Feature selection for multiclass support vector machines

Fatima Z Aazi, Rafik Abdesselam, Boujemâa Achchab, Abdeljalil Elouardighi

Feature Selection for Multiclass Support Vector Machines

Fatima Zahra Aazi*, Rafik Abdesselam, Boujemâa Achchab, Abdeljalil Elouardighi

LM²CE Laboratory, FSJES, Hassan 1st University, Settat, Morocco, and ERIC Laboratory, Lumière Lyon 2 University, Lyon, France. faazi@mail.univ-lyon2.fr
COACTIS Laboratory, ISH, Lumière Lyon 2 University, Lyon, France. Rafik.Abdesselam@univ-lyon2.fr
LM²CE, LAMSAD Laboratories, EST Berrechid, Univ. Hassan 1, Morocco. achchab@estb.ac.ma
LM²CE Laboratory, FSJES, Univ. Hassan 1, Settat, Morocco. jalilardighi@yahoo.fr

***Abstract* — In this paper, we present and evaluate a novel method for feature selection for multiclass support vector machines (MSVM). It consists in determining the relevant features using an upper bound of generalization error proper to multiclass case called the multiclass radius margin bound.**
A score derived from this bound will rank the variables in order of relevance, then, forward method will be used to select the optimal subset. The experiments are firstly conducted on simulated data to test the ability of the proposed method to find the relevant variables in the case where some variables are relevant for all classes, when they are relevant only for some classes and when the dimensions of data are very high. Afterward, two real cancer datasets publicly available will be used and some results will be compared with those of other methods of variable selection by MSVM.

Keywords- Classification; multiclass support vectors machines (MSVM); feature selection; multiclass radius-margin bound.

I. INTRODUCTION

In the problems related to genes expression profiles or text classification, the number of variables is usually very high compared to the number of observations, the importance of variables selection is justified by the possibility of existence of correlated, noise and / or redundant variables which give significant error rates. In this case, the variables selection essentially allows to improve the performances of forecasting or classification models by using only the variables that are important for the studied problem, reduce time and cost calculation and facilitate the understanding of the process generating information.

In the context of SVM, binary or multiclass, the developed models do not allow an automatic selection of variables and use all available ones.

In binary case, several approaches were been proposed to show the possibility of variable selection with SVM, these approaches can be grouped into two categories. The first consists in modifying the optimization program of SVM, so as to integrate the selection in the classification process. The second derives criteria from SVM to do selection.

Within the first category, several new forms of SVM were been proposed, the L_0SVM [1], L_1SVM [2,3], combination of L_0 and L_1 SVMs [4] and the infinite norm SVM [5] are examples of these forms. Similarly, by deriving criteria from SVM, various approaches were presented, including the recursive feature elimination algorithm SVM-RFE of Guyon et al. [6] using the margin as selection criterion and Rakotomamonjy's approach [7], considered as extension of SVM-RFE, using the upper bounds of generalization error specifics to SVM.

In multi-class case, as extension of the approaches of the first category, Wang and Shen [8,9] replaced the L_2-penalty in MSVM model of Lee et al. ($MSVM_{LLW}$) [10] by L_1-penalty (L_1MSVM). Similarly, Zhang et al. [11] proposed a sup norm penalty which is more efficient and easier to implement than that given by the L_1MSVM solution. Other methods were also been proposed in this context [12,13].

Moreover, and as extension of SVM-RFE, several techniques were presented, based either on a decomposition method, selecting variables for each pair of classes and then extend the results to multiclass case [14,15,16] or on a direct approach, considering all classes simultaneously [17,18].

However, in spite of the significant number of proposed extensions to multiclass case and their good performances compared to some existing techniques, no method is best or optimal [18,19] and the issue is still relevant. For this reason, and in order to contribute in this framework, we propose this article.

Indeed, studying the various extensions, we note that although the theoretical bases and good performances of Rakotomamonjy's approach [7] in selecting variables in binary case, no study, to our knowledge, has used an upper bound of generalization error *proper* to multiclass case to select the optimal subset of variables.

In this article, we propose a new method for ranking and selecting relevant variables in multiclass case, based on the upper bound of generalization error called radius margin bound (RM) [20]. This bound is specific to multiclass case and only applicable to hard margin $MSVM_{LLW}$ model [10] i.e. without training error and to $MSVM^2$ model of Guermeur et al. [20]. In this work we will use the first model.

The multiclass RM bound is presented as extension of the binary radius margin bound [21] while taking into account the characteristics of multiclass case. It was proposed by Guermeur et al. [20] for model selection. The contribution of this paper is to use it for model and variables selection.

The proposed method consists of three steps: firstly, and since we work with $MSVM_{LLW}$ model, we choose it's optimal parameters minimizing the multiclass RM bound in presence of all variables (model selection), then, we classify variables in order of relevance, and finally, proceeding by forward method, we choose the optimal subset minimizing the testing error, calculated on a test sample or by cross-validation.

The rest of the paper is organized as follow: section 2 presents the $MSVM_{LLW}$ model and the multiclass RM upper bound of generalization error. The proposed procedure for variable ranking and optimal subset' selection is given in section 3.

The data used, results of experiments and comparisons are presented in Section 4, followed by a general conclusion.

II. $MSVM_{LLW}$ MODEL AND RM BOUND

In the framework of multiclass SVM (MSVM), we are interested in q categories classification problems ($2 < q < \infty$). The goal is to estimate q decision function $f_k(x)$ and classify observations according to the classification rule:

$$\Phi_f(x) = \arg\max f_k(x); \qquad k=1, 2,\ldots,q.$$

The estimation of the decision functions is done using a set of pairs of independent and identically distributed observations $\{(x_i,y_i),\ i=1,\ldots,n\}$ called training set, where x is the description of an object belonging to the descriptions space X described by 'p' variables and Y the set of categories 'y' identified by their indices [1, q].

In this work, we will test the performances of the multiclass RM bound to perform variables selection for a hard margin $MSVM_{LLW}$ model. This section will briefly present the properties of this model and describes the RM bound.

A. *The $MSVM_{LLW}$ Model*

As all direct approaches [20, 22, 23], the $MSVM_{LLW}$ model solves the multiclass problem directly without decomposition, estimating 'q' decision functions simultaneously by solving one optimization program. It is considered as the most theoretically based of MSVM models as is the only one that implements asymptotically the Bayes decision rule [10].

The optimization problem is to solve, subject to the constraint $\sum_{k=1}^{q} f_k = 0$, the objective function of the form:

$$\min_f \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q} I(y_i \neq k)\left[f_k(x_i)+1\right]_+ + \lambda\sum_{k=1}^{q}\sum_{j=1}^{p} w_{kj}^2. \quad (1)$$

- The first term $I(y_i \neq k)\left[f_k(x_i)+1\right]_+$ represents the loss function, which measures the difference between estimations and reality. This term can also be written as $C\sum_{i=1}^{n}\sum_{k\neq y_j}\xi_{ik}$, with ξ_{ik} the slack variables and C the weight of these variables.

- The second term $\lambda\sum_{k=1}^{q}\sum_{j=1}^{p} w_{kj}^2$ ($\lambda \in R$ determined by cross-validation), measures the ability or the complexity of the hypothesis space, and also equal to the inverse of k separators' margins to maximize.

- $f_k(x) = \langle w_k, \Phi(x_i)\rangle + b_k$, $1 \le i \le n$, with
  - (w_k, b_k) the parameters of k^{th} separator to estimate.
  - $\Phi(x_i)$ the nonlinear transformation of x_i from original to feature space if data are not linearly separable. If not, $\Phi(x_i) = x_i$.

Problem solving is done using the Lagrangian, and the nonlinear transformation of data will be replaced by a kernel function.

B. *Multiclass Radius Margin Bound*

The multiclass RM upper bound of the generalization error that we will use is a direct extension of the two-class radius margin bound [21]. Used for model selection, it is considered as the easiest and the most popular of generalization error's upper bounds.

Guermeur et al. [20] demonstrate that the number of errors denoted L_m, resulting from the application of leave-one-out cross-validation procedure (LOOCV) for a hard margin q-category $MSVM_{LLW}$ trained on d_m, is upper bounded as follows:

$$L_m \le \frac{(q-1)^3}{q} D_m^2 \sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^* \qquad (2)$$

- m : the size of training sample,
- q : the number of categories,
- D_m : the diameter of the smallest sphere containing the dataset in original or feature space.
- α_{ik}^* : the Lagrange parameters resulting from the resolution of the optimization program (1).

Since the value given by LOOCV is an almost unbiased estimator of the generalization error, a variable is considered as relevant according to its influence on this error by measuring its contribution to minimize the second term of (2) which is the RM bound.

III. THE PROPOSED PROCEDURE FOR VARIABLE RANKING AND OPTIMAL SUBSET' SELECTION

The multiclass RM bound is generally used for model selection; it means to choose the optimal parameters of MSVM model. These parameters to optimize are: C, representing the weight of the training errors ξ_i, and σ the parameter of the kernel function if we decide to change the data space.

Note that a large value of C means a big weight of errors and thus get closer to a hard margin learning, and, conversely, a small value reflect acceptance of errors and therefore a soft margin learning.

The idea in this article is to use the multiclass RM bound to model and variables selection by ***combining*** the procedure of model selection for MSVM proposed by Guermeur et al. and ***an extension*** of the Rakotomamonjy's method of variable selection to the multiclass case.

The proposed procedure is based on a score called zero-order score proposed for two class problems [7], whose value will rank variables in order of relevance.

The zero-order score of a variable is the value of a criterion (the RM bound in our case) when this variable is removed.

We will consider a variable as most relevant when its suppression greatly increases the value of the bound and

therefore, contributes to the minimization of the generalization error.

The RM bound (2) depends on three factors: the number of categories q, the diameter D_m of the smallest sphere containing data and Lagrange parameters α_{ik}^*.

The first element 'q' is constant and independent of the number of variables, in contrast to the two other parameters D_m and α_{ik}^*. Indeed, an object is represented by its coordinates in space, so its position changes necessarily by removing a variable and thus the diameter of the sphere. Similarly, when removing a variable, data which are inputs to estimate the model change, and therefore α_{ik}^*, model's outputs, change too. Thus, the research of relevant variables will be based on the product:

$$D_m^2 \sum_{i=1}^{m} \max_{1 \le k \le Q} \alpha_{ik}^* \qquad (3)$$

Once the order of relevance of variables established, and given that an exhaustive research of optimal subset is very complicated even impossible for high dimensional data, we proceed by forward method incorporating a variable at a time in decreasing order of relevance and we choose the subset giving the minimum error rate calculated on a test sample or by cross-validation.

The proposed procedure for ranking and selecting relevant variables, for a hard margin $MSVM_{LLW}$ model, follows the three following steps:

Step 1: Choice of the model's parameters

In this step, we choose the parameters of the hard margin $MSVM_{LLW}$ model which minimize the RM bound, and therefore the generalization error, in presence of all variables. These parameters that will be used in the next step to rank variables.

The SVM method is based on the idea of finding a linear separator in a specific space, so if data are not linearly separable, i.e. a linear separator doesn't exist in original space, we move to a called feature space by projecting data in another space of higher dimension so as to find a linear separator. This transformation of data is done using kernel functions. Thus, to choose the optimal parameters of the hard margin $MSVM_{LLW}$ model, we first work in original space using a linear kernel. In this case, there will be only the parameter C to determine.

If we are unable to work without training error or if the required time is very important, we proceed to change the data space and work with a Gaussian kernel. In this case, we have to set the values of the two parameters C and σ (the parameter of the Gaussian kernel).

Step 2: Variables ranking

In this step, we rank variables in order of relevance according to the values of their zero-order scores. For this, we re-estimate, removing each time a variable, the $MSVM_{LLW}$ model and we compute the value of the product

$$D_m^2 \sum_{i=1}^{m} \max_{1 \le k \le Q} \alpha_{ik}^*$$

The variables with the highest scores are the most relevant given that their suppression increases the value of the multiclass RM bound and thus the value of the generalization error.

Step 3: Choice of the optimal subset of variables

After obtaining the order of relevance of variables, the last step is to choose the optimal subset. For this, we construct, by forward method, a sequence of models. The first one contains the first relevant variable, the second one contains the first two relevant variables and so on until all variables are integrated in decreasing order of relevance. Then we calculate the testing error rates. The model giving the minimum error rate is chosen as the best model with the optimal number of variables.

Note that for the first two stages, model selection and variables ranking, we must work without training error, as these two steps are based on the RM bound which is applicable to a hard margin $MSVM_{LLW}$ model. By contrast, it is not mandatory to do so in step 3, because we no longer use the RM bound, so we do simulations with combinations of values of C and σ until we find the values that minimize the testing error. Also, we insist on the idea that the biggest contribution of this article is in giving the order of relevance of variables which was not been done on multiclass case with direct approaches of MSVM before, this means that we can change the third step and use another method to select the optimal subset from the order given in the second step, here we use forward method but backward method or other procedures can also be used.

IV. RESULTS OF EXPERIMENTS AND COMPARISONS

In this section, we present the tests showing the ability of the score based on the RM bound to rank the variables and, therefore, to select the optimal subset. Seven datasets are considered, including five simulated databases and two real sets.

For all data sets, several simulations are conducted to find the parameters of the $MSVM_{LLW}$ model minimizing the RM bound (step 1) and to select the optimal subset (step 3).

Simulations and results have been obtained using the MSVMpack of Lauer et al. [24] allowing to train the $MSVM_{LLW}$ model and giving the parameters α_{ik}^* as output.

The diameter D_m of the smallest sphere containing data has been calculated using the hard margin SVDD algorithm [25].

A. *Simulated Data*

The used data are linearly separable in original or features space. For each case, n_1 observations are generated as training set and n_2 observations as testing set (if needed). Each observation is described by p variables (x^1, x^2, ..., x^p) with 2 are relevant and the others are noise variables.

The 2 relevant variables are generated from a mixture Gaussian. The remaining variables are independent and identically distributed generated from N(0 ,1).

1) Example 1 : Relevance for all classes

The data of this first example are those described by Zhang et al. [11], with n_1= 250 observations, n_2= 50000, q= 5 equally weighted classes and p=10 variables with (x^3, x^4,…, x^{10}) are 8 noise variables and (x^1, x^2) are relevant for all classes generated independently from $N(\mu_k , \sigma^2 I_2)$, with $\sigma = \sqrt{2}$ as follows:

$$\mu_k = 2(\cos([2k-1]\pi/q), \sin([2k-1]\pi/q)), \quad (3)$$

for each class k (k=1,2,..q).

To estimate the parameters of the hard margin $MSVM_{LLW}$ model in the presence of all variables, we first worked with a linear kernel. The results show that this kernel did not allow to train the model without error. We then tried a Gaussian kernel which gave a zero training error.

The model estimation using a Gaussian kernel requires setting the values of the parameters C and σ. For C, high values are used in order to penalize errors and therefore obtain a hard margin model. Simulations showed that the value C=1000 allows to work without error for different values of σ.

To set the value of σ, we conducted several simulations to select the value that, keeping zero training error, minimizes the generalization error via its upper bound

$$RM = \frac{(q-1)^3}{q} D_m^2 \sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$$

The term $(q-1)^3/q$ being constant, we choose the value that minimizes the product $D_m^2 \sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$.

The simulations' results are described in Table 1.

Table 1. Values of $D_m^2 \sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$ in terms of the values of σ

σ	$\sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$	D_m	$D_m^2 \sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$
0.5	312.4004	15.805	78 037.995
1.0	315.7785	14.027	62 131.820
1.5	413.478	9.180	34 841.310
2.0	771.132	5.984	**27 609.610**
2.5	1862.703	4.214	33 081.605
3.0	4315.325	3.220	44 741.290
4.0	19595.13	2.223	96 799.942
5.0	63546.20	1.780	201 390.617
6.0	114203.68	1.542	271 576.351
7.0	144505.89	1.398	282 480.114

The minimum of the upper bound is reached for σ = 2. The model will therefore be estimated with C=1000 and σ = 2.

The second step is to test the ability of the zero-order score to give the order of relevance of the variables.

To do this, we calculated, each time removing a variable, the $\sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$ and the diameter D_m of the smallest sphere enclosing data in the feature space, using the parameters chosen in step 1. The results are reported in table 2.

Table 2. Zero-order score of the 10 variables

Removed Variable	$\sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$	D_m	**Zero-order scores**
x^1	2010.701	4.841	**47 116.914**
x^2	1814.651	4.854	**42 759.475**
x^3	1158.179	5.580	36 063.672
x^4	1192.931	5.571	37 026.283
x^5	1137.751	5.504	34 470.632
x^6	1113.298	5.564	34 468.209
x^7	1173.368	5.504	35 543.780
x^8	1023.852	5.504	31 013.183
x^9	1070.409	5.554	33 023.519
x^{10}	1140.087	5.633	36 181.583

The most relevant variable is the one that maximizes the value of the zero-order score which is equal to the value of the product $D_m^2 \sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^*$ when the variable is removed.

The order of relevance of variables obtained according to table 2 is as follows:

$$x^1, x^2, x^4, x^{10}, x^3, x^7, x^5, x^6, x^9, x^8$$

The proposed score has successfully classified the first two variables which are most relevant in the two first ranges.

After ranking variables, we have estimated models to select the one that minimizes testing error and gives optimal sub-set of variables. For this, we built 10 databases (the first contains the first relevant variable, the second contains the first two relevant variables,…).

Training and testing errors obtained according to the used variables on a test sample of 50000 observations with C = 10 and σ = 2 are shown in Table 3.

Table 3. Training and testing errors

Used Variables	**Training error %**	**Testing error %**
(x^1)	56.80	61.18
(x^1, x^2)	34.00	**40.10**
(x^1, x^2, x^4)	29.60	41.33
(x^1, x^2, x^4, x^{10})	24.80	42.88
(x^1, x^2, x^4, x^{10}, x^3)	17.60	44.78
(x^1, x^2, x^4, x^{10}, x^3, x^7)	10.80	46.64
(x^1, x^2, x^4, x^{10}, x^3, x^7, x^5)	4.00	48.26
(x^1, x^2, x^4, x^{10}, x^3, x^7, x^5, x^6)	1.60	49.75
(x^1, x^2, x^4, x^{10}, x^3, x^7, x^5, x^6, x^9)	0.00	51.48
(x^1, x^2, x^4, x^{10}, x^3, x^7, x^5, x^6, x^9, x^8)	0.00	50.39

The minimum testing error 40.10% obtained using the subset of the first two relevant variables is lower than that obtained using all variables (50.39%).
Furthermore, Zhang et al. [11] have compared the results of 5 MSVM methods with feature selection, on the first data set used above, in terms of testing errors rates. The best testing error obtained, was 45.3% using the Supnorm method [11]. With our procedure, we obtained a better error rate: 40.1% using the first two relevant variables.

2) Example 2 : Relevance for some classes

In the simulation example in section 1), the two first variables are relevant for all classes, however, in reality, some variables might be important for one class, and not for another. In this section, we will study this case and see if our score is able to identify the relevant variables.
For this, we generate a second dataset with the same characteristics as the first one (n_1= 250 observations, p= 10 variables and q=5 equally weighted classes) except that the relevant variables x^2 and x^3 are as follow: x^2 is relevant only for the classes 2 and 4, x^3 is relevant only for the classes 1, 3, 5 and x^3 is more relevant than x^2 as its important for 3 classes. The 8 remaining variables are noise ones.
The numerous simulations (as done for the first example in table 1) have allowed to choose the type of kernel (Gaussian) and to set the values of parameters C and σ (C = 1000 and σ = 2) that allow to work without training error, require a reduced calculation time and minimize the value of the RM bound. Table 4 presents the order of relevance of variables according to the values of zero-order score and reveals that the score has successfully classified x^3 in the first position and x^2 in second position.

Table 4. Ranking of variables according to their zero-order scores

Removed Variable	$\sum_{i=1}^{m} \max_{1\le k\le Q} \alpha_{ik}^{*}$	$\mathbf{D_m}$	**Zero-order scores**	**Ranking**
x^1	1132.2971	5.326	32 129.899	8
x^2	**1672.0327**	**4.821**	**38 873.534**	**2**
x^3	**1991.1498**	**4.782**	**45 548.769**	**1**
x^4	1097.9851	5.434	32 428.987	7
x^5	1195.6819	5.349	34 214.171	3
x^6	1165.3138	5.381	33 752.091	5
x^7	1147.0951	5.414	33 629.598	6
x^8	1181.6572	5.377	34 176.575	4
x^9	1166.5453	5.359	30 639.041	10
x^{10}	1090.8932	5.371	31 466.781	9

3) Example 3 : Relevance in high dimensions

After having successfully ranked the 2 relevant variables in presence of 8 noise ones in sections 1 and 2, we proceed, here, to measure the effect of increasing the number of irrelevant variables on the performances of the score.
For this, we generate three datasets with 500, 1000 and 5000 variables, 250 observations, and 2 relevant variables (x^1, x^2). With x^1 is relevant only for the classes 1, 3, 5 and x^2 is relevant only for the classes 2 and 4.
To set the values of C and σ, we conduct several simulations using 3 values of C: 10000, 1000 and 100 and several values of σ to choose the best combination.
The results of variables ranking based on the values of their zero-order scores are presented in Figure 1.
According to Figure 1, the maximal values of zero-order score are those of the first two variables that we know they are the most relevant. Thus, the proposed score was been able to classify the first two variables in the first two positions in large dimensions. However, we observe that the value of σ the parameter of the Gaussian kernel increases while increasing the number of variables (σ = 24 for 500 variables, σ = 35 for 1000 variables and σ = 83 for 5000 variables). As mentioned in [26], large values of σ gradually reduce the kernel to a constant function, making it impossible to learn any non-trivial classifier. That means that in very large dimensions, we will need very large values of σ and thus the kernel will be no more able to learn correctly the MSVM model and thus we will no more have the correct ranking in order of relevance of variables.

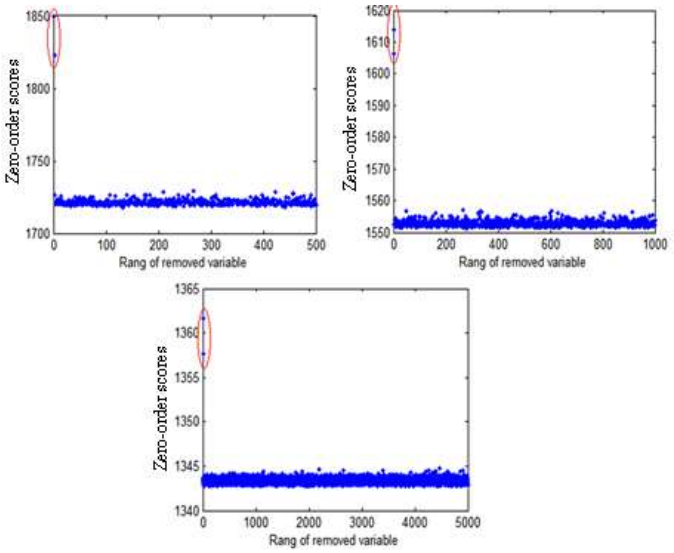

Figure 1. Zero order scores of variables

B. Real Data

1) Example 1

In this example, we apply our selection procedure to the real data set 'lung cancer' which is composed of 56 variables, 32 observations and 3 classes. These data, available at UCI repository, were used in the context of variable selection by multiclass SVM by LI et al. [27].
Using their proposed method, Li et al obtained, an average testing error rate of 45.8%, while, the application of Optimal Brain Damage method on this dataset gave an average testing error rate of 44.15%.
Our main goals here are: firstly, show the performances of our procedure to select the optimal subset and give a better testing rate relative to the case where all variables are conserved, and

secondly, compare the obtained testing error rate to those of the two methods described above.
For experiments, we use the 32 observations as a training set and we calculate testing error rates by LOOCV.
The linear kernel was able to give zero error rate, the optimal value of the parameter C is C = 1000.
We rank variables in order of relevance, then, we build 56 databases to select the optimal model. For each dataset, in order to minimize error rates, we try different values of C. Figure 2 shows the results of estimation of testing error rates by LOOVC according to values of C and the number of used variables in decreasing order of relevance.
The best testing error rate is 12.5% obtained using the first 10 variables in order of relevance with C=1.5. This rate is much better than that obtained in presence of all variables i.e. 56.25%. So we can confirm the effectiveness of the proposed procedure for ranking and selecting the optimal subset.
Furthermore, comparing this result to those of two variables selection methods by MSVM described before, we find that the achieved rate of 12.5% is much better.

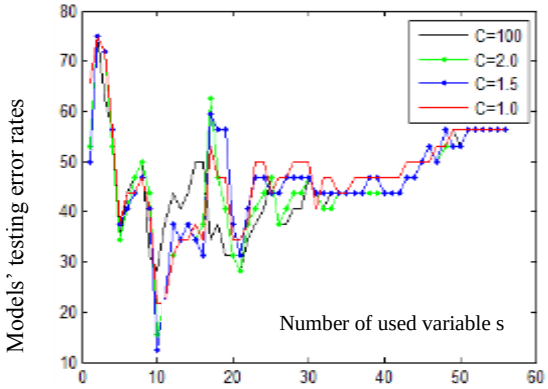

Figure 2. Testing error rates of the 56 models

2) Example 2

The second example on which we test the performances of our approach is the children cancer data set classifying the small round blue cell tumors (SRBCTs) into four classes, namely neuroblastoma (NB), rhabdomyosarcoma (RMS), non-odgkin lymphoma (NHL), et Ewing (EWS) using cDNA gene expression profiles.
(http://research.nhgri.nih.gov/microarray/Supplement/).
The data set includes 83 observations and 2308 variables. It was presented for the first time by Khan et al. [28] and has been used in the context of variable selection by multiclass SVM by Zhang et al. [11].
After model selection, we proceed to rank the 2308 features in order of relevance.
To estimate the testing error rates, we used 10 combinations of C and σ (C: 100, 10, 1 and σ: 1, 2, 3, 2.5) to select the best combination(s) giving the minimum testing error rate.
The results of the two best simulations in terms of testing error rates obtained by LOOCV according to the values of C, σ and the number of used variables in decreasing order of relevance are presented in Figure 3.
From Figure 3, we note that the proposed approach based on RM bound gave a zero error rate for the two simulations using the first 9 variables.
Comparing these results with those of previous studies on this dataset, we can see that they are far better in terms of the number of variables needed. Indeed, selecting variables for multiclass SVM using adaptive sup-norm regularization, Zhang et al. [11] obtained zero error rate, on a test sample, using 47 variables and using the L1 norm, 62 variables were required. Furthermore, using neural networks, getting a zero error rate was possible but using the first 96 genes [28].

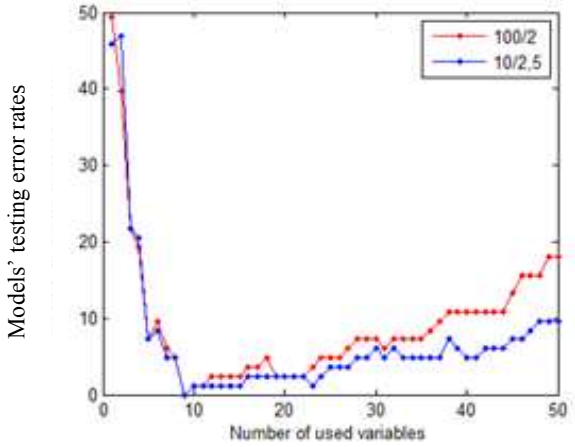

Figure 3. Testing error rates of the first 50 models

V. CONCLUSION AND PERSPECTIVES

The results of studies on variable selection by multiclass SVM show the effectiveness of using this technique to reduce dimensions and improve classification's accuracy.
In this paper, we proposed a new approach, based on the multiclass radius margin upper bound of generalization error, to give the order of relevance of variables and the optimal subset. As a result, the proposed method gives the correct order of relevance of variables for the simulated data, and significantly reduces the error rate for all used data sets.
In fact, one of the advantages of our method is that it uses the $MSVM_{LLW}$ model which is the most theoretically based of MSVM models and as a wrapper approach, selecting variables after the estimation of the model, takes into account the influence of each variable on the performances of the estimated model.
A constraint for the application of our procedure in very large dimensions consists in required computation time, which is important given the need to re-estimate the MSVM model for each variable in order to calculate the zero order scores, combined with the degradation of its performances given the need of high values of σ as explained in section III.3. To deal with these two problems, we propose to combine our approach with the MRMR filter method (Minimum Redundancy Maximum Relevance) which will filter a big number of noise variables before applying our approach to the selected

features. The results obtained on three high dimensional cancer datasets were very good in terms of the obtained testing error rate. It will be the extension of this work.

V. REFERENCES

[1] J. Weston, A. Elisseeff, B. Schlkopf and P. Kaelbling. "Use of the zero-norm with linear models and kernel methods". Journal of Machine Learning Research, 3:pp. 1439-1461. 2003.

[2] P.S. Bradley and O.L. Mangasarian. "Feature selection via concave minimization and support vector machines". Machine Learning Proceedings of the Fifteenth International Conference (ICML 98), pp. 82-90. Morgan Kaufmann. 1998.

[3] J. Zhu, S. Rosset, T. Hastie and R. Tibshirani. "1-norm support vector machines". Neural Information Processing Systems. MIT Press. 2003.

[4] Y. Liu and Y. Wu. "Variable selection via a combination of the L_0 and L_1 penalties". Journal of Computation and Graphical Statistics. 2007.

[5] H. Zou and M. Yuan. "The f∞-norm support vector machine". Statistica Sinica. 2008.

[6] I. Guyon, J.Weston, S. Barnhill, and V. N. Vapnik. "Gene selection for cancer classification using support vector machines". Machine Learning, 46(1-3) : pp. 389-422, 2002.

[7] A. Rakotomamonjy, "Variable Selection Using SVM-based Criteria". Journal of Machine Learning Research vol. 3, pp. 1357-1370. 2003.

[8] L. Wang and X. Shen, "Multi-category support vector machines, feature selection and solution path". Statistica Sinica 16. pp. 617-633. 2006.

[9] L. Wang and X. Shen, "On l1-norm multi-class support vector machines: methodology and theory". Journal of the American Statistical Association. pp.583–594. 2007.

[10] Y. Lee, Y. Lin, and G. Wahba. "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data". Journal of the American Statistical Association, 99(465):pp. 67-81, 2004.

[11] H.H. Zhang, Y.Liu, Y.Wu and J. Zhu, "Variable selection for the multicategory SVM via adaptive sup-norm regularization". Electronic Journal of Statistics, Vol. 2, pp. 149–167. 2008.

[12] J. Guo, "Class-specific Variable Selection for Multicategory Support Vector Machines", Statistics and its interface. 2011.

[13] J-T. LI and Y-M. JIA, "Huberized Multiclass Support Vector Machine for Microarray Classification". Acta Automatica Sinica, Vol. 36, No. 3. 2010.

[14] M-D. Shieh and C-C. Yang, "Multiclass SVM-RFE for product form feature selection", Expert Systems with Applications, vol 35. pp. 531–541. 2008.

[15] X-W. Chen , X. Zeng, and D.V. Alphen, "Multi-class feature selection for texture classification". Pattern Recognition Letters, 27. pp. 1685–1691. 2006.

[16] Y. Mao, X. Zhou, D. Pi, Y. Sun, and S.T.C. Wong, "Multiclass Cancer Classification by Using Fuzzy Support Vector Machine and Binary Decision Tree With Gene Selection". Journal of Biomedicine and Biotechnology. 2. pp 160–171. 2005.

[17]O. Chapelle and S. Keerthiy. "Multi-Class Feature Selection with Support Vector Machines". 2008.

[18] X. Zhou and D.P. Tuck. "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data". Bioinformatics Vol. 23 No. 9, pp. 1106–1114. 2007.

[19] P.M. Granitto and A. Burgos, "Feature selection on wide multiclass problems using OVA-RFE". Inteligencia Artificial vol 44. pp 27-34. 2009.

[20] Y. Guermeur and E. Monfrini, "A Quadratic Loss Multi-Class SVM for which a Radius–Margin Bound Applies". Informatica, Vol. 22, No. 1, pp.73–96. 2011.

[21] V.N. Vapnik. "The Nature of Statistical Learning Theory". Springer-Verlag, New York, 1995.

[22] J. Weston and C. Watkins. "Multi-class support vector machines". Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science. 1998.

[23] K. Crammer and Y. Singer. "On the algorithmic implementation of multiclass kernel based vector machines". Journal of Machine Learning Research, 2: pp. 265_292, 2001.

[24] F. Lauer and Y. Guermeur, "MSVMpack: a Multi-Class Support Vector Machine package". Journal of Machine Learning Research vol. 12, pp. 2293-2296. 2011.

[25] D M.J. Tax and R. P.W. Duin, "Support Vector Data Description", Machine Learning, 54, pp. 45–66, 2004

[26] J. Shawe-Taylor, N. Cristianini: Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.

[27] G-Z Li, J. Yang, G-P Liu and L. Xue. "Feature Selection for Multi-Class Problems Using Support Vector machines". Lecture Notes on Artificial Intelligence. 3173 (PRICAI2004), pp. 292-300. Springer. 2004.

[28] J. Khan, J.S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson and P. S. Meltzer. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks". Nature Medicine, 7. pp. 673-679. 2001.