Microporous and Mesoporous Materials 231 (2016) 82–91

Contents lists available at ScienceDirect

Microporous and Mesoporous Materials

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

The rational synthetic parameter analysis for subclasses of microporous aluminophosphates based on hierarchical feature selection model

Yuting Guo [a], Zhenhua Tian [b], Na Gao [c, d], Miao Qi [a], Jianzhong Wang [a, *]

[a] College of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China
[b] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[c] State Key Laboratory of Inorganic Synthesis and Preparative Chemistry, Changchun 130012, China
[d] Institute of Modern Agriculture, Jilin Economic Management Cadre College, Changchun 130012, China

ARTICLE INFO

Article history:
Received 25 January 2016
Received in revised form
20 April 2016
Accepted 14 May 2016
Available online 14 May 2016

Keywords:
AlPOs database
Aluminophosphates
Subclass
Rational synthesis
Feature selection

ABSTRACT

Open-framework aluminophosphates (AlPOs) is an important family of porous crystal materials. But the synthetic chemistry of this kind of materials is very complicated, and the synthesis mechanism has not been clearly understood yet. In this paper, we propose a Hierarchical Feature Selection Model (HFSM) composed of two layers to analyze the rational synthetic parameters for the subclass of microporous aluminophosphates (AlPOs) containing (6,8)-rings. In the first layer, we select a feature subset that could separate the (6,8)-ring-containing microporous AlPOs from other AlPOs. In the second layer, we further analyze which of these selected features are critical for the formation of each special subclass in (6,8)-ring-containing microporous AlPOs. With the optimal feature subset selected by the proposed model, we can obtain the highest accuracy rates as 94.28%, 94.03%, 91.27% and 92.20% for the classification of AEN, AWO, CHA and ERI, respectively. Extensive analysis is presented for the synthetic parameters selected by the hierarchical model, which could provide a useful guidance to the rational synthesis of such materials.

1. Introduction

Zeolites and related microporous materials have been widely applied in petroleum industry for catalysis, separation and ion-exchange [1,2]. These materials are formed by TO_4 tetrahedra (T infers to Si, P, Al, Ge, Ga, etc.) with a well defined regular pore system. As an important member of Zeolites and related microporous materials, open-framework aluminophosphate molecular sieve has caught much attention for the past 20 years [3–7]. However the process for synthesizing such materials is complicated and influenced by many parameters, such as gel composition, PH value, solvent, template etc. In order to provide guidance to rational synthesis of microporous inorganic materials, the group of State Key Laboratory of Inorganic Synthesis and Preparative Chemistry of Jilin University established AlPOs synthesis database internationally [8]. This database contains about 1700 synthetic records which belong to 17 classes according to the sizes of the pore rings.

Data mining is the transformation bridge from data and information to knowledge. With the rapid development of computer technology and artificial intelligence, data mining plays an important role in more and more application fields. In chemical researches, data mining has been widely applied to the original data processing and retrieval [9–11], statistical analysis [12–14] and parameterization of the molecular descriptors [15–17]. The establishment of the AlPOs synthesis database makes it possible that we can use data mining technology to study the synthetic parameters for the rational synthesis of AlPOs. Recently, several researchers have already adopted data mining techniques to analyze the impact of the synthetic parameters on the resulting structures based on the database established by Jilin University. In Ref. [18], Li et al. studied the relationship between synthetic parameters and rational formation of (6,12)-ring-containing AlPOs. The feature subset to be evaluated in Ref. [18] was obtained through an exhaustive searching strategy. They evaluated the classification performance of all the combinations of synthetic parameters using Support Vector Machines. The combined parameters which influenced most for distinguishing (6,12)-ring-containing AlPOs from non-(6,12)-ring-containing AlPOs were deemed as the optimal feature subset. In

* Corresponding author. School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China.
E-mail address: wangjz019@nenu.edu.cn (J. Wang).

http://dx.doi.org/10.1016/j.micromeso.2016.05.020
1387-1811/

Ref. [19], Yao et al. analyzed the affect of different synthetic parameters on the production of (6,12)-ring-containing AlPOs. In order to take the discriminant information of features into account, they calculated the importance degree of each feature by a fusing method which fused Fisher Score [20] and Mutual Information [21]. In Ref. [22], Qi et al. explored the relationship between the synthetic parameters and the specific resulting structure containing (6,12)-ring. In their method, the random subspace technique was first employed to pre-rank the synthetic parameters. Then, the fusion weights of synthetic parameters were obtained by Fisher score. At last, a sequential forward searching algorithm was utilized to select the most significant synthetic parameters based on the fusing results of previous two steps. In Ref. [23], Gao et al. discussed the impact of the parameters on the formation of (6,8)-ring-containing AlPOs and the subclasses of this ring-type based on their professional knowledge, and they validated their conclusions by Support Vector Machine (SVM).

Although the pioneering works mentioned above have made some achievements, there are still some limitations in them. In Refs. [18,19 and 22], the researchers only took the (6,12)-ring-containing AlPOs as prediction target to analyze the synthetic parameters. However, the (6,12)-ring-containing AlPOs actually consist of several subclasses. For example, both ATO and AFI are (6,12)-ring-containing AlPOs, but they are two kinds of molecular sieves since they are different in topological structures, and the relevant synthetic parameters which are important for their formation may be also different. Therefore, in order to analyze the synthetic parameters more reasonable, the diversity of the subclasses for a specific class of AlPOs must be taken into account. In Ref. [23], although the subclasses of (6,8)-ring-containing AlPOs have been considered, the optimal feature subsets for them was analyzed by professional knowledge of the domain experts rather than fully mining the data. Furthermore, the correlation among the selected features was neglected in all aforementioned works [18,19,22,23]. This may cause information redundancy since a good feature subset should be the one that contains features highly correlated with the class, while uncorrelated with the features each other [24].

Among various microporous aluminophosphates, AlPO molecular sieve with 8-ring channel is a typical kind of small pore materials used for gas separation and catalysis. For example, SAPO-34 (CHA zeotype) is an effective catalyst for the conversion of methanol to light olefins [25,26]. Thus, in this paper we focus on analyzing the formation parameters of the four important subclasses (AEN, AWO, CHA and ERI) of (6,8)-ring-containing AlPOs. In order to overcome the limitations of the previous works and better analyze the formation parameters of subclasses, we propose a Hierarchical Feature Selection Model (HFSM). Compared with the previous works in Refs. [18,19,22 and 23], the proposed model possesses two advantages: (1) The proposed model takes the subclasses of (6,8)-ring-containing AlPOs into consideration. Thus, with the proposed HFSM, the parameters which are critical for the formation of each specific subclass of (6,8)-ring-containing AlPOs can be well analyzed. (2) Since the correlations among the selected features are considered in our model, the feature subset selected by HFSM is more optimal than the previous works.

This paper is organized as follows. The material and method are presented in Section 2. The experimental results and analysis are shown in Section 3 and the conclusions are given in Section 4.

2. Materials and method

2.1. Data sets

Like other works in Refs. [18,19,22 and 23], we also use the microporous AlPOs database established by State Key Laboratory of Inorganic Synthesis and Preparative Chemistry of Jilin University in this study(http://zeobank.jlu.edu.cn/). By removing records which include missing values, we use the remainder 1279 records as the experimental samples. The experimental samples contain 332 (6,8)-ring-containing AlPOs which are composed of 18 zeotypes (or subclass). The number of four subclasses (AEN, AWO, CHA and ERI, as shown in Fig. 1) that possess the largest sample number of the (6,8)-ring-containing AlPOs are 26, 77, 100 and 36 respectively. The descriptions of the input synthetic parameters are shown in Table 1. The gel chemistry is crucial for the formation of microporous AlPOs. Therefore, four important features related to the molar concentrations of Al_2O_3, P_2O_5, solvent and the organic template in the starting gel are used as part of the input features for training the classifier [18]. As a result, the remaining seventeen synthetic parameters (or features) belonging to two classes (solvent and organic template) in Table 1 are analyzed in this work. Here, it should be pointed out that some other parameters such as synthetic temperature, time and pressure which may also crucial for the synthesis of zeolites are not considered in our paper and previous works in Refs. [18,19,22,23]. This is due to that some records in the microporous AlPOs database did not provide the values for them. Even for the records contain these parameters, most of them are given in a range form rather than precise value, which makes them hardly to be exactly analyzed by the data mining techniques.

2.2. The Hierarchical Feature Selection Model

AEN, AWO, CHA and ERI are four subclasses of (6,8)-ring-containing aluminophosphate molecular sieve. Although they all contain (6,8)-rings, their topological structures are very different from each other, as shown in Fig. 1. Thus, in order to better understand the rational synthesis of the four subclasses, it is necessary to select the important synthetic parameters for each of them.

Feature selection is one of the key steps in machine learning and pattern recognition problems. The aim of feature selection is to find the optimal feature subset that is necessary and sufficient for a specific task. Feature selection has several potential benefits, such as improving the accuracy of classification, avoiding the well-known “curse of dimensionality” problem, speeding up the training process and reducing storage demands. Specially, it can provide a better understanding and interpret ability for a domain expert [27,28].

In our study, let$X = [x_1, x_2,\ldots, x_n] \in \Re^{D\times n}$ be the entire AlPOs dataset containing n samples in D dimensional space. Suppose that these samples belong to C classes, we can denote the dataset as $\{X_1,\ldots,X_C\}$, in which X_i $(i = 1, \ldots,C)$ is the samples in the ith class. The original feature set is denoted by F, and each feature in the data set is F_m $(m = 1, \ldots, D)$. With the aim of analyzing the relationship between the synthetic parameters and the (6,8)-ring-containing AlPOs subclasses, we propose a Hierarchical Feature Selection Model as shown in Fig. 2.

From the flowchart in Fig. 2, we could see that the proposed feature selection model consists of two layers. In the first layer, we find the optimal feature subset Q which could separate the (6,8)-ring-containing AlPOs form non-(6,8)-ring-containing AlPOs. Then, in the second layer of the model, we further analyze which of the features in Q are important for the formation of AEN, AWO, CHA and ERI respectively.

2.2.1. The first layer of the Hierarchical Feature Selection Model

As discussed in Ref. [24], the correlation among features is a critical factor which should be taken into consideration in the feature selection process. For the AlPOs synthesis database analyzed in our study, there are some serious correlation relationships between the synthetic parameters. Taking the synthetic

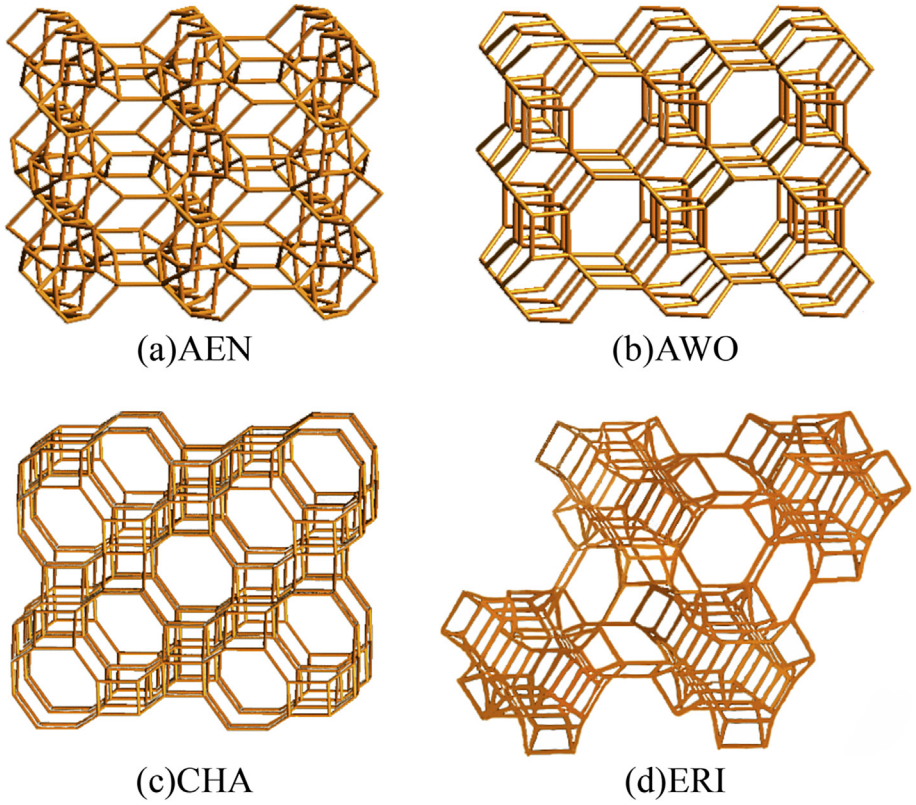


Fig. 1. Subclasses of the (6,8)-ring-containing AlPOs. (a) AEN-zeotype AlPOs, (b) AWO-zeotype AlPOs, (c) CHA-zeotype AlPOs and (d) ERI-zeotype AlPOs.

Table 1
Description of the input synthetic parameters (or features)[a].

ID	Description of parameter	ID	Description of parameter
F1	The molar of Al_2O_5	F12	The second longest distance of template
F2	The molar of P_2O_3	F13	The shortest distance of template
F3	The molar of solvent	F14	The Van der Waals volume
F4	The molar of template	F15	The dipole moment
F5	The density	F16	The ratio of C/N
F6	The melting point	F17	The ratio of N/(C + N)
F7	The boiling point	F18	The ratio of N/Van der Waals volume
F8	The dielectric constant	F19	The Sanderson electronegativity
F9	The dipole moment	F20	The number of free rotated single bond
F10	The polarity	F21	The maximal number of protonated H atoms
F11	The longest distance of template		

[a] F1–F4:Gel composition parameters; F5–F10: solvent parameters; F11–F21: organic template parameters.

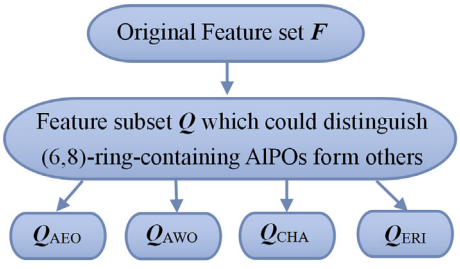


Fig. 2. Flowchart of the proposed model.

parameters in Table 1 as examples, the Pearson Correlation Coefficient between F17 and F18 is 0.95, and the Pearson Correlation Coefficient between F8 and F10 is 0.99. However, since the feature selection techniques employed in previous works (such as Fisher Score and Mutual Information in Refs. [19] and [22]) neglect the correlation among features, the features selected by them may contain some redundant information, which makes the selected feature subset far from optimal [40].

In order to overcome this limitation, the Maximum Weight and Minimum Redundancy (MWMR) [29] is utilized in the first layer of our model to find the optimal feature subset Q for distinguishing (6,8)-ring-containing AlPOs from other AlPOs in the AlPOs synthesis database. MWMR is a newly proposed feature selection framework which considers the weights and the correlations of the features simultaneously during the feature selection process. The objective function of MWMR can be defined as:

$$\max_{y}\left(\frac{y^T W}{d}-\frac{y^T R y}{d(d-1)}\right) \quad s.t.\ \sum_i y_i = d, y_i \in \{0,1\} \tag{1}$$

where d is the number of the selected features, y is the indicator vector indicating which features are selected into the subset Q. W and R denote the weights (or importance) and correlations of the features respectively. In Eq. (1), y^TW/d and $y^TRy/d(d-1)$ denote the mean of the weights and the mean of the correlation for the selected features. The constraints are used for restricting the number of selected features in the Q to be d. Thus, MWMR could select d features from the original feature set F to form the optimal feature subset Q in which the weights (or importance) of features are maximum while the correlations among the features are minimum.

Besides considering the correlation among features, another advantages of MWMR is that the importance and correlation of features are not restricted to a specific measurement. Therefore, we can choose any suitable method to estimate them for particular task. However, most existing methods for feature importance measurement cannot deal with the subclass of the data. For example, a fundamental assumption of well known Fisher Score is that class distributions are homoscedastic, which is rarely true in practice [30]. Thus, although the Fisher Score has been widely used in synthetic parameter analysis, it is not reasonable to employ it in our study which takes the subclasses of (6,8)-ring-containing AlPOs into account.

For the sake of estimating the importance of each feature in our problem more accurately, a new algorithm termed Subclass Discriminant Analysis Score (SDAS) is proposed in this paper. SDAS considers the subclass by measuring not only the intra-class dispersion of each subclass but also the inter-class dispersion between the subclass which belong to different classes. For n labeled samples belonging to C classes$\{X_1,...,X_C\}$, suppose that there exist several subclasses within each class,$\{X_{11},...,X_{ij},...,X_{CHc}\}$,where X_{ij} denotes the j-th subclass of the i-th class, H_i $(i = 1, \ldots, C)$ denotes the number of subclasses in the i-th class. Upon this assumption, Subclass Discriminant Analysis Score(SDAS) is proposed as:wher-

$$SDAS(F_m) = \frac{\sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{k=i+1}^{C}\sum_{l=1}^{H_k} p_{ij}p_{kl}\left(\mu_{ij}(F_m)-\mu_{kl}(F_m)\right)\left(\mu_{ij}(F_m)-\mu_{kl}(F_m)\right)}{\sum_{i=1}^{C}\sum_{j=1}^{H_i} p_{ij}\sigma_{ij}(F_m)} \tag{2}$$

ep_{ij} are the prior of the j-th subclass of the i-th class, $\mu_{ij}(F_m)$ and $\sigma_{ij}(F_m)$ are the mean and the variance of the j-th subclass of the i-th class for the m-th feature F_m. The numerator of the Eq. (2) emphasizes the scatter between subclasses of different classes, and the denominator of the Eq. (2) emphasizes the scatter within the subclasses. Thus, the more discriminative this feature is, the larger SDAS of this feature.

For MWMR, we utilize SDAS and Pearson Correlation Coefficient (PCC) [31] to estimate the weights and correlations of features, which we call MWMR(SDAS + PCC). Thus, according to the analysis about the MWMR and SDAS, the optimal feature subset Q selected by the first layer of our hierarchical model has two advantages. Firstly, through taking advantage of MWMR, the redundancy of features in Q can be effectively removed. Secondly, owing to the employment of SDAS for feature importance estimation, the features in Q could well preserve the separability of the subclasses in different classes and make the samples within the same subclass more compact.

2.2.2. The second layer of the Hierarchical Feature Selection Model

In the first layer of the proposed feature selection model, we have obtained the optimal feature subset Q which could distinguish (6,8)-ring-containing AlPOs from non-(6,8)-ring-containing AlPOs. Nevertheless, though the feature subset Q are the most important for the formation of (6,8)-ring-containing AlPOs, we cannot figure out which features in it are crucial for the subclass AEN, AWO, CHA and ERI. Therefore, in the second layer of HFSM, we will further analyze which features in Q are the most important for the synthesis of these four subclasses. The feature selection process in the first layer has fully considered the correlation among the selected features, so the correlations in Q are relatively small. Thus, in the second layer, we just use Gini Score [32], a simple and efficient feature selection method, to mining which features in Q affect most for the synthesis of AEN, AWO, CHA and ERI respectively.

Gini Score is a feature selection method based on Gini Index, which could measure the impurity level of the classes in the data set. The Gini index of the original data set is defined as:

$$GiniIndex(X) = 1 - \sum_{i=1}^{C} p_i^2 \tag{3}$$

where p_i is the probability of the sample belonging to the i-th class. Suppose n_i is the number of the sample belonging to the i-th class, then$p_i = n_i/n$. When all samples in the data set belong to the same class, the impurity level of this data set is taken the minimum value 0. If the data set is divided into C subsets $X_i'(i = 1, \ldots C)$ by different values of the m-th feature F_m, then total impurity level of the subsets is:

$$\sum_{i=1}^{C}\frac{n_i'}{n}GiniIndex(X_i') \tag{4}$$

where n_i' is the sample number of X_i'. The Gini score of F_m is the minimum total impurity level of the subsets when dividing the data set into C subsets by all the values of the m-th feature. The smaller the Gini score, the better the feature.

Finally, the complete-process of the proposed hierarchical model is summarized in Algorithm 1.

Algorithm 1. The hierarchical model for subclass synthesis parameter analysis
Input: Original feature set F.
Output: The optimal feature subset Q_{AEN}, Q_{AWO}, Q_{CHA} and Q_{ERI}.

1. Compute the weight W of the original feature set F by SDAS;
2. Compute the correlation matrix R of the original feature set F by PCC;
3. Select the optimal feature subset Q from the original feature set F using MWMR(SDAS + PCC);
4. Select $Q_{subclass}$ in Q using Gini Score.

Table 2
Confusion matrix.

Hypothesis	Actual positive	Actual negative
Hypothesis positive	True positive (*TP*)	False positive (*FP*)
Hypothesis negative	False negative (*FN*)	True negative (*TN*)

3. Experiments and discussions

3.1. Performance measures

In order to evaluate the performance of the proposed HFSM, we adopt two measures to compare the classification result of the features selected by our model and other methods in this paper.

The AlPOs which contain (6,8)-ring are deemed as positive samples, and the AlPOs which do not contain (6,8)-ring are deemed as negative samples. Suppose $n+$ and $n-$ are the numbers of positive samples and negative samples. With reference to the confusion matrix [33] in Table 2, the classification accuracy rate can be denoted as:

$$Acc - Rate = \frac{TP + TN}{n_+ + n_-} \tag{5}$$

During the feature selection in the two hierarchies, the numbers of positive and negative samples are imbalanced. So we also utilize the *F-measure* to evaluate the performances of our algorithm, which is denoted as:

$$F - measure = \frac{\left(1 + \beta^2\right) recall \times precision}{\beta \times recall + precision} \tag{6}$$

where $recall{=}TP/TP + FN$, $precision{=}TP/TP + FP$ andβ is a parameter to adjust the relative importance degree between recall and precision. In this study we set β as 1. The value of *F-measure* lies between 0 and 1, with value closer to 1 indicating better performance for imbalanced problem.

3.2. Synthetic parameter analysis of the (6,8)-ring-containing AlPOs subclass

In this section, we will validate the performance of the feature selection model by comparing the feature selection result in first layer and the second layer respectively. In the experiments, Nearest Neighbor (NN) classifier [34] and Naive Bayes (NB) classifier [35] are employed as prediction models for their advantage of simplicity.

3.2.1. The first layer of the model

We will apply MWMR(SDAS + PCC) to find the most distinguishable feature subset which could separate (6,8)-ring-containing AlPOs from non-(6,8)-ring-containing AlPOs. For comparison, we use MWMR(InforGain + PCC) and MWMR(Fisher + PCC) which represent the MWMR applying InforGain [36] and Fisher score to measure the importance of feature. In order to validate the effectiveness of the algorithm, we use 10-fold cross validation in the experiments. The *Acc-Rate* of the algorithms under various feature dimensions can be seen in Fig. 3, and feature selection results are given in Tables 3 and Table 4.

From Fig. 3, Tables 3 and 4, we can find through taking the subclass into consideration by utilizing SDAS to measure the importance of feature, MWMR (SDAS + PCC) is superior to MWMR (InforGain + PCC) and MWMR(Fisher + PCC) in most cases. Moreover, the feature subset with 9 features selected by MWMR(SDAS + PCC) obtains the highest *Acc-Rate* and *F-measure* with NN. It suggests that the features selected in the previous 9 dimensions by MWMR(SDAS + PCC) may take significant information for the formation of (6,8)-ring-containing AlPOs. In accordance with the result, the optimal feature subset Q is {F5, F9, F11, F12, F13, F14, F15, F17, F20}.

3.2.2. The second layer of the model

Next, we will apply Gini Score to find which features are more important to a special type of subclass (AEN, AWO, CHA and ERI) from the selected feature subset Q obtained in the first layer. We also apply *T*-test [37], Constraint Score [38] and Laplacian Score [39] in the experiments as comparison. We use 5-fold cross validation in the experiments to validate the effectiveness of the algorithm.

The synthetic parameter analysis of AEN

The *Acc-Rate* of different methods under various feature dimensions for the subclass AEN can be seen in Fig. 4, and the highest *Acc-Rate* and *F-measure* obtained by different feature selection methods are given in Table 5 and Table 6.

From the result, we could find that for the feature selection of subclass AEN, Gini Score obtains the highest *Acc-Rate* as 94.28% and *F-measure* as 0.7079 with Naïve Bayes when 3 features are selected. According to the result, the optimal feature subset for the formation of AEN is {F11, F14, F20}. For AEN with two 8-rings windows, the organic templates are usually located in the two-dimensional channels. The optimal feature subset for AEN includes 2 geometrical properties which represents that the size of template directly impact on the size of the framework. The feature selection result also includes F20 which implies that template features impact on the charge of framework to some extent.

The synthetic parameter analysis of AWO

For the subclass AWO, The *Acc-Rate* of different methods under various feature dimensions can be seen in Fig. 5, and the highest *Acc-Rate* and *F-measure* are given in Table 7 and Table 8.

From the result above, we can find that Gini Score obtains the highest *Acc-Rate* as 94.03% and *F-measure* as 0.8690 with Nearest Neighbor Classifier when 2 features are selected for the feature selection of subclass AWO. According to the result, the optimal feature subset for the formation of AWO is {F12, F17}. For AWO, the organic templates are usually located in the one-dimensional channel. Therefore one-dimensional channel may be more sensitive to the second longest distance of organic template. As the feature selection result shows the protonation ability of N of the template is also important for the formation of AWO.

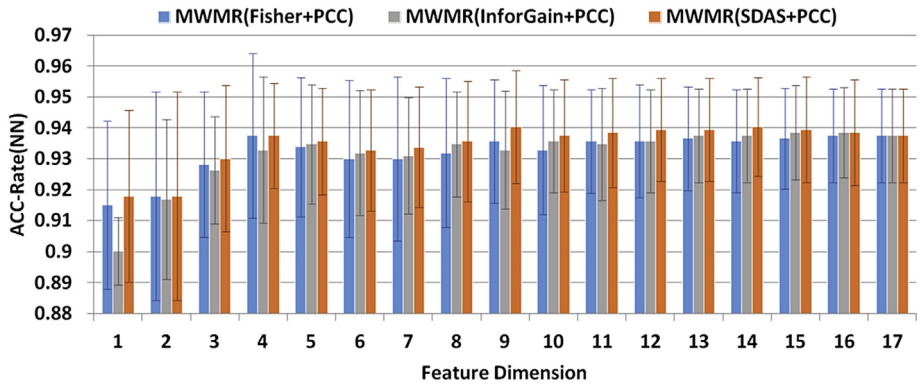
The synthetic parameter analysis of CHA

For the subclass CHA, the *Acc-Rate* of different feature selection methods under every dimension can be seen in Fig. 6, and the highest *Acc-Rate* and *F-measure* are given in Table 9 and Table 10.

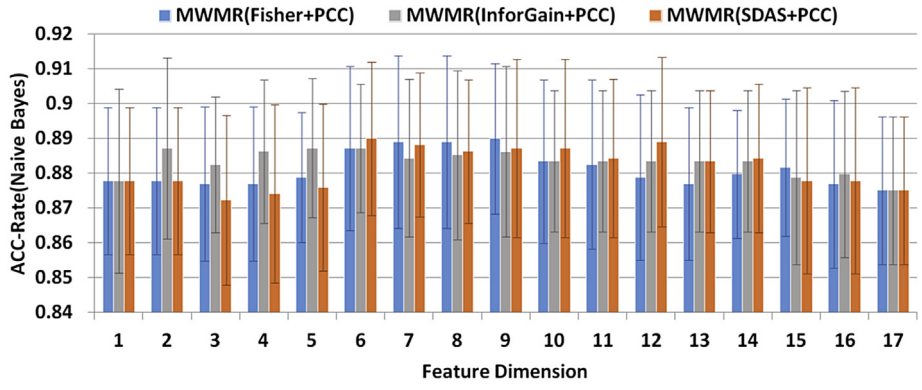
For the feature selection of subclass CHA, Gini Score obtains the highest *Acc-Rate* as 91.27% and *F-measure* as 0.8280 with Nearest Neighbor Classifier when 5 features are selected. According to the result, the optimal feature subset for the formation of CHA is {F20, F14, F12, F13, F11}. For CHA with a cha cage structure, the organic templates are usually located in the three-dimensional channels or cha cage. Since CHA contains a cha cage structure, all the geometry properties are included in this optimal feature subset. Cage structure plays a role in accommodating the organic template, therefore the number of free rotated sing bond is selected.

The synthetic parameter analysis of ERI

The feature selection result for the subclass ERI can be seen in Fig. 7, Table 11 and Table 12. For the subclass ERI, Constraint Score,



(a)



(b)

Fig. 3. The *Acc-Rate* obtained by MWMR with different weight measurements by (a) using NN as classifier and (b) using Naïve Bayes as classifier.

Table 3
The highest *Acc-Rate* (%) and the corresponding dimension obtained by the methods.

	NN		NB	
	Acc-Rate	*Dimension*	*Acc-Rate*	*Dimension*
MWMR (InfoGain + PCC)	93.85 (±0.023)	15	88.71 (±0.034)	6
MWMR (Fisher + PCC)	93.75 (±0.071)	4	**88.99** (±0.047)	9
MWMR (SDAS + PCC)	**94.03** (±0.033)	9	**88.99** (±0.049)	6

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 4
The highest *F-measure* obtained by optimal feature subset.

	NN		NB	
	F-measure	*Dimension*	*F-measure*	*Dimension*
MWMR (InfoGain + PCC)	0.8535 (±0.0011)	15	0.7108 (±0.0025)	6
MWMR (Fisher + PCC)	0.8516 (±0.0011)	4	**0.7155** (±0.0060)	9
MWMR (SDAS + PCC)	**0.8550** (±0.0017)	9	0.7075 (±0.0058)	6

The highest Acc-Rate and highest F-measure are highlighted in bold.

Laplacian Score and Gini Score obtain the highest Acc-Rate as 92.20% and F-measure as 0.6358 with Nearest Neighbor Classifier. However from the performance of these four methods of all above experiments and the experiments in this section, Gini Score is relatively more outstanding and stable. Therefore, we still use the result of Gini Score as the final result here. Based on the result of Gini Score, the optimal feature subset for the formation of ERI is {F11, F14, F20, F15, F17}. For this subclass, the organic templates are usually located in the three-dimensional channels. According to the outcome, the longest distance, the Van der Waals volume, the flexibility, the dipole moment, and the charge density parameters of template are important for its formation.

3.3. Compare with the previous work

In Ref. [23], some molecular engineering researchers have discussed the synthetic parameters of subclasses in the (6,8)-ring-containing AlPOs from the perspective of empirical knowledge.

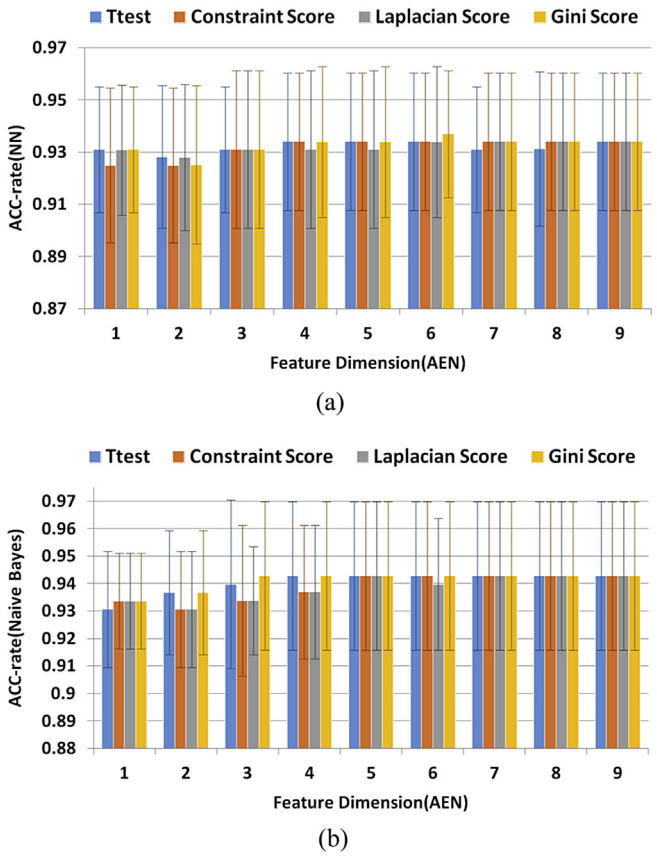


Fig. 4. The *Acc-Rate* obtained by different methods for the subclass AEN (a) using NN as classifier, (b) using Naïve Bayes as classifier.

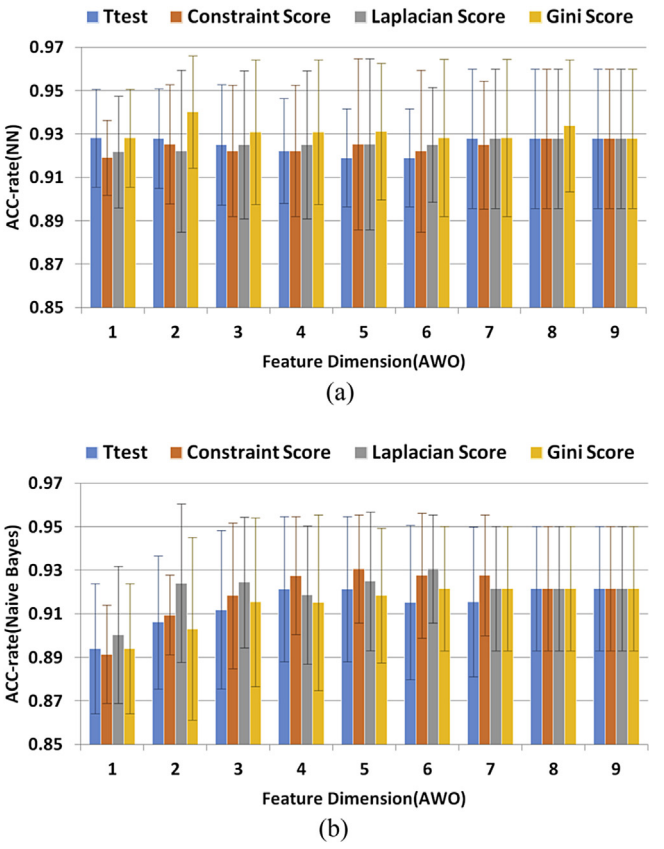


Fig. 5. The *Acc-Rate* obtained by different methods for the subclass AWO (a) using NN as classifier, (b) using Naïve Bayes as classifier.

Table 5
The highest *Acc-Rate* (%) obtained by different methods for subclass AEN.

	NN		NB	
	Acc-Rate	*Dimension*	*Acc-Rate*	*Dimension*
T-test	93.40 (±0.069)	4	**94.28** (±0.073)	4
Constraint Score	93.40 (±0.069)	4	**94.28** (±0.073)	5
Laplacian Score	9340 (±0.069)	7	**94.28** (±0.073)	5
Gini Score	**93.69** (±0.059)	**6**	**94.28** (±0.073)	**3**

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 7
The highest *Acc-Rate* (%) obtained by different methods for subclass AWO.

	NN		NB	
	Acc-Rate	*Dimension*	*Acc-Rate*	*Dimension*
T-test	92.82 (±0.051)	1	92.16 (±0.081)	8
Constraint Score	92.79 (±0.100)	8	93.06 (±0.061)	5
Laplacian Score	92.79 (±0.100)	7	**93.07** (±0.061)	6
Gini Score	**94.03** (±0.067)	**2**	92.16 (±0.081)	6

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 6
The highest *F-measure* obtained by different methods for subclass AEN.

	NN		NB	
	F-measure	*Dimension*	*F-measure*	*Dimension*
T-test	0.6248 (±0.0319)	4	**0.7079** (±0.0123)	4
Constraint Score	0.6181 (±0.0308)	4	**0.7079** (±0.0123)	5
Laplacian Score	0.6244 (±0.0308)	7	**0.7079** (±0.0123)	5
Gini Score	**0.6321** (±0.0335)	**6**	**0.7079** (±0.0123)	**3**

The highest Acc-Rate and highest F-measure are highlighted in bold.

Thus, we will compare our work based on Hierarchical Feature Selection Model (HFSM) with [23] in this section. The optimal feature subsets for the four subclasses obtained by Ref. [23] and our model are listed in Table 13. The *Acc-Rate* and *F-measure* obtained by the optimal feature subsets of the four subclasses selected by the two works are given in Table 14 and Table 15.

The optimal features for the formation of the subclasses of (6,8)-ring-containing AlPOs in Ref. [23] are deduced by molecular engineering researchers' professional knowledge and experience. So as we can see from Table 13, the number of features in their optimal feature subset is smaller than the number of features in optimal feature subset in this paper. However, the feature subsets obtained in this paper almost cover the optimal feature subsets deduced Ref. [23]. On the other hand [23], has not fully mined the data and the analyzing methods, thus their conclusions are not comprehensive. We can see from Tables 14 and 15 that the optimal

Table 8
The highest *F-measure* obtained by different methods for subclass AWO.

	NN		NB	
	F-measure	*Dimension*	*F-measure*	*Dimension*
T-test	0.8414 (±0.0019)	1	0.8268 (±0.0047)	8
Constraint Score	0.8414 (±0.0034)	8	**0.8460** (±0.0031)	5
Laplacian Score	0.8414 (±0.0034)	7	**0.8460** (±0.0031)	6
Gini Score	**0.8690** (±0.0028)	**2**	0.8264 (±0.0047)	6

The highest Acc-Rate and highest F-measure are highlighted in bold.

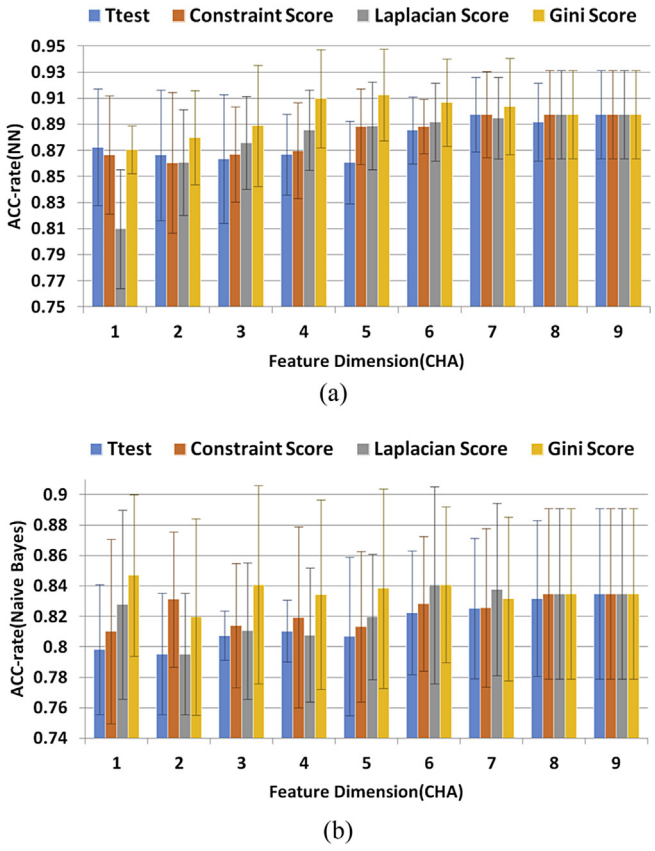


Fig. 6. The *Acc-Rate* obtained by different methods for the subclass CHA (a) using NN as classifier, (b) using Naïve Bayes as classifier.

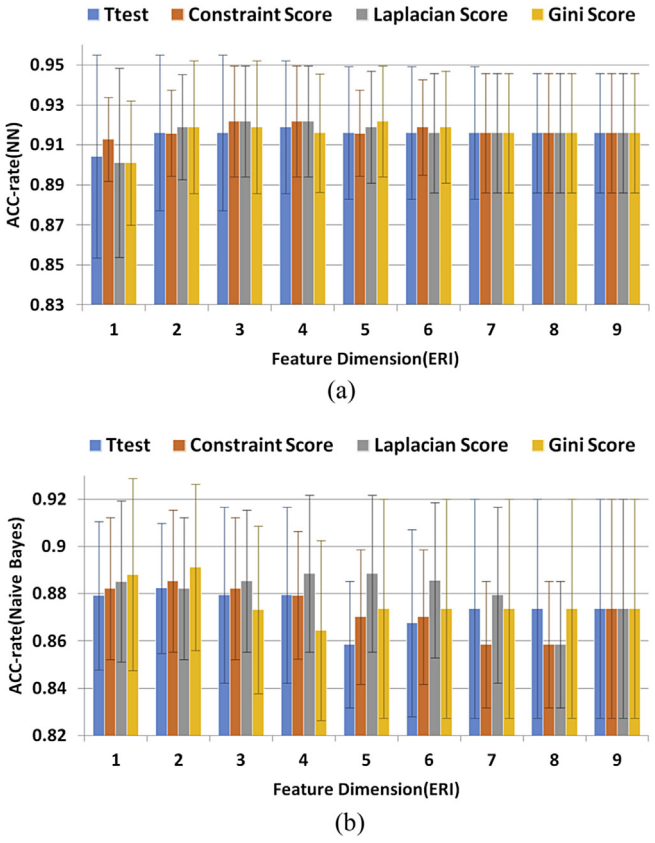


Fig. 7. The *Acc-Rate* obtained by different methods for the subclass ERI (a) using NN as classifier, (b) using Naïve Bayes as classifier.

Table 9
The highest *Acc-Rate* (%) obtained by different methods for subclass CHA.

	NN		NB	
	Acc-Rate	*Dimension*	*Acc-Rate*	*Dimension*
T-test	89.76 (±0.11)	9	83.49 (±0.31)	9
Constraint Score	89.76 (±0.11)	8	83.49 (±0.31)	8
Laplacian Score	89.76 (±0.11)	8	84.06 (±0.42)	6
Gini Score	**91.27** (±0.12)	5	**84.70** (±0.28)	1

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 11
The highest *Acc-Rate* (%) obtained by different methods for subclass ERI.

	NN		NB	
	Acc-Rate	*Dimension*	*Acc-Rate*	*Dimension*
T-test	91.90 (±0.11)	4	88.24 (±0.076)	2
Constraint Score	**92.20** (±0.077)	3	88.54 (±0.091)	2
Laplacian Score	**92.20** (±0.077)	3	88.86 (±0.11)	4
Gini Score	**92.20** (±0.077)	5	**89.13** (±0.12)	2

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 10
The highest *F-measure* obtained by different methods for subclass CHA.

	NN		NB	
	F-measure	*Dimension*	*F-measure*	*Dimension*
T-test	0.8059 (±0.0066)	9	0.7273 (±0.0140)	9
Constraint Score	0.8080 (±0.0066)	8	0.7273 (±0.0140)	8
Laplacian Score	0.8006 (±0.0066)	8	0.7312 (±0.0154)	6
Gini Score	**0.8280** (±0.0069)	5	**0.7328** (±0.0131)	1

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 12
The highest *F-measure* obtained by different methods for subclass ERI.

	NN		NB	
	F-measure	*Dimension*	*F-measure*	*Dimension*
T-test	0.6307 (±0.0406)	4	**0.4984** (±0.0465)	2
Constraint Score	**0.6358** (±0.0366)	3	**0.4984** (±0.0384)	2
Laplacian Score	**0.6358** (±0.0366)	3	**0.4984** (±0.0675)	4
Gini Score	**0.6358** (±0.0366)	5	**0.4984** (±0.0247)	2

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 13
The optimal feature subset comparisons.

	HFSM		Gao N. [23](2014)	
	The selected feature	*Dimension*	The selected features	*Dimension*
AEN	F2,F3,F4, F11,F14,F20	6	F11	1
AWO	F2,F3,F4, F12,F17	5	F3, F14	2
CHA	F2,F3,F4, F20,F14,F12,F13,F11	8	F3, F4, F20	3
ERI	F2,F3,F4, F11,F14,F20,F15,F17	8	F2, F3, F4, F14	4

Table 14
The highest *Acc-Rate* (%) and variance (%) obtained by the methods.

		NN	NB
AEN	Gao N. [23] (2014)	93.39 (±0.083)	**94.28** (±0.073)
	HFSM	**93.69** (±0.059)	**94.28** (±0.073)
AWO	Gao N. [23] (2014)	93.12 (±0.12)	91.56 (±0.062)
	HFSM	**94.03** (±0.067)	**92.16** (±0.081)
CHA	Gao N. [23] (2014)	86.41 (±0.13)	83.80 (±0.77)
	HFSM	**91.27** (±0.12)	**84.70** (±0.28)
ERI	Gao N. [23] (2014)	91.61 (±0.15)	88.22 (±0.13)
	HFSM	**92.20** (±0.077)	**89.13** (±0.12)

The highest Acc-Rate and highest F-measure are highlighted in bold.

Table 15
The highest *F-measure* (variance) obtained by optimal feature subset.

		NN	NB
AEN	Gao N. [23] (2014)	**0.6807** (±0.0099)	**0.7079** (±0.0123)
	HFSM	0.6321 (±0.0335)	**0.7079** (±0.0123)
AWO	Gao N. [23] (2014)	0.8565 (±0.0050)	0.8055 (±0.0039)
	HFSM	**0.8690** (±0.0028)	**0.8264** (±0.0047)
CHA	Gao N. [23] (2014)	0.7356 (±0.0039)	0.7065 (±0.0235)
	HFSM	**0.8280** (±0.0069)	**0.7328** (±0.0131)
ERI	Gao N. [23] (2014)	0.6263 (±0.0443)	0.2771 (±0.0710)
	HFSM	**0.6358** (±0.0366)	**0.4984** (±0.0247)

The highest Acc-Rate and highest F-measure are highlighted in bold.

feature subsets obtained by our work could reach higher *Acc-Rate* and *F-measure* than [23] in most instances.

4. Conclusions and future works

In this study, a Hierarchical Feature Selection Model composed of two layers is proposed. By our model, the synthetic parameters for the rational synthesis of the subclass of (6,8)-ring-containing AlPOs are analyzed during the feature selection process. Comprehensive experiments and extensive analysis are carried out to demonstrate the effectiveness of the proposed model. Taking advantage of this model, the optimal feature subsets for the formation of AEN, AWO, ERI and CHA are given, which could provide a useful guidance for rational synthesis for such materials.

In our future works, we would take account of more synthetic parameters in our feature selection model when the microporous AlPOs database is updated and refined. Furthermore, we will also try to synthesize some AlPO materials having new structures by the parameter analysis results obtained in our feature selection method.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61403078), the Fundamental Research Funds for the Central Universities (No. 2412016KJ035).

References

[1] H. Lee, S. Zones, M. Davis, Nature 425 (2003) 385–388.
[2] J.H. Yu, R.R. Xu, Chem. Soc. Rev. 25 (2006) 593–604.
[3] Y. Li, J.H. Yu, J.X. Jiang, Z.P. Wang, J.N. Zhang, R.R. Xu, Chem. Mater. 17 (2005) 6086–6093.

[4] J.Y. Li, J.H. Yu, W.F. Yan, Y.H. Xu, W.G. Xu, S.L. Qiu, R.R. Xu, Chem. Mater. 11 (1999) 2600–2606.
[5] J.H. Yu, J.Y. Li, K.X. Wang, R.R. Xu, K. Sugiyama, O. Terasaki, Chem. Mater. 12 (2000) 3783–3787.
[6] D. Zhou, J. Xu, J.H. Yu, L. Chen, F. Deng, R.R. Xu, J. Phys. Phys. Chem. B 110 (2006) 2131–2137.
[7] H.Z. Xing, J.Y. Li, W.F. Yan, P. Chen, Z. Jin, J.H. Yu, S. Dai, R.R. Xu, Chem. Mater. 20 (2008) 4179–4181.
[8] J.Y. Li, J.H. Yu, R.R. Xu, http://zeobank.jlu.edu.cn.
[9] D.A. Fletcher, R.F. McMeeking, D. Parkin, J. Chem. Inf. Comput. Sci. 36 (1996) 746–749.
[10] E. Perola, K. Xu, T.M. Kollmeyer, S.H. Kaufmann, F.G. Prendergast, Y.P. Pang, J. Med. Chem. 43 (2000) 401–408.
[11] S. Goto, T. Nishioka, M. Kanehisa, Bioinformatics 14 (1998) 591–599.
[12] C.A. Bennett, N.L. Franklin, Statistical Analysis in Chemistry and the Chemical Industry, John Wiley & Sons, New York, 1954.
[13] F.H. Allen, Acta Crystallogr. Sect. B 58 (2002) 380–388.
[14] W.D. Kraeft, D. Kremp, W. Ebeling, G. Röpke, Quantum Statistics of Charged Particle Systems, Akademie Verlag, Berlin, 1986.
[15] S. Maria, E. Lennart, J. Jörgen, M. Sjöström, S. Wold, J. Med. Chem. 41 (1998) 2481–2491.
[16] M. Karelson, V.S. Lobanov, A.R. Katritzky, Chem. Reviers 96 (1996) 1027–1044.
[17] Karelson Mati, Molecular Descriptors in QSAR/QSPR, John Wiley & Sons, New York, 2000.
[18] J.Y. Li, M. Qi, J. Kong, J.Z. Wang, Y. Yan, W.F. Huo, J.H. Yu, R.R. Xu, Y. Xu, Microporous Mesoporous Mater. 129 (2010) 251–255.
[19] M.H. Yao, M. Qi, J.S. Li, J. Kong, Microporous Mesoporous Mater. 186 (2014) 201–206.
[20] R.A. Fisher, Ann. Eugen. 7 (1936) 179–188.
[21] R. Steuer, J. Kurths, C.O. Daub, J. Weise, J. Selbig, Bioinformatics 18 (2002) S231–S240.
[22] M. Qi, J.S. Li, J.Z. Wang, Y.H. Lu, J. Kong, Ind. Eng. Chem. Res. 51 (2012) 16734–16740.
[23] N. Gao, Y. Yan, J.S. Li, J.Y. Li, Microporous Mesoporous Mater. 195 (2014) 174–179.
[24] Y. Saeys, I. Inza, P. Larranaga, Bioinformatics 23 (2007) 2507–2517.
[25] D.W. Lewis, G. Sankar, J.K. Wyles, J.M. Thomas, C.R.A. Catlow, D.J. Willock, Angewandte Chemie Int. Ed. Engl. 109 (1997) 2791–2793.
[26] J. Liang, H.Y. Li, S. Zhao, W.G. Guo, R.H. Wang, M.L. Ying, Appl. Catal. 64 (1990) 31–40.
[27] A.K. Jain, B. Chandrasekaran, Handbook of Statistices, vol. 2, 1982, pp. 835–855.
[28] P. Cunningham, Machine Learning Techniques for Multimedia, Springer, Berlin Heidelberg, 2008, pp. 91–112.
[29] J.Z. Wang, L.S. Wu, J. Kong, Y.X. Li, B.X. Zhang, Pattern Recognit. 46 (2013) 1616–1627.
[30] M. Zhu, A.M. Martinez, IEEE Trans. Pattern Analysis Mach. Intell. 28 (2006) 1274–1286.
[31] L.J. van't Veer, H.Y. Dai, M.J. van de Vijver, et al., Nature 415 (2002) 530–536.
[32] C. Gini, Variabilità e mutabilita, Bologna, 1912.
[33] P. Soda, Pattern Recognit. 44 (2011) 1801–1810.
[34] T.M. Cover, P.E. Hart, IEEE Trans. Inf. Theory 13 (1967) 21–27.
[35] I. Rish, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, IBM, New York, 2001, pp. 41–46.
[36] H. Liu, J. Li, L. Wong, Genome Inf. Ser. 23 (2002) 51–60.
[37] W.H. Press, S.A. Teukolsky, W.T. Vetterling, et al., Numerical Recipes in C (2nd ed.): the Art of Scientific Computing, Cambridge University Press, New York, 1992, p. 616.
[38] D.Q. Zhang, S.C. Chen, Z.H. Zhou, Pattern Recognit. 41 (2008) 1440–1451.
[39] X.F. He, C. Deng, N. Partha, Adv. Neural Inf. Process. Syst. (2005) 507–514.
[40] Y.T. Guo, J.Z. Wang, N. Gao, M. Qi, M. Zhang, J. Kong, Y.H. Lv, Int. J. Mol. Sci. 14 (2013) 22132–22148.