

矿井突水水源的多项逻辑回归判别

彭程*, 林靖涵

应急管理大学应急通信与控制工程学院, 河北 廊坊

收稿日期: 2026年8月10日; 录用日期: 2026年9月21日; 发布日期: 2026年9月30日

摘要

准确判别突水水源是矿山安全生产需要解决的关键问题之一。以袁店二矿突水水源判别问题为例, 选取归一化后的 Na^+ 、 K^+ 、 Ca^{2+} 、 Mg^{2+} 、 Cl^- 、 SO_4^{2-} 以及 HCO_3^- 离子浓度作为模型的输入, 使用带Armijo线搜索的最速下降法优化交叉熵损失函数, 建立突水水源判别的多项逻辑回归模型。计算结果表明, 得到的多项逻辑回归模型对55组训练样本的判别正确率为100%, 对8组测试样本的判别正确率为87.5%, 验证了多项逻辑回归是突水水源判别的一种有效工具。

关键词

突水水源判别, 多项逻辑回归, 最速下降法, Armijo线搜索

Multinomial Logistic Regression Discrimination of Mine Water Inrush Source

Cheng Peng*, Jinghan Lin

School of Emergency Communication and Control Engineering, University of Emergency Management,
Langfang Hebei

Received: August 10, 2026; accepted: September 21, 2026; published: September 30, 2026

Abstract

Identifying the water inrush source correctly is one of the key problems that must be solved for the safety of coal mining operations. The water inrush source discrimination problem of Yuandian No. 2 mine is taken as an example. The normalized ion concentrations of Na^+ + K^+ , Ca^{2+} , Mg^{2+} , Cl^- , SO_4^{2-}

*通讯作者。

文章引用: 彭程, 林靖涵. 矿井突水水源的多项逻辑回归判别[J]. 矿山工程, 2026, 14(5): 1411-1418.

DOI: 10.12677/me.2026.145140

and HCO_3^- are selected as the inputs of the model. The cross-entropy loss function is optimized using the steepest descent algorithm with the Armijo line search and the multinomial logistic regression model is established to discriminate the water inrush source. The computation results show that the discrimination accuracy of the obtained multinomial logistic regression model is 100% for the 55 training samples and 87.5% for the 8 testing samples, which indicate that multinomial logistic regression is an effective tool for mine water inrush source discrimination.

Keywords

Water Inrush Source Discrimination, Multinomial Logistic Regression, Steepest Descent Algorithm, Armijo Line Search

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

突水是矿山生产过程中可能出现的一种典型灾害, 准确判别突水水源对及时应对水害、减少损失有重要意义[1]。从模式识别的角度考虑, 突水水源判别是一类分类问题, 当前在突水水源判别中使用较多的分类方法包括贝叶斯判别[2] [3]、可拓分析[4] [5]、模糊判别[6] [7]、神经网络[8] [9]、支持向量机[10] [11]和偏最小二乘回归分析[12]等。多项逻辑回归, 又称为 Softmax 回归[13] [14], 是一种多分类器, 目前在突水水源判别中的应用相对较少。文献[15]在对数据进行主成分分析的基础上借助于 SPSS 软件得到了突水水源判别的多项逻辑回归模型, 但是该模型对训练样本的判别正确率偏低。本文使用 Matlab 软件计算得到了突水水源判别的多项逻辑回归模型, 并与文献[15]中的结果进行对比。

2. 多项逻辑回归原理

对于 K ($K > 2$) 分类问题, 多项逻辑回归模型的输入输出关系为

$$\mathbf{P}_w(\mathbf{x}) = \begin{bmatrix} p(y=1|\mathbf{x};\mathbf{w}) \\ p(y=2|\mathbf{x};\mathbf{w}) \\ \vdots \\ p(y=K|\mathbf{x};\mathbf{w}) \end{bmatrix} = \frac{1}{\sum_{l=1}^K e^{\mathbf{w}_l^T \mathbf{x}}} \begin{bmatrix} e^{\mathbf{w}_1^T \mathbf{x}} \\ e^{\mathbf{w}_2^T \mathbf{x}} \\ \vdots \\ e^{\mathbf{w}_K^T \mathbf{x}} \end{bmatrix}, \quad (1)$$

式中 $\mathbf{x} = [x_1, x_2, \dots, x_n, x_{n+1}]^T \in R^{n+1}$ 为输入向量, 其最后一维分量 $x_{n+1}=1$ 为截距项, $\mathbf{P}_w(\mathbf{x})$ 为输出向量, 其第 k ($k=1, 2, \dots, K$) 维分量 $p(y=k|\mathbf{x};\mathbf{w})$ 表示输入向量 $\mathbf{x}$ 属于第 k 类的概率, $y \in \{1, 2, \dots, K\}$ 为输入向量 $\mathbf{x}$ 可能的类型, $\mathbf{w} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_K^T]^T \in R^{K(n+1)}$ 为模型的参数向量。在模型建立以后, 以 $\mathbf{P}_w(\mathbf{x})$ 中最大的分量对应的类型作为分类结果。

为了建立多项逻辑回归模型, 需要利用训练样本估计出模型的参数向量 $\mathbf{w}$ 。假设训练样本的集合为 $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(M)}, y^{(M)})\}$, 可以通过优化式(2)定义的交叉熵损失函数

$$\min J(\mathbf{w}) = -\frac{1}{M} \left[\sum_{m=1}^M \sum_{k=1}^K 1(y^{(m)} = k) \log \frac{e^{\mathbf{w}_k^T \mathbf{x}^{(m)}}}{\sum_{l=1}^K e^{\mathbf{w}_l^T \mathbf{x}^{(m)}}} \right] \quad (2)$$

完成这一任务, 式中 $1(\cdot)$ 为示性函数, 其定义为:

$$1(y=k)=\begin{cases} 1 & y=k \\ 0 & y \neq k \end{cases}. \quad (3)$$

计算可得 $J(\mathbf{w})$ 的梯度:

$$\mathbf{g}(\mathbf{w})=\begin{bmatrix} \nabla_{\mathbf{w}_1} J(\mathbf{w}) \\ \nabla_{\mathbf{w}_2} J(\mathbf{w}) \\ \vdots \\ \nabla_{\mathbf{w}_K} J(\mathbf{w}) \end{bmatrix}, \quad (4)$$

式中

$$\nabla_{\mathbf{w}_k} J(\mathbf{w})=\frac{1}{M} \sum_{m=1}^M\left\{\mathbf{x}^{(m)}\left[p\left(y^{(m)}=k \mid \mathbf{x}^{(m)} ; \mathbf{w}\right)-1\left(y^{(m)}=k\right)\right]\right\}, k=1,2, \cdots, K. \quad (5)$$

本文使用最速下降法优化 $J(\mathbf{w})$, $\mathbf{w}$ 的更新公式为

$$\mathbf{w}_{t+1}=\mathbf{w}_t+\alpha_t \mathbf{d}_t, \quad (6)$$

式中 $\mathbf{w}_t$ 、 α_t 和 $\mathbf{d}_t=-\mathbf{g}(\mathbf{w}_t)$ 分别为第 t 步时参数向量 $\mathbf{w}$ 的值、步长和搜索方向。 α_t 由 Armijo 准则[16]确定, 即对于已知的 $\mathbf{w}_t$ 、 $\mathbf{g}(\mathbf{w}_t)$ 和 $\mathbf{d}_t$, 求出使式(7)成立的最小非负整数 r ,

$$J\left(\mathbf{w}_t+\beta^r \mathbf{d}_t\right) \leq J\left(\mathbf{w}_t\right)+\sigma \beta^r\left[\mathbf{g}\left(\mathbf{w}_t\right)\right]^T \mathbf{d}_t \quad (7)$$

式中常数 $\beta \in(0,1)$ 、 $\sigma \in(0,0.5)$, 则步长:

$$\alpha_t=\beta^r. \quad (8)$$

输入样本的不同分量数量级通常是不同的, 会给参数向量 $\mathbf{w}$ 的优化带来困难, 为此对输入样本进行归一化处理。设原始的训练样本为 $\left\{\left(\mathbf{u}^{(1)}, y^{(1)}\right),\left(\mathbf{u}^{(2)}, y^{(2)}\right), \cdots,\left(\mathbf{u}^{(M)}, y^{(M)}\right)\right\}$, 则多项逻辑回归模型中使用的输入向量 $\mathbf{x}^{(m)}(m=1,2, \cdots, M)$ 的第 $j(j=1,2, \cdots, n)$ 维分量 $x_j^{(m)}$ 可以由

$$x_j^{(m)}=\frac{u_j^{(m)}-u_j^{\min }}{u_j^{\max }-u_j^{\min }}, \quad m=1,2, \cdots, M \quad (9)$$

得到, 式中

$$\begin{aligned} u_j^{\max } &=\max \left(u_j^{(1)}, u_j^{(2)}, \cdots, u_j^{(M)}\right) \\ u_j^{\min } &=\min \left(u_j^{(1)}, u_j^{(2)}, \cdots, u_j^{(M)}\right) \\ j &=1,2, \cdots, n \end{aligned} \quad (10)$$

综上所述, 突水水源多项逻辑回归判别算法的流程为:

步骤 1: 按照式(9)和(10)对原始训练样本 $\left\{\left(\mathbf{u}^{(1)}, y^{(1)}\right),\left(\mathbf{u}^{(2)}, y^{(2)}\right), \cdots,\left(\mathbf{u}^{(M)}, y^{(M)}\right)\right\}$ 的输入 $\mathbf{u}^{(m)}(m=1,2, \cdots, M)$ 进行归一化处理, 得到优化时使用的输入向量 $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(M)}$ 。

步骤 2: 设置算法参数, 包括 Armijo 线搜索使用的参数 β 和 σ 以及最速下降法的最大更新步数 T 。

步骤 3: 设多项逻辑回归模型参数 $\mathbf{w}$ 的初值 $\mathbf{w}_0=\left[0,0, \cdots, 0\right]^T \in R^{K(n+1)}$, 令 $t=0$ 。

步骤 4: 根据式(4)、(5)计算梯度 $\mathbf{g}(\mathbf{w}_t)$, 令搜索方向 $\mathbf{d}_t = -\mathbf{g}(\mathbf{w}_t)$ 。

步骤 5: 根据式(7)和(8)确定步长 α_t 。

步骤 6: 令 $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t \mathbf{d}_t$, $t = t + 1$ 。若 $t < T$, 返回步骤 4。

步骤 7: 以 $\mathbf{w}_t$ 作为多项逻辑回归模型的参数。利用原始训练样本的 $u_j^{\min}$ 和 $u_j^{\max}$, 按照式(9)的方式对测试样本的输入进行处理, 进而利用式(1)计算测试样本对应的输出, 将最大概率对应的类型作为测试样本的类型。

3. 案例分析

3.1. 数据与判别方案

为了检验多项逻辑回归的效果, 使用文献[15]提供的袁店二矿突水水样数据进行计算和对比分析。水样来源于第四含水层水(记为类型 I)、煤系砂岩水(记为类型 II)和太原组灰岩水(记为类型 III), 共 63 组, 每组水样数据包含了 Ca^{2+} 等 7 种离子的浓度信息。选取表 1 中 8 组水样作为测试样本, 其余 55 组作为训练样本, 训练样本见文献[15]。

Table 1. Test samples

表 1. 测试样本

编号	离子浓度/(mg·L ⁻¹)							标记类型
	Na ⁺ +K ⁺	Ca ²⁺	Mg ²⁺	Cl ⁻	SO ₄ ²⁻	HCO ₃ ⁻	CO ₃ ²⁻	
1	331.18	36.06	32.61	285.57	249.80	446.82	6.93	I
2	353.03	34.79	32.22	306.16	241.57	485.30	1.39	I
3	184.99	50.16	37.97	156.06	197.57	314.46	0.00	III
4	298.84	25.33	9.60	208.13	76.97	368.54	0.00	III
5	237.43	16.50	10.00	170.32	26.70	413.25	0.00	II
6	272.95	12.08	2.45	155.24	2.60	504.23	0.00	II
7	320.93	10.20	2.91	209.89	69.19	403.88	26.98	II
8	339.32	7.13	11.28	231.06	77.76	457.01	15.97	II

根据水源判别依据的不同, 本文按照两种判别方案进行了多项逻辑回归建模。

方案 1, 选用除 CO_3^{2-} 之外的其他 6 种离子的浓度信息作为突水水源判别的依据。容易知道此时 $K = 3$, $n = 6$, 多项逻辑回归模型参数向量 $\mathbf{w} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \mathbf{w}_3^T]^T \in R^{21}$ 。

方案 2 为文献[15]的采用的判别方案, 使用全部 7 种离子的浓度信息作为判别的依据。此时 $K = 3$ 不变, $n = 7$, 故此种方案下多项逻辑回归模型的参数向量 $\mathbf{w} \in R^{24}$ 。

3.2. 方案 1 判别结果

设置参数 $\beta = 0.5$ 、 $\sigma = 0.2$ 、 $T = 20000$, 使用 Matlab 软件进行编程计算, 最速下降法优化得到的模型参数

$$\begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \boldsymbol{w}_3^T \end{bmatrix}^T = \begin{bmatrix} 9.378 & 7.080 & -16.459 \\ -11.380 & 5.473 & 5.907 \\ 10.507 & -20.623 & 10.116 \\ 5.998 & 5.936 & -11.934 \\ 2.184 & -11.914 & 9.730 \\ 7.357 & 4.925 & -12.282 \\ -9.557 & 9.871 & -0.314 \end{bmatrix}.$$

回代计算可知各训练样本的判别类型与标记类型均一致, 即多项逻辑回归模型对训练样本的判别正确率为 100%。

使用式(9)的方式处理 8 组测试样本并代入多项逻辑回归模型, 得到的模型输出及判别结果见表 2。由表 2 可知, 编号为 1 的测试样本对应的输出向量中第 1 维分量最大, 故可判定其属于类型 I。同理可判定编号为 2 和 3 的测试样本分别属于类型 I 和 III, 其余 5 组测试样本均属于类型 II。对照表 1 中给定的测试样本的标记类型可知编号为 4 的测试样本实际属于类型 III, 出现了误判, 而其余 7 组测试样本的判别类型与标记类型一致, 即该多项逻辑回归模型对 8 组测试样本的判别正确率为 87.5%。

Table 2. Discrimination results of test samples in scheme 1
表 2. 方案 1 对测试样本的判别结果

编号	多项逻辑模型输出	判别类型
1	[0.957 0.039 0.004] ^T	I
2	[0.958 0.042 0.000] ^T	I
3	[0.000 0.000 1.000] ^T	III
4	[0.000 1.000 0.000] ^T	II
5	[0.000 1.000 0.000] ^T	II
6	[0.000 1.000 0.000] ^T	II
7	[0.000 1.000 0.000] ^T	II
8	[0.000 1.000 0.000] ^T	II

3.3. 方案 2 判别结果

采用与方案 1 相同的算法参数, 计算得到的多项逻辑回归模型参数

$$\begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \boldsymbol{w}_3^T \end{bmatrix}^T = \begin{bmatrix} 11.939 & 4.733 & -16.672 \\ -13.622 & 6.545 & 7.077 \\ 10.043 & -19.242 & 9.199 \\ 7.968 & 3.315 & -11.283 \\ -1.281 & -7.326 & 8.606 \\ 5.497 & 6.566 & -12.063 \\ -5.065 & 6.095 & -1.030 \\ -8.289 & 8.319 & -0.030 \end{bmatrix}.$$

检验可知具有上述参数的多项逻辑回归模型对训练样本的判别正确率为 100%。将测试样本进行归一化处理后输入多项逻辑回归模型, 得到的模型输出以及判别结果见表 3。此时多项逻辑回归模型仍然仅误判了编号为 4 的测试样本, 即方案 2 得到的多项逻辑回归模型对 8 组测试样本的判别正确率也为 87.5%。

Table 3. Discrimination results of test samples in scheme 2
表 3. 方案 2 对测试样本的判别结果

编号	多项逻辑模型输出	判别类型
1	$[0.914 \quad 0.083 \quad 0.004]^T$	I
2	$[0.973 \quad 0.027 \quad 0.000]^T$	I
3	$[0.000 \quad 0.000 \quad 1.000]^T$	III
4	$[0.000 \quad 1.000 \quad 0.000]^T$	II
5	$[0.000 \quad 1.000 \quad 0.000]^T$	II
6	$[0.000 \quad 1.000 \quad 0.000]^T$	II
7	$[0.000 \quad 1.000 \quad 0.000]^T$	II
8	$[0.000 \quad 1.000 \quad 0.000]^T$	II

3.4. 结果对比

方案 1 和 2 得到的多项逻辑回归模型均将表 1 中编号为 4 的测试水样误判为属于类型 II, 其余训练和测试水样的判别类型与标记类型均是一致的, 即两种方案对 63 组水样的整体判别正确率均为 98.4%。这说明进行袁店二矿突水水源判别时, 可以忽略 CO_3^{2-} 离子的浓度信息, 使用 Ca^{2+} 等 6 种离子的浓度信息作为判别的依据。这样处理简化了模型, 同时可以减少水化学分析的工作量。

为了与文献[15]报道的结果进行比较, 使用方案 2 的判别依据, 63 组水样中, 仅将表 1 中后 4 组水样作为测试样本, 其余 59 组水样作为训练样本, 仍然采用与方案 1 相同的算法参数, 此时多项逻辑回归模型参数

$$\begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \mathbf{w}_3^T \end{bmatrix}^T = \begin{bmatrix} 6.610 & -4.844 & -1.766 \\ -10.883 & 6.537 & 4.346 \\ 12.077 & -12.672 & 0.595 \\ 8.165 & -0.241 & -7.924 \\ -1.207 & 3.818 & -2.611 \\ 11.945 & 12.924 & -24.869 \\ -7.378 & 10.312 & -2.934 \\ -11.085 & 2.743 & 8.343 \end{bmatrix}.$$

检验可知, 59 组训练样本中, 仅有表 1 中编号为 4 的训练样本被误判为属于类型 II, 其余 58 组训练样本的判别类型均与标记类型相同, 即此时多项逻辑回归模型对训练样本的判别正确率为 98.3%。将 4 组测试样本进行归一化处理后输入多项逻辑回归模型, 得到的模型输出以及判别结果见表 4。对照表 4 和表 1 可知多项逻辑回归模型对测试样本的判别正确率为 100%。

Table 4. Discrimination results of scheme 2 on test samples after changing sample division
表 4. 改变样本划分后方案 2 对测试样本的判别结果

编号	多项逻辑模型输出	判别类型
5	[0.000 0.966 0.034] ^T	II
6	[0.000 1.000 0.000] ^T	II
7	[0.000 1.000 0.000] ^T	II
8	[0.000 0.999 0.000] ^T	II

文献[15]使用 SPSS 软件进行了主成分分析和多项逻辑回归建模, 经回代计算其对 59 组训练样本和 4 组测试样本的判别正确率分别为 86.4%和 100%, 对比本文的计算结果可知多项逻辑回归模型实际上可以达到更高的判别正确率。

4. 结论

根据多项逻辑回归理论, 以水化学信息为依据, 使用最速下降法建立了突水水源判别模型。袁店二矿突水水源判别的结果表明, 多项逻辑回归模型对 63 组水样的整体判别正确率为 98.4%; 取 Na⁺+K⁺、Ca²⁺、Mg²⁺、Cl⁻、SO₄²⁻ 以及 HCO₃⁻ 离子浓度作为突水水源判别的依据是充分的, 不需要使用水样的 CO₃²⁻ 离子浓度作为多项逻辑回归模型的输入。

5. 展望

后续研究拟在统一交叉验证框架下引入支持向量机、随机森林及多层感知机等模型开展对比实验, 以进一步验证多项逻辑回归的适用性; 同时将对模型权重进行可视化解释, 分析各离子浓度对水源判别的贡献度, 并针对误判样本开展水化学机理层面的错误分析; 此外, 还将尝试引入 L1 正则化进行自动特征选择, 以及采用 L-BFGS、Adam 等优化算法替代最速下降法, 以期构建更为简洁高效的突水水源判别模型。

参考文献

[1] 崔中良, 黄保胜, 龙向前. 矿井突水水源判别研究现状及发展趋势[J]. 中国钨业, 2018, 33(4): 42-50.

[2] 郑剑, 陈陆望, 张杰, 等. 贝叶斯突水水源判别与反向水文地球化学模拟的含水层水力分析[J]. 合肥工业大学学报(自然科学版), 2025, 48(2): 203-211.

[3] 琚棋定, 胡友彪, 张淑莹. 基于主成分分析与贝叶斯判别法的矿井突水水源识别方法研究[J]. 煤炭工程, 2018, 50(12): 90-94.

[4] 胡友彪, 邢世平, 张淑莹. 基于可拓模型判别矿井突水水源[J]. 安徽理工大学学报(自然科学版), 2017, 37(6): 34-40.

[5] 赵静. 基于水化学和模糊综合评判的突水水源判别[J]. 陕西煤炭, 2022, 41(4): 99-103.

[6] 王心义, 赵伟, 刘小满, 等. 基于熵权-模糊可变集理论的煤矿井突水水源识别[J]. 煤炭学报, 2017, 42(9): 2433-2439.

[7] 董东林, 李祥, 林刚, 等. 突水水源的独立性权-模糊可变理论识别模型[J]. 煤田地质与勘探, 2019, 47(5): 48-53.

[8] 李迎雪, 郑禄林, 杨爱莲, 等. 基于 QPSO-BP 神经网络的矿井突水水源判识模型研究[J]. 贵州大学学报(自然科学版), 2025, 42(1): 114-124.

[9] 刘伟韬, 李蓓蓓, 杜衍辉, 等. 基于改进的 SSA-BP 神经网络的矿井突水水源识别模型研究[J]. 工矿自动化, 2024,

- 50(2): 98-105, 115.
- [10] 冯东梅, 吴健伟. 矿井突水水源的 SVM 识别方法[J]. 辽宁工程技术大学学报(自然科学版), 2017, 36(1): 23-27.
 - [11] 华星月, 邵良杉. 基于 KPCA-GWO-SVM 的矿井突水水源识别[J]. 煤矿安全, 2023, 54(2): 195-200.
 - [12] 李昂, 程迅, 吕璐娜, 等. 偏最小二乘回归分析法的突水水源判别模型[J]. 西安科技大学学报, 2025, 45(4): 771-783.
 - [13] 杨韞仪, 裴卉宁, 李书航, 等. 基于多项 Logistic 回归的座面材质预测模型[J]. 包装工程, 2021, 42(16): 182-187.
 - [14] 刘勇, 李肖, 施式亮, 等. 基于多项 Logistic 的危化品槽罐车火灾类型判别[J]. 中国安全科学学报, 2024, 34(1): 231-237.
 - [15] 张好, 姚多喜, 鲁海峰, 等. 基于主成分分析的多项 Logistic 回归模型的突水水源判别研究[J]. 高校地质学报, 2017, 23(2): 366-372.
 - [16] 马昌凤. 最优化方法及其 Matlab 程序设计[M]. 北京: 科学出版社, 2010.