International Journal of Security and Its Applications
Vol. 10, No. 3 (2016), pp.121-128
http://dx.doi.org/10.14257/ijsia.2016.10.3.11

Bagging eEP-based Classifiers for Junk Mail Classification

Yan Li[1] and Hua Zhou[2]

[1]*Henan Technical College of Construction , ZhengZhou,China,450064*
[2]*Zhongzhou University,ZhengZhou,China,450044*
hnzzliyan@sina.com

Abstract

The volume of junk emails on the Internet has grown tremendously in the past few years and is causing serious problems. Content-based filtering is one of mainstream technologies used so far. This paper has had a deep study in the content of emails and come up with a better idea to get the features which make it even convenient to e-mail classify as well. This paper uses the classification algorithm by Bagging eEP-based classifiers to the junk email examine, and carries out a new categorization and filtering algorithm BeEPJMC. The experiments show, the new feature extraction methods and the combination BeEP classification is a very efficient method of classification, and The classification efficiency of the algorithm BeEPJMC is higher than currently several better classification algorithm.

Keywords*: E-Mail Categorization, Feature Extraction,Bagging, Essential Emerging Patterns*

1. The Basic Concept

DB-based training data set contains N samples e-mail (T1, T2, ..., TN), is divided into two known types of C1, C2, and a given sample of each class mail. Classification in the mail, all the samples are all text. Although the text of the title, abstract and key words containing important classified information, but not all of the text contains such information. Therefore, assuming that all the text of this article contains only the contents of the message body, hereinafter referred to as message content.

We use Spam Corpus —PU Series corpus set. Its received from a provider of real-time e-mail. Corpus to retain only those e-mail the title and the body of the plain text content.

Many classification algorithms[1] have message content will be mapped to n-dimensional space, in which n equal to the contents of all e-mail appear in the number of different words. After the initial filtering, mail content in the different number of words remains high. Feature extraction task is to delete that information for the classification of small, "unimportant" words. After feature extraction, the message appears in the text known as the characteristics of the word. In the following discussion, sometimes referred to as the characteristics of items, the term "word", "Feature" and "item" will be mixed use.

Feature extraction, each message is a collection of items. So that W = (w_1, w_2,, ..., w_n) is the message content of the items appeared in The Complete Works. W subset $X \subseteq W$called itemsets. If the itemset X in T appear in the text, then T contains X.

Definition 1

a set of training data set D is a subset of DB. Itemset X in D on the degree of support sup_D (X) = $count_D(X)/|D|$, which $count_D$ (X) is a data set D contains a sample of the number of X, and | D | is the total number of D in the samples.

If D is a collection of C_i types of training samples, sup_D (X) recorded as sup_i (X), it is the itemset X in the category C_i training focused on the frequency samples.

Definition 2

ISSN: 1738-9976 IJSIA

Given two dierent classes of datasets D and D' ,the growthrate of an itemset X from D to D' is defined as GrowthRate(X)= $gr_{D'\to D}(X)$:

$$gr_{D'\to D}(X)=\begin{cases} 0 & \text{if} \quad sup_{D'}(X)=sup_D(X)=0 \\ \infty & \text{if} \quad sup_{D'}(X)=0, sup_D(X)\neq 0 \\ sup_D(X)/sup_{D'}(X) & otherwise \end{cases}$$

If the data sets D and D 'are non-spam and junk e-mail collection of samples, $gr_{D'\to D}(X)$ recorded as $gr_i(X)$, it is a set X from a non-junk mail to junk e-mail support (frequency) significant changes in the extent of the measure.

Denition 3

Given a growth rate threshold ρ >1, an itemset X is said to be ρ-Emerging Pattern[2](ρ-EP or simply EP)from a background dataset D to a target dataset D' . Itemset X is eEP (essential EP) of D , if (1) X is D, EP, (2) X in D, the support is not less than pre-specified minimum support threshold ξ, and (3) X subset of any really not satisfy the conditions (1) and (2).

When D and D 'are non-spam and junk e-mail when a collection of samples, D of the EP / eEP also known as spam EP / eEP. In fact, eEP is the "shortest possible", the most ability to express the EP. During the discussion after the feature extraction, we will discuss in detail the eEP based on the establishment of e-mail classifier.

2. The Pre-Processing and Feature Extraction of the E-mail Text

E-mail in the feature extraction, we first set of data pre-processing e-mail:

(1) an e-mail data sets together an e-mail messages to a large document, remove the message headers, message content, only in part. These e-mail documents to each message type label started, and then began to use -1 as the content of signs, followed by the message body, and finally to -11 as the end of each message;
(2) each repetition of the word removed from the e-mail message body to retain only a duplication of the word.

Feature Extraction for the Message:

(1) Statistics for each word in a normal e-mail and spam that exist in the frequency into the hash table;
(2) for the hash table appear in the zero-spam and email in the normal word zero times we have it mapped to a fixed two special symbols (special symbols can be used in place of any number), and then remove the e-mail the body of each repeat of the special symbols, special symbols so that each retained only one in the message body in order to shorten the length of the message body;
(3) According to a word in the greater difference between two types of messages in terms of its more important, more important the higher frequency of these two principles, for the hash table in a normal e-mail spam and there are times for the non-zero term, we Among several types of frequency for the larger number of frequency x, the smaller number of frequency y, proposed formula:

$$F(m)=\left(\frac{x}{x+y}\right)^{\alpha}(x-y)^{\beta}(\alpha>0,\beta>0)$$

α, β balance these two principles, One of value set δ , δ known as the balance factor, the balance of the different factor δ, by descending order of the results on the order of different threshold δ extract large in the c word as our word feature extraction (0 <c <1), the number of feature words set to 70, each time the

characteristics of the word is greater than the number of messages in order to select a different value of *c*;

(4) feature extraction of speech after the message body and words in the match to retain the characteristics of the word, by deleting the non-feature of the word, match result, the body of each message contains only the characteristics of the word (that is, Feature items).

3. EEP-based E-mail Classification and Filtering Algorithms BeEPJMC

Feature extraction, all of the messages are a collection of characteristics, and training data sets characteristic of DS is a collection of multiple sets.

A. Mining eEP

For the establishment of the e-mail-based classifier eEP, first of all need to dig eEP. The steps are as follows:

(1) get set minimum support threshold ξ and minimum growth rate ρ;

(2) for *i* = 1,2, Ci on behalf of two types of e-mail (spam and non-spam):

Training data set will be divided into categories Ci and Ci samples set;
Mining Mining eEP category Ci and the Ci-type eEP;

literature[6] gives the detailed steps eEP excavation, this article is no longer cumbersome. A large number of experiments showed that growth ξ = 1%. However, the minimum support threshold rates ρ depend on the data distribution. In general, for the easy classification smaller can take larger values (greater than 5), contrary of data sets, better value can be set up through repeated value should be taken (2~5). classifier, according to the classification of the samples tested from the accuracy to determine appropriate adjustments. See literature [5].

Some characteristics may not appear in any eEP, they do not work the classification of unknown samples.

Definition 4 Characteristics of the definition of *w* is a key feature of an effective, if *w* appears in at least one of eEP.

B. Sorting

eEP distinguish between a good performance. *X* is Ci-based category eEP, its growth rate of gri (*X*). This indicates that the focus in the classification of samples, *X*-type Ci samples in the frequency (support) is a non-Ci samples in the frequency of the gri (*X*) times. If the *X* in question appeared in *T* classification of mail, from a statistical point of view, *T* is the possibility of Ci-type *T* does not belong to Ci category gri (*X*) times.

E-mail to be classified in order to determine the type of *T*-owned, Ci each category eEP are trying to determine whether the *T*-type Ci. *X* is Ci-based category eEP. If *X* is not appear in *T*, then *X* can not determine whether the *T*-type Ci to judge. If *X* appears in *T*, *X* will be the probability $\frac{gr_i(X)}{gr_i(X)+1}$ Determine the type *T* belong to Ci, and to the probability $\frac{1}{gr_i(X)+1}$ Determine the type *T* does not belong to Ci.

In order to classify e-mail *T*, BeEPJMC combination of Ci to C_i-type and non-eEP of each category to determine the calculation of *T* scores are Ci-type, score (*T*, Ci). ultimately determine the type *T* belong. The PS (*T*, Ci) = (*X* | *X* is Ci category eEP, and *X* in *T* appear in), NS (*T*, Ci) = (*X* | *X* non-Ci category eEP, and *X* appear in *T*) . BeEPJMC the following steps to classify e-mail *T*:

a) the deletion of *T* in the absence of an effective focus on the characteristics of the word appears;

b) For i = 1, 2,

For PS (T, C_i) and NS (T, C_i);
By computing T under the category C_i is the score score (T, C_i)

$$score(T, C_i) = \sum_{X \in PS(T,C_i)} \frac{gr_i(X)}{gr_i(X)+1} + \sum_{X \in NS(T,C_i)} \frac{1}{gr_i(X)+1}$$

c) T was placed under the category of the highest scores.

C. Bagging eEP-based classifiers for classification

Given the training data set D and classifier number BeEPJMC m, according to the following ways algorithm firstly tectonic m a base classifier G_0 and G_1, G_2,... G_{m-1}:

(1) to determine the minimum support threshold ξ and minimum threshold ρ and growth in the training data set D on eEPs mining,., classifier G_0.
(2) for training data set, using the improved D Bagging sampling tectonic m - 1 self-help samples, D_B^1, D_B^2, …, D_B^{m-1}。 ... ,.
(3) I = 1, 2, and... M - 1, with ξ and ρ respectively for minimum support threshold and minimum threshold, the growth of samples of mining self-help eEPs, establish D_B^i and classifier Gi.

For each sample for classification and yankees classifier Gi（$0 \le i \le m$-1）were independent of prediction, and do eventually belongs to sample of each base classifier comprehensive judgement forecast results.

If the Gi can make sure to S predictions, said the foundation of classifier Gi samples S effective, or say to sample S failure. Gi With a single classifier CEEP algorithm, when not to S explicitly Gi, it no longer judge according to most of the decision to give up, but simply to S predictions, will decide S belongs to the power of the ability to distinguish with enough other yankees classifier.

When all the base classifier Gi （$0 \le i \le m$-1） were treated sample classification, after making independent prediction "S BeEPJMC will vote to determine the S belonging to the class. Voting rules are as follows:

(1) if the yankees classifier Gi samples of failure, Gi S not to vote.
(2) if the yankees classifier Gi samples S belong to predict, " the S belongs to C_k" is a contribution.
(3) BeEPJMC will sample S for the most votes, If not only the most votes of the sample under S, most of the most votes.

4. Analysis of Experimental Results and Evaluation

Experimental data sets using public spam corpus set PU series, provided by the Greek scholar Androutsopoulos. Its received from a provider of real-time e-mail. Corpus to retain only those e-mail the title and the body of the plain text content. Providers in order to protect the privacy of e-mail corpus, It will be different in different words in place of integers. PU Series corpus currently consists of PU1, PU2, PU3 and PUA four corpus. The average corpus of each PU is divided into 10, that is, part1 to part10. At present, we mainly corpus PU1, PU1 corpus of each check in a ten to 10 fold cross-validation (cross validation), PU Series corpus shown in Table 1.

Table 1. Spam Database (Unit: Letter)

Data set	Non-spam numbers	Spam numbers	Totall numbers	Remarks
PU1	618Pr	481Pr	1099	Encryption Forms
PU2	579Pr	142Pr	721	Encryption
PU3	2313Pr	1826Pr	4139	Encryption
PUA	571Pr	571Pr	1142	Encryption

Spam is usually classified using the performance evaluation of text classification relevant indicators. Specifically, based pos is the total number of spam, t_pos is the correct classification of spam (junk e-mail really) a few, and f_pos was wrongly classified as spam (false spam) number, then the following evaluation different indicators can be used to measure the spam filtering performance of the system:

(1) Recall : Recall that the rate of spam;

$$recall = \frac{t\_pos}{pos}$$

Recall reflects the filtration system's ability to find spam. The higher recall rate, "slipping through the net" less spam.

(2) Precision: that is, spam precision;

$$precision = \frac{t\_pos}{t\_pos + f\_pos}$$

The accuracy of the filter response system "to find the" junk e-mail capabilities, precision higher miscarriage of justice would be a legitimate message as spam the possibility of the smaller.

(3) Accuracy: that is, for all mail (including junk mail and legitimate e-mail) on the rate of the contractor. Accuracy, that is, for all mail (including junk mail and□Error rate: err = 1 legitimate e-mail) the rate of the sentence wrong.

(4) F-measure:

$$F = 2\frac{recall \times precision}{recall + precision}$$

F-measure is the recall rate and accuracy of harmonic average, it will recall rate and precision into a comprehensive indicator.

In addition, spam filtering is often used in the Fallout, Miss rate and so on.

In order to verify our proposed method of feature extraction and classification of CeEP after combining the classification system in the classification of the efficiency of e-mail, the paper has done three sets of experiments: (1) parameters for different values BeEPJMC filtering algorithm for classification and evaluation results ; (2) with parameters fixed to the assessment of changes in the growth rate of trend indicators; (3) Comparison with other algorithms.

Experimental environment for Pentium 4 CPU, 256MB RAM, 80GB hard drive, the operating system to Microsoft Windows XP, programming software for the Microsoft Visual C++. Net 5.0. Experiment using 10 fold cross-validation approach to the statistical results of the classification of mail. That data sets will be divided into ten mutually exclusive subsets of pay or "discount" DB_1, DB_2, ..., DB_{10}, roughly equal the size of each pack. Training and testing carried out 10 times. i times in the first, DB_i used as a test set, a subset of the remaining are used for training classifiers.

A. for different values of parameters BeEPJMC the results of the evaluation algorithm

Experiments found that Balancing factor, α and β values is very important because the balancing factor α and β values affect the order of characteristics. too small or too big can not extract the characteristics of a good item. *c* value of the selected test results for the equally important, the *c* values for different characteristics of the selected items have a great impact. After a large number of experiments show that, α and β, □ better select for α and β, countdown each other, and the value of 2 and 3. c the value of a good choice□ for between 0.30 to 0.60. And the growth rate r of the classification threshold evaluation of the results of the impact is not great. Therefore, the main consideration in

the experiment balance factor α and β value of the classification results. Figure 1 Figure 2, respectively, □better values are given access (α = 2, β= 1 / 2) and (α= 3, β = 1 / 3), when our results.

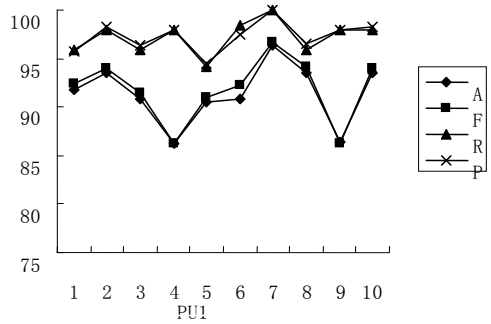

Figure 1. α，β were 2，1/2 PU1 Experimental Evaluation Results

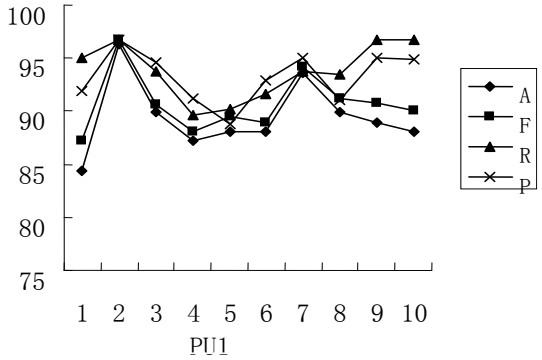

Figure 2. α，β were 3，1/3 PU1 Experimental Evaluation Results

Compare Figure 1, Figure 2 we can see that the balance factor, respectively 2,1/2 of α and β at the time of the recall rate and accuracy are respectively higher than the 3,1/3. however, accuracy and F-measure of 3,1/3 are respectively higher than the 2,1 / 2. Also found in PU1 corpus set part2 and part7 four evaluation criteria are the highest. respectively 2,1/2 α and β, the recall rate and precision of part7 were 100%. With the higher recall rate and the omission of the less spam, the higher the accuracy of miscarriage of justice would be a legitimate message as spam the possibility of the smaller. The two evaluation criteria to improve our spam classification and filtering efficiency is very important, so BeEPJMC algorithm to achieve a better classification's ability to spam.

B. Comparison with other algorithms

PU1 corpus in the same set, we use the current favorable Naive Bayesian (Nbayes) classification algorithm (reference[9] The evaluation results) as well as the more popular KNN algorithm (reference[10] The evaluation results), Decision Tree Algorithm (Decision Tree) (reference[8] The evaluation results) and Bayesian neural networks (Bayes network) algorithm (reference [11] The evaluation results) the results of BeEPJMC algorithm and compare the results of Table 2 below (Table 2 the results are listed in spam and non-spam classification of the results of the weighted average):

Table 2.Five Algorithms Classification and Filtering Evaluation Results

	A	*F*	*R*	*P*
Nbayes	0.9318	0.9415	0.9191	0.9578
KNN	0.977	0.9324	0.935	0.9298
Decision Tree	0.891	0.8805	0.882	0.879
Bayes Network	0.892	0.9362	0.9466	0.926
BeEPJMC	0.9532	0.9526	0.9538	0.9517

A representative of the accuracy, F on behalf of F-measure, R on behalf of the recall rate, P on behalf of precision.

From Table 2 we can see that the recall classification BeEPJMC and F-measure are the highest. The precision of BeEPJMC is the same as the highest precision of the Naive Bayesian algorithm. KNN algorithm is to the highest rate of accuracy. Decision tree of the evaluation index have reached the minimum. For spam classification algorithm the recall rate and the improvement of accuracy is particularly important, we can see BeEPJMC algorithm has a very high classification and spam filtering capabilities.

5. Prospects

In this paper, e-mail classification and filtering methods are discussed. A new feature extraction method are proposed. Combined with the basic exposure model based on (eEPs) classification algorithm CeEP, we realized the Bagging eEP-based classifiers for junk mail classification algorithm BeEPJMC. Experiments show, BeEPJMC is a very efficient method of e-mail classification and filtering. The next step will be to expand spam BeEPJMC used for other data sets. Spam BeEPJMC applied to other areas of data sets.

References

[1] P-N Tan, M Steinbach, V Kumar ,F ming, F hong jian. “Introduction to Data Mining. Beijing: posts and telecom press”, **(2006)**, pp. 259-293.
[2] H Alhammady and K Ramamohanarao, “The Application of Emerging Patterns for Improving the Quality of Rare-class Classification”, In Proc. of the 8th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD2004), Sydney, Australia, May **(2004)**, pp. 207-211.
[3] I. H. Witten and E. Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, **(2005)**.
[4] P Meesad;P Boonrawd;VNuipian, “A Chi-Square-Test for Word Importance Differentiation in Text Classification”;Proceedings of International Conference on Information and Electronics Engineering （ICIEE 2011;**(2011)**.
[5] J.R. Quinlan. Bagging, Boosting, and C4.5. In Proceeding of the 13th National Conference on Artificial Intelligence, AAAI/MIT Press. **(1996)** , pp. 725-730.
[6] F Ming, W Fang, “Mining Essential Emerging Patterns for Classification Computer Science” (Supplement) (in Chinese), **(2004)**, pp. 307~309.
[7] A Khan, B.Baharum Bahurdin, K Khan, “An Overview of E-Documents Classification[A]”;Proceedings of International Conference on Machine Learning and Computing(ICMLC 2009);**(2009)**
[8] A New Feature Selection Method Based on Distributional Information for Text Classification;Proceedings of the 2010 IEEE International Conference on Progress in Informatics and Computing;**(2010)**
[9] Z lei, Hu Y Sheng, C De Xuan *et al*, »An anti-spam filtering algorithm based on cost minimization. Huangzhong university science and technology(Nature Science Edition) ,**(2005).**vol. 12, no.33, pp. 352~355.
[10] K Mikawa, T Ishida; [A];Proceedings of 12th Asia Pacific Industrial Engineering & Management Systems Conference(APIEMS 2011);**(2011)**
[11] L Zhen, ZMing Tian. Spam filtering algorithm based on supervised Bayesian parameter estimation. Computer Application, vol. 26, no. (3), **(2006)**, pp. 558~561.

Author

Yan Li (1981.11-), female, master, lecturer, research direction: data mining, information security and cloud computing.

HuaZhou(1981.01-), female, master, lecturer, research direction: Network technology，datamining and cloud computing.

