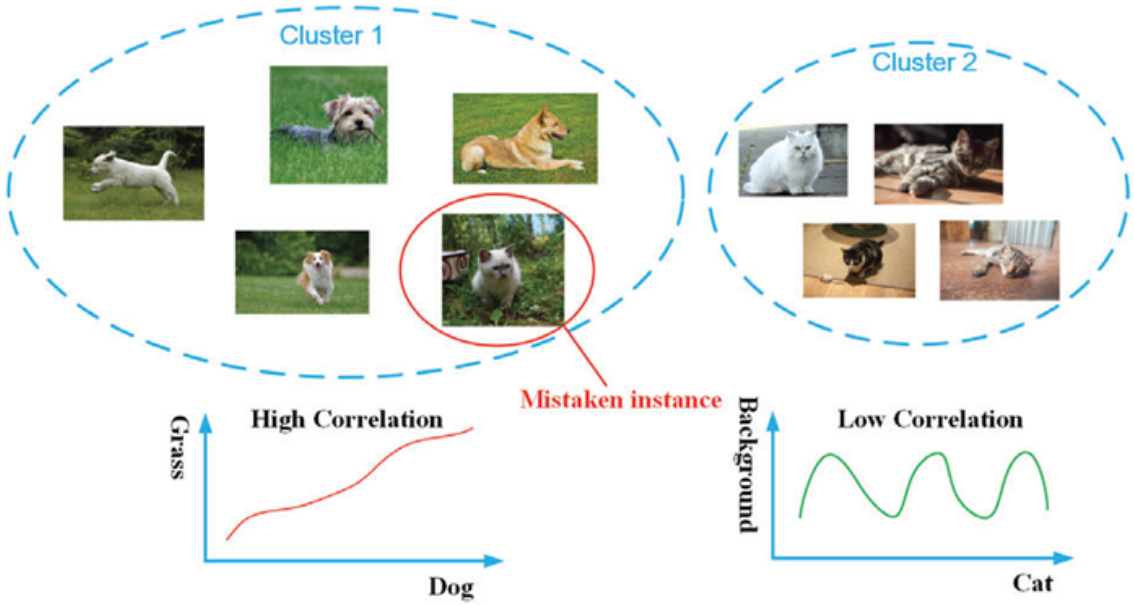
一种数据选择偏差下的去相关聚类方法

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

本博文是对Decorrelated clustering with data selection bias这篇文章的展开与叙述。现有的聚类算法大多没有考虑数据的选择偏差。然而，在许多实际应用中，人们不能保证数据是无偏的。选择偏差可能会导致特征之间产生意想不到的相关性，忽略这些意想不到的相关性会影响聚类算法的性能。因此，如何消除这些由选择偏差引起的非预期相关性是非常重要的，但在聚类过程中还没有被深入探讨。在本文中，提出了一种新的去相关正则化k -均值算法(DCKM)，用于有数据选择偏差的聚类。具体来说，去相关正则化器的目的是学习能够平衡样本分布的全局样本权值，从而消除特征之间的非预期相关性。同时，将学习到的权值与k-means相结合，使重新加权后的k-means聚类对数据的固有分布没有非预期的相关性影响。此外，本文还推导出了更新规则，以有效地推断DCKM中的参数。在真实数据集上的大量实验结果很好地证明了DCKM算法获得了显著的性能提升，表明在聚类时需要去除由选择偏差引起的非预期特征关联。

➢背景

- 通常在图像聚类过程中，数据选择偏差在现实情况中普遍存在。而数据选择偏差可能会导致图像特征之间的虚假关联。假设将一个虚假的特征被错误地识别为与一个重要的特征相关联，由于存在虚假相关，这种无意义的特征的作用将会在图像聚类中得到增强，从而使固有的数据分布无法得到揭示。
- 比如右图，在收集图像数据时，前景为目标对象的图片，如狗的图片和猫的图片，然而目标对象的背景却不相同，比如狗在草地上的图片比较多，而猫在草地上的图片比较少。因为这种偏差的存在，会造成狗的特征和草的特征的相关。如果在这样有偏差的数据上聚类，可能会使在草地上的猫的图片错分到狗的类中。
- 因此，对这些有偏差的数据进行聚类，会导致聚类效果下降。



Xiao Wang, Shaohua Fan, Kuang Kun, Chuan Shi, Jiawei Liu, Bai Wang. Decorrelated clustering with data selection bias. IJCAI 2020. (CCF-A)

➢ Decorrelation Regularized K-means

本文提出一种数据选择偏差下的去相关聚类方法，其中，方法包括：

- 获取存在偏差的多张图像，作为样本集；
- 基于样本集，联合优化加权后聚类算法和去相关正则项，得到最优加权后聚类算法，其中，最优加权后聚类算法是通过多次计算加权后聚类算法得到的，加权后聚类算法是通过使用去相关正则项学习得到的各样本权重，对聚类算法进行加权得到的；
- 各样本权重为通过使用去相关正则项，对样本集中的各图像，学习本次各样本权重；
- 通过在本次加权后聚类算法中包含的本次聚类中心和簇不是首次聚类中心和簇，并且本次聚类中心和簇与上次聚类中心和簇之间的差异小于阈值时，得到最优加权后聚类算法，以确定图像不受偏差影响的聚类中心和簇。

使用样本权重对聚类算法进行加权，得到加权后聚类算法，总体目标函数为：

$$\min_{\mathbf{w},\mathbf{F},\mathbf{G}} \sum_{i=1}^{n} \mathbf{w}_i \cdot ||\mathbf{X}_{i.} - \mathbf{G}_{i.} \cdot \mathbf{F}^T||_2^2,$$

加权的K-means

$$s.t. \quad \sum_{j=1}^{d} || \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w} \odot \mathbf{X}_{\cdot j})}{\mathbf{w}^T \cdot \mathbf{X}_{\cdot j}} - \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w} \odot (\mathbf{1}_n - \mathbf{X}_{\cdot j}))}{\mathbf{w}^T \cdot (\mathbf{1}_n - \mathbf{X}_{\cdot j})} ||_2^2 \leq \gamma_1,$$

去相关正则项

$$\mathbf{G}_{ik} \in \{0,1\}, \sum_{k=1}^{K} \mathbf{G}_{ik} = 1,$$

簇划分矩阵约束项

$$\mathbf{w} \succeq 0, ||\mathbf{w}||_2^2 \leq \gamma_2, (\sum_{i=1}^{n} \mathbf{w}_i - 1)^2 \leq \gamma_3.$$

权重约束项

Xiao Wang, Shaohua Fan, Kuang Kun, Chuan Shi, Jiawei Liu, Bai Wang. Decorrelated clustering with data selection bias. IJCAI 2020. (CCF-A)

➢ K-means and matrix factorization

众所周知，K-means是最具代表性的聚类算法之一。因此，为了验证去相关聚类的必要性，本文重点研究了K-means算法，并提出了一种新的去相关正则化的k-means算法。K-means将数据空间划分为一种称为Voronoi图的结构。此外，G正交非负矩阵分解(NMF)等价于松弛K-means聚类，即

$$\min_{\mathbf{F},\mathbf{G}} \sum_{i=1}^{n} ||\mathbf{X}_{i.} - \mathbf{G}_{i.} \cdot \mathbf{F}^T||_2^2,$$

$$s.t. \quad \mathbf{G}_{ik} \in \{0,1\}, \sum_{k=1}^{K} \mathbf{G}_{ik} = 1, \forall i = 1, 2, \cdots, n,$$

where $\mathbf{F} \in \mathbb{R}^{d \times K}$ is the cluster centroid matrix, $\mathbf{G} \in \mathbb{R}^{n \times K}$ is the cluster assignment matrix, each row of which satisfies the *1-of-K* coding scheme, i.e., if data point $\mathbf{X}_{i.}$ is assigned to k-th cluster, then $\mathbf{G}_{ik} = 1$; otherwise, $\mathbf{G}_{ik} = 0$.

➢ 单个特征去相关正则项

基于所有图像中目标特征的取值，将所有图像分为对照组和实验组，对照组和实验组均包含其余特征，对照组中目标特征的取值和实验组中目标特征的取值不同。

- 通过对照组的一阶矩公式，采用各图像的特征权重，确定对照组中各图像的特征几何分布，作为对照组分布，其中，对照组的一阶矩公式为：

$$\bar{\mathbf{X}}_{.-j} = \frac{\mathbf{X}_{.-j}^T \cdot \mathbf{X}_{.j}}{\mathbf{1}_n^T \cdot \mathbf{X}_{.j}}$$

- 通过实验组的一阶矩公式，采用各图像的特征权重，确定实验组中各图像的特征几何分布，作为实验组分布，其中，实验组的一阶矩公式为：

$$\hat{\mathbf{X}}_{.-j} = \frac{\mathbf{X}_{.-j}^T \cdot (\mathbf{1}_n - \mathbf{X}_{.j})}{\mathbf{1}_n^T \cdot (\mathbf{1}_n - \mathbf{X}_{.j})}$$

➢ Single/Global feature decorrelation regularizer

通过如下其余特征平衡项公式，采用平衡项调整各图像的特征权重，以确定对照组分布与实验组分布之间的距离达到最小时对应的最优调整后特征权重，作为样本权重，其中，单个其余特征平衡项公式为：

$$\mathbf{w}^j = \underset{\mathbf{w}^j}{\operatorname{argmin}} \left\| \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w}^j \odot \mathbf{X}_{\cdot j})}{\mathbf{w}^{jT} \cdot \mathbf{X}_{\cdot j}} - \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w}^j \odot (\mathbf{1}_n - \mathbf{X}_{\cdot j}))}{\mathbf{w}^{jT} \cdot (\mathbf{1}_n - \mathbf{X}_{\cdot j})} \right\|_2^2$$

样本权重

调整的对照组分布

调整的实验组分布

所有其余特征平衡项公式为：

$$\sum_{j=1}^{d} \left\| \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w} \odot \mathbf{X}_{\cdot j})}{\mathbf{w}^T \cdot \mathbf{X}_{\cdot j}} - \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w} \odot (\mathbf{1}_n - \mathbf{X}_{\cdot j}))}{\mathbf{w}^T \cdot (\mathbf{1}_n - \mathbf{X}_{\cdot j})} \right\|_2^2$$

全局样本权重

Xiao Wang, Shaohua Fan, Kuang Kun, Chuan Shi, Jiawei Liu, Bai Wang. Decorrelated clustering with data selection bias. IJCAI 2020. (CCF-A)

Decorrelated Clustering with Data Selection Bias

➢ Decorrelation Regularized K-means

在传统的k-means模型中，聚类质心F和聚类分配G是在原始特征x上学习的，但未预料到的高度相关的特征可能会混淆数据分布，导致聚类结果不满意。由于从去相关正则化器学习到的样本权值w能够对特征进行全局去相关，因此提出利用这些权值对K-means损失进行重新加权，并对加权的K-means损失和去相关正则化器进行联合优化。

$$\min_{\mathbf{w},\mathbf{F},\mathbf{G}} \sum_{i=1}^{n} \mathbf{w}_i \cdot ||\mathbf{X}_{i.} - \mathbf{G}_{i.} \cdot \mathbf{F}^T||_2^2,$$

加权的K-means

- 使用本次各样本权重对聚类算法进行加权，得到本次加权后聚类算法
- F为样本集的聚类的簇中心矩阵

$$s.t. \quad \sum_{j=1}^{d} || \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w} \odot \mathbf{X}_{\cdot j})}{\mathbf{w}^T \cdot \mathbf{X}_{\cdot j}} - \frac{\mathbf{X}_{\cdot -j}^T \cdot (\mathbf{w} \odot (\mathbf{1}_n - \mathbf{X}_{\cdot j}))}{\mathbf{w}^T \cdot (\mathbf{1}_n - \mathbf{X}_{\cdot j})} ||_2^2 \leq \gamma_1,$$

去相关正则项

- 用于约束去相关正则项的大小

$$\mathbf{G}_{ik} \in \{0, 1\}, \sum_{k=1}^{K} \mathbf{G}_{ik} = 1,$$

簇划分矩阵约束项

- Gik为第i个样本是否属于第k个簇
- 将样本集中样本Xi .分配给第k个簇，Gik＝1
- 将样本集中样本Xi .未分配给第k个簇，Gik＝0
- k为第k个簇，k为簇的序号，K为聚类总数

$$\mathbf{w} \succeq 0, ||\mathbf{w}||_2^2 \leq \gamma_2, (\sum_{i=1}^{n} \mathbf{w}_i - 1)^2 \leq \gamma_3$$

权重约束项

- 第一项为每个样本权重约束为非负数
- 第二项用于减少样本权重的方差以实现稳定性
- 第三项为所有样本中存在非0的权重，用于避免所有样本权重均为0

Xiao Wang, Shaohua Fan, Kuang Kun, Chuan Shi, Jiawei Liu, Bai Wang. Decorrelated clustering with data selection bias. IJCAI 2020. (CCF-A)

参考文献：

[1] Xiao Wang, Shaohua Fan, Kuang Kun, Chuan Shi, Jiawei Liu, Bai Wang. Decorrelated clustering with data selection bias. IJCAI 2020. (CCF-A)

[2] 王啸，石川，范少华. 一种数据选择偏差下的去相关聚类方法及装置[发明专利], 申请号: 2020105917421.

王啸老师个人主页：https://wangxiaocs.github.io/