基于局部相似度保留自表达学习的快速多视角离群点检测

研读笔记

中南大学彭汪祺 2021/6/30

1. 背景

离群点检测是数据挖掘领域中一个非常重要的课题。首先，通过论文的阅读，**我认为“离群点”是指数据中存在的一些不符合预期正常行为模式的数据点或者数据团，或者说它是混杂在某一体制源数据中的另一体制的数据，因此它的数据表现与其它正常数据点不同。**

离群点检测主要有两个作用：**第一，离群点检测过滤掉数据中的干扰信息，帮助研究人员或是算法模型更快地捕捉数据中存在的一些固有深层模式；第二，检测到的异常数据本身也通常包含丰富信息，需要特别提取出来进行分析。**例如信用卡诈骗检测、网络入侵检测等需要检测异常行为模式的应用中。

但是常用的离群点检测都是通过单一来源的数据进行处理，很少研究会涉及多个数据来源的场景。而事实上，在真实世界中，我们往往需要通过多视角数据来分析事物。这种多源数据，即多视角数据，相比单视角数据含有更丰富的信息，但也创造了新的离群点类型。例如，我们在了解一个旅游景点时，不仅要通过文字描述来了解，还需要通过图片作进一步认识，这就说明多视角数据能够为我们提供更多的信息，同时其中的离群点检测也更加复杂。

多视角数据的两大属性为一致性和互补性。所谓一致性，是因为不同视角的数据，都是在描述同一个对象，因此所表达的性状是一致的；而互补性，是由于不同视角的表现形式不同，其获取数据的来源也不同，因此必然存在不同的数据，这些数据可以在建立模型时进行互补。

2. 论文工作

这篇论文提出了一种基于局部相似度保留自表达学习的快速多视角离群点检测算法（SRLSP），SRLSP 融合了自表达学习和相似度学习，对类别离群点和属性离群点都有很好的检测性能，能够处理任意视角数量、无聚类结构、多种类型离群点、大规模数据集和在线计算场景这五种情况下的检测任务。具体有以下三种贡献：

(1) SRLSP 学习一个视角共享的相似度矩阵，并尝试使用这个矩阵来还原原始的数据特征；

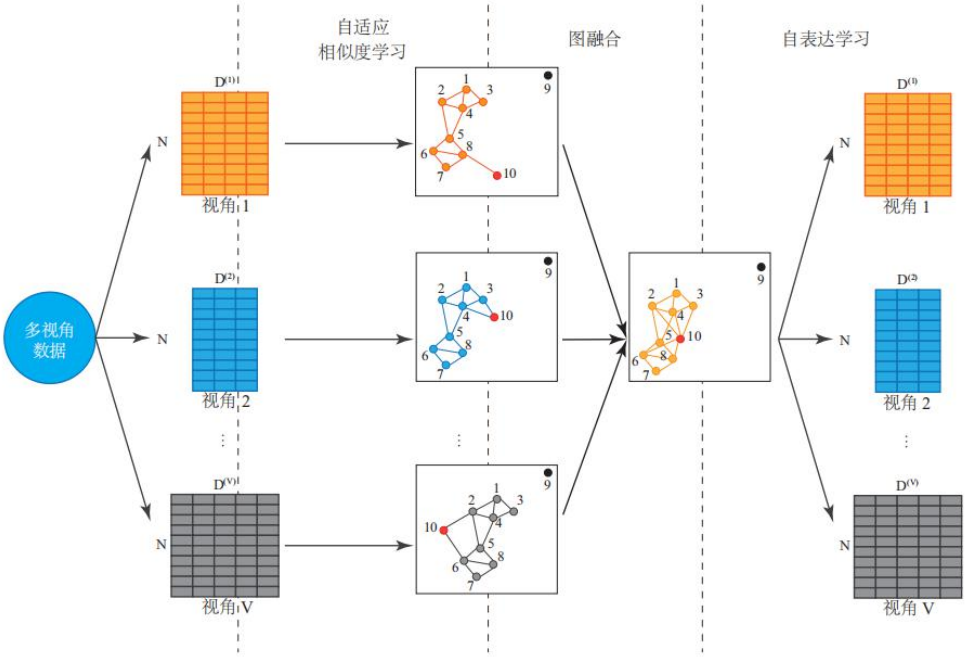
(2) 相似度学习的过程对数据集结构没有任何要求；

(3) SRLSP 是第一个能处理在线场景的多视角离群点检测算法。

3. 论文模型建立

3.1 模型介绍

论文介绍了两个子模型：l_2正则的自表达学习模型和图融合的自适应相似度学习，通过将两个子模型融合，两个子模型相互增益，得出基于局部相似度保留自表达学习的快速多视角离群点检测算法，该模型对于属性离群点和类别离群点都有很好的检测性能。整个模型运行的原理如下图所示：



SRLSP 模型首先将多个视角下的各个视角矩阵中，通过自适应相似度矩阵，构建出相似度矩阵，由于多视角的数据实质上都是在描述同一个实例对象，因此普通点在不同视角的相似度矩阵中，其邻居节点和拓扑结构都是一致的，而类别离群点在不同视角的的邻居节点和拓扑结构不同，将这些相似度矩阵通过图融合，检测不一致行为，可以检测出类别离群点 10。l_2正则的自表达学习对于检测属性离群节点有着相当好的效果，这里继续通过自表达学习来检测属性离群节点 9，同时恢复出各个视角的矩阵。

3.2 l_2正则的自表达学习子模型

该子模型分析单视角下的情况。设一组包含 N 个实例，V 个视角的数据集 X，它由 D 个特征描述。给出 N × D 单视角数据集 X，希望找到一个 N × N 的相似性矩阵 S，重构误差 N × D 的矩阵 E。重构过程可以通过以下式子来表现：

$$\boldsymbol{X}_i = \boldsymbol{S}_i\boldsymbol{X} + \boldsymbol{E}_i = \sum_{j=1}^{\mathrm{N}} \boldsymbol{S}_{i,j}\boldsymbol{X}_j + \boldsymbol{E}_i, \forall i = 1, \cdots, \mathrm{N}$$

自表达学习希望能够将重构误差矩阵减到最小，可以使用 F 范数来进行约束。为了调整矩阵 S 的表达能力，本文选用了l_2正则对矩阵 S 进行约束，目标函数如下所示：

$$\min_{S}\sum_{i=1}^{N}\left(\left\|\boldsymbol{X}_i-\sum_{j\in\mathcal{N}(i)}\boldsymbol{S}_{i,j}\boldsymbol{X}_j\right\|_2^2+\gamma\|\boldsymbol{S}_i\|_2^2\right)$$

其中，γ 和 k 作为输入参数。这里还需要用到 KNN 算法来计算点 i 的 k 邻居ℵ(i)，具体求解过程为：

(1) 取出 j ∈ ℵ(i)在S_i的第 j 个元素，组成向量$Z_i \in \mathbb{R}^{1\times k}$；

(2) 另外取出矩阵 X 相应的第 j 行，组成矩阵$X_{\aleph(i)} \in \mathbb{R}^{k\times D}$，从而可以将目标函数改为如下的形式：

$$\min_{\boldsymbol{Z}}\sum_{i=1}^{\mathrm{N}}\left(\left\|\boldsymbol{X}_i-\boldsymbol{Z}_i\boldsymbol{X}_{\mathcal{N}(i)}\right\|_2^2+\gamma\|\boldsymbol{Z}_i\|_2^2\right)$$

(3) 对Z_i求导，并令导数为零，这样就可以得到极值如下：

$$\boldsymbol{Z}_i=\boldsymbol{X}_i\boldsymbol{X}_{\mathcal{N}(i)}^T\left(\boldsymbol{X}_{\mathcal{N}(i)}\boldsymbol{X}_{\mathcal{N}(i)}^T+\gamma\boldsymbol{I}\right)^{-1}$$

这里 γ 发挥着重要的作用。当 γ 太小时，模型的表达能力过强，正常数据点和属性离群点都能完美表达，这样就难以检测出属性离群点，只有当 γ 设置合适时，模型要求重构向量中的权重较小，因此要想让属性离群点完美表达，需要配置较大的权重，而Z_i的较大元素会被这样就会被l_2正则给予较大的惩罚。这里研究者通过四个坐标点来说明了这个特点：

设置四个点的坐标分别为(1,2)，(1.5,3.5)，(0.3,1.5)，(4,9)。γ 为 0 时，得到的重构误差矩阵 E 为：

$$\begin{array}{ccccccc} \boldsymbol{X}_i & = & \boldsymbol{Z}_i & & \boldsymbol{X}_{\mathcal{N}(i)} & + & \boldsymbol{E}_i \\ \begin{bmatrix}0.3 & 1.5\end{bmatrix} & = & \begin{bmatrix}-2.4 & 1.8\end{bmatrix} & \begin{bmatrix}1 & 2\\ 1.5 & 3.5\end{bmatrix} & + & \begin{bmatrix}0 & 0\end{bmatrix} \\ \begin{bmatrix}4 & 9\end{bmatrix} & = & \begin{bmatrix}1 & 2\end{bmatrix} & \begin{bmatrix}1 & 2\\ 1.5 & 3.5\end{bmatrix} & + & \begin{bmatrix}0 & 0\end{bmatrix} \end{array}$$

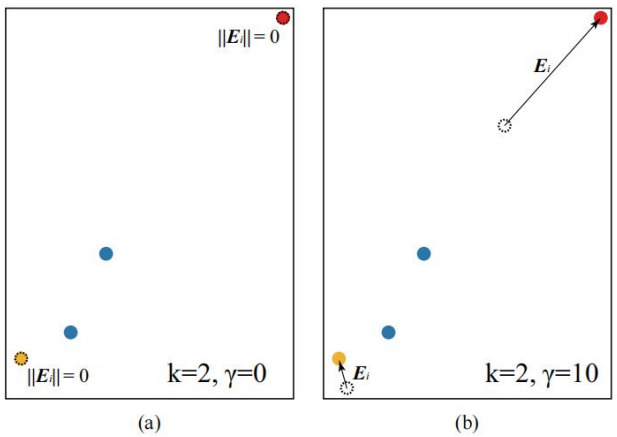
很明显，重构误差矩阵都为 0，无法区分。

设置 γ 为 10，重新得到的结果为：

$$\begin{array}{ccccccc} \boldsymbol{X}_i & = & \boldsymbol{Z}_i & & \boldsymbol{X}_{\mathcal{N}(i)} & + & \boldsymbol{E}_i \\ \begin{bmatrix}0.3 & 1.5\end{bmatrix} & = & \begin{bmatrix}0.1097 & 0.1946\end{bmatrix} & \begin{bmatrix}1 & 2\\ 1.5 & 3.5\end{bmatrix} & + & \begin{bmatrix}-0.1016 & 0.5995\end{bmatrix} \\ \begin{bmatrix}4 & 9\end{bmatrix} & = & \begin{bmatrix}0.7460 & 1.2718\end{bmatrix} & \begin{bmatrix}1 & 2\\ 1.5 & 3.5\end{bmatrix} & + & \begin{bmatrix}1.3463 & 3.0567\end{bmatrix} \end{array}$$

该结果可以通过权重来区分出属性离群点，效果较好。

该测试的具体图示如下：



(a) (b)

在顺利进行自表达过程的情况下，l_2正则项的使用可以方便我们直接从残差的角度进行分析，判断一个点是否有较大的概率是属性离群点。

3.3 多图融合的自适应相似度学习

多角度数据集处理得到 V 个视角的相似度矩阵后，我们需要考虑：如何利用多视角数据一致性，给这些相似度矩阵加上合适的约束；以及怎么从这些图里面达成共识，找出联合的一个图，方便后续的任务来使用呢？其中一种简单高效的方法称为多图融合。

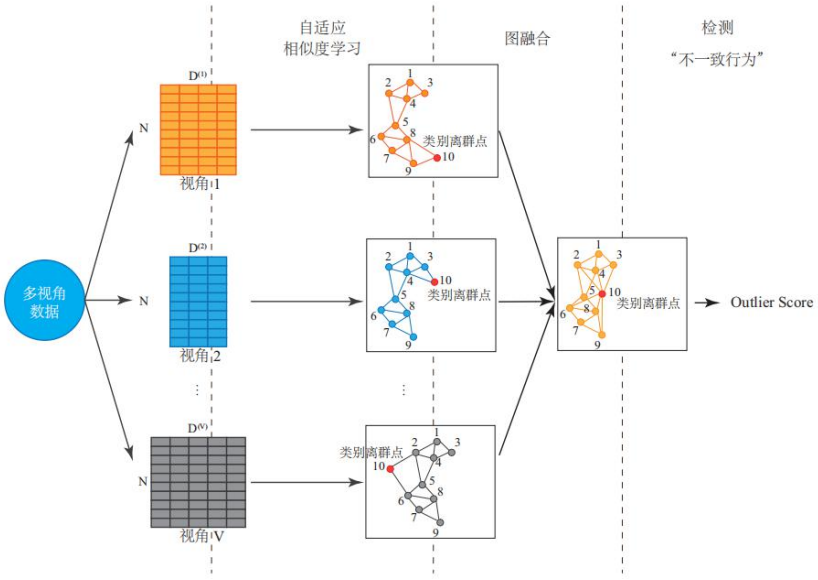
已知多视角数据的目标函数可以用下式来表示：

$$\sum_{v=1}^{(v)} \left\| S^{(v)} - S \right\|_F^2$$

我们对不同视角的数据使用自适应相似度学习，再使用多图融合的技术，就可以得到如下的目标函数式：

$$\min_{S,S^{(1)},\cdots,S^{(V)}} \sum_{v=1}^{\mathrm{V}} \left[\sum_{i,j=1}^{\mathrm{N}} \left(S_{i,j}^{(v)} \left\| \boldsymbol{X}_i^{(v)} - \boldsymbol{X}_j^{(v)} \right\|_2^2 + \gamma \boldsymbol{S}_{i,j}^{(v)^2} \right) + \lambda \left\| \boldsymbol{S}^{(v)} - \boldsymbol{S} \right\|_F^2 \right]$$

该目标函数通过在每个视角数据中使用自适应相似度学习来学出合适的视角相似度矩阵$S^{(v)}$，结合图融合过程，从而得到一个结合多视角一致性信息的共识图 S。对于该函数，一般使用坐标上升法进行优化，即$S^{(v)}$和 S 交替优化。这个过程大体框架如下图所示：



这个框架能用来处理类别离群点的“不一致行为”。正常数据点符合多视角数据“一致性”的属性，在大部分视角下具有类似的邻居结构，且与邻居的相似情况是高度近似的，因此能较一致地学出一个统一的全局相似关系。

3.4 模型融合

属性离群点和类别离群点模型的融合，实际上是一个相互增益的过程，这种融合将提高整体模型在多视角离群点检测的效果。

研究者提出的初步融合模型为：

$$\min_{\boldsymbol{S},\boldsymbol{S}^{(1)},\cdots,\boldsymbol{S}^{(\mathrm{V})}} \sum_{i=1}^{\mathrm{N}}\left[\sum_{v=1}^{\mathrm{V}}\left(\|\boldsymbol{X}_i^{(v)}-\sum_{j=1}^{\mathrm{N}}\boldsymbol{S}_{i,j}\boldsymbol{X}_j^{(v)}\|_2^2+\lambda\|\boldsymbol{S}_i-\boldsymbol{S}_i^{(v)}\|_2^2+\mu\sum_{j=1}^{\mathrm{N}}\|\boldsymbol{X}_i^{(v)}-\boldsymbol{X}_j^{(v)}\|_2^2\boldsymbol{S}_{i,j}^{(v)}\right)+\gamma\|\boldsymbol{S}_i\|_2^2\right]$$

$$\begin{aligned}
s.t.\ & diag(\boldsymbol{S})=\boldsymbol{0},\\
& diag(\boldsymbol{S}^{(v)})=\boldsymbol{0},\\
& \boldsymbol{S}_i^{(v)}\boldsymbol{1}=1,\\
& \boldsymbol{0}\leqslant\boldsymbol{S}_i^{(v)}\leqslant\boldsymbol{1},\\
& \forall v=1,\cdots,\mathrm{V},\\
& \forall i=1,\cdots,\mathrm{N}.
\end{aligned}$$

该融合过程中存在两个假设：

(1) 首先，需要假设多视角数据由一个共同的隐空间生成，并可以借由每个视角下的数据度量进行还原。事实上，这个还原过程类似于自动编码器的解码过程，模型的复杂程度决定了对数据复杂程度的容纳能力；

(2) 其次，另一方面，因为对数据模型进行了一致性的假设——多视角的数据都由同一个相同的隐空间生成，这个模型具有弱类别离群点的检测能力，多视角的邻居关系大致相似。

在上述函数中，需要使用所有的数据点进行数据还原和相似度计算。但是根据前面的分析，事实上只有每个点的邻居在这两个过程的计算中会发挥比较大的作用。因此，考虑到每个点 i 在每个视角 v 下的 k 邻居$\aleph(i,v)$，并且将这 V 个邻居集合到每个点在所有视角下的邻居集合$\aleph(i)$。在只考虑邻居的情况下，目标函数可以进一步表达为：

$$\min_{\boldsymbol{S},\boldsymbol{S}^{(1)},\cdots,\boldsymbol{S}^{(\mathrm{V})}} \sum_{i=1}^{\mathrm{N}}\left[\sum_{v=1}^{\mathrm{V}}\left(\|\boldsymbol{X}_i^{(v)}-\sum_{j\in\mathcal{N}(i)}\boldsymbol{S}_{i,j}\boldsymbol{X}_j^{(v)}\|_2^2+\lambda\|\boldsymbol{S}_i-\boldsymbol{S}_i^{(v)}\|_2^2+\mu\sum_{j\in\mathcal{N}(i)}\|\boldsymbol{X}_i^{(v)}-\boldsymbol{X}_j^{(v)}\|_2^2\boldsymbol{S}_{i,j}^{(v)}\right)+\gamma\|\boldsymbol{S}_i\|_2^2\right]$$

$$\begin{aligned}
s.t.\ & \boldsymbol{S}_{i,\neg\mathcal{N}(i)}=\boldsymbol{0},\\
& \boldsymbol{S}_{i,\neg\mathcal{N}(i)}^{(v)}=\boldsymbol{0},\\
& \boldsymbol{S}_i^{(v)}\boldsymbol{1}=1,\\
& \boldsymbol{0}\leqslant\boldsymbol{S}_i^{(v)}\leqslant\boldsymbol{1},\\
& \forall v=1,\cdots,\mathrm{V},\\
& \forall i=1,\cdots,\mathrm{N}.
\end{aligned}$$

此外，预先计算出点 i 和它的邻居$\aleph(i)$之间的距离$D_{i,j}$，通过下式得到向量$D_i^{(v)}$:

$$\boldsymbol{D}_{i,j}^{(v)}=\|\boldsymbol{X}_i^{(v)}-\left(\boldsymbol{X}_{\mathcal{N}(i)}^{(v)}\right)_i\|_2^2,\ j=1,\cdots,|\mathcal{N}(i)|$$

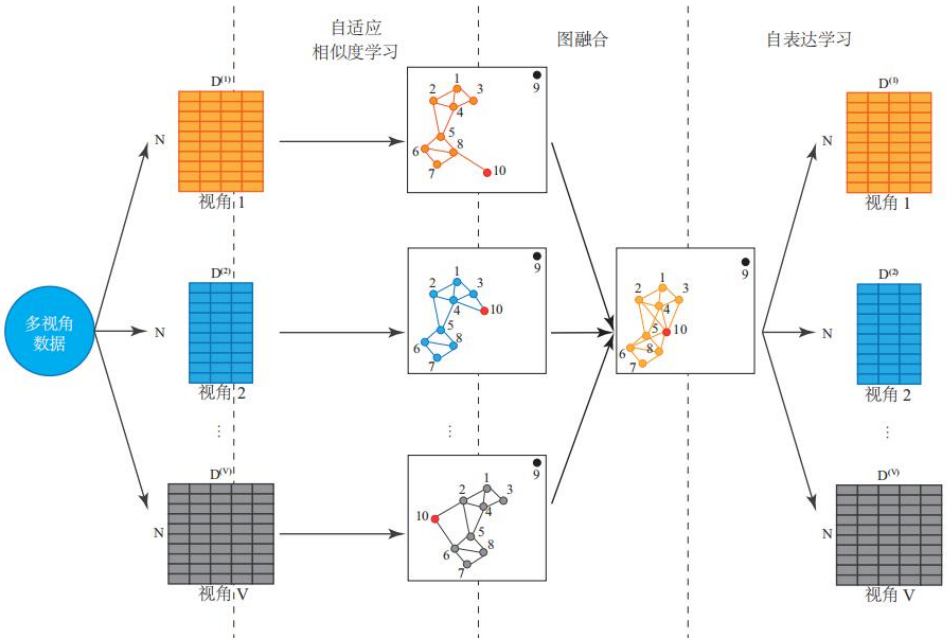
根据结果进行换元，得到模型最终目标函数如下：

$$\min_{\mathbf{Z},\mathbf{Z}^{(1)},\cdots,\mathbf{Z}^{(\mathrm{V})}} \sum_{i=1}^{\mathrm{N}}\left[\sum_{v=1}^{\mathrm{V}}\left(\|\boldsymbol{X}_i^{(v)}-\boldsymbol{Z}_i\boldsymbol{X}_{\mathcal{N}(i)}^{(v)}\|_2^2+\lambda\|\boldsymbol{Z}_i-\boldsymbol{Z}_i^{(v)}\|_2^2+\mu\boldsymbol{D}_i^{(v)}\boldsymbol{Z}_i^{(v)^T}\right)+\gamma\|\boldsymbol{Z}_i\|_2^2\right]$$

$$\begin{aligned}
s.t.\ & \boldsymbol{Z}_i^{(v)}\boldsymbol{1}=1,\\
& \boldsymbol{0}\leqslant\boldsymbol{Z}_i^{(v)}\leqslant\boldsymbol{1},\\
& \forall v=1,\cdots,\mathrm{V},\\
& \forall i=1,\cdots,\mathrm{N}.
\end{aligned}$$

总体来说，该模型通过自适应相似度学习$\mu\mathrm{D}_\mathrm{i}^{(\mathrm{v})}\mathrm{Z}_\mathrm{i}^{(\mathrm{v})^\mathrm{T}}$学习视角下的相似度关系$\mathrm{Z}_\mathrm{i}^{(\mathrm{v})}$，接着通过图融合过程$\lambda||\mathrm{Z}_\mathrm{i}-\mathrm{Z}_\mathrm{i}^{(\mathrm{v})}||_2^2$得到视角共享的相似度关系$\mathrm{Z}_\mathrm{i}$，最后使用自表达学习$||\mathrm{X}_\mathrm{i}^{(\mathrm{v})}-$

$Z_i X^{(v)}_{\aleph(i)}||_2^2$换元原始的多视角数据。基本流程图如下：



4. 优化求解和分析

由于相似度矩阵 S 的每一行都是独立的，因此可以分开单独优化，具体的优化问题可以转成如下式子：

$$\min_{\boldsymbol{s}} \boldsymbol{d}^T \boldsymbol{s} + \gamma \|\boldsymbol{s}\|_2^2$$

$$s.t.\ \boldsymbol{s}^T \mathbf{1} = 1,$$

$$0 \leqslant \boldsymbol{s} \leqslant 1$$

接下来需要讨论该问题的具体求解。

4.1 自适应相似度学习优化方法 1

上述问题是凸优化问题，这里可以套用 KKT 的条件式进行求解。的拉格朗日函数如下式所示：

$$L(\boldsymbol{x}, \lambda, \boldsymbol{v}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{\mathrm{M}} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{\mathrm{P}} h_i(\boldsymbol{x})$$

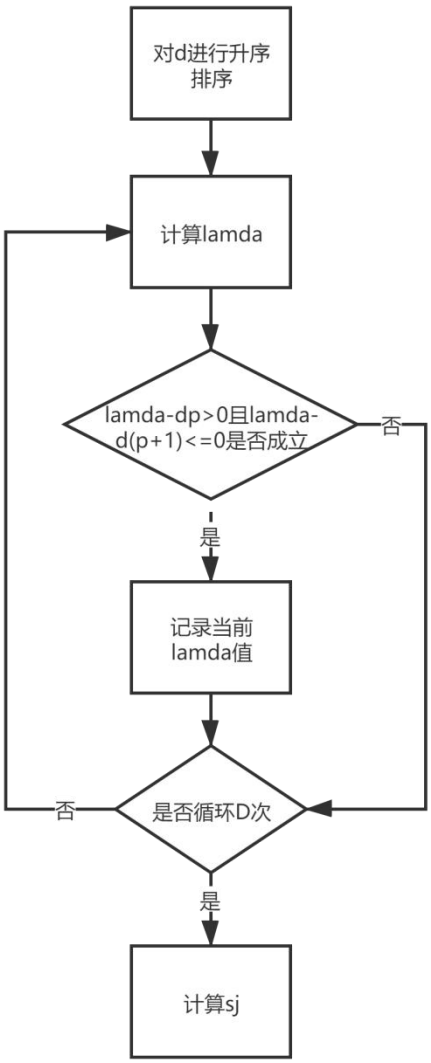
对应的 KKT 条件式为：

$$\frac{\partial L(\boldsymbol{s},\lambda,\boldsymbol{v})}{\partial \boldsymbol{s}} = \boldsymbol{d} + 2\gamma \boldsymbol{s} - \lambda \mathbf{1} - \boldsymbol{v} = 0,$$
$$\boldsymbol{s}^T \mathbf{1} - 1 = 0,$$
$$0 \leqslant \boldsymbol{s},$$
$$\boldsymbol{v} \geqslant 0,$$
$$\boldsymbol{v}_j \boldsymbol{s}_j = 0,\ j = 1, \cdots, \mathrm{D}.$$

得到三种情况：

$$\begin{cases} s_j = \dfrac{1}{2\lambda}(\lambda 1 + v - d), & \lambda - d_j > 0 \\ s_j = v_j = 0, & \lambda - d_j = 0 \\ s_j = 0, & \lambda - d_j < 0 \end{cases}$$

由此得出穷解法来优化自适应学习，流程图如下：



4.2 自适应相似度学习优化方法 2

该优化方法与方法 1 的一大区别在于，循环次数 p 不是通过枚举得到的，而是直接等于预先给定的 k，而λ是一个可学习变量。因此存在

$$k\boldsymbol{d}_k - \sum_{j=1}^{k} \boldsymbol{d}_j < 2\gamma \leqslant k\boldsymbol{d}_{k+1} - \sum_{j=1}^{k} \boldsymbol{d}_j$$

取 λ 的上界，得到下式：

$$\gamma = \frac{k}{2}\boldsymbol{d}_{k+1} - \frac{1}{2}\sum_{j=1}^{k} \boldsymbol{d}_j,$$

$$\lambda = \boldsymbol{d}_{k+1}.$$

继而可以得到 s 的求解公式：

$$\boldsymbol{s}_j = \frac{\left(\boldsymbol{d}_{k+1} - \boldsymbol{d}_j\right)_+}{k\boldsymbol{d}_{k+1} - \sum_{j=1}^{k} \boldsymbol{d}_j}$$

这就是第二种优化方法。

4.3 模型优化

优化主要采用坐标上升算法，**原因在于独特视角相似度与全局相似度的数据之间不存在依赖关系，因此可以彼此独立优化**。因此在坐标上升算法中，对此二者的变量集合涉及到的子问题进行优化求解。

首先固定全局相似度参数 Z，而更新独特视角参数$Z_i^{(v)}$：

子问题可以用下式来表示：

$$\min_{\boldsymbol{Z}_i^{(v)}} \left(\mu \boldsymbol{D}_i^{(v)} - 2\lambda \boldsymbol{Z}_i\right) \boldsymbol{Z}_i^{(v)T} + \lambda \| \boldsymbol{Z}_i^{(v)} \|_2^2$$

$$s.t.\ \boldsymbol{Z}_i^{(v)}\mathbf{1} = 1,$$

$$\mathbf{0} \leqslant \boldsymbol{Z}_i^{(v)} \leqslant \mathbf{1}.$$

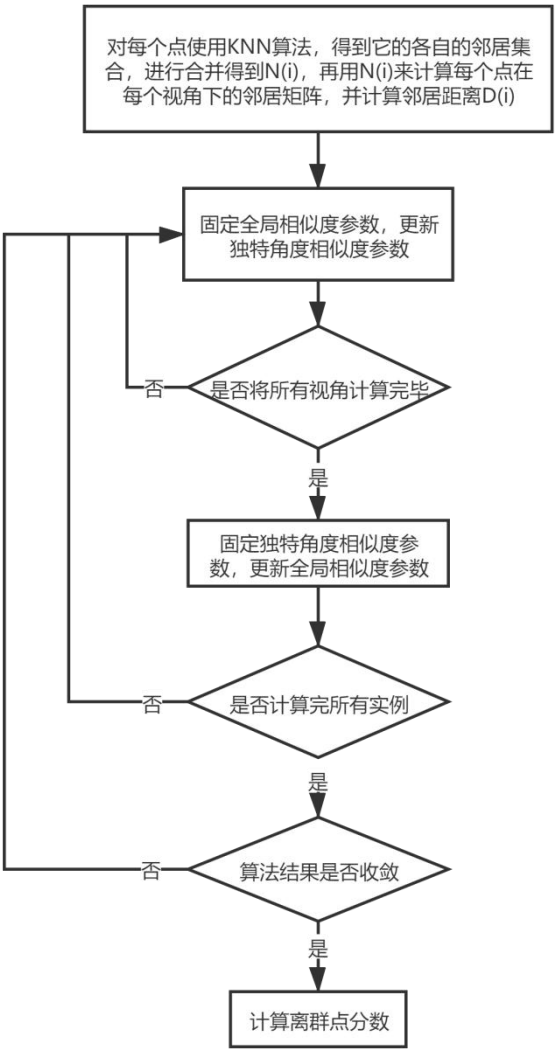
这可以用上述两种优化方法中的任意一种来进行优化。

接着，固定$Z_i^{(v)}$，而更新参数Z_i：

子问题同样可以用下式来表示：

$$Z_i = \left[\sum_{v=1}^{V}\left(X_i^{(v)} X_{\mathcal{N}(i)}^{(v)\ T} + \lambda Z_i^{(v)}\right)\right]\left(\sum_{v=1}^{V} X_{\mathcal{N}(i)}^{(v)} X_{\mathcal{N}(i)}^{(v)\ T} + \lambda V\mathbf{I} + \gamma\mathbf{I}\right)^{-1}$$

综上，用坐标上升法来求解问题的流程如下：

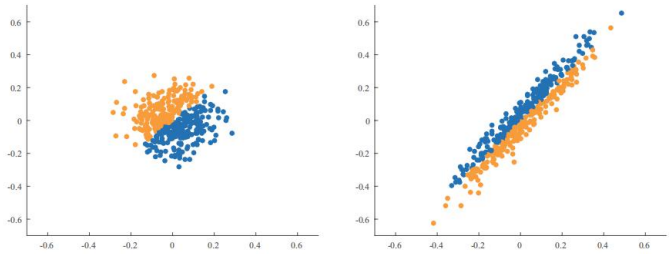


通过两个变量的子问题的最优化计算，目标函数保证一直下降，且原目标函数有下界，一定能够收敛到某个局部最优值。在这样一个并行交互更新的过程中，算法逐渐学习到数据集内在的模式，从而不断将多视角离群点区分开来。

5. 实验测试

5.1 人工数据集测试

首先采用 python 中的 sklearn 库，来生成无类别属性的数据集，人工数据集用于验证挑战 2——算法是否可以处理非聚类属性的数据集。这里生成了两个不同视角的下的数据集，如下图所示：



(a) 视角 1　　(b) 视角 2：由视角 1 的数据经过一个线性变换再加上一个随机扰动构成

该数据集中，类别-属性离群点的比例为N_A：N_C：N_{CA} = 5%：5%：5%。实验的评价指标为 AUC，AUC 是 ROC 曲线下的积，它代表区分程度的度量，表示算法模型能在多大程度上对样本进行正确区分，AUC 越大的算法表现越优秀。

SRLSP 模型与 OneClassSVM，HOAD，AP 等模型进行比较，得到在人工多视角数据集下的实验结果：

方法	AUC (mean ± std) ↑
OneClassSVM	0.842 ± 0.018
HOAD	0.468 ± 0.161
AP	0.660 ± 0.042
MLRA	0.790 ± 0.119
LDSR	0.825 ± 0.043
MODDIS	0.975 ± 0.011
SRLSP	**0.996 ± 0.005**

从结果中看来，SRLSP 模型的效果最佳。

5.2 UCI 数据集

UCI 数据集为当前多视角离群点检测算法常用的数据集，将在不同的离群点比例下进行实验，用于验证挑战 3——算法对 3 种多视角离群点的检测能力。UCI 数据集属于都是单视角数据集，研究者将每个单视角数据集的特征平均分成 2 份，当成 2 个视角的数据来处理，每个数据集只有一种类型的离群点，实验结果如下所示：

方法	$N_A : N_C : N_{CA}$ = 10% : 0% : 0%				
	iris	pima	zoo	ionosphere	letter
OneClassSVM	0.995 ± 0.002	0.028 ± 0.015	0.002 ± 0.005	0.454 ± 0.023	0.364 ± 0.005
HOAD	0.122 ± 0.236	0.955 ± 0.171	0.723 ± 0.244	0.562 ± 0.126	0.430 ± 0.094
AP	0.037 ± 0.036	0.001 ± 0.002	0.396 ± 0.101	0.483 ± 0.037	0.487 ± 0.012
MLRA	0.976 ± 0.027	0.934 ± 0.013	0.846 ± 0.066	0.637 ± 0.152	0.758 ± 0.029
LDSR	0.979 ± 0.029	0.994 ± 0.004	0.928 ± 0.018	0.727 ± 0.016	**0.996 ± 0.001**
MODDIS	0.758 ± 0.121	0.002 ± 0.003	0.894 ± 0.040	0.710 ± 0.022	0.051 ± 0.028
SRLSP	**1.000 ± 0.000**	**0.990 ± 0.002**	**0.979 ± 0.008**	**0.732 ± 0.007**	0.738 ± 0.007

方法	$N_A:N_C:N_{CA}=0\%:10\%:0\%$				
	iris	pima	zoo	ionosphere	letter
OneClassSVM	0.500 ± 0.080	0.505 ± 0.035	0.479 ± 0.093	0.533 ± 0.047	0.496 ± 0.025
HOAD	0.633 ± 0.100	0.514 ± 0.040	0.574 ± 0.073	0.449 ± 0.054	0.576 ± 0.022
AP	**0.966 ± 0.030**	0.492 ± 0.037	**0.933 ± 0.045**	**0.943 ± 0.024**	0.831 ± 0.014
MLRA	0.849 ± 0.061	0.664 ± 0.029	0.642 ± 0.097	0.801 ± 0.045	0.653 ± 0.021
LDSR	0.756 ± 0.121	0.627 ± 0.031	0.829 ± 0.064	0.834 ± 0.025	0.758 ± 0.024
MODDIS	0.819 ± 0.104	0.694 ± 0.030	0.782 ± 0.096	0.784 ± 0.034	0.852 ± 0.028
SRLSP	0.946 ± 0.043	**0.748 ± 0.030**	0.887 ± 0.054	0.922 ± 0.020	**0.925 ± 0.014**

方法	$N_A:N_C:N_{CA}=0\%:0\%:10\%$				
	iris	pima	zoo	ionosphere	letter
OneClassSVM	0.503 ± 0.064	0.508 ± 0.031	0.514 ± 0.093	0.555 ± 0.045	0.506 ± 0.021
HOAD	0.446 ± 0.150	0.357 ± 0.081	0.746 ± 0.021	0.429 ± 0.065	0.242 ± 0.117
AP	0.952 ± 0.022	0.436 ± 0.243	0.851 ± 0.067	**0.905 ± 0.018**	0.785 ± 0.021
MLRA	0.878 ± 0.039	0.744 ± 0.023	0.814 ± 0.054	0.725 ± 0.037	0.651 ± 0.032
LDSR	0.926 ± 0.038	**0.947 ± 0.015**	0.879 ± 0.044	0.785 ± 0.022	**0.957 ± 0.007**
MODDIS	0.866 ± 0.076	0.711 ± 0.085	0.856 ± 0.054	0.737 ± 0.016	0.821 ± 0.044
SRLSP	**0.981 ± 0.013**	0.809 ± 0.038	**0.930 ± 0.026**	0.795 ± 0.015	0.906 ± 0.015

从整体来看，SRLSP 在三种属性点上的综合表现优于另一种子空间类别方法 LDSR，其中对于属性离群点的处理能力更强，而 HOAD 算法在属性离群点上的检测一般。对于类别离群点，SRLSP 的表现也较为优秀，优于大部分算法。

5.3 真实多视角数据集

文中选用了三个真实数据集：

MSRC-v1 是剑桥大学微软研究院提供的一种场景识别数据集。我们选取了其中 7 个类别：tree，building，airplane，cow，face，bicycle，并使用 5 种视觉特征提取的方法生成每个视角的数据。

AWA 是一个关于动物属性的大规模数据集。我们选取了前 10 个类别：an telope，bat，beaver，bluewhale，bobcat，buffalo，chimpanzee，colie，cow，并从每个类别均匀选择 80 个实例。新数据集被称为 AWA-10。

Caltech101 是一个对象识别数据集。我们选取了其中 7 个类别：Faces，Mo torbikes，Dolla-Bill，Garfield，Snoopy，Stop-Sign，Windsor-Chair，并使用 6 种视觉特征提取的方法生成每个视角的数据。新数据集被称为 Caltech-7。

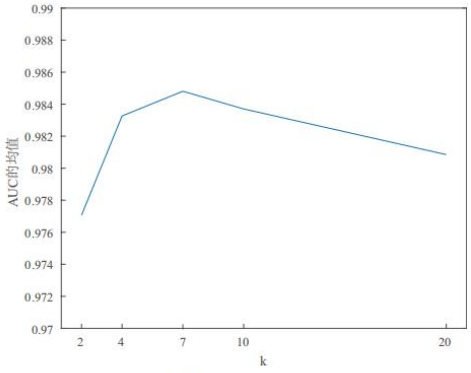
测试得到各个算法的效果如下：

方法	MSRC-v1	AWA-10	Caltech-7
OneClassSVM	0.680 ± 0.042	0.800 ± 0.024	0.789 ± 0.016
HOAD	0.374 ± 0.158	0.351 ± 0.119	0.356 ± 0.132
AP	0.508 ± 0.030	0.455 ± 0.024	0.454 ± 0.020
MLRA[1]	–	–	–
LDSR	0.973 ± 0.013	0.733 ± 0.060	0.943 ± 0.032
MODDIS	0.957 ± 0.013	0.813 ± 0.029	0.946 ± 0.008
SRLSP	**0.985 ± 0.006**	**0.938 ± 0.010**	**0.976 ± 0.004**

在真实数据集中，SRLSP 算法的性能非常出色，优于其它所有算法，证明了 SRLSP 在离群点检测中的优越性，更加适合实际应用。

5.4 参数敏感性实验

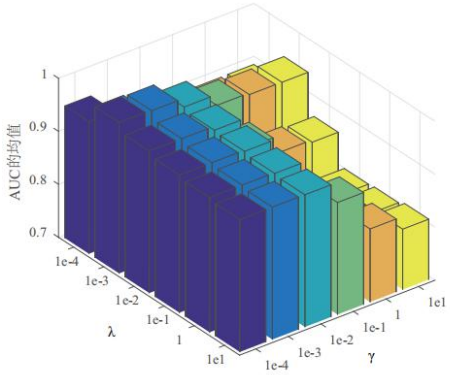
研究者对参数 k，λ 和 γ 进行了敏感度测试。首先固定 λ 和 γ，调整 k 得到的结果如下：



μ = 1, λ = 1e-3, γ = 1e-4

这里可以看出 k 与数据集结构有关，存在一个最优 k，使得检测效果最佳。

同理，固定 k，调整 λ 和 γ 得到的结果如下：



μ = 1, k = 7

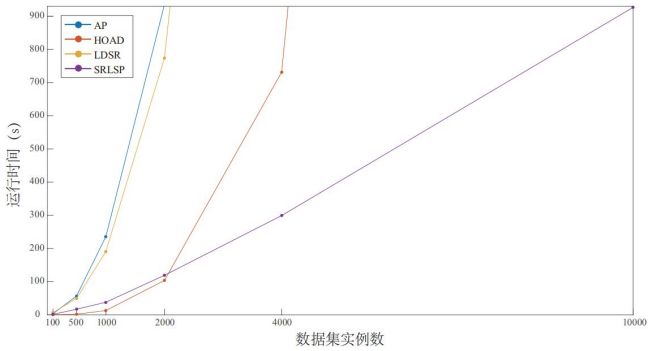
这里参数 λ 和 γ 与比例有关，λ 和 γ 越小，其检测准确度越高。但总体来说，整个模型对于参数不敏感。

5.5 运行时间对比实验

设置迭代收敛的条件如下：

$$\frac{|\mathrm{obj}^{t+1} - \mathrm{obj}^{t}|}{|\mathrm{obj}^{t}|} < 10^{-2}$$

不断增大数据集规模，检测不同算法达到收敛条件所需的时间，结果如下：



从这里可以看出，SRLSP 算法运行时间与数据集的规模成线性关系，明显优于其它算法，更容易应用到大规模的数据集上。

6. 总结

论文作者针对多视角离群点检测问题，**将自表达学习子模型与多图融合自适应相似度子模型融合，提出了 SRLSP 模型，它从局部相似度的角度出发，用邻居关系度量数据的一致性程度；同时从子空间的角度出发，用自表达学习进行数据还原，学习数据集的潜在结构**。

同时，从测试中可以看出，SRLSP 模型首先对多种离群点**具有较优秀的检测准确度**，并且能够处理多视角的数据集；而且**时间复杂度低**，其运行时间与数据规模基本成线性关系，更加适用于大规模数据应用；SRLSP 算法在三种真实多视角数据集中的性能测试均为最优，并且能方便地修改成在线版本，说明它**比其它算法更适合实际应用**。

7. 模型优化建议

(1) **从模型的运行示意图来看，它与自编码器非常类似，事实上该模型可以进一步扩展到神经网络上，采用一个编码器组件和解码器组件，其中编码器组件用于相似度学习，而解码器组件进行子空间学习，**该方法相对于论文中的子模型融合来说，可能具有更佳的性能与应用能力，并且可以减小参数的数量；

(2) 我认为可以进一步考虑集体离群点的问题。事实上离群点的意义与其规模密切相关。

如一家供应链公司，每天处理数以千计的订单和出货。如果一个订单的出货延误，则可能不是离群点，因为统计表明延误时常发生。然而，如果有一天有 100 个订单延误，则必须注意。这 100 个订单整体来看，形成一个离群点，尽管如果单个考虑，它们每个或许都不是离群点。

因此，模型在类似上述的应用场景下，需要考虑离群点的规模大小，异常行为的规模往往蕴含着大量信息，大规模的异常行为需要重点关注，而小规模的异常行为甚至可以忽略。如何在识别离群点的同时，对其规模大小进行判断，这可能是该模型优化的一个方向。