Enhancing E-mail Filtering Based on GRF

S.M. ELseuofi[1], W.A. Awad[2],S. A. El Hafeez[3] and R. M. El-Awady[4]

[1]Inf. System Dept., Ras El bar High inst.

[2]Math. &Comp.Sci.Dept. Science faculty, Port Said University

[3]Math. &Comp.Sci.Dept. Science faculty, Port Said University

[4]Electronic. &Communication.Dept.Faculty of Engineering MansouraUniversity

Abstract

The inferring of insignificant and repetitive features in the dataset can bring about poor expectations and misclassification process. Subsequently, selecting applicable feature subsets can help decrease the computational cost of feature measurement, accelerate learning process and enhance model interpretability.

Feature selection is an issue of worldwide computing optimization in machine learning in which subsets of relevant features are chosen to acknowledge powerful learning models. Rough sets Method in classification has demonstrated wasteful in its failure to deliver accurate and precise classification results about the large e-mail dataset while it likewise expends a ton of computational resources. In this study, we present GRF- Genetics Rough Filter-a hybrid of Genetic Algorithm-Rough set feature selection technique is developed to optimize the Rough set classification parameters, the prediction accuracy and computation time. Spam assassin dataset was used to validate the performance of the proposed system. GRF showed remarkable improvements over Neural Network, Rough Set and SVM methods in terms of classification accuracy.

Keywords: *E-mail classification, Genetic algorithm, Rough set, Machine learning.*

1. Introduction

E-mail has gotten to be greatly critical in our day by day life in view of high speed and low cost. Individuals are receiving an expanding measure of email both at work and also individual interchanges. Then again, we likewise get numerous messages from a lot of outsiders we don't have a clue. The vast majority of these messages are business promotion useless to the majority of us and frequently they are destructive e-mails containing viruses or malicious codes. These messages are called Junk email or Spam. Email spam targets singular clients with standard mail messages. On account of the ease connected with sending messages, spammers have the capacity send a huge number of messages every day over the web. Email addresses on spam records are frequently made by checking Usenet postings, taking Internet mailing records, lexicon assaults or hunting the web down locations, among others. Spam filtering is an automatic classification of incoming e-mail messages, to permit the exclusion of spam from incorporation in a group of legitimate messages for a particular user. Eras of spam filters have risen through the years to manage the spam issue. A large portion of these filters succeeded to some point in separating between spam and legitimate e-mails, however they oblige manual intercession. For instance content based methods oblige human endeavors to assemble lists of attributes and their scores. In the course of the most recent years, statistical filters have picked up more consideration as they find themselves able to change themselves; getting better and better with less manual intercession. The most popular statistical approach is the Bayesian filter, which appoints likelihood evaluations to messages. Even such filters have their limits as spammers still figure out how to avoid them by utilizing different abusing strategies [2]. Thus, novel methodologies are sought to manage continually expanding surge of spam and the tireless endeavors by spammers to break the current anti-spam filters. Information Filtering and Information Retrieval is quickly acquiring importance as the volume of electronically stored data becoming huge. E-mail Filtering is an important part of information filtering in that it categorizes emails within emails Data Set. On the other hand, a non-trivial obstacle in good email filtering is the huge amount of the data In most Information Retrieval techniques, each one email is described by a vector of extremely high dimension array of data, typically one value per word or pair of words in the message [1]. The vector coordinates are utilized as preconditions to a rule which decides what class the email belongs to. Email vectors ordinarily contain countless of dimensions [4], which renders the problem all but intractable for even the most powerful computers. The utilization of the cosine angle between two vectors [6] as a correlation metric further expands the quantity of operations to be performed for the classification of one email. This paper proposes a technique using Genetics-Rough Set Theory that can help cope with this situation. Given Data-Set of emails and a set of examples of classified emails, the method can rapidly place a minimal set of co-ordinate keywords to recognize classes of emails messages. Therefore, it significantly reduces the dimensionality of the keyword space. The resulting set of keywords (or preconditions) is commonly sufficiently enough to be understood by a human. This simplifies the production of knowledge-based systems, allowing easy editing of the rule bases.

2. Related work

There are some research work that applies machine learning methods in e-mail classification, M. Dredze, J. Blitzer, and F. Pereira. Used a rule based system to predict reply labels (needs reply, does not need reply) [15]. In this system they used relational features that rely on a user profile which included the number of sent and received emails from and to each user as well as the user's address book, rough set role indication email address and domain Document-specific features were the presence of question marks, request indicators such as question words (weighted using tf-idf scores), presence of attachment, document length, salutations, and the time of day. The system was tested on 2,391 manually labelled emails, coming from 4 students. On average it obtained a precision of 0.73 and recall of 0.64. Zhou, Bing.In this paper, a proposed model of multistage three-way email spam filtering based on principles of granular computing and rough sets [16]. three-way decision strategy used to filtering spam E-mails in which it divides incoming emails into three folders, namely, a mail folder consisting of emails that they accept as being legitimate, a spam folder consisting of emails that they reject as being legitimate, and a third folder consisting of emails that they cannot accept nor reject based on available information. The introduction of the third folder enable the system to reduce both acceptance and rejection errors. Many existing ternary approaches are essentially a single-stage process. O. Stephen, and A. Abimbola. They have combined the Genetic algorithm with the SVM to enhance the performance of SVM [17]. In its simplest form SVM can be used to represent a document in vector space where each feature (word) represents one dimension. Identical feature denotes same dimension. SVM did an acceptable performance after hybrid with the GA, while SVM computational time still needs to be improved. Boratyn, Grzegorz M.proposed a new method of Feature Selection for signal-like data. BSS-based extraction a new features reduces dimensionality and simplifies the attribute selection problem in the original space [18]. The classification is done in this paper by using the rough set method. Thus give the algorithm additional advantage is the possibility of analysis of the results by using any other kind of data, more effort need to be done to get satisfactory results. Thomas, Anju, D. Sugumar, and P. T. Vanathi. Proposed In this paper, a new algorithm that improves the quality of source separation by using dictionary learning technique for multichannel observations in both noisy and noiseless situations [19]. The dictionary is learned using single value decomposition algorithm and the result shows that the recovered image sources are more accurate. Luo, Qin, et al.proposes a method to optimize spam filtering rules using neural network and describes the design and implementation of an anti-spam system using the optimized rules [20]. Their system can automatically extract and learn the features of the mails and make dynamic adjustments to static rules. They compare the performance of our system with a famous rule-based spam filter-Spam Assassin and it is shown that our system has a better filtering performance.

3. Machine Learning algorithms

3.1. Brief Introduction to Rough Set Theory

Rough Set "RS" method has a great ability to process the decreases of data frameworks. In a data framework there may be a few attributes that are insignificant to the target idea (decision attribute), and some repetitive attributes. Reduction is expected to create straightforward helpful knowledge from it [21]. A reduction is the vital piece of a data framework. It is a minimal subset of condition attributes with respect to decision attributes. The Rough set theory is given as follows. A data framework is a couple S =< U,A >, where U = {x1, x2, ...xn} is a nonempty situated of items (n is the number of objects); A is a nonempty set of attributes, A = {a1, a2, ...am}(m is the number of attributes) such that a : U →Va for every a   A. The set Va is called the value set of a. A decision system is any information system of the form L = (U,A {d}), where d is the decision attribute and not belong to A. The components of A are called conditional attributes. Let S = < U, A > be an information system, then with any B ⊆ A there is associated an equivalence relation INDS(B): INDS(B) = $\{(x, x')^{\in}\ \mathrm{U}^2 | \forall a^{\in}\ Ba(x) =$

$a(x')\}$ INDS(B) is called the B-indiscernibility relation. The equivalence classes of B-indiscernibility relation are denoted [x]B. The objects in $\underline{B}X$ can be certainly classified as members of X on the basis of knowledge in B, while the objects in $\overline{B}X$ can be only classified as possible members of X on the basis of knowledge in B. Based on the lower and upper approximations of set $X \subseteq U$, the universe U can be divided into three disjoint regions, and we can define them as: $POS(X) = \underline{B}X$, $NEG(X) = U - \overline{B}X$, $BND(X) = \overline{B}X - \underline{B}X$ The equivalence classes of B-indiscernibility relation are denoted [x]B.

3.2. Brief Introduction to Genetic Theory

Genetic Algorithms "GA" is the type of algorithms that focused on and is routinely used to create useful solutions to optimization and search problems. The most fundamental idea is that the solid chromosome has a tendency to adjust and survive while the frail chromosome have a tendency to cease to exist That is, solution optimization is focused on natural evolution, and the Survival of the fittest concept. It have the ability to generate an initial population of practical solutions, and afterward recombine them in a special way to guide their hunt to just the most guaranteeing regions of the proposed solution. Every practical solution is can be present as a chromosome, and every chromosome is given a mean of fitness function and can be called as chromosome fitness. The chromosome fitness gives us indication and let us to decide its ability to survive and produce offspring. Probabilistic rules used by genetic algorithm in order to advance a solution from one generation to the next. The new solutions of generations are produced by genetic recombination operators like, First: Reproduction which is selecting the fittest to chromosome, Second: Crossover which is combining parent chromosomes to produce children chromosomes also Crossover combines the "fittest" chromosomes and passes the best genes to the next generation, Third: Mutation which is adding some genes in a chromosome, also Mutation guarantees the whole data set will be searched, and can drive the population out of the bad performance. The Most Important Parameters in GAs are Population Size, Evaluation Function, Crossover Method and Mutation Rate. Genetic algorithms are not generally utilized for classification issues straightforwardly because of the way of the algorithm. GAs models the principles of evolution and natural selection to rapidly look through a vast space of answers for an issue. Through crossover and mutation of the candidate solutions, potentially better solutions can be discovered [9]. The principle purpose of control for a GA is the fitness function. This function is designed for each problem given to the GA, and rates the solutions that it comes up with. This score is then utilized as part of the natural selection process. The fitness function is an extremely instinctive method for specifying the desired properties of the search results. While not straightforwardly used as a classifier, it can possibly help a hybridized classification system. The use of a GA has not been studied in email classification; however it can possibly be utilized as a part of hybrid system with Rough sets.

4. E-mail preprocessing

4.1. The structure of an email

In addition to the body message of an email, an email has another part called the header. The job of the header is to store data about the message and it contains numerous fields like the field (From) and (Subject), we decided to divide the email into 3 separate parts. The first part is the (Subject) that can be considered as the most important part in the email, it perceived that the majority of the new approaching messages have clear Subjects that can be utilized to unmistakably recognize whether that email is Spam or Ham. The second part is (From) which is the individual that taking the responsibility of the message, this field we store it in a database and use it after the decision of the classifier has been taken, that is the way to compare the field (From) stored in the database to the field (From) in the new incoming email, if they are the same so the decision of the new incoming email is Spam. The third part is the (Body) which is the main part of the message. Besides we applied two techniques in the preprocessing stage. Stopping is employed to remove common word. Case-change is employed to change the (Body) into small letters.

4.2. Feature Construction:

In this paper, our proposed methodology has two primary machine learning algorithms inserted in it; a genetic algorithm and Rough sets. It uses a hybrid approach to classification; the GA is used to search for a subset of features that would be best for the rough set to learn from. Because of the complexity of these algorithms, modifying and fine−tuning their parameters is not an exact science. The GA is used for feature selection, and will choose words which give The Rough set the best information about the different classes being used. This should make the Rough set quicker to train, as rough set training time is heavily dependent on the dimensionality of the inputs. It might also aid in the accuracy of the rough set predictions, having wiped out the less valuable words from the data vector. The features that are referred to from here on are words found inside the email collection. A common and intuitive representation found in text categorization is called the "List of words" representation. With this, the contents of an email are represented by refining the message into a table linking unique words to their relative frequencies. This is the representation that has been chosen for use in the hybrid classifier, since it is the technique most turned out to be viable. Natural language processing methods have far to go before they practical for use in email classification; more benefit can be gotten using the statistical analysis of individual words found in email..

4.3. Initial features Reductions

The system uses a genetic algorithm as the essential method for feature selection. However during development it rapidly got to be evident that tossing such a substantial number of features at the GA prompted to poor execution. The GA was not able to give back a reasonable feature selection given the 9000 or so candidate words. so as to give the GA some assistance, some initial reduction of the feature size needed to be done before the GA was allowed to do its work. The first step to reduce the huge number of the features was to remove words that only just seen once in the data set. These words are regularly useless random strings that can show up in various types of attachment data and html formatted email. It also includes a lot of numbers. Note that few numbers can at present be significant, such as college course codes for instance, so numbers are not eliminated entirely. When this step is carried out, it reduces a unique word count of order 9000 down to an order of 3000. A second useful step is to remove amazingly long words. Using the default configuration, words found in the body of the message that are bigger than 20 characters long are removed. This is only done for the body of the message since email addresses in the headers are ordinarily longer than 20 characters. This above and beyond diminishes of unique words down to around 1000. The last reduction step is to sort the remaining words according to the criteria selected, which will be word’s variance. After the list is sorted, the system cuts off the main 640 for the GA to work with. This is so that the GA can focus on the possible arrangements of words for the feature selection, and so it does not need to waste time considering low−value words. It is critical to pick this number to be a few times bigger than the size of the final feature selection; else the GA has nothing it can do. For example, when selecting 64 words, a top−selection of 320 is satisfactory. When the GA must select 256 words, a top−selection of 640 is more appropriate.

4.4. Chromosome structure

Each word in the e-mail can be represented as one bit in a binary chromosome for the GA. Hence the length of the chromosome is the length of the master index of word to consider, 640 in this situation. If a bit is set to 1, the selection represented by this chromosome includes this word, if it is 0, the selection does not include it. By recombining different sections of these bit strings, called crossover in GA terms, we can come up with new, potentially better selections. Chromosomes are scored according to a fitness function. Higher-scoring chromosomes are more likely to survive and are able to reproduction future generations. Reproduction includes mating with other chromosomes and exchanging genes via a 'crossover' function. Stronger chromosomes are more likely to be chosen to reproduction. One way to select parents is 'Roulette Wheel' selection. Mutation may occur, where a random change is made to the chromosome. Many mutations may be damaging, but some could enhance the ‘health’ of the chromosome..

5. Algorithms Implementation

5.1. Genetic Algorithm

```
1. Generate the number of generations = 10
2. Read the spam and ham Data Set
3. Mix the lines of spam randomly
4. Divide spam Data Set into 10 slices
5. Loop until 10th generation:
      a. Generate Number chromosomes based on
      the current slice of spam
      b. Score chromosomes
      c. Crossover the parents to form new
      offspring (child)
      d. Print results for the current generation
      e. Keep the fittest
      f. reproduction survivors
            I. 2 survivor's reproduction via a
            crossover function to create a child
            II. Find the sum of all
      chromosomes fitness in the population
            III. Use ‘Roulette Wheel’ selection
            top choose the 2nd parent
      f. Mutate some of the children by randomly
      deleting some genes
      g. Move to next slice of spam
6. Print Final results
```

5.2. Rough sets for spam filtering

In order to allow the straightforward application of rough set theory to classify the incoming email according to the feature vector that are getting out from the genetics algorithm we perform a decision table that, each received chromosome is represented by a set of words W = {W_1, . . ., $W_{n−1}$} together with its corresponding message class or Spam decision attribute D = {a_n}. Therefore, this feature vector containing all the terms existing in the Data set plus the class attribute stands for the attribute set A = W ∪ D = {W_1, W_2, . . ., $W_{n−1}$,W_n}. For Simplification purposes, Table 1 shows an example Data set containing a total count of six Chromosomes (m = 6) and 8 words (n = 8). In Table 1, chromosomes are represented as a feature vector in which the value assigned to each attribute Wi belonging to {W_1, . . ., $W_{n−1}$} is 1 when the message contains the term Wi, and 0 otherwise. Likewise, the value for the decision attribute (Spam Decision), W_n, is 1 for spam messages and 0 for legitimate ones. Therefore, a decision table is a pair S = (U, W), where U is a non-empty and finite set called the universe (e.g. all the chromosomes included in the Data set represented in Table 1), and W is the non-empty and finite set of Words previously defined. Using the example previously introduced in Table 1, A = C ∪ D = {100%free, Hot, Increase, offer, Urgent, Earn $, Cheap, Certified} ∪ {class}. By means of this characterization we define an equivalence relation, called indiscernibility relation, associated with every subset of attributes P ⊆ A. This relation is defined as shown in Expression (1).IND(P) = {(x, y) ∈ U2 : ∀ w ∈ P, w(x) = w(y)} (1)

Expression (1) establishes that, considering the attributes included in P, a chromosome x is indistinguishable from another one y (x, y ∈ U) if, and only if, they share the same values for all the attributes w_i included in P. By using the indiscernibility relation IND (P) from the set of attributes P, we can define the set of equivalence classes (basic categories) denoted by U/IND (P) or U/P. For instance, and considering P = {W_6, W_7, W_8} from Table 1, U/IND (P) = {{Chromosome_1}, { Chromosome_2}, { Chromosome_3, Chromosome_6}, { Chromosome_4}, { Chromosome_5}}. Equivalence classes defined through IND (P) are called basic categories of knowledge P, and are denoted by [x] IND (P). Therefore, emails e3 and e6 in our example are indistinguishable. Given a decision table S = (U, P), any set X ⊆ U can be defined by the use of two sets, called lower and upper approximations. The lower approximation, denoted by $\underline{P}X$, is the set of elements in U which can be classified with full certainty as elements of X using the set of attributes P, and is formally represented in Expression (2). $\underline{P}X$ = ∪{Y ∈ U/IND (P): Y ⊆ X} (2)in the example of Table 1 and using P = {W_7, W_7, W_8}, the lower approximation of set X = { Chromosome_4, Chromosome_6} is $\underline{P}X$ = { Chromosome_4}. Alternatively, the upper approximation, denoted by $\overline{P}X$ is the set of elements in U which can be possibly classified as elements in X. Expression (3) contains the definition of this concept. $\overline{P}X$ = ∪ {Y ∈ U/IND (P): Y ∩ X ≠ø} (3)in the example showed in Table 1, $\overline{P}X$ = { Chromosome_3, Chromosome_6, Chromosome_4}. A set X is rough regarding P if, and only if, $\overline{P}X \neq \underline{P}X$ Through the utilization of upper and lower approximations, we can define the positive, negative and borderline regions for a set X as respectively shown in Expression (4).

POSP (X) = $\underline{P}X$, NEGP (X) =U - $\overline{P}X$, BNDP (X) =$\overline{P}X$ - $\underline{P}X$ (4)

```
Begin
Read the Chromosomes
Split the Chromosomes from blank space
Insert all words in database
FOR each Word in the array
 Set X=0
        Select all the words present in the database
          WHILE words present in the database
            IF words in array matches database
            THEN
            Set W_fnd equals to the W_fnd of that
            word in the database and Set X= 1
            END IF
 END FOR

Initialize W1, W2, W3, W4, W5, W6, W7, W8, and
Spam equals to zero
    FOR i = 0 to Chromosome length
     IF Word equals “100%free” THEN
                SET $W_1$ = 1
                ELSE
       IF Word equals “Hot” THEN
                SET $W_2$ = 1
                ELSE
        IF Word equals “Increase” THEN
               SET $W_3$ = 1
                ELSE
         IF Word equals “Offer” THEN
               SET $W_4$ = 1
               ELSE
          IF Word equals “Urgent” THEN
               SET $W_5$ = 1
                ELSE
            IF Word equals “Earn $” THEN
               SET $W_6$ = 1
```

```
                    IF Word equals “Cheap” THEN
                             SET $W_7 = 1$
                             ELSE
                        IF Word equals “Certified” THEN
                            SET $W_8 = 1$
                          END IF
                   END IF
                END IF
              END IF
           END IF
        END IF
   END FOR
Insert the values of $W_1$, $W_2$, $W_3$, $W_4$, $W_5$, $W_6$, $W_7$, $W_8$
in the database.
       Display the dataset table
Stop
```

5.3. Rough Set Rules Generations:

By applying the previous code algorithm we get the following Rules:

Rule_1: 100%free = 0 && Hot = 0 && Increase = 0 && offer = 1 && Urgent = 1 && Earn $ = 1 && Cheap = 1 && Certified = 0 >>>>>>> Spam = 1.

Rule_2: 100%free = 0 && Hot = 1 && Increase = 1 && offer = 0 && Urgent = 0 && Earn $ = 0 && Cheap = 0 && Certified = 0 >>>>>>> Spam = 0.

Rule_3: 100%free = 1 && Hot = 1 && Increase = 1 && offer = 0 && Urgent = 0 && Earn $ = 0 && Cheap = 0 && Certified = 1 >>>>>>> Spam = 0.

Rule_4: 100%free = 0 && Hot = 0 && Increase = 0 && offer = 1 && Urgent = 0 && Earn $ = 0 && Cheap = 0 && Certified = 0 >>>>>>> Spam = 1.

Rule_5: 100%free = 0 && Hot = 0 && Increase = 0 && offer = 0 && Urgent = 0 && Earn $ = 1 && Cheap = 0 && Certified = 0 >>>>>>> Spam = 1.

Rule_6: 100%free = 1 && Hot = 1 && Increase = 0 && offer = 1 && Urgent = 0 && Earn $ = 0 && Cheap = 0 && Certified = 1 >>>>>>> Spam = 0.

6. Experiment Implementation

So as to test the execution of above GRF, we have to use data set of spam and legitimate emails, there are several collections of email freely accessible available to be used by researchers. SpamAssassinwill be used in this experiment "http://spamassassin.apache.org”, which contains 6000 emails with the spam rate 37.04%. Consequently we have divided the Data Set into training and testing sets keeping, in every such set, the same proportions of ham (legitimate) and spam messages as in the original example set. Each training set produced contained 62.96% of the original set; while each test set contain 37.04% as Table 1. In addition to the body message of an email, an email has another part called the header. The job of the header is to store information about the message and it contains many fields like the field (From) and (Subject), we decided to divide the email into 3 different parts. The first part is the (Subject) that can be considered as the most important part in the email, it noticed that most of the new incoming emails have descriptive Subjects that can be used to clearly identify whether that email is Spam or Ham.

The second part is (From) which is the person that taking the responsibility of the message, this field we store it in a database and use it after the decision of the classifier has been taken, that is the way to compare the field (From) stored in the database to the field (From) in the new incoming email, if they are the same so the decision of the new incoming email is Spam. The (Body) is the third part which is the main part of the message. Furthermore we applied two procedures in the preprocessing stage. Stopping is employed to remove common word. Case-change is employed to change the (Body) into small letters. The experiment is performed with the most frequent words in spam email.

Table 2: Data Set of Spam and Ham E-mails

Message collection	**Training Set**	**Testing Set**
Ham E-mails	2378	1400
Spam E-mails	1398	824
Total E-mails	3776	2224

Table 1: Example containing 6 chromosomes, 8 words and Decision

U	**Word_1**	**Word_2**	**Word _3**	**Word _4**	**Word _5**	**Word_6**	**Word _7**	**Word_8**	**Spam**
	100%free	Hot	Increase	Offer	Urgent	Earn $	Cheap	Certified	**Decision**
Chromosome_1	0	0	0	1	1	1	1	0	1
Chromosome_2	0	1	1	0	0	0	0	0	0
Chromosome_3	1	1	1	0	0	0	0	1	0
Chromosome_4	0	0	0	1	0	0	0	0	1
Chromosome_5	0	0	0	0	0	1	0	0	1
Chromosome_6	1	1	0	1	0	0	0	1	0

5.4. Performance evaluation

In order to test the performance of above mentioned methods, we used the most popular evaluation methods used by the spam filtering researchers. Spam Precision (SP), Spam Recall (SR), Accuracy (A). Spam Precision (SP) is the number of relevant documents identified as a percentage of all documents identified; this shows the noise that filter presents to the user (i.e. how many of the messages classified as spam will actually be spam)

$$A = \frac{\text{\# of e-mails correctly categorized}}{\text{Total \# of e-mails}} = \frac{N_{ham\rightarrow ham} + N_{spam\rightarrow spam}}{N_{ham} + N_{spam}}$$

Accuracy (A) is the percentage of all emails that are correctly categorized Where Nham→ham and Nspam→spam are the number of messages that have been correctly classified to the legitimate email and Spam email respectively; Nham→spam and Nspam→ham are the number of legitimate and spam messages that have been misclassified; Nham and Nspam are the total number of legitimate and spam messages to be classified.

$$SP = \frac{\text{\# of Spam Correctly Classified}}{\text{Total \# of messages}} = \frac{N_{spam\rightarrow spam}}{N_{spam\rightarrow spam} + N_{ham\rightarrow spam}}$$

$$SR = \frac{\text{\# of Spam Correctly Classified}}{\text{Total \# of messages}} = \frac{N_{spam\rightarrow spam}}{N_{spam\rightarrow spam} + N_{spam\rightarrow ham}}$$

Spam Recall (SR) is the percentage of all spam emails that are correctly classified as spam.

5.5. Performance Comparison

In order to do performance comparison of the proposed Hybrid system we run the same data onto three different machine learning algorithms. We summarize the performance result of the presented method in term of spam recall, precision and accuracy. Table 3 and Figure 1 summarize the results of the classifier. In term of accuracy we can find that the GRF Algorithm is the most accurate while the Rough Set Algorithm give us the lower accuracy, Support Vector Machine System and the Neural Network give us approximately the same lower percentage, while in term of spam precision we can find that the Neural Network method has the highest precision among the three algorithms while the Support Vector Machine has the worst precision percentage and the GRF Algorithm has a very competitive percent, and finally we can find that the recall is the less percentage among the three classifiers while the GRF Algorithm still has the highest performance but considered low when compared to accuracy while the Rough Set has the worst performance.

7. Conclusion and Future work

From the results of this study, it seems that there is considerable merit in using a hybrid approach to email classification GRF. By comparing performance with and without the aid of the GA for feature selection, it has been found that the proposed system GRF approach can attain the high accuracies needed for the intelligent filtering of email. GA is applied to optimize the feature subset selection and classification parameters for Rough set classifier. It eliminates the redundant and irrelevant features in the dataset, and thus reduces the feature vector dimensionality. This helps rough set to select optimal feature subset from the resulting feature subset. The resultant system achieves higher recognition rate using only few feature subset. GRF has shown a significant improvement over SVM and Neural Network Classifiers in terms of classification accuracy. Additionally, some work should be also carried out to address noise handling. Thus, we think that features should be regular expressions instead of words to handle noise and common misspellings.

Algorithm	Spam Recall (%)	Spam Precision (%)	Accuracy (%)
GRF	98.46	97.80	99.66
RS	92.36	94.56	94.7
SVM	95.00	93.12	96.90
NN	97.14	98.66	97.80

Table 3. Performance of three machine learning algorithms Compared with GRF

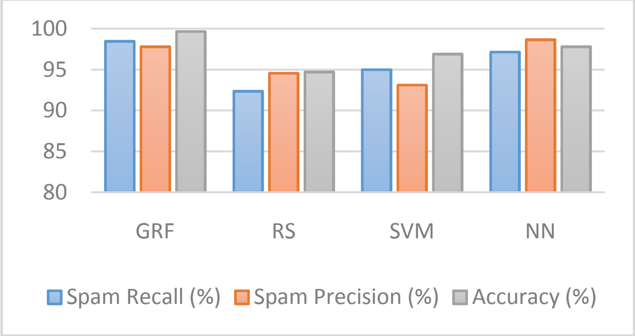

Figure 1. Spam Recall, Spam Precision and Accuracy curves of three classifiers

REFERENCES

[1] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, “Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis”, IET Computers & Digital Techniques, 2008

[2] Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali “Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification” Elsevier Computer Networks, 2009

[3] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch ”Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis” IEEE GLOBECOM, 2008

[4] Guzella, T. S. and Caminhas, W. M. ”A review of machine learning approaches to Spam filtering.” Expert Syst. Appl., 2009

[5] A.H. Mohammad, R.A. Zitar, Application of genetic optimized artificial immune system and neural networks in spam detection, Applied Soft Computing 11 (4) (2011) 3827–3845.

[6] El-Sayed M. El-Alfy, Radwan E. Abdel-Aal "Using GMDH-based networks for improved spam detection and email feature analysis"Applied Soft Computing, Volume 11, Issue 1, January 2011

[7] Li, K. and Zhong, Z., “Fast statistical spam filter by approximate classifications”, In Proceedings of the Joint international Conference on Measurement and Modeling of Computer Systems. Saint Malo, France, 2006

[8] Cormack, Gordon. Smucker, Mark. Clarke, Charles " Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011

[9] Almeida,tiago. Almeida, Jurandy.Yamakami, Akebo " Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers" Journal of Internet Services and Applications, Springer London , February 2011

[10] Yoo, S., Yang, Y., Lin, F., and Moon, I. “Mining social networks for personalized email prioritization”. In Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Paris, France), June 28 - July 01, 2009

[11] Cormack, Gordon. Smucker, Mark. Clarke, Charles " Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011

[12] S.X. Wu, W. Banzhaf, The use of computational intelligence in intrusion detection systems: a review, Applied Soft Computing 10 (1) (2010) 1–35.

[13] G. Lai, C. Chen, C. Laih, T. Chen, A collaborative anti-spam system, Expert System with Applications: An International Journal 36 (3) (2009) 6645–6653.

[14] Ishibuchi, H., & Nakashima, T. (2000), “Multi-objective pattern and feature selection by a genetic algorithm”, Proc. of Genetic and Evolutionary Computation Conference (Las Vegas, Nevada, U.S.A.), 1069-1076.

[15] M. Dredze, J. Blitzer, and F. Pereira. “Reply expectation prediction for email management”. In The

Second Conference on Email and Anti-Spam (CEAS), Stanford, CA, 2012.

[16] Zhou, Bing. "Multi-class decision-theoretic rough sets." International Journal of Approximate Reasoning , 2014.

[17] O. Stephen, and A. Abimbola, “Hybrid GA-SVM for Efficient Feature Selection in Email Classification”, vol. 3, no. 3,IISTE, 2012.

[18] Boratyn, Grzegorz M., et al. "Hybridization of Blind Source Separation and Rough Sets for Proteomic Biomarker Indentification." Artificial Intelligence and Soft Computing- Springer Berlin Heidelberg, 2010. 486-491.ICAISC 2010.

[19] Thomas, Anju, D. Sugumar, and P. T. Vanathi. "Blind Image Source Separation based on MMCA using Dictionary Technique." International Journal of Advanced Research in Electronics and Communication Engineering , 2013.

[20] Luo, Qin, et al. "Design and Implement a Rule-Based Spam Filtering System Using Neural Network." Computational and Information Sciences (ICCIS), 2011 International Conference on. IEEE, 2011.

[21] Glymin, Mawuena, and Wojciech Ziarko. Rough set approach to spam filter learning. Springer Berlin Heidelberg, 2007.