Toward Accurate Feature Selection Based on BSS-GRF

S.M. ELseuofi
Inf. System Dept. Ras El bar High inst., Damietta, Egypt

Wael Awad
Math. & Comp. Sci. Dept. Science faculty, Port Said University

Samy Abd El -Hafeez
Math. & Comp. Sci. Dept. Science faculty, Port Said University

R. M. El-Awady
Electronic. & Communication Dept. Faculty of engineering Mansoura University

***Abstract*—as of late, Feature extraction in email classification assumes a vital part. Many Feature extraction algorithms need more effort in term of accuracy. In order to improve the classifier accuracy and for faster classification, the hybrid algorithm is proposed. This hybrid algorithm combines the Genetics Rough set with blind source separation approach (BSS-GRF). The main aim of proposing this hybrid algorithm is to improve the classifier accuracy for classifying incoming e-mails.**

Keywords—rough set; Genetic; blind source separation; E-mail Filtering; Machine Learning

I. INTRODUCTION

Because of the expanding volume of undesirable email called as spam or Junk email, the clients and in addition Internet Service Providers are confronting a considerable measure of issues. Email spam additionally makes a significant issues to the security of arranged frameworks. Email classification is able to control the issue in a mixture of ways. Recognition and assurance of spam messages from the email delivery system permits end-clients to recapture a helpful method for correspondence. Many researches on content based email classification have been focused on the more complex classifier-related issues [4]. Presently, machine learning for email classification is a critical exploration issue. The accomplishment of machine learning methods in text content Classification has driven scientists to investigate learning algorithms in E-mail classification [6]. On the other hand, it is astounding that in spite of the expanding advancement of e-mail filtering innovations, the quantity of spam messages keeps on increasing quickly. Subsequently, novel methodologies are wanted to manage ever- expanding surge of spam and the industrious endeavors by spammers to break the current hostile to spam boundaries. Eras of spam filters have risen through the years to manage the spam issue. The greater part of these filters succeeded to some point in separating in the middle of spam and genuine messages, nonetheless they oblige manual intercession. For instance content based methods oblige human endeavors to fabricate arrangements of attributes and their scores [11]. Recently, statistical filters have picked up more consideration as they find themselves able to change themselves; improving and better with less manual intercession. The most mainstream measurable methodology is the Bayesian filter, which relegates probability evaluations to e-mails [12]. At the point when managing substantial scale datasets, it is regularly a reasonable need to look to decrease the span of the dataset, recognizing that by and large the examples that are in the information would at present exist if a delegate subset of cases were chosen. Further, if the privilege examples are chosen, the lessened dataset can frequently be less boisterous than the first dataset, creating predominant speculation execution of classifiers prepared on the decreased dataset. The quest for spam words in the approaching email can be seen as a feature selection problem that can be formed as takes after: Given N data points $x_i \in R^n$, $i = 1,\ldots\ldots,N$, with labels $y_i \in \{-1,1\}$ select an L-element subset of features $\{x_{ik} | K \in S, S \subseteq 1,\ldots\ldots,N\}$ while protecting or potentially enhancing discriminative capacity of a classifier. The quantity of pertinent features L is normally picked subjectively. The e-mails can be translated as signs in the Universe (U) that can be differentiated into statistically independent segments. Sizes of those parts indicated by ai represent the original points xi in a new feature space. Dimensionality of ai's is generally much littler than dimensionality of xi's making classification and feature selection problem easier. The new feature set will then be reduced to attributes relevant to the given classification. Each one property of ai is connected with a processed component that is still in the Universe (U) [15]. In this way, important features point to applicable parts where contrast between messages in Universe (U) can be watched. Optima of those parts for a specific class show values higher or lower than normally. The following sections present a Blind Source Separation (BSS) technique used to compute components and their magnitudes in each e-mail, Rough set tools used for reduction of the new feature set and classification of the incoming e-mail based on feature selection in Universe (U).

II. BLIND SOURCE SEPARATION ALGORITHM

In blind source separation (BSS), numerous perceptions are done by an array of words are handled so as to recover the beginning blending of the source signals. The term blind refers to the way that there is no particular data about the blending methodology or about the current source signals. Blind source separation (BSS) is the strategy that anybody can separate the first message or information from their mixtures without any learning about the blending methodology, yet utilizing some measurable properties of inactive or unique source message.

The perception of blind source separation is related to independent component analysis [3]. However, independent component analysis can be seen as a general-purpose tool replacing principal component analysis which implies it is appropriate to an extensive variety of issues. Some application of blind source separation is geophysical data processing, data mining, biomedical signal analysis and wireless communications [2].

Each sequence of attributes xi will be interpreted as signal and will be denoted by a column vector. It is expected here that each signal is a mixture of some underlying source of activity.

It is expected that each input signal is a linear combination of some statistically independent source [1].

$$X_i = Ma_i + e_i, \quad (1)$$

Where each column of M ∈ Rn x m is a base function Mj ∈ Rn , j = 1,2,….,m,ai ∈ Rm is a column vector of coefficients – magnitudes of each base functions in the signal xi and ei ∈ Rn represents noise or error of the model. M and a are unknown parameters that need to be evaluated. Statistical independence of the base functions can be fulfilled by minimizing shared information between the base functions. Thus M and a are evaluated by solving the following:

$$M, a = arg\ min\ a\ (arg\ min\ m\ (I(M1, M2, \ldots.., Mm) + \lambda||x - Ma||2)), \quad (2)$$

Where λ is a scaling factor, and I (M1, M2, ….., Mm) is a shared information between random variables M1, M2, ….., Mm defined as:

$$I(M1, M2, \ldots.., Mm) \sum_{j=1}^{m} H(Mj) - H(M1, M2, \ldots\ldots, Mm), \quad (3)$$

Where $H(Mj)$ is entropy of a random variableMj. Enhancement using (2) may be done with a linear regression algorithm [13]. The two amounts to be computed, M and a, make this problem complex. The minimization can be solved by Calculated only M:

$$M = arg\ min\ m\ (I(M1, M2, \ldots.., Mm) + \lambda||x - Ma||2), \quad (4)$$

Where the observed A of a in each step of the algorithm is the solution of the following:

$$A = arg\ min\ a\ ||x - Ma||2, \quad (5)$$

Where the value of M is an approximate solution of (4)

III. Genetic Algorithm

Genetic algorithms are concerning very much about evolution. Genetic algorithms (GA) is developed to be used in finding solution to a problem [14]. GA are a piece of evolutionary computing, which is quickly becoming area of artificial intelligence. It applies a series of genetic operators like selection, crossover, and mutation to a group of chromosomes where every chromosome give us answer to a problem. The starting set of chromosomes is chosen randomly from solution space. Genetic operators combine the genetic information of parent chromosomes to structure another era of the population; this methodology is known as reproduction. Every chromosome has an related fitness value, which measures its value as a solution to the problem. A chromosome representing a superior solution will have a higher fitness value. The chromosomes figured to duplicate based on their fitness value, therefore the chromosomes representing better solution have a higher possibility of survival. After numerous eras, a chromosome, which has the maximal fitness quality, is the best solution for the problem. The chromosome should in ought to somehow contain data about solution which it represents. The most utilized method for encoding is a binary string. every chromosome has one binary string. every bit in this string can represent some characteristic of the solution. Then again the entire string can represent a number [10]. After we have chosen what encoding we will utilize, we can make a step to crossover.by selecting genes from parent chromosomes and creates a new offspring. The least difficult route how to do this is to pick randomly some crossover point and everything before this point duplicate from a first parent and afterward everything after a crossover point duplicate from the second parent. After a crossover is performed, mutation should happen. This is to prevent falling all solutions in population into a neighborhood of solved problem. Mutation changes randomly the new offspring. For binary encoding we can switch a couple haphazardly picked bits from 1 to 0 or from 0 to 1.

IV. Rough Set Algoritm

Rough Set (RS) 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 [15] . 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 a data framework, then with any B ⊆ A there is related an equivalence relation INDS(B): INDS(B) = {(x,x^') U^2 |∀a B a(x)=a(x^')} INDS(B) is called the B-indiscernibility relation. The equality classes of B-indiscernibility relation are known as [x]B. The objects in ⎯BX can be positively classified as part of X on the basis of knowledge in B, while the objects in B ̅X can be just classified as possible part of X on the basis of knowledge in B. In view of the lower and upper approximations of set X ⊆ U, the universe U can be partitioned into three disjoint districts, and we can characterize them as: POS(X) = ⎯BX , NEG(X) = U − B ̅X , BND(X) = B ̅X − ⎯BX The comparability classes of B-indiscernibility relation are known as [x]B. indiscernibility is a binary equality relation that partitions as a given set of components (objects) into a specific number of disjoint comparability classes. An identicalness class of an component ai X comprises of all items ai X such that aiRai , where R demonstrates a binary relation [15]. Let IS = (Rm,A) be an information system of objects from universe Rm described by the set of attributes A , then with any B ⊆ A there is an related equivalence relation INDS(B):

$$INDS(B) = \{(x, x') \in U^2 | \forall a \in B\ a(x) = a(x')\}\ INDS(B) \quad (6)$$

Taking into account the idea of indiscernibility relation, a reduction in the space of attributes is possible. The thought is to

keep just those attributes that preserve the indiscernibility relation. The rejected attributes are redundant since their evacuation can't influence the classification. on the concept of indiscernibility relation, a reduction in the space of attributes is possible. The idea is to keep only those attributes that preserve the indiscernibility relation. The rejected attributes are redundant since their removal cannot affect the classification.

V. BSS-GRF Proposed Algorithm

Blind Source Separation-Genetic Rough Filter (BSS-GRF) is a proposed Hybrid algorithm that use the Blind source separation technique hybrid with Genetic algorithm to enhance the feature selection process and then the classification process done by rough set algorithm.

1) Discarding from the matrix X (incoming email) the column consisting of low value entered due to noise

2) Calculate the sub matrix.

3) Sort Words according to word ranks

4) Choose the number of generations (we'll use 10)

5) Read the spam and ham corpora

6) Randomly mix the lines of spam

7) Divide spam corpus into 10 slices

8) Loop until 10th generation:

a) Generate chromosomes based on the current slice of spam using the 'automatic' formula

b) Score chromosomes

c) Print results for the current generation

d) Keep the fittest 3rd

e) Reproduction survivors

- 2 survivor's reproduction via a crossover function to create a child
- 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

9) Print Final results

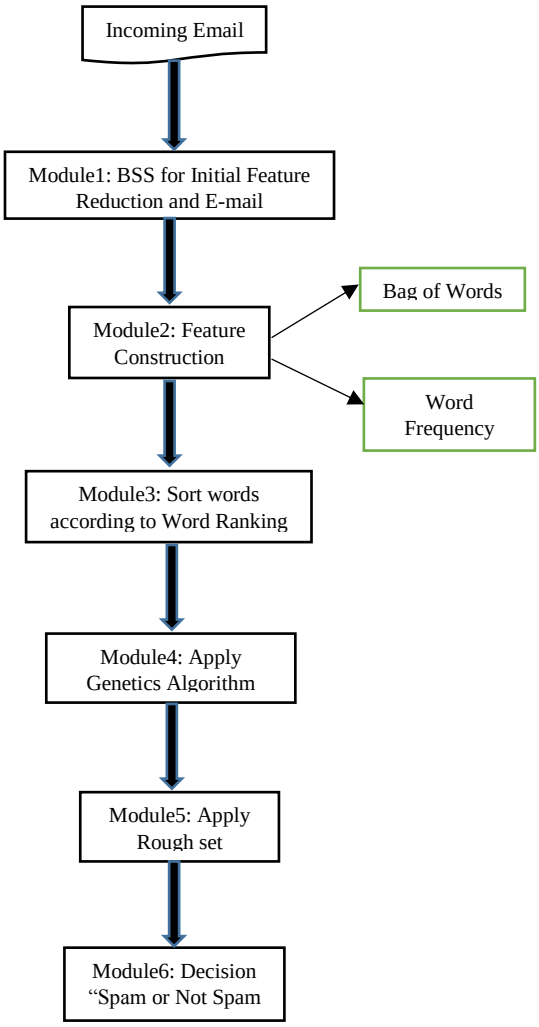

Fig. 1. BSS-GRF filtering Steps

VI. Experiment Implementation

In order to test the performance of above method, a huge database of spam and legitimate emails must be incorporated; there are few collections of email publicly available to be used by researchers. SpamAssassin (http://spamassassin.apache.org) will be utilized as a part of this experiment, which contains 6000 emails with the spam rate 37.04%.

Hence we have separated the database into training and testing sets keeping, in every such set, the same extents of ham (legitimate) and spam messages as in the first illustration set. Each training set produced contained 62.96% of the original set; while each test set contain 37.04% as Table 1.

TABLE I. DATABASE OF SPAM AND HAM E-MAILS

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

The idea here is the email is usually comes with a set of words, web links, that affect the classification process accuracy, these unwanted words can be removed manually or by special module. BSS-GRF (Blind Source Separation-Genetic Rough Filter) is presented to imbed BSS algorithm with GA to first: perform an email restoration process "an email without unwanted words". Second: GA perform the feature reduction process and word ranking that can be used for the classification process using Rough set method.

A. Performance Evaluation

In order to test the execution of above mentioned techniques, we used the most famous evaluation methods used by the spam filtering specialists. Spam Precision (SP), Spam Recall (SR), Accuracy (A). Spam Precision (SP) is the number of relevant documents identified as a percentage of all documents recognized; this shows the noise that filter presents to the user (i.e. what number of the messages named spam will really be spam)

$$SP = \frac{\text{No of Spam Correctly Classified}}{\text{Total \# of messages classifies as spam}} - \frac{N_{spam \rightarrow spam}}{N_{spam \rightarrow spam} + N_{ham \rightarrow spam}}$$

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

$$SR \quad \frac{\text{No of Spam Correctly Classified}}{\text{Total \# of messages}} = \frac{N_{spam \rightarrow spa}}{N_{spam \rightarrow spam} + N_{spam \rightarrow ham}}$$

Accuracy (A) is the percentage of all emails that are correctly Classified Where Nham→ham and Nspam→spam are

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

the number of e-mails that have been accurately 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.

B. Performance Comparison

In order to test the proposed Hybrid system we run the same data onto four different machine learning algorithms. We summarize the performance result of the presented method in term of spam recall, precision and accuracy. Table 2 and Fig.2 summarize the results of the classifier. Very competitive results can be seen from the BSS-GRF, in terms of spam recall, precision and accuracy the percentage here is much more than rough set . While in term of accuracy GRF[5] still has the high percent. Support Vector Machine System and the RS give us approximately the same lower percentage.

TABLE II. PERFORMANCE COMPARISON OF FOUR ALGORITHMS

Algorithm	Spam Recall (%)	Spam Precision (%)	Accuracy (%)
GRF	98.46	97.80	99.66
BSS-GRF	92.36	94.56	96.7
RS	92.00	90.12	94.90
SVM	95.00	93.12	96.90

VII. CONCLUSION AND FUTURE WORK

The prevoius results presented leads to new approach have to be taken by researchers in the future, Blind source separation show us a good result when it hybrid with Genetics in the purpose of feature separtion and reduction process. The presented method need more improvement in case of noise types, email corpora, more effort has to be done to improve the feature selection process in terms of accuracy, more classifiers type can be used to be hybrid with the BSS instead of the rough set method.

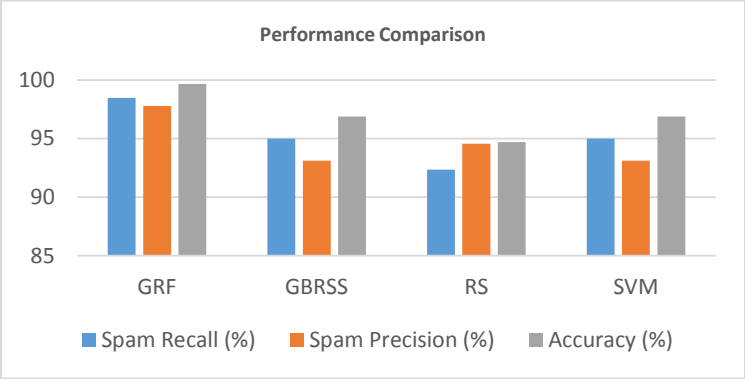

Fig. 2. Spam Recall, Spam Precision and Accuracy curves of four classifiers

REFERENCES

[1] Boratyn G.M., Smolinski T.G., Zurada J.M., Milanova M.G., Bhattacharyya S., and Suva L.J., Hybridization of Blind Source Separation and Rough Sets for Proteomic Biomarker Identification, Proc. of the 7th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2004), June 2004.

[2] A. Budipriyanto, Blind source separation based dynamic parameter identification of a multi-story moment-resisting frame building under seismic ground motions, Procedia Engineering, vol. 54, pp. 299-307, 2013.

[3] F. Gu, H. Zhang, and D. Zhu, "Blind separation of non-stationary sources using continuous density hidden Markov models," Digital Signal Process., vol. 23, no. 5, pp. 1549-1564, Sep. 2013.

[4] Karthika Renuka, D.; Hamsapriya, T.; Raja Chakkaravarthi, M.; Lakshmi Surya, P., "Spam Classification Based on Supervised Learning Using Machine Learning Techniques," Process Automation, Control and Computing (PACC), 2011 International Conference on , vol., no., pp.1,7, 20-22 July 2011

[5] S.M. ELseuofi, W.A. Awad, S. A. El Hafeez, R. M. El-Awady, Enhancing E-mail Filtering Based on GRF, International Journal of Computer Science Issues, Vol. 11, Issue 3, No 1, May 2014

[6] Md Rafiqul Islam and Wanlei Zhou. 2007. Architecture of adaptive spam filtering based on machine learning algorithms. In Proceedings of the 7th international conference on Algorithms and architectures for parallel processing (ICA3PP'07), Hai Jin, Omer F. Rana, Yi Pan, and Viktor K. Prasanna (Eds.). Springer-Verlag, Berlin, Heidelberg, 458-469

[7] A. ALmomani, T.-C. Wan, A. Altaher et al., "Evolving fuzzy neural network for phishing emails detection," Journal of Computer Science, vol. 8, no. 7, pp. 1099–1107, 2012.

[8] B. Biggio, G. Fumera, I. Pillai, F. Roli, A survey an experimental evaluation of image spam filtering techniques, Pattern Recognition Letters 32 (10) (2011) 1436–1446

[9] 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.

[10] 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.

[11] 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

[12] 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

[13] Kenney, J. F. and Keeping, E. S. (1962) "Linear Regression and Correlation." Ch. 15 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 252-285

[14] Eiben, A. E. et al (1994). "Genetic algorithms with multi-parent recombination". PPSN III: Proceedings of the International Conference on Evolutionary Computation. The Third Conference on Parallel Problem Solving from Nature: 78–87. ISBN 3-540-58484-6.

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