Random Under-Sampling Ensemble Methods for Highly Imbalanced Rare Disease Classification

Dong Dai, and Shaowen Hua

Abstract— **Classification on imbalanced data presents lots of challenges to researchers. In healthcare settings, rare disease identification is one of the most difficult kinds of imbalanced classification. It is hard to correctly identify true positive rare disease patients out of much larger number of negative patients. The prediction using traditional models tends to bias towards much larger negative class. In order to gain better predictive accuracy, we select and test some modern imbalanced machine learning algorithms on an empirical rare disease dataset. The training data is constructed from the real world patient diagnosis and prescription data. In the end, we compare the performances from various algorithms. We find that the random under-sampling Random Forest algorithm has more than 40% improvement over traditional logistic model in this particular example. We also observe that not all bagging methods are out-performing than traditional methods. For example, the random under-sampling LASSO is inferior to benchmark in our reports. Researchers need to test and select appropriate methods accordingly in real world applications.**

Index Terms— **imbalanced, rare disease, random under-sampling, random forest**

I. Introduction

Rare diseases have low prevalence rates and they are difficult to diagnose and identify. The ”Rare Disease Act 2002” defines rare disease to be less than 200,000 patients, or 1 in 1,500 [1]. By this nature, the rare disease dataset is extremely imbalanced. To predict or classify rare diseases from a large population is a challenging task due to low signal-to-noise ratio. This problem is called imbalanced learning in data mining field [2]. Classic statistical methods or standard machine learning algorithms are biased toward larger classes. Therefore, it is hard to positively identify rare disease patients who are in the minority class. If not treating and measuring properly, most of rare disease patients will be mis-classified as the other classes albeit the overall accuracy rate may appear to be high.

There are dedicated new algorithms and methods developed for imbalanced datasets classification. In this paper, we explore selecting and applying imbalanced machine learning algorithms to identify rare disease patients from real-world healthcare data. We will study the performance improvement from using imbalanced algorithms and compare the results to traditional methods. The empirical application we select is to identify Hereditary Angioedema (HAE) disease. HAE is a rare, genetic disease that causes episodic attacks of swelling under the skin. The prevalence of HAE is between 1 in every 10,000 and 1 in every 50,000 [3], [4].

Because it is rare, physicians have rarely encountered patients with this condition. On top of this, HAE attacks also resemble other forms of angioedema. Both make HAE hard to diagnose. Late or missed diagnosis can lead to incorrect treatment and unnecessary surgical intervention. Our objective in this application is to improve HAE patient identification rate using patient historical prescription and diagnosis information. Given the extremely unbalanced classes in the dataset, it is interesting to see how existing algorithms perform in such conditions. We will test a commonly used approach in imbalanced learning - Random Under Sampling method for ensemble learning with LASSO or Random Forest as element learners.

The remaining of the paper will be arranged in following ways. Section 2 describes the data source used in this empirical example. In section 3, we present the rules and methodologies in anonymous patient selection and variable construction. Next section introduces imbalanced learning algorithms. In the last two sections, the results are presented and discussed.

II. Data Source Description

To study the potential application of imbalance algorithms on rare disease identification, we select Hereditary Angioedema disease as an empirical use case. The data has been extracted from IMS longitudinal prescription (Rx) and diagnosis (Dx) medical claims data.

The Rx data is derived from electronic data collected from pharmacies, payers, software providers and transactional clearinghouses. This information represents activities that take place during the prescription transaction and contains information regarding the product, provider, payer and geography. The Rx data is longitudinally linked back to an anonymous patient token and can be linked to events within the data set itself and across other patient data assets. Common attributes and metrics within the Rx data include payer, payer types, product information, age, gender, 3-digit zip as well as the scripts relevant information including date of service, refill number, quantity dispensed and day supply. Additionally, prescription information can be linked to office based claims data to obtain patient diagnosis information. The Rx data covers up to 88% for the retail channel, 48% for traditional mail order, and 40% for specialty mail order.

The Dx data is electronic medical claims from office-based individual professionals, ambulatory, and general health care sites per year including patient level diagnosis and procedure

Dong Dai is Senior Manager with Advanced Analytics, IMS Health, Plymouth Meeting, PA 19462, USA (email: ddai@us.imshealth.com).

Shaowen Hua is Assistant Professor with School of Business, La Salle University, Philadelphia, PA, USA (email: hua@lasalle.edu).

information. The information represents nearly 65% of all electronically filed medical claims in the US. All data is anonymous at the patient level and HIPAA compliant to protect patient privacy.

III. Methodology

Since HAE disease is difficult to identify, many patients with this condition don't have a positive diagnosis ICD-9 code associate with them. Our objective was to build predictive models to find such patients using their past prescription records and other diagnosis histories. We extracted our training samples from Rx and Dx data sources described before.

A. Sample selection

HAE patients are selected as those with HAE diagnosis (ICD-9 CODE = 277.6) and at least one HAE treatment (prescription or procedure) during the selection period (1/1/2012 - 7/31/2015). The patient's index date is defined as the earliest date of HAE diagnosis or HAE treatment. Then the lookback period is defined as from earliest diagnosis or prescription date available from January 2010 till the day before index date. Using this rule, the selected HAE patients will have lookback periods with variable lengths. After further data cleaning by deleting the records without valid gender, age and region etc., we have 1233 HAE patients in the sample data.

Non-HAE patients are selected by randomly matching 200 non-HAE patients with similar lookback period for each one of the 1233 HAE patients. For example, if an HAE patient has a lookback period from 1/15/2013 to 2/5/2014, then a non-HAE patient is qualified to match and to be selected if he/she had clinical activity (any activity of prescription, procedure or diagnosis) from any day in January 2013 to any day in February 2014. Then the lookback period for the non-HAE patient will be defined as from earliest clinical activity date in January 2013 till latest clinical activity date in February 2014. This process of non-HAE patients matching to HAE patients has been done in a greedy manner. Specifically, non-HAE patient sample starts as an empty set, then 200 non-HAE patients matched to a given HAE patient are added to the non-HAE sample, and this approach is repeated for each HAE patient in the sample until we find 200 distinctive non-HAE patients for each HAE patient.

The above process generates 1233 HAE patients and 246600 (200 times 1233) non-HAE patients in the final patient sample.

B. Predictors for HAE

Literatures about HAE has been reviewed, from which a list of numerous potential HAE predictors are prepared. Specifically, three classes of clinical indications are derived for the lookback period for each patient, including prescriptions, procedures and diagnosis. Each clinical indication yields three predictors for the lookback period: flag (Yes/No: whether the patient had prescriptions, procedures or diseases), frequency (how many times the event of prescriptions, procedures or diseases occurred), average frequency (frequency divided by length of lookback period). These clinical predictors with demographic predictors (age, gender and region) compose the final predictors list.

TABLE I
Confusion Matrix

	Actual=1	Actual=0
Predicted=1	True Positive (TP)	False Positive (FP)
Predicted=0	False Negative (FN)	True Negative (TN)

TABLE II
Model Performance Metrics

Metric	**Definition**
AUC	Area under the ROC curve
AUPR	Area under the PR curve
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
True Positive Rate (TPR)	TP/(TP+FN)
False Positive Rate (FPR)	FP/(FP+TN)
Accuracy	(TP+TN)/(TP+FP+FN+TN)

C. Model Performance Evaluation

In a binary decision problem, a classifier labels data sample as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix (Table I, ”1” for positive class and ”0” for negative class). The confusion matrix has four categories: True Positives (TP) are examples correctly labeled as positives; False Positives (FP) refer to negative examples incorrectly labeled as positive; True Negatives (TN) correspond to negatives correctly labeled as negative; finally, False Negatives (FN) refer to positive examples incorrectly labeled as negative.

Based on the confusion matrix, we will be able to further define several metrics to evaluate model performance as listed in Table II.

In ROC space, one plots the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. The FPR measures the fraction of negative examples that are misclassified as positive. The TPR measures the fraction of positive examples that are correctly labeled. In PR space, one plots Recall on the x-axis and Precision on the y-axis. Recall is the same as TPR, whereas Precision measures that fraction of examples classified as positive that are truly positive.

D. Imbalanced Classification

Imbalanced data sets (IDS) correspond to domains where there are many more instances of some classes than others. Classification on IDS always causes problems because standard machine learning algorithms tend to be overwhelmed by the large classes and ignore the small ones. Most classifiers operate on data drawn from the same distribution as the training data, and assume that maximizing accuracy is the principle goal [5], [6].

Therefore, many solutions have been proposed to deal with this problem, both for standard learning algorithms and for ensemble techniques. They can be categorized into three major groups [7], [8]: (i) Data sampling: In which the training data are modified in order to produce a more balanced data to allow classifiers to perform in a similar manner to standard classification [9], [10]; (ii) Algorithmic modification: This procedure is oriented towards the adaptation of base learning methods to be more attuned to class imbalance issues [11]; (iii) Cost-sensitive learning: This type of solutions incorporate approaches at the data level, at the algorithmic level, or at both levels combined, considering higher costs for the misclassification of examples of the positive class with respect to the negative class, and therefore, trying to minimize higher cost errors [12], [13].

In this paper we implement the Random-Under-Sampling (RUS) (majority) approach. We first randomly under-sample the majority class data and combine them with the minority class data to build an artificial balanced dataset, upon which machine learning algorithms will be applied. This process is repeated for several iterations, with each iteration generating a model. The final model is an aggregation of models over all iterations (See Algorithm 1, similar to bagging [14]). In this paper, we apply RUS with LASSO [15] and Random Forest [16] (hereby denoted as ”Bagging LASSO” and ”Bagging RF” respectively). Specifically, we firstly perform a random under sampling of the majority pool (non-HAE) and combine it with all HAE patients to build a artificial balanced dataset, then LASSO or Random Forest is applied to the balanced sampled data; models are aggregated over iterations of random samples to learn the predictive pattern of HAE, while accounting for possible interactions among predictors. The conventional logistic regression (denoted by ”Logit” hereafter) is implemented as benchmark for model performance comparison.

The usual model performance metrics such as prediction accuracy and area under the ROC curve (AUC) are not appropriate for imbalanced classification problem. For example, the imbalance ratio is 1/200 (each one HAE patient has 200 matched non-HAE patients) in our data, a classifier which tries to maximize the accuracy of its classification rule may obtain an accuracy of 99.5% by simply ignoring the HAE patients, with the classification of all patients as non-HAE. Instead, we will use AUPR and Precision at various Recall levels for model performance comparisons in this paper. An important difference between ROC space and PR space is the visual representation of the curves. Looking at PR curves can expose differences between algorithms that are not apparent in ROC space [17].

For validation purpose, data is split into 80% for training and 20% for testing. Model performance for five-fold Cross-Validation on training data and further validation on testing data are both reported. The results and validation are presented in the next section.

IV. Results

With the 80% training data, we have $n = 986$ HAE patients and their 200 times matched non-HAE patients ($n = 197,200$). Then we perform five-fold cross-validation with the training data, specifically, we split all the training data into five folds, and for each given fold, we train a model with the remaining four folds and calculate the performance metrics on the given fold, the final performance outputs are metrics averaged over the five folds. Summary of the results are listed in Table III.

TABLE III
Model Performance (Five-fold Cross-Validation)

Metric	Logit	Bagging LASSO	Bagging RF
AUC	79.81%	83.03%	82.52%
AUPR	9.59%	9.21%	11.61%
Precision (Recall=50%)	4.49%	5.84%	5.99%
Precision (Recall=45%)	6.33%	7.23%	7.55%
Precision (Recall=40%)	7.86%	8.85%	10.04%
Precision (Recall=35%)	10.67%	10.10%	13.44%
Precision (Recall=30%)	13.61%	11.74%	16.98%
Precision (Recall=25%)	15.11%	13.81%	21.74%
Precision (Recall=20%)	18.46%	15.27%	24.69%
Precision (Recall=15%)	22.82%	18.95%	30.25%
Precision (Recall=10%)	29.02%	27.90%	31.83%
Precision (Recall=5%)	36.60%	34.02%	38.63%

We can see that Bagging RF has an improvement of 21% in terms of AUPR compared to Logistic Regression. And particularly for Recall level at 25%, Logistic Regression has a Precision of 15.11% while that of Bagging RF is 21.74%, which is more than 40% improvement over that of Logistic Regression.

And we can see PR curve (Figure 1) differentiates the algorithms better than ROC curve (Figure 2) (The dashed lines in two figures reflect the performance of a random classifier).

We further validate the model performances by applying the models trained from all training data to testing data, which has $n = 247$ HAE patients and their 200 times matched non-HAE patients ($n = 49,400$). A summary of the testing results are listed in Table IV. It further validates that Bagging RF outperforms Logistic Regression in this imbalanced classification problem.

The Bagging LASSO is inferior to Logistic Regression in our case. The reason could be that the high dimensional features have much correlation and in this case less sparse model is preferred.

In searching for better performance, we also test cost-sensitive learning methods such as weighted SVM and weighted LASSO on the same training data. Both of the results don’t show improvement over the standard or random-under-sampling ensemble methods.

V. Conclusion and Discussion

In this rare disease identification problem, we test under-sampling and cost-sensitive learning algorithms on a real

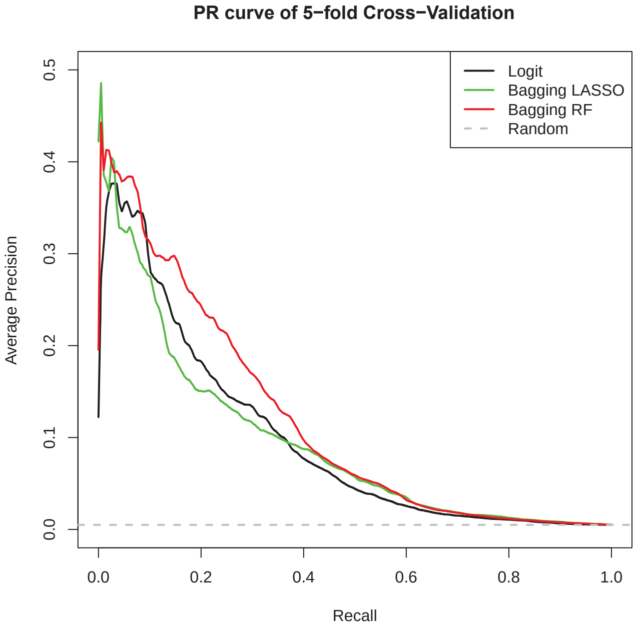

Fig. 1. Five-fold Cross-Validation PR curve

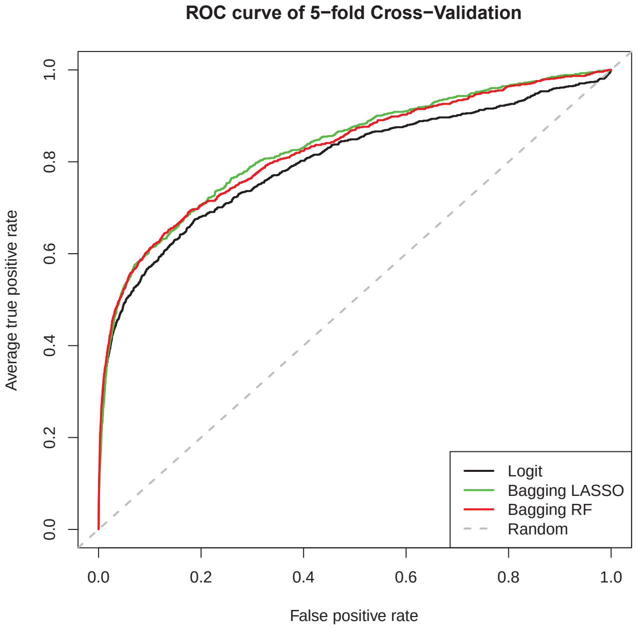

Fig. 2. Five-fold Cross-Validation ROC curve

TABLE IV
MODEL PERFORMANCE (TESTING)

Metric	Logit	Bagging LASSO	Bagging RF
AUC	82.77%	84.33%	85.12%
AUPR	13.02%	11.37%	14.09%
Precision (Recall=50%)	6.09%	5.63%	7.17%
Precision (Recall=45%)	8.35%	8.35%	8.43%
Precision (Recall=40%)	9.71%	9.02%	11.19%
Precision (Recall=35%)	12.63%	10.36%	14.48%
Precision (Recall=30%)	16.78%	12.85%	20.16%
Precision (Recall=25%)	21.91%	16.76%	22.55%
Precision (Recall=20%)	24.50%	20.00%	30.82%
Precision (Recall=15%)	24.03%	20.67%	37.76%
Precision (Recall=10%)	44.64%	37.31%	38.46%
Precision (Recall=5%)	54.55%	63.16%	57.14%

world patient data case. The training sample contains only 0.5% positive patients and it is a good real world example to demonstrate imbalanced learning challenge. We find that in this empirical example, random-under-sampling random forest method can boost precision at various recall levels compared to standard models. In the reported table, it has 40% improvement in AUPR over the benchmark model.

In real world application, due to the data high dimensionality and case complexity, it is not guaranteed that random-under-sampling Random Forest always to be the recommended method. Researchers should try various methods and select most appropriate algorithm based on performance.

APPENDIX

ACKNOWLEDGMENT

REFERENCES

[1] "Rare diseases act of 2002 public law 107280 107th congress."

Algorithm 1 A general framework of Random-Under-Sampling (RUS) ensemble learning

```
Let $S$ be the original training set.
for $k = 1, 2, \ldots, K$ do
    Construct subset $S_k$ containing instances from all classes with same number by executing the following: Random sample instances with (or without) replacement at the rate of $N_c/N_i$ where $N_c$ denote the desired sample size and $N_i$ denotes the original sample size of class $i$.
    Train a classifier $\hat{f}_k$ from subset $S_k$.
    Output final classifier $\hat{g} = \text{sign}\left(\sum_{k=1}^{K} \hat{f}_k\right)$.
end for
```

[2] H. He and Y. Ma, Eds., *Imbalanced learning: foundations, algorithms, and applications.* John Wiley and Sons., 2013.

[3] M. Cicardi and A. Agostoni, "Hereditary angioedema." *New England Journal of Medicine*, vol. 334, no. 25, pp. 1666–1667, 1996.

[4] L. C. Zingale, L. Beltrami, A. Zanichelli, L. Maggioni, E. Pappalardo, B. Cicardi, and M. Cicardi, "Angioedema without urticaria: a large clinical survey." *Canadian Medical Association Journal*, vol. 175, no. 9, pp. 1065–1070, 2006.

[5] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets." *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[6] S. Visa and A. Ralescu, "Issues in mining imbalanced data sets-a review paper." in *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*, vol. 2005, April 2005, pp. 67–73.

[7] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 4, pp. 463–484, 2012.

[8] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and

current trends on using data intrinsic characteristics." *Information Sciences*, no. 250, pp. 113–141.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique." *Journal of artificial intelligence research*, vol. 16, no. 1, pp. 321–357, 2002.

[10] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data." *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[11] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown." in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2001, pp. 204–213.

[12] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive." in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, August 1999, pp. 155–164.

[13] 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*. IEEE, November 2003, pp. 435–442.

[14] L. Breiman, "Bagging predictors." *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society., Series B (Methodological)*, pp. 267–288, 1996.

[16] L. Breiman, "Random forests." *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves." in *Proceedings of the 23rd international conference on Machine learning*. ACM, June 2006, pp. 233–240.