COMPARING DISTRIBUTIONS BY MEASURING DIFFERENCES THAT AFFECT DECISION MAKING

Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan erreault,Jiaming Song, Stefano Ermon

摘要

测量两个概率分布之间的差异是机器学习和统计学中的一个基本问题。我们提出了一种新的基于决策任务最优损失的差异类别——如果最优决策损失在混合分布上高于单独分布上，则两个分布是不同的。通过合理选择决策任务，推广了 jensen - shannon 散度和最大平均差异族。我们将我们的方法应用于双样本测试，在各种基准上，与竞争方法相比，我们获得了更好的测试能力。此外，建模者可以在通过决策损失比较分布时直接指定他们的首选项。我们应用这一性质来理解气候变化对不同经济活动的影响，并针对不同的决策任务选择特征。

关键词：散度；分布

1 引言

测量两个概率分布之间的差异是机器学习和统计学中的一个基本问题。提出了一种新的基于决策任务最优损失的差异类别——如果最优决策损失在混合分布上高于单独分布上，则两个分布是不同的。通过合理选择决策任务，推广了 jensen - shannon 散度和最大平均差异族。如图 1所示：

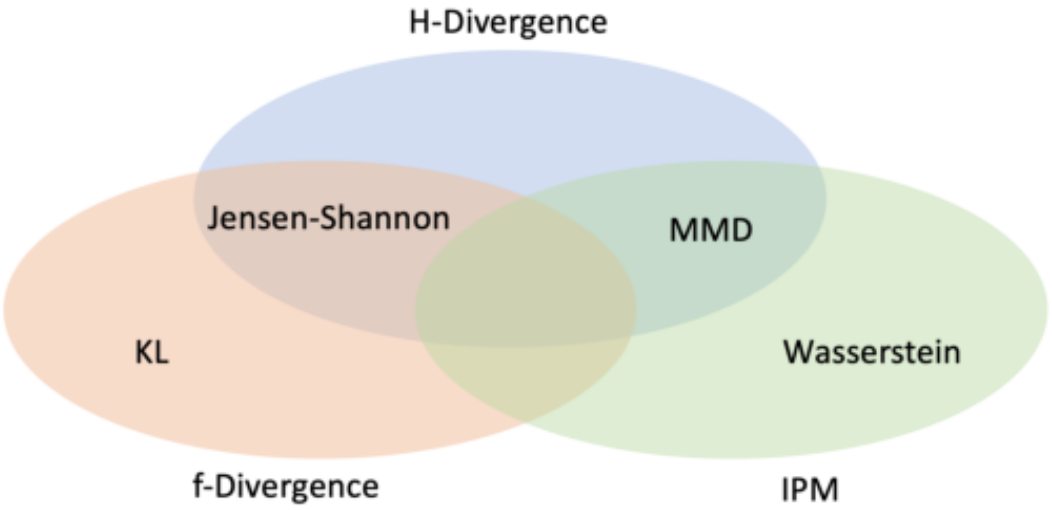

图 1: H 散度与现有散度之间的关系

2 相关工作

本次课程的论文复现工作拟通过将方法应用于双样本测试，在各种基准上，与竞争方法相比，获得了更好的测试能力。此外，建模者可以在通过决策损失比较分布时直接指定他们的首选项。应用这一性质来理解气候变化对不同经济活动的影响，并针对不同的决策任务选择特征。

3 本文方法

本文提出了一种新的方法，来比较 2 个分布的差异，并关注这种差异对于决策的影响。此方法为 H 散度：用于测 2 个概率分布间的差异；推广了一些众所周知的散度：比如詹森香农散度和最大均值散度族；本文证明了 H 散度的一些性质：通过使用有限样本集估计 H 散度的收敛结果。

4 复现细节

4.1 三个例子

用 3 个例子说明 H 散度的应用：双样本测试：表明在相同 I 类错误率下，H 散度比基于 MMD 的测试有更高测试功效；评估气候变化：表明从最小化损失角度，分布间差异是否足以影响不同实现中的决策；评估特征选择。

4.2 双样本检测实验

步骤：选用 4 个数据集：Blob、HDGM、HIGGS、MNIST；将每个数据集分成两个相等分区：调优超参数训练集、计算最终测试输出的验证集；每个排列测试用 100 个排列，运行每个测试 100 次来计算测试功率（正确输出 p!=q 的百分比）；重复整个实验 10 次，绘制并报告性能标准偏差。

5 实验结果分析

如图 2所示：获取 HIGGS 数据，选择的参数 n 不同，运行此实验如图 3所示：报告了平均测试功率，在 Higgs 上用比第二测试少两倍的样本实现了相同的测试功率

```
### HIGGS

Get the HIGGS data from here - https://drive.google.com/file/d/1sHIIFCoHbaukóMkbóe8a_tp1qnvuUOCc/view

To run HIGGS experiment, choose --n = 500,1000,1500,2500,4000,5000 and run
```
```
python higgs.py --n 500
```

图 2: HIGGS 实验运行步骤

N	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D	H-Div
1000	0.120±0.007	0.095±0.007	0.082±0.015	0.097±0.014	0.132±0.005	0.113±0.013	**0.240±0.020**
2000	0.165±0.019	0.130±0.019	0.183±0.026	0.232±0.032	0.291±0.017	0.304±0.012	**0.380±0.040**
3000	0.197±0.012	0.142±0.025	0.257±0.049	0.399±0.058	0.376±0.022	0.403±0.050	**0.685±0.015**
5000	0.410±0.041	0.261±0.044	0.592±0.037	0.447±0.045	0.659±0.018	0.699±0.047	**0.930±0.010**
8000	0.691±0.067	0.467±0.038	0.892±0.029	0.878±0.020	0.923±0.013	0.952±0.024	**1.000±0.000**
10000	0.786±0.041	0.603±0.066	0.974±0.007	0.985±0.005	1.000±0.000	1.000±0.000	**1.000±0.000**
Avg.	0.395	0.283	0.497	0.506	0.564	0.579	**0.847**

Table 1: Average test power ± standard error for N samples over the HIGGS dataset. The results on MNIST is similar and presented in Table 3, Appendix B.1.

图 3: HIGGS 实验结果

如图 4所示：获取 MNIST 数据，选择的参数 n 不同，运行此实验如图 5所示：报告了平均测试功率，在 MNIST 上，即使在 (Liu et al, 2020) 中评估的最小样本量上，也可以实现完美的测试能力。

```
### MNIST

Get the fake MNIST data from here - https://drive.google.com/file/d/13JpGbp7PEm4PfZ6VeqpFiy0lHfVpy5Z5/view

To run MNIST experiment, choose --n = 100,200,300,400,500 and run
```
```
python mnist.py --n 100
```

图 4: MNIST 实验运行步骤

N	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D	H-Div
200	0.414±0.050	0.107±0.018	0.193±0.037	0.234±0.031	0.188±0.010	0.555±0.044	**1.000±0.000**
400	0.921±0.032	0.152±0.021	0.646±0.039	0.706±0.047	0.363±0.017	0.996±0.004	**1.000±0.000**
600	1.000±0.000	0.294±0.008	1.000±0.000	0.977±0.012	0.619±0.021	1.000±0.000	**1.000±0.000**
800	1.000±0.000	0.317±0.017	1.000±0.000	1.000±0.000	0.797±0.015	1.000±0.000	**1.000±0.000**
1000	1.000±0.000	0.346±0.019	1.000±0.000	1.000±0.000	0.894±0.016	1.000±0.000	**1.000±0.000**
Avg.	0.867	0.243	0.768	0.783	0.572	0.910	**1.000**

Table 3: Average test power ± standard error for N samples over the MNIST dataset.

图 5: MNIST 实验结果

如图 6所示：获取 HDGM 数据，选择的参数 n、d 不同，运行此实验如图 7所示：报告了 HDGM 数据集的平均测试功率。左: 相同样本量 (4000)，不同数据维度的结果。右: 相同样本维数 (10)，不同样本容量的结果。此方法 (H-Div, 虚线的部分) 为几乎所有设置实现更好的测试功率。所有测试对于低数据维都具有较高的测试能力，但此方法对于高数据维的伸缩性更好。

```
### HDGM

To run HDGM experiment, choose --vtype = vjs, --n = 100,1000,1500,2500, and --d = 3,5,10,15,20 and run
```
```
python hdgm.py --exptype power --vtype vjs --n 100 --d 3
```

图 6: HDGM 实验运行步骤

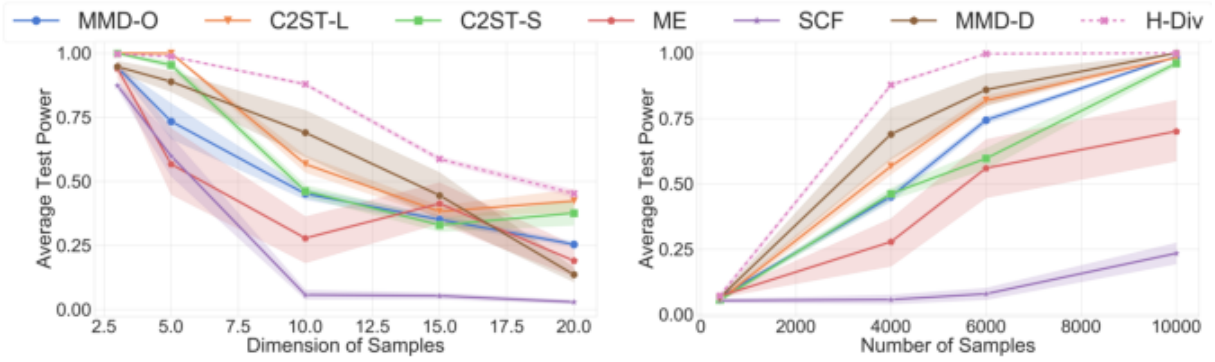

图 7: HDGM 实验结果

如图 6所示：运行此实验的步骤如图 7所示：左: 不同样本量且显著性水平 $\alpha = 0.05$ 时 Blob 数据集上的平均测试功率。此方法 (H-Div，虚线) 具有明显更好的测试能力，特别是对于小样本量的设置。右: 相同的图，显著性水平 $\alpha = 0.01$。

```
### Blob

To run Blob experiment with KDE
```
python blob_kde.py --exptype power --vtype vmin
```
To run Blob experiment with GMM
```
python blob_gmm.py --exptype power --vtype vmin
```
```

图 8: Blob 实验运行步骤

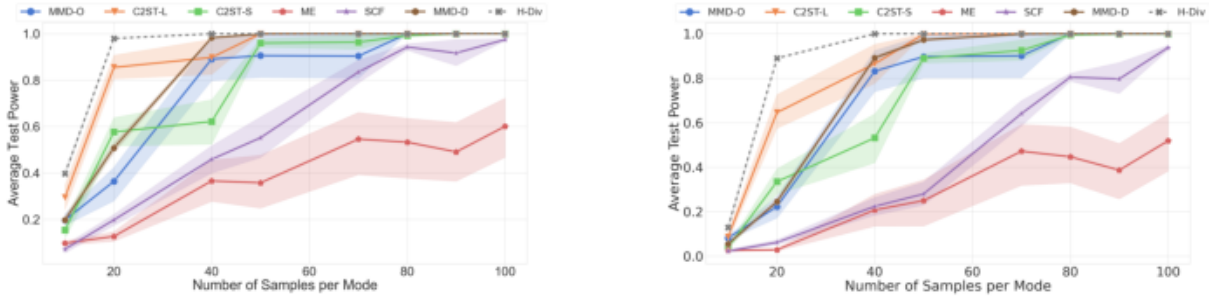

图 9: Blob 实验结果

如图 6所示：运行此实验的步骤如图 7所示：与农业相关的损失在不同地理位置的 H 散度示例图。颜色越深表示 H 散度越大。与 KL 等散度相比，H-散度衡量与不同社会和经济活动相关的变化 (通过选择适当的损失函数)。例如，尽管气候变化对高纬度或高海拔地区有显著影响，但这种变化与农业的相关性较小 (因为在这些地区几乎不可能进行农业活动)。

```
1. Run python vdiv_agriculture.py to get the vdivergences

2. Run plotting_map_crop.py to plot the vdivergence on the map
```

图 10: 评估气候变化实验运行代码

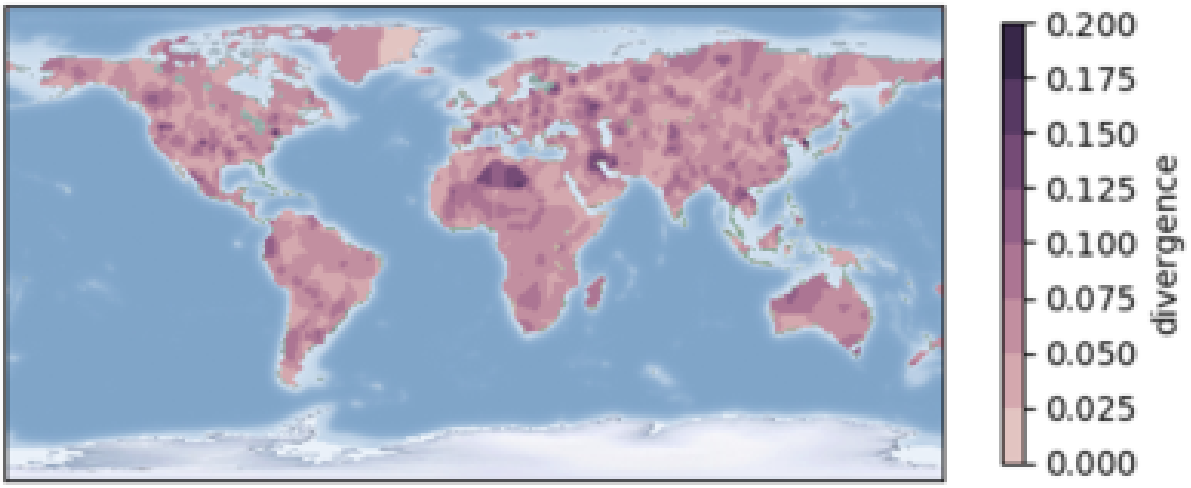

图 11: 评估气候变化实验结果

6 总结与展望

该论文提出了一类新的差异，可以比较基于最优损失的决策任务的两种概率分布。该论文作者证明，与各种基准上的竞争性基准相比，所提出的方法具有更好的测试能力。该论文提出的方法不仅思维巧妙，具有特殊的经验意义，它允许用户通过决策损失来比较分布时直接指定他们的偏好，这意味着可解释性水平将得到提高。实验证明 h 散度允许我们通过选择合适的动作 A 和 loss 来利用每种数据类型 (如图像、生物、文本) 的归纳偏差，从而提高测试能力证明了 h 散度的有效性，即决定两组样本是否来自相同的分布。