第 32 卷第 3 期
2016 年 3 月

无 机 化 学 学 报
CHINESE JOURNAL OF INORGANIC CHEMISTRY

Vol.32 No.3
457-463

基于数据挖掘方法的开放骨架磷酸铝定向合成参数分析

郭羽婷[1] 高 娜[2] 史瑞新[3] 齐 妙[1] 王建中*,[1]

([1]东北师范大学计算机科学与信息技术学院,长春 130117)

([2]吉林大学化学学院,长春 130012)

([3]吉林大学口腔医学院,长春 130021)

摘要:开放骨架磷酸铝化合物是多孔晶体材料的一个重要家族。然而,这类材料的合成受到反应原料、凝胶组成、溶剂、模板剂、结晶温度和结晶时间等多个因素的影响。本文以吉林大学"无机制备与合成化学国家重点实验室"建立的开放骨架磷酸铝合成反应数据库为研究对象,采用最大权重最小冗余特征选择算法(Maximum weight and minimum redundancy,MWMR),在充分考虑合成参数自身的重要程度和合成参数之间的相关关系的前提下,分析了溶剂、模板剂等合成参数对于合成含有(8,6)元环结构开放骨架磷酸铝的影响。通过大量实验验证了该方法在开放骨架磷酸铝合成参数分析中的有效性,分析了合成参数对产物生成的影响。实验结果表明模板剂的几何参数、模板剂中 C 原子和 N 原子的个数比,溶剂的偶极距等参数可能对于该类结构的合成具有较为重要的影响。

关键词:开放骨架磷酸铝;合成参数;数据挖掘;特征选择

中图分类号:O611.2 **文献标识码**:A **文章编号**:1001-4861(2016)03-0457-07

DOI:10.11862/CJIC.2016.075

Rational Synthetic Parameter Analysis of Open-framework AlPOs Based on Data Mining Method

GUO Yu-Ting[1] GAO Na[2] SHI Rui-Xin[3] QI Miao[1] WANG Jian-Zhong*,[1]

([1]*School of Computer Science and Information Technology, Northeast Normal University, Changchun* 130117, *China*)

([2]*College of Chemistry, Jilin University, Changchun* 130012, *China*)

([3]*School and Hospital of Stomatology, Jilin University, Changchun* 130012, *China*)

Abstract: Open-framework aluminophosphates (AlPOs) is an important family of the porous crystal materials. However, the synthesis of the Open-framework aluminophosphates is affected by many parameters, such as reaction material, gel composition, solvent, template agent, crystallization temperature and crystallization time etc. Based on the ALPOs synthesis database, which established by the State Key Laboratory of Inorganic Synthesis and Preparative Chemistry of Jilin University, the work in this paper concentrates on analyzing the relationship between the synthetic parameters and the final product. In order to take both the importance and correlation of the features into consideration in the synthetic parameter analysis, we apply Maximum Weight and Minimum Redundancy (MWMR) to analyze the impact of solvent parameters and template parameters for the rational synthesis of (8,6)-ring-containing AlPOs. The effectiveness of the method is demonstrated by extensive experiments. Furthermore, we also make some deep analyses about the relationship between the synthetic parameters and final products.The experimental results show that the geometric parameters of the of organic template, the n_C/n_N and the dipole moment of the solvent etc. may impact most for the final product of this kind of open-framework aluminophosphates.

Keywords: open-framework aluminophosphates; synthetic parameter; data mining; feature selection

收稿日期:2015-11-11。收修改稿日期:2015-12-23。

国家自然科学基金(No.61403078)资助项目。

*通信联系人。E-mail:wangjz019@nenu.edu.cn

开放骨架磷酸铝材料以其丰富的孔道结构、多样的元素组成在催化、吸附和分离等领域有着潜在的应用价值。然而,这类材料的合成受到多个合成参数的影响, 其结晶机理难以理解和难以建模,给定向合成带来巨大的挑战[1]。为了深入理解开放骨架磷酸铝材料的形成机理,吉林大学“无机制备与合成化学国家重点实验室”在国际上率先建立了开放骨架磷酸铝(AlPOs)合成反应数据库[2-3]。

数据挖掘技术可以从大量数据中提取或“挖掘”知识,是一种基于机器学习、统计学等的决策支持过程[4-5]。通过数据挖掘方法进行数据分析,可以发现重要的信息,对各个领域的研究均做出了较大的贡献。特征选择是一种重要的数据挖掘技术。特征选择是指根据某种评估标准, 选择出数量较少、评估效果较好的特征子集[6]。

通过特征选择技术可以深入分析数据本质,挖掘隐藏在大量数据中的潜在信息。相关领域研究者以吉林大学建立的磷酸铝合成反应数据库为研究对象,利用特征选择方法开展了合成参数分析相关的一系列研究。文献[7]通过一种穷尽搜索的策略分析了 11 个合成参数对于含有(12,6)元环结构 AlPOs 生成的影响。文献[8]采用基于决策树的特征选择方法分析了 26 个合成参数对于合成 $AlPO_4$-5 的影响。文献[9]将含有(12,6)元环结构的 AlPOs 作为研究对象,提出了一种基于融合学习与特征选择的分类方法,分析合成参数对于该类结构合成的影响。文献[10]提出了一种基于随机子空间、Fisher 得分和顺序前向搜索的特征选择模型,分析了合成参数与产物结构之间的关系。文献[11]根据经验知识对含有(8,6)元环结构 AlPOs 的合成参数进行了分析,并利用支持向量机对其结论作了验证。

已有工作充分证明了特征选择技术在磷酸铝合成参数分析中应用的有效性和可行性。但存在以下局限:已有工作在特征选择过程中没有考虑特征之间的相互依赖关系,即相关关系。文献[23]已验证了考虑特征之间相关关系在特征选择中的重要性。特征之间的相关关系一般用相关系数度量。两个特征之间的相关系数绝对值越大,它们之间的相互依赖关系就越强。在 AlPOs 合成反应数据库中,合成参数之间存在着比较严重的相关关系。如表 1 的合成参数中,F14 与 F20、F17 与 F18 的相关系数分别为 0.91、0.95,而 F8 与 F10 的相关系数尽高达 0.99。如果在特征选择过程中没有考虑特征之间的相关关系,最终结果将很有可能包含冗余信息,影响最终结论。

为了进一步完善有关 AlPOs 合成参数的分析工作,本文在充分的考虑了特征本身重要程度与特征之间相关关系的前提下,采用最大权重最小冗余算法 (Maximum Weight and Minimum Redundancy, MWMR)[12],深入挖掘各种参数对于 AlPOs 定向合成的影响,为定向合成实验设计提供指导性建议。

1 方法与实验

1.1 相关方法简介

1.1.1 最大权重最小冗余特征选择算法

假设输入数据包含 n 个样本,D 个特征。$W=$

表 1 合成参数描述

Table 1 Description of the synthetic parameters*

ID	Description of parameter	ID	Description of parameter
F1	Molar ratio of Al_2O_3	F12	Second longest distance of organic template
F2	Molar ratio of P_2O_5	F13	Shortest distance of organic template
F3	Molar ratio of solvent	F14	Van der Waals (VDW) volume
F4	Molar ratio of template	F15	Dipole moment
F5	Density of the solvent	F16	n_C/n_N
F6	Melting point of the solvent	F17	$n_N/(n_C+n_N)$
F7	Boiling point of the solvent	F18	N_{charge}/V_{VDM}
F8	Dielectric constant of the solvent	F19	Sanderson electronegativity
F9	Dipole moment of the solvent	F20	Number of free rotated single bond
F10	Polarity of the solvent	F21	Maximal number of protonated H atoms
F11	Longest distance of organic template		

*F1~F4: gel composition parameter; F5~F10: solvent parameter; F11~F21: organic template parameter.

$[w_1,w_2,\cdots w_D]^{\mathrm{T}}\in\Re^{D\times1}$ 表示每个特征重要程度或权重,其中,$w_i>0$,w_i 越大,相应的特征越重要。特征的权重可以使用任意经典的权重度量方法获得。$R\in\Re^{D\times D}$ 表示特征的相关矩阵。MWMR 算法的目的是从原始特征集的 D 个特征中选出 d 个特征形成一个新的特征子集 V,使 V 中 d 个特征的特征权重最大而特征之间的相关性最小。其目标函数定义为:

$$\text{maximize } f(y)=\left(\frac{y^{\mathrm{T}}W}{d}-\frac{y^{\mathrm{T}}Ry}{d(d-1)}\right)\qquad(1)$$

$$\text{subject to } \sum_i y_i=d,y_i\in\{0,1\}$$

其中,$y=[y_1,y_2,\cdots,y_D]^{\mathrm{T}}$ 是指示向量,$y_i=1$(或 0)表示第 i 个特征被选择到 (或没有被选择到)V 中。在公式(**1**)中,第一项表示选择到的 d 个特征的平均权重,第二项表示选择到的 d 个特征的平均相关系数, 而约束项则用来约束选择到的特征子集 V 中的特征为 d 个。因此,通过最大化公式(**1**)中的目标函数就可以保证 V 中所选的 d 个特征最为重要且冗余程度最小。

为了求解容易,放松 MWMR 约束条件,将公式(**1**)转变成公式(**2**):

$$\text{maximize } f(y)=\left(\frac{y^{\mathrm{T}}W}{d}-\frac{y^{\mathrm{T}}Ry}{d(d-1)}\right)\qquad(2)$$

$$\text{subject to } \sum_i y_i=d,y_i\in[0,1]$$

公式(**2**)将对 y 取值的约束放松到[0,1]。y 中元素值的大小代表其所对应的特征被选入 V 的概率。公式(**2**)中的目标公式是最大化一个二次函数,与标准二次规划问题相似。在这里 MWMR 引入类似于文献[13]的成对更新方法来求解公式(**2**)的最大化问题。成对更新策略是一个迭代更新的过程。在成对更新求解策略中,每次迭代更新只更新 y 中两个元素(y_i 和 y_j,$i\neq j$)的值。求解公式(**2**)的成对更新策略定义为:

$$y_l^{\text{new}}=\begin{cases}y_l & l\neq i,\ l\neq j;\\ y_l+\alpha & l=i;\\ y_l-\alpha & l=j;\end{cases}\qquad(3)$$

其中,α 为 y_i 和 y_j 更新的变化。更新后,计算更新 y_i 和 y_j 前后公式(**2**)的差值变化:

$$\begin{aligned}\Delta&=\left(\frac{y^{\text{newT}}W}{d}-\frac{y^{\text{newT}}Ry^{\text{new}}}{d(d-1)}\right)-\left(\frac{y^{\mathrm{T}}W}{d}-\frac{y^{\mathrm{T}}Ry}{d(d-1)}\right)\\&=\frac{(2R_{ij}-R_{ii}-R_{jj})\alpha^2}{d(d-1)}+(r_i(y)-r_j(y))\alpha\end{aligned}\qquad(4)$$

其中 $r_i(y)$ 为 $\frac{W}{y}-\frac{2Ry}{y(y-1)}$ 的第 i 个元素。公式(**4**)表示 Δ 和 α 的函数关系,MWMR 希望每次对 y_i 和 y_j 的更新可以使差值变化最大,这样公式(**4**)就可以快速的接近最大值。根据公式(**4**)和对指示向量 y 的约束条件,α 可以通过公式(**5**)求得:

$$\alpha=\begin{cases}\min(y_j,1-y_i) & \text{if } 2R_{ij}-R_{ii}-R_{jj}\geqslant0,\ r_i(y)>r_j(y)\\ \min(y_j,1-y_i,\ \dfrac{d(d-1)(r_j(d)-r_i(d))}{2R_{ij}-R_{ii}-R_{jj}}) & \text{if } 2R_{ij}-R_{ii}-R_{jj}<0,\ r_i(y)>r_j(y)\\ \min(y_j,1-y_i) & \text{if } 2R_{ij}-R_{ii}-R_{jj}>0,\ r_i(y)=r_j(y)\end{cases}\qquad(5)$$

通过使用公式(**3**)和公式(**5**)迭代更新 y 中成对元素值,就可以求得使公式(**2**)中目标函数最大值的 y[13]。实验验证[12]该求解方法的效率与精度均优于标准二次规划。

1.1.2　Fisher 得分

Fisher 得分[14]是一种依据 Fisher 准则给特征判别能力打分的特征选择方法。Fisher 准则在最大化类间离散程度的同时最小化类内离散程度。第 i 个特征 F_i 的 Fisher 得分定义为:

$$\text{Fisher }(F_i)=\frac{\sum_{j=1}^{T}n_j(m_i^j-m_i)^2}{\sum_{j=1}^{T}n_j(\sigma_i^j)^2}\qquad(6)$$

其中,T 为样本的类别总数,n_j 代表第 j($j=1,\cdots,T$)类样本的样本个数,m_i^j、σ_i^j 和 m_i 表示在第 i 个特征下第 j 类样本的均值、方差和样本的整体均值。公式(**6**)的分母和分子部分分别表示数据在第 i 个特征下各类的类内离散程度和类间离散程度。

1.1.3　Gini 得分

Gini 得分[15]是一种基于 Gini 指数(Gini Index)的特征选择方法。假设样本集 U 属于 T 个不同的类别,则 U 的 Gini 指数定义为:

$$\text{Gini indes }(U)=1-\sum_{i=1}^{T}p_i^2\qquad(7)$$

其中 p_i 是 U 中样本属于第 i 类的概率。Gini 指数也表示集合中样本所属类别的“不纯度”。当集合中所有样本都属于同一个类时,集合的“不纯度”为 0。对于第 i 个特征,遍历特征 F_i 的所有取值,按其不同取值将数据集 U 划分为 T 个子集,集合 U 划分后所有子集的最小 Gini 指数和即是该特征的 Gini 得分。

1.1.4　Relief-F 得分

Relief-F 得分[16]方法的主要思想是:一个重要的特征,可以使同类的样本距离近,而使不同类的样本距离远。根据该思想,每次随机地从原始样本集

中选择一个样本记为 S。Relief-F 得分是根据选中的样本 S 与和它在同一个类别的最近的样本 H(称为 nearest hit)的距离,和与 S 不属于同一类别的其它各个类中与 S 最近的样本(称为 $M(T)$)之间的距离来更新的。因此第 i 个特征 F_i 的权重更新公式如下:

$$\text{Relief F}(F_i)=\text{Relief F}(F_i)-\frac{f(F_i, S, H)}{l}+\sum_{T\neq T_s}\frac{f(F_i, S, M(T))}{l} \quad (8)$$

其中,$f(F_i,S,H)$是计算样本 S 与和 S 同类的最近样本 H 在特征 F_i 下的距离,$f(F_i,S,M(T))$是计算样本 S 和 S 不同类的那些最近邻样本 $M(T)$在特征 F_i 下的距离,l 为随机选择样本的次数。

1.1.5 单边秩和检验

秩和检验由 Wilcoxon 于 1945 年提出[17],是一种常用的假设检验方法。双边秩和检验可以检验 A、B 两组样本是否具有明显差异,而单边秩和检验则可以检验 A 组样本是否明显大于或明显小于 B 组样本。将观察值由小到大按次序排列后所编的次序号称为秩,用秩次号代替原始数据后,所得的某些秩次之和称为秩和,而秩和检验则是用统计量“秩和”进行的假设检验。单边秩和检验的过程如下:(1) 建立检验假设,确定检验水准 α。原假设 H_0 为:两组样本没有明显差异;备择假设 H_1 为:A 样本明显大于(或小于)B 组样本。(2) 把 A 组样本和 B 组样本混合起来,并按数值从小到大顺序编号,每个数据的编号即为它的秩。(3) 分别计算两组样本的秩和。n_1 为样本量较小的样本容量,n_2 是另一组样本的样本容量。Z_1 为样本量较小的一组的秩和,Z_2 为另外一组的秩和。(4) 确定统计量 Z:若 $n_1\neq n_2$,则 $Z=Z_1$;若 $n_1=n_2$,则 $Z=Z_1$ 或 $Z=Z_2$。(5) 根据检验统计量 Z,确定 p 值。如果 p 值小于或等于临界值 α,则原假设被拒绝。

1.2 实验数据

含有 8 元环结构的磷酸铝分子筛是比较典型的小孔分子筛,孔径尺寸大概处于 0.38~0.4 nm 之间,可被用于催化和气体分离[18-19]。开放骨架磷酸铝合成反应数据库大约包含 1 700 条合成反应数据。去除数据库中含有缺失项较多的数据后,本文使用剩余的 1 279 条磷酸铝合成反应数据作为实验样本。其中,365 条数据包含(8,6)元环结构,即该类开放骨架磷酸铝结构既包含 8 元环结构又包含 6 元环结构,如图 1 所示。本文选取 21 个合成参数(或特征)进行分析,如表 1。

文献[7]认为,凝胶组成是开放骨架磷酸铝合成至关重要的参数,因此文献[7]将凝胶组成参数作为分类器的默认输入,即在考量某参数对于数据的分类效果时,凝胶组成参数默认与待考量参数一起对数据进行分类。本文沿用文献[7]的参数取舍方法,将表 1 中 4 个凝胶组成参数作为分类模型的默认输入,而只具体分析其它 17 个合成参数。

1.3 实验过程

实验的样本容量为 1 279,样本维数为 17 维。在本文中,样本维数指描述每个合成样本的合成参数个数,即表 1 中 F5~F12 每个参数为一个维度。使用不同特征选择方法从原始特征集中选择 1 到 17

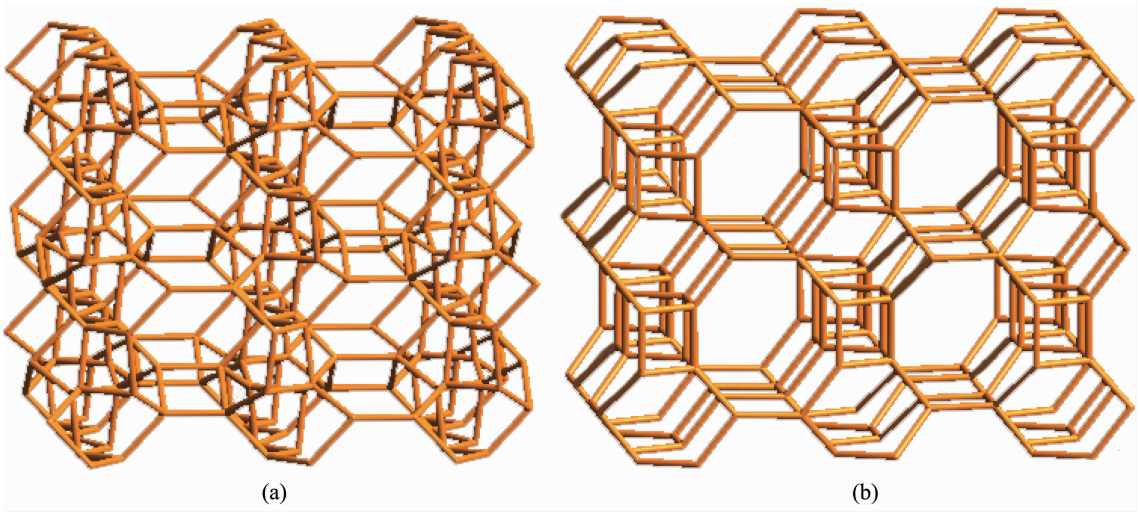

图 1 含有(8,6)元环结构的 AlPOs 举例: (a) 骨架结构为 AEN 的 AlPOs; (b) 骨架结构为 AWO 的 AlPOs

Fig.1 Examples of (8,6)-ring-containing AlPOs: (a) AEN-zeotype AlPOs; (b) AWO-zeotype AlPOs

维特征,即可以得到包含 1 到 17 维特征的 17 个特征子集。使用每个特征子集对数据进行分类,获得最好分类效果的特征子集则是最优特征子集。由于最优特征子集达到了对含有(8,6)元环结构 AlPOs 样本和其它样本分类的最好效果,因此其包含的特征对于该类合成的影响是较大的。

实验使用十折交叉验证 (10-fold Cross Validation)的方式,即将数据集分成 10 份(每份的样本容量大约为 128),轮流将其中 9 份(样本容量大约为 1 152)做训练,剩余的 1 份(样本容量大约为 128)做验证, 而最终根据 10 次结果的均值作为对算法分类精度的估计。在分类过程中,含有(8,6)元环 AlPOs 的样本容量为 365, 其它样本的样本容量为 914,两类样本的样本容量较为悬殊。因此,本文采用对类不平衡问题较不敏感的最近邻分类器 (Nearest Neighbor,NN)[20] 和支持向量机 (Support Vector Machine,SVM)[21]作为分类模型。实验中,SVM 的核函数采用径向基函数,因此需要优化的主要参数为惩罚系数 C 和径向基函数参数 γ。参数 C 和 γ 采用网格搜索法进行优化,即设定 $C\in[C_1,C_2]$,步长为 l_c,$\gamma\in[g_1,g_2]$, 步长为 l_g。然后使 SVM 遍历每对 $\{C',\gamma'\}$的取值,用训练样本训练 SVM 分类器,并用训练好的分类器对测试集分类,获得最好分类效果的参数被确定为最优参数。分类效果评价指标采用分类正确率(Acc-Rate)和 F-measure[22]。在 MWMR 算法中, 特征的重要程度分别采用 Fisher 得分、ReliefF 得分和 Gini 得分度量, 特征之间相关程度采用相关系数度量。MWMR-Fisher、MWMR-ReliefF、MWMR-Gini 分别表示以 Fisher 得分、ReliefF 得分、Gini 得分度量特征权重,以相关系数度量特征之间相关程度的MWMR。

2　结果与讨论

2.1　实验结果

我们将 MWMR-Fisher、MWMR-ReliefF、MWMR-Gini 与经典的 Fisher 得分、ReliefF 得分、Gini 得分对于含有(8,6)元环结构 AlPOs 特征选择的效果进行比较,并且比较以不同权重度量方法度量特征权重的 MWMR 的实验结果。

具体实验结果如下:(1) 通过比较 MWMR 与三种经典特征选择方法可以发现,由于考虑了特征之间的相关性,MWMR 取得了较其相应经典方法更好的分类效果;(2) 通过比较以不同权重度量方法度量特征权重的 MWMR 可以发现,MWMR-Fisher 选择 9 维特征、采用最近邻分类器达到了对数据进行分类的最高 Acc-Rate 90.89%和 F-measure 0.84。(3) 从以上实验结果可以看出,MWMR-Fisher 在选择 9 维特征时获得的特征子集可能对该类结构的合成具有较大的影响。根据实验结果我们得出对于含有(8,6)元环结构 AlPOs 合成较为重要的特征子集是:{F6,F9,F11,F12,F14,F15,F16,F17,F19}。

由于 MWMR 在特征选择过程中, 同时考虑了特征本身的重要程度和特征之间的相关程度,因此其取得了较好的实验效果。为了衡量每种方法所选出的最优特征子集所包含特征的相关程度,我们计算了各方法所选最优特征子集中每对特征之间相关系数的算数平均数, 即平均相关系数 (如表 2 所示)。从表 2 可以看出,MWMR 选择的最优特征子集的平均相关系数均要低于其相应的经典特征选择方法(如 MWMR-Fisher 和 Fisher)。

表 2　平均相关系数

Table 2　Mean of correlations among the optimal features

	NN	SVM
ReliefF	0.31	0.31
MWMR-ReliefF	0.30	0.29
Gini	0.32	0.32
MWMR-Gini	0.18	0.25
Fisher	0.38	0.47
MWMR-Fisher	0.29	0.20

2.2　与文献已有工作的比较

文献[11]中,无机化学分子工程学研究者根据经验知识对含有(8,6)元环结构 AlPOs 的合成参数做了一系列分析和验证。我们比较了本文得出的最优特征子集与文献[11]结论中的最优特征子集对数据的分类能力。当采用最近邻分类器作为分类模型时,文献[11]和 MWMR-Fisher 所选特征得到的 Acc-Rate、F-measure 分别是 85.65%、0.74 和 90.89%、0.84;当采用支持向量机作为分类模型时,文献[11]和 MWMR-Fisher 所选特征得到的 Acc-Rate、F-measure 分别是 84.13%、0.68 和 90.30%、0.82。从实验结果可以看出,MWMR-Fisher 选择的最优特征子集可以获得较文献[11]更好分类效果。

为了验证 MWMR-Fisher 选择的最优特征子集对于数据进行分类的优势,下面采用单边秩和检验验证 MWMR-Fisher 选出的最优特征子集在两种分

表 3 秩和检验的 p 值

Table 3 p-value of the rank sum test

	Classifier	p-value
AccRate	MWMR-Fisher vs [11] (NN)	0.002
	MWMR-Fisher vs [11] (SVM)	2.4836e–04
F-measure	MWMR-Fisher vs [11] (NN)	0.0016
	MWMR-Fisher vs [11] (SVM)	1.4181e–04

类模型下获得的 Acc-Rate 和 F-measure 是否明显高于文献[11]。在这个假设检验中,原始假设 H_0 为:采用 MWMR-Fisher 与文献[11]选择的最优特征子集对数据进行分类获得的 Acc-Rate 或 F-measure 没有明显差异,备择假设 H_1 为:采用 MWMR-Fisher 选择的最优特征子集对数据进行分类所获得的 Acc-Rate 或 F-measure 明显高于文献[11]。实验中,显著性水平 α 设为 0.05,表 3 列出了单边秩和检验结果。

从表 3 可以看出,在采用最近邻分类器和支持向量机作为分类模型时,p 都小于 0.05。因此本文的结论明显优于文献[11]。文献[11]仅仅从经验知识角度研究了含有(8,6)元环结构 AlPOs 的合成参数,并没有从数据本身及方法模型上做分析。因此,文献[11]选出的特征子集中特征数量较少,并不能完全涵盖对合成起重要作用的特征。

2.3 结果分析

MWMR-Fisher 选择 9 维特征、采用最近邻分类器时,可以获得对于含(8,6)元环结构 AlPOs 预测的最佳效果。因此,根据实验结果可以推断:溶剂的熔点、溶剂的偶极距、有机模板的最长距离、有机模板的次长距离、模板剂分子空间体积、模板剂分子极性、模板剂中 C 原子和 N 原子的个数比、模板剂中 N 原子与 C 加 N 原子个数比以及模板剂分子 Sanderson 电负性可能对该类结构的合成产生较为重要的作用。

MWMR 算法在选择一维特征时,仅仅考虑特征的重要程度,因此在第一维选择的特征是最重要的。当选择二维特征时,MWMR 同时考虑待选特征集中特征的重要性和待选特征与已选特征之间的相关关系,因此在第二维新进入最优特征子集的特征是第二重要的特征。以此类推,在遍历的选择 1~d 维特征时(d 为最优子集包含的特征个数),我们认为特征进入最优特征子集的顺序代表其相应的重要程度。那么,由 MWMR-Fisher 获得的最优特征子集将形成一个按由重要性从大到小降序排序的序列:F11,F16,F9,F19,F15,F6,F12,F17,F14。

从这个序列可以看出,有机模板剂的最长距离(F11)是最为重要的一个合成参数。使用该参数在最近邻分类器下对数据进行分类,Acc-Rate 可达 88.01%。模板剂中 C 原子和 N 原子的个数比(F16)这个参数在序列中排位第二,显示其重要程度仅次于有机模板剂的最长距离。而在观察 MWMR-Fisher 遍历选择 1 到 17 维特征的实验结果时发现,当在第二维 F16 加入最优特征子集后,Acc-Rate 曲线呈现出了非常明显的上升(从 88.01%到 89.87%)。模板剂中 C 原子和 N 原子的个数比(F16)这个参数描述的是模板剂分子的亲水性和疏水性,因此,可以推断模板剂分子的亲水性和疏水性对于该类结构的合成可能有较大的影响。排在重要性序列第三位的是溶剂的偶极距(F9),由此可以看出溶剂的极性参数也是较为重要的。而在化合实验中,溶剂极性的变化确实能导致最终产物的改变。其次,对该类结构形成影响较大的特征依次为模板剂分子 Sanderson 电负性(主要是分布在 N 原子上的电荷)(F19)、模板剂分子极性(F15)和溶剂的沸点(F6)等。从这个序列我们也可以看出,重要特征中共包含了 3 个模板剂的几何参数(有机模板的最长距离(F11),有机模板的次长距离(F12)和模板剂分子空间体积(F14)),因此,我们推断有机模板剂的几何属性对于该类结构的合成可能有着至关重要的作用。

3 总 结

本文将 MWMR 算法应用到开放骨架磷酸铝合成参数的分析问题当中。实验中,首先比较了采用不同特征权重度量方法的 MWMR 与相应过滤式特征选择方法对于开放骨架磷酸铝特征选择的效果,然后将本文的工作与有关开放骨架磷酸铝参数分析的已有文献工作做了对比。通过实验和对比,充分地验证了该算法在开放骨架磷酸铝合成反应数据库合成参数分析中的有效性,并挖掘了合成参数对于定向合成含有(8,6)元环结构开放骨架磷酸铝的影响,为其定向合成提供指导。

参考文献:

[1] XU Ru-Ren(徐如人), PANG Wen-Qin(庞文琴), YU Ji-Hong(于吉红), et al. *Chemistry-zeolite and Porous Materials*(分子筛与多孔材料化学). Beijing: Science Press, **2004**:1-23

[2] YAN Yan(颜岩), LI Ji-Yang(李激扬), QI Miao(齐妙), et al. *Sci. China, Ser. B Chem.*(中国科学 B 辑:化学), **2009**,**39**(11):1308-1313

[3] http://zeobank.jlu.edu.cn/

[4] Han J, Kamber M. *Data Mining*: *Concepts and Techniques*. San Francisco: Morgan Kaufman, **2001**.

[5] Witten H, Frank E. *Data Mining*: *Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufman, **2005**.

[6] Cord M, Cunningham P. *Machine Learning Techniques for Multimedia*. Berlin Heidelberg: Springer, **2008**:91-112

[7] Li J Y, Qi M, Kong J, et al. *Microporous Mesoporous Mater.*, **2010**,**129**:251-255

[8] HUO Wei-Feng(霍卫峰), GAO Na(高娜), YAN Yan(颜岩), et al. *Acta Phys. Chim. Sin.*(物理化学学报), **2011**,**27**(9):2111-2117

[9] Yao M H, Qi M, Li J S, et al. *Microporous Mesoporous Mater.*, **2014**,**186**:201-206

[10]Qi M, Li J S, Wang J Z, et al. *Ind. Eng. Chem. Res.*, **2012**, **51**(51):16734-16740

[11]Gao N, Yan Y, Li J S, et al. *Microporous Mesoporous Mater.*, **2014**,**195**:174-179

[12]Wang J Z, Wu L S, Kong J, et al. *Pattern Recognit.*, **2013**, **46**:1616-1627

[13]Liu H R, Yang X W, Latecki L J, et al. *Int. J. Comput. Vision*, **2012**,**98**(1):65-82

[14]Fisher R A. *Ann. Eugenics*, **1936**,**7**(2):179-188

[15]Gini C. *Variabilita e mutabilita*. Bologna: Tipografia di Paolo Cuppini, **1912**.

[16]Kononenko I. *Proceedings of the 7th European Conference in Machine Learning*. Berlin: Springer, **1994**:171-182

[17]Wilcoxon F. *Biometrics Bulletin*, **1945**,**1**(6):80-83

[18]Lewis D W, Sankar G, Wyles J K, et al. *Angew. Chem. Int. Ed. Engl.*, **1997**,**36**(23):2675-2677

[19]Padin J, Rege S U, Yang R T. *Chem. Eng. Sci.*, **2000**,**55**(20):4525-4535

[20]Cover T M, Hart P E. *IEEE Trans. Inf. Theory*, **1967**,**13**(1):21-27

[21]Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer, **1995**.

[22]Rijsbergen C. *Information Retrieval*. London: Butterworths, **1979**.

[23]Hall M. 17*th International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, **2000**:359-366