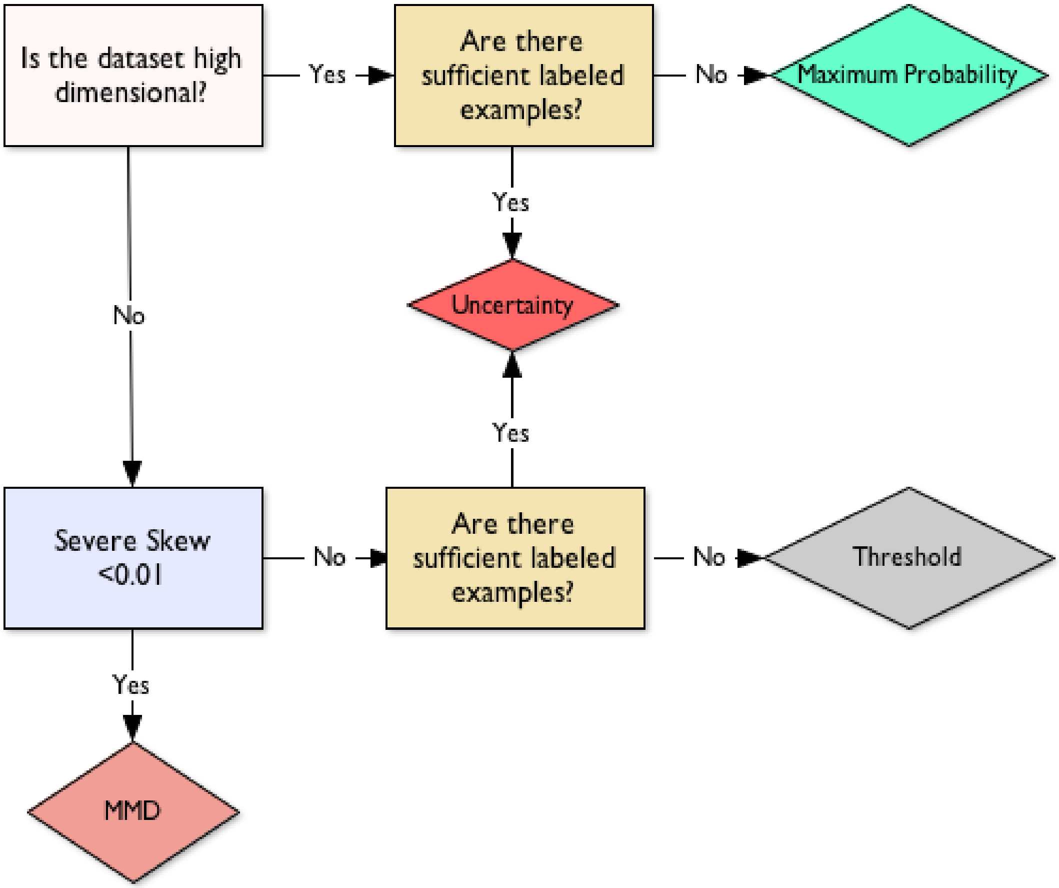

Is the dataset high dimensional?
Yes
Are there sufficient labeled examples?
No
Maximum Probability
Yes
Uncertainty
No
Yes
Severe Skew <0.01
No
Are there sufficient labeled examples?
No
Threshold
Yes
MMD

Figure 6.7: Algorithm: Guidelines to pick the right active learning strategy

Chapter 7

Conclusions and Future Directions

In this dissertation, the goal is to achieve high prediction performance on the minority (rare) classes under extreme class skew. For this purpose, we posit a new hypothesis, i.e., compact minority class(es), and study the high-class skew problem under three assumptions: disjunctive classes, overlapping classes and compact minority class. Disjunctive and overlapping classes assumptions have been investigated by prior research to some extent; however, compact minority class assumption has not been studied under skewed class learning.

This dissertation demonstrates that under the thesis hypothesis, i.e., compact minority class, new algorithms that leverage the cluster structure of the minority class outperform benchmark approaches, quantified by 10% improvement in F-1 score. Compactness hypothesis is satisfied on certain datasets that have been frequently used in imbalanced classification research; hence, the new methods proposed in this thesis outperform baseline approaches, such as cost-sensitive learning. However, as shown by our experiments with multiple kernel learning, on the WISE dataset, compactness assumption is not verified. Moreover, our findings from Chapter 4 reveal that on the WISE dataset, the compactness of the *majority* class is not verified either: methods that operate under the assumption of compact majority class perform poorly compared to the baseline methods.

These findings have led to further investigation on the effect of disjunctive classes, and class overlaps to the degradation of learning performance. Under the disjunctive class hypothesis, Cluster-Classify significantly outperforms state of the art baseline methods on most of the UCI datasets used in the experiments, as well as on the mortality prediction task of the WISE study. We have also shown empirically that on datasets with disjunctive classes, under-sampling from a single overlap region can lead to poor prediction performance and may not improve over random under-sampling.

Prior research in sampling has different views on samples that lie in the overlap region: some research viewed them as noise [95], [54], whereas others put more emphasis on them [98]. This thesis reveals that either view can be preferred depending on the characteristics of the dataset. On the WISE datasets, overlapping instances are likely to be noise, as shown by performance degradation by sampling from the overlap region. Similar degradation is not observed on the

UCI datasets, which tend to be less noisy than a clinical dataset for medical diagnosis.

In this dissertation, we also propose new active learning strategies to utilize under extreme class skew, when there are insufficient labeled minority class instances to learn accurate concept descriptions. Our strategies outperform a benchmark approach, uncertainty sampling, both on high-dimensional datasets and on smaller datasets, especially when the size of labeled set is small. We also propose an *unsupervised* active learning approach based on maximum mean discrepancy (mmd), and demonstrate that, if extreme class skew and a few minority class examples are anticipated in the unlabeled set, our new approach based on mmd is a better choice as a querying strategy compared to classifier-based active learning methods.

Previous research in active learning showed that under extreme class skew, querying based uncertainty is not an optimal choice [4], [3]. In this dissertation, we demonstrate that if the size of the labeled pool is big enough to learn the decision boundary reliably, active learning based on uncertainty performs well as it fine-tunes the decision boundary by selecting examples that are hard to learn. However, in the early iterations of active learning, when the decision boundary is not optimally placed, one should aim to balance the labeled set by sampling from regions where the probability of being in the minority class is maximal. On the Satimage and Mammography datasets, proposed methods such as active learning with threshold selection and maximum mean discrepancy outperform uncertainty sampling, especially under higher-skewed settings. On the 20 Newsgroups dataset, active learning with maximum probability performs the best, especially during the early stages of active learning, when the labeled set size is small.

To conclude, Tables 7.1 and 7.2 summarize the algorithms and their respective assumptions proposed in this dissertation and Figure 7.1 provides a guideline to pick the right strategy based on data characteristics. On the Satimage dataset, Cluster-Classify and POS Compact outperform benchmark methods over 28% and 12% in F-1 score respectively. Similarly on the Pendigits dataset, Cluster-Classify achieves a 27% improvement in F-1 score, compared to the baseline methods. On the Mammography dataset, Cluster-Classify improves upon the benchmark methods in F-1 performance by 1%. SMOTE is the best performing algorithm on the Pageblocks dataset, followed by Cluster-Classify, with F-1 scores of 64.2% and 62.5% respectively.

On the WISE dataset, Cluster-Classify achieves the highest AUROC score, 70.3%, in mortality prediction; it is followed by cost-sensitive logistic regression, which achieves 67.5% AUROC. However, random under-sampling performs better than the benchmark and the proposed methods in CHF, MI, and stroke prediction.

7.1 Future Work

Perhaps the main problem in learning under highly-skewed category distributions is the difficulty in understanding the intrinsic characteristics of a dataset. Knowing the causes of performance degradation in the accurate identification of minority classes in a particular task, would aid in using the right learning tools and algorithms. Previous research has shown that imbalanced learning

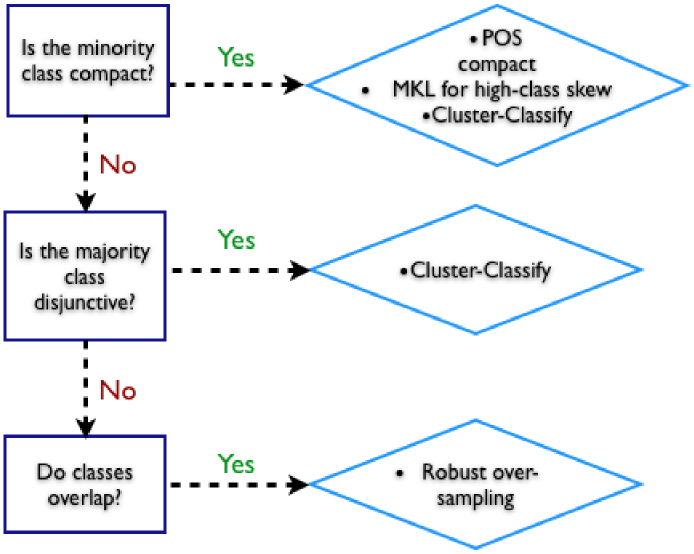


Figure 7.1: Guidelines to use the most suitable algorithm presented in this thesis

Algorithm	**Assumption**
POS compact	Compact minority class
MKL for high-class skew	Compact minority class
Cluster-Classify	Disjunctive majority class
Robust Under-sampling	Overlapping classes

Table 7.1: Assumptions of the learning algorithms introduced in this thesis

is challenging due to many factors, i.e., small disjuncts, class overlaps; however, a general framework that is robust against most of these major factors has not yet exist. In this thesis, we propose several algorithms that outperform benchmark methods, if their assumptions are satisfied. In the future, these algorithms or ideas can be combined to form a unified framework for imbalanced learning.

Alternatively, tests can be developed to determine if a dataset has small disjuncts or class overlaps. If it does not possess any of these aggravating factors, then simple methods such as random under-sampling should suffice to aid prediction performance under class imbalance. It is also important to determine the reasons behind class overlaps: if overlaps are caused by noisy examples, then sampling from the overlap region may actually degrade the prediction performance. Otherwise, one can either put more emphasis on the examples that lie in the overlap region, by perhaps increasing their weights in the learning algorithm, or perform a sampling algorithm that focuses on these *hard-to-learn* examples. How much sampling should be done on the overlap region should also be addressed in the future work. On the other hand, one can develop tests to determine how many disjuncts a dataset has, if any, and use this information to feed into an algorithm that is robust against disjunctive datasets, such as Cluster-Classify.

Algorithm	**When to use**
AL with maximum probability	The size of the training set is not big enough to learn accurate concept descriptions
AL with threshold selection	Skewness of the unlabeled set is not known apriori, need to adjust the threshold dynamically
AL with MMD	The number of minority class samples are expected to be very few, the dataset is expected to have extreme class skew

Table 7.2: Suggested settings to use the active learning algorithms presented in the thesis

In this thesis, we also show that if the dataset has compact minority class(es), then leveraging this property in the learning algorithm can lead to higher prediction performances. Clustering, or employing similarity-based methods to test whether the minority class is compact, can assist the researcher to use, or not the use, this property when tackling class imbalance. Under limited data, nearest neighbor based approaches can be used to detect local density around each example [38]. If minority class examples are found to be in dense clusters, one can conclude that the dataset has compactness property. Another approach can be to test whether minority class instances come from the same distribution, or share certain parameters.

In this dissertation, we show that for a given task, the best active learning strategy depends on the skew level, and the size of the labeled set under class imbalance. For certain datasets, the ideal active learning strategy varies based on the operating curve. Similar to the *passive* learning strategies we discussed above, knowing which querying strategies to apply before performing active learning is crucial. Alternatively, one can develop ensemble based active learning strategies that automatically switch depending on the operating curve, skew level, or the size of the labeled set.

Finally, in this dissertation, we are mainly concerned with binary classification; however the ideas, and the assumptions posed in this thesis can easily be translated to multi-class imbalanced learning.

Bibliography

[1] Scikit-learn. URL `http://scikit-learn.org/stable/`. 3.3.1, 4.2, 5.4.1

[2] Naoki Abe. An iterative method for multi-class cost-sensitive learning. In *In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3–11, 2004. 2.1

[3] Josh Attenberg and Foster Provost. 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 conference on Knowledge discovery and data mining*, KDD '10, pages 423–432, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0055-1. doi: http://doi.acm.org/10.1145/1835804.1835859. URL `http://doi.acm.org/10.1145/1835804.1835859`. 7

[4] Josh Attenberg and Foster Provost. Inactive learning?: difficulties employing active learning in practice. *SIGKDD Explor. Newsl.*, 12:36–41, March 2011. ISSN 1931-0145. doi: http://doi.acm.org/10.1145/1964897.1964906. URL `http://doi.acm.org/10.1145/1964897.1964906`. 1.4.2, 2.2, 6.2, 7

[5] C. Noel Bairey Merz, Sheryl F. Kelsey, Carl J. Pepine, Nathaniel Reichek, Steven E. Reis, William J. Rogers, Barry L. Sharaf, George Sopko, and for the WISE Study Group. The women's ischemia syndrome evaluation (wise) study: protocol design, methodology and feasibility report. *J Am Coll Cardiol*, 33(6):1453–1461, 1999. URL `http://content.onlinejacc.org/cgi/content/abstract/33/6/1453`. 1.4.1

[6] C. Blake, C. Merz. Uci repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/mlearn/MLRepository.html`. 1.4.3

[7] Michael Bloodgood and K. Vijay Shanker. Taking into account the differences between actively and passively acquired data: the case of active learning with support vector machines for imbalanced datasets. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 137–140, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL `http://portal.acm.org/citation.cfm?id=1620853.1620892`. 2.2

[8] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schlkopf, and Alex J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *IN ISMB*, page 2006, 2006. 5.3.1, 5.3.1, 6.5

[9] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression*

Trees. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984. 2.1

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882. URL `http://doi.acm.org/10.1145/1541880.1541882`. 2.3

[11] Nitesh V. Chawla, Ar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. Smoteboost: improving prediction of the minority class in boosting. In *In Proceedings of the Principles of Knowledge Discovery in Databases, PKDD-2003*, pages 107–119, 2003. 2.1

[12] Nitesh V. Chawla, David A. Cieslak, Lawrence O. Hall, and Ajay Joshi. Automatically countering imbalance and its empirical relationship to cost. *Data Min. Knowl. Discov.*, 17 (2):225–252, October 2008. ISSN 1384-5810. doi: 10.1007/s10618-008-0087-0. URL `http://dx.doi.org/10.1007/s10618-008-0087-0`. 1.5

[13] Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. *Discovery*, pages 1–12, 2004. URL `http://www.citeulike.org/user/rabio/article/1121487`. 2.1

[14] Andrew Moore Dan Pelleg. Active learning for anomaly and rare-category detection. In *Advances in Neural Information Processing Systems 18*, December 2004. 2.4

[15] Barnan Das, Narayanan Krishnan, and Diane Cook. Handling imbalanced and overlapping classes in smart environments prompting dataset. *Springer book on Data Mining for Service (to appear)*, 2012. 1.1

[16] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *In Neural Information Processing Systems*, 2005. 2.2, 6.1

[17] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 208–215, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390183. URL `http://doi.acm.org/10.1145/1390156.1390183`. 2.4

[18] Misha Deni and Thomas Trappenberg. Overlap versus imbalance. In *Advances in Artificial Intelligence*, volume 6085 of *Lecture Notes in Computer Science*, pages 220–231. Springer Berlin / Heidelberg, 2010. 1.1

[19] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 155–164. ACM Press, 1999. 2.1

[20] N. Drinkwater. Mstat, 2010. URL `http://www.mcardle.wisc.edu/mstat/download/download.html`. 3.3.2, 3.3.2

[21] Chris Drummond and Robert C. Holte. Exploiting the cost (in)sensitivity of decision tree splitting criteria. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pages 239–246. Morgan Kaufmann, 2000. 2.1

[22] Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets*

II, pages 1–8, 2003. 2.1

[23] Charles Elkan. The Foundations of Cost-Sensitive Learning. In *IJCAI*, pages 973–978, 2001. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.514`. 2.1

[24] Seyda Ertekin, Jian Huang, and C. Lee Giles. Active learning for class imbalance problem. In *In Proceedings of the 30th annual international ACM SIGIR con*, 2007. 2.2

[25] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Applications of Data Mining in Computer Security*. Kluwer, 2002. 2.3

[26] Wei Fan and Salvatore J. Stolfo. Adacost: misclassification cost-sensitive boosting. In *In Proc. 16th International Conf. on Machine Learning*, pages 97–105. Morgan Kaufmann, 1999. 2.1

[27] William Fithian and Trevor Hastie. Local case-control sampling: Efficient subsampling in imbalanced data sets. 2013. 2.1

[28] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. In *Machine Learning*, pages 133–168, 1997. 2.2

[29] Giorgio Fumera and Fabio Roli. Cost-sensitive learning in support vector machines. In *In VIII Convegno Associazione Italiana per LIntelligenza Artificiale*, 2002. 2.1

[30] V. García, R. A. Mollineda, J. S. Sánchez, R. Alejo, and J. M. Sotoca. When overlapping unexpectedly alters the class imbalance effects. In *Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part II*, IbPRIA '07, pages 499–506, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-72848-1. doi: http://dx.doi.org/10.1007/978-3-540-72849-8_63. URL `http://dx.doi.org/10.1007/978-3-540-72849-8_63`. 1.1

[31] Vicente García, José Salvador Sánchez, and Ramón A. Mollineda. Exploring the performance of resampling strategies for the class imbalance problem. In *Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part I*, IEA/AIE'10, pages 541–549, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-13021-6, 978-3-642-13021-2. URL `http://dl.acm.org/citation.cfm?id=1945758.1945822`. 2.1

[32] Mehmet Gonen and Ethem Alpaydin. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, July 2011. ISSN 1532-4435. 4.3.2

[33] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schlkopf, and Alexander Smola. A kernel method for the two sample problem. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 19*, pages 513–520. MIT Press, 2007. 5.3.1, 5.3.1

[34] Hongyu Guo. Learning from imbalanced data sets with boosting and data generation: The databoost-im approach. *SIGKDD Explorations*, 6:2004, 2004. 2.1

[35] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *Natural Computation, 2008. ICNC '08. Fourth International Conference on*, volume 4, pages 192 –201, oct. 2008. doi: 10.1109/ICNC.2008.871. 2.1

[36] Haibo He, Yang Bai, E.A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322 –1328, june 2008. doi: 10.1109/IJCNN.2008.4633969. 2.1

[37] Jingrui He and Jaime G. Carbonell. Rare class discovery based on active learning. In *ISAIM*, 2008. URL `http://dblp.uni-trier.de/db/conf/isaim/isaim2008.html#HeC08`. 2.4

[38] Jingrui He and Jaime G. Carbonell. Prior-Free Rare Category Detection. In *SDM*, pages 155–163, 2009. 2.4, 7.1

[39] Jingrui He, Hanghang Tong, and Jaime Carbonell. Rare category characterization. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 226–235, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4256-0. doi: 10.1109/ICDM.2010.154. URL `http://dx.doi.org/10.1109/ICDM.2010.154`. 1.1, 2.4, 4.1

[40] Xiaofeng He, Lei Duan, and Yiping Zhou andByron Dom. Threshold selection for web-page classification with highly skewed class distribution. In *18th International World Wide Web Conference (WWW2009)*, April 2009. URL `http://data.semanticweb.org/conference/www/2009/paper/121`. 1.1

[41] Shohei Hido, Hisashi Kashima, and Yutaka Takahashi. Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining*, 2(5-6):412–426, 2009. ISSN 1932-1872. doi: 10.1002/sam.10061. URL `http://dx.doi.org/10.1002/sam.10061`. 5.1

[42] Robert C. Holte, Liane E. Acker, and Bruce W. Porter. Concept learning and the problem of small disjuncts, 1995. 1.1

[43] Timothy M. Hospedales, Shaogang Gong, and Tao Xiang. Finding rare classes: Adapting generative and discriminative models in active learning. In *PAKDD (2)'11*, pages 296–308, 2011. 2.4

[44] Jason Van Hulse, Taghi M. Khoshgoftaar, Amri Napolitano, and Randall Wald. Feature selection with high-dimensional imbalanced data. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 507–514, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3902-7. doi: 10.1109/ICDMW.2009.35. URL `http://dx.doi.org/10.1109/ICDMW.2009.35`. 2.1

[45] Vikramaditya R. Jakkula and Diane J. Cook. Detecting anomalous sensor events in smart home data for enhancing the living experience. In *Artificial Intelligence and Smarter Living*, volume WS-11-07 of *AAAI Workshops*. AAAI, 2011. URL `http://dblp.uni-trier.de/db/conf/aaai/aisl2011.html#JakkulaC11`. 2.3

[46] Nathalie Japkowicz and Shaju Stephen. The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 6(5):429–449, October 2002. URL `http://portal.acm.org/citation.cfm?id=1293951.1293954`. 2.1

[47] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *SIGKDD*

Explor. Newsl., 6:40–49, June 2004. ISSN 1931-0145. doi: http://doi.acm.org/10.1145/1007730.1007737. URL http://doi.acm.org/10.1145/1007730.1007737. 1.1, 2.1, 5.1

[48] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-612-2. URL http://dl.acm.org/citation.cfm?id=645528.657646. 4.2

[49] Mahesh V. Joshi, Ibm T. J. Watson, and Ramesh C. Agarwal. Mining needles in a haystack: Classifying rare classes via two-phase rule induction, 2001. 1.1

[50] E. L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):pp. 457–481, 1958. ISSN 01621459. URL http://www.jstor.org/stable/2281868. 3.3.2

[51] Vijay Karamcheti, Davi Geiger, Zvi Kedem, and S. Muthukrishnan. Detecting malicious network traffic using inverse distributions of packet contents. In *Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, MineNet '05, pages 165–170, New York, NY, USA, 2005. ACM. ISBN 1-59593-026-4. doi: 10.1145/1080173.1080176. URL http://doi.acm.org/10.1145/1080173.1080176. 2.3

[52] William Klement, Szymon Wilk, Wojtek Michaowski, and Stan Matwin. Dealing with severely imbalanced data. 2009. 2.1

[53] Marius Kloft, Ulf Brefeld, Patrick Düessel, Christian Gehl, and Pavel Laskov. Automatic feature selection for anomaly detection. In *Proceedings of the 1st ACM workshop on Workshop on AISec*, AISec '08, pages 71–76, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-291-7. doi: 10.1145/1456377.1456395. URL http://doi.acm.org/10.1145/1456377.1456395. 2.3, 4.3.2, 4.3.3

[54] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997. 2.1, 5.5, 7

[55] Miroslav Kubat, Robert Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. In *Machine Learning*, pages 195–215, 1998. 1.1

[56] Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining anomalies using traffic feature distributions. In *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '05, pages 217–228, New York, NY, USA, 2005. ACM. ISBN 1-59593-009-4. doi: 10.1145/1080091.1080118. URL http://doi.acm.org/10.1145/1080091.1080118. 2.3

[57] Wenke Lee and Dong Xiang. Information-theoretic measures for anomaly detection. In *In Proceedings of the 2001 IEEE Symposium on Security and Privacy*, pages 130–143, 2001. 2.3

[58] Kingsly Leung and Christopher Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian confer-*

ence on Computer Science - Volume 38, ACSC '05, pages 333–342, Darlinghurst, Australia, Australia, 2005. Australian Computer Society, Inc. ISBN 1-920-68220-1. URL `http://dl.acm.org/citation.cfm?id=1082161.1082198`. 2.3

[59] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1994. 2.2

[60] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. URL `http://dl.acm.org/citation.cfm?id=188490.188495`. 2.2

[61] Shoushan Li, Shengfeng Ju, Guodong Zhou, and Xiaojun Li. Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 139–148, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390948.2390966`. 2.2

[62] Mei ling Shyu, Shu ching Chen, Kanoksri Sarinnapakorn, and Liwu Chang. A novel anomaly detection scheme based on principal component classifier. In *in Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03*, pages 172–179, 2003. 2.3

[63] Wei Liu and Sanjay Chawla. Class confidence weighted *k*nn algorithms for imbalanced data sets. In *PAKDD (2)*, pages 345–356, 2011. 2.1

[64] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory under-sampling for class-imbalance learning. In *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pages 965–969, 2006. doi: 10.1109/ICDM.2006.68. 5.1

[65] Yang Liu, Nitesh V. Chawla, Mary P. Harper, Elizabeth Shriberg, and Andreas Stolcke. A study in machine learning from imbalanced data for sentence boundary detection in speech, 2006. 1.1

[66] Jun Ma, Guanzhong Dai, and Zhong Xu. Network anomaly detection using dissimilarity-based one-class svm classifier. In *Proceedings of the 2009 International Conference on Parallel Processing Workshops*, ICPPW '09, pages 409–414, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3803-7. doi: 10.1109/ICPPW.2009.6. URL `http://dx.doi.org/10.1109/ICPPW.2009.6`. 2.3

[67] Marcus A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, 2003. 2.1, 2.1

[68] Hiroshi Mamitsuka and Naoki Abe. Active ensemble learning: Application to data mining and bioinformatics. *Systems and Computers in Japan*, 38(11):100–108, 2007. ISSN 1520-684X. doi: 10.1002/scj.10355. URL `http://dx.doi.org/10.1002/scj.`

10355. 2.2

[69] Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. In *In Proceedings of 21st International Conference on Machine Learning (ICML-2004*, pages 584–591. ACM Press, 2004. 2.2

[70] Lori Mosca, Emelia J. Benjamin, Kathy Berra, and et al. Effectiveness-based guidelines for the prevention of cardiovascular disease in women–2011 update: a guideline from the American Heart Association. *Journal of the American College of Cardiology*, 57(12): 1404–1423, March 2011. ISSN 1558-3597. URL http://dx.doi.org/10.1016/j.jacc.2011.02.005. 3.3.2

[71] George Nychis, Vyas Sekar, David G. Andersen, Hyong Kim, and Hui Zhang. An empirical evaluation of entropy-based traffic anomaly detection. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, IMC '08, pages 151–156, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-334-1. doi: 10.1145/1452520.1452539. URL http://doi.acm.org/10.1145/1452520.1452539. 2.3

[72] Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority report in fraud detection: Classification of skewed data, 2004. 1.1

[73] Foster Provost. Machine learning from imbalanced data sets 101, 2000. 2.1

[74] Nicholas Roy and Andrew Mccallum. Toward optimal active learning through sampling estimation of error reduction. In *In Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, 2001. 2.2

[75] Prithviraj Sen and Lise Getoor. Cost-sensitive learning with conditional markov networks. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 801–808, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143945. URL http://doi.acm.org/10.1145/1143844.1143945. 2.1

[76] Burr Settles. Active learning literature survey. Technical report, 2010. 2.2, 6.1

[77] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 287–294, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. doi: 10.1145/130385.130417. URL http://doi.acm.org/10.1145/130385.130417. 2.2

[78] Tom S.F.Haines and Tao Xiang. Active learning using dirichlet processes for rare class discovery and classification, 2011. 2.2

[79] Mai Shouman, Tim Turner, and Rob Stocker. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, 2 (3):220–223, 2012. 3.2.1

[80] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, December 2006. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248547.1248604. 4.3.3

[81] Sören Sonnenburg, Gunnar Rätsch, Sebastian Henschel, Christian Widmer, Jonas Behr,

Alexander Zien, Fabio de Bona, Alexander Binder, Christian Gehl, and Vojt&#x011B;ch Franc. The shogun machine learning toolbox. *J. Mach. Learn. Res.*, 11:1799–1802, August 2010. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1756006.1859911`. 4.4

[82] Jack W. Stokes, John C. Platt, Joseph Kravis, and Michael Shilman. Aladin: Active learning for statistical intrusion detection, 2008. 2.4

[83] Zeeshan Syed and John Guttag. Unsupervised similarity-based risk stratification for cardiovascular events using long-term time-series data. *J. Mach. Learn. Res.*, 999999:999–1024, July 2011. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2021026.2021034`. 2.3

[84] Songbo Tan. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Syst. Appl.*, 28(4):667–671, May 2005. ISSN 0957-4174. doi: 10.1016/j.eswa.2004.12.023. URL `http://dx.doi.org/10.1016/j.eswa.2004.12.023`. 2.1

[85] Lei Tang and Huan Liu. Bias analysis in text classification for highly skewed data. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM ’05, pages 781–784, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2278-5. doi: http://dx.doi.org/10.1109/ICDM.2005.34. URL `http://dx.doi.org/10.1109/ICDM.2005.34`. 1.1, 2.1

[86] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, January 2004. ISSN 0885-6125. 4.3.1

[87] M. Thottan and Chuanyi Ji. Anomaly detection in ip networks. *Signal Processing, IEEE Transactions on*, 51(8):2191 – 2204, aug. 2003. ISSN 1053-587X. doi: 10.1109/TSP.2003.814797. 2.3

[88] Katrin Tomanek and Udo Hahn. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, K-CAP ’09, pages 105–112, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-658-8. doi: http://doi.acm.org/10.1145/1597735.1597754. URL `http://doi.acm.org/10.1145/1597735.1597754`. 2.2

[89] Selen Uguroglu, Mark Doyle, Robert Biederman, and Jaime G. Carbonell. Cost-sensitive risk stratification in the diagnosis of heart disease. In *IAAI*, 2012. 1.4.1, 2.1

[90] Chawla Nitesh V., Bowyer Kevin W., Hall Lawrence O., and Kegelmeyer W. Philip. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=1622407.1622416`. 1.4.3, 2.1, 5.4

[91] A. Wagner and B. Plattner. Entropy based worm and anomaly detection in fast ip networks. In *Enabling Technologies: Infrastructure for Collaborative Enterprise, 2005. 14th IEEE International Workshops on*, pages 172 – 177, june 2005. doi: 10.1109/WETICE.2005.35. 2.3

[92] Gary M. Weiss. The Impact of Small Disjuncts on Classifier Learning. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.141.1134`.

1.1

[93] Gary M. Weiss and Haym Hirsh. A quantitative study of small disjuncts, 2000. 1.1, 5.1

[94] G.M. Weiss. The effect of small disjuncts and class distribution on decision tree learning, 2003. 1.1

[95] Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. 36(3):5718–5727, 2009. 2.1, 5.5, 7

[96] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 435 – 442, nov. 2003. doi: 10.1109/ICDM.2003.1250950. 2.1

[97] Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, pages 204–213. ACM Press, 2001. 2.1

[98] J. Zhang and I. Mani. Knn approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*, 2003. 2.1, 5.1, 5.5, 7

[99] Rui Zhang, Shaoyan Zhang, Sethuraman Muthuraman, and Jianmin Jiang. One class support vector machine for anomaly detection in the communication network performance data. In *Proceedings of the 5th conference on Applied electromagnetics, wireless and optical communications*, ELECTROSCIENCE'07, pages 31–37, Stevens Point, Wisconsin, USA, 2007. World Scientific and Engineering Academy and Society (WSEAS). ISBN 978-960-6766-25-1. URL `http://dl.acm.org/citation.cfm?id=1503549.1503556`. 2.3

[100] Zhaohui Zheng. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6:2004, 2004. 1.1

[101] Jingbo Zhu. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *In Proceedings of ACL*, pages 783–790, 2007. 2.2