Roy Howie and Lingyue Zheng Friday, November 10, 2016 ECE548 project 2

Task

Design an experiment that illustrates the problem with imbalanced class representation (see section 10.2 in the textbook). To address this problem, use the majority-class under sampling mechanism based on one-sided selection. For performance evaluation, use precision and recall (discussed in section 11.2).

Background

In machine learning, classification algorithms—such as nearest-neighbor and decision trees—are designed with the assumption that datasets are well-formed and evenly distributed, i.e., there is a relatively uniform distribution of classes in the domain.

However, this is often not the case. In many domains, the distribution of classes is extremely lopsided. This may be due to the natural distribution of data or due to how data has been recorded and documented. For example, consider the domain of airplane flights classified by two labels: `successful` and `unsuccessful`. Fortunately, due to modern innovation and technology, there are orders of magnitude more `successful` flights (ones which made it from point A to point B) than there are `unsuccessful` flights (those which crashed or never departed or had to make an alternate landing at point C along the way).

In this case, the dataset is naturally imbalanced. However, a machine does not understand the virtue of this lopsidedness, so it will need assistance is learning from the domain. There are three basic techniques for handling imbalanced datasets: majority-class under-sampling, one-sided Tomek link selection, and minority-class oversampling.

These work exactly as one would expect. Majority-class under-sampling randomly removes instances of the majority class from the dataset until both classes achieve parity; one-sided Tomek link selection finds Tomek links in the dataset and removes only the end corresponding to the majority class; minority-class oversampling introduces new, artificial examples of the minority class into the domain.

Majority-class under-sampling and one-sided Tomek link selection are best utilized when there is a large dataset under one's discretion, specifically when there is a great amount of instances of the majority class which can be discarded without fear of affecting the induced classifier. Minority-class oversampling is most appropriate for cases involving small datasets which most likely cannot handle the removal of further instances without affecting the classifier.

Furthermore, because of how the datasets are known to be imbalanced, instead of a simple accuracy rate, we will use two new methods of evaluation: precision and recall. Precision is the percentage of true positives (i.e., the ratio of correct positive classifications to all positive classifications). Recall is the probability a positive example will be correctly identified (the ratio of true positives to all examples which a perfect classifier would have identified as positive).

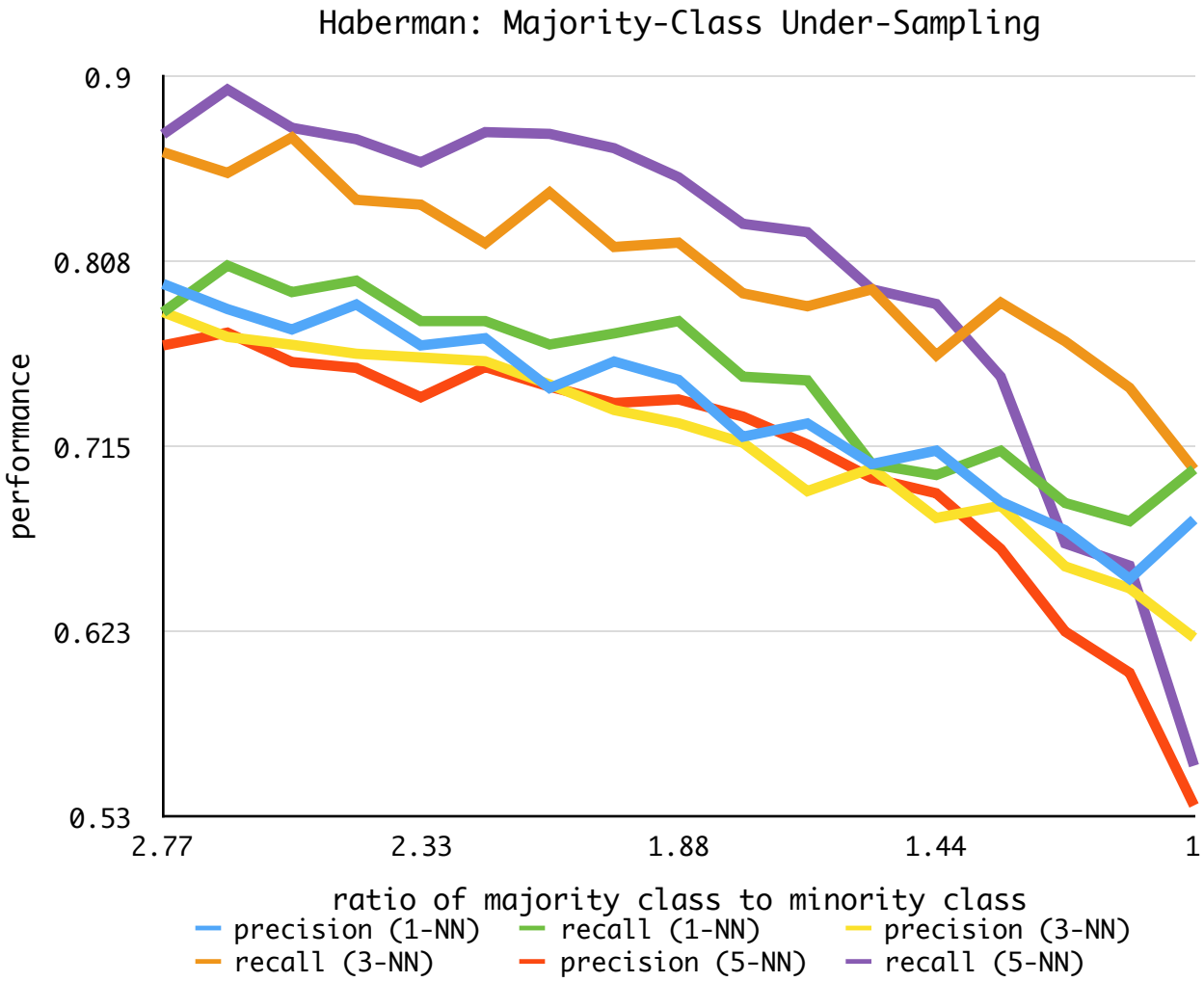
Goal

We will examine two datasets: the Haberman's Survival dataset and the Blood Transfusion dataset. We will transform the domain using two methods: majority-class under-sampling and one-sided Tomek link selection. We will then investigate the effects of these transformations on the ability of the Nearest-Neighbor algorithm to correctly classify the dataset. Performance will be evaluated using precision and recall.

Haberman's Survival Dataset

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. Here is the breakdown:

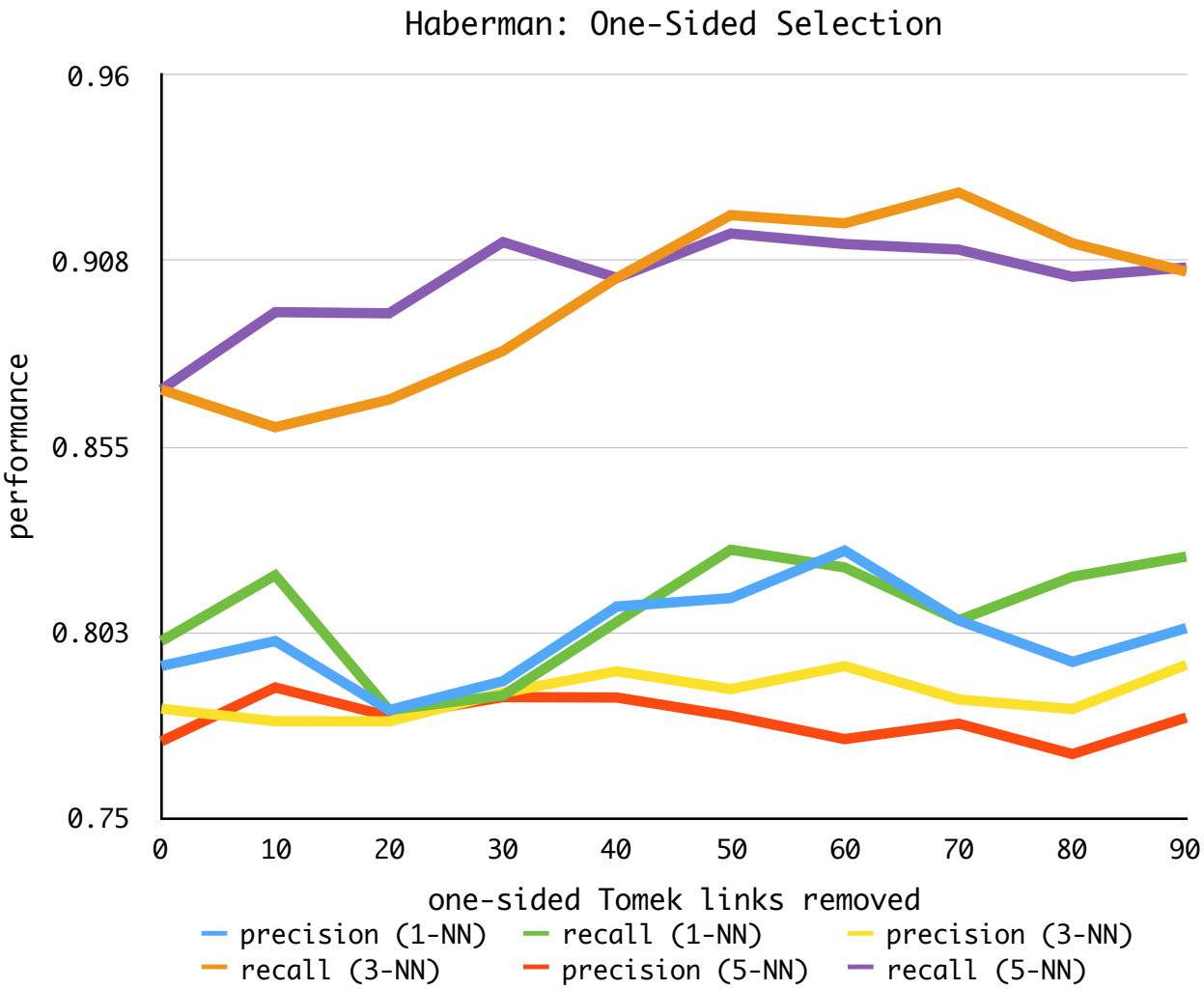
- 3 attributes (age of patient at time of operation, year of operation minus 1900, number of positive axillary nodes detected)
- 2 classes representing survival status: 1 if the patient survived 5 years or more, 2 if the patient died within 5 years
- 225 instances (73.5%) of class 1; 81 instances (26.5%) of class 2



The Haberman's Survival dataset did not fare well under majority-class under-sampling. There was a general downward trend in the ability of Nearest-Neighbor to accurately classify instances of the dataset. Under the evaluation of precision and recall, no classifier increased in performance over the duration of this dataset transformation. In general, there was approximately a 0.30 decrease in precision and recall for each classifier.

Particularly devastated was the ability of the 5-NN classifier to recall, which fell from 90% accuracy at the onset 55% at the time when the dataset had been leavened to an equal distribution of survivors and victims. Similarly, the ability of 3-NN to recall fell from 86% accuracy to 70%.

This method was completely unsuccessful in increasing the performance of the Nearest Neighbor classifier.

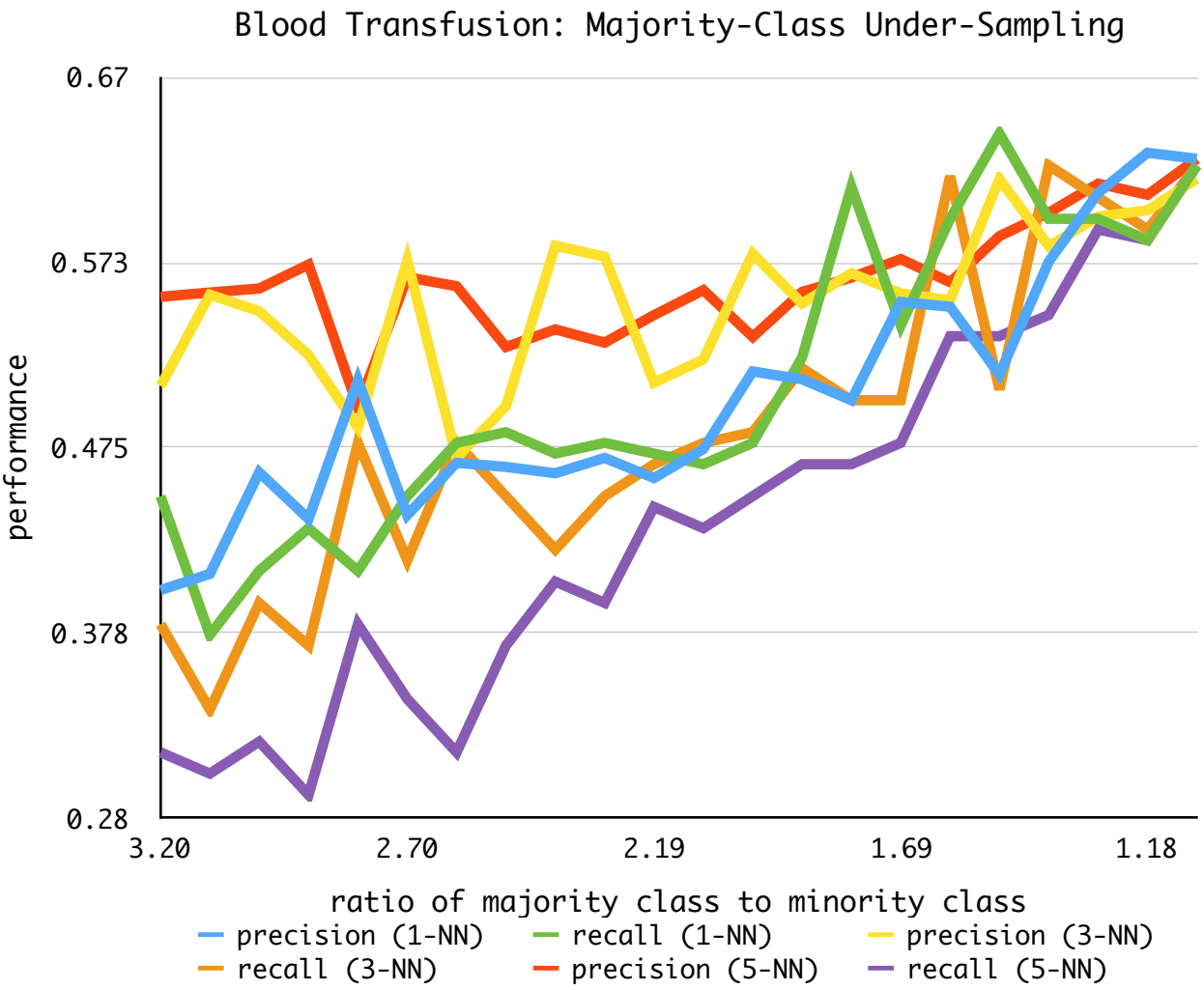


On the other hand, one-sided Tomek link removal had a generally positive or otherwise neutral effect on a classifier's performance. In particular, recall under the 3-NN and 5-NN classifiers increased by about 5% accuracy; all other classifiers fluctuated around 78% accuracy and saw little improvement under the dataset transformation.

Blood Transfusion Dataset

This dataset contains random instances from a blood donor database. The data was derived from visitors to the Blood Transfusion Service Center located in Hsin-Chu City, Taiwan. The center sends their blood transfusion and donated service bus to a single university in Hsin-Chu City to gather blood approximately every 3 months. This dataset is intended to demonstrate the effectiveness of the RFMTC (recency-frequency-monetary-time-classification). Here is the dataset breakdown:

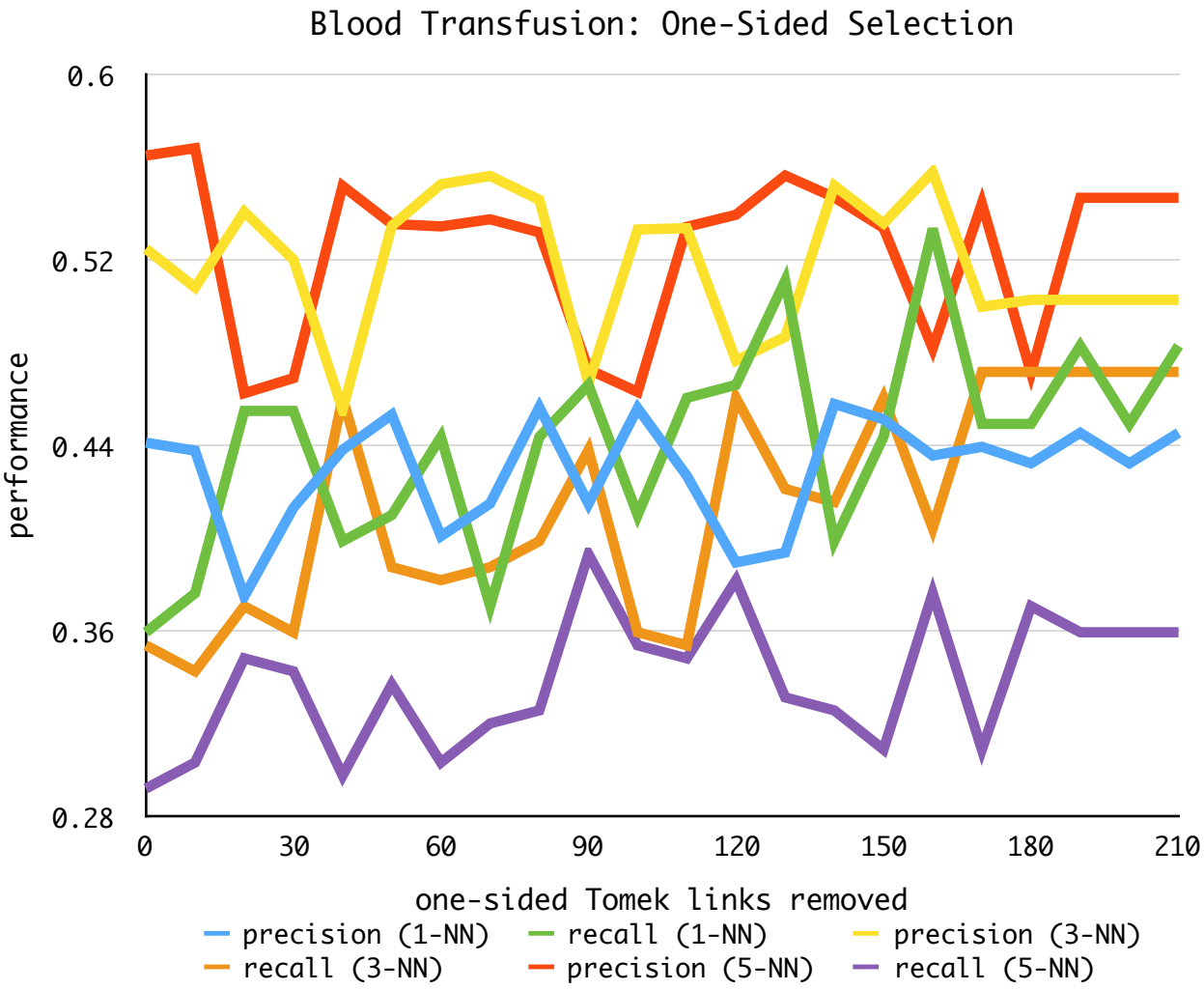
- 4 attributes: recency (months since last donation), frequency (total number of donations), monetary (total blood donated in c.c.), time (month since first donation)
- 2 classes: 1 if the patient donated blood in 2007, 0 otherwise
- 748 instances total with 178 (23.8%) of class 1 and 570 (76.2%) of class 0



Unlike the Haberman's Survival dataset, the Blood Transfusion dataset fared well under the majority-class under-sampling technique.

Most notable, all of the classifiers (1-NN, 3-NN, and 5-NN) converged to a 62% accuracy rate under precision and recall as the dataset went from a ratio of 3.2 instances of the majority class per instance of the minority class to a ratio of even parity. This involved, for instance, a doubling of accuracy from 30% to 62% in the case of 5-NN evaluated under recall.

This dataset transformation was particularly useful as the classifiers performed somewhat abysmally when evaluated using precision and recall. At the onset, only 3-NN and 5-NN evaluated under precision were correct 50% of the time; the other classifiers were in the 40 to 50% accuracy range when evaluated under precision or call. Majority-class under-sampling allowed 1-NN, 3-NN, and 5-NN to reach 62% accuracy under both metrics: a significant improvement.



Again, unlike the Haberman's Survival dataset, the Blood Transfusion dataset performed mediocrely after one-sided Tomek link removal.

Under all three classifiers, precision did not improve as the minority-class ends of a Tomek link were removed. Recall, however, did see a modest improvement under the dataset transformation. The 1-NN classifier saw a 13% increase, the 3-NN classifier a 12% increase, and the 5-NN classifier a 6% increase in recall capabilities.

Conclusion

As every engineer should know, no technique is a silver bullet. This was especially apparent in the cases we investigated.

The classifiers used on the Haberman Survival dataset saw an improvement in precision and recall under one-sided Tomek link removal but a massive deterioration in precision and recall under majority-class under-sampling. On the other hand, the classifiers of the Blood Transfusion dataset saw a great improvement in precision and recall when the dataset was transformed using majority-class under-sampling but only a modest improvement in recall when the dataset was transformed using one-sided Tomek link removal.

Hence, an engineer should test both approaches to see which is more appropriate for a dataset, as there is no way to necessarily know a priori.

If we were to repeat this experiment, we would explore the effect of these dataset transformations (majority-class under-sampling and one-sided Tomek link removal) on other classifiers, such as decision trees. This may prove to be an enlightening experience demonstrating different or otherwise unobserved capabilities of these techniques used to handle imbalanced domains.

Dataset Sources

https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival
http://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center