Text classification using a new classification model: L1-LS-SVM

Wei Liwei[1, a], Wei chuanshen[2,b], Xiao Qiang[3,c] and Zhang Ying[1,d]

[1] China National Institute of Standardization, Beijing 100191, China

[2]SINOPEC Luoyang Corporation, Luoyang 471012, China

[3]Beijing International Electric Engineering Ltd.Cor., Beijing 100041, China

[a]weilw@cnis.gov.cn, [b]weichsh@126.com, [c]aguyforeve@126.com, [d]zhangy@cnis.gov.cn

Keywords: LS-SVM; SVM; L1-LS-SVM; Text classification

Abstract. With the advent of big-data age, it is essential to organize, analyze, retrieve and protect the useful data or sensitive information in a fast and efficient way for customers from different industries and fields. Least squares support vector machine (LS-SVM) has an outstanding advantage of lower computational complexity than that of standard support vector machines. Its shortcomings are the loss of sparseness and robustness. Thus it usually results in slow testing speed and poor generalization performance. In this paper, a least squares support vector machine with L1 penalty (L1-LS-SVM) is proposed to deal with above shortcomings. A minimum of 1-norm based object function is chosen to get the sparse and robust solution based on the idea of basis pursuit (BP) in the whole feasibility region. A real Chinese corpus from Fudan University is used to demonstrate the effectiveness of this model. The experimental results show that L1-LS-SVM can obtain a small number of support vectors and improve the generalization ability of LS-SVM.

Introduction

With the advent of big-data age, besides the regular content resources with usability and validity generated by diverse applications, a plenty of malicious messages and behaviors are simultaneously distributed on the Internet that might interfere regular service, violate privacy, misguide people and even harm the social stability. Furthermore, all of these data are flowing and expanding dramatically. From the perspective of data management, it is essential to organize, analyze, retrieve and protect the useful data or sensitive information in a fast and efficient way for customers from different industries and fields. The sensitive information and malicious messages or behaviors are respectively expected to be found out for protection and to be classified, filtered and analyzed for tracing the attackers, protecting victims as well as invoking the intelligent defense systems to process data, learn knowledge and update model. Among the machine learning methods, since cluster analysis (unsupervised learning) and classification (supervised learning) are able to be employed for detecting, tracing, organizing and analyzing either available information or behavior patterns, they are suggested to be the effective ways and crucial techniques for maximizing the efficiency of information security and protection.

As an important machine learning method based on statistical learning theory, not only did the support vector machine (SVM) [1,2] is able to learning from small-scale samples effectively, but also resolve such practical problems as non-linearity, high dimensionality and local minima, etc. Recently SVM have received a lot of attention in the machine learning community because of their remarkable generalization performance. The SVM typically follows from the solution to a quadratic programming. Despite its many advantages, one problem is that the size of the matrix of the quadratic programming is directly proportional to the number of training points. Thus this greatly increases the computational complexity [3,4], especially for the problems which deal with mass data or need on-line computation. Least squares support vector machine just makes up for that shortcoming.

Least squares support vector machine (LS-SVM) [5] is equivalent to solve a set of linear equations instead of a quadratic programming. Because the ε –insensitive loss function used in SVM is replaced by a sum square error loss function, the inequality restriction is replaced by the equation restriction. Thus this makes the least squares support vector machine achieve lower computational complexity.

But there are some potential drawbacks for LS-SVM [6]. The first drawback is that the usage of the sum square error may lead to less robust estimates. Reference [6] presents a weighted LS-SVM to solve this issue. This method needs an interactive procedure to get optimal cost function and robust estimation gradually. The second drawback is that the sparseness of the data points is lost. The pruning method [7] introduces a procedure that the training samples be selected from a data set, and these training samples will introduce the smallest approximation error that can be omitted. Another method [8] deletes some columns of the coefficient matrix through a certain measure. When the final model is used to represent the original system, the performance would be hurt.

Focusing on the above-mentioned questions, we propose a new method to improve the sparseness and robustness of the LS-SVM. In this method, a l_1 norm representation is used as the object function. And LS-SVM is used to characterize the system as a set of linear equations with deficient rank just like the overcomplete problem in independent component analysis (ICA) [9]. So the solution with the minimum l_1 norm is got based on the idea of basis pursuit (BP) in the whole feasibility region [10,11]. BP is closely connected with linear programming. So the proposed method is called least squares support vector machine with linear programming formulation (L1-LS-SVM). Above contents are introduced in chapter 2. Then the performance of this method is examined by three examples.

This paper is organized as follows. In section 2, we give the L1-LS-SVM classifier formulations and then set up the corresponding solutions. Numerical test results represent in Section 3 shows that our L1-LS-SVM is of good sparse and robustness performance. Section 4 concludes the paper and introduces some future research directions.

L1-LS-SVM model

Algorithm

Like least squares support vector machine, the object function for the L1-LS-SVM is defined as:

$$\min J(\vec{w}, e) = \frac{1}{2}\|w\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{n} e_i^2 \tag{1}$$

$$y_i(w^T(k)\varphi(x_{i,k}) + b) = 1 - e_i,\ i = 1, \cdots, n,\ k = 1, \cdots, m \tag{2}$$

where $x_{i,k}$ denotes the k^{th} component of the input vector X_i. It can be overcomplete dictionaries such as wavelet.

The quadratic optimization problem can be solved by transforming Eq. 2 into:

$$y_i\left(\sum_{j=1}^{n}\sum_{k=1}^{m}\alpha_{j,k}y_j k(x_{i,k}, x_{j,k}) + b\right) + \sum_{k=1}^{m}\alpha_{j,k}/\gamma = 1,\ i = 1, \cdots, n \tag{3}$$

where $k(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \bullet \phi(\vec{x}_j)$ is called the kernel function.

Eq. 3 and the transformation formula can be written as the following matrix form:

$$A_1 * \begin{bmatrix} b \\ \vec{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{I} \end{bmatrix} \tag{4}$$

Where

$$A_1 = \begin{bmatrix} 0 & \vec{y}^T & \cdots & \vec{y}^T \\ \vec{y} & K_1 & \cdots & K_m \end{bmatrix}_{(n+1)\times(mn+1)}$$

$\vec{1}^T = [1, \cdots, 1]_{1\times n}$, $\vec{\alpha} = [\alpha_{1,1}, \cdots, \alpha_{n,1}, \alpha_{1,2}, \cdots, \alpha_{nm}]$ and

$$K_d = \begin{bmatrix} y_1 y_1 k(x_{1,d}, x_{1,d}) + \frac{1}{\gamma} & \cdots & y_1 y_n k(x_{1,d}, x_{n,d}) \\ & \vdots & \\ y_n y_1 k(x_{n,d}, x_{1,d}) & \cdots & y_n y_n k(x_{n,d}, x_{n,d}) + \frac{1}{\gamma} \end{bmatrix}, d = 1, \cdots, m.$$

The following equation is the standard form of LS-SVM:

$$\begin{bmatrix} 0 & \vec{y}^T \\ \vec{y} & K + \frac{I}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \vec{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{1} \end{bmatrix} \quad (5)$$

Compared Eq. 4 with the standard form of LS-SVM in Eq. 5, we can find that the kernel mapping is executed in each component and the Lagrange multiplier $\alpha_{i,k}$ can be seen as the weight for each component and sample other than only for each sample in other methods.

Then the output is obtained:

$$f(x) = \operatorname{sgn}\left(\sum_{j=1}^{n} \sum_{k=1}^{m} y_i \alpha_{i,k} k(x_{i,k}, x_{j,k}) + b \right) \quad (6)$$

Above function is equivalent to the sum of the sub-function in different elements:

$$f(x) = \operatorname{sgn}\left(\sum_{k=1}^{m} f_k(x) + b \right) = \operatorname{sgn}\left(\sum_{k=1}^{m} \left(\sum_{i=1}^{n} y_i \alpha_{i,k} k(x_{i,k}, x_k) \right) + b \right) \quad (7)$$

Where $f_k(x)$ represents the contribution for the output by each element.

Finding solutions

From Eq. 4, we can find that the new LS-SVM is equivalent to solve a deficient rank linear equation set just like the overcomplete problem in ICA. Because the matrix A is $n \times nm$, there are infinite solutions to Eq. 4. It brings us a chance and challenge to get sparse solutions. There are many approaches presented to resolve this problem, including the method of Frames (MOF) and basis pursuit (BP)[10,11].

Unlike MOF, BP replaces the l^2 norm with the l^1 norm:

$$\min \left\| \vec{\beta} \right\|_1 \quad (8)$$

$$\text{Subject to } A * \vec{\beta} = c \quad (9)$$

where $\vec{\beta} = \begin{bmatrix} b \\ \vec{\alpha} \end{bmatrix}$, $\begin{cases} A = A_1, c = \begin{bmatrix} 0 \\ \vec{1} \end{bmatrix}, & for \ \ classifier \\ A = A_2, c = \begin{bmatrix} 0 \\ \vec{y} \end{bmatrix}, & for \ \ regression \end{cases}$.

It is a very important character that $e_i = \sum_{k=1}^{m} \alpha_{i,k} / \gamma$. Because b is a constant, the minimum of $\left\| \vec{\beta} \right\|_1$ is equivalent to that of $\left\| \vec{\alpha} \right\|_1$. And from equation (6), we can conclude that:

$$\left\|\vec{e}\right\|_1 = \sum_{i=1}^{n}\left|e_i\right| = \frac{\sum_{i=1}^{n}\left|\sum_{k=1}^{m}\alpha_{i,k}\right|}{\gamma} \le \frac{\left|\sum_{i=1}^{n}\sum_{k=1}^{m}\alpha_{i,k}\right|}{\gamma} = \frac{\left\|\vec{\alpha}\right\|_1}{\gamma} \tag{10}$$

So the minimum of $\left\|\vec{\alpha}\right\|_1$ can guarantee $\left\|\vec{e}\right\|_1$ in a lower level. And it improves the robustness for the final solution. Of course, we can use other optimization forms or algorithms according to the requirements of the problems. The flexibility is just the most advantagesfor this method. So the new LS-SVM method is called least squares support vector machine with linear programming formulation.

Experiment and discussion

In this section, we use the typical experimental data: Chinese corpus that is collected by Fudan University Dr. Li Ronglu[12], which are shown in Table 1 We will report the results of our empirical analysis on the above presented L1-LS-SVM algorithm. The corpus including the training set and testing set. There are 1882 documents in the training set. The test set contains 934 documents which have no classification label. And the test set is divided into 10 classes. The proportion of training set text number and test set text number is two-to-one.

Table 1. The experimental databases

Text Class	Number of experimental set	
	Training set	Test set
Art	166	82
Computer	134	66
Economy	217	108
Education	147	73
Environment	134	67
medicine	136	68
Policy	338	167
Sports	301	149
Transportation	143	71
Military	166	83

In automatic text classification system, will be used in the experiment data is usually divided into two parts: the training set and testing set. The so-called training set is composed of a set of have finished classification (namely has a given category label) text, which used for summed up the characteristics of each category in structure classifier. According to the classification system settings, each class should contain a certain amount of training textThe test set is the collection of documents that used to test the effect of the classification. Each one of these texts was through the classifier classification, and then the classification results contrast to the correct decision. Thus we can evaluate the effect of classifier. But the test set is not participated in constructing the classifier.

In addition, three evaluation criteria measure the efficiency of text classification:

$$recall_i = \frac{a}{a+c} \tag{11}$$

$$precision_i = \frac{a}{a+b} \tag{12}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{13}$$

Where a is the positive example test documentations that are correctly classified as belongs to the number of this kind. b is the negative example test documentations that are be error classified for

belong to the number of this kind. c is the positive example test documentations that are be error classified for does not belong to the number of this kind.

Firstly, the data is pre-processed by the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System)[12]. In this method the Gaussian kernel is used, and the kernel parameter needs to be chosen. Thus the method has two parameters to be prepared set: the kernel parameter σ^2 and the coefficient γ. The recall precision and F1 for each category text using SVM-MK approach are shown in Table 2.

Table 2 Experimental results using L1-LS-SVM

Text Class	Evaluation index		
	Precision	Recall	F1
Art	100	96.59	98.61
Computer	98.73	98.91	98.87
Economy	95.37	94.76	94.95
Education	98.03	93.92	95.79
Environment	100	94.35	95.86
medicine	98.03	96.16	96.95
Policy	93.02	97.83	95.16
Sports	96.35	98.76	96.69
Transportation	98.73	95.32	96.91
Military	91.32	89.26	90.17

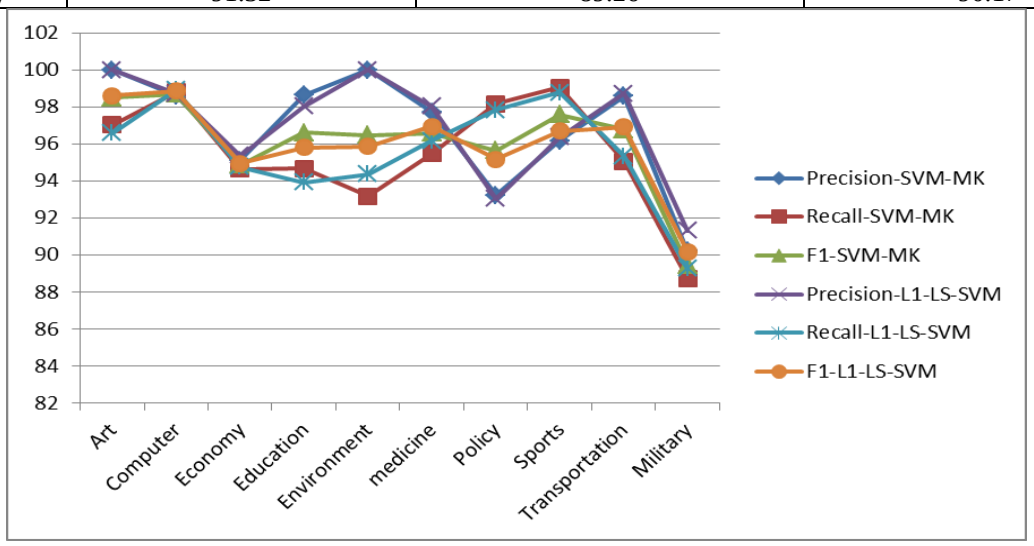

Fig.1 Comparison of results of improved SVM-MK and L1-LS-SVM

From Table 2 and Fig.1, we can conclude that the L1-LS-SVM model has better text classification capability in term of the recall, precision and the F1 in comparison with the improved SVM-MK models[13].This method is of better text classification performance in kinds of art, computer, economy, environment, sports, military and transportation. Compared to the traditional improved MK-SVM, L1-LS-SVM is not very good in classifying the kinds of education, policy. This may be because in removing relevant features of test results, and lost some information. So that the recall rate index is affected. This is also need to further improve. Consequently, the proposed L1-LS-SVM model can provide efficient alternatives in conducting text classification tasks.

Conclusions

This paper presents a novel L1-LS-SVM text classification model. By using the 1-norm, the L1-LS-SVM is equivalent to get the minimum of a sum absolute error in the feasibility region. So this method can improve the robustness and get the sparseness for the solution simultaneously. Another advantage is that it is equivalent to solve a linear programming and do not increase the computational burden that much. In addition, the output of the L1-LS-SVM can be viewed as a weighted sum for different components. This makes the output more understandable. And it provides efficient alternatives in conducting text classification tasks. Furthermore, empirical results show that the

L1-LS-SVM is very efficient in text classification. Generalizing the rules by the features that have been selected is another further work.

Acknowledgements

This work has been supported by the Basal Research Funds for central public research institutes (#222015Y-4006; #222016Y-4505; #222016Y-4504), and also supported by the science and technology support program(2014BAK07B00).

References

[1]V.Vapnik, *The nature of statistic learning theory*, Springer, New York(1995).

[2] V.Vapnik, *Statistic Learning Theory*, Willey, New York(1998).

[3]T.Gartner, P. A.Flach, WBCSVM: *Weighted Bayesian Classification based on support vector machine, 18th Int. Conf. on Machine Learning*, Willianstown, Carla E. Brodley, Andrea Pohoreckyj Danyluk, (eds.) (2001),p. 207–209.

[4] F.Jiang, *Research on Chinese Text Categorization based on Support Vector Machine*, Degree of Master paper, Chongqing University(2009).

[5] J.A.K.Suykens, Wandewalle J.: *Least squares support vector machine classifiers*, Neural Processing Letters, Vol. 9 (1999), p.293-300.

[6] J.A.K.Suykens, J.D.Branbater, L.Lukas, J.Wandewalle: *Weighted least squares support vector machine: robustness and sparseness approximation*, Neurocomputing, Vol. 48 (2002), p.85-105.

[7] Y.G.Li, C.Lin, W.D.Zhang: *Improved sparse least-squares support vector machine classifiers*, Neurocomputing, 2006 (69), p.1655-1658

[8] V.Jozsef, H.Gabor: *A sparse least squares support vector machine classifiers*, IEEE international conference on neural network(2004), p.543-548.

[9] A.Hyvarinen, J.Karhunen, E.Oja: I*ndependent Component Analysis*. Willey, New York(2001).

[10] S.S.Chen, D.L.Donoho, M.A.Sauders: *Atomic decomposition by basis pursuit*, SIAM review, Vol. 43 (2001), p.129-159.

[11] P.Georqiev, A.Cichocki: *Sparse component analysis of overcomplete mixture by improved basis pursuit method*, IEEE international symposium on Circuits and System(2004), p.37-40.

[12]Information on http://www.ics.uci. edu/~mlearn/ML Repositor y.html

[13] L.W.Wei, B.Wei, B.Wang: *Text Classification Using Support Vector Machine with Mixture of Kernel*, A Journal of Software Engineering and Applications, Vol. 5 (2012), p.231-236.