基于粗糙集和模糊神经网络的空气质量评价*

徐彩霞[1],李义杰[2]

(1.辽宁工程技术大学 研究生学院,辽宁 葫芦岛 125105;
2.辽宁工程技术大学 软件学院,辽宁 葫芦岛 125105)

摘 要:针对概率神经网络的输入量过多会影响其训练速度的问题,采用了基于分辨矩阵的粗糙集属性约简方法,删除不相关或不重要的指标。鉴于空气质量分级标准的模糊性,将模糊数学和概率神经网络结合起来,构建了模糊概率神经网络空气质量评价模型(FPNN),然后将约简后的指标值进行模糊化处理后输入到 PNN 神经网络进行智能训练。实例表明,该方法提高了收敛速度,评价结果客观可靠,具有一定的实用价值。

关键词:粗糙集;概率神经网(PNN);分辨矩阵;空气质量

中图分类号:TP183 **文献标识码**:A **文章编号**:1674-7720(2010)15-0096-04

Air quality evaluation based on rough sets and fuzzy neural network

XU Cai Xia[1], LI Yi Jie[2]

(1.Institute of Graduate, Liaoning Technical University, Huludao 125105, China;
2.College of Software Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: Aiming at the problem that too much input values inputs of probabilistic neural network affects the speed of their training, the paper uses rough set attribute reduction method based on discernable matrix to remove irrelevant or unimportant indexes. Considering the uncertainty of the standard of classification for air quality evaluation, A fuzzy probabilistic neural network model (FPNN) for air quality evaluation is proposed by combining fuzzy mathematics and probabilistic neural network. Then, the indexes after attributes reduction is processed with fuzzification, puts the results in PNN neural network for intelligent training.The trainning results indicates the proposed method improves the convergent speed and the outcomes proved to be objective and credible, therefore the evaluating method used is of some practical value.

Key words: rough sets; probabilistic neural network; discernable matrix; air quality

随着科技和经济的迅速发展,工业废气、机动车尾气、尘埃等急剧增加,成为空气质量下降的污染源,对人们的身体健康构成了严重威胁,因此采取控制和改善空气质量的有效措施,合理地进行空气质量评价及预测成为当前环境科学研究的重要内容之一。

常规的空气质量评价方法有:API 法、灰色聚类法、模糊综合评价法及模糊灰色模型等。但这些方法都存在着一些不足,如评价结果或多或少受主观因素的影响。近年来,有人把神经网络应用到空气质量评价上并取得了较好的效果。人工神经网络 ANN(Artificial Neural Nets)具有较强的非线性映射、自学习、自适应及容错能力,它能模拟大脑的思维,利用存储的网络信息对未知样本进行评价。

模糊数学是研究和处理自然界与信息技术中广泛存在的模糊现象的数学(其中的相对隶属度能很好地表示模糊概念的相对状态),但它很难表示时变知识和过程,而神经网络能够通过自学习功能来获得精确的或模糊的知识,两者的融合即模糊神经网络,弥补了模糊数学在学习方面的不足和神经网络在处理模糊数据方面的缺陷。

粗糙集理论是一种处理不完整和不确定知识的数学工具,它是 Z.Pawlak 于 1982 年提出的。粗糙集能有效地分析和处理不精确和不完整等各种不完备信息,并从

*基金项目:辽宁省教育厅基金项目(2009A350)

中发现隐含的规律。

本文把粗糙集理论和模糊概率神经网络知识运用到空气质量评价过程中，简化了网络模型，提高了评价效率和评价结果的客观性。

1 对粗糙集模糊概率神经网络的描述

1.1 粗糙集属性约简问题

知识约简是粗糙理论的重要内容之一，即求出信息系统的原有属性集合的一个极小子集，且该子集具有与原属性集合相同的分类能力，这样既保证了分类的质量，也提高了分类的速度。

定义 1 设知识库 $K=(U,F)$，F 是 U 上的等价关系族，对于每个子集 $X\subseteq U$ 和一个等价关系 $R\subseteq F$，且 $R\neq\phi$，定义两个子集，即 $\underline{R}(X)=\cup\{Y\in U/R|Y\subseteq X\}$，$\overline{R}(X)=\cup\{Y\in U/R|Y\cap X\neq\Phi\}$。分别称它们为 X 的 R 下近似集和 R 上近似集。

定义 2 设 $S=(U,C\cup D)$ 是一个决策表，D 的 C 正域记作 $POS_C(D)$，即 $POS_C(D)=\bigcup_{X\subseteq U/D}\underline{C}(X)$，$D$ 的 C 正域是论域 U 的所有使用类 U/C 所表达的知识能正确划入到决策类 U/D 之中的对象的集合。

定义 3 分辨矩阵由 Skowron 提出，其定义是：令 $S=(U,A,V,f)$ 是一个知识表达系统，其中 $U=\{x_1,x_2,\cdots,x_n\}$ 是论域；$A=C\cup D$ 是属性集合；子集 C 和 D 分别是条件属性集和决策属性集；$V=\cup V_a$，$V_a\in A$，V_a 表示属性值的集合；$f=U\times A\rightarrow V$ 是一个信息函数，对 $x_i\in U$，$a\in A$，有 $f(x_i,a)\in V_a$；$D(x)$ 是样本 x 在 D 上的值，则分辨矩阵记为 $\boldsymbol{M}=[m_{ij}]_{n\times n}$，第 i 行第 j 列的元素为 m_{ij}[1]：

$$m_{ij}=\begin{cases}\{a|a\in C\wedge f(x_i,a)\neq f(x_j,a)\}, & D(x_i)\neq D(x_j)\\ 0, & D(x_i)=D(x_j)\\ -1, & f(x_i,a)=f(x_j,a)\wedge D(x_i)\neq D(x_j)\end{cases}\tag{1}$$

其中 $n=|U|$。

1.2 指标相对隶属度矩阵

若空气质量有 b 个级别，c 项评价指标，则这 c 项指标对应的 b 级评价标准构成了空气质量评价的标准值矩阵：

$$\boldsymbol{Y}=(y_{ij})_{c\times b}\tag{2}$$

式中 y_{ij} 为第 i 项评价指标的第 j 级的评价标准值$(1\leqslant i\leqslant c,1\leqslant j\leqslant d)$。

令 m 为空气质量检测样本的个数，这 m 个检测样本数据构成了样本值矩阵 $\boldsymbol{X}$：

$$\boldsymbol{X}=(x_{ik})_{c\times m}\tag{3}$$

式中：x_{ik} 为第 k 个检测样本数据的第 i 项评价指标值，$(1\leqslant i\leqslant c,1\leqslant k\leqslant m)$。

空气污染程度的大小是个模糊概念，因此采用模糊数学理论中的相对隶属度来描述。令 p_{ij} 为第 i 项指标的第 j 级标准值的相对隶属度$(1\leqslant i\leqslant c,1\leqslant j\leqslant d)$，$p_{ij}$ 值的大小代表了空气污染的程度。再令 r_{ik} 为第 k 个检测样本数据的第 i 项指标的等级相对隶属度$(1\leqslant i\leqslant c,1\leqslant k\leqslant m)$。则标准指标相对隶属度矩阵 $\boldsymbol{P}=[p_{ij}]_{c\times d}$ 和检测样本指标相对隶属度矩阵 $\boldsymbol{R}=[r_{ik}]_{c\times m}$ 分别为[2]：

$$p_{ij}=\begin{cases}0, & j=1\\ (y_{ij}-y_{i1})/(y_{ib}-y_{i1}), & 1<j<b\\ 1, & j=b\end{cases}\tag{4}$$

$$r_{ij}=\begin{cases}0, & x_{ik}\leqslant y_{i1}\\ (x_{ik}-y_{i1})/(y_{ib}-y_{i1}), & y_{i1}<x_{ik}<y_{ib}\\ 1, & x_{ik}\geqslant y_{ib}\end{cases}\tag{5}$$

1.3 基于粗糙集模糊概率神经网络的空气质量评价模型的框架结构

与常用的 BP 神经网络相比，概率神经网络(PNN)是一种结构简单、训练速度快、非线性映射能力强且具有较好分类能力的神经网络模式。但若 PNN 有多个输入或大量的训练样本数据，分类结果的准确性就可能降低，同时也降低了网络的训练速度。因此需要运用粗糙集理论中的知识约简算法对属性进行约简，也就是在保证知识表达系统在分类能力不变的条件下，删除不重要或不相关的条件属性，减少 PNN 的输入神经元的个数，从而提高训练速度。

为了使整个评价模型的指标具有可比性，采用了模糊数学理论中的相对隶属度的知识，对约简后的评价标准数据进行预处理，并构造相对隶属度矩阵，这样就能较清晰地反映空气质量评价中的各指标的相对状态，并在此基础上构建模糊概率神经网络(FPNN)模型。

根据粗糙集和 FPNN 模型对问题分析的思路，空气质量评价模型的框架结构可以用图 1 所示的流程图描述。

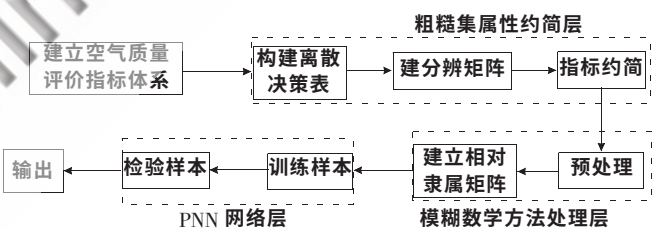

图 1 基于粗糙集的 FPNN 空气质量评价模型的框架结构

2 基于粗糙集 FPNN 空气质量评价模型的实例

2.1 指标体系的建立

指标体系的选择直接影响到评价结果的准确性，若评价指标太多，就会延长神经网络训练的时间，若指标太少，就可能降低评价结果的准确性。根据中华人民共和国国家标准(GB3095-1996《环境空气质量标准》及 2000 年的[2000]1 号文件)及我国空气污染的特点可知，影响我国空气质量的评价指标有：SO_2、NOx、TSP(悬浮颗粒物)、PM_{10}、DF(降尘)、NO_2、CO。

2.2 粗糙集属性约简

2.2.1 属性约简的步骤

粗糙集理论只能处理离散的数据，因此需要进行连续属性的值离散化，它可以由领域专家根据经验给出相

应的区间，也可以根据某种原则对空间进行划分，给出离散点进行离散化。本文采用后者。区分矩阵法是计算决策表属性约简的常用方法，但它没有充分考虑到数据的不相容度，只适用于相容决策表。下面给出最佳属性约简算法的步骤：

(1)连续数据的离散化；

(2)构造分辨矩阵 $M=[m_{ij}]_{n\times n}$；

(3)确定 D 的 C 正域 $POS_C(D)$，可按照文献[3]所提出的简便方法来快速确定 $POS_C(D)$；

(4)判断 C 中各属性 c_i 是否对 D 可约简，方法是当去掉属性 c_i 时，检验正域 $POS_C(D)\neq POS_{C-\{c_i\}}(D)$ 是否成立。若成立，则 c_i 不可约简，否则，c_i 可约简[4]；

(5)按步骤(3)~(4)遍历所有属性；

(6)所有不可约简的属性集合为约简后的指标，即条件属性 C 对于决策属性 D 的一个相对约简。

2.2.2 空气质量评价指标的约简

为了更清楚地了解空气的质量状况，在三级基础上增加一级，即将空气质量划分为四个等级，分别为：Ⅰ级、Ⅱ级、Ⅲ级和Ⅳ级。选取 10 个城市的数据，这 10 个城市污染差别显著，可以作为属性约简的样本(篇幅有限，此数据不再列出)。令 a_1、a_2、a_3、a_4、a_5、a_6、a_7 分别表示条件属性(空气质量评价指标)中的：SO_2、NOx、NO_2、PM_{10}(可吸入颗粒物)、TSP(总悬浮颗粒物)、CO、DF(降尘)。然后对属性值进行离散处理：令 x_{ik} 为第 i 个样本第 k 项指标值，y_{jk} 为第 k 项指标的第 j 级标准值，p_k 为所取样本的第 k 项指标离散处理后的值。当 $x_{ik}\leqslant y_{1k}$ 时，$p_k=0$；当 $y_{1k}<x_{ik}\leqslant y_{2k}$ 时，$p_k=1$；当 $y_{2k}<x_{ik}\leqslant y_{3k}$ 时，$p_k=2$；当 $y_{3k}<x_{ik}\leqslant y_{4k}$ 时，$p_k=3$；当 $x_{ik}\geqslant y_{4k}$ 时，$p_k=4$。其中我国空气质量分级标准如表 1 所示。决策属性 D 的属性值与空气质量等级的对应关系是：1——Ⅰ级，2——Ⅱ级，3——Ⅲ级，4——Ⅳ级，这样可得到离散化后得到的决策表如表 2 所示。然后根据公式(1)建立分辨矩阵(篇幅有限，不再显示)，应用属性约简的步骤(3)~(6)，最后得到约简后的指标是：SO_2、NO_2、TSP、PM_{10}。

表 1 空气质量分级标准[5]

标准	SO_2	NOx	NO_2	PM_{10}	TSP	CO	DF
Ⅰ级	0.02	0.05	0.04	0.04	0.08	0.80	2.40
Ⅱ级	0.06	0.08	0.08	0.10	0.20	1.50	3.00
Ⅲ级	0.10	0.10	0.08	0.15	0.30	2.00	4.00
Ⅳ级	0.30	0.30	0.10	0.20	1.00	2.40	6.50

注：除 DF 的单位为 t^2/km^2.月外，其他单位均为 mg/m^3

2.3 相对隶属度矩阵的建立及 FPNN 的仿真研究

在保证相同分类结果的情况下，粗糙集理论的属性约简去掉了不相关或不重要的属性，约简后的指标为：SO_2、NO_2、PM_{10} 和 TSP，这 4 个指标的值越小，表示空气受污染的程度越小，其分级标准见表 1，再按 2.2 节所述方法，构造出标准隶属度矩阵 P(篇幅有限，检测样本指标相对隶属度矩阵 R 略)。

表 2 离散化后的决策表

样本	a_1	a_2	a_3	a_4	a_5	a_6	a_7	D
1	0	0	0	0	0	1	0	1
2	1	0	0	0	1	1	1	1
3	0	1	1	0	0	0	0	1
4	1	1	0	1	1	1	2	2
5	2	1	1	0	1	1	1	2
6	3	1	2	2	2	1	2	3
7	2	2	2	2	1	3	2	3
8	2	1	1	2	3	2	1	3
9	4	3	3	2	2	1	3	4
10	4	2	2	3	3	2	2	4

$$P=\begin{array}{c}\begin{matrix}\text{Ⅰ级} & \text{Ⅱ级} & \text{Ⅲ级} & \text{Ⅳ级}\end{matrix}\\\begin{bmatrix}0 & 0.142\,86 & 0.285\,71 & 1\\0 & 0.666\,67 & 0.666\,67 & 1\\0 & 0.375\,00 & 0.687\,50 & 1\\0 & 0.130\,43 & 0.239\,13 & 1\\0 & 0.120\,00 & 0.200\,00 & 1\\0 & 0.437\,50 & 0.750\,00 & 1\\0 & 0.146\,34 & 0.390\,24 & 1\end{bmatrix}\end{array}\begin{matrix}SO_2\\NO_2\\PM_{10}\\TSP\\NOx\\CO\\DF\end{matrix}$$

为了使训练后的 FPNN 模型具有很好的分类质量和推广能力，且能充分反映空气质量标准各级指标标准值的意义，本文采用在标准指标相对隶属度矩阵 P 中应用内插法生成更多的样本。若内插样本 k 的指标 i 对评价级别 j 的相对隶属度为 P^*_{kj}，则内插样本 k 隶属于评价级别 j 的隶属度为 p^*_{kj}，满足 $\sum_{j=1}^{4}p^*_{kj}=1$。定义内插样本 k 对应的标准级别值为 T_k，则：

$$p^*_{kj}=p_{ij}+(p_{i(j+1)}-p_{ij})\times q/n \quad (1\leqslant i\leqslant 4,1\leqslant j\leqslant 4) \tag{6}$$

$$T_k=j+q/n \quad (1\leqslant j\leqslant n-1) \tag{7}$$

根据需要，取 $n=5$，共生成 16 个样本，将样本 1、2、4、6、7、9、10、11、13、15、16 作为训练样本，其余 5 个作为检验样本。SO_2、NO_2、PM_{10} 和 TSP 作为输入向量，本实验是基于 Matlab7.0 软件来实现整个模糊概率神经网络空气质量评价过程的，则输入层有 4 个神经元，径向基层的神经元个数同训练样本的个数相同，即为 11 个，将评价等级作为目标向量输出。本文空气质量评价共分 4 级，分别对应的等级值为 1、2、3、4，则有 4 个竞争神经元；经过参数寻优运算，确定高斯函数的平滑参数为 0.03~0.15 之间时效果最为理想。训练结果如表 3 所示，可见对于训练样本和检验样本，网络的判断率都达到了 100%。但指标约简后的神经网络模型结构简单，样本训练所用的时间更少。

由于影响空气质量的因素很多，导致了指标体系存在冗余，因此有必要进行指标约简。约简后的指标有：SO_2、NO_2、PM_{10} 和 TSP，这说明目前我国空气质量主要受

表 3 基于相对隶属度矩阵的 FPNN 训练和样本检验及其结果

样本	※◆SO_2	◆※NO_2	◆※PM_{10}	◆TSP	NOx	CO	DF	标准级别值	约简后 FPNN 判别值	约简前 FPNN 判别值
1	0	0	0	0	0	0	0	1.0	1	1
2	0.028 57	0.133 33	0.075 00	0.026 09	0.024 00	0.087 50	0.029 27	1.2	1	1
3*	0.057 14	0.266 67	0.150 00	0.052 17	0.048 00	0.175 00	0.058 54	1.4	1	1
4	0.085 72	0.400 00	0.225 00	0.078 26	0.072 00	0.262 50	0.087 80	1.6	2	2
5*	0.114 29	0.533 34	0.300 00	0.104 34	0.096 00	0.350 00	0.117 07	1.8	2	2
6	0.142 86	0.666 67	0.375 00	0.130 43	0.120 00	0.437 50	0.146 34	2.0	2	2
7	0.171 43	0.666 67	0.437 50	0.152 17	0.136 00	0.500 00	0.195 12	2.2	2	2
8*	0.200 00	0.666 67	0.500 00	0.173 91	0.152 00	0.562 50	0.243 90	2.4	2	2
9	0.228 57	0.666 67	0.562 50	0.195 65	0.168 00	0.625 00	0.292 68	2.6	3	3
10	0.257 14	0.666 67	0.625 00	0.217 39	0.184 00	0.687 50	0.341 46	2.8	3	3
11	0.285 70	0.666 67	0.687 50	0.239 13	0.200 00	0.750 00	0.390 24	3.0	3	3
12*	0.428 57	0.733 34	0.750 00	0.391 30	0.360 00	0.800 00	0.512 19	3.2	3	3
13	0.571 43	0.800 00	0.812 50	0.543 48	0.520 00	0.850 00	0.634 14	3.4	3	3
14*	0.714 28	0.866 67	0.875 00	0.695 65	0.680 00	0.900 00	0.756 10	3.6	4	4
15	0.857 14	0.933 33	0.937 50	0.847 83	0.840 00	0.950 00	0.878 05	3.8	4	4
16	1	1	1	1	1	1	1	4.0	4	4

*表示检验样本 ◆表示约简后的指标

这四种污染物的影响，为我国有关部门合理地制定空气污染防治措施提供了依据。模糊数学理论中的相对隶属度能够表明空气质量指标的相对状态，克服了采用最大隶属度时存在的只考虑极值、容易丢失中间信息的缺陷，将它和概率神经网络相结合，建立了模糊概率神经网络模型(FPNN)，该模型人为调节参数，使评价结果更客观合理，并且为了提高评价结果的质量，采用了在标准相对隶属度矩阵中进行插值的方法，生成更多的样本。仿真表明，指标约简后 FPNN 模型既保证了分类质量，也提高了收敛速度，实用性更强。当然本文所采用的空气质量评价方法也可以应用到其他领域中。

参考文献

[1] 史成东，陈菊红，胡健.基于粗糙集和神经网络的供应链绩效预测研究 [J]. 计算机工程与应用，2007，43(33)：203-206.

[2] 刘坤，刘贤赵.模糊概率神经网络模型在水质评价中的应用[J].水文，2007，27(1)：36-39.

[3] 汪小燕.基于分辨矩阵的论域划分方法[J].电脑学习，2007(4)：5-6.

[4] 李锦菊，沈亦钦.中美两国环境空气质量标准比较[J].环境监测管理与技术，2003，15(6)：24-26.

[5] 飞思科技产品研发中心. 神经网络理论与 MATLAB 7 [M].北京：电子工业出版社，2005：116-127.

(收稿日期：2010-02-02)

作者简介：

徐彩霞，女，1983 年生，硕士研究生，主要研究方向：数据库理论与研究。

李义杰，男，1954 年生，教授，硕士生导师，主要研究方向：数据库理论与研究。