

第四次上机实验报告

北京大学化学与分子工程学院 李梓焰 2101110396

摘要 本实验完成了以下工作：其一是对scikit-learn自带的手写数字测试集分别进行PCA与t-SNE降维分析，并比较两种降维方法的降维效果；其二是用K-Means聚类法，对不降维与经PCA降维后的scikit-learn自带的手写数字测试集进行聚类，比较降维至不同维度数对最终聚类结果的影响，结果表明，PCA降维后的最低维数至少要16，才能保证聚类结果与不降维时基本一致；其三是在结合不同的距离度量，采用t-SNE降维的前提下，用K-Means聚类法将分子聚类，并考察不同类别数K下聚类的表现，统计得到的distortion与轮廓系数随类别数K的变化表明，当K取3时，聚类的表现最佳

关键词 降维分析 手写数字识别 聚类分析 分子指纹

1 引言（略）

2 实验部分

2.1 仪器

2.1.1 硬件

Surface Pro（第5代，处理器参数：Intel® Core™ i5-7300U CPU @ 2.60GHz, 2.71 GHz, 2个内核，4个逻辑处理器；内存容量：8.00 GB）；自配台式电脑（处理器参数：Intel® Core™ i7-11700K CPU @ 3.60GHz, 3.60 GHz, 8个内核，16个逻辑处理器；内存容量：32.00 GB）

2.1.2 软件

操作系统：Surface Pro预装Windows 10家庭版，版本21H1；自配台式电脑预装Windows 10专业版，版本20H2

开发环境：Surface Pro使用的是Visual Studio 2019 Community, 64位Anaconda 3（版本号2021.05，含64位Python 3.8.8、Conda 4.10.1、NumPy 1.20.1、Pandas 1.2.4、SciPy 1.6.2、Scikit-learn 0.24.1、Matplotlib 3.3.4）；自配台式电脑使用的是Visual Studio 2022 Community, 64位Anaconda 3（版本号2021.11，含64位Python 3.9.7、Conda 4.10.3、NumPy 1.21.4、Pandas 1.3.4、SciPy 1.7.1、Scikit-learn 0.24.2、Matplotlib 3.4.3）

2.1.3 测试数据

按照上机实验要求，测试数据分为两部分，第一部分为scikit-learn自带的手写数字测试集，从scikit-learn库调用即可使用；第二部分为ZINC数据集中随机取出的10000个分子的SMILES与ECFP4指纹

2.2 实验过程

2.2.1 手写数字测试集的数据特征及可视化

从scikit-learn库中调用手写数字测试集，将数据集每个图像样本的64个像素值随机投影到正交的两个维度上，用不同颜色的散点代表不同数字，绘制出投影后测试集的分布情形。为更好地观察数据，输出的投影图不显示测试集中数字的模样，但在程序中，仍预留显示测试集部分数字的功能，以备不时之需。

2.2.2 手写数字测试集的PCA与t-SNE降维分析

分别用PCA和t-SNE的方法将每个图像样本的像素数据降维至2维，并采用与上一节相似的方法显示降维结果。此处对数据不做Min-Max归一化处理，但预留了Min-Max归一化选项以供选择。

2.2.3 手写数字测试集的K-Means聚类分析

采用K-Means聚类法，对未降维的手写数字测试集，以及降维至1、2、4、8、16、32的手写数字测试集，分别进行聚类，并用多个统计指标，如轮廓系数、homogeneity、completeness、V-measure等，分析聚类效果，以及聚类标签和实际标签的关联程度。

2.2.4 化合物的K-Means聚类分析

采用K-Means聚类法，将ZINC数据集中随机取出的10000个分子，分别划分为1至10类，划分时，采用余弦距离、Dice距离、Tanimoto距离（在分子指纹相似性比较中等同于Jaccard距离）衡量两分子的偏离程度。随后，根据聚类结果，计算欧几里得度量下的distortion（实际为各聚类的inertia，定义为各数据点与聚类中心的平方和），并作出distortion随类别数K变化的折线图，以便结合肘部原则，给出最佳K值的参考值；同时，按聚类结果，计算欧几里得度量与余弦度量下的轮廓系数，据此选取最佳K值。最后，用最佳K值的聚类结果，得出各聚类代表性化合物，并输出化合物的SMILES。

3 实验数据与结果分析

3.1 手写数字测试集的数据特征及可视化

按照2.2.1节的操作，笔者得到了手写数字测试集随机投影后的分布图，如图1所示，可以看到数字沿与坐标轴平行的方向分散排列，且分布区域非常广泛，且不同标签的数字的分布区域相互交织在一起，无法在保证划分后两两不重叠的前提下区分。这意味着，要对手写数字采取适当的降维分析，聚类分析才能获得较为正确的结果。

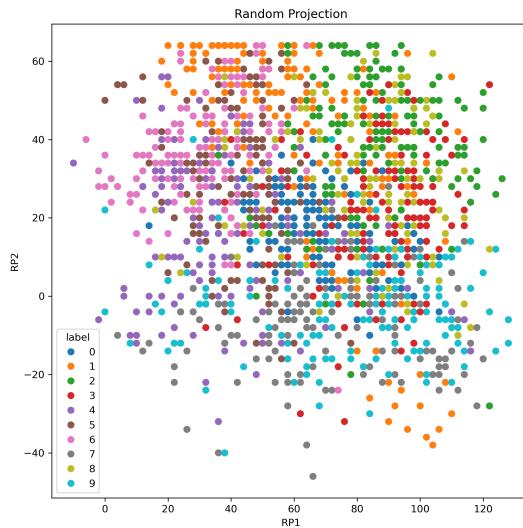


图1 手写数字测试集随机投影后的分布图

3.2 手写数字测试集的PCA与t-SNE降维分析

按照2.2.2节的操作，笔者首先采用PCA方法，将手写数字测试集数据降至二维，其分布图如图2左图所示。与图1相比，图2左图围着中心偏上的空腔环绕分布，且结合坐标轴比例尺可知，PCA降维后的数据分布范围较随机分布时更窄；更重要的是，数字0、4、6的重叠范围显著减小，甚至可以在重叠数据较少的情况下进行初步划分。上述结果表明，采用PCA方法降维后，数据的区分度有所上升。

然而，笔者也注意到，即使经过PCA方法降维，除0、4、6外的数字的分布区域仍然交织在一起，难以区分，为了更好地进行聚类，有必要采用更合适的降维方法，以进一步缩小数字的分布范围，同时增大不同数字的距离。因此，笔者采用t-SNE方法，将手写数字测试集数据降至二维，其分布图如图2右图所示。比照图1与图2左图可知，经t-SNE降维的数据，分布范围显著缩小，且不同数字聚类的距离显著增大，以至聚类之间基本无重叠数据，这意味着t-SNE能使同一标签的数据更加集中，而使不同标签的数据相互远离，为之后的聚类分析带来方便。

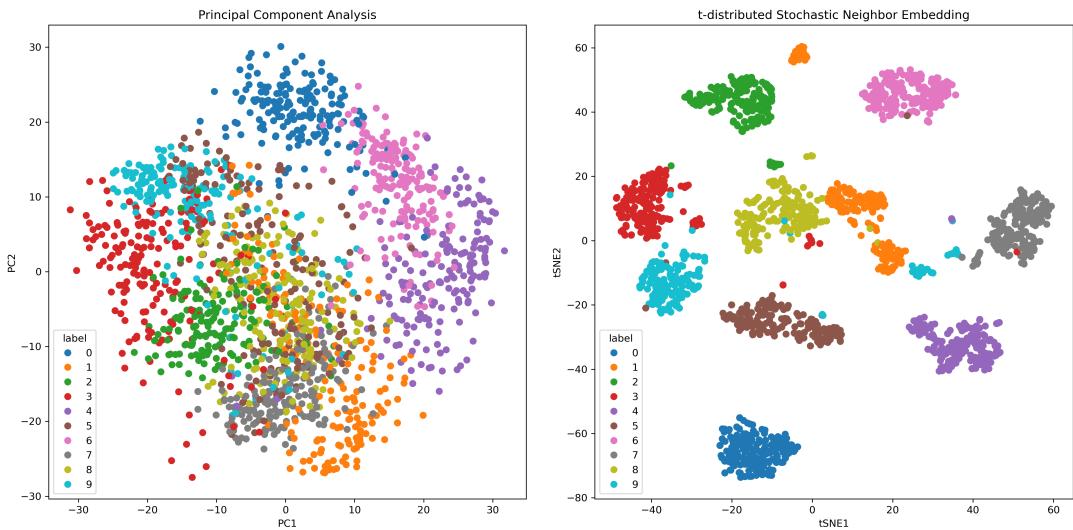


图2 降维分析手写数字测试集的分布图。左：PCA；右：t-SNE

3.3 手写数字测试集的K-Means聚类分析

按照2.2.3节的操作，笔者采用K-Means聚类法，对未降维的手写数字测试集（总计64维）进行聚类分析，结果如图3所示。比照图2右图后发现，尽管K-Means聚类时给出的标签不同于数字样本附带的标签，但两者之间基本上能建立起一一对应关系：K-Means的标签0-9，分别对应于原数据中的1与9、6、5、2、3、0、7、1与8、4、9（含一部分5与8）。

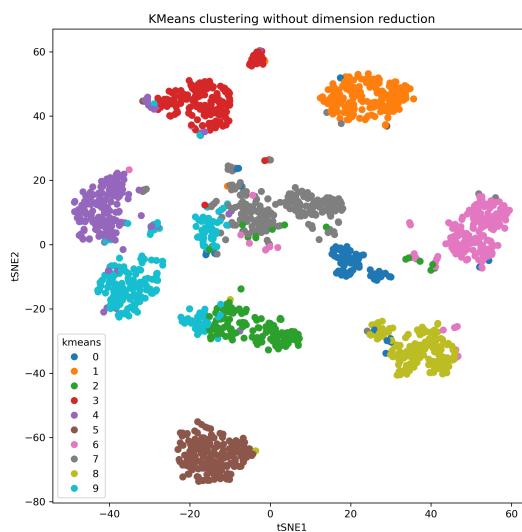


图3 未降维时，手写数字测试集的K-Means聚类结果

另一方面，统计得到的混淆矩阵（表1），也验证了根据聚类分布图与实际分布图对比得出的结论。

表1 未降维时，手写数字测试集的K-Means聚类混淆矩阵

实际数字\K-Means标签	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	177	0	0	1	0
1	55	2	1	24	1	0	0	99	0	0
2	2	0	0	148	13	1	3	8	0	2
3	0	0	2	0	156	0	6	7	0	12
4	5	0	0	0	0	0	10	2	164	0
5	0	1	136	0	2	0	0	0	2	41
6	1	177	0	0	0	1	0	2	0	0
7	2	0	5	0	0	0	170	2	0	0
8	6	2	7	3	2	0	3	99	0	52
9	20	0	7	0	6	0	8	0	0	139

接下来，为探究PCA降维对K-Means聚类结果的影响，笔者首先用PCA方法，将手写数字测试集分别压缩至1、2、4、8、16、32维，然后对压缩维度的数据进行K-Means聚类，得到的聚类分布图如图4所示。将图4与图3一并比较，可以发现：（1）无论降维至多少维度数，最后得到的K-Means聚类标签并不直接等于原数据点对应的数字；（2）用PCA方法降至16维及32维时，仅考虑K-Means聚类标签与数字的对应关系，而不考虑聚类标签的具体值时，除极少数数据点归类略有变化外，其余数据点的对应关系与不降维时基本一致；（3）继续降维至8维时数字3、9对应的聚类结果发生了巨大改变，由划分为两类变成基本合并为一类，表明用PCA将数据维度降低至8维时，部分数字形状信息出现丢失；（4）进一步降维至4维或以下时，t-SNE划分出的每个团簇中，开始出现不止一组的k-Means聚类标签，且每一种聚类标签都包含数量不小的数据点，表明降维的目标维度太低，会导致数字形状信息的大量丢失，从而导致K-Means聚类结果严重偏离实际标签。

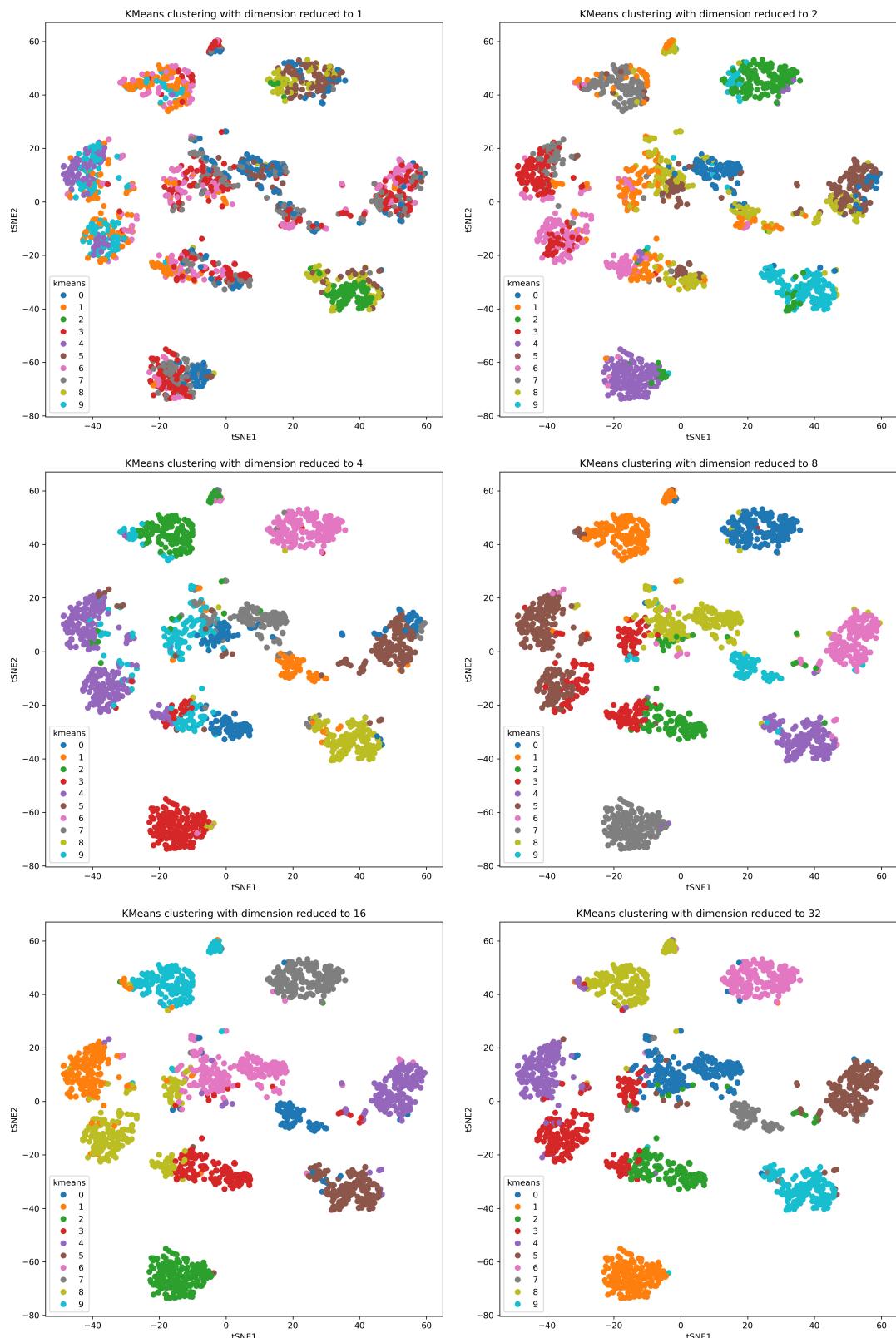


图4 PCA降维后，手写数字测试集的K-Means聚类结果，其中降维数依次为1、2、4、8、16、32

上述结论还得到了轮廓系数的佐证，该系数计算的是数据点到归属聚类的中心与其他聚类的中心的距离差均值，反映了聚类后数据的集中程度。上述聚类结果的轮廓系数如图5左图所示，可以看到不降维时，聚类的轮廓系数为0.50左右，降至32维与16维时，轮廓系数依然在0.50。转折点出现在目标维数达8维时，此时轮廓系数开始下降至0.45，之后随维数的降低，轮廓系数迅速下降，到目标维数达2维与1维时，轮廓系数已经跌至0.1与-0.1左右，表明此时的聚类标签与实际的数字已无任何关联。

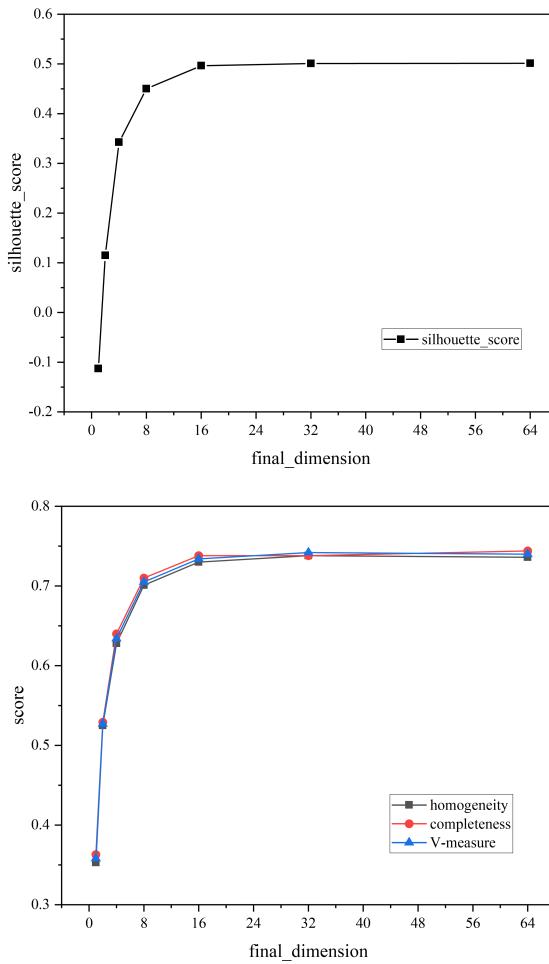


图5 手写数字测试集的K-Means聚类表现指标随PCA降维后的维数的变化（用Origin绘制）。左：轮廓系数；右：其他聚类指标，包括homogeneity、completeness、V-measure

除了轮廓系数，其他聚类指标，如homogeneity、completeness、V-measure，也反映出类似的趋势（见图5右图）：未降维时，或降维至32维与16维时，三种指标均在0.72至0.73上下略有涨落；降至8维时，三种指标开始下滑至0.70；继续降低原数据维度数，三种指标开始迅速下滑，至目标维数为2维与1维时，三种指标低至0.53与0.35左右，较未降维时的聚类结果有巨大下滑。综合以上分析，笔者认为，PCA降维后的最低维数至少要16，才能保证聚类结果与不降维时基本一致。

3.4 化合物的K-Means聚类分析

余弦距离、Dice距离、Tanimoto距离（Jaccard距离）可用于分子指纹相似性的度量，它们的定义为：

$$\begin{aligned} \text{Dis}_{\cos}(X, Y) &= 1 - \text{Sim}_{\cos}(X, Y) = 1 - \frac{X \cdot Y}{\|X\|_2 \|Y\|_2} \\ \text{Dis}_{\text{dice}}(X, Y) &= 1 - \text{Sim}_{\text{dice}}(X, Y) = 1 - \frac{2X \cdot Y}{\|X\|_1 + \|Y\|_1} = \frac{C_{TF} + C_{FT}}{2C_{TT} + C_{TF} + C_{FT}} \\ \text{Dis}_{\text{Tanimoto}}(X, Y) &= 1 - \text{Sim}_{\text{Tanimoto}}(X, Y) = 1 - \frac{X \cdot Y}{\|X\|_2^2 + \|Y\|_2^2 - X \cdot Y} = \frac{C_{TF} + C_{FT}}{C_{TT} + C_{TF} + C_{FT}} \end{aligned}$$

其中

$$X = (X_1, X_2, \dots, X_N) \quad \|X\|_1 = \sum_{i=1}^N |X_i| \quad \|X\|_2 = \sqrt{\sum_{i=1}^N X_i^2}$$

按照2.2.4节的操作，笔者分别采用上述距离度量，利用t-SNE将分子指纹压缩至2维，然后调用K-Means聚类方法，将压缩维度后的数据分别聚类为1-10个类别，并统计聚类结果的distortion与轮廓系数（以欧几里得度量和余弦度量表示），结果如图6、图7所示。

首先观察图6，可以发现无论采用何种距离度量，随聚类数K的增加，distortion均先迅速下降，待K=3后distortion下降趋势显著放缓，根据“肘部原则”可知，K=3为最佳聚类个数。

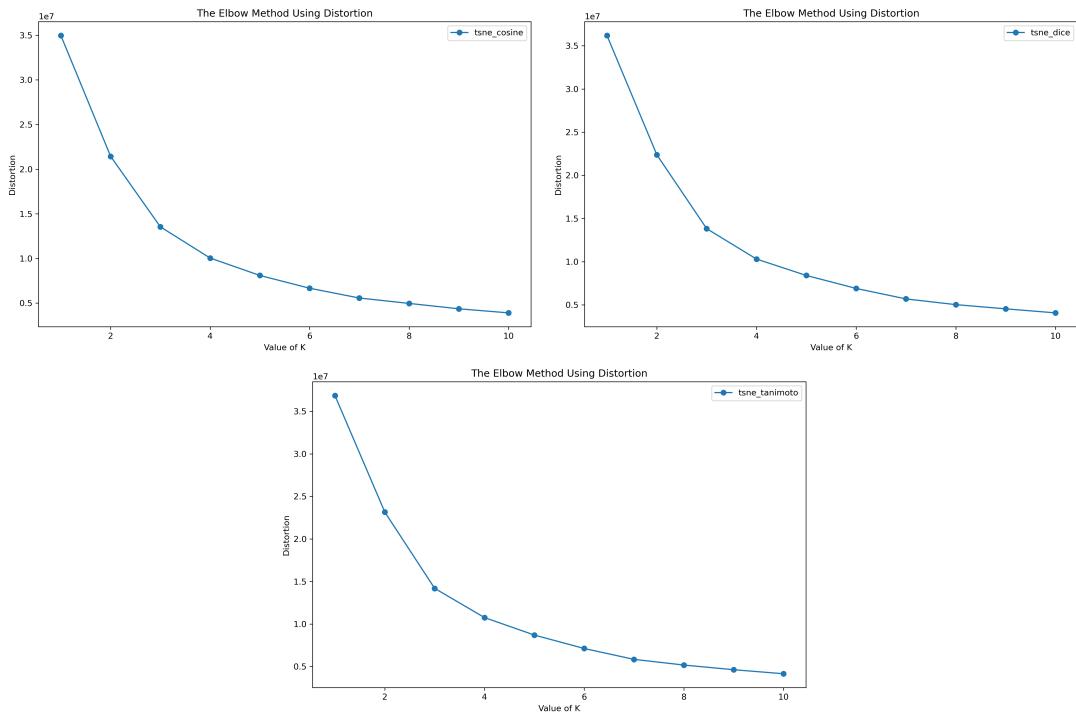


图6 结合不同的距离度量，采用t-SNE降至二维后，K-Means聚类的distortion随类别数K的变化结果，三种距离度量依次为余弦距离、Dice距离、Tanimoto距离

接下来是两种度量下的轮廓系数，如图7所示，考虑到轮廓系数只有在聚类数不少于2时有定义，这里只给出 $K \geq 2$ 的数据。相较于图6的单调下降，图7的增减变化较大，整体呈双峰状分布，其中第一个峰出现在 $K=3$ 处，且为考察区域的最大值；第二个峰除左图的tsne_cosine出现于 $K=6$ 外，其余图线均出现于 $K=7$ ，且该峰对应的轮廓系数小于 $K=3$ 处的轮廓系数。综上所述，笔者得出结论：K=3为最佳聚类个数。

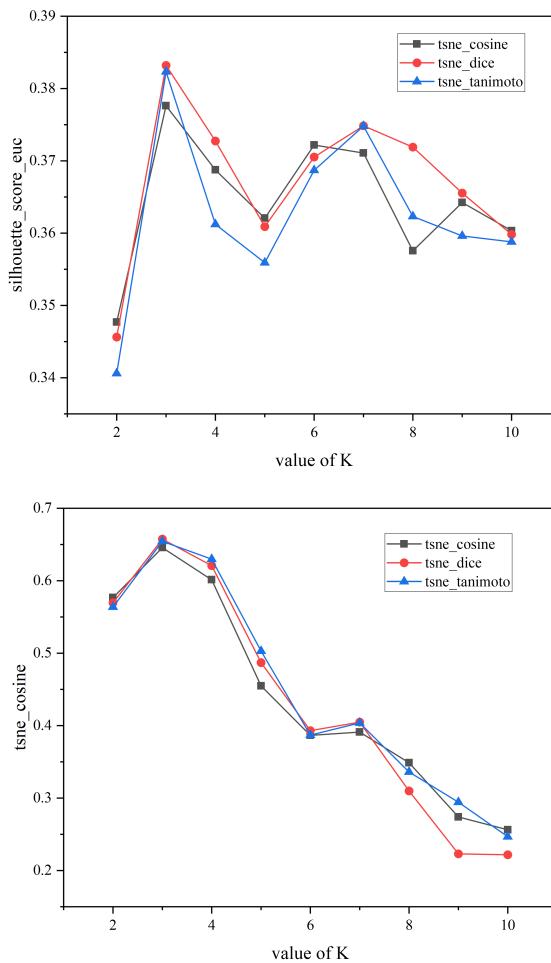


图7 结合不同的距离度量，采用t-SNE降至二维后，K-Means聚类的轮廓系数随类别数K的变化结果（均用Origin绘制）。左：用欧几里得度量计算轮廓系数；右：用余弦度量计算轮廓系数

根据这一结论，笔者给出了在两种不同的度量下，最接近聚类中心的分子的SMILES表达式，结果如表2所示。值得注意的是，无论t-SNE降维时采用何种度量方式，最后得到的分子均完全一致。

表2 欧几里得度量与余弦度量下，最接近聚类中心的分子的SMILES表达式

聚类标签	分子与聚类中心距离的度量	SMILES表达式		
		tsne_cosine	tsne_dice	tsne_tanimoto
1	欧几里得度量	C#CC[N+]1 = CN(c2ccc(C)cc2)Cc2cc(C)ccc21		
2	欧几里得度量	CC(C)C(= O)Nc1cccc(C(= O)Nc2cc([N+](= O)[O-])ccc2F)c1		
3	欧几里得度量	c1ccc(-c2csc(NCc3ccco3)n2)cc1		
1	余弦度量	CCOC1(O)C(= O)c2cccc2OC1(OCC)c1ccccc1		
2	余弦度量	Cc1ccccc1CSCC(= O)NN = Cc1ccc(C#N)cc1		
3	余弦度量	Cc1nc2sc3c(c2c(= O)n1CC(= O)NCc1ccc(Cl)cc1)CCC(C)C3		

3.5 实验结论

- (1) 不同的降维方法将极大影响手写数字测试集的聚类效果，其中，PCA方法优于随机投影，而t-SNE方法优于PCA方法；
- (2) 对手写数字测试集而言，PCA降维后的最低维数至少要16，才能保证K-Means聚类结果与不降维时基本一致；
- (3) 统计得到的distortion与轮廓系数随类别数K的变化表明，当K取3时，对ZINC化合物库的部分分子聚类的表现最佳。