CIT. Journal of Computing and Information Technology, Vol. 24, No. 4, December 2016, 369–382
doi: 10.20532/cit.2016.1003342

Microcanonical Annealing and Threshold Accepting for Parameter Determination and Feature Selection of Support Vector Machines

Seyyid Ahmed Medjahed[1], Tamazouzt Ait Saadi[2], Abdelkader Benyettou[1] and Mohammed Ouali[3]

[1]Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, Algérie
[2] University of Le Havre, France
[3]Computer Science Department, University of Sherbrooke, Canada

Support vector machine (SVM) is a popular classification technique with many diverse applications. Parameter determination and feature selection significantly influences the classification accuracy rate and the SVM model quality. This paper proposes two novel approaches based on: Microcanonical Annealing (MA-SVM) and Threshold Accepting (TA-SVM) to determine the optimal value parameter and the relevant features subset, without reducing SVM classification accuracy. In order to evaluate the performance of MA-SVM and TA-SVM, several public datasets are employed to compute the classification accuracy rate. The proposed approaches were tested in the context of medical diagnosis. Also, we tested the approaches on DNA microarray datasets used for cancer diagnosis. The results obtained by the MA-SVM and TA-SVM algorithms are shown to be superior and have given a good performance in the DNA microarray data sets which are characterized by the large number of features. Therefore, the MA-SVM and TA-SVM approaches are well suited for parameter determination and feature selection in SVM.

AACM CCS (2012) Classification: Computing methodologies → Feature selection

Computing methodologies → Machine learning → Machine learning algorithms → Feature selection

Theory of Computation → Theory and algorithms for application domains → Machine learning theory → Kernel methods → Support vector machines

Keywords: support vector machines, microcanonical annealing, threshold acceptance, parameter determination, feature selection

1. Introduction

Over the recent past years, support vector machines (SVMs) have become the reference for many classification problems. They have several advantages, we mention: their flexibility, their ability of generalization and their computational efficiency. The major problem in SVMs lies in the fact that they do not directly obtain the relevant features and the optimal parameters values.

The effectiveness of SVMs to obtain high classification rate and perfect quality of SVM model lies in two critical factors: feature selection and parameter determination [1]. The objectives of feature selection aim to reduce the number of features by removing irrelevant, noisy and redundant features. In addition, the determination of optimal value of parameters has an important role in reaching a high classification accuracy rate. There are two main parameters: the SVM model parameter and the kernel function parameter.

The adjustment of these parameters is a very interesting field of research. The classification accuracy rate largely depends on SVM parameter C (the regularization parameter) and the kernel function parameter which must be chosen carefully.

Several studies have been conducted in the domain of the parameters determination: grid search [2], [3] is the most widely used to determine the parameters of SVM and kernel function. Another approach, defined by Pai and Hong [4], combines genetic algorithms and the SVM to generate a set of parameter values for SVM. Also, Pain and Hong in [5], [6] presented a simulated annealing approach to obtain parameter values of SVM and test their approach on a real data set. Ren and Bai [7] developed

an approach to determine the optimal SVM parameters by using genetic algorithms and particle swarm optimization. These studies focused only on the determination of the parameters. Seyyid Ahmed Medjahed *et al.* [8] have proposed a new approach for parameter determination and feature selection called SA-SVM. This approach is based on simulated annealing and SVM.

For feature selection, many researchers are interested in developing a new method. The feature selection methods can be categorized as: Filter, Wrapper and Embedded methods. In [9] Chen and Hsien developed a latent semantic analysis (LSA) and a web page for feature selection (WPFA), combined with the SVM to screen features. Gold *et al.* [10] developed a Bayesian viewpoint of SVM classifiers to adjust the parameter values in order to determine the irrelevant features. Chapelle *et al.* [11] presented an automatically tuning multiple parameters and used the principal components to obtain features for the SVM technique. In [12] the authors adopt the accuracy rate of the classifier as the performance measure. Shon *et al.* [13], use genetic algorithms in screening the dataset features. Seyyid Ahmed Medjahed *et al.* [14] have proposed a new feature selection approach based on Gray Wolf Optimizer and a new objective function for band selection in hyperspectral image. In [15], the authors have used Binary Cuckoo Search which is a new optimization approach for hyperspectral band selection.

In this study, the problem of parameter determination and feature selection is defined as a combinatorial optimization problem. The goal is to reduce the number of features by eliminating the irrelevant features and to determine the optimal value of SVM parameters.

We propose two novel approaches called: MA-SVM (Microcanonical Annealing – Support Vector Machine) and TA-SVM (Threshold Accepting – Support Vector Machine). The first approach is based on Microcanonical Annealing (MA) and the second uses Threshold Accepting (TA). These algorithms are stochastic optimization algorithms which have never been tested in the context of feature selection and parameter determination. We slightly modified the MA and TA algorithms to be adapted for our objective.

Performance assessment was conducted on five medical datasets taken from the UCI Machine Learning Repository and two DNA microarray datasets largely used for cancer diagnosis (colon cancer and leukemia).

The first contribution of our approach is to combine the problems of parameter determination and feature selection in a single problem which interestingly obtains good results and establishes a relationship with SVM parameters and the selected features. Also, this combination provides a good generalization and a perfect quality of the SVM model.

The second contribution is to demonstrate the performances of two very effective optimization methods (MA and TA) which have never been applied in the context of feature selection and parameter determination.

This paper is organized as follows: in the next section, an overview of SVM is introduced. In Section 3, we discuss the problem of parameter determination and feature selection. Section 4 details the proposed approaches. In Section 5, we analyze the experimental results. Finally, concluding remarks are made in Section 6.

2. Overview of Support Vector Machine

Support vector machine is a popular classification method. Developed by Vladimir N. Vapnik in 1995 [16], [17], SVM aims to solve the binary classification problem by finding the optimal hyperplane which maximizes the margin between the instances of classes [18].

Mathematically, finding the optimal hyperplane is equivalent to solving the following primal optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ s.\,c. \quad & y_i\left(\langle w, x_i\rangle + b\right) \geq 1-\xi_i \\ & \xi_i \geq 0 \\ & i \in \{1,\ldots,N\} \end{aligned} \tag{1}$$

where, $X \subseteq \mathbb{R}^d$ is the instances set and $Y = \{-1, +1\}$ are the labels.

$\forall_i\, \xi_i \geq 0$ are called *slack variables* and they represent the distance between the wrong points and the hyperplane.

The parameter C controls the trade-off between the slack variables and the size of the margin. By introducing the Lagrange multipliers, we define the dual form of the problem (1) as follows:

$$\begin{aligned} \min_{\alpha} \quad & -\sum_{i=1}^{N}\alpha_i + \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right) \\ s.\,c. \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & \forall i \in \{1, \ldots, N\},\ 0 \le \alpha_i \le C \end{aligned} \tag{2}$$

$K(x_i, x_j)$ is the kernel function. Linear, Polynomial and Gaussian kernels are the widely used in the literature and are defined as follows:

Linear $K(x_i, x_j) = (x_i \cdot x_j)$
Polynomial $K(x_i, x_j) = [(x_i \cdot x_j) + 1]^d$ where $d \in \mathbb{N}$, $d \neq 0$
Gaussian $K(x_i, x_j) = \dfrac{-\left\|x_i - x_j\right\|^2}{2\sigma^2}$.

The problem (2) can be solved by using: Interior Point, Sequential Minimal Optimization, Trust Region, etc. In this work, we propose to use Sequential Minimal Optimization [19], [20], [21].

2.1. Sequential Minimal Optimization

Sequential Minimal Optimization (SMO) [19], [20], [21] is an optimization method used to solve quadratic problems. SMO is largely used for SVM and it reformulates the quadratic problem into small sub-problems.

The algorithmic scheme of SMO can be described as follows:

1. Firstly, we take a Lagrange multiplier α_1 that does not meet the conditions for *Karush-Kuhn-Tucker* (*KKT*) for the optimization problem.
2. The second step is to take a second multiplier α_2 and optimize the pair (α_1, α_2).
3. The steps 1 and 2 are repeated until convergence.

2.2. Parameter Determination and Feature Selection

Hyper-parameters of SVM (Parameter C and kernel function parameter) must be chosen carefully. The results depend directly on these parameters.

SVM parameter C represents the cost of the penalty. It creates a soft margin that permits some misclassifications. A large value of C increases the cost of misclassifying points and forces the creation of a more accurate model that might not generalize well. A small value of C produces unsatisfactory accuracy rate and makes the model useless [22].

The kernel parameter influences also the classification accuracy rate. For example, by using a Gaussian kernel, the parameter σ must be determined. The kernel parameter σ has a much stronger impact than parameter C on classification outcomes, because its value influences the partitioning outcome in the feature space. An excessive value for parameter σ leads to over-fitting, while a disproportionately small value results in under-fitting [22], [23].

In addition, feature selection is an important step in classification. It aims to select the optimal subset of features, which improves the generalization performance, computational efficiency, and feature interpretability. The features can be highly correlated and uninformative, which might decrease the classification accuracy rate and the quality of SVM model [22].

The aim of feature selection is to find the smallest feature subset that increases the classification accuracy rate. The optimal feature subset is not unique; it may be possible to achieve the same accuracy rate using different sets of features, because if two features are correlated, one can be replaced by the other. Note that feature subset selection chooses a set of features from the existing features, and does not construct new ones; there is no feature extraction or construction [24], [25].

3. The Proposed MA-SVM and TA-SVM Approaches

Most methods focus on solving the problem of parameter determination or the problem of feature selection, separately. These approaches do not address the problems of feature selection and parameter determination together.

The present study proposes to combine the problems of feature selection and parameter

determination in a single model. This combining improves the quality of SVM model and the generalization ability. Therefore, combining feature selection with parameter determination provides the values of SVM parameter C and kernel parameter in function of the relevant subset of selected features.

In the proposed approaches, we reformulate the problems of parameter determination and feature selection as a single combinatorial problem which can be defined as follows:

Considering the dataset $D = \{f_1, ..., f_t\}$ where f_i represents the feature of the dataset $\forall i = 1, ..., t$. We define the set $V = \{V_1, ..., V_t\}$, $V_i = \{0, 1\}$ $\forall i = 1, ..., t$, such as:

if $V_i = 1$	The feature f_i is selected and participates in the construction of the model
else if $V_i = 0$	The feature f_i is not selected and does not participate in the construction of the model

In this study, the objective function $E(X)$ of the proposed approaches is the classification error rate computed by using SVM. The goal is to find the optimal value of C, kernel parameter σ and V_i (SVM parameters and the optimal subset of features) that minimize the classification error rate. We are facing a combinatorial optimization problem which can be solved by using stochastic search techniques. The basic idea is to find the space where real valued energy function is minimized (finding the optimum).

3.1. Microcanonical Annealing

The microcanonical annealing (MA) is a variant of simulated annealing developed by Creutz [26] and it possesses properties close to those of simulated annealing [27]. The main difference with simulated annealing is the convergence towards the global optimum. The first is based on plateaus of temperature [28] and the second on decreasing plateaus of total energy.

The microcanonical annealing algorithm uses the algorithm of Creutz, which is based on the evaluation of a series of transitions to maximize the entropy for a total energy constant [29], [30]. This total energy is fixed beforehand [26].

The microcanonical annealing considers that the system is isolated (no heat exchange with its environment), thus it is based on decreasing levels of total energy related to the reduction of the kinetic energy at each step.

The total energy is defined as:

$$E_t = E(x_k) + E_c \; (E(x_k) \text{ is the energy function in step } k)$$

The total energy must be very high at the beginning of the algorithm and by reducing the total energy, the algorithm will converge to the global optimum. In this case, the kinetic energy E_c will play a similar role as the temperature in simulated annealing and it is constrained to be positive:

$$Pr\,(E_c = E) \propto exp\left(\frac{E}{KT}\right) \tag{3}$$

where K is the Boltzmann constant and T is the temperature in simulated annealing standard [26].

Equation (3) explains that the kinetic energy follows the Boltzmann distribution [31]. The MA algorithm is described as follows:

Algorithm 1. Microcanonical annealing.

1: initialize at a random state x_k; calculate the function $E(x_k)$
2: choose another state x_{k+1}; calculate the function $E(x_{k+1})$
3: **if** $\Delta E < 0$ **or** $\Delta E < E_c$ **then**
4: accept the transition and decrease the total energy
5: **else**
6: reject the transition
7: **end if**
8: repeat steps 2 to 7 until reaching equilibrium

Initially, the total energy is very high, in each plateau, the total energy is reduced. When $\Delta E = 0$, the algorithm accepts all transitions to the states of lower total energy. The movement toward higher total energy states is accepted only when $\Delta E = E_c$ (there must be a sufficient kinetic energy to compensate for the increase of potential energy, and thus to stay at constant energy).

Compared to simulated annealing, the advantage of the microcanonical annealing is that it does not require a random number generator for the acceptance or refusal of a configuration. Nevertheless, the MA algorithm is much faster than SA algorithm [31]. In addition, in case of large problems, several studies have shown that the results are very similar to those of simulated annealing, with a benefit to the microcanonical annealing in terms of computation time [31].

However, Creutz [26] noted that, in case of the problems of small size, the probability for the system to be trapped in metastable states is higher [26], [32].

In this first approach called MA-SVM, we adapt the microcanonical annealing to the problem of parameter determination and feature selection. Figure 1 shows the procedure of MA-SVM approach, the total energy is reduced in each plateau, and when the total energy is close to 0, the algorithm converges to the global optimum.

3.2. Threshold Accepting

Threshold accepting (TA) is a heuristic optimization algorithm. Developed by Duek and Scrubs [33], the threshold accepting is a variation of simulated annealing which simplifies the simulated annealing procedure by leaving out the probabilistic element in accepting worse solutions [34].

TA uses a deterministic threshold and a worse solution is accepted if the difference between the worse solution and the current solution is smaller than or equal to threshold [33]. Explicitly, TA algorithm uses a similar approach as simulated annealing, but instead of accepting new points that raise the objective function with a certain probability, it accepts all new points below a fixed threshold [35], [36]. The TA algorithm is described in Algorithm 2.

The threshold is systematically lowered like the temperature in simulated annealing standard [36].

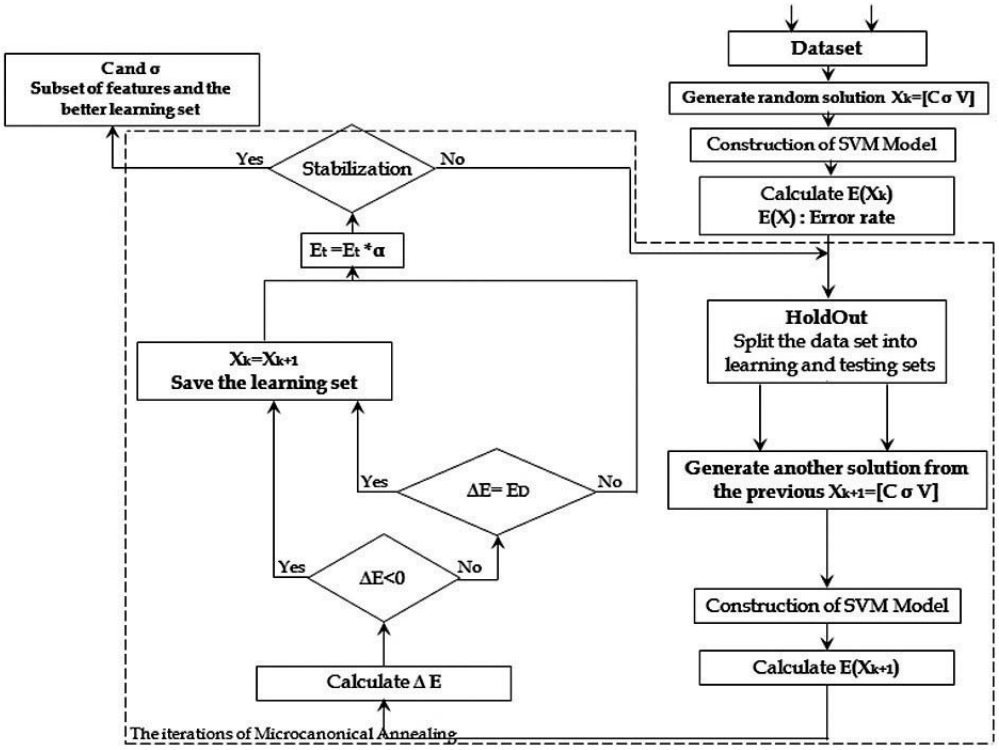

Figure 1. The procedure of the proposed MA-SVM approach.

Algorithm 2. Threshold accepting.

1: initialize $Numb_{Rounds}$ and $Numb_{Steps}$
2: randomly generate current solution x_k
3: calculate the function $E(x_k)$
4: **for** $r = 1$: $Numb_{Rounds}$
5: compute threshold sequence τ_r
6: **for** $i = 1$: $Numb_{Steps}$ **do**
7: generate another solution x_{k+1}
8: calculate the function $E(x_{k+1})$
9: Calculate $\Delta E = E(x_{k+1}) - E(x_k)$
10: **if** $\Delta E = < \tau_r$
11: $x_k = x_{k+1}$
12: accept the transition
13: **else**
14: reject the transition
15: **end if**
16: **end for**
17: **end for**

The key components of TA are the function that determines the lowering of the threshold during the course of the procedure, stopping criteria as well as the methods used to create initial and neighboring solutions. The main advantages of TA are its conceptual simplicity and its excellent performance on different combinatorial optimization problems [37].

In this second approach called TA-SVM, we adapt the threshold accepting algorithm to our problem. Figure 2 illustrates the operation of TA-SVM approach.

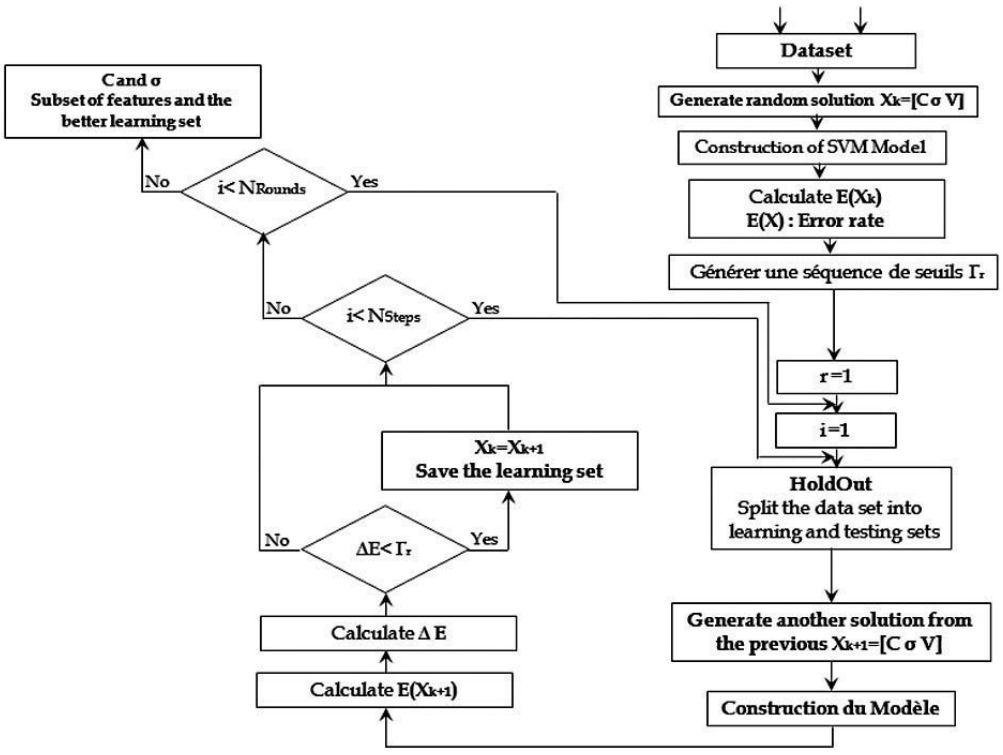

Figure 2. The procedure of proposed TA-SVM approach.

The second idea of this research is to integrate the training set selection in the problem by combining the holdout method with the MA and TA algorithms as shown in Figures 1 and 2.

The instances that constitute the training set represent a very important factor in building the model and obtaining a high classification accuracy rate. Some instances are considered as noise and their removal increases the classification accuracy rate. Therefore, we propose to incorporate the holdout method in each iteration of MA and TA algorithms. In each iteration, MA and TA algorithms split the dataset randomly into training and testing sets.

4. Experiment Results

4.1. Datasets

In this study, to evaluate performances of the proposed MA-SVM and TA-SVM approaches, two experimentations are done.

The first experimentation is conducted on five UCI Machine Learning Repository datasets (Breast Cancer, Cardiotocography, ILPD (Indian Liver Patient Dataset), Mammographic Mass and Vertebral Column) widely used in the literature. These datasets are taken from: http://archive.ics.uci.edu/ml/.

The format of UCI Machine Learning Repository datasets is arranged as shown in Table 1.

The second experimentation is done to support and validate our results. We have applied both approaches on two DNA microarray datasets: Colon Cancer and Leukemia. These datasets are widely used in the literature [38], [39], [40], [41], [42].

The colon cancer dataset contains expression levels of 2000 genes taken from over 62 different samples (40 negative and 22 positive) [43].

The leukemia dataset contains expression levels of 7129 genes taken from over 72 samples. Labels indicate which of the two variants of leukemia is present in the sample (AML: 25 samples and ALL: 47 samples) [44].

The colon cancer and leukemia datasets used for this experimentation are characterized by the large number of features (genes expression) and the analysis of all these genes is impossible. Nevertheless, the selection of a relevant feature subset has an important role.

4.2. Parameters setting

To conduct our experimentations and demonstrate the effectiveness of our approaches, the parameters of the proposed approaches are configured as follows:

For MA-SVM, the total energy E_t is set to 1000. This value will decrease slowly by following a geometric law with a ration $\alpha = 0.99$.

For TA-SVM, the value of $Numb_{Rounds}$ is equal to $1000 \times Domain\_size$.

MA-SVM and TA-SVM stop when the value of energy function reaches 0 (classification error rate = 0) or when the energy function stops evolving after a certain number of iterations (200 iterations).

SVM classifier is used with four kernel functions: Linear, Polynomial and Gaussian.

The dataset is randomly divided, by using the holdout method, into three sets: 60% of instances constitute the training set, 20% of instances constitute the testing set and 20% of instances are used for the validation.

Table 1. Datasets taken from the UCI Machine Learning Repository.

Datasets	Number of classes	Number of instances	Number of features
Breast Cancer	2	699	9
Cardiotocography	2	1831	21
ILPD	2	583	9
Mammographic Mass	2	961	5
Vertebral Column	2	310	6

Table 2. Classification accuracy rates obtained by MA-SVM and TA-SVM for each dataset.

Datasets	TA-SVM (%)			MA-SVM (%)		
	Linear	Polynomial	Gaussian	Linear	Polynomial	Gaussian
Breast Cancer	98.33	99.00	99.26	99.95	100	100
Cardiotocography	99.01	99.22	99.86	99.12	100	100
ILPD	70.45	70.43	75.33	70.00	73.89	75.47
Mammographic Mass	88.22	89.90	90.00	86.60	90.60	90.72
Vertebral Column	85.67	85.30	91.80	89.28	90.42	91.98

4.3. Results and discussion

The results obtained for UCI Machine Learning Repository datasets by using the proposed MA-SVM and TA-SVM approaches are summarized in the Table 2.

Table 2 describes the classification accuracy rate obtained by the proposed approaches for each dataset. The classification accuracy is computed by using the SVM classifier with Linear, Polynomial and Gaussian kernel functions. The first column in Table 2 represents the datasets. The second column is the classification accuracy rate obtained by TA-SVM and the third column is the classification accuracy rate obtained by MA-SVM.

For breast cancer dataset, the high accuracy is achieved by using MA-SVM with Polynomial and Gaussian kernel (100% of accuracy). The same remark is observed for cardiotocography dataset, MA-SVM achieved 100%. For ILPD dataset, MA-SVM using Gaussian kernel has reached 75.47% of accuracy and we noted 75.33% of accuracy obtained by TA-SVM with Gaussian kernel. Also, for Mammographic mass and Vertebral column datasets, MA-SVM using Gaussian kernel has achieved slightly higher accuracy, 91.98%, than TA-SVM with Gaussian kernel which has achieved 91.80%.

As seen in Table 2, both proposed approaches have produced a high classification accuracy rate with some advantage for MA-SVM. The Gaussian kernel and Polynomial kernel have produced quite similar results for breast cancer and cardiotocography datasets (100% using MA-SVM). For the rest of the datasets, the Gaussian kernel has achieved good accuracy compared to Polynomial kernel. Low classification accuracy is noted for the Linear kernel.

The analysis of the results has demonstrated that high accuracy is observed for the Gaussian kernel. This is why we focused the rest of our experimentations on the Gaussian kernel.

The number of iterations and computational time of MA-SVM and TA-SVM using the Gaussian kernel is described in Table 3.

Table 3 shows the number of iterations and the computational time of MA-SVM and TA-SVM. The first column contains the datasets.

The second column contains the number of iterations and the last column contains the computational time in seconds.

Table 3. Computation time and number of iterations performed by MA-SVM and TA-SVM.

TA-SVM (%)		
Datasets	**Number of iterations**	**Computation time (*S*)**
Breast Cancer	6084	1409.6
Cardiotocography	10341	7289.1
ILPD	9341	4593.4
Mammographic Mass	5576	1420.6
Vertebral Column	5400	1336

MA-SVM (%)		
Datasets	**Number of iterations**	**Computation time (*S*)**
Breast Cancer	1489	500
Cardiotocography	1002	1043.99
ILPD	7313	6335.27
Mammographic Mass	5000	3140.68
Vertebral Column	4560	2391.23

From Table 3, we clearly observe that MA-SVM is much faster that TA-SVM. The number of iterations performed by MA-SVM is smaller than that performed by TA-SVM.

The analysis of the results described in Tables 2 and 3, allows us to highlight, on the one hand, that the MA-SVM approach provides better results compared to those of the TA-SVM, and, on the other hand, that both approaches, MA-SVM and TA-SVM, gave us a better classification accuracy rate and provided the most appropriate C and σ values for all the datasets used in the experimentation.

Table 4 describes the number of selected features by the MA-SVM and TA-SVM using Gaussian kernel.

Table 4. Number of features selected by MA-SVM and TA-SVM approaches.

Datasets and the initial number of features	Number of selected features	
	TA-SVM	MA-SVM
Breast Cancer (9)	6	4
Cardiotocography (21)	14	13
ILPD (9)	5	3
Mammographic Mass (5)	2	4
Vertebral Column (6)	4	4

In the context of selected features, the results show that both proposed approaches, MA-SVM and TA-SVM, have selected the smaller relevant subset of features which has given the high classification accuracy rate.

The second experimentation is conducted on DNA microarray datasets: colon cancer and leukemia. In Table 5 we summarize the results obtained by the proposed MA-SVM and TA-SVM approaches by using the Gaussian kernel.

Table 5 shows the results obtained by the proposed MA-SVM and TA-SVM approaches using the Gaussian kernel. Table 5 contains the initial number of features, the number of selected features, the classification accuracy rate, the computational time and the number of iterations.

From Table 5 and for leukemia dataset, we noticed 99.98% of accuracy obtained by MA-SVM and 66.67% of accuracy obtained by TA-SVM. Also, the MA-SVM achieved 95.65% of accuracy on colon cancer dataset and 65.22% accuracy achieved by TA-SVM.

Table 5. The results obtained by MA-SVM and TA-SVM on the DNA microarray datasets.

Leukemia		
Methods	**TA-SVM**	**MA-SVM**
Number of initial features	7129	7129
Number of selected features	7129	2693
Classification accuracy rate (%)	66,67	99,98
Computation time (S)	2214.7	3354.57
Number of iterations	481	756

Colon cancer		
Datasets	**TA-SVM**	**MA-SVM**
Number of initial features	2000	2000
Number of selected features	2000	1015
Classification accuracy rate (%)	65.22	95.65
Computation time (S)	9793.4	6619.2
Number of iterations	1538	849

In this experimentation, the results obtained by the MA-SVM are quite appropriate and very satisfying compared to those obtained by the TA-SVM in terms of classification accuracy rate and computation time. It is clear that the number of selected features is much smaller than the initial one.

TA-SVM is not very effective on large datasets (large number of features) such as the microarray gene expression dataset. This is due to the fact that TA-SVM must compute the threshold set in each round, which is very intense in computation time.

The results obtained with the two datasets corroborate the conclusion that we have reached.

The proposed MA-SVM approach is compared to several approaches defined in previous works. Table 6 presents the classification accuracy rate obtained by MA-SVM and other approaches.

Table 6 presents the classification accuracy rate obtained by the proposed MA-SVM approach, compared to SVM-RFE, SVM-RFE-mRMR, PSO-SVM, and Relief. The results of SVM-RFE and SVM-RFE-mRMR are taken from [45]. For PSO-SVM and Relief, we have used the same training, testing and validation tests. We note that PSO-SVM is a wrapper approach based on particle swarm optimization [46] and SVM. Relief is a filter approach [47].

Table 6. Classification accuracy rate (%) obtained by MA-SVM and compared to other approaches.

Methods	Leukemia	Colon cancer
MA-SVM	99.98	95.65
SVM-RFE [45]	97.88	91.00
SVM-RFE-mRMR [45]	98.38	91.68
PSO-SVM	100	99.83
Relief	91.61	81.46

The higher classification accuracy rate is obtained by PSO-SVM, compared to other approaches. As seen in Table 6, MA-SVM has provided better results compared to SVM-RFE, SVM-RFE-mRMR and Relief.

5. Conclusion

In this paper, we present two novel approaches in feature selection and parameter determination applied for medical diagnosis. The approaches are called MA-SVM and TA-SVM. The first one is based on the Microcanonical Annealing algorithm, while the second one is based on the Threshold Accepting algorithm. The objective function to minimize is the classification error rate. The aims of this study are to determine the optimal parameter of the SVM and its kernel function. In addition, we attempt to select the optimal subset of features by removing the irrelevant and redundant features.

The experimentation is done under five UCI Machine Learning Repository datasets and two DNA microarray datasets widely used in the literature.

The results obtained by both approaches demonstrate that MA-SVM improves the classification accuracy rates by removing trivial or insignificant features and effectively finds the better parameters values. The TA-SVM is adapted for the datasets that contain a small number of features.

We concluded that MA-SVM provides satisfactory results and it is adapted to large problems.

Finally, we can say that MA-SVM is thus useful for parameter determination and feature selection in the SVM, regardless of the size of the problem.

References

[1] S. Lin, K. Ying, S. Chen and Z. Lee, "Particle Swarm Optimization for Parameter Determination and Feature Selection of Support Vector Machines", *Expert Systems with Applications*, vol. 35, p. 1817–1824, 2008. http://dx.doi.org/10.1016/j.eswa.2007.08.088

[2] J. Wang, X. Wu and C. Zhang, "Support Vector Machines Based on *k*-Means Clustering for Real-time Business Intelligence Systems", *Int. J. Business Intell. Data Mining*, no. 1, pp. 54–64, 2005. http://dx.doi.org/10.1504/IJBIDM.2005.007318

[3] C.-W. Hsu, C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification", Technical Report, University of National Taiwan, Department of Computer Science and Information Engineering, pp. 1–12, 2003.

[4] P. F. Pai and W. C. Hong, "Forecasting Regional Electricity Load Based on Recurrent Support Vector Machines with Genetic Algorithms", *Electric Power Syst. Res.*, vol. 74, no. 3, pp. 417–425, 2005. http://dx.doi.org/10.1016/j.epsr.2005.01.006

[5] P. P. Feng and W. C. Hong, "Support Vector Machines with Simulated Annealing Algorithms in Electricity Load Forecasting", *Energy Conversion Manage*, vol. 46, no. 17, pp. 2669–2688, 2005. http://dx.doi.org/10.1016/j.enconman.2005.02.004

[6] P. F. Pai and W. H. Chiang, "Software Reliability Forecasting by Support Vector Machines with Simulated Annealing Algorithms", *Journal of Systems and Software*, vol. 6, no. 6, pp. 747–755, 2006. http://dx.doi.org/10.1016/j.jss.2005.02.025

[7] Y. Ren and G. Bai, "Determination of Optimal SVM Parameters by Using GA/PSO", *Journal of Computers*, vol. 5, no. 8, pp. 1160–1169, 2010. http://dx.doi.org/10.4304/jcp.5.8.1160-1168

[8] S. A. Medjahed, M. Ouali, A. Benyettou and A. S. Tamazouzt, "An Optimization-Based Framework for Feature Selection and Parameters Determina-

tion of SVMs", *International Journal of Information Technology and Computer Science*, vol. 7, no. 5, pp. 1–9, 2015. http://dx.doi.org/10.5815/ijitcs.2015.05.01

[9] R.-C. Chen and C.-H. Hsieh", Web Pages Classification Based on a Support Vector Machine Using a Weighed Vote Schema", *Expert Syst. Appl.*, no. 31, pp. 427–435, 2006. http://dx.doi.org/10.1016/j.eswa.2005.09.079

[10] C. Gold and P. Sollich, "Bayesian Approach to Feature Selection and Parameter Tuning for Support Vector Machine Classifiers", *Neural Netw.*, vol. 18, pp. 693–701, 2005. http://dx.doi.org/10.1016/j.neunet.2005.06.044

[11] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing Multiple Parameters for Support Vector Machines", *Mach. Learn.*, no. 46, pp. 131–159, 2002. http://dx.doi.org/10.1023/A:1012450327387

[12] R. Kohavi and G. John, "Wrappers for Feature Subset Selection", *Artif. Intell.*, vol. 97, pp. 273–324, 1997. http://dx.doi.org/10.1016/S0004-3702(97)00043-X

[13] T. Shon, Y. Kim, C. Lee, and J. Moon, "A Machine Learning Framework for Network Anomaly Detection Using SVM and GA", *Proceedings of the IEEE Workshop on Information Assurance and Security*, no. 2, pp. 176–183, 2005. http://dx.doi.org/10.1109/IAW.2005.1495950

[14] S. A. Medjahed, T. A. Saadi, A. Benyettou and M. Ouali, "Gray Wolf Optimizer for Hyperspectral Band Selection", *Applied Soft Computing*, vol. 40, pp. 178–186, 2016. http://dx.doi.org/10.1016/j.asoc.2015.09.045

[15] S. A. Medjahed, A. S. Tamazouzt, A. Benyettou and M. Ouali, "Binary Cuckoo Search Algorithm for Band Selection in Hyperspectral Image Classification", *IAENG International Journal of Computer Science*, vol. 42, no. 3, pp. 183–191, 2015.

[16] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. http://dx.doi.org/10.1023/A:1022627411411

[17] M. Rychetsky, "Algorithms and Architectures for Machine Learning Based on Regularized Neural Networks and Support Vector Approaches", Shaker Verlag GmBH, Germany, 2001.

[18] W. S. Noble, "What is a support vector machine?", *Nature Biotechnology*, vol. 24, pp. 1565–1567, 2006. http://dx.doi.org/10.1038/nbt1206-1565

[19] J. C. Platt, "Improvements to Platt's SMO Algorithm for SVM Classifier Design", *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001. http://dx.doi.org/10.1162/089976601300014493

[20] J. C. Platt, B. Schölkopf, C. Burges and A. Smola, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization", *Advances in Kernel Methods – Support Vector Learning*, pp. 185–208, 1999.

[21] G. W. Flake and S. Lawrence, "Efficient SVM Regression Training with SMO", *Journal Machine Learning*, vol. 46, no. 3, pp. 271–290, 2002. http://dx.doi.org/10.1023/A:1012474916001

[22] S. Lin, Z. Lee, S. Chen and T. Tseng, "Parameter Determination of Support Vector Machine and Feature Selection Using Simulated Annealing Approach", *Applied Soft Computing*, vol. 8, pp. 1505–1512, 2008. http://dx.doi.org/10.1016/j.asoc.2007.10.012

[23] M. Pardo and G. Sberveglieri, "Classification of Electronic Nose Data with Support Vector Machines", *Sens. Actuators B: Chem.*, vol. 177, pp. 730–737, 2005. http://dx.doi.org/10.1016/j.snb.2004.12.005

[24] J. Kittler, "Feature Selection and Extraction", Academic Press, New York, vol. 3, pp. 59–83, 1986.

[25] L. Rendell and R. Seshu, "Learning Hard Concepts through Constructive Induction: Framework and Rationale", *Compu. Intell.*, vol. 6, pp. 247–270, 1990. http://dx.doi.org/10.1111/j.1467-8640.1990.tb00298.x

[26] M. Creutz, "Microcanonical Monte Carlo Simulation", *Phy. Rev. Letters*, vol. 50, pp. 1411–1414, 1983. http://dx.doi.org/10.1103/PhysRevLett.50.1411

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, "Equation of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953. http://dx.doi.org/10.1063/1.1699114

[28] S. Kirkpatrick, J. C. D. Gelatt and M. Vecchi, "Optimization by Simulated Annealing", *Science*, vol. 220, no. 4598, pp. 671–680, 1983. http://dx.doi.org/10.1126/science.220.4598.671

[29] L. Hérault and R. Horaud, "Figure-Ground Discrimination: A Combinatorial Optimization Approach", *IEEE Trans. On Patt. Analys. and Machine Intelligence*, vol. 15, pp. 899–914, 1993. http://dx.doi.org/10.1109/34.232076

[30] R. Eglese and R. Horaud, "Simulated Annealing: a Tool for Operational Research", *Euro. J. of Op. Research*, vol. 46, pp. 271–281, 1996. http://dx.doi.org/10.1016/0377-2217(90)90001-R

[31] S. T. Barnard, "Stochastic Stereo Matching over Scale", *International Journal of Computer Vision*, vol. 3, no. 7, pp. 17–32, 1989. http://dx.doi.org/10.1007/BF00054836

[32] E. Zahara, S. Fan and D. Tsai, "Optimal Multi-Thresholding Using a Hybrid Optimisation Approach", *Pattern Recognition Letters*, vol. 26, pp. 1082–1095, 2004. http://dx.doi.org/10.1016/j.patrec.2004.10.003

[33] G. Dueck and T. Scheuer, "Threshold Accepting: a General Purpose Optimization Algorithm Superior to Simulated Annealing", *Journal of Computational Physics*, vol. 90, pp. 161–175, 1990. http://dx.doi.org/10.1016/0021-9991(90)90201-B

[34] P. Moscato and J. Fontanari, "Stochastic Versus Deterministic Update in Simulated Annealing", *Physics Letters A*, vol. 146, pp. 204–208, 1990. http://dx.doi.org/10.1016/0375-9601(90)90166-l

[35] D. S. Lee, V. S. Vassiliadis and J. M. park, "A Novel Threshold Accepting Meta-Heuristic for the Job-Shop Scheduling Problem", *Computer and Operations Research*, vol. 31, pp. 2199–2213, 2004. http://dx.doi.org/10.1016/S0305-0548(03)00172-2

[36] A. Fachat, K. H. Hoffmann and A. Franz, "Simulated Annealing with Threshold Accepting or Tsallis Statistics", *Computer Physics Communications*, vol. 132, pp. 232–240, 2000. http://dx.doi.org/10.1016/S0010-4655(00)00153-3

[37] C.-J. Ting and H.-J. Wang, "A Threshold Accepting Algorithm for the Uncapacitated Single Allocation Hub Location Problem", *Journal of the Chinese Institute of Engineers*, vol. 37, no. 3. 2014. http://dx.doi.org/10.1080/02533839.2013.781797

[38] X. Li, S. Peng, J. C. B. Lu, H. Zhang and M. Lai, "SVM – t-RFE: A Novel Gene Selection Algorithm for Identifying Metastasis-Related Genes in Colorectal Cancer Using Gene Expression Profiles", *Biochemical and Biophysical Research Communications*, vol. 419, no. 2, pp. 148–153, 2012. http://dx.doi.org/10.1016/j.bbrc.2012.01.087

[39] M. Tong, K. Liu, C. Xu and W. Ju, "An Ensemble of SVM Classifiers Based on Gene Pairs", *Computers in Biologyand Medicine*, vol. 46, no. 6, pp. 729–737, 2013. http://dx.doi.org/10.1016/j.compbiomed.2013.03.010

[40] S. Shah and A. Kusiak, "Cancer Gene Search with Data-Mining and Genetic Algorithms", *Computers in Biology and Medicine*, vol. 37, pp. 251–261, 2002. http://dx.doi.org/10.1016/j.compbiomed.2006.01.007

[41] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines", *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002. http://dx.doi.org/10.1023/A:1012487302797

[42] R. Mallika and V. Saravanan, "An SVM Based Classification Method for Cancer Data Using Minimum Microarray Gene Expressions", *World Academy of Science, Engineering and Technology*, vol. 62, pp. 543–547, 2010.

[43] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack and A. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays", *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999. http://dx.doi.org/10.1073/pnas.96.12.6745

[44] O. M. Soufan, D. Kleftogiannis, P. Kalnis and V. B. Bajic, "DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm", *PLoS One*, vol. 10, no. 2, pp. 1–23, 2015. http://dx.doi.org/10.1371/journal.pone.0117988

[45] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR Filter for Gene Selection", *IEEE Transactions on Nanobioscience*, vol. 9, no. 1, pp. 31–37, 2010. http://dx.doi.org/10.1109/TNB.2009.2035284

[46] S.-W. Lina, K.-C. Yingb, S.-C. Chenc and Z.-J. Lee, "Particle Swarm Optimization for Parameter Determination and Feature Selection of Support Vector Machines", *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2008. http://dx.doi.org/10.1016/j.eswa.2007.08.088

[47] G. Brown, A. Pocock, M.-J. Zhao and M. Luján, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection", *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

Received: July 2016
Revised: November 2016
Accepted: December 2016

Contact addresses:
Seyyid Ahmed Medjahed
Université des Sciences et de la Technologie
d'Oran Mohamed Boudiaf
USTO-MB, BP 1505, El M'naouer
31000 Oran
Algérie
e-mail: seyyid.ahmed@univ-usto.dz

Tamazouzt Ait Saadi
University of Le Havre
France
e-mail: tamazouzt.ait.saadi@univ-lehavre.fr

Abdelkader Benyettou
Université des Sciences et de la Technologie
d'Oran Mohamed Boudiaf
USTO-MB, BP 1505, El M'naouer
31000 Oran
Algérie
e-mail: aek.benyetou@univ-usto.dz

Mohammed Ouali
Computer Science Department
University of Sherbrooke
J1K2R1
Canada
e-mail: mohammed.ouali@usherbrooke.ca

Seyyid Ahmed Medjahed graduated with MSc in Engineering of Data and Knowledge from Oran University, Mathematics and Computer Science, Oran Algeria. He started teaching as Assistant Professor in 2012 at Relizane University Center, Algeria. His area of research interests includes: machine learning and image processing.

Tamazouzt Ait Saadi has been a teacher and researcher at Mostaganem University, Algeria since September 2003. She obtained her certificate of Engineer in Computer Science in 1989. In 2003, she received her MSc and continued her doctoral studies at Havre University, France.

Abdelkader Benyettou received his BSc in Engineering in 1982 from the Institute of Telecommunications of Oran and the MSc degree in 1986 from the University of Sciences and Technology of Oran, Algeria. In 1987, he joined teh Computer Sciences Research Center of Nancy, France, where he worked until 1991 on Arabic speech recognition by expert systems and received his PhD in Electrical Engineering in 1993 from the University of Sciences and Technology of Oran. His interests are in the area of speech and image processing, Arabic speech recognition, neural networks and machine learning. He has been the director of the Signal-Speech-Image – SIMPA Laboratory, Department of Computer Sciences, Faculty of Sciences, University of Sciences and Technology of Oran, since 2002.

Mohammed Ouali received his PhD in Real Time Informatics, Robotics and Automatic Control from Mines ParisTech, France, and the PhD in Mathematics and Computer Science from Sherbrooke University, Canada in 1999. Before joining teh academia, dr. Ouali had spent more than 15 years in the industry as a senior vision, pattern recognition and big data analysis.