Universal Learning Technology: Support Vector Machines

By Vladimir VAPNIK*

ABSTRACT This paper describes the Support Vector Machine (SVM) technology, its relation to the main ideas of Statistical Learning Theory, and shows a universal nature of SVMs. It also contains examples that show a high level of generalization ability of SVMs.

KEYWORDS Statistical Learning theory, Support Vector Machines, Pattern recognition

1. INTRODUCTION

The problem of learning and generalization is one of the oldest in the natural science. Its discussion started more than two thousand years ago when philosophers for the first time started to analyze phenomena of Nature. However, in this paper we will discuss only one classical principle introduced more than five hundred years ago, the so called Occam razor principle.

Our main discussion will start from the 1930th when three important events took place: K. Popper introduced the necessary condition for generalization, the so-called principle of non-falsifiability; A. N. Kolmogorov introduced axiomatization of theory probability; and statistics; R. Fisher introduced classical model of applied statistics. Unfortunately the classical model of applied statistics does not capture very important situations when one is trying to construct a predictive rule using a restricted number of examples with a large number of features (learning from a small number of examples in high dimensional space).

A model of high dimensional learning is based on different theory, the so-called Statistical Learning Theory (or VC theory) that was constructed to overcome the "curse of dimensionality" of classical statistics.

· Two Models of Learning: Generative Model and Predictive Model

The basic mathematical model of learning can be described as follows (see **Fig. 1**): there exists an object (say a human being) that can classify observed vectors x_i into two categories by indexing them with scalar y_i that can take only two values 1 and –1. Let us make an agreement that $y_i = 1$ means that vector x_i belongs to specific category of interest and value $y_i = -1$ means that it does not belong to such category.

Suppose that Black Box (BB) classifies ℓ examples as follows

$$(x_1, y_1), \ldots, (x_\ell, y_\ell) \qquad (1)$$

The problem is to create a computer program that using these examples constructs such a rule that classifies new (unseen) vectors x_j approximately as well as BB.

Example: A doctor examines pictures of mammograms and classifies corresponding patients into two categories: patients that have breast cancer $y_i = 1$ and patients that have no breast cancer $y_i = -1$. Since the pictures have a digital description x_i that means that doctor defines set (1). The problem is using a computer to construct such a rule that classifies new (unseen) vectors x_j approximately as well as the doctor (BB) does.

Generative model of learning: The main question here is: Which mathematical problem has to solve the computer to construct a desired decision rule. The idea of the classical statistics approach is: Try to recover which rule uses the Black Box to classify data. Generally speaking, the Black Box

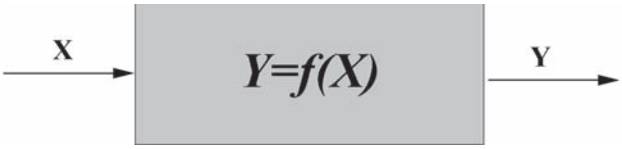

Fig. 1 Basic mathematical model of learning.

*NEC Laboratories America, Inc.

generates classification y_i according to some conditional probability function $p(y|x)$. The classical statistics suggests to estimate the rule by recovering a conditional probability function in space of data x. That means the classical approach is based on recovering Generative Model.

In the 1960s when the first fast computers appeared it was realized that it was very difficult to estimate conditional probability for the following two reasons:

1) Description of a probability density function in high dimension spaces requires a lot of parameters that should be estimated.
2) Estimating density function is an ill-posed problem and therefore requires a lot of data (per parameter) to estimate function reasonably well.

This fact gave a reason to declare a "curse of dimensionality" in the classical approach. The methods of estimating classification rules in high-dimensional cases required too many examples and therefore could not be justified within the framework of classical statistics if the ratio of the number examples to the number of parameters were small (say less than 10).

Predictive model of learning: Therefore, in the 1970s an alternative approach to the learning based on Statistical Learning Theory (or the VC theory) was developed [1,2].

The alternative approach gives up the ambitious goal to estimate the rule used by the Black Box (the generative model of data). Instead, it suggests estimating the predictive model from a given set of models

$$y=f(x,\alpha),\ \alpha\in\Lambda. \tag{2}$$

Statistical Learning Theory suggests choosing from the set of rules (2) one that minimizes the number of mis-classifications with the Black Box. It address two questions:

1) How to choose from a given set of admissible rules (2) the best one.
2) How to construct a set of functions that has a function which approximates Black Box rule well.

In the next section we describe answers to these questions. In the meantime, let us discuss the difference between the two settings: the classical one and SLT.

Suppose that the Black Box uses the following classification rule: the data that are above the line (**Fig. 2**) belong to the class of interest while data below this line do not.

Consider the separating curve presented on this figure. This curve is very different from the line that the Black Box uses. It is not a good estimation of the Black Box rule. Therefore from the classical point of view this solution is unacceptable.

However, using this curve one can make predictions that are not much worse than predictions given by the Black Box rule. Therefore from a Statistical Learning Theory point of view it is a good approximation.

A classical solution is more demanding and therefore requires more information to be solved well. In fact, a classical solution requires solving ill posed problems where among a set of functions that explain empirical data well one has to find one that approximates the Black Box function well. Statistical Learning Theory gives up this ambitious goal by looking for any rule that makes a good prediction. Therefore, it can be applied to some problems where classical methods cannot be used.

Remark: Two approaches used by classical statistics and SLT can be considered as a reflection of two directions in the Philosophy of Science: Realism and Instrumentalism. Realism declares that the goal of Science is to find real law of Nature, while Instrumentalism declares that the goal of Science is to create instruments for dealing with Nature (in particular for prediction).

2. STATISTICAL LEARNING THEORY

Statistical Learning Theory provides a complete answer to the following question: When using observations (1) one can (in a given set of admissible rules)

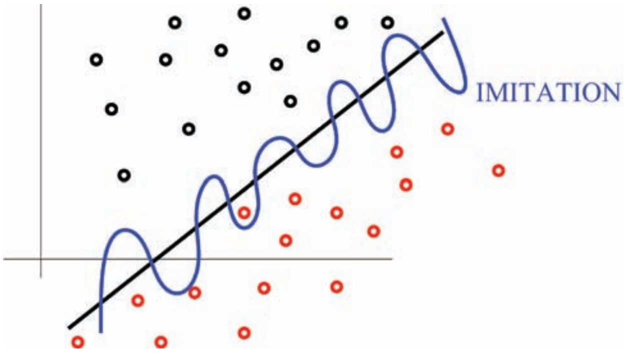

Fig. 2 Classification rule by Black Box.

find the rule that guarantees (on unseen test examples) the smallest number of mis-classifications with respect to the Black Box rule. In other words, when one can achieve the generalization.

It was shown that two and only two factors are responsible for the generalization:

1) How well the chosen decision rule classified the training data (1).
2) What is the capacity (diversity) of the admissible set of functions (2) from which the rule was chosen.

That is: ① How many misclassification does the chosen rule makes on training examples (1). ② How diverse is a set of admissible functions.

Several concepts of diversity (capacity) of the sets of functions were introduced that are important for different mathematical settings of the concept of generalization. The most important among them are: the VC entropy that describes generalization for a given environment and the VC dimension that describes generalization for all possible environments. In the Section 2.2 we will introduce the concept of VC dimension that plays an important role in SVM theory. VC dimension is an integer number h and can be introduced for any set of functions.

Below, using the concept of VC dimension h of admissible set of functions we will describe the main results of SLT.

Proposition 1: In order to obtain generalization it is necessary and sufficient that VC dimension of the set h of admissible rules be finite.

If VC-dimension is infinite than there exists a rule that separates training data well and makes a wrong prediction.

Proposition 2: For any fixed number of training data ℓ with probability $1-\eta$ the inequality

$$P(test\ err.) \leq Fr(train.\ err.) + \Phi\left(\frac{\mathrm{h}}{\ell}, \frac{-1\mathrm{n}\eta}{\ell}\right) \qquad (3)$$

holds true, where $P(test\ err.)$ is probability of misclassification of test examples, $Fr(train.\ err.)$ is frequency of misclassification of the training examples, and $\Phi(\cdot)$ is the known function of confidence interval.

This inequality is a basis for creating efficient algorithms.

2.1 Structural Risk Minimization

Inequality (3) inspired the following method of minimizing probability of the predictive error. Consider the structure defined on the admissible set of functions f(x, α), $\alpha \in \Lambda$

$$S_1 \in \ldots . \in S_{\mathrm{n}}$$

where S_h is subset of admissible functions with VC dimension equal h. Let us choose such subset S_* and such function $f(x, \alpha_0^*)$ in this subset that minimize the right hand side of inequality (3). Such method for choosing desired function we call Structural Risk Minimization (SRM) method.

The structural risk minimization method possesses the following remarkable property:

Proposition 3: Let structure be defined on a sufficiently large set of functions such that the VC dimension of entire set of functions is infinite. Then SRM method is strongly universally consistent.

That means that with increasing number of observation the SRM method converges with probability one to the best possible solution independent of the set of function on which the structure was constructed.

The SVM technology is a realization of the SRM principle.

2.2 Definition of the VC Dimension

Consider a set of vectors

$$x_1, \ldots, x_\ell \,. \qquad (4)$$

If vectors (4) are in general position, there exist exactly 2^ℓ different ways to divide this set into two subsets.

We say that vectors (4) cannot falsify a set of indicator functions $f(z, \alpha), \alpha \in \Lambda$ if all 2^ℓ separations of (4) are possible by this set of indicators.

Definition: A set of indicator functions $f(z, \alpha)$, $\alpha \in \Lambda$ has VC dimension h if:

- There exist h vectors that cannot falsify this set.
- Any $h+1$ vectors falsify this set.

Figure 3 shown that three vectors do not falsify set of lines in a plain and any four vectors falsify. The VC dimension of the set of line is three.

Remark: VC dimension is a mathematical reflection of K. Popper concept of non-falsifiability which in the philosophy of Science is considered as a necessary condition for the generalization.

2.3 Occam's Razor or Non-Falsifiability

According to VC theory the necessary and sufficient conditions of generalization for methods that minimize the number of training errors depend on how diverse the set of function was from which one chooses a desired solution. It requires controlling the VC dimension.

A classical statistics approach inspired by Occam razor principle:

- Entities (features) should not be multiplied beyond necessity.

requires to control the number of parameters.

Since the value of VC dimension (but not the number of parameters) forms the necessary and sufficient conditions and defines the generalization bounds for the predictive learning (see Proposition 1 and Proposition 2) the crucial question is:

- Does VC dimension coincides with the number of free parameters?

The answer is no. VC dimension can coincide with the number of free parameters of a set of admissible functions, can be much larger than the number of free parameters, or can be much smaller. The last case has important (both theoretical and practical) implications. In a set of functions with large (even infinite) number of free parameters, but with small VC dimension one can obtain a good generalization (since according to Proposition 2 the generalization depends on VC dimension). In such situations, one can achieve in high dimensional cases good generalization using not too large number of examples.

Therefore, it is very important to describe a set of functions where the VC dimension is much smaller than the number of parameters. It so happens that such set of functions can be constructed based on the set of linear functions.

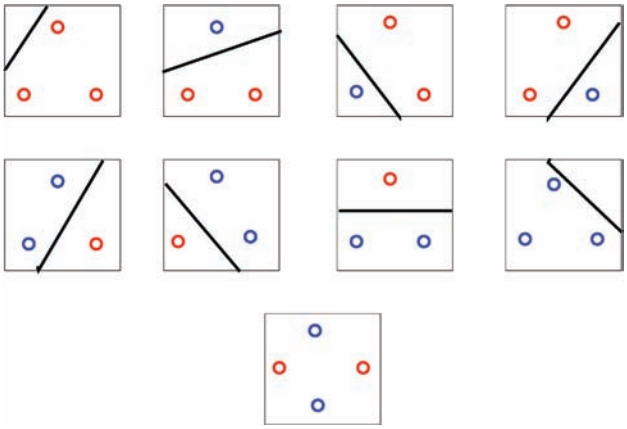

Fig. 3 Three vectors do not falsify set of lines in a plain and any four vectors falsify.

2.4 Large Margin Hyperplanes

Let us consider the set of separating hyperplanes

$$y=\text{sign}\{(w,x)+b\} \tag{5}$$

where $x \in R^n$ belongs to the n dimensional space, w is a n dimensional vector of free parameters, b is a scalar, and sign(u) is an indicator function: it equals 1 if $u \geq 0$ and equals -1 if $u < 0$. Therefore, the set of function (5) has $n + 1$ free parameters. One can prove that VC dimension of such set of functions equals $n + 1$ (coincides with the number of free parameters).

Now let us consider a set of hyperplanes separating data with the margin equals Δ

$$y=\text{sign}\Delta\{(w,x)+b\} \tag{6}$$

Here value $y = 1$ if

$$(w,x)+b\geq\Delta,$$

$y = -1$ if

$$(w,x)+b<-\Delta,$$

and $y = 0$ (or undefined) otherwise.

Proposition 4: Suppose that $|x| \leq R$ then the VC dimension of the set of separating hyperplane with the margin Δ has a bound

$$h \leq \min\left\{\frac{R^2}{\Delta^2}, n\right\}+1 \tag{7}$$

where min{a, b} is the smallest value from these two.

Note that if Δ is sufficiently small, the VC dimension of Δ margin separating hyperplanes coincides with the VC dimension of the set of separating hyperplanes and equals the number of free parameters n + 1.

If the margin is large, the VC dimension can be much smaller than the number of free parameters. We will use this fact to construct SVMs.

3. MAIN IDEAS OF SVMs

To construct large margin separating rules we will act in opposite to the Occam razor recommendation. We will increase the number of free parameters. We will map n dimensional input vectors $x \in X^n$ in a Hilbert (infinite) dimensional space $z \in Z^\infty$

$$x \longrightarrow z$$

creating from the training set (1) a new training set

$$(y_1, z_1), \ldots, (y_\ell, z_\ell) \tag{8}$$

Then we will construct Δ-margin separating hyperplane in space Z.

Let our data be bounded by the value R. Without restriction of generality we assume that $R = 1$. In this case (according to Proposition 4) the VC dimension of the set of linear (in Z space) Δ-margin classifiers

$$y=\mathrm{sign}\{(w,z)+b\}$$

is bounded by the value $h \le 1/\Delta^2$, if equality

$$(w,w) = 1/\Delta^2$$

holds true. By choosing the value of margin Δ, one can control the VC dimension of the admissible set. To minimize the guarantee error bound (3) one has to choose from the set of Δ margin separating hyperplanes one with the smallest number of training errors.

This idea leads to a simple optimization problem: Minimize the functional

$$R = (w,w) + C \sum_{i=1}^{\ell} \xi_i \tag{9}$$

(where C is a constant that defines h) subject to constraints

$$y_i((\mathrm{w,z})+\mathrm{b}) \ge 1-\xi_i, \qquad i=1, \ldots, \ell \tag{10}$$

This problem has a simple solution: The optimal Δ-margin separating hyperplane (in Z space) has a form

$$y = \mathrm{sign}\left[\sum_{i=1}^{\ell} y_i \alpha_i^0 (z_i, z_j) + b \right] \tag{11}$$

where α_i^0 $i = 1, \ldots, \ell$ is the solution of the following optimization quadratic optimization problem. Maximize the positive definite quadratic form

$$R = \sum_{i=1} \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (z_i, z_j) \tag{12}$$

subject to box constraints

$$0 \le \alpha_i \le C, \quad {}_1=1, \ldots, \ell \tag{13}$$

and one equality constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{14}$$

3.1 Kernel Trick

In the previous section we described the solution to the classification problems in high dimension input space X. We mapped our input vectors even in higher dimensional space Z and constructed in this space the large margin separating hyperplane.

The idea was that in high dimensional image space Z, the ratio of the radius of the sphere to the value of the margin can be chosen to be small. This will imply a small VC dimension and this (but not the dimensionality of the space) guarantees a good generalization.

However, in a form (11) - (14) the obtained solution is non-constructive. It requires explicit mapping vectors x_i into Z space and calculating in this space inner products between two vectors. Note however that image vectors z_i appear in both equations only in inner products. This makes it possible to calculate the inner product without constructing image vectors z. The following remarkable proposition (due to Mercer, 1909) holds true.

Proposition 5: Let vectors $x \in X$ be mapping into elements $z \in Z$ of a Hilbert space H

$$x \longrightarrow z \tag{15}$$

and let (z_i, z_j) be an inner product of elements z_j and z_j of space H. Then

- For any mapping (15) there exists positive definite (PD) function* $K(x, x^*)$ such that

$$(z_i, z_j) = K(x_i, x_j) \,. \tag{16}$$

- For any PD function $K(x_i, x_j)$ there exists a mapping (15) that (16) holds.

According to this proposition mapping is completely equivalent to the choice of positive definite kernel function $K(x_i, x_j)$ which we also call a similarity measure. Note, that one can choose appropriate similarity measure without any knowledge about a corresponding mapping (15).

Using similarity measure kernel $K(x_i, x_j)$ one can

*Function $K(x, x^*)$ is called positive definite if for any x_1, ..., x_ℓ the determinant of the Gram matrix is non-negative.

$$\left| K(x_i, x_j) \right| \ge 0, \quad i=1, \ldots, \ell\, , j=1, \ldots, \ell$$

rewrite equations (11)-(14) in the following constructive form. The optimal Δ-margin separating hyperplane (in Z space) has a form

$$y = \mathrm{sign}\left[\sum_{i=1}^{\ell} y_i \alpha_i^{0} K(x_i, x_j) + b\right] \qquad (17)$$

where α_i^0 , $i = 1, \ldots, \ell$ is the solution of the following quadratic optimization problem. Maximize the positive definite quadratic form

$$R = \sum_{i=1}\sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (18)$$

subject to box constraints (13) and one equality constraint (14).

Examples of two popular kernels are:
Polynomial kernel of degree d:

$$K(x_i, x_j) = [1+(x_i, x_j)]^{d}$$

Exponential kernel:

$$K(x_i, x_j) = \exp\left\{-\left|\frac{x_i - x_j}{\sigma}\right|^{q}\right\}, \qquad 0 \le q \le 2$$

Note that different similarity measures specify different coefficients in equations. They do not change equations. Therefore, SVMs form universal generalization method. They use the same equations for different real life problems. Depending on specific similarity measure they can be applied to any learning problem of interest (say such as stock market prediction, or medical diagnostics, or prediction of elements of weather). Specific similarity measure is choosing for the problem of interest and the construction of a rule is based the same method of solving equations that does not depend on similarity measure.

It is very important to note that similarity measure can be constructed not only for vectorial input spaces X. It can be constructed for abstract elements. This fact is playing important role in many applications (such as text analysis or bioinformatics).

3.2 Properties of SVMs

SVMs possess the following remarkable properties:

1) They always converge to the best possible solutions.
   SVMs execute the SRM minimization inductive principle, which is strongly universally consistent (see Proposition 3). Its structure is defined by different values of Δ-margin in Hilbert (infinite dimensional) space. Therefore one can construct set of Δ-margin separating hyperspace with any VC dimension.
2) They minimize guarantee bounds (3) for a finite number of observations.
   SVMs were constructed to minimize right hand side of inequality (3), which requires one to construct Δ-margin separating hyperplane with a large margin.
3) They have a standard way to incorporate singularities of real-life problems using appropriate similarity measure $K(x_i, x_j)$ between two vectors x_i, x_j.
4) The similarity measure can be defined for nonvectorial data (e.g. chemical formulas, or poetry texts, or political situations).
5) They have a universal generalization engine (simple QP solver).
6) They construct non-linear decision rules using linear technology.

Combinations of these properties is unique in applied analysis.

It was shown in many empirical studies that SVMs generalization engine possesses the state-of-the-art generalization properties in solving real life problems.

Below we describe results of applying the SVM generalization engine to the problem of classification of proteins into 56 classes [3] and compare this classification with other methods. In **Fig. 4**, the x-axes

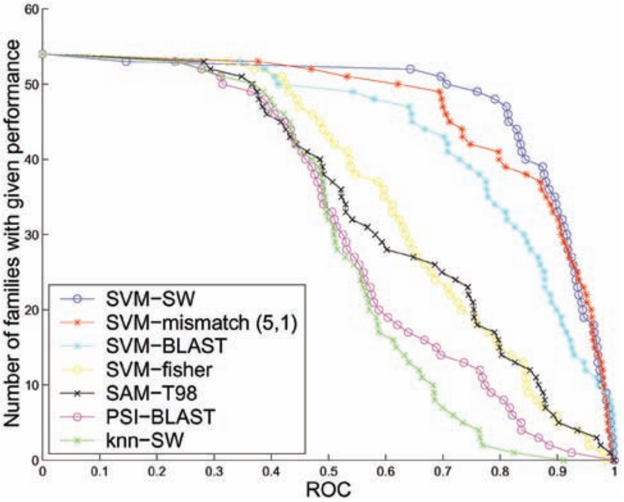

Fig. 4 Comparison of the classical bioinformatics algorithms.

define the value of ROC score (a measure that defines precision of classification; the best precision gives score equal to 1). The y-axes define how many classes is classified with ROC score that is not worse than x.

The ideal outcome is defined by the line $y = y(x) = 1$ for all $0 \leq x \leq 1$.

Figure 4 compares the classical bioinformatics algorithms with one that uses the SVM generalization engine. Input for algorithms is strings of (possible) different size that characterize proteins. Output is number of class.

All algorithms depend on two concepts: appropriate similarity measure for two strings (called alignment score in bioinformatics literature) and generalization engine. For example knn-SW means k-nearest neighbor generalization method and SW similarity measure or SVM-SW means SVM generalization engine SW similarity measure. Figure 4 shows a big advantage of the SVM over other methods used.

4. TRANSDUCTIVE INFERENCE

Statistical Learning Theory considers two distinctive types of inferences: inductive, and transductive [1,2].

The goal of transductive inference is to estimate the values of an unknown predictive function at a given point of interest (but not in the whole domain of its definition). The point is that, by solving less demanding problems, one can achieve more accurate solutions. A general theory of transduction was developed where it was shown that the bounds of generalization for transductive inference are better than the corresponding bounds for inductive inference[2].

1) Prediction of molecular bioactivity for drug discovery[4].

The CUP-2001 competition on data analysis methods required the construction of a rule for predicting molecular bioactivity using data given by the DuPont Pharmaceutical Company. The data belonged to a binary 139,351 dimensional space, which contained a training set of 1,909 vectors, and a test set of 634 vectors.

Below the results are given for the winner of the competition (among 119 competitors that used traditional approaches), SVM-inductive inference and SVM transductive inference.

· Winner's accuracy	68.1%
· SVM inductive mode accuracy	74.5%
· SVM transductive mode accuracy	82.3%

It is remarkable that the jump in performance obtained due to a new philosophy of inference (transductive instead of inductive) was larger than the jump resulting from the reinforcement of the technology in construction of inductive predictive rules.

2) Text categorization[5].

In the text categorization problem the replacement of inductive inference by transductive inference reduced the error rate from 30% to 15%.

Remark: The discovery of transductive inference and its advantage over inductive inference is not just a technical achievement, but a breakthrough in the philosophy of generalization.

Until now, the traditional method of inference was the inductive-deductive method, where using available information one defines a general rule first, and then using this rule deduces the answer one needs. That is, first one goes from particular to general and then from general to particular.

In the transductive mode one provides direct inference from particular to particular, avoiding the illposed part of the inference problem (inference from particular to general).

5. CONCLUSION

The construction of SVMs solved important problems of machine intelligence: creation of a universal generalization engine that possesses high performance and can be used for solving different learning problems.

With the creation of SVMs it became clear that problems of learning contain different components not just generalization. It includes problem of choice of similarity measure, choice of appropriate invariants, and many others. It was also shown that the existence duality - similarity measure in input space and separating hyperplane in image (mapping) space — allows solving such problems within the framework of SVMs.

Recently, a discussion started about creating “cognitive computers,” that can interact with customers, learn from customer and Internet patterns of behavior, and learn to perform intelligent tasks. In such computers, one of the blocks should be a universal generalization engine. SVMs can be considered as a good candidate for such a block.

REFERENCES

[1] V. Vapnik, “The Nature of Statistical Learning Theory,” Springer, 1995.

[2] V. Vapnik, “Statistical Learning Theory,” J. Wiley, 1998.

[3] C. Leslie, E. Eskin, et al., “Mismatch string kernels for discriminative protein classification,” Bioinformatics, **20**, 4, 2004.

[4] J. Weston, F. Peres-Cruz, O. et al., “Feature selection and transduction for prediction of molecular bioactivity for drug design,” Bioinformatics, **19**, 3, 2003.

[5] T. Joachims, “Learning to Classify Text Using Support Vector Machines,” Kulwer Academic Publishers, 2002.

Received February 8, 2005

* * * * * * * * * * * * * * *

Vladimir VAPNIK is one of the creators of Statistical Learning Theory and Support Vector technology. He is the author of seven monographs and more than 100 articles. Till 1990 he worked in Russian Academia of Science, then from 1990 to 2002 in Bell-Labs (AT&T Labs). Since 2002 he has been working for NEC Laboratories America.

* * * * * * * * * * * * * * *