P-Order Twin Support Vector Machine

Xu Ma, Qiaolin Ye＊

College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China

＊Corresponding author: yqlcom@njfu.edu.cn

# ABSTRACT

Twin Support Vector Machine(TWSVM) [1] is an effective classifier especially for binary data, which is defined by squared distance in the objective. It is well known that distance is susceptible to outliers, which can lead to errors. It is desirable to develop a revised TWSVM. In this paper, a new robust twin support vector machine via p-order optimized algorithm was proposed. We improved the TWSVM algorithm by iterative method. Theoretical support shows that iterative method is effective in the solution to improve TWSVM via p-order instead of distances. A large number of experiments show that p-order Twin Support Vector machine (pTWSVM) can process the noise data effectively and has a better accuracy.

1. Introduction

Support vector machine [2] has been a vital method for pattern classification in the last decade. The standard Support vector machine devotes to get an optimal separating hyper plane that has the max margin between the two data sets to reduce generalization error. An advantage of SVM is that it can regulate the trade-off between structural complexity and empirical risk.

In 2001, G.Fungand and O.L.Mangasarian proposed a algorithm termed as PSVM [3] that two parallel planes are pushed apart as far as possible to classify points [3] . Instead of solving a quadratic or a linear program，PSVM only need to solve a single system of linear equations. The formulation of PSVM make the solution of SVM comes to fast and effective.Not only does it hold the advantage of high speed, PSVM also has comparable accuracy to traditional SVM classifiers.

In 2006, O.L. Mangasarian and E.W.Wild proposed a nonparallel plane classifier via generalized eigenvalue [4]. The novel approach to classification problems cuts down the requirement that the bounding generated by SVMs be parallel in the input space. The nonparallel proximal planes can be easily obtained by solving the classical generalized eigenvalue problem.

Different from PSVM and GEPSVM, a new nonparallel plane classifier termed as the Twin Support Vector Machine (TWSVM) was proposed by Jayadeva in 2007. TWSVM obtains nonparallel planes around which the data points of the corresponding class get clustered. It solves a pair of quadratic programming problems. Each of the two quadratic programming problems has the formulation of a typical SVM, except that not all data points are used in the constraints of either problem at the same time.

In order to deal with noise data problem effectively, a new Robust Twin Support Vector Machine named as R-TWSVM was proposed by Zhiquan Qi etc. R-TWSVM classifies via second order cone programming formulations. Moreover, only inner products about inputs in the dual problems, this leads to an advantage that kernel trick can be applied directly in the for nonlinear cases and also avoids solving extra inverse matrices.

Aiming at the defect that TWSVM cannot solve the imbalanced data, Yuan-Hai Shao proposed an efficient weighted Lagrangian twin support vector machine(WLTSVM). Two advantages are introduced in the paper. One is using graph to reserve the proximity information, and the other is weighting biases in the Lagrangian TWSVM formulations to relief the problem of imbalanced data classification.

In this paper, we are absorbed in the problem of robust TWSVM on data set with outlier data samples. In classical TWSVM, it is willing to minimize the distance with the squared distance. As we know, squared outliers distance will expand the error distance of samples. From this point，we hold the notion that distance with a lower orders can emphasis the percentage of normal points distance. A p-order is used for the improvement of TWSVM that p ought to be lower than 2.

The p-order twin support vector machine (pTWSVM) method is focus on the following problems:

1. The modification of the TWSVM objective with p-order distance. This point deals to change the squared order distance into order distance. The step not only modifies the objective function but also the inequality constraints.

2)The formulation of proposed algorithm. To solve the problem we have proposed, an effective iterative algorithm has been implemented in this paper. With this algorithm, only a few iterations are needed and we will get the optimal result.

3) The proof of the algorithm convergence. The convergence of the iterative algorithm has rigorous theoretical inference with several lemmas.

4)The nonlinear kernel classifier. In order to extend the formulation to nonlinear classifiers, we also consider the kernel tricks in the problem and solve the kernel surfaces.

The paper is organized as follows: We introduces related work in the section 2. Section 3 dwells on our theoretical work for the new method in detail, including the improvement and related proof. Section 4 is about the extension on nonlinear kernel. Section 5 deals with the experiment and Section 6 summarize this paper.

# 2.Related Work

This section is mainly about the definition of some vectors. In the paper, the vectors are all column vectors. A row vector will be defined by transposing a column vector via a prime superscript T. We suppose A represents the matrix of positive classes and B represents the matrix of negative ones. Let the number of points in positive and negative classes be given by m1 and m2, vectors are all in the real space . Therefore, the size of A and B are m1\*n and m2\*n respectively and the number of all patterns will be m, m=m1+m2. For a matrix A, Ai is the ith row of A which is a row vector in ,the of Ai will be denoted by . In addition, e1 and e2 are vectors of ones of appropriate dimension for A and B. I denotes the identity matrix of arbitrary dimension.

## 2.1GEPSVM

The proximal support vector machine via generalized eigenvalues(GEPSVM) is a great classifier for binary classification problem which the distance of patterns is measured by . The goal of GEPSVM classifier is to obtain two nonparallel planes:

and . (1)

Let’s take the plane as an example. In order minimize the Euclidean distance of the plane from positive classes A and maximize the Euclidean distance of the plane from negative classes B, we have to solve the following optimization problem:

(2)

This formulation can be converted to Rayleigh Quotient form as follows:

, (3)

where G and H are symmetric matrices in real space and z is the classification plane. The G, H and z are donated as:

,

, (4)

.

The solution of (3) can be obtained by solving the generalized eigenvalue problem via the properties of Rayleigh Quotient [5, 10] [GEPSVM, The Symmetric Eigenvalue problem].

. (5)

It’s easy to get the minimum (2) when z is the value of the eigenvector corresponding to the smallest eigenvalue . Therefore, we can achieve the plane which is close to patterns of positive class and far away from patterns of negative class. And vice versa, by the same method we can get another value.

## 2.2TWSVM

Suppose we have data points of n-dimensional belongs to two classes represented by matrices A and B respectively. The TWSVM devotes to obtaining two nonparallel hyper planes which each plan is as close as possible to one type points and as far as possible to the rest.

The TWSVM can be obtained by solving the following pairs of quadratic programming problems:

(6)

(7)

where are parameters. The two nonparallel planes can be obtained by :

（8）

We can classify the point X by comparing the geometrical margin to the two planes respectively.

3.P-Order Twin Support Vector Machine

3.1 Optimization Algorithm to the Proposed Method

Form the TWSVM, it clearly shows that the squared distance in the formulas. It may be not satisfied the for the problem. The result we obtained could be affected by the outliers pronouncedly. That is, p-order is a good method for instead of squared distance. If we can find an appropriate p, the algorithm can emphasize norm data and overlook outliers best. Now, we can find that what the p-order value is to obtain a balance between the norm data and the outliers. Assuming squared distance is a benchmark, if ，the distance of data will be shortened and the outlier data samples will be alleviated. The paper holds the notion that the percentage of outliers decides the p value.

The improvement of TWSVM can be obtained by solving the following problem:

(9)

（10）

The Lagrange function of the problem is:

(11)

where are the vectors of Lagrange multipliers.

To solve the problem ,a good approach is splitting the distance to squared and (p-2)-th order :

(12)

Denote

(13)

the Lagrange function can be written as:

(14)

The derivative on every parameter, i.e., the Karush-Kuhn-Tucker(K.K.T) necessary and sufficient optimality conditions for the problem is:

(15)

(16)

(17)

(18)

(19)

(20)

(21)

(22)

Form ,, we have

(23)

We define

(24)

Notice that can be signified as:

(25)

Combining (15) and (16) leads to

(26)

this can be written as:

(27)

i.e.:

(28)

Although is always positive semidefinite , it is possible that it may not be well conditioned in some situations. So the problem can be regularized by introducing a regularization term as follows:

(29)

where and is an identity matrix of appropriate dimensions.

Using the Lagrange function and the K.K.T. conditions above, we obtain the Wolfe dual of p-th order TWSVM as follows:

(30)

Similarly, the other one’s dual is:

(31)

We can obtain the optimal via an iterative algorithm. In each iteration, are calculated with the current calculated . The iteration produce is repeat until converges. The iteration is started with a initialized . the are re-changed adaptively during each iteration.

The algorithm to solve the problem:

|  |
| --- |
| Input : Training data , parameter .  Give out .  Initialize .  While converge do   1. Calculate ; 2. Calculate via dual function; 3. Update , add regularization term if necessary;   End  Output . |

The another one is similarly like the process above.

3.2 Convergence Analysis

To prove the convergence of the new algorithm, we need the following lemmas that proved by Hua Wang in his paper:

Lemma: For any nonzero vectors A, B, when , the following inequality holds:

(32)

In order to use this lemma in this paper, we can transform it into the following form:

(33)

Considering the objective function, suppose , and the updated is B. According the objective function, we know that

(34)

i.e.

(35)

Connect the lemma, it can be obtained that:

(36)

(37)

(38)

Combing the (37) and (38), we arrive at

(39)

Note that A is , and B is updated . Inequality (39) indicates that will converge after each iteration. According , so the is convergent in each iteration. is a positive number, this leads to it will converge to an optimal value via an iterative approach. The convergence of also means that ’s convergence.

4.The Nonlinear Kernel Classifier

In order to extend our new method to nonlinear classifiers, we have modified the new algorithm by using the kernel method.

As we know, kernel-generated surfaces for TWSVM

,and (40)

where

and K is an appropriately chosen kernel. Note that if the K is a linear kernel like , it will degenerate into an ordinary plane.

We construct an optimization problem KPTWSVM as follows:

(41)

where is a parameter. Next, we define a Lagrange function L by the above formula:

(42)

To solve the problem, we split the distance into two parts:

(43)

In this formula, can be represented by . The Lagrange function is updated as follows:

(44)

We obtain the K.K.T. conditions for KPTWSVM as follows:

(45)

(46)

(47)

(48)

(49)

(50)

(51)

(52)

Combing (45) and (46), we obtain

(53)

Let

, (54)

and the augmented vector . Then the formula can be solved as:

(55)

i.e.:

(56)

The Wolfe dual of KPTWSVM is given by

(57)

In a similar manner, the another KPTWSVM kernel-generated surface can be obtained by solving a new dual function.

Once the two KPTWSVM problems are solved to obtain the surfaces, a new data can be classified in a manner similar to the linear case.

In the actual experiments, if the number of patterns is large, then the rectangular kernel technique can be used to reduce the dimensionality of KPTWSVM. In the linear case, a regularization term always be useful.

5.Experimental Results

5.1 Binary data

In order to directly compare the differences between TWSVM and pTWSVM, we did experiments on an artificial data set. A simple data set was constructed, with ten points distributed over and respectively. For class 1, the points are (0,10), (1,9), (2,8), (3,7) etc. For class 2, the pints are (0,0), (1,1), (2,2), (3,3) and so on. The two classes of points are strictly binary data. In a two-dimensional Cartesian coordinate system, there should be two lines perpendicular to each other. Although PTWSVM is committed to improving the robustness of TWSVM, it should has the same accuracy as TWSVM in the case of no outliers. Moreover, since there is no noise, the algorithm should only need to iterate once to achieve the final convergence results.



(TWSVM) (pTWSVM)

fig1 binary data experiments pictures

The above diagram shows that the two algorithms have good classification effect on binary data sets and the classification surfaces are almost the same. This result is in line with our expectations conjecture.

Next we add a little noise to the data set to test the robustness of the two algorithms.



（TWSVM） （PTWSVM） （TWO）

fig2 binary data with outliers experiments pictures

Form the picture1 and picture2 we can find that the classification surfaces are similar in terms of structure. Picture3 shows that pTWSVM provides a better classification. This proves that pTWSVM is much less susceptible to noise than TWSVM and has good robustness.

5.2 Study the p value of the new proposed method

The new method leads to a problem that what the value of p is. Considering the objective function, we hold the notion that the p’s value is under the influence of outliers. In order to get a higher accuracy, the greater proportion of noise, p value is smaller，and vice versa. Formula 11？perspicuously indicate that p value directly affect the result of the formula. Splitting the formula into two parts: the outliers functional margin and the normal data’s functional margin. The role of p value is to emphasize the proportion of the two.

We experiment with several benchmark data set as examples. Considering the objective function, we hold the notion that the parameter p value can directly affect experiment accuracy. Our goal is to obtain the best parameter value to get the best classification accuracy. We vary p of the proposed objective in the range of 0.1 to 2 to study its impacts to the classification performance. Through the experimental data, we simulate the corresponding correct rate curve.



fig3 accuracy with different p value

Pictures show that the accuracy of pTWSVM on different data sets will increase with increasing p value. When p is about 1.3, the accuracy reaches the maximum. When , the accuracy will be slightly decreased. This can be attributed that, when p is big, the distance measurement will be compromised. This conclusion implies p should not be too big to invalidate the distance measurement in the Euclidean space. Upon the result in fig3, empirically, we set in all our subsequent experiments.

## 5.3 Comparison of accuracy

In this section, we compare the pTWSVM algorithm with other algorithm by experiments. We evaluate the proposed method on several widely used benchmark datasets in machine learning studies. The descriptions of the datasets are given in table 1.

Table1: Data sets descriptions

|  |  |  |
| --- | --- | --- |
| Data sets | Number | Dimension |
| Heart | 270 | 13 |
| Australian | 690 | 14 |
| Pima | 768 | 8 |
| Sonar | 208 | 60 |
| Spect | 267 | 44 |
| germ | 1000 | 24 |
| Monk1 | 561 | 6 |
| cancer | 683 | 9 |
| Ionodata | 351 | 34 |
| splice | 1519 | 60 |
| cmc | 642 | 9 |
| blood | 748 | 4 |
| haberman | 306 | 3 |
| liver | 327 | 6 |
| diabetes | 768 | 8 |

Table 2:Test Set Accuracy with a Linear Kernel

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SVM | PSVM | GEPSVM | TWSVM | PTWSVM |
| Heart | 0.8444 | 0.8407 | 0.8593 | 0.8519 | 0.8741 |
| Australian | 0.8551 | 0.8580 | 0.7362 | 0.8041 | 0.8653 |
| pima | 0.7969 | 0.7707 | 0.6771 | 0.7073 | 0.8013 |
| sonar | 0.7117 | 0.7840 | 0.7019 | 0.8068 | 0.9038 |
| spect | 0.7000 | 0.7671 | 0.6150 | 0.6578 | 0.8289 |
| germ | 0.7200 | 0.7660 | 0.7040 | 0.7020 | 0.7660 |
| monk1 | 0.6644 | 0.6651 | 0.7014 | 0.6667 | 0.6829 |
| cancer | 0.9795 | 0.9604 | 0.7251 | 0.9591 | 0.9708 |
| ionodata | 0.9261 | 0.8603 | 0.8239 | 0.8750 | 0.8750 |
| splice | 0.7320 | 0.8360 | 0.7620 | 0.5981 | 0.5981 |
| CMC | 0.6916 | 0.7511 | 0.6822 | 0.5794 | 0.7259 |
| blood | 0.7941 | 0.7733 | 0.6123 | 0.8182 | 0.8476 |
| haberman | 0.2745 | 0.7351 | 0.7582 | 0.7255 | 0.6601 |
| liver | 0.6646 | 0.6970 | 0.4512 | 0.6098 | 0.7073 |
| Diabetes | 0.7969 | 0.7695 | 0.6771 | 0.7292 | 0.8047 |

The table 2 compares the performance of the pTWSVM classifier with that of some other SVMs. The experiments of each algorithm were implemented by using MATLAB R2014b. All classifiers are trained with a linear kernel. Optimal values of the parameters were obtained by using a tuning set comprising of 10 percent of the data set.

Form the table we can find that pTWSVM performs best on the vast majority of data sets compared to several other algorithms. And a detail is that pTWSVM behavers better than TWSVM on any dataset, or the same. In addition, from the table, we can find the standard deviation of the new method is always smaller than the standard deviation of other methods are small. This implies that our proposed new method has better robustness and our algorithm has higher stability. This is in line with our expectations.

The experimental results indicate that pTWSVM is not only effective, but also can be a better choice on most data sets. In addition, the accuracy of pTWSVM is very close to the best.

## 5.4 Robustness Against Outliers Samples

Since the main advantage of the new proposed pTWSVM algorithm dedicated to process noisy samples, we will focus on the processing of the data sets with outliers in the following experiments.

First, we construct a noise matrix . The noise matrix will be involved in the benchmark data set and they will constitute the data sets with outliers , where is the noise factor that can determine the degree of data contamination. The differences in accuracy of each algorithm will be obtained by comparing the performances of the origin data and contaminated data. To get a deep association, we take different value in experiment. The following pictures summarizes the performance of different algorithms on some benchmark datasets with different values of .





fig4 accuracy with different noise factor value

Form the pictures above, we can get the following points:

Fist, the proposed pTWSVM method is consistently better then TWSVM method on the experimental data sets, which demonstrate that the proposed new methods is able to effectively improve the clustering accuracy on noisy data with outlier data samples. This also shows that the new pTWSVM method in the practical application will achieve better results.

Second, no matter what the noise factor value is, the accuracy of pTWSVM always be higher then the accuracy of TWSVM. Although the improvements by pTWSVM method over the comparing methods on the original benchmark data sets without noise are mediocre as shown in Table2，the improvements by our new method on the contaminated data with outlier data samples are considerably large. For example, on the heart data set with outliers, the average pTWSVM accuracy of different value is 0.7481, and TWSVM accuracy is 0.6633. So our proposed method improves the clustering accuracy over the TWSVM method by . In contrast, the improvement of

clustering accuracy on the same data set under the noiseless condition is about 4.47% =(0.8667-0.8296)/0.8296 . The same observations can be seen on all the other experimental data sets, which show that the proposed method has better capability to cluster on contaminated data.

Third, the pictures show that the change in accuracy of pTWSVM is flat and does not change much. This clearly indicate that the new proposed pTWSVM method is faster and easier to stabilize than original TWSVM method. The feature confirms pTWSVM method’s robustness against outlier data samples.

6.Conclusions

We have proposed a robust TWSVM based on the -th order of distance, which formulated a non-smooth non-convex minimization problem. Compare to the squared distance，the -th order TWSVM has better accuracy and it is very robust against outlier data samples. The new proposed method takes much more challenging optimization problem than that in the traditional TWSVM. To solve the problem, we introduced an efficient iterative algorithm and provided the rigorous theoretical analysis on the convergence of our algorithm.

# References

[1] Khemchandani R, Jayadeva, Chandra S. Fuzzy Twin Support Vector Machines for Pattern Classification[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(5):905-910.

[2] Cortes C, Vapnik V. Support-Vector Networks[J]. Machine Learning, 1995, 20(3):273-297.

[3] Fung G, Mangasarian O L. Proximal support vector machine classifiers[C]// ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2001:77--86.

[4] Mangasarian O L, Wild E W. Multisurface proximal support vector machine classification via generalized eigenvalues[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006, 28(1):69-74.