International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 629 - 632, 2015, ISSN: 2319–8656

# Spam Detection in Social Networks Using Correlation Based Feature Subset Selection

Sanjeev Dhawan
Department of Computer Science & Engineering,
University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra-136119,
Haryana, India

Meena Devi
Department of Computer Science and Engineering
University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra-136119,
Haryana, India

**Abstract:** Bayesian classifier works efficiently on some fields, and badly on some. The performance of Bayesian Classifier suffers in fields that involve correlated features. Feature selection is beneficial in reducing dimensionality, removing irrelevant data, incrementing learning accuracy, and improving result comprehensibility. But, the recent increase of dimensionality of data place a hard challenge to many existing feature selection methods with respect to efficiency and effectiveness. In this paper, Bayesian Classifier with Correlation Based Feature Selection is introduced which can key out relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of our method is presented through broad.

**Keywords**: Bayesian Classifier, Feature Subset Selection, Naïve Bayesian Classifier, Correlation Based FSS, Spam, Non-Spam

## 1. INTRODUCTION

It is impossible to tell exactly who was the first one to come upon a simple idea that if you send out an advertisement to a number of people, then at least one person will react to it no matter what is the proposal. E-mail provides a very good way to send these millions of advertisements at no cost for the sender, and this unfortunate fact is nowadays extensively exploited by several organizations. As a result, the e-mailboxes of millions of people get cluttered with all this so-called unsolicited bulk e-mail also known as "spam" or "junk mail". Being incredibly cheap to send, spam causes a lot of problems to the Internet community: large amounts of spam-traffic between servers cause delays in delivery of solicited email, people with dial-up Internet access have to spend bandwidth downloading junk mail. Sorting out the unwanted messages takes time and introduces a risk of deleting normal mail by mistake. Finally, there is quite an amount of pornographic spam that should not be uncovered to children. A number of ways of fighting spam have been proposed. There are "social" methods like legal measures (one example is an anti-spam law introduced in the US) and plain personal participation (never respond to spam, never publish your e-mail address on WebPages, never forward chain-letters. . .). There are 60 "technological" ways like blocking spammer's IP-address (blacklist), e-mail filtering etc.. Unluckily, till now there is no perfect method to get rid of spam exists, so the amount of spam mail keeps increasing. For example, about 50% of the messages coming to my personal mailbox are unsolicited mail. For blocking spam at the moment Automatic e-mail filtering appears to be the most effective method and a tough competition between spammers and spam-filtering methods is going on: the better the anti-spam methods get, so do the tricks of the spammers. Several years ago most of the spam could be reliably handle by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines. To overcome these spammers began to specify random sender addresses and to append random characters to the end of the message subject. Spam filtering rules adjusted to consider separate words in messages could deal with that, but then junk mail with specially spelled words (e.g. B-U-Y N-O-W) or simply with misspelled words (e.g. BUUY NOOW) was born. To fool the more advanced filters that relies on word frequencies spammers append a large amount of "usual words" to the end of a message. Besides, there are spams that contain no text at all (typical are HTML messages with a single image that is downloaded from the Internet when the message is opened), and there are even self-decrypting spams (e.g. an encrypted HTML message containing JavaScript code that decrypts its contents when opened). So, as you see, it's a never-ending battle. There are two basic approaches to mail filtering knowledge engineering (KE) and machine learning (ML). In the former case, a set of rules is created according to which messages are categorized as spam or legitimate mail. A typical rule of this kind could look like "if the Subject of a message contains the text BUY NOW, then the message is spam". A set of such rules should be created either by the user of the filter, or by some other authority (e.g. the software company that provides a particular rule-based spam-filtering tool).The major drawback of this method is that the set of rules must be constantly updated, and maintaining it is not convenient for most users. The rules could, of course, be updated in a centralized manner by the maintainer of the spam filtering tool, and there is even a peer-2-peer knowledgebase solution, but when the rules are publicly available, the spammer has the ability to adjust the text of his message so that it would pass through the filter. Therefore it is better when spam filtering is customized on a per-user basis. The machine learning approach does not require specifying any rules explicitly. Instead, a set of pre-classified documents (training samples) is needed. A specific algorithm is then used to "learn" the classification rules from this data. The subject of machine learning has been widely studied and there are lots of algorithms suitable for this task. This article considers some of the most popular machine learning algorithms and their application to the problem of spam filtering. More-or-less self-contained descriptions of the algorithms are presented and a simple comparison of the performance of my implementations of the algorithms is given. Finally, some ideas of improving the algorithms are shown.

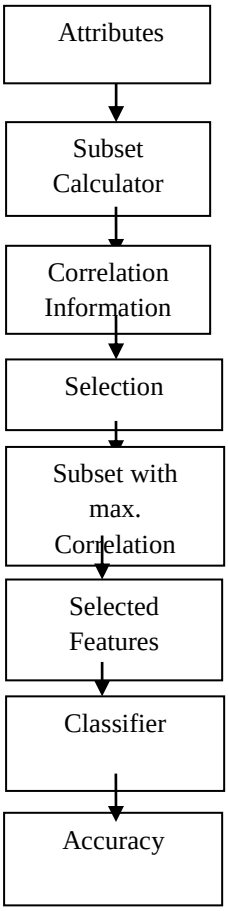
## 2. CHALLENGES IN SPAM DETECTION

One of the barriers to legislation against spam is the fact that not everyone uses exactly the same definition. It doesn't help that laws may be made at different levels even within the

same country, let alone laws in different countries. With so many different and sometimes conflicting laws, prosecution can be very difficult. Another barrier both to legislation and practical filtering is that email is not designed in such a way that the sender can always be traced easily. There is no authentication of the sender built in to the protocol used by email, leaving it possible for people to forge sender information. This makes it hard to trace back and prosecute the sender, or to avoid receiving messages from a known spammer in the future. There are several proposals to adapt this protocol like Microsoft's "Caller ID for email". Spam changes with time as new product are introduced and seasons change. For example, Christmas-themed spam is not usually sent in June. But beyond that, there are targeted changes happening in spam. Perhaps the largest problem of spam filtering is that spammers have intelligent beings working to ensure that "direct email marketing" (the marketing term for spam) is seen by as many potential customers as possible. Many anti-spam tools are freely available online, which means that spammers have access to them too, and can learn how to get through them. This makes spam detection a co-evolutionary process, much like virus detection: both sides change to gain an advantage, however temporarily. Although it does change, spam is not completely volatile. Terry Sullivan found that while spam does undergo periods of rapid changes, it also has a core set of features which are stable for long periods of time. Spam changes from person to person. This is partly due to targeting on the part of the address harvesters, who try to guess the interests of the recipients so that the response rate will be higher. But more importantly, legitimate mail also varies from person to person. In theory it should be possible to discover spam without much attention to the legitimate mail. However, the great success of classifiers which use both, such as Graham's Bayesian classifier and the CRM114 discriminator [Yer04], implies that use of data from both legitimate and spam email is very beneficial. One final thing to note in the difficulty of spam classification is that all mistakes in classification are not equal. False negatives, messages that have accidentally been tagged as non-spam, are usually seen by the user. They may be annoying, but are usually easy to deal with. However, false positives, messages that have been accidentally tagged as spam, tend to be more problematic. When a single legitimate message is in a pile of spam, it is much easier to miss seeing it. (A typical user will not read all spam, but instead scans subject and from lines quickly to see if anything legitimate stands out.) While there is relatively little impact if a person receives a single spam, missing a real message which might be important is much more dangerous. One research firm suggests that companies lose $3 billion dealing with false positives.

## 3. PROPOSED WORK

In previous work various spam detection algorithm have been proposed ranging from text based to feature based using classifiers such as naïve bayes, SVM, ANN, kNN and decision tree etc. However Naïve Bayesian Method is utilized by 99% of the company. The reason for this is their classification efficiency. But these probabilistic methods take in consideration all the feature of the spam making the overall accuracy ranging from 65 to 74 %. So we require a more efficient method to improve spam detection and false alarm reduction. The feature subset algorithm tries to formulate the vector space of the features by filtering of subset selecting the most prominent feature of spam and removing unwanted features. The filtering allows the reduction in search space and noise. After filtering using FSS we have applied attribute selection based naïve Bayesian probabilistic classifier and achieved 17-20% more accuracy.




## 4. FEATURE SUBSET SELECTION

Feature subset selection is used for identifying and removing as much irrelevant and redundant information as possible and thus it reduces the dimensionality of the data and may allow learning algorithms to run faster and more effectively. In some cases, accuracy on future classification can be improved; in others, the result is a more compact, well interpreted representation of the aimed concept.

## 5. CORRELATION BASED FSS

CFS algorithm relies on a heuristic for assessing the cost or merit of a subset of features. This heuristic takes into account the usefulness of individual features for forecasting the class label along with the level of intercorrelation among them. The hypotheses on which the heuristic is based is:

Sound feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.

Features are relevant if their values vary systematically with category membership. A feature is useful if it is correlated with or forecaster of the class; otherwise it is irrelevant. Empirical grounds from the feature selection literature show that, along with irrelevant features, redundant information

should be wiped out as well. A feature is said to be redundant if one or more of the other features are highly correlated with it. The above definitions for relevance and redundancy lead to the idea that best features for a given classification are those that are highly correlated with one of the classes and have an insignificant correlation with the rest of the features in the set.
If the correlation between each of the components in a test and the outside variable is known, and the inter-correlation between each pair of components is given,then the correlation between a composite consisting of the summed components and the outside variable can be predicted from

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k - (k-1)\overline{r_{ii}}}} \qquad (5.1)$$

Where

$r_{zc}$ = correlation between the summed components and the outside variable.

$k$ = number of components (features).

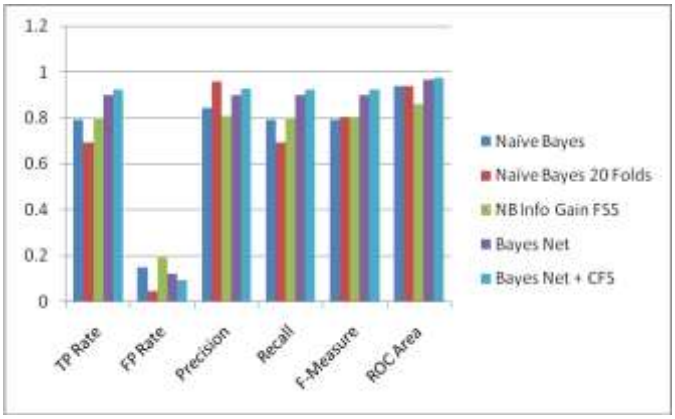
$\overline{r_{zi}}$ = average of the correlations between the components and the outside variable.

$\overline{r_{ii}}$ = average inter-correlation between components.

Equation 5.1 represents the Pearson's correlation coefficient, where all the variables have been standardized. The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them. Thus, equation 5.1 shows that the correlation between a composite and an outside variable is a function of the number of component variables in the composite and the magnitude of the inter-correlations among them, together with the magnitude of the correlations between the components and the outside variable. Some conclusions can be extracted from (5.1):

- The higher the correlations between the components and the outside variable, the higher the correlation between the composite and the outside variable.
- As the number of components in the composite increases, the correlation between the composite and the outside variable increases.
- The lower the inter-correlation among the components, the higher the correlation between the composite and the outside variable.

## 6. CLASSIFICATION RESULTS




| Classifier | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Correct |
|---|---|---|---|---|---|---|---|
| Naïve Bayes | 0.793 | 0.152 | 0.842 | 0.793 | 0.794 | 0.937 | 79.3871 |
| Naïve Bayes 20 Folds | 0.692 | 0.046 | 0.959 | 0.692 | 0.804 | 0.937 | 79.5262 |
| NB Info Gain FSS | 0.8 | 0.196 | 0.808 | 0.8 | 0.802 | 0.861 | 80.0478 |
| Bayes Net | 0.9 | 0.123 | 0.9 | 0.9 | 0.899 | 0.965 | 89.9587 |
| Bayes Net + CFS | 0.924 | 0.096 | 0.925 | 0.924 | 0.924 | 0.974 | 92.4147 |

## 7. CONCLUSION AND FUTURE SCOPE

Feature subset selection (FSS) plays a vital act in the fields of data excavating and contraption learning. A good FSS algorithm can efficiently remove irrelevant and redundant features and seize into report feature interaction. This also clears the understanding of the data and additionally enhances the presentation of a learner by enhancing the generalization capacity and the interpretability of the discovering mode. An alternative way employing a classifier on a corpus of e-mail memos from countless users and a collective dataset.
In this work we have worked on improving SPAM detection based on feature subset selection of Spam data set. The Feature Subset selection methods such as Info Gain Attribute selection and Correlation based Attribute Selection can be perceived as the main enhancement to Naïve Bayesian/ probabilistic methods. We have analyzed the Probabilistic SPAM Filters and attained more than 92% of success in filtering SPAM.
However many open issues still remain open such as, the system deals only with content as it has been translated to plain text or HTML. Since some spam is sent where most of the message is in an image, it would be worth looking at ways in which images and other attachments could be examined by the system. These could include algorithms which extract text from the attachment, or more complex analysis of the information contained within the attachment. We can also work on a technique to recognize web junk e-mail according to finding these boosting pages in place of web spam page itself. We will begin from a small set of spam seed pages to get a hold of boosting pages. Then web junk e-mail pages are supposed to be identified making use of boosting pages. We can also work on a better larger dataset; the system should be tested over a longer period than the one-year one available in the public domain.

## 8. REFERENCES

[1] Hayati, Vidyasagar, Potdar, and Pedram, "Evaluation of spam detection and prevention frameworks for email and image spam: a state of art," In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, ACM, pp. 520-527, 2008.

[2] Becchetti, Luca, Carlos Castillo, Debora Donato, Ricardo Baeza-Yates and Stefano Leonardi, "Link analysis for web spam detection," ACM Transactions on the Web (TWEB), vol. 2, no. 1, 2008.

[3] Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, And Efstathios Stamatatos, "Words Vs. Character N-Grams For Anti-Spam Filtering," International Journal on Artificial Intelligence Tools, pp. 1–20, 2006.

[4] Joshua Attenberg, Kilian Weinberger, Anirban Dasgupta, Alex Smola, and Martin Zinkevich, "Collaborative Email-Spam Filtering with the Hashing Trick," CEAS, 2009.

[5] Tu Ouyang, Soumya Ray, Michael Rabinovich and Mark Allman," Can network characteristics detect spam effectively in a stand-alone enterprise?," In Passive and Active Measurement, (Springer Berlin Heidelberg, 2011), pp. 92-101, 2011.

[6] Rushdi Shams and Robert E. Mercer,"Classifying Spam Emails using Text and Readability Features," IEEE 13th International Conference on Data Mining (ICDM), pp. 657-666, 2013.

[7] Lei Yu, Huan Liu," Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution" Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[8] Liumei Zhang, Jianfeng Ma, and Yichuan Wang, "Content Based Spam Text Classification: An Empirical Comparison between English and Chinese," 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS), IEEE, pp. 69-76, 2013.

[9] Igor Santos, Carlos Laorden, Borja Sanz, and Pablo Garcia Bringas, "JURD: Joiner of Un-Readable Documents to reverse tokenization attacks to content-based spam filters", Consumer Communications and Networking Conference (CCNC), IEEE, pp. 259-264, 2013.

[10] De Wang, Danesh Irani, and Calton Pu, " A study on evolution of email spam over fifteen years," IEEE 2013 9th International Conference on In Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), pp. 1-10, 2013.

[11] Bujang, Yanti Rosmunie, and Husnayati Hussin, "Should we be concerned with spam emails? A look at its impacts and implications," 2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M), IEEE, pp. 1-6 2013.

[12] Manek, Asha S., D. K. Shamini, Veena H. Bhat, P. Deepa Shenoy, M. Chandra Mohan, K. R. Venugopal, and L. M. Patnaik, "ReP-ETD: A Repetitive Preprocessing technique for Embedded Text Detection from images in spam emails," 2014 IEEE International Advance Computing Conference (IACC), pp. 568-573, 2014.

[13] Bosma, Maarten, Edgar Meij, and Wouter Weerkamp, "A framework for unsupervised spam detection in social networking sites, Advances in Information Retrieval," Springer Berlin Heidelberg, pp. 364-375, 2012.

[14] Dave, Vacha, Saikat Guha, and Yin Zhang, "Measuring and fingerprinting click-spam in ad networks," In Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication, ACM, pp. 175-186, 2012.

[15] Karthika Renuka and Visalakshi, "Latent Semantic Indexing Based SVM Model for Email Spam Classification," Journal of Scientific & Industrial Research, vol. 73, pp. 437-442,July 2014.