噪声标签—学习报告

袁荣

* 收获与总结

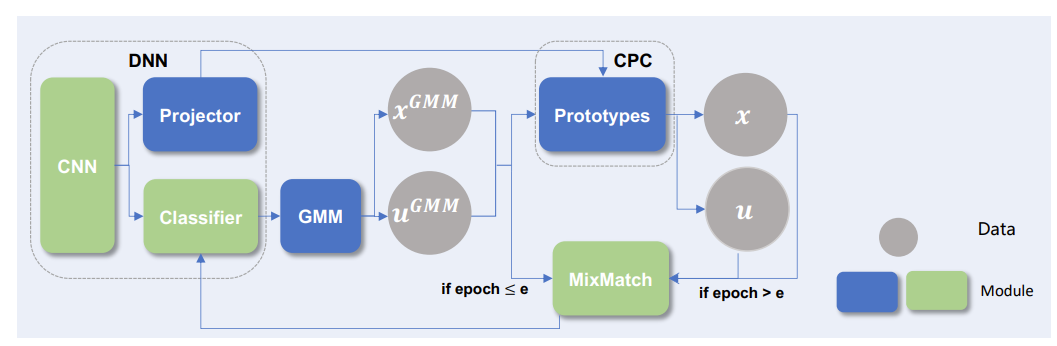
1. 收获：

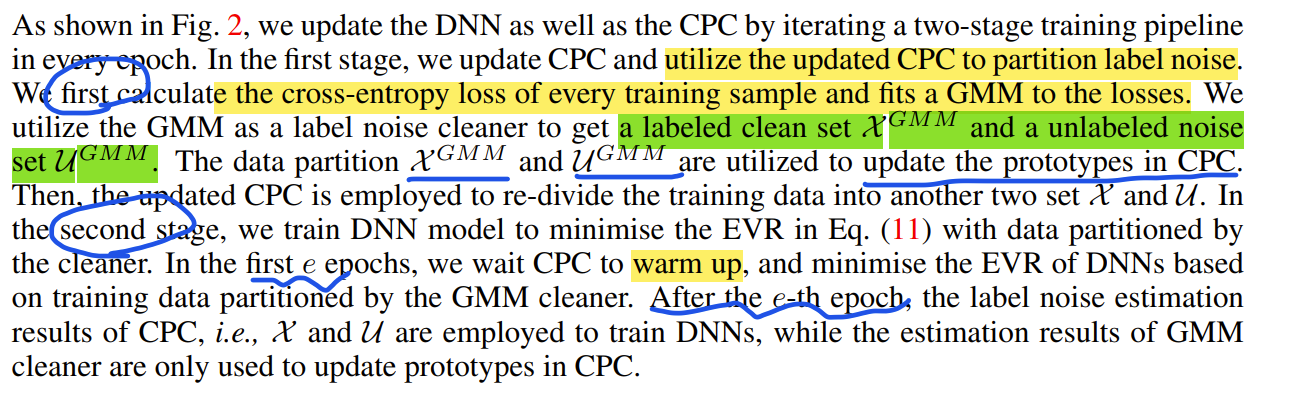
第一遍：摘要 + 结论 + 图 + 表

体会：以往的噪声标签学习往往都是通过一个由无监督学习训练出来的label cleaner（其实个人感觉就是分类器）对训练样本进行分类，具体划分为：纯净的且有标签的集合以及混淆的且没有标签的集合。但由于label cleaner的训练源于每个样本的training loss，如果此时训练数据集取自不同的类（即分布不同），比如一半来自于动物，一半来自于植物，这就将会导致label cleaner的划分标准出现差错。鉴于此，文章提出Class Prototype – Based Cleaner的解决方法，其最大的特点是更加关注数据集的分布情况，能更好地对训练样本进行分类。最后，文章通过在CIFAR-10等数据集上做实验，得出比以往SOTA（SOTA实际上就是State of the arts 的缩写，指的是在某一个领域做的Performance最好的model）解决方法效果更好的结论。

第二遍：算法 + 实验

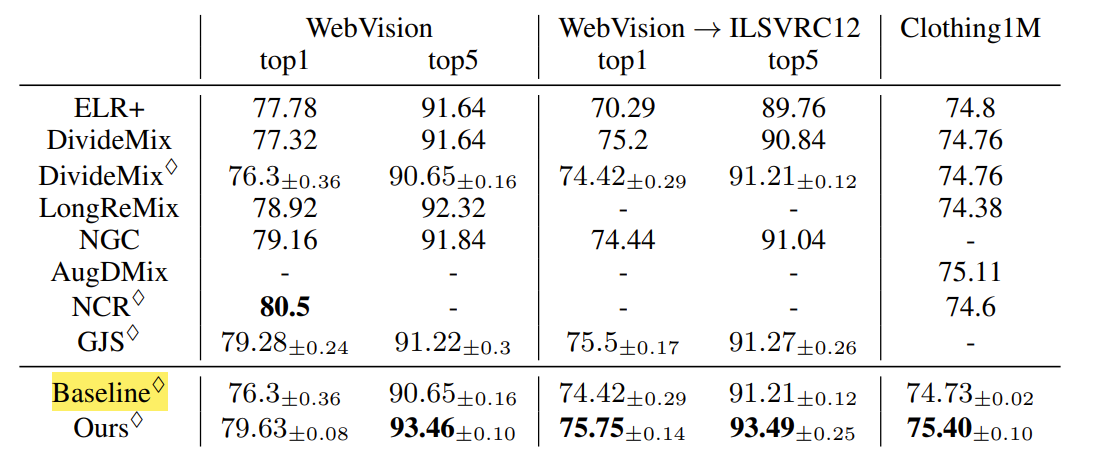
体会：算法部分重点讲解了以下framework，能完全明白各个阶段训练的过程。





接下来就是具体讲解class\_prototype\_based cleaner的提出过程，这一部分从理论基础（数学本质）上推导了cleaner所利用的损失函数同时证明其有效性，但由于表达式复杂，感觉看得云里雾里的，很头大。

实验部分首先介绍了Clothing1M和WebVision1.0两个数据集，接着分析在DivideMix框架下采用ResNet系列算法对数据集运算所得到的结果，但由于了解较少，表格中的数据（如ELR+等等）不知道是什么，还比如高亮部分（Baseline）是取自哪个？最后便是CPC与各种算法在真实世界噪声与人工合成噪声领域的对比分析。

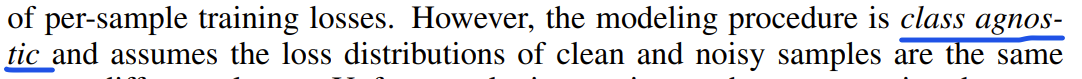


在Analysis中，为分析CPC是否为一种噪声标签学习的好方法，文章对class-agnostic GMM、naive class-aware GMM，以及由他们各自训练出来的CPC进行了横向与纵向分析，并成功验证了CPC的提出是高效可行的。然后再保持训练框架相同的情况下，文章将CPC与其余cleaner进行了比较，仍然得出CPC\_awr效果更好的结论。最后为考察CPC学习到的prototype质量如何，计算了预测值与真实值的距离（感觉用相似度解释会更好），然后选取距离最小的作为结果与MoPro比较，实验发现CPC学习到的prototype准确度更好。

第三遍：提出问题 + 解决问题

* 存在的问题

1. 划线部分没能完全理解，还有例如embedding也不是很清楚，甚至标题的prototype都不太明白该怎么理解



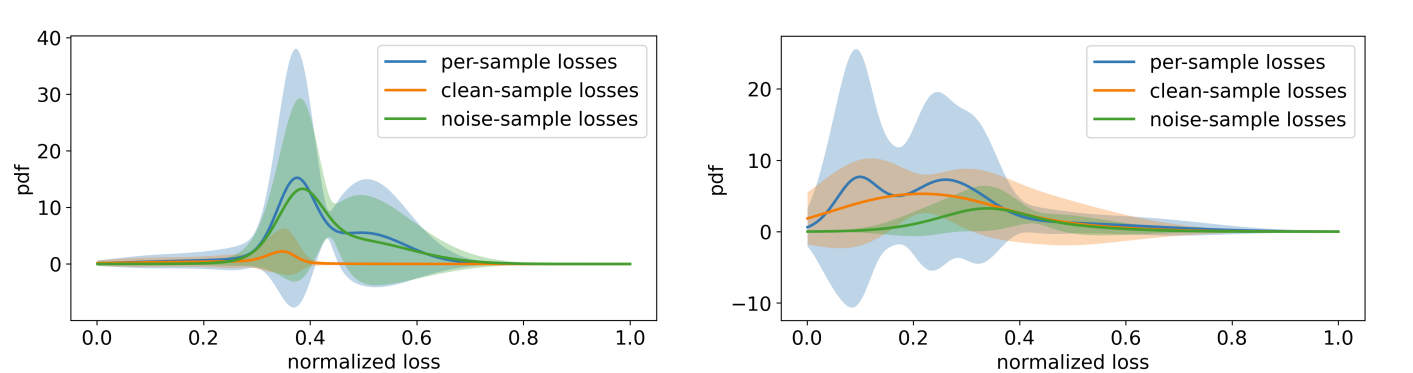
--------------------------------------------------------------------------------------------

Class agnostic 可以理解为对于样本数据中类的划分比较模糊

Prototype 类原型，例如狗有不同品种，但整个大类就是狗

Embedding （嵌入） 表示一种降维的映射

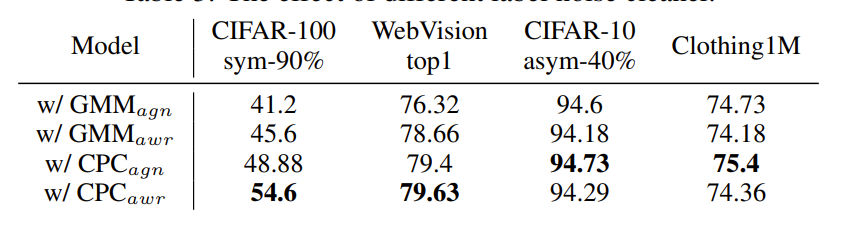
1. 图的包络线不明白是什么意思



--------------------------------------------------------------------------------------------

曲线表示所有类别的平均概率密度，而轮廓部分（包络线）表示其95%的置信区间

1. 表格中具体内容（除了CIFAR系列数据集之外）都不是很能看懂，例如：



--------------------------------------------------------------------------------------------

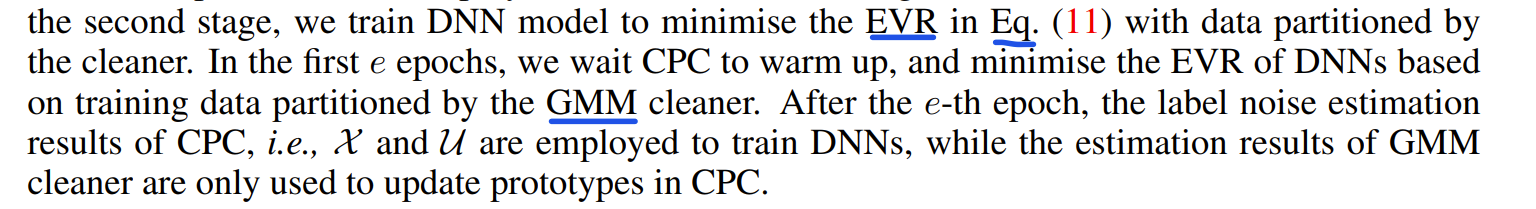
GMMagn --- original class-agnostic GMM cleaner

GMMawr --- naive class-aware GMM cleaner

CPCagn --- trained based on conventional class-agnostic GMM cleaner

CPCawr --- trained based on class-aware GMM cleaner

1. 大多专业术语或者其对应的缩写词汇没了解过（例如GMM，EVQ）



------------------------------------------------------------------------------------------------

GMM --- Gaussian Mixture Model

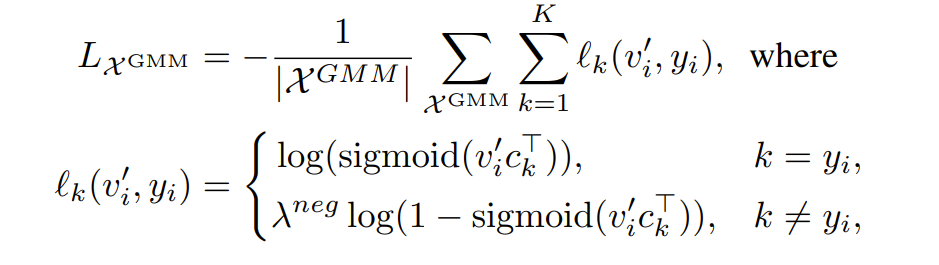
EVR ---empirical vicinal risk 经验邻近风险

PS：

经验风险（Empirical risk，ER）

邻域风险（Vicinal Risk，VR）

1. “经验风险最小化”即“训练误差最小化”。
2. 关于“泛化性”，通常可以通过使用大规模训练数据来提高，但是实际上，获取有标签的大规模数据需要耗费巨大的人工成本，甚至有些情况下根本无法获取数据。解决这个问题的一个有效途径是“邻域风险最小化”，即通过先验知识构造训练样本的邻域值。一般的做法就是传统的数据增强方法，比如加噪、翻转、缩放等，但是这种做法很依赖于特定的数据集和人类的先验知识。
3. 论文提到的数学表达式具体不知道是怎么推导出来的，例如：



------------------------------------------------------------------------------------------------

数学公式推导取自相关参考文献