Journal of Hazardous Materials 166 (2009) 853–859

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine

Siavash Riahi[a,b,*], Eslam Pourbasheer[b], Mohammad Reza Ganjali[b], Parviz Norouzi[b]

[a] Institute of Petroleum Engineering, Faculty of Engineering, University of Tehran, Tehran, Iran
[b] Center of Excellence in Electrochemistry, Faculty of Chemistry, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 15 May 2008
Received in revised form 2 September 2008
Accepted 26 November 2008
Available online 3 December 2008

Keywords:
Chemometrics
QSRR
Genetic algorithms
Support vector machine
Essential oils

ABSTRACT

The quantitative structure-retention relationship (QSRR) of the essential oil components against the gas chromatography retention index (RI) was studied. The genetic algorithm (GA) was employed to select the variables that resulted in the best-fitted models. After the variables were selected, the linear multivariate regressions [e.g. the multiple linear regression (MLR), the partial least squares (PLS)] as well as the nonlinear regressions [e.g. the polynomial PLS (poly-PLS), the support vector machine (SVM)] were utilized to construct the linear and nonlinear QSRR models. The obtained results using SVM were compared with those of MLR, PLS and poly-PLS, exhibiting that the SVM model demonstrated a better performance than that of the other models. The relative standard error SE (%) of the training set and the test set for the SVM model was 1.96 and 4.25, and the square correlation coefficients were 0.987 and 0.962 respectively, while the square correlation coefficient of the cross validation (Q^2) on the SVM model was 0.963, revealing the reliability of this model. The resulting data indicated that SVM could be used as a powerful modeling tool for the QSRR studies. This is the first research on the QSRR of the essential oil compounds against the retention index using the SVM.

1. Introduction

Essential oils are sometimes used to flavor compounds in food and have antimicrobial activities. Also, they are toxic to humans including: carcinogenicity, reproductive and developmental toxicity, neurotoxicity as well as acute toxicity. This applies whether taken internally, applied to the skin or simply inhaled. As with most medicinal drugs, either of a "synthetic" or a "natural" origin, the compounds present in essential oils have the potential to create serious, and even fatal toxic effects if ingested in overly large quantities, or used incorrectly [1,2].

In addition, essential oils play a determining role as natural flavoring compounds in both the perfumery and the pharmaceutical industries. In Japan, *Citrus sudachi* is a famous sour citrus fruit on account of its unique pleasant citrus odor and its savory taste [3–5]. It is cultivated in Tokushima prefecture on Shikoku Island. Commonly, *Citrus sudachi* is used as seasoning in foods or as flavoring in alcoholic beverages. The essential oil components of this natural product include: alcohols, organic acids, aldehydes, ketones, esters, aromatic compounds and terpenes—some of which have shown antibacterial activity [6]. All these compounds have been identified by gas chromatography-mass spectrometry (GC-MS). Nevertheless, the mass spectra do not always present enough evidence for the structure elucidation and a prediction model should be used to verify the molecular structure. These methodologies, called quantitative structure retention relationships (QSRR), permit the generation of useful equations for the prediction of retention indices for molecules that are similar but different from those used to develop the model [7]. Seeking the quantitative relationship between the molecular structure and the gas chromatographic retention indices has been a basic task in chemistry. Correlations between the GC retention indices and the molecular structures can provide more profound insights into the interactions between the eluents and the stationary phases from a theoretical viewpoint.

Recently several QSPR/QSRR studies on the retention relationship of essential oil components have been reported. Kovats gas chromatographic retention indices for both apolar (DB-1) and polar (DB-Wax) columns for 48 compounds from Ylang-Ylang essential oil by Olivero et al. [8], capillary column gas chromatographic retention time for natural sterols (trimethylsilyl ethers) from olive oil by Acuña-Cueva et al. [9], Kovats retention indices of terpenes by Hemmateenejad et al. [10], retention indices of pyrazines by Stanton et al. [7].

This work relates the quantitative structure investigation on the essential oils, extracted from the *Citrus sudachi* fruit and their RI relationship. In the QSAR/QSRR studies, there are some techniques

* Corresponding author. Tel.: +98 21 61112788; fax: +98 21 66405141.
E-mail address: riahisv@khayam.ut.ac.ir (S. Riahi).

0304-3894/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2008.11.097

which can be applied for the model construction, such as the multiple linear regression (MLR), the partial least squares (PLS). Also, the nonlinear regressions can be applied like the polynomial PLS (poly-PLS), used for the inspection of the linear and nonlinear relation between the interested property and the molecular descriptors, respectively. MLR yields models that are simpler and easier to interpret than PLS, because these methods perform regression on the latent variables that do not have any physical meaning. However, due to the collinearity between the structural descriptors, MLR is not able to extract useful information from the structural data. As a consequence, an overfitting problem is encountered. PLS is a factor analytical technique which uses factors, or latent variables to create a target matrix used for calibration. PLS is suitable if there are fewer factors in the target matrix than the number of the factors which are originally present in the data matrix. In PLS, the combination step and the regression stage are amalgamated with the decomposition step and the production of the latent variables, so that the eigenvectors of the data matrix are extracted in a sequence congruent with the eigenvectors of the target matrix [11]. Normal or linear PLS uses a linear function to regress the scores of the descriptors matrix on the scores of the retention indexes matrix to find the inner relation. The polynomial PLS employs a nonlinear function, in this case using a squared function to find this inner relation. A good explanation of PLS and nonlinear PLS is given in the paper by Wold et al. [12]. This paper describes the PLS process with the similarities and differences between the linear and nonlinear methods.

The support vector machine (SVM) is a new algorithm developed by the machine learning community [13,14]. The SVM approach automatically controls the flexibility of the resulting classifier on the training data. Accordingly, by the design of the algorithm, the deteriorating effect of the input dimensionality on the generalization ability is largely suppressed. Due to its remarkable generalization performance, SVM has attracted attention and gained extensive application, such as; pattern recognition problems [15,16], drug design [17], QSAR [18–21] and quantitative QSPR analysis [22,23]. In most of these cases, the performance of the SVM modeling either matches or is significantly better than that of the traditional machine learning approaches.

The main aim of the present work was to establish a new QSRR model for predicting the retention index property of the organic compounds, derived from the essential oil of *Citrus sudachi* using the SVM techniques. The performance of this model was compared with those obtained by the MLR, PLS and poly-PLS methods. This is the first research on QSRR of the essential oil compounds against the retention index, using SVM.

2. Materials and methods

2.1. Data set

The data set used in this study was taken from the work of Mookdasanit et al. [24] and is presented in Table 1. This set contains the retention index property of *Citrus sudachi* essential oil compounds, which were measured at the same conditions with the HP5 column (30 m×0.32 mm i.d.; Hewlett Packard, CA). The retention index of the compounds fell in the range of 800 for Hexanal and 1752 for α-Sinensal, at the mean value of 1249.

2.2. Equipment

A Pentium IV personal computer (CPU at 3.06 GHz) with a Windows XP operating system was used. The geometry optimization was performed with HyperChem (Version 7.0 Hypercube, Inc). The Dragon 2.1 software was utilized [25] to calculate the molecular descriptors. The SPSS software (version 11.50, SPSS, Inc.) was

Table 1
The data set and the corresponding observed and predicted RI values by SVM for the training and test set.

Number	Name	RI (Exp)	RI (SVM)	E (%)[a]
Training set				
1	Terpinolene	1086	1029	−5.25
2	*cis*-pinocamphone	1173	1112	−5.2
3	Limonen-8,9-oxide	1199	1166	−2.75
4	*trans*-dihydro carvone	1204	1199	−0.42
5	Camphene	951	949	−0.21
6	Pinene	980	978	−0.2
7	Terpinene	1017	1015	−0.2
8	p-cymene	1024	1022	−0.2
9	(E)-β-ocimene	1048	1046	−0.19
10	*cis*-sabinene hydrate	1068	1066	−0.19
11	*trans*-sabinene hydrate	1097	1095	−0.18
12	Thujone	1114	1112	−0.18
13	Terpinen-4-ol	1177	1175	−0.17
14	Terpineol	1189	1187	−0.17
15	Citronellol	1229	1227	−0.16
16	*cis*-carveol	1230	1228	−0.16
17	Citronellal	1153	1155	0.17
18	Dill ether	1186	1188	0.17
19	*cis*-dihydro carvone	1197	1199	0.17
20	2,6-dimethyl-5-heptenal	1094	1096	0.18
21	Perillene	1101	1103	0.18
22	*cis*-limonen-1,2-oxide	1134	1136	0.18
23	*trans*-p-mentha-2-en-1-ol	1141	1143	0.18
24	Limonene	1030	1032	0.19
25	1,8-cineole	1032	1034	0.19
26	6-methyl-5-hepten-2-one	990	992	0.2
27	Diisopropyl disulfide	1018	1020	0.2
28	Heptanol	968	970	0.21
29	Ethyl acetate	810	812	0.25
30	α-thujene	930	933	0.32
31	Linalool	1100	1107	0.64
32	*trans*-carveol	1218	1226	0.66
33	Fenchone	1087	1096	0.83
34	(Z)-β-ocimene	1042	1052	0.96
35	p-cymen-8-ol	1185	1209	2.03
36	Phellandrene	1003	1024	2.09
37	Campholenal	1126	1160	3.02
38	Myrcene	991	1046	5.55
39	*cis*-linalool oxide	1074	1138	5.96
40	Hexanal	800	857	7.13
41	*trans*-gama-Bisabolen	1533	1433	−6.52
42	Carvacrol	1296	1214	−6.33
43	Piperitone	1253	1183	−5.59
44	Humulene	1454	1425	−1.99
45	Thymol	1293	1268	−1.93
46	Geranyl acetone	1454	1428	−1.79
47	δ-cadinene	1517	1490	−1.78
48	(E,E)-α-farnesene	1508	1495	−0.86
49	Perilla alcohol	1296	1285	−0.85
50	β-Cedrene	1418	1407	−0.78
51	β-Caryophyllene	1413	1407	−0.42
52	*trans*-Carvone oxide	1280	1278	−0.16
53	α-terpinen-7-al	1283	1281	−0.16
54	Undecanal	1308	1306	−0.15
55	*cis*-carvyl acetate	1362	1360	−0.15
56	β-Pathchoulene	1381	1379	−0.14
57	β-Chamigrene	1474	1472	−0.14
58	Cubebol	1512	1510	−0.13
59	Elemol	1549	1547	−0.13
60	Spathulenol	1575	1573	−0.13
61	Caryophyllene oxide	1578	1576	−0.13
62	Cedrol	1596	1594	−0.13
63	Humulene oxide	1601	1599	−0.12
64	α-muurolol	1646	1644	−0.12
65	Patchouli alcohol	1659	1657	−0.12
66	α-sinensal	1752	1750	−0.11
67	Geranial	1270	1269	−0.08
68	Citronellyl acetate	1354	1354	0
69	β-eudesmol	1649	1649	0
70	*trans*-gama-cayophyl	1403	1405	0.14
71	γ-elemene	1433	1435	0.14
72	(E)-β-farnesene	1457	1459	0.14
73	Drima-7,9(11)-diene	1469	1471	0.14

Table 1 (*Continued*)

Number	Name	RI (Exp)	RI (SVM)	E (%)[a]
74	Neryl acetate	1366	1368	0.15
75	α-ylangene	1372	1374	0.15
76	Cumin aldehyde	1238	1240	0.16
77	Perillaaldehyde	1272	1274	0.16
78	α-copaene	1370	1379	0.66
79	Epicubebol	1491	1510	1.27
80	β-elemene	1387	1435	3.46
Test set				
1	β-sesquiphellandrene	1520	1384	−8.95
2	(E)-Nerolidol	1564	1491	−4.67
3	Dodecanal	1407	1343	−4.55
4	Terpinene	1060	1022	−3.58
5	p-mentha-1-en-9-ol	1289	1244	−3.49
6	α-cedr-8(15)-en-9-ol	1643	1604	−2.37
7	δ-elemene	1337	1319	−1.35
8	Sabinene	974	965	−0.92
9	Carvone	1241	1240	−0.08
10	Germacrene D	1479	1478	−0.07
11	*trans*-limonen-1,2-oxide	1136	1136	0
12	Myrtenal	1191	1191	0
13	Phellandrene	1030	1043	1.26
14	*cis*-p-mentha-2-en-1-ol	1121	1138	1.52
15	Decanal	1206	1238	2.65
16	Pinene	934	961	2.89
17	*trans*-linalool oxide	1088	1138	4.6
18	Nonanal	1105	1172	6.06
19	Octanal	1004	1078	7.37
20	(E)-2-hexenal	852	922	8.22

[a] Relative error.

employed for the simple MLR analysis. The PLS and GAs evaluations were carried out using the PLS-Toolbox Version 2.0 for use with Matlab from the Eigenvector Research Inc. The SVM toolbox was developed by Gunn [26,27].

2.3. Descriptors calculation and selection

The first step to obtain a QSRR model was to encode the structural features of the molecules, which were named molecular descriptors. The molecular descriptors, used to search the best model for the retention indexes of these compounds, were calculated with the Dragon program on the basis of the minimum energy molecular geometries. These geometries were optimized with the aid of the HyperChem package, based on the AM1 semiempirical method. The calculated descriptors were first analyzed for the existence of constant or near-constant variables. The detected ones were then removed. In addition, to decrease the redundancy existing in the descriptor data matrix, the descriptors correlation with each other and with the retention index of the molecules was examined. Afterwards, the collinear descriptors (i.e. $r > 0.9$) were detected. Among the collinear descriptors, the one presenting the highest correlation with the RI property was retained. The other descriptors were removed from the data matrix. Then, the remaining descriptors were collected in an $n \times m$ data matrix (D), where $n = 100$ and $m = 325$ are the numbers of the compounds and the descriptors, respectively.

MLR and PLS were utilized as linear techniques, whereas poly-PLS and SVM were employed as nonlinear feature mapping techniques for the construction of the QSRR models in this work. Since the PLS, poly-PLS and SVM methods cannot select the most significant descriptors from the pool of the calculated molecular descriptors, it would be necessary to use a variable selection method. In the present work, the genetic algorithm (GA) variable subset selection method [28,29] was used for the selection of the most relevant descriptors from the pool of the remaining 325 descriptors. These descriptors would be used as inputs of the MLR, PLS, poly-PLS and SVM.

2.4. Genetic algorithm

Nowadays, GA is wellknown as an interesting and most widely used variable selection method. GA is a stochastic method to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation.

To select the most relevant descriptors, the evolution of the population was simulated [30–32]. The population of the first generation was selected randomly. Each individual member in the population, defined by a chromosome of binary values, represented a subset of descriptors. The number of the genes at each chromosome was equal to the number of the descriptors. A gene was given the value of 1, if its corresponding descriptor was included in the subset; otherwise, it was given the value of zero. The number of the genes with the value of 1 was kept relatively low to have a small subset of descriptors [33]. As a result, the probability of generating 0 for a gene was set greater (at least 60 %) than the value of 1. The operators used here were crossover and mutation. The application probability of these operators was varied linearly with a generation renewal (0–0.1 % for mutation and 60–90 % for crossover). The population size was varied between 50 and 250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90 % of the generations took the same fitness.

2.5. Support vector machine (SVM)

SVM, developed by Vapnik and Cortes [34] as a novel type of machine learning method, is gaining popularity due to its many attractive features, and promising empirical performance. SVM demonstrates a great advantage. It can adopt the structure risk minimization (SRM) principle, being superior to the traditional empirical risk minimization (ERM) principle. The conventional neural networks utilize the ERM principle. On the one hand, SRM minimizes an upper bound of the generalization error on the Vapnik–Chernoverkis (VC) dimension. On the other hand, ERM minimizes the training error. With reference to the regression approximation, we supposed that there is a given set of data points $G = \{(x_i, d_i)\}_i^n$ (x_i is the input vector, d_i is the desired value and n is the total number of the data patterns), drawn independently and identically from an unknown function. With three distinct characteristics, SVMs can approximate the function. Firstly, the regression is assessed in a set of linear functions. Secondly, the regression assay is defined as the risk minimization problem, regarding the ε-insensitive loss function. Thirdly, the risk based on the SRM principle is minimized, where the structure elements are defined by the inequality constant of $(1/2)||\omega||^2 \leq$. With the form of the function (1) below, the linear function is formulated at the high dimensional feature space.

$$y = f(x) = w\phi(x) + b \tag{1}$$

In this function, $\phi(x)$ is the high dimensional feature space, being nonlinearly mapped from the input space x. The first and second above-mentioned characteristics are reflected in the minimization of the regularized risk function (2) of SVMs. With the help of function (2), the estimation of the w and b coefficients is performed. The use of this risk function involves two targets; (i) to find a function, displaying the highest ε deviation from the actual values in all the training data points and (ii) to find a function, which is simultaneously as flat as possible.

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

Table 2
The statistical parameters of different constructed QSRR models.

	Training set			Test set			Cross-validation
	R^2	RMSE	F	R^2	RMSE	F	Q^2
MLR	0.949	48.26	226.53	0.931	60.42	21.14	0.936
PLS	0.944	50.28	207.73	0.913	66.95	16.96	0.922
Poly-PLS	0.953	45.99	248.33	0.932	56.92	26.96	0.941
SVM	0.987	24.70	872.08	0.962	51.43	27.52	0.963

$$L_\varepsilon(d,y) = \begin{cases} |d-y| - \varepsilon, & |d-y| \geq \varepsilon, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

As far as the first term, $C(1/n)\sum_{i=1}^{n} L_\varepsilon(d_i, y_i)$, of function (2) is concerned, it is called empirical error (risk) and it is calculated by the ε-insensitive loss function (3). The function (3) is capable of using the sparse data points to represent the designed function (1). Additionally, the second term of the function (2), $(1/2)\|\omega\|^2$, is named regularized term. Finally, ε is called the SVMs tube size and C is the regularization constant, determining the trade-off between the empirical error and the regularized term. The introduction of the ξ and ξ^* positive slack variables results in equation (4), to the following constrained function:

$$\text{Minimize } R_{\text{SVMs}}(\omega, \xi^*) = \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (4)$$

In equation (4), i stands for the data sequence, with $i = 1$ being the most recent observation and $i = 1$ being the earliest observation. Decision function (5) takes the form below, after introducing the Lagrange multipliers and exploiting the optimality constraints:

$$f(x, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)K(x, x_i) + b \quad (5)$$

In equation (5), a_i and a_i^* are the introduced Lagrange multipliers. With the utilization of the Karush–Kuhn–Tucker (KKT) conditions, only a limited number of coefficients will not be zero among a_i and a_i^*. The related data points could be referred to the support vectors. For equation (5), K refers to the kernel function, including the linear, polynomial, splines and radial basis function.

With respect to the support vector regression, the function which is broadly employed is the Gaussian radial basis function (6):

Radial Basis Function (RBF):

$$k(\bar{x}_i, \bar{x}_j) = \exp\left(-\gamma\left\|\bar{x}_i - \bar{x}_j\right\|^2\right) \quad (6)$$

3. Results and discussion

For the selection of the most important descriptors, GA was run many times with different initial sets of population. At the end, a population of good models was obtained. Among these models, one model presented the highest statistical quality and it was used repeatedly in comparison with the other models.

The descriptors, selected by this method, were used to construct some linear and nonlinear models with the employment of the MLR, PLS, poly-PLS and SVM techniques.

3.1. MLR analysis

The statistical parameters of the GA–MLR model, constructed by the selected descriptors, are depicted in Table 2. The methods for the calculations of these descriptors and their meaning have been explained in the Handbook of Molecular Descriptors by Todeschini et al. [35].

Table 3
Details of the constructed GA–MLR model.

Descriptor description	Symbols	Coefficient	MF[a]
	Constant	−650.89 (±90.09)	–
Number of 10-membered rings	nR10	106.23 (±19.63)	19.92
Mean information content vertex degree magnitude	IVDM	490.93 (±30.14)	1828.54
Radial Distribution Function - 2.0/weighted by atomic Sanderson electronegativities	RDF020e	35.97 (±5.47)	133.24
H autocorrelation of lag 8/weighted by atomic masses	H8m	−7025.61 (±726.74)	−22.66
Unsaturation index	Ui	−72.32 (±14.80)	−98.3
Fragment-based polar surface area	PSA	3.89 (±0.59)	49.18

$N = 80$, $R^2 = 0.949$, RMSE = 48.26, $F = 226.53$.
[a] MF refer to the mean effect value.

The six descriptors, which were selected by GA, are; the number of the 10-membered rings (nR10), the mean information content vertex degree magnitude (IVDM), the Radial Distribution Function - 2.0/weighted by atomic Sanderson electronegativities (RDF020e), the H autocorrelation of lag 8/weighted by atomic masses (H8m), the Unsaturation index (Ui) and the Fragment-based polar surface area (PSA).

The obtained correlation matrix between these descriptors showed the capability of the QSRR regression models to predict the retention index accurately, which is not associated with the collinearity between the variables.

To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed with the equation below, displayed in the last column of Table 3.

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}} \quad (7)$$

MF_j represents the mean effect for the considered descriptor j, β_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptors for each molecule and, eventually, m is the descriptor number in the model. The MF value indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign exhibits the variation direction in the values of the activities as a result of the increase (or reduction) of these descriptor values.

One of the constitutional descriptors, appearing in the model, is nR10. As it can be apparent from Table 3, the nR10 mean effect has a positive sign. This sign suggests that the retention index is directly related to this descriptor. Subsequently, the increase of the 10-membered ring number of the molecules results in its retention index increasing.

IVDM is the second descriptor, appearing in the model. It is one of the topological descriptors (2D). This index was proposed as a measure of the molecular complexity together with some other information indices derived from the distance matrix. It is clear from Table 2 that this descriptor has a positive sign, illustrating a greater mean effect value than that of the other descriptors. Therefore, this descriptor had a significant effect on the retention mechanism of the essential oil molecules.

The third descriptor is RDF020e, which is one of the radial distribution function (RDF) descriptors. RDF in this form meets all the requirements for the 3D structure descriptors. It is independent of the atom number (i.e. the size of a molecule), it is unique regarding the three-dimensional arrangement of the atoms and it is invariant against the translation and rotation of the entire molecule. Additionally, the RDF descriptors can be restricted to specific atom types

or distance ranges to represent specific information in a certain three-dimensional structure space (e.g. to describe the steric hindrance or the structure/activity properties of a molecule). RDF020e displays a positive sign, which indicates that the retention index is directly related to this descriptor.

The fourth descriptor is H8m, which was weighted by atomic mass. The H8m mean effect has a negative sign (Table 3), which reveals that the retention index is inversely related to this descriptor. Hence, it was concluded that by increasing the molecular mass the value of this descriptor increased, causing a reduction in its retention index.

The fifth descriptor is Ui, which was the unsaturation index related to the unsaturated bonds. In line with Table 3, the Ui mean effect displays a negative sign, showing that the retention index is inversely related to this descriptor. Accordingly, it was concluded that by increasing the unsaturated bonds in the molecules the value of this descriptor increased, leading to a decrease in its retention index.

The final descriptor is PSA, which was the polar surface area (includes all the O, N, S atoms and the covalently bonded Hs). The PSA mean effect demonstrates a positive sign, revealing that the retention index is directly related to this descriptor. Thus, when the polar surface area of the molecules is increased, the retention index also increases. For this model (MLR), the values of R^2 and RMSE for the training and test set are 0.949 and 48.26, as well as 0.931 and 60.42, respectively. The Q^2 value of the leave-one-out cross-validation is 0.936.

3.2. PLS and poly-PLS analysis

The resulting MLR equations could describe the structure retention relationships well. However, due to the collinearity problem in the MLR analysis, the collinear descriptors were removed before the MLR model development. Therefore, some information was discarded in the MLR analysis. On the contrary, the factor analysis-based methods (e.g. the PLS regression) could handle the collinear descriptors, and consequently better predictive models were attained by PLS. Hence, to model the structure-retention index relationships in a better way, PLS and poly-PLS were also employed as a linear and nonlinear methods in this study [36,37]. In the case of PLS and poly-PLS, before the statistical analysis, the descriptors were scaled to zero mean and unit variance (auto-scaling procedure). The number of the significant factors for the PLS and poly-PLS algorithms were determined using the cross-validation method [38,39]. The best PLS and poly-PLS models contained six selected descriptors in a three latent variable space. In general, the number of components (latent variables) was smaller than the number of independent variables in the PLS analysis. The statistical parameters obtained by these models for the training and test sets are summarized in Table 2. It can be observed that the modeling with poly-PLS is slightly better than that with MLR and PLS.

3.3. SVM analysis

After the establishment of the MLR, PLS and poly-PLS model, SVM was used to develop a model by the training set compounds, based on the same subset of descriptors. The LOO (Leave One Out) cross-validation method implied in SVM was used to build a model. The SVM performance for regression depends on the combination of several factors, such as the kernel function type, the capacity parameter C, ε of the ε-insensitive loss function and its corresponding parameters.

Initially, the selection of the kernel function should take place as it determines the sample distribution in the mapping space. RBF is broadly used in many studies, owing to its good general performance and the small number of parameters to be adjusted

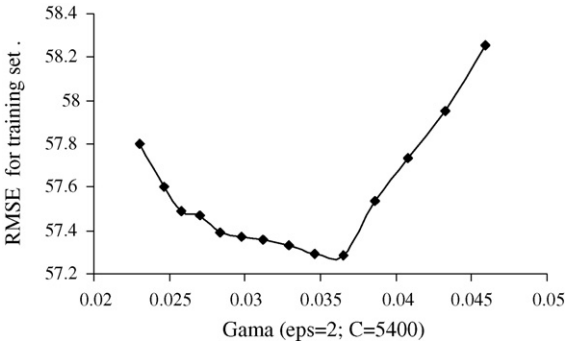

Fig. 1. The gamma(γ) vs. RMSE for the training set ($\varepsilon = 2$; $C = 5400$).

[40]. In this study, RBF was employed with R having the form of: $\exp(-\gamma \times |u - v|^2)$, γ is a kernel parameter, while u and v are two independent variables.

Moreover, the corresponding parameters (e.g. γ of the kernel function) strongly influence the number of the support vectors, having a close relation with the SVM performance and the training time. The extremely high number of support vectors could lead to overfitting and increase the training time. With respect to γ, it controls the amplitude of the RBF function and accordingly, it controls the SVM generalization ability. Fig. 1 depicts the plot of γ versus RMSE on the LOO cross-validation. It is clear that the optimal γ value was 0.037.

Regarding the ε-insensitive parameter, it can prevent the entire training set meeting boundary conditions. In this way, the sparsity possibility in the dual formulation solution is provided. The optimum ε value is significantly affected by the noise type present in the data, which is usually unknown. Fig. 2 illustrates the RMS error of the LOO cross-validation on different epsilon. In agreement with this figure, the most favorable value was equal to 2.

Finally, the influence of the capacity parameter C was investigated. The capacity parameter C controls the trade-off between the margin maximization and the training error minimization. When the C value is too low, then insufficient stress will be placed on fitting the training data. When the C value is too high, then the algorithm will over-fit the training data. Nevertheless, according to Ref. [41], the prediction error was not frequently affected by the C parameter. The learning process can be stabilized, if initially a high C value is selected. Fig. 3 presents the plot of RMSE versus the C value with the values of $\gamma = 0.037$ and $\varepsilon = 2$. The optimum C value was equivalent to 5400.

As a consequence, the best choices regarding the γ, ε and C values were 0.037, 2 and 5400. For the optimal model, the cross-validated coefficient Q^2 value was 0.963. It provided an RMSE value of 24.7 binding affinity units for the training set, 51.4 for the test set, whereas the corresponding correlation coefficients (R^2) were 0.987 and 0.962, respectively (Table 2).

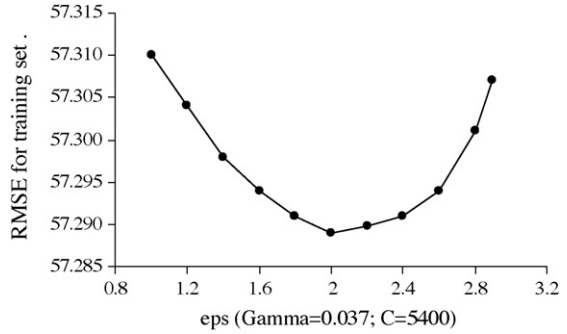

Fig. 2. The epsilon vs. RMSE for the training set (Gamma(γ) = 0.037; $C = 5400$).

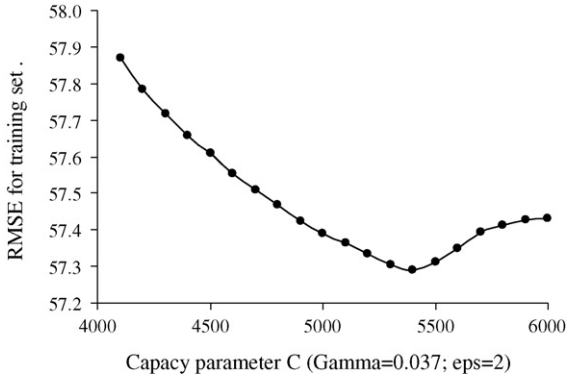

Fig. 3. The capacity parameter *C* vs. RMSE for the training set (Gamma(γ) = 0.037; ε = 2).

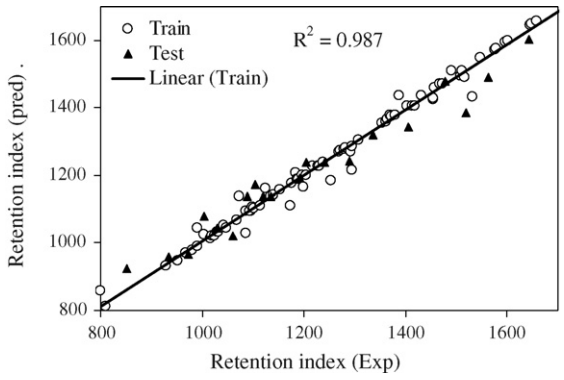

Fig. 4. Predicted vs. experimental retention index (RI) (SVM).

The calculated retention index, obtained from the SVM predictive model, is listed in Table 1. Fig. 4 exhibits the predicted versus the experimental values of the retention index for the training and test sets with the SVM method. Table 2 shows the statistical parameters of the results, attained by four model studies for the same set of essential oil compounds. The RMSE values of the SVM model for the training and test data set were much lower than those of the models proposed in the other methods. The square correlation coefficient (R^2), given by the SVM model, was higher than that of the MLR, PLS and poly-PLS methods. Furthermore, the results of the *F*-test were obtained (Table 2). From this table, it can be noticed that the SVM model gives the highest *F* values, so this model provides the most satisfactory results, compared with the results obtained from the MLR, PLS and poly-PLS methods. Consequently, this SVM approach currently constitutes the most accurate method for predicting the retention index of the *Citrus sudachi* essential oil components.

4. Conclusion

In the present study, two linear methods (MLR and PLS) and two nonlinear methods (poly-PLS and SVM) were used to construct a quantitative relation between the retention index of some essential oil components and their calculated descriptors.

The results obtained by SVM were compared with the results obtained by MLR, PLS and poly-PLS. The results demonstrated that SVM was more powerful in the retention index prediction of the essential oil compounds than MLR, PLS and poly-PLS. A suitable model with high statistical quality and low prediction errors was eventually derived. This model could accurately predict the retention index of these components that did not exist in the modeling procedure. It was easy to notice that there was a good prospect for the SVM application in the QSRR modeling.

References

[1] S.G. Deans, K.P. Svoboda, The antimicrobial properties of marjoram (*Origanum majorana* L.) volatile oil, Flavour Fragrance J. 5 (1990) 187–190.
[2] C.P. Dionigi, D.F. Millie, P.B. Johnsen, Effects of farnesol and the off-flavor derivative geosmin on streptomyces-tendae, Appl. Environ. Microbiol. 57 (1991) 3429–3432.
[3] H. Sugisawa, R.H. Yang, C. Kawabata, H. Tamura, Volatile constituents in the peel oil of sudachi (*Citrus sudachi*), Agric. Biol. Chem. 53 (1989) 1721–1723.
[4] H. Tamura, R.H. Yang, H. Sugisawa, Aroma profiles of peel oils of acid citrus, in: R. Teranishi, R.G. Buttery, H. Sugisawa (Eds.), Bioactive Volatile Compounds from Plants, ACS Symposium Series, vol. 525, American Chemical Society, Washington DC, 1993, pp. 121–136.
[5] H. Tamura, A. Padrayuttawat, T. Tokunaga, Seasonal change of volatile compounds of *Citrus sudachi* during maturation, Food Sci. Technol. Res. 5 (1999) 156–160.
[6] I. Kubo, H. Muroi, A. Kubo, Naturally-occurring antiacne agents, J. Nat. Prod. 57 (1994) 9–17.
[7] D.T. Stanton, P.C. Jurs, Computer-assisted prediction of gas chromatographic retention indices of pyrazines, Anal. Chem. 61 (1989) 1328–1332.
[8] J. Olivero, T. Gracia, P. Payares, R. Vivas, D. Diaz, E. Daza, P. Geerlings, Molecular structure and gas chromatographic retention behavior of the components of Ylang-Ylang oil, J. Pharm. Sci. 86 (1997) 625–630.
[9] R. Acuna-Cueva, F. Hueso-Urena, N.A.I. Cabeza, S.B. Jimenez-Pulidoa, M.N. Moreno-Carretero, J.M.M. Martos, Quantitative structure-capillary column gas chromatographic retention time relationships for natural sterols (trimethylsilyl ethers) from olive oil, JAOCS 77 (2000) 627–630.
[10] B. Hemmateenejad, K. Javadnia, M. Elyasi, Quantitative structure-retention relationship for the Kovats retention indices of a large set of terpenes: a combined data splitting-feature selection strategy, Anal. Chim. Acta 592 (2007) 72–81.
[11] E. Malinowski, Factor Analysis in Chemistry, second ed., Wiley, Chichester, 1991.
[12] S. Wold, N. Kettaneh-Wold, B. Skagerberg, Nonlinear PLS modeling, Chemometr. Intell. Lab. Syst. 7 (1989) 53–65.
[13] V. Vapnik, Estimation of Dependencies Based on Empirical Data, Springer, Berlin, 1982.
[14] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Disc. 2 (1998) 121–167.
[15] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inf. Comput. Sci. 43 (2003) 1882–1889.
[16] H.X. Liu, R.S. Zhang, F. Luan, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Diagnosing breast cancer based on support vector machines, J. Chem. Inf. Comput. Sci. 43 (2003) 900–907.
[17] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Comput. Chem. 26 (2001) 5–14.
[18] S. Riahi, M.F. Mousavi, M. Shamsipur, Prediction of selectivity coefficients of a theophylline-selective electrode using MLR and ANN, Talanta 69 (2006) 736–740.
[19] S. Riahi, R. Ganjali, P. Norouzi, F. jafari, Application of GA-MLR, GA-PLS and the DFT quantum chemical (QM) calculations for the prediction of the selectivity coefficients of a histamine-selective electrode, Sens. Actuators B 132 (2008) 13–19.
[20] A.A. Moosavi-Movahedi, S. Safarian, G.H. Hakimelahi, G. Ataei, D. Ajloo, S. Panjehpour, S. Riahi, M.F. Mousavi, S. Mardanyan, N. Soltani, A. Khalafi-Nezhad, H. Sharghi, H. Moghadamnia, A.A. Saboury, QSAR analysis for ADA upon interaction with a series of adenine derivatives as inhibitors, Nucleos. Nucleot. Nucl. 23 (2004) 613–624.
[21] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, QSAR study of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: an inhibitor of AP-I and NF-kappa B mediated gene expression based on support vector machines, J. Chem. Inf. Comput. Sci. 43 (2003) 1288–1296.
[22] S. Riahi, M.R. Ganjali, E. Pourbasheer, P. Norouzi, QSRR studies on gas chromatography retention index of essential oil compounds based on GA-MLR, Chromatographia 67 (2008) 917–922.
[23] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs, J. Chem. Inf. Comput. Sci. 44 (2004) 161–167.
[24] J. Mookdasanit, H. Tamura, T. Yoshizawa, T. Tokunaga, K. Nakanishi, Trace volatile components in essential oil of *Citrus sudachi* by means of modified solvent extraction method, Food. Sci. Technol. Res. 9 (2003) 54–61.
[25] R. Todeschini, V. Consonni, M. Pavana, [Online] available: http://www.disat.unimib.it/vhm/.
[26] [Online] available: http://www.isis.ecs.soton.ac.uk/isystems/kernel/.
[27] S.R. Gunn, Support Vector Machines for Classification and Regression;, University of Southampton, UK, 1997.
[28] S. Riahi, M.R. Ganjali, A. Beheshti, A. Mohammadi, P. Norouzi, Partition Coefficient Prediction of a Large Set of Various Drugs and Poisons by a Genetic Algorithm and Artificial Neural Network, J. Chin. Chem. Soc. 55 (2008) 345–355.
[29] S. Riahi, A. Beheshti, M.R. Ganjali, P. Norouzi, A Novel QSPR Study of normalized migration time for drugs in capillary electrophoresis by new descriptors; quantum chemical investigation, Electrophoresis 29 (2008) 4027–4035.

[30] J. Hunger, G. Huttner, Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks, J. Comput. Chem. 20 (1999) 455–471.
[31] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali, P. Norouzi, Exploring QSARs for antiviral activity of 4-alkylamino-6-(2-hydroxyethyl)-2- methylthiopyrimidines by support vector machine, Chem. Biol. Drug Des. 72 (2008) 205–216.
[32] C.L. Waller, M.P. Bradley, Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies, J. Chem. Inf. Comput. Sci. 39 (1999) 345–355.
[33] J. Aires-de-Sousa, M.C. Hemmer, J. Casteiger, Prediction of H-1 NMR chemical shifts using neural networks, Anal. Chem. 74 (2002) 80–90.
[34] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[35] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany, 2000.
[36] R. Leardi, Application of genetic algorithm-PLS for feature selection in spectral data sets, J. Chemometr. 14 (2000) 643–655.
[37] R. Leardi, A.L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometr. Intell. Lab. Syst. 41 (1998) 195–207.
[38] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202.
[39] C. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1997.
[40] S. Riahi, M.R. Ganjali, E. Pourbasheer, P. Norouzi, Comparative study of the derivative and partial least squares methods applied to the spectrophotometric simultaneous determination of atorvastatin and amlodipine from their combined drug products, Curr. Pharm. Anal. 3 (2007) 268–272.
[41] W.J. Wang, Z.B. Xu, W.Z. Lu, X.Y. Zhang, Determination of the spread parameter in the Gaussian kernel for classification and regression, Neurocomputing 55 (2003) 643–663.