文章编号:1000-582X(2010)03-119-05

SOM 网络与 SVM 在水质富营养化评价中的对比

石　欣,熊庆宇,雷璐宁
(重庆大学 自动化学院,重庆 400044)

摘　要:针对复杂水环境中的富营养化评价问题,利用三峡库区水体富营养化监测数据,对自组织映射神经网络和支持向量机模型在解决该评价问题上的性能表现进行对比研究。实验结果表明,2 种模型均有较快的计算速度和较高的精度,但与自组织映射网络模型相比,支持向量机模型具有更好的稳定性和抗干扰能力,在参数选择上更为简单。

关键词:支持向量机;自组织映射;神经网络;对比;评价;富营养化
中图分类号:TP183　　文献标志码:A

A comparative study of eutrophication evaluation models based on SOM neural network and SVM

SHI Xin, XIONG Qing-yu, LEI Lu-ning
(Automation Department, Chongqing University, Chongqing 400044, P. R. China)

Abstract: Considering the eutrophication evaluation problems of complicated water environment, using the eutrophication inspection data of the Three Gorges Reservoir district to compare of the eutrophication evaluation models' performances based SOM neural network and SVM. The results shows that both models have fast calculation speed and high precision, but compared to SOM neural network model, the SVM model has better stability and anti-jamming capability, also its preferences are simpler.

Key words: support vector machine; self organizing maps; neural networks; comparison; evaluation; eutrophication

近年来,随着废水排放的增加,化肥、合成洗涤剂和农药等化学品的大量使用,中国湖泊、水库和江河富营养化问题日益严重[1]。然而,水体富营养化程度受到多种因素的影响,评价因子与富营养化评价等级间关系复杂,各等级间关系模糊,各评价方法大都有其适用条件和局限性,因此至今尚未形成一个统一确定的评价模型[2]。

自组织映射神经网络(Self Organizing Maps, SOM)是一类无监督学习的神经网络模型,可以对外界未知环境或样本空间进行学习或者模拟,具有较强的处理复杂非线性问题的能力。水环境是一个复杂的非线性系统,其中包含许多未知因素[3-4],因此,利用 SOM 网络对水体进行富营养化的分析评价是值得研究的。支持向量机(Support Vector Machine, SVM)是 20 世纪 90 年代中期由 Vapnik 等人基于统计学理论提出的一种新的机器学习方法。它能够较好地解决有限样本、非线性、高维数以及神经网络方法所面临的结构选择、局部极小点等实际问题[5-6]。目前,利用 SOM 网络和 SVM 模型对水质评价问题的研究已有许多,但对两者在解决

收稿日期:2009-11-10
基金项目:国家科技重大专项(2009ZX07528-003-09);重庆市科技重大攻关项目(CSTC,2006AA7024)
作者简介:石欣(1978-),男,重庆大学博士研究生,主要从事智能信息处理等研究。
熊庆宇(联系人),男,重庆大学教授,博士生导师,(E-mail) xiong03@cqu.edu.cn。

水质评价问题时的各项性能缺乏详细的分析研究。笔者尝试利用水质数据对2种方法进行对比实验，比较2种模型在解决水体富营养化评价问题中的适应性和性能。

1 SOM神经网络模型

SOM网络模型由输入层和输出层组成，输入层各神经元负责接收数据，并通过权向量将外界信息汇集到输出层的各神经元[7-10]。

图1为SOM网络结构，设训练样本维数为 N，训练样本集为 $\boldsymbol{X}=[X_1,X_2,\cdots,X_k,\cdots,X_N]^{\mathrm{T}}$；输入层有 i 个神经元；输出层神经元 j 的权值向量为 $\boldsymbol{W}_j=[w_{j1},w_{j2},\cdots,w_{jN}]^{\mathrm{T}}$，$j=1,2,\cdots,P$，其中 P 是输出层神经元的总数。对 $j=1,2,\cdots,P$，比较内积 $W_j^{\mathrm{T}}X$，选择具有最大内积的神经元，由此决定了兴奋神经元的拓扑邻域中心的位置。

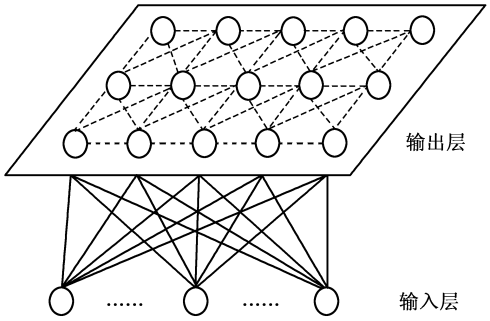

图1 SOM网络结构

基于内积 $W_j^{\mathrm{T}}X$ 最大化的最优匹配准则，等价于向量 X 和 W_j 的欧式距离的最小化。用标号 $i(X)$ 标识最优匹配输入向量 $\boldsymbol{X}$ 的神经元，可根据下列条件决定 $i(X)$：

$$i(X)=\arg\min_j \|X-W_j\|,j=1,2,\cdots,P,$$

满足这个条件的特定神经元 i 被称为获胜神经元。

获胜神经元的拓扑邻域函数为

$$h_{j.i(x)}(n)=\exp\left(-\frac{d_{j.i}^2}{2\sigma^2(n)}\right),n=0,1,\cdots,$$

式中：$d_{i.j}$ 为获胜神经元 i 和兴奋神经元 j 之间的侧向距离；σ 为拓扑邻域的“有效宽度”。

假定在时刻 n 神经元 j 的权值向量为 $w_j(n)$，更新权值向量 $w_j(n+1)$ 在时刻 $n+1$ 被定义为：

$$w_j(n+1)=w_j(n)+\eta(n)h_{j.i(x)}(n)(x-w_j(n)),$$

其中，$\eta(n)$ 是算法的学习率参数。网络中获胜神经元 i 的拓扑邻域中的所有神经元按照上式进行权值的调整。

2 支持向量机模型

设训练样本输入为 $x_i,i=1,\cdots,n$，对应的期望输出为 $y_i\in\{-1,+1\}$，

$$(x_i,y_i),y_i\in\{-1,+1\},i=1,\cdots,n。$$

为使超平面对所有样本正确分类并且具备最大的分类间隔，要求超平面方程 $wx+b=0$ 满足如下约束条件：

$$x_i\cdot w+b\geqslant+1,y_i=+1,$$
$$x_i\cdot w+b\leqslant-1,y_i=-1,$$

式中：w 为权值向量；b 为一个常量。

这即是一个关于 w 和 b 的二次最优化问题。

$$\begin{cases}\min\limits_w \dfrac{1}{2}\|w\|^2+C\sum\limits_{i=1}^n\xi_i,\\ y_i(wx+b)\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,\cdots,n。\end{cases}$$

引入拉格朗日乘子 α_i，将上述问题转化为一个“对偶”问题：

$$\begin{cases}\min\limits_{\alpha_i} L(\alpha)=\sum\limits_{i=1}^n\alpha_i-\dfrac{1}{2}\sum\limits_{i,j=1}^n\alpha_i\alpha_j y_i y_j(x_i,x_j),\\ \sum\limits_{i=1}^n\alpha_i y_i=0,0\leqslant\alpha_i\leqslant C,\end{cases}$$

式中 C 为惩罚因子。

非线性映射 Φ 通过核函数 $K(x_i,x_j)$ 来实现，若 $K(x_i,x_j)=\Phi(x_i)\cdot\Phi(x_j)$ 满足 Mercer 条件，则有：

$$\begin{cases}\min\limits_{\alpha_i} L(\alpha)=\sum\limits_{i=1}^n\alpha_i-\dfrac{1}{2}\sum\limits_{i,j=1}^n\alpha_i\alpha_j y_i y_j K(x_i,x_j),\\ \sum\limits_{i=1}^n\alpha_i y_i=0,0\leqslant\alpha_i\leqslant C。\end{cases}$$

设最优解为 $\alpha_i^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_n^*)$，根据 Karush-Kuhn-Tucker 最优化条件（KKT 条件），这个优化问题的解必须满足：

$$\alpha_i^*\{y_i[(w_0x_i)+b_0]-1\}=0,i=1,\cdots,n。$$

因此，对多数样本 α_i 将为零，取值不为零的 α_i，即当 $y_i[(w_0x_i)+b_0]=1$ 时，对应的样本即支持向量（SV），在几何上离最优超平面最近，求解上述问题，得到决策函数：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^n\alpha_i^* y_iK(x_i,x)+b)。$$

在实际运用中的核函数有很多，用得较多的有多项式核函数、径向基核函数、Sigmoid 核函数等[5,11-14]。考虑到径向基核函数能够将样本非线性地映射到更高维空间，且数值限制条件和参数数目相对较少，对模型选择的复杂度影响较小[14]，因此，笔者选择径向基核函数进行实验。

水质富营养化评价属于多分类问题，笔者采用支持向量机一对一（one-against-one）分类算法[13]，该算法在 k 类训练样本中构造所有可能的二类分类器，每类仅在 k 类中的二类训练样本上训练，构造 $N=k(k-1)/2$ 个分类器，学习过程采用投票决策法[14]，若当前学习训练结果表明测试样本 x 属于第 i 类，则对第 i 类的分数加 1，否则第 j 类的票数加 1，最后取最大分数的那类为 x 的等级。

3　对比实验

3.1　数据准备

根据王明翠等人提出的《湖泊富营养化评价方法及分级标准》[15]，笔者选取高锰酸盐指数（COD_{Mn}）、总磷（TP）、总氮（TN）、叶绿素 a（chla）、透明度（SD）5 项指标作为富营养化评价指标。水体富营养化评价级别分为贫营养、中营养、轻度富营养、中度富营养、重度富营养 5 个级别，笔者以数字 1～5 分别代表各营养级别。

采用三峡库区不同站点所测水质数据作为训练与测试样本，分别对支持向量机和 SOM 网络进行学习训练和验证。

3.2　训练阶段

为了测试 2 种模型的分类性能，笔者采用 300 组水质数据作为训练集，每组数据包含 5 个水质指标。

针对 SOM 网络模型，笔者将学习率参数分别设置为 0.5 和 0.9，网络结构经测试设为 7×6 进行实验，模型的性能与训练集的训练时间、迭代次数和训练误差等参数相对应。

如表 1 所示，根据训练参数的不同，模型的性能参数有较小的变化。在学习率参数不变的情况下，随着迭代次数的增加，模型的训练时间显著增长，两者的增长趋势呈 1∶1的关系，即训练时间与迭代次数以相同的倍数增长，而训练误差下降约 10%。在迭代次数相同的情况下，随着学习率的增大，训练时间逐渐减少，且迭代次数越大，训练时间减少的百分比从 0 上升到 18.8%，即迭代次数越大，学习率参数的变化对训练时间的影响越大，但由表中可以看出，学习率的增大对训练误差的影响很小，小于 3%。综上可以看出，SOM 模型参数之间相互影响，因此，在其参数的选择上较为复杂。在输入确定情况下，SOM 模型随着迭代次数的增加，训练时间会显著增长，但误差却随之减少，即 SOM 模型的计算精度要以训练时间为代价获得。

表 1　300 个数据集的训练结果（SOM-NN）

学习率	迭代次数	训练时间/s	训练误差
0.5	100	4	0.120
	1000	44	0.100
	5000	202	0.090
0.9	100	4	0.120
	1000	38	0.097
	5000	164	0.089

代表 SVM 模型性能的相关参数如表 2 所示。表中列出了当径向基函数中 $\sigma=0.7$ 时，不同 C 值条件下各参数的值。

表 2　300 个数据集的训练结果（SVM）

惩罚因子 C	支持向量数	训练时间/s	训练误差
1	18	130.04	0
10	18	148.49	0
100	18	153.21	0
200	18	161.17	0

由表 2 可以看出，随着 C 值的增大，训练时间增长了约 23.9%，两者的增长趋势呈 161∶1的关系。可见 C 值对训练时间的影响较小，训练时间的大小应主要由训练集的大小来决定。而支持向量数和训练误差没有产生变化。

综上对 SOM 模型和 SVM 模型的分析，可以看出，针对相同的训练样本集，SOM 模型的计算精度受学习率和迭代次数的影响，以训练时间为代价获得；而 SVM 模型的计算精度与训练时间无关，训练时间亦主要由训练集决定，受惩罚因子 C 的影响较小。SOM 模型性能受参数影响比 SVM 模型要大得多，因此，SOM 模型参数的选择要更为复杂。

3.3　测试阶段

另取 10 组水质样本，对训练后的 SOM 模型和 SVM 模型进行测试，验证 2 种模型是否具有较好的泛化性能，并将评价结果与综合营养状态指数法的评价结果进行对比，检验其分类的准确性，如表 3 所示。

表 3　评价结果对比

评价方法	评价结果									
TSI	1	1	4	2	1	3	4	2	3	2
SOM	1	1	5	2	1	4	4	2	3	2
SVM	1	1	4	2	1	3	4	2	3	2

从表3中通过比较可以看出，SVM模型对水体富营养化程度的判定与综合营养指数法判定的结果完全一致，而SOM网络模型判定的结果有较小差异。SOM网络模型对10个测试数据的分类精度为80%，支持向量机对10个测试数据的分类精度达到了100%，其分类精度大大高于SOM模型。

由于水质监测数据常常受到监测仪器等外界环境的影响，因此，检验模型的抗干扰能力是十分必要的。为了检验SOM网络和SVM在处理不良样本时的抗干扰能力，采用4组加入了白噪声的测试样本，对SOM模型和SVM模型在处理受干扰样本时的性能进行测试，分别取10、20、40、80 dB 4种信噪比进行统计研究。测试结果如表4所示。

表4 抗干扰检验结果

SNR/dB	模型	最小值/%	识别率最大值/%	平均值/%
10	SOM	69.57	77.05	73.87
	SVM	90.47	96.80	92.65
20	SOM	79.47	83.68	82.13
	SVM	98.68	99.74	99.21
40	SOM	97.10	99.47	98.58
	SVM	100.00	100.00	100.00
80	SOM	98.68	99.74	99.21
	SVM	100.00	100.00	100.00

由表4可以看出，SVM模型对白噪声有显著的免疫能力，测试样本完全不受各种程度噪声的影响，这说明该模型具有很好的抗干扰能力。

对SOM网络模型，可以看到随着SNR的减低，识别率明显下降。当SNR小于或等于20 dB时，模型的免疫能力受到局限。但是当SNR大于或等于40 dB时，测试样本没有受到太大的影响。因此，该模型在SNR大于或等于20 dB时，具有一定的抗干扰力，但相对SVM模型而言，它对噪声更为敏感，其抗干扰能力低于SVM模型。

4 结 语

通过实验，SOM网络模型和SVM模型均表现出了较高的分类精度，证明将这2种方法应用在水质富营养化评价分类问题上是可行的，然而相比之下，SVM模型在分类精度上的表现更好。2种方法在训练过程中都有较快的计算速度，但SOM网络模型的计算速度受参数影响较大，而SVM模型的计算速度在不同参数条件下，变化较小，相对稳定，使得SOM网络模型在参数选择上较SVM模型更为复杂。在面对外界干扰时，SOM网络模型对噪声的反应较为敏感，而SVM模型在这一问题上表现出较好的抗干扰能力。实验证明：SVM模型用于富营养化评价的性能表现比SOM网络模型更为突出。

参考文献：

[1] 任黎，董增川，李少华. 人工神经网络模型在太湖富营养化评价中的应用[J]. 河海大学学报，2004，32(2)：147-150.
REN LI，DONG ZENG-CHUAN，LI SHAO-HUA. Application of artificial neural network model to assessment of Taihu Lake eutrophication[J]. Journal of Hehai University (Natural Sciences)，2004，32(2)：147-150.

[2] 金相灿. 中国湖泊富营养化[M]. 北京：中国环境科学出版社，1990.

[3] SIMON HAYKIN. 神经网络原理[M]. 北京：机械工业出版，2004：285-347.

[4] 雷璐宁，石为人，范敏. 基于改进的SOM神经网络在水质评价分析中的应用[J]. 仪器仪表学报，2009，30(11)：1621-1626.
LEI LU-NING，SHI WEI-REN，FAN MIN. Water quality evaluation analysis based on improved SOM neural network [J]. Chinese Journal of Scientific Instrument，2009，30(11)：1621-1626.

[5] 邓乃扬，田英杰. 数据挖掘中的新方法：支持向量机[M]. 北京：科学出版社，2004：214-218.

[6] HUANG Z，CHEN H，HSU C J，et al. Credit rating analysis with support vector machines and neural networks：a market comparative study[J]. Decision Support Systems，2004，37(4)：543-558.

[7] LAU K W，YIN H，HUBBARD S. Kernel self-organizing maps for classification [J]. Neurocomputing，2006，69(16/18)：2033-2040.

[8] YAO X，LE L. Research on comparison of credit risk evaluation models based on SOM and LVQ neural network[C]// Proceedings of the 7th World Congress on Intelligent Control and Automation，June 25-27，2008，Chongqing，China. Chongqing：[s. n.]，Institute of Electrical and Electronics Engineers Inc，2008：2230-2235.

[9] KOHONEN T. The self-organizing map[J]. Neurocomputing，1990，78(9)：1464-1481.

[10] CHANG K，GAO J L，YUAN Y X，et al. Research on water quality comprehensive evaluation index for water supply network using SOMρ[C]// 2008International Symposium on Information Science and Engineering，

December 20-22, 2008, Shanghai, China. [S. l.]: IEEE, 2008: 621-624.

[11] XINHA Z, ZJENBO L. CHUNYU K. Underwater acoustic targets classification using support vector machines[C]// Proceedings of the IEEE International Conference on Neural Networks and Signal Processing, December, 14-17,2003, Ningxia, China. [S. l.]: IEEE, 2003: 932-935.

[12] BOUAMAR M, LADJAL M. Multi-sensor system using Support Vector Machines for water quality classification[C]// 9th IEEE International Symposium on Signal Processing and its Applications, Feb 12-15. 2007, Sharjah. [S. l.]: IEEE, 2007:12-15.

[13] OMAK E, ARSLAN A. A new training method for support vector machines: Clustering k-NN support vector machines[J]. Expert Systems with Applications, 2008, 35(3): 564-568.

[14] 刘坤，刘贤赵，孙瑾，等. 基于支持向量机的水环境质量综合评价[J]，中国环境监测，2007，23(3)：81-84.
LIU KUN, LIU XIAN-ZHAO, SUN JIN, et al. Comprehensive assessment of water environmental quality based on support vector machine [J]. Environmental Monitoring in China, 2007, 23(3): 81-84.

[15] 王明翠，刘雪芹，张建辉. 湖泊富营养化评价方法及分级标准[J]. 中国环境监测，2002,18(5):47-49.
WANG MING-CUI, LIU XUE-QIN, ZHANG JIAN-HUI. Evaluate method and classification standard on lake eutrophication[J]. Environmental Monitoring in China, 2002, 18(5):47-49.

（编辑　王维朗）

(上接第 **102** 页)

[12] 陈杰，吴亦红. 碱性过硫酸钾消解测定城市污泥中总氮[J]. 环境监测管理技术，2005，17(1)：35-36.
CHEN JIE, WU YI-HONG. Alkaline potassium per sulfate digestion determination of total nitrogen in municipal sludge [J]. Environmental Monitoring Management and Technology, 2005, 17(1): 35-36.

[13] 封勇，陈杰，吴亦红. 城市污泥中总氮的测定方法[J]. 河北化工，2005(6)：74-75.
FENG YONG, CHEN JIE, WU YI-HONG. The measurement method of TN in municipal sludge[J]. Hebei Chemical Industry, 2005(6): 74-75.

[14] 张志军，李定龙. 污泥样品中总氮、总磷的联合测定[J]. 江苏工业学院学报，2006，18(3)：37-39.
ZHANG ZHI-JUN, LI DING-LONG. Joint determination of total nitrogen and total phosphor in sludge samples[J]. Journal of Jiangsu Polytechnic University, 2006, 18(3): 37-39.

[15] 周旭红，曹晓辉. 污泥中总磷测定方法的探讨[J]. 浙江化工，2005，36(2)：41-42.
ZHOU XU-HONG, CAO XIAO-HUI. Study on the method of total phosphor in pollute soil[J]. Zhejiang Chemical Industry, 2005, 36(2): 41-42.

[16] MCBRIDE M B, RICHARDS B K, STEENHUIS T. Bioavailability and crop uptake of trace elelnents in soil columns amended with sewage sludge products[J]. Plant and Soil, 2004, 262(1/3): 71-84.

[17] FLYHAMMAR P. The use of sequebtial extraction on anaaerobically degraded municipal solid waste [J]. Science of the Total Environment, 1998, 212(2/3): 203-215.

[18] SCANCAR J, MILACIC R, STRAZAR M, et al. Total metal concentration and partitioning of Cd, Cr, Cu, Ni and Zn in sewage sludge[J]. Science of the Total Environment, 2000, 250(1/3): 9-19.

[19] MAIER E A, GRIEPINK B, MUNTAU H, et al. Certification of the total contents of the aqua regia soluble contents of Cd, Co, Cu, Pb, Mn, Ni, and Zn in a sewage sludge[J]. Fresenius' Journal of Analytical Chemistry, 1993,347(8):588-591.

（编辑　张　苹）