LSSVM-based social spam detection model

Xiaolei Yang and Yidan Su and JinPing Mo
Guangxi University of Finance and Economics,Nanning,China
Guangxi University, Nanning,China
Guangxi University of Finance and Economics,Nanning,China
314569693@qq.com,qinhua@gxu.edu.cn, 17782974@qq.com

Keywords: social spam; social bookmark system; lssvm ; detection model

Abstract. To Resolve the garbage tag issue in Folksonomy, Lssvm algorithm for social spa m detection model (least Squares support vector machine classifiers) was proposed. The met hod of inequality change the constraints in the traditional support vector machine into equa lity constraints, and take the empirical function of the squared error loss function as the E xperience function in training set. so that the quadratic programming problem convert QP i nto solving linear equations, it was improving solution the speed of solution and accuracy o f convergence.The experimental results show that we have got higher classification accuracy and less predict time than traditional svm detection methods based on least squares support vector machine algorithm garbage tag detection model.

Introduction

The scholars in domestic have attempted to use SVM[1] for garbage tag detection.when we were using svm for solving the QP(quadratic programming)problem, the variable dimen sion equal to the number of training samples so the number of matrix elements is the Squa re of the number of training samples. When the data size is reacheing a certain level, usi ng the SVM algorithm to solve is very hard because it is too complicated.

With the the structure of the optimization problem, Lssvm[4] algorithm(leastSquares supp ort vector machine classifiers) respectively,have used the error factor and quadratic terms,fo r the objective function, while the inequality constraints and equality constraints.have being adopted for the constraints of the form . In LS-SVM method, the objective function of op timization problem using the error Squared term, as well as equality constraints,puting QP i nto a set of linear equations solving, making the Lagrange multiplier and the error term is proportional to direct consequence of making the final decision function.as result of it, all s amples were related to the final objective function.

The experiments show that the lslsvm algorithm on garbage tag detection model for hand ling large-scale high-dimensional data sets on training time and classification accuracy has i mproved significantly, and in the small training sample also reflects the good trans-resistanc e.

Lssvm algorithm classification model

Suppose the training sample set $T=\{(x_k, y_k) \mid k=1,2,3,\cdots,n\}$, $x_k \in R^n, y_k \in R$, x_k is nput data and y_k is output data. Optimization problem can be described as in the original space (w space):

$$\min \quad C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) + \frac{1}{2} w_i^2 \qquad (1)$$

Subject to

Published by Atlantis Press, Paris, France.

$$\begin{cases} f(x) = (w, x) + b \\ y_i - f(x_i) - e \le \zeta_i \\ f(x_i) - y_i - e \le \zeta_i^* \\ \zeta_i^* \ge 0 \quad \zeta_i^* \ge 0 \end{cases} \tag{2}$$

We use the error sum of QPuares care instead of slack variables C, and replace the ineq uality constraints by equality constraints,we get the LSSVM regression optimization problem:

$$\min_{w,b,e} J(w,e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{k=1}^{N} e_k^2 \tag{3}$$

Constraints: $y_k = w^T \varphi(x_k) + b + e_k, k = 1, 2, \cdots \mathrm{N}$,and $\varphi(\,) : \mathrm{R}^{\mathrm{n}} \to R^m$ is Nuclear space mapping function.

Weight vector: $w \in R^m$, The weight vector error vector: $e_k \in R$,b is the bias vector, loss function J is the amount of error and rules , γ is Adjustable function, The purpose of the nuclear space of the mapping function is to extract the characteristics of a sample in the original space,and it is mapped to a vector in high dimensional space to solve the problem of the original space linear inseparable from the original space.. According to the function (1) we construct the Lagrange function [6]:

$$L(w,b,e,\alpha) = J(w,e) - \sum_{k=1}^{N} \alpha_k \{ w^T \varphi(x_k) + b + e_k - y_k \}. \tag{4}$$

Using the Lagrange multiplier $\alpha_k \in R$, to optimize the (2)

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0, w = \sum_{k=1}^{N} \alpha_k \varphi(x_k); \\ \dfrac{\partial L}{\partial b} = 0, \sum_{k=1}^{N} \alpha_k = 0; \\ \dfrac{\partial L}{\partial e_k} = 0, \alpha_k = \gamma e_k; \\ \dfrac{\partial L}{\partial \alpha_k} = 0, w^T \varphi(x_k) + b + e_k - y_k = 0 \end{cases} \tag{5}$$

k=1,2,3,…,N.

Matrix equation is:

$$\begin{pmatrix} 0 & 1_v^T \\ 1_v & \Omega + \dfrac{1}{\gamma} I \end{pmatrix} \begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix}$$

$y = (y_1, y_2, \cdots, y_n)$; $1_v = (1, 2, \cdots, 1)$, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$; $\Omega_{kl} = \varphi(x_k)^T \varphi(x_l)$

$k, l = 1, 2, 3 \cdots, N$; (6)

According to Mercer conditions, there is a mapping function φ and nuclear function $K(\bullet, \bullet)$,(10) is advaning:

$$K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l) \tag{7}$$

LS-SVM least QPuares support vector machine's functions, is estimated to be:

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \tag{8}$$

The b is solved by (4),we are using the Gauss RBF kernel :

Published by Atlantis Press, Paris, France.

$$K(x_k, x_l) = \exp\left(-\frac{\| x_k - x_l \|^2}{2\sigma^2}\right) \quad (9)$$

The relations between the formal definition of Folksonomy in [7] as follows:

Definition 1:definition of Folksonomy relationship is a tuple,F:= (U,T,R,Y, <), U, T, R is finite, their elements are called users, tags, and resources, Y is a ternary relationship, Th at is Y U × T × R, Its elements are called tag assignments (TAS), Refers to a superset of the relationship between a user-specific label a subset of / tags,that is < U × T × R。

Definition 2: For a given user, u∈U,Pu is the constraints of F on u,that is Pu: = (Tu, Ru,Iu, < u), and Iu: = { (t,r) ∈T × R | (u,t ,r) ∈Y} ,Tu: = π1(Iu) ,Ru: = π2(Iu) , < u: ={ (t1,t2) ∈T × T |(u,t1,t2) ∈ ＜ } 。π is the projector, πi projection of the i.

Tu --- posted over the label set by user u

Ru --- provided a collection of resources.by user u

Iu-- the collection between resources and user u tags

＜ u-User u labels a subset of / tags superset relationship

Characteristics of the user model in Garbage label detection

Algorithm is described as follows:

Input: the original user data set X, the original test set Y, the classification table of the test set, the kernel function K (ui, uj) and penalty parameter C. Output: Classification and Prediction results set R, the detection accuracy P

a) the original data set X and the original test set Y must be pro-precessed; the invalid or incomplete data should be deleted;

b)For the X and Y, the user signs over the labels were combined with the corresponding resources,Respectively, to generate the text set X 'and Y'

c) the text set X 'and Y' in the data should be splited one by one entry segmentation, t hen to build dictionary D-the formation of the vector framework;

d) (6) is used to convert the in data the X 'and Y' into vector form, respectively, to gen erate a user feature vector corresponding to the training set F, the user feature vector test s et F ':

e) To put the kernel function K (ui, uj) and the penalty parameter C into the algorithm then we get the decision function f (x).

f) Put The characteristics of each user in the F 'vector generation into the f (x),then calc ulate the classification symbol, If f (x) is equal to 0 to determine if the user put the peopl e as spam label; if equal to a user as normal user. Class standard calculation of the result set for the classification forecasting results set R.

g) we compare the test set classification table data with R,then calculate the detection ac curacy

Experiments

MATLAB2009a is used for experiment,The experimental hardware environment is: CPU P4, 3.0GHz, 1GB of memory. All experiments run for 15 times to take the average. this paper, the data set from the binary classification test data set: synth、bc、sonarall、haberman 、gisette_scale. the synth sonarall and haberman is the real data sets from the data set used LibSVM [8]. gisette_scale data set is provided from from the set the BibSonomy PKDD2009. The data set is collected from well-known social bookmarking sites BibSonomy. The site is a system based on Folksonomy framework, the source data taken in this paper consists of two data files (tas, bookmark), tas file contains the relationship between the records of the users, tas_id, label and corresponding bookmark_id bookmark file contains resources, resource description, bookmark_idand the corresponding tas_id relationship records. Two data files, then to pick by tas_id and bookmark_id.

Published by Atlantis Press, Paris, France.

The experimental program is divided into two groups, first group is relatively small training set, respectively, using svm and lssvm of algorithms to classify, while the second group classification when the training set is relatively large.

Table1. The first group:

Data set	the number of training set	the number of samples	dimensions
synth	180	1000	2
bc	50	256	10
sonarall	80	208	60
haberman	50	306	3
gisette_scale	300	500	4972

Table 2 .The second group:

Data set	the number of training set	the number of samples	dimensions
synth	250	1000	2
bc	100	256	10
sonarall	130	208	60
haberman	100	306	3
gisette_scale	500	500	4972

Experimental results and analysis

In order to verify the effect of the detection model,we compare the prediction time and classification accuracy rate using lssvm with SVM Algorithm. The first set of training data set for small-scale samples, the second group of relatively large-scale sample training set data, derived from data for 10 experiments averaged contrast to the results shown in Table 3:

Table 3 .Predict the time and classification accuracy comparison(first group)

First groop	Training time (seconds)		Accuracy (%)	
dataset	SVM	LSSVM	SVM	LSSVM
synth	0.018	0.0531	85.3	85.70
bc	0.008	0.010	54.29	60.54
sonarall	0.0165	0.0136	50.96	65.38
haberman	0.008	0.006	35.94	45.09
gisette_scale	4.515	0.410	47.50	65.50

Table 4. Predict the time and classification accuracy comparison(second group)

Second group	SVM	LSSVM	SVM	LSSVM
dateset	Training time (seconds)		Accuracy (%)	
synth	0.0385	0.055	89.7	90.50
bc	0.017	0.014	73.04	73.04
sonarall	0.028	0.013	46.63	63.94
haberman	0.012	0.013	40.19	44.11
gisette_scale	9.471	0.9723	47.00	65.51

It can be seen from Table 3, two-dimensional data set of synth and three-dimensional da ta sets haberman in the training set of 180 and 50 under the conditions of testing, using sv m algorithm training time is 0.018s and 0.008s, while the use of lssvm algorithm training t ime is 0.0531s and 0.006s, the training accuracy using svm algorithm accuracy of 85.3% an d 35.94% respectively, while the use of lssvm algorithm accuracy of 85.70% and 45.09%, f rom here we can see that when the data dimension is relatively low, lssvm algorithm comp ared to the svm algorithm in predicting the time and classification accuracy only a minor u pgrade or flat.

When the dimension increases, we use the 10-dimensional data sets bc and 60-dimension al data sets sonarall to do test. training time was 0.008s and 0.0165s with lssvm algorithm, while the use of svm algorithm training time was 0.010s and 0.0136s; using the sVM algor ithm,the training precision was 54.29% and 50.96%,,whlie using lssvm algorithm was 60.5 4% and 65.38%. Thus, when the dimension increases, the training time was shortened, espe cially the accuracy significantly increased.

Published by Atlantis Press, Paris, France.

We compare svm algorithm whth lssvm of algorithms under the High-dimensional label data set gisette_scale. The training time is 4.515s using svm algorithm,while training time is 4.515s with lssvm. The lssvm training time is close to the less10 times than svm. Svm al gorithm in terms of training accuracy, the svm is 47.50% ,while lssvm is 60.50% . It can be seen that the training accuracy when using algorithm of lssvm was 10 percent above sv m algorithm . we can see that when the dimension of data is higher, lssvm has improved significantly especially in predicting the time.These conclusions in Table 4, the same expres sion, not repeat them here.

In addition, when the training set size also affects both svm algorithm and lssvm of algo rithm performance, we compare in Table 3 and Table 4, when the increase in the number of samples, using svm algorithm for low-dimensional data set the synth, bc, haberman, and accuracy increased by 4.418.75 and 4.25 percentage points, while the lssvm algorithm 4.8, 12.5 and 0.98 percentage points, reflected in high-dimensional data with high dimensional d ata sets sonarall,.and in gisette_scale test,The float between two kinds of data sets was 4.33 percent and 0.98 percent using svm whlie the float is 0.5 percent and 0.1 percent using ls svm. Therefore, the algorithm lssvm is more stable than svm algorithm.It can also be seen that when the training set is smaller, lssvm algorithms has better performance compare with svm algorithms.

Summary

we were using LSVSM to build the user model ,and then divided the users of the sites into two classes by LSSVM,of which one was the normal,the other was spammer.The meth od of inequality change the constraints in the traditional support vector machine into equal ity constraints, and take the empirical function of the Squared error loss function as the E xperience function in training set. so that the quadratic programming problem convert QP in to solving linear equations, it was improving solution the speed of solution and accuracy o f convergence.The experimental results show that we have got higher classification accuracy and less predict time than traditional svm detection methods based on least QPuares suppo rt vector machine algorithm garbage tag detection model..So cut off the social spam by red ucing the spammer.The result of the experiment shows that the classification accuracy of L SSVM-based social spam detection model is higher than Traditional svm.

Fund: the National Natural Science Foundation (project number: 61063032); the Ministry of Education Humanities and social science research project (project number: 11YJAZH080)

Acnowledgment

This research has been partly supported by nation natural science fund project(61063032), ministry of education social science research project (11YJAZH080).

References

[1] KIM CJ,HWANG KB.Naive Bayes classier learning with featureselection for spam dete ction in social bookmarking [c]// Lecture Notes in Computer Science.Berlin: Springer-Verlag, 2008.

[2] Tan Xi, Xia Ningxia, Suyi Dan, based on support vector machine garbage tag detection model [J]. Application Research of Computers, 2010,27(10): 40:46

[3] GRAMME P,CHEVALIER J F.Rank for spam dsetection[c]//Lecture Notes in Computer Science.Berlin: Springer-Verlag,2008.

[4] Van Gestel, T., Suykens, J.A.K, Baesens, B, Viaene, S, Vanthienen, J.,Dedene, G., De Moor, B. VandewalleJ. Benchmarking least QPuares support vector machine classifiers", Ma ch. Learning, vol 54, pp.5-32, 2003

[5] ADKOUR A,HEFNI T,HEFNY A,et al.Using semantic featuresto detect spamming in so cial bookmarking systems[c]// LectureNotes in Computer Science.Berlin: Springer-Verlag,200 8.

[6] HOTHO A,JASCHKE R,SCHMITZ C,et al.Emergent semantics in BibSonomy[M],Lisko wsky: GI Jahrestagung,2006: 305-312.

[7] SALTON G,McGILL M J.Introduction to modern information retrieval [m].New York:M cGraw-Hill,1983:1-12.

[8] http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/

[9] BROADLY.Social spam definition[EB/OL](2008-7-21) .http://www.bryanchen.com/2008/0 7/21/social-spam/.

[10] Kuh, A.,De Wilde P. "Commentson pruning error minimization in least QPuares suppor t vector machines".IEEE Trans". Neural Networks, vol 18(2). 2007.

Published by Atlantis Press, Paris, France.