

# 用人工神经网络预测无限稀释活度系数

王秀丽 孙 伟 闫 岩 宋海华\*

(天津大学化工学院 天津 300072)

**摘 要** 用 BP 神经网络, 以分子连接性指数等作为输入参数, 对各种溶剂的无限稀释活度系数进行预测, 平均相对误差约 9.25%。与 UNIFAC 法比较, 人工神经网络方法要优于 UNIFAC 法。

**关键词** 活度系数 分子连接性指数 人工神经网络

## Infinte Dilution Activity Coefficient Predictions Via Artificial Neural Network

Wang Xiuli, Sun Wei, Yan Yan, Song Haihua\*

(School of Chemical Engineering, Tianjin University, Tianjin 300072)

**Abstract** The infinite dilution activity coefficients of solvents are calculated by using BP neural network and molecular connectivity index, and the average relative error is within 9.25%. The prediction with artificial neural network is more accurate than that with UNIFAC model.

**Key words** Activity coefficient, Molecular connectivity indice, Artificial neural network

溶质在溶剂中的无限稀释活度系数是化工分离过程的重要热力学性质之一。除实验测定外, 常用的计算方法是基团贡献法<sup>[1]</sup>。基团贡献法假定分子中一个基团的贡献基本上不受其它基团的影响, 只体现分子加和性, 具有一定的局限性。人工神经网络(ANN)技术作为一种新型的信息处理和计算系统, 在模式识别和非线性函数关系等领域取得了显著成功。利用 ANN 建立物质的定量结构-性质关系(QSPR), 预测物质的理化性质, 如沸点<sup>[2]</sup>、气液平衡数据<sup>[3]</sup>, 都取得了良好的结果。本文利用 ANN 技术, 建立分子连接性指数等参数与无限稀释活度系数之间的定量结构-性质关系。

### 1 理论部分

#### 1.1 分子描述参数的选择

无限稀释活度系数是与分子结构有关的热力学性质, 因此分子描述参数的选择既要体现分子的立体结构, 又要体现分子的电子结构。基于以上考虑, 选择分子连接性指数、偶极距和摩尔体积作为描述参数。

分子连接性指数是 Kier 等<sup>[4]</sup>根据拓扑理论在 Randic 的分子分枝指数的基础上发展起来的, 一般可表示为:

---

王秀丽 女, 31 岁, 博士生, 现从事化工分离研究。\*联系人

2002-05-13 收稿, 2002-09-05 修回

$${}^mX_t^v = \sum_{j=1}^{n_m} \prod_{i=1}^{m+1} (d_i^v)_j^{-\frac{1}{2}} \quad (1)$$

$t$  表示子图的类型,  $m$  表示指数  $X$  的阶,  $n_m$  表示阶为  $m$  的  $t$  类型子图的数目,  $j$  是子图序号,  $d_i^v$  表示第  $i$  个原子的点价, 上标  $v$  表示不饱和度、杂原子校正。

由上式可以得到分子的各阶指数, 通过神经网络筛选参数, 即由输入层和隐含层之间的权重来决定描述符的取舍, 最终保留了 4 个特征参数:  ${}^1X_p^v$ ,  ${}^2X_p^v$ ,  ${}^3X_p^v$ ,  ${}^4X_{pc}^v$ 。其中  ${}^1X_p^v$  与分子的大小和表面积有关,  ${}^2X_p^v$  可有效区分同分异构体,  ${}^3X_p^v$  与分子的弯曲程度有关,  ${}^4X_{pc}^v$  与取代基的长度和类型有关。

由于采用分子连接性指数作为输入参数难以全面地描述分子特征, 因此选择偶极距作为电子结构参数, 用于反映分子的局部原子或基团的性质; 选择摩尔体积作为立体参数, 用于描述分子的立体效应。

## 1.2 BP 神经网络

本文采用 BP 神经网络模型, 以对数 S 型(logsig)函数为传递函数, 学习算法采用自适应自动调节的 Levenberg-Marquardt 优化方法。网络以训练集的均方差(MSE)小于  $10^{-3}$  或训练次数达到 10000 作为收敛判据。输入层有 13 个节点, 分别对应于溶质和溶剂的分子连接性指数、偶极距、分子的摩尔体积和温度。输出层节点数为 1, 对应于无限稀释活度系数。由于使用一个隐层和足够多的节点可以实现任意精度的输入-输出映射关系, 所以仅选择一个隐层, 而隐层节点数则通过优选确定为 4 个。所用程序用 MATLAB 语言编写。

## 1.3 数据集的组成

表 1 数据集中溶质的分类

Tab.1 Classes of solutes in data library

类别	化合物
烃	戊烷, 己烷, 庚烷, 辛烷, 壬烷, 2-甲基戊烷, 异辛烷, 2,4-二甲基戊烷, 2,5-二甲基己烷, 2,3,4-三甲基戊烷, 2,2,4-三甲基戊烷, 环己烷, 乙基环己烷 戊烯, 1-己烯, 1-辛烯, 环己烯 苯, 甲苯, 对二甲苯, 邻二甲苯, 间二甲苯, 乙苯 二氯甲烷, 氯仿, 二氯乙烯, 氯苯
醇	甲醇, 乙醇, 1-丙醇, 2-丙醇, 1-丁醇, 2-丁醇, 2-甲基-1-丙醇, 2-甲基-2-丙醇
醚	乙醚, 异丙醚, 甲基叔丁醚, 乙基叔丁醚, 异戊醚
醛	甲醛, 丁醛, 戊醛
酮	丙酮, 2-丙酮, 2-丁酮, 2-戊酮, 3-戊酮
酯	乙酸甲酯, 乙酸乙酯, 乙酸丁酯
其它	水, 14-二氧杂环己烷, 噻吩, 吡啶

表 2 数据集中溶剂的分类

Tab.2 Classes of solvents in data library

类别	化合物
烃	丁烷, 戊烷, 己烷, 庚烷, 辛烷, 壬烷, 癸烷, 十二烷, 十三烷, 十六烷, 2,2,4-三甲基戊烷, 2,3,4-三甲基戊烷, 环己烷, 甲基环己烷 1-辛烯, 1-己烯, 1-癸烯 苯, 甲苯, 乙苯, 1,2-二甲苯, 1,3-二甲苯, 1,4-二甲苯 氯仿, 四氯化碳, 1,2-二氯乙烯, 氯苯

醇	甲醇, 乙醇, 丁醇, 辛醇, 1-丙醇, 2-丙醇, 1-丁醇, 1-戊醇, 1-辛醇, 苯乙醇, 乙二醇, 二甘醇
酮	丙酮, 2-丁酮, 2-戊酮, 2-己酮, 2-庚酮, 环己酮
酸	甲酸, 1,2-亚丙基碳酸
酯	乙酸甲酯, 乙酸乙酯, 乙酸丁酯, 三乙酸丙酯, 苯甲酸乙酯, 己内酯, 邻苯二甲酸二乙酯, 邻苯二甲酸二乙酯
酰胺	二甲基乙酰胺, <i>N</i> -甲基甲酰胺, <i>N</i> -甲基乙酰胺, <i>N</i> -乙基甲酰胺, <i>N,N</i> -二甲基甲酰胺, <i>N,N</i> -二甲基乙酰胺, <i>N,N</i> -二乙基乙酰胺
其它	硝基甲烷, 硝基丙烷, 硝基苯, 苯甲腈, 戊二腈, 二甲亚砜, 环丁酮, 乙醛, 苯甲醚, 苯酚, 三乙基胺, 咪喃, 吡啶, 1,4-二氧杂环己烷, <i>N</i> -甲基吡咯烷酮, <i>N</i> -甲基-2-吡啶酮, 2-吡咯烷酮, <i>N</i> -甲酰基吗啉

数据集由 55 个溶质, 80 个溶剂的分子连接性指数、偶极距、分子摩尔体积、温度、无限稀释活度系数组成, 共 1365 组数据。这些数据分为训练集、有效验证集和预测集三部分, 分别包括 1178、134、53 组数据。为保证预测的精度, 将溶质和溶剂按官能团分类(见表 1~2), 每一个集合的数据在选择时尽量包含所有种类溶质和溶剂的组合类型。分子连接性指数由式(1)计算得到, 分子的偶极距、摩尔体积由文献[5~7]查找, 无限稀释活度系数来自文献[8~27], 测定条件为  $1.01 \times 10^5$  Pa, 283.15~373.15K。

## 2 结果和讨论

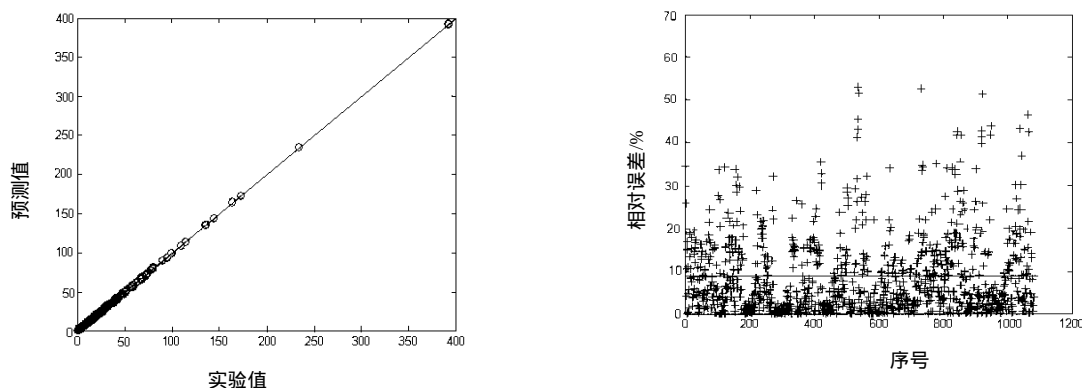


图 1 训练集中  $g^\infty$  实验值与预测值的关系及相对误差

Fig.1 Relation of experimental and predicted  $g^\infty$  and the relative errors in training set

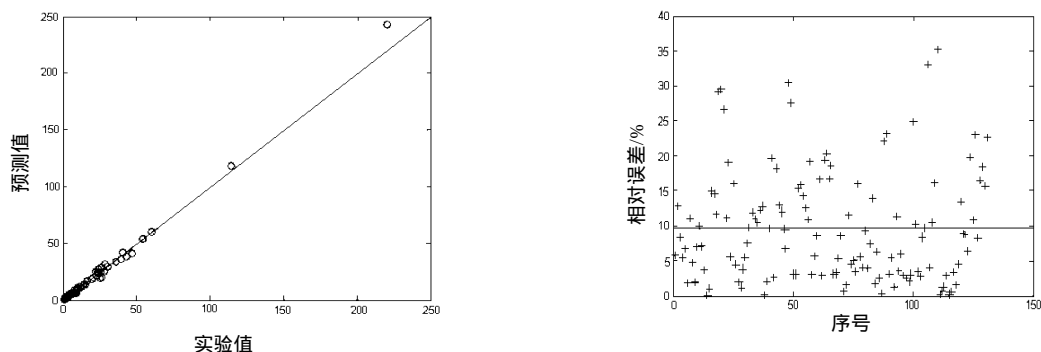


图 2 有效验证集中  $g^\infty$  实验值与预测值的关系及相对误差

Fig.2 Relation of experimental and predicted  $g^\infty$  and the relative errors in validation set

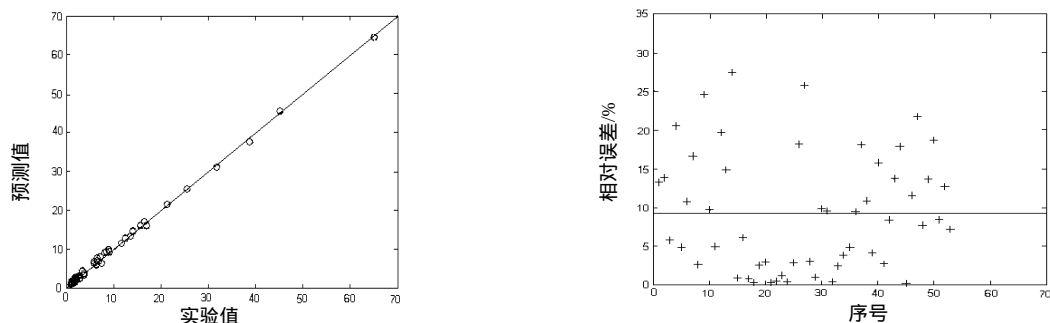
图 3 预测集中  $g^{\infty}$  实验值与预测值的关系及相对误差Fig.3 Relation of experimental and predicted  $g^{\infty}$  and the relative errors in test set

图 1~图 3 所示分别为无限稀释活度系数在训练集、有效验证集、预测集的预测效果。从图中可以看出, 实验值和预测值的相关性较好, 平均相对误差均在 9% 左右(见表 3)。但是, 仍存在相对误差偏高的数据。对比三个集合发现, 在训练集中相对误差较大数据对应的溶质-溶剂组合类型, 相应地在有效验证集和预测集中, 溶质在该种类溶剂中  $g^{\infty}$  的相对误差也较大。如以 *N*-甲基吡咯烷酮、*N*-甲酰基吗啉为溶剂的数据组, 在三个集合中的相对误差分别在 45%、25%、20% 左右。对训练集中误差较大的数据组进行分析, 发现主要原因是因为训练集中该类的样本数目较少, 神经网络在训练过程中未能完全提取其特征。随着可以得到的数据的补充, 网络的预测性能会逐渐改进, 预测误差会逐渐减小。

表 3 ANN 预测结果

Tab.3 Errors predicted by ANN

数据集名称	平均绝对误差	平均相对误差/%
训练集	0.3184	8.82
有效验证集	0.8985	9.64
测试集	0.4086	9.25

### 3 ANN 方法与 UNIFAC 比较

为了说明计算结果的准确性, 特选经典的 UNIFAC<sup>[1]</sup>(Universal Quasi-Chemical Functional Group Activity Coefficient)法与之比较。用 ANN 方法与 UNIFAC 方法计算几种烷烃在二甲基甲酰胺中的无限稀释活度系数, 得到的相对误差见表 4。可以看出, 用 ANN 方法计算无限稀释活度系数要优于 UNIFAC 法。

表 4  $g^{\infty}$  的实验值与预测值比较Tab.4 Comparison between experimental and predicted  $g^{\infty}$  values

溶质	实验值	相对误差/%	
		UNIFAC	本文
正己烷	17.11	14.1	0.5
正辛烷	30.97	4.0	0.16
正壬烷	40.94	1.9	0.07
正癸烷	62.52	8.4	9.7
平均值		7.1	2.61

## 4 结论

用神经网络来表达分子结构和无限稀释活度系数之间的非线性关系是可行的, 由于可以得到的数据有限, 预测结果不是很理想。同为结构-性质关系, 人工神经网络以其在表达非线性问题方面的特定优势, 要优于从统计理论计算体系无限稀释活度系数的 UNIFAC 方法。

## 参考文献

- [1] G J Pieretti, C H Deal, E L Derr. Ind. Eng. Chem., 1959,51(1): 95~99.
- [2] G Espinosa, D Yaffe, Y Cohen et al. J. Chem. Inf. Comput. Sci., 2000,40(3):859~879.
- [3] R Sharma, O Singhal, R Ghosh et al. Comput. Chem. Eng., 1999, 23(3): 385~390.
- [4] L B Kier, L H Hall. New York: Academic Press, 1976:8.
- [5] J A Dean. Lange's Handbook of Chemistry. 15th Edition. New York: McGraw-Hill, 1999:1.76~1.342,5.105~5.129.
- [6] 卢焕章. 石油化工基础数据手册. 北京:化学工业出版社, 1984:138~952.
- [7] 马沛生. 石油化工基础数据手册续编. 北京:化学工业出版社, 1993:158~1262.
- [8] C B Castells, D I Eikens, P W Carr. J. Chem. Eng. Data, 2000,45(2): 369~375.
- [9] U Weidlich, J Gmehling. J. Chem. Eng. Data, 1987,32(2): 138~142.
- [10] U Weidlich, H J Rohm, J Gmehling. J. Chem. Eng. Data, 1987,32(4): 450~453.
- [11] C Knoop, D Tiegs, J Gmehling. J. Chem. Eng. Data, 1989,34(2): 240~247.
- [12] M Schiller, J Gmehling. J. Chem. Eng. Data, 1992,37(4): 503~508.
- [13] C Möllmann, J Gmehling. J. Chem. Eng. Data, 1997, 42(1): 35~40.
- [14] D Gruber, D Langenheim, J Gmehling. J. Chem. Eng. Data, 1997,42(5): 882~885.
- [15] D Gruber, D Langenheim, J Gmehling. J. Chem. Eng. Data, 1998,43(2): 226~229.
- [16] D Gruber, M Topphoff, J Gmehling. J. Chem. Eng. Data, 1998,43(6): 935~940.
- [17] M Topphoff, D Gruber, J Gmehling. J.Chem.Eng.Data, 1999,44(6): 1355~1359.
- [18] M Topphoff, D Gruber, J Gmehling. J. Chem. Eng. Data, 2000,45(3): 484~486.
- [19] M Krummen, D Gruber, J Gmehling. J. Chem. Eng. Data, 2000,45(5): 771~775.
- [20] Y G Dobrjakov, I M Balashova, G Maurer. J. Chem. Eng. Data, 2000,45(2): 185~193.
- [21] M Juang, G K Morgan, D W Arnold. J. Chem. Eng. Data, 1989,34(2): 161~165.
- [22] G Tse, S I Sandler. J. Chem. Eng. Data, 1994,39(2): 354~357.
- [23] G Hradetzky, M Wobst, H Vopel et al. Fluid Phase Equilibria, 1990,54: 133~145.
- [24] L Dallinga, M Schiller, J Gmehling. J. Chem. Eng. Data, 1993,38(1): 147~155.
- [25] A Bermudez, G Foco, S B Bottini. J. Chem. Eng. Data, 2000,45(6): 1105~1107.
- [26] A Nikolie, D Vastag, M R Tarjani et al. J. Chem. Eng. Data, 1994,39(3): 618~620.
- [27] G Foco, A Bermudez, S Bottini. J. Chem. Eng. Data, 1996,41(5): 1071~1074.