Microporous and Mesoporous Materials 186 (2014) 201–206

Contents lists available at ScienceDirect

Microporous and Mesoporous Materials

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

A novel classification method based on the ensemble learning and feature selection for aluminophosphate structural prediction

Minghai Yao [a,b], Miao Qi [a,*], Jinsong Li [a], Jun Kong [b,c,*]

[a] *School of Computer Science and Information Technology, Northeast Normal University, Key Laboratory of Intelligent Information Processing of Jilin Universities, Changchun 130117, China*
[b] *School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China*
[c] *Key Laboratory for Applied Statistics of MOE, Northeast Normal University, Changchun 130024, China*

ARTICLE INFO

Article history:
Received 2 May 2013
Received in revised form 30 September 2013
Accepted 9 December 2013
Available online 13 December 2013

Keywords:
Zeolites
Polymer processing
Feature selection
Ensemble learning
Simulation

ABSTRACT

In this paper, a novel classification algorithm based on the ensemble learning and feature selection is proposed for predicting the specific microporous aluminophosphate ring structure. The proposed method can select the most significant synthetic factors for the generation of (6, 12)-ring-containing structure. First, the clustering method is employed for making each training subset contains all the structural characteristics of samples. Then, the method takes full account of the discrimination and class information of each feature by calculating the scores. Specially, the scores are fused for getting a weight for each feature. Finally, we select the significant features according to the weights. The result of feature selection will help to predict the (6, 12)-ring-containing AlPO structure well. Moreover, we compare our method with several classical feature selection methods and classification method by theoretical analysis and extensive experiments. Experimental results show that our method can achieve higher predictive accuracy with less synthetic factors.

1. Introduction

Zeolite materials are an important class of crystalline inorganic microporous solids formed by TO4 tetrahedra (T infers Si, P, Al, Ge, Ga, etc.) with a well defined regular pore system. The most interesting features of zeolites lie in their variable chemical compositions of the pore wall, as well as the tunable pore diameters and pore shape. These excellent characters endow zeolites with wide applications in catalysis, adsorption, separation, ion exchange and other fields [1–3]. According to the number of the pore ring, zeolites are classified as small, medium, large, and extra-large pore structure with the pore window delimited by 8, 10, 12 and more than 12 T-atoms. Extra-large pore zeolites are drawing more and more attention because they can process bigger molecule as desire in the fields mentioned above. In recent years, Corma and Baumes et al. have been engaged in research about the synthesis of microporous materials. They have found a lot of factors that affect the synthesis of microporous materials [4–7]. In literature [8], Corma and co-workers summed up the research situation of extra-large pore molecular sieve materials from the structure, stability, catalysis and so on. Microporous aluminophosphate as an important branch of molecular sieve materials, has been widespread concern by researchers at home and abroad. At present, 60 kinds of microporous aluminophosphate structures are known, where twelve-ring pore size of 0.73 nm is a typical representative and has important applications in adsorption and catalytic fields.

However, the crystallization kinetics of such materials is rather complicated. In general, there are many factors that influence the crystallization kinetics and the final crystalline phases, such as reaction raw materials, the gel composition, the reaction pH, the organic template agent, solvent, etc. Therefore, the rational synthesis of new microporous materials remains a significant challenge in the field of inorganic chemistry. In order to mine the relationships between the synthetic factors and the resulting structures, and further guide the rational synthesis of AlPO materials, Yu and co-workers have built the AlPO synthesis database including about 1600 items for scholars [9–10]. Each reaction data records the synthesis conditions including gel molar, temperature and time, solvent and template type, and the structural characteristics of the product. This database can provide a research platform for the rational design and synthesis of microporous materials [11]. In the past few years, AlPO molecular sieve has been used as a target to probe the relationships between synthetic factors and the resulting framework structures [12–16]. Li and co-workers adopted Support Vector Machine (SVM) to predict (6, 12)-ring-containing microporous AlPO's, which gave the best combination

* Corresponding authors at: School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China. Tel.: +86 431 84536326.

E-mail addresses: qim801@nenu.edu.cn (M. Qi), kongj435@nenu.edu.cn (J. Kong).

1387-1811/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.micromeso.2013.12.008

of synthetic factors based on brute-force method [12]. In literature [13], Partial Least Squares and Logistic Discrimination were used to predict the formation of microporous aluminophosphate $AlPO_4$-5. In addition, four re-sampling methods were proposed to deal with the problem of class imbalance. Li and co-workers proposed an $AlPO_4$-5 prediction system based on C5.0 combined with Fisher score [14].

As mentioned above, using feature selection methods and data mining techniques can better find the interaction between the synthetic condition and the specified product. In this paper, a classification algorithm based on the ensemble learning and feature selection is proposed, which can provide helpful analysis to the rational synthesis of microporous aluminophosphate. In our method, the cluster analysis technique is used to cluster training samples. Moreover, multi-classifier fusion mechanism is employed for improving classification performance. In particular, a new feature selection method is proposed to explore the significant factors for specific structure. The method combines the generation method for training and testing sets in cluster analysis approaches, ensures the diversification of training sets, solves the problem with sample imbalance. By improving the feature selection method, we explore the main factors which affect the results of the synthetic in the synthesis process for AlPO. Therefore, we obtain the AlPO synthesis prediction model which has higher prediction accuracy.

In order to demonstrate the effectiveness and superiority of the proposed method, we compare our method with several classical feature selection methods and classification method on the basis of prediction accuracy through extensive experiments. From the view of data processing, the proposed method ensures the richness of the data structure information for sampling training samples and considers both the discrimination and class information of features for feature selection. Moreover, it can deal with the problems with the class imbalance and the redundancy among features. The proposed method adopts the idea of ensemble learning, which can construct a prediction model having higher prediction accuracy.

This paper is organized as follows. Section 2 introduces the feature selection methods, the FCM clustering algorithm and ensemble learning. Section 3 describes the idea of the classification method and feature selection method. Section 4 is comparison experiments. Section 5 is results and discussions. Finally, conclusions are given in Section 6.

2. Related works

2.1. Feature selection

The need for feature selection (FS) often arises in machine learning and pattern recognition problems. FS has been one of the key steps in mining high-dimensional data for decades. The idealized definition of feature selection is to find the minimally sized feature subset that is necessary and sufficient for a specific task. FS has several potential benefits, such as improving the accuracy of classification, avoiding the well-known "curse of dimensionality", speeding up the training process and reducing storage demands. Specially, it can provide a better understanding and interpretability for a domain expert [17]. Generally, FS techniques are classically grouped into two classes: filter based methods and wrapper based methods [18–19]. A filter method assesses the quality of a given subset of features using solely characteristics of that subset without any learning algorithm. In contrast, the wrapper method evaluates the adequacy of a subset of features based on the performance of some classifier operating with the selected features. Wrapper based methods are more expensive computationally. In this study, we are particularly interested in the filter methods and propose a novel feature selection method by considering both discrimination and class information.

2.2. FCM clustering algorithm

The fuzzy c-means (FCM) algorithm is proposed by Bazdek, which is an improvement of the hard k-means algorithm [20]. It assigns a class membership to a data point, depending on the similarity of the data point to a particular class relative to all other classes. The FCM objective function of data set into c clusters is:

$$J_m(\mu, \upsilon) = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}d^2(x_j, \upsilon_i) \quad s.\ t.\sum_{i=1}^{c}\mu_{ij} = 1, \tag{1}$$

where $X = (x_1, x_2, \ldots x_j, \ldots x_n)$ is a data matrix with the size of $p \times n$, p represents the dimension of each feature, and n represents the number of data points, m presents the index of fuzziness, and υ_i is the fuzzy cluster centroid of the ith cluster. Using the Euclidean norm, the distance metric d measure the similarity between a data point x_j and a cluster centroid υ_i in the feature space:

$$d^2(x_j, \upsilon_i) = \|x_j - \upsilon_i\|^2. \tag{2}$$

The objective function is minimized when data points are close to the centroids of their clusters and assigned high membership values, and low membership values are assigned to data points far from the centroids. Letting the first derivatives of J_m with respect to μ and υ equal to zero yields, the two necessary conditions for minimizing J_m as follows:

$$\mu_{ij} = \left(\sum_{k=1}^{c}\left(\frac{d(x_j, \upsilon_i)}{d(x_j, \upsilon_k)}\right)^{2/(m-1)}\right)^{-1} \tag{3}$$

and

$$\upsilon_i = \frac{\sum_{j=1}^{n}\mu_{ij}^{m}x_j}{\sum_{j=1}^{n}\mu_{ij}^{m}}. \tag{4}$$

The FCM algorithm proceeds by iterating the two necessary conditions until a solution is reached. Each data point will be associated with a membership value for each class after FCM clustering. By assigning the data point to the class with the highest membership value, a segmentation of the data could be obtained.

2.3. Ensemble learning

Ensemble learning improves generalization performance of individual learners by combining the outputs of a set of diverse base classifiers. Previous theoretical and empirical researches have shown that formation of ensemble is always more accurate than individual components in the ensemble, if and only if individual members are both accurate and diverse [21].

Lots of methods have been developed for constructing classification ensemble. The most popular techniques include the Random subspace methods [22], Bagging [23] and Boosting [24]. Random subspace method was first introduced by the literature [25], which is based on a random sampling for original feature components to obtain different feature subsets. In recent years, it has been applied to feature selection, clustering and other areas. Both Bagging and Boosting train the base classifiers by resampling training sets. These classifiers are usually combined by simple majority voting in the final decision rule. One difference between Bagging and Boosting lies in that the former obtains a bootstrap sample by uniformly sampling with replacement from original training set, while the latter resamples the training data by emphasizing more on samples that are misclassified by previous classifiers. Recently, besides classification ensemble, there also appears clustering