

模式识别 Fisher 判别分析

22009200601 汤栋文

2024 年 10 月 14 日

目录

| | |
|------------------------------|---|
| 摘要 | 2 |
| 1. 方法 | 2 |
| 1.1 数据集 | 2 |
| 1.2 线性判别分析 | 2 |
| 1.3 分类器选择 | 2 |
| 1.4 交叉验证 | 2 |
| 1.5 性能度量 | 3 |
| 2. 实验 | 3 |
| 1.1 Fisher 准则对分类准确率的影响 | 3 |
| 1.2 Fisher 准则对分类稳定性的影响 | 4 |
| 1.3 Fisher 准则对分类稳定性的影响 | 4 |
| 3. 结论 | 5 |

摘要

Fisher 准则是用于分类问题的两种密切相关的方法。它们都旨在找到一个投影方向，使得类内散度最小、类间散度最大。从而使得不同类别的数据点在这个方向上尽可能地分开。本报告在多个数据集和多种分类器下，探究了以下两个内容：

1. 分析了 Fisher 准则在多个数据集多种分类器上的有效性。
2. 发现了 Fisher 准则可能有利于降低分类器对数据的敏感度。

本报告在多个数据集和分类头上证明了 Fisher 准则总是能够提高分类准确率，降低冗余信息的干扰，提取出与类别最相关的信息。并且对数据进行可视化，直观地对比使用 Fisher 降维前后的差异。最后本报告分析了 Fisher 准则对分类器在数据敏感度上的影响。

1. 方法

1.1 数据集

本研究使用了 sonar 数据集和 iris 数据集。其中 sonar 数据集包含了 60 个特征（声纳信号返回的强度），并且每个样本都标记为岩石 (R) 或矿石 (M)。iris 数据集包含了鸢尾花属植物三个种类共 150 个样本的测量数据，每种类型的鸢尾花各有 50 个样本，每个样本记录了四个特征：萼片长度、萼片宽度、花瓣长度、花瓣宽度，单位均为厘米。

为了消除特征间的量纲影响，我们对原始特征进行了标准化处理，即减去平均值并除以标准差，使得每个特征具有零均值和单位方差。这也使得各种方法之间的对比完全公平。

1.2 线性判别分析

本研究应用 Fisher 准则进行降维，分别将 60 维和 4 维的原始数据降维到最佳的 1 维上。这样可以最大化类间散度的同时最小化类内散度，提取与类别最相关的信息并降低冗余信息的影响，不仅有助于减少计算复杂度，还可以突出那些能够最好地区分不同类别的特征。

1.3 分类器选择

为了评估 Fisher 准则对分类准确率的影响，我们选择了三种不同的分类器：朴素贝叶斯 (Gaussian Naive Bayes)、逻辑回归 (Logistic Regression)、K 近邻 (K-Nearest Neighbors)。这些分类器覆盖了概率模型、线性模型、基于样本的学习方法，因此能够提供多样化的视角来评价特征降维的效果。

1.4 交叉验证

对于每种分类器，我们分别使用了两种不同的交叉验证策略来估计模型性能：

留一法 (Leave-One-Out)：这种方法下，每次迭代中仅有一个样本被用作测试集，其余所有样本构成训练集。这种方法提供了无偏的性能估计，虽然这种方法计算量消耗较大，但在 sonar 数据集和 iris 数据集上，计算量可以接受。

随机数据集划分：这是一种随机划分数据集的方法，在此我们设定了 1000 随机的重复分

割，每次保留 20% 的数据作为测试集。这种策略可以提供关于模型稳定性的额外信息，因为它允许我们观察多次不同的随机分割下的性能变化。

1.5 性能度量

对于每种配置，我们都计算了分类准确率作为主要的性能指标。此外，在随机数据集划分策略下，我们也记录了 1000 次随机划分的准确率的标准差，以此来反映模型稳定性。

最后，我们还利用降维后的数据进行了可视化。对于原始数据，我们使用 PCA 降维到二维进行可视化。对于降维后的一维数据，我们通过添加以随机数为噪声的第二维度，避免一绘图时数据点过于密集，以便于观察和对比。通过绘制散点图直观地展示了基于 Fisher 准则的 LDA 前后数据点分布的可分情况。这样的可视化有助于更好地理解 Fisher 准则如何影响数据结构及分类边界。

2. 实验

1.1 Fisher 准则对分类准确率的影响

表格 1: sonar 数据集上的分类准确率 (acc) 和训练和完成一次推理的总时间 (time) 实验结果，LDA time 指的是对数据应用 Fisher 准则进行降维所需要的时间。其中 KNN 的 k 值选择最优值 k=8。

| sonar | LDA time (ms) | classifier | acc (%) | time (ms) |
|---------|---------------|------------|--------------|------------|
| w/ LDA | 3.5 (+3.5) | Bayers | 89.9 (+22.6) | 0.8 (-0.2) |
| | | LogisticR | 89.9 (+12.5) | 1.4 (-4.7) |
| | | KNN (k=8) | 91.3 (+3.8) | 1.5 (-7.0) |
| w/o LDA | - | Bayers | 67.3 | 1.0 |
| | | LogisticR | 77.4 | 6.1 |
| | | KNN (k=8) | 82.7 | 8.5 |

表格 2: iris 数据集上的分类准确率 (acc) 和训练和完成一次推理的总时间 (time) 实验结果，LDA time 指的是对数据应用 Fisher 准则进行降维所需要的时间。其中 KNN 的 k 值选择最优值 k=5。

| iris | LDA time (ms) | classifier | acc (%) | time (ms) |
|---------|---------------|------------|-------------|------------|
| w/ LDA | 0.5 (+0.5) | Bayers | 98.7 (+3.4) | 0.9 (-0.0) |
| | | LogisticR | 98.0 (+2.7) | 1.5 (-0.2) |
| | | KNN (k=5) | 98.7 (+4.0) | 1.2 (-0.0) |
| w/o LDA | - | Bayers | 95.3 | 0.9 |
| | | LogisticR | 95.3 | 1.7 |
| | | KNN (k=5) | 94.7 | 1.2 |

结论：

1. Fisher 准则能够有效去除与类别不相关的冗余信息，提高分类准确率。
2. 降维过程需要计算时间。当数据自身维数低时，引入额外的计算时间；当数据自身维数高时，虽然引入的额外的计算时间，但分类器的计算量可显著下降，总体时间可能减少。

1.2 Fisher 准则对分类稳定性的影响

表格 3: sonar 数据集上 1000 次随机数据集划分的平均分类准确率 (acc) 和分类准确率的标准差 (std)。

| sonar | classifier | mean acc (%) | std (x1000) |
|---------|------------|--------------|-------------|
| w/ LDA | Bayers | 89.4 (+22.7) | 1.8 (-2.9) |
| | LogisticR | 89.1 (+15.7) | 1.9 (-1.4) |
| | KNN (k=8) | 88.5 (+23.6) | 2.2 (-3.1) |
| w/o LDA | Bayers | 66.7 | 4.7 |
| | LogisticR | 73.4 | 3.3 |
| | KNN (k=8) | 64.9 | 5.3 |

表格 4: iris 数据集上 1000 次随机数据集划分的平均分类准确率 (acc) 和分类准确率的标准差 (std)。

| iris | classifier | mean acc (%) | std (x1000) |
|---------|------------|--------------|-------------|
| w/ LDA | Bayers | 97.1 (+3.4) | 1.6 (-1.7) |
| | LogisticR | 92.6 (+1.6) | 7.5 (+3.0) |
| | KNN (k=5) | 97.5 (+7.3) | 1.3 (-3.2) |
| w/o LDA | Bayers | 93.7 | 3.3 |
| | LogisticR | 91.0 | 4.5 |
| | KNN (k=5) | 90.2 | 4.5 |

结论：

- 在大多数情况下，Fisher 准则有助于降低分类器对数据的灵敏度。
- 基于 Fisher 准则的 LDA 方法有降低某些分类器上分类稳定性的风险。

1.3 Fisher 准则对分类稳定性的影响

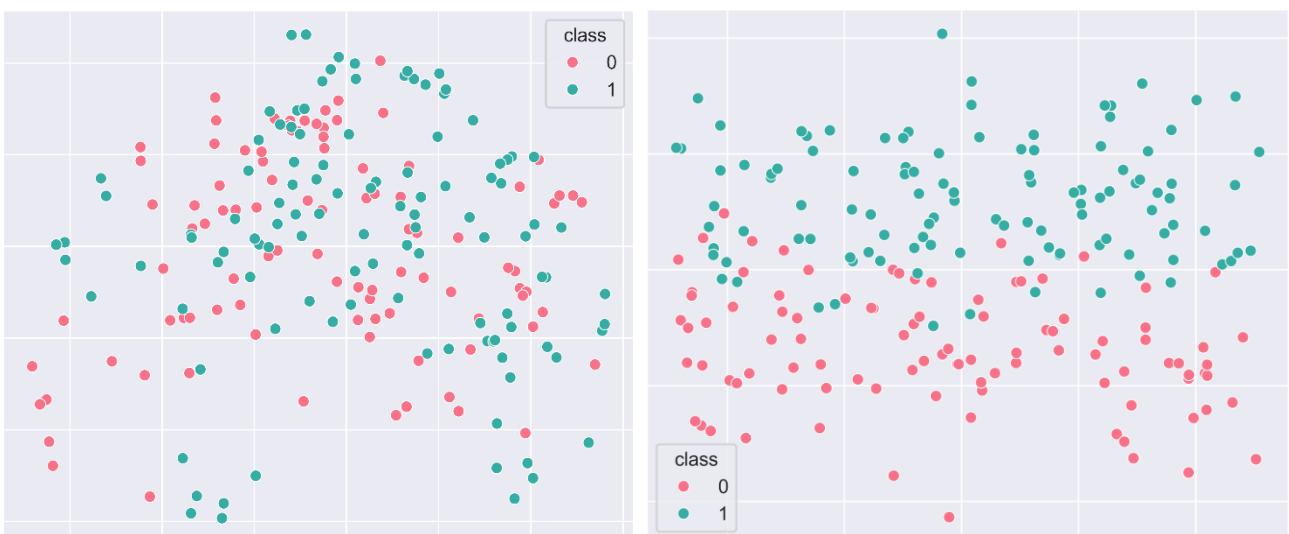


图 1：左：sonar 数据集原始数据使用 PCA 降维到 2 维的可视化。右：LDA 后的数据点可视化。

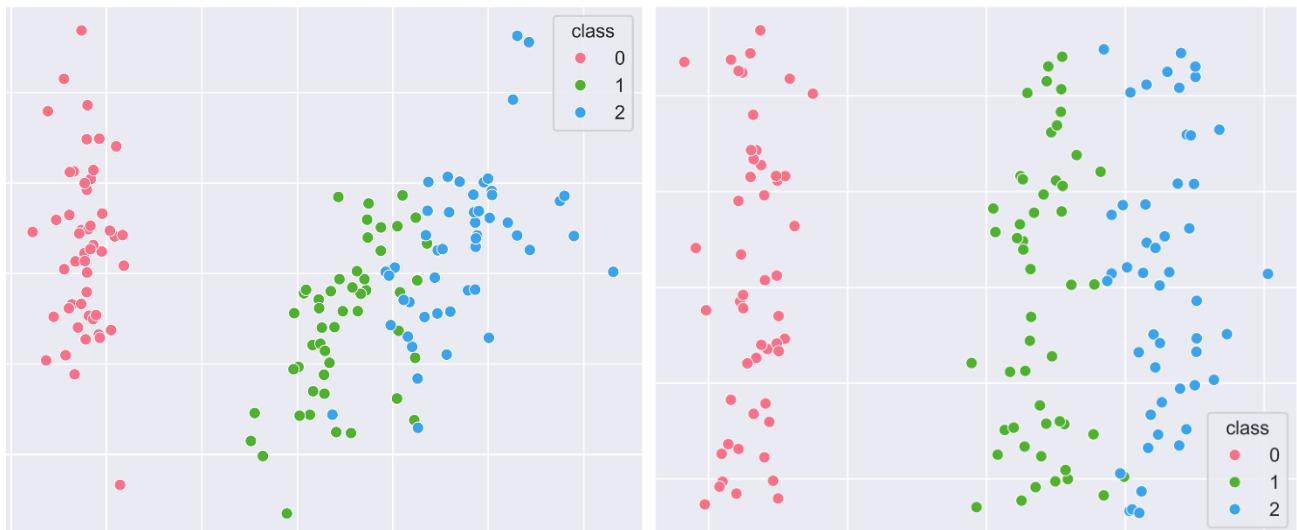


图 2: 左: iris 数据集原始数据使用 PCA 降维到 2 维的可视化。右: LDA 后的数据点可视化。

3. 结论

综上所述, Fisher 准则不仅能够显著提高分类准确率, 还能在大多数情况下增强模型的稳定性。然而, 它也可能在特定条件下影响某些分类器的稳定性。未来的工作可以进一步探索 Fisher 准则与其他分类器的结合, 以及在更复杂的数据集上的表现, 以期获得更全面的性能提升。

附录

Code:

https://github.com/MTDoven/Machine-Learning/tree/main/pattern_recognition/LDA

Reference:

<https://blog.csdn.net/AugustMe/article/details/97015004>

<https://blog.csdn.net/nmbnn8821750/article/details/137092656>

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html

<https://aistudio.baidu.com/datasetdetail/169732>

<http://archive.ics.uci.edu/ml/index.php>