LLP-AAE 算法在金融风险识别领域的应用

范子祎,杨 欢
(对外经济贸易大学 信息学院,北京 100029)

摘 要:通过结合机器学习中的标签比例学习问题与生成模型算法,提出了一种新的深度学习算法,称为对抗性自动编码器算法。考虑到金融领域的数据具有天然的获取成本高、保密性大与客户可分类的特点,故选择把对抗性自动编码器算法实践于银行的风险检测中。在标签比例学习中,数据集为包层面的标签比例数据,将这种类型的数据作为输入,对抗性自动编码器依赖编码器解码器等的对抗训练来捕获原始数据分布,并得到足够逼真的合理重构分布。同时施加一种对抗学习机制作为一种正则化,该部分对抗使得编码器学习足够充分与多维,进一步提升了算法的效果。与现存的方法相比,提出了具有可扩展性的深度学习解码器,拓宽了应用范围情景,侧重将标签比例学习领域的深度学习方法应用于金融类型数据之中。

关键词:对抗性自动编码器;标签比例学习;金融风险识别;弱监督学习
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2021)0151-04

Adversarial Auto-encoding for Learning from Label Proportions on Financial Risk Identification

FAN Zi-yi, YANG Huan
(School of Information, University of International Business and Economics, Beijing 100029, China)

Abstract: Combining a weakly-supervised method, learning from label proportions (LLP) and a generative model, adversarial auto-encoders (AAE), we propose an effective deep learning algorithm of tabular data, called LLP via adversarial auto-encoders (AAE-LLP). Considering that the data in the financial field are naturally characterized by high acquisition cost, high confidentiality and customer classification, the adversarial autoencoder algorithm is chosen to be applied in the risk detection of banks. In the tag proportion learning, the data set is the tag proportion data at the packet level, and this type of data is taken as the input. Adjective autoencoders rely on the antagonism training of encoders and decoders to capture the original data distribution, and get a reasonable reconstruction distribution that is realistic enough. At the same time, an antagonistic learning mechanism is applied as a kind of regularization, which makes the encoder learning sufficient and multi-dimensional, and further improves the effect of the algorithm. Compared with the existing methods, a deep learning decoder with extensibility is proposed to expand the scope of application scenarios, and the deep learning method in the field of label proportional learning is applied to financial type data.

Key words: AAE; LLP; financial risk identification; weak supervised learning

0 引 言

金融行业的数据信息十分复杂,客户每天日常的交易行为会产生大量类型的数据。但是,由于隐私保护,获取的金融数据往往是不完整的。特别是在信贷市场,基于用户交易行为汇总的数据无法判断客户是否有违约风险。传统的监督学习或半监督学习方法基于数据的每一个属性,利用大量的标注数据学习得到一个分类器,从而对未知的样本进行分类。然而,在实际情况中,因为隐私保护等问题,获取完整标签的数据是不太现实的,而且识别到每一位客户的风险极其困难,更多时候只能够得到所有样本中每一类样本所占的比例。因此,希望用一种弱监督的学习方法解决这类问题,能够基于无标注的数据求解客户风险识别。标签比例学习[1-2]作为一种新兴的弱监督学习方法,可以将数据集分成不同的包。对每个包内的样本标签是未知的,但可以根据每类样本在每个包的比例求解出一个分类器,从而预测每个样本的类别。其在总统竞选、人口统计、商业客户识别[4]等领域有着广泛的应用[3],这促使利用标签比例学习解决金融领域中信贷客户风险识别问题。

收稿日期:2020-08-18
作者简介:范子祎(1996-),女,硕士,研究方向为弱监督机器学习-比例标签学习。

在深度学习出现之前,有几种浅模型用来求解标签比例学习。一种是概率估计模型,利用概率估计预测标签比如 MeanMap[2] 和 Laplacian MeanMap[3],但是这种基于 MCMC 算法的统计方法受制于过高的计算复杂度,以及存在先验条件假设的限制问题。当存在很多不确定的原始数据分布时,往往产生错误的分类结果,或者和原始数据分布偏差较大;后来,有学者提出用支持向量机求解标签比例学习。Invcal 和 alter-∝SVM[6] 考虑了潜在的未知实例标签和已知的包的标签比例,得到一个大边界的分类器;以及在模型的精度和泛化能力有所提升的 Linear Twin SVM[7] 和 LLP-NPSVM[8]。虽然这种方法有效避免了对数据集假设的局限性,但是基于 SVM 的方法缺乏可收缩性,受到混合优化问题的约束,这意味着该方法不能精确地解决一类问题的大规模实例[9]。深度学习出现之后,DLLP[10] 首次利用 DNN 解决标签比例学习问题;LLP-GAN[11] 从生成模型的隐式表示学习出发,将生成对抗网络的判别器作为特征提取器学习比例信息,拓展了标签比例学习的应用环境。Multi-class LLP[12] 研究了基于标签比例学习非二进制分类模型,弥补了之前的研究空白,证明了标签比例学习在大规模数据处理的可行性。深度学习能够利用分层结构[20],构建多个隐层处理复杂的高维数据[13],大容量金融表格数据也能够满足深度学习对于模型训练集规模的要求,避免出现过拟合问题。

文中研究工作基于以下三个方面:

第一,LLP-GAN 表明了生成模型生成对抗网络应用于图像识别标签比例学习问题的可行性,文中利用对抗性自动编码模型进一步推广标签比例学习在深度学习中的应用,即学习金融领域的表格数据。在生成模型中,生成对抗网络[14-15] 利用一个捕获数据分布的生成模型 G 和一个判断样本真假的判别模型 D 的极大极小化博弈,求得最终二者达到平衡的唯一解。基于反向传播算法还有一个应用广泛的生成模型变分编码器[16],利用特征识别网络和变分近似推理来拟合一个近似推理模型,使用所提出的下界估计器对难以处理的后验进行估计。但是二者各有缺陷,变分自动编码基于蒙特卡洛抽样进行反向传播时,计算 KL 散度需要获得先验分布的确切的函数形式。对抗性自动编码只需要对模型施加一定的分布来使得近似分布和真实分布匹配[17]。生成对抗网络需要对神将网络的输出层施加像素级的分布,而对抗性自动编码通过模型中的编码器获取数据分布,只需要用一个更简单的分布就能得到更好的测试性。

第二,文中工作将标签比例学习和深度学习算法对抗性自动编码相结合,希望在仅有标签比例的弱监督环境下,模型既能够处理大批量的数据,也能够保证原始数据的内在特征[13],而这是以往基于金融数据使用决策树等方法进行客户分类的缺陷。

第三,和已有的标签比例学习算法相比,在对不需要样本标注的弱监督问题处理上,LLP-AAE 能够通过模型中的编码器从高维数据中得到低维表示形式,发现原始数据的分布特征和潜在属性,得到最终的分类结果。

1 相关研究

LLP-GAN[11] 的研究,证明了生成模型在弱标签比例问题上的实用性。而重点是将更具竞争力的生成模型对抗性自动编码应用于表格数据的标签比例学习问题[19]。对抗性自动编码可以将自动编码器转换为生成模型,依靠自动编码器训练来捕捉数据分布。对抗性训练就像一个正则化机制,在从头开始训练自动编码器的同时塑造数据分布。在对抗性自动编码中,自动编码器的训练有两个目标:传统重建损失和对抗性训练损失。为了使潜变量的聚集与先验分布相匹配,对抗性自动编码利用对抗性训练来匹配聚集的后验分布和任意先验分配。另一方面,对抗性自动编码不需要像变分自动编码一样获取先验分布的精确函数形式[17]。变分自动编码使用 KL 散度惩罚来对自动编码器的隐藏码向量施加先验分布。在蒙特卡罗抽样,为了通过 KL 散度进行反向传播,需要获得先验分布的精确函数形式。而对抗性自动编码采用了一种对抗性的训练方法,即匹配潜变量的后验概率和先验分布,只需要从先验分布中进行抽样就能使近似分布与真实分布相匹配分配。对抗性自动编码依赖于自动编码器训练,以更简单的分布捕捉数据分布,从而产生更好的测试性能[18]。

2 AAE-LLP 模型

在模型中,目标是利用生成对抗网络来训练一个概率自动编码器。对抗性自动编码可以进行变分推理来匹配潜变量的后验。

AAE-LLP 由一个自动编码器和两个独立的对抗网络组成,如图 1 中展示的 AAE-LLP 框架所示。设 x 为输入向量,y' 和 z' 为自动编码器的潜在变量。将先验分布 $p(y)$ 和 $p(z)$ 加在两个潜变量上。假设数据能由两个变量刻画,一个是类变量,遵循分类分布 $p(y)=\mathrm{Cat}(y)$,另一个是遵循高斯分布 $p(z)=N(z\mid 0,I)$ 的连续变量。然后从编码器得到编码分布 $q(y,z\mid x)$,从解码器得到解码分布 $p(x\mid y,z)$,然后从全连接层得到 $q(y\mid x)$ 和线性激活函数层的 $q(z\mid x)$ 分布。设 $P_d(x)$ 为数据分布,基于编码分布,定义了两个聚集后

验分布 $q(y)$ 和 $q(z)$,如下所示:

$$q(y)=\int_x q(y\mid x)P_d(x)\mathrm{d}x \tag{1}$$

$$q(z)=\int_x q(z\mid x)P_d(x)\mathrm{d}x \tag{2}$$

全部的对抗网络部分可以看作对于自动编码器的一种正则化,即引导 z' 来逼近 z , y' 来逼近 y 。对于对抗网络的第一部分,设置了一个类别分布 $\mathrm{Cat}(y)$,确保潜在类变量 y' 的聚集后验分布遵循分类分布。另一个对抗网络上施加高斯分布,并确保潜在变量 z' 的聚集后验分布与高斯分布相匹配。

利用比例信息指导无监督学习,即生成器 G 将从真实数据生成的标签比例 q 作为 AAE-LLP 的弱监督信息。该模型在每一个包上依次进行无监督重构、对抗性正则化和弱监督学习的训练。对应下面的章节中,三个训练阶段定义三种损失函数。

2.1　重构损失

假设编码器的网络是 g_η ,解码器是 f_σ ,在处理每一个分包时无监督的重构部分,解码器更新 g_η 与 f_σ 来最小化重构损失 L_{recon} 。

$$L_{\mathrm{recon}}=L(x,x')=L(x,g_\eta(f_\sigma(X)))=E_{x\sim p_d}\parallel X-g_\eta(f_\sigma(X))\parallel^2 \tag{3}$$

2.2　对抗损失

在对抗正则部分,设置了两个判别器与一个生成器,此时编码器同时充当了生成器。首先,输入训练集中每一个包 $B=\{x(1),x(2),\cdots,x(N)\}$ 的特征向量 x 生成不同包的类别比例标签,得到 $L=\{(B,P)\}_{i=1}^{n}$ 。经过编码器后,通过假设的潜变量空间 y' 和 z' 得到了 $q(y')$ 和 $q(z')$ 。根据每一个包的比例信息来设置类别分布 $\mathrm{cat}(y)$,而潜变量 z 则考虑施加一个高斯噪声 $N(0,1)$ 。

两个判别器 $D_1(y)$ 和 $D_1(z)$ 是为了判断输入是来自真实数据集还是来自生成器 G 的假数据,它们会在得到生成器的输入后进行同步更新。下一步,对抗网络中的生成器 G 进行更新,以生成足够逼真的数据来迷惑判别器。

考虑 P 是一种能将数据空间分成 n 个不相交子集的划分,即 $P_i(y)(p(y)=\mathrm{Cat}(y))$, $i=1,2,\cdots,n$ 是 P 中元素的边缘分布。在对抗网络中,当生成器固定时,以变量 y 为例,对于判别器 D_1 可以得到 $P_{D_1}(k\mid y)$ 。

$$P_{D_1}(1\mid y)=\frac{\sum_{i=1}^{n}P_i(y)}{\sum_{i=1}^{n}P_i(y)+q(y')} \tag{4}$$

$$P_{D_1}(0\mid y)=\frac{q(y')}{\sum_{i=1}^{n}P_i(y)+q(y')} \tag{5}$$

在以上公式中, $q(y')$ 是生成数据的边缘分布,因为生成器也充当了编码器的作用,所以得到 $P_{g_1}=q(y')$ 。

判别器 $D_1(y)$ 的最优化目标是:

$$\max_{D_1}V(G,D_1(y))=E_{x\sim p_d}E_{y\sim p_{g_1}}[\log p_{D_1}(0\mid y)]+E_{y\sim p(y)}[\log p_{D_1}(1\mid y)] \tag{6}$$

生成器 G 针对类别变量 y 的最优化目标是:

$$L(G_1)=\min_G V(G,D_1(y))=\min_G E_{y\sim p_{g_1}}[\log p_{D_1}(0\mid y)] \tag{7}$$

对于潜变量 z ,判别器 $D_2(z)$ 的最大化目标是:

$$\max_{D_2}V(G,D_2(z))=E_{x\sim p_d}E_{y\sim p_{g_2}}[\log p_{D_2}(0\mid y)]+E_{y\sim p(z)}[\log p_{D_2}(1\mid y)] \tag{8}$$

G 的最优化目标是:

$$L(G_2)=\min_G V(G,D_2(y))=\min_G E_{y\sim p_{g_2}}[\log p_{D_2}(0\mid y)] \tag{9}$$

为了解决生成对抗网络生成器的不稳定性,及其带来的训练过度问题,使用了特征匹配这个方法[19],由此得到了新的判别器目标:

$$L(G)=\parallel E_{x\sim p_d}f(x)-E_{x\sim p_g}f(G(z))\parallel_2^2 \tag{10}$$

其中, $f(x)$ 指判别器的某一中间层。

2.3　比例损失

假设编码器的每一个输出结果为 $\check{P}_i^j=p_\theta(y\mid x_i^j)$,其中 θ 是网络参数。然后得到了包级别的比例信息,在第 i 个包上表示为 $\bar{P}_i$:

$$\bar{P}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\widetilde{p_i^j}=\frac{1}{N_i}\sum_{j=1}^{N_i}p_\theta(y\mid x_i^j) \tag{11}$$

在比例学习阶段,根据生成的类样本 y' 计算包级别的标签,得到比例 q 。为了与真实数据的包级别标签比例 p 保持一致,自动编码器更新 q 以最小化标签比例上的交叉熵代价。认为 $p_i(y)$ 是第 i 个包中 y 类的比例。根据抽样方法的思想和 Jensen' s 不等式,得到:

$$\begin{aligned}L_{\mathrm{prop}}&=-\mathrm{CE}(p,q)=-p\log[E_{x\sim P_d}E_{y\sim P_{g_1}}q]=\\&\sum_{i=1}^{n}\sum_{y=1}^{Y}p_i(y)\log[\frac{1}{N_i}\sum_{j=1}^{N}p_{g_1}(y\mid x_i^j)]\cong\\&\sum_{i=1}^{n}\sum_{y=1}^{Y}p_i(y)\log[\int p_d^i(x)p_{g_1}(y\mid x)\mathrm{d}x]\geqslant\\&\sum_{i=1}^{n}\sum_{y=1}^{Y}p_i(y)E_{x\sim p_d^i}[\log p_{g_1}(y\mid x)]\end{aligned} \tag{12}$$

3　实　验

3.1　数据准备

之前的研究者在 UCI 标准数据集德国信用数据集上开展了大量的分类实验,以期更好地识别有风险

类型客户,但其中很少有用标签比例学习的方法进行的实验。

之所以选择这个数据集和信用评分区域,是因为相信这是标签比例学习最适合使用的领域之一。个人信用信息难以收集和分类已经成为一个通识。一方面,个人信用信息具有高度机密性,有时无法获取。另一方面,一个人的每一次申请或交易都可能导致一种趋势,来描述他的逾期风险。因此,这些不同的特征之间有着客观的内在的联系,却并不是很容易被发掘。从这一点来看,对抗性自动编码算法中的潜变量空间可以很好地匹配这种情况。第三,德国信用数据集共有 1 000 个记录,21 个特征,仅仅 1 000 条记录是不够丰富的,不足以让机器充分地学习,这就满足了生成模型的功能,它可以生成看起来像真实数据的假数据,在某些迭代轮次时,甚至有能力混淆判别器。一旦把这些足够逼真的数据输入到本模型中,神经网络的多感知层可以感知不存在的新数据,从中学习更为丰富的数据变化。

在这个数据集中,有不同类型的不同特征,包括现有支票账户的状态、月内期限、信用历史、目的、储蓄账户、现在的就业时间、个人身份和性别、其他担保人、财产、其他分期付款计划、住房情况、工作、电话、是否外籍工人、信贷金额,分期付款率占可支配收入的百分比、现居住地、年龄(以年计)、在我行现有信贷额、有责任提供担保的人数、海关类别。把这些特征分开看,没有一个可以作为判断一个客户逾期的依据,单独来看这些属性都很普通,当它们汇聚在一起时却有了新的信息。比如住在同一个地区、拥有相同的信用额度/收入、面临同样的住房状况的人有很大的可能具备相同的信用情况。这使得“包”有了特别的优势。

3.2 实验步骤

在本实验中,选择随机分包来获得这种标签比例学习算法的包比例标签,并将袋子大小定为 30。然后将数据集拆分为 800 条训练集,200 条测试集,训练 50 个轮次。设置了一个动态学习率,当轮次小于 30 时,判别器 D_y 、判别器 D_z 和生成器 G 的学习率为 0.1,解码器 DE 为 0.2。当轮次大于 30 小于 50 时,所有网络的学习率将变为 0.05。为了提高生成器的质量,采用了特征匹配技术。不同网络的架构参考对抗性自动编码。

3.3 实验结果

在这个数据集上使用有监督的对抗性自动编码作为对照实验,这样就可以明确问题的难度和极限,最终得到 27% 的错误率。与其他方法相比,对抗性自动编码在 10 次运行中得到的最好结果为每组 16 个样本时,错误率为 27% 。本算法结果虽然不能超越有监督的准确性,但在考虑个人信息敏感性的同时,能更好地适应不完全信息,降低获取信息的成本。

表 1 德国银行信用卡数据结果对比

算法	分包				
	4	16	32	64	128
随机森林	78%(±0.2)				
监督学习 AAE	73%(±0.1)				
标签比例学习 AAE	68%(±0.3)	73%(±0.3)	66%(±0.6)	64%(±0.2)	59%(±0.6)

4 结束语

提出了对抗性自动编码器网络(AAE)在标签比例学习问题域上的实现,它由四个网络学习器组成,一个生成器、两个鉴别器和一个解码器。主要动机是对抗性自动编码算法在有监督问题或半监督问题上,在多个问题域内有优秀的实验表现,因此在具备有监督和半监督问题的部分共性下,它能否在标签比例学习问题上取得很好的效果。最后证明了该算法可以移植和使用,具有很大的适用空间。

然而,该算法还有一些值得优化的空间,比如不同的损失函数可以被加权,这样就可以控制不同网络学习器的重要性。以及受限于实验设备,无法进行实验来获得本算法在图片类问题上的表现。

参考文献:

[1] WANG Z,FENG J. Multi-class learning from class proportions[J]. Neurocomputing,2013,119:421-428.

[2] QUADRIANTO N,SMOLA A J,CAETANO T S,et al. Estimating labels from label proportions[C]//Proceedings of the 25th international conference on machine learning. Helsinki, Finland:ACM,2009:776-783.

[3] PATRINI G,NOCK R,RIVERA P,et al. (Almost) no label no cry[C]//Neural information processing systems. Canada:NIPS foundation,2014:190-198.

[4] 石 勇,马福海,齐志泉,等. 基于比例标签学习的商业银行重要基金客户识别研究[J]. 数学的实践与认识,2017,47(19):291-302.

[5] 肖燕珊,梁 飞,刘 波. 基于弱标签的多示例迁移学习方法[J]. 计算机应用研究,2021,38(1):125-128.

[6] YU F X,LIU D,KUMAR S,et al. ∝SVM for learning with label proportions[C]//Proceedings of the 30th international conference on machine learning. Atlanta, GA, USA:JMLR.org,2013:III-504 - III-512.

[7] WANG B, CHEN Z, QI Z. Linear twin SVM for learning from label proportions[C]//IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. Singapore:IEEE,2015:56-59.

[8] QI Z,WANG B,MENG F,et al. Learning with label propor-

(下转第 158 页)