# Project Short Description

Submitted by: Anupam Gangotia

Project Title: Email Classifier

Problem Description: Today everyone receives hundreds of mail every day, and organizing these emails becomes a chaos for the receiver. This Problem becomes even worse with the presences of Spam Mails. Electronic spamming is a major problem these days for email users. Spam email is an email sent to somebody without consent and its content can cause unease and distress. Few characteristics for spam emails are:

* The spam email is unsolicited.
* Always sent in bulk and the sender of spam email doesn’t target recipients personally. Thus, the addresses of recipients often are guessed and the same spam email is sent to numerous people at the same time.

These spam emails have already caused many problems such as consuming network bandwidth, wasting recipient time and so on. To resolve these problems, classiﬁcation of spam email from legitimate email has become very important. Recently, many machine learning and data mining techniques have been applied in spam email classiﬁcation, such as Naive Bayes, Decision Tree.

Proposed Solution: This Project is aims at building an Email classifier that using a test data set classifies the emails into two following categories (Spam, Atheism, Medical, Autos, Sports). If we denote the set of all e-mail messages by M, we may state that we search for a function

f : M → {Spam,Sports,Business,Technology,Entertainment}.

We shall look for this function by training machine learning algorithms on a set of pre-classified messages

{(m1, c1), (m2, c2), . . . , (mn, cn)}, mi ∈ M, ci ∈ { Spam, Atheism,Medical,Autos,Sports }.

To classify the emails, we have used a hybrid approach that combines Decision Tree and Naïve Bayes classifiers together. The naïve bayes classifier performs reasonably well. Its biggest weakness lies in the assumption that the attributes are independent of each other. Decision Trees classifiers, unless pruned, even for a small number of attributes require prohibitively large memory and running times.

Therefore our approach combines the Decision Tree Classifiers' propensity to separate out dependent attributes, and the effective classification by the Naïve Bayes Classifier on independent attributes. The idea is simple. In the learning phase, the Hybrid Classifier grows a tree exactly like the Decision Tree. The only difference is at the leaves, where a naive Bayes learner is now implemented. This Naive Bayes learner learns only on the training examples that arrive at that particular leaf, using only those attributes that have not been used by the Decision Tree along the path from the root to the leaf. During the inference phase, just as in Decision Trees, the attribute values of the test example determine the path that it takes down the tree and hence the particular leaf node that it reaches. The decision at the leaf node is taken by the Naive Bayes classifier based on the attributes of the test which have still not been considered.

Attributes To be considered while training the Classifier:

1. Sender (From Field)
2. Other Recipents(Cc, Bcc)
3. Subject of the email
4. Date
5. Content of the email

Example of Text classification using Naïve Bayes: This example considers only two classes spam and not spam. Following table shows the training dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Docs** | | | | **Spam** |
| big | Money |  |  | Yes |
| win | big | Big |  | Yes |
| get | grant |  |  | No |
| big | great | funding |  | No |

Test document is “win big”.

P(Spam=Yes|w=”win”^w=”big”)= P(w=”win”| Spam=Yes). P(w=”big”| Spam=Yes). P(Spam=Yes)

=(1/5).(3/5).(1/2)=(3/50)

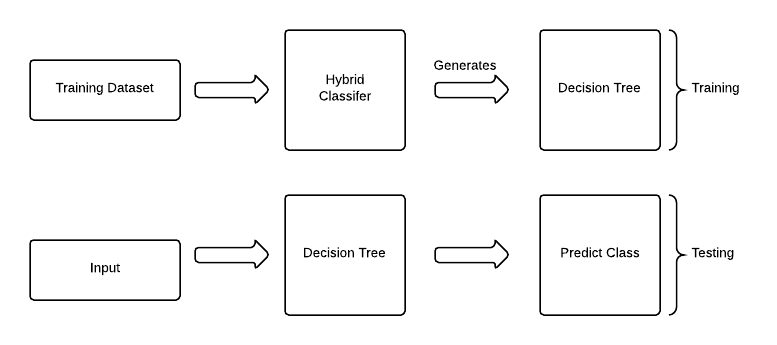
P(Spam=No|w=”win”^w=”big”)= P(w=”win”| Spam=No). P(w=”big”| Spam=No). P(Spam=No)

=(1/5).(1/5).(1/2)=(1/50)

Since P(Spam=Yes|w=”win”^w=”big”)> P(Spam=No|w=”win”^w=”big”), Therefore Naïve bayes will classify this test document as Spam=Yes.

Programming Tools: No Open Source Libraries. Solution provided in Java 1.7.

### Architectural Diagram:



### Refrence Papers:

1. Scaling Up the Accuracy of Naive Bayes Classifiers,a DecisionTree Hybrid by Ron Kohavi.

2. Machine Learning Techniques in Spam Filtering by Konstantin Tretyakov.

3. An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization by Jyh-Jian Sheu.

4.A Novel Hybrid Approach to Machine Learning by Muthukaruppan Annamalai, Ankur Jain, Vaishnavi Sannidhanam