BROOF: Exploiting Out-of-Bag Errors, Boosting and Random Forests for Effective Automated Classification Online Appendix [*]

Thiago Salles Marcos Gonçalves Victor Rodrigues
Fed. Univ. of Minas Gerais
Computer Science Department
Belo Horizonte, Brazil
{tsalles, mgoncalv, victor.rodrigues}@dcc.ufmg.br

Leonardo Rocha
Fed. Univ. of São João Del-Rei
Computer Science Department
São João Del-Rei, Brazil
lcrocha@ufsj.edu.br

1. EXPLORED DATASETS—DETAILS

Due to the blind-review process, we temporarily made available here the online appendix. After the review process, we shall move this appendix to a definitive institutional web page.

In the following, we detail all the datasets explored in this work.

1.1 Topic Categorization

In order to evaluate BROOF under the topic categorization setting, we explored the following datasets:

20 Newsgroups (20NG) a classical textual dataset with roughly $20,000$ labeled documents gathered from newsgroups. Each document is classified into one of 20 categories. Each category has approximately $1,000$ examples.

4 Universities (4UNI) (aka WEBKB) this dataset contains Web pages collected from Computer Science departments of four universities by the Carnegie Mellon University (CMU) text learning group. There is a total of $8,277$ web pages, classified into 7 categories (such as student, faculty, course and project web pages).

Reuters (REUT) this is a classical text collection, composed by news articles collected and annotated by Carnegie Group, Inc. and Reuters, Ltd. We consider here a set of $13,327$ articles, classified into 90 categories.

ACM-DL (ACM) a subset of the ACM Digital Library with $24,897$ documents containing articles related to Computer Science. We considered only the first level of the taxonomy adopted by ACM, whereas each document is assigned to one of 11 classes.

MEDLINE (ML) a subset of the MedLine dataset, with $861,454$ documents classified into 7 distinct classes related to Medicine. This collection was obtained from [3]. In that work the authors considered the first level of the taxonomy so that each document article is classified under only one category, avoiding dealing with multilabel cases.

UniRCV1. The Reuters Corpus Volume 1 (RCV1) is a dataset with $804,427$ English language news stories. We considered the complete *topics* taxonomy comprised of 103 classes. However, as a multi-label dataset, the multi-label cases need special treatment, such as score thresholding, etc. (see [2] for details), in order to be properly consumed by unilabel classifiers. As our current focus is on unilabel tasks, to allow a fair comparison among the other datasets (which are also unilabel) and all baselines (which also focus on unilabel tasks), we decided to remove the documents assigned to more than one class from RCV1, deriving a new dataset which we call *UniRCV1*. This collection has 101 classes and about 20% less documents. Nevertheless, as we shall see, the effectiveness levels obtained by our method and the best baselines are still compatible with those of the original multilabel RCV1.

The details regarding each topic categorization dataset (size, number of features and class distribution) can be found in Table 1.

1.2 Sentiment Analysis

In order to evaluate BROOF under the sentiment analysis setting, we considered twelve datasets of messages labeled as positive and negative from many domains, including messages from social networks, movie and product reviews, opinions and comments in news articles. The explored datasets are:

Amazon consists of a set of product reviews form *amazon.com.*

BBC a set of messages from comments in the BBC and Runners World forum from SentiStrength research [4].

Debate consists of tweets about the 2008 U.S. Presidential debate.

Digg user provided comments on web content aggregated in *digg.com.*

MySpace a set of messages crawled from the Myspace network, used in SentiStrength research.

NYT includes sentence-level snippets from a set of New York Times opinion editorials.

Tweets a set of tweets from VADER work [1] which were crawled from Twitter's public timeline (with varied times and days of posting).

Twitter this dataset consists of human labeled messages used in the SentiStrength research.

[*]This work was partially supported by CNPq, CAPES, FINEP, FAPEMIG and INWEB.

SIGIR'15, August 09–13, 2015, Santiago, Chile
Copyright 2015 ACM 978-1-4503-3621-5/15/08...$15.00.
http://dx.doi.org/10.1145/2766462.2767747.

Dataset	Size	# Features	Class Distribution						
			# Classes	Minor Class	1^o Quartile	Median	Mean	3^o Quartile	Major Class
20 Newsgroups (Newsgroups)	18805	61050	20	628	955	979.5	940.2	990	999
4 Universities (Web)	8277	40195	7	137	343	930	1182	1382	3759
Reuters (News)	13327	19590	90	2	8	29	148.1	91	3964
ACM-DL (Computer Science)	24897	56499	11	63	761	2041	2263	3278	6562
UniRCV1 (News)	652909	46120	101	3	401	1656	6464	6725	62943
MEDLINE (Medicine)	861454	268783	7	1843	36196	44089	123065	143568	455994

Table 1: Statistics Summary for each Reference Dataset.

Yelp consists of a set of business and services reviews from the greater Phoenix, AZ metropolitan area.

Youtube a set of user provided comments on video content.

The details regarding each sentiment dataset (size, number of features and class distribution) can be found in Table 2.

1.3 Microarray Analysis

In order to validate the effectiveness of BROOF in microarray analysis tasks, we consider here six microarray gene expression datasets for the task of predicting the presence of specific cancer types or the ausence of cancer.

9tumors this dataset consists of samples regarding nine human tumor types.

Brain1 this dataset consists of gene expression microarray data regarding five human brain tumor types.

Brain2 a set of samples refering to four malignant glioma types.

DLBCL a set of samples with gene expression information regarding diffuse large b-cell lymphomas (DLBCL) and follicular lymphomas.

Leukemia gene expression profiles characterizing AML, ALL, and mixed-lineage leukemia (MLL).

Prostate samples consisting of prostate tumor and normal tissues.

The details regarding each microarray dataset (size, number of features and class distribution) can be found in Table 3.

References

[1] C. J. Hutto and E. Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. Oh, editors, *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press, 2014.

[2] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *JMLR.*, 5:361–397, 2004.

[3] L. Rocha, F. Mourão, A. Pereira, M. A. Gonçalves, and W. Meira Jr. Exploiting temporal contexts in text classification. In *Proc. CIKM*, pages 243–252, 2008.

[4] M. Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength 1, 2013.

Dataset	Size	# Features	Class Distribution		
			# Classes	Minor Class	Major Class
Amazon	1237	2347	2	617	620
BBC	729	6861	2	93	636
Debate	1487	2926	2	740	747
Digg	775	3236	2	206	569
MySpace	825	2703	2	131	694
NYT	1237	5340	2	616	621
Tweets	1248	3638	2	623	625
Twitter	2272	8330	2	938	1334
Yelp	4999	24508	2	2499	2500
Youtube	2396	7278	2	756	1640

Table 2: Statistics Summary for each Reference Dataset.

Dataset	Size	# Features	Class Distribution						
			# Classes	Minor Class	1^o Quartile	Median	Mean	3^o Quartile	Major Class
9tumors	60	5726	9	2.00	6.00	7.00	6.67	8.00	9.00
Brain1	90	5920	5	4.00	6.00	10.00	18.00	10.00	60.00
Brain2	50	10367	4	7.00	12.25	14.00	12.50	14.25	15.00
DLBCL	77	5469	2	19.00	28.75	38.50	38.50	48.25	58.00
Leukemia	72	11225	3	20.00	22.00	24.00	24.00	26.00	28.00
Prostate	102	10509	2	50.00	50.50	51.00	51.00	51.50	52.00

Table 3: Statistics Summary for each Reference Dataset.