UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Faculty of Computer Science and Information Technology

Improving Email Spam Detection using Support Vector Machine Based on Reweighted Select Features

by

Mohamed Sirelkhatim Osman

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Science

Supervisor

Dr. Ali Ahmed

January 2016

Abstract

Spam is unsolicited or bulk electronic messages used to promote material that is very often not wanted or applicable to recipient. The content of spam can also be immoral, illegal and offensive; this is may cause time-consuming to handle, influencing bandwidth and misusing storage space.

Many spam detection techniques based on machine learning algorithms have been proposed.

Support Vector Machine (SVM) has been commonly used in e-mail spam classification, one of the main problems of spam filtering its high data dimensionality of the feature space due to the massive number of e-mail dataset.

In this thesis Support Vector Machine based on feature selection through reweight process was applied to enhancing the classification accuracy.

Thesis results show that feature selection and feature reweighting have improved the Accuracy of Spam detection.

المستخلص

الرسائل غير المرغوب فيها هي رسائل الكترونية تستعمل للتر ويج لمنتجات أو مواد في أغلب الأحيان لَيست مطلوبة أَو قابلة للتطبيق لدى المستلم. محتوى الرسائل غير المرغوب فيها يمكن أيضاً أَن يكون غير شرعي، لا أخلاقي أوقد يكون هجوم؛ هذا قد يسبب تضييع للوقت لمعالَجتة، يؤثر على سعة الناقل ويسيئ إستعمال سعةَ التخزين.

الكثير من تقنيات الكشف عن الرسائل غير المرغوب فيها المبنية علي خوارزميات تعلم الآلة تم أقتراحها.

أستخدمت عموما في تصنيف الرسائل غير المرغوب فيها، (Support Vector Machine)
إحدى المشكلات الرئيسية التي تواجه تقنيات الكشف عن الرسائل غير المرغوب فيها هي الأبعاد المتعددة لقاعدة البيانات أو الخصائص الكثيرة للرسائل الإلكترونية نسبة للعدد الهائل من رسائل البريد الإلكتروني.

في هذه الإطروحة (Support Vector Machine) المستندة علي إختيار الخصائص المهمة ثم إعادة أوزانها قدمت لتَحسين دقة التصنيف.

نتائج الإطروحة أظهرت أن إختيار الميزة و إعادة وزن الميزات حسنا دقةَ كشف الرسائل غير المرغوب فيها.

1.1 An Overview

Spam is generally defined as “unsolicited, usually commercial, email sent to a large number of recipients” (S. M. Lee, D. S. Kim, J. H. Kim, and J. S. Park, 2010) . Spam is also known as unsolicited or junk email. Just like the junk email you get at home advertising everything from credit cards to local restaurants; email spam operates in the same way (T. Bogers and A. Van den Bosch, 2008). Spammers send out hundreds of thousands, and sometimes tens of millions, of emails to unsuspecting email recipients. These spam emails are usually trying to sell something. While most people delete these spam email messages without even reading them, a small percentage of email recipients open and read the email messages and sometimes even buy the products being sold. This is what makes it profitable for the spammers. It costs very little to send an email message. Therefore only a small percentage of people who receive spam need to make a purchase to make it profitable for the spammers.

An electronic message is spam if (A) the recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients, and (B) the recipient has no verifiably granted deliberate, explicit, and still-revocable permission for it to be sent. (T. Subramaniam, H. A. Jalab, and A. Y. Taqa, 2010)

The problem named spam has come to existence with the widespread usage of electronic mail (email) which not only wastes the time of the users, but also brings about other problems such as influencing bandwidth and misusing storage space (A. Beiranvand, A. Osareh, and B. Shadgar, 2012).

Spam detection methods try to identify likely spam either manually or automatically, and then act upon this identification by either deleting the spam content or visibly marking it as such for the user.

Spam filters are email programs that attempt to organize email according to criteria that the user specifies. The ultimate goal is to filter out all unwanted email. Spam filters use a variety of techniques to determine which emails are spam. Most spam filters offer the user a variety of options for how to handle spam. Users can choose to have the emails automatically deleted, sent to a spam folder or delivered to their normal inbox marked as spam.

Most spam filters will turn off all links contained in emails deemed to be spam as a protection. These links can be turned back on, however, if the user determines the email is not spam.

Most email programs come equipped with basic spam filter features. However, if users desire greater protection or control over spam, they can purchase spam filtering software.

Spam filters yield outstanding results. Laboratory testing shows that a content-based learning filter can correctly classify all but a few spam messages out of a hundred and all but a few thousand non-spam messages out of a thousand.

There is some evidence that similar results may be achieved in practice either by machine learning methods or by other methods like blacklisting, grey listing, and collaborative filtering. The controlled studies necessary to measure the effectiveness of all types of filters—and combinations of filters—have yet to be conducted. We argue that understanding and improving the effectiveness of spam filters is best achieved through a combination of laboratory and field studies, using common measures and statistical methods. According to Araúzo-Azofra and Benítez (A. Beiranvand, A. Osareh, and B. Shadgar, 2008) and Nizamani et al. (S. Nizamani, Memon, N., Wiil, 2012) by using feature selection methods one can improve the accuracy, applicability, and understand ability of the learning process.

Feature selection is the process of finding an optimal subset of features that contribute significantly to the classification (A. R. Behjat, A. Mustapha, H. Nezamabadi-pour, M. Sulaiman, and N. Mustapha, 2012). Selecting a small subset of features can decrease the cost and the running time of a classification system. It may also increase the classification accuracy because irrelevant or redundant features are removed.

Recently, the number of undesirable messages coming to e-mail has strongly increased (Alguliyev and Nazirova, 2012). According to Symantec Intelligence Report in September 2012 the percentage of spam in e-mail traffic was increased by 2.7 percentage points from August and averaged 75%, in addition to Kaspersky Lab annual report the total amount of spam in mail traffic was 78.5% (Bulletin, 2012; Wood, 2012).

As problem sizes continue to scale up with the explosive growth of using e-mail, essential research is required to enhance the classification efficiency and effectiveness (accuracy) (Forman, 2003; Sanasam et al., 2010).

E-mail classification is supervised learning problem and classification is very important methods to cancel this problem of spam. Recent research shows that spam classification is usually processed by statistical theory and Machine Learning (ML) algorithms, to differentiate between non-spam and spam e-mail (Fagbola et al., 2012; Guzella and Caminhas, 2009; Saad et

al., 2012). ML methods are able to extract the knowledge from a group of emails supplied and using the gained information in the categorization of newly received e-mail (Saad et al., 2012). The aim of ML is to improve the performance of the computer program through experience so as to make best decisions and solve problems in an intelligent way by using illustration data (Salehi and Selamat, 2011).

In the field of ML, Feature Selection (FS) is an importance topic to select a subset of features among the full features and then reduce the high data dimensionality, lead to show the best performance in classification accuracy (Yun et al., 2007).

Usually, the performance of feature selection algorithms has been measured by comparing the performance of classification algorithms before and after feature selection.

1.2 Problem Definition

The major problem of e-mail classification is the high dimensionality of feature space, (Shang et al., 2007). which is great preventive problem for many of the machine learning algorithm. For these reason we need a reducing stage of dimensions. and study the effect of re-weight of features.

1.3 Research Objective

1- To apply feature selection method and select the best features that represent the e-mail and remove the rest noise features.
2- To apply the re-weighting process by giving more weight values to the important selected features.
3- To apply SVM classifier based on re-weighted selected features.

1.4 Scope of Research

The dataset used in this thesis is spambase and it's available in Hewlett-packard labs.

1.5 Methodology

This section briefly explains the methodology adopted in this research, it is also discusses the research steps taken in comparing the ability of classification with and without feature selection method through re-weight process.

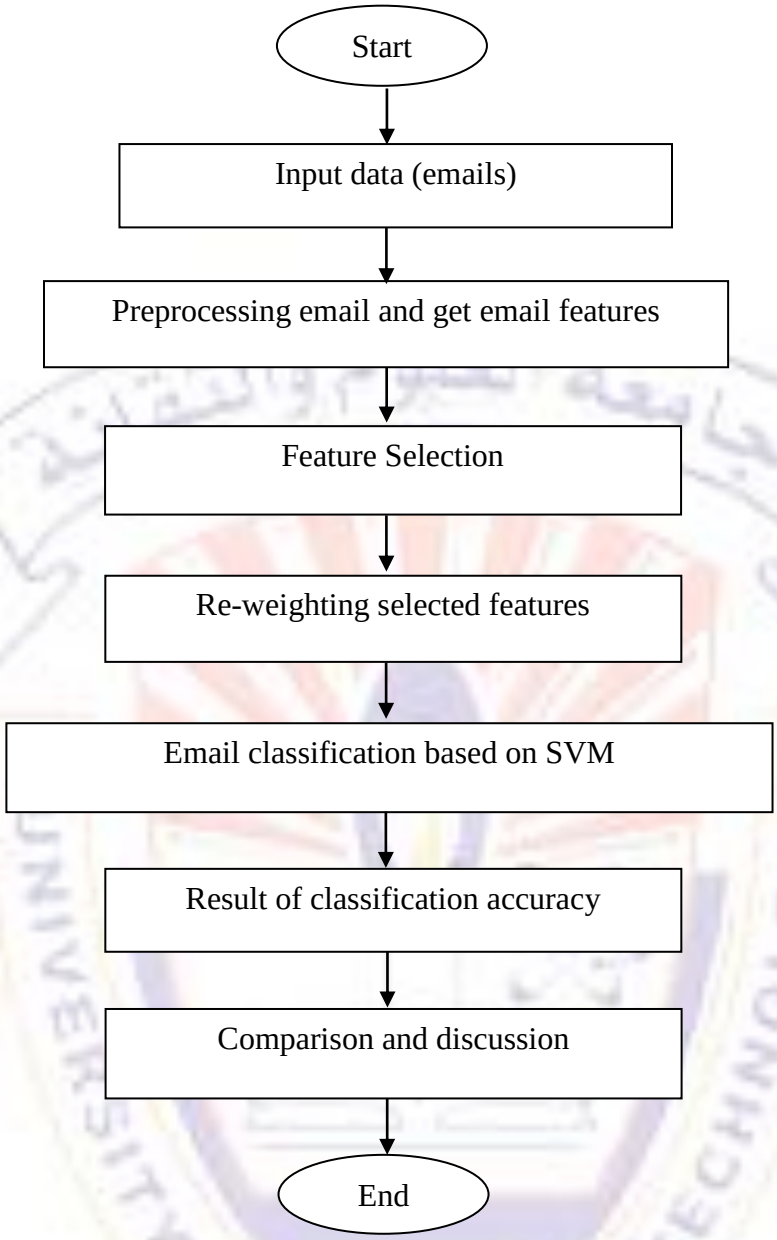

Figure (1.1): Overview of Research Methodology

1.6 Thesis structure

This research consists of five chapters starting chapter one introduction. Chapter two literature review it's includes introduction to information security, spam technique or spam filter, provides backgrounds of Support Vector Machine, feature selection. Chapter three is about research methodology, it's including preprocessing data and the experiment phases and tools. Chapter four contains experimental results and analysis. Chapter five contains Conclusions and recommendations.