计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):38−44 [doi: 10.15888/j.cnki.csa.006948]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 EM 和 GMM 的朴素贝叶斯岩性识别①

赵　铭[1], 金大权[2], 张　艳[3], 高世臣[1], 仲婷婷[1]

[1](中国地质大学 (北京) 数理学院, 北京 100083)
[2](中国石油长庆油田公司第四采气厂, 西安 710016)
[3](中国地质大学 (北京) 地球物理与信息技术学院, 北京 100083)
通讯作者: 张　艳, E-mail: 632119714@qq.com

摘　要: 朴素贝叶斯分类器可以应用于岩性识别. 该算法常使用高斯分布来拟合连续属性的概率分布, 但是对于复杂的测井数据, 高斯分布的拟合效果欠佳. 针对该问题, 提出基于 EM 算法的混合高斯概率密度估计. 实验选取苏东 41-33 区块下古气井的测井数据作为训练样本, 并选取 44-45 号井数据作为测试样本. 实验采用基于 EM 算法的混合高斯模型来对测井数据变量进行概率密度估计, 并将其应用到朴素贝叶斯分类器中进行岩性识别, 最后用高斯分布函数的拟合效果作为对比. 结果表明混合高斯模型具有更好的拟合效果, 对于朴素贝叶斯分类器进行岩性识别的性能有不错的提升.

关键词: 概率密度估计; EM 算法; 朴素贝叶斯分类器; 岩性识别

引用格式: 赵铭,金大权,张艳,高世臣,仲婷婷.基于 EM 和 GMM 的朴素贝叶斯岩性识别.计算机系统应用,2019,28(6):38–44. http://www.c-s-a.org.cn/1003-3254/6948.html

Naive Bayesian Lithology Recognition Based on EM and GMM

ZHAO Ming[1], JIN Da-Quan[2], ZHANG Yan[3], GAO Shi-Chen[1], ZHONG Ting-Ting[1]

[1](School of Science, China University of Geosciences, Beijing 100083, China)
[2](Fourth Gas Production Plant, Petro China Changqing, Xi'an 710016, China)
[3](School of Geophysics and Information Technology, China University of Geosciences, Beijing 100083, China)

Abstract: Naive Bayesian classifier can be applied to lithologic identification. The Gaussian distribution is often used to fit the probability distribution of continuous attributes, but it is not effective for complex logging data. To solve this problem, a hybrid Gaussian probability density estimation based on EM algorithm is proposed. Logging data of the lower ancient gas Wells in the block 41-33 of Sudong are selected as training samples, and data of 44-45 Wells are selected as test samples. The experiment uses the mixed Gaussian model based on EM algorithm, to estimate the probability density of logging data variables at first, and then applies it to the Naive Bayes classifier for the lithology identification. Finally, the fitting effect of the single Gaussian distribution function was used as the comparison. The results reveal that the mixed Gaussian model has a better fitting effect and the performance of the Naive Bayes classifier for the lithology identification could be improved through this way.

Key words: probability density estimation; EM algorithm; Naive Bayesian classification; lithology identification

1　引言

贝叶斯网络源于概率统计学, 作为数据挖掘和机器学习的重要方法之一, 被人们广泛的应用. 朴素贝叶斯 (Naive Bayes) 分类方法是贝叶斯网络的简化, 具有

① 收稿时间: 2018-11-10; 修改时间: 2018-12-04, 2019-01-07; 采用时间: 2019-01-18; csa 在线出版时间: 2019-05-25

坚实的理论基础, 和其他分类方法相比, 展现出高速度和高效率, 被广泛应用于模式识别, 数据挖掘以及机器学习中[1]. 朴素贝叶斯分类方法基于条件独立性假设, 即假设一个变量对分类的影响独立于其他变量. 当独立性假设成立时, 与其它分类方法相比, 朴素贝叶斯方法理论上具有最小的误分类率. 在实际的应用中, 对于连续变量的数据. 我们通常假设变量服从高斯分布, 通过 EM 算法求得各个变量所服从高斯分布的均值和方差, 从而可以得到变量不同取值的概率作为后验概率. 再根据贝叶斯定理, 构造朴素贝叶斯分类器, 从而实现对数据分类的结果. 而混合高斯模型 GMM 是指多个高斯分布函数的线性组合. 理论上, GMM 模型可以拟合出任意变量的分布. 使用混合高斯模型代替原有的高斯分布作为变量的概率密度函数, 可以提升连续变量的概率密度拟合效果, 从而改进了朴素贝叶斯分类器对连续型数据的分类能力.

2 朴素贝叶斯

2.1 贝叶斯方法

贝叶斯方法提供了一种通过概率进行推理的手段. 它假定待考查的变量遵循某种概率分布, 且可根据这些概率以及已经观察到的数据进行推理, 从而做出最优的决策[2–5]. 我们通过贝叶斯定理的公式来介绍这一方法:

$$P(h|D)=\frac{P(h)P(D|h)}{P(D)} \tag{1}$$

当给定训练集合 D, 假设空间 H 中的最有可能假设可以通过贝叶斯公式来计算.

其中, $P(h)$表示还没有进行训练前, 假设 h 拥有的初始概率, 即h的先验概率, 它通常根据关于 h 是一正确假设的概率的背景知识. 在没有先验知识的情况下, 通常可以认为候选假设服从均匀分布, 即把每一个候选假设赋予相同的概率. $P(D)$表示将要观察的训练实例集 D 的先验概率, 即在没有确定某一假设成立时 D 的概率, 通常可以用全概率公式求出. $P(h|D)$表示给定训练实例集 D 时 h 成立的概率, 即h的后验概率, 通常理解为在看到训练实例集 D 后, h成立的置信度.

当变量属性是离散型时, 类的先验概率$P(h)$可以通过训练集的各类样本出现的次数来估计. 当变量属性是连续型时, 有两种方法来估计属性的后验概率 $P(h|D)$. 第一种方法是把每一个连续的变量属性离散化, 然后用相应的离散区间替换连续属性值, 但这种方法不好控制离散区间划分的粒度. 第二种方法是, 可以假设连续变量服从某种概率分布, 然后使用训练数据估计分布的参数, 高斯分布通常被用来表示连续属性的类条件概率分布.

2.2 朴素贝叶斯

朴素贝叶斯, 简单来说就是对于给出的待分类项, 假设各个属性之间是相对独立的, 求解在此项出现的条件下各个类别的概率最大值. 然后将其归类于所求解出的最大值所属的类别. 在属性相对独立的假设下, 朴素贝叶斯分类器具有简单的星型结构. 每个属性结点只有唯一的父类结点, 这意味着, 当类给定时, 属性之间条件独立[6].

$$P(c|x)=\frac{P(c)P(x|c)}{P(x)}=\frac{P(c)}{P(x)}\prod_{i=1}^{d}P(x_i|c) \tag{2}$$

其中, d为属性数目, x_i为x在第i个属性上的取值.

对于所讨论的所有类别来说, $P(x)$都是相同的, 故所得判别准则如下:

$$h(x)=\arg\max P(c)\prod_{i=1}^{d}P(x_i|c) \tag{3}$$

即, 所判类别为属于赋予先验概率为权重的概率乘积的最大值.

在分类器中, 我们对每个属性条件概率$P(x_i|c)$的估计是首要的计算部分, 只有求出条件概率才能进行贝叶斯分类的判别[7]. 在本文中, 我们分别用高斯模型和混合高斯模型来进行概率密度估计, 再构造朴素贝叶斯分类器进行对比.

3 概率密度估计

3.1 高斯混合模型

当贝叶斯分类器选取连续变量的时候, 需要知道各个变量的概率密度函数. 一般情况下, 我们通常假设各个变量服从高斯概率分布. 然而, 测井数据中的各个变量通常不能完全服从高斯概率分布, 拟合效果误差较大. 针对这种情况, 本文考虑使用混合高斯概率模型 (GMM) 来拟合各个测井数据的概率密度分布.

混合高斯模型的数学模型为:

$$f(x)=\sum_{i=1}^{m}\varepsilon_i*Guass(\mu_i,\sigma_i) \tag{4}$$

其中, ε_i是表示第i个高斯项的权重或者称为混合系数, 且$\sum_{i=1}^{m}\varepsilon_i=1$. $Guass(\mu_i,\sigma_i)$表示高斯密度函数, μ_i和σ_i分别为高斯密度函数的均值和方差. GMM 模型使用的高斯模型的个数称为 GMM 高斯模型的阶数[7,8]. 通常情况下, GMM 模型进行概率密度估计的阶数不易过大或者过小. 阶数过大会导致参数估计过程难以收敛, 阶数过小会导致参数估计误差较大. 本文考虑选取五阶 GMM 模型进行概率密度估计.

3.2 EM 算法

采用 GMM 模型进行概率密度估计, 便要对 GMM 模型中进行参数估计, 通常可以采用极大似然估计法获得参数, 然而极大似然估计需要知道观测数据由哪个高斯分模型产生. 如果不清楚观测数据由哪个分模型产生, 即不确定每个数据所属的分类. 这就意味着需要使用隐变量来进行参数估计, 针对这种情况选取 EM 算法解决 GMM 模型的参数估计问题. 本文选取的测井数据并不知道每个数据所属的分类, 所以选取 EM 算法来估计 GMM 模型的参数.

EM 算法以极大似然估计为基本思想, 采用迭代的方法进行参数估计. EM 算法的流程可以分为 E 步骤和 M 步骤. 首先要初始化分布参数θ; 然后重复 E、M 步骤直到收敛[9–11]:

E 步骤: 根据参数θ初始值或上一次迭代所得参数值来计算出隐性变量的后验概率 (即隐性变量的期望), 作为隐性变量的估计值:

$$Q_i(\mathrm{z}^{(i)}) := p(z^{(i)}|x^{(i)};\theta). \tag{5}$$

M 步骤: 将似然函数最大化以获得新的参数值:

$$\theta := \arg\max_{\theta}\sum_{i}\sum_{z^{(i)}}Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})} \tag{6}$$

4 实例分析

研究数据来自苏里格气田 41-33 区块下古气井的测井曲线. 该地区岩性为复杂的碳酸盐, 主要有 7 种岩性, 分别是石灰岩、白云质石灰岩、泥质石灰岩、白云岩、灰质白云岩、泥质白云岩和泥岩. 根不同的测井参数及其不同的组合形式可以识别不同的岩性. 选取的测井参数不同, 岩性识别的效果具有很大的差异. 因此, 根据不同测井参数区分不同岩性的敏感性分析[12,13], 结合人工判别岩性的经验, 最终确定自然伽马 (GR), 补偿中子 (CNL)、密度 (DEN)、声波时差 (AC)、光电吸收截面指数 (PE) 和深侧向电阻率 (RLLD) 六个测井参数作为朴素贝叶斯分类器的分类属性.

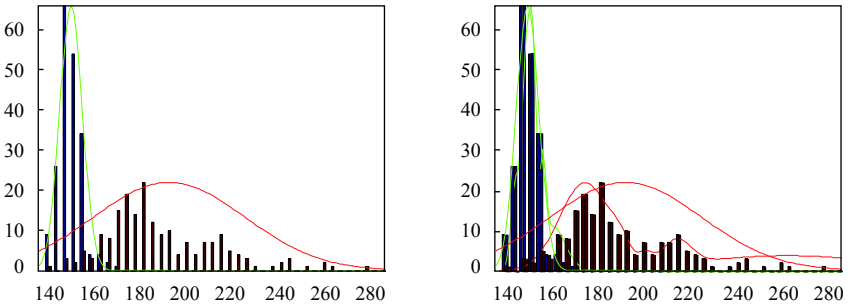
分别选取石灰岩、白云质石灰岩、泥质石灰岩、白云岩、灰质白云岩、泥质白云岩和泥岩各 200 个样本, 共 1400 条样本作为测试集. 其中深侧向电阻参数取值范围过大, 结合先前的处理经验, 对其进行对数处理 (log10). 对六个测井曲线参数进行量纲化, 避免不同量纲对实验结果造成不良影响. 经过上述处理过的数据, 作为实验的训练集.

针对实验选用的训练集, 首先分别用高斯模型和混合高斯模型对选取的 6 个测井参数进行概率密度估计, 然后对比概率密度估计效果. 高斯模型主要是对每种岩性的不同测井参数的均值和方差进行 EM 算法迭代估计, 得到每种岩性的不同测井参数的均值和方差, 从而得到高斯模型的参数, 以此作为先验信息构造朴素贝叶斯分类器. 而混合高斯模型是用 EM 算法迭代每种岩性的不同测井参数的均值, 方差以及每个高斯模型的权重, 从而得到混合高斯模型的参数, 并以此作为先验信息构造朴素贝叶斯分类器. 当朴素贝叶斯分类器处理连续属性时, 通常假设连续属性服从某种分布, 这里分别用高斯分布和混合高斯分布作为连续属性的概率密度分布函数. 同时对不同概率密度模型作用下的朴素贝叶斯分类器分类效果作对比, 选训练集中的白云岩和泥岩中的 AC 测井参数, 来对两种不同的概率密度函数估计效果进行分析, 并根据两种概率密度函数的曲线分析分类器的分类效果. 概率密度估计效果如图 1 所示.

在图 1 中, 根据所选取的数据, 左边蓝色直方图和右边红色直方图分别代表了白云岩、泥岩数据真实的分布, 图中绿色和红色的线分别代表白云岩和泥岩的拟合的概率密度曲线, 图 1(a) 和图 1(b) 分别为高斯模型拟合效果图和混合高斯模型拟合效果图.

为了更好地比较高斯模型和混合高斯模型的概率密度拟合效果, 引入“误判区”这个概念. 图 2 给出两个等概率类别的例子, 同时给出了最简单情况下x的函数$p(x|\omega_i)$, i=1,2 的变化情况. x_0处的虚线是将特征空间分为R_1, R_2两个区域. 根据贝叶斯决策规则, 对于R_1区域的所有x值, 分类器都判定属于ω_1, 而对于R_2区域的所有x值, 都判定属于ω_2. 但是, 从图中可以判定错误是避免的. 错误率P_e的计算公式为:

$$2P_e = \int_{-\infty}^{x_0} p(x|\omega_2)dx + \int_{x_0}^{+\infty} p(x|\omega_1)dx \tag{7}$$



(a) 单高斯拟合效果 (b) GMM 拟合效果

图 1 白云岩、泥岩 AC 估计效果对比

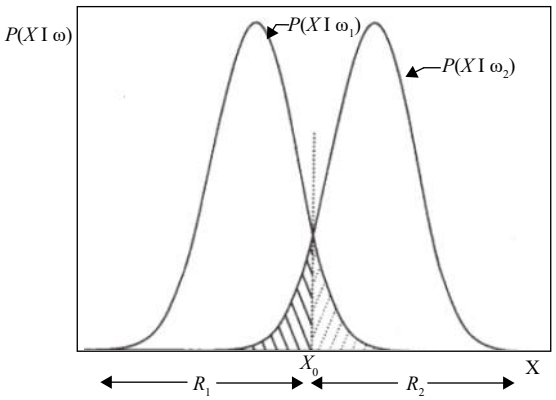


图 2 由两个等概率类别的贝叶斯分类器形成的 R_1 和 R_2 两区域的例子

式 (7) 和图 2 中的阴影部分的面积相等. 因此, 我们把两条概率密度曲线交汇的阴影部分的面积称为误判区[14].

根据图 1, 从概率密度函数的拟合效果上来看, 混合高斯模型拟合的概率密度曲线比高斯模型拟合的概率密度曲线更贴近代表真实分布的直方图. 所以混合高斯模型拟合出来的概率密度曲线更符合测试集数据的真实分布情况. 其次, 两种岩性的测井参数概率密度曲线与坐标轴所围的面积, 分别代表根据 AC 属性来判断属于白云岩和泥岩的样本. 两条概率密度曲线交汇部分与横轴所围面积代表误判区. 误判区的面积越小, 代表两种岩性基于当前样本的分离度越高. 因此为了提高朴素贝叶斯分类器的分类的准确率, 在选取不同的分布模型拟合样本的真实分布时, 应该选择误判区的面积小的分布模型. 从图 1 中可以看出, 混合高斯模型中, 绿色和红色两种岩性的概率密度曲线交汇处与坐标轴围成的面积相比于高斯模型来说更小, 因此选用混合高斯模型作为朴素贝叶斯分类器连续属性的分布函数时, 往往能取得更好的分类效果.

对于 1400 条训练样本, 我们分别采用高斯模型和混合高斯模型的概率密度估计方法对训练集数据进行概率密度估计. 根据 EM 算法得到的高斯模型均值和方差, 混合高斯模型的均值、方差和权重, 做出不同测井参数的概率密度曲线. 针对估计出的 6 个测井曲线属性概率密度函数, 构造朴素贝叶斯分类器, 记录训练样本分类的准确率.

图 3(a)–图 3(f) 从左向右分别依次为假设 AC、CNL、DEN、PE、GR、RLLD 服从高斯概率分布, 采用 EM 算法迭代估计出来的概率密度函数的均值和方差, 从而做出的概率密度函数的图像.

图 4(a)–图 4(f) 从左向右分别依次为假设 AC、CNL、DEN、PE、GR、RLLD 服从混合高斯概率分布, 采用 EM 算法迭代估计出来的概率密度函数的均值、方差以及每个高斯模型的权重, 从而做出的概率密度函数图像.

对比两个图像可以看出, 采用混合高斯概率密度模型估计出的函数模型更符合实际测井曲线资料的真实分布, 具有更好的拟合效果, 不同岩性的测井参数的概率密度曲线交汇部分与横轴所围成的面积更小, 即分类的误判区面积更小. 因此基于 GMM 模型的朴素贝叶斯分类器分类效果应该更好.

根据估计出来的 6 个属性的概率密度函数, 构造朴素贝叶斯分类器. 针对 1400 条训练样本进行训练, 统计分类的正确率, 即岩性识别的正确率, 根据单高斯模型得到的分类正确的样本数为 1106, 分类准确率为 79%, 根据混合高斯模型得到的分类正确的样本数为 1176, 准确率为 84%. 可见, 混合高斯拟合的变量概率密度对于朴素贝叶斯分类器的分类准确性有一定的提升.

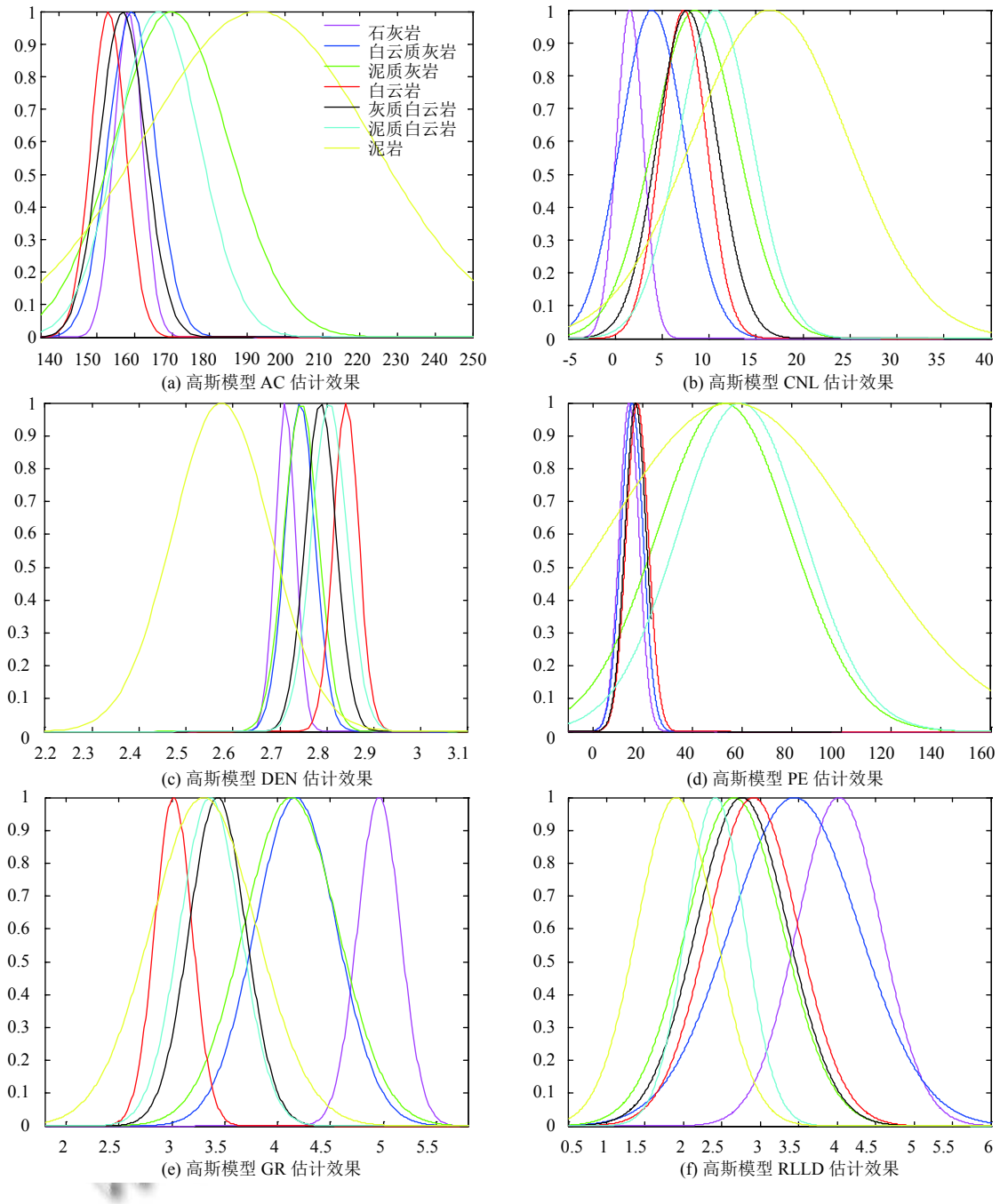


图 3 高斯概率密度估计效果对比

选取 41-33 区块下井号为 44-45 的古气井测井曲线作为测试样本. 选取 44-45 井的 557 条测井曲线数据, 同样选取自然伽马 (GR), 补偿中子 (CNL)、密度 (DEN)、声波时差 (AC)、光电吸收截面指数 (PE) 和深侧向电阻率 (RLLD) 六个属性作为分类指标属性, 其中电阻率仍然进行对数处理 (log10). 测试集的岩性识别效果如图 5 所示.

从图 5 可以看出, 本次测试使用三种方法进行岩性识别, 钻井岩性代表数据真实的岩性, 7 种岩性分别用不同的颜色表示出来, 通过和钻井岩性一列的颜色进行对比, 可以看出岩性识别效果的优劣. 分别采用中心距离判别法, 高斯模型的朴素贝叶斯和 GMM 模型的朴素贝叶斯三种方法进行测试. 根据钻井岩性对比三种方法的识别结果, 通过对比三种方法识别结果和

钻井岩性的颜色可以看出, 采用中心距离判别法进行岩性识别的效果较差, 因为只根据测井数据的均值来进行分类, 选择距离均值距离最近的类别作为分类的类别, 误判区较大. 而传统朴素贝叶斯岩性识别效果要远优于中心距离判别法, 主要是因为在概率密度曲线拟合的过程中, 考虑了均值和方差共同的影响效果, 因而岩性识别效率得到了提升. 基于混合高斯模型的朴素贝叶斯分类器分类效果比传统朴素贝叶斯效果分类更好, 主要因为在概率密度拟合的过程中, 相比于高斯模型, 混合高斯模型能够更好地拟合测井数据的实际分布, 减小分类的误判区, 因而所得到的岩性识别效率最高.

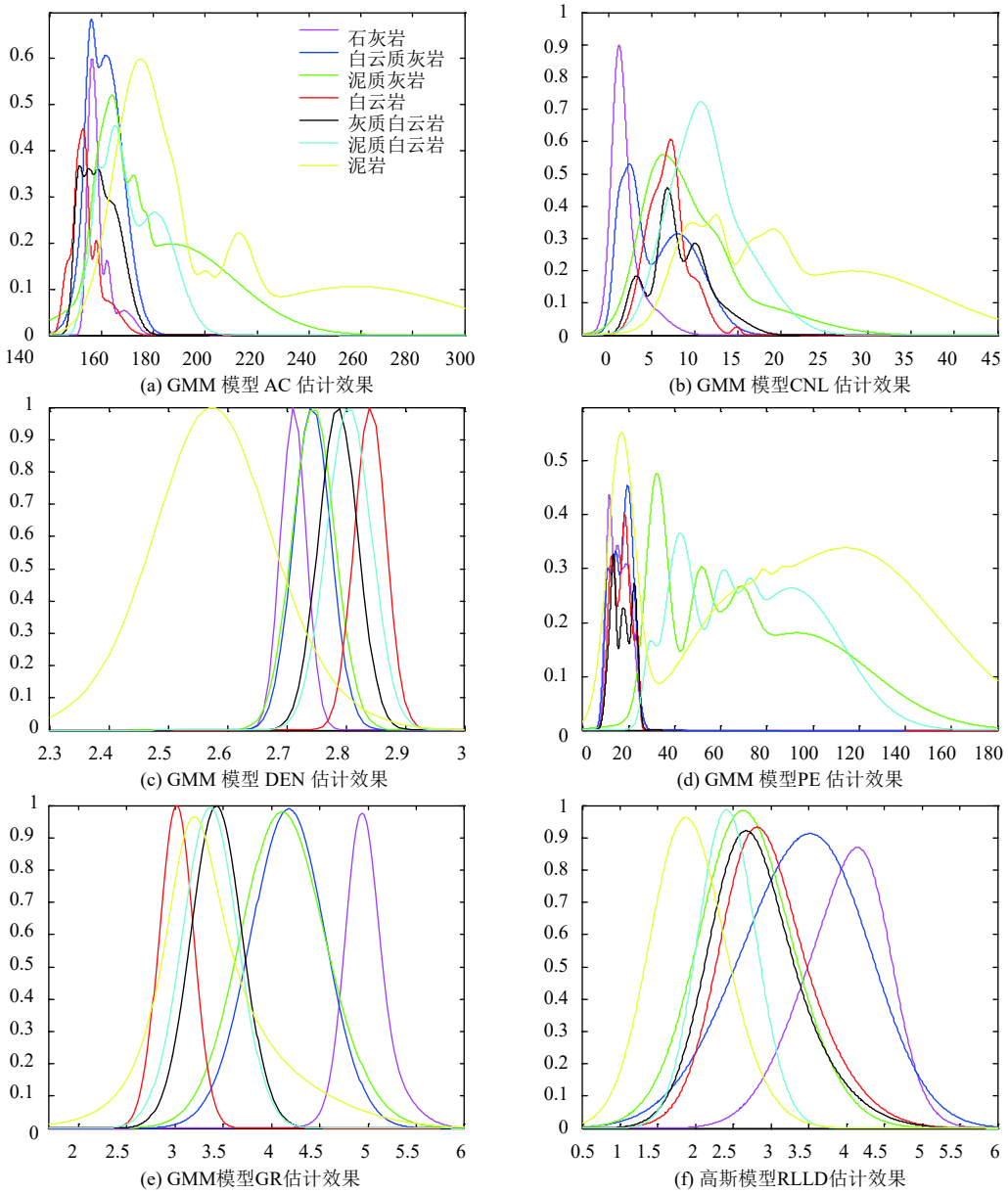


图 4 混合高斯概率密度估计效果对比

5 总结

本文提出了一种基于 EM 和 GMM 的朴素贝叶斯分类器模型用于岩性识别. 通过对测井曲线参数对不同岩性的敏感度分析, 选取了 AC, CNL, DEN, PE, GR, RLLD 六个参数作为朴素贝叶斯的分类变量. 通过 EM 算法进行参数迭代, 使用混合高斯模型来拟合每个

分类变量的真实概率分布, 构建贝叶斯分类器, 从而实现岩性识别. 相比于传统朴素贝叶斯分类器, 混合高斯模型比高斯模型具有更好的拟合效果, 不同岩性之间的误判区也更小. 在训练集样本中基于混合高斯模型的朴素贝叶斯分类器岩性识别准确率为 84%, 传统朴素贝叶斯分类器的准确率为 79%, 因此基于混合高斯模型的朴素贝叶斯分类器可以提升分类器的分类效果. 但是, 用于构建朴素贝叶斯分类器的变量现实中并不是完全独立的, 这会影响分类器的分类效果. 若想得到更好的分类效果, 可以借助一些专家经验, 预估各个分类变量之间的条件依赖, 或者通过贝叶斯网络结构学习算法构建贝叶斯网络, 用贝叶斯网络进行分类, 这样岩性识别的准确率会进一步提升.

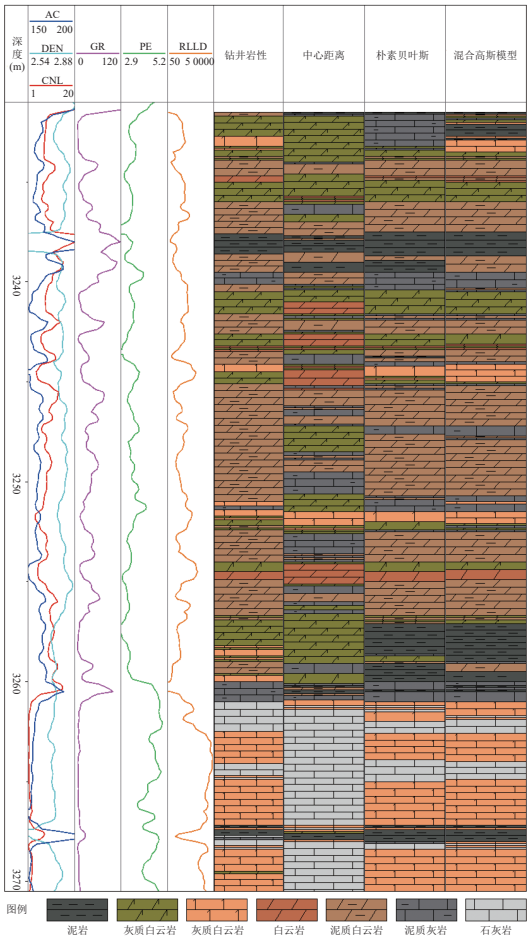


图 5 测试集岩性识别结果

参考文献

1 周志华. 机器学习. 北京: 清华大学出版社, 2016. 147–154.

2 彭兴媛, 刘琼荪. 不同类变量下属性聚类的朴素贝叶斯分类算法. 计算机应用, 2011, 31(11): 3072–3074.

3 金展, 范晶, 陈峰, 等. 基于朴素贝叶斯和支持向量机的自适应垃圾短信过滤系统. 计算机应用, 2008, 28(3): 714–718.

4 李晶辉, 张小刚, 陈华, 等. 一种改进隐朴素贝叶斯算法的研究. 小型微型计算机系统, 2013, 34(7): 1654–1658. [doi: 10.3969/j.issn.1000-1220.2013.07.041]

5 王玮, 陈恩红, 王煦法. 基于贝叶斯方法的知识发现. 小型微型计算机系统, 2000, 21(7): 703–705. [doi: 10.3969/j.issn.1000-1220.2000.07.009]

6 秦锋, 任诗流, 程泽凯, 等. 基于属性加权的朴素贝叶斯分类算法. 计算机工程与应用, 2008, 44(6): 107–109. [doi: 10.3778/j.issn.1002-8331.2008.06.033]

7 钟金琴, 辜丽川, 檀结庆, 等. 基于分裂 EM 算法的 GMM 参数估计. 计算机工程与应用, 2012, 48(34): 28–32, 59. [doi: 10.3778/j.issn.1002-8331.1206-0419]

8 徐定杰, 沈忱, 沈锋. 混合高斯分布的变分贝叶斯学习参数估计. 上海交通大学学报, 2013, 47(7): 1119–1125.

9 Hobolth A, Jensen JL. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Statistical Applications in Genetics and Molecular Biology, 2005, 4(1): 18.

10 Taheri S, Mammadov M. Learning the naive Bayes classifier with optimization models. International Journal of Applied Mathematics and Computer Science, 2013, 23(4): 787–795. [doi: 10.2478/amcs-2013-0059]

11 Jiang LX, Wang DH, Cai ZH, *et al*. Survey of improving naive Bayes for classification. Proceedings of the 3rd International Conference on Advanced Data Mining and Applications. Harbin, China. 2007. 134–145.

12 袁照威, 段正军, 张春雨, 等. 基于马尔科夫概率模型的碳酸盐岩储集层测井岩性解释. 新疆石油地质, 2017, 38(1): 96–102.

13 高世臣, 张丹. 多参数概率融合法在叠前地震储层预测中的应用--以苏里格气田苏 194 区块为例. 油气地质与采收率, 2015, 22(6): 61–67. [doi: 10.3969/j.issn.1009-9603.2015.06.011]

14 Theodoridis S, Koutroumbas K. 模式识别. 李晶皎, 王爱侠, 王骄, 译. 北京: 电子工业出版社, 2010: 8–10.