Support Vector Machines-Based Quantitative Structure−Property Relationship for the Prediction of Heat Capacity

C. X. Xue,[†] R. S. Zhang,[†,‡] H. X. Liu,[†] M. C. Liu,[†] Z. D. Hu,*[,†] and B. T. Fan[§]

Department of Chemistry and Department of Computer Science, Lanzhou University, Lanzhou 730000, China, and Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France

Received February 17, 2004

The support vector machine (SVM), as a novel type of learning machine, for the first time, was used to develop a Quantitative Structure−Property Relationship (QSPR) model of the heat capacity of a diverse set of 182 compounds based on the molecular descriptors calculated from the structure alone. Multiple linear regression (MLR) and radial basis function networks (RBFNNs) were also utilized to construct quantitative linear and nonlinear models to compare with the results obtained by SVM. The root-mean-square (rms) errors in heat capacity predictions for the whole data set given by MLR, RBFNNs, and SVM were 4.648, 4.337, and 2.931 heat capacity units, respectively. The prediction results are in good agreement with the experimental value of heat capacity; also, the results reveal the superiority of the SVM over MLR and RBFNNs models.

1. INTRODUCTION

The heat capacity of a substance is a measure of how well the substance stores heat. Whenever we supply heat to a material, it will necessarily cause an increase in the material's temperature. The heat capacity is defined as the amount of heat required to raise the temperature of a unit of mass of a substance by a unit change in temperature, so that $c = \Delta Q/(m\Delta T)$, where c is the specific heat capacity in J/(kg °C), ΔQ is the change in heat content in Joules, m is the mass in kg, and ΔT is the change in temperature in °C.[1] The heat capacity of the compounds is a subject of interest in terms of understanding the fundamental chemical and physical processes in combustion chemistry. It has also received attention in recent years from the point of view of safety in chemical industrial processes. Experimental heat capacity data are desirable, but due to the advancement of technology in discovery or synthesis of new compounds, there is often a significant gap between the demand for such data and their availability. Of the millions of known substances, heat capacity values are only recorded for a few thousand. Moreover, for some toxic, explosive, or radioactive compounds the experimental determination of the heat capacity is extremely difficult. Hence a reliable theoretical method for predicting the heat capacity is desired.

The quantitative structure−property relationship (QSPR) approach has become very useful in the prediction of many physicochemical properties. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from the structure alone and are not dependent on any experimental properties. Once the structure of a compound is known, any descriptor can be calculated no matter whether it is synthesized or not. So once a reliable model is established, we can use this method to predict the property of compounds. This study can tell us which of the structural factors may play an important role in the determination of a property. The QSPR approach is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physical or chemical properties, can be correlated with a change in molecular features of the compounds termed descriptors. After the calculation of the molecular descriptors, linear methods, such as multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) or nonlinear methods, e.g. neural networks, can be used in the development of a mathematical relationship between the structural descriptors and the property.

Machine learning techniques have been applied to the QSPR analysis since the late 1980s, mainly in response to increased accuracy demands. The most popular neural networks model is the back-propagation (BP) neural networks due to its simple architecture yet powerful problem-solving ability. However, the BP neural networks suffers from a number of weaknesses which include the need for a large number of controlling parameters, difficulty in obtaining a stable solution, and the danger of overfitting. Other problems with the use of neural networks concern the reproducibility of results, due largely to random initialization of the networks and variation of stopping criteria.[2] Genetic algorithms can suffer in a similar manner. The stochastic nature of both population initialization and the genetic operators used during training can make results hard to reproduce.[3] Owing to the reasons outlined above, there is a continuing need for the application of more accurate and informative techniques in QSPR analysis.

The support vector machine (SVM) is a new algorithm developed from the machine learning community. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application, such as pattern recognition problems,[4−6] drug design,[7] quantitative

* Corresponding author phone: +86−931-891-2578; fax: +86−931-891-2582; e-mail: huzd@lzu.edu.cn.
[†] Department of Chemistry, Lanzhou University.
[‡] Department of Computer Science, Lanzhou University.
[§] Université Paris 7-Denis Diderot.

10.1021/ci049934n CCC: $27.50
Published on Web 05/13/2004

structure−activity relationship (QSAR),[8] and QSPR analysis.[9−11] Nevertheless, to the best of our knowledge there is no prediction of heat capacity by the QSPR approach based on SVM.

In this work, for the first time, SVM was used for the prediction of heat capacity at 298.15 K of a diverse set of 182 compounds using descriptors calculated by the software CODESSA.[12] The aim was to establish a QSPR model that could be used for the prediction of heat capacity of a diverse set of compounds from their molecular structures alone, to show the flexible modeling ability of SVM and, at the same time, to seek the important structural features related to the heat capacity of compounds. MLR and radial basis function networks (RBFNNs) methods were also utilized to establish quantitative linear and nonlinear relationship to compare with the results obtained by SVM.

2. EXPERIMENTAL SECTION

2.1. Data Preparation. The heat capacity values of 182 compounds were collected from the database and used for this study.[13] The compounds include hydrocarbons, chlorocarbons, alcohols, acids, ketones, aldehydes, ethers, esters, amines, nitriles, sulfide, and thios. A complete list of the compounds' names and corresponding experimental heat capacities was given in Table 1. The data set was randomly divided into two subsets: the training set (2,3,4,6,7,8,10...) and the test set (1,5,9...) (136 and 46 points, respectively). The training set was used to adjust the parameters of the models, and the test set was used to evaluate its prediction ability. Leave-one-out (LOO) cross-validation was performed on the training set to select the parameters of RBFNNs and SVM.

2.2. Descriptor Calculation. All structures of the molecules were drawn with the HyperChem program and exported in a file format suitable for MOPAC.[14] The final geometries were obtained with the semiempirical PM3 method in the MOPAC 6.0 program.[15] All the geometries had been fully optimized without symmetry restrictions. In all cases frequency calculations had been performed in order to ensure that all the calculated geometries correspond to true minima. The resulted geometry was transferred into software CODESSA that can calculate constitutional, topological, geometrical, electrostatic, and quantum-chemical descriptors. The constitutional descriptors reflect the molecular composition of the compound without using the geometry or electronic structure of the molecule. The topological descriptors describe the atomic connectivity in the molecule. The geometrical descriptors describe the size of the molecule and require 3D-coordinates of the atoms in the given molecule. The electrostatic descriptors reflect characteristics of the charge distribution of the molecule. The quantum-chemical descriptors add important information to the conventional descriptors.

3. METHODOLOGY

3.1. Feature Selection and Regression Analysis. Once descriptors were generated, in this work, the correlation analysis of descriptors was performed first. In the process of correlation analysis, pairwise correlations between descriptors were examined so that only one descriptor was retained from a pair contributing similar information (correlation coefficients greater than 0.85). After the correlation analysis of the descriptors, descriptor-screening methods were used to select the most relevant descriptor to establish the models for prediction of the molecular property. Here, the forward stepwise regression method was used to choose the subset of the molecular descriptors. Forward stepwise regression starts with no model terms, and at each step it adds the most statistically significant term (the one with the highest *F*-statistic or lowest *P*-value) until there are none left.

After the descriptor was selected, multiple linear regression was used to develop the linear model of the property of interest, which takes the form below:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (1)$$

In this equation, Y is the property, that is, the dependent variable, X_1-X_n represents the specific descriptor, while b_1-b_n represents the coefficients of those descriptors, and b_0 is the intercept of the equation. The statistical evaluation of the data was obtained by the software SPSS.

3.2. Radial Basis Function Neural Networks Theory. The theory of RBFNNs has been extensively presented in the paper of Yao et al.[16,17] Here only a brief description of the RBFNNs principle was given. The RBFNNs consist of three layers: the input layer, the hidden layer, and the output layer. The input layer does not process the information; it only distributes the input vectors to the hidden layer. Each neuron on the hidden layer employs a radial basis function (RBF) as a nonlinear transfer function to operate on the input data. In general, there are several radial basis functions: linear, cubic, thin plate spline (TPS), Gaussian, multiquadratic, and inverse multiquadratic. The most often used RBF is the Gaussian function that is characterized by a center (c_j) and width (r_j). In this study, Gaussian was selected as the radial basis function. The operation of the output layer is linear, which is given in eq 2

$$y_k(x) = \sum w_{kj}h_j(x) + bk \qquad (2)$$

where y_k is the kth output unit for the input vector x, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and h_j is the notation for the output of the jth RBF unit.

The training procedure when using RBF involves selecting centers, width, and weights. In this paper, the forward subset selection routine was used to select the centers from training set samples.[18,19] The adjustment of the connection weight between the hidden layer and the output layer was performed using a least-squares solution after the selection of centers and width of radial basis functions.

3.3. Support Vector Machines. 3.3.1. Structural Risk Minimization.[20,21] Previous approaches to statistical learning have tended to be based on finding functions to map vector-encoded data to their respective classes. The conventional minimization of the empirical risk over the training data does not, however, imply a good generalization to the novel test data. Indeed, there could be a number of different functions which all give a good approximation to a training set. It is nevertheless difficult to determine a function which best captures the true underlying structure of the data distribution. Structural risk minimization (SRM) aims to address this

Table 1. Compounds and the Predicted Results of the Heat Capacity (J K^{-1} Mol^{-1})

no.	name	heat capacity	MLR[a]	RBFNNs[b]	SVM[c]	no.	name	heat capacity	MLR[a]	RBFNNs[b]	SVM[c]
1[d]	methane	35.695	33.940	32.859	35.220	76	2-butene, (Z)-	80.150	87.298	85.953	85.523
2	methyl chloride	40.740	43.198	41.321	42.724	77[d]	1-butyne	81.820	80.360	78.747	78.316
3	methylene chloride	50.950	55.602	54.339	53.248	78	1-butene	85.560	87.847	86.564	86.127
4	chloroform	63.521	68.926	65.452	66.950	79	2-butene, (E)-	87.670	87.772	86.481	86.023
5[d]	carbon tetrachloride	82.888	84.111	72.643	83.012	80	1-propene, 2-methyl-	88.090	90.592	88.379	88.775
6	carbon monoxide	29.141	31.318	31.404	31.756	81[d]	isobutane	96.650	100.336	98.730	98.505
7	carbon dioxide	37.135	39.439	40.608	38.574	82	butane	98.490	97.168	96.452	95.579
8	methyl alcohol	44.101	42.440	43.623	44.502	83	butane, 1-chloro-	107.940	108.139	107.770	107.970
9[d]	formic acid	45.801	43.767	45.387	43.541	84	butane, 2-chloro-	110.220	110.281	109.28	109.958
10	methylamine	50.050	47.203	48.229	48.738	85[d]	propane, 2-chloro-2-methyl-	111.950	115.254	112.76	114.230
11	ethylene	42.883	44.149	41.901	43.909						
12	acetylene	44.036	36.795	35.804	45.922	86	furan	65.400	65.182	68.176	67.715
13[d]	ethane	52.487	53.637	50.415	52.594	87	furan, 2,3-dihydro-	74.310	74.802	74.024	74.092
14	ethene, chloro-	53.680	56.431	55.348	53.864	88	cyclobutanone	74.310	76.074	76.812	78.704
15	ethene, 1,2-dichloro-, (Z)-	64.890	66.495	64.238	62.883	89[d]	furan, tetrahydro-	76.634	84.024	80.161	80.556
16	ethyl chloride	65.640	66.057	64.445	63.587	90	2-oxetanone,	84.410	78.190	83.385	81.208
17[d]	ethene, 1,2-dichloro-, (E)-	66.560	66.688	64.388	63.583	91	gamma-butyrolactone	86.100	85.285	86.316	85.491
18	ethene, 1,1-dichloro-	67.123	69.695	67.210	67.128	92	1,3-dioxane	89.400	94.065	90.671	90.739
19	ethane, 1,1-dichloro-	76.320	79.673	76.552	77.964	93[d]	1,4-dioxane	92.100	94.149	90.739	90.585
20	ethane, 1,2-dichloro-	77.320	76.657	74.190	73.953	94	2-butenal	93.920	87.974	88.369	88.014
21[d]	tetrachloroethylene	94.919	90.931	77.904	89.971	95	cyclobutanol	94.470	85.931	86.698	86.309
22	ethylene oxide	47.850	43.559	50.428	48.045	96	1-butanol	108.030	106.866	109.030	108.569
23	ketene	51.750	50.597	51.832	48.980	97[d]	ethanol, 1,1-dimethyl-	113.630	114.111	111.140	113.572
24	acetaldehyde	55.320	55.314	56.450	54.992	98	ethyl acetate	113.640	109.717	108.230	108.709
25[d]	acetic acid	63.440	67.743	65.853	66.453	99	ethoxy ethane	119.460	107.809	107.780	116.866
26	methyl formate	64.380	65.442	64.056	63.044	100	pyrrole	71.600	69.931	71.358	72.315
27	ethanol	65.210	65.117	65.607	65.385	101[d]	(Z)−2-butenenitrile	83.850	91.827	89.899	80.484
28	dimethyl ether	65.570	65.523	63.583	63.406	102	(E)−2-butenenitrile	86.740	92.096	90.140	90.883
29[d]	1,2-ethanediol	77.990	74.685	74.813	74.478	103	1-butanamine	113.900	112.327	114.150	114.418
30	acetonitrile	52.220	53.020	54.409	52.104	104	2-butanamine	120.300	114.191	116.720	115.101
31	dimethylamine	70.500	70.292	70.423	70.891	105[d]	2-propanamine, 2-methyl-	120.920	119.266	118.010	118.930
32	thiirane	53.320	49.635	55.351	53.717	106	thiophene, tetrahydro-	90.860	90.266	91.398	89.515
33[d]	ethanethiol	73.008	71.500	73.099	71.892	107	diethyl sulfide	116.570	113.532	117.930	117.130
34	dimethyl sulfide	74.060	71.599	73.208	71.970	108	1-propanethiol, 2-methyl-	118.830	115.463	120.240	118.423
35	disulfide, dimethyl	94.220	88.860	90.835	90.123	109[d]	2-butanethiol	119.700	115.607	120.440	118.554
36	acetyl chloride	67.860	69.073	67.828	67.391	110	propane, 2-(methylthio)-	120.000	115.870	120.780	118.812
37[d]	urea, methyl-	88.700	86.850	84.301	86.051	111	2-propanethiol, 2-methyl-	121.130	120.483	120.990	122.912
38	cyclopropene	52.900	45.268	51.496	48.828	112	1,3-cyclopentadiene	75.400	76.144	74.891	75.241
39	cyclopropane	55.600	54.311	56.208	55.256	113[d]	cyclopentene	81.280	85.482	81.253	81.853
40	allene	59.030	61.708	60.269	59.180	114	cyclopentane	82.800	95.084	88.354	89.770
41[d]	propyne	60.730	59.142	57.847	56.778	115	cyclobutane, methylene-	87.400	87.497	84.969	84.986
42	propene	64.320	66.974	65.130	64.860	116	1,3-pentadiene, (Z)-	97.180	99.464	98.549	98.431
43	propane	73.600	76.106	74.026	74.118	117[d]	1,4-pentadiene	98.240	99.412	98.500	98.369
44	propane, 1-chloro-	85.300	86.692	85.311	85.062	118	2-pentene, (Z)-	98.800	108.921	109.000	98.212
45[d]	propane, 2-chloro-	87.560	90.125	87.813	88.374	119	1,3-pentadiene, (E)-	99.060	99.532	98.619	98.489
46	oxetane	61.541	63.813	64.398	64.121	120	2,3-pentadiene	99.900	103.085	102.110	102.648
47	cyclopropanone	64.300	55.493	61.009	60.113	121[d]	1,2-pentadiene	101.000	103.657	102.650	103.305
48	1,3-dioxolane	71.000	73.249	72.991	73.129	122	1,3-butadiene, 2-methyl-	102.690	100.684	99.057	99.320
49[d]	beta-propiolactone	71.240	65.136	70.270	68.993	123	2-butene, 2-methyl-	105.020	110.005	109.460	108.905
50	acetone	75.020	79.420	76.476	78.895	124	1,2-butadiene, 3-methyl-	105.250	105.213	103.520	104.586
51	2-propen-1-ol	76.020	76.682	77.235	76.815	125[d]	cyclopropane, 1,1-dimethyl-	106.370	100.324	100.760	97.931
52	propanal	80.730	87.766	86.474	80.032						
53[d]	2-propenoic acid	81.800	78.417	76.178	77.036	126	2-pentene, (E)-	108.900	108.724	108.790	107.932
54	1,3,5-trioxane	81.900	82.852	82.289	82.094	127	1-butene, 2-methyl-	109.960	110.500	110.010	109.471
55	1-propanol	85.560	85.928	86.283	86.715	128	butane, 2-methyl-	118.900	120.424	120.750	119.504
56	acetic acid, methyl ester	86.030	88.312	85.933	86.438	129[d]	pentane	120.070	118.196	118.750	117.528
57[d]	isopropyl alcohol	89.320	89.171	86.619	88.866	130	propane, 2,2-dimethyl-	120.820	124.813	124.260	123.340
58	ethane, methoxy-	93.300	86.342	84.878	94.486	131	pentane, 1-chloro-	130.580	129.439	129.230	130.431
59	acrylonitrile	63.940	64.260	64.989	62.691	132	2H−pyran, 3,4-dihydro-	92.200	95.340	91.644	91.706
60	1-propanamine	91.170	91.022	91.294	92.181	133[d]	cyclopentanone	95.330	96.113	97.630	96.602
61[d]	2-propanamine	97.550	94.306	92.526	94.326	134	2H−pyran, tetrahydro-	99.100	104.817	99.354	98.402
62	thietane	68.620	69.902	71.181	71.474	135	cyclopentanol	105.430	105.812	108.730	103.496
63	1-propanethiol	94.890	92.553	94.841	94.399	136	2-pentanone	125.900	120.740	125.680	122.521
64	ethane, (methylthio)-	95.060	92.811	95.159	94.707	137[d]	3-pentanone	129.870	119.471	124.580	131.726
65[d]	2-propanethiol	96.150	95.510	95.261	96.288	138	1-pentanol	130.700	128.471	131.070	131.758
66	urea, *N*,*N*′-dimethyl-	103.800	107.537	109.300	108.815	139	pyridine	78.230	80.979	81.204	82.931
67	urea, *N*,*N*-dimethyl-	107.200	109.485	108.58	109.879	140	1H−pyrrole, 1-methyl-	90.890	92.813	93.662	94.149
68	urea, ethyl-	115.700	108.422	109.080	118.706	141[d]	(E)−2-pentenenitrile	106.100	106.567	108.950	108.476
69[d]	cyclobutene	64.410	65.308	65.381	64.955	142	(Z)−2-pentenenitrile	106.100	106.640	109.040	109.016
70	cyclobutane	70.600	74.707	70.984	72.007	143	pentanenitrile	116.540	115.683	120.720	118.689
71	methylenecyclopropane	72.930	66.971	68.447	67.613	144	butanenitrile, 2-methyl-	121.800	116.226	120.930	118.503
72	1-methylcyclopropene	74.680	67.208	68.600	67.802	145[d]	propanenitrile, 2,2-dimethyl-	124.220	121.206	122.560	122.466
73[d]	2-butyne	78.020	79.932	78.323	77.867						
74	1,2-butadiene	79.480	82.479	80.804	80.749	146	thiophene, 2-methyl-	95.370	92.974	92.953	95.048
75	1,3-butadiene	79.810	78.289	76.658	76.155	147	thiophene, 3-methyl-	95.790	93.177	93.131	95.043

Table 1 (Continued)

no.	name	heat capacity	MLR[a]	RBFNNs[b]	SVM[c]	no.	name	heat capacity	MLR[a]	RBFNNs[b]	SVM[c]
148	ethyl propyl sulfide	139.200	135.130	138.150	141.078	165[d]	2-hexyne	119.650	122.393	122.440	122.995
149[d]	butane, 1-(methylthio)-	139.790	134.493	137.570	139.282	166	cyclobutane, ethyl-	122.800	117.437	110.460	118.365
150	1-pentanethiol	141.210	134.838	137.880	139.360	167	2-butene, 2,3-dimethyl-	123.600	133.233	133.410	122.452
151	2-butanethiol, 2-methyl-	143.390	139.783	143.920	142.724	168	1-hexyne	125.950	122.697	122.700	123.192
152	propane, 2-methyl-2-(methylthio)-	143.800	140.217	144.380	143.432	169[d]	2-pentene, 3-methyl-, (E)-	126.600	130.988	131.310	130.903
						170	2-pentene, 3-methyl-, (Z)-	126.600	131.245	131.530	131.192
153[d]	benzene	94.100	89.269	86.429	94.947	171	1-hexene	130.830	129.577	129.960	129.259
154	1,4-cyclohexadiene	94.100	96.330	92.379	92.588	172	1-butyne, 3,3-dimethyl-	131.310	127.221	126.300	125.824
155	1,3-cyclohexadiene	94.200	96.716	92.662	92.966	173[d]	pentane, 3-methylene-	133.600	130.256	130.640	130.011
156	cyclopentene, 4-methyl-	100.000	107.683	104.360	102.903	174	butane, 2,3-dimethyl-	139.400	141.700	140.560	140.004
157[d]	cyclopentene, 3-methyl-	100.000	108.044	104.650	102.935	175	pentane, 3-methyl-	140.100	140.288	138.850	138.877
158	bicyclo[3.1.0]hexane	100.500	95.349	100.400	92.113	176	butane, 2,2-dimethyl-	141.500	143.422	142.080	140.166
159	cyclopentene, 1-methyl-	101.000	107.514	104.220	102.701	177[d]	hexane	142.600	139.956	138.230	139.595
160	cyclohexane	105.300	115.230	107.700	107.680	178	1-butene, 2,3-dimethyl-	143.500	132.813	133.020	141.825
161[d]	cyclopentane, methyl-	109.500	116.769	111.760	108.911	179	phenol	103.220	100.008	100.700	99.857
162	1,3,5-hexatriene, (Z)-	110.170	110.826	109.900	110.723	180	toluene	103.700	112.238	110.490	109.928
163	1,3,5-hexatriene, (E)-	110.620	111.226	110.260	111.200	181[d]	styrene	151.290	122.248	119.670	149.550
164	3-hexyne	119.500	122.364	122.410	122.960	182	naphthalene	133.020	132.444	132.710	137.580

[a] Predicted heat capacity by MLR. [b] Predicted heat capacity by RBFNNs. [c] Predicted heat capacity by SVM. [d] Test set.

problem and provides a well-defined quantitative measure for the capacity of a learned function to generalize over unknown test data. Due to its relative simplicity, the Vapnik-Chervonenkis (VC) dimension in particular has been adopted as one of the more popular measures for such a capacity. By choosing a function with a low VC dimension and minimizing its empirical error to a training data set, SRM can offer a guaranteed minimal bound on the test error.

3.3.2. Theory of SVM for Regression.[22] The foundation of Support Vector Machines (SVM) has been developed by Vapnik, and they are gaining popularity due to many attractive features and promising empirical performance.[21,23] The formulation embodies the Structural Risk Minimization (SRM) principle, which has been shown to be superior to the traditional Empirical Risk Minimization (ERM) principle, employed by conventional neural networks. SRM minimizes an upper bound on VC dimension ("generalization error"), as opposed to ERM that minimizes the error on the training data. It is the difference that equips SVM with good generalization performance, which is the goal in statistical learning. Originally, SVM were developed for pattern recognition problems[24] and now, with the introduction of ϵ-insensitive loss function, SVM have been extended to solve nonlinear regression estimation. The estimated function is a linear expansion in terms of functions defined on a certain subset of the data (support vectors), and the final number of coefficients in such an expansion does not depend on the dimensionality of the space of input variables. These two properties make SVM an especially useful technique for dealing with very large data sets in a high-dimensional space.

Compared to other neural network regressors, there are three distinct characteristics when SVM are used to estimate the regression function. First of all, SVM estimate the regression using a set of linear functions that are defined in a high-dimensional space. Second, SVM carry out the regression estimation by risk minimization where the risk is measured using Vapnik's ϵ-insensitive loss function. Third, SVM use a risk function consisting of the empirical error and a regularization term which is derived from the structural risk minimization principle of converging to the global optimum and not to a local optimum.

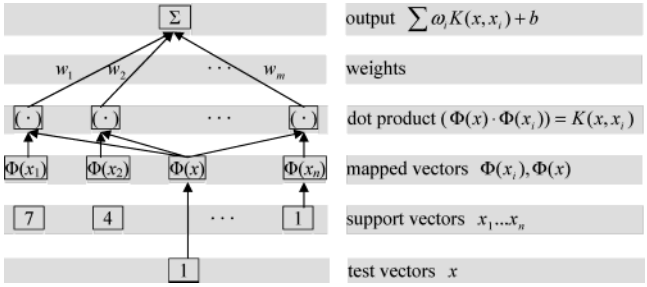

Figure 1. Architecture of a regression machine constructed by the support vector algorithm.[22]

Figure 1 contains a graphical overview over the different steps in the regression stage of SVM. In support vector regression (SVR), the basic idea is to map the data x into a higher-dimensional feature space F via a nonlinear mapping Φ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i,d_i)\}_i^n$ (x_i is the input vector, d_i is the desired value, and n is the total number of data patterns), and SVM approximate the function using the following equation

$$y = f(x) = w\Phi(x) + b \tag{3}$$

where $\Phi(x)$ is the high-dimensional feature space which is nonlinearly mapped from the input space x. The coefficients w and b are estimated by minimizing

$$R_{\mathrm{SVMs}}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\epsilon(d_i,y_i) + \frac{1}{2}||w||^2 \tag{4}$$

$$L_\epsilon(d,y) = \begin{cases} |d-y| - \epsilon & |d-y| \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In the regularized risk function given by eq 4, the first term $C(1/n)\sum_{i=1}^{n} L_\epsilon(d_i,y_i)$ is the empirical error (risk). They are measured by the ϵ-insensitive loss function given by eq 5. This loss function provides the advantage of enabling one to use sparse data points to represent the decision function given by eq 3. The second term $1/2||w||^2$, on the other hand, is the regularization term. C is referred to as the regularized

constant, and it determines the tradeoff between the empirical risk and the regularization term. Increasing the value of C will result in the relative importance of the empirical risk with respect to the regularization term to grow. ϵ is called the tube size, and it is equivalent to the approximation accuracy placed on the training data points. Both C and ϵ are user-prescribed parameters.

To obtain the estimations of w and b, eq 4 is transformed to the primal function given by eq 6 by introducing the positive slack variables ξ_i and ξ_i^* as follows:

$$\text{Minimize } R_{\text{SVMs}}(w,\xi^{(*)}) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{Subject to } \begin{cases} d_i - w\Phi(x_i) - b_i \le \epsilon + \xi_i \\ w\Phi(x_i) + b_i - d_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (6)$$

Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, the decision function given by eq 3 has the following explicit form

$$f(x,a_i,a_i^*) = \sum(a_i - a_i^*)K(x,x_i) + b \quad (7)$$

where the kernel function K corresponds to $K(x,x_i) = \phi(x)^T\phi(x_i)$. One has several possibilities for the choice of this kernel function, including linear, polynomial, splines, and radial basis function. The elegance of using kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly. In the support vector regression, a commonly used kernel function is the Gaussian Radial Basis Function.

The overall performances of RBFNNs and SVM were evaluated in terms of the root-mean-square (rms) error which was defined as below

$$\text{rms} = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \quad (8)$$

where y_k is the desired output and $\hat{y}_k$ is the actual output of the model, and n_s is the number of compounds in the analyzed set.

3.4. RBFNNs and SVM Implementation and Computation Environment. All calculation programs implementing RBFNNs were written in M-file based on the basis MATLAB script for RBFNNs. All calculation programs implementing SVM were written in R-file based on the R script for SVM and compiled using an R1.7.1 compiler.[25] The scripts were run on a Pentium IV PC with 256M RAM.

4. RESULTS AND DISCUSSION

4.1. Results of MLR. About 600 descriptors were calculated by the CODESSA program. After the correlation analysis of the descriptors, the pool of descriptors was reduced to 227. The stepwise regression routine was used to develop the linear model for the prediction of the heat capacity using calculated structural descriptors. The best linear model contains 4 molecular descriptors. The regression coefficients of the descriptors and their physical-chemical meaning were listed in Table 2. The linear correlation coefficient value of each of the two descriptors is <0.85, which means the descriptors were independent in this MLR analysis. The predicted results were given in Table 1. This model gave an rms error of 4.268 heat capacity units for the training set, 5.794 for the test set, and 4.648 for the whole set, and the corresponding correlation coefficients (R) were 0.988, 0.975, and 0.985, respectively. Figure 2 showed these predicted versus experimental heat capacity.

Table 2. Descriptors, Coefficients, Standard Error, and T-Values for the Linear Model[a]

chemical meaning	descriptor	coeff	SE	beta	T-value
intercept	(constant)	0.882	1.911		0.461
molecular volume	MV	0.678	0.064	0.502	10.520
number of rings	NR	−12.697	0.942	−0.236	−13.476
number of atoms	NA	2.762	0.267	0.407	10.348
Randic index (order 2)	RI2	4.262	1.048	0.115	4.068

[a] $R = 0.988$; $R^2 = 0.975$; SE of the estimate = 4.268; rms = 4.189; $n = 136$; F =1295.787.

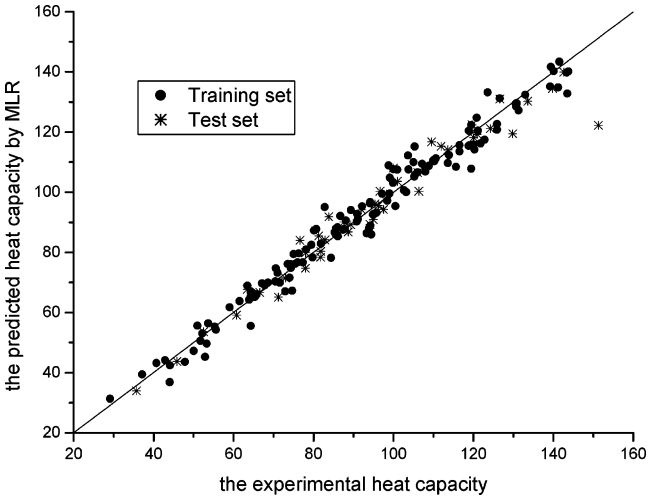

Figure 2. Predicted versus experimental heat capacity (MLR).

4.2. Result of RBFNNs. From Table 2, it can be seen that the model of MLR was not sufficiently accurate (rms = 4.189, SE = 4.268) and showed the factors influencing the heat capacity were complex and not all of them were a linear correlation with the heat capacity. So, we built the nonlinear prediction models by RBFNNs and SVM to further discuss the correlation between the molecular structure and the heat capacity based on the same descriptor set.

After the establishment of a linear model, RBFNNs were used to develop a nonlinear model based on the same subset of descriptors. Each minimum error on the LOO cross-validation was plotted versus the width (Figure 3), and the minimum was chosen as the optimal conditions.

Through the above process, the optimum width and the best number of hidden layer units were selected as 2.0 and 20, respectively. From the best network, the inputs in the test set were presented with it, and the results with RBFNNs were obtained. They were shown in Table 1 and Figure 4. The network gave an rms error of 3.422 for the training set, 6.310 for the prediction set, and 4.337 for the whole set, and the corresponding correlation coefficients (R) were 0.987, 0.992, and 0.973, respectively.

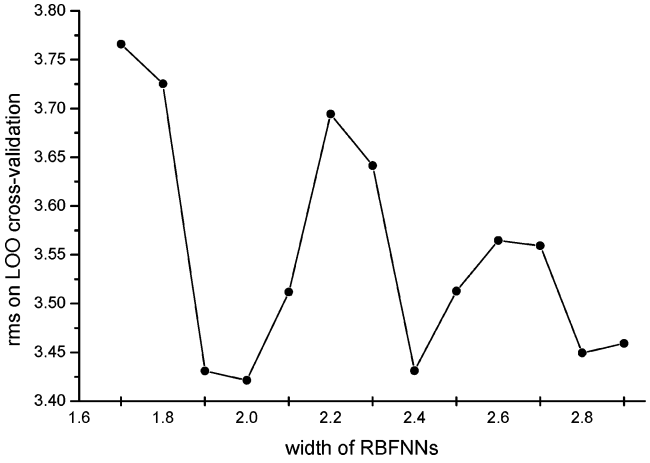

Figure 3. The width of RBFNNs versus rms error on LOO cross-validation.

4.3. Result of SVM. 4.3.1. Selection of the Parameters of the SVM. Analysis of the results obtained by RBFNNs, it can be seen that the model constructed by RBFNNs was not sufficiently accurate and the prediction ability was bad (the rms error for the test set was 6.310), so after the establishment of nonlinear model by RBFNNs, the support vector machines were used to develop an accurate nonlinear model based on the same subset of descriptors.

Similar to other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are capacity parameter C, ϵ of ϵ-insensitive loss function, the kernel type K, and its corresponding parameters. In this work, LOO cross-validation was performed for parameters selection,[26,27] which probably is the current best-performing approach to the SVM design problem.[28] C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. To make the learning process stable, a large value should be set up for C.

The kernel type is another important parameter. For regression tasks, the Gaussian kernel is commonly used. The form of the Gaussian function is as follows

$$\exp(-\gamma^*|u - v|^2)$$

where γ is a constant, the parameter of the kernel, and u and v are two independent variables. γ controls the amplitude of the Gaussian function and, therefore, controls the generalization ability of SVM. Each rms error on the LOO cross-validation was plotted versus γ (Figure 5), and the minimum was chosen as the optimal conditions. In this case: $\gamma = 0.008$.

The optimal value for ϵ depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ϵ, there is the practical consideration of the number of resulting support vectors. ϵ-insensitivity prevents the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value of ϵ is critical from theory. To find an optimal ϵ, the rms on

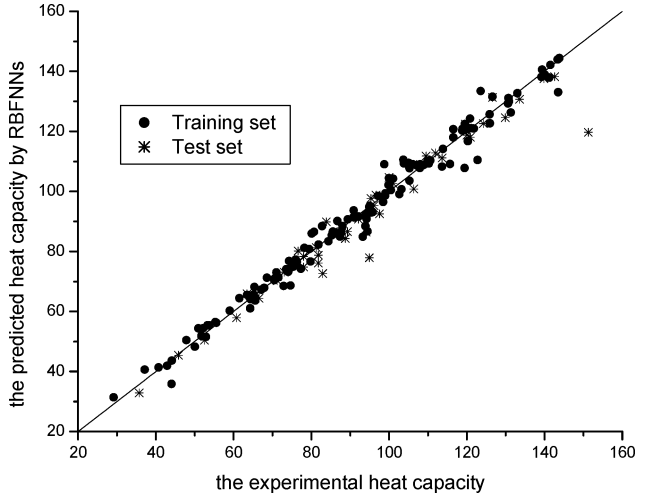

Figure 4. Predicted versus experimental heat capacity (RBFNNs).

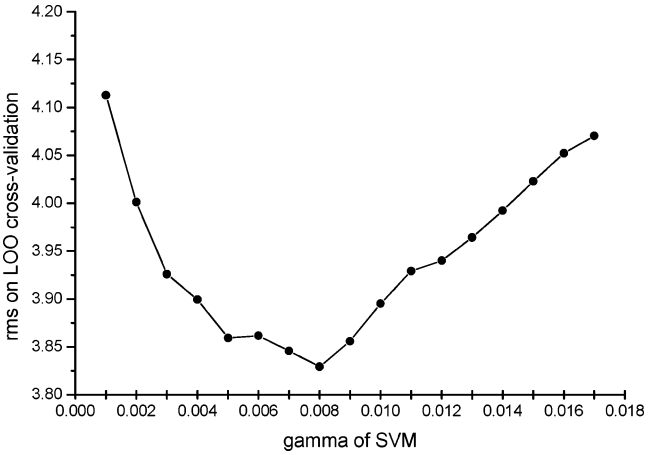

Figure 5. The gamma versus rms error on LOO cross-validation ($C = 100$, $\epsilon = 0.1$).

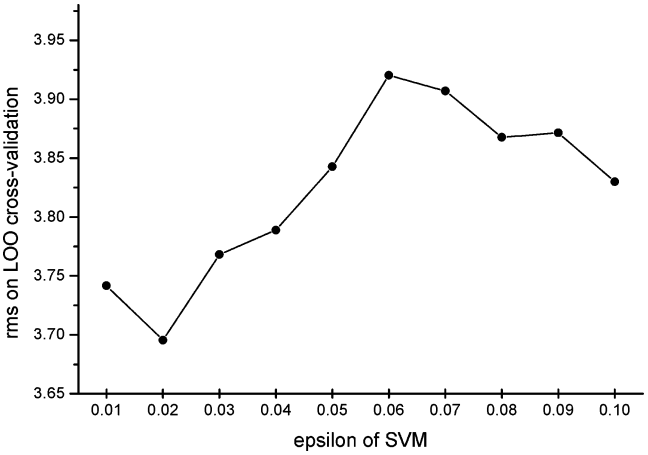

Figure 6. The epsilon versus rms error on validation set ($C = 100$, $\gamma = 0.008$).

LOO cross-validation on different ϵ was calculated. The curve of rms versus the epsilon was shown in Figure 6. The optimal ϵ was found as 0.02.

The last important parameter is the regularization parameter C, of which the effect on the rms was shown in Figure 7. From Figure 7, the optimal C was found as 100.

4.3.2. The Predicted Results of SVM. Through the above process, the γ, ϵ, and C were fixed to 0.008, 0.02, and 100, respectively, when the support vector number of the SVM

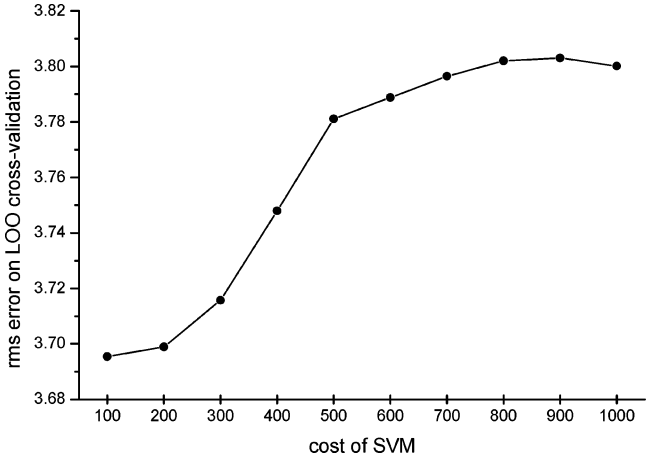

Figure 7. The C versus rms error on validation set ($\gamma = 0.008$, $\epsilon = 0.02$).

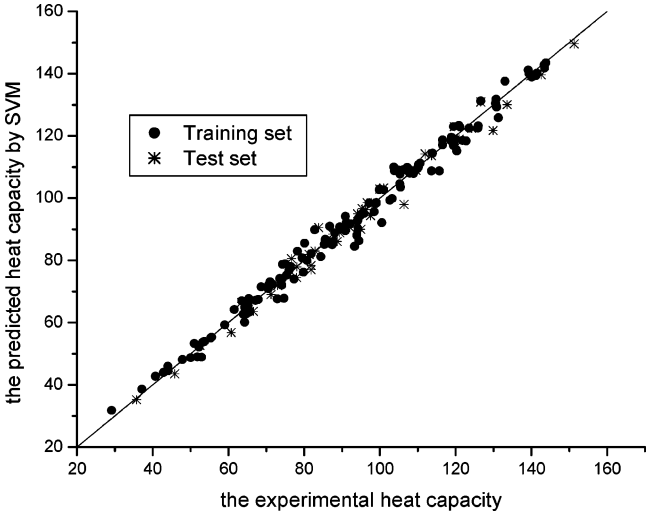

Figure 8. Predicted versus experimental heat capacity(SVM).

model was 19, the predicted results of the optimal SVM were shown in Table 1 and Figure 8. The model gave an rms of 2.880 for the training set, 3.078 for the prediction set, and 2.931 for the whole set, and the corresponding correlation coefficients (R) were 0.994, 0.993 and 0.994, respectively. The performance of SVM is better than MLR and RBFNNs models in Table 1.

4.4. Discussion of the Input Parameters and the Results. By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the heat capacity of the compounds. Of the four descriptors, two are constitutional, one is topological, and one is geometrical descriptor. According to the beta values (Table 2), the more relevant descriptor is a geometrical descriptor: molecular volume (MV). MV is a bulk property, which describes the size of a molecule and relates to the dispersion interaction among molecules; this parameter receives a positive regression coefficient in the regression indicating that the larger the molecular volume is, the higher the heat capacity is. The constitutional descriptors include the number of rings (NR) and the number of atoms (NA). NR receives a negative coefficient in the regression, and this indicates that increasing the number of rings leads to a low heat capacity. So, the heat capacity of noncyclic compound is higher than that of cyclic compound. NA receives a positive coefficient indicating that the heat capacity increases with the increasing of the number of the atoms. The inclusion of topological descriptors: the Randic index (order 2) (RI2), which encodes the size, shape, and degree of branching in the compound and also relates to the dispersion interaction among molecules. It receives a positive coefficient in the regression model indicating the heat capacity increases with increasing the RI2 of the molecule.

Analysis of the results obtained indicated that the models we proposed correctly represent the structural-property relationships of these compounds and that molecular descriptors calculated solely from structures can represent the structural features of the compounds responsible for their heat capacity. Moreover, it seems that the prediction ability of SVM is better than MLR and RBFNNs models. The root cause that SVM can obtain the best results is that SVM adopts the Structural Risk Minimization principle.

5. CONCLUSION

The support vector machine, as a novel type of learning machine, for the first time, was used to develop a QSPR model for the prediction of the heat capacity of a diverse set of 182 compounds based on descriptors calculated from the molecular structure alone. MLR and RBFNNs were also utilized to establish quantitative linear and nonlinear relationships to compare with the results obtained by SVM. Very satisfactory results were obtained with the proposed methods. The models proposed could identify and give some insight into factors that are likely to govern the heat capacity of the compounds. Additionally, nonlinear models using SVM based on the same set of descriptors produced even better models with a good predictive ability than the two other MLR and RBFNNs models. This study of the QSPR model shows that the SVM is a very promising tool in the prediction of heat capacity and exhibits a high speed of leaning when compared with RBFNNs. The training procedure is also simple when using SVM because there are fewer parameters having to be optimized, and only support vectors are used in the generalization process. Besides, the SVM exhibits the better whole performance due to embodying the Structural Risk Minimization principle and some advantages over the other techniques. Furthermore, the proposed approach can also be extended to other QSPR or QSAR investigations.

ACKNOWLEDGMENT

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Program PRA SI 02-03). The authors also thank the R Development Core Team for affording the free R1.7.1 software.

REFERENCES AND NOTES

(1) Sears, F. W.; Salinger, G. *Thermodynamics, Kinetic Theory, and Statistical Thermodynamics*, 3rd ed.; Addison-Wesley: Reading, MA, 1975.

(2) Manallack, D. T.; Livingstone, D. J. Neural networks in drug discovery: have they lived up to their promise? *Eur. J. Med. Chem.* **1999**, *34*, 95−208.

(3) Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(4) Pang, S. N.; Kim, D.; Bang, S. Y. Membership authentication in the dynamic group by face classification using SVM ensemble. *Pattern Recognit. Lett.* **2003**, *24*, 215−225.

(5) Liu, H. X.; Zhang, R. S.; Luan, F.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Diagnosing breast cancer based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 900−907.
(6) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nNondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.
(7) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(8) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR study of ethyl 2-[(3-Methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: an inhibitor of AP-1 and NF−KB mediated gene expression based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288−1296.
(9) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* **2004**, *43*, 161−167.
(10) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. An accurate QSPR study of O−H bond dissociation energy in substituted phenols based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 669−677.
(11) Xue, C. X.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Study of the quantitative structure-mobility relationship of carboxylic acids in capillary electrophoresis based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2004**, in press.
(12) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Version 2.0 Reference Manual, 1995−1997.
(13) http://srdata.nist.gov/cccbdb/.
(14) HyperChem, Release 4.0 for Windows, Hypercube, Inc., 1995.
(15) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange;* QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
(16) Yao, X. J.; Liu, M. C.; Zhang, X. Y.; Hu, Z. D.; Fan, B. T. Radial basis function network-based quantitative structure−property relationship for the prediction of Henry's law constant. *Anal. Chim. Acta* **2002**, *462*, 101−117.
(17) Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217−225.
(18) Orr, M. J. L. *Introduction to Radial basis function networks, center for cognitive science*; Edinburgh University: 1996.
(19) Orr, M. J. L. *MATLAB routines for subset selection and ridge regression in linear neural networks, Center for cognitive science*; Edinburgh University: 1996.
(20) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **1998**, *2*(2), 1−47.
(21) Vapnik, V. *Estimation of Dependences Based on Empirical Data*; Springer: Berlin, 1982.
(22) Smola, A. J.; Schölkopf, B. *A tutorial on support vector regression*; NeuroCOL2 Technical report series, NC2-TR-1998-030; October, 1998.
(23) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(24) Burges, C. J. C. A tutorial of support vector machines for pattern recognition. http://svm.research.bell-labs.com/SVMdoc.html, 1998.
(25) Venables, W. N. D.; Smith, M.; the R Development Core Team. R manuals 2003.
(26) Schölkopf, B.; Burges, J.; Smola, A. *Advances in kernel methods: Support vector machine*; MIT Press: Cambridge, MA, 1999.
(27) Cherkassky, V.; Mulier, F. *Learning from data: Concepts, theory, and methods*; Wiley: New York, 1998.
(28) Anguita, D.; Ridella, S.; Rivieccio, F.; Zunino, R. Hyperparameter design criteria for support vector classifiers. *Neurocomputing* **2003**, *55*, 109−134.

CI049934N