第 56 卷 第 4 期
2020 年 7 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 56, No. 4
July, 2020

DOI:10.13232/j.cnki.jnju.2020.04.014

储层预测的代价敏感主动学习算法

汪 敏[1],赵 飞[1],闵 帆[2*]

(1.西南石油大学电气信息学院,成都,610500;2.西南石油大学计算机科学学院,成都,610500)

摘 要:传统的储层预测需要耗费大量的时间且对研究人员的专业能力要求极高,采用人工智能方法实现储层预测可以有效地改善预测效率.然而,因为环境、设备等原因导致油气井数据中存在大量属性值缺失,大大降低了储层识别精度.针对属性值缺失造成分类困难的问题,提出一个统一评估和动态选择的代价敏感主动学习算法(Active Learning Algorithm with Unified Evaluation and Dynamic Selection,ALES):(1)考虑各种代价的设置和计算,包括误分类代价、属性代价、标签代价和样本代价;(2)使用softmax回归实现对属性值和标签价值的统一评估;(3)提出一种具有排列组合和贪婪策略的最优获取方案,实现属性值和标签的动态选择.采用三个真实测井数据进行实验,显著性实验分析证明了ALES的有效性及其相对于监督代价敏感分类算法和缺失填补算法的优越性.

关键词:主动学习,代价敏感,不完备数据,统一评估,动态选择

中图分类号:TP181 **文献标识码:**A

Reservoir prediction through cost-sensitive active learning

Wang Min[1], Zhao Fei[1], Min Fan[2*]

(1. School of Electrical Information, Southwest Petroleum University, Chengdu, 610500, China; 2. Institute for Artificial Intelligence, School of Computer Science, Southwest Petroleum University, Chengdu, 610500, China)

Abstract: For oil and gas industry, traditional reservoir prediction usually takes a lot of time and requires researchers to have high expertise, while using artificial intelligence to realize reservoir prediction effectively improves the efficiency of prediction. However, due to environmental and equipment reasons, there are a large number of missing attribute values in oil and gas well data, which greatly reduce the accuracy of reservoir identification. To solve the problem of classification difficulty due to the lack of attribute values, we propose a cost-sensitive active learning algorithm with unified evaluation and dynamic selection (ALES). First, we consider the setting and calculation of various costs, including misclassification costs, attribute costs, label costs and sample costs. Second, we use softmax regression to achieve a unified evaluation of attribute values and label values. Third, we propose an optimal acquisition scheme with permutation and greedy strategies to achieve dynamic selection of attribute values and labels. The experiments used three actual logging interpretation data. The results of significance test verify the effectiveness of ALES and its superiority to the state-of-the-art supervised cost-sensitive classification algorithms and missing filling algorithms.

Key words: active learning, cost-sensitive, incomplete data, unified evaluation, dynamic selection

在石油工业中,测井是整个石油生产开采中的重要环节,测井数据的完整性对地下煤层气(如干层、气层等)的判断具有重要意义.传统的储层预测往往需要耗费大量时间,并且对研究人员的

基金项目:四川省青年科技创新研究团队项目(2019JDTD0017),教育部高等教育司产学合作协同育人项目(201801140013,201801006094)
收稿日期:2020-04-29

*通讯联系人,E-mail:minfanphd@163.com

专业能力要求较高. 新的人工智能方法可以有效提高储层预测的效率和准确性,然而在实际生产过程中,由于测井环境的复杂以及人为因素导致采集的数据含有大量缺失值,而大量缺失数据的存在会大大降低后续储层预测的准确性.

因此,在储层预测前,通常会对缺失数据进行处理. 流行的方法是缺失值插补[1],包括经典的回归[2]和相关分析等方法,人工神经网络[3]和遗传规划[4]也有助于设计复杂的方案. 另一种方法是主动特征获取(Active Feature Acquisition, AFA)[5],这是属性值严重缺失时最可靠的方法[6]. 在这种情况下,缺失的值可以根据请求以一定的代价获得,例如运行附加的诊断过程. 这些算法在一定程度上改善了数据的可用性和可学习性. 然而,对于储层预测这种属性缺失同时标签稀缺的场景,如何对属性值的价值和标签价值进行统一评估,获取关键的属性值和标签仍是需要考虑的关键问题.

因此,本文提出一个新的代价敏感主动学习问题,并设计一种动态评估和增量学习算法(Active Learning Algorithm with Unified Evaluation and Dynamic Selection,ALES)来解决此问题. 首先,考虑到不同的输入、输出和各种代价设置,定义一个五元组的不完备代价敏感信息系统数据模型. 输入包括不完备数据集、属性代价和教师代价,输出包括查询的关键属性值和标签以及预测的样本,优化目标是总代价最小.

其次,提出一种统一评估和动态选择关键属性值和标签的方法. 使用softmax回归来获得每个样本预测为每个类别的概率;然后计算预期的误分类代价和属性填补代价,通过排列组合和贪婪策略获得最优的属性值填补方案;最后,选择代价最小的一个关键样本来执行相应的预测或查询,并增量更新训练模型.

第三,开发一种新的动态评估和增量学习算法(ALES). 图1是ALES算法的框架图,通过单个样本x_i来展示动态评估和增量学习过程,并迭代完成整个测试集的属性值和样本选择. 采用最小总代价的优化策略可以获取要查询的关键属性值和标签,最终实现对所有实例的分类. 评价指标为平均代价.

算法在三个真实的测井数据集上进行测试实验. 将ALES算法与流行的分类器和最新的代价敏感学习算法进行比较,然后使用Friedman检验和Nemenyi假设检验来验证ALES与对比算法之间的显著性差异. 结果表明,就平均代价而言,ALES优于这些对比算法.

1 问题描述

本节介绍三种典型的主动学习问题,包括固定查询个数的主动学习、代价敏感主动学习和不完备数据的代价敏感主动学习.

许多复杂的学习任务,标记样本非常困难,既耗时又昂贵. 主动学习能制定策略动态选择关键样本,与专家交互,有效降低所需训练样本数量[7]. 根据Settles[8]的说法“主动学习试图查询专

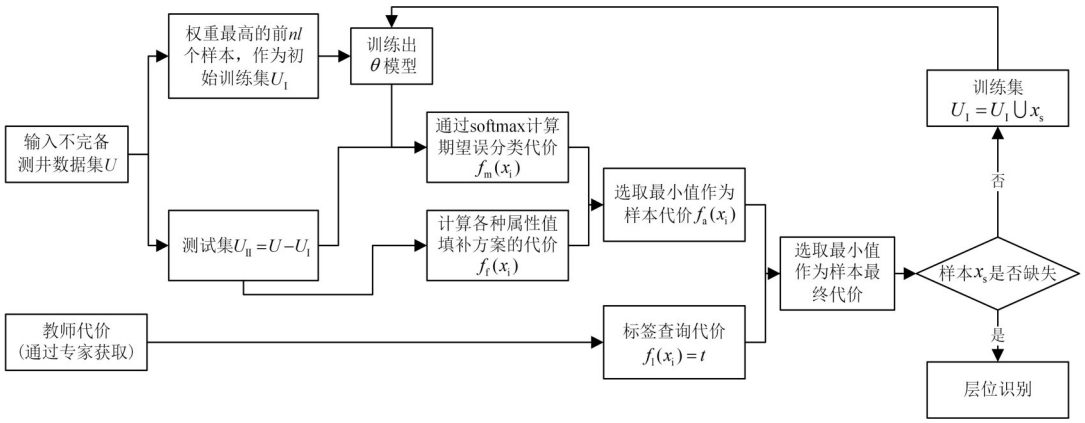

图1 ALES算法框图

Fig. 1 The algorithm framework of ALES

家获取未标记样本标签来克服标记瓶颈",主动学习的根本问题是确定如何选择最关键的样本.有两个主要标准,即信息性[9]和代表性[10-11].

1.1　固定查询个数的主动学习　在某些应用场景中,专家提供固定数量的标签,例如,考虑这样一种情况,某任务总预算为10000元,专家每标注一个标签花费10元,因此总标签的数量为1000个.

定义1　决策系统(DS)是一个三元组:

$$S=(U,C,D) \tag{1}$$

其中,U被称为非空的有限样本集合,C是条件属性的集合,D是决策属性的集合.在主动学习中,三元组决策系统通常作为数据输入.

问题1考虑固定查询个数的主动学习.输入是决策系统$S=(U,C,D)$和专家提供的nl个标签.输出包括训练子集U_{I}和目标子集U_{II}.与监督学习不同,训练子集U_{I}的真实标签通过查询由专家提供.对于目标子集U_{II},标签由相关策略预测.优化目标是最大预测精度.

问题1　固定查询个数的主动学习

输入:决策系统$S=(U,C,D)$,其中D的值未知,专家提供的nl个标签.

输出:训练集$U_{\mathrm{I}}\subset U$,预测标签$U_{\mathrm{II}}=U-U_{\mathrm{I}}$.

优化目标:最大化U_{II}的预测准确度.

1.2　代价敏感主动学习　对于代价敏感主动学习,教师代价和误分类代价应纳入决策系统.

定义2　考虑教师代价和误分类代价敏感的决策系统(TMC-DS)是五元组:

$$S=(U,C,D,M,t) \tag{2}$$

其中,M是误分类代价矩阵,而t是教师代价.

考虑代价时,问题1演变为代价敏感主动学习问题(问题2).输入为TMC-DS:$S=(U,C,D,M,t)$,其中D值未知.输出包括查询样本集U_{I}和预测样本集U_{II}.这里$t\times\left|U_{\mathrm{I}}\right|$是总教师代价,而$\sum_{i=1}^{\left|U-U_I\right|}M(l_i,y_i)$是总误分类代价.优化目标是使总代价$t\times\left|U_{\mathrm{I}}\right|+\sum_{i=1}^{\left|U-U_I\right|}M(l_i,y_i)$最小.

问题2　代价敏感主动学习

输入:TMC-DS:$S=(U,C,D,M,t)$,其中标签未知.

输出:查询样本$U_{\mathrm{I}}\subset U$,以及U_{II}的预测标签.

优化目标:$\min\left(t\times\left|U_{\mathrm{I}}\right|+\sum_{i=1}^{\left|U-U_I\right|}M(l_i,y_i)\right)$.

1.3　不完备数据的代价敏感主动学习

定义3　不完备的代价敏感信息系统(ICS-DS)是六元组:

$$S=(X,y,W,M,t,a) \tag{3}$$

其中,$X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times m}$是数据矩阵,$y$是标签矢量,$W$是与$X$大小相同的指示矩阵.

如果x_{ij}缺失,则$\omega_{ij}=0$,否则为1.a是单个缺失属性值查询代价.表1是一个不完备的代价敏感信息系统,其中$X=\{x_1,x_2,\cdots,x_{15}\}$.这里的条件属性是数值,缺失值用*表示.

表1　不完备信息系统

Table 1　An incomplete information system

U	c_1	c_2	c_3	c_4
x_1	*	3.5	1.4	0.2
x_2	*	3.0	1.4	0.2
x_3	*	*	*	0.2
x_4	4.6	3.1	1.5	*
x_5	5.0	*	1.4	*
x_6	7.0	3.2	4.7	*
x_7	6.4	3.2	4.5	1.5
x_8	6.9	3.1	4.9	1.5
x_9	*	2.3	4.0	*
x_{10}	*	2.8	4.6	1.5
x_{11}	6.3	3.3	*	2.5
x_{12}	*	*	5.1	1.9
x_{13}	7.1	3.0	5.9	2.1
x_{14}	6.3	2.9	5.6	1.8
x_{15}	*	3.0	5.8	2.2

问题3考虑不完备代价敏感决策系统(ICS-DS).缺失的属性值和标签都能以代价获取.输出为查询的属性值集合A_r,查询的样本标签集合X_r,以及预测的标签集合X_t.优化目标是总代价最小.总代价包括三部分:属性代价、标签代价和误分类代价.

问题3　不完备信息系统代价敏感主动学习

输入:ICS-IS:$S=(X,y,W,M,t,a)$.

输出:查询的属性值集合A_r,查询的样本标签集合X_r,预测的标签集合X_t.

优化目标:最小平均代价.

2 改进的算法

本文提出统一评估和动态选择的代价敏感主动学习算法,实现属性值和标签的动态查询. 首先,根据训练集构建模型并采用softmax回归训练参数$\boldsymbol{\theta}$;其次,使用评估函数$f_a(x_s)$和$f_l(x_s)$实现属性值和标签的统一评估和动态选择,获得属性值和标签的代价估算,并以最小总代价动态选择样本x_s;最后对训练集$\boldsymbol{U}_1$增量更新. 对于每一次选择的样本x_s进行相应的属性评估,判定该样本标签是预测或直接查询. 如果样本x_s不再缺失属性,它将被添加到训练集$\boldsymbol{U}_1$中并重新训练$\boldsymbol{\theta}$模型. 这样,所有标签将被查询或预测,并获得最终的总代价.

2.1 统一评估和动态选择的代价敏感主动学习

2.1.1 优化方法 数据集用$\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$表示,其中每个样本$\boldsymbol{x}=[x_{i1},x_{i2},\cdots,x_{im}]^T$是一个$m$维向量. 首先,由于所有样本都没有标签,因此选择一些具有最大密度和最大距离的代表性样本,构建初始训练集;其次构建一个概率模型以获得用于计算分类概率的参数$\boldsymbol{\theta}$;最后定义一个优化问题来获得属性值和标签. 在每次迭代中,主动学习从未标记的集合中选择一个样本x_s,获取其属性值并预测标签或直接查询标签,直到每个样本获得标签迭代才会终止. 给定代价函数f^*,选择总代价最小的无标签样本,即

$$s^*=\underset{nl<s<n}{\operatorname{argmin}}\left|f^*(x_s)\right| \tag{4}$$

其中,$f^*(x_s)=\min\left(f_a(x_s),f_l(x_s)\right)$.

2.1.2 模型构建及参数更新 该部分是算法的第一阶段,主要包括选择具有代表性的样本以构成初始训练集、构建模型和参数更新.

2.1.2.1 构建初始训练集 定义一个新的指标,即样本权重γ. 首先,使用CFDP[12]算法中的方法来计算局部密度ρ. 样本x_i的局部密度ρ_i定义为:

$$\rho(x_i)=\sum_{x_j}\chi\left(dist(x_i,x_j)-d_c\right) \tag{5}$$

其中,当$x<0$时,$\chi(x)=1$,否则$\chi(x)=0$,d_c是截止距离,$dist(x_i,x_j)$是两个样本之间的距离.

其次,计算最小距离δ. 通过计算样本x_i与密度更高的其他任何样本之间的最小距离来测量δ,即:

$$\delta(x_i)=\begin{cases}\max\limits_{x_j}\left(dist(x_i,x_j)\right) & \rho(x_i)\\ \min\limits_{j:\rho(x_i)>\rho(x_j)}\left(dist(x_i,x_j)\right) & \text{otherwise}\end{cases} \tag{6}$$

最后,计算样本权重$\gamma(x_i)$:

$$\gamma(x_i)=\rho(x_i)\cdot\delta(x_i) \tag{7}$$

选择具有最高$\gamma(x_i)$的前nl个样本作为代表样本,构建初始训练集$\boldsymbol{U}_1$.

2.1.2.2 构建模型 建立概率预测模型,通过softmax回归获得$P\left(y_j\middle|x_i;\theta\right)$. 给定任何样本$x_i$,属于$y_j$的条件概率为:

$$P\left(y_j\middle|x_i;\theta\right)=\frac{e^{\theta_j^Tx_i}}{\sum_{l=1}^{k}e^{\theta_l^Tx_i}} \tag{8}$$

2.1.2.3 参数更新 首先确定损失函数$J(\theta)$,损失函数$J(\theta)$表示预测值和真实值之间的偏差. 代价函数是:

$$J(\theta)=-\frac{1}{nl}\left[\sum_{i=1}^{nl}\sum_{j=1}^{k}1\{y_i=j\}\lg\frac{e^{\theta_j^Tx_i}}{\sum_{l=1}^{k}e^{\theta_l^Tx_i}}\right] \tag{9}$$

其中,$i\in\{1,2,\cdots,nl\}$,$j\in\{1,2,\cdots,k\}$,i表示第i个样本,j表示类别. $\{\cdot\}$是指示性函数,即当括号内参数为true时,结果为1,否则结果为0. 其次,通过最小化损失函数获得最优模型参数$\boldsymbol{\theta}$. 使用迭代优化算法(例如梯度下降法[13]或拟牛顿法[14])来求解$J(\theta)$. 经过一些推导后,获得代表损失函数偏导数的梯度:

$$\nabla\theta_jJ(\theta)=-\frac{1}{nl}\sum_{i=1}^{nl}\left[x_i\left(\left\{y_i=j\right\}-P\left(y_i=j\middle|x_i;\theta\right)\right)\right] \tag{10}$$

为求解参数θ_j,使用迭代式:

$$\theta_j:=\theta_j-\alpha\nabla_{\theta_j}J(\theta) \tag{11}$$

其中,α是步长. 最后,通过更新$\boldsymbol{U}_1$,求解损失函数$J(\theta)$以更新参数$\boldsymbol{\theta}$.

2.1.3 统一评估和动态选择属性值及标签 统一评估和动态选择属性值及标签的关键在于各种代价的计算. 考虑五个代价函数,即误分类代价$f_m(x)$、属性填补代价$f_f(x)$、属性代价$f_a(x)$、标签代价$f_l(x)$和样本代价$f(x)$. 下文详细介绍五种代价的计算方法.

2.1.3.1 期望误分类代价函数$f_m(x)$ 期望误

分类代价在样本决策中起着至关重要的作用. 利用分类概率,可以获得期望误分类代价函数$f_m(x)$. 计算该函数主要包括以下三个步骤. 首先,使用参数$\boldsymbol{\theta}$计算假设函数$h_\theta(x)$. 对于每个输入的样本x,假设函数给出每个类别j的概率值,即$P\left(y_j\middle|x;\theta\right)$. 假设函数为:

$$h_\theta(x)=\begin{bmatrix}p(y)=1|x;\theta\\p(y)=2|x;\theta\\\vdots\\p(y)=k|x;\theta\end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x}}\begin{bmatrix}e^{\theta_1^T x}\\e^{\theta_2^T x}\\\vdots\\e^{\theta_k^T x}\end{bmatrix}\tag{12}$$

其中,$\{\theta_1,\theta_2,\cdots,\theta_k\}\in R^{m+1}$是模型参数,并使用$\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x}}$归一化概率分布.

其次,获得预测概率$P(x)$:

$$P(x)=\max_{1\leqslant j\leqslant k}P\left(y_j\middle|x;\theta\right)\tag{13}$$

最后,利用误分类概率,获得期望误分类代价,即:

$$f_m(x_i)=\left(1-\max_{1\leqslant j\leqslant k}P\left(y_j\middle|x_i\right)\right)\cdot M_{ij}\tag{14}$$

2.1.3.2　属性填补代价函数$f_f(x)$　对于样本x,是否查询一个或多个属性值以获得最小代价?查询哪些属性?由于不同的属性估算方案具有不同的代价,因此定义属性填补函数$f_f(x)$. 根据预期的误分类代价、属性查询代价a和已填补的m'个属性值,属性填补函数为:

$$f_f(x_f)=\left(1-\max_{1\leqslant j\leqslant k}P\left(y_j\middle|x_f\right)\right)\cdot M_{ij}+m'a\tag{15}$$

其中,x_f表示填补了单个或多个属性后的样本.

2.1.3.3　属性代价函数$f_a(x)$　利用排列组合和贪婪策略,设计属性代价函数$f_a(x)$以获得最优的属性填补方案及其代价. 首先,使用加权平均法获得期望的填补值. 期望的填补值为:

$$\bar{x}=\frac{\bar{x}_1k_1+\bar{x}_2k_2+\cdots+\bar{x}_nk_n}{k_1+k_2+\cdots+k_n}\tag{16}$$

其次,通过排列组合策略,获得多种属性值的填补方案. 缺失属性的填补方案数量为:

$$c=\sum_{i=0}^{m'}C_{m'}^{i}\tag{17}$$

其中,m'是样本x_i缺失属性的数量.

第三,通过贪婪策略,搜索所有填补方案中代价最小的一个. 属性代价$f_a(x)$为:

$$f_a(x)=\min_{1\leqslant f\leqslant c}\left(1-P(x_f)\right)M+m'a\tag{18}$$

2.1.3.4　标签代价函数$f_l(x)$　定义查询标签代价函数,即:

$$f_l(x_i)=t\tag{19}$$

2.1.3.5　样本代价$f(x)$　对于任何样本x,获得其最优属性代价$f_a(x)$和标签代价$f_l(x)$. 因此,将两者中的最小值作为样本代价,即:

$$f^*(x)=\min\left(f_a(x),f_l(x)\right)\tag{20}$$

2.2　算法描述　本节首先描述ALES算法的伪代码,其次介绍平均代价计算公式,最后分析算法时间复杂度.

2.2.1　ALES算法伪代码　算法1描述了ALES算法,该算法迭代地选择属性值以预测标签或直接查询标签.

算法1 不完备信息系统代价敏感主动学习算法(ALES)

输入:$S=(\boldsymbol{X},\boldsymbol{y},\boldsymbol{W},\boldsymbol{M},t,a)$

输出:预测标签$\boldsymbol{L}=[l_i]_{n\times 1}$

1. $\boldsymbol{U}_{\mathrm{I}}=\varnothing,\boldsymbol{U}_{\mathrm{II}}=\boldsymbol{U},[l_i]_{n\times 1}\leftarrow -1$;//初始化

//步骤1. 选取初始训练集U_{I}

2. $\boldsymbol{U}_{\mathrm{I}}\leftarrow select(\boldsymbol{U},nl)$;

3. $\boldsymbol{U}_{\mathrm{II}}\leftarrow\boldsymbol{U}-\boldsymbol{U}_{\mathrm{I}}$;

//步骤2. 样本属性值和标签的统一评估和动态选择

4. while (true) do

5. $[\boldsymbol{\theta}]_{k\times(m+1)}\leftarrow softmaxTrain(\boldsymbol{U}_{\mathrm{I}})$;//训练$\theta$模型,其中$k$为训练集中的类别个数

6. for ($i\leftarrow 1$ to $\left|\boldsymbol{U}_{\mathrm{II}}\right|$) do

7. $\{s_f,f_a\}\leftarrow computeAttributeCost(\boldsymbol{X})$;//获取最优的属性查询方案

8. if $\left(f_a(x_i)<t\right)$ then

9. $f(x_i)\leftarrow f_a(x_i)$;

10. else

11. $f(x_i)\leftarrow t$;

12. end if

13. end for

14. $x_s\leftarrow\underset{1\leqslant i\leqslant|U_{\mathrm{II}}|}{\operatorname{argmin}}(f)$;选择总代价最小的样本$x_s$

//步骤3. 对样本x_s分类,并更新训练集

15. $l_s\leftarrow classify(x_s,s_f)$;

16. if (x_j is complete) then

17. $\boldsymbol{U}_{\mathrm{I}}\leftarrow\boldsymbol{U}_{\mathrm{I}}\cup x_s$;//更新训练集

18. $\boldsymbol{U}_{\mathrm{II}} \leftarrow \boldsymbol{U}_{\mathrm{II}} - x_s$;
19. end if
20. if ($\boldsymbol{U}_{\mathrm{II}} == \varnothing$) then
21. break;
22. end if
23. end while
24. return $\boldsymbol{L} = [l_i]_{n\times 1}$;

第1行对应初始化阶段. 专家标记的样本集为 $\boldsymbol{U}_{\mathrm{I}} = \varnothing$,未分类的样本集为 $\boldsymbol{U}_{\mathrm{II}} = \boldsymbol{U}$. 将所有样本的标签初始化为−1. 第2行选择 k 个代表性样本来构成初始训练集 $\boldsymbol{U}_{\mathrm{I}}$. 第3行更新数据 $\boldsymbol{U}_{\mathrm{II}}$.

第4至14行统一评估和动态选择属性值及标签. 第5行通过softmax回归获得模型参数 $\boldsymbol{\theta}$. 第7至13行计算样本代价 $f(x)$. 第7行计算属性代价 $f_a(x)$并获得最佳查询方案. 第8至12行比较属性代价 $f_a(x)$和标签代价 $f_l(x)$. 两者中最小的是样本代价 $f(x)$. 第14行选择 $\boldsymbol{U}_{\mathrm{II}}$中的代价最小的样本 x_s.

第15至19行将所选样本 x_s 进行分类并更新训练集. 第15行使用给定方案 s_f 对样本 x_s 进行分类. 对于 x_s,如果查询的是关键属性值则使用属性填补方案 s_f 预测标签,否则直接查询真实标签. 最后,如果所有样本都获得标签即 $\boldsymbol{U}_{\mathrm{II}} == \varnothing$,则循环终止. 第24行返回所有样本的标签.

2.2.2 平均代价计算 ALES算法的优化目标是最小化平均代价,如式(21)所示:

$$\min\left(\frac{1}{n}\times\left(a\times\left|\boldsymbol{A}_r\right|+t\times\left|\boldsymbol{X}_r\right|+\sum_{i=1}^{|\mathrm{X_t}|}\boldsymbol{M}(l_i,y_i)\right)\right) \tag{21}$$

代价的计算包含三部分,分别为属性查询代价、标签查询代价以及误分类代价.

2.2.3 复杂度分析 表2分析了ALES算法的时间复杂度. 令 m 和 n 分别为属性和样本的数量. 算法1的时间复杂度为:

$$O(mn^2)+O(n^2)+O(m'cn)=O(mn^2) \tag{22}$$

其中,m' 是一个样本中的属性缺失个数,$m' \leqslant m$.

表2 算法1的复杂度计算

Table 2 Complexity calculation of algorithm 1

行	复杂度	描述
第2行	$O(mn^2)$	选择初始训练集
第3行	$O(n^2)$	训练 θ 模型
第6~19行	$O(m'tn)$	迭代选择属性值和标签
总计	$O(mn^2)+O(n^2)+O(m'tn)=O(mn^2)$	

3 实验结果与分析

3.1 数据集 实验采用三个真实的测井数据集,包括某油气田公司滇黔川地区天然气井数据、美国Hugoton油气田的井下数据和Panoma油气田井下数据,具体信息如表3所示. 数据的属性包括伽马射线、电阻率、光电效应、中子密度空隙率等属性指标,最终的储层预测目标为干层或气层等决策信息. 实验采用完全随机缺失(Missing Completely at Random,MCAR)的方法,并且设置缺失率为10%~50%. 实验中相应的代价设置:属性查询代价 $a=0.2$,标签查询代价 $t=1$,误分类代价 $\boldsymbol{M}(l_i,y_i)=2$.

表3 数据集信息

Table 3 Information of datasets

序号	名字	样本数	属性数	类别数
1	Well_01	301	11	4
2	Well_02	408	7	2
3	Well_03	4149	7	2

3.2 实验设计 将ALES算法与朴素贝叶斯(NB),k最近邻(kNN),J48,CALF[15],GESI[16]和BPCA[17]等算法进行对比,获得每种算法的平均代价. 其中NB,kNN和J48三种监督型代价敏感分类算法使用Weka[18]的内置代码进行测试. CALF,GESI和BPCA是缺失值填补的算法. CALF提出一种基于协同过滤加权预测的主动学习填补算法. GESI提出一种新颖的非参数单插补广义回归神经网络集成算法. BPCA提出一种基于双聚类的贝叶斯主成分分析方法. 在双聚类中,识别出与缺失样本最相关的基因和实验条件,并在这些双聚类上运行BPCA来估算缺失值. 实验中三个真实测井数据通过MCAR的方法,将属性值随机缺失,分别产生缺失率为10%~50%的五个数据集.

3.3 实验结果及分析 表4对比了缺失率为50%时ALES算法和其他六种算法的平均代价,表中的黑体字表示每个数据集的最佳结果. 可以看出,ALES算法优于现有的六种分类算法. 各个算法的平均代价结果,通过KEEL[19]软件内置

表4　不同缺失率下ALES算法和其他六种对比算法的平均代价比较

Table 4　The average cost of ALES and other six algorithms on different missing ratios

				10%			
	NB	kNN	J48	CALF	GESI	BPCA	ALES
Well_01	0.5343	0.2990	0.9412	0.3578	0.4118	0.3137	**0.1485**
Well_02	1.0365	1.3223	1.3688	0.9023	1.3223	0.8837	**0.7601**
Well_03	0.6802	0.3230	0.2940	**0.2304**	0.2473	0.3789	0.2438
MeanRank	5.38	4.25	5.63	3.13	4.5	3.88	**1.25**
				30%			
	NB	kNN	J48	CALF	GESI	BPCA	ALES
Well_01	0.5245	0.6618	0.7255	0.6294	0.6716	0.6176	**0.2490**
Well_02	1.2027	1.3156	1.0963	0.8731	1.1362	0.7973	**0.7734**
Well_03	0.4584	0.4550	0.3900	0.4465	0.4001	0.4738	**0.3745**
MeanRank	4.63	5.38	4.38	3.88	4.63	4.13	**1.00**
				50%			
	NB	kNN	J48	CALF	GESI	BPCA	ALES
Well_01	0.8775	0.8333	0.7990	0.4480	0.8137	0.5637	**0.3814**
Well_02	1.5349	1.4684	1.2292	0.9422	1.4219	0.8704	**0.7794**
Well_03	0.7669	0.7525	0.7178	0.6512	0.9347	0.7115	**0.5261**
MeanRank	6.13	5.38	4.13	2.88	5.38	3.13	**1.00**

的检验方法进行结果分析,得到平均排名和性能分析.

平均排名是通过弗里德曼(Friedman)方法获得的.当有两个以上相关算法时,Friedman检验[20]是最著名的非参数检验.通过显著性分析,ALES算法的平均排名为1.00,在真实测井数据集的测试中排名第一.ALES优于现有的监督分类器算法和缺失值填补算法.

根据弗里德曼(Friedman)统计,可以拒绝"所有算法都具有相同性能"的假设.统计结果表明,这些算法的性能明显不同.使用事后Nemenyi检验在$\alpha=0.05$的显著性水平上进一步比较算法.表5是最后的测试结果,可见ALES算法明显优于其他六种算法.根据Friedman检验计算出的p值排名结果,首先缺失值填补算法普遍要比传统的监督算法的平均代价要小,而在缺失值填补算法中,ALES算法比其他的填补算法平均代价小,填补效果更好,分类精度较高.

图2显示了ALES和六种算法在10%,20%,30%,40%和50%缺失率下的平均代价对比.对于三个真实测井数据集,ALES算法的平均代价曲线明显低于其他算法,这表明ALES算法在各个缺失率下都有不错表现.

表5　数据缺失50%时ALES和六种对比算法的post-hoc对比

Table 5　Post-hoc of ALES and other six algorithms when the datasets are missing at 50%

算法	$z=(R_0-R_i)/SE$	p
ALES vs. NB	3.3551	0.0008
ALES vs. kNN	2.8641	0.0042
ALES vs. GESI	2.8641	0.0042
ALES vs. J48	2.0458	0.0408
ALES vs. BPCA	1.3911	0.1642
ALES vs. CALF	1.2275	0.2196

4　结　论

在属性缺失的情况下实现储层的准确预测是一个困难而且有意义的问题,本文从数据模型、代价敏感的优化方法和算法设计三个层面研究了该问题.数据模型考虑了不完备的数据、属性查询代价和标签查询代价;优化方法定义了代价函数,来获得属性值和标签的统一评估;算法设计给出

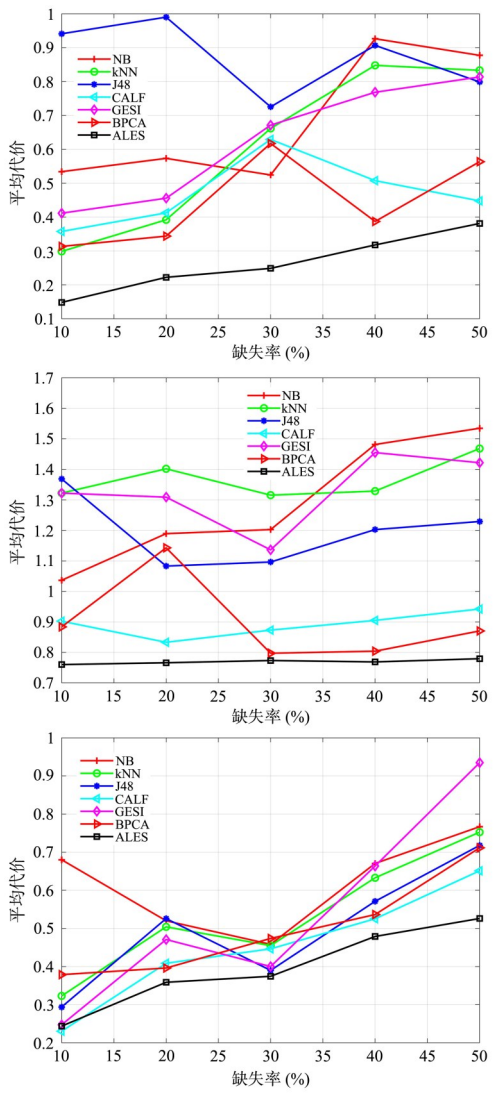

图 2 不同缺失率时 ALES 和六种算法的平均代价对比(从上至下分别对应:Well 01;Well 02;Well 03)

Fig.2 The average cost of ALES and six algorithms on different missing ratio (Well 01, Well 02 and Well 03 from up to dowm)

了各种输入、输出和优化目标.实验结果表明,ALES算法可以以较低的代价填补缺失值,也可以更好地识别含气层.同时,显著性分析的结果证明ALES算法优于其他监督算法和填补算法.

参考文献

[1] Zahin S A, Ahmed C F, Alam T. An effective method for classification with missing values. Applied Intelligence, 2018, 48(10): 3209－3230.

[2] Zhang J, Clayton M K, Townsend P A. Missing data and regression models for spatial images. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(3): 1574－1582.

[3] Silva-Ramírez E L, Pino-Mejías R, López-Coello M, et al. Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks, 2011, 24(1): 121－129.

[4] Azadeh A, Asadzadeh S M, Jafari-Marandi R, et al. Optimum estimation of missing values in randomized complete block design by genetic algorithm. Knowledge-Based Systems, 2013, 37: 37－47.

[5] Melville P, Saar-Tsechansky M, Provost F, et al. Active feature-value acquisition for classifier induction// The 4th IEEE International Conference on Data Mining. Brighton, United Kingdom: IEEE, 2004: 483－486.

[6] Kwon O, Sim J M. Effects of data set features on the performances of classification algorithms. Expert Systems with Applications, 2013, 40(5): 1847－1857.

[7] Min F, Liu F L, Wen L Y, et al. Tri-partition cost-sensitive active learning through kNN. Soft Computing, 2019, 23(5): 1557－1572.

[8] Settles B. Active learning. San Rafael: Morgan & Claypool Publishers, 2012: 1－114.

[9] Tong S, Koller D. Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research, 2002, 2(1): 45－66.

[10] Wang M, Min F, Zhang Z H, et al. Active learning through density clustering. Expert Systems with Applications, 2017, 85: 305－317.

[11] Wang M, Fu K, Min F, et al. Active learning through label error statistical methods. Knowledge-Based Systems, 2020, 189: 105140.

[12] Rodriguez A, Laio A. Machine learning. clustering by fast search and find of density peaks. Science, 2014, 344(6191): 1492－1496.

[13] Allcock J, Zhang S Y. Quantum machine learning. National Science Review, 2019, 6(1): 26－28.

[14] Dennis J E, Moré J J. Quasi-newton methods, motivation and theory. SIAM Review, 1977, 19(1): 46－89.

[15] 黄帷,闵帆,任杰.基于协同过滤加权预测的主动学

习缺失值填补算法.南京大学学报(自然科学),2018,54(4):758－765.(Huang W,Min F,Ren J. Missing value imputation with active learning based on collaborative filtering weighted prediction. Journal of Nanjing University (Natural Science),2018,54(4):758－765.)

[16] Gheyas I A,Smith L S. A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing,2010,73(16－18):3039－3065.

[17] Meng F C,Cai C,Yan H. A bicluster-based bayesian principal component analysis method for microarray missing value estimation. IEEE Journal of Biomedical and Health Informatics,2014,18(3):863－871.

[18] Holmes G,Donkin A,Witten I H. WEKA:A machine learning workbench//Proceedings of ANZIIS'94:Australian New Zealnd Intelligent Information Systems Conference. Brisbane,Australia:IEEE,1994:357－361.

[19] Triguero I,González S,Moyano J M,et al. KEEL 3.0:an open source software for multi-stage analysis in data mining. International Journal of Computational Intelligence Systems,2017,10(1):1238－1249.

[20] Reyes O,Altalhi A H,Ventura S. Statistical comparisons of active learning strategies over multiple datasets. Knowledge-Based Systems,2018,145:274－288.

(责任编辑　杨可盛)