Email Spam Filtering

Ananda Kishore Sirivella, UFID: 9951-5080

ABSTRACT

In recent times, Electronic mail are being used a frequent to communicate by millions of people on the daily basis. Over the decade, the exploration of email advertising and increasing in unsolicited bulk e-mail has created a need for a reliable anti-spam filter. A Junk email could be an unsolicited advertisement, phishing scam, email spoofing, commercial advertisement, .etc. With day to day upgrade in email spam patterns, we look for a machine learning algorithm for the anti-spam filtering. In the project, we shall explore, implement and compare some of the most popular machine learning methods (Naïve Bayes classifier, K nearest neighbor classifier & SVM). In the project, I'll be going through the motivation of using these algorithms or the basic science behind them, working of the aforementioned algorithms, their accuracies and other such factors. Towards the end of the project, I aim to achieve at a statistical conclusion over the effectiveness of each classifiers over different email spam patterns, knowing which classifiers fits best for a what type pattern of spam.

PROJECT PLAN

A. *Exploration Stage - I*

Initial literature search over the various complexities of spam e-mails, and its various patterns. This phase emphasizes more on the deeper understanding of the problem. Targeted by March 13, 2016.

Involves elaboration over the types of email spams, its definitions and possible patterns like keywords, special character distribution,.etc.

B. *Exploration stage – II*

Initial literature search over the concepts of machine learning, and exploration of some of its popular methods (Naïve Bayes classifier, K nearest neighbor classifier & SVM). This phase emphasizes on the deeper understanding of the machine learning concepts. Targeted by March 19, 2016.

C. *Exploration stage – III*

Tool exploration for the implementation of the machine learning methods and a dataset for the results evaluation. This phase would involve learning python, and its libraries for the algorithm implementations. Targeted by March 27, 2016

Involves exploring the UCI Spambase dataset, and understanding its diversity. More of understanding the test set.

D. *Development Stage*

Implementation of the machine learning methods in python. And preliminary unit testing of the implementation. Targeted by April 12, 2016.

Involves converting machine learning concepts into a executable code. A phase of software development of the classifier.

E. *Review Stage*

Parsing the data set, collection, comparison and exclusive analysis of the results. This phase emphasizes on the result comparison, evaluation of each classifier pros and cons with respect to the variation of e-mail spam patterns. Targeted by April 19, 2016.

Involves evaluation of classifier correctness, time analysis, and its effectiveness over recognition of patterns.

F. *Final Stage*

Documentation of the project, and its results. This phase emphasizes on the preparing report, presentation and video of the project. Targeted by April 24, 2016.

REFERENCES

[1] Sergios Theodoridis , *Machine Learning: A Bayesian and Optimization Perspective*, 1st edition, Academic Press, 2015.

[2] Richard O. Duda, Peter E. Hart, David G. Stork, *Pattern Classification*, 2nd Edition, Wiley-Interscience, October 2000.

[3] David G. Stork, Elad Yom-Tov, *Computer Manual in MATLAB to accompany Pattern Classification*, 2nd Edition, Wiley-Interscience, April 2004.

[4] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, 1st Edition, Springer, October 1, 2007.

[5] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer, February 9, 2009.

[6] Drucker, H., Wu, D., Vapnik, V.N, *Support Vector Machines for Spam Categorization*. IEEE Transactions on Neural Networks 10(5), 1048 – 1054 (September 1999)

[7] Hinneburg, C.C.A.A., Keim, D.A.: *What is the nearest neighbor in high dimensional spaces?* In: Proc. of the International Conference on Database Theory (ICDT), pp. 506 –515. Morgan Kaufmann, Cairo, Egypt (September 2000)

[8] Bickel, P.J., Ritov, Y., Zakai, A., *Some theory for generalized boosting algorithms.* Journal of Machine Learning Research 7, 705 –732 (2006)

[9] en.wikipedia .org

[10] paulgraham.com/spam.html

[11] archive.ics.uci.edu/ml/datasets/Spambase