Microporous and Mesoporous Materials 185 (2014) 1–6

Contents lists available at ScienceDirect

Microporous and Mesoporous Materials

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

Sample-expand method for predicting the specified structure of microporous aluminophosphate

Miao Qi [a], Zhanmin Qin [a], Na Gao [b], Jun Kong [c,*], Yuting Guo [a], Yinghua Lu [a,*]

[a] School of Computer Science and Information Technology, Northeast Normal University, Key Laboratory of Intelligent Information Processing of Jilin Universities, Changchun 130117, China
[b] State Key Laboratory of Inorganic Synthesis and Preparative Chemistry, College of Chemistry, Jilin University, Changchun 130012, China
[c] Key Laboratory for Applied Statistics of MOE, Changchun 103024, China

ARTICLE INFO

Article history:
Received 22 April 2013
Received in revised form 23 August 2013
Accepted 7 October 2013
Available online 19 October 2013

Keywords:
Zeolite synthesis
Sample-expand
K-nearest neighbor
Support vector machines
Prediction

ABSTRACT

Imbalanced data sets often exist in many real-world fields and this problem has got more and more attention in recent years. In this paper, a sample-expand method is proposed as data pre-processing procedure to improve the predictive performance of the zeolite synthesis on imbalance data set. First, the data pre-processing is implemented for expanding samples by exploring the marginal structure of the given data set using *k*-nearest neighbor algorithm (KNN). Then, the expanded data set is input to support vector machines (SVM) for classification. Finally, *Q* times *n*-fold cross-validations procedure (CVs) is adopted to assess the prediction performance. The advantage of the data pre-processing is that it can obtain stable data set for establishing the training model and abide by the classification criteria of SVM, such that the improved predictive performance is achievable. Moreover, other classical machine learning methods are also presented to accomplish the prediction task. Compared experimental results demonstrate that SVM method can reach very satisfactory predictive accuracy on the pre-processing data set. Specially, the phase diagram of gel composition is provided as a guiding role for subsequent rational synthesis experiments.

1. Introduction

Over the past few decades, rational synthesis and design of chemical materials have attracted a lot of attention. In synthetic chemistry, one important goal is to product some chemical materials with desired structures and characteristics under certain synthetic conditions. However, the process of synthesis is rather complex since there are a large number of synthetic factors affecting the results, such as source materials, crystallization temperature, and time. Accordingly, predicting the specified structure under the specified synthetic condition is a challenging and significant task. Understand the relationship between the synthetic factors and the specified characteristic can help chemical researchers to explore the chemical synthesis mechanism, overcome the malpractice of discovering problem by extensive experiments and reduce the experimental cost.

With the rapid development of computer science and technology, many machine learning and data mining methods have been applied to chemical research. After obtaining a large number of experimental data, various modeling methods have been developed in various areas of chemistry [1–6]. Feed-forward neural networks (FFNNs) and associative neural network (ASNN) were used to predict ^{1}H NMR chemical shift first time [1]. Various regression methods, multi-layer perceptron (MLP) and SVM were described to predict the ITQ-21 zeolite phase crystallinity [2]. MIA-QSPS combined with partial least squares (PLS) were used to predict predicted the ^{13}C chemical shifts [3]. Generalized regression neural network combined with gray relational analysis and principal component analysis (PCA) was adopted to establish nonlinear model for calculating homolysis bond dissociation energies (BDE) of Y—NO bond [4]. Genetic algorithm (GA) was employed to select effective chemical features and SVM was adopted to establish the proper predictive model for promoter sequences [5]. In [6], GA was first used to select the most informative molecular descriptors. Then, multiple linear regression (MLR) was carried out to build the relationship between selected molecular descriptors and chemical shift. Experimental results indicated that the carbon-13 NMR chemical shifts property can be successfully modeled with theoretical molecular descriptors.

In this paper, we present a novel predictive model based on KNN and SVM methods for microporous aluminophosphate sieves. Moreover, the generalization of the proposed model is also given for guiding the subsequent experiments. The target of this paper is to predict the specified structure under the specified synthetic condition in rational synthesis. In order to avoid the influence of

[*] Corresponding authors. Tel.: +86 18686501175 (J. Kong).
E-mail address: 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.2013.10.009

imbalance data set and improve the predictive results, a novel sample-expand method is first proposed. This pre-processing procedure can find the marginal samples such that the training model can well depict internal structure characteristics of the whole training samples. Then, back propagation neural network (BP), and classification and regression tree (CART) methods are compared with SVM for predicting the specified structure. Finally, extensive comparison results indicate the effectiveness and superiority of the proposed method. Specially, the phase diagrams of gel composition are provided as a generalization of the proposed method.

2. Aluminophosphate molecular sieve

Aluminophosphate molecular sieves $AlPO_4$-n was synthesized by Wilson et al., [7] in 1982. The common structural feature of microporous aluminophosphates (AlPOs) is characterized by alternating connection of Al-centered polyhedra and P-centered tetrahedra. Its discovery not only enriches the molecular sieve family, but also attracts the attention of many scholars due to their significant applications in petroleum industry. Consequently, research on the rational synthesis of AlPOs is extremely important. In the basis of the publication of scientific achievements, Xu and Yu et al., [8] built up the AlPO synthesis database for exploiting and analyzing synthesis mechanism, which contains over 1600 reaction data for ca. 230 AlPO structures. Each reaction data is composed of the basic information about references, source materials, solvent, template, synthetic condition, and structural characteristics of the product. In this AlPO synthesis database, AlPOs contain 17 types of ring structure according to the pore ring sizes of channel structures. Among them, AlPOs with (6,12)-rings have a dozen and represent a major class. Some researches have been reported for predicting this specified structure [9,10] or the type of product [11,12] on this database, aiming to provide a new route and insight for rational synthesis of zeolites. In particular, four resampling methods have been depicted to solve the problem of class imbalance [11]. Experimental results indicated these methods achieved better predictive performances than some common resampling methods. For improving the database and providing good research platform, Li et al., [13] employed a series of imputing methods to estimate the missing values. Experimental results indicated that the back-propagation neural network imputation (BPimpute) method performed best. More recently, the syntheses of aluminophosphate molecular sieves AFI have been carried out based on the results of BPimpute method [14]. However, there is no relevant report about the prediction of the specified structure under the synthetic condition with specified template and solvent. Moreover, how to generalize the predictive model according to existing data has also not be presented on this database.

3. Machine learning methods

3.1. K-nearest neighbor algorithm (KNN)

KNN is one of the simplest classification algorithms. It first calculates the distances between a query sample and all the other samples, and takes K samples with the smallest distances. Then the query sample is classified by assigning the label most frequently among the K nearest samples. In general, the Euclidean distance is used as the distance metric in KNN. In addition to classification, KNN is also used as an important tool for data analysis.

3.2. Back propagation neural network (BP)

BP was proposed by Rumelhart and McCelland in 1986 [15], which is a multi-layer feed-forward neural network model and used widely as a supervised classification method. It makes the sum of the squares of errors minimum by adjusting the weights and thresholds of the network through error back propagation. The learning rule of BP is based on gradient descent. The topological structure of BP includes input layer, hidden layer, and output layer.

3.3. Support vector machine (SVM)

The main idea of SVM is to map data into a higher-dimensional space for realizing the linear separation. The target of training a SVM is to find the separating hyperplane with the largest margin. The support vectors are the training samples that define the optimal hyperplane. Informally speaking, it is very crucial to find the support vectors because they are most informative for classification. Within the past decades, it has been extensively applied in variously fields due to outstanding performances [16–18].

3.4. Classification and regression tree (CART)

CART is a nonparametric technique that constructs the prediction model with samples, and can select the most important variables in determining the class [19]. Moreover, a CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole training learning samples. The process is then repeated for each of the resulting data fragments.

3.5. Oversampling methods

For two-class problem, the classifier might overfit the samples belong to majority class when the numbers of two-class samples are not balanced. The oversampling methods can deal with this problem effectively, especially when the number of minority class is fewer and all samples should be used for classification as much as possible. The idea of oversampling is to generate some samples for the minority class by some Interpolation technologies. Random oversampling [20] is a very simple and effective method which replicates the samples from minority class randomly until the two classes are balanced. A guided resampling method [21] was given by enlarging the under-represented clusters through unsupervised learning methods. Synthetic Minority Over-sampling Technique (Smote) [22] is a heuristic oversampling method which forms new samples by interpolating between the nearest neighbor samples of minority. In our previous work, two guided oversampling methods based on fuzzy C-means were presented [11], which considered the between-class imbalance and the within-class imbalance simultaneously.

4. Experiments and discussion

4.1. Experimental data

The experimental data are selected from AlPO synthesis database, which is available at the IZA website <http://zeobank.jlu.edu.cn/>. There are 1279 items after deleting the ones with missing values, each of which is formed with 21 factors of gel compositions, solvent, and organic template [10]. The gel compositions are described as molar ratios: P_2O_5/Al_2O_3, Solvent/Al_2O_3, Template/Al_2O_3. The numbers of solvent and templates types are 13 and 125, respectively. Our target is to predict the specified structure in product under the specified conditions with the same solvent and the same template. In this work, the specified structure concerns (6,12)-rings in AlPOs. Water and triethylamine are selected as