ISSN: 0067-2904
GIF: 0.851

Spam Filtering based on Naïve Bayesian with Information Gain and Ant Colony System

Huda Adil Abd Algafore*, Soukaena Hassan Hashem
Department of Computer Science, University of Technology, Baghdad, Iraq

Abstract

This research introduces a proposed hybrid Spam Filtering System (SFS) which consists of Ant Colony System (ACS), information gain (IG) and Naïve Bayesian (NB). The aim of the proposed hybrid spam filtering is to classify the e-mails with high accuracy. The hybrid spam filtering consists of three consequence stages. In the first stage, the information gain (IG) for each attributes (i.e. weight for each feature) is computed. Then, the Ant Colony System algorithm selects the best features that the most intrinsic correlated attributes in classification. Finally, the third stage is dedicated to classify the e-mail using Naïve Bayesian (NB) algorithm. The experiment is conducted on spambase dataset. The result shows that the accuracy of NB with IG-ACS is better than NB with IG only.

Keywords: Ant Colony System, Feature Selection, Information Gain, Naïve Bayesian and Spam Filtering System.

نظام تصفية الرسائل الالكترونية الغير مرغوب فيها بتهجين طريقة اختيار الخواص بأستخدام كسب المعلومات ونظام مستعمرة النمل

هدى عادل عبد الغفور*، سكينة حسن هاشم
قسم علوم الحاسبات، الجامعة التكنولوجية، بغداد، العراق

الخلاصة

يقدم هذا البحث نظام مقترح هجين لتصفية الرسائل الالكترونية غير المرغوب بها والذي يتالف من نظام مستعمرة النمل مع نظام الافتراضية البسيط. هدف النظام المقترح تصنيف الرسائل الالكترونية الغير مرغوب بها بدقة عالية .النظام الهجين المقترح يتكون من ثلاث مراحل متعاقبة. في المرحلة الاولى يتم احتساب كسب المعلومات (IG) لكل خاصية . ثم تقوم خوارزمية نظام مستعمرة النمل باختيار افضل الخواص التي تكون مترابطة ترابطا جوهريا في عملية التصنيف الرسائل الالكترونية. اخيرا ،الخطوة الثالثة يتم بها تصنيف الرسائل الالكترونية باستخدام خوارزمية نظام النظرية الافتراضيه البسيط. التجارب اجريت على بيانات spambase. النتائج اظهرت دقة التصنيف الرسائل الالكترونية لنظام الافتراضية البسيط مع نظام مستعمرة النمل افضل من نظام الافتراضية البسيط مع كسب المعلومات.

1. Introduction

Some types spam filters are designed to work as manual patterns. They consist of matching rules which are required to be adjusted to each incoming e-mail message. Their mission requires experience and time. Moreover, the features of all unwanted messages (e.g. offered products and frequent terms)

*Email: Asola.adel95@yahoo.com

change from time to time, that requires the rules to be updated. However, significant advantages must be presented by any system that offers automatic separating of spam and not spam e-mails[1]. The Naïve Bayes classifier is used for classifying e-mails, words probabilities that plays the master role here. Whether any words occur always as in spam e-mail but not in ham e-mail, hence this e-mail maybe spammed. NB classifier technique became so common method in e-mail filtering software. This filter must be trained to categorize e-mails affectively [2]. The Bayesian classifiers determine attributes (spam common keywords or phrases). The classifier assigns probabilities for them [3]. Each word has a specific occurrence probability in spam or ham e-mail in its dataset. If the total probabilities of words exceeded a certain limit, then the filter will mark the e-mail to one of the two categories, either spam or ham. Mostly, all the statistics-based spam filters employ Bayesian probability computation to gather the individual token's statistics into one outcome. Bayesian filtering decision depends on this outcome [2]. E-mail is very significant issue and exposed to many risks; one of most important risk of them is the spam e-mail. This research proposes a way to solve this kind of risk using NB classifier for filtering the e-mails.

Ant Colony Optimization (ACO) is a field of interest within "Swarm Intelligence (SI)". In nature, it can be noticed that it is easy for real ants to find the shortest path between their nest and the food source without using visible information or a universal model, to be environmentally adapted. Pheromone deposition is the principal factor that enables the real ant to find the shortest path during a specific time period. Each ant is probabilistically prioritize to follow a way that is rich in this chemical [4]. Ant Colony System (ACS) is used for selection a subset of features with the aim to reduce the dimensionality of feature set and present a better accuracy result in classifying e-mails. Selection of features is essential to reduce the computation cost and improve the performance of classification [5].

2. Related works

This section explains some of the related work to the proposed spam filtering system. Christina V. et al., 2010, presented various types of spam filters that identify whether the incoming message is spam or useful such as (legislation, content scanning, white list / black list, keyword matching, mail header analysis, postage and Bayesian analysis). While they are still overwhelmed with spam e-mail. This does not mean that the used filter system is not efficient, but it means that spam filter is not acclimated to the changes that have been made by spammers. In their work, they applied supervised techniques of machine learning to filter the spam e-mails. They used some learning techniques namely; "Multilayer Perceptron, C4.5 Decision tree classifier, Naïve Bayes Classifier", which are used to learn the spam e-mails features. This model is constructed by training to distinguish spam e-mails from useful e-mails [6].

Awad W. A. and ELseuofi S. M., 2011, presented the growing size of spam e-mail, which inspire the need for efficient anti-spam filters. They presented some of popular machine learning methods such as Support Vector Machine (SVM), Bayesian classification, Artificial Neural Network (ANN), Rough sets, Artificial immune system and K-Nearest Neighbor (k-NN) to classify spam e-mail. Algorithms characterization was introduced, and their performance on the spam corpus and Spam Assassin was compared [2]. Eberhardt J.,2014, presented that spammer constantly discover new ways to bring spammy content to the public. This is done by e-mail, social network and advertising products. Estimation of the Messaging Anti-Abuse Working Group reported that ninety percentage of e-mails in U.S.A were spam. Although there is no method capable of blocking all spam. Spam filters are improving at finding spam and deleting it. Bayesian approach have been used in text categorization and an early used method in spam filter. He tested two optimization on Naïve Bayesian text categorization and spam filtering. He showed that NB filter could be applied for more accurate text classifier and some modification could increase the accuracy results [7].

3. Description of the Proposed Spam Filtering System

Figure-1 depicts NB classifier with the three cases; NB classifier with all features, NB classifier with Information Gain feature selection and NB classifier with IG and ACS feature selection.

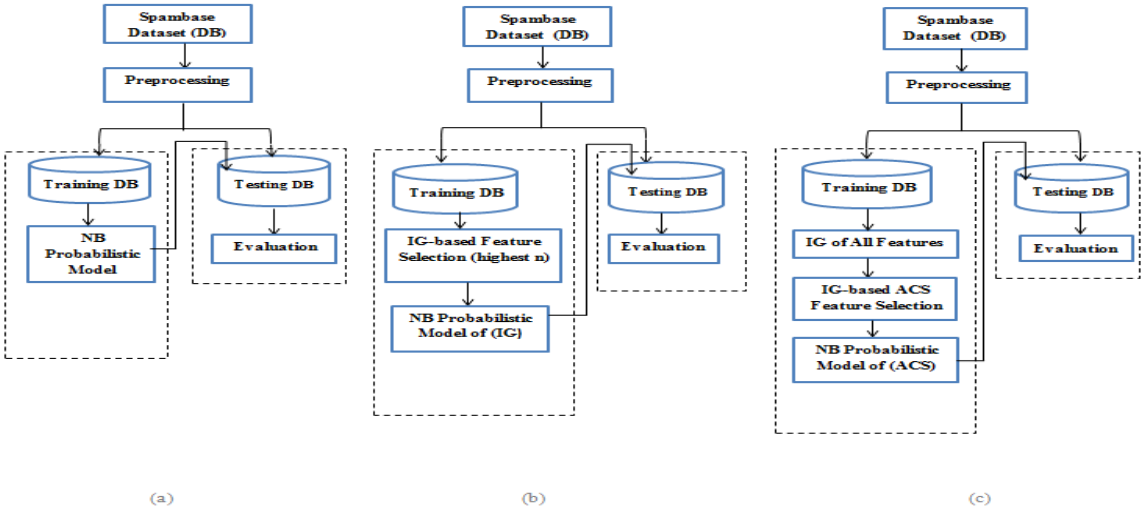

Figure 1- Block diagram of (a. NB Classifier, b. NB with IG Classifier and c. NB with ACS Classifier)

The system is an offline NB classifier for spam filtering. First, the spambase dataset is prepared. Then, the traditional NB classifier with all features of spambase dataset is applied as depicted in Figure-1a Moreover, NB classifier is applied on a subset of features selected by information gain (IG) algorithm as depicted in Figure-1b Finally, NB classifier is applied on a subset of features that are selected using information gain (IG)-based Ant Colony System (ACS) as depicted Figure-1c Generally, ACS selects only the best combination of features. In this research, the ACS is used to select the worst features in order to be turned off.

3.1. Dataset Description (Spambase Dataset)

The dataset of this system is a spambase dataset which contains more than four thousands of records and 57 features in addition to e-mail class type. Spambase dataset is a dataset of features that describes the incoming e-mail characteristic. The dataset is prepared for data mining and is used by many researchers [9-11]. Normalization min-max process is applied to the values of these features to set them in a uniform range between [0, 1]. After the Normalization stage, the spambase dataset is divided into two datasets: the first one is training dataset that it, consists of (3007). The second dataset is testing dataset that is used for classification and consists of (1044) records. Moreover, another training dataset is prepared which consists of (1507) records of e-mails. Each record consist of (57) columns that represent e-mail features in addition to the class type (spam and non-spam).

3.2. Naïve Bayesian Classifier

This research applies Naive Bayesian rules to training dataset, results of NB classifier probabilities are used to classify the testing dataset into spam or not. NB classifier consists of two stages they are: training stage and classification stage (test), both of them depend on the message content which is represented by features, more explanation is shown in Algorithm (1).

Algorithm (1): Spam Naïve Bayesian Filter

Input: Spambase training and testing dataset.

Output: Classification based on NB.

begin

Step 1 : Training

Spam_Probability = No. of spams in training dataset / The total No. of e-mails

NSpam_ Probability = No. of nspams in training dataset / The total No. of e-mails

For each Feature (F) in Training dataset **do**

Spam_ Probability (F)=Spam_counter(F)/No. of spams in training dataset

NSpam_ Probability (F)=NSpam_counter (F)/ No. of nspams in training dataset

End For

Step 2 : Testing (Classification)

For each Feature (F) in Testing dataset **do**

For each E-mail (E) in Testing dataset **do**

If value of (F) is Found in training dataset then **do**

Result_Spam*= Spam_ Probability (F)
Result_NSpam*= NSpam_ Probability (F)
Else
Get two nearest feature probabilities
Get the average of these features
Result_Spam*= Spam_ Probability (F) of the average
Result_NSpam*= NSpam_ Probability (F) of the average
End if
End For
Result_Spam* = Spam_ Probability
Result_NSpam* = NSpam_ Probability
If Result_Spam > Result_NSpam then **do**
Result = Spam
Else
Result = NSpam
End if
End For
Step 3 : Evaluating NB Classifier
End

Several problems occur in NB classification process:
The first problem is some elements are zero in either spam class or non-spam class, the probability of that element is zero as well. Use Laplace smooth to solve zero probability problems by adding one to each counter of zero probability elements.

$$\textbf{zero probability}(\textbf{element}|\textbf{spam}) = \frac{\textbf{spam count}+1}{\textbf{class spam count}+\textbf{the total frequncies of the feature}} \quad (1)$$

$$\textbf{zero probability}(\textbf{element}|\textbf{non}-\textbf{spam}) = \frac{\textit{not spam count}+1}{\textit{class non spam count}+\textit{the total frequncies of the attribute}} \quad (2)$$

Second problem occurs when value exists in a feature of test dataset, but does not exist in training dataset. The proposal solves the problem by taking the arithmetic average of probability of the nearest two values.

3.3. Naïve Bayesian Classifier with IG for Spam Filtering

This classifier calculates (IG) of each feature by calculating all elements entropy of feature with both classes (spam and non-spam) and the entropy of them. IG selects subset of features with high information gain and turn off features with low information gain. IG feature selection is used to reduce the size of features, reduce classification compution and improving NB classification results. More explanation is given in Algorithm (2).

Algorithm (2) Naïve Bayesian with IG Spam Filtering
Input: Spambase training and testing datasets.
Output: NB Classifier which is trained and tested with selected features resulting from IG.
Begin
Step 1: Feature Selection
Total Entropy = Entropy ($\frac{\text{spam E-mail counter}}{\text{total E-mail counter}}, \frac{\text{non-spam E-mail counter}}{\text{total E-mail counter}}$) in training dataset.
For each Feature (F) in Training dataset **do**
For each Element (El) in Training dataset **do**
T=F_El_Spam counter + F_El_NSpam counter
Feature_Entropy + = $\frac{\text{T}}{\text{Total E-mails count}} *$ Entropy (F_El_Spam counter, F_El_NSpam counter)
End For
IG_Feature = Total Entropy- Feature_Entropy
End For
Step 2: Classification
Apply NB classifier as in Algorithm (1) to the selected features using information gain.
End

3.4. Naïve Bayesian with ACS for spam filtering

ACS algorithm is an ingredient of the Ant Colony Optimization (ACO). Naïve Bayesian with ACS Classifier uses Ant Colony System as a feature selection. ACS selects the worst features by turn off those are with the highest transition rule and maintain the best features that improves the accuracy result of NB. For more explanation see Algorithm (3). Initial pheromone (Ph_0)=1/57, No. of node=n=57, the connection between nodes (weight) is calculated by using Eq.(3) that represents Heuristic measure

$$\eta_{ij} = \left(\frac{1}{information\ gain_i} + \frac{1}{information\ gain_j}\right) / 2 \quad (3)$$

No. of ants=5, first node is randomly selected, No. of iterations =10, "Q" "a random value between (0, 1)", "Q_0" is assumed to be 0.5, Beta is the attractive measure of move =1, ρ, ρ_1: evaporate value = 0. 1, DT: 1/St, St: length of selected subset.

The following example illustrates how ACS is applied on sample of 6 features and their IGs is shown in Table-1. These features are represented as nodes; which are fully connected as shown in Figure-2. The connection between nodes(weight) is calculated by using Eq.(3)

Table 1- Information Gain for 6 features

Feature No.	Feature	Information gain
0	"word_freq_make"	0.0270575185097344
1	"word_freq_address"	0.0718784366690677
2	"word_freq_all"	0.0812229229640578
3	"word_freq_3d"	-0.00677980625135199
4	"word_freq_our"	0.0775579928265888
5	"word_freq_over"	0.0550068441918068

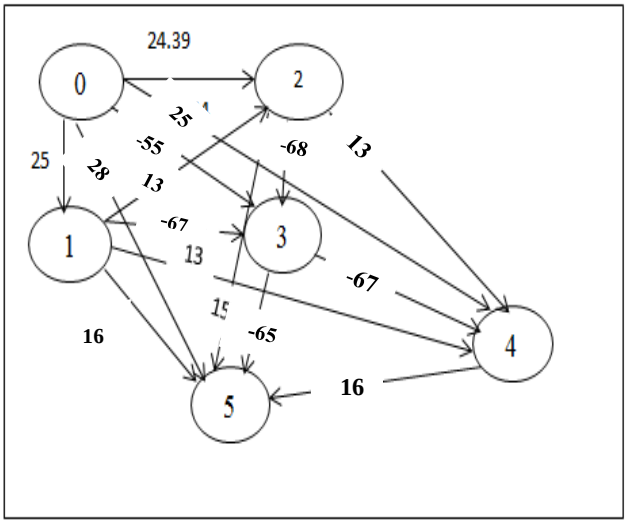

Figure 2- ACS with sample of 6 features

Algorithm (3) Naïve Bayesian with ACS for spam filtering

Input: Spambase training and testing datasets and IG for each of 57 features.

Output: Classification based on NB which is trained and tested with maintained features resulting from ACS .

Begin

Step1 :Feature Selection

1. Initialize (Q_0, Q, Ph_0, ρ , ρ_1 ,n, Dt , start node, No. of Ants, No. of iterations).
2. Distribute Ants on some nodes randomly.
3. Take nodes from the path of the Ant.
4. Check Q_0 , Q.

4.1Calculate Information Gain using Algorithm (2).

4.2 Heuristic measure =$(\frac{1}{\text{InformatinGain}_i} + \frac{1}{\text{InformationGain}_j})/2$ is used as distance cost.

4.3 If $Q \leq Q_0$ an Ant applies Exploitation transition rule
= max[pheromone $*$ heuristic measure] to select the next node.

4.4 Otherwise, an Ant applies Exploration transition rule
= max$[\frac{\text{pheromone} * \text{heuristic measure}}{\text{all nodes(pheromone} * \text{heuristic measure)}}]$ to select the next node.

5. Local updating pheromone rule is :

$$\text{Ph}_{ij}(t) = (1-\rho)\,\text{Ph}_{ij}(t) + \rho\frac{1}{n + \text{length of the route between two nodes}}$$

6. Global updating pheromone rule is :

$$\text{Ph}_{ij}(t+1) = (1-\rho_1)\text{Ph}_{ij}(t) + \rho_1 \text{Dt}$$

7. Eliminate of the Best worst features.

Step2 : Classification

8. Apply NB Classifier Algorithm (1) to the maintained features on testing dataset.

End

4. Evaluation Measurements of Classification and Experimental Works

Three classification models have been experimented. These models have been trained and tested on the same spambase dataset. The evaluation process estimates validity and accuracy of these constructed models as in Eq.(4).

Accuracy evaluates the classifier efficiency through its correct predictions percentage [11].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Where

1. True Positive (TP): infected e-mail that is correctly categorized as spam
2. False Positive (FP): e-mail that is incorrectly categorized as spam.
3. True Negative (TN): e-mail that is correctly categorized as e-mail.
4. False Negative (FN): infected e-mail that is incorrectly categorized as e-mail [12].

Table-2 clarifies the accuracy result of testing 1044 sample when two training dataset are used namely 1507 sample and 3007 sample.

Table 2-Training samples and accuracy results

No. of Training samples	No. of Testing samples	Accuracy		
1507	1044	NB	NB with IG	NB with ACS
		86.87%	86.78%	91%
3007		89.75%	94.44%	94.7%

The training sample size affects the classifier performance as shown in Table-2 As long as the training sample size is increased, the accuracy results are increased. The accuracy results increased in ascending order in the three classifiers whereas NB with ACS classifier is the highest one of them.

Table 3- Second Training sample and Accuracy results

No. of Training samples	**No. of Testing samples**		**Accuracy**	
3007	1044	Feature affect	NB with IG	NB with ACS
		spam	93%	93.58%
		Non-spam	84.57%	88.98%

Total features number is 57 features where 30 of them are affect as "non-spam" and 27 are affect as "spam". Non-spam features are the total sum of the feature probabilities as non-spam. Their total sum is higher than the total sum of that feature probabilities as spam and vice versa for spam features.

The highest accuracy result of NB with IG classifier is obtained from selecting (50 from 57) features (select means turning off these features), (24 affect as non-spam, 26 affect as spam) features. NB with IG gives a higher accuracy result by selecting these features because there is a statistical correlation between these features.

Selecting only (24) that affect as non-spam features by NB with IG classifier leads to (84.57%) as accuracy result, while selecting (26) that affect as spam features leads to (93%) as accuracy result. Our aim is improving the accuracy of classification results by feature selection.

The highest accuracy result of NB with ACS classifier is obtained from selecting (31 from 57 mentioned above) features, (18 affect as non-spam, 13 affect as spam) features. Selecting only (18) that affect as non-spam features by NB with IG and ACS classifier leads to (93.58%) as accuracy result, while selecting (13) that affect as spam features leads to (88.98%) as accuracy result.

The line graph in Figure-3 shows the accuracy result percentages that are achieved by (NB, NB with IG and NB with ACS) classifiers in classifying testing dataset.

Axes of Figure-3 (where X axis represents the selected features to be turned off and Y axis represents the accuracy results of classifiers), the selected number of features starts by (5) and ends with (55) as shown in Figure-3 The accuracy result of NB classifier is "represented by the orange line in Figure 3.". The accuracy result did not change because there is no feature selection stage. The accuracy results of NB with IG classifier is "represented by the red line in Figure 3.", NB with ACS classifier accuracy results are "represented by the yellow line in Figure 3.". As shown in the figure NB with IG classifier selects 10 features lead to decrease the accuracy result percentage, while NB with ACS selects 10 features lead to increase the accuracy percentage. That means the 10 selected features by NB with ACS classifier are better than that are selected by NB with IG classifier.

NB with ACS classifier selects 20 features leads to increase accuracy result because of the statistical correlation between features. Whereas NB with IG classifier selects 20 features, leads to increase the accuracy result more than when NB with IG classifier selecting 10 features, but still NB with ACS classifier presented better accuracy result in selecting 20 features. For instance in Figure 3., NB ACS classifier selects 31 features that present a better accuracy result than the 31 features that were selected by NB with IG classifier.

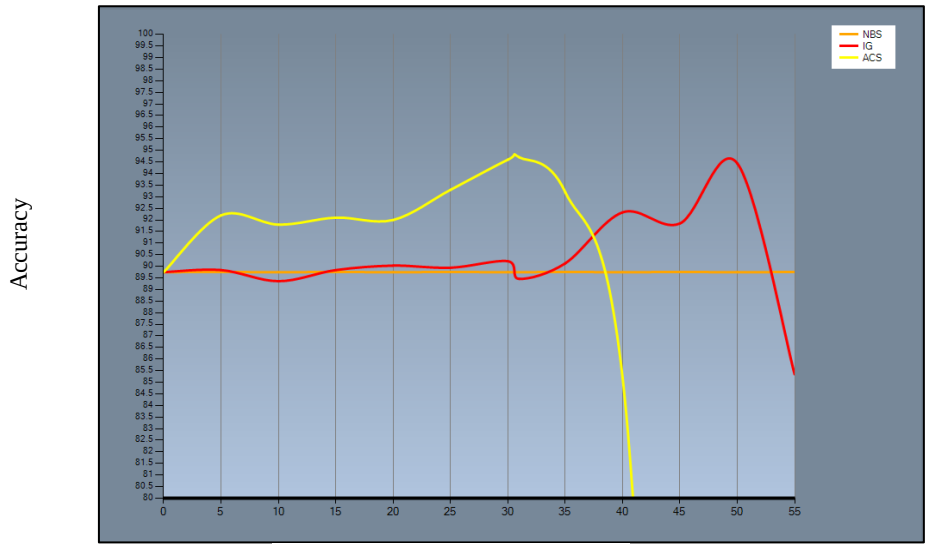

Figure 3- The accuracy with selected features of three classifiers

Table- 4 shows the highest accuracy results that are achieved by using NB with IG and NB with ACS classifiers. The first and second columns present selected features (non-spam, spam) by using NB with IG classifier, while third and last columns present NB with ACS classifier selected features.

Table 4- Selected features using NB with IG and NB with ACS

NB with IG classifier				**NB with ACS classifier**			
Non-spam No.=24		**Spam No.=26**		**Non-spam No.=18**		**Spam No.=13**	
0	17	3	36	1	22	11	
1	18	11	37	2	44	27	
2	19	25	38	4	54	28	
4	21	26	40	5	55	29	
5	22	27	41	7	56	30	
7	39	28	42	8		31	
8	44	29	43	9		34	
9	53	30	45	10		35	
10	54	31	46	12		36	
12	55	32	47	14		38	
13	56	33	48	16		41	
14		34	49	17		49	
16		35	50	19		50	

Intersected features are features that are selected from the highest accuracy cases achieved by using NB with ACS and IG based NB classifiers. Table-5 presents 31 intersected features which are 18 as non-spam and 13 as spam. Selecting only 18 as non-spam of the intersected features leads to (93.58%) as accuracy result, while selecting 14 as spam of the intersected features leads to (88.98%) as accuracy result. The accuracy achieved by selecting all intersected features is (94.73%). As a result, the selection of these features improves the accuracy of the results, that means these features are statistically correlated. The fact is that these features that have been selected from both classifiers are most likely unnecessary features.

Table 5- Intersected features from NB with IG and NB with ACS

Intersected Features			
Non-spam No.=18		**Spam No.=13**	
1	14	11	38
2	16	27	41
4	17	28	49
5	19	29	50
7	22	30	
8	44	31	
9	54	34	
10	55	35	
12	56	36	

Overall, a better accuracy result can be obtained by increasing the sample size. From the result, one can be noticed that the number of non-spam features is more than that of spam features so the classifiers concentrate on them. The best accuracy result is achieved by NB with ACS classifier.

5. Conclusions

- Naive Bayesian classifier gives good accuracy result of classification.
- The accuracy of Naïve Bayesian is increased by proposed IG and ACS feature selection algorithms. The accuracy of NB with IG classifier is better than those obtained by using NB classifier. The accuracy of NB with ACS classifier is the best results by selecting a set of unnecessary features which reduces the classification computation and gives the highest accuracy result.
- Classifiers performance is better when training sample size is increased due to more features elements will be exist.
- The randomization in ant colony system may lead to bad classification result depending on the first node that is randomly selected.
- NB with ACS classifier implements the exploration transition rule, and very rare leads to obtain less accurate results.

References

1. Androutsopoulos I., Koutsias J., Chandrinos K.V., Paliouras G., and Spyropoulos C.D. **2000**. An Evaluation of Naive Bayesian Anti-Spam Filtering, 11th European Conference on Machine Learning, Barcelona, Spain, pp: 9-17.
2. Awad W.A. and ELseuofi S.M.**2011**. Machine Learning Methods for Spam E-Mail Classification, *International Journal of Computer Science & Information Technology (IJCSIT)*, 3(1).
3. Chakraborty N. and Patel A.**2012**. E-mail Spam Filter Using Bayesian Neural Networks, *International Journal of Advanced Computer Research*, 2(1).
4. Jensen R., Abraham A., Grosan C., and Ramos V. **2006**. *Swarm Intelligence in Data Mining*, book studies in computational intelligence, 34.
5. Deriche M. **2009**. Feature Selection Using Ant Colony Optimization, 6th International Multi-Conference on Systems, Signals and Devices.
6. Christina V., Karpagavalli S. and Suganya G. **2010**. E-mail Spam Filtering Using Supervised Machine Learning Techniques, *International Journal on Computer Science and Engineering*, 2(9), pp: 3126-3129.
7. Eberhardt J. **2014**. Bayesian Spam Detection, Division of Science and Mathematics University of Minnesota, UMM CSci Senior Seminar Conference, USA 56267, December.
8. UCI Machine Learning Repository, archive.ics.uci.edu /ml/datasets/ Spambase
9. Cepek M., and Snorek M. **2013.** Inductive Preprocessing Technique for Knowledge Discovery Demonstrations on Real Worlds Datasets, International Conference in Inductive Modelling ICIM.
10. Alqatawna J., Faris H., Jaradat K., Al-Zewairi M. and Adwan O. **2015**. Improving Knowledge Based Spam Detection Methods: The effect Of Malicious Related Features in Imbalance Data Distribution, *Int. J. Communications, Network and System Sciences.* 8(8).
11. Costa E. P., Lorena A. C., Carvalho A. C. P. L. F. and Freitas A. A. **2007**. A Review of Performance Evaluation Measures for Hierarchical Classifiers, 3rd Association for the Advancement of Artificial Intelligence.
12. Hossin M., and Sulaiman M.N.**2015**. A Review on Evaluation Metrics for Data Classification Evaluations, *International Journal of Data Mining & Knowledge Management Process (IJDKP*) 5(2).