Journal of Hazardous Materials 179 (2010) 715–720

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

QSPR prediction of flash point of esters by means of GFA and ANFIS

Aboozar Khajeh [a], Hamid Modarress [b,*]

[a] Islamic Azad University, Birjand Branch, Birjand, Southern Khorasan, Iran
[b] Amirkabir University of Technology, Department of Chemical Engineering, 424, Hafez Avenue, 15914 Tehran, Iran

ARTICLE INFO

Article history:
Received 18 October 2009
Received in revised form 10 March 2010
Accepted 13 March 2010
Available online 19 March 2010

Keywords:
Flash point
Ester
QSPR
Genetic function approximation (GFA)
Adaptive neuro-fuzzy inference system (ANFIS)

ABSTRACT

A quantitative structure property relationship (QSPR) study was performed to develop a model for prediction of flash point of esters based on a diverse set of 95 components. The most five important descriptors were selected from a set of 1124 descriptors to build the QSPR model by means of a genetic function approximation (GFA). For considering the nonlinear behavior of these molecular descriptors, adaptive neuro-fuzzy inference system (ANFIS) method was used. The ANFIS and GFA squared correlation coefficient for testing set was 0.969 and 0.965, respectively. The results obtained showed the ability of developed GFA and ANFIS for prediction of flash point of esters.

1. Introduction

Ester compounds are a sort of important medical and chemical materials that use in many foods, cosmetics, medicines and chemicals. The flammability characteristics of esters are very important for safety considerations in storage, processing, and handling. Flash point is one of the major quantities used to characterize the fire and explosion hazard of liquids [1,2]. Flash point is the lowest temperature, corrected to the standard atmospheric pressure of 760 mm Hg (101.3 kPa), at which application of a test flame causes the vapor of a specimen to ignite under specified test conditions [3]. In other words, flash point is the temperature at which the vapor pressure divided by the pressure of the atmosphere is equal to the lower flammability limit (LFL) expressed in mole fraction [3]. The flash point can be determined by open or closed-cup methods [4]. Experimental flash point data is scarce in the literature, thus the development of theoretical prediction methods which are desirably convenient and reliable for predicting flash point is required.

One of the successful approaches for prediction of flash point is quantitative structure property relationships (QSPR) [5–9]. By utilizing QSPR approach, Gharagheizi et al. [5] used a new collection of 79 functional groups to correlate flash point temperature (FP) of 1378 pure compounds. Katritzky et al. [6] developed a general three-parameter QSPR model for prediction of the flash point of a diverse set of 271 compounds based on the multi-parameter regression. By using of the experimental boiling point as a descriptor, their correlation improved resulting a value for squared correlation coefficient $R^2 = 0.9529$. In a subsequent work Katritzky et al. [7] by using geometrical, topological, quantum mechanical and electronic descriptors predicted the flash points of 758 organic compounds by linear and nonlinear methods. Pan et al. [8] developed a back-propagation (BP) neural network based QSPR model for the prediction of flash points of 92 alkanes using group bond contribution method. Quantitative structure property relationship (QSPR) and topological indices have been used by Patal et al. [9] to predict flash point of different classes of solvents by multiple linear regression and back-propagation neural network.

QSPR is a mathematical method that relates simple and complex physicochemical properties of various compounds from numerical descriptors derived from molecular structures. The advantage of this approach over other methods lies in the fact that it requires only the knowledge of chemical structure, and is not dependent on any experimental properties [10]. In QSPR modeling different computational techniques, such as multiple linear regression (MLR) [10,11], partial least square analysis (PLS) [12], Multilayer perceptrons (MLP) neural network [9,13], radial basis function (RBF) neural network [11,14] and support vector machine (SVM) [15,16] have been used.

Neuro-fuzzy as an intelligent computational method is one of the most popular research fields which use ANNs theory in order to determine fuzzy inference properties by processing data samples. A specific approach in neuro-fuzzy development is the adaptive neuro-fuzzy inference system (ANFIS), which has shown significant

* Corresponding author. Tel.: +98 21 64543176.
E-mail address: hmodares@aut.ac.ir (H. Modarress).

0304-3894/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2010.03.060

results in modeling complex nonlinear systems with estimation speed, simplicity, error free and capacity to learn from examples [17–19].

The aim of this work is to build new QSPR model that could be used for predicting flash point of ester compounds from their molecular structure. In this work, after obtaining the most statistically significant descriptors by means of genetic algorithm (GA)-based variable-selection approach, the nonlinear behavior of these molecular descriptors for predicting flash point of ester was studied by means of a hybrid subtractive clustering ANFIS.

2. Materials and methods

2.1. Data set

In this work, the flash point dataset of 95 esters was taken from [20]. Flash point values of these compounds were in the range from −2 to 421°F. The data set was randomly divided into two groups: a training set of 76 compounds and a test set of 19 compounds. The training set was used for model generation and the test set was used for evaluation of the prediction ability of obtained model.

2.2. Molecular descriptors

To obtain QSPR model for each molecule more than 1000 molecular descriptors were calculated utilizing Dragon software developed by the Milano Chemometrics and QSAR research group [21]. These descriptors can be classified into several groups: constitutional descriptors, topological descriptors, connectivity indices, information indices, 2D autocorrelations, Burden eigenvalues descriptors, eigenvalue-based indices, geometrical descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, atom-centred fragments, molecular properties.

2.3. Genetic function approximation (GFA)

The GFA apply the genetic algorithms to the problem of function approximation [22]: given a large number of potential factors influencing a response to find the subset of terms that correlates best with the response. It works in the following way: first of all a particular number of equations (e.g. 50) are generated randomly, then pairs of parent are selected from the present population, with probabilities proportional to their fitness and crossovers are performed and progeny equations are generated. The goodness of each progeny equation is assessed by various score such as:

R-square:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} \tag{1}$$

where SSR is the sum of squares of regression, and SST is the total sum of squares.

Adjusted *R*-square:

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSR}/(n-p)}{\mathrm{SSE}/(n-1)} \tag{2}$$

where SSE is the sum of squares of errors, n is the number of data points from which the model is built, and p is the number of parameters in a regression model.

Friedman's lack of fit (LOF):

$$\mathrm{LOF} = \frac{\mathrm{SSE}}{[1-(c+(dp/n))]^2} \tag{3}$$

where c is the number of basis functions (other than the constant term), d is a user defined smoothness factor, p is the number of features in the model, and n is the number of data points from which the model is built.

If the fitness of new progeny equation is better, then it is preserved.

2.4. Adaptive neuro-fuzzy inference system (ANFIS)

The ANFIS is a multilayer feed-forward neural network which uses neural network learning algorithms and fuzzy reasoning in order to combines the advantages of both neural and fuzzy inference.

Fuzzy logic modeling techniques can be classified into three categories, namely the linguistic (Mamdani-type) [23], the relational equation, and the Takagi-Sugeno-Kang (TSK) [24]. Based on the TSK model, an adaptive network based fuzzy inference system (ANFIS) has been introduced by Jang [25]. In a TSK model with a rule base of M rules, each giving p antecedents, the ith rule can expressed as:

Rule i: if x_i is F_1^i and ... and x_p is F_p^i, then:

$$y^i(X) = c_0^i + c_1^i x_1 + c_2^i x_2 + \ldots + c_p^i x_p = C_i X \tag{4}$$

where $i = 1, 2, \ldots, M$, $c_j^i (j = 0, 1, \ldots, p)$ are the consequent parameters, $y^i(X)$ is the output of the ith rule, and F_k^i ($k = 1, 2, \ldots, p$) are fuzzy sets.

The overall output, $y(X)$, of the model is obtained by combining the outputs from the M rules in the following prescribed way:

$$y(X) = \frac{\sum_{i=1}^{M} f^i(X) y^i(X)}{\sum_{i=1}^{M} f^i(X)} = \frac{\sum_{i=1}^{M} f^i(X)(c_0^i + c_1^i x_1 + \ldots + c_p^i x_p)}{\sum_{i=1}^{M} f^i(X)} \tag{5}$$

where the $f_i(X)$ are rule firing level (strengths), defined as:

$$f^i(X) = T_{k=1}^{p} \mu_{F_k^i}(x_k) \tag{6}$$

in which T denoted a T-norm, usually minimum or product.

In ANFIS architecture, a FIS is described in a layered, feed-forward network structure, where some of the parameters are represented by adjustable nodes and the others as fixed nodes. The ANFIS structure contains five layers described below:

In the first layer, all the nodes are adjustable nodes. They generate fuzzy membership grades of the inputs and outputs of this layer are given by:

$$O_{1,i} = \mu_{A_i(x)}, \quad i = 1, 2 \tag{7}$$

$$O_{1,i} = \mu_{B_{i-2}(y)}, \quad i = 3, 4 \tag{8}$$

where $\mu_{A_i(x)}$ and $\mu_{B_{i-2}(y)}$ can adopt any fuzzy membership function.

The second layer consist of fixed nodes represent the T-norm operators that combine the possible input membership grades in order to compute the firing strength of the rule. The outputs of this layer are given by:

$$O_{2,i} = w_i = \mu_{A_i(x)} \mu_{B_i(y)}, \quad i = 1, 2 \tag{9}$$

The output signal w_i so-called the firing strength of a rule.

The third layer implements a normalization function and the outputs of this layer can be represented as:

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2 \tag{10}$$

In the fourth layer, the nodes are adjustable nodes and every node i has the following function:

$$O_{4,i} = \bar{w}_i f_i = w_i(p_i x + q_i y + r_i), \quad i = 1, 2 \tag{11}$$

where $\bar{w}_i$ is the output of layer 3, and $\{p_i, q_i, r_i\}$ is the parameter set.

The fifth layer represents the aggregation of the outputs performed by weighted summation. The output is computed as:

$$O_{5,i} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{w_1 + w_2} \tag{12}$$

Table 1
The five molecular descriptors used in Eq. (13).

ID	Molecular descriptor	Type	Definition
1	D/Dr05	Topological descriptors	Distance/detour ring index of order 5
2	IVDM	Information indices	Mean information content on the vertex degree magnitude
3	IC0	Information indices	Information content index (neighborhood symmetry of 0-order)
4	MATS5p	2D autocorrelations	Moran autocorrelation – lag 5/weighted by atomic polarizabilities
5	G2s	WHIM descriptors	2st component symmetry directional WHIM index/weighted by atomic electrotopological states

Table 2
Correlation matrix of the five descriptors used in QSPR model.

	D/Dr05	IVDM	IC0	MATS5p	G2s
D/Dr05	1				
IVDM	−0.14248	1			
IC0	0.212743	−0.47576	1		
MATS5p	0.003483	0.181737	0.257629	1	
G2s	0.47399	−0.33704	0.366479	0.227338	1

Table 3
Premise parameters.

	Input 1 [σ,c]	Input 2 [σ,c]	Input 3 [σ,c]	Input 4 [σ,c]	Input 5 [σ,c]
Rule 1	[2.204 −4.353E−011]	[0.4124 2.253]	[0.08873 1.156]	[0.3891 −0.09068]	[0.1447 0.2092]
Rule 2	[2.204 −4.353E−011]	[0.4068 3.03]	[0.06989 1.324]	[0.4097 −0.03908]	[0.1431 0.2406]
Rule 3	[2.204 −4.353E−011]	[0.3984 3.775]	[0.1382 1.439]	[0.3943 0.00086]	[0.1474 0.2998]

Table 4
Consequent parameters.

Consequent parameters	C_1^i	C_2^i	C_3^i	C_4^i	C_5^i	C_0^i
Rule 1	0.005274	235.7	85.41	−22.56	−262.1	−663
Rule 2	4.082E−007	222.9	39.05	−22.77	390.1	−738.1
Rule 3	12.16	264	165.5	−103	123.6	−856.2

3. Results and discussion

Based on the GFA (17000 iterations, *R*-square score, 50 population size, 50% mutation probability) the following equation was derived by using Materials Studio Software of Accelrys Inc. [26] with five descriptors:

$$\mathrm{FP} = 14.823500795\left(\frac{D}{Dr05}\right) + 206.42085581(IVDM) + 267.756013902(IC0) - 34.357883234(MATS5p) + 94.539715611(G2s) - 885.235147504 \quad (13)$$

D/Dr05 is a topological descriptor which is a distance/detour ring index of order 5. Positive coefficient of this descriptor suggests that existence of five-member ring group causes an increase in the flash point of ester components. *IVDM* based on the partition of vertices according to the vertex degree magnitude and is a measure of molecular complexity together with some other information indices derived from the distance matrix [27]. This descriptor has a significant effect on flash point value than that of the other descriptors. *IC0* is an information content index which is neighborhood symmetry of 0-order calculated from the molecular formula [27]. It reflects the branching and atom composition diversity of a molecule. Increase in this descriptor increases the flash point. The 2D autocorrelation descriptor *MATS4p*, based on Moran Autocorrelation of topological Structure, describe how a considered property is distributed along a topological molecular structure. When this descriptor increases the flash point decreases. *G2s* is 2st component symmetry directional WHIM index which weighted by atomic electrotopological states [27]. The coefficient for the *G2s* descriptor is positive, meaning that increasing the molecular symmetry can lead to enhancement of flash point value. The molecular descriptors and their physical meanings are presented in Table 1.

The correlation matrix of the selected descriptors is given in Table 2, which shows that the five descriptors are independent of each other and could be used to develop a QSPR model.

In this work, the ANFIS model on the basis of the subtractive clustering algorithm with inputs and outputs similar to GFA was developed. The subtractive clustering algorithm is a modification of the mountain method of Yager and Filev [28] and was introduced by Chiu [29]. This method generates an ANFIS structure using the clustering algorithm subtractive clustering and generates an FIS with the minimum number of rules required to distinguish the fuzzy

Table 5
Squared correlation coefficient (R^2) and root mean squares error (RMSE) for GFA and ANFIS methods.

Data set	GFA		ANFIS	
	R^2	RMSE	R^2	RMSE
Training set	0.9648	16.26704	0.9754	13.58811
Test set	0.9646	21.75769	0.9691	20.16043
Total	0.9642	17.50351	0.9732	15.13268

Table 6
Comparison between the presented models and previous models.

No.	Model	R^2	RMSE
1	Tetteh et al. [31]	0.9326	13.1
2	Katritzky et al. [6]	0.9529	11.2
3	Katritzky et al. [7]	0.878	–
4	Gharagheizi and Alamdari [32]	0.9669	12.7
5	Gharagheizi et al. [5]	0.9757	11.206
6	Current work (GFA)	0.9642	17.50351
7	Current work (ANFIS)	0.9732	15.13268

Table 7
Reported flash points in Ref. [20] and predicted flash points by QSPR model using GFA and ANFIS methods for the esters and calculated descriptors for QSPR model.

		Reported flash points	Predicted flash points		Descriptors			
			GFA	ANFIS	IVDM	IC0	MATS5p	G2s
Train								
1	Benzyl acetate	215	209.63590	200.210	3.391	1.357	−0.302	0.224
2	Benzyl benzoate	311	308.72830	291.090	3.962	1.296	−0.297	0.200
3	Sec-butyl acetate	60	83.23307	79.239	2.842	1.295	−0.333	0.250
4	Tert-butyl acetate	61	53.42044	53.261	2.753	1.295	0.00	0.250
5	Butyl acrylate	102	125.61830	118.700	3.078	1.313	−0.036	0.240
6	Butyl benzoate	223	241.25000	234.660	3.642	1.297	−0.212	0.213
7	Tert-butyl butanoate	116	109.41890	108.890	3.128	1.239	0.250	0.273
8	Butyl formate	64	74.92512	71.159	2.752	1.333	−0.300	0.263
9	Sec-butyl formate	78	85.97113	84.767	2.689	1.333	−1.000	0.263
10	Butyl methacrylate	126	139.52650	139.270	3.197	1.281	0.000	0.231
11	Decyl acetate	220	223.76100	226.960	3.748	1.167	−0.093	0.208
12	Decyl formate	211	212.45790	212.910	3.668	1.182	−0.114	0.213
13	Diallyl maleate	248	265.79630	257.670	3.719	1.46	0.867	0.235
14	Dibutyl maleate	284	274.37410	283.880	3.923	1.352	0.905	0.200
15	Dibutyl sebacate	352	377.45640	367.500	4.404	1.235	−0.164	0.183
16	Diethyl carbonate	77	94.45461	82.444	2.896	1.415	0.600	0.250
17	Diethylene glycol ethyl ether acetate	224	235.29900	227.760	3.516	1.379	−0.143	0.218
18	Diethyl maleate	199	201.84060	197.050	3.482	1.459	1.250	0.218
19	Diethyl oxalate	252	217.06540	239.020	3.197	1.485	−0.667	0.231
20	Diethyl phthalate	322	300.93050	303.930	3.906	1.429	0.630	0.200
21	Diethyl succinate	194	191.39810	192.980	3.482	1.42	1.250	0.218
22	Dimethyl maleate	235	236.91340	234.910	3.197	1.53	1.222	1.000
23	1,2-Dimethylpropyl acetate	109	104.18490	102.520	2.983	1.265	−0.357	0.240
24	2,2-Dimethylpropyl formate	100	98.92504	90.686	2.807	1.295	−1.000	0.250
25	2,2-Dimethylpropyl propanoate	142	123.86400	124.940	3.128	1.239	−0.286	0.231
26	Dimethyl terephthalate	313	302.89590	292.510	3.700	1.483	−0.222	0.208
27	Dodecyl acetate	240	256.90070	266.300	3.948	1.143	−0.065	0.200
28	Dodecyl butanoate	319	286.89760	301.570	4.124	1.124	−0.048	0.193
29	Dodecyl propanoate	299	276.10170	299.930	4.039	1.133	−0.056	0.236
30	2-Ethoxyethyl acetate	130	137.75860	126.150	3.078	1.379	0.125	0.240
31	Ethylacetoacetate	135	146.81040	145.160	3.031	1.433	0.000	0.240
32	Ethyl acrylate	60	55.80975	50.691	2.689	1.4	0.400	0.263
33	Ethyl benzoate	190	205.61600	197.530	3.391	1.357	−0.185	0.224
34	Ethyl cyanoacetate	230	188.88110	220.150	2.896	1.689	−0.013	0.250
35	Ethylene carbonate	305	305.00000	305.000	2.522	1.571	0.000	0.679
36	Ethylene glycol diacetate	190	182.70750	188.680	3.197	1.485	0.333	0.231
37	Ethyl formate	25	−8.10189	8.5227	2.250	1.435	0.000	0.301
38	2-Ethylhexyl acetate	190	180.40090	178.210	3.482	1.198	−0.160	0.218
39	Ethyl isobutanoate	57	60.35072	56.375	2.842	1.295	0.333	0.250
40	Ethyl isovalerate	95	92.00100	94.298	3.031	1.265	0.286	0.240
41	Ethyl lactate	115	110.79420	117.840	2.842	1.415	−0.200	0.250
42	Ethyl methacrylate	70	75.61281	66.299	2.842	1.352	0.333	0.250
43	1-Ethylpropyl acetate	112	114.09310	111.470	3.031	1.265	−0.357	0.240
44	Glyceryl triacetate	280	312.92270	306.640	3.771	1.501	0.042	0.204
45	Heptyl acetate	169	166.68520	164.750	3.384	1.217	−0.185	0.224
46	Hexyl formate	134	133.49670	129.610	3.125	1.265	−0.357	0.240
47	Isobutyl acrylate	88	122.54760	114.470	3.031	1.313	−0.229	0.240
48	Isobutyl butanoate	135	134.01850	135.130	3.197	1.239	−0.167	0.231
49	Isobutyl formate	84	85.97113	84.767	2.689	1.333	−1.000	0.263
50	Isobutyl isobutyrate	100	125.34880	127.210	3.155	1.239	−0.167	0.231
51	Isopentyl acetate	77	103.06420	103.160	3.031	1.265	−0.036	0.240
52	Isopentyl formate	105	86.75234	81.673	2.896	1.295	−0.111	0.250
53	Isopentyl isovalerate	171	166.68250	167.030	3.447	1.198	0.029	0.218
54	Isopentyl propanoate	138	128.28070	131.370	3.197	1.239	0.000	0.231
55	Methyl acetate	14	−27.50540	−14.831	2.156	1.435	0.000	0.301
56	Methyl acrylate	26	36.70287	59.801	2.446	1.459	0.000	0.279
57	Methyl benzoate	181	192.24620	177.170	3.246	1.392	−0.375	0.231
58	Methyl butanoate	57	37.87009	29.554	2.689	1.333	0.400	0.263
59	Methyl chloroacetate	125	125.33010	120.310	2.446	1.79	0.000	0.279
60	Methyl dodecanoate	242	241.08780	247.370	3.852	1.155	−0.077	0.204
61	Neopentyl acetate	109	97.99229	97.072	2.953	1.265	−0.357	0.240
62	Nonyl acetate	186	206.05890	206.210	3.637	1.182	−0.114	0.213
63	Octyl acetate	188	186.83510	184.850	3.516	1.198	−0.143	0.218
64	Octyl butanoate	220	223.76100	226.960	3.748	1.167	−0.093	0.208
65	Octyl formate	175	174.52920	172.130	3.422	1.217	−0.185	0.224
66	Pentyl butanoate	162	166.68520	164.750	3.384	1.217	−0.185	0.224
67	Pentyl propanoate	146	149.51050	145.190	3.239	1.239	−0.250	0.273
68	Propyl acetate	59	61.92061	63.139	2.689	1.333	−0.300	0.263
69	Propyl acrylate	101	102.01440	92.698	2.896	1.352	−0.111	0.250
70	Propyl methacrylate	120	115.91650	109.270	3.031	1.313	−0.036	0.240
71	Undecyl acetate	256	241.08780	247.370	3.852	1.155	−0.077	0.204
72	Undecyl butanoate	299	272.41460	284.750	4.039	1.133	−0.056	0.197
73	Undecyl formate	234	229.74720	233.340	3.777	1.167	−0.093	0.208

Table 7 (*Continued*)

		Reported flash points	Predicted flash points		Descriptors			
			GFA	ANFIS	IVDM	IC0	MATS5p	G2s
74	Undecyl propanoate	278	256.90070	266.300	3.948	1.143	−0.065	0.2
75	Vinyl formate	−2	17.33493	28.000	2.250	1.53	0.000	0.301
76	Vinyl propionate	34	55.80975	50.691	2.689	1.4	0.400	0.263
Test								
77	Butyl acetate	71	86.75234	81.673	2.896	1.295	−0.111	0.250
78	Butyl butanoate	128	136.95040	139.410	3.239	1.239	0.000	0.231
79	Butyl propanoate	90	117.39850	105.460	3.078	1.265	−0.036	0.289
80	Dibutyl phthalate	315	349.02070	335.830	4.246	1.342	0.562	0.188
81	Diethyl malonate	199	231.49430	220.030	3.346	1.451	−0.476	0.224
82	Dimethyl phthalate	295	274.65370	270.750	3.700	1.483	0.600	0.208
83	Dioctyl phthalate	421	421.80280	423.190	4.753	1.234	0.604	0.172
84	Ethyl acetate	25	15.28239	29.917	2.446	1.379	0.000	0.279
85	Ethyl butanoate	78	71.49744	67.016	2.896	1.295	0.333	0.250
86	2-Ethylhexyl acrylate	160	208.11160	203.580	3.605	1.211	−0.140	0.213
87	Ethyl propanoate	54	37.87009	29.554	2.689	1.333	0.400	0.263
88	Isobutyl acetate	64	83.23307	79.239	2.842	1.295	−0.333	0.250
89	Isobutyl propanoate	117	109.69530	108.230	3.031	1.265	−0.229	0.240
90	Isopropyl acetate	36	38.60873	40.625	2.626	1.333	0.000	0.263
91	3-Methylbutyl butanoate	156	155.33020	159.410	3.346	1.217	0.019	0.261
92	Nonyl formate	189	193.85340	191.920	3.550	1.198	−0.143	0.218
93	Pentyl formate	114	113.18830	103.500	2.950	1.295	−0.556	0.250
94	Tetradecyl formate	319	277.36870	290.090	4.063	1.133	−0.056	0.197
95	Vinyl acetate	17	36.702870	59.801	2.446	1.459	0.000	0.279

qualities associated with each of the clusters. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system's behavior [30]. The advantage of the subtractive clustering algorithm is the fact that the number of clusters does not need to be specified in advance and the algorithm itself determines the number of clusters. However, four other parameters, the cluster radius, the squash factor, accept and reject ratio need to be set. In this method, the total number of fuzzy rules is only related to the number of clusters. Hence, it will be a correct choice to use this algorithm for solving the problems with the large number of input dimension. The Gaussian membership function defined in Eq. (14) used in the ANFIS model.

$$f(x;\sigma,c) = \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right) \tag{14}$$

where c and σ are parameters of the membership function, governing the Gaussian functions accordingly.

Hybrid learning rule is used to train the model according to input/output data pairs. A hybrid algorithm can be divided into forward pass and a backward pass. The forward pass of the learning algorithm stop at nodes at layer 4 and the consequent parameters are identified by least squares method. In the backward pass, the error signals propagate backward and the premise parameters are undated by gradient descent. It has been proven that this hybrid algorithm is highly efficient in training the ANFIS [25]. The premise and consequent parameters that have been optimized by hybrid algorithm are given in Tables 3 and 4. All ANFIS calculations were carried out using Matlab mathematical software with fuzzy logic toolboxes for windows running on a personal computer.

To compare the prediction abilities of the methods, two statistical parameters namely root mean squares error (RMSE) and squared correlation coefficient (R^2), were calculated and the results are shown in Table 5.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{\exp} - y_i^{\text{calc}})^2}{\sum_{i=1}^{n}(y_i^{\exp} - \bar{y})^2} \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i^{\exp} - y_i^{\text{calc}})^2}{n}} \tag{16}$$

where $y_i^{\exp}$, y_i^{calc}, and $\bar{y}$ are values of experimental, calculated and average of calculated property and n is the number of compounds in dataset. The results show the ability of GFA and ANFIS methods and also indicate that ANFIS model is more accurate than GFA model.

Although, no similar work has been done by the approaches employed in this work for this group of compounds, but a comparison between this work and other QSPR works for estimating the flash point of pure compounds included esters have been made in Table 6.

The values of the descriptors for QSPR model and the flash points reported in Ref. [20] and the flash points predicted by QSPR model for training and test set are presented in Table 7. ANFIS method used five descriptors that were selected because they maximize the performance of the GFA model, but as can be seen in this table, these descriptors are efficient for ANFIS. Therefore, five selected descriptors and ANFIS method are convenient for prediction of flash point of ester. It is worth noting that the *D*/*Dr05* descriptor for ethylene

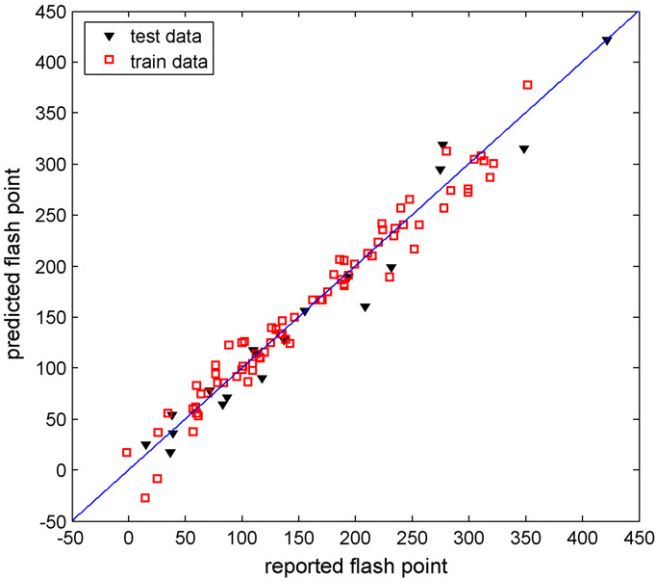

Fig. 1. Predicted flash points by GFA method vs. reported flash points [20].

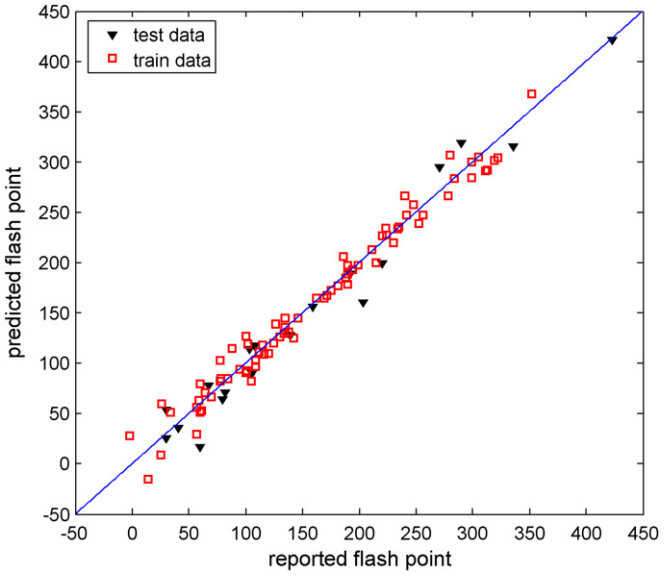

Fig. 2. Predicted flash points by ANFIS method vs. reported flash points [20].

carbonate with a five-member ring group has the value of 12.467 where as for the other compounds this descriptor has the value of zero. However selecting *D*/*Dr05* descriptor among more than 1000 descriptors indicates the important effect of a five-member ring group in flash point calculations.

The plot of predicted and reported flash points is shown in Figs. 1 and 2 that indicate excellent correlation between the reported and predicted values and confirms the good predictive ability of QSPR model.

4. Conclusions

The QSPR model of the flash point of some esters was successfully developed based on various molecular descriptors by using genetic function approximation (GFA) and adaptive neuro-fuzzy inference system (ANFIS) methods. The squared correlation coefficient of 0.969 for ANFIS and 0.965 for GFA for testing set show that these methods have good predictive ability and robustness for estimating flash point of esters.

References

[1] H.J. Liaw, T.A. Wang, A non-ideal model for predicting the effect of dissolved salt on the flash point of solvent mixtures, J. Hazard. Mater. 141 (2007) 193–201.
[2] H.J. Liaw, C.T. Chen, V. Gerbaud, Flash-point prediction for binary partially miscible aqueous-organic mixtures, Chem. Eng. Sci. 63 (2008) 4543–4554.
[3] D. Kong, D.J. Ende, S.J. Brenek, N.P. Weston, Determination of flash point in air and pure oxygen using an equilibrium closed bomb apparatus, J. Hazard. Mater. 102 (2003) 155–165.
[4] J.C. Jones, J. Godefroy, A reappraisal of the flash point of formic acid, J. Loss Prev. Process Ind. 15 (2002) 245–247.
[5] F. Gharagheizi, R.F. Alamdari, M.T. Angaji, A new neural network-group contribution method for estimation of flash point temperature of pure components, Energy Fuels 22 (2008) 1628–1635.
[6] A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, QSPR analysis of flash points, J. Chem. Inf. Comput. Sci. 41 (2001) 1521–1530.
[7] A.R. Katritzky, I.B. Stoyanova-Slavova, D.A. Dobchev, M. Karelson, QSPR modeling of flash points: an update, J. Mol. Graph. Model. 26 (2007) 529–536.
[8] Y. Pan, J. Jiang, Z. Wang, Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network, J. Hazard. Mater. 147 (2007) 424–430.
[9] S.J. Patel, D. Ng, M.S. Mannan, QSPR flash point prediction of solvents using topological indices for application in computer aided molecular design, Ind. Eng. Chem. Res. 48 (2009) 7378–7387.
[10] J. Xu, B. Chen, W. Xu, S. Zhao, Ch. Yi, W. Cui, 3D-QSPR modeling and prediction of nonlinear optical responses of organic chromophores, Chemometr. Intell. Lab. Syst. 87 (2007) 275–280.
[11] H.F. Chen, Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression, Anal. Chim. Acta 609 (2008) 24–36.
[12] L. Li, S. Xie, H. Cai, X. Bai, Z. Xue, Quantitative structure–property relationships for octanol–water partition coefficients of polybrominated diphenyl ethers, Chemosphere 72 (2008) 1602–1606.
[13] R. Wang, J. Jiang, Y. Pan, H. Cao, Y. Cui, Prediction of impact sensitivity of nitro energetic compounds by neural network based on electrotopological-state indices, J. Hazard. Mater. 166 (2009) 155–186.
[14] H. Modarresi, H. Modarress, J.C. Dearden, QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach, Chemosphere 66 (2007) 2067–2076.
[15] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine, J. Mol. Graph. Model. 27 (2008) 188–196.
[16] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines, J. Hazard. Mater. 151 (2008) 603–609.
[17] M.A. Akcayol, Application of adaptive neuro-fuzzy controller for SRM, Adv. Eng. Softw. 35 (2004) 129–137.
[18] A. Khajeh, H. Modarress, B. Rezaee, Application of adaptive neuro-fuzzy inference system for solubility prediction of carbon dioxide in polymers, Expet. Syst. Appl. 36 (2009) 5728–5732.
[19] A. Khajeh, H. Modarress, Prediction of solubility of gases in polystyrene by adaptive neuro-fuzzy inference system and radial basis function neural network, Expet. Syst. Appl. 37 (2010) 3070–3074.
[20] C.L. Yaws, Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds, Norwich, New York, 2003.
[21] http://www.michem.disat.unimib.it/chm/.
[22] D. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.
[23] E.H. Mamdani, S. Assilian, An experiment in linquistic synthesis with a fuzzy logic controller, Internat. J. Man-Mach. Stud. 7 (1975) 1–13.
[24] M. Sugeno, Industrial Applications of Fuzzy Control, Elsevier, Amsterdam, 1985.
[25] J. Jang, ANFIS: adaptive network-based fuzzy inference systems, IEEE Trans. Syst. Man Cybern. 23 (3) (1993) 665–685.
[26] http://accelrys.com.
[27] R. Todeschini, V. Consonni, in: R. Manhold, H. Kubinyi, H. Temmerman (Eds.), Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[28] R. Yager, D. Filev, Approximate clustering via the mountain method, IEEE Trans. Syst. Man Cybern. 24 (8) (1994) 1279–1284.
[29] S.L. Chiu, Fuzzy model identification based on cluster estimation, J. Intell. Fuzzy Syst. 2 (3) (1994) 267–278.
[30] Y.L. Loukas, Adaptive neuro-fuzzy inference system: an instant and architecture-free predictor for improved QSAR studies, J. Med. Chem. 44 (2001) 2772–2783.
[31] J. Tetteh, T. Suzuki, E. Metcalfe, S. Howells, Quantitative structure–property relationships for the estimation of boiling point and flash point using a radial basis function neural network, J. Chem. Inf. Comput. Sci. 39 (1999) 491–507.
[32] F. Gharagheizi, R.F. Alamdari, Prediction of flash point temperature of pure components using a quantitative structure–property relationship model, QSAR Comb. Sci. 27 (2008) 679–683.