Report for WebSpam 2008 Classification Modeling

Yuchun Tang
Secure Computing Corporation
4800 North Point Parkway, Suite 300
Alpharetta, GA 30022, USA

Yuanchen He
Secure Computing Corporation
4800 North Point Parkway, Suite 300
Alpharetta, GA 30022, USA

Sven Krasser
Secure Computing Corporation
4800 North Point Parkway, Suite 300
Alpharetta, GA 30022, USA

{ytang, yhe, skrasser}@securecomputing.com

ABSTRACT

We apply Random Forests classification algorithm on WebSpam 2008 data. Under 7-fold cross validation, the highest Area under ROC value is 0.82045.

Categories and Subject Descriptors

I.5.m [**Computing Methodologies**]: Pattern Recognition—*Miscellaneous*

General Terms

Webspam 2008, Random Forests

1. EXPERIMENTS

1.1 Feature Extraction

Organizers provide four groups of pre-computed features, which are obvious features, link based features, transformed link based features and content based features. We combine them together before classification modeling. There are altogether 277 features.

1.2 Classification Modeling

We decide to model the problem as a classification problem instead of a regression problem. Only nonspam and spam samples in SET1 are used for modeling. We select random forests for classification modeling because it demonstrates good performance on many other complex classfication tasks [1]. The R randomForest package, available at `http://cran.r-project.org/web/packages/randomForest/index.html`, is used in our experiments.

1.3 Result Analysis

The random forests classifier includes 500 trees, and is modeled with default parameters. Under 7-fold cross validation, the highest Area under ROC value is 0.82045. The ROC curve is shown in Fig. 1.

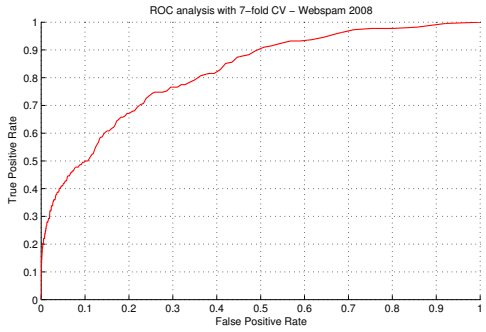

Figure 1: ROC curve

2. REFERENCES

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.