Extensions of the SVM Method to the Non-Linearly Separable Data

Luminita STATE[1], Catalina COCIANU[2], Cristian USCATU[2], Marinela MIRCEA[2]
[1]Department of Mathematics and Informatics,
University of Pitesti, Pitesti, Romania[1]
[2]Department of Economic Informatics and Cybernetics,
The Bucharest University of Economic Studies, Bucharest, Romania [2]
lstate@clicknet.ro, ccocianu@ase.ro, cristiu@ase.ro, mmircea@ase.ro

The main aim of the paper is to briefly investigate the most significant topics of the currently used methodologies of solving and implementing SVM-based classifier. Following a brief introductory part, the basics of linear SVM and non-linear SVM models are briefly exposed in the next two sections. The problem of soft margin SVM is exposed in the fourth section of the paper. The currently used methods for solving the resulted QP-problem require access to all labeled samples at once and a computation of an optimal solution is of complexity $O(N^2)$. Several approaches have been proposed aiming to reduce the computation complexity, as the interior point (IP) methods, and the decomposition methods such as Sequential Minimal Optimization – SMO, as well as gradient-based methods to solving primal SVM problem. Several approaches based on genetic search in solving the more general problem of identifying the optimal type of kernel from pre-specified set of kernel types (linear, polynomial, RBF, Gaussian, Fourier, Bspline, Spline, Sigmoid) have been recently proposed. The fifth section of the paper is a brief survey on the most outstanding new techniques reported so far in this respect.
Keywords: *Support Vector Machines, Soft Margin Support Vector Machines, Kernel functions, Genetic Algorithms*

1 Introduction

Support Vector Machines (SVMs) belong to the class of most effective and popular classification learning tools [1], [2]. The learning problem for SVMs can be briefly described as follows. Let us denote by *S* a system of unknown input-output dependency, the unknown dependency being of deterministic/non-deterministic, linear/non-linear type. Besides, it is possible that the output is influenced by the observable input as well as a series of unobservable latent factors. Being given the lack of information about the input-output dependency of *S*, the most reasonable modeling should be in probabilistic terms. Unfortunately, in real world problems, there is no information about the underlying joint probability distribution corresponding to the (possible) non-linear dependency $y = f(x)$ between the high dimensional space of inputs *x* and the output space of *S*. The estimates of the unknown input-output dependency are obtained by a supervised distribution-free method, on the basis of a finite size training set consisting of input-output pairs of observations. The SVM methods belong to the classification class in the sense that the output space of it is a two-valuate domain, conventionally denoted by $\{-1,1\}$. Accordingly, a SVM can be viewed as a classifier discriminating between the inputs coming from two classes and the training set corresponds to a sequence of labeled inputs.

In spite of the fact that initially the people involved in the field of statistical machine learning believed that the SVM approaches are mostly of a theoretical value, the developments based on SVMs proved significant qualities from applicative perspective. So far a tremendous volume of efforts have been invested in research concerning SVMs, leading to a long list of publications in this area.

From mathematical point of view, the core problem of learning SVM is a quadratic programming problem [1], [3]. The research in the SVMs area focused mainly on designing fast algorithms for solving the QP optimization problem, refining the concepts aiming to extend the SVMs for discriminating between non-separable classes, and on developing

DOI: 10.12948/issn14531305/17.2.2013.14

mixture models resulted by combining the SVM with boosting type techniques [4].

The main aim of the paper is to briefly investigate the most significant topics of the currently used methodologies of solving and implementing SVM-based classifier. Following a brief introductory part, the basics of linear SVM and non-linear SVM models are briefly exposed in the next two sections. The problem of soft margin SVM is exposed in the fourth section of the paper.

The currently used methods for solving the resulted QP-problem require access to all labeled samples at once and a computation of an optimal solution is of complexity $O(N^2)$. Several approaches have been proposed aiming to reduce the computation complexity, as the interior point (IP) methods, and the decomposition methods such as Sequential Minimal Optimization – SMO, as well as gradient-based methods to solving primal SVM problem. Several approaches based on genetic search in solving the more general problem of identifying the optimal type of kernel from pre-specified set of kernel types (linear, polynomial, RBF, Gaussian, Fourier, Bspline, Spline, Sigmoid) have been recently proposed. The fifth section of the paper is a brief survey on the most outstanding new techniques reported so far in this respect.

2 The Basic Linear SVM Model

Let us assume that the inputs of *S* are represented by the values of *n* pre-specified attributes, that is the input space can be taken as $\mathbb{R}^n$, therefore the sequence of observations on input-output dependency of *S* can be represented as $\mathscr{S} = \left\{\left(x_i, y_i\right), x_i \in \mathrm{R}^n, y_i \in \{-1,1\}, 1 \le i \le N\right\}$, where for each component $\left(x_i, y_i\right) \in \mathscr{S}$, y_i is the output of *S* as the response to the input x_i.

We say that $\mathscr{S}$ is linearly separable if there exists a hyperplane that correctly separates the positive inputs from the negative ones. Obviously, since $\mathscr{S}$ is a finite set, if it is linearly separable, then the family of correctly separable hyperplanes is infinite. From intuitive point of view, being given the fact that the only information concerning the unknown input-output dependency of *S* is represented by the finite set $\mathscr{S}$, in order to assure good generalization capacities, the hyperplane should be as equidistant as possible to the positive and negative examples. The linear SVM implements a linearly parameterized classification decision rule, corresponding to a hyperplane almost equidistant to the subsamples labeled by 1 and -1 respectively. The classification decision rule is given by

$$\mathrm{h}_{w^*,b^*}: \mathbb{R}^n \to \{-1,1\} \quad , \quad \mathrm{h}_{w^*,b^*}(x) = \begin{cases} 1, w^{*T} x + b^* \ge 0 \\ -1, w^{*T} x + b^* < 0 \end{cases},$$

where the parameters (w^*, b^*) should be such that the hyperplane of equation $h_{b^*,w^*}(x) = 0$ separates the positive and the negative training examples from $\mathscr{S}$ with the largest "gap" between them (optimal margin linear classifier).

From mathematical point of view, an optimal margin classifier is a solution of the quadratic programming (QP) problem [1]

$$\begin{cases} \text{minimize } \frac{1}{2}\|w\|^2 \\ y_i\left(w^T x_i + b\right) \ge 1, \quad 1 \le i \le N \end{cases} \tag{1}$$

The dual problem of (1) is a QP problem on the objective function $\theta_d(\alpha) = \min_{w,b} L(w,b;\alpha)$

$$\begin{cases} \text{maximize } \theta_d(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \alpha_i \ge 0, 1 \le i \le N \end{cases} \tag{2}$$

If $\alpha^* = \left(\alpha_1^*, \alpha_2^*, ..., \alpha_N^*\right)^T$ is a solution of (2), then the optimal value of the parameter *w* is $w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$. If $\propto_i^* > 0$ for some *i*,

DOI: 10.12948/issn14531305/17.2.2013.14

then x_i is called a support vector. The bias term b cannot be determined by solving the SVM problem (1), a convenient choice of b being expressed in terms of the support vectors and w^* as follows.

According to the Karush-Kuhn-Tucker (KKT) complementarity conditions

$$\alpha_i^*\left(1-y_i\left(w^{*T}x_i+b\right)\right)=0, 1\le i\le N$$

hence the value of the parameter b should be such that $y_i\left(w^{*T}x_i+b\right)=1$ holds for any support vector x_i. Also, taking into account the constraints of (1), the value of the bias b should be set such that $y_i\left(w^{*T}x_i+b\right)\ge 1$ holds for all examples, that is $1-\min\limits_{\substack{i\\ y_i=1}} w^{*T}x_i \le b \le -1-\max\limits_{\substack{i\\ y_i=-1}} w^{*T}x_i$. Taking b as the middle of the interval, the parameters of the classification decision rule h_{b^*,w^*} are (w^*, b^*) where

$$b^*=-\frac{1}{2}\left\{\max_{\substack{i\\ y_i=-1}} w^{*T}x_i+\min_{\substack{i\\ y_i=1}} w^{*T}x_i\right\} \quad (3)$$

3 The Non-Linear SVM

Usually, in real world problems, there is no enough evidence to set suitable models for the classes of interest, the whole information concerning them being contained in the set of samples

$$\mathscr{S}=\left\{\left(x_i, y_i\right), x_i\in \mathrm{R}^n, y_i\in\{-1,1\}, 1\le i\le N\right\}$$

and it is either very difficult or even impossible to check whether $\mathscr{S}$ is linearly separable. Moreover, even when $\mathscr{S}$ happens to be linearly separable, there are no reasons to assume that the provenance classes are also linearly separable. Consequently, in case the provenance classes are not linearly separable, the use of any classification decision rule based on a linear-type approach would lead to poor results when it classifies new test data.

In order to cope with such a possibility, a non-linear transform of the given data onto a new space are hoped to provide more information about the provenance classes, therefore the parameters of a classification decision rule would be better tuned to separate the data coming from these classes, the ideal case being to find a non-linear transform such that in the new space the classes are linearly separable. Obviously, being given the finite type description of the classes represented by $\mathscr{S}$, it is impossible to guarantee that the classes are indeed linearly separable in the new space, therefore we at most could hope that $\mathscr{S}$ becomes linearly separable. In such a case, the main problem is to formulate an option concerning the functional expression of a particular non-linear transform without increasing significantly the computational complexity.

From mathematical point of view, the non-linear transform is a vector valued function $g:\mathbb{R}^n\to\mathcal{F}$, the image of $\mathscr{S}$ in the space $\mathcal{F}$ being given by the set of new representations of the given data $\mathscr{S}_g=\left\{\left(g\left(x_i\right), y_i\right), x_i\in \mathrm{R}^n, y_i\in\{-1,1\}, 1\le i\le N\right\}$. The transform g is referred as a feature extractor, and $\mathcal{F}$ is called the feature space, its dimension being not necessarily finite.

Assuming that the $\mathscr{S}_g$ is at least "almost linearly separable", it appears quite natural to use a linear classifier in the feature space, the separating surface in $\mathcal{F}$ between the images of the provenance classes being a hyperplane of equation $w^T g(x)+b=0$. Consequently, we get a non-linear classifier of particular type, where the decision rule combines a parameterized expression of linear type to a non-linear dependency of the values of the initial attributes of the form,

$$h_{b,w}:\mathrm{R}^n\to\{-1,1\},$$

$$h_{b,w}(x)=\begin{cases}1, & w^T g(x)+b\ge 0\\ -1, & w^T g(x)+b<0\end{cases}.$$

Note that the expression of $h_{b,w}$ can be viewed as a combination of a linear filter de-

DOI: 10.12948/issn14531305/17.2.2013.14

fined by the parameters $w \in \mathrm{R}^m, b \in \mathrm{R}$ and the non-linear filter represented by *g*.

The performance of the resulted classifier is essentially determined by the quality of the feature extractor *g*, the main problem becoming the design of a particular informational feature extractor. Another problem is related to the computational complexity involved by the estimation process of the classifier parameters and the classification of new data. The "kernel trick" provides a solution to these problems. It consists in selecting a function *K* that "covers" the explicit functional expression of *g*, therefore the evaluation of $h_{b,w}$ is performed exclusively in terms of *K*. Since *g* is "hidden" by *K*, the resulted feature space cannot be explicitly known, therefore its dimension may be even infinite. The core result in approaches of this type is the celebrated theorem due to Mercer [3], [5]. According to this results, if $K:\mathbf{R}^n \mathrm{x} \mathbf{R}^n \to [0,\infty)$ is a continuous symmetric function, the existence of a function *g* such that for any $x, x' \in \mathbf{R}^n$, $K(x,x') = g(x)^T g(x')$ holds, is guaranteed in case *K* satisfies a set of quite general additional conditions.

Some of the most frequently used kernels are presented in Table 1.

Table 1. Examples of kernels

	$K(x,x')$
Linear	$x^T x'$
Polynomial of degree *d*	$(x^T x'+1)^d, \quad d \geq 1$
Gauss RBF	$\exp\left(-\gamma \|x-x'\|^2\right), \gamma > 0$
Exponential RBF	$\exp\left(-\gamma \|x-x'\|\right), \gamma > 0$

Since $\mathscr{S}_g$ is finite, in case it is linearly separable in the space $\mathcal{F}$, there are an infinite number of classifiers $h_{b,w}$ that separate the given data without errors. Let us assume that for a selected kernel *K*, $\mathscr{S}_g$ is linearly separable. Then we could search for a linear classifier in $\mathcal{F}$ that offers the best generalization capacity in the sense that it still classifies at least "almost correctly", new, unseen yet examples. This requirement may be formulated as the task to determine the parameters (w,b) such that the hyperplane of equation $h_{b,w}(x)=0$ is as equidistant as possible to all images of the training data in the feature space, therefore it is aimed to separate the examples of $\mathscr{S}_g$ with the largest "gap" between positive and negative examples. Such a classifier is referred as an optimal margin classifier.

Stated in mathematical terms, the problem is formulated as follows. Let *K* be a kernel and *g* be the induced feature extractor, $K(x,x') = g(x)^T g(x')$. An optimal margin classifier is a solution of the constrained QP problem [3],

$$\begin{cases} \text{minimize } \frac{1}{2}\|w\|^2 \\ y_i\left(w^T g(x_i)+b\right) \geq 1, \quad 1 \leq i \leq N \end{cases} \tag{4}$$

its corresponding dual problem being the constrained QP problem imposed on the objective function $Q(\alpha)$,

DOI: 10.12948/issn14531305/17.2.2013.14

$$\begin{cases} \max Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_j, x_i) \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \alpha_i \geq 0, 1 \leq i \leq N \end{cases} \quad (5)$$

According to the developments presented in the previous section, if $w^* = \sum_{i=1}^{N} \alpha_i^* y_i g(x_i)$, and $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)^T$ is a solution of (5), then the optimal parameters are

$$b^* = -\frac{1}{2} \left\{ \max_{\substack{i \\ y_i=-1}} \sum_{j=1}^{N} \alpha_j^* y_j K(x_j, x_i) + \min_{\substack{i \\ y_i=1}} \sum_{j=1}^{N} \alpha_j^* y_j K(x_j, x_i) \right\}. \quad (6)$$

Note that although apparently the parameters depend on the hidden feature extractor *g*, the resulted classifier is based exclusively on the values of the particular selected kernel,

$$h_{w^*,b^*}: \mathbb{R}^n \to \{-1,1\}, \; h_{w^*,b^*}(x) = \begin{cases} 1, \sum_{j=1}^{N} \alpha_j^* y_j K(x_j, x) + b^* \geq 0 \\ -1, \sum_{j=1}^{N} \alpha_j^* y_j K(x_j, x) + b^* < 0 \end{cases} \quad (7)$$

4 Soft Margin SVM

The aim of the developments presented in this section is to present a modified approach in order to cope with cases when the particular kernel fails to extract enough information from data to discriminate without errors between the positive and negative examples, that is $\mathscr{S}_g$ is not linearly separable in $\mathcal{F}$. In such a case we could search for a classifier $h_{b,w}$ that classifies at least "as correct as possible" the data. This idea can be formulated in mathematical terms as follows.

Let g be a particular feature extractor and $h_{w,b}: \mathbb{R}^n \to \{-1,1\}$,

$$h_{w,b}(x) = \begin{cases} 1, \sum_{j=1}^{N} \alpha_j y_j K(x_j, x) + b \geq 0 \\ -1, \sum_{j=1}^{N} \alpha_j y_j K(x_j, x) + b < 0 \end{cases}$$

a classifier of parameters *w* and *b*.

We include in the model a set of slack variables $\xi_1, ..., \xi_N$, defined by $\xi_i = \max\{0, 1 - y_i(w^T g(x_i) + b)\}$. Obviously, for any misclassified example (x_i, y_i), the value of ξ_i expresses the magnitude of the error committed by $h_{b,w}$ with respect to (x_i, y_i). The overall importance of the cumulated errors is expressed as $F\left(\sum_{i=1}^{N} \xi_i^t\right)$, where F is a convex and monotone increasing function and $t>0$ is a weight parameter.

An optimality criterion can be expressed in terms of an objective function that combines additively $\|w\|^2$ with the overall effect of the errors, for instance $\frac{1}{2}\|w\|^2 + CF\left(\sum_{i=1}^{N} \xi_i^t\right)$.

In this case an optimal classifier (w^*, b^*) is a solution of the constrained QP-problem

DOI: 10.12948/issn14531305/17.2.2013.14

$$\begin{cases} \text{minimize } \frac{1}{2}\|w\|^2 + CF\left(\sum_{i=1}^{N}\xi_i^t\right) \\ y_i\left(w^T g(x_i)+b\right)\geq 1-\xi_i, \quad 1\leq i\leq N \\ \xi_i \geq 0, \quad 1\leq i\leq N \end{cases} \qquad (8)$$

where C is a conventionally selected constant.

Unfortunately, stated in this general form, the problem (8) cannot be solved, but, for particular functional expressions of F and the weight parameter t, its solution can be computed explicitly. The simplest model uses F(u)=u and t=1, the problem (8) becoming the constrained QP-problem

$$\begin{cases} \text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i, \\ y_i\left(w^T g(x_i)+b\right)\geq 1-\xi_i, \quad 1\leq i\leq N \\ \xi_i \geq, \quad 1\leq i\leq N \end{cases} \qquad (9)$$

Using similar arguments as in case of (1), the dual QP-problem of (9) is

$$\begin{cases} \text{maximize } Q(\alpha)=\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_j,x_i) \\ \sum_{i=1}^{N}\alpha_i y_i = 0 \\ \alpha_i \geq 0, 1\leq i\leq N \end{cases}$$

(10)

The parameters of an optimal hyperplane are

$$w^* = \sum_{i=1}^{N}\alpha_i^* y_i g(x_i)$$

$$b^* = -\frac{1}{2}\left\{\max_{\substack{i \\ y_i=-1}}\sum_{j=1}^{N}\alpha_j^* y_j K(x_j,x_i) + \min_{\substack{i \\ y_i=1}}\sum_{j=1}^{N}\alpha_j^* y_j K(x_j,x_i)\right\}$$

where $\alpha^* = (\alpha_1^*,\alpha_2^*,...,\alpha_N^*)^T$ is a solution of (10) and the expression of the decision function of the classifier is given by (7).

5 Methods of Learning Parameters of SVMs

As it was pointed out in the previous sections, the support vector machines are a class of linear or kernel-based binary classifiers that attempt to maximize the minimal distance between each member of the class and separating surface. In most cases, the task of learning a support vector machine is cast as a constraint quadratic programming problem. The currently used methods for solving the resulted QP-problem require access to all labeled samples at once and a computation of an optimal solution is of complexity $O(N^2)$. Several approaches have been proposed aiming to reduce the computation complexity, as the interior point (IP) methods [6], and the decomposition methods such as Sequential Minimal Optimization – SMO [7], as well as gradient-based methods to solving primal SVM problem. These methods exhibit convergence rate independent of the number of samples, which particular useful in case of large datasets.

A long series of generalizations and improvements have also been recently proposed by many authors. For instance, in [8] a parallel version of SMO is proposed to accelerate the SVM training. Also, boosting algorithms were proved to be closely related to the primal formulation for SVM [9].

IP methods cast SVM learning formulated as a QP-problem subject to linear constraints by replacing the constraints with a barrier function yielding to a sequence of unconstrained problems which can be optimized efficiently using Newton or Quasi-Newton methods. To overcome the quadratic memory requirement of IP methods, several decomposition methods such as SMO [7], SVM^{light} [10], and SVM-Perf [11] switch to the dual representation of the SVM QP optimization problem and employ an active set of constraints thus working on a subset of dual variables. The algorithms belonging to this family are fairly simple to implement and entertain good asymptotic convergence properties, but the time complexity is typically super linear in the training set size N. Moreover, since decomposition methods aim to maximize the dual objective function, they often result in a rather slow convergence rate to the optimum of the primal objective function.

The SMO algorithm [7] allows to solve the SVM-QP dual problem without extra-matrix

DOI: 10.12948/issn14531305/17.2.2013.14

storage. The idea is to use the Osuna's theorem [12] for decomposing the overall QP problem into smaller size QP sub-problems, the smallest size optimization problem being solved at each step.

Unconstrained gradient methods were very common in solving optimization problems until the emergence of the ultra-fast IP methods. While gradient –based methods are known to exhibit slow convergence rate, the computational demands imposed by large scale classification and regression problems of high dimension feature space revived the theoretical and applied interest in gradient methods.

A refined method combining gradient ascent algorithm with decomposition scheme including heuristic parameters for solving the dual problem of nonlinear SVM was introduced in [13], [14]. The proposed refinement consists of the use of heuristically established weights in correcting the search direction at each step of the learning algorithm that evolves in the feature space. The use of weights is justified by the idea of getting better tuning effect to the particular training sequence. The tests pointed out good convergence properties and, moreover, the proposed modified variant proved higher convergence rates as compared to the Platt's SMO algorithm.

The main objectives of the research were to evaluate the influence of magnitude of the exponential RBF kernel parameter on the number of iterations required to obtain significant accuracy, as well as on the magnitude of the inter-sample distance, and sample variability and separability degrees. The experimental analysis aimed also to derive conclusions on the recognition rate as well as on the generalization capacities. All linear classifiers proved almost equal recognition rate and generalization capacities, the difference being given by the number of iteration required for learning the separating hyperplanes.

The tests pointed out that the variation of the recognition rates depends also on the inner structure of the classes from which the learning data come as well as on their separability degree. Consequently, the results are encouraging and entail future work toward extending these refinements to multi-class classification problems and approaches in a fuzzy-based framework.

The Pegasos (Primal Estimated sub-GrAdient SOlver for SVM) algorithm [15] is an improved stochastic sub-gradient method that uses fixed size subsamples of the training set to compute approximate sub-gradient, two concrete algorithms that are closely related to the Pegasos algorithm being the NORMA algorithm [16] and a stochastic gradient algorithm proposed by Zhang [17]. At it is reported in [15], on the basis of a large series of tests, the Pegasos algorithm is substantially faster than SVM-Perv.

Boosting is a meta-algorithm for supervised learning that combines several weak classifiers that can label examples only slightly better than random guessing into a single strong classifier with far better classification accuracy. Some of the most successful boosting methods in problems as text recognition, filtering, feature selection and face recognition are AdaBoost and its variants [18], [19].

Recently, a new boosting type algorithm based on Pegasos and stochastic gradient descent-based SVM training method was proposed and its performance was experimentally evaluated for both the linear and the kernel-based case [4]. The algorithm is a two-phases SVM allowing the use of gradient descent-based methods without the need to fine-tune the kernel parameters. A long series of tests proved that the algorithm is much more efficient than the kernel-based SVM algorithms, both in terms of computing and storage requirements, due to the fact that each weak classifier requires only a single inner product calculation, while the evaluation of kernel expansion terms involved by the use of NORMA and Pegasos algorithms are substantially more computationally expensive to achieve the same accuracy levels. Also, the combination of boosting and online SVM training has the potential to create efficient algorithms that outperform standard training algorithms when the kernel parameters are not known.

DOI: 10.12948/issn14531305/17.2.2013.14

Moreover, one of the core problem in improving the efficiency of the classifier is to identify the optimal types of kernels and for each type of kernel its optimal parameters and then apply the standard techniques for solving the resulted QP problem. In other words, in these approaches, the problem is to tune the type of kernel together with its parameters to the particular problem at hand. In [14] is reported an experimental analysis on the parameter γ of the RBF type kernels, the tests being performed on simulated data. Several approaches based on genetic search in solving the more general problem of identifying the optimal type of kernel from pre-specified set of kernel types (linear, polynomial, RBF, Gaussian, Fourier, Bspline, Spline, Sigmoid) were reported in [20] and [21].

A new class of approaches contains algorithms, referred as Genetic Algorithms-SVM (GA-SVM or GSVM), and Hybrid Genetic Algorithms SVM (HGA-SVM). In the novel HGA-SVM model [20], the type of kernel and the parameters of SVM are dynamically optimized by implementing an evolutionary process, the approach simultaneously determining the appropriate type of kernel function and optimal kernel parameter values for optimizing the SVM model to fit various datasets. The types of kernel functions (RBF kernel, polynomial kernel and linear kernel) together with all the values of the parameters are directly encoded into the chromosomes using integers and real-valued numbers respectively. Therefore each chromosome is represented by a triple whose entries are the particular type of kernel function, and the first and second parameter values in this particular chromosome in population respectively, the type of the kernel being represented by an integer number, the second and the third parameters coded in terms of real valued numbers. The proposed model can implement either the roulette-wheel or the tournament method for chromosome selection. The chromosomes are modified using the crossover operator and boundary mutation method introduced by Adewuya [22], the method being of elitist type in the sense that only the one best chromosome in each generation is allowed to survive in the succeeding generation.

In [21] the GSVM algorithm was applied for effective detection of the Doppler heart sounds. The GSVM algorithm is a genetic algorithm-based SVM classification technique defined in terms of a kernel function type, kernel function parameters, and the soft margin constant C that represents the penalty parameter of support vector machine classifier. The proposed model uses a 28-bits chromosome, grouped as follows. The genes of the first group are the kernel function type represented by 3 bits, the value of the C parameter (3 bits), the value of Gaussian kernel parameter (7 bits). The genes belonging to the second set represent the value of the polynomial kernel parameter (2 bits), the value of the Sigmoid kernel parameter (2 bits), the value of the Bspline kernel parameter (2 bits). The next group of genes encode the value of the RBF (7 bits) and the value of the Fourier kernel parameter (2 bits). The fitness function used in GSVM is based on classification accuracy of the trained SVM classifier, the initial population consisting of 30 chromosomes, each of chromosomes having 28 bits. The new populations are generated in the search initiated by the GSVM algorithm using the crossover and mutation operations. The most important 15 chromosome in population are saved for composing the next population, the chromosomes that have low fitness values being eliminated. A subsample of 40% portion of the optimum chromosomes are randomly selected and subjected to crossover operator. Therefore 10 chromosomes are subjected to crossover operator, 5 bits of the each of the random 2 chromosomes are randomly selected and replaced each other, yielding to a 10 new chromosomes. The bit inversion method is used as a mutation operator and it is applied to 0.4% portion of the total bits numbers of other 5 chromosomes.

6 Conclusions

Support Vector Machines are maybe the most effective and popular classification

DOI: 10.12948/issn14531305/17.2.2013.14

learning tool, the task of learning a SVM being cast as a constrained QP-problem. The respective dual problem is also a constrained QP-problem whose solution can be approximated by an adaptive learning scheme to assure the maximization of the objective function. One of the main benefits of SVMs is their ability to incorporate and construct non-linear predictors using kernels which satisfy Mercer's conditions, the common approach for solving the optimization problem for SVM when kernels are employed being to switch to the dual problem and find the optimal set of dual variables. The performance of the resulted classifier is essentially conditioned by the quality of the feature extractor induced by the selected kernel. The most frequently used kernels belong to polynomial class, or are of exponential type, as for instance Gaussian kernels.

Some of the trends in optimizing the learning process of SVM-based classifier aim to design hybrid architectures and to develop methods "tuned" to the particular problem by including special tailored genetic algorithms.

References

[1] V. Vapnik, *Statistical learning theory*, John Wiley, New York, 1998.

[2] J. Shawe-Taylor and N. Cristianini, *Support vector machines and other kernel-based learning methods*, Cambridge University Press, UK, 2000.

[3] S. Abe, "Support vector machines for pattern classification," *Advances in Pattern Recognition*, 2010, pp. 1-473.

[4] V. Yugov and I. Kumazava, "Online Boosting Algorithm Based on Two-Phase SVM Training," *ISRN Signal Processing*, Volume 12, 2012.

[5] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Proc. Roy. Soc. London Ser. A* , 83, 1908, pp. 69–70.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[7] J. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," In: *Advances in Kernel Methods – Support Vector Learning*, Cambridge, MA, MIT Press, 1999.

[8] L.J. Cao, S.S., Keerthi, C.J. Ong, P. Uvaraj, X.J. Fu and H.P. Lee, "Developing parallel sequential minimal optimization for fast training support vector machine," *Neurocomputing* 70, 2006, pp. 93-104.

[9] G. Ratsch, B. Schőlkopf, S. Mika and K.R. Muller, "SVM and boosting one class," *Technical Report*, 2000.

[10] T. Joachims, "Making large-scale SVM learning practical," *Advances in Kernel Methods – Support Vector Learning*, 1998.

[11] T. Joachims, "Training linear SVMs in linear time," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[12] E. Osuna, R. Freund and F. Girosi, "An improved training algorithm for support vector machines," in *Proc. the IEEE Workshop. Neural Networks for Signal Processing*, 1997.

[13] C. Cocianu, L. State and P. Vlamos, "A New Method for Learning the Support Vector Machines," in *Proc. the 6th International Conference on Software and Data Technology*, 2011, pp. 365-370.

[14] C. Cocianu and L. State, *Kernel-Based Methods for Learning Non-Linear SVM*, *Economic Computation and Economic Cybernetics Studies and Research*, no. 1/2013, ISSN 0424-267X.

[15] S. Shalev-Shwartz, Y. Singer and N. Srebro, "Pegasos: primal estimated sub-GrAdient sOlver for SVM," in *Proceedings of the 24th ACM International Conference on Machine Learning (ICML '07)*, 2007.

[16] J. Kivinen, A.J. Smola and R.C. Williamson, "Online learning with kernels," *IEEE' TSP*, 52, 2002.

[17] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in ICML '04 *Proceedings of the 21st International Conference on Machine Learning*, 2004.

DOI: 10.12948/issn14531305/17.2.2013.14

[18] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, 1997.
[19] R.E. Schapire, Y. Freund, P. Bartlett and W.S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Annals of Statistics*, vol. 6, no. 5, 1998.
[20] C.H. Wu, Y. Ken and T. Huang, "Patent classification system using a new hybrid genetic algorithm support vector machine," *Applied Soft Computing* 10, 2010.
[21] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier", *Expert Systems with Applications* 36, 2009, 10618-10626
[22] A.A. Adewuya, *New Methods in Genetic Search with Real-Valued Chromosomes*, M.S. MIT, MA, 1996

Luminiţa STATE, Professor, PhD, currently working with University of Pitesti, Department of Mathematics and Computer Science. Competence areas: artificial intelligence, machine learning, statistical pattern recognition, digital image processing. Research in the fields of machine learning, pattern recognition, neural computation. Author of 15 books and more than 120 papers published in national and international journals.

Cătălina-Lucia COCIANU, Professor, PhD, currently working with Academy of Economic Studies, Faculty of Cybernetics, Statistics and Informatics, Department of Informatics in Economy. Competence areas: machine learning, statistical pattern recognition, digital image processing. Research in the fields of pattern recognition, data mining, signal processing. Author of 15 books and more than 90 papers published in national and international journals.

Cristian Răzvan USCATU, Associated Professor, PhD, currently working with the Academy of Economic Studies, Faculty of Cybernetics, Statistics and Informatics, Department of Informatics in Economy. Competence areas: computer programming, data structures. Author/co-author of 10 books and more than 40 papers published in national and international journals.

Marinela MIRCEA, Lecturer, PhD, currently working with Academy of Economic Studies, Faculty of Cybernetics, Statistics and Informatics, Department of Informatics in Economy. Competence areas: information system, Business Intelligence. Research in the fields of information system, Business Intelligence, classification techniques. Author of 6 books and more than 50 papers published in national and international journals.

DOI: 10.12948/issn14531305/17.2.2013.14