Internet Electronic Journal of Molecular Design

August 2002, Volume 1, Number 8, Pages 418–427

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Milan Randić on the occasion of the 70th birthday
Part 4

Guest Editor: Mircea V. Diudea

Support Vector Machines for Cancer Diagnosis from the Blood Concentration of Zn, Ba, Mg, Ca, Cu, and Se

Ovidiu Ivanciuc

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–1157

Received: April 27, 2002; Revised: May 25, 2002; Accepted: June 17, 2002; Published: August 31, 2002

Citation of the article:
O. Ivanciuc, Support Vector Machines for Cancer Diagnosis from the Blood Concentration of Zn, Ba, Mg, Ca, Cu, and Se, *Internet Electron. J. Mol. Des.* **2002**, *1*, 418–427, http://www.biochempress.com.

Copyright © 2002 BioChem Press

Internet Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machines for Cancer Diagnosis from the Blood Concentration of Zn, Ba, Mg, Ca, Cu, and Se[#]

Ovidiu Ivanciuc*

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–1157

Received: April 27, 2002; Revised: May 25, 2002; Accepted: June 17, 2002; Published: August 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (8), 418–427

Abstract

Motivation. Machine learning techniques, mainly artificial neural networks, clustering and classification algorithms, have recently received considerable attention as successful methods for modeling medical data. Using a wide variety of mathematical equations, machine learning algorithms are able to generate predictive models for different cancer types.

Method. Support vector machines (SVM) are a new machine learning algorithm that found numerous applications in bioinformatics, cheminformatics, computational biology, and structure–activity relationships. In this study we have investigated the application of SVM for cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se. The SVM model with the best prediction power was identified by a leave–10%–out cross–validation procedure, using the dot, polynomial, radial basis function, neural, and anova kernels.

Results. Extensive simulations demonstrate that the classification performances of SVM depend strongly on the kernel type and various parameters that control the kernel shape. The best prediction results were obtained with a dot kernel with seven support vectors. The anova kernel offered comparable predictions, but with 24 support vectors.

Conclusions. Support vector machines represent a powerful and flexible classification algorithm, with many potential applications in modeling medical data. The results reported in the present study demonstrate such an application in the cancer diagnosis.

Keywords. Cancer diagnosis; support vector machines; machine learning; kernel algorithm; classification algorithm.

1 INTRODUCTION

Machine learning techniques, mainly artificial neural networks, clustering and classification algorithms, have recently received considerable attention as successful methods for modeling medical data [1–8]. Using a wide variety of mathematical equations, machine learning algorithms are able to generate predictive models for different cancer types. Support vector machines (SVM)

[#] Dedicated to Professor Milan Randić on the occasion of the 70th birthday.

* Correspondence author; E–mail: ivanciuc@netscape.net.

represent a new class of machine learning algorithms that found numerous applications in various classification and regression models. In this study we have investigated the application of SVM for the cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se, using a data set previously explored in Refs. [6] and [9]. The influence of the kernel type on the SVM performances was extensively explored using various kernels, namely the dot, polynomial, radial basis function, neural, and anova kernels.

2 MATERIALS AND METHODS

Support vector machines were developed by Vapnik [10–12] as an effective algorithm for determining an optimal hyperplane to separate two classes of patterns [13–23]. In the first step, using various kernels that perform a nonlinear mapping, the input space is transformed into a higher dimensional feature space. Then, a maximal margin hyperplane (MMH) is computed in the feature space by maximizing the distance to the hyperplane of the closest patterns from the two classes. The patterns that determine the separating hyperplane are called support vectors.

This powerful classification technique was applied with success in medicine, computational biology, bioinformatics, and structure–activity relationships, for the classification of: microarray gene expression data [24], translation initiation sites [25], genes [26], cancer type [27–30], pigmented skin lesions [31], HIV protease cleavage sites [32], GPCR type [33], protein class [34], membrane protein type [35], protein–protein interactions [36], protein subcellular localization [37–39], protein fold [40], protein secondary structure [41], specificity of GalNAc–transferase [42], DNA hairpins [43], organisms [44], aquatic toxicity mechanism of action [45], carcinogenic activity of polycyclic aromatic hydrocarbons [46], structure–odor relationships for pyrazines [47].

In this study we have investigated the application of SVM for cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se. The 74 experimental data reported in Table 1 were taken from the literature [6,9], and consist of 32 data from cancer patients (class +1) and 42 data from normal individuals (class –1). All SVM models from the present paper for the classification of pyrazines into three aroma classes were obtained with mySVM [48], which is freely available for download. Links to Web resources related to SVM, namely tutorials, papers and software, can be found in BioChem Links [49] at http://www.biochempress.com. Before computing the SVM model, the input vectors were scaled to zero mean and unit variance. The prediction power of each SVM model was evaluated with a leave–10%–out cross–validation procedure, and the capacity parameter C took the values 10, 100, and 1000. We present below the kernels and their parameters used in this study.

The dot kernel. The inner product of x and y defines the dot kernel:

$$K(x,y) = x \cdot y \tag{1}$$

Table 1. Blood concentration of Zn, Ba, Mg, Ca, Cu, and Se for cancer patients (class +1) and normal individuals (class –1)

No	Zn	Ba	Mg	Ca	Cu	Se	Class
1	0.65	0.005	19.9	78.4	0.58	0.088	+1
2	0.63	0.012	20.7	81.4	1.02	0.066	+1
3	0.52	0.032	19.4	74.1	0.68	0.059	+1
4	0.66	0.007	23.7	86.5	1.01	0.07	+1
5	0.64	0.023	20.4	78.4	0.94	0.073	+1
6	0.67	0.026	20.2	85.6	1.09	0.071	+1
7	0.67	0.022	19.4	85.1	0.84	0.052	+1
8	0.67	0.006	19.6	76.7	0.85	0.081	+1
9	0.73	0.013	17.8	74.7	0.84	0.074	+1
10	0.51	0.010	16.4	77.2	0.88	0.084	+1
11	0.54	0.017	18.6	74.7	1.14	0.081	+1
12	0.70	0.009	21.6	78.8	0.97	0.071	+1
13	0.41	0.013	17.4	60.1	0.69	0.075	+1
14	0.55	0.017	20.8	71.2	0.98	0.083	+1
15	0.58	0.012	21.7	71.4	0.74	0.068	+1
16	0.46	0.007	18.2	68.3	0.81	0.096	+1
17	0.44	0.035	21.1	71.6	1.31	0.057	+1
18	0.54	0.013	22.5	79.5	0.86	0.075	+1
19	0.48	0.006	18.3	71.9	0.76	0.046	+1
20	0.49	0.034	17.7	68.9	0.73	0.088	+1
21	0.47	0.021	15.2	66.3	1.00	0.067	+1
22	0.45	0.163	16.9	65.6	0.80	0.067	+1
23	0.49	0.008	15.6	63.0	0.74	0.072	+1
24	0.43	0.143	15.3	57.0	0.83	0.049	+1
25	1.76	0.243	12.5	52.1	0.64	0.082	+1
26	0.70	0.008	13.8	63.8	0.84	0.052	+1
27	2.20	0.067	15.6	65.8	0.96	0.066	+1
28	0.58	0.032	9.2	41.8	0.98	0.087	+1
29	1.09	0.010	10.8	42.3	0.60	0.076	+1
30	0.55	0.201	14.3	60.0	0.57	0.065	+1
31	0.60	0.228	18.8	72.6	0.99	0.069	+1
32	1.08	0.238	20.2	76.4	0.95	0.060	+1
33	1.91	0.224	27.3	74.6	2.60	0.052	–1
34	0.67	0.175	18.6	65.6	1.33	0.074	–1
35	1.13	0.148	16.6	63.5	0.94	0.038	–1
36	0.88	0.145	20.1	59.4	1.37	0.045	–1
37	0.54	0.033	16.1	49.6	1.46	0.049	–1
38	1.03	0.052	16.4	42.5	1.59	0.042	–1
39	0.92	0.039	16.8	64.8	1.54	0.043	–1
40	0.76	0.042	16.4	54.0	1.69	0.085	–1
41	1.61	0.078	16.0	49.9	1.18	0.055	–1
42	1.56	0.044	10.2	57.2	1.35	0.049	–1
43	0.84	0.051	16.5	48.2	1.05	0.055	–1
44	0.70	0.051	14.2	41.0	0.64	0.031	–1
45	0.73	0.024	15.2	36.0	1.14	0.069	–1
46	0.69	0.048	18.6	44.9	1.91	0.079	–1
47	1.01	0.031	17.8	46.9	0.75	0.099	–1
48	0.83	0.049	18.4	34.4	0.86	0.126	–1
49	0.30	0.002	6.5	15.3	0.43	0.074	–1
50	0.61	0.037	19.4	49.4	2.03	0.055	–1
51	0.53	0.032	17.3	45.1	0.85	0.037	–1
52	0.51	0.026	18.6	54.8	1.21	0.022	–1
53	2.40	0.046	15.8	53.0	1.20	0.065	–1
54	0.52	0.031	19.6	41.0	0.75	0.051	–1
55	0.35	0.008	17.7	36.8	1.10	0.040	–1
56	0.56	0.028	19.5	43.7	1.06	0.069	–1

Table 1. (Continued)

No	Zn	Ba	Mg	Ca	Cu	Se	Class
57	0.32	0.024	11.1	30.5	0.40	0.081	–1
58	0.75	0.035	20.2	50.7	0.94	0.081	–1
59	1.98	0.036	17.5	53.6	0.57	0.074	–1
60	0.22	0.046	9.9	35.5	0.45	0.059	–1
61	0.33	0.018	13.6	34.9	0.66	0.061	–1
62	0.97	0.036	17.8	48.3	0.72	0.047	–1
63	0.78	0.027	18.3	46.9	0.49	0.075	–1
64	0.32	0.028	10.8	41.2	0.66	0.034	–1
65	0.48	0.024	20.9	49.5	1.20	0.125	–1
66	0.54	0.033	16.1	51.2	1.17	0.061	–1
67	0.58	0.029	15.5	44.8	2.74	0.046	–1
68	0.66	0.026	16.4	39.8	1.08	0.068	–1
69	0.69	0.046	14.0	47.4	1.07	0.058	–1
70	1.32	0.041	18.0	49.8	0.43	0.056	–1
71	0.27	0.036	16.0	45.0	1.32	0.047	–1
72	0.41	0.050	19.9	56.5	1.35	0.056	–1
73	0.47	0.035	12.3	40.1	1.73	0.057	–1
74	1.90	0.030	15.7	43.0	1.44	0.039	–1

The polynomial kernel. The polynomial of degree d (values 2, 3, 4, and 5) in the variables x and y defines the polynomial kernel:

$$K(x,y) = (x \cdot y + 1)^d \tag{2}$$

The radial kernel. The following exponential function in the variables x and y defines the radial basis function kernel, with the shape controlled by the parameter γ (values 0.5, 1.0, and 2.0):

$$K(x,y) = \exp(-\gamma \| x - y \|^2) \tag{3}$$

The neural kernel. The hyperbolic tangent function in the variables x and y defines the neural kernel, with the shape controlled by the parameters a (values 0.5, 1.0, and 2.0) and b (values 0, 1.0, and 2.0):

$$K(x,y) = \tanh(ax \cdot y + b) \tag{4}$$

The anova kernel. The sum of exponential functions in x and y defines the anova kernel, with the shape controlled by the parameters γ (values 0.5, 1.0, and 2.0) and d (values 1, 2, and 3):

$$K(x,y) = \left(\sum_i \exp(-\gamma(x_i - y_i)) \right)^d \tag{5}$$

3 RESULTS AND DISCUSSION

Similarly with other multivariate statistical models, the performances of SVM classifiers depend on the combination of several parameters, and the kernel type is the most important one. Because the use of SVM models in chemometrics, structure–activity studies, and QSAR is only in the beginning, there are no clear guidelines on selecting the most effective kernel for a certain classification problem. Another important problem in SVM applications is the selection of the input numerical indices that can discriminate the investigated patterns. For the moment, this is an unexplored problem, and in this study we have used the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se from [6,9] without any attempt of removing indices with low influence on the classification performance.

The statistical results obtained in the SVM experiments are presented in Table 2. The calibration of the SVM models, performed with the whole set of 74 patterns from Table 1, is characterized by the following statistics: SV, number of support vectors; BSV, number of bounded support vectors; +/+, number of +1 patterns (cancer patients) predicted in class +1; +/–, number of +1 patterns predicted in class –1; –/–, number of –1 patterns (normal individuals) predicted in class –1; –/+, number of –1 patterns predicted in class +1; CAa, accuracy. The high flexibility of multivariate statistical models in approximating a wide range of mathematical functions comes with a significant danger, namely overfitting. Using sophisticated kernels, SVM can be calibrated to perfectly discriminate two populations of patterns, but only a cross–validation test can demonstrate the potential utility of an SVM model. For each SVM model we present in Table 2 the following leave–10%–out (L10%O) cross-validation statistics: ASV, average number of support vectors; ABSV, average number of bounded support vectors; TRa, training accuracy; TEa, test accuracy.

The first group of SVM models from Table 2, experiments 1–3, was obtained with the dot kernel. As can be seen from the results of experiments 2 and 3, the dot kernel is able to perfectly separate the two classes of patterns, using only seven support vectors, and with good leave–10%–out (L10%O) cross–validation results, namely TEa = 0.93. The SVM model from experiment 3 is determined by four +1 patterns (cancer patients) (*i.e.*, **24**, **25**, **28**, and **29**) and by three –1 patterns (normal individuals) (*i.e.*, **34**, **35**, and **60**). These seven patterns can be used to predict the cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se.

The dot kernel is the simplest kernel used in our SVM experiments, and one would expect that by using more complex kernel functions the classification performances of the SVM model would increase. However, our experiments performed with the polynomial, radial basis function, neural, and anova kernels clearly show that the prediction statistics obtained with these functions are lower than those obtained with the simple dot kernel. The results obtained with the polynomial kernel (Table 2, experiments 4–15) show that a perfect separation of the two classes is obtained in calibration, but the number of support vectors is large (between 19 and 30) and the L10%O results are worse than those obtained with the dot kernel (TEa takes values between 0.87 and 0.89).

The SVM models obtained with the radial basis function kernel (Table 2, experiments 16–24) have TEa between 0.82 (experiments 22–24) and 0.89 (experiments 16–18), indicating that the L10%O predictions are of lower quality than those obtained with the dot kernel. The main deficiency of the SVM models obtained with the radial kernel is the large number of support vectors, between 51 (experiments 16–18) and 70 (experiments 22–24).

Table 2. Results for SVM Modeling of the Cancer Diagnosis.[a]

No	*C*	*K*			SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
1	10	D			12	5	32	0	42	0	1.00	10.8	4.1	0.99	0.93
2	100				7	0	32	0	42	0	1.00	7.0	0.0	1.00	0.93
3	1000				7	0	32	0	42	0	1.00	7.0	0.0	1.00	0.93
			d												
4	10	P	2		19	0	32	0	42	0	1.00	18.3	0.0	1.00	0.87
5	100		2		19	0	32	0	42	0	1.00	18.3	0.0	1.00	0.87
6	1000		2		19	0	32	0	42	0	1.00	18.3	0.0	1.00	0.87
7	10		3		28	0	32	0	42	0	1.00	24.1	0.0	1.00	0.89
8	100		3		28	0	32	0	42	0	1.00	24.1	0.0	1.00	0.89
9	1000		3		28	0	32	0	42	0	1.00	24.1	0.0	1.00	0.89
10	10		4		30	0	32	0	42	0	1.00	26.1	0.0	1.00	0.88
11	100		4		30	0	32	0	42	0	1.00	26.1	0.0	1.00	0.88
12	1000		4		30	0	32	0	42	0	1.00	26.1	0.0	1.00	0.88
13	10		5		28	0	32	0	42	0	1.00	25.6	0.0	1.00	0.87
14	100		5		28	0	32	0	42	0	1.00	25.6	0.0	1.00	0.87
15	1000		5		28	0	32	0	42	0	1.00	25.6	0.0	1.00	0.87
			γ												
16	10	R	0.5		51	0	32	0	42	0	1.00	47.8	0.0	1.00	0.89
17	100		0.5		51	0	32	0	42	0	1.00	47.8	0.0	1.00	0.89
18	1000		0.5		51	0	32	0	42	0	1.00	47.8	0.0	1.00	0.89
19	10		1.0		62	0	32	0	42	0	1.00	56.8	0.0	1.00	0.85
20	100		1.0		62	0	32	0	42	0	1.00	56.8	0.0	1.00	0.85
21	1000		1.0		62	0	32	0	42	0	1.00	56.8	0.0	1.00	0.85
22	10		2.0		70	0	32	0	42	0	1.00	63.5	0.0	1.00	0.82
23	100		2.0		70	0	32	0	42	0	1.00	63.5	0.0	1.00	0.82
24	1000		2.0		70	0	32	0	42	0	1.00	63.5	0.0	1.00	0.82
			a	*b*											
25	10	N	0.5	0.0	21	18	23	9	33	9	0.76	17.2	14.3	0.79	0.77
26	100		0.5	0.0	17	14	25	7	35	7	0.81	16.1	14.3	0.78	0.77
27	1000		0.5	0.0	17	14	25	7	35	7	0.81	17.4	15.4	0.77	0.79
28	10		1.0	0.0	18	18	25	7	33	9	0.78	18.7	16.1	0.76	0.81
29	100		1.0	0.0	18	15	25	7	34	8	0.80	17.5	15.4	0.77	0.78
30	1000		1.0	0.0	18	15	25	7	34	8	0.80	18.6	16.0	0.76	0.84
31	10		2.0	0.0	20	18	23	9	33	9	0.76	20.0	18.0	0.74	0.85
32	100		2.0	0.0	20	17	23	9	34	8	0.77	19.1	16.5	0.75	0.81
33	1000		2.0	0.0	21	19	23	9	32	10	0.74	18.6	16.3	0.75	0.88
34	10		0.5	1.0	30	30	22	10	25	17	0.64	26.0	24.3	0.64	0.62
35	100		0.5	1.0	30	30	22	10	25	17	0.64	25.7	23.9	0.65	0.62
36	1000		0.5	1.0	30	30	22	10	25	17	0.64	25.6	23.7	0.65	0.62
37	10		1.0	1.0	30	28	18	14	28	14	0.62	26.1	24.5	0.63	0.61
38	100		1.0	1.0	28	28	18	14	26	16	0.59	25.3	23.8	0.63	0.66
39	1000		1.0	1.0	28	28	18	14	26	16	0.59	25.8	23.9	0.64	0.59
40	10		2.0	1.0	24	22	21	11	31	11	0.70	23.2	21.3	0.68	0.73
41	100		2.0	1.0	24	22	21	11	31	11	0.70	23.3	21.2	0.68	0.76
42	1000		2.0	1.0	25	22	21	11	31	11	0.70	23.4	21.2	0.68	0.76
43	10		0.5	2.0	32	32	22	10	19	23	0.55	29.8	28.7	0.56	0.65
44	100		0.5	2.0	30	30	22	10	19	23	0.55	28.2	26.7	0.58	0.57
45	1000		0.5	2.0	30	30	22	10	19	23	0.55	28.4	26.6	0.58	0.58
46	10		1.0	2.0	32	32	20	12	19	23	0.53	30.2	28.7	0.55	0.54
47	100		1.0	2.0	32	30	17	15	27	15	0.59	29.6	28.6	0.55	0.54
48	1000		1.0	2.0	32	30	17	15	27	15	0.59	29.0	28.4	0.55	0.54
49	10		2.0	2.0	30	30	19	13	24	18	0.58	27.2	26.1	0.60	0.62
50	100		2.0	2.0	30	30	19	13	24	18	0.58	26.7	25.8	0.61	0.59
51	1000		2.0	2.0	30	30	19	13	24	18	0.58	26.9	26.1	0.60	0.57

Table 2. (Continued)

No	*C*	*K*	γ	*d*	SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
52	10	A	0.5	1	18	0	32	0	42	0	1.00	16.6	0.0	1.00	0.85
53	100		0.5	1	18	0	32	0	42	0	1.00	16.6	0.0	1.00	0.85
54	1000		0.5	1	18	0	32	0	42	0	1.00	16.6	0.0	1.00	0.85
55	10		1.0	1	21	0	32	0	42	0	1.00	20.2	0.0	1.00	0.81
56	100		1.0	1	21	0	32	0	42	0	1.00	20.2	0.0	1.00	0.81
57	1000		1.0	1	21	0	32	0	42	0	1.00	20.2	0.0	1.00	0.81
58	10		2.0	1	23	0	32	0	42	0	1.00	22.9	0.0	1.00	0.82
59	100		2.0	1	23	0	32	0	42	0	1.00	22.9	0.0	1.00	0.82
60	1000		2.0	1	23	0	32	0	42	0	1.00	22.9	0.0	1.00	0.82
61	10		0.5	2	24	0	32	0	42	0	1.00	22.3	0.0	1.00	0.91
62	100		0.5	2	24	0	32	0	42	0	1.00	22.3	0.0	1.00	0.91
63	1000		0.5	2	24	0	32	0	42	0	1.00	22.3	0.0	1.00	0.91
64	10		1.0	2	31	0	32	0	42	0	1.00	29.2	0.0	1.00	0.91
65	100		1.0	2	31	0	32	0	42	0	1.00	29.2	0.0	1.00	0.91
66	1000		1.0	2	31	0	32	0	42	0	1.00	29.2	0.0	1.00	0.91
67	10		2.0	2	40	0	32	0	42	0	1.00	36.9	0.0	1.00	0.91
68	100		2.0	2	40	0	32	0	42	0	1.00	36.9	0.0	1.00	0.91
69	1000		2.0	2	40	0	32	0	42	0	1.00	36.9	0.0	1.00	0.91
70	10		0.5	3	27	0	32	0	42	0	1.00	26.7	0.0	1.00	0.89
71	100		0.5	3	27	0	32	0	42	0	1.00	26.7	0.0	1.00	0.89
72	1000		0.5	3	27	0	32	0	42	0	1.00	26.7	0.0	1.00	0.89
73	10		1.0	3	42	0	32	0	42	0	1.00	38.4	0.0	1.00	0.89
74	100		1.0	3	42	0	32	0	42	0	1.00	38.4	0.0	1.00	0.89
75	1000		1.0	3	42	0	32	0	42	0	1.00	38.4	0.0	1.00	0.89
76	10		2.0	3	52	0	32	0	42	0	1.00	48.9	0.0	1.00	0.91
77	100		2.0	3	52	0	32	0	42	0	1.00	48.9	0.0	1.00	0.91
78	1000		2.0	3	52	0	32	0	42	0	1.00	48.9	0.0	1.00	0.91

[a] The table reports the experiment number Exp, capacity parameter *C*, kernel type *K* (dot D; polynomial P; radial basis function R; neural N; anova A) and corresponding parameters, calibration results (SV, number of support vectors; BSV, number of bounded support vectors; +/+, number of +1 patterns (cancer patients) predicted in class +1; +/–, number of +1 patterns predicted in class –1; –/–, number of –1 patterns (normal individuals) predicted in class –1; –/+, number of –1 patterns predicted in class +1; CAa, accuracy), and cross-validation results (ASV, average number of support vectors; ABSV, average number of bounded support vectors; TRa, training accuracy; TEa, test accuracy).

The fourth group of SVM models was obtained with the neural kernel (Table 2, experiments 25–51). Although we have explored a wide range of values for the parameters *a* (values 0.5, 1.0, 2.0) and *b* (values 0, 1.0, 2.0), the classification results are bad. The neural kernel is not able to separate the two classes in calibration, while in prediction TEa takes values between 0.54 (experiments 46 and 47) and 0.88 (experiment 33). Despite its success as a transfer function in artificial neural networks, the hyperbolic tangent is not a very useful kernel in SVM models, as found also in other investigations [45–47].

The last group of SVM models was obtained with the anova kernel (Table 2, experiments 52–78). This kernel separates perfectly the two classes in calibration, while in prediction TEa takes values between 0.81 (experiments 55–57) and 0.91 (experiments 61–69 and 76–78). Compared with the dot kernel, the number of support vectors is much larger, between 18 (experiments 52–54) and 52 (experiments 76–78). Considering the number of support vectors, the best SVM models are obtained in experiments 61–63, with TEa = 0.91, 24 support vectors in calibration, and an average of 22.3 support vectors in prediction. Although TEa is very close to the value obtained with the dot kernel, namely 0.93, the number of support vectors is more than three times greater (dot kernel SVM models were obtained with 7 support vectors in calibration and prediction).

4 CONCLUSIONS

Support vector machines represent a new class of machine learning algorithms that can have significant applications in the design of chemical libraries, in chemometrics, and in structure–activity models. The possibility to discriminate clusters separated by non–linear surfaces, the unique solution for the class separation, and the fast optimization are three important advantages of SVM. In this study we have investigated the application of SVM for the cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se, using a data set previously explored in Refs. [6] and [9]. The influence of the kernel type on the SVM performances was extensively explored using various kernels, namely the dot, polynomial, radial basis function, neural, and anova kernels.

The role of a classifier is to learn the classification rule from training patterns and then to apply the rule to new patterns in order to obtain reliable predictions. Therefore, for a classifier, one of the most important properties is its generalization ability or its ability to make correct predictions for patterns not used in the calibration phase. In this investigation, the prediction power of each SVM model was evaluated with a leave–10%–out cross–validation procedure. After experimenting with various kernels and associated parameters, our results clearly demonstrate that the performance of the SVM classifier is strongly dependent on the kernel shape.

The best predictions were obtained with a dot kernel with seven support vectors, namely four +1 patterns (cancer patients) (*i.e.*, **24**, **25**, **28**, and **29**) and three –1 patterns (normal individuals) (*i.e.*, **34**, **35**, and **60**). These seven patterns can be used to predict the cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se. Although the dot kernel represents a simple separation hyperplane, its predictions are better than those obtained with the polynomial, radial basis function, and neural kernels. Only the anova kernel gives predictions close to those obtained with the dot kernel, but with three times more support vectors. The neural kernel constantly gives bad classification results, both in calibration and prediction. Despite its success as a transfer function in artificial neural networks, the hyperbolic tangent is not a very useful kernel in SVM models, as found also in other investigations [45–47].

Contrary to the almost general misconception, our results clearly indicate that the dot polynomial can give better predictions than more complex kernels. Because the development of an SVM model is an empirical process, various kernels and associated parameters must be investigated in order to identify the SVM with the best prediction power.

Supplementary Material

The mySVM model files for experiments 2 and 3 are available as supplementary material.

5 REFERENCES

[1] B. Kovalerchuk, E. Triantaphyllou, J. F. Ruiz, V. I. Torvil, and E. Vityaev, The Reliability Issue of Computer–Aided Breast Cancer Diagnosis, *Comput. Biomed. Res*. **2000**, *33*, 296–313.
[2] J. E. Montie and J. T. Wei, Artificial Neural Networks for Prostate Carcinoma Risk Assessment, *Cancer* **2000**, *88*, 2655–2660.
[3] D. J. Sargent, Comparison of Artificial Neural Networks with Other Statistical Approaches. Results from Medical Data Sets, *Cancer* **2001**, *91*, 1636–1642.
[4] R. W. Veltri, M. C. Miller, A. W. Partin, E. C. Poole, and G. J. O'Dowd, Prediction of Prostate Carcinoma Stage by Quantitative Biopsy Pathology, **2001**, *91*, 2322–2328.
[5] P. Boracchi, E. Biganzoli, and E. Marubini, Modelling Cause–Specific Hazards with Radial Basis Function Artificial Neural Networks: Application to 2233 Breast Cancer Patients, *Statist. Med.* **2001**, *20*, 3677–3694.
[6] A. S. Al–Ammar and R. M. Barnes, Supervised Cluster Classification using the Original *n*–Dimensional Space without Transformation into Lower Dimension, *J. Chemometrics* **2001**, *15*, 49–67.
[7] A. Ciampi and F. Zhang, A New Approach to Training Back–Propagation Artificial Neural Networks: Empirical Evaluation on Ten Data Sets from Clinical Studies, *Statist. Med.* **2002**, *21*, 1309–1330.
[8] C. Stephan, K. Jung, H. Cammann, B. Vogel, B. Brux, G. Kristiansen, B. Rudolph, S. Hauptmann, M. Lein, D. Schnorr, P. Sinha, and S. A. Loening, An Artificial Neural Network Considerably Improves the Diagnostic Power of Percent Free Prostate–Specific Antigen in Prostate Cancer Diagnosis: Results of a 5–Year Investigation, *Int. J. Cancer* **2002**, *99*, 466–473.
[9] E. Zhu and X. Wang, *Chem. J. Chinese Univ*. **1993**, *14*, 621–625.
[10] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979.
[11] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
[12] V. Vapnik, *Statistical Learning Theory*, Wiley–Interscience, New York, 1998.
[13] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowledge Discov.* **1998**, *2*, 121–167.
[14] B. Schölkopf, K. –K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers, *IEEE Trans. Signal Process.* **1997**, *45*, 2758–2765.
[15] V. N. Vapnik, An Overview of Statistical Learning Theory, *IEEE Trans. Neural Networks* **1999**, *10*, 988–999.
[16] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods*: *Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
[17] N. Cristianini and J. Shawe–Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
[18] K.–R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, An Introduction to Kernel–Based Learning Algorithms, *IEEE Trans. Neural Networks* **2001**, *12*, 181–201.
[19] C.–C. Chang and C.–J. Lin, Training ν–Support Vector Classifiers: Theory and Algorithms, *Neural Comput*. **2001**, *12*, 2119–2147.
[20] I. Steinwart, On the Influence of the Kernel on the Consistency of Support Vector Machines, *J. Machine Learning Res.* **2001**, *2*, 67–93, http://www.jmlr.org.
[21] A. Ben–Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, Support Vector Clustering, *J. Machine Learning Res.* **2001**, *2*, 125–137, http://www.jmlr.org.
[22] R. Collobert and S. Bengio, SVMTorch: Support Vector Machines for Large–Scale Regression Problems, *J. Machine Learning Res.* **2001**, *1*, 143–160, http://www.jmlr.org.
[23] O. L. Mangasarian and D. R. Musicant, Lagrangian Support Vector Machines, *J. Machine Learning Res.* **2001**, *1*, 161–177, http://www.jmlr.org.
[24] M. P. S. Brown, W. Noble Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D. Haussler, Knowledge–Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines, *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 262–267.
[25] A. Zien, G. Ratsch, S. Mika, B. Schölkopf, T. Lengauer, and K. R. Muller, Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites, *Bioinformatics* **2000**, *16*, 799–807.
[26] R. J. Carter, I. Dubchak, and S. R. Holbrook, A Computational Approach to Identify Genes for Functional RNAs in Genomic Sequences, *Nucleic Acids Res*. **2001**, *29*, 3928–3938.
[27] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, Support Vector Machine

Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, *Bioinformatics* **2000**, *16*, 906–914.
[28] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures, *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 15149–15154.
[29] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines, *Machine Learning* **2002**, *46*, 389–422.
[30] C.–H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub, Molecular Classification of Multiple Tumor Types, *Bioinformatics* **2001**, *17*, S316–S322.
[31] S. Dreiseitl, L. Ohno–Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions, *J. Biomed. Informat.* **2001**, *34*, 28–36.
[32] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein, *J. Comput. Chem.* **2002**, *23*, 267–274.
[33] R. Karchin, K. Karplus, and D. Haussler, Classifying G–Protein Coupled Receptors with Support Vector Machines, *Bioinformatics* **2002**, *18*, 147–159.
[34] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Prediction of Protein Structural Classes by Support Vector Machines, *Comput. Chem.* **2002**, *26*, 293–296.
[35] Y.–D. Cai, X.–J. Liu, X. Xu, and K.–C. Chou, Support Vector Machines for Predicting Membrane Protein Types by Incorporating Quasi–Sequence–Order Effect, *Internet Electron. J. Mol. Des.* **2002**, *1*, 219–226, http://www.biochempress.com.
[36] J. R. Bock and D. A. Gough, Predicting Protein–Protein Interactions from Primary Structure, *Bioinformatics* **2001**, *17*, 455–460.
[37] S. J. Hua and Z. R. Sun, Support Vector Machine Approach for Protein Subcellular Localization Prediction, *Bioinformatics* **2001**, *17*, 721–728.
[38] Y.–D. Cai, X.–J. Liu, X.–B. Xu, and K.–C. Chou, Support Vector Machines for Prediction of Protein Subcellular Location, *Mol. Cell Biol. Res. Commun.* **2000**, *4*, 230–233.
[39] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Support Vector Machines for Prediction of Protein Subcellular Location by Incorporating Quasi–Sequence–Order Effect, *J. Cell. Biochem.* **2002**, *84*, 343–348.
[40] C. H. Q. Ding and I. Dubchak, Multi–Class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics* **2001**, *17*, 349–358.
[41] S. J. Hua and Z. R. Sun, A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach, *J. Mol. Biol.* **2001**, *308*, 397–407.
[42] Y.–D. Cai, X.–J. Liu, X.–B. Xu, and K.–C. Chou, Support Vector Machines for Predicting the Specificity of GalNAc–Transferase, *Peptides* **2002**, *23*, 205–208.
[43] W. Vercoutere, S. Winters–Hilt, H. Olsen, D. Deamer, D. Haussler, and M. Akeson, Rapid Discrimination Among Individual DNA Hairpin Molecules at Single–Nucleotide Resolution Using an Ion Channel, *Nat. Biotechnol.* **2001**, *19*, 248–252.
[44] C. W. Morris, A. Autret, and L. Boddy, Support Vector Machines for Identifying Organisms – A Comparison with Strongly Partitioned Radial Basis Function Networks, *Ecological Model*. **2001**, *146*, 57–67.
[45] O. Ivanciuc, Support Vector Machine Identification of the Aquatic Toxicity Mechanism of Organic Compounds, *Internet Electron. J. Mol. Des.* **2002**, *1*, 157–172, http://www.biochempress.com.
[46] O. Ivanciuc, Support Vector Machine Classification of the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons, *Internet Electron. J. Mol. Des.* **2002**, *1*, 203–218, http://www.biochempress.com.
[47] O. Ivanciuc, Structure–Odor Relationships for Pyrazines with Support Vector Machines, *Internet Electron. J. Mol. Des.* **2002**, *1*, 269–284, http://www.biochempress.com.
[48] S. Rüping, mySVM, University of Dortmund, http://www–ai.cs.uni–dortmund.de/SOFTWARE/MYSVM/.
[49] BioChem Links, http://www.biochempress.com.