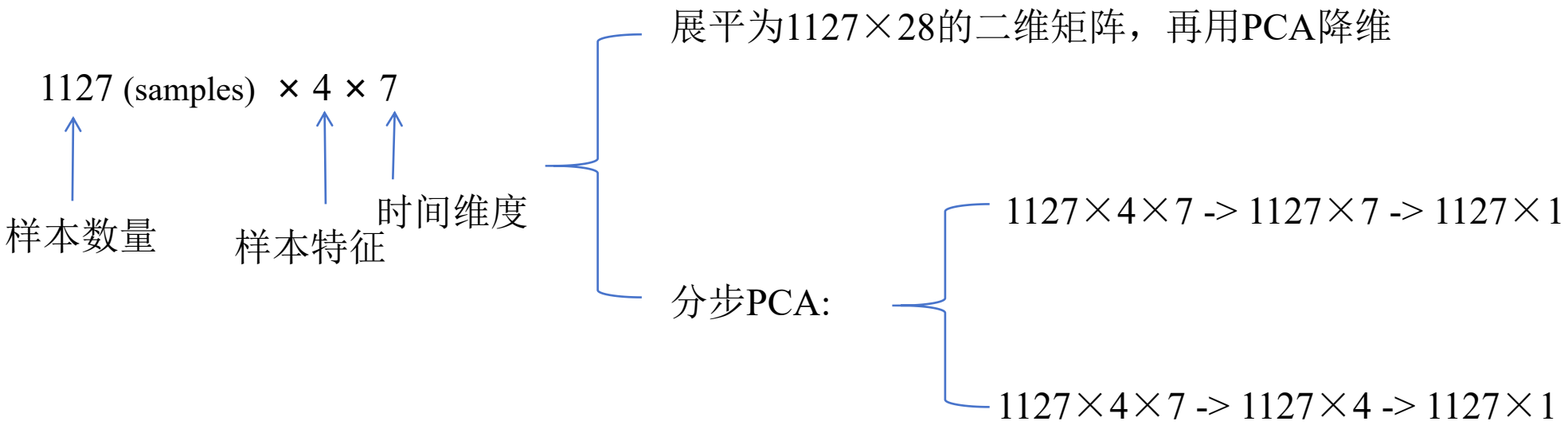
同济大学 TONGJI UNIVERSITY

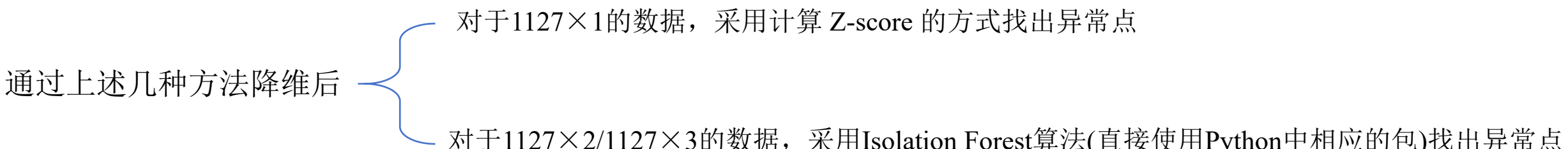
面向磁浮轨道异常检测的大数据分析框架研究

小组汇报

2024-7-18　　By 刘震

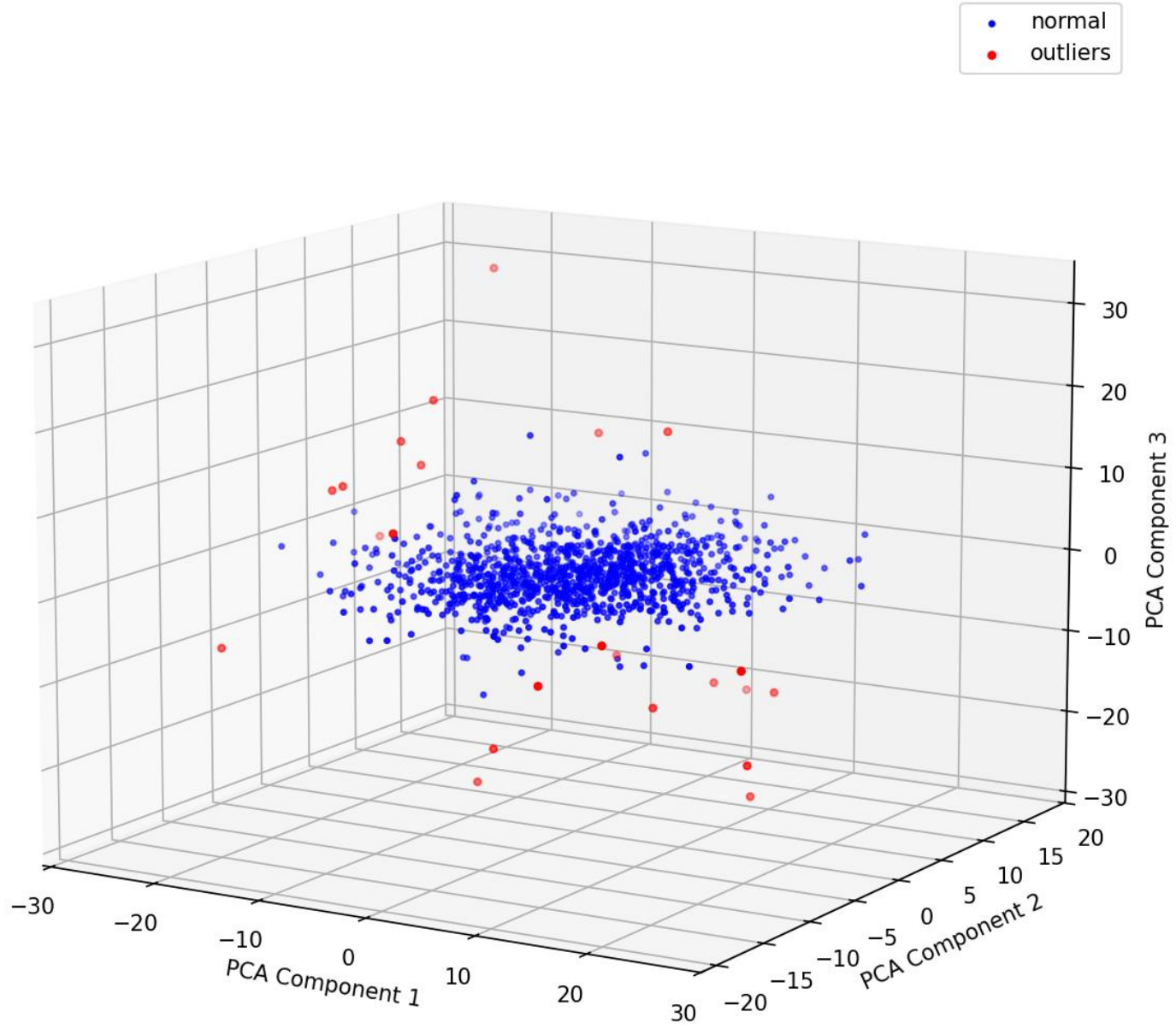
Part I: PCA (Principal Component Analysis) 数据处理





设定2%-3%的数据为异常值

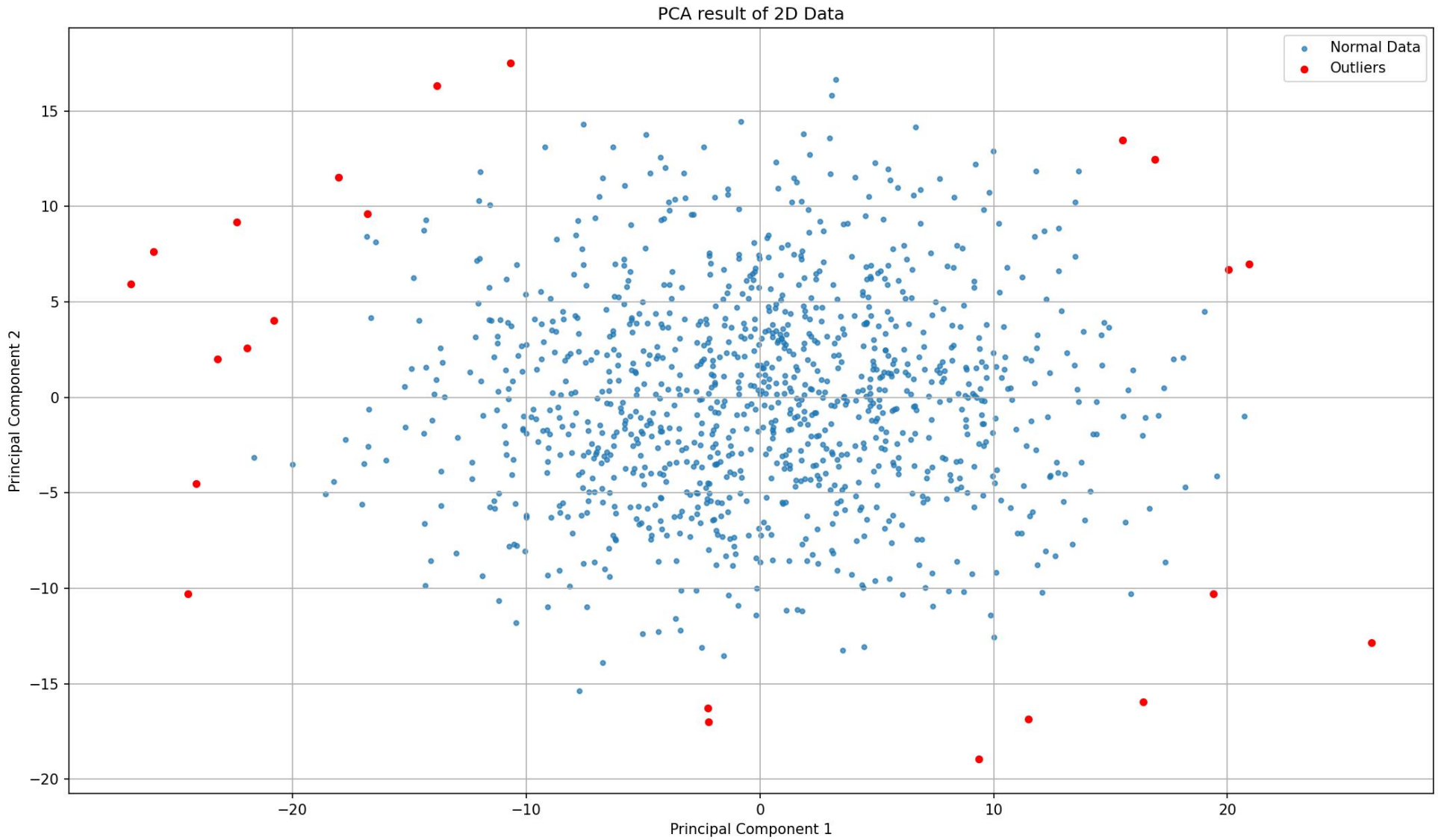
取前三个主成分，可视化结果：（其中蓝色的点为正常点，红色点为异常点）



标红异常点索引

2, 6, 7, 8, 11, 38, 41, 44, 47, 73, 74, 183, 729, 923, 1034, 1036, 1037, 1114, 1120, 1121, 1123, 1124, 1126

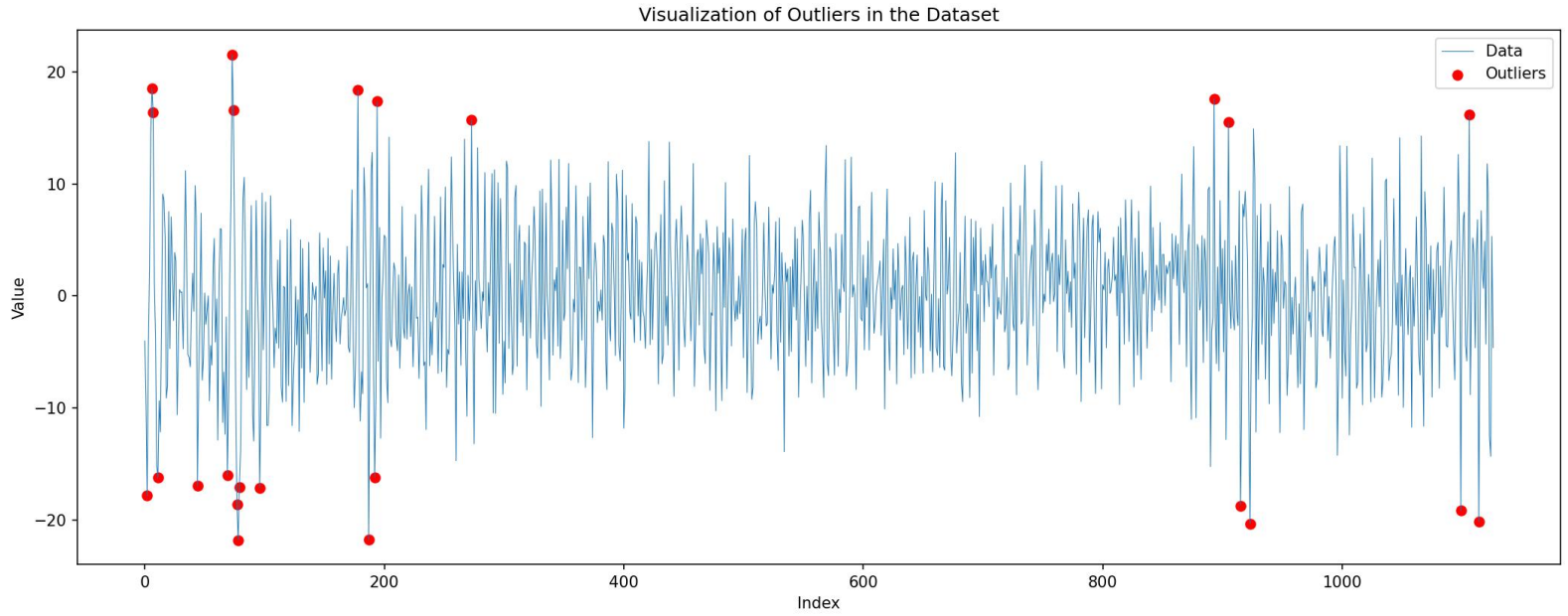
取前两个主成分，可视化结果：（其中蓝色的点为正常点，红色点为异常点）



标红异常点索引：

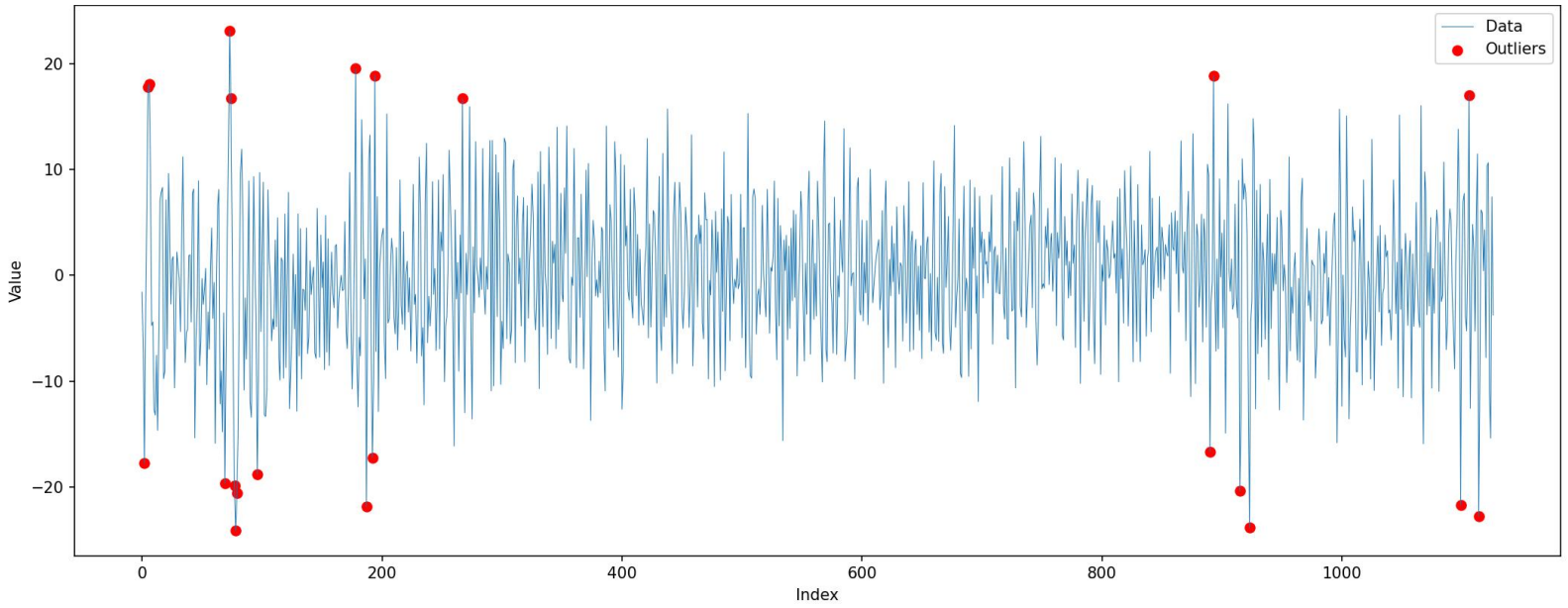
47, 61, 69, 73, 74,
77, 78, 79, 178,
183, 187, 264, 273,
369, 729, 893, 923,
926, 1000, 1034,
1099, 1114, 1124

取第一个主成分，可视化结果：（其中蓝色的点为正常点，红色点为异常点）



标红异常点索引：

3, 7, 8, 12, 45, 70, 74, 75, 78, 79, 80, 97, 179, 188, 193, 195, 274, 894, 906, 916, 924, 1100, 1107, 1115



标红异常点索引：

3, 6, 7, 70, 74, 75, 78, 79, 80, 97, 179, 188, 193, 195, 268, 891, 894, 916, 924, 1100, 1107, 1115

PCA：检测结果准确率

检测方式：将14个标签文件中有超过4个文件被标记的点取出来（共33个），记作明显异常点；超过2个文件被标记的点取出来（共179个），记作普通异常点。与降维后检测异常的结果进行比对：

	明显异常点	普通异常点
1维	17（77.2%）	19（86.3%）
2维	8（27.5%）	13（44.8%）
3维	6（26.1%）	22（95.6%）

其中效果较好的有：

降维至1维（分步PCA，再计算Z-score）的方法与结果符合的最好：明显异常点中有17个点符合（86.3%）

降维至3维（展开为1127*28的矩阵后PCA，利用Isolation Forest检测异常点）：普通异常点有22个符合（95.6%）

Part II：UMAP（Uniform Manifold Approximation and Projection）数据处理

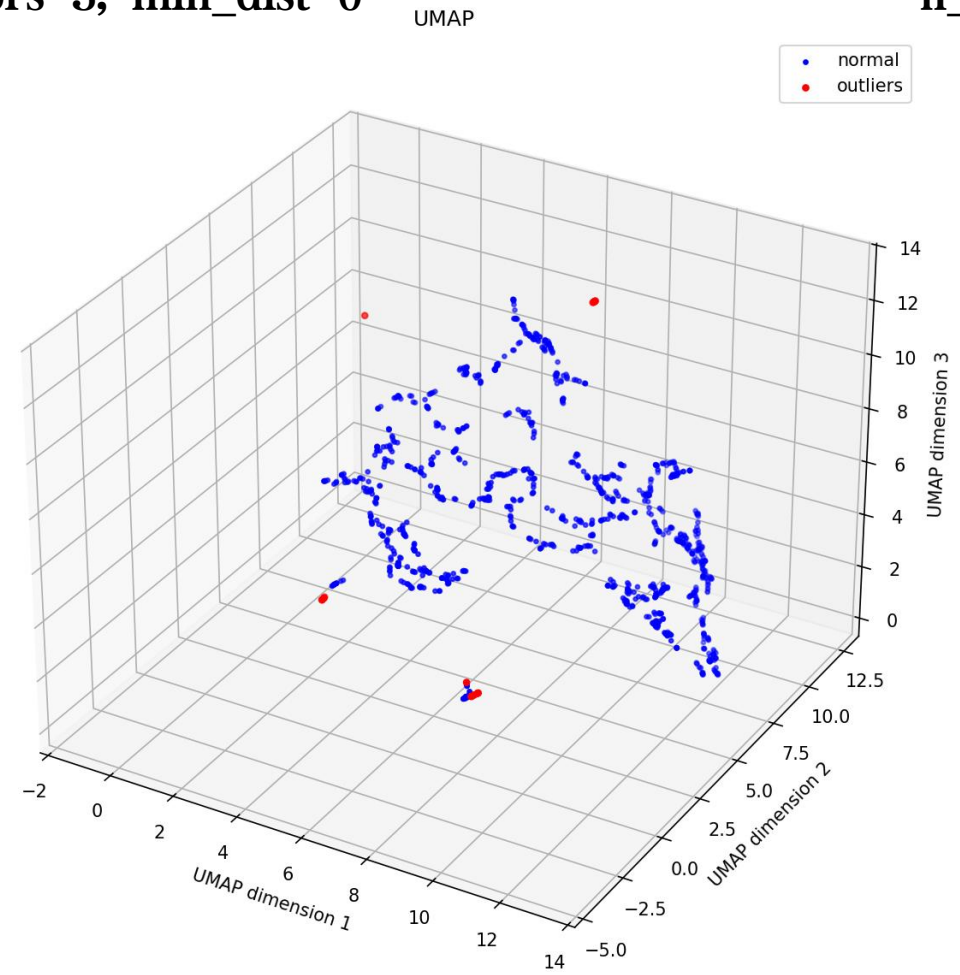
UMAP**降维中有三个重要参数：**n_neighbors, min_dist, metric

分别取n_neighbors=3，15，100;
min_dist=0，0.5;
metric=euclidean(**欧几里得距离**)

进行降维、可视化

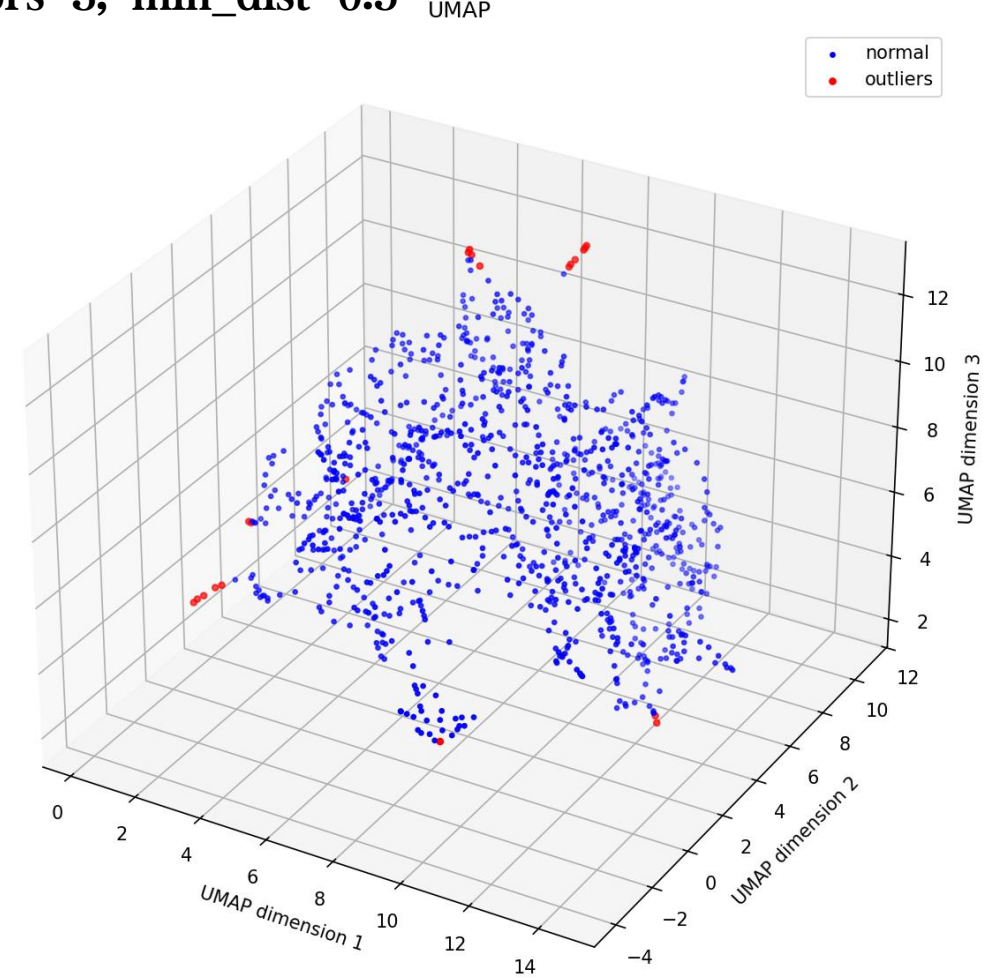
降维至3维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

n_neighbors=3, min_dist=0



9, 48, 64, 70, 73, 78, 79, 80, 243, 258, 436, 445, 512, 528, 730, 844, 918, 924, 926, 1037, 1038, 1074, 1100

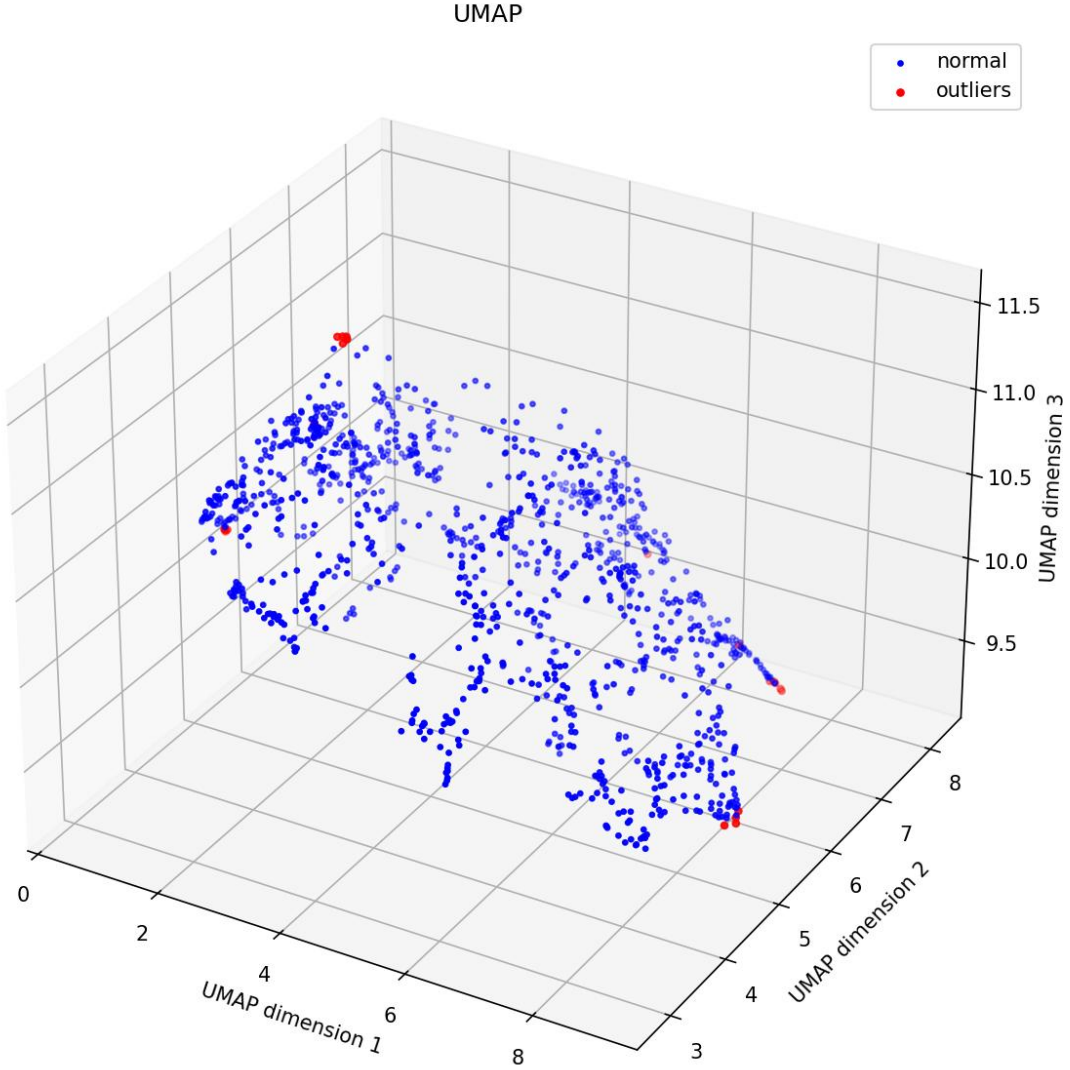
n_neighbors=3, min_dist=0.5



48, 64, 70, 73, 79, 80, 92, 122, 173, 243, 258, 276, 406, 445, 514, 528, 545, 730, 756, 797, 918, 924, 1100

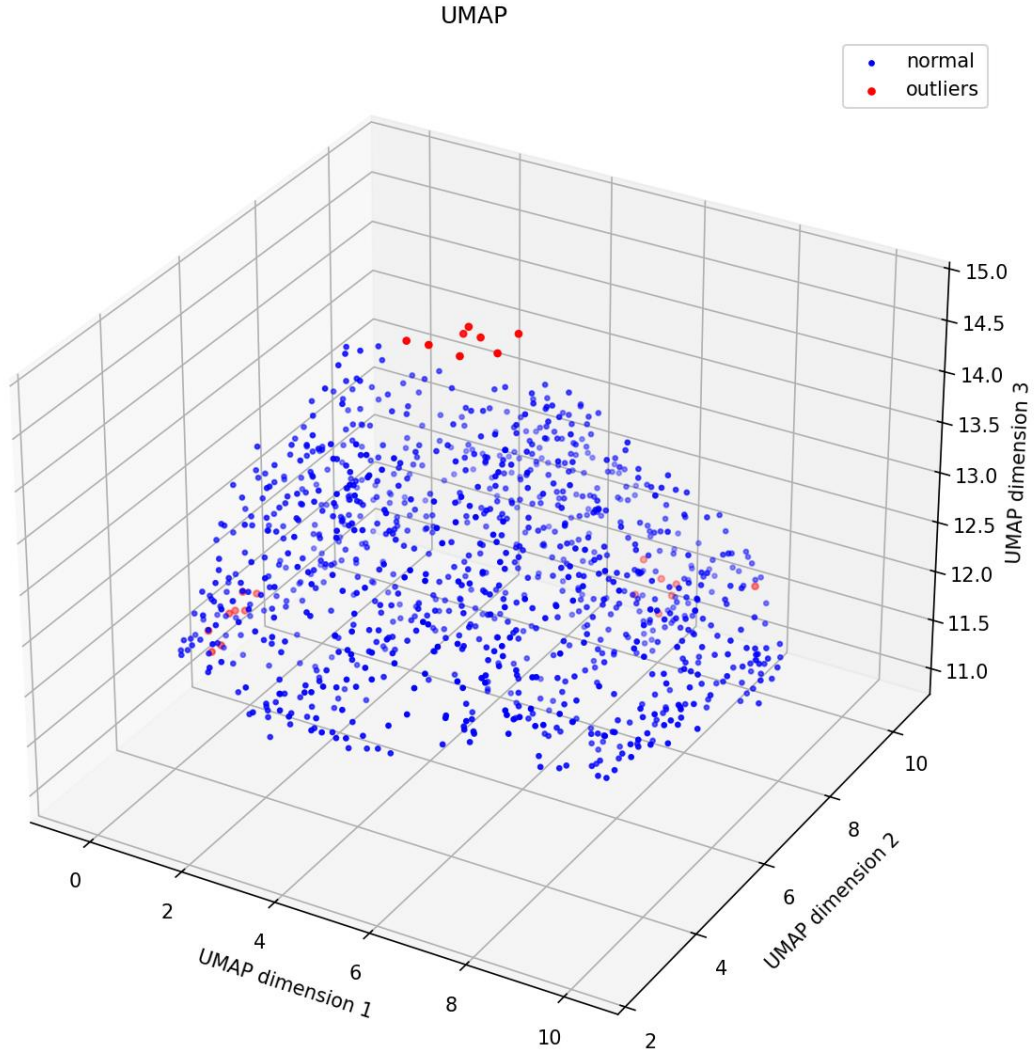
降维至3维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

n_neighbors=15, min_dist=0



79, 80, 111, 179, 195, 257, 274, 340, 372, 404, 422, 661, 861, 924, 927, 1008, 1021, 1022, 1053, 1095, 1100, 1107, 1121

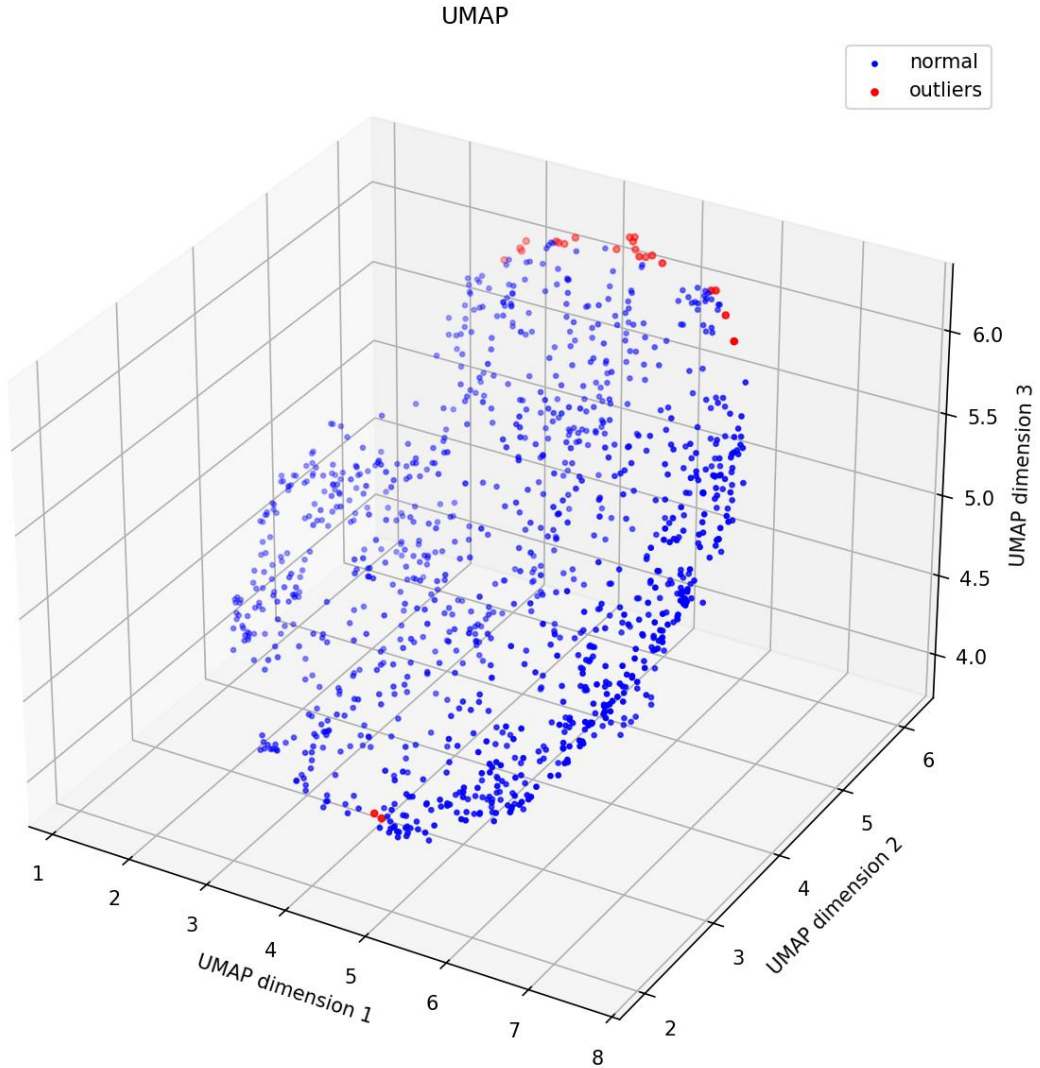
n_neighbors=15, min_dist=0.5



8, 9, 36, 37, 42, 43, 70, 79, 80, 97, 188, 257, 404, 428, 756, 861, 872, 916, 924, 1037, 1038, 1100, 1117

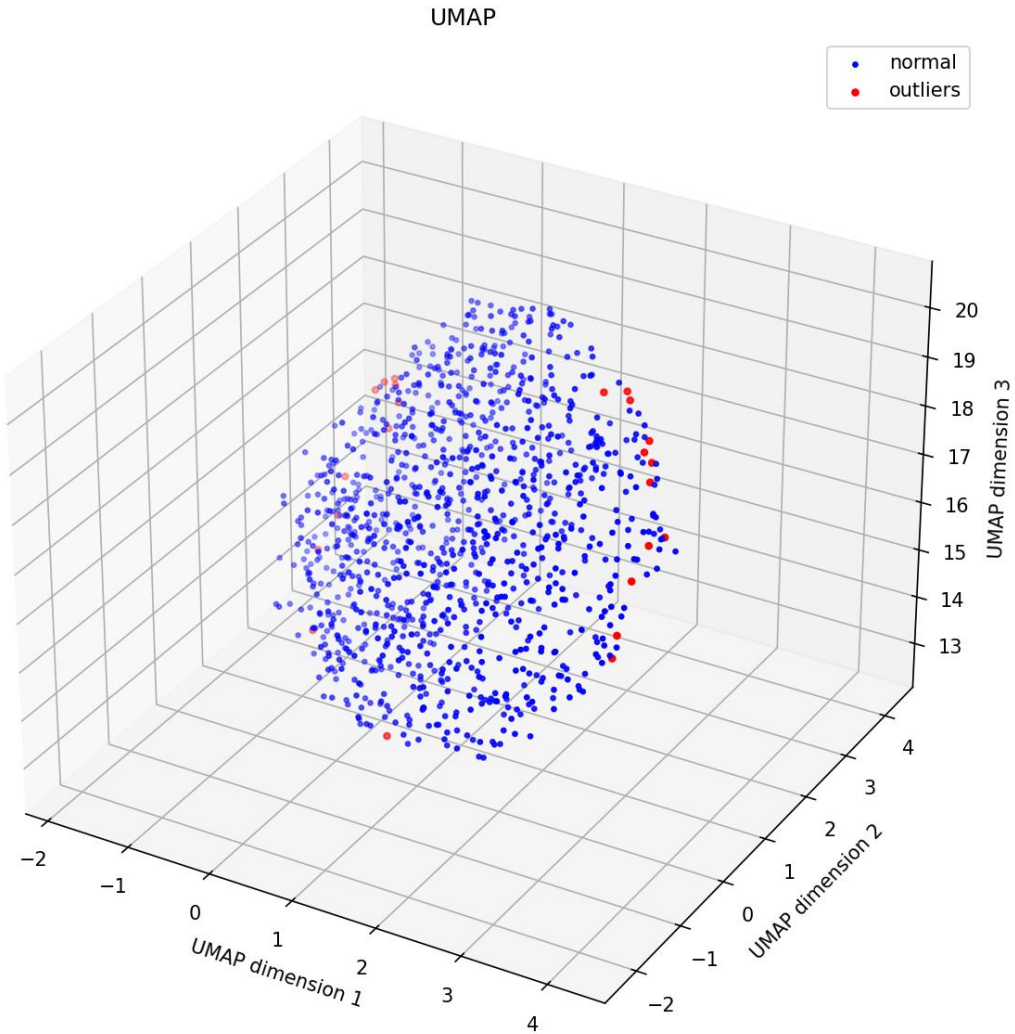
降维至3维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

n_neighbors=100, min_dist=0



26, 260, 264, 265, 353, 372, 392, 404, 655, 734, 776, 798, 806, 861, 872, 876, 909, 920, 1070, 1093, 1096, 1116, 1127

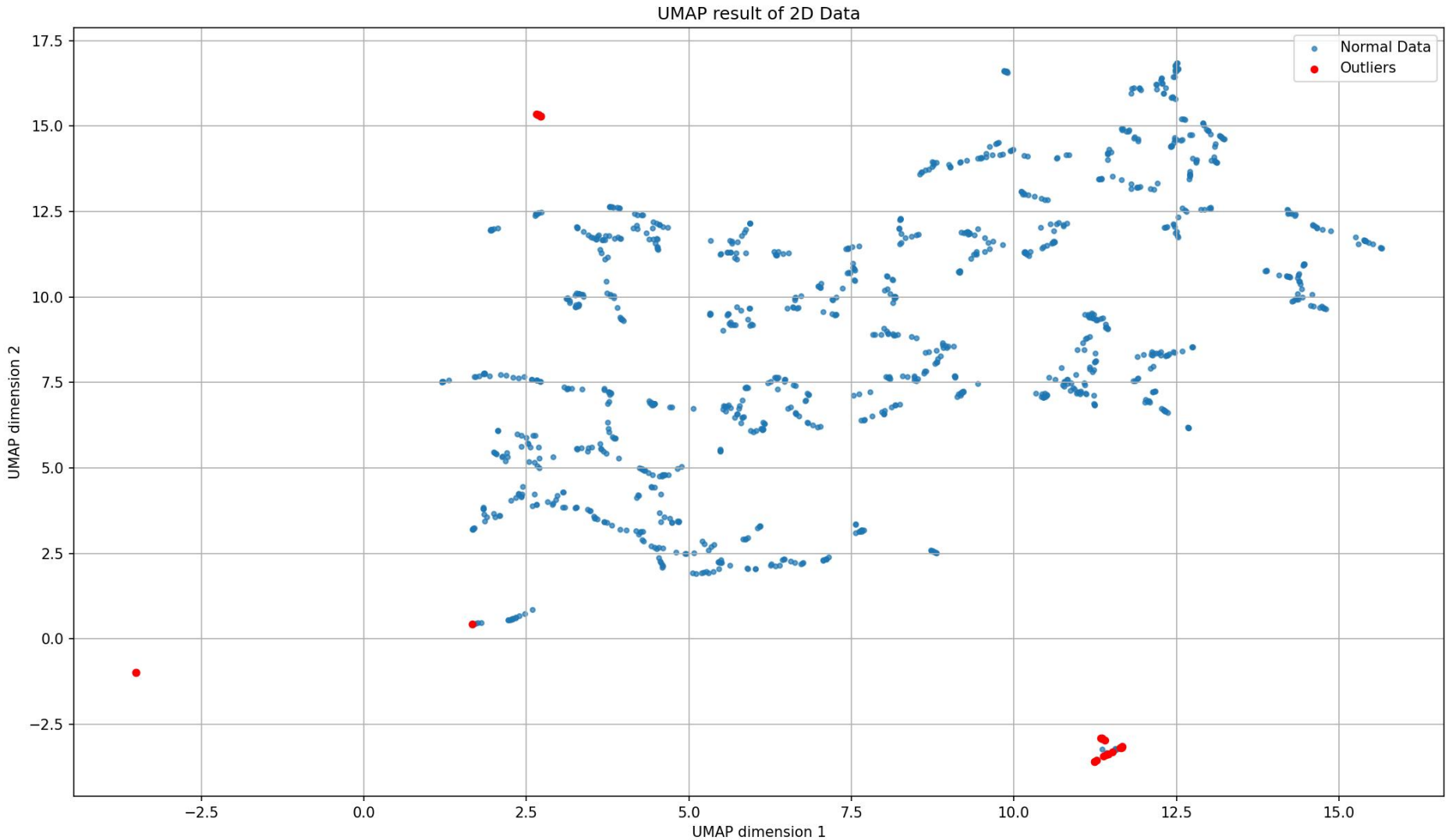
n_neighbors=100, min_dist=0.5



62, 79, 80, 92, 179, 188, 257, 266, 270, 274, 422, 506, 591, 667, 879, 916, 917, 924, 927, 1067, 1100, 1107, 1115

降维至2维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

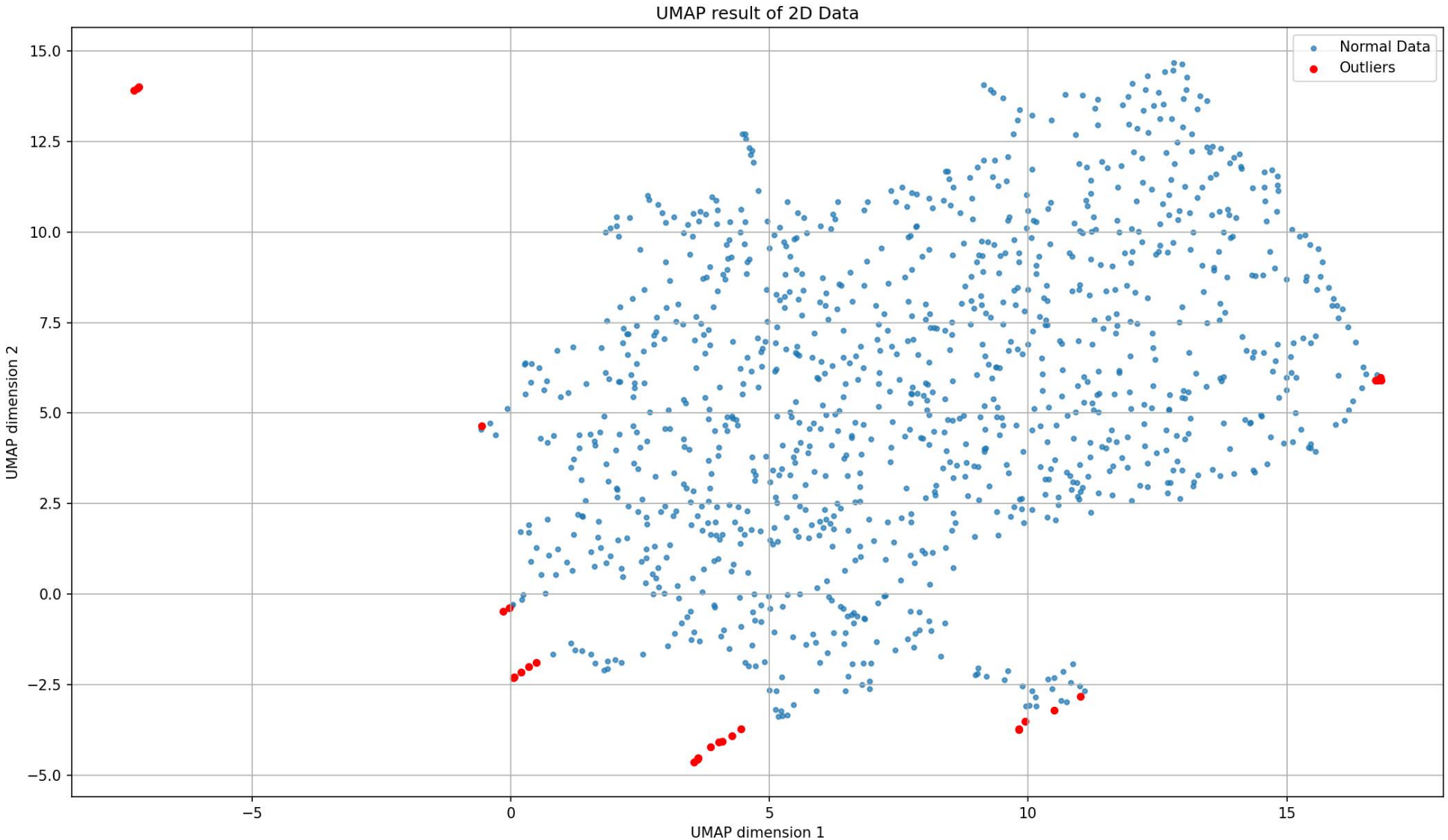
n_neighbors=3, min_dist=0



8, 47, 63, 72, 139, 147, 242, 257, 416, 417, 444, 506, 511, 527, 729, 733, 843, 881, 917, 923, 925, 933, 1031, 1036, 1037, 1073, 1083, 1092, 1115

降维至2维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

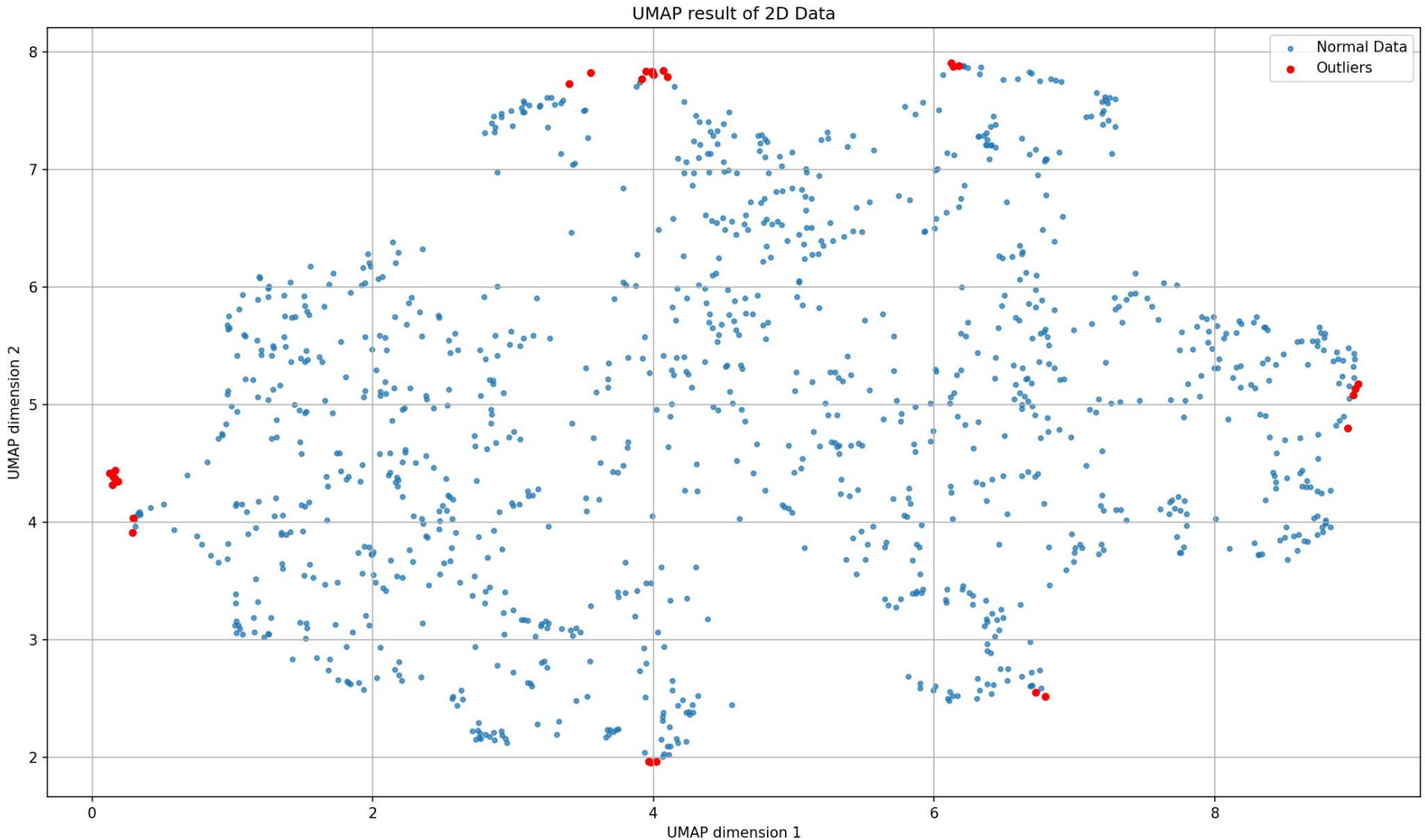
n_neighbors=3, min_dist=0.5



8, 38, 39, 69, 72,
78, 79, 91, 147,
275, 427, 527, 755,
796, 860, 917, 923,
941, 1034, 1035,
1036, 1037, 1039,
1040, 1064, 1065,
1077, 1099, 1124

降维至2维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

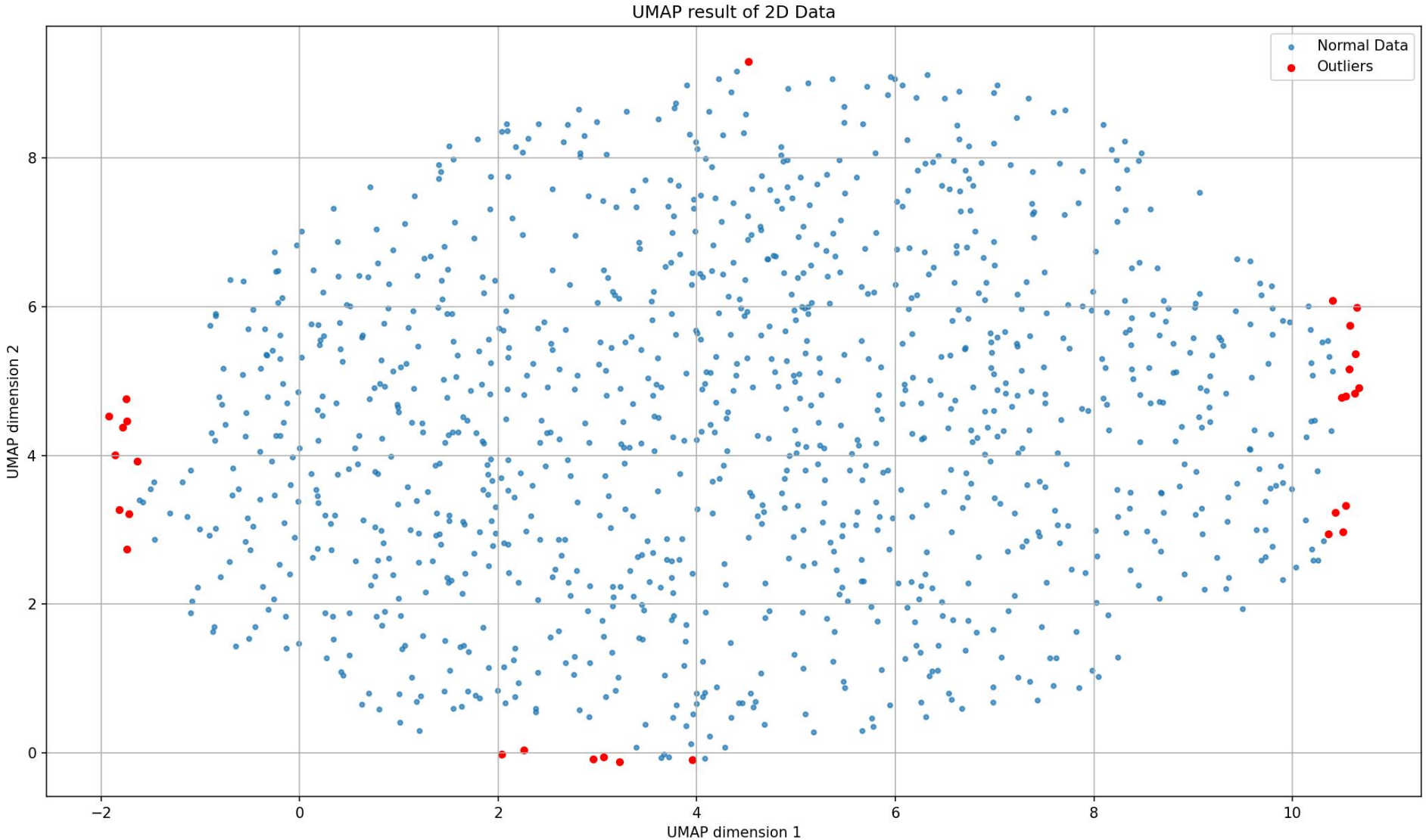
n_neighbors=15, min_dist=0



69, 77, 78, 79, 182, 194, 242, 301, 302, 369, 371, 403, 438, 502, 654, 677, 729, 783, 797, 805, 860, 871, 875, 908, 915, 923, 1092, 1099, 1114

降维至2维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

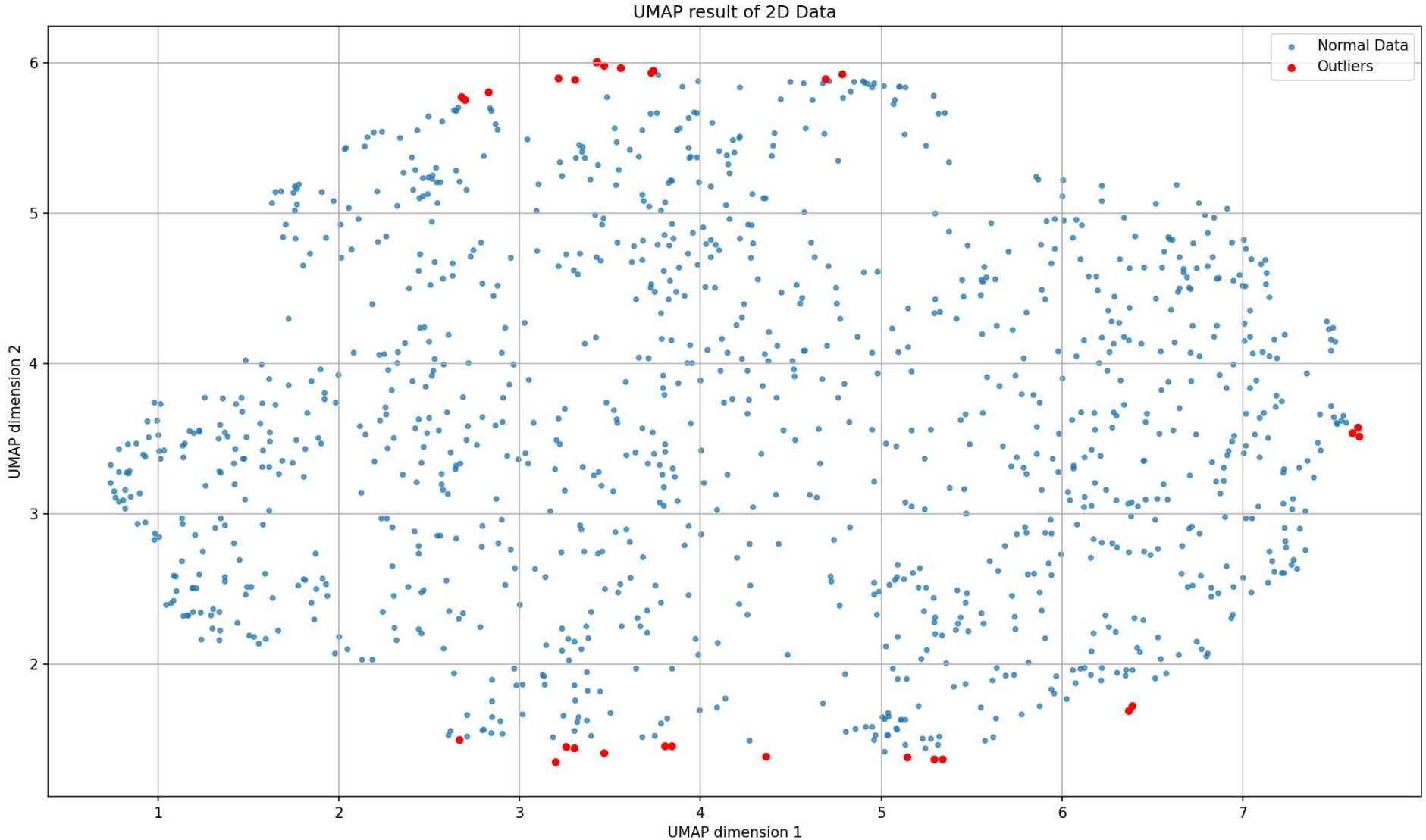
n_neighbors=15, min_dist=0.5



69, 77, 78, 79, 125, 178, 182, 187, 194, 267, 290, 301, 360, 369, 388, 438, 505, 569, 677, 875, 893, 905, 915, 923, 997, 1066, 1099, 1106, 1114

降维至2维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

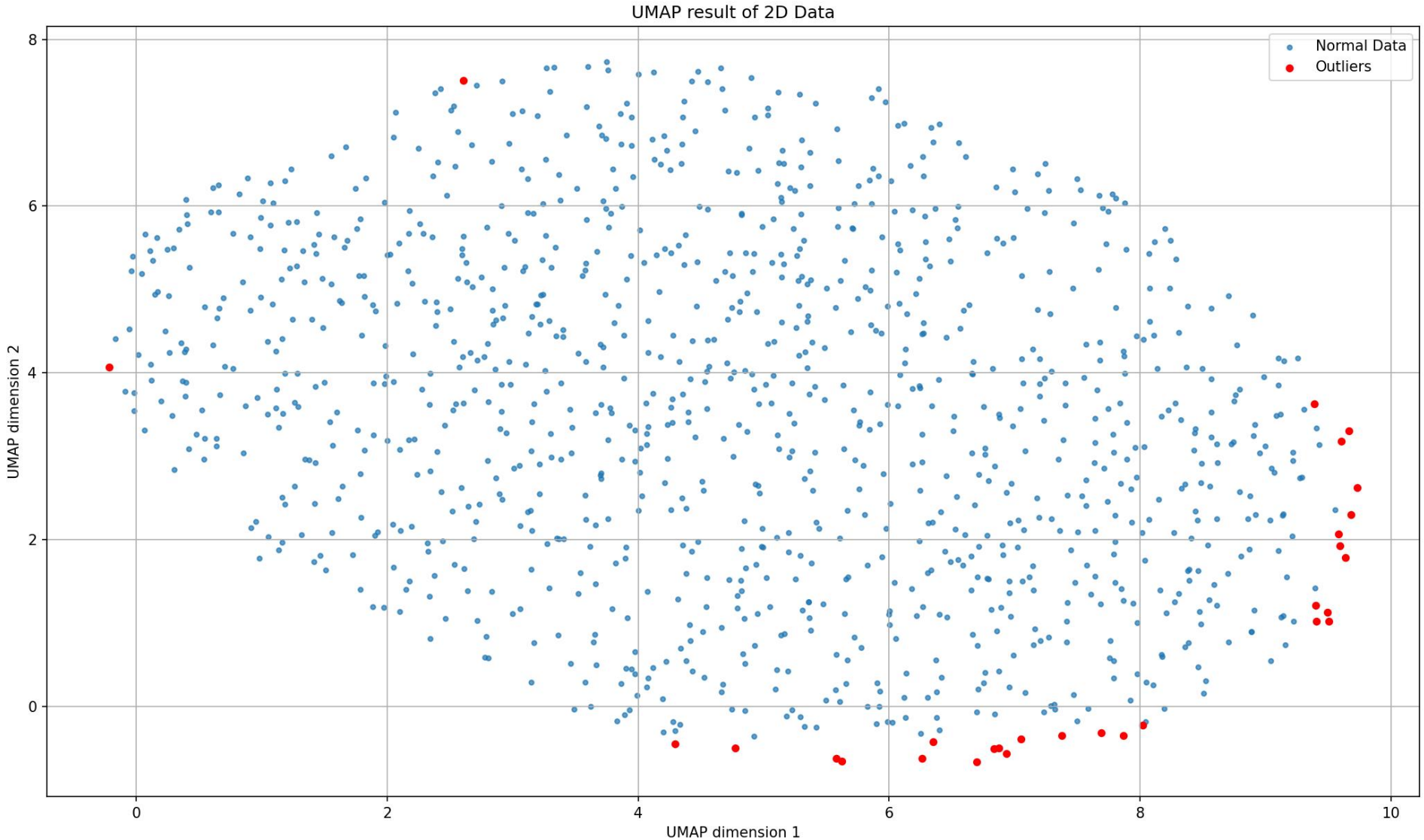
n_neighbors=100, min_dist=0



25, 182, 192, 265, 286, 301, 306, 364, 369, 373, 388, 391, 403, 419, 502, 654, 655, 684, 729, 751, 783, 797, 805, 871, 875, 908, 915, 925, 1114

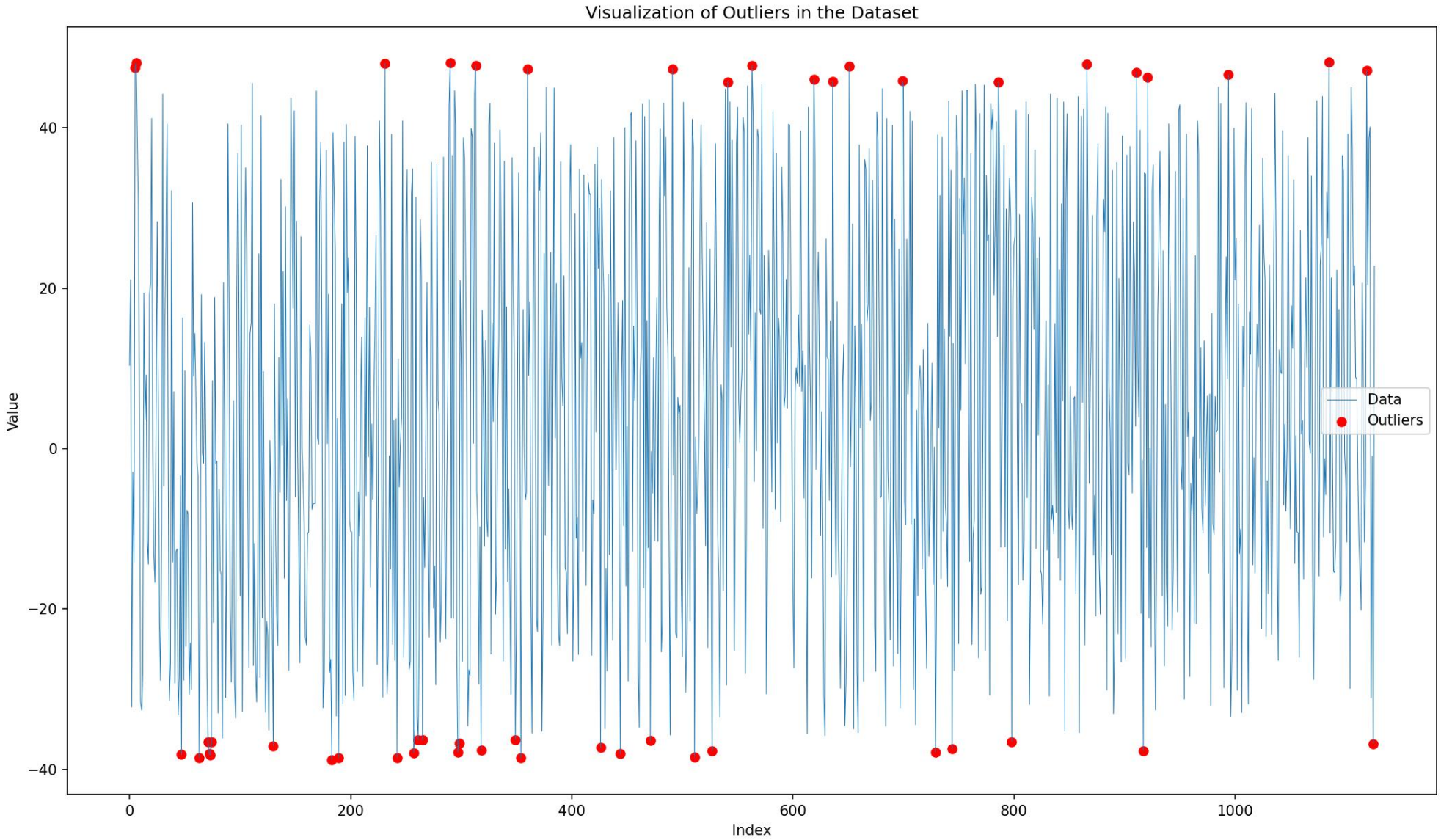
降维至2维，可视化结果：（其中蓝色的点为正常点，红色点为异常点）

n_neighbors=100, min_dist=15



27, 61, 69, 77, 78, 79, 96, 125, 182, 238, 301, 338, 368, 369, 401, 421, 505, 534, 683, 794, 878, 890, 903, 915, 916, 922, 948, 1058, 1099

降维至1维，可视化结果：（其中红色点为异常点）



6, 7, 48, 64, 72, 73,
74, 75, 131, 184,
190, 232, 243, 258,
262, 266, 291, 298,
299, 314, 319, 350,
355, 361, 427, 445,
472, 492, 512, 528,
542, 564, 620, 637,
652, 700, 730, 745,
787, 799, 867, 912,
918, 922, 995, 1086,
1120, 1126

UMAP：检测结果准确率

检测方式：将14个标签文件中有超过4个文件被标记的点取出来（共33个），记作明显异常点；超过2个文件被标记的点取出来（共179个），记作普通异常点。与降维后检测异常的结果进行比对：

	明显异常点	普通异常点
1维	5 (10.4%)	16（33.3%）
2维	5（17.2%）	13（44.8%）
3维	10（43.4%）	16（69.5%）

其中效果最好的是：

降维至3维（展开为1127*28的矩阵后UMAP(n_neighbors=15, min_dist= 0)，利用Isolation Forest检测异常点）：普通异常点有16个符合（69.5%）

反复调整n_neighbors和min_dist的值，发现UMAP数据降到一、二维时准确率非常低

和PCA的结果相比，UMAP准确率较低

结论：UMAP方法降维此数据集的效果不如PCA

未解决的问题：

怎样更准确的衡量降维手段的好坏?

后续想法：

寻找信息损失量更小的降维手段

探索更适合本题数据集的异常值检测方法

同济大学
TONGJI UNIVERSITY

Thank you for listening!

2024-7-18　　By 刘震