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 **2**

2 Description of Data **2**

2.1 Data Preprocessing 2

2.2 Top 10 Words Per Year 4

2.3 Word Cloud 6

3 Previous Studies **8**

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 Data Preprocessing

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:

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

We combine all emails and extracted the information including date, from, to, subject, content, number of cc, and number of bcc. In the process of data preprocessing, we faced some challenges to get the information of year, weekday, and hour at which the email was sent. We tried to use the string methods in python but end up finding regular expression is more powerful to extract the weekday and month. The other 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.

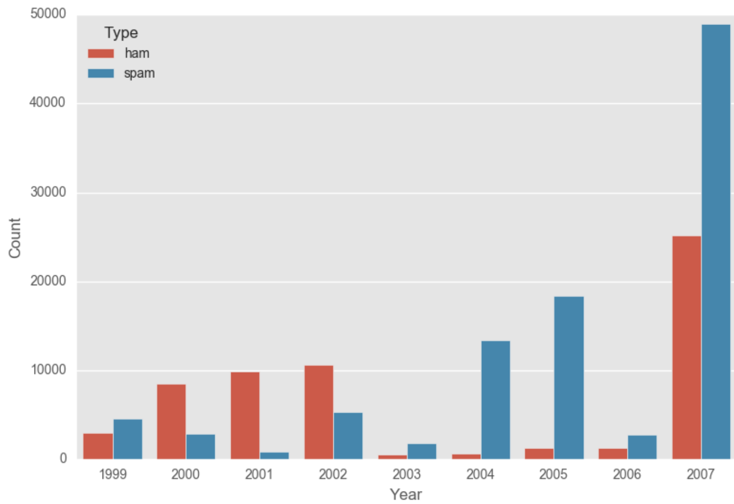


Figure 1: Amount of Email From 1999 to 2007

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

After finishing the data cleaning up, we delete the emails with year not between 1999 to 2007 to prevent the situation that the date of the emails is after when the data was collected and that the date of email is so early that the email are still not common used. There are 159981 email with 60951 of them are ham and the other 98930 are spam. From Table 1 and Figure 1, 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. Figure 2 shown the amount of the email sent each hour and each day. There is a peak for sending spam email at around 12:00 to 15:00. However, ham emails were usually sent between 8:00 to 20:00. Also, according to the right plot of Figure 2, ham emails tend to be sent during weekdays and has lower proportion in weekend but spam email seems to be balance each day.

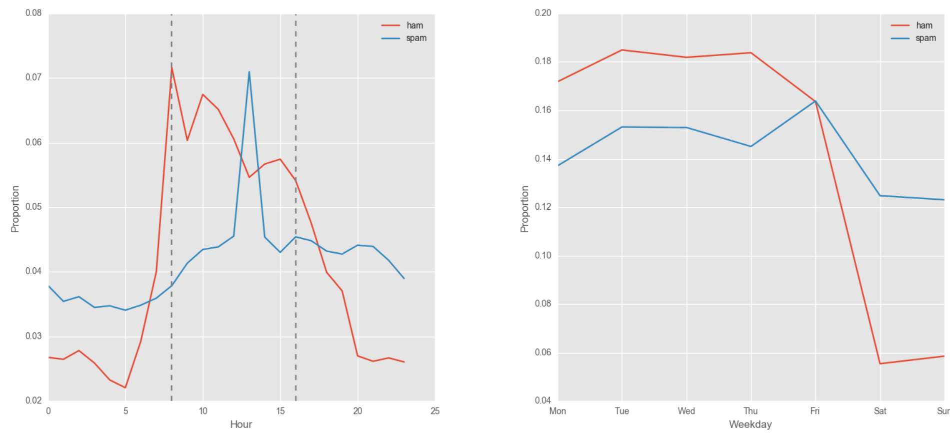


Figure 2: Amount of Email Hourly and On Different Weekday

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.3, 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 appreared 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. 4), 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 3: Top 10 Frequent Words Of Spam Email

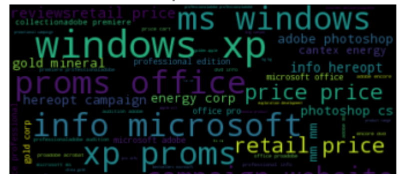
	1999	2000	2001	2002	2003	2004	2005	2006	2007
0	board	ect	enron	list	dmdx	dmdx	putdup	node	samba
1	use	enron	company	linux	edu	mail	dmdx	nodes	source
2	hb	hou	said	com	mit	putdup	interval	network	new
3	handyboard	subject	ect	new	mail	file	edu	time	help
4	like	vince	energy	data	cert	use	obj	information	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 4: Top 10 Frequent Words Of Ham Email

2.3 Word Cloud

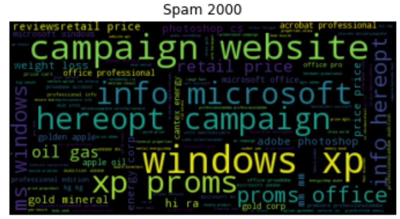
Ham 1999

Spam 1999



Ham 2000

Spam 2000




Ham 2001
original message
dow jones
hou ect
year
enron enron
natural gas
new york
ect ect
Spam 2001
retail price
proms office
ms windows
windows xp
xp proms
info microsoft
mg pills
Ham 2002
dbcaps data
linux ie
unknown database
database alias
time
alias dbcaps
Spam 2002
email address
name address
government grant
legal notice
city state
credit card
web site
Ham 2003
ribbon campaign
mit edu
campaign html
jonathan ascii
Spam 2003
info hereopt
mg pill
xp proms
windows xp
hereopt campaign
pills mg
Ham 2004
digitalvox digitalvox
ribbon campaign
jonathan ascii
putdup putdup
campaign html
problem
Spam 2004
windows xp
xp professional
retail price
microsoft office
Ham 2005
endobj obj
psy psych
putdup putdup
psych arizona
Spam 2005
investment advice
windows xp
adobe photoshop
office xp
xp professional
Ham 2006
peer peer
number
paper
system
node
time
data
information
Spam 2006
proms office
yahoo
windows xp
info microsoft
ms windows
campaign website
Ham 2007
samba source
math ethz
branches samba
stat math
reproducible code
Spam 2007
groupe desjardins
retail price
mg pill
pills mg

From the wordclouds, we see that Microsoft and windows XP seem to be a consistent theme in spam. However, we do see a shift less on windows and more towards sales and products during the later half. For ham, based on the words that show up, we may think that the email data originate from an education or technical source due to terms like edu and systems.

3 Previous Studies

[10pt, letterpaper, titlepage]article graphicx enumerate

Introduction

It is important to study what people did and create our study on top of it. In this section, we focus on summarize previous studies of spam/ham e-mail filtering and their machine learning methods. At the end of this section, we point out the new methods and new data we use in this project to show our understanding.

Previous Studies

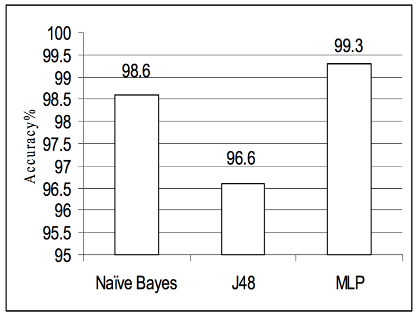
People already came up with the idea of spam/ham e-mail filtering before 2004. In the paper *Machine Learning Techniques in Spam Filtering* written by Konstantin in 2004, the experiment used four main methods : Naive Bayes, K-NN, Perceptron, and SVM. And compare accuracy rate of each methods. In this basic practice, they found the Perceptron method has the highest accuracy rate, 98.5% with a corpus of 1099 messages.

Algorithm	$N_{L\to S}$	$N_{S\to L}$	P	F_L	F_S	G
Naïve Bayes ($\lambda = 1$)	0	138	87.4%	0.0%	28.7%	1.56
k-NN ($k = 51$)	68	33	90.8%	11.0%	6.9%	1.61
Perceptron	8	8	98.5%	1.3%	1.7%	1.75
SVM	10	11	98.1%	1.6%	2.3%	1.74

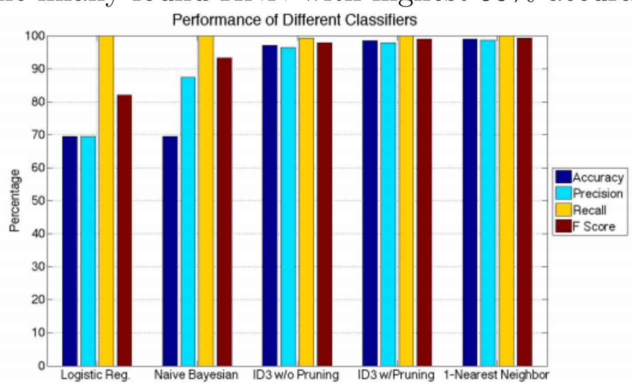
In 2010, *Email Spam Filtering using Supervised Machine Learning Techniques* written by V.Christina, they used Naive Bayes, J-48(Decision Tree) and Multilayer Perceptron. And they found out MLP performed the best with 99.3% accuracy rate when experimenting with a corpus of 1500 messages.

COMPARATIVE RESULTS OF THE CLASSIFIERS

Evaluation Criteria	Naïve Bayes	J48	MLP
Training time (secs)	0.15	0.20	138.05
Correctly Classified Instances	1479	1449	1490
Prediction Accuracy (%)	98.6	96.6	99.3
False Positive (%)	5	4	1



In 2016, *Spam Mail Detection using Classification* written by Parhat and Gambhir used Naive Bayes, SVM and J-48(Decision Tree). And they found out Naive Bayes performed the best with 76% accuracy rate in their experiment.
And *Email Spam Detection* written by Ge and Lauren, used the corpus from TREC 2007 with 1000 messages. They tried logistic regression, Naive Bayesian, Decision Tree and K-NN. The finally found KNN with highest 99% accuracy rate.



Here is the summary of methods in each previous studies by year.

	NB	KNN	SVM	Decision Tree	MLP (Neural Network)	Logistic	Best Model
2004	v	v	v		v Perceptron		MLP 98.5%
2010	v			v C4.5	v		MLP 99.3%
2016	v	v	v	v J48	v	v	KNN 99%

Our works

1. **Use multiple data source**: In each paper, they mainly use a single year of corpus data. In our project, we tried to source different e-mail and integrate them. The format of each data source is different thus hard to clean. And we successfully got to manage a huge data set.

2. **Try 6 methods at the same time**: Previous studies compare accuracy rate with

different methods, but they didn't compare them all at a time. So we studied the methods from 2004 to 2016, and apply all of possible methods with adequate tuning parameters to compare them

3. **Apply 1-Gram and 2-Gram**: Each paper marked that data processing step is important to a good result. Here, we introduced bag of words of 1-Gram and 2-Gram methods in the feature engineering part. And we can see different result of accuracy rate in the following section.