2012 IEEE 24th International Conference on Tools with Artificial Intelligence

Label Space Transfer Learning

Samir Al-Stouhi
Department of Computer Engineering
Wayne State University
Detroit, MI 48202
Email: s.alstouhi@wayne.edu

Chandan K. Reddy
Department of Computer Science
Wayne State University
Detroit, MI 48202
Email: reddy@cs.wayne.edu

David E. Lanfear
Heart and Vascular Institute
Henry Ford Hospital
Detroit, MI 48202
Email: dlanfea1@hfhs.org

***Abstract*—Small datasets pose a tremendous challenge in machine learning due to the few available training examples compounded with the relative rarity of certain labels which can potentially impede the development of a representative hypothesis. We define "Rare Datasets" as ones with low samples/features ratio and a skewed label distribution. Since a generalized training model can not be theoretically guaranteed, a method to leverage similar data is needed. We propose the first algorithm that utilizes transfer learning for the label space, present theoretical verification of our method and demonstrate the effectiveness of our framework with several real-world experiments. In addition, we formally describe what constitutes a "Rare Dataset" and present a detailed characterization of related methods.**

Keywords: Rare class, transfer learning, class imbalance, AdaBoost, Weighted Majority Algorithm, healthcare.

I. Introduction

Standard machine learning methods require a training dataset with adequate training examples and a relatively balanced ratio of labels. Several applications challenge these assumptions since there could be a limited number of training examples and a large number of dimensions. Text and image datasets are high-dimensional datasets that often occupy a large hypothesis space requiring complex models with a high VC dimension [1] and a large number of training examples. This problem is compounded when the dataset is highly skewed and unevenly represented where the majority of samples belong to an overrepresented (majority) class and only a few examples belong to an underrepresented (minority) class [2]. A heart failure re-hospitalization prediction problem is presented in this paper where the dataset has a small number of examples and an imbalanced label space.
The concept of compensating for the skew within the label space belongs to the domain of "Imbalanced Learning" and the concept of extracting knowledge from an auxiliary dataset to compensate for the overall lack of samples belongs to a family of methods known as "instance-based transfer learning". We aim to construct a hypothesis and uncover the separating hyperplane with only a handful of training examples with data that is complex in both the feature and label spaces. The complexity of the data and the rarity of training examples prohibit hypothesis construction by human experts or standard algorithms and thus we present a "last resort" sort of solution that can be applied when nothing else suffices.

II. Formal Definition of "Rare Classes"

We will refer to Figure 1 for an overview of four different types of datasets and the associated data-mining methods.

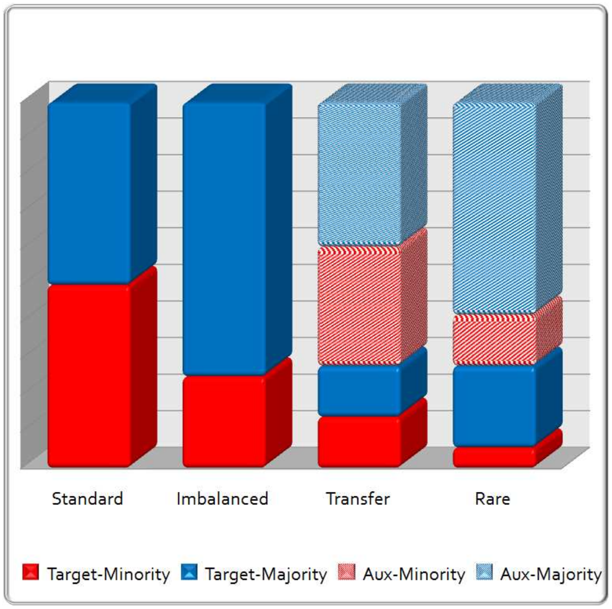


Figure 1: Different Types of Datasets

A. Standard Dataset

Standard datasets are the most studied in the machine learning domain. Standard datasets are balanced where relatively equal numbers of samples belong to each label with a label of approximately one. Standard datasets have an adequate number of examples to construct a representative model. "Probably Approximately Correct (PAC)" learning theory [3] estimates the minimal required number of samples needed to develop a hypothesis. PAC is applied to uncover if the ratio of the dimensions to the number of training samples is too large. If that ratio is

1082-3409/12 $26.00 © 2012 IEEE
DOI 10.1109/ICTAI.2012.103

exceedingly high, the hypothesis space would be too large and prone to model over-fitting. PAC gives a theoretic relationship between the number of samples needed in terms of the size of hypothesis space and the number of dimensions. The simplest example is a binary data set with binary classes and d dimensions with hypothesis space of size 2^{2^d}, requiring $O\left(2^n\right)$ samples [4].

B. Imbalanced Dataset

A dataset is considered imbalanced if it is skewed and unevenly represented where most of its samples belong to an overrepresented (majority) class and only few samples belong to an underrepresented (minority) class [2]. Since traditional learning methods are optimized to maximize accuracy, they generally fail to develop a proper hypothesis when overwhelmed by a majority class of examples [5]. In imbalanced datasets, the data distribution of the majority class significantly dominates the instance space.

The authors in [2] present a comprehensive review of the methods that are applied for imbalanced data. They outline different sampling methods, cost-sensitive learning methods, kernel-based learning methods, and active learning methods. A popular class imbalance algorithm, SMOTE [6], generates an arbitrary number of synthetic minority examples to shift the classifier learning bias toward the minority class. SMOTEBoost [6] is an extension work based on this idea where the synthetic procedure was integrated with adaptive boosting. Similar methods such as Borderline SMOTE [7] and Adaptive Synthetic Sampling [8] have been proposed. Sun et al. [9] proposed a cost-sensitive boosting algorithm to undertake the class imbalance problem with multiple classes. Most of the class imbalanced classification algorithms attach a cost to the minority class as **it is assumed that the minority class is more important than the majority**.

C. Transfer Dataset

Transfer learning datasets are balanced datasets where classification on the dataset of interest (referred to as target set) is improved by including a similar and possibly larger auxiliary dataset (referred to as the source set). Such knowledge transfer can be gained by integrating relevant source samples into the training model or by mapping the source set training models to the target models. The knowledge assembled can be transferred across domain tasks and domain distributions **with the assumption that they are mutually relevant, related, and similar**.

Pan and Yang [10] present a comprehensive survey of transfer learning methods and discuss the relationship between transfer learning and other related machine learning techniques. Methods for transfer learning include an adaptation of Gaussian processes to the transfer learning scheme via similarity estimation between source and target tasks [11]. A SVM framework was proposed by Wu and Dietterich [12] where scarcity of target data is offset by abundant low-quality source data. Pan, Kwok, and Yang [13] used a low-dimensional mapping space to reduce the distribution difference between source and target domains by exploiting Borgwardt's Maximum Mean Discrepancy Embedding (MMDE) method [14], which was originally designed for dimensionality reduction. Pan et al. [15] proposed a more efficient feature-extraction algorithm, known as Transfer Component Analysis (TCA), to overcome the computationally expensive cost of MMDE. Several boosting-based algorithms have been modified for transfer learning and will be more rigorously analyzed in this paper.

D. Rare Dataset

A rare dataset is a small and imbalanced dataset where a similar auxiliary and possibly larger dataset is available to improve classification. Researchers have addressed the subject of rare data as they identified such datasets and acknowledged the need for transfer learning methods for improved classification.

Shuli et a.l [16] define what constitutes a rare dataset and present an overview of the work in the field. Weiss [17] presents a great overview of the problems encountered when handling rare data and addresses the issues encountered by researchers evaluating such datasets. Different solutions are outlined for handling absolute and relatively rare data with a discussion of solutions for segmentation, bias and noise associated with these datasets. In his latest work, Weiss [18] singles out the usage of additional data when the absolute number of samples is rare as a technique that warrants additional research. In [19], an end-to-end investigation of rare categories in imbalanced data sets in both the supervised and unsupervised settings is presented and transfer learning for rare datasets is singled out as a possible future solution.

E. Summary of Methods

Table I presents a summary of the machine learning domains that are applied to handle the datasets outlined in Figure 1. Classification of rare classes exists as a fringe field of research between different machine learning domains and we attempt to address this problem in this paper. Our main contributions:

1) Formally define what constitutes a "Rare Dataset".
2) Present a survey of related methods and highlight the need for a new type of algorithms to solve a niche but important problem that is not addressed in current literature.
3) Propose the first transfer learning algorithm optimized for transfer within the label space.
4) Submit experimental results for real-world datasets to demonstrate the effectiveness of our framework.

The paper is organized as follows: In Section II, we formally defined the problem and discussed the related works. Our algorithm is proposed in Section III along with theoretical foundation. Experimental results with related discussions are given in Section IV. The last two sections conclude our discussion with possible future directions.

Table I: Summary of the different datasets' attributes

Learning Domain	**Satisfactory Size**	**Balanced Dataset**
Standard	Yes	Yes
Imbalanced	Yes	No
Transfer	No	Yes
Rare	No	N0

III. Proposed Label-Transfer Algorithm

A. Boosting-based Transfer Learning

Consider a domain (D) comprised of feature space (X). We can specify a mapping function to map the feature space to the label space as "$X \to Y$" where $Y \in \{-1, 1\}$. Let us denote the domain with auxiliary data as the source domain set (X_{src}) and denote (X_{tar}) as the target domain set that needs to be mapped to the label space (Y_{tar}).
Boosting-based transfer learning methods apply ensemble training to both source and target instances with an update mechanism that incorporates only the source instances that are useful for target instance classification. These methods perform this form of mapping by giving more weight to source instances that improve target training and vice-versa. TrAdaBoost [20] is the first transfer learning method to use boosting as a best-fit inductive transfer learner. TrAdaBoost trains the base classifier on the weighted source and target set in an iterative manner. The source instances that are not correctly classified on a consistent basis would converge and would not be used in the final classifier's output since that classifier only uses boosting iterations $\frac{N}{2} \to N$. Dynamic-TrAdaBoost [21] improved TrAdaBoost by correcting for the bias induced by normalization causing early convergence. Boosting has been extended to many transfer problems including regression transfer [22] and multi-source learning [23]. TransferBoost [24] is used for boosting when multiple source tasks are available. TransferBoost calculates an aggregate transfer term for every source task as the difference in error between the target-only task and the target plus each additional source task. AdaBoost was also extended in [25] for concept drift with a fixed cost that is incorporated into the source instances' update mechanism via AdaCost [9]. This cost is pre-calculated using probability estimates as a measure of relevance between source and target distributions. A source task that is unrelated to the target task will exhibit negative transferability and its instances' weights would be diminished by a fixed [25] or dynamic rate [24] within AdaBoost's update mechanism.

Table II: Summary of the Notations

Notation	**Description**
X	feature space, $X \in \mathbb{R}^d$
Y	label space = $\{-1, 1\}$
d	number of features
F	mapping function $X \to Y$
D	domain
src	source (auxiliary) instances
tar	target instances
maj	majority class,
min	minority class
ε^t	classifier error at boosting iteration "t"
w	weight vector
N	number of iterations
n	number of source instances
m	number of target instances
t	index for boosting iteration
$\ddot{f}^t$	weak classifier at boosting iteration "t"
$\mathbf{I}$	Indicator function

B. Algorithm Description

The pseudo code of "Label-Transfer" is presented in Algorithm 1. The framework is based on AdaBoost[26], which is a meta algorithm that combines a cascade of weak classifiers for an optimal feasible solution. AdaBoost generates a weighted set of additive weak classifiers to construct a committee capable of non-linear approximation. The weak classifier on line 7 is trained with the weighted target and source instances to discover the hyperplane that forms the classification decision boundaries. In Algorithm 1, the decision boundaries are utilized to build a model for the target instances and to control the weight of the auxiliary instances that are best fit for training. The target instances' weights are updated using only the target's accuracy rate which is calculated on line 8 and is incorporated into the target instances update mechanism after calculating β_{tar} on line 11 and using it for AdaBoost's update mechanism on line 14 as:

$$w^{t+1}_{tar_i} = w^{t}_{tar_i} \beta_{tar}^{\mathbb{I}\left[y_{tar_i} \neq \ddot{f}^t_i\right]} \tag{1}$$

Given a majority label, $Y_{src-maj} \in \{+1\}$, Sensitivity(Sen) is the label dependent accuracy measure of the source instances' majority and is calculated on line 9 as:

$$Sen^t_{src} = \sum_{j=1}^{n} \frac{\left[w^j_{src-maj}\right] \mathbb{I}\left[y_{src-maj_j} = \ddot{f}^t_j\right]}{\sum\limits_{i=1}^{n} \left[w^i_{src-maj}\right]} \tag{2}$$

Specificity (Spc) is the label dependent accuracy measure of the source instances' minority instance, $Y_{src-min} \in \{-1\}$, and is calculated on lines 10 as:

$$Spc^t_{src} = \sum_{j=1}^{n} \frac{\left[w^j_{src-min}\right] \mathbb{I}\left[y_{src-min_j} = \ddot{f}^t_j\right]}{\sum\limits_{i=1}^{n} \left[w^i_{src-min}\right]} \tag{3}$$

As per the transfer learning paradigm, the source distribution is considered relevant and target instances can benefit from incorporating relevant source instances. The label dependent accuracy is applied on line 12 to update the source majority weights as:

$$w^{t+1}_{src-maj_i} = \left(1 + Sen^t_{src}\right) w^t_{src_i} \beta_{src}^{\left[y_{src_i} \neq \ddot{f}^t_i\right]} \tag{4}$$

The source minority weights are updated on line 13 as:

$$w^{t+1}_{src-min_i} = \left(1 + Spc^t_{src}\right) w^t_{src_i} \beta_{src}^{\left[y_{src_i} \neq \ddot{f}^t_i\right]} \tag{5}$$

The update mechanism in equations 4 and 5 is borrowed from the Weighted Majority Algorithm(WMA)[27] with a label-dependent dynamic cost factor. WMA is a meta-learning algorithm that constructs an additive set of weak learners, where the number of mistakes for n source samples ($n\varepsilon^{WMA}$) is bounded by the number of mistakes made by the best performing of the N weak classifiers ($n\varepsilon^{best}$) as:

$$n\varepsilon^{WMA} \le \frac{n\varepsilon^{best} \ln\left(\beta^{-1}_{src}\right) + \ln(N)}{1 - \beta_{src}} \tag{6}$$

Label-Transfer works within the transfer learning paradigm but applies an update mechanism optimized to balance the distribution of labels within the label space. While SMOTE synthetically generates minority samples, Label-Transfer infuses samples from an auxiliary domain that fit the outcome of the weak classifiers where transfer learning is shifted to the label space by controlling the weight convergence using a label-dependent error. All samples are initially given equal weights and minority labels are initially misclassified by weak classifiers optimized for accuracy. These classifiers tend to achieve high accuracy by binning most data to the majority label when all instances are equally weighted. The label dependent costs control the convergence of the source instances as weights converge slower for labels with initial high error rates (minority labels) and vice versa. As minority labels get higher normalized weights with every boosting iteration, the classifiers would subsequently construct more balanced separating hyperplanes. Since only the $\frac{N}{2} \to N$ weak classifiers are used for the final output, the expectation is that a more balanced mix of source weights would have been constructed in the initial $1 \to \frac{N}{2}$ boosting iterations. This is the first transfer learning algorithm to optimize with label information and shift the instance transfer from the feature-space to the label-space.

IV. Experimental Results on Real-World Datasets

A. Experimental Setup

AdaBoost [26] with target instances was applied as the standard boosting classifier. We applied SMOTE [6] to the target data before boosting to compare with an imbalanced method (SMOTE-AdaBoost). For transfer learning, we used Dynamic-TrAdaBoost [21] as the reference algorithm. Thirty boosting iterations was experimentally proven sufficient for training. Small and imbalanced datasets can

Algorithm 1 Label-Transfer

```
Require:
  ▷ Source Majority $D_{src-maj} = \{x_{src-maj_i}, y_{src-maj_i}\}$
  ▷ Source Minority $D_{src-min} = \{x_{src-min_i}, y_{src-min_i}\}$
  ▷ Target Majority $D_{tar-maj} = \{x_{tar-maj_i}, y_{tar-maj_i}\}$
  ▷ Target Minority $D_{tar-min} = \{x_{tar-min_i}, y_{tar-min_i}\}$
  ▷ $Y_{maj} \in \{+1\}, Y_{min} \in \{-1\}$
  ▷ $\text{Max iterations} : N, \text{Base learner} : \ddot{f}$
  ▷ $\text{Source samples} : n, \text{Target samples} : m$
Ensure: $\text{Target Classifier Output} : \left\{ \dot{f} : X \to Y \right\}$
    $\dot{f} = sign\left[\prod_{t=\frac{1}{2}}^{N}\left({\beta^{t}_{tar}}^{-\ddot{f}^t}\right) - \prod_{t=\frac{1}{2}}^{N}\left({\beta^{t}_{tar}}^{-\frac{1}{2}}\right)\right]$
Procedure:
 1: Initialize the target weight vector:
        $w_{tar} = \{w_{tar-maj} \cup w_{tar-min}\}$
 2: Initialize the source weight vector:
        $w_{src} = \{w_{src-maj} \cup w_{src-min}\}$
 3: $Set\,\beta_{src} = \frac{1}{1+\sqrt{\frac{2\ln(n)}{N}}}$
 4: Set $D$ as the combined dataset with $w$ weights:
        $D = \{D_{src-maj} \cup D_{src-min} \cup D_{tar-maj} \cup D_{tar-min}\}$
 5: for $t = 1$ to $N$ do
 6:     Normalize Weights: $w = \frac{\mathrm{w}}{\sum_{i}^{n} \mathrm{w}_{src_i} + \sum_{j}^{m} w_{tar_j}}$
 7:     Find the candidate weak learner $\ddot{f}^t : X \to Y$ that
        minimizes error for the D weighted according to $w$
 8:     Calculate label-independent accuracy of $D_{tar}$:
            $Acc^t_{tar} = \sum_{j=1}^{m} \frac{\left[w^j_{tar}\right]\mathbb{1}\left[y_{tar} = \ddot{f}^t_j\right]}{\sum_{j=1}^{m}\left[w^j_{tar}\right]}$
 9:     Calculate label dependent accuracy of $D_{src-maj}$:
            $Sen^t_{src} = \sum_{j=1}^{n} \frac{\left[w^j_{src-maj}\right]\mathbb{1}\left[y_{src-maj_j} = \ddot{f}^t_j\right]}{\sum_{i=1}^{n}\left[w^i_{src-maj}\right]}$
10:     Calculate label dependent accuracy of $D_{src-min}$:
            $Spc^t_{src} = \sum_{j=1}^{n} \frac{\left[w^j_{src-min}\right]\mathbb{1}\left[y_{src-min_j} = \ddot{f}^t_j\right]}{\sum_{i=1}^{n}\left[w^i_{src-min}\right]}$
11:     Set $\beta^t_{tar} = \frac{Acc^t_{tar}}{1-Acc^t_{tar}}$
12:     Update Source Majority Weights $(i \in maj)$:
            $w^{t+1}_{src-maj_i} = \left(1 + Sen^t_{src}\right) w^t_{src_i} \beta_{src}^{\left[y_{src_i} \neq \ddot{f}^t_i\right]}$
13:     Update Source Minority Weights $(i \in min)$:
            $w^{t+1}_{src-min_i} = \left(1 + Spc^t_{src}\right) w^t_{src_i} \beta_{src}^{\left[y_{src_i} \neq \ddot{f}^t_i\right]}$
14:     Update Target Weights:
            $w^{t+1}_{tar_i} = w^t_{tar_i} {\beta^t}_{tar}^{\mathbb{1}\left[y_{tar_i} \neq \ddot{f}^t_i\right]}$
15: end for
```

prematurely converge during training; For a fair comparison, we terminated learning and restarted if any algorithm did not reach 30 boosting iterations.

Base Learner $\left(\ddot{f}\right)$: We did not use decision stumps as weak learners since the majority of training data belongs to the source and we need to guarantee an error rate of less than 0.5 on the target to avoid early boosting termination (as mandated by AdaBoost). For example, decision stumps on data with 95% source and 5% target or 95% majority and 5% minority is not guaranteed (and will certainly not work for many boosting iterations) to get an error rate of less than 0.5 on minority target instances that compromise a small subset of the training data. We used a strong classifier, classification trees, and applied a top-down approach where we trimmed the tree at the first node that achieved a target error rate that is less than 0.5.

Cross Validation: We used standard cross validation methods when the number of target minority samples was sufficient. If the target dataset had too many samples to fit the definition of a rare dataset, we used a fraction for training and left the remainder for testing. We randomly selected non-intersecting training and testing sets when available samples were not sufficient for standard cross validation. We ran each experiment 30 times and reported the average accuracy to reduce bias.

B. Dataset Description

We tested several healthcare related datasets and provided a detailed description of these datasets in Table III.

HealthCare: Heart failure (HF) continues to be an enormous public health problem despite the many advances in its pharmacotherapy over the past 25 years [28]. HF is associated with significant clinical and economic burden and the high rate of hospitalization is a major contributor of the estimated cost for 2009 of \$37.2 billion [29]. Re-hospitalization for heart failure occurs in one-in-five patients within 30 days of discharge in patients over 65 years old. HF patient data was collected at Henry Ford Health System (HFHS) in Detroit. Using administrative resources at HFHS, we identified patients with a discharge diagnosis of heart failure (9th Edition/Revision International Classification of Diseases [ICD-9]) between January 1, 2000 and June 30, 2008. The index of hospitalization was the first inpatient admission during the period of observation. There were 8913 unique patients who had a first hospitalization with primary HF diagnosis. Patients were monitored until they reached an endpoint (death or re-hospitalization), or were censored at the earlier of either dis-enrollment or final follow up on December 31, 2008. New models were needed to assess the risk of re-hospitalization. Similar to the population demographics of heart failure studies [30], the data for re-hospitalization is skewed as it reflects the local populations' ethnic demographics and the generally older age of HF patients. The lack of certain demographics required a new set of methods that leverages available demographics and generalizes to demographics with few samples.
This dataset was split across different demographics. Three different demographics were used to test the datasets with three different distributions. Features included demographic and medical diagnostic conditions. The task is to predict if a patient will be re-admitted after being released from the hospital for Heart Failure. We reported the average results with 50 minority samples.

CMC*[1]*: This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The task is to predict if a non-Muslim woman is employed based on her demographic and socio-economic characteristics. Only 24% of the population is non-Muslim and only 4.8% of the population is non-Muslim and also employed.

Parkinson*[1]*: This dataset is composed of a range of biomedical voice measurements from people with early-stage Parkinson's disease. The goal is to predict if a patient's score on the Unified Parkinson's Disease Rating Scale (UPRDS) is high (UPRDS$\geq$10) or low (UPRDS$<$10).

C. Experimental Results

1) F-Measure Analysis: F-measure results are presented in Table IV with the following significance tests:

- Tested the null hypothesis that the F-measure performance is **not** significantly better than AdaBoost and applied the Friedman Test with $p < 0.01$. SMOTE-AdaBoost and Label-Transfer were able to reject the hypothesis for all datasets.
- Performed paired t-tests with $\alpha = 0.01$ to test the null hypothesis that F-measure performance was **not** improved over SMOTE-AdaBoost. For all datasets, Label-Transfer rejected the hypothesis.

Table IV: Comparison of F-Measure values on Real-world datasets.

	Target AdaBoost	SMOTE AdaBoost	Dynamic TrAdaBoost	Rare Transfer
HealthCare (Race)	0.205	0.257	0.055	**0.328**
HealthCare (Age)	0.180	0.229	0.063	**0.269**
HealthCare (Sex)	0.194	0.236	0.037	**0.301**
CMC (Religion)	0.276	0.325	0.188	**0.378**
Parkinson (Sex)	0.404	0.552	0.702	**0.749**

[1] http://archive.ics.uci.edu/ml/

Table III: Description of the Datasets

Dataset	Features	Source Majority	Source Minority	Target Majority	Target Minority
HealthCare (Race)	Numeric : 2 Nominal : 20	African-American Not Re-hospitalized 4468 (78%)	African-American Re-hospitalized 1026 (18%)	Caucasian Not Re-hospitalized ≈183 (3.2%)	Caucasian Re-hospitalized ≈50 (0.87%)
HealthCare (Age)	Numeric : 1 Nominal : 21	Over 50 Not Re-hospitalized 4513 (75%)	Over 50 Re-hospitalized 1182 (20%)	Under 50 Not Re-hospitalized ≈241 (4.0%)	Under 50 Re-hospitalized ≈50 (0.84%)
HealthCare (Sex)	Numeric : 2 Nominal : 20	Male Not Re-hospitalized 3366 (76%)	Male Re-hospitalized 818 (18%)	Female Not Re-hospitalized ≈211 (4.8%)	Female Re-hospitalized ≈50 (1.1%)
CMC (Religion)	Numeric : 5 Nominal : 3	Muslim Un-employed 955 (74%)	Muslim Employed 298 (23%)	Non-Muslim Un-employed 15 (0.02%)	Non-Muslim Employed 7 (0.006%)
Parkinson (Sex)	Numeric : 19 Nominal : 0	Male UDPRS≥10 3732 (89%)	Male UDPRS<10 276 (8%)	Female UDPRS≥10 112 (0.03%)	Female UDPRS<10 13(0.003%)

2) Specificity Analysis: Specificity results are presented in Table V with the following significance tests:

- Tested the null hypothesis that the specificity performance is **not** significantly better than AdaBoost and applied the Friedman Test with p $<$ 0.01. SMOTE-AdaBoost and Label-Transfer were able to reject the hypothesis for all datasets.
- Performed paired t-tests with $\alpha = 0.01$ to test the null hypothesis that specificity performance was **not** improved over SMOTE-AdaBoost. For all datasets, Label-Transfer rejected the hypothesis.

Table V: Comparison of Specificity values on Real-world datasets.

	Target AdaBoost	SMOTE AdaBoost	Dynamic TrAdaBoost	Rare Transfer
HealthCare (Race)	0.178	0.279	0.033	**0.411**
HealthCare (Age)	0.150	0.244	0.039	**0.321**
HealthCare (Sex)	0.167	0.262	0.021	**0.341**
CMC (Religion)	0.249	0.321	0.120	**0.471**
Parkinson (Sex)	0.306	0.550	0.748	**0.792**

3) G-Mean and AUC Analysis: G-Mean results are presented in Table VI and AUC results in Table VII as evidence that boosting specificity, reducing error for the minority label, did not degrade the overall performance of the classifier.

Table VI: Comparison of G-Mean values on Real-world datasets

	Target AdaBoost	SMOTE AdaBoost	Dynamic TrAdaBoost	Rare Transfer
HealthCare (Race)	0.382	0.456	0.145	**0.533**
HealthCare (Age)	0.355	0.440	0.164	**0.478**
HealthCare (Sex)	0.374	0.444	0.117	**0.510**
CMC (Religion)	0.422	0.467	0.331	**0.488**
Parkinson (Sex)	0.518	0.715	0.841	**0.874**

Table VII: Comparison of AUC values on Real-world datasets

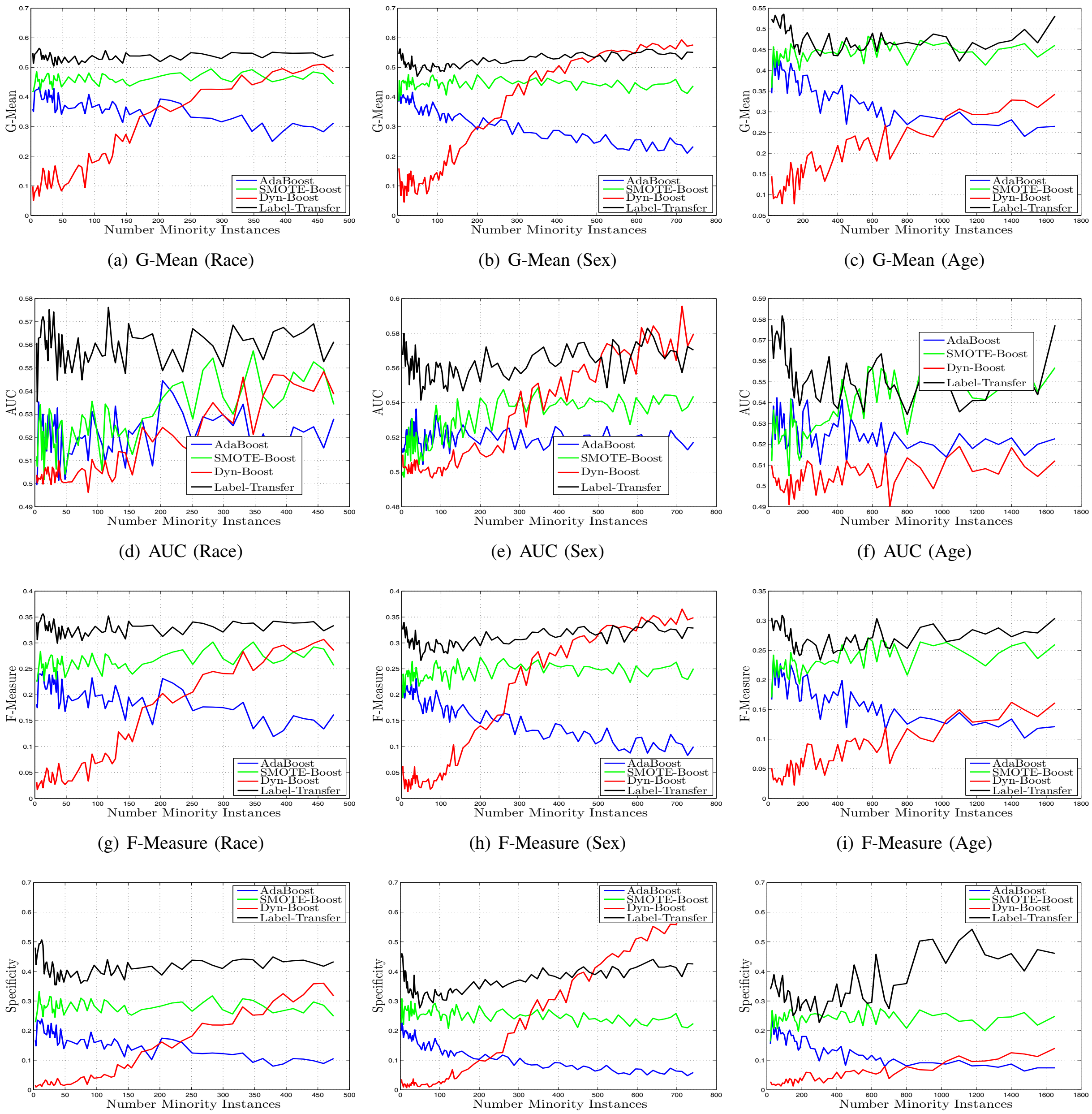
	Target AdaBoost	SMOTE AdaBoost	Dynamic TrAdaBoost	Rare Transfer
HealthCare (Race)	0.519	0.521	0.504	**0.559**
HealthCare (Age)	0.526	0.532	0.503	**0.555**
HealthCare (Sex)	0.520	0.517	0.502	**0.560**
CMC (Religion)	0.506	0.510	**0.524**	0.513
Parkinson (Sex)	0.649	0.761	0.862	**0.885**

D. Performance with different target instances

We present the F-measure, G-Mean, AUC and Specificity plots of the HealthCare dataset in Figure 2. We plot the measures using different demographics and a varied the number of target samples. Our algorithm compensated for the lack of minority data while Dynamic-TrAdaBoost only performed well once the minority samples' size reached a significant level. The graphs demonstrate that our algorithm minimized the minority labels error without degrading the classifier's overall performance.

V. Discussion and Extensions

A true extension would be to use this algorithm to induce a SMOTE type of balance. We did not do so in this paper to have a fair comparison and to illustrate the effectiveness of our method. Rather than generating synthetic samples, as in SMOTE, we can integrate source instances that retained the maximum weight after N boosting iterations. Such instances would have relatively higher weight because they fit the target instances' hypothesis space and thus fit the target's feature distribution. Another more advanced approach is to generate synthetic samples from the combination of weighted samples after N boosting iterations. Both proposed methods can be used to generate majority and minority



(a) G-Mean (Race) (b) G-Mean (Sex) (c) G-Mean (Age)

(d) AUC (Race) (e) AUC (Sex) (f) AUC (Age)

(g) F-Measure (Race) (h) F-Measure (Sex) (i) F-Measure (Age)

(j) Specificity (Race) (k) Specificity (Sex) (l) Specificity (Age)

Figure 2: G-mean, AUC, F-measure, and Specificity results on different demographics at different minority samples.

samples that can be integrated with the target instances to construct a standard set that is balanced and have enough samples for training. Once a standard set is constructed, any standard machine learning algorithm can be quickly and efficiently applied.

VI. Conclusion

We formally defined what constitutes a "Rare Class" and identified a niche area of research not covered in current machine learning domains. We presented an overview of related fields and motivated the necessity for more research with actual scenarios. We proposed the first transfer learning method optimized for the label space with theoretical verification. We demonstrated the effectiveness of our framework with demographics data. Future work could remove classification from the current algorithm and it can serve as an intermediate step for label space transfer learning to extract samples from the auxiliary domain to augment the training set.

VII. Acknowledgements

This work is supported in part by the US National Science Foundation grant IIS-1231742.

References

[1] V. N. Vapnik and Ya, “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities,” *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[2] H. He and E. Garcia, “Learning from imbalanced data,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263–1284, 2009.

[3] M. J. Kearns and U. V. Vazirani, *An introduction to computational learning theory*. MIT Press, 1994.

[4] T. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.

[5] F. Provost, “Machine learning from imbalanced data sets 101,” in *of the Am. Assoc. for Artificial Intelligence Workshop*, 2000.

[6] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[7] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-smote: A new over-sampling method in imbalanced data sets learning,” in *International Conference on Intelligent Computing*, vol. 3644, 2005, pp. 878–887.

[8] H. Haibo, B. Yang, E. A. Garcia, and L. Shutao, “Adasyn: Adaptive synthetic sampling approach for imbalanced learning,” in *of the IEEE International Joint Conference on Neural Networks*, 2008, pp. 1322–1328.

[9] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, “Cost-sensitive boosting for classification of imbalanced data,” *Pattern Recognition*, vol. 40, no. 12, pp. 3358 – 3378, 2007.

[10] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[11] B. Cao, S. J. Pan, Y. Zhang, D. Yeung, and Q. Yang, “Adaptive transfer learning,” in *proceedings of the AAAI Conference on Artificial Intelligence*, 2010, pp. 407–412.

[12] P. Wu and T. G. Dietterich, “Improving svm accuracy by training on auxiliary data sources,” in *proceedings of the twenty-first international conference on Machine learning*, 2004, pp. 871–878.

[13] S. J. Pan, J. T. Kwok, and Q. Yang, “Transfer learning via dimensionality reduction,” in *proceedings of the national conference on Artificial intelligence*, 2008, pp. 677–682.

[14] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schlkopf, and A. J. Smola, “Integrating structured biological data by kernel maximum mean discrepancy,” *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[15] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” in *proceedings of the 21st international jont conference on Artifical intelligence*, 2009, pp. 1187–1192.

[16] H. Shuli, Y. Bo, and L. Wenhuang, “Rare class mining: Progress and prospect,” in *of the Chinese Conference on Pattern Recognition, CCPR 2009*, 2009, pp. 1–5.

[17] G. M. Weiss, “Mining with rarity: a unifying framework,” *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.

[18] G. Weiss, “Mining with rare cases,” in *Data Mining and Knowledge Discovery Handbook*, 2010, pp. 747–757.

[19] J. He, “Rare category analysis,” Ph.D. dissertation, Carnegie Mellon University, 2010.

[20] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” in *proceedings of the international conference on Machine learning*, 2007, pp. 193–200.

[21] S. Al-Stouhi and C. K. Reddy, “Adaptive boosting for transfer learning using dynamic updates,” in *ECML/PKDD (1)*, 2011, pp. 60–75.

[22] D. Pardoe and P. Stone, “Boosting for regression transfer,” in *proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 863–870.

[23] Y. Yao and G. Doretto, “Boosting for transfer learning with multiple sources,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1855–1862.

[24] E. Eaton and M. desJardins, “Set-based boosting for instance-level transfer,” in *proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 422 –428.

[25] A. Venkatesan, N. Krishnan, and S. Panchanathan, “Cost-sensitive boosting for concept drift,” in *proceedings of the 2010 International Workshop on Handling Concept Drift in Adaptive Information Systems*, 2010.

[26] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in *proceedings of the Second European Conference on Computational Learning Theory*, 1995, pp. 23–37.

[27] N. Littlestone and M. K. Warmuth, “The weighted majority algorithm,” in *proceedings of the 30th Annual Symposium on Foundations of Computer Science*, 1989, pp. 256–261.

[28] P. A. McCullough, E. F. Philbin, J. A. Spertus, S. Kaatz, K. R. Sandberg, and W. D. Weaver, “Confirmation of a heart failure epidemic: findings from the resource utilization among congestive heart failure (reach) study,” *Journal of the American College of Cardiology*, vol. 39, no. 1, pp. 60–69, 2002.

[29] M. Writing Group, D. Lloyd-Jones, Adams *et al.*, “Heart disease and stroke statistics 2009 update,” *Circulation*, vol. 119, no. 3, pp. 480–486, 2009.

[30] H. Krum and R. E. Gilbert, “Demographics and concomitant disorders in heart failure,” *The Lancet*, vol. 362, no. 9378, pp. 147–158, 2003.