Internet *net* Electronic Journal of
Molecular Design

May 2002, Volume 1, Number 5, Pages 269–284

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Milan Randić on the occasion of the 70th birthday
Part 1

Guest Editor: Mircea V. Diudea

Structure–Odor Relationships for Pyrazines with Support Vector Machines

Ovidiu Ivanciuc

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

Received: January 12, 2002; Accepted: April 5, 2002; Published: May 31, 2002

Citation of the article:

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

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Structure–Odor Relationships for Pyrazines with Support Vector Machines[#]

Ovidiu Ivanciuc*

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

Received: January 12, 2002; Accepted: April 5, 2002; Published: May 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (5), 269–284

Abstract

Motivation. The flavor class prediction of chemical compounds can be efficiently performed with structure–odor relationships (SOR), leading to a better understanding of the mechanism of odor perception. SOR models for various odor classes were developed with a wide variety of structural descriptors and statistical equations.

Method. We have investigated the application of support vector machines (SVM) for the classification of 98 tetra–substituted pyrazines representing three odor classes, namely 32 green, 23 nutty, and 43 bell–pepper. The chemical structure of the pyrazines was encoded by five theoretical descriptors, namely the sum of electrotopological indices, the number of carbon atoms of the substituent R_2, the charge on the first atom of the substituent R_4 computed with an *ab initio* method (Hartee–Fock with a 3–21G basis set), and the molecular surface of the substituents R_1 and R_3.

Results. Three sets of SVM experiments were performed for the classification of pyrazines, each one considering the classification of one class of compounds against the compounds from the remaining two classes. The SVM models were computed with the dot, polynomial, radial basis function, neural, and anova kernels. The leave–10%–out cross–validation results represent the main criterion for selecting the best SVM model that has the highest prediction power. The results obtained demonstrate that the SVM classification of pyrazines in aroma classes depends strongly on the kernel type and various parameters that control the kernel shape. In general, the neural kernel gives the worst results. The best predictions were obtained with the polynomial kernel of degree 2 for the green and bell–pepper classes, and with the anova kernel ($\gamma = 0.5$ and $d = 1$) for the nutty pyrazines.

Conclusions. The classification of chemical compounds in odor classes with SOR models can be efficiently made with support vector machines. The solution of the SVM model is a unique hyperplane that guarantees a maximum separation between two classes of chemical compounds. This hyperplane can be computed very fast and represents the solution of a quadratic programming problem, but the classification results depend on the kernel type and structural descriptors. The identification of the optimum predictive kernel and elimination of the overfitted SVM models requires extensive cross–validation experiments.

Keywords. Structure–odor relationships; pyrazine; support vector machines; machine learning; kernel algorithm.

1 INTRODUCTION

Various techniques of molecular design can significantly help fragrance researchers to find relationships between the chemical structure and the odor of organic compounds [1–3]. A wide variety of structural descriptors (molecular fragments, topological indices, geometric descriptors, or

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

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

quantum indices) and a broad selection of qualitative or quantitative statistical equations were used to model and predict the aroma and its intensity for various classes of organic compounds [3–14]. Besides providing an important guidance for the synthesis of new fragrances, structure–odor relationships (SOR) offer a better understanding of the mechanism of odor perception.

Support vector machines (SVM) represent a new class of machine learning algorithms for classification and regression [15–28] with numerous applications in medicine, bioinformatics and chemistry [29–51]. In this paper we present the first application of support vector machines for the aroma classification, using literature data [14] for 98 tetra–substituted pyrazines representing three odor classes, namely 32 green, 23 nutty, and 43 bell–pepper.

2 MATERIALS AND METHODS

2.1 Chemical Data

A database of 98 tetra–substituted pyrazines (see Figure 1 for the general structure and Table 1 for the substituents and aroma classes) representing three odor classes, namely 32 green, 23 nutty, and 43 bell–pepper, was taken from literature [14]. The chemical structure of the 98 pyrazines was encoded by five theoretical descriptors, namely the sum of electrotopological indices, the number of carbon atoms of the substituent R_2, the charge on the first atom of the substituent R_4 computed with an *ab initio* method (Hartee–Fock with a 3–21G basis set), and the molecular surface of the substituents R_1 and R_3 [14]. These five structural descriptors were used in a neural network model to separate the pyrazines into aroma classes [14]. While the descriptors used to quantify the chemical structure are very important in SVM models, no algorithm is presently available for the efficient selection of a group of effective structural descriptors that allow an optimum separation into classes of the chemical compounds.

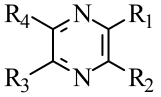

Figure 1. General structure of the pyrazines.

2.2 Support Vector Machines

Support vector machines were developed by Vapnik [15–17] as an effective algorithm for determining an optimal hyperplane to separate two classes of patterns [18–28]. 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.

Table 1. Structure and Aroma Class (Green = 1; Nutty = 2; Bell–Pepper = 3) for the 98 Pyrazines

No	R_1	R_2	R_3	R_4	Class
1	$N(CH_3)_2$	H	H	$CH_2CH(CH_3)_2$	1
2	OC_4H_9	H	H	H	1
3	OC_6H_5	H	$CH(CH_3)_2$	H	1
4	SC_2H_5	H	$(CH_2)_2CH(CH_3)C_2H_5$	H	1
5	$N(CH_3)_2$	CH_3	H	H	1
6	OCH_3	CH_3	$CH_2CH(CH_3)_2$	H	1
7	OCH_3	CH_3	$CH_2CH(CH_3)C_2H_5$	H	1
8	OCH_3	CH_3	$CH_2CH_2CH(CH_3)C_2H_5$	H	1
9	OC_2H_5	CH_3	$CH(CH_3)C_2H_5$	H	1
10	OC_2H_5	CH_3	$CH_2CH(CH_3)_2$	H	1
11	OC_2H_5	CH_3	$CH_2CH(CH_3)C_2H_5$	H	1
12	OC_2H_5	CH_3	$(CH_2)_2CH(CH_3)C_2H_5$	H	1
13	OC_6H_5	CH_3	$CH_2CH(CH_3)_2$	H	1
14	OC_6H_5	CH_3	$(CH_2)_2CH(CH_3)C_2H_5$	H	1
15	SCH_3	CH_3	$CH_2CH(CH_3)_2$	H	1
16	SCH_3	CH_3	$CH_2CH(CH_3)C_3H_7$	H	1
17	SC_2H_5	CH_3	$CH_2CH(CH_3)C_2H_5$	H	1
18	OCH_3	$COH(CH_3)_2$	CH_3	H	1
19	OCH_3	$COH(CH_3)_2$	H	CH_3	1
20	OCH_3	$COCH_3$	H	CH_3	1
21	OCH_3	$COCH_3$	OCH_3	CH_3	1
22	H	C_2H_5	H	CH_3	1
23	C_2H_5	C_2H_5	H	H	1
24	H	$CH(CH_3)_2$	CH_3	CH_3	1
25	H	C_4H_9	H	H	1
26	H	$CH_2CH(CH_3)_2$	H	H	1
27	SCH_3	$CH_2CH(CH_3)_2$	H	H	1
28	H	C_5H_{11}	H	H	1
29	H	C_5H_{11}	CH_3	CH_3	1
30	OCH_3	C_5H_{11}	H	H	1
31	CH_3	$(CH_2)_2CH(CH_3)_2$	CH_3	H	1
32	OCH_3	C_7H_{15}	H	H	1
33	CH_3	H	CH_3	H	2
34	OCH_3	H	H	H	2
35	OCH_3	H	H	CH_3	2
36	OC_2H_5	H	H	H	2
37	SCH_3	H	H	H	2
38	SCH_3	H	H	CH_3	2
39	SC_2H_5	H	H	H	2
40	CH_3	CH_3	H	H	2
41	CH_3	CH_3	CH_3	H	2
42	CH_3	CH_3	CH_3	CH_3	2
43	$NHCH_3$	CH_3	H	H	2
44	OCH_3	CH_3	H	H	2
45	OCH_3	CH_3	H	CH_3	2
46	OC_2H_5	CH_3	H	H	2
47	SCH_3	CH_3	H	H	2
48	SC_2H_5	CH_3	H	H	2
49	H	C_2H_5	H	H	2
50	H	C_2H_5	CH_3	CH_3	2
51	CH_3	C_2H_5	H	CH_3	2
52	CH_3	C_2H_5	CH_3	H	2
53	SC_2H_5	C_2H_5	H	H	2
54	H	$CH_2CH(CH_3)_2$	CH_3	CH_3	2
55	SC6H5	C_8H_{17}	H	H	2
56	CH_3	C_3H_7	H	H	3
57	OCH_3	C_3H_7	H	H	3

Table 1. (Continued)

No	R_1	R_2	R_3	R_4	Class
58	SCH_3	C_3H_7	H	H	3
59	CH_3	$CH(CH_3)_2$	H	H	3
60	OCH_3	$CH(CH_3)_2$	H	H	3
61	OCH_3	$CH(CH_3)_2$	H	CH_3	3
62	OCH_3	$CH(CH_3)_2$	CH_3	H	3
63	OCH_3	$CH(CH_3)_2$	OCH_3	CH_3	3
64	OCH_3	$CH(CH_3)_2$	CH_3	OCH_3	3
65	OCH_3	$CH(CH_3)_2$	OCH_3	$CH(CH_3)_2$	3
66	SCH_3	$CH(CH_3)_2$	H	H	3
67	OCH_3	C_4H_9	H	H	3
68	SC_2H_5	C_4H_9	H	H	3
69	CH_3	$CH_2CH(CH_3)_2$	H	H	3
70	OCH_3	$CH_2CH(CH_3)_2$	H	H	3
71	OCH_3	$CH_2CH(CH_3)_2$	H	CH_3	3
72	OCH_3	$CH_2CH(CH_3)_2$	CH_3	H	3
73	OCH_3	$CH_2CH(CH_3)_2$	CH_3	CH_3	3
74	OCH_3	$CH(CH3)C_2H_5$	H	H	3
75	OC_2H_5	C_5H_{11}	H	H	3
76	SCH_3	C_5H_{11}	H	H	3
77	SC_2H_5	C_5H_{11}	H	H	3
78	OCH_3	$(CH_2)_2CH(CH_3)_2$	H	H	3
79	OCH_3	$CH_2CH(CH_3)C_2H_5$	H	H	3
80	OCH_3	$(CH_2)_3CH=CH_2$	H	H	3
81	OCH_3	$(CH_2)_2CH=CHCH_3$ (E)	H	H	3
82	OCH_3	$(CH_2)_2CH=CHCH_3$ (Z)	H	H	3
83	OCH_3	C_6H_{13}	H	H	3
84	OCH_3	$(CH_2)_3CH(CH_3)_2$	H	H	3
85	OCH_3	$CH_2CH(CH_3)C_3H_7$	H	H	3
86	OCH_3	C_8H_{17}	H	H	3
87	OC_2H_5	C_8H_{17}	H	H	3
88	SCH_3	C_8H_{17}	H	H	3
89	SC_2H_5	C_8H_{17}	H	H	3
90	OCH_3	$C_{10}H_{21}$	H	H	3
91	OC_2H_5	$C_{10}H_{21}$	H	H	3
92	OCH_3	CH_3	OCH_3	CH_3	3
93	OCH_3	C_2H_5	H	H	3
94	OCH_3	$CH(CH_3)C_3H_7$	H	H	3
95	OCH_3	$(CH_2)_6CH(CH_3)_2$	H	H	3
96	OCH_3	$CH_2CH(CH_3)C_6H_{13}$	H	H	3
97	OCH_3	$CH_2CH(CH_3)_2$	H	$CH_2CH(CH_3)_2$	3
98	OC_2H_5	$CH_2CH(CH_3)_2$	H	H	3

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 [29], translation initiation sites [30], genes [31], cancer type [32–35], pigmented skin lesions [36], HIV protease cleavage sites [37], GPCR type [38], protein class [39], membrane protein type [40], protein–protein interactions [41], protein subcellular localization [42–44], protein fold [45], protein secondary structure [46], specificity of GalNAc–transferase [47], DNA hairpins [48], organisms [49], aquatic toxicity mechanism of action [50], carcinogenic activity of polycyclic aromatic hydrocarbons [51].

All SVM models from the present paper for the classification of pyrazines into three aroma classes were obtained with mySVM [52], which is freely available for download. Links to Web resources related to SVM, namely tutorials, papers and software, can be found in BioChem Links [53] at http://www.biochempress.com. Three groups of SVM experiments were performed for the classification of pyrazines, each one considering the classification of one class of compounds against all remaining compounds. Group 1 discriminates the 32 green compounds (class +1) against the remaining 66 compounds, group 2 discriminates the 23 nutty compounds (class +1) against the remaining 75 compounds, and group 3 discriminates the 43 bell–pepper compounds (class +1) against the remaining 55 compounds. 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:

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

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 in structure–activity studies 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.

Table 2. SVM Modeling Results for the Green Aroma (Class +1)[a]

Exp	*C*	*K*			SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
1	10	D			44	38	16	16	66	0	0.84	40.1	33.0	0.82	0.80
2	100				44	38	16	16	66	0	0.84	40.4	32.5	0.82	0.80
3	1000				44	38	16	16	66	0	0.84	44.0	32.1	0.82	0.80
			d												
4	10	P	2		28	10	26	6	65	1	0.93	27.1	8.5	0.93	0.85
5	100		2		31	9	26	6	65	1	0.93	27.6	7.5	0.93	0.85
6	1000		2		30	9	27	5	65	1	0.94	27.6	7.4	0.94	0.86
7	10		3		33	9	28	4	65	1	0.95	30.5	6.7	0.96	0.76
8	100		3		29	4	31	1	64	2	0.97	27.4	2.7	0.99	0.78
9	1000		3		27	0	32	0	66	0	1.00	25.9	0.0	1.00	0.79
10	10		4		31	2	31	1	66	0	0.99	28.1	1.6	0.99	0.74
11	100		4		33	0	32	0	66	0	1.00	28.3	0.0	1.00	0.74
12	1000		4		33	0	32	0	66	0	1.00	28.3	0.0	1.00	0.74
13	10		5		33	0	32	0	66	0	1.00	30.7	0.0	1.00	0.72
14	100		5		33	0	32	0	66	0	1.00	30.7	0.0	1.00	0.72
15	1000		5		33	0	32	0	66	0	1.00	30.7	0.0	1.00	0.72
			γ												
16	10	R	0.5		54	9	26	6	65	1	0.93	48.6	8.1	0.93	0.79
17	100		0.5		46	9	30	2	64	2	0.96	43.3	6.4	0.96	0.77
18	1000		0.5		38	0	32	0	66	0	1.00	37.8	0.0	1.00	0.75
19	10		1.0		59	9	26	6	65	1	0.93	56.1	7.2	0.94	0.78
20	100		1.0		53	1	31	1	66	0	0.99	48.9	0.9	0.99	0.78
21	1000		1.0		49	0	32	0	66	0	1.00	47.2	0.0	1.00	0.78
22	10		2.0		75	4	31	1	66	0	0.99	68.1	3.0	0.99	0.77
23	100		2.0		69	0	32	0	66	0	1.00	64.4	0.0	1.00	0.77
24	1000		2.0		69	0	32	0	66	0	1.00	64.4	0.0	1.00	0.77
			a	*b*											
25	10	N	0.5	0.0	38	35	14	18	49	17	0.64	29.8	26.7	0.71	0.73
26	100		0.5	0.0	32	32	21	11	38	28	0.60	28.9	25.9	0.71	0.73
27	1000		0.5	0.0	32	32	21	11	38	28	0.60	28.6	25.4	0.71	0.72
28	10		1.0	0.0	38	38	17	15	33	33	0.51	31.6	28.6	0.68	0.72
29	100		1.0	0.0	38	38	17	15	33	33	0.51	31.1	28.3	0.68	0.73
30	1000		1.0	0.0	38	38	17	15	33	33	0.51	31.0	28.2	0.68	0.72
31	10		2.0	0.0	40	37	13	19	48	18	0.62	32.8	29.9	0.67	0.67
32	100		2.0	0.0	39	37	13	19	48	18	0.62	32.4	29.5	0.67	0.68
33	1000		2.0	0.0	36	32	16	16	50	16	0.67	32.4	29.8	0.66	0.68
34	10		0.5	1.0	53	51	7	25	40	26	0.48	46.8	44.9	0.49	0.46
35	100		0.5	1.0	53	51	7	25	40	26	0.48	46.7	44.6	0.49	0.46
36	1000		0.5	1.0	53	50	7	25	41	25	0.49	46.6	44.8	0.50	0.47
37	10		1.0	1.0	46	46	16	16	30	36	0.47	43.9	42.4	0.51	0.54
38	100		1.0	1.0	46	46	16	16	30	36	0.47	43.5	41.7	0.52	0.53
39	1000		1.0	1.0	46	46	16	16	30	36	0.47	43.5	41.7	0.52	0.54
40	10		2.0	1.0	47	44	10	22	44	22	0.55	34.6	32.7	0.61	0.64
41	100		2.0	1.0	46	44	10	22	44	22	0.55	33.1	31.4	0.61	0.64
42	1000		2.0	1.0	46	44	10	22	44	22	0.55	33.0	31.1	0.61	0.64
43	10		0.5	2.0	50	50	14	18	26	40	0.41	48.0	46.4	0.46	0.43
44	100		0.5	2.0	50	50	14	18	26	40	0.41	46.8	45.4	0.46	0.43
45	1000		0.5	2.0	50	50	15	17	26	40	0.42	47.0	45.6	0.46	0.44
46	10		1.0	2.0	54	52	6	26	40	26	0.47	47.0	46.2	0.48	0.45
47	100		1.0	2.0	54	52	6	26	40	26	0.47	46.0	45.2	0.48	0.46
48	1000		1.0	2.0	53	51	7	25	40	26	0.48	46.3	45.3	0.48	0.45
49	10		2.0	2.0	48	48	18	14	26	40	0.45	44.6	42.9	0.51	0.49
50	100		2.0	2.0	48	45	9	23	44	22	0.54	44.9	43.1	0.49	0.53
51	1000		2.0	2.0	48	45	9	23	44	22	0.54	43.5	42.1	0.50	0.55
			γ	*d*											
52	10	A	0.5	1	38	20	24	8	63	3	0.89	36.1	18.9	0.89	0.80
53	100		0.5	1	40	18	25	7	64	2	0.91	35.1	13.6	0.91	0.84
54	1000		0.5	1	39	12	26	6	64	2	0.92	35.0	10.0	0.92	0.82
55	10		1.0	1	40	18	25	7	63	3	0.90	38.9	16.1	0.90	0.83
56	100		1.0	1	39	14	26	6	64	2	0.92	36.4	10.7	0.92	0.82
57	1000		1.0	1	40	9	28	4	66	0	0.96	34.6	6.1	0.96	0.77
58	10		2.0	1	46	18	25	7	63	3	0.90	40.8	14.8	0.91	0.82

Table 2. (Continued)

Exp	*C*	*K*	γ	*d*	SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
59	100	A	2.0	1	41	7	29	3	65	1	0.96	37.8	5.7	0.96	0.76
60	1000		2.0	1	36	3	30	2	66	0	0.98	33.9	1.9	0.99	0.70
61	10		0.5	2	43	9	26	6	66	0	0.94	38.0	7.2	0.94	0.72
62	100		0.5	2	37	4	31	1	65	1	0.98	32.8	2.5	0.99	0.73
63	1000		0.5	2	33	0	32	0	66	0	1.00	31.8	0.0	1.00	0.70
64	10		1.0	2	42	7	31	1	65	1	0.98	39.1	4.4	0.98	0.71
65	100		1.0	2	38	0	32	0	66	0	1.00	34.2	0.0	1.00	0.74
66	1000		1.0	2	38	0	32	0	66	0	1.00	34.2	0.0	1.00	0.74
67	10		2.0	2	45	1	31	1	66	0	0.99	40.4	0.9	1.00	0.71
68	100		2.0	2	41	0	32	0	66	0	1.00	40.2	0.0	1.00	0.70
69	1000		2.0	2	41	0	32	0	66	0	1.00	40.2	0.0	1.00	0.70
70	10		0.5	3	38	2	31	1	66	0	0.99	34.8	1.5	1.00	0.73
71	100		0.5	3	36	0	32	0	66	0	1.00	33.7	0.0	1.00	0.74
72	1000		0.5	3	36	0	32	0	66	0	1.00	33.7	0.0	1.00	0.74
73	10		1.0	3	40	0	32	0	66	0	1.00	37.5	0.0	1.00	0.70
74	100		1.0	3	40	0	32	0	66	0	1.00	37.5	0.0	1.00	0.70
75	1000		1.0	3	40	0	32	0	66	0	1.00	37.5	0.0	1.00	0.70
76	10		2.0	3	51	0	32	0	66	0	1.00	47.0	0.0	1.00	0.72
77	100		2.0	3	51	0	32	0	66	0	1.00	47.0	0.0	1.00	0.72
78	1000		2.0	3	51	0	32	0	66	0	1.00	47.0	0.0	1.00	0.72

[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 (green aroma) predicted in class +1; +/–, number of +1 patterns predicted in class –1; –/–, number of –1 patterns (nutty and bell–pepper compounds) 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).

Table 3. SVM Modeling Results for the Nutty Aroma (Class +1). For Notations see Table 2

Exp	*C*	*K*			SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
1	10	D			29	23	19	4	71	4	0.92	26.3	20.5	0.93	0.89
2	100				29	23	19	4	71	4	0.92	26.3	20.3	0.93	0.89
3	1000				29	23	19	4	71	4	0.92	26.3	20.3	0.93	0.89
			d												
4	10	P	2		22	5	21	2	74	1	0.97	19.9	4.0	0.97	0.89
5	100		2		21	4	22	1	74	1	0.98	18.7	2.6	0.98	0.86
6	1000		2		20	1	22	1	75	0	0.99	18.5	1.0	0.99	0.88
7	10		3		20	2	22	1	74	1	0.98	19.0	1.4	0.99	0.85
8	100		3		19	0	23	0	75	0	1.00	17.5	0.0	1.00	0.83
9	1000		3		19	0	23	0	75	0	1.00	17.5	0.0	1.00	0.83
10	10		4		22	0	23	0	75	0	1.00	20.2	0.0	1.00	0.85
11	100		4		22	0	23	0	75	0	1.00	20.2	0.0	1.00	0.85
12	1000		4		22	0	23	0	75	0	1.00	20.2	0.0	1.00	0.85
13	10		5		26	0	23	0	75	0	1.00	22.8	0.0	1.00	0.88
14	100		5		26	0	23	0	75	0	1.00	22.8	0.0	1.00	0.88
15	1000		5		26	0	23	0	75	0	1.00	22.8	0.0	1.00	0.88
			γ												
16	10	R	0.5		40	5	22	1	73	2	0.97	36.6	4.2	0.97	0.89
17	100		0.5		38	3	22	1	74	1	0.98	34.6	2.1	0.99	0.89
18	1000		0.5		33	0	23	0	75	0	1.00	32.1	0.0	1.00	0.87
19	10		1.0		53	3	22	1	73	2	0.97	50.1	2.4	0.98	0.89
20	100		1.0		52	0	23	0	75	0	1.00	48.7	0.0	1.00	0.86
21	1000		1.0		52	0	23	0	75	0	1.00	48.7	0.0	1.00	0.86
22	10		2.0		74	1	23	0	75	0	1.00	67.4	0.7	1.00	0.88
23	100		2.0		71	0	23	0	75	0	1.00	66.4	0.0	1.00	0.88
24	1000		2.0		71	0	23	0	75	0	1.00	66.4	0.0	1.00	0.88
			a	*b*											
25	10	N	0.5	0.0	24	21	12	11	65	10	0.79	21.7	18.8	0.78	0.77
26	100		0.5	0.0	22	20	13	10	65	10	0.80	20.4	17.4	0.80	0.79
27	1000		0.5	0.0	22	20	13	10	65	10	0.80	20.6	17.8	0.80	0.78
28	10		1.0	0.0	27	25	10	13	63	12	0.74	23.3	20.2	0.77	0.75
29	100		1.0	0.0	27	25	10	13	63	12	0.74	22.9	20.0	0.76	0.73

Table 3. (Continued)

Exp	*C*	*K*	*a*	*b*	SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
30	1000	N	1.0	0.0	27	25	10	13	63	12	0.74	22.8	20.5	0.73	0.74
31	10		2.0	0.0	25	23	11	12	64	11	0.77	24.3	21.7	0.72	0.69
32	100		2.0	0.0	26	23	11	12	64	11	0.77	24.5	22.2	0.71	0.69
33	1000		2.0	0.0	26	23	11	12	64	11	0.77	24.5	22.2	0.71	0.69
34	10		0.5	1.0	34	34	21	2	27	48	0.49	32.1	30.4	0.62	0.49
35	100		0.5	1.0	34	34	21	2	26	49	0.48	32.1	30.1	0.63	0.55
36	1000		0.5	1.0	34	34	21	2	26	49	0.48	32.1	29.9	0.65	0.59
37	10		1.0	1.0	35	33	6	17	59	16	0.66	30.5	29.2	0.61	0.57
38	100		1.0	1.0	36	33	6	17	59	16	0.66	30.3	29.0	0.61	0.58
39	1000		1.0	1.0	36	33	6	17	59	16	0.66	30.3	28.8	0.62	0.61
40	10		2.0	1.0	34	32	7	16	59	16	0.67	30.6	29.2	0.63	0.62
41	100		2.0	1.0	34	32	7	16	59	16	0.67	30.5	29.1	0.63	0.61
42	1000		2.0	1.0	34	32	7	16	59	16	0.67	30.4	29.0	0.63	0.61
43	10		0.5	2.0	37	35	5	18	58	17	0.64	32.3	31.5	0.54	0.56
44	100		0.5	2.0	36	34	6	17	58	17	0.65	32.1	31.3	0.53	0.56
45	1000		0.5	2.0	36	34	6	17	58	17	0.65	32.0	31.2	0.53	0.56
46	10		1.0	2.0	36	36	22	1	23	52	0.46	33.0	32.2	0.53	0.49
47	100		1.0	2.0	36	34	6	17	58	17	0.65	32.8	32.0	0.52	0.50
48	1000		1.0	2.0	36	34	6	17	58	17	0.65	32.7	31.7	0.54	0.49
49	10		2.0	2.0	35	33	6	17	59	16	0.66	30.7	29.2	0.61	0.59
50	100		2.0	2.0	35	33	6	17	59	16	0.66	30.5	29.1	0.62	0.59
51	1000		2.0	2.0	35	33	6	17	59	16	0.66	30.5	29.1	0.61	0.59
			γ	*d*											
52	10	A	0.5	1	23	9	20	3	73	2	0.95	21.2	7.7	0.96	0.92
53	100		0.5	1	23	3	22	1	73	2	0.97	19.2	3.1	0.97	0.87
54	1000		0.5	1	22	3	22	1	74	1	0.98	18.9	1.6	0.98	0.86
55	10		1.0	1	27	7	22	1	73	2	0.97	22.7	4.8	0.97	0.91
56	100		1.0	1	20	4	22	1	74	1	0.98	19.5	2.1	0.98	0.86
57	1000		1.0	1	19	0	23	0	75	0	1.00	17.6	0.0	1.00	0.88
58	10		2.0	1	25	5	22	1	73	2	0.97	22.4	3.3	0.97	0.89
59	100		2.0	1	22	0	23	0	75	0	1.00	20.4	0.0	1.00	0.81
60	1000		2.0	1	22	0	23	0	75	0	1.00	20.4	0.0	1.00	0.81
61	10		0.5	2	23	3	22	1	73	2	0.97	22.0	2.3	0.98	0.85
62	100		0.5	2	24	0	23	0	75	0	1.00	22.5	0.0	1.00	0.83
63	1000		0.5	2	24	0	23	0	75	0	1.00	22.5	0.0	1.00	0.83
64	10		1.0	2	24	1	23	0	75	0	1.00	23.5	0.6	1.00	0.85
65	100		1.0	2	24	0	23	0	75	0	1.00	23.3	0.0	1.00	0.85
66	1000		1.0	2	24	0	23	0	75	0	1.00	23.3	0.0	1.00	0.85
67	10		2.0	2	32	0	23	0	75	0	1.00	29.9	0.0	1.00	0.87
68	100		2.0	2	32	0	23	0	75	0	1.00	29.9	0.0	1.00	0.87
69	1000		2.0	2	32	0	23	0	75	0	1.00	29.9	0.0	1.00	0.87
70	10		0.5	3	26	0	23	0	75	0	1.00	23.8	0.0	1.00	0.86
71	100		0.5	3	26	0	23	0	75	0	1.00	23.8	0.0	1.00	0.86
72	1000		0.5	3	26	0	23	0	75	0	1.00	23.8	0.0	1.00	0.86
73	10		1.0	3	29	0	23	0	75	0	1.00	27.7	0.0	1.00	0.86
74	100		1.0	3	29	0	23	0	75	0	1.00	27.7	0.0	1.00	0.86
75	1000		1.0	3	29	0	23	0	75	0	1.00	27.7	0.0	1.00	0.86
76	10		2.0	3	39	0	23	0	75	0	1.00	37.5	0.0	1.00	0.88
77	100		2.0	3	39	0	23	0	75	0	1.00	37.5	0.0	1.00	0.88
78	1000		2.0	3	39	0	23	0	75	0	1.00	37.5	0.0	1.00	0.88

Table 4. SVM Modeling Results for the Bell–Pepper Aroma (Class +1). For Notations see Table 2

Exp	*C*	*K*		SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
1	10	D		48	42	34	9	46	9	0.82	43.5	37.6	0.83	0.74
2	100			48	42	34	9	46	9	0.82	43.1	36.9	0.82	0.74
3	1000			48	42	34	9	46	9	0.82	43.3	36.8	0.82	0.74
			d											
4	10	P	2	27	9	43	0	52	3	0.97	25.3	7.8	0.97	0.88
5	100		2	27	8	43	0	52	3	0.97	25.2	6.3	0.97	0.84
6	1000		2	31	8	43	0	52	3	0.97	26.6	5.9	0.97	0.85
7	10		3	27	5	43	0	52	3	0.97	25.7	3.6	0.98	0.79
8	100		3	29	3	43	0	53	2	0.98	25.1	2.1	0.98	0.76

Table 4. (Continued)

Exp	*C*	*K*	*d*		SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
9	1000	P	3		26	0	43	0	55	0	1.00	23.4	0.0	1.00	0.78
10	10		4		25	1	43	0	55	0	1.00	23.9	0.7	1.00	0.81
11	100		4		25	0	43	0	55	0	1.00	23.8	0.0	1.00	0.81
12	1000		4		25	0	43	0	55	0	1.00	23.8	0.0	1.00	0.81
13	10		5		29	0	43	0	55	0	1.00	27.5	0.0	1.00	0.82
14	100		5		29	0	43	0	55	0	1.00	27.5	0.0	1.00	0.82
15	1000		5		29	0	43	0	55	0	1.00	27.5	0.0	1.00	0.82
			γ												
16	10	R	0.5		43	7	43	0	52	3	0.97	41.9	5.9	0.97	0.89
17	100		0.5		40	5	43	0	52	3	0.97	38.0	3.8	0.98	0.83
18	1000		0.5		33	0	43	0	55	0	1.00	31.1	0.0	1.00	0.84
19	10		1.0		57	5	43	0	52	3	0.97	53.9	4.1	0.97	0.84
20	100		1.0		52	1	43	0	54	1	0.99	48.4	0.9	0.99	0.88
21	1000		1.0		51	0	43	0	55	0	1.00	48.4	0.0	1.00	0.87
22	10		2.0		76	2	43	0	54	1	0.99	67.4	1.3	0.99	0.86
23	100		2.0		68	0	43	0	55	0	1.00	63.3	0.0	1.00	0.85
24	1000		2.0		68	0	43	0	55	0	1.00	63.3	0.0	1.00	0.85
			a	*b*											
25	10	N	0.5	0.0	35	32	26	17	40	15	0.67	32.2	29.2	0.67	0.66
26	100		0.5	0.0	35	32	26	17	40	15	0.67	32.1	29.7	0.67	0.67
27	1000		0.5	0.0	36	33	27	16	38	17	0.66	31.9	29.4	0.67	0.67
28	10		1.0	0.0	38	35	26	17	37	18	0.64	34.3	31.6	0.65	0.66
29	100		1.0	0.0	38	35	26	17	37	18	0.64	33.2	30.5	0.65	0.64
30	1000		1.0	0.0	38	35	26	17	37	18	0.64	33.0	30.1	0.66	0.65
31	10		2.0	0.0	38	36	25	18	37	18	0.63	34.6	32.5	0.64	0.62
32	100		2.0	0.0	38	36	25	18	37	18	0.63	34.2	32.0	0.64	0.62
33	1000		2.0	0.0	38	36	25	18	37	18	0.63	33.9	31.7	0.64	0.65
34	10		0.5	1.0	48	46	20	23	32	23	0.53	40.5	38.5	0.57	0.58
35	100		0.5	1.0	46	43	21	22	34	21	0.56	39.4	37.4	0.58	0.60
36	1000		0.5	1.0	46	43	21	22	34	21	0.56	41.6	39.7	0.55	0.55
37	10		1.0	1.0	40	38	24	19	36	19	0.61	33.3	31.7	0.64	0.63
38	100		1.0	1.0	38	38	23	20	36	19	0.60	33.0	31.1	0.64	0.63
39	1000		1.0	1.0	34	34	28	15	35	20	0.64	32.9	31.0	0.64	0.63
40	10		2.0	1.0	44	42	22	21	35	20	0.58	32.9	30.8	0.65	0.66
41	100		2.0	1.0	43	41	22	21	35	20	0.58	32.8	30.9	0.65	0.68
42	1000		2.0	1.0	43	41	22	21	35	20	0.58	32.5	30.7	0.65	0.67
43	10		0.5	2.0	52	50	18	25	30	25	0.49	47.7	46.4	0.48	0.48
44	100		0.5	2.0	52	50	18	25	30	25	0.49	46.8	45.8	0.47	0.47
45	1000		0.5	2.0	52	50	18	25	30	25	0.49	46.8	45.8	0.47	0.47
46	10		1.0	2.0	46	46	24	19	30	25	0.55	43.9	42.9	0.51	0.57
47	100		1.0	2.0	46	46	24	19	30	25	0.55	43.0	42.4	0.52	0.58
48	1000		1.0	2.0	46	46	24	19	30	25	0.55	43.3	42.4	0.51	0.58
49	10		2.0	2.0	40	40	23	20	35	20	0.59	33.3	31.5	0.64	0.63
50	100		2.0	2.0	36	34	26	17	38	17	0.65	32.8	31.4	0.64	0.61
51	1000		2.0	2.0	40	40	23	20	35	20	0.59	32.4	31.2	0.64	0.64
			γ	*d*											
52	10	A	0.5	1	31	16	41	2	49	6	0.92	29.0	13.4	0.93	0.87
53	100		0.5	1	28	11	42	1	51	4	0.95	26.3	8.9	0.95	0.84
54	1000		0.5	1	25	6	43	0	52	3	0.97	23.7	4.3	0.97	0.82
55	10		1.0	1	32	14	41	2	50	5	0.93	29.6	11.0	0.93	0.84
56	100		1.0	1	28	5	42	1	51	4	0.95	25.8	5.2	0.97	0.82
57	1000		1.0	1	24	2	43	0	53	2	0.98	24.1	1.8	0.98	0.82
58	10		2.0	1	31	8	42	1	52	3	0.96	29.6	7.4	0.96	0.84
59	100		2.0	1	29	3	43	0	53	2	0.98	27.0	1.9	0.98	0.81
60	1000		2.0	1	26	1	43	0	55	0	1.00	24.0	0.3	1.00	0.78
61	10		0.5	2	31	5	43	0	52	3	0.97	28.1	3.7	0.97	0.84
62	100		0.5	2	28	1	43	0	54	1	0.99	25.3	0.9	0.99	0.80
63	1000		0.5	2	29	0	43	0	55	0	1.00	25.6	0.0	1.00	0.83
64	10		1.0	2	31	2	43	0	54	1	0.99	29.4	1.5	0.99	0.82
65	100		1.0	2	29	0	43	0	55	0	1.00	26.3	0.0	1.00	0.82
66	1000		1.0	2	29	0	43	0	55	0	1.00	26.3	0.0	1.00	0.82
67	10		2.0	2	36	1	43	0	55	0	1.00	33.9	0.8	1.00	0.85
68	100		2.0	2	35	0	43	0	55	0	1.00	31.8	0.0	1.00	0.84

Table 4. (Continued)

Exp	*C*	*K*	γ	*d*	SV	BSV	+/+	+/–	–/–	–/+	CAa	ASV	ABSV	TRa	TEa
69	1000	A	2.0	2	35	0	43	0	55	0	1.00	31.8	0.0	1.00	0.84
70	10		0.5	3	29	1	43	0	55	0	1.00	26.6	0.8	1.00	0.80
71	100		0.5	3	30	0	43	0	55	0	1.00	26.5	0.0	1.00	0.81
72	1000		0.5	3	30	0	43	0	55	0	1.00	26.5	0.0	1.00	0.81
73	10		1.0	3	33	0	43	0	55	0	1.00	29.1	0.0	1.00	0.83
74	100		1.0	3	33	0	43	0	55	0	1.00	29.1	0.0	1.00	0.83
75	1000		1.0	3	33	0	43	0	55	0	1.00	29.1	0.0	1.00	0.83
76	10		2.0	3	39	0	43	0	55	0	1.00	37.5	0.0	1.00	0.85
77	100		2.0	3	39	0	43	0	55	0	1.00	37.5	0.0	1.00	0.85
78	1000		2.0	3	39	0	43	0	55	0	1.00	37.5	0.0	1.00	0.85

Each group of SVM models consists of 78 experiments. Table 2 presents the statistical results for group 1 (class +1 for green compounds), Table 3 collects the SVM results for group 2 (class +1 for nutty compounds), and Table 4 offers the results for group 3 (class +1 for bell–pepper compounds). The calibration results reported in Tables 2, 3 and 4 are: SV, number of support vectors; BSV, number of bounded support vectors; +/+, number of +1 patterns predicted in class +1; +/–, number of +1 patterns predicted in class –1; –/–, number of –1 patterns predicted in class –1; –/+, number of –1 patterns predicted in class +1; CAa, accuracy. Using complex non–linear 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 and avoid overfitting. For each SVM model we present in Tables 2, 3 and 4 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. As implemented in mySVM, *C* is scaled by 1/number of training examples.

The group 1 of experiments discriminates between the 32 green compounds (class +1) against the remaining 66 nutty and bell–pepper compounds. The statistical results for these 78 SVM models presented in Table 2 show that the calibration and prediction results vary widely with the kernel function. The results obtained with the dot kernel (Table 2, experiments 1–3) do not depend on the value of *C*, with TEa = 0.80. However, the number of support vectors is quite large, and better prediction results are obtained with the polynomial and anova kernels. In the models obtained with the polynomial kernel (Table 2, experiments 4–15) TEa takes values between 0.72 and 0.86, with the best results obtained in the experiment 6, with a polynomial kernel of degree 2 and *C* = 1000. For the experiment 6, in the SVM model calibration five green compounds are classified in the class –1 (**5**, **23**, **27**, **30**, and **32**) and the nutty compound **54** is classified as green. A clear overfitting effect is observed for the polynomial kernel, with calibration results improving when the polynomial degree increases from 2 to 5 (CAa increases from 0.93 to 1), while the L10%O cross–validation TEa decreases from 0.86 to 0.72 with the increase of the polynomial degree. This is an obvious demonstration of the fact that SVM models can be overfitted when too complex kernels are used. The L10%O cross–validation test is a reliable method for locating the SVM model with the best prediction power, although other cross–validation partitioning of this group of compounds can offer equally good guiding in selecting the best kernel.

The next group of SVM models (Table 2, experiments 16–24) was obtained with the radial basis function kernel. While the calibration results are good (CAa takes values between 0.93 and 1), the prediction results are low (TEa takes values between 0.77 and 0.79), a clear sign of overfitting. The results obtained with the neural kernel (Table 2, experiments 25–51) have the lowest statistics from the group of SVM models developed for the green compounds, with low calibration (CAa takes values between 0.41 and 0.64) and prediction (TEa takes values between 0.43 and 0.73) statistics. It is clear that the neural kernel is not a good candidate for the classification of green compounds.

The last group of SVM models for the classification of the green compounds was obtained with the anova kernel (Table 2, experiments 52–78), with overall good calibration (CAa between 0.89 and 1) and L10%O prediction (TEa between 0.70 and 0.84) statistics. When γ and d increase, the calibration results increase, while the prediction statistics decrease, showing that overfitted SVM models are obtained with high values for γ and d. The best results obtained with the anova kernel (Table 2, experiment 53, with $C = 100$, $\gamma = 0.5$, $d = 1$, CAa = 0.91, and TEa = 0.84) are very close to those obtained with the best SVM model for the classification of green aroma (experiment 6), indicating that this kernel is an interesting alternative for the polynomial kernel for this classification problem. The calibration of the SVM model from the experiment 53 results in seven green compounds classified in the class –1 (**1**, **5**, **23**, **27**, **30**, **31**, and **32**) and two nutty compounds (**50** and **54**) are classified as green. Six from these nine compounds (**5**, **23**, **27**, **30**, **32**, and **54**) were not correctly separated in experiment 6, indicating that this group of molecules is difficult to characterize with this SAR model. If we consider that the experimental classification of these compounds is correct, only additional structural descriptors can improve the SVM model and allow a correct classification of all molecules.

The group 2 of experiments considers the classification of the 23 nutty compounds (class +1) against the remaining 75 green and bell–pepper compounds (class –1). A global examination of the statistical results for these 78 SVM models presented in Table 3 reveals some interesting trends regarding the predictive power of each kernel, which roughly decreases in the following order: anova, dot, radial basis function, polynomial, and neural. The results obtained with the dot kernel (Table 3, experiments 1–3) do not depend on the value of C, with CAa = 0.92 and TEa = 0.89. Using 29 support vectors, this simple kernel gives a surprisingly good classification, compared with the other, more complex, kernels. The polynomial kernel (Table 3, experiments 4–15, CAa between 0.97 and 1, TEa between 0.83 and 0.88) has overall good results, with the best predictions for experiment 4 (C =10, CAa = 0.97, TEa = 0.89) which has the same prediction statistics with the dot SVM, but the model has a lower number of support vectors (SV = 22) and better calibration statistics with only three classification errors (compounds **5**, **46**, and **54**). As observed in other SAR studies [50,51], the classification performance in calibration for the polynomial kernel increases with the increase of the degree *d*, offering a complete separation for experiments 8–15. Usually, the prediction statistics decrease when *d* increases, but for the separation of the nutty compounds the minimum prediction is obtained for experiments 8 and 9, and then TEa increases to 0.88 when *d* =5.

The SVM models for the nutty aroma obtained with the radial basis function kernel (Table 3, experiments 16–24) are fairly good, with CAa between 0.97 and 1, and TEa between 0.86 and 0.89. However, the number of support vectors is significantly larger compared with the polynomial kernel, making these SVM models of little practical interest. For example, the experiment 4 (polynomial kernel, *d* = 2, TEa = 0.89) has a calibration SVM model with 22 support vectors while in the experiment 17 (radial kernel, TEa = 0.89) the SVM model has 38 support vectors. Although we have investigated a large number of neural kernel SVM models for the classification of nutty aroma (Table 3, experiments 25–51) this group of SAR models has low calibration (CAa takes values between 0.48 and 0.80) and prediction (TEa takes values between 0.49 and 0.78) statistics. Similarly with the results obtained for the green aroma, these statistical indices indicate that the neural kernel is the worst function in the classification of the nutty compounds.

Good SVM models for the nutty aroma were obtained with the anova kernel (Table 3, experiments 52–78), with overall high calibration (CAa between 0.95 and 1) and L10%O prediction (TEa between 0.81 and 0.92) statistics. When γ and *d* increase, the calibration statistics improve and a perfect separation is obtained between nutty and non–nutty compounds. However, TEa decreases (although not monotonically), indicating that for large γ and *d* the SVM models are slightly overfitted. The best two SVM models for the classification of the nutty aroma are obtained with the anova kernel (Table 3, experiment 52, with C = 10, γ = 0.5, d = 1, CAa = 0.95, and TEa = 0.92; experiment 55, with C = 10, γ = 1, d = 1, CAa = 0.97, and TEa = 0.91), a result that adds further evidence to our previous findings indicating the anova kernel as a good candidate for highly predictive SVM models. The calibration from the experiment 52 gives an SVM model with 23 support vectors in which three nutty compounds classified in the class –1 (**46**, **53**, and **54**) and two green compounds (**5** and **22**) are classified as nutty. The SVM model from experiment 55 has only three classification errors (compounds **5**, **22**, and **54**) but the number of support vectors increases to 27 and TEa is slightly lower.

The group 3 of experiments considers the classification of the 43 bell–pepper compounds (class +1) against the remaining 55 green and nutty compounds (class –1). A comparison of the statistical results from Table 4 shows that the best prediction is obtained in experiment 16 (C = 10, radial kernel, γ = 0.5, CAa = 0.97, and TEa = 0.89) followed by experiment 4 (C = 10, polynomial kernel, d = 2, CAa = 0.97, and TEa = 0.88) and experiment 21 (C = 1000, radial kernel, γ = 1, CAa = 1, and TEa = 0.87). However, if we consider also the number of support vectors in each model, then experiment 4 (SV = 27 and ASV = 25.3) gives a better SVM model than experiment 16 (SV = 43 and ASV = 41.9). When comparable statistical results are obtained, the SVM model with fewer support vectors must be preferred. In experiments 4 and 16, the calibration SVM model classifies the green compounds **27**, **30** and **32** as having a bell–pepper aroma, while in experiment 21 all compounds are correctly classified. The neural kernel (Table 4, experiments 25–51) gives the worst results for the classification of bell–pepper compounds (CAa between 0.49 and 0.67, and TEa between 0.47 and 0.68). Considering that we have investigated a fairly large number of SVM models with a neural kernel that span a wide selection of values for the parameters *a* and *b*, the conclusion of our experiments is that the neural kernel is not fit for the classification of green, nutty, and bell–pepper aroma.

Fairly good SVM models for the classification of bell–pepper aroma were obtained with the anova kernel (Table 4, experiments 52–78), with statistical results close to those obtained with the polynomial and radial kernels (CAa between 0.92 and 1, and TEa between 0.81 and 0.78). The best SVM model obtained with the anova kernel (Table 4, experiment 52, with $C = 10$, $\gamma = 0.5$, $d = 1$, CAa = 0.92, and TEa = 0.87) has prediction results close to those from experiment 4, but with more support vectors, *i.e.* 31 instead 27. Also, the number of calibration errors is larger, with two bell–pepper compounds (**92** and **93**) classified in the class –1, and with six green compounds (**18**, **19**, **27**, **30**, **31**, and **32**) classified as having a bell–pepper aroma. The results from Table 4 indicate that the separation surface between the bell–pepper compounds and the remaining 55 pyrazines can be approximated with sufficient precision by a polynomial of degree 2 kernel, while more complex kernel functions decrease the calibration or prediction statistics. An increase in the performances of the SVM model can be obtained by investigating other sets of structural descriptors, while more complicated separation functions have little to add.

4 CONCLUSIONS

Support vector machines represent a new class of machine learning algorithms that can have significant applications in structure–activity relationships, chemometrics, and design of chemical libraries. In the SVM approach, two clusters of patterns are optimally separated with a hyperplane that maximizes the separation between the two classes. Using various kernels, a non–linear mapping transforms the input space into a higher dimensional feature space, and then a quadratic programming algorithm determines a unique maximal margin hyperplane. 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 classification models for the classification of 98 tetra–substituted pyrazines representing three odor classes, namely 32 green, 23 nutty, and 43 bell–pepper [14]. The chemical structure of the 98 pyrazines was encoded by five theoretical descriptors, namely the sum of electrotopological indices, the number of carbon atoms of the substituent R_2, the charge on the first atom of the substituent R_4 computed with an *ab initio* method (Hartee–Fock with a 3–21G basis set), and the molecular surface of the substituents R_1 and R_3 [14]. The SVM applications in structure–activity models, chemometrics, and chemical libraries clustering are only in the beginning and for the moment there are no clear rules on selecting the most efficient parameters that control the SVM performances, namely the kernel and the set of structural descriptors that are essential for the SVM model. We have explored the influence of the kernel type on the SVM performances by testing various kernels, namely the dot, polynomial, radial basis function, neural, and anova kernels. Because there is no simple algorithm for descriptor selection in SVM models, we have used the theoretical indices from [14].

The results obtained demonstrate that the SVM classification of pyrazines in aroma classes depends strongly on the kernel type and various parameters that control the kernel shape. From our SVM experiments we have selected as best models those that have the best statistics in the leave–10%–out cross–validation test. Another important parameter that must be monitored in an SVM study is the number of support vectors, and when SVM models with close L10%O statistics have been obtained, we have preferred the SVM models with a lower number of support vectors. The best predictions were obtained with the polynomial kernel of degree 2 for the green and bell–pepper classes, and with the anova kernel ($\gamma = 0.5$ and $d = 1$) for the nutty pyrazines. In general, the neural kernel gives the worst results, while the radial basis function kernel gives good results for the separation of nutty and bell–pepper aroma, but with a much larger number of support vectors than the polynomial and anova kernels. The L10%O statistics show that more complex kernels tend to overfit, as clearly indicated by the decrease of prediction statistics when the degree of the polynomial kernel increases from 2 to 5. In this study we have not addressed the important problem of selecting significant descriptors in SVM models. In QSAR studies it is generally accepted that it is more important to screen a wide variety of structural descriptors instead of using too sophisticated mathematical models. The same is true for SVM models, and the improvement of the classification of pyrazines in aroma classes can come from other sets of structural descriptors.

Supplementary Material

The mySVM model files for the classification of pyrazines with green, nutty, and bell–pepper aroma are available as supplementary material.

5 REFERENCES

[1] G. Frater, J. A. Bajgrowicz, and P. Kraft, Fragrance Chemistry, *Tetrahedron* **1998**, *54*, 7633–7703.

[2] P. Kraft, J. A. Bajgrowicz, C. Denis, and G. Frater, Odds and Trends: Recent Developments in the Chemistry of Odorants, *Angew. Chem.–Int. Edit.* **2000**, *39*, 2981–3010.

[3] K. J. Rossiter, Structure–Odor Relationships, *Chem. Rev.* **1996**, *96*, 3201–3240.

[4] M. Chastrette, D. Zakarya, and J. F. Peyraud, Structure Musk Odor Relationships for Tetralins and Indans Using Neural Networks (on the Contribution of Descriptors to the Classification), *Eur. J. Med. Chem.* **1994**, *29*, 343–348.

[5] M. Chastrette, C. El Aïdi, and J. F. Peyraud, Tetralin, Indan and Nitrobenzene Compound Structure–Musk Odor Relationship Using Neural Networks, *Eur. J. Med. Chem.* **1995**, *30*, 679–686.

[6] M. Chastrette, D. Cretin, and C. El Aïdi, Structure–Odor Relationships: Using Neural Networks in the Estimation

of Camphoraceous or Fruity Odors and Olfactory Thresholds of Aliphatic Alcohols, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 108–113.

[7] D. Zakarya, D. Cherqaoui, M. Esseffar, D. Villemin, and J. M. Cense, Application of Neural Networks to Structure Sandalwood Odour Relationships, *J. Phys. Org. Chem.* **1997**, *10*, 612–622.

[8] D. Cherqaoui, M. Esseffar, D. Villemin, J. M. Cense, M. Chastrette, and D. Zakarya, Structure Musk Odour Relationship Studies of Tetralin and Indan Compounds using Neural Networks, *New J. Chem.* **1998**, *22*, 839–843.

[9] D. Zakarya, M. Chastrette, M. Tollabi, and S. Fkih–Tetouani, Structure–Camphor Odour Relationships using the Generation and Selection of Pertinent Descriptors Approach, *Chemom. Intell. Lab. Syst.* **1999**, *48*, 35–46.

[10] A. S. Dimoglo, P. F. Vlad, N. M. Shvets, and M. N. Coltsa, Structure–Ambergris Odour Relationship Investigation in a Mixed Series of Decalin and Non–Decalin Compounds: The Electronic– Topological Approach, *New J. Chem.* **2001**, *25*, 283–288.

[11] K. Audouze, F. Ros, M. Pintore, and J. R. Chrétien, Prediction of Odours of Aliphatic Alcohols and Carbonylated Compounds using Fuzzy Partition and Self Organising Maps (SOM), *Analusis* **2000**, *28*, 625–632.

[12] R. D. M. C. Amboni, B. S. Junkes, R. A. Yunes, and V. E. F. Heinzen, Quantitative Structure–Odor Relationships of Aliphatic Esters using Topological Indices, *J. Agric. Food Chem.* **2000**, *48*, 3517–3521.

[13] G. Buchbauer, C. T. Klein, B. Wailzer, and P. Wolschann, Threshold–Based Structure–Activity Relationships of Pyrazines with Bell–Pepper Flavor, *J. Agric. Food Chem.* **2000**, *48*, 4273–4278.

[14] B. Wailzer, J. Klocker, G. Buchbauer, G. Ecker, and P. Wolschann, Prediction of the Aroma Quality and the Threshold Values of Some Pyrazines using Artificial Neural Networks, *J. Med. Chem.* **2001**, *44*, 2805–2813.

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

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

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

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

[19] 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.

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

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

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

[23] 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.

[24] C.–C. Chang and C.–J. Lin, Training ν–Support Vector Classifiers: Theory and Algorithms, *Neural Comput.* **2001**, *12*, 2119–2147.

[25] 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.

[26] 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.

[27] 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.

[28] O. L. Mangasarian and D. R. Musicant, Lagrangian Support Vector Machines, *J. Machine Learning Res.* **2001**, *1*, 161–177, http://www.jmlr.org.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.
[36] 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.
[37] 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.
[38] R. Karchin, K. Karplus, and D. Haussler, Classifying G–Protein Coupled Receptors with Support Vector Machines, *Bioinformatics* **2002**, *18*, 147–159.
[39] 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.
[40] 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.
[41] J. R. Bock and D. A. Gough, Predicting Protein–Protein Interactions from Primary Structure, *Bioinformatics* **2001**, *17*, 455–460.
[42] S. J. Hua and Z. R. Sun, Support Vector Machine Approach for Protein Subcellular Localization Prediction, *Bioinformatics* **2001**, *17*, 721–728.
[43] 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.
[44] 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.
[45] C. H. Q. Ding and I. Dubchak, Multi–Class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics* **2001**, *17*, 349–358.
[46] 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.
[47] 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.
[48] 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.
[49] 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.
[50] 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.
[51] 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.
[52] S. Rüping, mySVM, University of Dortmund, http://www–ai.cs.uni–dortmund.de/SOFTWARE/MYSVM/.
[53] BioChem Links, http://www.biochempress.com.