Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking

Chanju Kim and Kyu-Baek Hwang

School of Computing, Soongsil University, Seoul 156-743, Korea
cjkim@ml.ssu.ac.kr, kbhwang@ssu.ac.kr

Abstract. Social bookmarking systems such as BibSonomy and del.icio.us have become increasingly popular with the prevalent use of internet. These systems provide powerful infrastructure solutions for semantic annotation and information sharing, promoting diverse kinds of internet-based activities, e.g., web exploration, creating and joining web-based communities, and buying recently published volumes. This usefulness is also gaining the attention of malicious users a.k.a. spammers. For instance, these spammers abuse social bookmarking systems for biasing web search results and advertising improperly. Manual spam detection is not scalable due to the vast amount of related information. In this paper, we propose a machine learning-based approach to automatic spam detection. In specific, a set of relevant features, i.e., the number of posts and posted tags for each user are extracted from training data. The extracted tags are sorted by mutual information. Then, the tags, having high mutual information value and used in test data, are chosen for the classification task. In our experiments, naive Bayes classifiers with varying numbers of selected features were learned from a subset of the given training dataset and evaluated on a separate validation set for finding the optimal parameter setting. Finally, the learned results from the entire training dataset with the best setting was applied to the real test dataset for the challenge.

1 Introduction

The performance of social bookmarking systems [4] can be severely degraded by malicious users, i.e., spammers. The spammers post irrelevant and misleading information for their private benefit. The irrelevant and misleading information in public repositories not only makes the repositories untrustful but also wasting their resources in processing the vast amount of unnecessary information. Thus, it is crucial to discriminate spamming from proper posting for the success of social bookmarking systems.

Several challenges exist in this spam filtering problem. One is the enormous amount of data. The number of users of a social bookmarking system usually amounts to several tens of thousands. Also, the number of posts could amount to several millions. To make matters worse, the given data for spam filtering could be highly skewed. For instance, the ratio between spammers and active users is about one to twelve in the given training dataset for the first task of ECML PKDD Discovery Challenge 2008 (`http://www.kde.cs.uni-kassel.de/ws/rsdc08/`). Another difficulty arises from the fact that the characteristics of the spam filtering data are gradually changing as time goes by. In other words, the distribution of a feature variable might be largely different according to the time period over which it is estimated.

We addressed the above problems using naive Bayes classifiers learning with an enhanced feature selection method. More specifically, tags for spam detection are chosen based on their mutual information value as well as their usage in test period. In our experiments, the suggested feature selection method was shown to generally outperform the conventional feature selection method solely based on mutual information.

The paper is organized as follows. In Section 2, we describe the given task and explain the proposed method for tackling the problem. The experimental results for parameter setting and the performance of the proposed approach on the test dataset is given in Section 3. Finally, conclusions are drawn in Section 4.

2 The Method

The given task is to discriminate spammers from active-users (non-spammers) based on their posting information such as the tags, the dates, the urls, the descriptions, and the related information to the published volumes. The training dataset consists of 31,715 users (including both spammers and active users) and was gathered during the period from January 1989 through March 2008. Among the 31,715 users, 2,467 are active users and the others are spammers. The dataset is comprised of seven tables, i.e., tas, tas_spam, bookmark, bookmark_spam, bibtex, bibtex_spam, and user.

We formulated the given task as a supervised learning problem in which each user corresponds to a data example and the target variable denotes whether the user is spammer or not. As feature variables, the number of bookmark postings, the number of bibtex postings, and the tags were deployed. Here, the tag variable denotes whether a user have ever posted a specific tag or not.

Because there exist a tremendous amount of posted tags (more than 425,000), an appropriate number of tags should be selected for avoiding the overfitting problem and reducing the computational cost. We harnessed the mutual information [1] for tag selection. The mutual information between a tag and the target variable is calculated as follows.

$$I(Tag_i; Target) = \sum_{tag_i, target} \hat{P}(Tag_i, Target) \log \frac{\hat{P}(Tag_i, Target)}{\hat{P}(Tag_i) \cdot \hat{P}(Target)}, \qquad (1)$$

where Tag_i is a binary variable denoting whether the ith tag is used or not, $Target$ corresponds to the target variable, and $\hat{P}(\cdot)$ denotes the probability value estimated from a given dataset. Here, the summation is taken over all possible configurations of the two variables, Tag_i and $Target$.

As a classifer, the naive Bayes classifier [5] was adopted because of its computational efficiency as well as its optimality for classification tasks even when the conditional independence assumption is invalid [2]. Actually, we have also tried other famous classification methods including artificial neural networks, support vector machines, tree-augmented naive Bayes classifiers, and decision trees. Their performance in our problem setting was much worse than that of the naive Bayes classifier although the experimental results are not shown here. The naive Bayes classifier in our problem setting is simply formulated by the following equation.

$$\begin{aligned} P(Target|F_1, F_2, ..., F_n) &= \frac{P(Target) \cdot P(F_1, F_2, ..., F_n|Target)}{P(F_1, F_2, ..., F_n)} \\ &= \frac{P(Target) \cdot P(F_1|Target) \cdot P(F_2|Target) \cdot ... \cdot P(F_n|Target)}{P(F_1, F_2, ..., F_n)}, \end{aligned} \quad (2)$$

where F_i corresponds to the ith feature variable. The feature variables include Tag_i defined as in Equation (1), the number of bookmark postings, and the number of bibtex postings. Tag_i's are binary. Other two feature variables were discretized by the supervised discretization method of Weka [6]. The method is based on the approach proposed by [3].

One of the challenges in spam detection lies in the fact that the tag usage pattern is continuously changing. For example, some tags chosen from a training dataset by mutual information might not exist in a separate test dataset. In this case, such tags cannot tell a test example is spammer or not because they have never been used in the test dataset. To mitigate this problem, we propose an enhanced feature selection method, considering both the mutual information and whether the tag is used in the test period as follows.

1. Remove the tags which do not exist in the test dataset.

2. Select a pre-specified number of tags from the remaining tags according to their mutual information values.

3 Experimental Evaluation

To evaluate the proposed approach and find the optimal parameter value[1], we reserved the data examples from the latest two months from the given training dataset. Hence, the training period is from January 1989 to January 2008 and the validation period is from February to March of the same year. The numbers of postings during these periods are shown in Table 1. The numbers of spammers and active users[2] during the same periods are shown in Table 2.

In order to empirically find the optimal number of tags for classification, we experimented with varying numbers of selected tags from 100 to 3,000. We also

[1] Here, the parameter value denotes the number of tags.

[2] It should be noted here that some users exist in both the training dataset and the validation dataset.

Table 1. The number of postings in the given training dataset.

	Non-spam	Spam	Total
Training period	260,271	1,264,539	1,524,820
Validation period	8,421	362,266	370,687
Total	268,692	1,626,805	1,895,497

Table 2. The number of users in the given training dataset.

	Active users	Spammers	Total
Training period	2,466	29,248	31,714
Validation period	656	10,610	11,266

compared the conventional mutual information-based feature selection method with the proposed one. The experimental results are shown in Fig. 1.

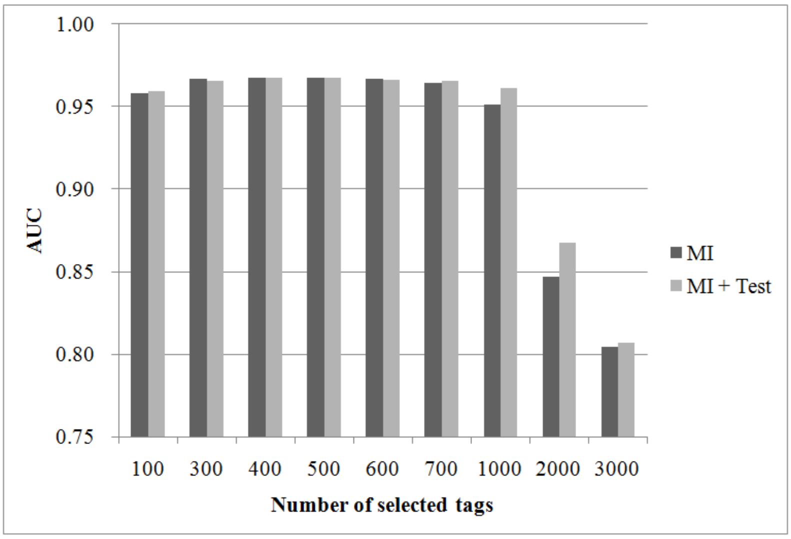

Fig. 1. Comparison of the feature selection methods with varying numbers of selected tags. MI: the conventional method. MI + Test: the proposed method.

From the results, we can observe that the proposed method improves the performance of naive Bayes classifiers in general. Also, the classification per-

Table 3. Comparison of the feature selection methods when the number of selected tags is less than 1,000. MI: the conventional method. MI + Test: the proposed method. Here, the performance is measured by the AUC.

# of Tags	100	300	400	500	600	700
MI	0.95793	**0.96709**	**0.96771**	0.96769	**0.96689**	0.96445
MI + Test	**0.95911**	0.96580	0.96739	**0.96775**	0.96598	**0.96590**

formance decreases as the number of selected tags exceeds 1,000. One possible explanation for this phenomenon is that the large number of selected tags causes the overfitting problem. In fact, the number of extreme prediction values, i.e., zero and one, increases as the number of tags grows.

The detailed comparison of the feature selection methods when the number of selected tags is less than 1,000 is given in Table 3. In this case, the classification performance obtained by the proposed feature selection method is similar to that by the conventional one. We conjecture that this is because the tags with very high mutual information are not so much different in our training and validation datasets.

We applied our spam detection method to predicting spammers in the final test dataset. The final results are shown in Fig. 2.

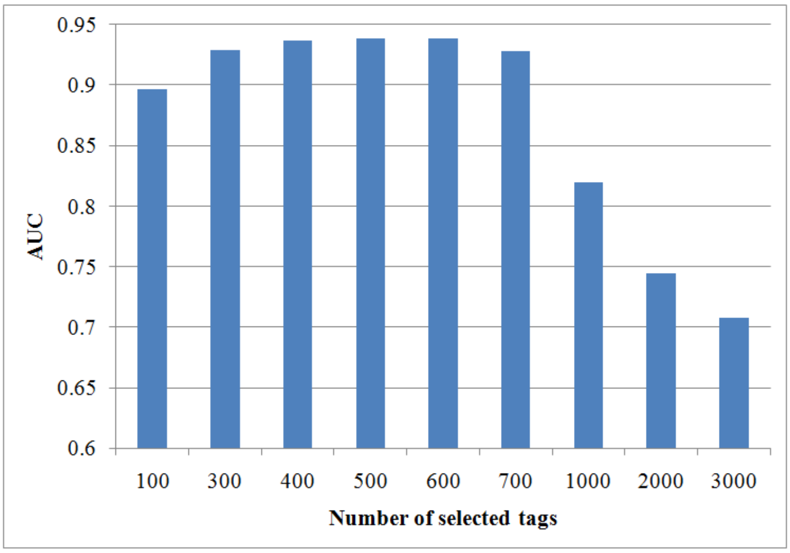

Fig. 2. Classification performance of the proposed method on the real test dataset with varying numbers of selected tags.

Interestingly, the optimal number of selected tags from Table 3 and Fig. 2 is the same. In both cases, the best classification performance was obtained by considering 500 tags. The best result in Fig. 2 is 0.93899.

4 Conclusions

We described a machine learning-based approach for spam detection. As a classifier, the naive Bayes classifier was employed because of its simplicity and efficiency. The number of bookmark postings, the number of bibtex postings, and the tags were considered as feature variables for the classification. For the tag selection, mutual information as well as the term's usage in test period were taken into account. The proposed feature selection method was shown to outperform the conventional mutual information-based approach in general. The number of selected tags was empirically optimized through the validation experiments. Through the experiments on the final test dataset, we have shown that our empirical choice of the optimal number of selected tags from the training dataset was meaningful. One of the directions for future work would be to combine the interrelationship between tags into our approach for more enhanced classification performance.

Acknowledgements

This work was supported in part by the Seoul Development Institute through Seoul R&BD Program (GS070167C093111) and in part by the Ministry of Knowledge Economy of Korea through Ubiquitous Computing and Network (UCN) Project (21st Century Frontier R&D Program).

References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience (1991)
2. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29(2/3), 103–130 (1997)
3. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027. (1993)
4. Heymann, P., Koutrika, G., Garcia-Molina, H.: Can social bookmarking improve web search? In: First ACM International Conference on Web Search and Data Mining. (2008)
5. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: Tenth National Conference on Artificial Intelligence, pp. 223–228. (1992)
6. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann, San Francisco (2005)