IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

Comparison of classification algorithms like Neural Network (NN), Support Vector Machine (SVM), and Naïve theorem (NB) and Back propagation TECHNIQUE for automatic email classification

Sathyabhama.N[1], Dr.Saravan.S.V[2] and Vengada Prabhama[3]

[1] **Department of MCA, SNS College of Technology, Coimbatore, Tamilnadu, PIN: 641035, India**

[2] **Principal, Christ the King College of Engineering, Coimbatore, Tamilnadu, PIN: 641035, India**

[3] **Department of MCA, SNS College of Technology, Coimbatore, Tamilnadu, PIN: 641035, India**

Abstract

This paper proposes a replacement email classification model using a supervised technique of multi-layer neural network to implement back propagation technique. Backpropagation adjusts the loads in Associate in Nursing quantity proportional to the error for the given unit (hidden or output) increased by the weight and its input. The coaching method continues till some termination criterion, like a predefined mean-squared error, or a most range of interations. Email has become one altogether the fastest and therefore the best styles of communication. However, the increase of email users with high volume of email messages might result in un-structured mail boxes, email congestion, email overload, unprioritised email messages, and resulted at intervals the dramatic increase of email classification management tools throughout the past few years. Our aim is to the use of empirical Analysis to select out Associate in Nursing optimum, novel assortment of choices of a users' email contents that modify the speedy detection of the foremost important words, phrases in emails.

Keywords: *Datamining, Classification, Support Vector Machine, Backpropagation,Neural network.*

1. Introduction

Backpropagation is the most popular of the MLP learning algorithms. The backpropagation algorithm can be defined as follows. For a test set, propagate one test through the MLP in order to calculate the output. Then compute the error, which will be the difference of the expected value and the actual value. Finally, backpropagate this error through the network by adjusting all of the weights; starting from the weights to the output layer and ending at the weights to the input layer. The backpropagation algorithm is a typical supervised learning algorithm, where inputs are provided and propagated forward to generate one or more outputs. Given the output, the error is calculated using the expected output. The error is then used to adjust the weights. There are two types of error functions for backpropagation. The first error function is used for output cells, and the second is used on for hidden cells. These functions are defined as follows respectively.

Note that in both equations, u is the output of the given cell, otherwise known as its activation. Y is the expected of correct result. Finally, w represents all of the weights connecting the hidden cell to all inputs cells. The activation or transfer function g to be used will be the standard sigmoid squashing function. While g represents the sigmoid, g' represents the first derivative of the sigmoid.

At this point, given our test input and expected result, we have the error calculated for each output and hidden node. The next step is to use this error to adjust the corresponding weights for the node. We will use the following equation for this purpose, which utilizes the error previously calculated for the node (whether hidden or output).

For the given error E and activation or cell output, ui, we multiply by a learning rate p and add this to the current weight. The result is a minimization of the error at this cell, while moving the output cell activation closer to the expected output.

Email has been an efficient and popular communication mechanism as the number of Internet users' increases. Therefore, email management has become a very important and growing drawback for people and organizations as a result of it's susceptible to misuse. One with all the issues that square measure most dominant is disordered email message, engorged and un-structured emails in mail boxes. It should be terribly arduous to search out archived email message, explore for previous emails with specific contents or options once the mails don't seem to be well structured and arranged.

Schuff et al expressed that "Emails are widely used to synchronize real-time communication, which is inconsistent with its primary goals". Email messages square measure designed to be sent, accumulate in the repository and be sporadically collected and browse by a receipt, that lends itself to the main points of a vacation or a meeting's forthcoming agenda. Since the general public think about emails for potency and effectiveness of communication, mail boxes could become engorged. Messages vary from static organization information to conversations with such a broad horizon of messages. Users could realize it tough to rank and with success method the contents of recent incoming messages. Additionally it should be tough to search out an antecedently archived message within the mailbox. Kushmerick expressed that "the ubiquitousness of email and its convenience as information management tools create it unlikely that users' behavior can amend as falling information measure and disk storage costs additional cut back the inducement to steer removed from victimization email as a document storage system". At this stage new effective technique for managing data in email, reducing email overloads is developed by classifying emails supported the importance of words within the email messages and etymologizing the nearest categories the mail might belong to either: crucial, urgent, vital, vital and not vital.

Email classification presents challenges owing to the giant and numerous variety of options within the data set and enormous variety of meals. Pertinency in email datasets with existing classification techniques was restricted as a result of the massive variety of options makes most mails indistinguishable. In several emails datasets, solely a tiny low share of the entire options could also be helpful in classifying mills, and victimization all the options could adversely have an effect on performance. The standard of coaching dataset decides the performance of each the e-mail classification algorithm and have choice algorithms. A perfect coaching dataset for every specific class can embody all the vital terms and their attainable distribution within the class. The classification algorithms like Neural Network (NN), Support Vector Machine (SVM), and Naïve theorem (NB) square measure presently utilized in numerous data sets and showing an honest classification result as experimented by Young . There are, however, very few studies in applying back propagation techniques (BPT) for email classifications. The most disadvantage of BPT is that they need hefty time for parameter choice and coaching. On the opposite hand, previous analysis has shown that back propagation in neural networks (NNs) is able to do terribly correct results, that square measure typically additional correct than those of the symbolic classifiers. NNs are with success applied in several planet tasks. During this paper we tend to gift BACK PROPAGATION TECHNIQUE a NNs-based system for automatic email classification.

2. Literature Survey

Email Classification Challenges:
The characteristics of emails take issue considerably and as a consequence email classification poses sure challenges, infrequently encountered in text or document classification. a number of the variations and challenges are:
1. Every users' mailbox is totally different and is consistently increasing. Email contents vary from time to time as new messages square measure intercalary and previous messages square measure deleted. A classification theme that may adapt to varied email characteristics is vital.
2. Manual classification of emails relies on personal preferences and thus the standards used might not be as straightforward as those used for text classification. This distinction must be taken into consideration by any technique planned for email classification.
3. The data content of emails varies considerably, and alternative factors, like the topic field, sender, CC field, BCC field, the person email is addressed to, play a very important role in the classification. This can be in

distinction to documents, that square measure richer in content leading to easier identification of topics or context.
4. Emails may be classified into folders and will even be classified into subfolders inside a folder. The variations within the emails classified to subfolders could also be a strictly linguistics (e.g., meeting with motion folder, conference expenses inside motion folder and plenty of more).

Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurones. This is true of ANNs as well.

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyse. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

Support Vector Machine

Support Vector Machines (SVMs), a new generation learning system based on recent advances in statistical learning theory. SVMs deliver state-of-the-art performance in real-world applications such as text categorisation, hand-written character recognition, image classification, biosequences analysis, etc., and are now established as one of the standard tools for machine learning and data mining. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane.

Naïve theorem (NB)

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Naïve Bayesian classifiers assume that the effect of an attribute value of a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved, in this sense is considered “naïve”

Major headings are to be column centered in a bold font without underline. They need be numbered. "2. Headings and Footnotes" at the top of this paragraph is a major heading.

3. Existing methods and drawbacks

There is very little exploration into the issues of categorizing and grouping emails into folders however less add classification of emails supported the activities of users (based on what the users do). One in all the common existing strategies used for email classifications is to archived messages into folders with a read of reducing the quantity of data objects a user should method at any given time. This can be a manual classification resolution, however, this {can be} AN inadequate resolution as folder names don't seem to be essentially a real reflection of their content and their creation and maintenance can impose a major burden on the user . Yukon planned a replacement email classification model employing a linear neural network trained by the Perception Learning algorithmic program (PLA) and a nonlinear neural network trained by Back Propagation Neural Network (BPNN). A linguistics Feature house (SFS) technique was additionally introduced during this classification model.

The disadvantages of rule-based system square measure that they're difficult for non technical users as a result of writing the principles need some level of programming expertise. Bifrost AN email classifier and an example email management system avoids this problem by belongings user outline all filtering rules with an easy graphical interface. Additionally planned a replacement approach by mechanically assessing incoming messages and creating recommendations before emails reach the user’s Inbox, that the priority system classifies every message as of either high or low importance supported its expected utility to the user. Whereas Yukun designed a system “that mechanically filter spam emails by victimization the principal partial analysis (PCA) and therefore the Self Organized Feature Map (SOFM). In their schema, every email is delineated by a series of matter and non-textual options. To cut back the quantity of matter options, PCA is employed to pick out the foremost relevant options. Finally the output of the PCA and therefore the non-textual options ought to be inputted into a well-trained SOFM to classify (spam or normal)" and in outline Boone describe Re: Agent system cluster similar messages supported existing folder structure provided by the user whereas it learns conception and sell policies for future message classification supported these folders examples.

Depending upon the mechanism used, email classification schemes may be loosely classified into: i) Rule primarily based classification, ii) data Retrieval {based|based mostly|primarily primarily based} classification and iii) Machine Learning based classification techniques.

Rule {based|based mostly|primarily primarily based} Classification: Rule based classification systems use rules to classify emails into folders. William Cohen] uses the murderer learning algorithmic program to induce 'keyword recognizing rules' for email classification. Ishmail is another rule-based classifier integrated with the Emacs mail program email.

Information Retrieval primarily based Classification:

Segal and Kephart have used the TF-IDF classifier because the means that for classification in SwiftFile, enforced as AN add-on to Lotus Notes. It predicts 3 probably destination folders for each incoming email. The TF-IDF classifier performs well within the absence of an oversized coaching set and additionally once the number of coaching information will increase, adding to the heterogeneousness of a folder.

Email Classification

Classification of text in AN email message is AN example of supervised learning that seeks to make a probabilistic model of a perform that maps emails to categories. In supervised learning of text in email messages, wherever a complete email dataset represents one example of emails to be classified, a learning algorithmic program is conferred with a group of already classified, or tagged, examples. This set is termed the coaching set. variety of classified emails from the coaching set square measure removed before model building to be used for testing the model's performance. This set is thought because the testing set. to raised live the classification accuracy of our model, many models square measure designed from totally different partitions of the examples to coaching and testing sets. The classification error is then averaged over every model. This method is termed n -times cross validation wherever "n" is that the variety of times the instance set is partitioned off. we tend to manufacture one thousand models for analysis victimisation this method and that we obtained 1000- times cross validation. currently our model has been created, it had been wont to predict the classification of future email messages. The accuracy of our models square measure mostly dependent on: The performance of our back propagation algorithmic program. The vital word choice victimisation data retrieval.

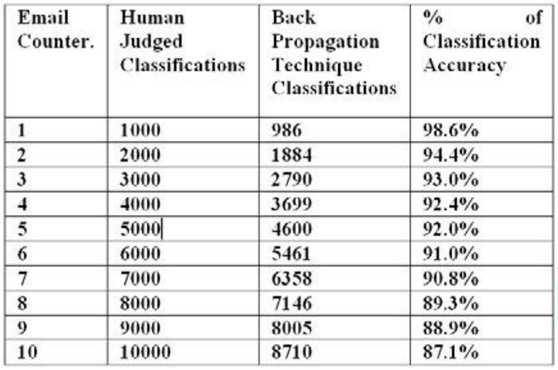
The "representativeness" of the coaching information with relevancy recently noninheritable email information to be classified.

The additional representative, the coaching information, the higher the performance. a bigger variety of coaching examples is commonly higher, as a result of a bigger sample is probably going to be additional reflective of the particular distribution of the information as a full.

Email Message Transformation

Classification machine learning algorithms care for numerical quantities as inputs. A tagged example is commonly a vector of numeric attribute values with one or more additional connected labels. For a few strategies, like naïve theorem, number counts of the attribute values in square measure all that's needed. Attributes in these cases may be nominal sorts however they're still ultimately born-again to numeric counts before process.
They're square measure several approaches to the current method – one in all the foremost common is to treat every email content as a "bag of words." every distinctive word constitutes AN attribute (a position in our example vector). The quantity of occurrences of a word in an exceedingly an email (frequency of occurrence) is that the attribute's price for that email message. Email messages square measure so delineated as vectors of numeric attributes wherever every attribute price is that the frequency of the prevalence of a definite term. This set of email message vectors is commonly stated as a vector house. Algorithms that care for such representations square measure aforesaid to be victimization vector house models of the information. Moreover, some sorts of words, or series of words, could also be desirable for learning. As an example, usually nouns or noun phrases square measure most popular. Part-of-speech identification algorithms and lexical/semantic dictionaries square measure usually wants to give extra data concerning terms. Also, quite common words like "and" and "the" square measure usually filtered out victimization stop word to enhance performance. All of those transformations square measure performed before any learning takes place. The quantity of steps concerned

during this pre-processing may be quite varied and infrequently constitutes the majority of the model building method.



Email Counter.	Human Judged Classifications	Back Propagation Technique Classifications	% of Classification Accuracy
1	1000	986	98.6%
2	2000	1884	94.4%
3	3000	2790	93.0%
4	4000	3699	92.4%
5	5000	4600	92.0%
6	6000	5461	91.0%
7	7000	6358	90.8%
8	8000	7146	89.3%
9	9000	8005	88.9%
10	10000	8710	87.1%

4. Conclusions

In this paper, we tend to study a way to generate correct email classes. We tend to analyze the characters of emails and study the e-mail spoken language structure, that we tend to argue haven't been sufficiently investigated in previous analysis on email classification exploitation back propagation technique. we tend to build a completely unique structure: Our classification relies on heuristic technique with the used of Term Frequency Inverse Document Frequency (TF-IDF) to work out what words in a very corpus of email messages may well be a lot of favourable to use in a very question, we tend to additionally implement a neural network primarily based system for machine-driven email classification into user outlined “word classes” Associate in Nursingd our BPT enforced was able to learn technique in an associative learning approach, during which the network is trained by providing it with input and matching output patterns.

References

[1] Schuff, D., O. Turetke, D. Croson, F 2007, ‘Managing Email Overload: Solutions and Future Challenges’, *IEEE Computer Society, vol. 40, No. 2, pp*. 31-36.
[2] Kushmerick, N., Lau, T. 2005, ‘*Automated Email Activity Management: An Unsupervised learning Approach’, Proceedings of 10th International Conference on Intelligent User Interfaces*, ACM Press, pp. 67-74.
[3] Helfman, J., Isbell, C. 1995, ‘Ishmail: Immediate Identification of Important Information’, AT&T Labs.
[4] Ramos, J. (2002). *Using TF-IDF to Determine Word Relevance in Document Queries,* Department of Computer Science, Rutgers University, Piscataway, NJ, 08855.
[5] Yun, F.Y., Cheng, H.L., Wei, S. (2008). *Email Classification Using Semantic Feature Space, Proceedings* of the 2008 International Conference on Advanced Language Processing and Web Information Technology, IEEE Computer Society Washington, DC, USA, pp.32-37.
[6] Yukun, C., Xiaofeng, L., Yunfeng, L. (2007). *An E-mail Filtering Approach Using Neural Network*, Springer Berlin, pp. 688-694.
Web address

1. http://www.cs.cmu.edu/~awm/10701/
2. http://www.cs.cmu.edu/~awm/10701/project/data.html
3. http://www.thearling.com/dmintro/dmintro_2.htm
4. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf