Threshold Selection for Web-Page Classification with Highly Skewed Class Distribution

Xiaofeng He, Lei Duan, Yiping Zhou, Byron Dom
Yahoo! Inc.
2811 Mission College Blvd.
Santa Clara, CA 95054 {xhe,leiduan,zhouy,bdom}@yahoo-inc.com

ABSTRACT
We propose a novel cost-efficient approach to threshold selection for binary web-page classification problems with imbalanced class distributions. In many binary-classification tasks the distribution of classes is highly skewed. In such problems, using uniform random sampling in constructing sample sets for threshold setting requires large sample sizes in order to include a statistically sufficient number of examples of the minority class. On the other hand, manually labeling examples is expensive and budgetary considerations require that the size of sample sets be limited. These conflicting requirements make threshold selection a challenging problem. Our method of sample-set construction is a novel approach based on stratified sampling, in which manually labeled examples are expanded to reflect the true class distribution of the web-page population. Our experimental results show that using false positive rate as the criterion for threshold setting results in lower-variance threshold estimates than using other widely used accuracy measures such as F1 and precision.

Categories and Subject Descriptors
H.4.m [Information Systems]: Miscellaneous

General Terms: Algorithms, Measurement

Keywords: Web-page classification, binary classifier, threshold selection, skewed class distribution, stratified sampling

1. INTRODUCTION
Web-page classification is an important process in the construction of web-search indexes. Classifiers provide necessary information for crawler, indexer and relevance ranking functions to select, index and rank high quality web search results. A spam classifier can help on filtering web spam in search results. A product review-page classifier can help on serving more focused product-review results for certain product-related queries. Many of these prediction tasks are instances of binary classification. Most classifiers work by first computing a score (often associated to the probability that the web page is a positive instance of the associated class) and then applying a threshold to the computed score. Part of the process of training such a classifier is that of determining the optimal threshold value.

The problems we face involve huge numbers of web pages and class distributions highly skewed toward the negative classes. In nearly all of the problems we address the fraction of positives is between 0.1% and 10%. Constructing a validation set via simple random sampling will likely result in too few positive cases unless a prohibitively large sample is acquired. This sparseness of positive cases results in a high degree of uncertainty in threshold values when they are estimated based on selection criteria such as precision, recall, F1 measure and their variants.

In this paper, we propose a novel method of threshold selection for skewed class distributions by leveraging a *stratified* sampling approach and using false positive rate (***fpr***) as the objective criterion. The sampled data are labeled by human editors and then expanded or weighted to reflect the true distribution in the population. Different threshold selection criteria are evaluated on both *semi-synthetic* and real data. Among these criteria, we found that false positive rate produced the most stable estimates. This criterion limits the classification errors at a pre-specified level, which is a desired property in applications such as spam filtering where false positives need be kept low. Space limitations preclude the inclusion here of all details, results and references. They are available in [1]. A survey and evaluation of threshold-selection methods can be found in [2].

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. SAMPLING AND SELECTION
To estimate the threshold for a classifier, we need to build a validation set by sampling web pages and labeling them manually. The procedures for constructing this validation set and computing threshold estimates based on it must address the imbalanced class distribution, while both producing threshold estimates of the required accuracy and keeping the manual-classification/labeling requirements to within budgetary constraints.

2.1 Sampling Techniques
To address the challenges outlined we use stratified random sampling techniques, which include proportional stratified random sampling (i.e. sampling same percentage of data from each stratum) and disproportional stratified random sampling (i.e. sampling same number of data form each stratum). In this work disproportional stratified sampling was chosen to address the highly skewed class distribution.

To generate strata, we first divided the population into constant-width bins by score. Then we apply the following heuristics: 1) merge adjacent bins until the margin of error for estimating the fraction of positives reaches an acceptable level; (2) merge adjacent bins with similar fractions of positives; (3) merge bins unlikely to contain the threshold (based on prior knowledge). The number to sample from each stratum was determined based on its estimated impact on threshold selection.

2.2 Sample Size
We use *sampling margin of error* as the criterion for strata sample-size determination. Assume N data points are randomly

sampled from a population. Let p_1 be the percentage of observed positives, and p_2 be the percentage of observed negatives. The standard error of the percentage can be calculated from standard deviation as: $SE = \sqrt{N} \times std = \sqrt{N} \times \sqrt{p_1 \times p_2}$. Then the margin of error of sampling with 95% confidence interval can be calculated as: $2 \times SE / N = 2 \times \sqrt{p_1 \times p_2 / N}$.

2.3 Sample Set Expansion

After web pages are sampled and labeled, the set is expanded so that its class distribution is close to the population distribution. We perform this expansion by duplicating the labeled pages, so that each stratum's page count in the sample set is proportional to the stratum's page count in the total population.

2.4 Threshold Selection Criteria

Most common classification accuracy/error metrics treat all cases equally and are therefore dominated by the most common class (negatives in our applications). This can be a problem for data with an imbalanced class distribution. The *Receiver Operating Characteristic* (ROC) graph does not depend on class distribution, but the precision-recall curve can exhibit a significant change with the change of class distribution. In our experiments 3 most promising threshold selection criteria are compared, including F_1 measure, precision and false positive ratio (***fpr***). In general, the choice of threshold selection criteria also depends on the *costs* of different types of errors (e.g. false positives versus false negatives) in the particular application.

3. EXPERIMENTAL ANALYSIS

We use semi-synthetic datasets in the experiments. First we randomly sample from Yahoo web database 2 sets of pages along with each page's 2 classification scores. The first set contains 1 million pages, and the second one contains 2 million pages. One score is from spam classifier and the other is from the soft-404 error-page classifier. All classification scores are from 0 to 255. To simulate the true labels for these pages, we create 10 score bins by dividing the score range equally. In each bin we randomly assign class labels to the contained pages with a probability estimated for this bin. The same process with 20 score bins is applied to both datasets, resulting in 4 semi-synthetic datasets for each classifier to be used as the population to sample from. Our simulation of true labels does not impact threshold selection, since thresholds and performance metrics depend on class distribution (i.e. the ratio of positive count to negative count) at each score point, instead of true class label for each data point.

To measure the sensitivity of our selection scheme to strata-boundary placement, we apply stratified sampling using 3 sets of strata boundaries for each of the 4 datasets. In the first experiment, we sample 1000 pages from each stratum. Assuming positive ratio is 5%, at 95% confidence level, we would get 1.378% margin of error with sample size 1000. In the second experiment, we sample a different number of pages from each stratum. Each stratum has 500 to 1800 pages in the sample, and 4000 pages in total are sampled. In the third experiment, we sample 4000 pages purely randomly. We repeat each experiment for 50 times, and calculate the mean and standard deviation of the selected thresholds based on F_1 measure, precision of 90% and ***fpr*** of 0.2%.

In the first experiment, we find thresholds selected based on the criterion of ***fpr*** have smaller variance and mean values closer to true thresholds compared to other criteria. An example of this can be seen in Figure 1, which shows that a significantly narrower distribution of threshold values is obtained using ***fpr*** as a criterion as opposed to precision. ***fpr***-based thresholds also vary less with different strata boundaries. The criterion of F_1 measure is the worst among all 3 criteria. In the second experiment, we have the similar observation as in the first experiment, except the standard deviations are consistently larger for all 3 criteria than in the first experiment. From the third experiment, we find all criteria generate thresholds with large variance and mean values considerably distant from true values compared to the first and second experiments.

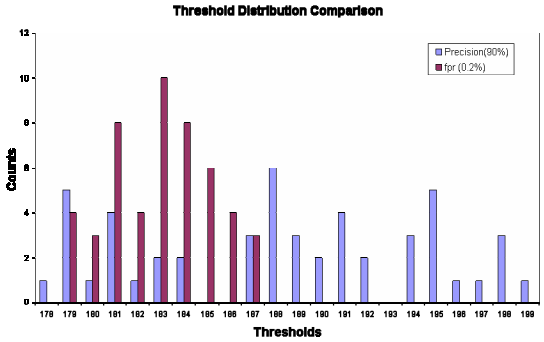

Figure 1: Threshold distribution for sample of 1 Million pages.

We learn from above results that 1) threshold selection based on ***fpr*** is less sensitive than the other 2 criteria to the means of randomly selecting samples from the strata; 2) threshold selection based on ***fpr*** is less sensitive to both the strata boundary placement and to sampling error than the other 2 criteria; 3) ***fpr*** is less sensitive to the amount of data sampled from each stratum than other 2 criteria; 4) pure random sampling is worse than the stratified random sampling approaches in the first two experiments.

4. CONCLUSIONS

We have described and demonstrated the effectiveness of a novel technique for threshold selection in binary classification problems with highly imbalanced class distributions. Use of this technique enables the optimization of the trade-off between classification accuracy obtained and the number of manual judgments required. The technique is based on a stratified sampling scheme designed specifically for threshold selection. As part of this work we have also determined that using ***fpr*** as the criterion for threshold selection results in the most stable estimates.

REFERENCES

[1] X. He, L. Duan, Y. Zhou and B. Dom, Threshold selection for web-page classification with highly skewed class distribution, *Yahoo! Labs Research Report YL-2009-001*, 2009

[2] Y. Yang, A Study on Thresholding Strategies for Text Categorization, Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval, 2001