ECML-PKDD Discovery Challenge 2006 Overview

Steffen Bickel

Humboldt-Universität zu Berlin, School of Computer Science
Unter den Linden 6, 10099 Berlin, Germany
bickel@informatik.hu-berlin.de

Abstract. The Discovery Challenge 2006 deals with personalized spam filtering and generalization across related learning tasks. In this overview of the challenge we motivate and describe the problem setting and the evaluation measure. We give details on the construction of the data sets and discuss the results.

1 Introduction

The Discovery Challenge 2006 is about personalized spam filtering and generalization across related learning tasks. People spend an increasing amount of time for reading messages and deciding whether they are spam or non-spam. Some users spend additional time to label their received spam messages for training local spam filters running on their desktop machines. Email service providers want to relieve users from this burden by installing server-based spam filters. Training such filters cannot rely on labeled messages from the individual users, but on publicly available sources, such as newsgroup messages or emails received through "spam traps" (spam traps are email addresses published visually invisible for humans but get collected by the web crawlers of spammers).

This combined source of training data is different from the distributions of the emails received by individual users. When learning spam filters for individual users from this type of data one needs to cope with a discrepancy between the distributions governing training and test data and one needs a balance between generalization and adaptation. The generalization/adaptation can rely on large amounts of unlabeled emails in the user's inboxes that are accessible for server-based spam filters. Utilizing this unlabeled data a spam filter can be adapted to the properties of specific user's inboxes but when little unlabeled data for a user are available a generalization over multiple users is advised.

The Discovery Challenge 2006 covers this setting, labeled training data collected from publicly available sources are provided. The unlabeled inboxes of several users serve as test data. The inboxes differ in the distribution of emails. The goal is to construct a spam filter for each single user that correctly classifies its emails as spam or non-spam. A clever way of utilizing the available sets of unlabeled emails from different users is required.

This overview is organized as follows. In Section 2, we discuss the problem setting and define the evaluation measure. We describe the data sets in Section

3. Section 4 gives an overview of the participants and summarizes the results. In Section 5 we discuss the different approaches and Section 6 concludes.

2 Problem Setting and Evaluation Measure

In the problem setting of the challenge the inboxes of several users are given and the goal is to correctly classify the messages in each inbox as spam or non-spam. No labeled training examples from the inboxes are available, instead, one common set of labeled data is given. The labeled data and the inboxes are governed by different distributions. A learning algorithm cannot rely only on the labeled data because the bias between training data and inboxes hinders learning of a correct classification model for the inboxes. The unlabeled data in the inboxes need to be used to adapt to their distributions.

The individual distributions of the inboxes are neither independent (identical spam messages are sent to many users), nor are they likely to be identical: distributions of inbound messages vary greatly between (professional, recreational, American, Chinese, . . .) email users. A learning algorithm can exploit the similarity of the inboxes.

There are two different tasks that differ in the number of inboxes and the proportion of labeled to unlabeled data (see Section 3).

Usually, cross-validation is used for tuning parameters of a classification model. In our case, cross-validation cannot be used because the emails in the inboxes are unlabeled. We provide a second set of labeled training data and inboxes for parameter tuning. The difference between the tuning set and the evaluation set is that the emails in the inboxes of the tuning set are labeled. The feature representation of the tuning data differs from the evaluation data (different dictionary). This means, the tuning data can not be used to augment the training data.

The problem setting differs from the standard setting of semi-supervised learning in three ways,

- there is a bias between training and evaluation data, the training and test data are governed by different distributions,
- several distinct but similar unlabeled inboxes are given, a multi-task learning or a transfer learning approach can be used for modeling and exploiting the similarity between inboxes,
- the number of labeled emails is larger than the number of unlabeled examples for a single inbox (task A).

The evaluation criterion for the challenge is the AUC value. The AUC value is the area under the ROC curve (Receiver Operating Characteristic curve). A ROC curve is a plot of true positive rate vs. false positive rate as the prediction threshold sweeps through all the possible values. The area under this curve has the nice property that it specifies the probability that, when we draw one positive and one negative example at random, the decision function assigns a higher value to the positive than to the negative example.

We compute AUC values for each inbox separately and average over all inboxes of the task. The winner for each task is the participant with the highest average AUC value. There is an additional creativity award for each task for the most interesting solutions in terms of non-straightforward approaches, innovative ideas, and assumed high impact.

3 Data Sets

The composition of the labeled training set is the same for both tasks, they differ in number of emails. 50% of the labeled training data contain spam emails sent by blacklisted servers of the Spamhaus project (www.spamhaus.org). 40% are non-spam emails from the SpamAssassin corpus and 10% are non-spam emails sent from about 100 different subscribed English and German newsletters. Table 1 summarizes the composition of the labeled training data for both tasks. The labeled data of the tuning set has the same size and composition as the actual training data but with different emails.

	task A	task B
emails sent from blacklisted servers	2000	50
SpamAssassin emails	1600	40
newsletters	400	10
total	4000	100

Table 1. Composition of labeled training data.

Evaluating the filters with respect to the personal distributions of messages requires labeled emails from distinct users. We construct different inboxes using real but disclosed messages. As non-spam part of the inboxes we use messages received by distinct Enron employees from the Enron corpus [9] cleaned from spam. Each inbox is augmented with spam messages from distinct spam sources. Some spam sources are used for multiple inboxes, in those cases all available emails from this source were sorted by date and split into different consecutive subsets. Because of the topic drift the distribution of the emails in the different parts differs.

The two tasks differ in the number and size of inboxes, task A has 3 and task B 15 evaluation inboxes. The size of the inboxes in task A is 2500 and in task B 400. Tables 2 and 3 summarize the composition of the evaluation and the tuning inboxes for task A and B. Each inbox consists of 50% spam and 50% non-spam emails.

The messages are preprocessed and transformed into a bag-of-words representation. We provide feature vectors with term frequencies. Our preprocessing uses charset-, MIME-, base64-, URL- (RFC 1738), and subject line-decoding (RFC 2047). Our tokenization takes care of HTML tags, following the X-tokenizer proposed by Siefkes et al. [3].

inbox ID	evaluation/ tuning	non-spam/ Enron user	spam source
0	eval	Farmer	Dornbos spam trap, part 1 (www.dornbos.com)
1	eval	Lokay	Dornbos spam trap, part 2 (www.dornbos.com)
2	eval	Sanders	spam trap of Bruce Guenter, part 1 (www.em.ca/∼bruceg/spam)
3	eval	Bass	personal spam of Richard Jones, part 1 (www.annexia.org/spam)
4	eval	Campbell	personal spam of Tobias Scheffer, part 1
5	eval	Dasovich	spam collection of SpamArchive.org, part 1
6	eval	Germany	spam collection of SpamArchive.org, part 2
7	eval	Kean	personal spam of Paul Wouters, part 1 (www.xtdnet.nl/paul/spam)
8	eval	Mann	Dornbos spam trap, part 3 (www.dornbos.com)
9	eval	Nemec	Dornbos spam trap, part 4 (www.dornbos.com)
10	eval	Rogers	spam trap of Bruce Guenter, part 2 (www.em.ca/∼bruceg/spam)
11	eval	Scott	spam trap of Bruce Guenter, part 3 (www.em.ca/∼bruceg/spam)
12	eval	Shackleton	personal spam of Richard Jones, part 2 (www.annexia.org/spam)
13	eval	Shapiro	personal spam of Tobias Scheffer, part 2
14	eval	Symes	spam collection of SpamArchive.org, part 3
0	tune	Lay	personal spam of Paul Wouters, part 2 (www.xtdnet.nl/paul/spam)
1	tune	Taylor	spam trap of Bruce Guenter, part 4 (www.em.ca/∼bruceg/spam)

Table 2. Composition of the evaluation and tuning inboxes for task A.

4 Participation and Results

57 teams from 19 different countries participated in the challenge. 26 participants submitted their results for evaluation, 20 teams have an academic and 6 teams a commercial background. Not all teams submitted results for both tasks. We averaged the AUC values for all inboxes as described above and determined the ranking. We conducted significance tests using a significance level of 5% to test the null hypothesis that the second rank has a higher AUC value than the first. The test statistic is computed as described in Hanley and McNeil [7]. For task A

inbox ID	evaluation/ tuning	non-spam/ Enron user	spam source
0	eval	Beck	spam trap of Bruce Guenter (www.em.ca/∼bruceg/spam)
1	eval	Kaminski	spam collection of SpamArchive.org
2	eval	Kitchen	personal spam of Tobias Scheffer (www.em.ca/∼bruceg/spam)
0	tune	Williams	Dornbos spam trap, part 3 (www.dornbos.com)

Table 3. Composition of the evaluation and tuning inboxes for task B.

we could not reject the null hypothesis for rank two and three, this means there is no statistically significant difference between them and they are all ranked first. For task B we could reject the null hypothesis for the second rank, this means there is one winner.

Table 4 and 5 show the first five ranks for task A and task B, respectively. Figure 1 displays the distribution of AUC over all ranks. Some participants report higher results in their workshop paper because they improved their algorithms after the submission deadline.

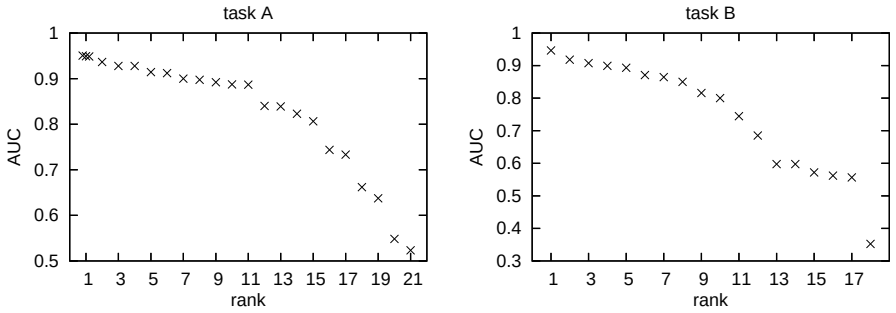

Fig. 1. Distribution of AUC performance dependent on rank over all participants for task A (left) and task B (right).

We selected the solution of Bernhard Pfahringer (University of Waikato, New Zealand) for the Spam Filtering Creativity Award - task A/B, we decided to award one team for both tasks instead of one for each task because most teams used the same algorithm for both tasks. Details on his algorithm are given in the next section.

5 Discussion

The teams approached the problem in very different ways but most of the participants used variants of semi-supervised learning techniques. Among the semi-supervised algorithms were graph-based algorithms [2] , large-margin-based

rank	avg. AUC	team
1	0.9507	Khurram Junejo, Mirza Yousaf, Asim Karim *Lahore University of Management Sciences, Pakistan*
1	0.9491	Bernhard Pfahringer *University of Waikato, New Zealand*
1	0.9487	Kushagra Gupta, Vikrant Chaudhary, Nikhil Marwah, Chirag Taneja *Inductis India Pvt Ltd*
2	0.9365	Nikolaos Trogkanis *National Technical University of Athens, Greece* Georgios Paliouras *National Center of Scientific Research "Demokritos" Greece*
3	0.9278	Chao Xu, Yiming Zhou *School of Computer Science and Engineering,* *Beijing University, China*
4	0.9277	Lalit Wangikar, Mansi Khanna, Ankush Talwar Nikhil Marwah, Chirag Taneja *Inductis India Pvt Ltd*
5	0.9144	Dimitrios Mavroeidis, Konstantinos Chaidos, Stefanos Pirillos, Dimosthenis Christopoulos, Michalis Vazirgiannis *DB-NET Lab, Informatics Dept., Athens University EB, Greece*

Table 4. First five ranks for task A.

rank	avg. AUC	team
1	0.9465	Gordon Cormack *University of Waterloo, Canada*
2	0.9183	Nikolaos Trogkanis *National Technical University of Athens, Greece* Georgios Paliouras *National Center of Scientific Research "Demokritos" Greece*
3	0.9074	Kushagra Gupta, Vikrant Chaudhary, Nikhil Marwah, Chirag Taneja *Inductis India Pvt Ltd*
4	0.8992	Dyakonov Alexander *Moscow State University, Russia*
5	0.8933	Wenyuan Dai *Apex Data & Knowledge Management Lab,* *Shanghai Jiao Tong University*

Table 5. First five ranks for task B.

methods [1, 4, 10], self-training approaches [6, 8], positive-only learning [10], and multi-view learning methods [4]. The assumption in most of those algorithms is that the unlabeled data is drawn from the same distribution as the labeled data. This assumption is violated in our case, but nevertheless semi-supervised learning reduces the error compared to methods that do not utilize the unlabeled data.

Bernhard Pfahringer the winner of the creativity award accounts for the bias between training and evaluation data in two ways [2]. Firstly, whenever a pre-

diction for some evaluation email is needed, his algorithm transforms the whole training set by only selecting those features which are actually present in the evaluation email (i.e. have a non-zero value). A classification model is trained using this transformed training set and that model's prediction is used for the evaluation example in question. This procedure forces the learner to concentrate on the features that are actually present in the evaluation example. This idea of filtering non-existent features is similar to the approach of Steinberg and Golovnya [11]. Secondly, Pfahringer uses a learning algorithm by Zhou et al. [5] that is one of the best known graph-based semi-supervised learning algorithms. The algorithm of Zhou et al. originally suffers from a cubic runtime complexity in the number of examples. Pfahringer develops a variant of this algorithm with linear complexity. The tremendous reduction in runtime and memory requirements make the algorithm applicable for large data sets.

Trogkanis and Paliouras [10], ranked second in both tasks, are very cautious when transferring knowledge from labeled to biased unlabeled data. Their approach is almost unsupervised. A classifier trained an the labeled data is allowed to label only a very few unlabeled emails with high confidence. In the subsequent step the labeled data is ignored and a semi-supervised algorithm is applied only to the inbox emails.

Two teams developed models that account for the similarity of inboxes with transfer learning. Participant Mohammad Al-Hasan (Rensselaer Polytechnic Institute) first measures the pairwise cosine similarity of all emails between all inboxes. In a second step a self-training-like learning algorithm learns separate classifiers for all inboxes in parallel. In each self-training iteration the most confident previously unlabeled email for each inbox is labeled together with the most similar email from one other inbox. With this approach confident decisions from one inbox are transfered to other inboxes. Trogkanis and Paliouras [10] use semi-supervised learning and augment the unlabeled data of one inbox by a weighted set of the unlabeled emails of all other inboxes. Gordon Cormack, ranked first in task B, even ignores the separation of emails into inboxes and pools all inboxes into one unlabeled set for semi-supervised training.

6 Conclusion

Most of the participants obtained lower classification errors by utilizing the data from the unlabeled inboxes in addition to the labeled data. Those results indicate that server-sided spam-filtering can be improved by personalization using unlabeled inboxes.

The results of the participants show that a wide range of semi-supervised learning algorithms can improve the classification performance for the problem setting of the challenge. Most semi-supervised learning algorithms make the implicit assumption that the training and test data are drawn from the same distribution. This assumption is violated in our case. It is an open problem to develop semi-supervised learning methods that account for a bias between

training and test data. We assume that such methods could further improve the benefit of spam filter personalization.

Some participants used transfer learning to account for the similarity between the inboxes. In their experiments the knowledge transfer between the inboxes improved the classification performance. Algorithms that did not exploit the similarity between the inboxes but used more sophisticated semi-supervised methods received higher scores in the overall ranking. This raises the question whether the best semi-supervised approaches can be integrated into a transfer or multi-task learning framework and whether this further improves the classification performance.

To our knowledge there is no spam filtering software for practical settings that utilizes unlabeled examples within a learning framework, also, personalization in server-sided spam filtering algorithms is widely disregarded. In this respect the results of the Discovery Challenge are encouraging. A real application in a server-sided setting poses additional challenges regarding the scalability of the methods.

Acknowledgment

The Discovery Challenge 2006 has been supported by Strato Rechenzentrum AG and by the German Science Foundation DFG under grant SCHE540/10-2.

References

1. Kyriakopoulou A. and Kalamboukis T. Text classification using clustering. In *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.
2. Pfahringer B. A semi-supervised spam mail detector. In *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.
3. Siefkes C., Assis F., Chhabra S., and Yerazunis W. Combining winnow and orthogonal sparse bigrams for incremental spam filtering. In *Proceedings of the European Conference on Principle and Practice of Knowledge Discovery in Databases*, 2004.
4. Mavroeidis D., Chaidos K., Pirillos S., Christopoulos D., and Vazirgiannis M. Using tri-training and support vector machines for addressing the ecml-pkdd 2006 discovery challenge. In *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.
5. Zhou D., O. Bousquet, Lal T., Weston J., and Schölkopf B. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 2003.
6. Cormack G. Harnessing unlabeled examples through iterative application of dynamic markov modeling. In *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.
7. Hanley J. and McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843, 1983.
8. Junejo K., Yousaf M., and Karim A. A two-pass statistical approach for automatic personalized spam filtering. In *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.

9. B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning*, 2004.
10. Trogkanis N. and Paliouras G. TPN^2: Using positive-only learning to deal with the heterogeneity of labeled and unlabeled data. In *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.
11. Dan Steinberg and Mikhaylo Golovnya. Identifying spam with predictive models. In *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.