Survey of Supervised and Unsupervised Algorithms in Email Management

Jainesh Patel, Neha R Soni

Abstract— Increase in internet speed provides an environment to grow the use of emails. Almost all organizations now communicate with the help of emails, which makes emails an important part of human life, as it helps in daily communication in formal or informal way. The rapid growth and use of email increases the problem of email overload. The email management is the task of managing emails into different categories or classes or folders according to some criteria or condition. The email management task is talked with the help of machine learning algorithms. This paper surveyed Naïve Bayes, Support Vector Machine, K-Nearest Neighbor from supervised and K-Means, Fuzzy C-Means, Evolving Email Clustering Method from unsupervised algorithms for email management. These algorithms are discussed in brief along with their working strategy and their performance for email domain.

Index Terms— Email Overload, Email Management, Supervised Algorithms, Unsupervised Algorithms, Email Categorization. Inbox, Gmail's Inbox

1 INTRODUCTION

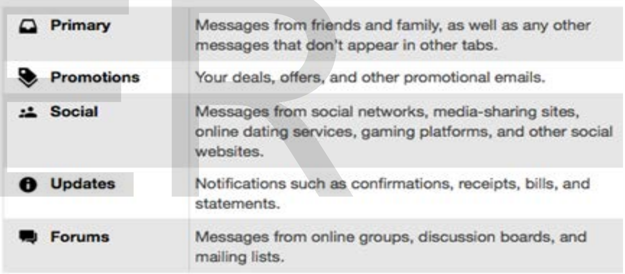
In today's competitive world information is most valuable assets. To sustain and compete in the real world people needs to exchange lots of information from various sources. Email is one of the widely used facility for information exchange, provided by email service providers with the help of internet. People can share any information to others around the globe by using emails. Increasing an internet speed in recent years also make drastic increase in use of an email.

According to an estimate [1], the number of email messages sent daily has reached around 3.4 billion in 2012, resulting in the evolution of problem called Email Overload. The incoming emails are of different types [13], proper handling of an individual email is required according to its importance.

Most of the email service providers provide the facility of categorizing the emails as legitimate and illegitimate i.e. Spam and non-spam. But to handle email overload problem there is necessity to organize incoming non-spam emails into different class or category or folders according to some criteria or condition present in email.

The well-known real time example of email categorization is Gmail's inbox [12], which categorizes incoming emails into different categories as shown in Figure 1. Gmail mainly categorizes incoming emails into four categories namely "Promotions", "Social", "Updates" and "Forums". The fifth category named "Primary" contains emails that are not fall under any of the above category.

The task of automatic handling of incoming emails is commonly known as email mining, email foldering, email categorization or email management. Machine learning algorithms are used for email management both in supervised and unsupervised manner also named as classification and clustering respectively in data mining community.



Primary	Messages from friends and family, as well as any other messages that don't appear in other tabs.
Promotions	Your deals, offers, and other promotional emails.
Social	Messages from social networks, media-sharing sites, online dating services, gaming platforms, and other social websites.
Updates	Notifications such as confirmations, receipts, bills, and statements.
Forums	Messages from online groups, discussion boards, and mailing lists.

Figure 1 Gmail's Categories for email management [12]

Naïve Bayes [3, 8], Support Vector Machine [3, 8], K-Nearest Neighbor [7, 8] from classification domain and K-Means[2, 9], Fuzzy C-Means[9], Evolving Email Clustering Algorithm[9] from clustering domain had been applied and used for email management.

This paper surveys many classification and clustering algorithms applied for email management. Section 2 describes different supervised algorithms and Section 3 describes different unsupervised algorithms.

2 CLASSIFICATION ALGORITHMS

Working of classification algorithms in email categorization is same as in spam detection, but the main difference between the two is number of classes that a classifier has to predict. In spam detection emails are classified in only two classes, "spam" and "non-spam". While in email categorization, emails usually have more than two classes, like "Work", "Social", "Entertainment" etc according to need of the user. Many classification algorithms have been used for this task

- *Jainesh is currently pursuing masters degree program in computer engineering in SVIT-Vasad, Gujarat, India, PH-08980439511. E-mail: jaineshpatel13@gmail.com*
- *Neha Soni is Assistant Professor at SVIT-VASAD ,Gujarat, India . E-mail: neha_ripal@yahoo.co.in*

and they are discussed here.

2.1 Naïve Bayes Classifier

Naive Bayes is the simple Bayesian classification algorithm and a statistical analysis algorithm, which works on numeric data.
Naive Bayesian classifier assumes that the effect of an attribute value on a given class in independent of the values of the other attributes. This assumption is called conditional independence. [4]. It considers email terms or words chosen at random from the total terms present in each folder. The words chance of being inserted is proportional to the probability of finding that word in the category being considered. The classifier determines the likelihood of conditional probability of particular email being inserted into one of the category, and inserts email into category that has the highest probability.
In Naive Bayes classifier [9], each class or category C¬j of emails is associated with the probability distribution of features in which the probability of a feature is proportional to its frequency in the category. For this TF-ID [3] scheme is used.
Given an email A represented as feature vector:
W= [w_1,w_2,w_3,…,w_n]
the probability that A belongs to category Cj is calculated as

$$\Pr(Cj|A) = \Pr(A|Cj)\,\Pr(Cj)\,/\Pr(A) \qquad (1)$$

Assuming each feature of A is drawn from the distribution associated with Cj independently, Pr(A|Cj) can be estimated as $\prod_i$ Pr(w_i|Cj). Thus to determine which class A should belong to, the class Cj is found with the maximum probability Pr(Cj) $\prod_i$ Pr(wi|Cj).
An incoming email will be assigned to a category that has the highest probability to insert it.
Bayesian classifiers have also exhibited high accuracy and speed when applied to large database [4].

2.2 Support Vector Machine Classifier

Support Vector Machines (SVM) uses nonlinear mapping to transform the original training data to a higher dimension. With an appropriate nonlinear mapping to sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors and margins [4].
The "one versus rest" methodology is applied to use the SVM classifiers to classify emails to more than two categories. For n (n>2) categories, the SVM classifier is applied n times. Each time, a SVM classifier decides whether the uncategorized email belongs to a certain category Cj or not, and the probability of the email to belong to the category is also calculated. For each email, the categories are ranked according to the probabilities that the email is assigned to.
Tang, Pei and Luk [3] had compared the performance of SVM classifier for email categorization. They concluded that number of emails does not affect classifier performance but number of categories does.

2.3 K-Nearest Neighbor Classifier

K-Nearest Neighbour is the simplest classification algorithm, in which input consists of K closest training examples in feature space and output depends on a class membership. An object is classified by a majority contribution of its neighbours, with the object being assigned to the class most common among its K-Nearest Neighbours. K-NN algorithm is sensitive to local structure of data [4].
Nearest neighbour simply treats the feature vector as a vector in n−dimensional space, and finds the nearest matching vector in terms of distance. This is calculated in the usual Pythagorean $a^2+b^2=c^2$ way, but generalised to n dimensions [8].
To find the closest objects a number of similarity measures are used among which the most popular is Euclidean distance calculated as,

$$D(pi, pj) = \sqrt{(xj - xi)^2 + (yj - yi)^2} \qquad (2)$$

Where, pi and pj represents the points or objects in space having coordinates (xi, yi) and (xj, yj) respectively.
The main strength of the KNN algorithm is that it provides good accuracy on many domains with the fast learning phase. But it is slow during instance classification because all the training instances have to be visited and the accuracy degrades with increase of noise in training data.
Matthew, Chung [7] uses KNN with resemblance for email categorization in which they considered phrases are in shingle and concluded that KNN gives nearly 97% accuracy which is far better than Naïve Bayes Classifier. Zhao and Jianguo [14] had compared sVSM, Bayesian and KNN classifier and concluded that KNN outperforms than other two.

3 CLUSTERING ALGORITHMS

Clustering algorithms are used for grouping of similar objects. They can be used for email management to reduce email overload by grouping of similar emails. K-means, Fuzzy C Algorithm, Email Evolving Clustering Method of clustering had been used for email management.

3.1 K-Means Clustering Algorithm

The K-Means algorithm takes input parameter k, and partitions a set of n objects into k cluster so that the resulting intracluster similarity is high but intercluster similarity is low [4]. Here, cluster similarity is measured as a mean of the objects in a cluster.
Application of K-Means clustering algorithm for email categorization works as:
Randomly two emails are selected from set of emails and they are assigned as initial cluster centers. Other emails are assigned to their closest cluster center. The degree of closeness can be decided with different distance metrics like Euclidian distance, Cosine Similarity etc.
After this setup there is need to find new cluster center which will be the arithmetic mean of all emails in existing clusters. Then for each new incoming email, it is compared with these

new cluster centers and is assigned to the one which has minimum distance.

Limitation of using K-Means clustering algorithm is that it is sensitive to initialization parameter and to the presence of outliers.

G. Soni and C.I.Ezeife [2] had compared K-Means, K-Means++, Kernel-selected and their own approach and results are shown.

3.2 Fuzzy C-Means Algorithm

In hard clustering like K-Means, data is divided into distinct cluster, where each data element is belong to exactly one cluster. In soft clustering like Fuzzy C-Means data elements can belong to more than one cluster, according to membership function.

Fuzzy C-Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. In Fuzzy C-Means, the centroid of a cluster is the mean of all data points, weighted by their degree of membership to the cluster. The membership is a number between 0 and 1. In Fuzzy C-Means each cluster is represented by a cluster prototype, the center of the cluster and the membership degree of an email to each cluster is measured with the distance between each email and cluster prototype [9]. Thus closer cluster prototype depends on closest email contents, which tends to greater membership degree of email in cluster.

Taiwo, Shikun and Rinat [9] had compared K-Means and FCM algorithms and concluded that FCM gives better results than K-Means for email management.

3.3 Evolving Email Clustering Algorithm

The Evolving Email Clustering Method (EECM) is developed with fuzzy inference system and separates the email input sample space based on similarity of email contents to create fuzzy rules [9]. The per-defined function is created based on contents of email messages to determine the group that the email belongs.

The EECM is distance based clustering algorithm, where group centers are denoted by evolved emails in the dataset. EECM provides good similarity measure. For any such group the maximum distance, between an sample point, which belongs to one group and is the farthest from this group centre, and its group centre, is less than or equal to a threshold value, which has been set as a grouping parameter. This parameter would affect the number of email groups to be created.

For each newly created group its group center is located and group radius is initially set with a value 0. After presenting incoming emails to already created groups, their group center's position and radius will be updated. The group will not be updated if its group radius will reach a special value called threshold value.

Taiwo, Shikun and Rinat [9] performed EECM clustering for email grouping and has been proven to be better algorithm in good performance than K-Means and FCM. They were able to achieve 98% accuracy in email grouping.

4 Conclusion

This paper surveys many classification and clustering algorithms used in email management. Email management itself is difficult task and there are many issues concerning with the email domain.

This paper surveys Naïve Bayes, Support Vector Machine and K-Nearest Neighbour algorithms from classification domain and K-Means, Fuzzy C-Means and Evolving Email Clustering Method from clustering domain with brief working methodology and performance scenario.

References

[1] Radicati, S., Hoang, Q.: "Email statistics report", 2012-2016. The Radicati Group, Inc., London (2012).

[2] Gunjan Soni, C.I. Ezeife., "An Automatic Email Management Approach Using Data Mining Techniques", Springer-Verlag Berlin Heidelberg (2013) pp. 260-267.

[3] Guanting Tang, Jian Pei, Wo-Shun Luk, "Email mining : tasks, common techniques, and tools", Springer (2013).

[4] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques 2nd Edition, Morgon Kaufmann, 2006.

[5] Soni, Gunjan, Electronic Theses and Dissertations, "An automatic email mining approach using semantic non-parametric K-Means ++ clustering" (2013)., Paper 4864.

[6] Katakis, Tsoumakas, Vlahavas, "E-mail Mining: Emerging Techniques for E-mail Management"

[7] Matthew Chang,Chung Keung Poon, "Using phrases as features in email classification", The Journal of Systems and Software, Elsevier (2009).

[8] Saurabh Khatri, Emmanuel M., "Review on Classification Algorithms in Email Domain", IJARS (2013).

[9] Taiwo Ayodele, Shikun Zhou, Rinat Khusainov, "Evolving Email Clustering Method for Email Grouping: A Machine Learning Approach", IEEE (2009)

[10] M. Hayat, J. Basiri, A. Shakery, "Content-Based Concept Drift Detection for Email Spam Filtering", 5th International Symposim on Telecommunications, IEEE (2010)

[11] Jose M., G. Castillo, M. Garcia, R. Bueno, "A Comparative Study on Feature Selection and Adaptive Strategies for Email Foldering", IEEE (2011).

[12] A new inbox that puts you back in control, http://www.gmailblog.blogspot.in/2013/05/a-new-inbox-that-puts-you-back-in.htm

[13] How Mature Is Your Email-Marketing Program? 14 Types Of Emails to Help Boost ROI,

[14] http://www.fulcrumtech.net/resources/14-email-types/

[15] Zhao Lu, Jianguo Ding, "An Efficient Semantic VSM based Email Categorization Method", International Conference on Computer Application and System Modelling (ICCASM 2010), IEEE(2010).