计算机应用数学（上）
利用无标签数据增强机器学习模型

XXX

September 27, 2019

Abstract

随着大数据时代的到来，机器学习日渐展现自己的威力。分类问题是机器学习中的经典问题，在大数据时代也得到了广泛应用。但是随着数据量的增大，数据标注的成本越来越高昂，使得监督模型的训练只能在少量标注样本中进行。因此，学界针对利用无标签数据增强监督学习模型进行了广泛而深入的研究。

1 引文

机器学习，与计算统计联系非常紧密，是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的学科。机器学习研究的对象是数据，依据数据是否有标注，机器学习可分为监督学习、半监督学习和无监督学习等。

分类任务是机器学习中的一个子任务，在垃圾邮件检测、网络入侵检测、光学字符识别及计算机视觉领域得到了广泛应用。二分类是分类任务的一个子类，即给定样本，判断该样本属于正类（Positive Class）和负类（Negative Class）的哪一类。常见的二分类机器学习算法有感知机算法、逻辑斯蒂回归及支持向量机等。最近几年，在计算机视觉分类任务中，卷积神经网络异军突起，分类性能十分优越。

常见的监督学习算法，要求样本标注，通常监督学习算法的性能随着数据量的增加能够有所提升。然而随着大数据时代的到来，数据量爆发式增长，样本标注的代价随之变得非常高昂，因而监督学习的模式只能利用少量样本的标注数据。为了利用无标签数据提升监督学习算法的性能，学界对此进行了广泛而深入的研究。

利用无标签数据增强机器学习模型的方法，即半监督学习，面对的数据只有部分有标注。其中，若有标注的样本仅有正例样本，则又称为Positive and Unlabeled Learning，下文称PU Learning。PU Learning的问题比常见的半监督问题更困难一些，因此其算法可对应地扩展到半监督学习上。本文针对PU Learning等半监督学习的最新进展进行综述。

2 基础概念

半监督学习与监督学习类似，它的任务是学习一个模型，使模型能够对任意给定的输入，对其相应的做出一个好的预测。不同的是，在半监督学习中，给定的训练数据集$T = \{(x_1, y_1), \ldots, (x_k, y_k), x_{k+1}, \ldots, x_n\}$中只有部分数据有标注，因此在半监督学习的模型训练时，损失函数需要考虑到无标签数据的损失估计。若损失函数不考虑无标签数据，即无标签数据不参与模型训练，则半监督问题转化为少量样本数据的监督学习问题。

机器学习模型的假设空间包含所有可能的条件概率分布或决策函数，假设空间的模型一般有无穷多个。假设空间用F表示，X和Y分别是输入空间和输出空间的变量，则此时F为

$$F = \{f|Y = f_\theta(X), \theta \in \mathrm{R}^n\} \tag{2.1}$$

参数向量θ属于n维欧式空间R^n，称为参数空间。监督学习在假设空间F中选取模型f作为决策函数，用损失函数度量预测错误的程度，记作$L(Y, f(X))$。由于(X, Y)是随机变量，遵循联合分布$P(X, Y)$，所以损失函数的期望是

$$R_{exp}(f) = E_P[L(Y, f(X))] = \int_{X \times Y} L(y, f(x))P(x, y)\mathrm{d}x\mathrm{d}y \tag{2.2}$$

这是理论上的平均损失，称为风险函数或期望损失。然而联合分布$P(X, Y)$未知，因此风险函数不可计算。在监督学习中，给定数据集$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，可以计算模型

关于训练数据集的平均损失2.3，称为经验风险。

$$R_{emp} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \tag{2.3}$$

根据大数定律，当样本容量N趋于无穷时，经验风险$R_{emp}(f)$趋于期望风险$R_{exp}(f)$。因此，在实践中，通常用带策略的经验风险来估计期望风险。

而在半监督学习中，无标签数据因为没有标签，不可以直接估计它们的经验风险。考虑到机器学习模型的性能能够随着数据的增加而得到提升，为了利用无标签数据提升机器学习模型的性能，学者们研究了各种方法将无标签数据一起加入训练。

在[1, 2] 中，Zhou和Lee 用少量样本训练的分类器预测无标签数据的标签从而得到伪标签，将其作为真实标签加入训练集迭代训练。Kiryo在[3] 研究了PU Learning 问题下根据无标签数据估计负例样本损失的方法，该方法需要先估计正例样本占全体样本的比例，再利用经验风险公式推导得出无标签数据估计负例样本损失的算法。Xie 在[4] 中展示了AUC优化的结果：线性分类器下，针对AUC优化的损失函数不需要获得无标签数据的标签。Dong将目光转向了多标签问题，通过考察实例（instance）和标签（label）的相关性，在[5] 中，将缺失标签的预测形式化为优化问题。Zhang在[6]中提出了一个简单的想法：在多分类问题中，让无标签数据的预测结果不偏向于任何一类。尽管这个想法没什么新意，Zhang在[6]中给出了实验和理论两方面的证明，该方法在监督任务上有正则的效果，对模型进行了约束，有利于减小泛化误差。

3 半监督学习的分类算法

3.1 为无标签数据赋予伪标签

半监督学习相较于监督学习的难点在于，如何处理无标签数据。利用少量有标注的样本训练一个可用的分类器，然后预测无标签数据的标签，再将其加入训练，这是一个自然的想法。然而，如果得到的分类器不足够好，预测的伪标签会带来噪声，反而会降低分类器的性能。Zhou[1]提出了tri-training的算法，其不要求充分且冗余的视图，可以简便的应用于常见的数据挖掘场景。Tri-training利用三个独立的分类器为无标签数据标注和输出最终的预测结果，只要遵从多数分类器的选择即可。

用L标记样本容量为$|L|$的有标注数据集，U标记样本容量为$|U|$的无标注数据集。在co-training模式下，每个分类器标注的置信度都需要测量出来。而在tri-training中，额外有两个分类器，比如，h_1和h_3，一个分类器h_3利用L进行初始训练。接着，对任一个分类器，如果另外两个分类器对某无标签样本的标注一致，则该样本的标注可以用来训练该分类器，而无需测量分类器的标注置信度。例如，无标注数据集U中的样本x，在h_2和h_3下的标注一致，则x的标注可以用于h_1的训练。容易看出，如果h_2和h_3在x上的预测正确，那么h_1可得到一个合理的新的样本，否则h_1会得到一个有噪声标签的样本。但是，即使在最差的情况下，只要新标注样本充分多，分类结果噪声率的增加也是可以容忍的。

在tri-training的每轮训练中，分类器h_2和h_3选取U中的某些样例标注用于h_1的训练。因为分类器在tri-training的训练中不断精炼，每一轮选取的无标注样本数量可能不相同。用L^t和L^{t-1}分别标记在t和$t-1$轮给h_1提供标注的样本集。那么，在t和$t-1$轮用于h_1的训练集分别是$L \cup L^t$和$L \cup L^{t-1}$，样本容量分别是m^t和m^{t-1}。

记η_L为L的分类噪声比例，即L中的无标注的样本数量为$\eta_L|L|$。记$\hat{e}_1^t$为第t轮中$h_2 \& h_3$的误分类率的上界。于是，在L^t中误分类的样本数量为$\hat{e}_1^t|L^t|$。因此，t轮分类噪声率为

$$\eta^t = \frac{\eta_L|L| + \hat{e}_1^t|L^t|}{|L \cup L^t|}. \tag{3.1}$$

此时，记$u^t = m^t(1-2\eta^t)^2$为一常数，若$u^t > u^{t-1}$，则$\epsilon^t < \epsilon^{t-1}$。代入可得

$$u^t > u^{t-1} \Leftrightarrow |L \cup L^t|(1 - 2\frac{\eta_L|L| + \hat{e}_1^t|L^t|}{L \cup L^t})^2 > |L \cup L^{t-1}|(1 - 2\frac{\eta_L|L| + \hat{e}_1^{t-1}|L^{t-1}|}{L \cup L^{t-1}})^2. \tag{3.2}$$

考虑到η_L通常非常小，假定$0 \le \hat{e}_1^t, \hat{e}_1^{t-1} < 0.5$。若$|L^{t-1}| < |L^t|$，则3.2左式的第一项比右式的对应项大。若$\hat{e}_1^t|L^t| < \hat{e}_1^{t-1}|L^{t-1}|$，则左式的第二项比右式的对应项大。这些约束如3.3，用于决定一个无标注样本是否应标注参与训练。

$$0 < \frac{\hat{e}_1^t}{\hat{e}_1^{t-1}} < \frac{|L^{t-1}|}{|L^t|} < 1. \tag{3.3}$$

相比于Zhou[1]，Lee[2]的伪标签法更为经验性，其给无标注数据赋予伪标签，即选择预测的最大概率的类别作为它的真实类别，这种方法的效果和熵正则化相等。这种伪标签方法适合类

间以低密度分隔的数据集。除此之外，Lee[2]的方法还使用了去噪的自编码器和Dropout技术。去噪的自编码器的基本想法是自编码器学习到的表征应对输入模式的部分污损时应更健壮。

$$h_i = s(\sum_{j=1}^{d_v} W_{ij}\tilde{x}_j + b_i) \tag{3.4}$$

$$\hat{x}_j = s(\sum_{i=1}^{d_h} W_{ij}h_i + a_j) \tag{3.5}$$

其中$\tilde{x_j}$是第j项输入的污损版，$\hat{x_j}$是重建后的第j项值。自编码器通过最小化x_j和$\hat{x_j}$的重建损失进行表征学习。Dropout是应用于深度神经网络监督学习的一种技术。在每个样本通过网络激活时，隐藏节点以0.5的概率随机发射。

$$h_i^k = drop(s^k(\sum_{j=1}^{d^k} W_{ij}^k h_j^{k-1} + b_i^k)), k = 1, \cdots, M \tag{3.6}$$

其中以0.5的概率$drop(x) = 0$，否则$drop(x) = x$。这项技术，在每次权重更新时训练了一个子模型，训练过程类似于bagging，可以减小神经网络对训练数据的过拟合。

伪标签技术将无标签数据的目标类别当作真实的标签，Lee[2]选取最大预测概率的类作为每个无标签样本的类别，并在fine-tuning阶段配合Dropout使用伪标签。在fine-tuning阶段，经过预训练的网络以监督学习的方式，同时利用有标注和无标注数据参与训练。但是因为有标注和无标注数据的数量相差较大，损失函数中两项的损失需要加以权衡，否则会降低网络的性能。3.8中$\alpha(t)$ 就是调节有标注样本和伪标签样本损失的函数，n是mini-batch中有标签样本的数量，n'代表无标签样本。$\alpha(t)$的合理调节对网络的性能非常重要，如果过高，甚至会扰乱有标注样本的训练。然而$\alpha(t)$ 过小的时候，就难以利用无标注样本增强训练效果。Lee[2]采用了确定性的退火过程，$\alpha(t)$缓慢地增大，期望以此帮助优化过程避免较差的局部最小。

$$y_i' = \begin{cases} 1 & \text{if } i = \text{argmax}_{i'} f_{i'}(x) \\ 0 & \text{otherwise} \end{cases}. \tag{3.7}$$

$$L = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C} L(y_i^m, f_i^m) + \alpha(t)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C} L({y'}_i^m, {f'}_i^m) \tag{3.8}$$

3.2 对经验风险的估计

在2.3中，我们用经验风险估计期望风险，根据大数定律，当样本容量N趋于无穷时，经验风险$R_{emp}(f)$趋于期望风险$R_{exp}(f)$。而无标签数据因为没有标签，不能采用2.3的形式进行经验风险估计，因此Xie[4]和Kiryo[3]分别从AUC优化和PU Learning的角度，提出了利用无标签数据更好地估计经验风险的算法。

在监督学习中，正例、负例和普通样本记作

$$\begin{aligned} X_P &:= x_i{}_{i=1}^{n_P} \overset{i.i.d}{\sim} p_P(x) := p(x|y=+1), \\ X_N &:= x'_j{}_{j=1}^{n_N} \overset{i.i.d}{\sim} p_N(x) := p(x|y=-1), \\ X &:= x_k{}_{k=1}^{n_P+n_N} \overset{i.i.d}{\sim} p(x). \end{aligned} \tag{3.9}$$

Xie[4]假定使用线性模型$f(x) = w^\top x$，非线性的部分则需对输入空间做非线性特征变换。AUC优化的目标函数具有特殊性，等价于随机取正例样本，输出的预测结果排在随机取的负例样本前面的概率，它可以形式化为

$$\text{AUC} = 1 - \mathop{\mathbb{E}}_{x\in X_P}[\mathop{\mathbb{E}}_{x'\in X_N}[\ell_{01}(w^\top(x-x'))]]. \tag{3.10}$$

最大化AUC等价于最小化下式的AUC风险。为了避免混淆，记监督学习的AUC风险为PN-AUC风险。

$$R_{PN} = \mathop{\mathbb{E}}_{x\in X_P}[\mathop{\mathbb{E}}_{x'\in X_N}[\ell_{01}(w^\top(x-x'))]]. \tag{3.11}$$

在半监督学习的设定中，我们可以拿到无标签数据集。无标签的实例从正例和负例的混合分布中采样得到:

$$X_U := \{x''_k\}_{k=1}^{n_U} \overset{i.i.d}{\sim} p(x) = \theta_P p_P(x) + \theta_N p_N(x), \tag{3.12}$$

其中θ_P和θ_N是正类和负类的先验概率。Xie[4]证明，优化将无标签数据当作负类，同正例数据计算得到的AUC风险，与优化一个无偏的AUC风险等价。

定理 3.1 PU-AUC风险R_{PU}，以无标签数据为负类，计算其和正类数据的AUC风险和NU-AUC风险R_{NU}，即以无标签数据为正类，计算其和负类数据的AUC风险，等价于带有一个线性变换的监督形式的PN-AUC风险，其中，R_{PU}和R_{NU}分别定义为：

$$R_{PU} = \mathop{\mathbb{E}}_{x\in X_P}[\mathop{\mathbb{E}}_{x''\in X_U}[\ell_{01}(w^\top(x-x''))]], \tag{3.13}$$

$$R_{NU} = \mathop{\mathbb{E}}_{x''\in X_U}[\mathop{\mathbb{E}}_{x'\in X_N}[\ell_{01}(w^\top(x''-x'))]]. \tag{3.14}$$

根据期望的线性性质，可以得到：

$$\begin{aligned} R_{PU} &= \mathop{\mathbb{E}}_{x\in X_P}[\mathop{\mathbb{E}}_{x''\in X_U}[\ell_{01}(w^\top(x-x''))]] \\ &= \mathop{\mathbb{E}}_{x\in X_P}[\theta_P \mathop{\mathbb{E}}_{\bar{x}\in X_P}[\ell_{01}(w^\top(x-\bar{x}))] + \theta_N \mathop{\mathbb{E}}_{x'\in X_N}[\ell_{01}(w^\top(x-x'))]] \\ &= \frac{1}{2}\theta_P + \theta_N \mathop{\mathbb{E}}_{x\in X_P}[\mathop{\mathbb{E}}_{x'\in X_N}[\ell_{01}(w^\top(x-x'))]]. \end{aligned}$$

注意到在$X_P \times X_P$上成对的期望风险是对称的，因此它是个常数。于是

$$R_{PU} = \theta_N R_{PN} + \frac{1}{2}\theta_P. \tag{3.15}$$

类似地，可以得到：

$$R_{NU} = \theta_P R_{PN} + \frac{1}{2}\theta_N. \tag{3.16}$$

综上，R_{PU}和R_{NU}等价于带了一个线性变换的监督形式的AUC风险得到证明。因为$\theta_P + \theta_N = 1$，综合3.15和3.16，得到

$$R_{PU} + R_{NU} - \frac{1}{2} = R_{PN} \tag{3.17}$$

3.17表明，只要在有无标签数据的同时，也有正例和负例数据，那么不需要知道类别的先验概率θ_P和θ_N，就可以计算出对无偏的AUC风险R_{PN}的估计。

Xie[4]根据3.17设计了两个算法：SAMULT和SAMPURA。SAMULT基于3.17将无标签数据分别看作正例数据和负例数据计算经验风险，然后合并两项得到无偏的AUC风险估计。不仅如此，既然不需要估计整个数据集的标签，那么可以从无标签数据集中自助采样若干个数据集，依据3.17训练多个分类器，就可以得到SAMULT的集成版本SAMPURA。因为两种算法非常相似，本文只介绍SAMULT算法。

基于3.17，可以得出如下的半监督形式的AUC优化问题，同时利用有标签数据和无标签数据：

$$\hat{R}_{PNU} = \gamma\hat{R}_{PN} + (1-\gamma)(\hat{R}_{PU} + \hat{R}_{NU} - \frac{1}{2}), \tag{3.18}$$

其中$\gamma \in [0,1]$是一个平衡两项风险的参数，并且

$$\hat{R}_{PN} = \frac{1}{n_P n_N}\sum_{x\in X_P}\sum_{x'\in X_N}\ell(w^\top(x-x')),$$

$$\hat{R}_{PU} = \frac{1}{n_P n_U}\sum_{x\in X_P}\sum_{x''\in X_U}\ell(w^\top(x-x'')),$$

$$\hat{R}_{NU} = \frac{1}{n_N n_U}\sum_{x\in X_N}\sum_{x''\in X_U}\ell(w^\top(x''-x')).$$

因为0-1损失为离散值且难以优化，实践中用平方损失$\ell(z) = (1-z)^2$代替0-1损失。Gao和Zhou[7]证明了平方损失与AUC渐近一致。为了减少过拟合的风险，在风险估计项上再加上ℓ_2的正则化项：

$$\min_w \hat{R}_{PNU}(w) + \lambda||w||^2, \tag{3.19}$$

其中λ是调节风险估计和正则项的参数。实践中，3.18中的常数项省略后不改变优化的问题。可以得到优化问题3.19的解析解：

$$\hat{w} = (\gamma H_{PN} + (1-\gamma)(H_{PU} + H_{NU}) + \lambda I_d)^{-1}(\gamma h_{PN} + (1-\gamma)(h_{PU} + h_{NU})), \tag{3.20}$$

其中

$$
\begin{aligned}
h_{PN} &= \frac{1}{n_P} X_P^\top 1_{n_P} - \frac{1}{n_N} X_N^\top 1_{n_N}, \\
h_{PU} &= \frac{1}{n_P} X_P^\top 1_{n_P} - \frac{1}{n_U} X_U^\top 1_{n_U}, \\
h_{NU} &= \frac{1}{n_U} X_U^\top 1_{n_U} - \frac{1}{n_N} X_N^\top 1_{n_N}, \\
H_{PN} &= \frac{1}{n_P} X_P^\top X_P - \frac{1}{n_P n_N} X_P^\top 1_{n_P} 1_{n_N}^\top X_N - \frac{1}{n_P n_N} X_N^\top 1_{n_N} 1_{n_P}^\top X_P + \frac{1}{n_N} X_N^\top X_N, \\
H_{PU} &= \frac{1}{n_P} X_P^\top X_P - \frac{1}{n_P n_U} X_P^\top 1_{n_P} 1_{n_U}^\top X_U - \frac{1}{n_P n_U} X_U^\top 1_{n_U} 1_{n_P}^\top X_P + \frac{1}{n_U} X_U^\top X_U, \\
H_{NU} &= \frac{1}{n_U} X_U^\top X_U - \frac{1}{n_U n_N} X_U^\top 1_{n_U} 1_{n_N}^\top X_N - \frac{1}{n_U n_N} X_N^\top 1_{n_N} 1_{n_U}^\top X_U + \frac{1}{n_N} X_N^\top X_N,
\end{aligned}
$$

X_P，X_N和X_U分别是正例、负例和无标签的实例矩阵，1_d是d维的全为1的向量，I_d是d维的单位矩阵。

当只能获取到正例和无标签的数据时，3.18中的$\hat{R}_{PN}$和$\hat{R}_{NU}$为0，SAMULT退化为只优化$\hat{R}_{PU}$的特殊形式，这也正是PU Learning所面对的问题。3.21与将无标签数据当作负类的AUC优化目标等价。而Kiryo[3]则从经验风险最小化的角度给出了一个更好的结果，在深度神经网络和卷积神经网络上展现出了PU Learning问题下的当前最佳的性能。

$$
\min_w \hat{R}_{PU}(w) + \lambda ||w||^2 \tag{3.21}
$$

Kiryo[3]提出的non-negative PU Learning基于[8, 9]提出的从仅有正例和无标签的数据中对风险进行无偏估计的算法。问题如3.9所示，正例和无标签数据分别独立的从$p_p(x) = p(x|Y = +1)$和$p(x)$中采样得到，即正例数据记为$X_p = \{x_i^p\}_{i=1}^{n_p} \sim p_p(x)$且无标签数据记为$X_u = \{x_i^u\}_{i=1}^{n_u} \sim p(x)$。记$\pi_P = p(Y = +1)$为正例先验概率，则$\pi_N = p(x|Y = -1) = 1 - \pi_P$。Kiryo[3]假定$\pi_P$已知，因为根据[10]可以从仅有正例和无标签的数据中估计出正例先验概率。

无偏的PU Learning依赖于无偏的风险估计。记$g : \mathbb{R}^d \to \mathbb{R}$为任意决策函数，$\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$为损失函数，那么$\ell(t, y)$代表真实值为$y$时预测输出为$t$带来的损失。记$R_p^+(g) = \mathbb{E}[\ell(g(X), +1)]$，$R_n^-(g) = \mathbb{E}_n[\ell(g(X), -1)]$，其中$\mathbb{E}_p[\cdot] = \mathbb{E}_{X \sim p_p}[\cdot]$且$\mathbb{E}_p[\cdot] = \mathbb{E}_{X \sim p_n}[\cdot]$。那么，$g$的风险就是$R_n^-(g) = \mathbb{E}_{(X,Y) \sim p(x,y)}[\ell(g(X), Y)] = \pi_p R_p^+(g) + \pi_n R_n^-(g)$。在PN Learning（即训练数据既有正例又有负例的学习模式）中，因为X_p和X_n都存在，$R(g)$可以直接以下式近似:

$$
\hat{R}_{pn}(g) = \pi_p \hat{R}_p^+(g) + \pi_n \hat{R}_n^-(g), \tag{3.22}
$$

其中$\hat{R}_p^+(g) = \frac{1}{n_p} \sum_{i=1}^{n_p} \ell(g(x_i^p), +1)$，$\hat{R}_n^-(g) = \frac{1}{n_n} \sum_{i=1}^{n_n} \ell(g(x_i^n), -1)$。在PU Learning中，$X_n$无法获取，但是$R_n^-(g)$可以间接地估计出来。记$R_p^-(g) = \mathbb{E}_p[\ell(g(X), -1)]$，$R_u^-(g) = \mathbb{E}_{X \sim p(x)}[\ell(g(X), -1)]$。因为$\pi_n p_n(x) = p(x) - \pi_p p(x)$，可以得到$\pi_n R_n^-(g) = R_u^-(g) - \pi_p R_p^-(g)$，这样$R(g)$就可以得到间接的估计:

$$
\hat{R}_{pu}(g) = \pi_p \hat{R}_p^+(g) - \pi_p \hat{R}_p^-(g) + \hat{R}_u^-(g), \tag{3.23}
$$

其中$\hat{R}_p^-(g) = \frac{1}{n_p} \sum_{i=1}^{n_p} \ell(g(x_i^p), -1)$，$\hat{R}_u^-(g) = \frac{1}{n_u} \sum_{i=1}^{n_u} \ell(g(x_i^u), -1)$。3.22和3.23估计的经验风险对于所有常用的损失函数都是无偏且一致的。

直觉上，无偏的PU Learning算法得益于变换$\pi_n R_n^-(g) = R_u^-(g) - \pi_p R_p^-(g)$。当我们用$N$项数据$\{x_i^n\}_{i=1}^{n_n}$近似$\pi_n R_n^-(g)$时，收敛的速率是$\mathcal{O}_p(\frac{\pi_n}{\sqrt{n_n}})$，其中$(O)_p$表示概率的阶；当用$P$项数据$\{x_i^p\}_{i=1}^{n_p}$和$U$项数据$\{x_i^u\}_{i=1}^{n_u}$近似$R_u^-(g) - R_p^-(g)$时，收敛得到速率是$\mathcal{O}_p(\frac{\pi_p}{\sqrt{n_p}} + \frac{1}{\sqrt{n_u}})$。当决策函数集$\mathcal{G} = g|||g||_\infty \leq C_g$其中$C_g > 0$是一个常数时，那么对于任意的$n$和$q(x)$，$\mathcal{R}_{n,q}(\mathcal{G}) = \mathcal{O}(1)$，所有的界都变得平凡；但$\mathcal{G}$不满足条件时，$\hat{R}_{pu}(g)$就不满足界，有可能发散到$-\infty$。因此，为了得到高质量的$\hat{g}_{pu}$估计，$\mathcal{G}$不能过于复杂，也就是模型$g$不能非常灵活。

在训练MNIST的多层感知机用于手写数字分类时，Kiryo[3]发现，以uPU指代无偏的PU Learning算法，PN指代PN Learning算法:

(A) 在训练数据上，uPU和PN算法损失都在下降，uPU损失下降的速度比PN更快;

(B) 在测试数据上，PN算法的损失在下降，而uPU开始时候比PN的损失低，但是后期比PN要高。

总结而言，uPU算法的过拟合问题很严重，这也证明了为了获得高质量的$\hat{g}_{pu}$估计，模型g不能过于灵活。

Kiryo[3]在实验中发现，$\hat{R}_{pu}(\hat{g}_{pu})$持续下降直到小于0.但是风险估计是个非负函数，因此小于0是的uPU风险估计需要得到修正，特别地，由于$\hat{R}_p^+(g) \geq 0$恒成立，$\pi_n R_n^-(g) = R_u^-(g) - \pi_p R_p^-(g) \geq 0$在训练时不总是成立的。基于这个观察，Kiryo[3] 提出了非负的风险估计算法:

$$\tilde{R}_{pu}(g) = \pi_p \hat{R}_p^+(g) + max\{0, \hat{R}_u^-(g) - \pi_p \hat{R}_p^-(g)\}. \tag{3.24}$$

大规模的PU算法如1所示，在Kiryo[3]的实验中，non-negative PU Learning的算法在深度神经网络和卷积神经网络上表现性能良好，克服了无偏PU Learning过拟合的问题，达到了PU Learning问题上目前最佳的效果。

Algorithm 1 Large-scale PU Learning based on stochastic optimization

Require:
training data (X_p, X_u);
hyper parameters $0 \leq \beta \leq \pi_p \sup_t \max_y \ell(t, y)$ and $0 \leq \gamma \leq 1$;
Ensure:
model parameter θ for $\hat{g}_{pu}(x;\theta)$ or $\tilde{g}_{pu(x;\theta)}$;
let $\mathcal{A}$ be an external SGD-like stochastic optimization algorithm
while no stopping criterion has been met **do**
  Shuffle (X_p, X_u) into N mini-batches, and denote by (X_p^i, X_u^i) the i-th mini-batch
  for $i = 1$ to N **do**
    if $\hat{R}_u^-(g; X_u^i) - \pi_p \hat{R}_p^-(g; X_p^i) \geq -\beta$ **then**
      Set gradient $\nabla_\theta \hat{R}_{pu}(g; X_p^i, X_u^i)$
      Update θ by $\mathcal{A}$ with its current step size η
    else
      Set gradient $\nabla_\theta(\pi_p \hat{R}_p^-(g; X_p^i) - \hat{R}_u^-(g; X_u^i))$
      Update θ by $\mathcal{A}$ with a discounted step size $\gamma\eta$
    end if
  end for
end while

3.3 多标签的半监督学习

Dong[5]研究的半监督多标签问题，形式上与二分类的半监督问题不同，此处一个实例的多个标签，缺失了部分或者全部。通过考虑缺失的标签在实例和标签上的两种相似性，Dong[5]提出了双凸的优化问题，可以用高效的块坐标下降算法求解。

在原始的监督学习的多标签设定中，我们拿到一个训练数据集$\{(x_i, y_i)\}_{i=1}^m$。每个实例x_i表示一个d维的实值向量。标签y_i表示一个n维的二值标签向量，1表示这个实例属于对应维的概念，否则就是-1.所有的标签由标签空间$Y = \{-1, 1\}^n$组成。也就是，我们有一个实例矩阵$X = [x_1, x_2, \cdots, x_m]'$，每一行表示一个实例，和一个完整的标签矩阵$Y \in -1, 1_{m \times n}$，其中$Y_{ij} = 1$表示第$i$个实例有第$j$个标签，$Y_{ij} = -1$就表示第$i$个实例没有第$j$个标签。记$S_i = \{j | Y_{ij} = 1, j = 1, \cdots, n\}$实例$x_i, \forall i = 1, \cdots, m$的相关标签的全集。

在半监督学习的弱标签学习设定中，我们有相同的实例矩阵X。但是完整的标签矩阵Y无法获得，我们只能拿到标签出现的矩阵$C \in \{0, 1\}^{m \times n}$，其中$C_{ij} = 1$表示第$i$个实例有第$j$个标签（与$Y_{ij} = 1$相同），而$C_{ij} = 0$时，对应的标签$Y_{ij}$可能为-1或者1。半监督学习的目标，就是从$\{X, C\}$中学习一个预测的标签矩阵$\hat{Y} \in \{-1, 1\}^{m \times n}$近似$Y$。

正式地，记G_I为有标签和无标签实例的带权的邻接图。G_I上的每个顶点对应一个实例x_i，x_i和x_p的一条边表示，x_i是x_p的一个k近邻，或者x_p是x_i的一个k近邻。定义一个$m \times m$的稀疏矩阵S，表示相邻的实例的相似性:

$$S_{ip} = \begin{cases} \frac{1}{z_i} \exp(-\frac{||x_i - x_p||_2^2}{2\sigma^2}) & \text{if } p \in \mathcal{N}_i \\ 0 & \text{otherwise} \end{cases}.$$

其中$\mathcal{N}_i$是第i个实例的k近邻的实例集合。$z_i = \sum_{p \in \mathcal{N}_i} \exp(-\frac{||x_i - x_p||_2^2}{2\sigma^2})$，因此$\sum_{p \in \mathcal{N}_i} = 1$。为了减少有标签和无标签实例的$k$近邻计算花费，Dong[5]采用kd树用于搜索每个实例的k近邻，并且用多标签维度缩减方法减小维度灾难的影响。

在半监督的弱标签学习中，可观察到的实例的相关标签集是不完整的，因此不能像计算实例相似度一样直接计算标签的相似度矩阵L，于是需要学习这个矩阵。为了后续讨论的简便，先假定标签相似度矩阵L已经给出。

为了估计预测的标签矩阵$\hat{Y}$，从光滑性假设的角度有两种主要的方法。第一种，从实例相似度的角度，一个实例的相关标签集可以从它的近邻中推导出，即$\hat{Y}_{ij} \approx \sum_{p \in \mathcal{N}_i} S_{ip} \hat{Y}_{pj}$。第二种，从标签相似度的角度，训练实例的每个确定标签的标记都可以从相近的标签标记中推导出，即$\hat{Y}_{ij} = \sum_{q \in \hat{N}_j} \hat{Y}_{iq} L_{qj}$，其中$\hat{N}_j$是第j个标签的标签集的$k$近邻。显然，预测的标签矩阵不仅与实例相似度有关，也与标签相似度有关。于是就可以把实例相似度和标签相似度的光滑性表达为：

$$\hat{Y}_{ij} \approx \sum_{p \in \mathcal{N}_i} \sum_{q \in \hat{\mathcal{N}}_j} S_{ip} \hat{Y}_{pq} L_{qj} \tag{3.25}$$

接着，添加一个新的正则项

$$\begin{aligned} \Omega(\hat{Y}, S, L) &= \sum_{ij} (\hat{Y}_{ij} - \sum_{p \in \mathcal{N}_i} \sum_{q \in \hat{\mathcal{N}}_j} S_{ip} \hat{Y}_{pq} L_{qj})^2 \\ &= ||\hat{Y} - S\hat{Y}L||_F^2 \end{aligned} \tag{3.26}$$

其中$||M||_F^2 = tr(MM')$，$tr(\cdot)$是一个矩阵的迹。

正式地，记XW和$X\bar{W}$为两个线性的多标签模型，其中$W, \bar{W} \in \mathcal{R}^{d \times n}$是系数矩阵。第一个模型$XW$初始化用于预测观察到的相关标签，即$C_{ij} = 1$的元素，优化的目标形式化为$||(XW) \circ C - C||_F^2$，其中$\circ$表示Hadamard积。第二个模型初始化用于预测标签出现矩阵C中未确定的元素，即$C_{ij} = 0$的元素，优化的目标形式化为$||(X\bar{W} \circ (E - C) + (E - C))||_F^2$，其中$E_{m,n}$是全为1的矩阵。综上，可以得出目标公式为发现$W$，$\bar{W}$和标签相似度矩阵$L$使得下式的目标最小化：

$$\begin{aligned} \min_{W, \bar{W}, L} & ||(XW) \circ C - C||_F^2 + \alpha \Omega(U, S, L) + \\ & \beta ||(X(W - \bar{W})) \circ (E - C)||_F^2 + \\ & \zeta ||(X\bar{W} \circ (E - C) + (E - C))||_F^2 \\ \text{s.t.} & U = (XW) \circ C + (X\bar{W}) \circ (E - C) \end{aligned} \tag{3.27}$$

其中α, β, ζ是参数。$U = (XW) \circ C + (X\bar{W}) \circ (E - C)$是两个模型的集成预测。目标式的最优化求解可以通过块坐标下降进行。

4 总结

机器学习中的监督算法在工业界已经得到了广泛应用，例如，垃圾邮件过滤、网络入侵检测、商品推荐等。从2012年的深度学习爆发以来，以卷积神经网络为代表的深度学习在计算机视觉、自然语言处理、神经机器翻译等领域大显神通。大数据时代最大的好处是每天海量增长的数据，坏处则是这些数据很难得到有效的标注，弱监督、半监督等领域关注的问题正是在噪声标注、错误标注或标签缺失、大量无标签数据的场景下设计有效的机器学习算法，完成分类、回归等各项任务。

本文对半监督学习的一些算法进行了简介和综述。在面对无标注数据时，Zhou[1]和Lee[2]采用伪标签的方式通过已有的分类器给无标注数据加上标签参与训练；Xie[4]则利用AUC优化目标的特殊性，使得无标注数据既当作正例与负类计算AUC，又当作负例与正类计算AUC，从而得到无偏的AUC估计。Kiryo[3]则基于已有的无偏PU Learning框架，发现过拟合的问题，加以修正，从而使得深度网络也可以应用PU Learning的算法。Dong[5]面对多标签问题，既考虑了实例之间的相似度，又考虑了标签之间的相似度，从而提出了更优的多标签预测算法。

Zhou[1]推导出了伪标签参与训练的决策条件，然而这也限制了可利用的无标签数据的数量。Xie[4]和Dong[5]的不足在于只适用于线性模型，在面对复杂问题时难起作用。Lee[2]的伪标签更经验性，依赖于调参。Kiryo[3]适用于深度网络，是一个优秀的想法，实验结果也很优秀，但是在面对深度网络mini-batch的训练设计时，估计的损失风险会有很大的波动。另外，正例的先验概率估计也并不是一项简单的任务。综上，半监督学习确实取得了一些进展，而大规模可应用的算法仍待研究。

参考文献

[1] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," IEEE Transactions on knowledge and Data Engineering, vol. 17, no. 11, pp. 1529–1541, 2005.

[2] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in International Conference on Machine Learning, 2013.

[3] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in Advances in Neural Information Processing Systems, pp. 1674–1684, 2017.

[4] Z. Xie and M. Li, "Semi-supervised auc optimization without guessing labels of unlabeled data," pp. 4310–4317, 2018.

[5] H.-C. Dong, Y.-F. Li, and Z.-H. Zhou, "Learning from semi-supervised weak-label data," pp. 2926–2933, 2018.

[6] X. Zhang and Y. LeCun, "Universum prescription: Regularization using unlabeled data.," in AAAI, pp. 2907–2913, 2017.

[7] G. Wei and Z. Zhi-Hua, "On the consistency of AUC pairwise optimization," in Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, pp. 939–945, 2015.

[8] M. C. du Plessis, G. Niu, and M. Sugiyama, "Analysis of learning from positive and unlabeled data," in Advances in Neural Information Processing Systems, pp. 703–711, 2014.

[9] M. Du Plessis, G. Niu, and M. Sugiyama, "Convex formulation for learning from positive and unlabeled data," in International Conference on Machine Learning, pp. 1386–1394, 2015.

[10] M. Christoffel, G. Niu, and M. Sugiyama, "Class-prior estimation for learning from positive and unlabeled data," in Asian Conference on Machine Learning, pp. 221–236, 2016.