文章编号:1672-3961(2012)05-0091-05 DOI:10.6040/j.issn.1672-3961.2012.05.016

基于粗糙集理论的瓦斯灾害信息特征提取技术

李慧[1,2],胡云[1,3],李存华[1]
(1.淮海工学院计算机工程学院,江苏 连云港 222005;2.中国矿业大学信电学院,江苏 徐州 221008;
3.南京大学信息工程学院,江苏 南京 110004)

摘要:为了准确预测煤与瓦斯突出的危险性,建立有效的煤矿瓦斯预警支持系统,针对煤矿瓦斯灾害的特点,本研究提出了一种新颖的基于粗糙集的瓦斯灾害特征提取算法。该算法首先利用维数化简技术对瓦斯灾害信息矩阵进行优化,并在此基础上,利用信息论中熵的概念和最大熵原理构建瓦斯灾害信息特征提取模型。通过实际应用,证实了粗糙集理论在瓦斯灾害特征提取与瓦斯灾害预测中的有效性和实用性。

关键词:粗糙集理论;煤矿瓦斯;特征提取;信息熵
中图分类号:TP274 **文献标志码**:A

The technique of gas disaster information feature extraction based on rough set theory

LI Hui[1,2], HU Yun[1,3], LI Cun-hua[1]
(1. Department of Computer Science, Huaihai Institute of Technology, Lianyungang 222005, China;
2. School of Information & Electrical Engineering, China University of Mining & Technology, Xuzhou 221008, China;
3. Department of Information Engineering, Nanjing University, Nanjing 110004, China)

Abstract: In order to accurately predict coal and gas outburst danger and to establish an effective early-warming support system of gas in coal mine, a high efficient gas disaster feature extraction algorithm based on rough set was proposed in view of the characteristics of coal mine gas disaster. The algorithm first refined the gas disaster information matrix by using dimensionality reduction, then the entropy and max entropy in the concept of rough set theory were used to establish data mining model of gas disaster prediction. The effectiveness and practicality of rough set theory in the prediction of gas disaster and feature extraction was confirmed through practical application.

Key words: rough set theory; coal mine gas; feature extraction; information entropy

0 引言

煤矿瓦斯灾害长期以来一直危害着煤矿安全和矿工的生命安全,如何有效预测瓦斯涌出量是众多学者长期关注的热点。预防瓦斯灾害的有效方法是在大量的瓦斯监测信息中有效地提取瓦斯灾害的特征,及早识别、发现、捕获瓦斯灾害信息,为预防灾害提供依据。特征提取是完成瓦斯灾害判断、识别的关键技术。煤矿瓦斯灾害的特征提取及识别技术是煤矿安全领域所要研究的主要技术之一[1-6]。

胡永梅等利用灰色系统模型预测瓦斯涌出量[7]。但瓦斯涌出量与其影响因素之间存在着非常复杂的非线性关系,这些研究成果不能清晰地反映出各因素对矿井瓦斯涌出量影响的动态模糊的本构关系。近几年,利用神经网络与数值计算相结合已成为瓦斯涌出量预测研究的一个新趋势[8-11]。杨智懿等利用BP网络建立了工作面瓦斯涌出量的神

收稿日期:2012-05-06
基金项目:江苏省自然科学基金资助项目(11KJB520001);江苏省海洋资源研究院科技开放基金资助项目(JSIMR11B12)
作者简介:李慧(1979-),女,江苏连云港人,讲师,硕士研究生,主要研究方向为人工智能,数据挖掘与信息检索. E-mail:shufanzs@126.com

经网络预测模型[8]，但是BP网络存在很难克服的收敛速度慢、训练时间长、容易陷入局部极小值等缺点；KARACAN利用有监督的人工神经网络对美国长壁工作面矿井的瓦斯涌出量进行了预测研究[12]；薛鹏骞等利用小波神经网络建立了矿井瓦斯涌出量的预测模型[13].

由于对瓦斯灾害信息的正确识别是建立在有效的特征提取基础之上的，可见对瓦斯灾害的特征参数集合提取的特征越多、越精细，则瓦斯灾害识别的概率就越高。因此，本研究提出了一种新颖的基于粗糙集的瓦斯灾害特征提取算法。该算法首先利用维数简化技术对瓦斯灾害信息矩阵进行优化，并在此基础上利用信息理论中的熵的概念和最大熵原理构建瓦斯灾害信息特征提取模型。

1 问题描述与相关定义

通过分析煤矿瓦斯监测数据，可以生成一组用矩阵形式表示的基本特征，即原始特征。设有 m 个瓦斯灾害信息采集通道，每道选取 n 个样本数据，可以构成一个 $m \times n$ 个监测数据，其矩阵表示形式如下：

$$\boldsymbol{R} = (x_{ij})\ (i = 1,2,\cdots m;\ j = 1,2,\cdots n)。$$

利用粗糙集理论进行数据推理需要对数据进行预处理。通常需要把数据或知识表示成信息表的形式，构建起瓦斯灾害特征知识库。本研究将瓦斯灾害信息矩阵映射为一个不完备信息系统，下面先给出不完备信息系统的概念。

将一个四元组 $S = \langle U, A, V, f\rangle$ 定义为信息系统，其中 U 代表非空对象集合；A 表示非空的属性集合，由条件属性 C 和决策属性 D 构成，即 $A = C \cup D$；V 代表属性值的集合；f：$U \times A \rightarrow V$ 代表信息函数，它指定 U 中每个对象 x 的属性值。对象 x_i 在属性 a 上的取值用 $a(x_i)$ 表示。如果信息系统 S 中至少存在一个对象其属性值是缺省的，则称 S 为不完备信息系统。

将给定的瓦斯灾害的特征知识映射为一个不完备信息系统 $S = \langle U, A, V, f\rangle$：$U = \{x_1, x_2, \cdots, x_m\}$ 表示瓦斯灾害（对象）的集合，对于任何一个子集 $X \subseteq U$ 可称为 U 中的一个概念或范畴；$A = C \cup D$ 为属性集合，子集 C 和子集 D 分别表示条件属性和决策属性；若用 V_{a_i} 表示属性 a_i 的取值范围，则 $V = \bigcup_{a_i \in A} V_{a_i}$ 代表属性值的集合；f：$U \times A \rightarrow V$ 表示一个信息函数，并指定 U 中每个对象 x 的属性值。

2 基于粗糙集的瓦斯灾害特征提取算法

为了实现从瓦斯灾害信息特征中提取出对瓦斯灾害信息识别最有效的特征，以实现降低特征空间的维数，从而提高瓦斯灾害预警的准确度，需要对瓦斯灾害信息特征矩阵进行优化，维数化简是一种较好的方法[14-20]，本研究采用奇异值分解技术[15-17]对瓦斯灾害信息特征矩阵进行降维。

2.1 特征空间的维数约简

矩阵的奇异值是矩阵的固有属性，矩阵的奇异值具有非常好的稳定性，当矩阵的元素发生小的变动时，奇异值的变化很小。矩阵的奇异值按递增顺序排列，一般前面的一些奇异值就能详细地刻画出矩阵的特性。奇异值分解是一种常用的矩阵分解技术[4-5]，可以实现将一个 $m \times n$ 的矩阵 $\boldsymbol{R}$ 分解成3个矩阵。

$$\boldsymbol{R} = \boldsymbol{T}_0 \boldsymbol{S}_0 \boldsymbol{D}_0',\ \boldsymbol{S}_0 = \mathrm{diag}(\sigma_1, \cdots, \sigma_r), \quad (1)$$

其中，$\sigma_1 \geqslant \cdots \geqslant \sigma_r \geqslant 0$，$\boldsymbol{T}_0$ 和 $\boldsymbol{D}_0$ 分别是 $m \times r$ 和 $n \times r$ 的正交矩阵，r 是矩阵 $\boldsymbol{R}$ 的秩（$r \leqslant \mathrm{nub}(m,n)$）。$\boldsymbol{S}_0$ 是一个 $r \times r$ 的对角矩阵，其主对角线上的元素恰为 r 个非零且按照从大到小顺序排列的序列，称这些对角元素为矩阵 $\boldsymbol{R}$ 的奇异值（singular value）。对于矩阵 $\boldsymbol{R} = \boldsymbol{T}_0 \boldsymbol{S}_0 \boldsymbol{D}_0'$，$\boldsymbol{T}_0, \boldsymbol{S}_0, \boldsymbol{D}_0$ 通常是满秩的。矩阵的奇异值分解允许存在一个简化的近似矩阵。由于矩阵的奇异值为非递增序列，只保留其前面 k 个项，舍去后面数值较小的项，这样就可以将 $\boldsymbol{S}_0$ 简化为仅有 k 个单值的矩阵（$k < r$），得到一个新的对角矩阵 $\boldsymbol{S}$，达到特征降维的目的。如果矩阵 $\boldsymbol{T}_0, \boldsymbol{D}_0$ 通过这种简化技术得到矩阵 $\boldsymbol{T}, \boldsymbol{D}$，那么得到重构的矩阵 $\boldsymbol{R}_k = \boldsymbol{TSD}'$，$\boldsymbol{R}_k \approx \boldsymbol{R}$。奇异值分解能够生成与初始矩阵 $\boldsymbol{R}$ 最近似的一个。

本研究将奇异值分解应用到瓦斯灾害信息的特征提取系统中，通过奇异值分解技术实现对瓦斯灾害信息特征矩阵进行维数简化后，大大降低了特征空间的维数。

2.2 基于最大熵的瓦斯灾害信息特征提取

最大熵原理的基本思想是：在只掌握关于未知分布的部分知识时，应该选取符合这些知识但熵值最大的概率分布，即选择一个统计模型，满足所有已知的事实，对未知的事实则不做任何假设. 这个原则体现在对参数 $p(y|x)$ 的估计上，其中 x 是瓦斯灾害事件发生的条件，y 是瓦斯灾害事件。则 x 和 y 的联合概率，记为 $p(x,y)$。

特征选择是在抽样数据的基础上,抽样数据表示为$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$。其中,$x_i$ 表示决策属性,y_i 是分类属性,是由专家提供的类标号。在训练数据的基础上,可以用概率分布的极大似然对训练样本进行表示,即

$$\bar{p}(x,y)\equiv\frac{\mathrm{freq}(x,y)}{\sum_{x,y}\mathrm{freq}(x,y)},\tag{2}$$

其中 $\mathrm{freq}(x,y)$ 表示(x,y)在样本中出现的次数。

定义 1 特征函数　有时简称特征,一般情况下是一个二值函数$f(x,y)\rightarrow\{0,1\}$,例如对于瓦斯灾害信息特征提取问题,定义特征函数为

$$f(x,y)=\begin{cases}1, & (x=\text{瓦斯突出特征})\wedge(y=\text{异常振动});\\ 0, & \text{其它类型瓦斯灾害。}\end{cases}$$

对于特征函数f_i,其相对于经验概率分布 $\bar{p}(x,y)$的期望为

$$E_p f_i=\sum_{x,y}\bar{p}(x,y)f_i(x,y)。\tag{3}$$

对于特征函数f_i,其相对于模型概率分布 $p(y|x)$的期望为

$$E_p f_i=\sum_{x,y}\bar{p}(x)p(y|x)f_i(x,y)。\tag{4}$$

定义 2 约束　令对于特征函数对于经验概率分布的数学期望与对于模型确定概率分布的数学期望相等,即限定所求模型的概率为在样本中观察到的事件的概率:

$$E_{\bar{p}} f_i=E_p f_i。\tag{5}$$

则将式(5)称为约束。

假设存在 n 个特征$f_i(i=1,2,\cdots,n)$,则模型属于约束所产生的模型集合可形式化地表示为

$$P=\{p\,|\,E_p f_i=E_p f_i,\ i=1,2,\wedge,n\}。\tag{6}$$

而满足约束条件的模型有很多,模型的目标就是产生在约束集下具有最均匀分布的模型。若在可选的概率分布 P 中选择模型,则具有最大熵的模型 p^* 即为所选模型,可形式化表示为

$$p^*(y|x)=\arg\max\{-\sum(\bar{p}(x)p(y|x))\log(\bar{p}(x)p(y|x))\}。\tag{7}$$

采用拉格朗日乘子算法求解这个最优解,得

$$p^*(y|x)=\frac{1}{Z(x)}\exp(\sum_i\lambda_i f_i(x,y)),\tag{8}$$

其中 $Z(x)$是归一化因子:

$$Z(x)=\sum_y\exp(\sum_{i-1}^{n}\lambda_i f_i(x,y))。$$

λ_i 是特征参数,代表每个特征函数的重要性。如果通过在训练集上进行学习得到 λ_i 的值,则可以求得最优解,完成最大熵模型的构建。λ_i 和 $p(y|x)$可以采用迭代算法进行确定,具体算法如下:

算法 1　参数 λ_i 和 $p(y|x)$的确定

输入:瓦斯灾害信息的原始特征函数 $f_1,f_2,\cdots,f_n$;经验分布 $\bar{p}(x,y)$

输出:最优特征参数 λ_i^* 和模型 p_λ^*

算法步骤如下:

① $\lambda_i=0,i=1,2,\cdots,n$;

② for $i=1$ to n,解下列方程:

$$\sum_{x,y}\bar{p}(x)p(y|x)f_i(x,y),$$

$$\exp(\Delta\lambda_i f^{\#}(x,y))=E_P f_i,$$

得

$$\Delta\lambda_i=\frac{1}{M}\log\frac{E_P f_i}{E_{P_\lambda} f_i},$$

其中 $M=\sum_{i=1}^{n}f_i(x,y)$;

③ $\lambda_i\leftarrow\lambda_i+\Delta\lambda_i$;

④ 如果 λ_i 未收敛,转到第② 步。

3　实验结果与分析

3.1　实验数据

为了验证本研究提出算法的有效性,下面对基于最大熵原理的特征提取模型进行实验研究。本实验所研究的实测数据来自徐州矿务集团张双楼矿区,其中对煤层瓦斯含量、压力、温度、湿度、煤质密度、电介常数等监测数据的获取来源于矿区的日常监测数据。

所取的监测数据包括 28 面指标:GAS-D(大巷瓦斯浓度),GAS-H(回采工作面瓦斯浓度),GAS-J(掘进工作面瓦斯浓度),WIN-C(井底车场风速),TE-M(煤层温度),TE-H(回采工作面温度),HU-M(煤层含水量),HU-H(回采工作面湿度),WIN-F(工作面风速),Q-WJ(Q 值),GAS-P(煤层瓦斯压力),SW-E(声发射频率),WIN-H(回采工作面风速),KC-MTD(开采方法),KC-SP(开采速度),GAS-BIN(瓦斯初始散放速度). 整个数据集中包括不完整记录在内的总共有 2 403 个记录,全部指标都齐全的记录有 1 320 个。

3.2　特征提取实验

3.2.1　瓦斯灾害特征提取实验

利用最大熵原理求取瓦斯灾害信息包含特征选择和参数估计。特征选择是选出对分类对象有明显表征作用的属性;参数估计是用最大熵原理对每一个特征进行参数估计,使每个特征对应于一个特征参数。特征参数用来反映决策属性与分类属性之间的关联强度。

选取实验数据中的 60% 作为训练样本,40% 作

为预测样本,共分为突出、危险和不突出3类来对最大熵模型进行验证。分析因素为监测数据中的15个指标属性,通过前文提出的奇异值分解技术,去掉不相关或弱相关的属性,并进行属性概化,再按本研究提出的算法1对$f(x)$进行建模。特征提取后的指标属性及预测结果如表1所示。从表1中的数据可以看出,在经过特征提取的数据集上再应用本研究提出的最大熵原理计算出的预测类别与实际类别逼近效果良好。

表1 特征提取后的指标属性及预测结果

Table 1 The results of index attribution and prediction after feature extraction

序号	瓦斯压力	放散速度	地质构造	煤层厚度	煤层倾角	预测类别	实际类别
1	2.58	6.37	6	3.5	6.2	突出	突出
2	0.67	2.78	4	4.5	4.5	危险	突出
3	3.16	3.12	5	8.1	12.3	危险	危险
4	1.24	6.65	3	3.9	9.5	不突出	不突出
5	0.98	5.44	1	5.3	8.5	突出	突出
6	4.28	1.68	4	2.4	2.4	突出	突出
7	3.55	3.28	4	2.8	6.8	不突出	不突出
8	4.53	5.12	2	6.5	4.7	突出	突出
9	2.55	6.53	6	4.1	6.0	突出	突出
10	0.88	2.56	5	4.8	4.9	危险	突出

再以煤层温度特征(TE-M)作为实验对象进行验证,实验结果如图1所示。图中横坐标表示时间,纵坐标表示瓦斯压力值。从图1可以看出,矿区温度的监测曲线与应用最大熵原理进行的预测曲线基本吻合。

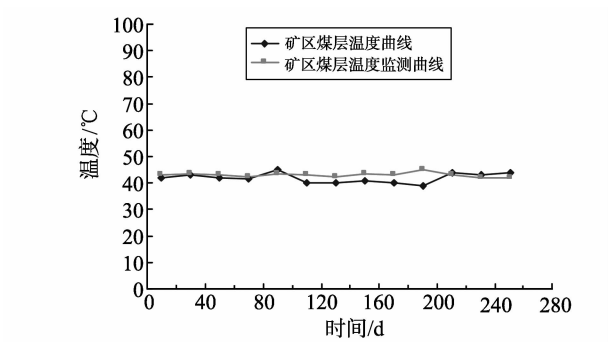

图1 矿区温度监测与最大熵拟合分布曲线

Fig. 1 The fitting distribution curve of mining temperature monitoring and maximum entropy

(2) 模型有效性验证实验

重新选择一组数据集对本研究所建立的最大熵模型进行有效性验证,采用的误差标准为

$$E=\frac{\sum_{K=1}^{n}[y_e(k)-y_m(k)]^2}{2}, \qquad (6)$$

其中,$y_e(k)$表示验证输出,$y_m(k)$表示实际输出,n表示检测次数。通过仿真可以证明基于最大熵原理求得的预测模型的准确率能达到0.85,其仿真所得的误差曲线如图2所示。

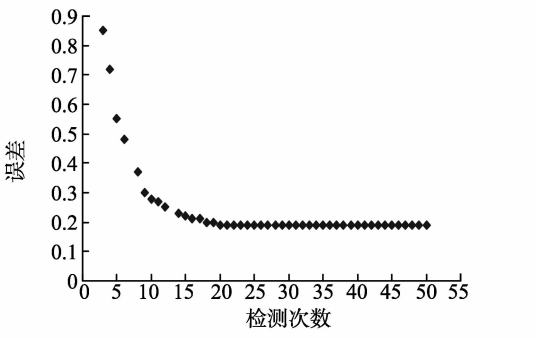

图2 瓦斯监测模型的误差曲线

Fig. 2 The error curve of gas monitoring model

4 结语

通过上述分析研究可知,最大熵方法的特点是在研究的问题中尽量把问题与信息熵联系起来,再把信息熵最大作为一个有益的假设用于所研究的问题中。这种方法所得到的结果符合煤矿实际情况,因此最大熵方法是适合煤矿瓦斯灾害特征提取的有效方法。本研究丰富了系统可靠性预警和评价手段,为煤矿事故预防与营救等工作提供了有效的决策支持。

参考文献:

[1] 张翔,肖小玲,徐光榷. 基于最大熵估计的支持向量机概率建模[J]. 控制与决策, 2006, 21(7):767-769.
ZHANG Xiang, XIAO Xiaoling, XU Guangyou. Robabilistic outputs for support vector machines based on the maximum entropy estimation[J]. Control and Decision, 2006, 21(7):767-769.

[2] 张鹏, 崔文利. 基于粗糙集优化神经网络结构的启发式算法[J]. 控制工程, 2009, 16(1):142-145.
ZHANG Peng, CUI Wenli. A heuristic algorithm based on rough set for design of neural network structure[J]. Control Engineering of China, 2009, 16(1):142-145.

[3] 谢国民, 付华, 董晶. 基于粗糙集和神经网络的煤矿瓦斯预报方法研究[J]. 计算机测量与控制, 2011, 19(4):793-795.
XIE Guomin, FU Hua, DONG Jing. Study of coal mine gas forecast method based on rough sets and neural network[J]. Computer Measurement & Control, 2011, 194):793-795.

[4] 付华,赵丹,周放. RS-RBF信息融合在瓦斯监测中的应用研究[J]. 传感器与微系统, 2009, 28(12):30-33.
FU Hua, ZHAO Dan, ZHOU Fang. Research on applica-

tion of RS-RBF information in gas monitoring[J]. Transducer and Micro System Technologies, 2009, 28(12): 30-33.

[5] 邵良杉. 基于粗糙集理论的煤矿瓦斯预测技术[J]. 煤炭学报, 2009, 34(3):371-376.
SHAO Liangshan. Disaster prediction of coal mine gas based on rough set theory[J]. Journal of China Coal Society, 2009, 34(3):371-376.

[6] 付华, 许振良. 煤矿瓦斯灾害特征提取与信息整合技术研究[D]. 辽宁:辽宁工程技术大学, 2006:58-83.
FU Hua, XU Zhenliang. Research on disaster feature extraction and information fusion of coal mine gas[D]. Liaoning:Liaoning Technical University, 2009, 34(3):371-376.

[7] 胡永梅. 灰色系统 GM(1,1)模型在煤矿瓦斯涌出量预测中的应用[J]. 能源与环境, 2008, 1(4):45-46.
HU Yongmei. The application of grey system model GM (1, 1) in mine gas emission prediction[J]. Energy and Environment, 2008, 1(4):45-46.

[8] 杨智懿, 熊亚选, 张乾林. 工作面瓦斯涌出量的神经网络模型预测研究[J]. 煤炭工程, 2004(10):73-75.
YANG Zhiyi, XIONG Yaxuan, ZHANG Qianlin. Research on the prediction of gas emission in working face based on neural network[J]. Coal Engineering, 2004, 1(10):73-75.

[9] 杨敏, 李瑞霞, 汪云甲. 煤一瓦斯突出的粗神经网络预测模型研究[J]. 计算机应用与工程, 2010,46(6):241-244.
YANG Min, LI Ruixia, WANG Yunjia. New method for prediction coal or gas outburst based on RSNN neural network[J]. Computer Engineering and Applications, 2010, 46(6):241-244.

[10] 王历,高阳,王巍巍. 预测状态表示综述[J]. 山东大学学报:工学版,2010,40(4):23-29.
WANG Li, GAO Yang, WANG Weiwei. Survey on predictive representations of state[J]. Journal of Shandong University: Engineering Science, 2010, 40(4): 23-29.

[11] 邱道宏, 张乐文, 崔伟,等. 基于趋势检查法的遗传神经网络模型及工程应用[J]. 山东大学学报:工学版, 2010, 40(3):113-119.
QIU Daohong, ZHANG Lewen, CUI Wei, et al. A genetic neural network model based on a trend examination method and engineering application[J]. Journal of Shandong University: Engineering Science, 2010, 40(3): 113-119.

[12] KARACAN C O. Modeling and prediction of ventilation methane emissions of U. S. long wall mines using supervised artificial neural networks[J]. International Journal of Coal Geology, 2008, 73(3-4):371-378.

[13] 薛鹏骞, 吴立锋, 李海军. 基于小波神经网络的瓦斯涌出量预测研究[J]. 中国安全科学学报, 2006, 16(2):22-25.
XUE Pengqian, WU Lifeng, LI Haijun. Predicting the amount of gas emitted based on wavelet neural network [J]. China Safety Science Journal, 2006, 16(2):22-25.

[14] XU Yan, LI Jintao, WANG Bin. A category resolve power-based feature selection method [J]. Journal of Software, 2008, 19(1):82-89

[15] ZHANG Sheng, WANG Weihong, Ford James, et a1. Using singular value decomposition approximation for collaborative filtering[C]// Proceedings of 7th IEEE International Conference on E-cormmerce Technology. California, USA: IEEE, 2005: 257-264.

[16] JEAN G. Singular value decomposition(SVD) and polar form[J]. Geometric Methods and Applications, 2011, 1(38):367-385.

[17] AHMAD A M. A new digital image watermarking scheme based on Schur decomposition[J]. Multimedia Tools and Applications, 2012, 59(3):851-883.

[18] FU Hua, HUA Ming. Application of data fusion to environmental measurement in coal mine[C]// Proceedings of the 3th International Symposium on Precision Measurements. California, USA: AAAI, 2006: 1-6.

[19] LILIAN B L. The logical dimension of argumentation [J]. Giving Reasons, 2011, 2(20):81-116.

[20] DON L. The interpretive dimension of economics: science, hermeneutics, and praxeology[J]. The Review of Austrian Economics, 2011, 24(2):91-128.

(编辑:胡春霞)