AI Communications 29 (2016) 583–593
DOI 10.3233/AIC-160707
IOS Press

Feature selection for multiclass support vector machines

F.Z. Aazi [a,b] R. Abdesselam [c], B. Achchab [a,*] and A. Elouardighi [d]
[a] *LAMSAD Laboratory, EST Berrechid, Hassan 1st University, Morocco*
[b] *ERIC Laboratory, Lumière Lyon 2 University, France*
[c] *COACTIS Laboratory, ISH, Lumière Lyon 2 University, Lyon, France*
[d] *LM2CE Laboratory, FSJES, Hassan 1st University, Settat, Morocco*

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 the 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 score to give the correct order of relevance of variables and the ability of the proposed method to find the subset giving a better error rate than the case where all features are used. Afterward, four real datasets publicly available will be used and the results will be compared with those of other methods of variable selection by MSVM.

Keywords: Discrimination, Multiclass Support Vectors Machines (MSVM), variables selection, hard margin MSVM models, multiclass radius-margin bound

1. Introduction

In a classification problem, the relevant variables are not known a priori. The importance of selection is justified by the possibility of existence of correlated, noise and/or redundant variables which usually give significant error rates [34,36]. Indeed, the variables selection essentially allows to improve the performances of classification models by using only the variables that are important for the studied problem, reduce time and cost of calculation and facilitate understanding of the process generating information.

These advantages are to be exploited for large dimensions problems and especially when the number of variables is very large compared to the number of observations. This is the case, for example of the problems related to DNA microarray gene expression profiles, where genes are considered as variables and the number of observations is generally low for cost reasons. Indeed, some studies suggest that only a small number of genes is sufficient [22,35] and for a binary classification problem, 50 genes are usually sufficient [7].

*Corresponding author. E-mail: achchab@estb.ac.ma.

There are generally three categories of methods for variables selection [2,12,17]: Filter, Wrapper and Embedded. In the first category, the selection is made a priori before the estimation of forecasting model, it consists in testing each variable independently of others and then order them according to a given criterion. The Wrapper methods select variables after developing model and thus take into account the influence of variables on the performances of the model. The last category (Embedded) incorporates the selection of variables during the learning process.

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 proposed to show the possibility of variable selection with SVM. These approaches can be grouped into two categories. The first one, containing embedded methods, consists in modifying the optimization program of SVM, so as to integrate the selection in the classification process. The second one, derives criteria from SVM to do selection (wrapper approaches).

Within the first category, several new forms of SVM were been proposed, the L_0SVM [32], L_1SVM [3,39], combination of L_0 and L_1 SVMs [23] and the F_∞-norm SVM [40] are examples of these forms. Sim-

ilarly, by deriving criteria from SVM, various approaches were presented, including the Recursive Feature Elimination algorithm SVM-RFE of Guyon et al. [13] using the margin as a selection criterion and Rakotomamonjy's approach [26], considered as extension of SVM-RFE, using the upper bounds of the generalization error specifics to SVM.

In the multiclass case, as extension of the approaches of the first category, Wang and Shen [30,31] replaced the L_2-penalty in the MSVM model of Lee et al. ($MSVM_{\mathrm{LLW}}$) [19] by the L_1-penalty (L_1MSVM). Similarly, Zhang et al. [37] 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 [11,21].

Moreover, and as extension of SVM-RFE, several techniques were presented, based either on a decomposition method, selecting the variables for each pair of classes and then extend the results to the multiclass case [5,24,27] or on a direct approach, considering all classes simultaneously [4,38].

However, in spite of the significant number of the proposed extensions to the multiclass case and their good performance compared to some existing techniques, no method is best or optimal [8,38] and the issue is still relevant.

Studying the various extensions, we note that although the theoretical bases and the good performance of Rakotomamonjy's approach [26] in selecting the relevant attributes in the binary case, no study, to our knowledge, has used an upper bound of the generalization error proper to the multiclass case to select the optimal subset of variables.

In this paper, we propose a new method for ranking and selecting the relevant variables in the multiclass case (assigning one class to each example), based on the upper bound of the generalization error called the Radius Margin bound (RM) [10]. This bound is specific to the multiclass case and only applicable to the hard margin $MSVM_{\mathrm{LLW}}$ model [19] i.e., without training error and to the $MSVM^2$ model of Guermeur et al. [10].

The multiclass RM bound is presented as an extension of the binary radius margin bound [29] while taking into account the characteristics of the multiclass case. It was proposed and used by Guermeur et al. [10] for model selection. In this paper, we will use it for model and variables selection for a hard margin $MSVM_{\mathrm{LLW}}$ model.

The proposed method consists of three steps: firstly, and since we work with the $MSVM_{\mathrm{LLW}}$ model, we choose its optimal parameters that minimize the multiclass RM bound in the presence of all variables (model selection); secondly, we rank the variables in order of relevance, and finally, proceeding by forward method, we choose the optimal subset minimizing the error calculated on a validation set.

The rest of the paper is organized as follows: Section 2 presents the $MSVM_{\mathrm{LLW}}$ model and the RM upper bound of the 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 and some perspectives of this work.

2. $MSVM_{\mathrm{LLW}}$ 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 functions $f_k(x)$ and to classify the observations according to the classification rule:

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

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 is the set of categories 'y' identified by their indices $[1, q]$.

Several approaches were proposed in the context of MSVM, belonging either to the category of decomposition or direct methods [9].

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

2.1. The $MSVM_{\mathrm{LLW}}$ model

As all direct approaches [6,10,33], the $MSVM_{\mathrm{LLW}}$ 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 [19] as it is the only one that implements asymptotically the Bayes decision rule.

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

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

- The first term $I(y_i \neq k)[f_k(x_i) + 1]_+$ represents the loss function, which measures the difference between the estimation and the reality. This term can also be written as $C \sum_{i=1}^{n} \sum_{k \neq y_i} \xi_{ik}$, with ξ_{ik} are the slack variables.
- The second term $\lambda \sum_{k=1}^{q} \sum_{j=1}^{p} w_{kj}^2$ ($\lambda \in \mathbf{R}$, determined by cross-validation), measures the ability or the complexity of the hypothesis space, also equal to the inverse of the k separators' margins to maximize.
- $f_k(x) = \langle w_k, \Phi(x_i) \rangle + b_k$, $1 \leqslant i \leqslant n$, with
  - (w_k, b_k) the parameters of the kth separator to estimate.
  - $\Phi(x_i)$ the nonlinear transformation of x_i from the original space to the feature space if the data are not linearly separable. If not, $\Phi(x_i) = x_i$.

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

2.2. The 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. Used for model selection, it is considered as the easiest and the most popular of the generalization error's upper bounds.

Guermeur et al. [10] demonstrate that the number of errors denoted L_n, resulting from the application of leave-one-out cross-validation procedure for a hard margin q-category $MSVM_{\mathrm{LLW}}$ model trained on d_n, is upper bounded as follows:

$$L_n \leqslant \frac{(q-1)^3}{q} D_n^2 \sum_{i=1}^{n} \max_{1 \leqslant k \leqslant q} \alpha_{ik}^*, \quad (2)$$

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

Since the value given by leave-one-out cross validation 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.

3. The proposed procedure for variables ranking and optimal subset' selection

The multiclass RM bound is generally used for model selection; it means to choose the optimal parameters of the MSVM model. These parameters to optimize are: C, representing the weight of the training errors and σ the parameter of the Gaussian kernel.[1]

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 extend the Rakotomamonjy's method of variable selection to the multiclass case using the multiclass RM bound while selecting model using this same bound.

The proposed procedure is based on a score called zero-order score proposed for two class problems [26], whose value will rank variables. 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 of the smallest sphere containing data D_n and the 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_n 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 outputs, change too. Thus, the

[1] Considered as the best choice [15] if we decide to change the data space.

research of the relevant variables will be based on the product:

$$D_n^2 \sum_{i=1}^{n} \max_{1 \leqslant k \leqslant q} \alpha_{ik}^*. \tag{3}$$

Once the order of relevance of the variables is established, and given that an exhaustive research of the optimal subset is very complicated even impossible for high dimensional data, we proceed to choose the optimal subset using the forward method. For this, we construct a sequence of models/subsets by incorporating each time a variable in decreasing order of relevance and we choose the subset giving the error rate minimum calculated on a validation set.

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

Step 1 (Choice of the parameters of $MSVM_{\text{LLW}}$ model)**.** In this step, we choose the parameters of the hard margin $MSVM_{\text{LLW}}$ model which minimize the multiclass 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 the 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 the data is done using the kernel functions [14,29]. Thus, to choose the optimal parameters of the hard margin $MSVM_{\text{LLW}}$ model, we work, first, in the original space using a linear kernel. In this case, there will be only the value of the parameter C to determine. If we are unable to work without training error, we proceed to change the space of the data and work with a Gaussian kernel as first choice. In this case, we have to set the values of the two parameters C and σ.

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 $MSVM_{\text{LLW}}$ model and we compute the value of the product (3).

The variables with 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)**.** The last step is to choose the optimal subset. For this, we construct, using 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 we integrate all the variables in decreasing order of relevance. Then, we calculate the error rates on a validation set. 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, parameters 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 $MSVM_{\text{LLW}}$ model. By contrast, it is not mandatory to do so in Step 3, because we no longer use the RM bound. Indeed, we conduct simulations with combinations of values of C and σ until we find the values that minimize the validation error.

Also, we insist on the idea that the main contribution of this article is in giving the order of relevance of variables which was not been done on the multiclass case with direct approaches of MSVM before. That means that we can change the third step and use another method to select the optimal subset from the order given in second step, here we use forward method but backward method or other procedures can also be used.

4. Experimental results and comparisons

In this section, we present the tests showing the ability of the score based on the RM bound to rank variables and, therefore, to select the optimal subset. Six datasets are considered, including two simulated databases and four real sets. For all data sets, several simulations are conducted to find the parameters of the $MSVM_{\text{LLW}}$ model minimizing the multiclass 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. [18] allowing to train the $MSVM_{\text{LLW}}$ model and giving the parameters α_{ik}^* as output. The diameter D_n of the smallest sphere containing the data has been calculated using the hard margin Support Vector Data Description (SVDD) algorithm [1,28].

4.1. Simulated data

The used data are linearly separable in original or features space. Each observation is described by P variables $(x^1, x^2, \ldots, x^P)$ where 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 $N(0, 1)$.

4.1.1. Example 1

The data of this first example are generated as described by Zhang et al. [37], with $n_1 = 250$ observations as training set, $n_2 = 1000$ observations as validation set and $n_3 = 50000$ observations as testing set, $q = 5$ equally weighted classes (each class has the same number of observations) and $P = 10$ variables with $(x^3, x^4, \ldots, x^{10})$ are 8 noise variables and (x^1, x^2) are relevant for all classes generated from a mixture Gaussian as follow: for each class k, the two variables are generated independently from $N(\mu_k, \sigma^2 I_2)$, with $\sigma = \sqrt{2}$ and for $k = 1, 2, \ldots, q$:

$$\mu_k = 2\left(\cos\left([2k-1]\frac{\pi}{q}\right), \sin\left([2k-1]\frac{\pi}{q}\right)\right).$$

To estimate the parameters of the hard margin $MSVM_{\mathrm{LLW}}$ model in the presence of all variables (based on the training set), 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 to set 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 learning without error for different values of σ.

To set the value of σ, we have conducted several simulations to select the value that, keeping zero training error, minimizes the generalization error via its upper bound RM (2). The term $\frac{(q-1)^3}{q}$ being constant, we choose the value that minimizes the product (3). The simulations results are described in Table 1.

Table 1
The values of $D_n^2 \sum_{i=1}^n \max_{1\leqslant k\leqslant q} \alpha_{ik}^*$ in terms of the values of σ

σ	$\sum_{i=1}^n \max_{1\leqslant k\leqslant q} \alpha_{ik}^*$	D_n	$D_n^2 \sum_{i=1}^n \max_{1\leqslant k\leqslant q} \alpha_{ik}^*$
0.5	312.4516	15.8043	78042.9108
1.0	314.5438	14.0641	62216.7636
1.5	403.8787	9.3509	35315.4766
2.0	744.4999	6.0851	27567.9379
2.5	**1419.4191**	**4.3059**	**26318.3012**
3.0	3249.519	3.2828	35019.4164
3.5	7532.2469	2.6707	53726.0107
4.0	14059.3651	2.2835	73316.7771
4.5	27448.5771	2.0225	112286.639
5.0	48089.0095	1.8343	161809.899

The minimum of the upper bound is reached for $\sigma = 2.5$. The model will therefore be estimated with $C = 1000$ and $\sigma = 2.5$.

The second step is to test the ability of the zero-order score to give the order of relevance of variables. For this purpose, we calculate, each time removing a variable, the $\sum_{i=1}^n \max_{1\leqslant k\leqslant q} \alpha_{ik}^*$ and the diameter D_n of the smallest sphere enclosing the data in features space, using the parameters chosen in Step 1. The results are reported in Table 2.

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 $D_n^2 \sum_{i=1}^n \max_{1\leqslant k\leqslant q} \alpha_{ik}^*$ when the variable is removed. The order of relevance of variables obtained according to Table 2 is as follows:

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

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

Table 2
Zero-order scores of the 10 variables

Removed variable	$\sum_{i=1}^n \max_{1\leqslant k\leqslant q} \alpha_{ik}^*$	D_n	Zero order score of variables $(D_n^2 \sum_{i=1}^n \max_{1\leqslant k\leqslant q} \alpha_{ik}^*)$
x^1	4704.2294	3.4967	**57519.5574**
x^2	4482.486	3.5298	**55849.5973**
x^3	2201.1227	4.0958	36926.1656
x^4	2545.6538	4.0479	41712.1738
x^5	2227.1454	4.0309	36188.5102
x^6	2154.1154	4.0386	35135.0135
x^7	2345.0198	4.0536	38533.8045
x^8	2320.8622	4.0376	37835.9383
x^9	2082.4165	4.0631	34378.2499
x^{10}	2166.5158	4.1308	36968.3719

After ranking variables, we estimate the models to select the one that minimizes the validation error (using the validation set) and gives the optimal subset. For this, we build 10 databases: the first contains the first relevant variable, the second contains the first two relevant variables... etc. And in order to select the optimal parameters, we try different values of C and σ for each model.

The results of seven combinations of C and σ are presented in Fig. 1. The best subset for all combinations contains the first two variables.

The best combination giving the minimum error rate is $C = 100$ and $\sigma = 2$. Table 3 shows the errors obtained with those values according to the number of used variables on the validation set.

The last step is to calculate the testing error rate on the test sample of 50000 observations using the set of the two first variables and the parameters chosen from validation ($C = 100$ and $\sigma = 2$). The final training sample contains data used in training and validation ($N = 1250$). The testing error rate resulting is 39.57%. Using all features the testing error rate is 53.74%.

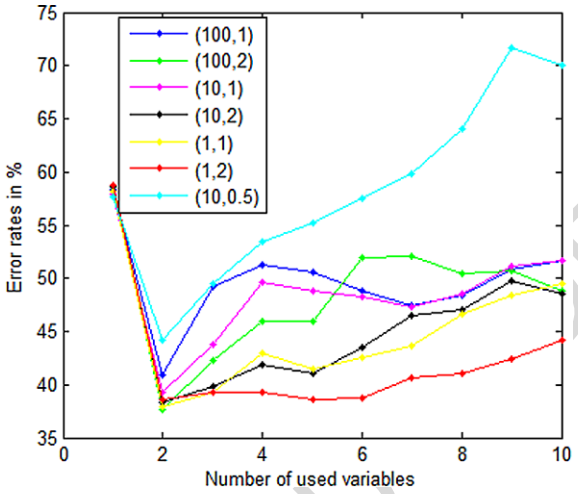

Fig. 1. Validation error rates in terms of the number of used variables for seven combinations of C and σ.

Table 3
Validation error rates with $C = 100$ and $\sigma = 2$

Used variables	Validation error in %
(x^1)	58.6
(x^1, x^2)	**37.7**
(x^1, x^2, x^4)	42.3
(x^1, x^2, x^4, x^7)	45.9
$(x^1, x^2, x^4, x^7, x^8)$	46.0
$(x^1, x^2, x^4, x^7, x^8, x^{10})$	52.0
$(x^1, x^2, x^4, x^7, x^8, x^{10}, x^3)$	52.1
$(x^1, x^2, x^4, x^7, x^8, x^{10}, x^3, x^5)$	50.5
$(x^1, x^2, x^4, x^7, x^8, x^{10}, x^3, x^5, x^6)$	50.7
$(x^1, x^2, x^4, x^7, x^8, x^{10}, x^3, x^5, x^6, x^9)$	48.8

4.1.2. Example 2

In the simulation example in Section 4.1.1, 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, $n_2 = 1000$ observations, $n_3 = 50000$ observations, and $P = 10$ variables) 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 it's important for 3 classes. The 8 remaining variables are noise ones.

The numerous simulations (as done in the first example (Table 1)) have allowed to choose the type of kernel (Gaussian) and to set the values of parameters C and σ ($C = 1000$ and $\sigma = 2$) 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 5 presents the obtained error rates using $C = 10$ and $\sigma = 2$ according to the number of used variables on the validation set ($n_2 = 1000$ observations). The best subset is the one containing the first two relevant variables.

The last step is to calculate the testing error rate on the test sample of 50000 observations using the set of the two first variables and the parameters chosen from validation ($C = 10$ and $\sigma = 2$). The final training sample contains data used in training and validation ($N = 1250$). The testing error rate resulting is 40.16% which is better than that obtained using all features (45.24%).

4.2. Real data

We divide this subsection on three parts, in the first example we will compare the performances of our approach according to the number of variables with those given by maximum relevance approach [25] using a real dataset from UCI repository. In the second and third examples, we will compare the classification results given by our approach to those of other methods of variables selection by multiclass SVM.

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

Removed variable	$\sum_{i=1}^{n} \max_{1\leqslant k\leqslant q} \alpha_{ik}^{*}$	D_n	Zero order score of removed variable	Ranking
x^1	1132.2971	5.3268	32129.8990	8
x^2	1672.0327	4.8217	**38873.5347**	**2**
x^3	1991.1498	4.7828	**45548.7694**	**1**
x^4	1097.9851	5.4346	32428.9876	7
x^5	1195.6819	5.3492	34214.1712	3
x^6	1165.3138	5.3818	33752.0906	5
x^7	1147.0951	5.4145	33629.5977	6
x^8	1181.6572	5.3779	34176.5754	4
x^9	1066.5453	5.3597	30639.0415	10
x^{10}	1090.8932	5.3707	31466.7816	9

Table 5
Validation error rates with $C = 10$ and $\sigma = 2$

Used variables	Validation error in %
(x^3)	58.70
(x^3, x^2)	**38.90**
(x^3, x^2, x^5)	41.50
(x^3, x^2, x^5, x^8)	40.30
$(x^3, x^2, x^5, x^8, x^6)$	42.50
$(x^3, x^2, x^5, x^8, x^6, x^7)$	42.60
$(x^3, x^2, x^5, x^8, x^6, x^7, x^4)$	44.60
$(x^3, x^2, x^5, x^8, x^6, x^7, x^4, x^1)$	44.50
$(x^3, x^2, x^5, x^8, x^6, x^7, x^4, x^1, x^{10})$	45.30
$(x^3, x^2, x^5, x^8, x^6, x^7, x^4, x^1, x^{10}, x^9)$	45.30

4.2.1. Example 1

In this section we compare the performances in classification of our approach to those of the well known ranking approach maximum relevance consisting in ranking variables according to the relevance to the target class using mutual information [25].

For this, we will use the waveform data set from UCI repository containing 5000 observations, 40 variables and 3 classes. All of the 40 attributes include noise, the first 21 are important for class separation and the last 19 are noise variables with mean 0 and variance 1.

After having ranked the 40 variables in order of relevance by maximum relevance approach and radius margin based approach, on a training set of 3000 observations, we compare the performances of the two methods according to the number of used variables in decreasing order of relevance, using a validation set of 1000 observations.

The minimum validation error rate using the maximum relevance method is 33.5% obtained using the first five variables among the 21 relevant ones, however, our approach based on the radius margin bound gave a better error rate 28.7% using the first six variables among the 21 relevant ones.

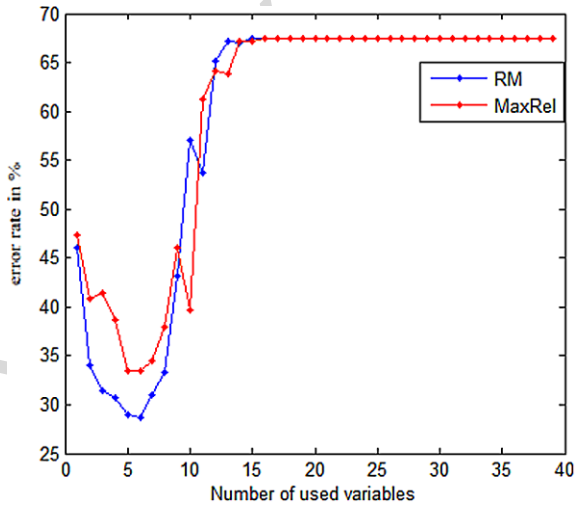

Fig. 2. Validation error rates in terms of the number of used variables.

In the testing step, based on the remaining 1000 observations, the five variables of the maximum relevance approach give an error rate of 33.1% and the six features of our approach give 28.7% (see Fig. 2).

4.2.2. Example 2

The large dimensional dataset on which we test the performances of our approach is the children cancer data set [16] classifying the small round blue cell tumors (SRBCTs) into four classes, namely neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), et Ewing (EWS) using cDNA gene expression profiles.

http://research.nhgri.nih.gov/microarray/Supplement/.

The data set includes 63 observations for training, 20 blind observations for testing and 2308 variables

(genes). It was presented for the first time by Khan et al. [16] and has been used in the context of variable selection by multiclass SVM by Zhang et al. [37].

We split the 63 observations into 50 for training and 13 for validation, then, we standardize the training set to have zero mean and unit variance. The validation and test sets are standardized based on the training set parameters.

Afterward, in order to reduce the computation time, we select 200 genes by the filter approach MRMR (Maximum Relevance Minimum redundancy) [25], known for its good performance in terms of reducing the number of variables in large dimensions. These 200 genes on which we apply our approach.

After model selection, we proceed to rank the 200 features in order of relevance, then, using forward method, we select the best subset giving the error rate minimum on the validation set and the best combination of C and σ.

Table 6 summarizes the results obtained with the selected combination ($C = 10$ and $\sigma = 2$) and the best subset containing the first 9 genes.

The proposed approach based on the multiclass RM bound predicts correctly the classes of the unseen 20 test observations using the first 9 genes.

Comparing these classification results to those of some previous studies on this data set, we deduce that they are far better in terms of the number of genes needed and the recognition rates obtained. Indeed, selecting variables for multiclass SVM using adaptive sup-norm regularization (Adp-supII), Zhang et al. [37], obtained one error, using the first 47 genes and using the L1 norm, 63 genes were required. Furthermore, using Neural Networks (NN), getting a zero error rate was possible using the first 96 genes [16].

Table 6
Classification and selected genes from SRBCT data

	Validation		Testing error rate
	Obtained error	Selected genes (Image Id)	
MRMR + RM	0/13	325182	0/20
		143306	
		814444	
		878652	
		813707	
		796258	
		842820	
		1435862	
		784224	

However, some studies criticize the use of filters as pre-processing step and claim that some times the good results are due to the use of these filter.

In order to verify that the results obtained are not due to the preprocessing step selecting 200 genes by MRMR filter, we apply directly our approach based on the RM bound to the 2308 genes.

As before, we select the parameters minimising the RM bound in presence of all variables, then we rank them in order of relevance and we select, based on the validation set, the best subset and the best combination of C and σ using forward method. The results are presented in Table 7.

The results of classification show the effectiveness of our approach to select a very reduced subset of variables which give good classification rates on unseen data in large dimension.

A summary of the classification results for SRBCT dataset is presented in Table 8.

4.2.3. *Example 3*

For this last example, we will test the performances of our approach on two real datasets from UCI repository and we will compare the obtained testing errors to those obtained by the method of Li et al. presented in [20].

The datasets are soybean-s and stepp-order data containing both 47 samples, 35 variables and 4 classes. For comparison, we will apply the same steps as Li et al.

Table 7
Classification and selected genes from the 2308 original ones

	Validation		Testing error rate
	Obtained error	Selected genes (Image Id)	
RM	0/13	325182	1/20
		770394	
		629896	
		377461	
		365826	
		796258	

Table 8
Classification results and selected genes for SRBCT dataset

Method	Number of selected genes	Testing error rate
L1 [37]	63	1/20
Adp-supII [37]	47	1/20
NN [16]	96	0/20
RM	6	1/20
MRMR + RM	9	0/20

Table 9
Classification and selected variables

	Validation		Testing error rate	Error rate with all features
	Obtained error	Number of selected variables		
Soybean-s	0	7.25	0.0217 ± 0.0251	0.2608 ± 0.2534
Stepp-order	0	7.25	0.0326 ± 0.0416	0.1630 ± 0.2111

For this, we first transform the attributes into the interval $[-1, 1]$, then, we split the datasets into two equal parts in terms of the number of observations per class, the first part for training and validation and the second for testing. Such split is performed several times.

For each dataset, after having chosen the parameters minimizing the RM bound for the hard margin $MSVM_{\mathrm{LLW}}$, we rank the variables in order of relevance based on the training sets, then, we select the optimal subsets and the parameters using the validation sets. Finally, the testing error is calculated on the test sample. Table 9 lists the results obtained.

The numbers of the selected variables are given in average from the different splits. The results in terms of the testing error rates give the average errors with their corresponding standard deviation.

In comparison to the case where all the features are used, the results confirm that our approach can increase significantly the recognition rate and reduce the number of variables.

Furthermore, comparing these results to those of Li et al., we note that the proposed method produce a slightly better result for the soybean dataset. For the stepp-order data, the obtained error rate is not far from that obtained in [20].

5. Conclusion and perspectives

The results of the studies on variable selection by multiclass SVM show the effectiveness of using this technique to reduce the dimensions and to improve the classifications accuracy. In this paper, we proposed a new approach, based on the multiclass radius margin upper bound of the generalization error, to give the order of relevance of the 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 $MSVM_{\mathrm{LLW}}$ model which is the most theoretically based of MSVM models [19] and as a wrapper approach, selecting the variables after the estimation of the model, takes into account the influence of each variable on the performance of the estimated model.

A constraint for the application of our procedure in very large dimensions consists in the 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. To deal with this problem, as done in the second real example, we propose to combine our approach with an appropriate filter method which will filter a big number of noise variables before applying the radius margin bound. The results obtained on some high dimensional cancer datasets were very good in terms of the obtained testing error rate.

Acknowledgements

We would like to acknowledge support for this work from CNRST-Ministry of Higher Education, Scientific Research And Executive Training of Morocco and Rhône-Alpes Region/France by according to Aazi F.Z, respectively, a research scholarship and accueil-DOC 2012 and 2013 scholarships. We wish to especially thank Hassan 1st University and express our gratitude for its financial support and finally, the authors would like to thank Professors Y. Guermeur and F. Lauer for their valuable help and for providing them with additional informations on their works.

References

[1] A. Ben-Hur, D. Horn, H.T. Siegelmann and V. Vapnik, Support vector clustering, *Journal of Machine Learning Research* **2** (2001), 125–137.

[2] A. Blum and P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence* **97**(1–2) (1997), 245–271. doi:10.1016/S0004-3702(97)00063-5.

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

[4] O. Chapelle and S. Keerthiy, Multi-class feature selection with support vector machines, in: *Proceedings of the American Statistical Association*, ASA, Denver, CO, USA, 2008.
[5] X.-W. Chen, X. Zeng and D.V. Alphen, Multi-class feature selection for texture classification, *Pattern Recognition Letters* **27** (2006), 1685–1691. doi:10.1016/j.patrec.2006.03.013.
[6] K. Crammer and Y. Singer, On the algorithmic implementation of multiclass kernel based vector machines, *Journal of Machine Learning Research* **2** (2001), 265–292.
[7] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286** (1999), 531–537. doi:10.1126/science.286.5439.531.
[8] P.M. Granitto and A. Burgos, Feature selection on wide multiclass problems using OVA-RFE, *Inteligencia Artificial* **44** (2009), 27–34.
[9] Y. Guermeur, SVM multiclasses, théorie et applications. Habilitation à diriger des recherches, Université Nancy 1, 2007.
[10] Y. Guermeur and E. Monfrini, A quadratic loss multi-class SVM for which a radius margin bound applies, *Informatica* **22** (2011), 73–96.
[11] J. Guo, Class-specific variable selection for multicategory support vector machines, *Statistics and Its Interface* **4** (2011), 19–26. doi:10.4310/SII.2011.v4.n1.a3.
[12] I. Guyon, A. Elisseeff and L. Kaelbling, An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**(7–8) (2003), 1157–1182.
[13] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning* **46**(1–3) (2002), 389–422. doi:10.1023/A:1012487302797.
[14] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Springer, 2001.
[15] S.S. Keerthi and C.J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, *Neural Computation* **15** (2003), 1667–1689. doi:10.1162/089976603321891855.
[16] J. Khan et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine* **7** (2001), 673–679. doi:10.1038/89044.
[17] R. Kohavi and G. John, Wrapper for feature subset selection, *Artificial Intelligence* **97**(1–2) (1997), 273–324. doi:10.1016/S0004-3702(97)00043-X.
[18] F. Lauer and Y. Guermeur, MSVMpack: A multi-class support vector machine package, *Journal of Machine Learning Research* **12** (2011), 2269–2272.
[19] 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** (2004), 67–81. doi:10.1198/016214504000000098.
[20] G.-Z. Li, J. Yang, G.-P. Liu and L. Xue, Feature selection for multi-class problems using support vector machines, in: *PRICAI2004*, Lecture Notes on Artificial Intelligence, Vol. 3173, Springer, 2004, pp. 292–300.
[21] J.-T. Li and Y.-M. Jia, Huberized multiclass support vector machine for microarray classification, *Acta Automatica Sinica* **36** (2010), 399–405.
[22] W. Li and Y. Yang, How many genes are needed for a discriminant microarray data analysis?, in: *Critical Assessment of Techniques for Microarray Data Mining Workshop*, 2000, pp. 137–150.
[23] Y. Liu and Y. Wu, Variable selection via a combination of the L0 and L1 penalties, *Journal of Computation and Graphical Statistics* **16** (2007), 782–798. doi:10.1198/106186007X255676.
[24] 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** (2005), 160–171. doi:10.1155/JBB.2005.160.
[25] H. Peng, F. Long and C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(8) (2005), 1226–1238. doi:10.1109/TPAMI.2005.159.
[26] A. Rakotomamonjy, Variable selection using SVM-based criteria, *Journal of Machine Learning Research* **3** (2003), 1357–1370.
[27] M.-D. Shieh and C.-C. Yang, Multiclass SVM-RFE for product form feature selection, *Expert Systems with Applications* **35** (2008), 531–541. doi:10.1016/j.eswa.2007.07.043.
[28] D.M.J. Tax and R.P.W. Duin, Support vector data description, *Machine Learning* **54** (2004), 45–66. doi:10.1023/B:MACH.0000008084.60811.49.
[29] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
[30] L. Wang and X. Shen, Multi-category support vector machines, feature selection and solution path, *Statistica Sinica* **16** (2006), 617–633.
[31] L. Wang and X. Shen, On L1-norm multi-class support vector machines: Methodology and theory, *Journal of the American Statistical Association* **102** (2007), 583–594. doi:10.1198/016214506000001383.
[32] J. Weston, A. Elisseeff, B. Schölkopf and M. Tipping, Use of the zero-norm with linear models and kernel methods, *Journal of Machine Learning Research* **3** (2003), 1439–1461.
[33] 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.
[34] E. Xing, M. Jordan and R. Karp, Feature selection for high-dimensional genomic microarray data, in: *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 601–608.
[35] M. Xiong, X. Fang and J. Zhao, Biomarker identification by feature wrappers, *Genome Res.* **11** (2001), 1878–1887.
[36] Y. Yang and J.O. Pederson, A comparative study on feature selection in text categorization, in: *Proceedings of the Fourteenth International Conference on Machine Learning*, Vol. 412, 1997.
[37] 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* **2** (2008), 149–167. doi:10.1214/08-EJS122.
[38] X. Zhou and D.P. Tuck, MSVM-RFE: Extensions of SVM-RFE for multiclass gene selection on DNA microarray data, *Bioinformatics* **23** (2007), 1106–1114. doi:10.1093/bioinformatics/btm036.

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

[40] H. Zou and M. Yuan, The F_∞-norm support vector machine, *Statistica Sinica* **18** (2008), 379–398.