Research On Spam Filter Based On Improved Naive Bayes and KNN Algorithm

BiyiRen[1, a], YuliangShi[2,b]

[1]School of Software, Beijing University of Technology, Beijing100124, China;

[2]School of Software, Beijing University of Technology, Beijing100124, China.

[a]renbiyi@126.com, [b]shiyl@bjut.edu.cn

Keywords: spam filter, Naive Bayes, KNN.

Abstract. In the field of data mining and pattern recognition, classification is a very important core technology. This paper present two kinds of improved classification algorithm.Using the improved Naive Bayes (NB) and KNN algorithm structure classifier to filter normal mail and spam. Improved NB algorithm can dynamically adjust the threshold k, reduces the mail mistake rate.Center vector method is introduced into the similarity calculation formula of KNN, better reflect the interrelation between the text and categories. Finally, improvedNB algorithm and KNN algorithm make comparison and analysis, it is concluded that the effective experimental results.

Introduction

With the development of Internet and information technology, E-mail has become a widely popular way of communication in people's lives. Email provide convenient and quick, but spam has become more rampant. spam disrupts people's normal communication, caused the attention of people from all walks of life. According to the Internet society of China issued"in the first quarter of 2014 China anti-spam survey report"points out, the average number of user receive spam E-mail per week is 11.4, spam accounts for 26.5% of the total number of e-mails. So the effective spam filtering method has very important practical significance.The rest of this paper is organized as follows: Section 2 describes the core spam filter algorithm, Section 3 presents experiments and evaluation.The summary is showed in Section 4.

Spam FilterAlgorithm

Naive BayesAlgorithm.

Principle. Bayes theorem is a British mathematician Thomas Bayes in a paper published in 1763.Bayesian theory assumes that: if the result of the event are uncertain, the only way to quantify it is the probability of the event. Bayesian theory can be expressed in a mathematical formula, that is the Bayesian formula.

The Bayesian formula:assuming that experiment E, the sample space is S, A is a event of E, $\{B_1, B_2 \ldots B_n\}$ is the division of S, B_n is mutually exclusive events, and $P(A_i) > 0$, $P(B_i) > 0$, thenfor any event A.

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{\infty} P(A|B_i)}, (, (\ i = 1,2,3...n\ ) \quad (1)$$

Naive Bayes classifier is simple and effective, withoutregard to dependencies feature items, so the amount ofcalculation is reduced[1].This simple classifier relies on two assumptions: one is each attribute is conditionally independent from the other attributes given the class and another is all the attributes have influence on the class [2].

Assuming that A_x present eigenvalue set $\{a_1,a_2,...a_n,\}$, assumed every eigenvalue is independent. $P(A_j)$ is belong to the probability of spam, P(B) is a probability of text category. Judging whether a email is spam or not,based on Bayesian formula and the full probability formula:

$$P(B_j|A_x) = \frac{P(B_j)P(A_x|B_j)}{P(A_x)} \propto P(B_j)P(A_x|B_j),(\ j = 1,2,....n\ ) \quad (2)$$

$P(A_x|B_j) = P(a_1|B_j) * P(a_2|B_j) *\ldots* P(a_n|B_j) = \prod_{j=1}^{n} P(a_j|B_j)$ (3)

$P(B_j)$ is the priori probability, $P(A_x|B_j)$ is the conditional probability, $P(A_x)$ unchanged in the same article.Assumed features is independent in between. $P(B_j)$ and $P(A_x|B_j)$can be estimated from the training set.

Improvement. In the real classification of email, mail classification not only should consider how to make the right judgments, andconsider to make the wrong judgment will bring what consequences. Email filter will be two kinds of circumstance of classification error: One is normal mail isclassified as junk mail. Another is spam is classified as mormal mail. That normal mail is classified as spam may cause serious consequences of the user. Therefore, we need a filtering algorithm can minimize the loss as far as possible.

In the front section, the Naive Bayesian formula calculate a mail which to beclassified.The probability of A_x belongs to the legitimate mail and spam are:$P(B_1|A_x)$(B_1 as spam class) and

$P(B_0|A_x)$ (B_0 is the normal mail class).In traditional methods, generally when $P(B_1|A_x)>P(B_0|A_x)$, determine the A_x mail as spam, or convicted of normal mail. However, this judgment is not accurate, will lead to a higher rate of miscarriage of justice.

In order to more accurately identify the spam.Assuming that when $\frac{P(B_1|A_x)}{P(B_0|A_x)} > \theta$, that is when the probability of A_x is spam is θ times bigger than the probability of the normal mail,it will be considered as spam. Otherwise it will be considered as normal mail.Whenthe bigger the θ value, the greater the possibility of it as spam.

$\frac{P(B_1|A_x)}{P(B0|A_x)} = \frac{P(B_1|A_x)}{1-P(B_1|A_x)} > \theta \Rightarrow P(B_1|A_x) > \frac{\theta}{1+\theta} = k$(4)

That is when $P(B_1|A_x) > k$, the A_x isconsidered as spam.According to the Eq. 4,we get a k value, the k values range between 0 and 1, so through setting the k values you can judge mail, reduced the amount of calculation.Finally we can obtain relatively satisfactory results.

KNN Algorithm.

Principle.K nearest neighbor method (KNN) by the Cover and Hart, is a kind of lazy, supervised, and machine learning method based on the instance. KNN classification method is an instant based learning algorithm that categorized objects based on closest feature space in the training set [3].

The method in the paper is: the similarity of K texts according to the formula for sum, will belong to the same class, then for each class sort sum. Put the unclassified text to the classification of the larger similarity's sum. Formula is as follows:

$T_j(d) = \sum_{i=1}^{k} T_j(d_i)Sim(d, d_i)$(5)

K is the num of selected text, T_j (d_i) indicates whether the text di belong to C_j (if belong to, the value is 1, otherwise the value is 0), Sim (d, d_i) is angle cosine formula:

$Sim(d_1, d_2) = \frac{\sum_{i=1}^{n} W_{1i}W_{2i}}{\sqrt{\sum_{i=1}^{n} W_{1i}^2 \sum_{i=1}^{n} W_{2i}^2}}$(6)

W_{1i} and W_{2i} indicate the weight of feature of d_1 and d_2 in the text vector. Two vector Angle is smaller, the greater the cosine value. The text of the two vectors represent is more likely to belong to the same category and vice versa.

According to the Rocchio formula,the center vector method presented by the Hull at the earliest.It is a kind of classification method using the vector space model.It can reflect to the relevance of the classification between text and a category. The algorithm idea is: represent Unclassified text into a text vector,and then calculate the class centre vector distance of the text with the training text set,assigning the text to the class with the smallest distance .Using inner product of vectors method to calculate distance,that is:

$C_i = \frac{1}{n}\sum_{k-1}^{n} d_{ik}$(7)

n represent the number of class i ,d_{ik} represent the k texts in class i.

Unclassified text and class center vector reflect the relevant degree of text and the categories. Similarity was calculated by the vector inner product, namely the vector distance, notes for VD

(Vector Distance, VD). This paper use the ratio of features number of two texts appear and the maximum features number of each text, it work together with VD. As the cosine of the Angle adjustment factor. we use the CM (Common and Maximum, CM) present the ratio.

The improved similarity calculation formula is:

$$\mathrm{Sim}(d_1, d_2) = \left(\sqrt{\mathrm{VD}} + \mathrm{CM}\right) * \frac{\sum_{i=1}^{n} W_{1i} W_{2i}}{\sqrt{\sum_{i=1}^{n} W_{1i}^2 \sum_{i=1}^{n} W_{2i}^2}} \quad (8)$$

Experiments and Evaluation

Naive Bayes Algorithm Experiments. For the improved Naive Bayes(INB) algorithm, I can get ak value that was derived by the formula (4),the k values range between 0 and 1,so through set the k values to the judge email classification, reduced the amount of calculation.After a lot of experimental data, found:

(1) when the $0<k\ll0.39$,the recall rate of the system was 100%, which can ensure the system to judge the mail is really legal mail, there is no spam wrongly as legal mistakes.

(2) when the $0.6\ll k<1$,the accuracy of the system is 100%, which ensure the judging of system for junk mail are real spam, there is no legal email judge as spam.

(3) when the $0.39<k<0.5$,the k value is more close to 0.39, junk mail was more less judge as legal email;

(4) when the $0.5<k<0.6$,the k value is more close to 0.6,the less legal email was mistaken to spam.

Thus, when k = 0.6, It can make the legal mail was mistaken as spam in the minimum degree.

Table 1 corresponding effect is shown inFig.1 to Fig.3:

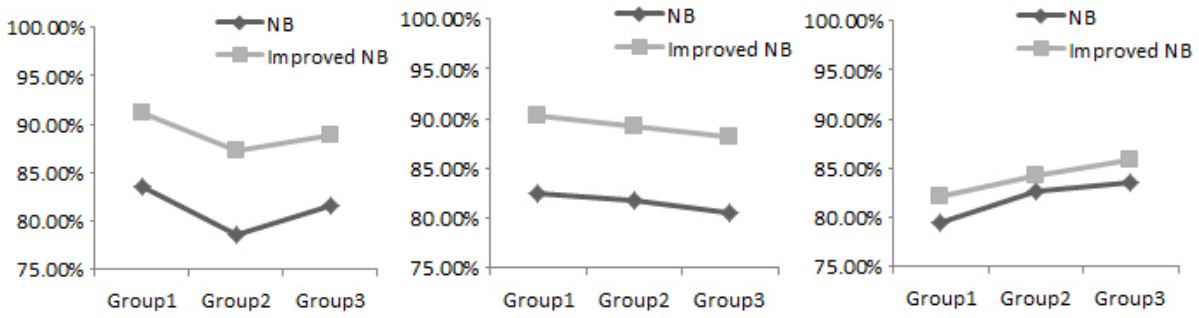

Fig. 1 The **recall** contrastFig. 2The **precision**contrastFig. 3The accuracycontrast

According to the experimental results, we can see that this paper mail filtering system has higher recall, precision and accuracy, the system performance is relatively stable, thus proving the validity and high efficiency of the improved algorithm. As shown in Fig.1 to Fig.3,the recall and precision of email filtering system improved by 5% than previous, accuracy improved by 3% than before.

KNN Algorithm Experiments. Respectively using KNN and its improved algorithm classifying test sample set, through comparing the three sets of classification results show the effectiveness and feasibility of the improved KNN(IKNN).

Table 2 corresponding effect is shown inFig.4 to Fig.6:

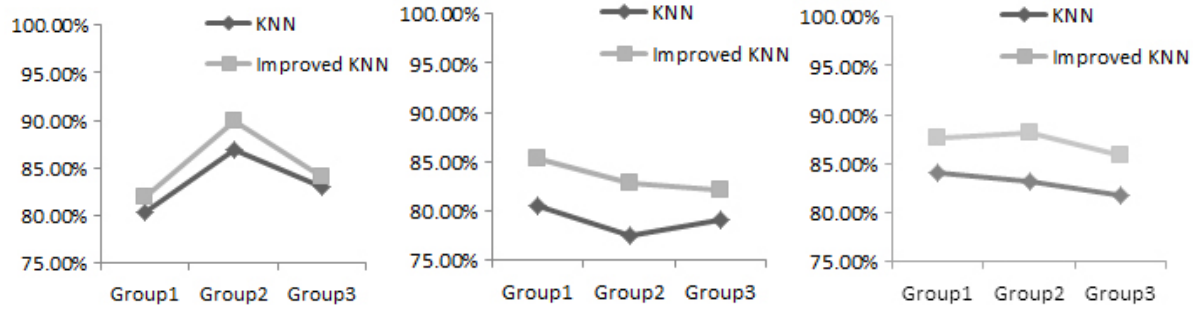

Fig.4The **recall**contrastFig.5The **precision**contrastFig.6The accuracycontrast

We can see from Fig4 to Fig6, if evaluate the classification effect on the whole, the effect of improved KNN classification is about 3% higher than using KNN in the recall,**precision,** accuracy. So improved KNN has better classification effect.

Summary

Based on Naive Bayes and KNN algorithm of spam filters is currently one of the more efficient spam filtering technology, it has been widely used in the field of spam filtering. The paper aiming at the defect of Naive Bayes proposed to improved algorithm, this algorithm can by adjusting the k value, to reduce the probability of legal mail was wrongly as spam.At the same time, by introducing the idea of class center vector method and the number of the features in two texts to improve the ability classification of KNN.

It is necessary to research a effective text classification method, which will play a better role in spam filtering. Next, we still need a lot of experiments.

References

[1]. Lin Li; Chi Li, "Research and Improvement of a Spam Filter Based on Naive Bayes," in Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on , vol.2, no., pp.361-364, 26-27 Aug. 2015.

[2]. de Campos, L.M.; Cano, A.; Castellano, J.G.; Moral, S., "Bayesian networks classifiers for gene-expression data," in Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on , vol., no., pp.1200-1206, 22-24 Nov. 2011.

[3]. Gayathri, K.; Marimuthu, A., "Text document pre-processing with the KNN for classification using the SVM," in Intelligent Systems and Control (ISCO), 2013 7th International Conference on , vol., no., pp.453-457, 4-5 Jan. 2013.

[4]. Harisinghaney, A.; Dixit, A.; Gupta, S.; Arora, A., "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm," in Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on , vol., no., pp.153-155, 6-8 Feb. 2014.

[5]. Ying Wang; Hao Wang; Kui Yu; Hongliang Yao, "L1 regularized ordering for learning Bayesian network classifiers," in Natural Computation (ICNC), 2011 Seventh International Conference on , vol.3, no., pp.1522-1526, 26-28 July 2011.

[6]. Han-Bing Yan; Ya-Shu Liu, "Spam filter based on incremental Bayes arithmetic," in Electrical and Control Engineering (ICECE), 2011 International Conference on , vol., no., pp.6111-6114, 16-18 Sept. 2011.

[7]. Harisinghaney, A.; Dixit, A.; Gupta, S.; Arora, A., "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm," in Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on , vol., no., pp.153-155, 6-8 Feb. 2014.

[8]. Lijun Wang; Xiqing Zhao, "Improved KNN classification algorithms research in text categorization," in Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on , vol., no., pp.1848-1852, 21-23 April 2012.