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 email. Some papers tried to compare different methods and found the best one. 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 they used. Moreover, the data set they used typically are experimental and smaller. 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, KNN, Multi Layer Perceptron (MLP), and Logistics by calculating accuracy rate and cost time. In addition, we would use two kinds of data prepossessing (unit-gram and bi-gram tfidf) to increase complexity in our project.

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

Table 1: Proportion of Ham Email and Spam Email From 1999 to 2007

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 (Table 1) 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. To see whether there are differences in email across the email, we will examine the top words by year.
The main challenge of cleaning up the emails come from trying to remove the html and css elements, such that when we tokenize we wouldn't end up with tag elements as the most frequent words. Although we cannot remove all html and css, we did manage to get rid of most of it by the removal of contents between brackets, parathesis, and curly braces as well as words that begin with a period.

2.2 Top 10 Words Per Year

One of the main purposes for this project is to explore whether the keywords for spam and ham email changed by year. In this section, we counted the appear frequency of each word as a vector by year respectively. Sort the frequency and find out the top 10 frequent words per year. We would like to explore that whether the frequent words changed by year. Moreover, in the next section, we will use word cloud to visualize the results we found out.

According to the Figure.1, although keywords did not change year by year, we still have found out that there might have a difference in 2002. Before 2001, some keywords apprered repeatedly, such as "microsoft", "adobe", "windows", and "free". It seems before 2001, in our data set, the spam emails mostly are related to computer and Microsoft topics. From 2002, "stock", "business", "money", and "com" become top frequent keywords. We categorize those as economy and Internet topics.

For the ham email (Firure. 2), there is not specific topic for each year. We cannot conclude any specific topic or gap for year. However, overall keywords mainly focus on acadmic, such as "edu", "university", "data". We inferred that ham email data may mostly come from academic organizations.

	1999	2000	2001	2002	2003	2004	2005	2006	2007
0	price	price	price	free	price	price	company	com	pills
1	company	company	free	email	company	company	statements	price	desjardins
2	info	info	com	click	pills	email	information	yahoo	mg
3	gold	gold	company	mail	mg	money	adobe	net	item
4	microsoft	adobe	save	money	item	professional	business	org	price
5	adobe	windows	website	business	info	information	com	company	save
6	windows	microsoft	microsoft	list	save	com	price	gold	votre
7	office	campaign	money	information	gold	new	professional	hotmail	online
8	save	hi	like	time	stock	mail	time	info	vous
9	xp	office	adobe	new	click	time	email	aol	like

Figure 1: Top 10 Frequent Words Of Spam Email

	1999	2000	2001	2002	2003	2004	2005	2006	2007
0	board	ect	enron	list	dmdx	dmdx	putdup	node	samba
1	use	enron	company	linux	edu	mail	dmdx	nodes	source
2	hb	hou	said	com	mit	putdup	interval	network	new
3	handyboard	subject	ect	new	mail	file	edu	time	help
4	like	vince	energy	data	cert	use	obj	information	branches
5	edu	cc	new	use	use	jonathan	mail	peer	code
6	thanks	pm	power	like	list	list	list	file	list
7	using	com	subject	time	information	digitalvox	university	new	com
8	know	thanks	com	net	time	wrote	endobj	message	data
9	time	gas	gas	message	ms	time	time	mail	use

Figure 2: Top 10 Frequent Words Of Ham Email

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