Inter*net* Electronic Journal of
Molecular Design

May 2003, Volume 2, Number 5, Pages 348–357

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 9

Guest Editor: Jun–ichi Aihara

Support Vector Machines Classification of Black and Green Teas Based on Their Metal Content

Ovidiu Ivanciuc

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

Received: November 21, 2002; Accepted: February 3, 2002; Published: May 31, 2003

Citation of the article:

O. Ivanciuc, Support Vector Machines Classification of Black and Green Teas Based on Their Metal Content, *Internet Electron. J. Mol. Des.* **2003**, *2*, 348–357, http://www.biochempress.com.

Internet Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machines Classification of Black and Green Teas Based on Their Metal Content[#]

Ovidiu Ivanciuc*

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

Received: November 21, 2002; Accepted: February 3, 2002; Published: May 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (5), 348–357

Abstract

Motivation. Green and black teas are made from the processed leaves of *Camellia sinensis*. The metal content (Zn, Mn, Mg, Cu, Al, Ca, Ba, and K) of commercial tea samples, determined by inductively coupled plasma atomic emission spectroscopy, can be used in pattern recognition models to discriminate between the two tea types.

Method. We have investigated the application of SVM (support vector machines) for the classification of 44 tea samples (26 black tea and 18 green tea) based on the metal content. An efficient algorithm was tested for the selection of input parameters for the SVM models, in order to find the minimum metal profile that provides a good separation of the two classes.

Results. Using the hierarchical descriptor selection procedure, the initial group of eight metals was reduced to a set of three metals, namely Al, Ba, and K. The classification of the green and black teas was done with the dot, polynomial, radial basis function, neural, and anova kernels. The calibration and leave–20%–out cross–validation results show that the statistical performances of SVM models depend strongly on input descriptors, kernel type and various parameters that control the kernel shape. Several SVM models obtained with the anova kernel offered the best results, all with no error in calibration and one error in prediction (for a green tea sample).

Conclusions. The hierarchical descriptor selection algorithm is an effective procedure to identify the optimum set of input variables for an SVM model. Using the Al, Ba, and K content determined with the inductively coupled plasma atomic emission spectroscopy, a highly predictive SVM model was developed for the classification of green and black teas.

Keywords. Support vector machines; SVM; tea classification.

1 INTRODUCTION

Tea is obtained from the processed leaves of *Camellia sinensis*. Green tea is obtained by drying and roasting the leaves, while for obtaining black tea the leaves are additionally fermented. The Oolong tea is obtained when the fermentation is partial. In a recent study, the concentration of eight metals (Zn, Mn, Mg, Cu, Al, Ca, Ba, and K) was determined by inductively coupled plasma atomic

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

* Correspondence author; E–mail: ovidiu_ivanciuc@yahoo.com.

emission spectroscopy (ICP–AES) for 48 samples of commercial tea (26 black tea; 18 green tea; 4 Oolong tea) [1]. The concentrations of these metals were used as chemical descriptors in pattern recognition models to discriminate between the three tea types. Linear discriminant analysis (LDA) and multi–layer feedforward artificial neural networks (ANN) were used for the classification of these 48 tea samples. Support vector machines (SVM) 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 classification of 44 black and green tea samples based on the metal content. 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. A new algorithm for selecting relevant structural descriptors in SVM models was tested with good results in reducing the input space.

2 MATERIALS AND METHODS

2.1 Chemical Data

The metal content (Zn, Mn, Mg, Cu, Al, Ca, Ba, and K) of 44 commercial tea samples, expressed as mg/kg in dry basis, was used as chemical descriptors for the SVM model. The experimental values for 44 samples (26 black tea, class +1; 18 green tea, class –1) were taken from literature [1]. Because only 4 samples of Oolong tea were determined, we did not consider this tea type in our SVM model. In Table 1 we present the Al, Ba, and K content of the 44 samples, together with their experimental classification into black or green tea.

2.2 Structure–Toxicity Models with Support Vector Machines

Support vector machines were developed by Vapnik [2–4] as an effective algorithm for determining an optimal hyperplane to separate two classes of patterns [5–11]. 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 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 [12], translation initiation sites [13], genes [14], cancer type [15–18], pigmented skin lesions [19], HIV protease cleavage sites [20], GPCR type [21], protein class [22], membrane protein type [23], protein–protein interactions [24], protein subcellular localization [25–27], protein fold [28], protein secondary structure [29], specificity of GalNAc–transferase [30], DNA hairpins [31], aquatic toxicity mechanism of action [32,33], carcinogenic activity of polycyclic aromatic hydrocarbons [34], structure–odor relationships for pyrazines [35], cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se [36].

Table 1. Metal content (mg kg^{-1} dry basis) of tea samples and type (class +1, black tea; class –1, green tea)

No	Al	Ba	K	Class
1	625	17.0	15100	+1
2	701	19.6	15037	+1
3	719	19.2	15071	+1
4	776	15.4	14683	+1
5	770	16.2	15254	+1
6	840	15.2	15521	+1
7	856	15.4	15268	+1
8	1377	15.0	14804	+1
9	932	14.4	14802	+1
10	1483	14.0	13796	+1
11	950	13.8	14714	+1
12	1019	13.0	14448	+1
13	823	20.0	10151	+1
14	969	14.2	15693	+1
15	938	19.2	12578	–1
16	975	26.8	14049	–1
17	941	16.0	15506	+1
18	883	16.2	14095	+1
19	910	24.0	12582	–1
20	778	12.0	14162	+1
21	625	33.2	16137	–1
22	821	23.4	16212	+1
23	1725	31.8	9011	+1
24	1126	19.6	14937	+1
25	593	26	16902	–1
26	971	36.4	17844	–1
27	1150	20.8	14263	+1
28	1046	18.4	16728	+1
29	1012	21.0	9906	+1
30	383	3.8	20481	–1
31	1427	22.4	10997	+1
32	1130	26.8	10071	+1
33	167	19.0	17837	+1
34	769	15.2	24264	–1
35	831	25.2	18670	–1
36	757	19.4	23922	–1
37	767	13.6	20260	–1
38	685	24.8	23534	–1
39	833	25.2	24146	–1
40	703	21.8	24619	–1
41	1129	29.6	19772	–1
42	1105	30.6	24251	–1
43	1682	31.8	20932	–1
44	659	14.6	23346	–1

In this study we have investigated the application of SVM for the classification of black and green tea using as chemical descriptors the concentration of Zn, Mn, Mg, Cu, Al, Ca, Ba, and K was determined by ICP–AES. All SVM models from the present paper for the classification of polar and nonpolar pollutants were obtained with mySVM [37], which is freely available for download. Links to Web resources related to SVM, namely tutorials, papers and software, can be found in BioChem

Links [38] 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–20%–out (L20%O) 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:

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

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, and 2):

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

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):

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

2.3 Descriptor Selection in Support Vector Machines

All studies that develop QSAR models from a large set of structural descriptors use a wide range of algorithms for selecting significant descriptors. Currently, there is no widely accepted algorithm for selecting the best group of descriptors for an SVM model. Because an exhaustive test of all combinations of descriptors requires too large computational resources, we have used a heuristic method for descriptor selection.

This heuristic algorithm starts from the set of 8 chemical descriptors from [1] (namely, the concentration of Zn, Mn, Mg, Cu, Al, Ca, Ba, and K) and generates an optimal set of descriptors by applying the following steps:

(1) Starting from the complete group of *N* descriptors, all SVM models with one descriptor each are computed. For each descriptor or group of descriptors, 78 experiments were performed using the

dot, polynomial, radial basis function, neural, and anova kernels, with various parameters (see Eqs. (1)–(5) and Table 2). The prediction performances of each SVM experiment are evaluated with the L20%O cross–validation procedure, and the accuracy index AC is computed for each experiment, namely AC = (TP + TN)/(TP + FP + TN + FN), where TP is the true positive number, FP is the false positive number, TN is the true negative number, and FN is the false negative number. The descriptor that gives the maximum prediction AC is selected for further experiments.

(2) Using the descriptor selected in step (1) and each of the remaining $N - 1$ descriptors, pairs of descriptors are tested in SVM models. The pair of descriptors with the maximum prediction AC is selected for further experiments.

(3) In each step, a new descriptor is selected, namely the one that, together with the descriptors selected in previous steps, gives the maximum prediction AC. The process stops when prediction AC does not increase by adding a new descriptor, or when a certain maximum number of descriptors are selected.

3 RESULTS AND DISCUSSION

The results of the descriptor selection algorithm show that SVM models obtained with the concentration of Al, Ba, and K give the maximum prediction AC = 0.98, with only one error in the L20%O prediction, namely sample **15**, which is predicted as black tea. Because adding another descriptor does not increase the prediction AC, we will discuss only SVM models obtained with these three metal concentrations.

The SVM statistical results obtained with the Al, Ba, and K concentrations are presented in Table 2. The calibration of the SVM models was performed with the whole set of 44 compounds (26 black tea, SVM class +1; 18 green tea, SVM class –1). The calibration results reported in Table 2 are: TP_c, true positive in calibration, the number of +1 patterns (nonpolar compounds) computed in class +1; FN_c, false negative in calibration, the number of +1 patterns computed in class –1; TN_c, true negative in calibration, the number of –1 patterns (polar compounds) computed in class –1; FP_c, false positive in calibration, the number of –1 patterns computed in class +1; SV_c, number of support vectors in calibration; BSV_c, number of bounded support vectors in calibration; AC_c, calibration accuracy. Using complex kernels, SVM models can be calibrated to perfectly discriminate two populations of patterns, but only a cross–validation prediction test can demonstrate the potential utility of an SVM model. For each SVM model we present in Table 2 the following leave–20%–out cross–validation statistics: TP_p, true positive in prediction; FN_p, false negative in prediction; TN_p, true negative in prediction; FP_p, false positive in prediction; SV_p, average number of support vectors in prediction; BSV_p, average number of bounded support vectors in prediction; AC_p, prediction accuracy.

Table 2. Results for SVM classification of black and green tea using Al, Ba, and K concentration as input data. [a]

Exp	*C*	*K*			TP_c	FN_c	TN_c	FP_c	SV_c	BSV_c	AC_c	TP_p	FN_p	TN_p	FP_p	SV_p	BSV_p	AC_p
1	10	D			24	2	15	3	14	10	0.89	24	2	14	4	11.6	7.6	0.86
2	100				24	2	15	3	14	10	0.89	24	2	14	4	12.0	7.4	0.86
3	1000				24	2	15	3	14	10	0.89	24	2	14	4	12.0	7.6	0.86
			d															
4	10	P	2		25	1	16	2	13	3	0.93	24	2	14	4	11.6	2.2	0.86
5	100		2		25	1	16	2	13	3	0.93	23	3	13	5	10.6	1.4	0.82
6	1000		2		25	1	16	2	13	3	0.93	23	3	13	5	10.4	1.0	0.82
7	10		3		26	0	18	0	13	0	1.00	23	3	11	7	10.6	0.0	0.77
8	100		3		26	0	18	0	13	0	1.00	23	3	11	7	10.6	0.0	0.77
9	1000		3		26	0	18	0	13	0	1.00	23	3	11	7	10.6	0.0	0.77
10	10		4		26	0	18	0	13	0	1.00	23	3	12	6	11.0	0.0	0.80
11	100		4		26	0	18	0	13	0	1.00	23	3	12	6	11.0	0.0	0.80
12	1000		4		26	0	18	0	13	0	1.00	23	3	12	6	11.0	0.0	0.80
13	10		5		26	0	18	0	14	0	1.00	23	3	13	5	10.4	0.0	0.82
14	100		5		26	0	18	0	14	0	1.00	23	3	13	5	10.4	0.0	0.82
15	1000		5		26	0	18	0	14	0	1.00	23	3	13	5	10.4	0.0	0.82
			γ															
16	10	R	0.5		25	1	17	1	20	2	0.95	23	3	14	4	18.6	1.8	0.84
17	100		0.5		26	0	18	0	15	0	1.00	24	2	11	7	15.2	0.0	0.80
18	1000		0.5		26	0	18	0	15	0	1.00	24	2	11	7	15.2	0.0	0.80
19	10		1.0		26	0	18	0	24	1	1.00	23	3	13	5	22.4	0.2	0.82
20	100		1.0		26	0	18	0	24	0	1.00	23	3	13	5	22.4	0.0	0.82
21	1000		1.0		26	0	18	0	24	0	1.00	23	3	13	5	22.4	0.0	0.82
22	10		2.0		26	0	18	0	33	0	1.00	22	4	15	3	28.2	0.0	0.84
23	100		2.0		26	0	18	0	33	0	1.00	22	4	15	3	28.2	0.0	0.84
24	1000		2.0		26	0	18	0	33	0	1.00	22	4	15	3	28.2	0.0	0.84
			a	*b*														
25	10	N	0.5	0.0	22	4	13	5	14	12	0.80	19	7	12	6	12.0	9.4	0.70
26	100		0.5	0.0	22	4	12	6	14	10	0.77	17	9	14	4	11.4	8.8	0.70
27	1000		0.5	0.0	22	4	12	6	13	10	0.77	17	9	13	5	11.4	8.4	0.68
28	10		1.0	0.0	19	7	11	7	14	14	0.68	16	10	12	6	12.6	10.0	0.64
29	100		1.0	0.0	19	7	11	7	14	14	0.68	15	11	12	6	12.4	10.2	0.61
30	1000		1.0	0.0	19	7	12	6	14	14	0.70	15	11	12	6	12.0	9.8	0.61
31	10		2.0	0.0	20	6	11	7	17	13	0.70	19	7	12	6	12.4	10.8	0.70
32	100		2.0	0.0	20	6	11	7	16	13	0.70	17	9	13	5	12.4	11.0	0.68
33	1000		2.0	0.0	20	6	11	7	16	13	0.70	17	9	13	5	12.4	11.0	0.68
34	10		0.5	1.0	18	8	10	8	18	16	0.64	14	12	10	8	15.0	13.6	0.55
35	100		0.5	1.0	16	10	10	8	16	16	0.59	15	11	10	8	13.6	12.2	0.57
36	1000		0.5	1.0	16	10	10	8	16	16	0.59	15	11	10	8	13.6	12.2	0.57
37	10		1.0	1.0	18	8	10	8	18	16	0.64	17	9	9	9	14.6	13.8	0.59
38	100		1.0	1.0	18	8	10	8	18	16	0.64	16	10	9	9	14.6	13.4	0.57
39	1000		1.0	1.0	18	8	10	8	18	16	0.64	16	10	9	9	14.8	13.4	0.57
40	10		2.0	1.0	19	7	10	8	18	15	0.66	18	8	8	10	14.4	13.2	0.59
41	100		2.0	1.0	19	7	10	8	18	15	0.66	15	11	7	11	14.6	13.2	0.50
42	1000		2.0	1.0	19	7	10	8	18	15	0.66	15	11	7	11	14.6	13.2	0.50
43	10		0.5	2.0	18	8	9	9	20	18	0.61	15	11	11	7	15.6	15.2	0.59
44	100		0.5	2.0	8	18	13	5	18	18	0.48	15	11	10	8	14.0	12.4	0.57
45	1000		0.5	2.0	8	18	13	5	18	18	0.48	16	10	11	7	12.0	11.2	0.61
46	10		1.0	2.0	17	9	9	9	22	20	0.59	18	8	9	9	16.0	14.4	0.61
47	100		1.0	2.0	8	18	11	7	20	20	0.43	13	13	8	10	15.6	14.4	0.48
48	1000		1.0	2.0	8	18	10	8	18	18	0.41	16	10	8	10	14.8	13.6	0.55
49	10		2.0	2.0	14	12	9	9	20	20	0.52	16	10	6	12	16.4	14.2	0.50
50	100		2.0	2.0	13	13	9	9	20	20	0.50	16	10	7	11	16.4	14.2	0.52
51	1000		2.0	2.0	13	13	9	9	20	20	0.50	16	10	6	12	16.4	14.2	0.50

Table 2. (Continued)

Exp	*C*	*K*	γ	*d*	TP$_c$	FN$_c$	TN$_c$	FP$_c$	SV$_c$	BSV$_c$	AC$_c$	TP$_p$	FN$_p$	TN$_p$	FP$_p$	SV$_p$	BSV$_p$	AC$_p$
52	10	A	0.5	1	25	1	17	1	13	5	0.95	24	2	17	1	11.2	3.6	0.93
53	100		0.5	1	26	0	18	0	11	0	1.00	22	4	16	2	10.0	0.2	0.86
54	1000		0.5	1	26	0	18	0	11	0	1.00	22	4	16	2	10.0	0.0	0.86
55	10		1.0	1	26	0	18	0	14	1	1.00	23	3	16	2	12.4	1.0	0.89
56	100		1.0	1	26	0	18	0	14	0	1.00	23	3	16	2	11.2	0.0	0.89
57	1000		1.0	1	26	0	18	0	14	0	1.00	23	3	16	2	11.2	0.0	0.89
58	10		2.0	1	26	0	18	0	16	0	1.00	24	2	17	1	14.0	0.0	0.93
59	100		2.0	1	26	0	18	0	16	0	1.00	24	2	17	1	14.0	0.0	0.93
60	1000		2.0	1	26	0	18	0	16	0	1.00	24	2	17	1	14.0	0.0	0.93
61	10		0.5	2	26	0	18	0	12	0	1.00	24	2	16	2	12.8	0.0	0.91
62	100		0.5	2	26	0	18	0	12	0	1.00	24	2	16	2	12.8	0.0	0.91
63	1000		0.5	2	26	0	18	0	12	0	1.00	24	2	16	2	12.8	0.0	0.91
64	10		1.0	2	26	0	18	0	22	0	1.00	26	0	17	1	19.0	0.0	0.98
65	100		1.0	2	26	0	18	0	22	0	1.00	26	0	17	1	19.0	0.0	0.98
66	1000		1.0	2	26	0	18	0	22	0	1.00	26	0	17	1	19.0	0.0	0.98
67	10		2.0	2	26	0	18	0	29	0	1.00	25	1	17	1	23.8	0.0	0.95
68	100		2.0	2	26	0	18	0	29	0	1.00	25	1	17	1	23.8	0.0	0.95
69	1000		2.0	2	26	0	18	0	29	0	1.00	25	1	17	1	23.8	0.0	0.95
70	10		0.5	3	26	0	18	0	19	0	1.00	24	2	15	3	16.4	0.0	0.89
71	100		0.5	3	26	0	18	0	19	0	1.00	24	2	15	3	16.4	0.0	0.89
72	1000		0.5	3	26	0	18	0	19	0	1.00	24	2	15	3	16.4	0.0	0.89
73	10		1.0	3	26	0	18	0	24	0	1.00	23	3	17	1	21.4	0.0	0.91
74	100		1.0	3	26	0	18	0	24	0	1.00	23	3	17	1	21.4	0.0	0.91
75	1000		1.0	3	26	0	18	0	24	0	1.00	23	3	17	1	21.4	0.0	0.91
76	10		2.0	3	26	0	18	0	31	0	1.00	24	2	16	2	26.4	0.0	0.91
77	100		2.0	3	26	0	18	0	31	0	1.00	24	2	16	2	26.4	0.0	0.91
78	1000		2.0	3	26	0	18	0	31	0	1.00	24	2	16	2	26.4	0.0	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 (TP$_c$, true positive in calibration; FN$_c$, false negative in calibration; TN$_c$, true negative in calibration; FP$_c$, false positive in calibration; SV$_c$, number of support vectors in calibration; BSV$_c$, number of bounded support vectors in calibration; AC$_c$, calibration accuracy) and L20%O prediction results (TP$_p$, true positive in prediction; FN$_p$, false negative in prediction; TN$_p$, true negative in prediction; FP$_p$, false positive in prediction; SV$_p$, average number of support vectors in prediction; BSV$_p$, average number of bounded support vectors in prediction; AC$_p$, prediction accuracy).

The first group of SVM models computed with the Al, Ba, and K concentrations were obtained with the dot kernel, with AC$_c$ = 0.89 and AC$_p$ = 0.86 (experiments 1–3). Also, the neural kernel, with AC$_c$ between 0.41 and 0.80 and AC$_p$ between 0.48 and 0.70, is not a good candidate for the SVM models that discriminate black and green tea. The calibration AC$_c$ increases to 1 for almost all SVM models having a polynomial, radial basis function, and anova kernel. In the L20%O prediction, the best results were obtained with the anova kernel: polynomial kernel, AC$_p$ between 0.77 and 0.86; radial kernel, AC$_p$ between 0.80 and 0.84; anova kernel, AC$_p$ between 0.86 and 0.98.

The best predictions were obtained in experiments 64–66 with the anova kernel, AC$_c$ = 1, and AC$_p$ = 0.98. Only one error was obtained in the L20%O prediction, namely sample **15**, which is predicted as black tea in all three experiments. The LDA and ANN classification models reported in the literature [1] use as input descriptors the concentration of all eight metals and have fairly good AC values: LDA, calibration 0.95 and prediction 0.90; ANN, calibration 1 and prediction 0.95. These results are not directly comparable with the results reported in this paper, because in Ref. [1]

the Oolong tea samples were considered and different cross–validation methods have been used. The results obtained with SVM use a much lower number of chemical descriptors and give slightly better prediction statistics. The major finding is the possibility to reduce the input space from eight to three metal concentrations, which shows that the SVM descriptor selection algorithm is very effective.

4 CONCLUSIONS

Support vector machines represent an attractive new class of machine learning algorithms that can have significant applications in developing structure–activity models, chemometrics, and design of chemical libraries. In this study we have investigated the application of SVM (support vector machines) for the classification of 44 tea samples (26 black tea and 18 green tea) based on the metal concentration. An efficient algorithm was tested for the selection of input parameters for the SVM models, in order to find the minimum metal concentration profile that provides a good separation of the two classes. Using the hierarchical descriptor selection procedure, the initial group of eight metals was reduced to a set of three metals, namely Al, Ba, and K. The classification of the green and black teas was done with the dot, polynomial, radial basis function, neural, and anova kernels. The calibration and leave–20%–out cross–validation results show that the statistical performances of SVM models depend strongly on input descriptors, kernel type and various parameters that control the kernel shape. Several SVM models obtained with the anova kernel offered the best results, all with no error in calibration and one error in prediction (for a green tea sample).

The experiments reported in this paper show that the hierarchical descriptor selection algorithm is an effective procedure to identify the optimum set of input variables for an SVM model. Using the Al, Ba, and K content determined with the inductively coupled plasma atomic emission spectroscopy, an SVM model obtained with the anova kernel can effectively discriminate between green and black tea samples.

Supplementary Material

The mySVM model files for experiments 64, 65, and 66 are available as supplementary material.

5 REFERENCES

[1] M. Á. Herrador and A. G. González, Pattern Recognition Procedures for Differentiation of Green, Black and Oolong Teas According to Their Metal Content from Inductively Coupled Plasma Atomic Emission Spectrometry, *Talanta* **2001**, *53*, 1249–1257.

[2] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[4] V. Vapnik, *Statistical Learning Theory*, Wiley–Interscience, New York, 1998.

[5] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowledge Discov.* **1998**, *2*, 121–167.

[6] 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.

[7] V. N. Vapnik, An Overview of Statistical Learning Theory, *IEEE Trans. Neural Networks* **1999**, *10*, 988–999.

[8] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods*: *Support Vector Learning*, MIT Press, Cambridge, MA, 1999.

[9] N. Cristianini and J. Shawe–Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines, *Machine Learning* **2002**, *46*, 389–422.

[18] 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.

[19] 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.

[20] 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.

[21] R. Karchin, K. Karplus, and D. Haussler, Classifying G–Protein Coupled Receptors with Support Vector Machines, *Bioinformatics* **2002**, *18*, 147–159.

[22] 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.

[23] 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.

[24] J. R. Bock and D. A. Gough, Predicting Protein–Protein Interactions from Primary Structure, *Bioinformatics* **2001**, *17*, 455–460.

[25] S. J. Hua and Z. R. Sun, Support Vector Machine Approach for Protein Subcellular Localization Prediction, *Bioinformatics* **2001**, *17*, 721–728.

[26] 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.

[27] 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.

[28] C. H. Q. Ding and I. Dubchak, Multi–Class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics* **2001**, *17*, 349–358.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] O. Ivanciuc, Aquatic Toxicity Prediction for Polar and Nonpolar Narcotic Pollutants with Support Vector

Machines, *Internet Electron. J. Mol. Des.* **2003**, *2*, 195–208, http://www.biochempress.com.

[34] 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.

[35] O. Ivanciuc, Structure–Odor Relationships for Pyrazines with Support Vector Machines, *Internet Electron. J. Mol. Des.* **2002**, *1*, 269–284, http://www.biochempress.com.

[36] 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.

[37] S. Rüping, mySVM, University of Dortmund, http://www–ai.cs.uni–dortmund.de/SOFTWARE/MYSVM/.

[38] BioChem Links, http://www.biochempress.com.