논문 2010-2-3

로지스틱 회귀 분석을 이용한 스팸 필터링의 특징 축소

Features Reduction using Logistic Regression for Spam Filtering

정용규*, 이범준**

Yong-Gyu Jung, Bum-Joon Lee

요 약 오늘날의 스팸 메일이 메일 서버와 네트워크 저장장치의 대부분을 차지함으로 인해 네트워크 부하와 같은 부정적인 문제가 발생하고 있으며 사용자 입장에서는 스팸을 삭제하기 위한 시간과 자원 소모 같은 문제를 가지고 있다. 자동 스팸 메일 필터링은 문제 해결위한 필수적인 요소로 부각 되었다. 대표적인 방법은 나이브 베이지안 방법과 달리 PCA를 통하여 많은 차원을 가지는 스팸 테이터 집합을 몇 개의 주축으로 차원을 축소 시켜 연차 처리의 부담을 줄이고 특정 집으로 분류를 위한 로지스틱 회귀 분석 방법을 사용하여 스팸 필터링을 하였다. 이를 통하여 속도와 성능 두가지의 성과를 얻을 수 있었다.

Abstract Today, The much amount of spam that occupies the mail server and network storage occurs the lack of negative issues, such as overload, and for users to delete the spam should spend time, resources have a problem. Automatic spam filtering on the incidence to solve the problem is essential. A lot of Spam filters have tried to solve the problem emerged as an essential element automatically. Unlike traditional method such as Naive Bayesian, PCA through the many-dimensional data set of spam with a few spindle-dimensional process that narrowed the operation to reduce the burden on certain groups for classification Logistic regression analysis method was used to filter the spam. Through the speed and performance, it was able to get the positive results.

Key Words : Logistic Regression Analysis, Feature Reduction, Principal Component Analysis, Spam mail

I. Introduction

All of the spam E-mail accounts for 90% of the moment, and even junk mail is being forwarded. The amount of spam that occupies much of the mail server and network storage capacity of the lack of negative issues, such as hatching, and for users to delete the spam should spend time, resources have a problem. Automatic spam filtering on the incidence to solve the problem is essential. Automatic spam email filtering for spam is subject to the learning dataset, the process of feature extraction and classification will be processed through. Previous studies looking at the spam filtering using SVM, artificial neural network, the decision tree with category and rule-based Classifier through Categories has been researched as An upper abuse, pharmaceutical products and related words, HTML is the same color and are included in the rules. But spammers sense of these rules, the problem can be avoided. And the class conditional probability calculations using effective, reliable document classification most widely used method, the basic document classification Naive Bayesian Classifier.

In this paper, PCA (Principal Component Analysis) through the many different dimensions that a lot of spam data set contains information on the dimensions

*종신회원, 을지대학교 의료IT마케팅학과
**정회원, 을지대학교 의료산업학부 의료전산학전공
접수일자 2010.4.1, 수정일자 2010.4.19

shrink by a couple of operations to reduce the burden of handling classified into specific groups for the logistic regression analysis using the spam filtering improves the classification performance and speed.

II. Related research

1. PCA

PCA to the characteristics of a multidimensional vector data consisting of information about maintaining a high level in low-dimensional multivariate data processing, is the one of ways to reduce the dimension.

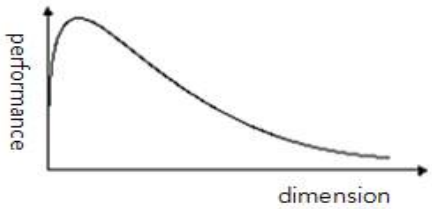

그림 1. 차원의 저주
Figure 1. Dimensional Curve

Dimensions can be reduced by multivariate data corresponding to the main component of the spindle by a statistical method to obtained feature vectors.

Distributed to the largest and the spindle and spindle No. U_1 and Article U_2 of the division spindle in two different spindle has to be vertical.

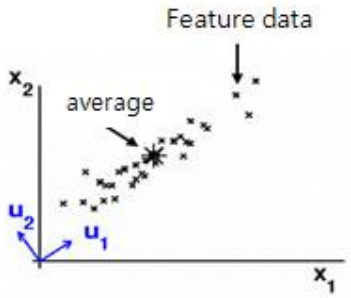

그림 2. 특징 벡터의 분포
Figure 2. Distribution of feature vectors

By the end of the axis by converting the feature vector is relocated to

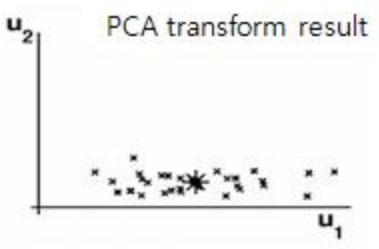

그림 3. 주축 변환 결과
Figure 3. Spindle conversion results.

Through PCA for dimension reduction can be divided into five stages. The first, covariance matrix calcurated

$$\mu = \frac{1}{N}\sum_{n=1}^{N} X \tag{1}$$

$$\sum = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^T \tag{2}$$

N is total number of feature X and μ is mean of feature X

The second, eigenvector is represented as follows

$$\sum = U \Lambda U^T \tag{3}$$

$\sum$ is an n-by-n covarianc matrix, U is an lengh n eigvectors vector and Λ is scalar. the n values Λ that satisfy the equation are the eigenvalues.

The third, maximum eigenvalue is selected

$$(\lambda_{1,\ldots}\lambda_n) \tag{4}$$

The fourth, eigenvalues associated with the selected vector obtained as a unique transformation matrix W

$$W = [U_1, \ldots U_n] \tag{5}$$

The fifth way to convert the feature vectors

$$y = W^T x \tag{6}$$

2. Logistic Regression Analysis

Linear regression analysis of existing categories of the dependent variable for analysis purposes simplifies the type extends letting the decision to form a binary category Continuous data and a number of different categories of data are converted into binary data, the operation is usually performed by One belongs to a group situation, using values for the predictor variables can be classified into certain groups.

Logistic regression analysis, how to go through a two-step process. The first step is the probability of belonging to each group was estimated to calculate the probability of belonging to group 1 as P (Y = 1) gets the estimate. The next step for each observed value for any one group classified as a reference value by applying the category of those likely to category.

Logit function using the formula (1) and uses the same standard formula. x is independent variable and β is weight value of independent variable

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q)}} \qquad (7)$$

3. Rule-based Classifier

E-mail filtering, an important part of the unconscionable reliability is a problem with the sender and receiver. E-mail from someone you trust if you are unable to read messages from people you trust, but if the whole message is considered spam. The issue can not know all the sender's reputation as a regular mail client using the address book and trusted senders are identified. Learning when to store a list of spammers and instead use the address book, and two on the list are duplicated by the sender if the sender does not filter. If the sender is not on the list of the sender does not know whether to trust the sender can not be filtered.

To represent the contents of the mail if the subject line 'advertising' the phrase used only for filtering of spam recently, '/advertising','advertising/'as a shortcut to avoid filtering because these partly special character between words to be checked and filtered.

The size distribution as a general mail filtering to remove the HTML content of the body have been accomplished by so much so that the body size of 50KB or more, such as sales and distribution of pirated programs are listed as food distribution is considered spam. In addition to the form or the attachments, instantly, the sender of the email address domain, contained in the body in the form of a Moongo time mail is sent, the sender ID in the form of emails, you can see the list filtering, etc., but also large potential incorrectly classified.

4. Naive Bayesian Classifier

Naive Bayesian Classifier is the simplest model. It's "Naive" called, all properties in a document within a given class is independent assumption. This assumes that most of the problems as in the real world, despite the obvious lie often Naive Bayesian Classifier seems a good performance. but the function approximation accuracy in the high is insufficient. This is a large number of independent property. Separately for each property can be learned, it makes learning easy.

Caused by the frequency of words in the document represents the document. Not consider the order of the above words, the occurrence of each word in the document in the document to calculate the probability, caused by multiplying the probability of a word can be considered. Here we separate the occurrence of the word "event" can be understood as the document may be thought of as a set of events. This event is the model is polynomial. E-mail when you perform the classification using the polynomial model Classifieds.

III. PCA and logistic regression analysis

Conventional logistic regression analysis applied to the spam filtering performance as 92% in the category showing high performance, but to perform real-time speed 2.72 seconds quickly to deal with the spam filter does not fit.

This paper to solve these problems and made up multidimensional data dimension reduction, while maintaining the information processing operations by reducing the burden of sorting through logistic regression improves performance. Figure 3 using the PCA is a dimension reduction step.

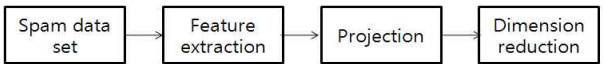

그림 4. PCA를 통한 차원 축소 단계
Figure 4. PCA dimension reduction steps

Dimensional dataset, a logistic regression analysis, collapsed over the category, the steps are displayed. Features probability calculate the value of the formula in step (1) After calculating the value of p compared to clams in the category, the category will be compared with reference value. Figure 4. Logistic regression analysis is a step through the category.

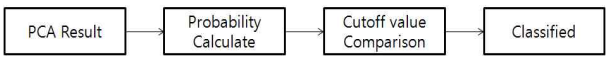

그림 5. 로지스틱 회귀분석을 통한 분류 단계
Figure 5. Logistic regression analysis with step category

Ⅳ. Experimental methods and results

1. Experimental methods

Naive Bayesian spam filtering performance and the performance of the proposed method in this paper for comparison with data collected from the e-mail 'spambase.data' was used for the experiment was using weka.

Algorithms

Input : S
Output : C_{spam},$C_{non-spam}$
  //pre-processing
  S' = PCA(S)
    // Logistic regression
   pValue = $P(S'_i = 1)$
      if(pValue > Cutoff value)
        return C_{Spam}
      else
        return $C_{Non-Spam}$

그림 6. 동작 알고리즘
Figure 6. Operating algorithm

$S=\{S_1, S_2, \ldots, S_{4601}\}$ which consists of a set of data S_i that $S_i=\{x_1, x_2, \ldots . x_{58}\}$ has been configured. Often indicate the occurrence x_i of the word appeared. Dataset is classified into two classes, Class C, '0 'or '1' has a binary value. '1 'If the Spam, '0' means that one is the Non-Spam. To suggest, as shown in Figure 4 shows the algorithm.

표 1. 분류 성능 비교표
Table 1. Category Performance Comparison

	Naive Bayesian	Proposed Method
Correct	79.2871%	84.5789%
Incorrect	20.7129%	15.2141%
Spped	0.23 seconds	0.22 seconds

2. Experimental Results

The entire process from data preprocessing with PCA dimensions gradually reduce gamyeo logistic regression analysis was conducted to pre-process the entire first 58 properties in the properties of the dimension was reduced to 48. Figure 5 shows the accuracy category, according to the number of dimensions. Reduce the number of dimensions, but losing more accurate than the Naive Bayesian methods are showing high performance.

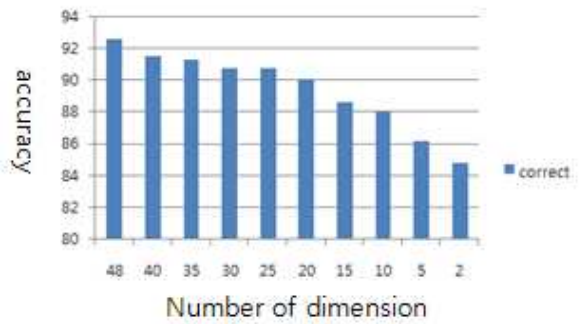

그림 7. 정 분류 변화율
Figure 7. Change Correctly Classified

Figure 6 is based on the number of dimensions represents the rate of change is performed. For the first 48 of the properties do not have the speed difference

between the level before collapse, but the number of dimension reduction performed faster as you can see in.

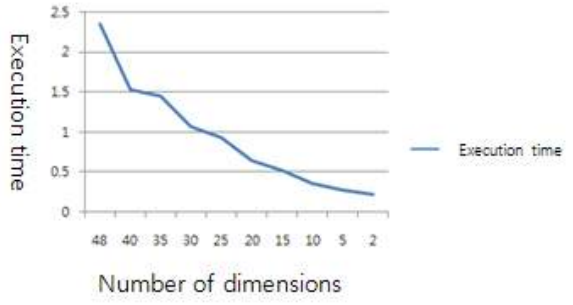

그림 8. 수행 속도 변화률
Figure 8. Change execution speed

V. Conclusion

Using logistic regression analysis of existing spam filtering, the classification performance increases to high speed depending on the number dimensions. Dataset representative of an existing algorithm, Naive Bayesian classification algorithm performs slower than had any problems. In this paper, applying the PCA performed using logistic regression analysis, the speed aspect was trying to solve. Naive Bayesian, the rate of 79.2871% of the dragon looked jeongbun the data of the proposed technique to 84.5789% despite two-dimensions. Classification performance was also appeared to perform best when viewed in terms of speed of 0.23 and 0.22 seconds showed a similar result. Both methods have similar performance, showing that one in terms of overall performance, the proposed technique has been excellence. Value of Refined reduce error rates by adjusting the formula by a higher performance category will be considered in future.

References

[1] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," AAAI Technical Report WS-98-05, 1998

[2] Vikas P. Deshpande, Robert F. Erbacher, and Chris Harris“ An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques” Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY 20-22 June 2007

[3] Toby Segaran, “Programming collective intelligence”, O'REILLY, 2007

[4] Ian H.Witten, Frank Eibe, "Data Mining: Practical Machine Learning Tools and Techniques" Morgan Kaufmann, 2000

[5] Pang-Ning Tan & Michael Steinbach & Vipin Kumar, “Introduction to Data Mining”, ELSEVIER, 2006

[6] H. Drucker, D. Wu, and V. N. Vapnik., “Support Vector Machines for Spam Categorization”, IEEE Trans. on Neural networks, 1999.

[7] D. Mertz, “Spam Filtering Techniques. Six approaches to eliminating unwanted e-mail.”, Gnosis Software Inc., September, 2002. Ciencias Físicas, Universidad de Valencia, 1992.

[8] M. Vinther, “Junk Detection using neural networks”, MeeSoft Technical Report, June 2002. Available: http://logicnet.dk/reports/JunkDetection/JunkDetection.htm.

[9] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. “Indexing By Latent Semantic Analysis”, Journal of the American Society For Information Science, 41, 391-407. (1990)

[10] Jiawei Han, Micheline Kamber, “Data mining - Concepts and Techniques”, Morgan Kaufmann Publishers, 2001.

저자 소개

정 용 규(종신회원)

- 1981년 서울대학교 (이학사)
- 1994년 연세대학교 (공학석사)
- 2003년 경기대학교 (이학박사)
- 1999년~현재 을지대학교 교수
- 2001~현재 ISO/TC154K위원장
- 2005~현재 산업표준(KS)심의위원

<주관심분야: 임상데이터마이닝, 의료정보시스템, 전자거래표준>

이 범 준(정회원)

- 2007년~현재 을지대학교 의료산업학부 의료전산학전공

<주관심분야: 임상데이터마이닝, 의료영상인식, 생체정보전달 >