浙江科技学院学报，第 25 卷第 3 期，2013 年 6 月
Journal of Zhejiang University of Science and Technology
Vol. 25 No. 3, Jun. 2013

doi：10.3969/j.issn.1671-8798.2013.03.005

证据理论组合模型用于三维荧光光谱水质分析

李津蓉，武晓莉
（浙江科技学院 自动化与电气工程学院，杭州 310023）

摘　要：为了充分利用三维荧光光谱信息，提高分析模型对水质综合指标的预测精度及稳健性，提出了一种基于D-S(Dempster-Shafe)证据理论的模型组合方法。传统的模型组合方法主要依赖于对某个单一的预测性能评价指标进行优化来确定组合权值，没有对多个性能指标进行综合性考虑。将三维荧光光谱看作多个不同波长的激发光照射下的发射光谱，基于每个发射光谱建立子模型，将多个性能指标作为证据，确定子模型的概率分配函数，然后基于 DS 证据合成理论对子模型的可信度进行评价并确定组合权值。实验以 40 个不同来源的地表水样本为研究对象，建立组合分析模型，对总有机碳值(TOC)进行定量分析。实验证明，组合模型相对于子模型而言各个评价指标均有较大提高，充分说明了该方法的有效性。

关键词：三维荧光光谱；证据理论；总有机碳；定量分析
中图分类号：O433.4；R123.1　　**文献标志码**：A　　**文章编号**：1671-8798(2013)03-0180-06

Water quality analysis by three-dimensional fluorescence spectra based on multi-models combined by evidential theory

LI Jinrong, WU Xiaoli
(School of Automation and Electrical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China)

Abstract: A multi-models combining method based on D-S (Dempster-Shafe) evidential theory was proposed to optimize the use of three-dimensional fluorescence spectral information so as to improve the precision and robustness of water quality analytic model. The combined weights in traditional combined method for forecasting are obtained with optimizing a single prediction performance index, with no considering of different aspects of the prediction model. The three-dimensional fluorescence spectrum of a sample can be taken as a number of emission spectra issuing from exciting light with different wavelength. So multiple analytic sub-models were built based on emission spectra and several performance indexes were used to provide evidences

收稿日期：2013-04-01
基金项目：浙江省教育厅科研计划项目(Y201018267)
作者简介：李津蓉(1977—　)，女，天津市人，讲师，博士研究生，主要从事光谱信号分析与处理研究。

from different views for calculating the probability assignment functions of sub-models. And then the credibility of every sub-model was evaluated based on D-S theory to determine the combinational weight. Forty surface water samples of variable origin and total organic carbon (TOC) value were used as research objects and combined model was built to predict the TOC values. The experimental results showed that the combined model improved appreciably at different performance indexes compared with the sub-models.

Key words: 3-D fluorescence spectra; evidential theory; total organic carbon; quantitative analysis

自然水体中有机物的种类繁多，组成复杂，且浓度变化范围很大，在实际中通常采用生化需氧量(BOD)、化学需氧量(COD)和总有机碳(TOC)等综合指标来对水体中的有机物总量进行描述。三维荧光光谱技术作为一种非破坏性、高灵敏度的快速检测技术，近年来在水质污染分析领域中得到广泛的研究与应用。目前，大多数基于三维荧光的水质分析研究均针对相同来源的水样本[1-4]，相同来源意味着相同或相似的溶解有机物(DOM)组成，且 DOM 的浓度仅在较小范围内变化，这使得有机物综合指标与某些位置的荧光强度或荧光强度之和呈现较强的线性关系[5-6]。但不同来源水体中有机物的种类变化多样，且浓度变化范围很大，这给基于荧光技术的水质分析带来较大困难，使得水质综合指标的定量分析模型的精确性难以保证[7]。因此，如何充分且合理地利用三维荧光信息是保证分析模型精确度的关键问题。

组合模型方法首先建立多个不同的子模型，然后通过对各子模型的预测结果进行一定的加权求和得到最终的预测结果。其优势在于可以通过不同的模型来考虑不同的影响因素，进而从不同的角度进行建模预测，达到充分利用信息的目的，而难点在于如何确定其组合权值。

组合模型的预测性能应优于每个子模型，而评价一个预测模型的性能包括多方面指标，如预测均方误差、复相关系数、最大误差及误差方差等。线性组合的组合回归法[8]和岭回归法[9]是通过最小化组合预测误差来确定组合权值；方差－协方差优选组合预测法[10]是通过最小化组合预测误差的方差来确定组合权值。这些方法在确定组合权值时仅考虑了模型某一个方面的性能优化，若要从多方面保证组合模型的预测性能得到提升，则需要在计算组合权值的过程中对多个性能指标进行同步优化。证据理论是一种处理不确定性问题的完整理论体系，利用经验和知识对不同证据建立信任度函数，再由其合成公式对不同的证据进行综合，从而达到对信息从多角度全面利用的目的。因此，这种方法在专家系统、信息融合等领域中得到了广泛应用[11-13]。

基于以上考虑，本研究通过对三维荧光光谱数据中不同激发波长下的发射光谱建立多个子模型，然后利用多个不同的预测模型评价指标来获取每个子模型的信任度，最后利用证据理论的数据融合原理对子模型的可信度进行综合处理，得到组合权值，进而建立组合模型预测水质综合指标 TOC。

1　实验样品与仪器

用于水质分析的 40 个水样采集于某市不同河流及湖泊。根据地表水环境质量标准(GB 3838—2002)要求，将所有采集后水样自然沉降 30 min 后取上层非沉降部分进行分析。对每一个水样都分别测量其 TOC 参考值和三维荧光光谱。三维荧光光谱测定在 HITACHI 公司的 F-4500 型荧光光谱仪上进行，采用激发波长范围 λ_{Ex}＝220～400 nm，激发波长间隔 10 nm，激发狭缝宽度 2.5 nm；发射波长范围 λ_{Em}＝250～700 nm，间隔 5 nm，发射狭缝宽度 5.0 nm；波长扫描速度 2 400 nm/min；光电倍增管电压 700 V。采用 SHIMADZU 公司的 TOC-VCSH 型总有机碳分析仪测定每个水样的 TOC 分析值。图 1 给出了其中两个水样的三维荧光指纹图，图 2 给出了 40 个水样的 TOC 分析值的变化曲线。从图 1 和图 2可以看出，由于水样的来源不同，三维荧光峰的个数、位置及强度均有较大差别，同时 TOC 值的变化范围也很大。

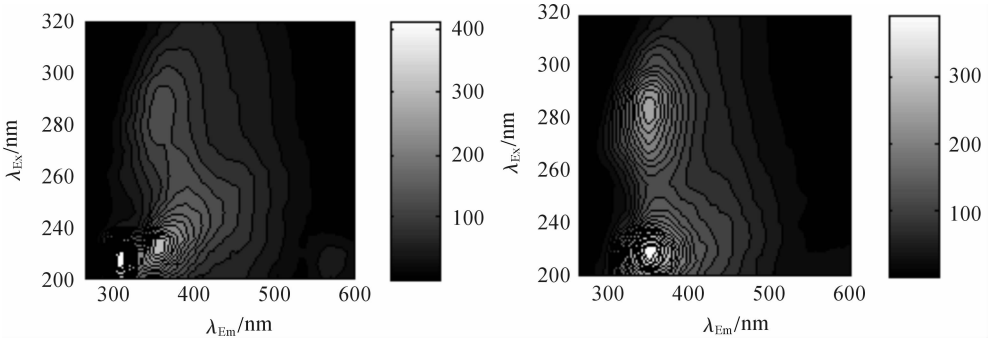


图 1 两个水样本的三维荧光

Fig. 1 EEM fluorescence spectra of 2 water samples

2 组合预测模型的建立

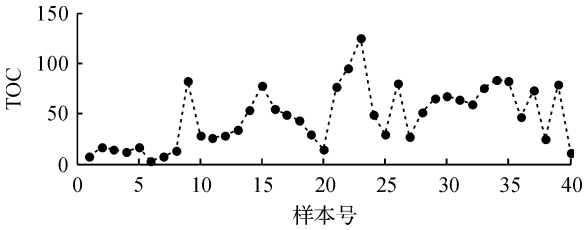


图 2 40 个水样本的 TOC 实验值

Fig. 2 TOC changing curve of 40 water samples

每个水样的三维荧光数据可看作多条在不同波长的激发光照射下的发射光谱，令 $S_i(\lambda_j)$ 表示第 i 个水样在波长 λ_j 的激发光照射下的发射光谱，T_i 表示第 i 个水样的 TOC 化验值。

步骤一：子模型的建立。假设训练样本集容量为 n，每个样本的 EEM 矩阵维数为 $k \times l$，k 表示激发光波长个数，l 表示发射光波长个数。对 n 个训练样本，分别采用 k 个激发波长下的荧光发射光谱及相应的化学分析值 $T = [T_1, T_2, \cdots, T_n]^{\mathrm{T}}$ 建立预测分析子模型。建模算法采用最小二乘支持向量机(LS-SVM)，第 j 个子模型 $f_j(\cdot)$ 的 TOC 估计值为 $\hat{T}_j = [\hat{T}_{j,1}, \cdots, \hat{T}_{j,n}]$，即

$$\hat{T}_{j,i} = f_j(S_i(\lambda_j)) \tag{1}$$

步骤二：选择并计算预测子模型的性能评价指标。对子模型准确性的评价直接决定了子模型在组合模型中的权值，但在不同的评价指标下，模型的评价结果也会有所不同。选择侧重不同的评价指标可以从多个方面对子模型进行观察和评价，可以更加合理地对子模型的预测性能给予定义。本研究采用校正误差均方根(SEC)、复相关系数(R_SEC^2)和校正误差标准差(STDC)[14]三个性能指标作为确定子模型可信度的证据，SEC 反映了训练样本中水样的 TOC 参考值与子模型对 TOC 的估计值之间的平均吻合程度，越小则说明模型的平均准确性越高；R^2 反映了 TOC 参考值与估计值之间的相关性的强弱，越接近于 1，则相关性越强；STDC 反映了误差的空间离散程度，越小则说明模型的误差水平越稳定，即稳健性越高。第 j 个子模型的三个评价指标的定义分别如式(2)～(4)所示。

$$\mathrm{SEC}_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(T_i - \hat{T}_{j,i})^2} \tag{2}$$

$$R\_\mathrm{SEC}_j^2 = 1 - \frac{\sum_{j=1}^{n}(T_i - \hat{T}_{j,i})^2}{\sum_{i=1}^{n}(T_i - \overline{T})^2} \tag{3}$$

$$\mathrm{STDC}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{T}_{j,i} - \overline{T})^2} \tag{4}$$

式中：T_i 表示第 i 个校正样本的 TOC 化验值；$\overline{T}$ 表示校正样本的 TOC 化验值均值；$\hat{T}_{j,i}$ 表示第 j 个子模型对第 i 个校正样本的 TOC 估计值。

步骤三：根据步骤二给出的模型评价指标，确定每个子模型在不同证据下的概率分配函数。根据证据理论，对于识别框架 Θ，定义了一个集函数 $m: 2^{\Theta} \to [0,1]$，若 m 满足：

$$\sum_{A\subset\Theta} m(A) = 1, \text{且 } m(\varphi) = 0 \tag{5}$$

则称 m 为识别框架 Θ 上的基本概率分配函数，它表示对结论 A 的信任程度。根据证据理论令 $M_{1,j}$、$M_{2,j}$ 和 $M_{3,j}$ 分别表示第 j 个子模型在评价指标 SEC、R_SEC^2 和 STDC 下的概率分配函数，其定义分别如式(6)～(8)所示。

$$M_{1,j} = \frac{\dfrac{1}{\mathrm{SEC}_j}}{\sum_{j=1}^{k}\dfrac{1}{\mathrm{SEC}_j}} \tag{6}$$

$$M_{2,j} = \frac{\left[\dfrac{1}{(1-R\_\mathrm{SEC}_j^2)}\right]}{\sum_{j=1}^{k}\dfrac{1}{(1-R\_\mathrm{SEC}_j^2)}} \tag{7}$$

$$M_{3,j} = \frac{\dfrac{1}{\mathrm{STDC}_j}}{\sum_{j=1}^{k}\dfrac{1}{\mathrm{STDC}_j}} \tag{8}$$

步骤四：对证据进行融合并确定组合模型的权值。将步骤三中所提供各个子模型的评价指标作为模型可靠性的证据，利用 D-S 证据推理方法，对各个子模型的概率分配函数进行合成计算，将综合所有证据所得到的子模型的可信度作为子模型在组合模型中的权值。根据 D-S 理论，可将对水样 TOC 值的分析看作一个识别框架 Θ，在识别框架 Θ 上第 j 个子模型的性能指标可看作该子模型可信度的证据，当有多个性能指标为子模型提供证据时，可根据 Dempster 证据合成规则对多个证据下的概率分配函数进行融合，得到最终对该子模型的信任度，并将信任度作为第 j 个子模型在合成模型中的权值 W_j。证据融合公式如式(9)所示：

$$W_j = M_{1,j} \oplus M_{2,j} \oplus M_{3,j} = \frac{1}{K}\prod_{i=1}^{3} M_{i,j} \tag{9}$$

其中，$K = \sum_{j=1}^{k}\prod_{i=1}^{3} M_{i,j}$。

步骤五：建立对 TOC 值的组合预测模型。根据步骤四中所计算的组合权值对步骤一中所建立的 k 个子模型进行线性组合，得到组合模型 $F(\cdot)$，即

$$F(\cdot) = \sum_{j=1}^{k} W_j f_j(\cdot) \tag{10}$$

3　结果与讨论

3.1　建立组合预测模型

将 40 个水样随机划分出 12 个样本作为测试集，剩余 28 个样本作为训练集用于建立 TOC 分析模型。采用 LS-SVM 对训练集样本的 19 个激发波长下的荧光发射光谱建立 19 个 TOC 分析子模型。由于组合模型的预测效果不仅取决于组合权值的选择，更依赖于单一子模型的预测精度。因此，子模型的选择是建立组合模型的第一步，选择子模型时应依照下列原则[15]：

1）精度。保证单一子模型的预测精度，精度太差的模型对于组合预测结果没有意义。

2）相关性。子模型的预测误差的相关性应尽可能小，这样有利于子模型在组合模型中产生互补作用。

3）数量。子模型的数量不能太少，太少了组合效果不明显；而太多了则计算量过大，在实际应用中不可行。

根据以上三个原则，选择 λ_{Ex}＝{220 230 260 270 280 290 310}nm 的 7 个发射光谱子模型进行组合，不同波长的激发光所对应子模型的性能评价指标如表 1 所示。根据性能指标 SEC，R^2_SEC 和 STDC，预测效果最好的子模型所对应的激发光波长分别为 270、220、230 nm，三个性能指标的最优值在表 1 中用

黑体字标出。

根据性能评价指标得到每个子模型的概率分配函数 M_{1j}、M_{2j}和 M_{3j}($j=1,\cdots,7$),然后根据 D-S 证据融合式(8)对三个评价指标给出的证据进行融合,得到每个子模型的可信度,即组合权值 W_j($j=1,\cdots,7$),计算结果如表 2 所示。

表 1 子模型的预测性能指标

Table 1 Performance indexes of sub-models

λ_{Ex}/nm	SEC	R^2_SEC	STDC
220	9.331 4	**0.897 6**	9.420 1
230	9.452 0	0.884 0	**8.132 1**
260	12.118 3	0.809 3	12.349 1
270	**8.361 6**	0.890 2	8.520 9
280	10.917 1	0.845 3	11.125 1
290	11.787 1	0.819 6	12.011 7
310	14.173 2	0.790 4	13.100 9

表 2 子模型的概率分配函数和组合权值

Table 2 Probability assignment functions and combination weights of sub-models

λ_{Ex}/nm	M_1	M_2	M_3	W
220	0.161 9	0.197 2	0.156 8	0.220 3
230	0.159 8	0.174 1	0.181 7	0.222 4
260	0.124 6	0.105 9	0.119 6	0.069 5
270	0.180 6	0.183 9	0.173 4	0.253 5
280	0.138 4	0.130 5	0.132 8	0.105 6
290	0.128 1	0.111 9	0.123 0	0.077 7
310	0.106 6	0.096 4	0.112 8	0.051 0

根据组合权值将 7 个子模型进行线性组合,组合模型的 SEC、R_SEC^2 和 STDC 分别为 7.632 4、0.935 1和 7.880 3,比评价指标最好的子模型分别提高了 8.7%、4.2%和 3.1%。

3.2 组合模型的预测结果及评价

对 12 个测试样本使用 **3.1** 节建立的组合模型对 TOC 值进行分析,并与每个子模型的预测分析结果进行比较。评价指标包括预测标准误差(SEP),预测误差的复相关系数(R_SEP^2)和预测误差标准差(STDP),其定义分别如式(11)~(13)所示。

$$\mathrm{SEP}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(T_i-\widetilde{T}_i)^2} \tag{11}$$

$$R\_\mathrm{SEP}^2=1-\frac{\sum_{i=1}^{n}(T_i-\widetilde{T}_i)^2}{\sum_{i=1}^{n}(T_i-\overline{T})^2} \tag{12}$$

$$\mathrm{STDP}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{T}_i-\overline{T})^2} \tag{13}$$

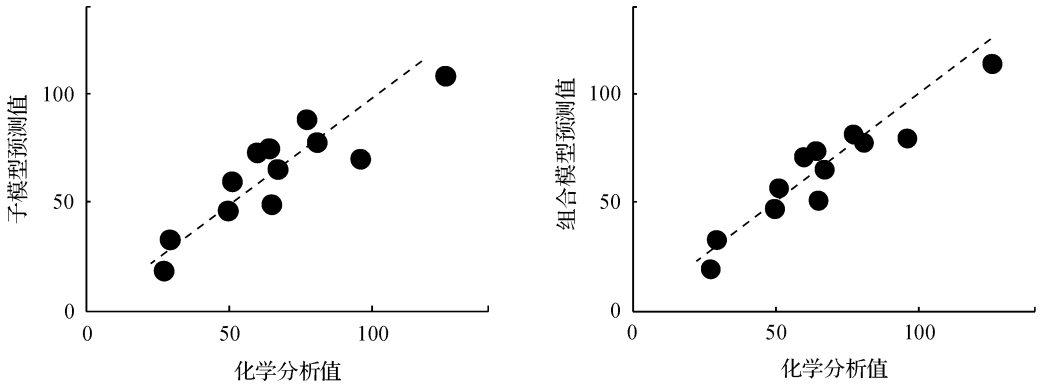
式中:T_i 表示第 i 个检测样本的 TOC 化验值;$\overline{T}$ 表示检测样本 TOC 化验值的均值;$\widetilde{T}_i$ 表示分析模型第 i 个检测样本的 TOC 估计值。

表 3 给出了组合模型和子模型的预测误差性能比较。从表 3 可以看出,组合模型的每个预测性能指标相对于最优指标的子模型都有了进一步的提高,其中 SEP 相对于该指标的最优子模型提高了 10%,R_SEP^2 相对于该指标的最优子模型提高了 5.82%,STDP 相对于该指标的最优子模型提高了 12.35%。这说明基于 D-S 证据理论对多个性能指标给出的证据进行融合所得到的组合模型能够同时吸收不同子模型的优势,得到一个综合预测性能较强的分析模型。图 3 给出了测试集中水样的 TOC 化学分析值分别与 $\lambda_{Ex}=220$ nm 的子模型和组合模型预测值的对比图。从图 3 可以看出,与子模型相比,组合模型的稳健性和精确度均得到了较大提高,达到了实际分析精度的要求。

表 3 子模型和组合模型的预测结果

Table 3 Prediction results of sub-models and combined model

模型	λ_{Ex}/nm	SEP	R_SEP^2	PEV
子模型	220	11.242 3	0.828 7	12.560 7
	230	12.367 0	0.803 4	13.326 6
	260	21.595 9	0.711 2	18.642 9
	270	14.780 7	0.677 3	11.465 3
	280	17.143 0	0.806 0	13.207 6
	290	15.812 9	0.630 7	13.007 7
	310	16.423 3	0.601 7	14.560 6
组合模型	10.776 2	0.876 9	10.038 9	



(a) 激发光波长220 nm处的子模型　　(b) 组合模型

图 3　TOC 的化学分析值与模型预测值的比较

Fig. 3　Comparison of chemical analysis values with model prediction values for TOC

4　结　语

本研究基于不同激发光波长下的荧光发射光谱对综合水质评价指标 TOC 建立多个子模型，然后将三个模型预测能力的性能指标作为子模型可靠性的证据，计算在三个证据下每个子模型的概率分配函数，并通过 DS 对三个性能指标所提供的证据进行合成计算，得到每个子模型的组合权值。最后根据所得到的组合权值将子模型的预测结果进行加权求和，得到组合模型的 TOC 预测值。实验证明，基于 DS 证据理论所得到的组合模型可以综合多个子模型的优势，与单一子模型相比具有更高的精确度和稳定性。

参考文献：

[1] Bieroza M, Baker A, Bridgeman J. Relating freshwater organic matter fluorescence to organic carbon removal efficiency in drinking water treatment[J]. Science of the Total Environment, 2009, 407(5): 1765-1774.

[2] Murphy K R, Hambly A, Singh S, et al. Organic matter fluorescence in municipal water recycling schemes: toward a unified PARAFAC model[J]. Environmental Science & technology, 2011, 45(7): 2909-2916.

[3] Yao X, Zhang Y L, Zhu G W, et al. Resolving the variability of CDOM fluorescence to differentiate the sources and fate of DOM in lake taihu and its tributaries[J]. Chemosphere, 2011, 82(2): 145-155.

[4] 黎司，吉芳英，周光明，等. 三峡库区水体溶解有机质的荧光光谱特性[J]. 分析化学，2009，37(9)：1328-1332.

[5] 陈茂福，吴静，律严励，等. 城市污水的三维荧光指纹特征[J]. 光学学报，2008，28(3)：578-582.

[6] Lee S, Ahn K H. Monitoring of COD as an organic indicator in waste water and treated effluent by fluorescence excitation-emission (FEEM) matrix characterization [J]. Water Science and Technology: A Journal of the International Association on Water Pollution Research, 2004, 50(8): 57-63.

[7] 王志刚，刘文清，张玉钧，等. 不同来源水体有机综合污染指标的三维荧光光谱法与传统方法测量的对比研究[J]. 光谱学与光谱分析，2007，27(12)：2514-1517.

[8] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[9] 夏陆岳，俞立. 基于 SNNs-RR 的聚丙烯熔融指数软测量[J]. 化工学报，2008，59(7)：1631-1634.

[10] Wang M Y, Lan W T. Combined forecast process: combining scenario analysis with the technological substitution model[J]. Technological Forecasting and Social Change, 2007, 74(3): 357-378.

[11] Dymova L, Sevastianov P, Bartosiewicz P. A new approach to the rule-base evidential reasoning: Stock trading expert system application[J]. Expert Systems with Applications, 2010, 37(8): 5564-5576.

[12] Sevastianov P, Dymova L, Bartosiewicz P. A framework for rule-base evidential reasoning in the interval setting applied to diagnosing type 2 diabetes[J]. Expert Systems with Applications, 2012, 39(4): 4190-4200.

[13] 文成林，周哲，徐晓滨. 一种新的广义梯形模糊数相似性度量方法及在故障诊断中的应用[J]. 电子学报，2011，39(S1)：1-6.

[14] 廖顺宝，张赛. 属性数据空间化误差评价指标体系研究[J]. 地球信息科学学报，2009，11(2)：176-182.

[15] 曾鸣，冯义，刘达，等. 基于证据理论的多模型组合电价预测[J]. 中国电机工程学报，2008，28(16)：84-89.