Primal-Dual Framework for Feature Selection using Least Squares Support Vector Machines

Raghvendra Mall &
Johan A.K. Suykens
ESAT-STADIUS, KU Leuven
rmall@esat.kuleuven.be

Mohammed El Anbari &
Halima Bensmail
Qatar Computing Research
Institute
{melanbari,hbensmail}@qf.org.qa

ABSTRACT

Least Squares Support Vector Machines (LSSVM) perform classification using L_2-norm on the weight vector and a squared loss function with linear constraints. The major advantage over classical L_2-norm support vector machine (SVM) is that it solves a system of linear equations rather than solving a quadratic programming problem. The L_2-norm penalty on the weight vectors is known to robustly select features. The zero-norm or the number of non-zero elements in a vector is an ideal quantity for feature selection. The L_0-norm minimization is a computationally intractable problem. However, a convex relaxation to the direct zero-norm minimization was proposed recently. In this paper, we propose a combination of L_2-norm penalty and the convex relaxation of the L_0-norm penalty for feature selection in classification problems. We propose a primal-dual framework for feature selection using the combination of L_2-norm and L_0-norm penalty resulting in closed form solution. A series of experiments on microarray data and UCI data demonstrates that our proposed method results in better performance.

1. INTRODUCTION

Least Squares Support Vector Machines (LSSVM) [1] is an alternative to the standard support vector machines (SVM) [2]. It is a widely used tool for classification and regression problems. Given a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where the input $x_i \in \mathbb{R}^d$ is a vector with d features and the class label $y_i \in \{-1, +1\}$, the LSSVM finds an optimal hyperplane to separate the two classes using the following optimization problem:

$$\begin{aligned} \min_{w,e_k,b} \quad & \frac{1}{2}\lambda||w||^2 + \frac{1}{2}\sum_{k=1}^{N} e_k^2 \\ \text{such that} \quad & e_k = y_k - w^\intercal \phi(x_k) - b, k = 1, \ldots, N, \end{aligned} \tag{1}$$

where λ is a regularization constant, e_k is the error corresponding to the k^{th} point and b is the bias term. Here $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{n_h}$ is a mapping to a high dimensional feature space as in the standard SVM [2] case. Throughout this paper we use the *linear kernel*. This means that $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ or $\phi(x) = x$. This allows to have interpretable models as the feature space is known beforehand. Finally the classifier in the primal is defined as: $y(x) = \text{sign}[w^\intercal \phi(x) + b]$.

The corresponding dual classifier is defined as: $y(x) = \text{sign}[\sum_{k=1}^{N} \alpha_k K(x_k, x) + b]$. Here $K(x_k, x_j) = \phi(x_k)^\intercal \phi(x_j)$, K is a positive definite kernel function and α_k are the Lagrange multipliers which can be positive or negative due to equality constraints. The KKT conditions lead to $w = \sum_{k=1}^{N} \alpha_k \phi(x_k)$ and $e_k = \frac{1}{\gamma}\alpha_k$. The second KKT condition makes the LSSVM solutions non-sparse i.e. each data point is considered as a support vector. Thus, the LSSVM formulation in [1] has the form of a *penalty+loss* with the λ playing the role of regularizer.

The major advantage of the LSSVM formulation over a standard SVM is that the equality constraints and the squared loss function leads to solving a system of linear equations instead of a quadratic programming (QP) problem as in the case of classical SVM. It is widely known [1, 3] that solving a system of linear equations is computationally easier than solving QPs. In this paper we take this into consideration and the proposed method always solves a system of linear equations and have closed form solutions.

The zero-norm defined as $||w||_0 = \text{card}\{w_i | w_i \neq 0\}$ counts the number of non-zero elements in the vector w. When the zero-norm is minimized it results in very sparse models. Recently, the zero norm has been receiving a lot of attention in the machine learning community [4, 5, 6, 7]. The minimization of the zero-norm is a computationally intractable problem as shown in [8]. This is because the zero-norm minimization is non-convex and NP-hard problem. However, recently a direct zero-norm optimization method was proposed in [9] which can achieve the true zero-norm asymptotically under Bayesian interpretation. This is closely related to the concept of Automatic Relevance Determination (ARD) for feature selection [11].

1.1 Motivations & Contributions

The role of L_2-norm in feature selection for SVM classifiers is to select the similar set of features upon different randomizations of the data. This leads to robustness in selection of features [10]. The L_2-norm penalty also results in shrinkage. It fits the coefficients toward zero but cannot make the coefficients *exactly* zero. So, in this paper we combine the L_2-norm penalty along with the convex relaxation for direct zero-norm penalty as formulated in [9, 6] for feature selection using LSSVM classifiers. The proposed method selects groups of essential features for classification and eliminates

The 19th International Conference on Management of Data (COMAD), 19th-21st Dec, 2013 at Ahmedabad, India.

the unnecessary variables. We propose a primal-dual framework for sparse feature selection using a combination of L_2-norm and L_0-norm penalty while taking into consideration both the cases when $N \gg d$ and when $d \gg N$. Due to space limitations we refer the readers to [4, 9, 12, 13, 14, 15, 17, 18, 19, 20] for related work.

2. PROPOSED METHOD

The direct zero-norm optimization method results in an iterative convex formulation for L_0-norm based classifiers [6, 9]. It results in a local minimum to the non-convex zero-norm problem with good predictive capabilities and sparsity in both the feature and input space [9, 6]. However, the L_0-norm penalty doesn't guarantee the selection of the same set of variables for different randomizations. Thus, we use the L_2-norm penalty in combination with L_0-norm penalty along with a squared loss function in our formulation.

2.1 Primal Formulation

We pre-process the dataset $\mathcal{D}$ to be mean-centered and have unit norm along each dimension d. Since the data is mean-centered we don't have the intercept term b. The constrained optimization problem for the proposed approach at iteration t is given by:

$$\min_{w^{(t)}, e_k} \quad \frac{1}{2}\lambda||w||^2 + \frac{1}{2}w^\intercal \Lambda^{(t-1)} w + \frac{1}{2}\sum_{k=1}^{N} e_k^2 \qquad (2)$$
$$\text{such that} \quad e_k = y_k - w^\intercal x_k, k = 1, \ldots, N,$$

where λ is the regularization parameter and $\Lambda^{(t-1)} = \text{diag}(\frac{1}{|w_1^{(t-1)}|^2}, \ldots, \frac{1}{|w_d^{(t-1)}|^2})$. The $w^\intercal \Lambda^{(t-1)} w$ term in the optimization function is the convex relaxation to the $||w||_0$ minimization. The L_0-norm penalty term is the same as that introduced in [9, 6]. After elimination of e_k in (2), we can obtain the following convex unconstrained optimization problem:

$$\min_{w^{(t)}} \quad \frac{1}{2}\lambda||w||^2 + \frac{1}{2}w^\intercal \Lambda^{(t-1)} w + \frac{1}{2}\sum_{k=1}^{N} (y_k - w^\intercal x_k)^2 \qquad (3)$$

The solution to (3) at each iteration t can be obtained by directly differentiating the convex optimization function in (3) w.r.t to w. It results in a iteratively weighted ridge regression [21] like solution:

$$w^{(t)} = (\lambda \mathcal{I} + \Lambda^{(t-1)} + \mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal Y \qquad (4)$$

where $\mathbf{X} = [x_1, x_2, \ldots, x_N]^\intercal$ and $Y = [y_1, y_2, \ldots, y_N]^\intercal$. This solution corresponds to the primal and is more appropriate for the case when $N \gg d$. The final classifier in the primal is then defined as: $y(x) = \text{sign}[w^{(t)\intercal} x]$.

Since the proposed approach follows an iterative procedure to a local minimum, it is needed to have a good starting value. We initially solve the LSSVM problem to obtain the weight vector $w^{(0)}$. The regularization parameter λ is also obtained by solving the LSSVM problem via coupled simulated annealing (CSA) [22]. Thus, the initial value of $\Lambda^{(0)} = \text{diag}(\frac{1}{|w_1^{(0)}|^2}, \ldots, \frac{1}{|w_d^{(0)}|^2})$. The L_0-norm penalty doesn't introduce any additional tuning parameters as in [9], performs direct zero-norm objective minimization and is advantageous over AROM and FSV methods.

2.2 Dual Formulation

One of the KKT conditions of LSSVM provides the connection between the primal weight vector w and the dual Lagrange multipliers α_k. The relation is given by $w = \sum_{k=1}^{N} \alpha_k x_k = \mathbf{X}^\intercal \alpha$ where $\alpha = [\alpha_1, \ldots, \alpha_N]^\intercal$. In the case when the number of points in the dataset is much less than the number of features in the dataset i.e. $d \gg N$, it is more suitable to solve the problem in the dual. Given the connection between w and α, replacing α in (3) results in the following convex unconstrained optimization problem:

$$\min_{\alpha^{(t)}} \quad \frac{1}{2}\lambda \alpha^\intercal \mathbf{X}\mathbf{X}^\intercal \alpha + \frac{1}{2}\alpha^\intercal \mathbf{X} \Lambda^{(t-1)} \mathbf{X}^\intercal \alpha + \frac{1}{2}\sum_{k=1}^{N} (y_k - \alpha^\intercal \mathbf{X} x_k)^2 \qquad (5)$$

where λ is the regularization parameter and $\Lambda^{(t-1)} = \text{diag}(\frac{1}{|w_1^{(t-1)}|^2}, \ldots, \frac{1}{|w_d^{(t-1)}|^2})$. The $\alpha^\intercal \mathbf{X} \Lambda^{(t-1)} \mathbf{X}^\intercal \alpha$ term in the optimization function is the convex relaxation to the $||w||_0$ minimization. The solution to (5) at each iteration t can be obtained by directly differentiating the convex optimization function in (5) w.r.t to α. In (5), we can replace $\mathbf{X}\mathbf{X}^\intercal$ by the kernel matrix K as it is the linear kernel case. The solution to (5) is given by:

$$\alpha^{(t)} = (\lambda K + \mathbf{X} \Lambda^{(t-1)} \mathbf{X}^\intercal + K K^\intercal)^{-1} K^\intercal Y \qquad (6)$$

Once we obtain the solution vector $\alpha^{(t)}$ for iteration t, we recalculate the weight vector $w^{(t)} = \sum_{k=1}^{N} \alpha_k^{(t)} x_k$ and re-evaluate $\Lambda^{(t)}$ using the aforementioned procedure. The initial coefficients $\alpha^{(0)}$ and the regularization parameter λ are obtained by solving the LSSVM classifier in the dual. The initial weight vectors $w^{(0)} = \sum_{k=1}^{N} \alpha_k^{(0)} x_k$ and $\Lambda^{(0)} = \text{diag}(\frac{1}{|w_1^{(0)}|^2}, \ldots, \frac{1}{|w_d^{(0)}|^2})$.

2.3 Stopping Criteria

The iterative procedure proposed for the primal and dual formulation is executed till we either reach convergence or we reach a maximum number of iterations (*max iterations*). In case of the primal, we define a threshold $\theta = \frac{||w^{(t)} - w^{(t-1)}||_2^2}{d}$. For the dual this threshold is defined as $\theta = \frac{||w^{(t)} - w^{(t-1)}||_2^2}{N}$. We continue the iterative procedure until this threshold θ reaches machine precision (denoted by ϵ). Empirically, we observed that generally 5 to 10 iterations suffice. Once the iterative procedure stops, we follow the setup in [4] and select the top r features from the weight vector such that $||w||_0 = r$.

3. EXPERIMENTAL RESULTS

In this section, we compare our proposed L_2-norm and L_0-norm (L2+L0) penalty based feature selection method with FSV method [15] and L_1-SVM [17] from the LibLinear library (`http://www.csie.ntu.edu.tw/~cjlin/liblinear/`) in the primal as these methods have formulations in the primal. We compare the proposed approach with AROM method [4] and Recursive Feature Elimination (RFE) [23] in the dual as these methods are computationally cheaper in the dual. We utilize the implementation of the aforementioned methods from the matlab toolbox of Spider (`http://www.kyb.tuebingen.mpg.de/bs/people/spider/index.html`). We also compare the proposed methodology with a primal-dual formulation of direct zero-norm minimization based LSSVM (D-L0) [9] and the original LSSVM [1].

3.1 Experiments

We demonstrate our results on 4 microarray gene datasets in the dual. Out of these 4 datasets, two datasets are cancer microarray datasets namely Colon and Leukemia which are obtained from UCI repository [24]. The other two microarray datasets are obtained from `http://featureselection.asu.edu/datasets.php`. We also illustrate the effectiveness of our proposed approach over 6 datasets in the primal. These datasets are also obtained from the UCI repository.

Algorithm 1: Primal-Dual framework for feature selection using LSSVM

Data: $\mathcal{D} = \{(x_i, y_i) : x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$ for classification, $i = 1, \ldots, N\}$.
Result: The optimal feature vector $w \in \mathbb{R}^d$ s.t. $|w_i| \geq 0$, $i = 1, \ldots, r$.

if $N \gg d$ **then**
  Solve LSSVM classifier in primal to obtain $w^{(0)}$ and λ.
  Initialize $\Lambda^{(0)} = \text{diag}(\frac{1}{|w_1^{(0)}|^2}, \ldots, \frac{1}{|w_d^{(0)}|^2})$, $\theta = \inf$ & $cnt = 0$.
  while $\theta > \epsilon$ *and* $cnt < max\ iterations$ **do**
    Solve (4) to obtain $w^{(t)}$.
    Calculate $\Lambda^{(t)} = \text{diag}(\frac{1}{|w_1^{(t)}|^2}, \ldots, \frac{1}{|w_d^{(t)}|^2})$.
    Estimate $\theta = \frac{||w^{(t)} - w^{(t-1)}||_2^2}{d}$.
    Increment cnt to $cnt + 1$.
else if $d \gg N$ **then**
  Solve the LSSVM classifier in the dual to obtain $\alpha^{(0)}$ and λ.
  Calculate $w^{(0)} = \sum_{k=1}^{N} \alpha_k^{(0)} x_k$.
  Initialize $\Lambda^{(0)} = \text{diag}(\frac{1}{|w_1^{(0)}|^2}, \ldots, \frac{1}{|w_d^{(0)}|^2})$, $\theta = \inf$ & $cnt = 0$.
  while $\theta > \epsilon$ *and* $cnt < max\ iterations$ **do**
    Solve (6) to obtain $\alpha^{(t)}$.
    Estimate $w^{(t)} = \sum_{k=1}^{N} \alpha_k^{(t)} x_k$.
    Calculate $\Lambda^{(t)} = \text{diag}(\frac{1}{|w_1^{(t)}|^2}, \ldots, \frac{1}{|w_d^{(t)}|^2})$.
    Evaluate $\theta = \frac{||w^{(t)} - w^{(t-1)}||_2^2}{d}$.
    Increment cnt to $cnt + 1$.
Sort the final weight vector w based on its absolute values.
Select the top r features s.t. $||w||_0 = r$ and set rest to 0.

We randomly partition the dataset into 80% as the training set and 20% as the test set. In order to estimate the value of the hyper-parameter λ, we perform 50 cross-validations of LSSVM using CSA [22]. We first use the training set for feature selection by specifying a given number of features (r). After obtaining the desired weight vector w, classification is performed over the test set using this reduced weight vector. All the experiments are conducted on a PC with 4 Gb RAM, 3Ghz CPU using Matlab 2009a.

3.2 Dual Experimental Results

We evaluate the predictive performance of various feature selection methods in the dual on the 4 microarray gene datasets as shown in Figure 1. From Figure 1 we can observe that the $L_2 + L_0$-norm penalty based proposed approach results in lower or equal error estimates than the original LSSVM in most cases for different value of r. This justifies the need of feature selection before prediction is done. For all the datasets, the L_2-norm and L_0-norm penalty (L2+L0) based method and the direct L_0-norm (D-L0) based method perform better than AROM, RFE, L_1-norm SVM and standard LSSVM for different values of r with the exception of **GLI** dataset. For the **GLI** dataset, the AROM, RFE and L_1-norm SVM performs better but they are computationally more expensive methods. In general, between the proposed approach (L2+L0) and D-L0, our method performs better for all the 4 microarray datasets.

Data	Method	Largest Common Feature subset size									
		100	**300**	**500**	**700**	**900**	**1100**	**1300**	**1500**	**1700**	**1900**
	L2+L0	**26**	**66**	**120**	**196**	**290**	**395**	**508**	**648**	**870**	**1412**
C	**D-L0**	2	13	41	83	145	236	350	502	751	1327
O	**AROM**	3	12	39	78	148	225	352	498	753	1330
L	**RFE**	7	18	37	90	135	240	346	498	771	1350
	L1	8	16	36	84	140	238	351	501	768	1346
	LSSVM	15	52	87	150	214	322	435	568	810	1354
		100	**800**	**1500**	**2200**	**2900**	**3600**	**4300**	**5000**	**5700**	**7100**
	L2+L0	**17**	**229**	**491**	**766**	**1134**	**1487**	**1866**	**2325**	**3165**	**7037**
L	**D-L0**	9	193	434	676	1022	1370	1795	2267	2859	6882
E	**AROM**	10	183	431	667	1015	1410	1850	2238	2901	6866
U	**RFE**	8	178	429	669	1018	1312	1750	2256	2714	6737
	L1	7	169	422	671	1001	1332	1772	2301	2702	6797
	LSSVM	13	219	454	704	1032	1376	1773	2205	2788	6872
		100	**2300**	**4500**	**6700**	**8900**	**11100**	**13300**	**155000**	**17700**	**22100**
	L2+L0	**100**	229	856	1905	**3400**	**5420**	**7798**	**10715**	**14026**	**21920**
G	**D-L0**	2	380	671	1066	1610	2469	3653	5357	7572	20853
L	**AROM**	12	278	651	1166	1810	2579	3573	5735	8127	19959
I	**RFE**	8	292	701	1256	1610	2456	3842	5912	7601	20129
	L1	8	288	699	1244	1700	2501	3678	5882	7812	20259
	LSSVM	4	**452**	**1137**	**1938**	2899	4048	5305	6785	8633	20749
		100	**2100**	**4100**	**6100**	**8100**	**10100**	**12100**	**14100**	**16100**	**18100**
	L2+L0	**10**	34	210	545	1064	1902	2989	4545	6778	**10772**
S	**D-L0**	7	54	215	534	1017	1664	2584	3926	5906	9849
M	**AROM**	6	33	201	526	999	1676	2612	4010	6091	9958
K	**RFE**	5	32	212	536	1010	1767	2588	3992	5990	10100
	L1	6	28	221	522	1009	1812	2489	3891	6019	9845
	LSSVM	6	**293**	**741**	**1308**	**2001**	**2908**	**3918**	**5157**	**6995**	10340

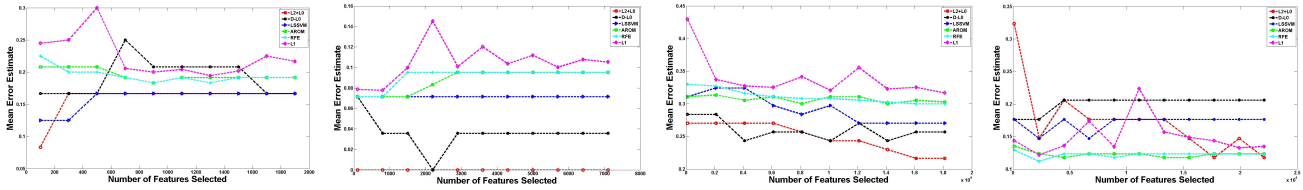
Table 1: Comparison of largest common feature set sizes over 10 randomizations for different feature selection methods in the dual corresponding to various values of r

Table 1 contains information about the largest common subset size over 10 randomizations for different feature selection methods. This indicates the features that appeared consistently during each randomization for a given value of r s.t. $||w||_0 = r$. Higher values indicate the presence of a set of variable which is consistently being selected. Thus, it corresponds to the *robustness* of the proposed approach. We observe that the proposed (L2+L0) approach is more robust in selection of features than the other methods for Col and Leu (cancer microarray datasets). However, for the SMK dataset, the LSSVM method shows more *robustness* in general. As we mentioned earlier that it is the L_2-norm penalty which leads to robustness in selection of similar sets of variables, the standard LSSVM formulation also uses L_2-norm penalty and hence shows this *robustness*.

3.3 Primal Experimental Results

We conducted experiments on 6 UCI datasets in the primal i.e. when $N \gg d$. Table 2 demonstrates the effectiveness of the proposed approach in comparison to methods like L_1 SVM, FSV method, direct zero-norm based LSSVM and the standard LSSVM. Table 2 contains information about the value of r corresponding to which each method performs best in terms of predictive power.

From Table 2 we observe that the proposed approach (L2+L0) outperforms other methods in terms of accuracy for 3 datasets. It showcases that our method leads to maximum sparsity w.r.t. feature selection. However, for the Musk1 (Mus) dataset feature selection is not beneficial. This can be observed from Table 2 since the best results correspond to LSSVM for $r = 166$. We only highlight those re-



(a) Colon Dataset (b) Leukemia Dataset (c) SMK-CAN-187 Dataset (d) GLI-85 Dataset

Figure 1: Results of various feature selection methods for different subset size on several microarray datasets. The red line corresponds to proposed L2+L0 method.

Datasets	L2+L0			D-L0			L1			FSV			LSSVM		
	Err	Time	r	Err	Time	r	Err	Time	r	Err	Time	r	Err	Time	r
BC	0.02 ± 0.01	0.01 ± 0.0	8	0.02 ± 0.01	0.01 ± 0.0	8	0.04 ± 0.01	0.11 ± 0.01	8	0.03 ± 0.01	424.3 ± 20	10	0.02 ± 0.01	0.01 ± 0.0	8
GER	**0.32 ± 0.03**	0.01 ± 0.01	**7**	0.33 ± 0.03	0.02 ± 0.01	9	0.36 ± 0.01	0.15 ± 0.02	16	0.33 ± 0.02	1040 ± 13.0	16	0.29 ± 0.02	0.01 ± 0.01	16
Mus	0.25 ± 0.04	0.04 ± 0.0	**16**	0.19 ± 0.06	0.04 ± 0.0	166	0.28 ± 0.03	0.14 ± 0.02	166	0.18 ± 0.03	10.5 ± 0.9	166	**0.17 ± 0.03**	0.03 ± 0.0	166
Son	**0.24 ± 0.05**	0.01 ± 0.0	**4**	0.24 ± 0.14	0.01 ± 0.0	32	0.25 ± 0.06	0.12 ± 0.02	60	0.26 ± 0.05	1.44 ± 0.11	32	0.28 ± 0.07	0.01 ± 0.0	60
Tit	**0.21 ± 0.02**	0.0 ± 0.0	3	0.22 ± 0.02	0.0 ± 0.0	3	0.22 ± 0.0	0.12 ± 0.03	3	-	-	-	0.22 ± 0.02	0.0 ± 0.0	3
TN	0.02 ± 0.0	0.02 ± 0.0	20	0.02 ± 0.0	0.02 ± 0.0	20	0.02 ± 0.0	0.18 ± 0.03	20	-	-	-	0.02 ± 0.0	0.02 ± 0.0	20

Table 2: Performance comparison over 6 datasets in the primal

sults which are unique and correspond to best performance and least number of features used. We also infer that the FSV method is computationally quite expensive and is infeasible for datasets like Tit and TN. Hence in Table 2 the results aligning to the FSV method for these datasets are represented by '-'.

4. CONCLUSION

In this paper we proposed a combination of L_2-norm penalty and the convex relaxation of the L_0-norm penalty for feature selection in classification problems. The proposed method was formulated in a primal-dual framework by iteratively solving a system of linear equations. It is computationally easier than standard QP-based SVM solvers. The L_2-norm penalty helped in robustly selecting variables during each randomization whereas the L_0-norm penalty reduced the noisy feature coefficients to zero. We demonstrated the efficiency of the proposed approach on 10 real world datasets and evaluated it against several state-of-the-art feature selection based SVM classifiers.

Acknowledgments

This work was supported by Research Council KUL, ERC AdG A-DATADRIVE-B, GOA/10/09MaNet, CoE EF/05/006, FWO G.0588.09, G.0377.12, SBO POM, IUAP P6/04 DYSCO, COST intelliCIS and by Qatar Computing Research Institute.

5. REFERENCES

[1] Suykens, J.A.K., Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, **1999**, *9(3)*, 293-300.

[2] Vapnik, V. The Nature of Statistical Learning Theory. *Springer-Verlag*, **1995**, New York.

[3] Neto, A.R.R, Barreto, G.A. A novel heuristic for Building Reduced-Set SVMs using the Self-Organizing Map. *Advances in Computational Intelligence*, **2001**, *6691*, 97-104.

[4] Weston, J., Elisseeff, A. Schölkopf, Tipping, M. Use of Zero Norm with Linear and Kernel Methods. *Journal of Machine Learning Research*, **2003**, *3*, 1439-1461.

[5] Candes, E.J., Wakin, M.B., Boyd, S. Enhancing sparsity by Reweighted L_1 minimization. *Journal of Fourier Analysis and Applications, special issue on Sparsity*,**2008**, *14(5)*, 877-905.

[6] Huang, K., Zheng, D., Sun, J., Hotta, Y., Fujimoto, K., Naoi, S. Sparse Learning for Support Vector Classification. *Pattern Recognition Letters*, **2010**, *31(13)*, 1944-1951.

[7] Lopez, J., De Brabanter, K., Suykens, J.A.K. Sparse LSSVMs with L_0 minimization. *In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*,**2011**, 189-194.

[8] Amaldi, E., Kann, V. On the approximability of minimizing non zero variables or unsatisfied relations in linear systems. *Theoritical Computer Science*, **1998**, *209*, 237-260.

[9] Huang, K., King, I., Lyu, M.R. Direct Zero-norm Optimization for Feature Selection. *In Proceedings of the* 8^{th} *International Conference on Data Mining*, *2008*, 845-850.

[10] Bousquet, O., Elisseeff, A. Stability and Generalization. *Journal of Machine Learning Research*, **2002**, *2*, 499-526.

[11] Li, Y., Campbell, C., Tipping, M. Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, **2002**, *18(10)*, 1332-1339.

[12] Ming, Y., Lin, Y. Model selection and Estimation in Regression with Grouped Variables. *Journal of Royal Statistical Society*, **2006**, *68*, 49-67.

[13] Friedman, J., Hastie, T., Tibshirani, R. A note on group lasso and a sparse group lasso. *arXiv:1001.0736*, **2010**.

[14] Hastie, T., Tibshirani, R., Friedman, J. Elements of Statistical Learning. *Springer series in Statistics*, **2001**.

[15] Bradley, P., Mangasarian, O. Feature Selection via concave minimization and support vector machines. *Machine Learning Proceedings of the Fifteenth International Conference on Machine Learning*, **1998**, 82-90.

[16] Song, M., Breneman, C., Bi, J., Sukumar, N., Bennett,K., Cramer, S., Tugcu, N. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of Chemical Information and Computer Sciences*, **2002**, *42(6)*, 1347-1357.

[17] Zhu, J., Rosset, S., Hastie, T., Tibshirani, R. 1-norm SVMs. *Neural Information Processing Systems*, **2003**, *16*.

[18] Zou, H., Hastie, T. Regularization and variable selection via the elastic net. *Journal of Royal Statistics*, **2005**, *67*, 301-320.

[19] Zou, H. The Adaptive Lasso and its Oracle Properties. *Journal of American Statistical Association*, **2006**, *101(476)*, 1418-1429.

[20] Zou, H., Zhang, H.H. On the adaptive elastic net with diverging number of parameters. *Annals of Statistics*, **2009**, *37(4)*, 1733-1751.

[21] Hoerl, A.E., Kennard, R.W. Ridge Regression: Biased Estimation for Non-othrogonal Problems. *Technometrics*, **1970**, *12(1)*, 55-67.

[22] Xavier de Souza, S., Suykens, J.A.K., Vandewalle, J., Bolle, D. Coupled Simulated Annealing for Continuous Global Optimization. *IEEE Transactions on Systems, Man and Cybertics*, **2010**, *40(2)*, 320-335.

[23] Guyon, I., Weston, J., Barnhill, S., Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, **2002**, 389-422.

[24] Blake, C.L., Merz, C.J. Repository of machine learning databases. *University of California*, **1998**, `http://www.ics.uci.edu/~mlearn/mlrepository.html`.