

Article

## AIPOs Synthetic Factor Analysis Based on Maximum Weight and Minimum Redundancy Feature Selection

Yuting Guo <sup>1,2,3</sup>, Jianzhong Wang <sup>1,3,\*</sup>, Na Gao <sup>4</sup>, Miao Qi <sup>3</sup>, Ming Zhang <sup>1,3,\*</sup>, Jun Kong <sup>1,3</sup> and Yinghua Lv <sup>2,\*</sup>

<sup>1</sup> College of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, Jilin, China; E-Mails: guoyt484@nenu.edu.cn (Y.G.); kongjun@nenu.edu.cn (J.K.)

<sup>2</sup> Faculty of Chemistry, Northeast Normal University, Changchun 130024, Jilin, China

<sup>3</sup> Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130117, Jilin, China; E-Mail: qim801@nenu.edu.cn

<sup>4</sup> State Key Laboratory of Inorganic Synthesis and Preparative Chemistry, Changchun 130012, Jilin, China; E-Mail: gaona0431@163.com

\* Authors to whom correspondence should be addressed; E-Mails: wangjz019@nenu.edu.cn (J.W.); zhangm545@nenu.edu.cn (M.Z.); luyh@nenu.edu.cn (Y.L.); Tel./Fax: +86-431-8453-6326 (J.W.).

*Received: 25 September 2013; in revised form: 23 October 2013 / Accepted: 23 October 2013 /*

*Published: 8 November 2013*

---

**Abstract:** The relationship between synthetic factors and the resulting structures is critical for rational synthesis of zeolites and related microporous materials. In this paper, we develop a new feature selection method for synthetic factor analysis of (6,12)-ring-containing microporous aluminophosphates (AIPOs). The proposed method is based on a maximum weight and minimum redundancy criterion. With the proposed method, we can select the feature subset in which the features are most relevant to the synthetic structure while the redundancy among these selected features is minimal. Based on the database of AIPO synthesis, we use (6,12)-ring-containing AIPOs as the target class and incorporate 21 synthetic factors including gel composition, solvent and organic template to predict the formation of (6,12)-ring-containing microporous aluminophosphates (AIPOs). From these 21 features, 12 selected features are deemed as the optimized features to distinguish (6,12)-ring-containing AIPOs from other AIPOs without such rings. The prediction model achieves a classification accuracy rate of 91.12% using the optimal feature subset. Comprehensive experiments demonstrate the effectiveness of the proposed algorithm, and deep analysis is given for the synthetic factors selected by the proposed method.

**Keywords:** AlPOs; data mining; feature selection; rational synthesis

---

## 1. Introduction

As an important class of crystalline materials, zeolites and related microporous materials have been widely used in the petroleum industry for catalysis, separation and ion-exchange [1,2]. Following the discovery of the aluminophosphate molecular sieves  $\text{AlPO}_4\text{-}n$  ( $n$  denotes the structure type) in 1982, a large variety of open-framework aluminophosphates with different structure types have been synthesized and open-framework aluminophosphate materials has become an important member of the porous crystal material family. Recently, the rational synthesis of microporous inorganic materials has attracted extensive attention [3–10]. However, since the synthesis of such materials is typically carried out in a gel medium under hydrothermal/solvothermal conditions by using alkali metal ions or organic amines/ammoniums as the templates or structure-directing agents (SDAs) [11], it is very complicated and influenced by many factors. Therefore, in order to provide guidance to rational synthesis of microporous inorganic materials, the researchers of State Key Laboratory of Inorganic Synthesis and Preparative Chemistry of Jilin University established an international ALPO synthesis database [12] based on a large number of synthesis experiments and collections from the papers.

With the rapid development of computer technology and artificial intelligence, data mining plays an increasingly important role in more and more research areas. The goal of data mining is to find the implied knowledge from the given data. The applications of data mining techniques in chemical science have shown their feasibility for numeric calculation, simulation and data analysis. Nowadays, one of the most widely used data mining techniques in chemical science is feature selection. Feature selection is usually used as a preprocessing step in machine learning that can select the most important features for particular tasks by seeking the potential information hidden in the data. Recently, several feature selection methods were successfully applied in chemical data analysis. Pichler [13] developed an interactive feature selection method based on KNN (K Nearest Neighbor) to classify doublet/singlet patterns from the same Stationary Electrode Polarography (SEP) data. Liu evaluated the performance of the methods as Information Gain, Mutual Information,  $\chi^2$ -Test (CHI), Odds Ratio (OR) and GSS Coefficient (GSS) for finding the optimal feature subset in drug discovery; the features were firstly ranked according to the scores obtained by different feature selection methods and then the top-ranking features were used for classification task [14]. Teramoto and Fukunishi proposed a supervised consensus scoring (SCS) method for docking and virtual. In SCS, a series of scoring functions including PLP,  $F$ -Score, LigScore, DrugScore, LUDI,  $X$ -Score, AutoDock, PMF,  $G$ -Score, ChemScore and  $D$ -Score were integrated to form a complementary scoring function, which could compensate for the deficiencies of each scoring method [15]. In addition, a Mutual Information Gain algorithm was utilized to generate a feature subset which excluded features having weak correlation with the target variable, and then the selected features were input into a Genetic Programming model to analyze QSAR (Quantitative Structure Activity Relationship) data [16]. In a further study, 649 bitter and 13,530 randomly selected molecules from the MDL Drug Data Repository (MDDR) were analyzed by Information Gain, and the selected features were then classified by Naive Bayes classifier to identify

the bitterness of small molecules [17]. Feature selection methods also have been applied to AIPOs database analysis. Li *et al.* evaluated the classification performance produced by different combinations of synthetic features (11 features in total) using Support Vector Machines (SVM), and then checked which individual or combined features effected most for distinguishing the two classes of AIPOs. They found that suitable template parameters were of vital importance to the classification performance [18]. Huo *et al.* [19], measured the importance of the various synthetic features (26 features in total) of AIPOs by *F*-Score [20], and sorted the features in descending order according to their importance degree. The features were then added into Decision Tree (DT) model orderly to test their discriminative abilities. They regarded the feature subset that could reach the best classification performance as the optimal subset. Through their experiments, they found that T1\_Distance2 (the second longest distance of organic template) was the determinant factor to distinguish AIPO<sub>4-5</sub> from other types of aluminophosphate molecular sieves. Although the pioneering works in [18] and [19] have shown that the feature selection techniques can be applied for AIPOs database analysis effectively, there were also some limitations in them. Firstly, the feature subset evaluated in [18] was generated using an exhaustive searching strategy, which made it hard to be scaled to high-dimensional AIPOs data. Secondly, the optimal feature subsets in both [18] and [19] were evaluated by a specific classifier (DT in [19] and SVM in [18]). Thus, the classifiers need to be trained and tested many times in the feature selection procedure, which made them very time-consuming. Finally, the correlation among the selected features was neglected in both [18] and [19]. Some researchers [21] have pointed out that a good feature subset should be the one that contains features highly correlated with the class, while uncorrelated with each other. Therefore, ignoring the correlation among the selected features might cause the problem of “information redundancy”, which hinders optimal results from the selected features.

In order to overcome the limitations of the previous works, a new feature selection algorithm based on maximum weight and minimum correlation criterion is proposed in this paper. The proposed method not only considers the importance of the feature, but also takes the correlation among the selected features into account. Thus, through the proposed method, we can select the optimal feature subset in which the features are maximally relevant to the synthetic structure while the redundancy among these selected features is minimal. In the experiments, three feature evaluation algorithms (Fisher score, ReliefF score and Gini score) are combined with redundancy measurement method (Pearson correlation coefficient) to test the performance of our method. Compared with other feature selection methods [18,19] for AIPOs database analysis, our method possesses the following advantages. (1) The optimal feature subset generated by our method does not depend on any classifier. Thus, the feature selection procedure does not need to train any classifier, which makes our algorithm more efficient; (2) The feature selection procedure of our algorithm is a pair-wise updating optimization process, so it can be easily scaled to high-dimensional AIPOs data; (3) The proposed method takes the correlation among features into consideration. Thus, it can obtain better results than other state of the art feature selection methods.

## 2. Results and Discussions

In this section, we first compare the performance of the proposed algorithm with other classical scoring feature selection methods that neglect the correlation among features during the feature

selection process. Then, the feature selection results obtained by the proposed algorithm are analyzed and compared with the previous works [18,19]. At last, we also compare the proposed algorithm with several state of the art feature selection approaches, such as Constraint score [22], MRMR [23,24] and FCBF [25].

In the experiments, the Nearest Neighbor and Naive Bayes classifiers are employed as prediction models for their advantage of simplicity. In order to validate the effectiveness of the algorithm comprehensively, we use 10-fold cross validation in the experiments.

### 2.1. Performance Measures

The synthetic records used in the experiments contain 398 (6,12)-ring-containing AIPOs and 852 AIPOs without such rings. For the purpose of distinguishing the (6,12)-ring-containing AIPOs from others, we deem the former as positive samples and the latter as negative samples respectively. It is obvious that the numbers of positive and negative samples are imbalanced in this study. So besides the classification accuracy rate, we also utilize the  $F$ -measure to evaluate the performances of the proposed algorithm.

Suppose  $n_+$  and  $n_-$  are the numbers of positive samples and negative samples. With reference to the confusion matrix [26] in Table 1, the classification accuracy rate (Acc\_Rate) and  $F$ -measure can be denoted as:

$$\text{classification accuracy rate} = \frac{TP + TN}{n_+ + n_-} \quad (1)$$

and

$$F\text{-measure} = \frac{(1 + \beta^2)\text{recall} \times \text{precision}}{\beta \times \text{recall} + \text{precision}} \quad (2)$$

where  $\text{recall} = \frac{TP}{TP + FN}$ ,  $\text{precision} = \frac{TP}{TP + FP}$ , and  $\beta$  is a parameter to adjust the relative importance degree between recall and precision. In this work, we set  $\beta$  as 1. The value of  $F$ -measure lies between 0 and 1, with value closer to 1 indicating better performance for imbalanced problems.

**Table 1.** Confusion matrix.

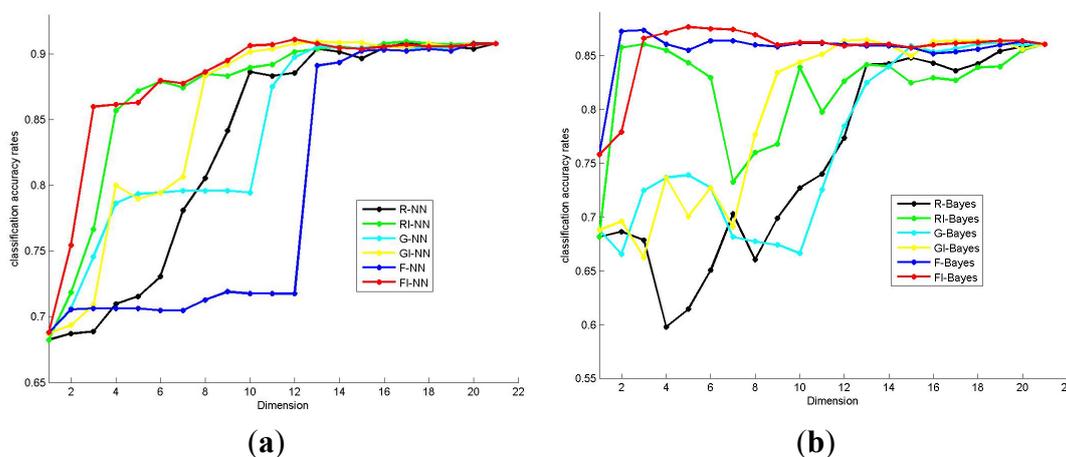
Hypothesis	Actual positive	Actual negative
Hypothesise positive	True positive ( $TP$ )	False positive ( $FP$ )
Hypothesise negative	False negative ( $FN$ )	True negative ( $TN$ )

### 2.2. Effectiveness of the Proposed Method

We will firstly verify the effectiveness of the proposed method by comparing it with some classical scoring feature selection methods without considering the correlation among features. In this experiment, Fisher score [27], ReliefF score [28] and Gini score [29] are applied to measure the importance of the feature, while Pearson Correlation Coefficient (PCC) is applied to measure the correlation among features. The classification accuracy rate of different methods under various feature dimensions can be seen in Figure 1. In this figure, FI (Fisher improve) denotes the proposed method

that utilizes the Fisher score and Pearson Correlation Coefficient to estimate the importance and correlation of features. RI (ReliefF improve) denotes the proposed method that utilizes the ReliefF score and Pearson Correlation Coefficient to estimate the importance and correlation of features. And GI (Gini improve) denotes the proposed method that utilizes the Gini score and Pearson Correlation Coefficient to estimate the importance and correlation of features. *F*, *R* and *G* denote original Fisher score, ReliefF score and Gini score respectively. The best classification accuracy rates obtained by these methods are listed in Table 2.

**Figure 1.** Comparison of the original and proposed feature selection methods. (a) Using Nearest Neighbor as classifier; (b) Using Naive Bayes as classifier.



**Table 2.** Highest classification accuracy rates reached by the original and the proposed feature selection methods.

Method	Nearest Neighbor		Naive Bayes	
	Highest Acc Rate	Dimension	Highest Acc Rate	Dimension
F	0.9080	20	0.8736	3
FI	0.9112	12	0.8767	5
R	0.9088	17	0.8608	21
RI	0.9096	17	0.8608	3
G	0.9080	21	0.8624	19
GI	0.9096	13	0.8648	13

From Figure 1 and Table 2, it can be seen that through taking the correlation among the selected features into consideration, the proposed algorithm can outperform the classical scoring feature selection methods. When the Nearest Neighbor classifier is utilized, the best classification accuracy rates obtained by FI, RI and GI are 91.12%, 90.96% and 90.96% respectively. When the Naive Bayes classifier is employed, the best classification accuracy rates obtained by FI, RI and GI are 87.67%, 86.08% and 86.48%. Moreover, it also should be noted that the dimensions of the optimal feature subset obtained by the proposed algorithm are less than the classical scoring feature selection algorithms in most cases.

In this paper, the numbers of the positive samples and negative samples are imbalanced, so we take the *F*-measure to evaluate the performance of proposed method. As shown in Table 3, the proposed

algorithm is superior to the classical scoring feature selection methods for the class imbalance problem especially when the Fisher score is utilized to evaluate the importance of feature.

**Table 3.** Highest  $F$ -measure reached by the original and the improved feature selection methods.

Method	Highest $F$ -measure (Nearest Neighbor)	Highest $F$ -measure (Naive Bayes)
F	0.8144	0.7817
FI	0.8586	0.8071
R	0.8585	0.7851
RI	0.8599	0.7851
G	0.8518	0.7640
GI	0.8579	0.8003

From above experimental results, we can find that the performance of Fisher score combined with PCC in the proposed algorithm is superior to ReliefF score and Gini score combined with PCC, since the optimal feature subset generated by FI is in a lower dimension and wins the highest classification accuracy rate as well. So in the next experiment, we will focus on analyzing the features selected by FI.

### 2.3. Analysis of the Feature Selection Results

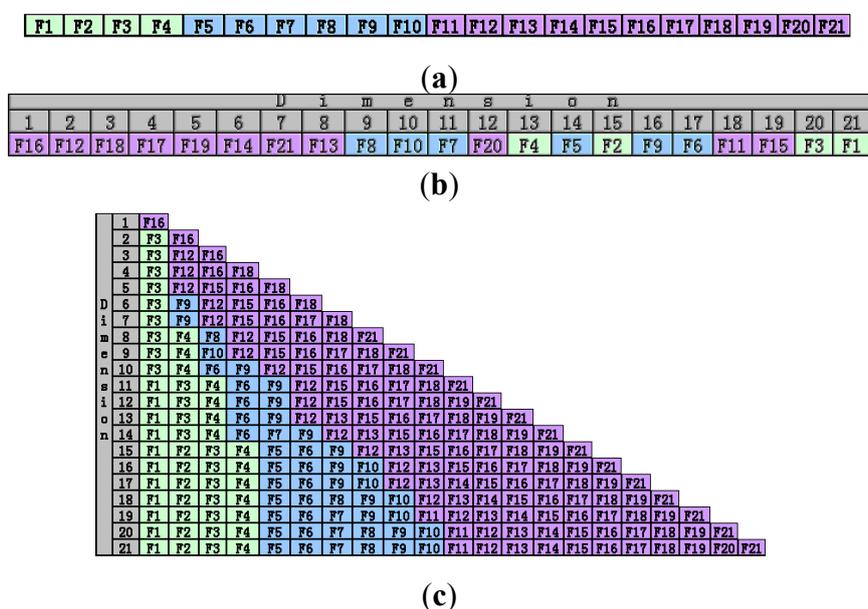
In this part, we make some analysis about the feature selection result obtained by the proposed method (Fisher score combined with PCC) and compare our feature selection result with the previous works [18,19].

Here, let's reconsider the performances of the proposed method (FI-NN) and the original Fisher score (F-NN) in Figure 1a. Firstly, we can find that the classification performances of the proposed method are superior to the original Fisher score under most dimensions. This means that by reducing the redundancy from selected features, the proposed method can select more optimal feature subsets for distinguishing the (6,12)-ring-containing AlPOs. Secondly, it can be observed that when the dimension of the selected features increases from 1 to 12, the classification performance of the FI shows a dramatic upward trend, and reaches its peak point at dimension 12. However when the dimensions of the selected features are larger than 12, the classification performance presents a tiny downward trend. This suggests that the features selected in the previous 12 dimensions may take significant information for separating the (6,12)-ring-containing AlPOs from others. Thirdly, classification accuracy rate sharply increases from about 75.5% to 86.2% when F12 (the second longest distance of organic template) is added to the optimal feature subset at dimension 3. This phenomenon indicates that the second longest distance of organic template is a very crucial factor for the rational synthesis of (6,12)-ring-containing AlPOs.

Figure 2 shows the feature selection results obtained by Fisher score and the proposed method (FI). In this figure, in order to distinguish different types of features more clearly, we assign different colors to different categories of features (as shown in Figure 2a). The features marked with green color belong to gel composition, the features marked with blue color and purple color belong to solvent and organic template. Figure 2b,c illustrate feature subsets selected by the original Fisher score and the proposed algorithm (FI). In Figure 2b, the features are sorted in descending order according to their Fisher scores. Since the features with higher Fisher scores are more important, if we want to obtain a

feature subset that contains  $k$  features, we just need to select the first  $k$  features in the descending sequence mentioned above. Figure 2c demonstrates the features selected by the proposed algorithm (FI) under every dimension (the selected features are sorted in ascending order according to their ID). The features selected by FI under dimension 12 which could lead to the highest classification accuracy rate are F1 (the molar amount of  $\text{Al}_2\text{O}_3$ ), F3 (the molar amount of solvent), F4 (the molar amount of template in the gel composition), F6 (the melting point), F9 (the dipole moment), F12 (the second longest distance of organic template), F15 (the dipole moment), F16 (the ratio of  $C/N$ ), F17 (the ratio of  $N/(C + N)$ ), F18 (the ratio of  $N/\text{Van der Waals volume}$ ) and F21 (the maximal number of protonated H atoms).

**Figure 2.** The features selected by Fisher and Fisher combined with PCC. (a) Different category synthesis factors are represented as different color; (b) features selected by Fisher score; (c) features selected by Fisher score combined with PCC in our algorithm.



There is a remarkable phenomenon in Figure 2b that the first 8 features selected by Fisher score are all marked with purple color, which means they all belong to organic template. Although the organic template factors are significant for AIPOs synthesis, these factors are not sufficient to distinguish AIPOs with different structures effectively. From Figure 1a, we can find that when the first two features F16 (ratio of  $C/N$ ) and F12 (second longest distance of organic template) are selected, the classification accuracy rate of the classifier could reach about 71%. However, after the other 6 template features with higher Fisher score (F18 (the ratio of  $N/\text{Van der Waals volume}$ ), F17 (the ratio of  $N/(C + N)$ ), F19 (the Sanderson electronegativity), F14 (the Van der Waals volume), F21 (the maximal number of protonated H atoms), F13 (the shortest distance of organic template)) are added into the selected feature subset gradually, the classification accuracy rate of the classifier is almost unchanged. Failure of the feature selection described above is caused by information redundancy, or the correlation among the selected features. Since the first 8 features in Figure 2b come from the same category (organic template), they are far from orthogonal and cannot improve the performance of the classification task. In other words, although the first 8 template features in Figure 2b have higher

Fisher score values, selecting them all into the feature subset does not enable the addition of new information into the selected feature subset. This clarifies the importance of accounting for redundancy during the feature selection process.

Li *et al.* found the optimal feature subset was consisted of 8 features that obtained the highest classification accuracy rate of 82.44% by SVM classifier [18]. However, their feature selection process was an exhaustive searching strategy, so the entire process was extraordinarily time consuming. In their study, the correlation among the selected features was not at all mentioned. Thus, the correlations between some of the selected features were very high, for example: the correlation between F7 (boiling point) and F8 (dielectric constant) was 0.8370; the correlation between F7 (boiling point) and F9 (dipole moment) was 0.8306; the correlation between F8 (dielectric constant) and F9 (dipole moment) was 0.9512. Huo *et al.* worked out that a feature subset consisting of 19 features was the best combination for predicting AIPOs, with the highest *AUC* of 90% and the highest classification accuracy rate of 88.18% [19]. Nevertheless, since the correlation among features was also neglected in their study, there were some highly correlated features in their optimal feature set too. For example, the correlation between F8 (dielectric constant) and F10 (polarity) was 0.9849, while the correlation between F14 (Van der Waals volume) and F20 (number of free rotated single bond) was 0.9073.

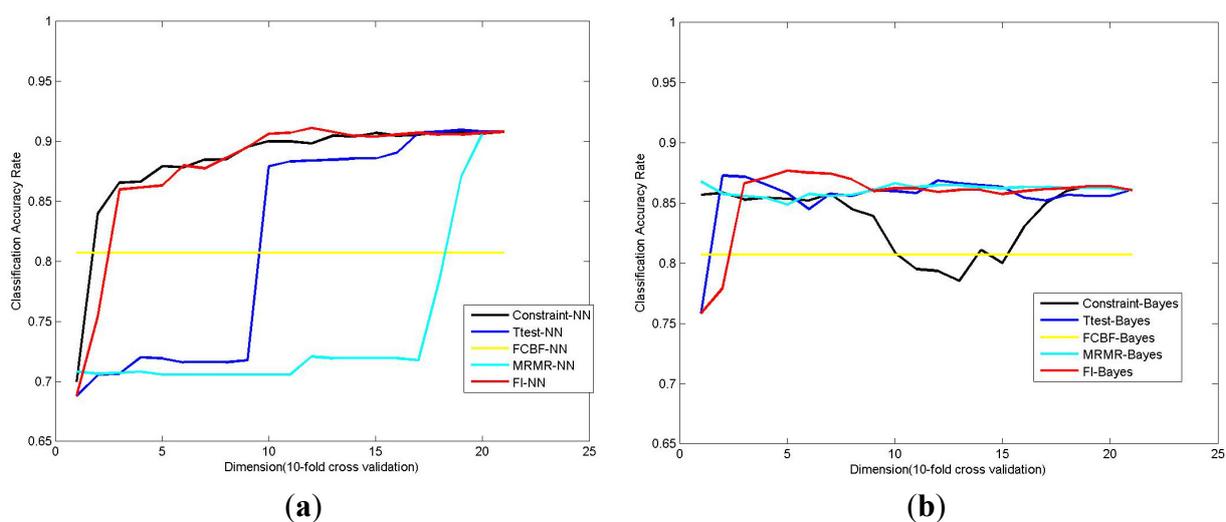
In the proposed method, we take into account the correlation among the selected features in the feature selection process. So, as shown in Figure 1a the classification accuracy rate curve of FI presents a distinctive uptrend before getting to the peak point, and when features belonging to a new category are added to the selected feature set at dimension 2 and 6, the curve appears obviously ascending. In the optimal feature set produced by this study, the molar amount of  $\text{Al}_2\text{O}_3$ , solvent and template are gel composition features; melting point and dipole moment are solvent features; the second longest distance of organic template, the dipole moment, the ratio of  $C/N$ , the ratio of  $N/(C + N)$ , the ratio of  $N/\text{Van der Waals volume}$  and the maximal number of protonated H atoms are organic template features. Since the selected features by the proposed algorithm are comprehensive, we obtain the highest classification accuracy rate as 91.12% using Nearest Neighbor classifier, which is much simpler than the classifiers employed in [18] and [19]. Xu *et al.* pointed out that synthesis of microporous aluminophosphate was carried out in a gel medium under hydrothermal/solvo-thermal conditions by using the templates as structure-directing agents [11]. Gel composition is the material basis for producing chemical reaction, solvent provides the reaction environment, and template plays a role of structure-directing. Among the optimal features, F12 (second longest distance of organic template) is the most important feature. In the rational synthesis of microporous materials, the geometric factor of the organic template plays a vital role to affect the shape and the pore size of an AIPO structure. For open-framework AIPOs with (6,12)-rings, the organic templates are usually located in the one-dimensional 12-ring channels, thus their longest direction is extended along the channels. Therefore, the second longest distance of the organic templates is determinative to the window size of the channels [18]. From Figure 2, we can see that the optimal feature subset selected by our method contains 12 features belonging to three categories. However, the original Fisher score only selected features from two categories at dimension 12. Moreover, we can find that the second longest distance of the organic templates (F12) is selected by our method. These observations indicate that the proposed method is more consistent with the prior knowledge of synthetic chemists.

Compared with the methods in previous works [18,19], the proposed method has the following advantages. Firstly, it is independent of any classifier. Thus, as can be seen from the experimental results, the performances of our method are superior to other classical feature selection algorithms under both Nearest Neighbor and Naive Bayes classifiers. Secondly, the proposed method takes the correlations among the selected features into consideration. Therefore, it can remove the redundant information from the selected feature subset. However, we should point out that there also exists an inconvenient point in the proposed method. Since various feature scoring and correlation measurement algorithms can be incorporated into our method, there may be a need to conduct experiments to verify which combination of feature scoring and correlation measuring algorithms can obtain the best feature selection result.

#### 2.4. Comparisons with Other Feature Selection Methods

In this subsection, we compare the performance of the proposed method with some other state of the art feature selection methods including *T*-test [30], Constraint score [22], MRMR [23,24] and FCBF [25]. Among these methods, *T*-test and Constraint score are univariate feature selection methods that select features by the weights or importance degrees of features, while both FCBF and MRMR are multivariable feature selection methods that take the correlation among the selected features into consideration. We compare their performances under various dimensions on the AIPs dataset (Figure 3). Here, it should be noted that since the number of selected features cannot be predefined in FCBF, we are unable to test its performance under every dimension. Thus, only the average classification accuracy rate of 10-fold cross validation of FCBF is shown in Figure 3. The best classification accuracy rates obtained by these methods are listed in Table 4.

**Figure 3.** Performance comparison of the proposed algorithm and some popular feature selection methods. (a) Using nearest neighbor as classifier; (b) Using Naive Bayes as classifier.



**Table 4.** Highest classification accuracy rates reached by different feature selection methods.

Method	Nearest Neighbor		Naive Bayes	
	Highest Acc Rate	Dimension	Highest Acc Rate	Dimension
Constraint Score	0.908	21	0.8639	19
Ttest	0.9096	19	0.8728	2
FCBF	0.8072	/	0.8584	/
MRMR	0.908	21	0.868	1
Our algorithm (FI)	0.9112	12	0.8767	5

From Figure 3 and Table 4, we can find that the proposed algorithm outperforms other feature selection methods since it could get higher classification accuracy rate under relatively lower dimension, especially when the Nearest Neighbor is utilized for classification. However, it also can be observed that the proposed algorithm does not win over other algorithms by a very large margin in some cases. Therefore, like the experiments in Section 2.2, the *F*-measure is also employed here to evaluate the performances of different algorithms. From the *F*-measure values obtained by different algorithms in Table 5, we can see that the performance of the proposed algorithm is much better than other algorithms. These experimental results are consistent with Section 2.2.

**Table 5.** Optimal *F*-measure values reached by different feature selection methods.

Method	<i>F</i> -measure (Nearest Neighbor)	<i>F</i> -measure (Naive Bayes)
Constraint Score	0.8388	0.7588
Ttest	0.8046	0.7825
FCBF	0.5416	0.7730
MRMR	0.7723	0.7721
Our algorithm (FI)	0.8586	0.8071

### 3. Materials and Method

#### 3.1. Data Sets

The microporous aluminophosphate dataset used in this paper comes from the database of AIPOs synthesis established by the State Key Laboratory of Inorganic Synthesis and Preparative Chemistry of Jilin University (<http://zeobank.jlu.edu.cn/>). This database contains 1600 synthetic records in all. After removing the records that contain missing values (about 29% of the total), we use the remainder 1250 records in our experiment. In these records, 398 (6,12)-ring-containing AIPOs are deemed as positive samples, while 852 non-(6,12)-ring-containing AIPOs are deemed as negative samples. In this study, 21 synthetic features (or factors) belonging to three categories (Gel composition, Solvent and Organic template) are concerned (shown in Table 6). For more details about the definitions and meanings of the synthetic factors in Table 6, see [31].

**Table 6.** Description of the input synthetic factors.

Category	ID	Description
Gel composition	F1	The molar amount of Al <sub>2</sub> O <sub>3</sub> in the gel composition
	F2	The molar amount of P <sub>2</sub> O <sub>5</sub> in the gel composition
	F3	The molar amount of solvent in the gel composition
	F4	The molar amount of template in the gel composition
Solvent	F5	The density
	F6	The melting point
	F7	The boiling point
	F8	The dielectric constant
	F9	The dipole moment
	F10	The polarity
Organic template	F11	The longest distance of organic template
	F12	The second longest distance of organic template
	F13	The shortest distance of organic template
	F14	The Van der Waals volume
	F15	The dipole moment
	F16	The ratio of C/N
	F17	The ratio of N/(C + N)
	F18	The ratio of N/Van der Waals volume
	F19	The Sanderson electronegativity
	F20	The number of free rotated single bond
	F21	The maximal number of protonated H atoms

### 3.2. The Proposed Algorithm

Formally, suppose  $D = [d_1, d_2, \dots, d_n] \in R^{m \times n}$  is the input dataset that contains  $n$  samples in  $m$  dimensional space (For the microporous aluminophosphate dataset utilized in this study, the values of  $m$  and  $n$  in  $D$  are 21 and 1250, respectively). We can denote each row vector of  $D$  by  $P_i$  ( $i = 1, \dots, m$ ), which is corresponding to a feature. The aim of the proposed feature selection algorithm is to select  $k$  ( $k < m$ ) features from the original feature set to form a feature subset  $U$  in which the importance of the features are maximizing and the correlations among the features are minimizing.

Let  $S = [s_1, s_2, \dots, s_m]^T \in R^{m \times 1}$  be the positive weight of each feature which reflects its importance, where  $s_i$  is the weight of the  $i$ th feature ( $i = 1, \dots, m$ ). In this study, the weights of features can be obtained by any classical feature evaluation method (such as Fisher score, ReliefF score and Gini score), and the features with larger weights are more important. Let  $C \in R^{m \times m}$  be the correlation matrix, where  $C_{ij} \geq 0$  ( $i \neq j$ ) indicates the correlation between the  $i$ th and  $j$ th features. Since the self-correlation of the synthetic factor is meaningless, we assign the diagonal elements  $C_{ii}$  ( $i = 1, 2, \dots, m$ ) to be 0.  $f = [f_1, f_2, \dots, f_m]^T$  is an indicator vector, where  $f_i = 1$  means that the  $i$ th feature is selected into the subset  $U$ , and  $f_i = 0$  means the  $i$ th feature is not selected. The objective function of the proposed feature selection algorithm can be defined as:

$$\begin{aligned} & \max_f \left( \frac{f^T S}{k} - \frac{f^T C f}{k(k-1)} \right) \\ & \text{s.t. } \sum_i f_i = k, f_i \in \{0,1\} \end{aligned} \tag{3}$$

In Equation (3),  $\frac{f^T S}{k}$  stands for the average weight of the selected features,  $\frac{f^T C f}{k(k-1)}$  stands for the average correlation among the selected features, and the constraints are used for restricting the number of selected features in the  $U$  to be  $k$ . Thus, maximizing Equation (3) can ensure that the selected features in  $U$  are most important and least redundant. However, Equation (3) is a quadratic integral programming problem and it is hard to be solved [32]. Therefore, in our study, we relax the constraint of  $f_i \in \{0,1\}$  to  $f_i \in [0,1]$ , and convert the objective function in Equation (3) to:

$$\begin{aligned} & \max_f \left( \frac{f^T S}{k} - \frac{f^T C f}{k(k-1)} \right) \\ & \text{s.t. } \sum_i f_i = k, f_i \in [0,1] \end{aligned} \tag{4}$$

### 3.3. Solution

In this section, a pair-wise updating algorithm similar to that found in [32] is introduced to solve the maximization problem in Equation (4).

The Lagrangian function of Equation (4) can be derived as:

$$L(f, \lambda, \mu, \beta) = \left( \frac{f^T S}{k} - \frac{f^T C f}{k(k-1)} \right) - \lambda \left( \sum_i f_i - k \right) + \sum_i \mu_i f_i + \sum_i \beta_i (1 - f_i) \tag{5}$$

Where  $\lambda$ ,  $\mu_i$  and  $\beta_i$  are Lagrangian multipliers. Based on the Karush-Kuhn-Tucker (KKT) conditions [33], the solution that maximizes the Equation (4) must satisfy the first-order necessary conditions as:

$$\begin{cases} \left( \frac{S}{k} - \frac{2Cf}{k(k-1)} \right)_i - \lambda + \mu_i - \beta_i = 0 \\ \sum_i \mu_i f_i = 0 \\ \sum_i \beta_i (1 - f_i) = 0 \end{cases} \tag{6}$$

where  $\left( \frac{S}{k} - \frac{2Cf}{k(k-1)} \right)_i$  is the  $i$ th element of vector  $\frac{S}{k} - \frac{2Cf}{k(k-1)}$ . Because  $f_i$ ,  $\mu_i$  and  $\beta_i$  are all non-negative,

$\sum_i \mu_i f_i = 0$  means that if  $f_i > 0$ , then  $\mu_i = 0$ . Similarly,  $\sum_i \beta_i (1 - f_i) = 0$  means that if  $f_i < 1$ , then  $\beta_i = 0$ .

Thus, according to the relationship between  $\left( \frac{S}{k} - \frac{2Cf}{k(k-1)} \right)_i$  and  $\lambda$ , the KKT conditions can be rewritten as:

$$\left( \frac{S}{k} - \frac{2Cf}{k(k-1)} \right)_i \begin{cases} \leq \lambda & f_i = 0 \\ = \lambda & f_i \in (0,1) \\ \geq \lambda & f_i = 1 \end{cases} \tag{7}$$

Here, since  $\left(\frac{S}{k} - \frac{2Cf}{k(k-1)}\right)_i$  could reflect the relationship between the feature's weight and its average correlation with other features in  $U$ , we call it the reward of  $i$ th feature, and denote it by  $r_i(f)$ . According to the value of  $\left(\frac{S}{k} - \frac{2Cf}{k(k-1)}\right)_i$ , we can partition the feature set into three subsets,  $U_1 = \{P_i \mid f_i=0\}$ ,  $U_2 = \{P_i \mid f_i \in (0,1)\}$  and  $U_3 = \{P_i \mid f_i=1\}$ . From the constraints of  $f$  in Equation (4), it can be found that if a feature is in subset  $U_1$  or  $U_2$ , the value of its corresponding element in  $f$  can be increased. On the contrary, if a feature is in subset  $U_2$  or  $U_3$ , the value of its corresponding element in  $f$  can be decreased.

The pair-wise updating strategy to solve Equation (4) is defined as:

$$f_l^{new} = \begin{cases} f_l & l \neq i, l \neq j; \\ f_l + \alpha & l = i; \\ f_l - \alpha & l = j; \end{cases} \tag{8}$$

That is, only the values of two elements in  $f$  ( $f_i$  and  $f_j$ ,  $i \neq j$ ) are updated in each iteration of our algorithm. After updating  $f_i$  and  $f_j$ , the change of Equation (4) is:

$$\begin{aligned} \Delta &= \left(\frac{f^{newT}S}{k} - \frac{f^{newT}Cf^{new}}{k(k-1)}\right) - \left(\frac{f^TS}{k} - \frac{f^TCf}{k(k-1)}\right) \\ &= \frac{f^{newT}S - f^TS}{k} + \frac{f^TCf - f^{newT}Cf^{new}}{k(k-1)} \\ &= \frac{(s_i - s_j)\alpha}{k} + \frac{(2C_{ij} - C_{ii} - C_{jj})\alpha^2 + 2(e_jCf - e_iCf)\alpha}{k(k-1)} \\ &= \frac{(2C_{ij} - C_{ii} - C_{jj})\alpha^2 + (k-1)(s_i - s_j)\alpha + 2(e_jCf - e_iCf)\alpha}{k(k-1)} \end{aligned} \tag{9}$$

where  $e_i$  is a row vector with only the  $i$ th element equal to 1, and 0 otherwise. So, Equation (9) can be further converted as:

$$\begin{aligned} \Delta &= \frac{(2C_{ij} - C_{ii} - C_{jj})\alpha^2}{k(k-1)} + \frac{(k-1)s_i\alpha - 2e_iCf\alpha}{k(k-1)} - \frac{(k-1)s_j\alpha - 2e_jCf\alpha}{k(k-1)} \\ &= \frac{(2C_{ij} - C_{ii} - C_{jj})\alpha^2}{k(k-1)} + \left(\frac{s_i}{k} - \frac{2e_iCf}{k(k-1)}\right)\alpha - \left(\frac{s_j}{k} - \frac{2e_jCf}{k(k-1)}\right)\alpha \\ &= \frac{(2C_{ij} - C_{ii} - C_{jj})\alpha^2}{k(k-1)} + \left(\frac{S}{k} - \frac{2Cf}{k(k-1)}\right)_i\alpha - \left(\frac{S}{k} - \frac{2Cf}{k(k-1)}\right)_j\alpha \\ &= \frac{(2C_{ij} - C_{ii} - C_{jj})\alpha^2}{k(k-1)} + (r_i(f) - r_j(f))\alpha \end{aligned} \tag{10}$$

With the aim of maximizing  $\Delta$ , according to Equation (10) and the constraints of  $f$ ,  $\alpha$  can be computed as:

$$\alpha = \begin{cases} \min(f_j, 1 - f_i) & \text{if } 2C_{ij} - C_{ii} - C_{jj} \geq 0 \text{ and } r_i(f) > r_j(f) \\ \min\left(f_j, 1 - f_i, \frac{k(k-1)(r_j(f) - r_i(f))}{2C_{ij} - C_{ii} - C_{jj}}\right) & \text{if } 2C_{ij} - C_{ii} - C_{jj} < 0 \text{ and } r_i(f) > r_j(f) \\ \min(f_j, 1 - f_i) & \text{if } 2C_{ij} - C_{ii} - C_{jj} > 0 \text{ and } r_i(f) = r_j(f) \end{cases} \tag{11}$$

Note that in the updating algorithm above, only the situation that  $r_i(f) \geq r_j(f)$  is considered. If  $r_i(f) < r_j(f)$ , exchange  $i$  and  $j$  to implement the algorithm.

By iteratively updating the values of pair-wise elements in  $f$  and computing  $\alpha$  using Equations (8) and (11), the objective function in Equation (4) can be increased and reach its maximum [32]. The implementation details of the proposed feature selection method are summarized in Algorithm 1.

**Algorithm 1.** The feature selection process of the proposed method.

Input: The original data sample  $D$ .  
 Output: The indicator vector  $f$ .

1. Compute scores of features  $S$  and correlation matrix  $C$ .
2. Initialize  $f$ ,
3. Do
4. Select  $P_i \in U_1 \cup U_2$  which has the largest reward  $r_i(f)$ ;
5. Select  $P_j \in U_2 \cup U_3$  which has the smallest reward  $r_j(f)$ ;
6. if  $r_i(f) > r_j(f)$   
 Compute  $\alpha$  using Equation (11), and then update  $f_i$  and  $f_j$  according to Equation (8);
7. else if  $r_i(f) = r_j(f)$
8. if  $2C_{ij} - C_{ii} - C_{jj} > 0$   
 Compute  $\alpha$  using Equation (11), and then update  $f_i$  and  $f_j$  according to Equation (8);
9. else if  $2C_{ij} - C_{ii} - C_{jj} = 0$   
 Check whether there exist a  $P_o \in U_1 \cup U_2$  and a  $P_x \in U_2 \cup U_3$  such that  $2C_{ox} - C_{oo} - C_{xx} > 0$  and  $r_o(f) = r_x(f)$ . If the pair  $(P_o, P_x)$  can be found, Compute  $\alpha$  using Equation (11), and then update  $f_o$  and  $f_x$  according to Equation (8); Otherwise,  $f$  is a solution of Equation (4);
10. end if
11. end if
12. until  $f$  is a solution of Equation (4).

As can be seen in Algorithm 1, a heuristic strategy is adopted in each iteration of the pair-wise updating algorithm to increase the objective function maximally. In this strategy, a pair of elements in  $f$  whose values should be updated is selected according to the rewards of their corresponding features. In other words, the element whose value should be increased in each iteration is selected as the one whose corresponding feature has the largest reward in subset  $U_1$  or  $U_2$ , and the element whose value should be decreased in each iteration is selected as the one whose corresponding feature has the smallest reward in subset  $U_2$  or  $U_3$ . From Equation (10), we can find that the increase of the objective function in Equation (4) can be maximized by this method. The solution of proposed algorithm is obtained when the value of Equation (4) cannot be further increased.

#### 4. Conclusions

In this study, a novel feature selection method based on maximum weight and minimum redundancy criterion is proposed. Comprehensive experiments and deep analysis based on the microporous

aluminophosphates (AIPOs) database demonstrate the effectiveness of the proposed algorithm. This work also demonstrates the feasibility of feature selection techniques in chemical data analysis. By taking advantage of the proposed algorithm, we investigate the relationship between synthetic factors and rational synthesis of microporous materials. The classification result with a classification accuracy rate of 91.12% shows that a number of synthetic factors including the molar amount of  $\text{Al}_2\text{O}_3$ , the molar amount of solvent, the molar amount of template in the gel composition, the melting point, the dipole moment, the second longest distance of organic template, the dipole moment, the ratio of  $C/N$ , the ratio of  $N/(C + N)$ , the ratio of  $N/\text{Van der Waals volume}$  and the maximal number of protonated H atoms play vital roles for rational synthesis of (6,12)-ring-containing AIPOs. Among these optimal synthetic factors, the second longest distance of organic template, which is the geometric size of the organic template, plays the most important role in the prediction. This work provides *a priori* knowledge and a useful guidance for rational synthesis experiments of such materials.

In future studies, we will gradually add more synthetic features (or factors) into the database to investigate their influences for the synthesis of AIPOs.

### Acknowledgments

This work was supported by the Fund of Jilin Provincial Science & Technology Department (No.201115003), the Fundamental Research Funds for the Central Universities (No.11QNJJ005), the Science Foundation for Post-Doctor of Jilin Province (No.2011274), the Young Scientific Research Fund of Jilin Province Science, the Technology Development Project (No.201201070, 201201063), and the Fund of Key Laboratory of Symbolic Engineering MOE (No.93K172012K13).

### Conflicts of Interest

The authors declare no conflict of interest.

### References

1. Hyunjoo, L.; Zones, S.I.; Davis, M.E. A combustion-free methodology for synthesizing zeolites and zeolite-like materials. *Nature* **2003**, *425*, 385–388.
2. Yu, J.H.; Xu, R.R. Insight into the construction of open-framework aluminophosphates. *Chem. Soc. Rev.* **2006**, *25*, 593–604.
3. Li, Y.; Yu, J.H.; Liu, D.H.; Yan, W.F.; Xu, R.R.; Xu, Y. Design of zeolite frameworks with defined pore geometry through constrained assembly of atoms. *Chem. Mater.* **2003**, *15*, 2780–2785.
4. Li, Y.; Yu, J.H.; Wang, Z.P.; Zhang, J.N.; Guo, M.; Xu, R.R. Design of chiral zeolite frameworks with specified channels through constrained assembly of atoms. *Chem. Mater.* **2005**, *17*, 4399–4405.
5. Li, Y.; Yu, J.H.; Xu, R.R.; Baerlocher, C.; McCusker, L.B. Combining structure modeling and electron microscopy to determine complex zeolite framework structures. *Angew Chem.* **2008**, *120*, 4473–4477.
6. Li, Y.; Yu, J.H.; Jiang, J.X.; Wang, Z.P.; Zhang, J.N.; Xu, R.R. Prediction of open-framework aluminophosphate structures using the automated assembly of secondary building units method with Lowenstein's constraints. *Chem. Mater.* **2005**, *17*, 6086–6093.

7. Ren, X.Y.; Li, Y.; Pan, Q.H.; Yu, J.H.; Xu, R.R.; Xu, Y. A crystalline germanate with mesoporous 30-ring channels. *J. Am. Chem. Soc.* **2009**, *131*, 14128–14129.
8. Li, J.Y.; Li, L.; Liang, J.; Chen, P.; Yu, J.H.; Xu, Y.; Xu, R.R. Template-designed syntheses of open-framework zinc phosphites with extra-large 24-ring channels. *Cryst. Growth Des.* **2008**, *8*, 2318–2323.
9. Li, J.Y.; Yu, J.H.; Yan, W.F.; Xu, Y.H.; Xu, W.G.; Qiu, S.L.; Xu, R.R. Structures and templating effect in the formation of 2D layered aluminophosphates with  $\text{Al}_3\text{P}_4\text{O}_{16}^{3-}$  stoichiometry. *Chem. Mater.* **1999**, *11*, 2600–2606.
10. Yu, J.H.; Li, J.Y.; Wang, K.X.; Xu, R.R.; Sugiyama, K.; Terasaki, O. Rational synthesis of microporous aluminophosphates with an inorganic open framework analogous to  $\text{Al}_4\text{P}_5\text{O}_{20}\text{HC}_6\text{H}_{18}\text{N}_2$ . *Chem. Mater.* **2000**, *12*, 3783–3787.
11. Xu, R.R.; Pang, W.Q.; Yu, J.H.; Huo, Q.S.; Chen, J.S. *Chemistry of Zeolites and Related Porous Materials: Synthesis and Structure*; John Wiley and Sons: Singapore, 2007.
12. Li, J.Y.; Yu, J.H.; Xu, R.R. Database of AIPO Syntheses. Available online: <http://zeobank.jlu.edu.cn> (accessed on 26 October 2011).
13. Pichler, M.A.; Perone, S.P. Computerized pattern recognition applications to chemical analysis: Development of interactive feature selection methods for the K-nearest neighbor technique. *Anal. Chem.* **1974**, *46*, 1790–1798.
14. Liu, Y. A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823–1828.
15. Teramoto, R.; Fukunishi, H. Supervised consensus scoring for docking and virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 526–534.
16. Venkatraman, V.; Dalby, A.R.; Yang, Z.R. Evaluation of mutual information and genetic programming for feature selection in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1686–1692.
17. Rodgers, S.; Glen, R.C.; Bender, A. Characterizing bitterness: Identification of key structural features and development of a classification model. *J. Chem. Inf. Model.* **2006**, *46*, 569–576.
18. Li, J.Y.; Qi, M.; Kong, J.; Wang, J.Z.; Yan, Y.; Huo, W.F.; Yu, J.H.; Xu, R.R.; Xu, Y. Computational prediction of the formation of microporous aluminophosphates with desired structural features. *Microporous Mesoporous Mater.* **2010**, *129*, 251–255.
19. Huo, W.F.; Gao, N.; Yan, Y.; Li, J.Y.; Yu, J.H.; Xu, R.R. Decision trees combined with feature selection for the rational synthesis of aluminophosphate  $\text{AlPO}_{4-5}$ . *Acta Phys. Chim. Sin.* **2011**, *27*, 2111–2117.
20. Chen, Y.W.; Lin, C.-J. Combining SVMs with various feature selection strategies. *Feature Extraction* **2006**, *207*, 315–324.
21. Hall, M.A. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. In *Machine Learning: Proceedings of the International Conference*; Morgan Kaufmann: San Francisco, CA, USA, 2000; pp. 359–366.
22. Zhang, D.Q.; Chen, S.C.; Zhou, Z.-H. Constraint score: A new filter method for feature selection with pairwise constraints. *Pattern Recogn.* **2008**, *41*, 1440–1451.
23. Ding, C.; Peng, H.C. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinf. Comput. Biol.* **2005**, *3*, 185–205.

24. Peng, H.C.; Long, F.H.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
25. Yu, L.; Liu, H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington, DC, USA, 21–24 August 2003.
26. Soda, P. A multi-objective optimization approach for class imbalance learning. *Pattern Recogn.* **2011**, *44*, 801–1810.
27. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
28. Kononenko, I. Estimating Features: Analysis and Extension of RELIEF. In Proceedings of the 6th European Conference on Machine Learning, Catania, Italy, 6–8 April 1994; pp. 171–182.
29. Breiman, L.; Friedman, J.H.; Olshen, R.A. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, USA, 1984.
30. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517.
31. Yan, Y.; Li, J.Y.; Qi, M.; Zhang, X.; Yu, J.H.; Xu, R.R. Database of open-framework aluminophosphate syntheses: Introduction and application (I). *Sci. China Ser. B* **2009**, *52*, 1734–1738.
32. Liu, H.R.; Yang, X.W.; Latecki, L.J.; Yan, S.C. Dense neighborhoods on affinity graph. *Int. J. Comput. Vis.* **2012**, *98*, 65–82.
33. Kuhn, W.; Tucker, A. Nonlinear Programming. In Proceedings of the Second Berkeley Symposium, Berkeley, CA, USA, 31 July–12 August 1950; pp. 481–492.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).