The Research on Text Classification algorithm based on SVM

Chenjie Zhang [a], Hanping Hu [b]

Changchun University of Science and Technology, Changchun 130022, China

[a]custzcj@163.com, [b]custhhp@163.com

Abstract. This article is based on Support Vector Machine Theory, with a SVM software package is used to realize a data classifier, data texts is the core aim of classification, namely, find the optimal hyper-plane to divide those data. Along with the change of data distribution and amount, the linear classifier without kernel function and the nonlinear classifier with Gaussian kernel function, they are both adopted here, and the result of the simulation is analyzed. Besides, in this article, including the comparison of the nonlinear classifier with Gaussian kernel function due to the different width of the radial basis and different kernel methods.

Keywords: SVM, kernel function, SVM software package, kernel width.

1. Introduction to text classification

Text classification, this text can be news reports, web pages, email, academic papers and other, category is often about the contents of the text, such as politics, economy, sports and so on; there are also on the characteristics of the text, such as positive opinions, negative opinions; also can be determined on the basis of the application, such as spam and non spam. Text classification is initially information retrieval system , which was originally applied to the field of literature classification, library classification, document and patent classification, rule acquisition of text for indexing and classification. Automatic classification technology appear for pattern recognition, machine learning, data mining, mathematics, statistics, and other related fields provides the research with the help of the classification algorithm has been more and more widely used.

The process of text classification can be divided into two parts, training and classification. The purpose of the training is to construct a classification model, which is used to classify the samples and classes. Classification is based on the training results of the unknown sample classification, given the process of category identification. Text classification problem solving is a three step process: the first step, establish a text representation model, described in advance of data or set of concepts; the second step, through the analysis of the described by attributes sample (or instance, objects, etc.) to construct classification model; the third step, the unknown category text input classification model and obtains the sample categories. To build a classification model and the analysis of data tuple to form the training data set, this process is also called supervised learning. The task of automatic text classification is : in a given classification system, according to the content of the text automatically to the text relevance to the category.

From a mathematical point of view, text classification is a process of mapping, which will not mark the text of the category to be mapped to the existing category, which is expressed as a mathematical formula:

$$f : A \rightarrow B \tag{1-1}$$

In the formula, A represents a collection of classes to be classified, and B represents a collection of classes that have been identified in the classification system.

2. Text classification algorithm

The task of automatic text categorization is to automatically assign text to a predefined category. There are a lot of methods of text classification, which are commonly used in Naive Bayesian, neural network, K- algorithm, decision tree and support vector machine. Which support vector machine based on statistical learning theory as the basis, to avoid the problems in traditional classification algorithms sample infinite, has a good generalization performance, accuracy also showed obvious advantages. At present, it has been successfully used in the field of pattern recognition.

Support vector machine (SVM) is based on statistical learning theory of VC dimension theory and structural risk minimization principle of a kind of new machine learning method, it is on the basis of finite sample information, the model complexity and learning ability in seeking the best compromise, to expect to get a good generalization ability. Support vector machine algorithm aimed at finding a hyperplane h, the hyperplane can separate from the training data and with the category boundary along the vertical in the plane direction of the maximum distance, so SVM algorithm is also known as the maximum margin algorithm. Samples from most of the samples are not support vectors, reduce or remove these samples on the classification results, so that only the category boundary sample categories to decide classification results, with strong adaptation ability and higher accuracy. In addition, SVM algorithm from samples tends to infinity theory of constraints, the automatic classification of small samples with higher precision.

Here to define a kernel: kernel (Kernel) is a function K, to all $x, z \in X$,meet $K\left(x, z\right) = \left\langle \mathrm{F}\left(x\right), \mathrm{F}\left(z\right)\right\rangle$,here $x \rightarrow F$ is the mapping from the input space to a feature space.

When using kernel functions, the need to build a nonlinear learner is divided into two steps:

(1) uses a nonlinear mapping to transform data into a feature space.

(2) using linear classifiers in feature space.

And the use of kernel function can combine the two steps, directly in the original space of inner product operation, then rose to high dimensional space and avoid directly in the high dimensional space are calculated, and results with and without the use of kernel functions are completely equivalent.

With the said processing method of SVM is a kernel function selection, the key lies in the inner product of the vector space after, through data are mapped to a high dimensional space. To solve the problem in the original space linear inseparable.

SVM provides a problem independent of the dimension of the characterizations of the complexity function method. It has been introduced into high dimensional feature space, the input space of nonlinear decision boundary transformation into high-dimensional feature space a linear decision boundary, using dual kernel of kernel function to solve the numerical optimization of quadratic programming problem.

The common kernel functions are polynomial kernel, Gauss kernel and linear kernel,Depending on the classification problem, different kernel functions can be used.

3. SVM classification application

3.1 About the SVM package

In MATLAB programming to the text of the SVM classification, if it is a text classification, as already mentioned, then you need to construct text classification dictionary using the Java platform, custom segmentation rules, such as to the use of stop words such as, but within these contents and process in the designated range; if text data or directly coordinates are used, it is much easier. So in this paper directly gives the text data in the form of Cartesian coordinates or array and discuss the classification.

The program package for the SVM and kernel method toolbox, namely support vector machine and kernel function toolbox, and I in this paper, the simulation program will be used in this package. This package contains the file type for part directly open the mat auto file types, others for the M documents can be directly by MATLAB software to open, which is all optimal hyperplane (support vector machine) and the kernel function of individual program.

The basic support vector classification file is the file svc.m, in which there are building a kernel matrix part this part is to build a bridge between the kernel function and support vector is low dimensional data is mapped to high-dimensional data..Another part is the optimal hyper plane construction, everything here is initialization and variable settings for the hyperplane to facilitate for different data especially linear form is not very strong data for classification, and can better deal with classification problems.

SVM algorithm about data linearly separable, nonlinear and using the Gaussian kernel classification, the use of the SVM matlab itself comes with the program package, SVM using the built-in function:

$$svmStruct = svmtrain\left(Training, Group\right) \tag{3-1}$$

The training for training of input samples, which produce the training set; group as the sample by a linear or nonlinear classifier obtained after the classification, the output is a classifier svm Struct. In relation to the kernel function and kernel function and calculation method are optional, for example, but must appear in pairs, then classifier and testing sample substitution svm classify in that can obtain the classification results. This is the general principle of the whole package.

3.2 Linear classification

In the linear classification, this paper will take two groups: the control group a group of ten points, another group of 20 points, the purpose is to understand the complexity of text data of linear classifier, namely, the non use of the kernel function of the classifier in classification effects.

Linear classification procedures are as follows:

```
clear all; close all;clc;
sp=[3,7; 6,6; 4,6; 5,6.5; 4,7] % positive sample points
nsp=size(sp);
sn=[1,2; 3,5; 7,3; 5,4.2; 6,2.7] % negative sample points
nsn=size(sn)
sd=[sp;sn]
lsd=[true true true true true false false false false false]
Y = nominal(lsd)
figure(1);
subplot(1,2,1)
plot(sp(1:nsp,1),sp(1:nsp,2),'b+');
hold on
plot(sn(1:nsn,1),sn(1:nsn,2),'r*');
subplot(1,2,2)
svmStruct = svmtrain(sd,Y,'showplot',true);
```

The results of classification are shown in Figure 1. It can be seen that the results of the linear classifier are satisfactory when the sample is given. The point of the circle is the support vector. From figure 3.5, we can see that TP=10, FP=0, then the accuracy of the classification P=1. Of course, the amount of data is so small, the accuracy of the calculation is not much significance.

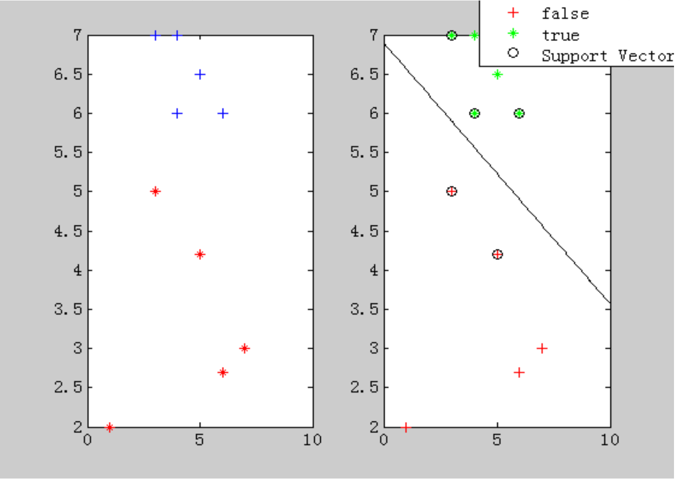

Figure 1.Linear classifier classification results for 10 training data points

The two program in the program based on the classification of points, or write up to ten:

sp=[3,7; 6,6; 4,6; 5,6.5; 4,7; 7,4; 4,3; 9,2; 5,7; 8,8]

sn=[1,2; 3,5; 7,3; 5,4.2; 6,2.7; 1,2; 3,2; 7,0; 5,2; 4,9]

Classification results are shown in Figure 2 below.

The same classifier can be seen from the above two graphs, that is, the linear classifier, for the simple and a little complicated classification of training data text. In the second classification results, TP=8, FP=1, then at this time of the classification accurate rate P = 0.89, accurate rate was decreased by 11%, so you can be sure of is the limitations of linear classifier is still very large, so to solve this problem, it is necessary to use nonlinear classifier.

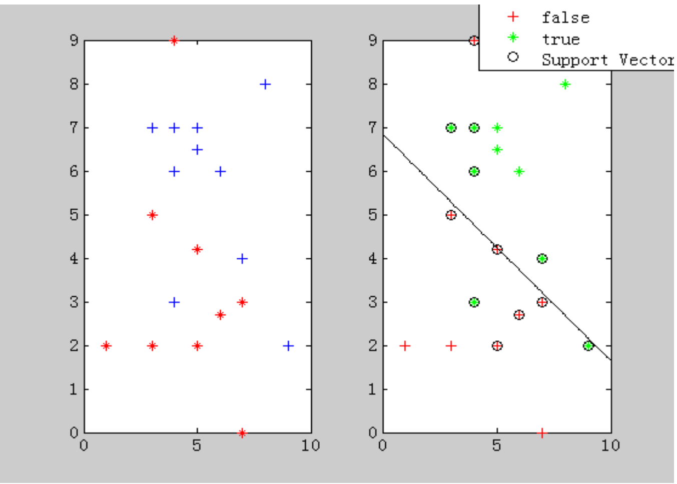

Figure 2.Linear classifier classification results for 20 training data points

3.3 Nonlinear classification

In the process of writing a nonlinear classifier, we still use the same 20 training data points, the same as the above second procedures, the purpose is to compare with the linear classifier. In order to highlight the advantages of the nonlinear classifier, I use the Gauss kernel function directly.

The nonlinear classification procedures are as follows:

```
clear all;close all;clc;
sp=[3,7; 6,6; 4,6; 5,6.5; 4,7; 7,4; 4,3; 9,2; 5,7; 8,8] % positive sample points
nsp=size(sp);
sn=[1,2; 3,5; 7,3; 5,4.2; 6,2.7; 1,2; 3,2; 7,0; 5,2; 4,9] % negative sample points
nsn=size(sn)
sd=[sp;sn]
lsd=[true true true true true true true true true true false false false false false false false false false false]
Y = nominal(lsd)
figure(1);
subplot(1,2,1)
plot(sp(1:nsp,1),sp(1:nsp,2),'b+');
hold on
plot(sn(1:nsn,1),sn(1:nsn,2),'r*');
subplot(1,2,2)
svmStruct = svmtrain(sd,Y,'Kernel_Function','rbf','rbf_sigma',0.6,'method','SMO','showplot',true);
% svmStruct = svmtrain(sd,Y,'Kernel_Function','quadratic','showplot',true);
% use the trained svm (svmStruct) to classify the data
sC=svmclassify(svmStruct,sd,'showplot',true)% sC is the classification result vector
```

Results of nonlinear classification are shown in Figure3. We can see with a kernel based nonlinear classifiers to deal with the complexity of the training text data is definitely better than a linear classifier, in the result, TP=9, FP=0, P = 1, the accurate rate instantly reaches the maximum value, that there is no false false alarm data, but not all statistically correct data, which involved the recall rate, and because of the FN=1. So to 0.9, the recall rate is not too high, 10% error can see for statistical data, we only pay attention to more is the classification of positive data. Although the recall rate is not very high, but still can be identified with kernel function especially like Gaussian kernel function such superiority of kernel function with stronger nonlinear classifier classification accuracy to better than linear classifier.

The following two conclusions can be drawn from the above experimental procedure:

(1)the classification of linear data is often easier to deal with the classification of nonlinear data;

(2)Linear classifier to deal with a range of far less than a wide range of nonlinear classifiers, linear classifier and nonlinear classifier essence is the use of kernel function or not, kernel function in nonlinear data processing has great advantages.

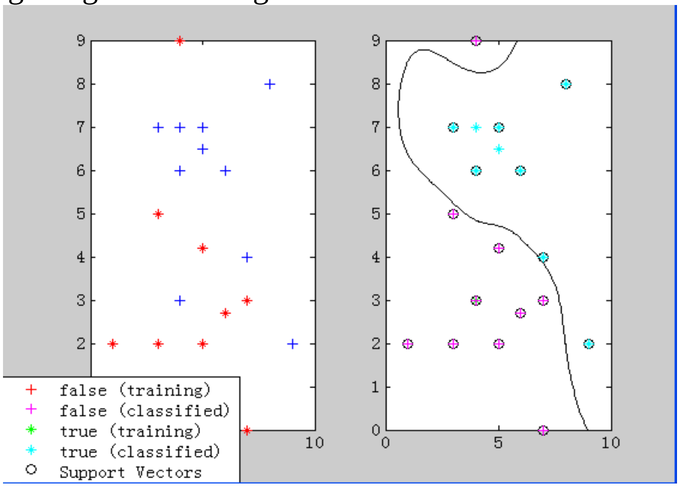

Figure 3.Classification results of 20 training data points with Gauss kernel function

4. Conclusion

In this paper, the data text is implemented based on SVM classification algorithm. And achieved the following results:

(1) The use of a linear classifier to classify a small amount of regular data;

(2) The with Gaussian kernel based nonlinear classifiers to classify the large number of irregular distribution of data, and the classification results were analyzed. The classification results show that: the error formula is calculated in accordance with high accuracy, and with the linear classifier were compared.

(3) An additional discussion is made on the kernel width and kernel function method, and it is found that the improved classification results are better.

In today's information age, with the increase of people's demand for automatic information processing, it is believed that the research on the SVM will quickly make the rapid progress and wide application.

Reference

[1]R. Ghani, S. Slattery and Yiming Yang.Hypertext Categorization using Hyperlink Patterns and Meta Data [A].The Eighteenth International Conference on Machine Learning [C]. 2001, 178-185.

[2]Church, K.W. Lisa, F.Rau. Commercial Applications of Natural Language Processing[J]. Communication of ACM, 1995, 38(11): 71-78.