University of California Davis
Department of Statistics

Machine Learning (STA-208)

Ying-Chen Chou
Chia-Hui Shen
Jiahui Tan
Pei-Ying Ling

Contents

1 Introduction

With the progress of the times, email has been one of the most efficient communication media nowadays. However, with the ubiquitous use of email, useless messages and advertisements spread widely which caused email misuse.To improve this issue, lots of people did the researches for some approaches to construct filters by using machine learning. Most researches mainly focus on adjusting classical methods to make filters more efficient.

To have better understanding, we referred from lots of papers about detection of spam e-mail. Some papers tried to compare different methods and found the best methods. However, the results in different papers sometimes are different because most of them did not compare the same methods at the same time. Another reason may come from different data set in they used. Moreover, the data set they used typically is experimental. Therefore, we want to make a overall comparison for the methods they tried and apply those methods on larger data set we combined. We would compare four methods: SVMs, Decision Trees, Naive Bayesian,MLP, and Logistics by calculating accuracy rate and time. In addition, we would use two kinds of data prepossessing (unit-gram and bi-gram tfidf) to add complexity of our comparisons.

Moreover, our data set contains over than 150,000 email from 1999 to 2007. We supposed that the keyword per year would change because of the innovated technology or social cognition. We would like to discuss keywords in spam email by year to explore the characteristics in each year.

2 Description of Data

2.1 Raw Data

There are only a few public sources for email data to which almost everyone trying to do similar analysis will turn to. For our project, we downloaded datasets from the following three locations going between the years 1999-2007.

1. Enron-Spam datasets
2. SpamAssassin data
3. TREC email corpus

Year	Ham	Spam	Ham %	Spam %	Total
1999	2978	4611	0.392410	0.607590	7589
2000	8512	2851	0.749098	0.250902	11363
2001	9872	848	0.920896	0.079104	10720
2002	10663	5280	0.668820	0.331180	15943
2003	545	1773	0.235116	0.764884	2318
2004	627	13420	0.044636	0.955364	14047
2005	1309	18418	0.066356	0.933644	19727
2006	1226	2730	0.309909	0.690091	3956
2007	25219	48999	0.339796	0.660204	74218

From the table showing the distribution of emails between years, we can see that there is a disproportional number of emails between each year as well as between spam and ham groups. The imbalance dataset may be worrisome for our classifiers. However, there is not much we can do about this situation. Maybe if we can see that there is not too much variation between ham and spam emails across years or there is no time obvious time effect on the emails, then combining the different years wouldn't be a problem.

2.2 Top 50 Words Per Year

2.3 Word Cloud

2.4 Mathematics

3 Previous Studies

4 Method

4.1 Feature Engineering

4.2 Fit Models

4.3 Conclusion and exploration