Research of a Spam Filter based on Improved Naive Bayes

Ye Yuan[1,a] ,Shouzheng Li[2,b] and Yuanyuan Wang[3,c] , Chao Liu[4,d] , Weimiao Feng[5,e] ,and Min Yu[6,f]

[1]Harbin Engineering University, Harbin , China, 150001

[2]Harbin Engineering University, Harbin , China, 150001

[3]Northeast Forestry University, Harbin , China, 150001

[4]Institute of Information Engineering,Chinese Academy of Sciences,Beijing,China,100093

[5]Institute of Information Engineering,Chinese Academy of Sciences,Beijing,China,100093

[6]Institute of Information Engineering,Chinese Academy of Sciences,Beijing,China,100093

[a]569093011@qq.com,[b]26407166@qq.com, [c] stdong0451@163.com, [d]2443532866@qq.com, [e]330227346@qq.com, [f]272523869@qq.com

Keywords: Naive Bayes,SVM,spam mail,trim

Abstract. In spam filtering filed, Naive Bayes algorithm is one of the most popular algorithms, but its conditional independent assumption makes its reliance on training sets of sample space distribution. In order to improve the accuracy rate, sample space became so complex, resulting in the algorithm of time complexity is increased and the internal stability is poor. In order to solve above problems, this paper proposes a modified using support vector machine(SVM) of the Native Bayes algorithm :SVM-NB.First,SVM constructs an optimal separating hyperplane for training set in the sample space at the junction two types of collection,then according to its similarities and differences between the neighboring class mark for each sample to reduce the sample space also increase the independence of classes of each samples,finally using Naive Bayesian classification algorithm for mails. The simulation results show that the algorithm reduces the sample space complexity, fast to get the optimal classification feature subset,effectively improve the classification speed and accuracy of spam filtering.

Introduction

Along with the explosive growth of network information, E-mail turns to one of the most popular way of communication in daily life.But then recoiled as spam problems, according to the latest survey data which is showed by China send junk union[1], the user email receive 35.0 mail per week on average, but Spam accounted for 41% of the total, spam affected people's life and work seriously. Therefore, security and reliability of the mail system become the focus of people's attention.

Naive Bayes algorithm model

The classification principle of Naive Bayes text is to solve the probability value $(P_1, P_2, ..., P_n)$ of the vector $X(x_1, x_2, ..., x_n)$ which belongs to the category $C(C_1, C_2, ..., C_j)$, and P_j is the probability of $X(x_1, x_2, ..., x_n)$ which belongs to C_j, the $\max(P_1, P_2, ..., P_n)$ corresponding category is the category text X belongs to the category, therefore, classification problem is described as solving the maximum value of the equation(1).

$$P(c_j \mid x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n \mid c_j)}{P(c_1, c_2, ..., c_n)} \quad (1)$$

$P(c_j)$ is the probability in the training text which belongs to the category c_j.

$P(x_1, x_2, ..., x_n \mid c_j)$,if the text to be classified belongs to the category c_j, the probability c_j included in the category of vector $(x_1, x_2, ..., x_n)$.

$P(c_1, c_2, ..., c_n)$ is the joint probability of given all categories.

Obviously, the denominator $P(c_1, c_2, ..., c_n)$ is a known constant for any given category, so to simplify the formula(1) for solving the maximum number of formula(2).

$$c_{NB} = \arg\max_{c_j \in C} P(x_1, x_2, ..., x_n \mid c_j) P(c_j) \tag{2}$$

According to the hypothesis of Naive Bayes, the text feature vector attributes $x_1, x_2, ..., x_n$ identically distributed, and the joint probability distribution is equal to the product of the probability distribution of each attribute, that is

$$P(x_1, x_2, ..., x_n \mid c_j) = \prod P(x_i \mid c_j) \tag{3}$$

So

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod P(x_i \mid c_j) \tag{4}$$

Mentioned above in the Naive Bayes algorithm and its improved algorithm,their use is the basic principle of Naive Bayes, just relax the assumption of independence, but in fact attributes which are mutually dependent attributes are still present in training set among. It can be seen from formula(4) when the final the probability of the text category has been calculated, these algorithms are encountered some bottlenecks, because it used conditional independence assumption, the actual property which is not independent from each other limits the performance of the algorithm, especially in accuracy and recall rate. So, is there an algorithm that can be applied the conditions of independence assumptions to the real world？ If an algorithm based on all those involved to calculate the properties of the sample set according to whether the associated process, that is, if there is a relationship and not independent between two attributes, determine whether the two attributes of the same category,and then deal with these two attributes based on the algorithm, this is the improved naive Bayesian algorithm NB - TSVM proposed in this paper.

Naive bayes algorithm based on SVM algorithm

（1）Support Vector Machine（SVM）

SVM because of the significant generalization ability and it is highly favored by people, the principle is constructed a hyperplane in the eigenspace, so that the width between the two types of structure to achieve the maximum distance that is the hyperplane structured by the distance is farthest, but must also make the class of the wrong points to minimize the punishment, it is the essence of SVM quadratic optimization problem.

In the case of separable training set, the SVM constructs an optimal hyperplane,

$$(w \quad x) + b = 0 \tag{5}$$

It makes the following sample set $(x_i, y_i), i = 1,2,...,n; x \in R^d, y \in \{+1,-1\}$,satisfy the constraint conditions

$$y_i \left[(w_i \quad x_i) + b \right] - 1 \geq 0, i = 1,...,n \tag{6}$$

And makes the punishment function minimum, namely

$$\phi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w \quad w) \tag{7}$$

By solving the optimization problem to be optimal hyperplane as follows：$\sum_{i=1}^{n} y_i a_i^0 (x \quad x_i) + b = 0$, a_i^0 is a Lagrange multiplier.

When the traning set can not be divided, the introduction of relaxation factor $\varepsilon_i \geq 0$ and the penalty parameter C , The minimization function is $\phi(w) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \varepsilon_i$ under the constraint condition

$y_i[(w_i \quad x_i)+b] \geq 1-\varepsilon_i, i=1,...,n$, the classification rules just take $I(x)=\mathrm{sgn}\left(\sum_{i=1}^{n} y_i a_i^0 (x \quad x_i)+b\right)$.

The introduction of kernel function is one of the major characteristics of the SVM algorithm, it is often difficult to differentiate low dimensional space vector set, then naturally thought map the low dimensional space to high dimension space, but the attendant will increase computational complexity, however, the kernel function cleverly solves the the problem. $K(x,y)=\phi(x) \quad \phi(y)$, ϕ represents a kind of mapping,as long as the appropriate selection of kernel function, we can obtain the classification function corresponding of higher dimensional space function $I(x)=\mathrm{sgn}\left(\sum_{i=1}^{n} y_i a_i^0 (\phi(x),\phi(x_i))+b\right)$,the $\phi(x)$ is higher dimensional vector than x (we don't need to know the specific form of ϕ),due to $K(x,y)=\phi(x) \quad \phi(y)$ only relates to x,y, and doesn't involve the high dimensional operation, so there is no increase in computational complexity.

（2）Improved Naive Bayes TSVM-NB

As mentioned above, the use of the premise condition of the Naive Bayes is that the attribute of the training set is independent of each other, With the principle of support vector machine is that you can find a perfect hyperplane, the two categories of the boundaries of the mixing will not appear if the distance between the two categories reaches maximum. However, in actual application, it has a serious impact on the recall rate and accuracy of the classification of the Naive Bayes algorithm because of this independence assumption is not true. In this paper,it is proposed a kind of improved Naive Bayes algorithm TSVM-NB by using the Support Vector Machine(SVM) pruning techniques to reduce the overlapping between properties, enhance its independence, and combining with the advantages of Naive Bayes algorithm which is the classification speed.

First of all, the training set by the Naive Bayes algorithm for the initial training to obtain the initial training classes and categories of binding of each vector in the training set, and then trim the training set by using the following algorithm.

Find the nearest neighbor of each vector point, and then for each vector point, keep the point if this point and its nearest neighbor belong to the same, or delete the point if this point and its nearest neighbor belong to the heterogeneous.

What is the nearest neighbor? How to find the nearest neighbors? Euclidean distance is used as the distance between the two vectors, namely set the two vectors as

$x_i(x_i^1,x_i^2,...,x_i^n)$, $x_j(x_j^1,x_j^2,...,x_j^n)$

Then the distance between x_i and x_j are defined as

$$D(x_i,x_j)=\sqrt{\sum_{k=1}^{n}(x_i^k,x_j^k)^2} \qquad (8)$$

The closest vector to a vector is the nearest neighbor of it.

We give the implementation method of above：give a training set which has been trained for the first time by Naive Bayesian algorithm $(x_1,y_1),(x_2,y_2),...,(x_m,y_m),x_i \in R^n, y_i \in \{-1,1\}, i=1,2,...,m$,the training set is expressed as a matrix

$$TR_{m\times(n+1)}=(XY), \mathrm{X}=\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_m \end{pmatrix}, \mathrm{Y}=\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{pmatrix}$$

Input： $X(x_1,x_2,...,x_m)$, $Y(y_1,y_2,...,y_m)$ are vectors in sample training sets；

Output：the sample classification vector $V(v_1,v_2,...,v_m)$ which is trained by TSVM.

Calculate the distances between every two vector, what's more, the distance of itself is infinite.

For i=1 to m

{

For j=1 to m

If(i!=j)

Calculate the $D(x_i,x_j)$

}

Find the nearest neighbor for each vector.

For i=1 to m

{

$MinD_x = \infty$

For j=1 to m

If ($D(x_i,x_k) < D(x_i,x_j)$)

$MinD_x = D(x_i,x_j)$

}

Determine whether the class mark of each vector and its nearest neighbor are consistent, if they are not consistent, then delete this vector.

For i=1 to m

{

If ($MinD_x = D(x_i,x_j)$)

if ($L_{xi} != L_{xj}$)

L represent the class mark of two vectors

Delete vector x_i

}

Classify the mail with clipped training sets with NB algorithms.

Algorithm implementation in spam filtering

The step and processes of simple implementation is mentioned above. In the following figure 2 shows add the classify algorithm to the classification, namely The concrete realization method of the classification, as shown in figure 2.

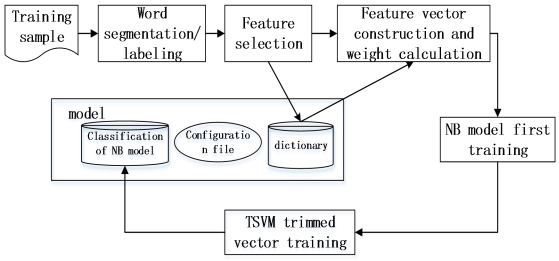

Fig.2 Spam filtering process

1） The automatic segmentation and text labeling are realized by ICTCLAS Chinese word segmentation system which is developed by Institute of Computing Technology Chinese Academy of Sciences with a large amount of normal mail and spam as the training sets.

2）Feature select to use the method of information gain, in the global scope where doesn't distinguish the spam and normal mail, calculate IG value of each feature X, and then sorted by IG value size, select the required number as a characteristic successively. Construct the feature vectors which are represent that email after finished the choice.

3）After construct the feature vectors, it should be done firstly to train the feature vector by using Naive Bayes algorithm, then the initial training set and its category of feature vector.

4） Clipping the feature vector of 3) with TSVM aims at reducing independence constraint among the attribute, namely reduce the dimension to reduce redundancy from the characteristic vector set. Then the clipped training set will be get.

5）In the end, Naive Bayesian algorithm classify the mail according to the clipped training set.

Conclusions

Based on support vector machine (SVM) algorithm and Naive Bayes algorithm (Naive Bayes), this paper aim at the conditional independence from each attribute which is the limit of Naive Bayes algorithm, use SVM to find the optimal plane, cut overlapping attributes, enhance the independence of properties, come up with improved Naive Bayes algorithm TSVM -NB, and evaluate it according to the accuracy and recall rate of the spam system's evaluation parameters. After a large number of experiments, it can be proved that this algorithm will improve the accuracy and recall on solving the spam to a certain extent.

To improve the algorithms mainly applies to the data set which is seriously crisscross or overlap among the attribute vector, namely in the case of classification is not particularly easy, if aliasing is relative weak among the attribute vectors, it will be hard to embody the advantage of improving the algorithms.

With the development of science and technology, spam is not just limited to the text form. There is a variety of forms such as the rubbish pictures, rubbish video and audio, the algorithm studied in this paper is just for the junk mails which are in the text form, how to filter pictures, video and audio efficiently will be in the next step research.

References

[1]http://www.anti-spam.org.cn/

[2]Yoon JiWon, Kim Hyoungshick,HUh, Jun Ho. Hybrid Spam Filtering for Mobile Communication. Computer & Security,2010,29(4):446-459

[3]He Haibo,Garcia Edwardo A. Learning form Imbalanced Data[J]. IEEE Transactions on Knowledge and Data Engineering. 2009 , 21(9):1263-1284

[4]Wu Xindong,Kumar Vipin,Quinlan J.Ross. Top 10 algorithms in Data Mining[J]. Knowledge and Information System. 2008 14(1):1-37

[5]Ruggieri.S. Efficient C4.5.IEEE Transactions on Knowledge and Data Engineering[J].2002,14(2):438-444

[6]Scholkopf.B, Mika.S, Burges.C, Knirsch.P, Muller.KR, Ratsch.G, Smola.A. Input space versus feature space in kernel-based methods[J]. IEEE Transactions on Neural Network,1999,10(5):1000-1017

[7]Friedman N,Geiger D,Goldszmidt M.Bayesian network classifiers.Machine Learning,1997,29(2-3):131-163