UMAP：原理、公式、应用与实战

2025 年 9 月 17 日

1 引言

UMAP（Uniform Manifold Approximation and Projection）基于流形学习与拓扑数据分析，是一种常用的非线性降维算法。它在原空间构建加权邻接图，保留局部连通性，再优化低维嵌入以还原这些邻域关系，生成适合探索分析与聚类诊断的可视化布局。

2 原理与公式

2.1 邻接图构建

对每个样本 $\mathbf{x}_i$，UMAP 选取 k 个最近邻，并使用平滑指数核计算边权：

$$\mu_{ij} = \exp\left(-\frac{\max(0, d(\mathbf{x}_i, \mathbf{x}_j) - \rho_i)}{\sigma_i}\right), \tag{1}$$

其中 d 为距离度量，ρ_i 确保至少存在距离为零的邻居，σ_i 用于归一化局部连通性。将有向边对称化得到模糊拓扑表示：

$$\mathbf{W} = \mu + \mu^\top - \mu \odot \mu^\top. \tag{2}$$

2.2 低维优化

UMAP 通过最小化高、低维模糊集合之间的交叉熵，学习低维嵌入 $\mathbf{y}_i$。嵌入空间的连接强度采用可微曲线：

$$\nu_{ij} = \frac{1}{1 + a\|\mathbf{y}_i - \mathbf{y}_j\|_2^{2b}}, \tag{3}$$

参数 a, b 根据距离分布拟合。损失函数为

$$C = \sum_{(i,j)} \left[w_{ij} \log \frac{w_{ij}}{\nu_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - \nu_{ij}} \right], \tag{4}$$

采用随机梯度下降对采样边进行优化。

2.3 超参数与实践要点

关键参数包括邻居数$_neighbors$ min_dist $spectral\ andom\ PCA\ UMAP$

3 应用与技巧

- **单细胞数据分析**：识别基因表达中的多样细胞类型与稀有群体。
- **文本与嵌入评估**：在语句或文档嵌入上观察语义簇是否清晰。
- **异常诊断**：结合时间或元数据信息，突出异常点或过渡态。

- **实用建议**：先做特征标准化，尝试多组$_neighbors/min_dist$ $t-SNE\ PCA$

4 Python 实战

脚本 `gen_t_umap_figures.py` 对合成数据标准化后，在不同邻居数下运行 UMAP，并计算衡量邻域保持度的 trustworthiness 指标，输出嵌入图与评分曲线。

Listing 1: 脚本 gen$_{tu}map_figures.py$

```python
import umap
from sklearn.manifold import trustworthiness

neighbors_list = [10, 30, 50]
embeddings = {}
trust_scores = []
for n in neighbors_list:
    reducer = umap.UMAP(n_neighbors=n, min_dist=0.1, metric="euclidean"
        ,
                        init="spectral", random_state=42)
    embedding = reducer.fit_transform(points)
    embeddings[n] = embedding
    trust_scores.append(trustworthiness(points, embedding, n_neighbors
        =15))
```

5 实验结果

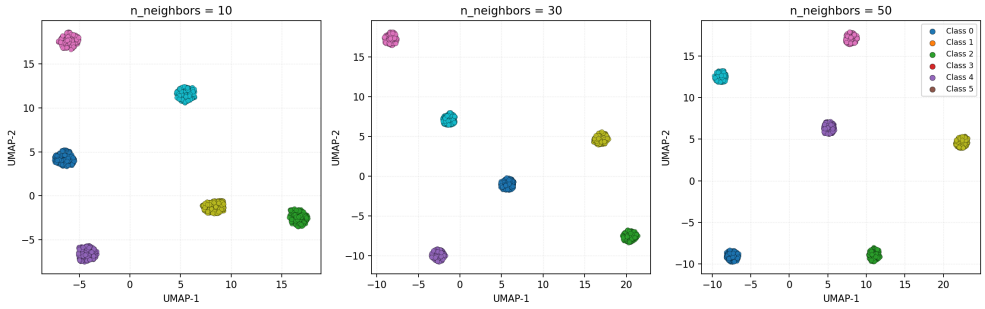

图 1: 不同邻居数下的 UMAP 嵌入，可根据类别着色

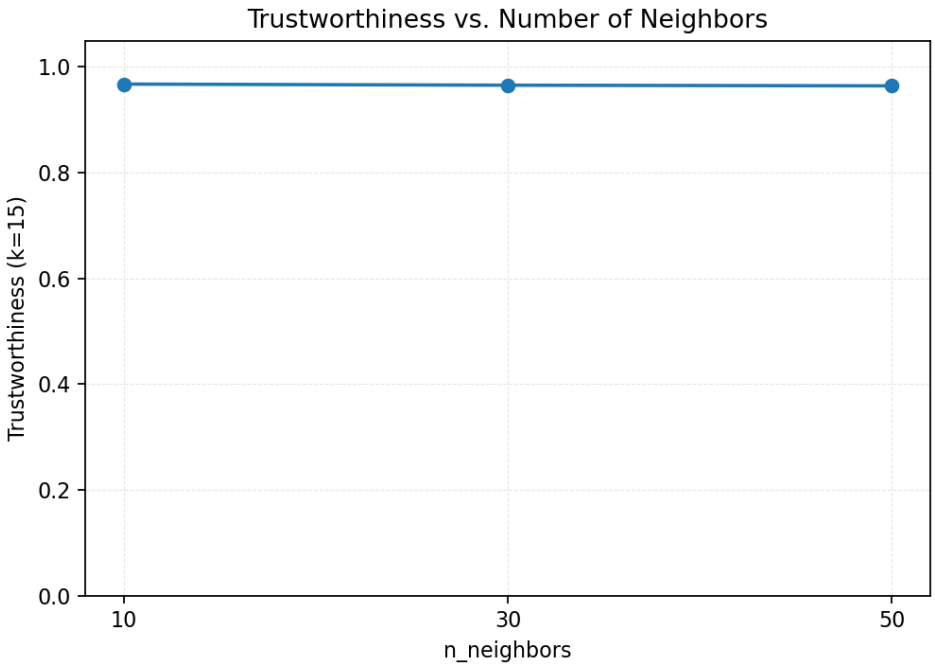

图 2: trustworthiness 指标随邻居数量变化的曲线

6 总结

UMAP 通过模糊邻域建模与交叉熵优化兼顾局部细节与全局布局。合理调整邻居数、最小距离和距离度量，可在探索分析中获得稳定、易解读的嵌入图。示例展示了如何比较不同参数下的结果并利用 trustworthiness 曲线评估嵌入质量。