The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 101–108

Cancer Related Gene Identification via p-norm Support Vector Machine*

Jun-Yan Tan[1] Chun-Hua Zhang[2,†] Nai-Yang Deng[1,‡]

[1]College of Science, China Agricultural University, Beijing 100083, China
[2]Information School, Renmin University of China, Beijing 100872, China

Abstract This paper focuses on the feature selection in classification via a new version of support vector machine (SVM) named p-norm support vector machine $(0 < p < 1)$. Different from the 2-norm in the standard linear SVM, the p-norm of the normal vector of the decision plane is used which leads to more sparse solution. By using the successive linear algorithm, we can get an approximate local optimal solution to our p-norm SVM. In addition, the lower bounds for the absolute value of nonzero components in every local optimal solution is established, which provides theoretical direction for the elimination of zero components in any numerical solution. The numerical experiments show that the p-norm SVM is effective in selecting relevant features, compared with the popular 1-norm SVM, 0-norm SVM and support vector machine-recursive feature elimination based (SVM-RFE).

1 Introduction

Feature selection is a problem pervasive in many applications of data mining. Restricting the input space to a small subset of input variables has obvious benefits in terms of data storage, computational requirements, and cost of future data collection. Furthermore, it often provides better data or model understanding. Especially, the classical feature selection in biology is to choose the genes that are relevant to some disease, which can be helpful for drug design and disease treatment. This paper focuses on feature section in classification via a new version of support vector machine (SVM) and its application to the gene selection in biology.

Several methods have been proposed for feature selection in the SVM framework. Guyon *et al.* (2002) [1] proposed support vector machine-recursive feature elimination based (SVM-RFE). The SVM-RFE method ranks all the features according to some score function and eliminates several features with the lowest scores in each iteration. This process is repeated to reach the highest classification accuracy. The performance of SVM-RFE is sensitive to the number of features eliminated in each iteration, different numbers lead to different accuracy and different feature subset. In addition, [3] proposed the 1-norm SVM, which imposed the absolute value penalty of the normal vector of the separating plane. In order to get more sparse classifiers, [4, 5] proposed the feature selection

---

*This work is supported by the Key Project of the National Natural Science Foundation of China (No. 10631070), the National Natural Science Foundation of China (No.10971223)

†Corresponding author. E-mail: zhangchunhua@ruc.edu.cn

‡Corresponding author. E-mail: dengnaiyang@cau.edu.cn

via concave minimization (FSV), which can automatically select features by the 0-norm penalty of the number of features. Different from the above methods, another related method, called as SCAD SVM was proposed by Zhang [7] which formulated the SVM as a regularization problem with smoothly clipped absolute deviation(SCAD) penalty on the directional vector of the separating plane. The inconvenience of using SCAD SVM is that there are too many parameter in the model. Another revision of SVM was considered by Wang [8], he proposed a hybrid huberized support vector machine (HHSVM) where the huberized hinge loss function and the 1-norm penalty are applied. HHSVM encouraged the highly correlated features to have similar weights and to be selected or removed together. But the HHSVM was not so sparse, it selected more features than 1-norm SVM.

Recently, p-norm $(0 < p < 1)$ attracted great attention in the optimization framework, the idea that using p-norm can find spare solution is considered in [9, 10, 11, 12]. This paper proposes p-norm support vector machine, which replaces the 2-norm penalty by the p-norm $(0 < p < 1)$ penalty in the objective function of the primal problem in the standard linear SVM. The p-norm SVM conducts feature selection and classification simultaneously. However, there are two difficulties in solving p-norm SVM model: (i). It is impossible to solve the primal problem via its dual problem and the primal problem itself is hard to be solved, because it is neither differentiable nor convex; (ii). Feature selection needs to find the nonzero components of the solution to the primal problem. However, usually algorithms can only provide an approximate solution where nonzero components in the solution can not be identified theoretically. For example, if the numerical results show that the solution have many entries with small values, can we consider these entries to be zero entries?

For the difficulty (i), the primal problem is equivalent to a differentiable one and solved by a successive linear approximation algorithm (SLA) ([2]), which yields a stationary point of the primal problem. Furthermore, for the difficulty (ii), the lower bounds for the absolute value of nonzero entries in every local optimal solution is established, which can eliminate zero entries in any numerical solution. Lastly, the performance of p-norm SVM is illustrated on both of the simulation datasets and real datasets. Experimental results show that p-norm SVM can select a small subset of relevant features with good prediction performance, compared with 1-norm SVM and 0-norm SVM. Especially, according to our results on two real microarray datasets, most of the genes that have been proved to be really relevant to the cancers in other literatures, are chosen by p-norm SVM, which indicates that our method is very effective in the genes selection.

Now we describe our notation. All vectors are column vectors unless transposed to a row vector by a super script $\top$. For a vector x in R^n, $[x]_i (i = 1, 2, \cdots, n)$ denotes the i-th component of x. $|x|$ denotes a vector in R^n of absolute value of the components of x. $\|x\|_p$ denotes that $(|[x]_1|^p + \cdots + |[x]_n|^p)^{\frac{1}{p}}$. Strictly speaking, $\|x\|_p$ is not a general norm when $0 < p < 1$, but we still follow this term p-norm, because the forms are same except that the values of p are different. $\|x\|_0$ is the number of nonzero components of x. For two vectors $x \in R^n$ and $y \in R^n$, $(x \cdot y)$ denotes the inner product of x and y.

This paper is organized as follows. In section 2, the p-norm SVM for feature section is introduced. In section 3, the SLA is proposed to solve p-norm SVM. In section 4, the lower bounds for the absolute value of nonzero entries in any local optimal solution of p-norm SVM is established. In section 5, numerical experiments are given to demonstrate the effectiveness of our method. We conclude this paper in section 6.

2 *p*-norm support vector machine

For feature selection, *p*-norm SVM is an embedded method in which training data are given to a learning machine, which returns a predictor and a subset of features on which it performs predictions. In fact, feature selection is performed in the process of learning.

Consider the classification problem with the training set T is given by

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\}, \tag{1}$$

where $x_j \in R^n$ and $y_j \in \{1, -1\} (j = 1, \cdots, l)$. The feature vector

$$g_i = ([x_1]_i, [x_2]_i, \cdots, [x_l]_i)^{\mathrm{T}}, (i = 1, \cdots, n) \tag{2}$$

denotes the values of i-th feature in all inputs.

Replacing $\frac{1}{2}\|w\|_2^2$ in the objective function of the standard SVM by $\|w\|_p^p$, we have

$$\min_{w,b,\xi} \quad \|w\|_p^p + C\sum_{i=1}^{l} \xi_i \, , \tag{3}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) \geq 1 - \xi_i \, , \; i = 1, \cdots, l \, , \tag{4}$$

$$\xi_i \geq 0 \, , \; i = 1, \cdots, l \, , \tag{5}$$

then, the linear *p*-norm support vector machine is described as follows:

Algorithm1: p-norm SVM

(1) Select the parameters $C (C > 0)$ and $p (0 < p < 1)$; using the set given by (1), construct the primal problem (3)-(5);

(2) Find the global solution (w^*, b^*, ξ^*) to the problem (3)-(5);

(3) Select the feature set: $\{i | [w^*]_i \neq 0, i = 1, \cdots, n\}$;

(4) Construct the decision function $f(x) = \mathrm{sgn}((\tilde{w}^* \cdot \tilde{x}) + b^*)$, where the components of $\tilde{w}^*$ are nonzero components of w^* and the components of $\tilde{x}$ are also corresponding to nonzero components of w^*.

Note that, in the Algorithm 1, there are two difficulties (i) and (ii) that have been addressed in Section 1, so the following sections will consider them respectively.

3 The SLA algorithm for the problem (3)-(5)

Consider the problem (3)-(5), the objective function is not differentiable, because of the absolute value in the first item. In order to make this problem smooth, the variable $v = ([v]_1, \cdots, [v]_n)^\top$ is introduced to eliminate the absolute value from the problem (3)-(5), which leads to the following equivalent problem:

$$\min_{w,b,\xi,v} \quad \|v\|_p^p + C\sum_{i=1}^{l} \xi_i \, , \tag{6}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) \geq 1 - \xi_i \, , \; i = 1, \cdots, l \, , \tag{7}$$

$$\xi_i \geq 0 \, , \; i = 1, \cdots, l \, , \tag{8}$$

$$-v \leq w \leq v, \tag{9}$$

where, $\|v\|_p^p = [v]_1^p + \cdots + [v]_n^p$ due to the constraint (9). Furthermore, we note that the problem (6)-(9) is differentiable, but not convex. In fact, it is the minimization of a concave objective function over at polyhedral set. Even though it is difficult to find a global solution to this problem, a fast successive linear approximation (SLA) algorithm ([2]) terminates finitely at a stationary point which satisfies the necessary optimality condition for problem (6)-(9). For convenience, we state the SLA algorithm below.

Algorithm2: SLA for the problem (3)-(5)

(1) Select the parameter C ($C > 0$) and p ($0 < p < 1$), start with a random ($w^{(0)}$, $b^{(0)}$, $\xi^{(0)}$, $v^{(0)}$) and let $k = 1$;

(2) Find the solution $(w^{(k)}, b^{(k)}, \xi^{(k)}, v^{(k)})$ to the following linear programming:

$$\begin{aligned} \min_{w,b,\xi,v} \quad & p(v^{(k-1)})^{p-1} v + C\sum_{i=1}^{l} \xi_i \ , && (10) \\ \text{s.t.} \quad & y_i((w \cdot x_i) + b) \geq 1 - \xi_i \ , \ i = 1, \cdots, l \ , && (11) \\ & \xi_i \geq 0 \ , \ i = 1, \cdots, l \ , && (12) \\ & -v \leq w \leq v, && (13) \end{aligned}$$

where $(v^{(k-1)})^{p-1} = ([v^{(k-1)}]_1^{p-1}, \cdots, [v^{(k-1)}]_n^{p-1})^\top$;

(3) If $|p(v^{(k)})^{p-1}(v^{(k)} - v^{(k-1)}) + C\sum_{i=1}^{l}(\xi^{(k)} - \xi^{(k-1)})| = 0$, then let $w^* = w^{(k)}, b^* = b^{(k)}, \xi^* = \xi^{(k)}$ and stop; otherwise, let $k = k + 1$ and go to step 2.

4 The Lower bounds for nonzero components in solutions

Using a similar strategy in [9], we get the following theorem 1, which can be used to identify nonzero components in any local optimal solutions to the problem (3)-(5), even though the Algorithm 2 can only find the approximate local optimal solution.

Theorem 1 For any local optimal solution (w^*, b^*, ξ^*) to the problem (3)-(5), if $[w^*]_i \in (-L_i, L_i)$, then $[w^*]_i = 0 (i = 1, 2 \cdots n)$, where $L_i = (\frac{p}{C\sqrt{l}\|g_i\|_2})^{\frac{1}{1-p}}$ and g_i is defined in (2).

Proof: Suppose $\|w^*\|_0 = k$. Without loss of generality, let $w^* = ([w^*]_1, [w^*]_2, \cdots, [w^*]_k, 0, 0 \cdots 0)^T$ and $z^* = ([w^*]_1, [w^*]_2, \cdots, [w^*]_k)^T$. For the new training set $\widetilde{T} = \{(\widetilde{x}_1, y_1), \cdots, (\widetilde{x}_l, y_l)\}$, where $\widetilde{x}_i = ([x_i]_1, [x_i]_2, \cdots, [x_i]_k)^\top$,we consider the optimization problem

$$\begin{aligned} \min_{z,b,\boldsymbol{\xi}} \quad & \|z\|_p^p + C\sum_{i=1}^{l} \xi_i \ , && (14) \\ \text{s.t.} \quad & y_i((z \cdot \widetilde{x}_i) + b) \geq 1 - \xi_i \ , \ i = 1, \cdots, l \ , && (15) \\ & \xi_i \geq 0 \ , \ i = 1, \cdots, l \ . && (16) \end{aligned}$$

The Lagrange function of (14)-(16) is: $L(z, b, \xi, \alpha, \zeta) = \|z\|_p^p + C\sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i(y_i(z^T\widetilde{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{l} \zeta_i \xi_i$. According to the KKT condition, if (z^*, b^*, ξ^*) is a local optimal

Table 1: Simulated datasets

Datasets	Methods	No. of selected features	Percent of relevant features(%)	Average error(%)	Precision	Sensitivity
$n = 20$	p-norm SVM	**3.34**	**89.8**	**3.86**	0.948	**0.954**
	0-norm SVM	5.79	51.8	3.88	**0.956**	0.954
	1-norm SVM	15.7	28.4	5.39	0.937	0.946
	SVM-RFE	4	75	11.2	0.849	0.892
$n = 50$	p-norm SVM	**4**	**74.1**	4.39	0.944	0.959
	0-norm SVM	4.2	72.1	**3.18**	**0.965**	**0.962**
	1-norm SVM	21.7	13.8	7.41	0.915	0.921
	SVM-RFE	5	60	18.5	0.784	0.791

solution of (14)-(16), then

$$\nabla_z L(z,b,\xi,\alpha,\zeta) = p|z^*|^{p-1}\text{sgn}(z^*) - \sum_{i=1}^{l} \alpha_i y_i \widetilde{x}_i = 0, \tag{17}$$

$$\nabla_{\xi_i} L(z,b,\xi,\alpha,\zeta) = C - \alpha_i - \zeta_i = 0, i = 1,\cdots,l, \tag{18}$$

$$\zeta_i \geq 0, \alpha_i \geq 0, i = 1,\cdots,l. \tag{19}$$

By (17), we have $p|z^*|^{p-1}$=$|\sum_{j=1}^{l} \alpha_j y_j \widetilde{x}_j|$, which is equivalent to $p|[z^*]_i|^{p-1}$=$[|\sum_{j=1}^{l} \alpha_j y_j \widetilde{x}_j|]_i$ =$|(\alpha_1 y_1 \cdots \alpha_l y_l)g_i|$, $i = 1,\cdots,k$. By the Cauchy-Schwarz inequality, (18) and (19), we have $|(\alpha_1 y_1 \cdots \alpha_l y_l)g_i| \leq \|(\alpha_1 y_1 \cdots \alpha_l y_l)\| \|g_i\| \leq C\sqrt{l}\|g_i\|$. Therefore, $|[z^*]_i| \geq (\frac{p}{C\sqrt{l}\|g_i\|})^{\frac{1}{1-p}}$, which means that for any local optimal solution (w^*, b^*, ξ^*) of (3)-(5), we have $[w^*]_i \in (-L_i, L_i) \Rightarrow [w^*]_i = 0, i = 1,2,\cdots,n.\Box$

According to Theorem 1, we can identify the nonzero components of the local optimal solution to (3)-(5). Based on the Algorithm 2 and the Theorem 1, the new algorithm is established as follows:

Algorithm 3

(1) Select the parameters $C(C > 0)$ and $p(0 < p < 1)$; using the set given by (1), construct the optimization problem (3)-(5);

(2) Using the Algorithm 2 to get the local optimal solution (w^*, b^*, ξ^*) to (3)-(5);

(3) Compute $L_i = (\frac{p}{C\sqrt{l}\|g_i\|})^{\frac{1}{1-p}} (i = 1,\cdots,n)$ and select the feature set: $\{i | [w^*]_i| > L_i, i = 1,\cdots,n\}$;

(4) Construct the decision function $f(x) = \text{sgn}((\tilde{w}^* \cdot \tilde{x}) + b^*)$, where the components of $\tilde{w}^*$ are nonzero components of w^* and the components of $\tilde{x}$ are also corresponding to nonzero components of w^*.

In the following section, our experiments are conducted according to the algorithm 3.

5 Numerical experiments

In this section, some experiments on simulated datasets and real datasets are conducted, by comparing p-norm SVM with 0-norm SVM, 1-norm SVM and SVM-RFE.

Simulated datasets

Table 2: Real datasets

Datasets	No. of features	Methods	No. of selected features	Average errors(%)	Precision	Sensitivity
Arcene	10000	p-norm SVM	**6.5**	17.6	**0.877**	0.764
		0-norm SVM	27.4	24.6	0.709	0.736
		1-norm SVM	12.9	**19.8**	0.844	0.763
		SVM-RFE	70	16.6	0.797	**0.836**
Colon caner	2000	p-norm SVM	**4.6**	16.1	0.75	0.811
		0-norm SVM	13.2	14.5	0.760	0.814
		1-norm SVM	15.6	**13.5**	**0.773**	**0.886**
		SVM-RFE	64	14.5	0.771	0.795
Prostate cancer	12600	p-norm SVM	**8.3**	**2.9**	**0.976**	**0.971**
		0-norm SVM	10.5	4.7	0.947	0.941
		1-norm SVM	17.9	3.5	0.962	0.97
		SVM-RFE	50	5.1	0.961	0.934

We generate the simulated data set using the method described in [18]

- Independently generate 100 stochastic vectors $x_i \in R^n, i = 1, 2, \cdots, 100$ and $[x_i]_j \sim N(0,1), j = 1, 2, \cdots, n$;
- The outputs are determined by the hyperplane $g(x) = 4[x]_1 + 2[x]_2 + 4[x]_3 - 0.6$, which means that the output of an input x_i is "+1" if $g(x_i) \geq 0$ and is "-1" if $g(x_i) < 0$.

We consider the dimensionality of the inputs $n = 20, 50$. The experiments will repeat 100 times and the average test errors are computed. Note that, in Algorithm 3, the performance of p-norm SVM depends on the parameters C and p. Therefore, C and p should be adjusted properly. In our experiments, the best value of $C \in [2^-5, 2^5]$ and $p \in (0,1)$ is chosen by ten-fold cross validation.

According to Algorithm 3, 100 experiments are conducted for every datasets. Note that, in Algorithm 3, the performance of p-norm SVM depends on the parameters C and p. Therefore, C and p should be adjusted properly. In our experiments, the best value of C and p is chosen by ten-fold cross validation.

Our experimental results are illustrated in Table 1. Obviously, p-norm SVM performs the best. In Table 1, the data in 4th column shows the percentage of the number of the right features over the number of the selected features, which means the bigger the value, the better the result. The average error, average precision and sensitivity are computed by 100 experiments. It's easy to see that p-norm SVM selects the least features with the best accuracy. Along with the increasing of the number of redundant features, the performance of p-norm SVM is still the best in selecting relevant features.

Real datasets

To test our method on real-world data, one NIPs benchmark data set('arcene' [13]) and two real microarray data sets ('colon cancer' [1] and 'prostate cancer'[14]) are used. According to Algorithm 3, the ten-fold cross validations on 3 datasets are conducted, and the results are listed in Table 2. We can see that p-norm SVM performs better than the other three methods in selecting features and the test error of of p-norm SVM is lower.

Table 3: The genes selected by p-norm SVM

	Gene	Entrez	Discription
Colon cancer	377	Hsa.36689	H.sapiens mRNA for GCAP-II/uroguanylin precursor
	765	Hsa.692	Human cysteine-rich protein (CRP) gene
	1870	Hsa.1660	Peptidyl-prolyl CTS-TRANS isomerase
	1325	Hsa.3016	S-100P Protein
	1423	Hsa.1832	SMOOTH MUSCLE ISOFORM (HUMAN)
	1873	Hsa.404	Human MXI1 mRNA
Prostate cancer	6185	Hs.823	hepatoma mRNA for serine protease hepsin
	2839	Hs.281587	Human (clone CTG-A4) mRNA sequence
	6390	Hs.95420	Homo sapiens cDNA, 3' end
	10672	Hs.146355	Human c-abl gene, complete cds
	6145	Hs.81047	untitled
	10234	Hs.30250	Homo sapiens short form transcription factor C-MAF mRNA

Therefore, p-norm SVM is effective in selecting relevant features, compared with 1-norm SVM, 0-norm SVM and SVM-RFE.

The genes selected by our method are listed in Table 3. Most of them have been identified as tumor or tissue specific genes. For example, the gene 377, 1423, 765, 1870 in colon cancer dataset, have been proved to be relevant to colon cancer in [8]. For the prostate cancer dataset, all selected genes except the gene 6145 are said to be relevant to the prostate cancer in [19]. We should pay more attention to the gene 6145, it is selected by p-norm SVM that means it might be relevant to the prostate cancer.

6 Conclusion

Feature selection is very important in many applications of data mining. This paper focuses on the feature selection in classification via a new version of SVM named p-norm SVM. Different from the 2-norm in the standard linear SVM, p-norm $(0 < p < 1)$ of the normal vector of the decision plane is used which leads to more sparse solution. By using the successive linear algorithm, we can get an approximate local optimal solution to the primal problem of the p-norm SVM. Furthermore, the lower bounds for the absolute value of nonzero components in every local optimal solution is established, which provide theoretical direction for the elimination of zero components in any numerical solution. The numerical experiments show that the p-norm SVM is effective in selecting relevant features, compared with 1-norm SVM, 0-norm SVM and SVM-RFE.

References

[1] Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46:** , 389-422,2002.

[2] Bradley P.S. ,Mangasarian O.L., and Street W.N., Feature selection via mathematical programming. INFORMS Journal on Computing, 1998.

[3] Bradley P.S., Mangasarian O.L. Feature selection via concave minimization and support vector machines. *In Proc. 13th ICML*,82-90,1998.

[4] Fung G., Mangasarian O. L. Data Selection for Support Vector Machine Classifiers, *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 20-23, 2000.

[5] Zhu J., Rosset S., Hastie T. Tibshirani R. 1-norm svms, Advances in Neural *Information Processing Systems* 16, 2003.

[6] Mangasarian O.L., Wild E.W. Feature selection for nonlinear kernel support vector machines.*IEEE Seventh International Conference on Data Mining (ICDM'07)*,2007.

[7] Zhang H.H., Ahn J.Y., Lin X.D.and Park C.W. Gene selection suing suport vector machines with non-convex penalty. *bioinformatics*,**22:** 88-96,2006.

[8] Wang L., Zhu J., Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, **23:** 2507-2517,2008.

[9] Chen X.J.,Xu F.M., and Ye Y.Y., Lower bound theory of nonzero entries in solutions of l_2-l_p minimization, http://www.polyu.edu.hk/ama/staff/xjchen/LpnormJan28.pdf.

[10] Bruckstein A.M., Donoho D.L. and Elad M., From sparse sulutions of systems of equations to sparse modeling of signals and images, *SIAM Reviewer*, **51**,34-81, 2009.

[11] Fan J. and Li R., Varible selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96**, 1348-1360, 2001.

[12] Xu Z., Zhang H. Wang Y. and Chang X., $L_{\frac{1}{2}}$ regularizer, *Science in China Series F-Inf Sci,52*,1-9, 2009.

[13] http://www.nipsfsc.ecs.soton.ac.uk/datasets/

[14] Singh D., Febbo P., Ross K., Jackson D., Manola J., Ladd C., Tamayo P., Renshaw A., D'Amico A., Richie J., Lander E., Loda M., Kantoff P., Golub T.,Sellers W. Gene expression correlates of clinical prostate cancer behavior. *cancer cell*,**1:**,203-209,2002.

[15] Neumann J., Schnörr C., Steidl G. Combined SVM-based feature selection and classification. *Mach. Learn.*. **61:** 129-150,2005.

[16] Golab J., Interleukin 18–interferon gamma inducing factor–a novel player in tumor immunotherapy?, Cytokine., 12:332-338, 2000.

[17] Dinarello C. A., Interleukin-18, Methods., 19:121-132, 1999.

[18] Zhang H.H., Ahn J., Lin X.D., Park C. Variable selection for svm via shrinkage methods. Available at http://math.uc.edu/ linxd/paper/9.pdf

[19] Lai Y.L.,Genome-wide co-expression based prediction of differential expressions. *Bioinformatics*, 24 666-673,2008.