

基于 PU 与生成对抗网络的 POI 定位算法

田继伟 王劲松 石 凯

(天津理工大学计算机科学与工程学院 天津 300384)
(天津市智能计算及软件新技术重点实验室(天津理工大学) 天津 300384)
(计算机病毒防治技术国家工程实验室(天津理工大学) 天津 300457)
(jiwei.tian@foxmail.com)

Positive and Unlabeled Generative Adversarial Network on POI Positioning

Tian Jiwei, Wang Jinsong, and Shi Kai

(School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384)
(Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology (Tianjin University of Technology), Tianjin 300384)
(National Engineering Laboratory for Computer Virus Prevention and Control Technology (Tianjin University of Technology), Tianjin 300457)

Abstract With the rapid popularization of smart mobile devices, people rely more and more on location-based social networking service (LBSNS). Due to the high cost of data acquisition, point of interest (POI) positioning based on small data collection has become a big challenge. Recent research focuses on received signal strength (RSS) and simultaneous localization methods. Although there has been some research on POI positioning, the existing approaches do not discuss the problem of insufficient positive training samples. Based on the truthful positive data and a large amount of unlabeled data, a novel approach, called positive and unlabeled generative adversarial network (puGAN), is proposed. Firstly, we use positive and unlabeled method along with the generative adversarial network to effectively mine the hidden features of data. Secondly, based on the hidden features, we calibrate the positive data and unlabeled data, then treat them as the input of the discriminator. Finally, with the minimax of generator and discriminator, a POI-discriminator model is obtained. We evaluate the new method by analyzing ROC curve and the relationship between training error and testing error. The results of experiments show that the method we proposed can effectively solve the problem of insufficient positive samples and outperforms the traditional models of POI positioning, including one-class classifier, SVM and neural network.

Key words data mining; point of interest (POI); positioning; positive and unlabeled; generative adversarial network (GAN)

摘 要 随着智能移动设备的快速普及,人们对基于位置的社交网络服务的依赖性越来越高.但是,由于数据采集成本昂贵以及现有数据采集技术的缺陷,基于小样本数据挖掘的兴趣点(point of interest,

POI)定位已经成为了一种挑战.尽管已经有一些 POI 定位方面的研究,但是现有的方法不能解决正样本数据不足的问题.提出一种基于 PU 与生成对抗网络(positive and unlabeled generative adversarial network, puGAN)的模型,采用 PU 学习和生成对抗网络相结合的方式挖掘数据的隐藏特征,生成伪正样本弥补数据不足的问题,并校正无标签样本数据的分布,从而训练出有效的 POI 判别模型.通过分析 ROC 曲线以及训练误差和测试误差在迭代过程中的变化和关系来比较不同模型在实验场景下的效果.结果表明,puGAN 模型可以有效解决数据样本不足的问题,进而提高 POI 定位的准确性.

关键词 数据挖掘;兴趣点;定位;PU;生成对抗网络

中图法分类号 TP391

随着互联网的蓬勃发展,个性化推荐技术已经成为学术界的研究热点之一.近年来,个性化推荐技术已经成功应用于多个领域,尤其是在广告推荐领域,它带来了巨大的商业价值.但是,个性化推荐技术的数据来源于人们在网络上产生的行为,这些网络行为往往是低成本的,由这些低成本的行为推荐来的广告虽然具有很大价值,但确有局限性.例如,经常在网络上浏览汽车的人虽然有很大可能性接受汽车类的广告推荐,但是相比之下,那些去 4S 店的人有更大可能性购买汽车,从而给商家带来真正的利润.考虑到推荐技术的这种局限性,对兴趣点(point of interest, POI)的定位和推荐技术就变得非常重要.另外,随着智能移动设备的快速普及,人们对基于位置的社交网络服务(location-based social networking service, LBSNS)的依赖性也越来越高.基于数据挖掘的 POI 定位和推荐技术已经成为一个重要课题.高榕等人^[1]将 POI、社交关联和评论信息融合起来,提出了一种 GeoSoRev 兴趣点推荐模型.Chen 等人^[2]研究了 LBSNS 中的连续个性化 POI 推荐问题,将时间关系融合到 POI 推荐中.Gao 等人^[3]将内容信息融合进了 POI 推荐系统中.Yin 等人^[4]考虑了用户在不同区域往往有不同的兴趣,研究了跨区域的 POI 推荐系统.但是,文献[1-4]的 POI 推荐系统都无法离开 POI 定位技术的建设.

目前,国内外许多研究已经利用无线网络来进行 POI 定位.其中,使用最多的方法是基于无线信号接收强度(received signal strength, RSS)的位置指纹技术^[5-7],该技术已经广泛应用于移动通信领域,且可以在任何具有 WiFi 发射设备的地方推广应用.但是由于用户隐私问题,有 POI 标记的数据难以收集.一般来说,采集来的数据可以分为 2 类:一类是占比很大的没有标记 POI 的数据;另一类是占比很小的标记了 POI 的数据,即正样本数据.这就造成了正样本数据不足的问题,极大地影响了 POI 定位的准确

性.目前很少有研究致力于解决这一普遍存在的问题.Matos 等人^[8]提出了一种模拟定位的方式去进行 POI 的识别.然而,这种识别系统的误差会随着系统计算量的增大而增加.此外,Moghtadaiee 等人^[9]提出了一种短距离 RSS 误差模型来提高 POI 定位效果,但累积误差的问题仍然无法得到有效解决.

本文基于现有数据,提出了一种基于 PU 与生成对抗网络(positive and unlabeled generative adversarial network, puGAN)的模型来进行 POI 的有效定位与识别.实验表明,本文提出的 puGAN 模型比传统的 POI 定位和分类模型具有更好的效果.本文工作的主要贡献有 2 个方面:

- 1) 将 PU(positive and unlabeled)学习^[10-11]和生成对抗网络(generative adversarial network, GAN)模型相结合,并应用到 POI 定位算法中,以此解决数据样本不足的问题,从而提高 POI 定位的准确性;
- 2) 理论证明了 puGAN 模型的普遍适用性.

1 相关工作

到目前为止,大部分 POI 定位模型都有着数据采集不足的缺陷.有 2 种方式可以解决这个问题:一种方式是覆盖所有的采集点采集充足的数据,但这种方式往往伴随着巨大的人力、财力和时间成本,通常来说是不可接受的;另一种方式是基于少量正样本和大量无标签样本,使用 PU 学习方法.下面简单介绍 PU 学习算法.

1.1 PU 学习算法

PU 学习作为半监督学习^[12]的一个重要分支,逐渐成为了研究界的热点问题.PU 算法对无标签样本的使用情况,可以概括成 2 个方面:

- 1) 基础模型的研究.通过单类模型(one-class, OC)对正样本数据进行学习,但是这种方式通常会

浪费较多的数据特征,从而导致模型欠拟合.或者使用 SVM 等二分类模型,它们虽然能够利用到所有的数据特征,但无法避免无标签样本噪声过大而可能导致的欠拟合问题.

2) 基于高可信样本的挖掘.这种方法的核心是优先找到高可信的正样本或者负样本,从而基于此样本进行学习和分类.任亚峰等人^[13]使用这种 PU 学习的方式,从大量无标签样本中识别出少量高可信的负样本,进行了识别虚假评论的研究.

PU 学习虽然能够利用少量正样本以及大量无标签样本进行学习,但是正样本和无标签样本的样本量相差巨大,影响了模型学习的准确性和快速性.本文在利用 PU 学习算法的同时引入了 GAN 模型,一方面生成伪正样本弥补正样本不足的缺陷,另一方面学习无标签样本的分布,提高分类模型的准确性.下面简单介绍 GAN 模型.

1.2 GAN 模型

2014 年,Goodfellow 等人^[14]提出了 GAN,并进行了一系列研究^[15-16],解决了如何使机器学习的训练成果向着理想方向前进的问题.Odena^[17]首次将 GAN 模型与半监督学习相结合来提高生成器的质量.

如图 1 所示,GAN 模型针对正样本提出了生成器与判别器的概念.生成器通过不断学习真实样本数据的分布生成伪样本,从而尽可能欺骗判别器.而判别器则通过学习真实样本和伪样本的差异,尽可能区分出数据来源.

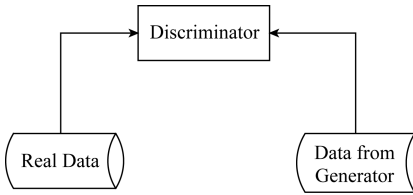


Fig. 1 The illustration of GAN model

图 1 GAN 模型图

生成器和判别器不断博弈的过程就是数据不断生成和分类的过程.实验证明,GAN 模型在训练样本不足的情况下,能够有效学习出原有数据的分布特征.与此同时,模型的判别器也能够准确地区分出样本是来源于生成器还是来源于真实的训练数据集.

另外,2017 年,Arjovsky 等人^[18]提出了一种观点,即 GAN 中利用 JS 散度会导致纳什不均衡的问题,这会使模型训练不稳定.为了解决这个问题,他提出了 WGAN 模型,这种模型在提升稳定性与准

确性上有很好的表现.本文提出的 puGAN 模型也使用了这一思想.

不同于传统的 GAN 模型,本文不仅将 GAN 模型应用于 POI 定位中,而且在模型中分别对正样本和无标签样本都引入了生成器和判别器.

2 基于 PU 与生成对抗网络的模型

本文将第 1 节的 PU 学习与生成对抗网络模型相结合,提出了基于 PU 与生成对抗网络的模型.下面将详细介绍 puGAN 模型的原理及理论分析.

2.1 puGAN 模型

在本文中, X_p, X_u 分别表示正样本数据集和无标签样本数据集; $p_p(x), p_n(x), p_u(x), p(x)$ 分别表示正样本、负样本、无标签样本以及整个数据集合的分布; G_p, G_u 分别表示 puGAN 模型中正样本生成器和无标签样本生成器; D_p, D_u 分别表示正样本判别器和无标签样本判别器; D_{pu} 表示正样本/无标签样本判别器.本文假设数据分布遵循的规则为

$$p(x) = p_p(x) + p_n(x),$$

$$p_u(x) = \theta_p p_p(x) + \theta_n p_n(x),$$

其中 θ_p, θ_n 都是 $0 \sim 1$ 之间的小数,并且满足 $\theta_p + \theta_n = 1$.

基于上面的假设,本文设计了 puGAN 模型.如图 2 所示,描述了 puGAN 模型的训练过程.其中,判别器 D_u 通过对真实的无标签样本与无标签样本生成器 G_u 生成的数据进行判别分类,从而调整学习得到无标签样本数据的分布特征.同理,判别器 D_p 通过对正样本进行判别分类,训练得到正样本数据的分布特征.

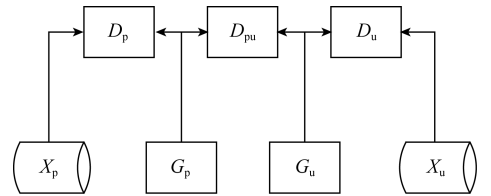


Fig. 2 The illustration of puGAN model

图 2 puGAN 模型图

最终,puGAN 模型的目标函数为

$$\begin{aligned} \min_{\{G_p, G_u\}} \max_{\{D_p, D_u, D_{pu}\}} V(G, D) = & \\ \alpha \min_{\{G_p\}} \max_{\{D_p\}} V_{G_p, D_p}(G, D) + & \\ \beta \min_{\{G_u\}} \max_{\{D_u\}} V_{G_u, D_u}(G, D) + & \\ \gamma \min_{\{G_p, G_u\}} \max_{\{D_{pu}\}} V_{G_p, G_u, D_{pu}}(G, D), & \quad (1) \end{aligned}$$

其中, α, β, γ 均为大于 0 的参数.

判别器希望提高自己的辨别能力, 能够准确分辨出样本是来源于真实样本集还是来源于生成器生成. 而生成器则不断提升自己的“伪造”技术, 希望能够生成可以“以假乱真”的数据. 根据零和博弈的原则, 将目标函数进行推导, 可以得到

$$V_{G_u, D_u}(G, D) = E_{x \sim p_u(x)} \lg(D_u(x)) + E_{z \sim p_{zu}(z)} \lg(1 - D_u(G_u(z))), \quad (2)$$

$$V_{G_p, D_p}(G, D) = E_{x \sim p_p(x)} \lg(D_p(x)) + E_{z \sim p_{zp}(z)} \lg(1 - D_p(G_p(z))), \quad (3)$$

$$V_{G_p, G_u, D_{pu}}(G, D) = E_{x \sim p_p(x)} \lg(D_{pu}(G_p(x))) + E_{z \sim p_u(z)} \lg(1 - D_{pu}(G_u(z))). \quad (4)$$

在此训练过程中, 正样本判别器 D_p 、无标签样本判别器 D_u 、正样本/无标签样本判别器 D_{pu} 、正样本生成器 G_p 和无标签样本生成器 G_u 均由深度神经网络训练得到. 整个模型的训练过程如算法 1 所示.

算法 1. puGAN 模型训练算法.

输入: 正样本数据集和无标签样本数据集;

输出: 正样本/无标签样本判别器模型.

① FOR 训练次数

② FOR 迭代次数

③ 从均匀分布 $p_{zu}(z)$ 中抽取 $\{z_u^{(1)}, z_u^{(2)}, \dots, z_u^{(m)}\}$ 共 m 个随机数作为无标签样本生成器的输入;

④ 从均匀分布 $p_{zp}(z)$ 中抽取 $\{z_p^{(1)}, z_p^{(2)}, \dots, z_p^{(m)}\}$ 共 m 个随机数作为正样本生成器的输入;

⑤ 从无标签样本的先验分布 $p_u(x)$ 中随机抽取 $\{x_u^{(1)}, x_u^{(2)}, \dots, x_u^{(m)}\}$ 共 m 个样本;

⑥ 从正样本的先验分布 $p_p(x)$ 中随机抽取 $\{x_p^{(1)}, x_p^{(2)}, \dots, x_p^{(m)}\}$ 共 m 个样本;

⑦ 计算无标签样本判别器的梯度值, 并以此梯度进行梯度下降计算

$$\nabla_{\theta_{D_u}} \frac{1}{m} \sum_{i=1}^m [\lg D_u(x_u^{(i)}) + \lg(1 - D_u(G_u(z_u^{(i)})))];$$

⑧ 计算正样本判别器的梯度值, 并以此梯度进行梯度下降计算

$$\nabla_{\theta_{D_p}} \frac{1}{m} \sum_{i=1}^m [\lg D_p(x_p^{(i)}) + \lg(1 - D_p(G_p(z_p^{(i)})))];$$

⑨ 计算正样本/无标签样本判别器的梯度值, 并以此梯度进行梯度下降计算

$$\nabla_{\theta_{D_{pu}}} \frac{1}{m} \sum_{i=1}^m [\lg D_{pu}(G_p(z_p^{(i)})) + \lg(1 - D_{pu}(G_u(z_u^{(i)})))];$$

⑩ END FOR

⑪ 再次从均匀分布 $p_{zu}(z)$ 中抽取 $\{z_u^{(1)}, z_u^{(2)}, \dots, z_u^{(m)}\}$ 共 m 个随机数作为无标签样本生成器的输入;

⑫ 再次从均匀分布 $p_{zp}(z)$ 中抽取 $\{z_p^{(1)}, z_p^{(2)}, \dots, z_p^{(m)}\}$ 共 m 个随机数作为正样本生成器的输入;

⑬ 计算无标签样本生成器以及正样本生成器的梯度值, 并以此梯度更新生成器

$$\nabla_{\theta_{G_u}} \frac{1}{m} \sum_{i=1}^m [\lg(1 - D_u(G_u(z_u^{(i)})))];$$

$$\nabla_{\theta_{G_p}} \frac{1}{m} \sum_{i=1}^m [\lg(1 - D_p(G_p(z_p^{(i)})))];$$

⑭ END FOR

在此训练过程中, 噪声普遍存在于正样本和无标签样本数据中, 因此, 本文分别对正样本和无标签样本引入了生成器与判别器, 从而调整并学习得到 2 个样本的分布特征. 随后, 使用无标签样本生成器和正样本生成器分别生成伪无标签样本和伪正样本数据, 并根据此数据进行分类学习, 最终得到了 POI 定位的分类模型.

2.2 理论分析

命题 1. 对于给定的 G_u 与 G_p , 最优化的分类器 D_u^*, D_p^*, D_{pu}^* 为

$$D_u^*(x) = \frac{p_u(x)}{p_u(x) + p_{gu}(x)}, \quad (5)$$

$$D_p^*(x) = \frac{p_p(x)}{p_p(x) + p_{gp}(x)}, \quad (6)$$

$$D_{pu}^*(x) = \frac{p_p(x)}{\phi_u p_{gu}(x) + \phi_p p_{gp}(x)}. \quad (7)$$

证明. 对于任何一个给定的生成器, 分类器 D_u, D_p, D_{pu} 的估值函数方程为

$$V(G_u, D_u) = \int_x p_u(x) \lg(D(x)) dx + \int_z p_z(z) \lg(1 - D(G(z))) dz, \quad (8)$$

$$V(G_p, D_p) = \int_x p_p(x) \lg(D(x)) dx + \int_z p_z(z) \lg(1 - D(G(z))) dz, \quad (9)$$

$$V(G_p, G_u, D_{pu}) = \int_x p_p(x) \lg(D(G(x))) dx + \int_z p_u(z) \lg(1 - D(G(z))) dz. \quad (10)$$

通过合并最优估值函数,可以得到结论:

$$\begin{aligned}
 C(G) = \min_{\{G_p, G_u\}} \max_{\{D_p, D_u, D_{pu}\}} & V(G, D) = \alpha [E_{x \sim p_u(x)} \lg(D_u^*(x)) + E_{x \sim p_{zu}(x)} \lg(1 - D_u^*(G_u(x)))] + \\
 & \beta [E_{x \sim p_p(x)} \lg(D_p^*(x)) + E_{x \sim p_{zp}(x)} \lg(1 - D_p^*(G_p(x)))] + \\
 & \gamma [E_{x \sim p_p(x)} \lg(D_p^*(G_p(x))) + E_{x \sim p_u(x)} \lg(1 - D_u^*(G_u(x)))] = \\
 & \alpha [E_{x \sim p_u(x)} \lg \frac{p_u(x)}{p_u(x) + p_{gu}(x)} + E_{x \sim p_{zu}(x)} \lg \frac{p_{gu}(x)}{p_u(x) + p_{gu}(x)}] + \\
 & \beta [E_{x \sim p_p(x)} \lg \frac{p_p(x)}{p_p(x) + p_{gp}(x)} + E_{x \sim p_{zp}(x)} \lg \frac{p_{gp}(x)}{p_p(x) + p_{gp}(x)}] + \\
 & \gamma [E_{x \sim p_p(x)} \lg \frac{p_p(x)}{p_{gu}(x) + p_{gp}(x)} + E_{x \sim p_u(x)} \lg \frac{p_{gu}(x) + p_{gp}(x) - p_p(x)}{p_{gu}(x) + p_{gp}(x)}]. \quad (11)
 \end{aligned}$$

证毕.

定理 1. 当且仅当 $p_u = p_{gu}$, $p_p = p_{gp}$, 且 $2p_p = \phi_u p_{gu} + \phi_p p_{gp}$ 时, 目标函数可以达到全局最小值.

证明. 当满足上述条件时, 可以得到:

$$E_{x \sim p_u(x)} [-\lg 2] + E_{x \sim p_{zu}(x)} [-\lg 2] = -\lg 4, \quad (12)$$

$$E_{x \sim p_p(x)} [-\lg 2] + E_{x \sim p_{zp}(x)} [-\lg 2] = -\lg 4, \quad (13)$$

$$E_{x \sim p_{zp}(x)} [-\lg 2] + E_{x \sim p_{zu}(x)} [-\lg 2] = -\lg 4. \quad (14)$$

将式(12)~(14)代入式(11), 可以得到:

$$\begin{aligned}
 C(G) = & \alpha [-\lg 4 + JS(p_p \parallel p_{gp})] + \\
 & \beta [-\lg 4 + JS(p_u \parallel p_{gu})] + \\
 & \gamma [-\lg 4 + JS(p_{gp} \parallel p_{gu})], \quad (15)
 \end{aligned}$$

其中, JS 为分布之前的 JS 散度.

所以, 当满足条件 $p_u = p_{gu}$, $p_p = p_{gp}$ 且 $2p_p = \phi_u p_{gu} + \phi_p p_{gp}$ 时, 可以取得目标函数的全局最优值.

证毕.

3 实例分析

3.1 数据集

本文使用的是脱敏后某高校校园无线网络接入日志数据, 日志提供了对设备信息加密后的唯一标识及扫描得到的 WiFi 信息列表. 具体的 WiFi 数据格式为

$$\langle id_1, sig_1 | id_2, sig_2 | \dots | id_m, sig_m \rangle,$$

其中, id_i 表示第 i 个发射 WiFi 信号设备的标识, sig_i 为第 i 个 WiFi 所对应的信号强度.

数据集共包含 176 449 条数据, 覆盖了 210 个 WiFi 接入点和 150 个 POI. 其中, 正样本 15 888 条, 无标签样本 160 561 条. 本文将有连接行为的数据标记为该 WiFi 接入点的正样本数据, 并与对应的 POI 进行关联, 将与其他 POI 关联的正样本数据作为该 POI 的负样本数据, 将无连接行为的数据作为无标签数据. 通常来说, 一个 POI 可能对应多个 WiFi 接入点, 同时每个 WiFi 接入点对应着多条正

样本和无标签样本数据. 通过对数据集进行初步分析, 可以知道正样本大约只有总数据量的 9%, 而剩余的 91% 均为无标签样本. 本文又对正样本数据进行深入统计和分析, 得到了它的统计规律如图 3 所示:

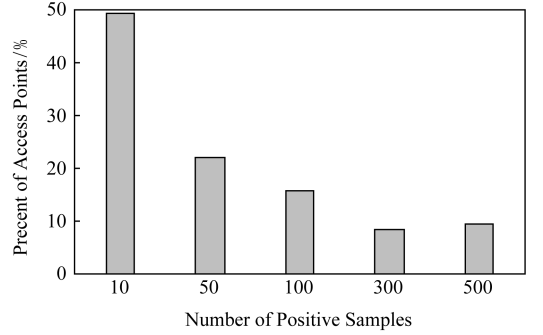


Fig. 3 The statistics of positive samples

图 3 正样本数据统计图

从图 3 可以发现, 47% 的 WiFi 接入点拥有的正样本数在 10 条以下, 而正样本数量在 300 条以上的 WiFi 接入点仅仅占比 9%. 由此可见, 过半的 WiFi 接入点的正样本缺失情况严重. 此情况在一个 POI 具有多个 WiFi 接入点的时候尤为明显.

在实验中, 由于采集得到的负样本在物理空间中通常偏于一侧, 为了避免训练结果出现偏差, 本文使用正样本和负样本作为评估集, 使用正样本和无标签样本作为训练集.

在训练过程中, 由于数据本身特征维度相对较高, 在样本偏少的情况下, 传统的分类模型通常会因训练不足而难以得到理想的样本分布情况. 为了证明 puGAN 模型的有效性, 本文进行了 2 组对比实验: 第 1 组实验对比了 OC、SVM^[19]、神经网络(neural network, NN)^[20]以及 puGAN 这 4 种模型的效果; 第 2 组实验对比了 GAN 模型在不同使用条件下的效果. 下面将分别阐述这 2 组实验.

3.2 传统模型与 puGAN 模型对比

实验通过调整模型不同的超参数进行交叉验证训练,得到了不同准确率下对应的数据召回率.根据此准确率和召回率,使用 ROC 曲线来对模型效果进行比较与分析.如图 4 所示是不同模型在实验数据集上的表现.

图 4 中,横轴表示模型的召回率,纵轴表示模型的准确率,ROC 面积越大表示模型效果越好.

从图 4 可以看出,单类模型的 ROC 面积为 0.761,在准确率为 80%时召回率不足 60%,效果明显低于其他的几类模型.SVM 与神经网络(NN)模型在 ROC 曲线的评估上效果相当,当准确率为 80%时召回率均为 77%左右.而 puGAN 模型在准确率为 80%时召回率达到了 85%左右,对应的 ROC 面积也明显高于其他模型.

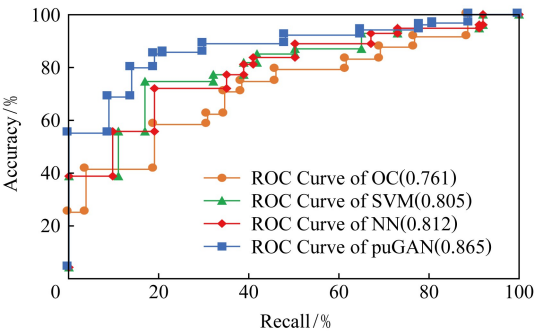


Fig. 4 Comparison of ROC curves among OC, SVM, NN and puGAN

图 4 OC,SVM,NN 和 puGAN 的 ROC 曲线对比图

此外,我们还观察了这 4 种模型的训练误差和测试误差随迭代次数增加的变化,这种变化反应模型的拟合情况,如图 5 所示:

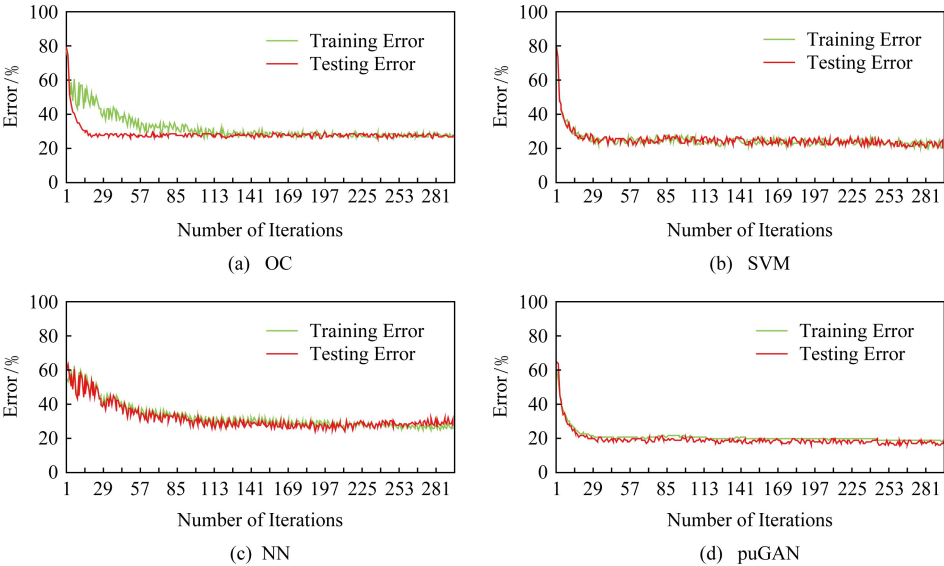


Fig. 5 Training error and Testing error among OC, SVM, NN and puGAN

图 5 OC,SVM,NN 和 puGAN 训练误差和测试误差分析对比图

由图 5 可知,单类模型明显存在着欠拟合的情况,经过多轮的迭代,训练误差始终维持在 26%左右,整体效果较差.SVM 和神经网络的表现明显优于单类模型,训练误差维持在 21%左右.但随着迭代次数的增加,由于神经网络结构复杂,它的测试误差不断上升,出现了过拟合的情况.puGAN 模型在训练初始阶段,收敛速度会略慢于 SVM 以及神经网络,但在最终效果却明显优于上述 2 类模型,且模型表现稳定,训练误差和测试误差均维持在 15%左右.

综上,由于单类模型在训练过程中只利用了正样本进行学习,而忽略了大量的无标签样本,所以整体实验效果均不佳.SVM 与神经网络均使用了正样

本和无标签样本进行二分类学习,SVM 模型结构较神经网络简单,因此收敛速度较快,在准确率和召回率上二者差别不大.puGAN 对正样本和无标签样本进行学习,生成伪正样本来弥补数据不足的问题,同时生成伪无标签样本来提高判别器对正样本的识别能力.因此,从 ROC 曲线以及模型拟合情况来看,puGAN 的效果明显优于传统的机器学习模型.

3.3 3 种 GAN 模型对比

第 2 组实验中,我们对比分析了 puGAN 模型与另外 2 种未校正模型在准确率、召回率以及训练误差、测试误差等指标上的表现情况.其中未校正的模型包括未对正样本建模(本文称 GAN1)以及未对

无标签样本建模(本文称 GAN2)这 2 种情况.

如图 6 所示,puGAN 模型的 ROC 面积略高于 GAN1 和 GAN2 这 2 种未校正的模型.在这 3 种模型中,GAN1 模型面积最小,其值只有 0.835.该模型由于未能够对正样本进行分布上的学习,使得训练得到的模型在准确率以及召回率上的表现不佳,得到的分类面也偏于无标签样本数据一方.GAN2 模型虽然也受到了训练样本不均匀的影响,但是在准确率和召回率上的表现却略好于 GAN1 模型,原因是 GAN2 模型对正样本进行了建模和重采样,使其受到的正样本数据不足的影响略低于 GAN1 模型.

此外,我们也同样观察了上述 3 种模型训练误差和测试误差的变化,并对它们的拟合情况进行了分析.如图 7 所示为 3 种模型的训练误差以及测试误差随着迭代次数增加的变化情况.从图 7 中可以看出,GAN1 模型在训练初期存在过拟合的情况,但随

着迭代次数的增加,过拟合逐渐缓解,效果趋于稳定. GAN2 模型在训练过程中整体效果较为稳定,但是误差较大,始终维持在 20%以上.而 puGAN 模型最终训练误差和测试误差均在 15%左右且效果稳定.

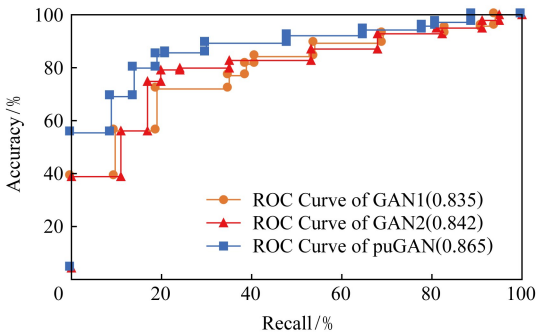


Fig. 6 Comparison of ROC curves among GAN1, GAN2 and puGAN

图 6 GAN1,GAN2 和 puGAN 的 ROC 曲线图

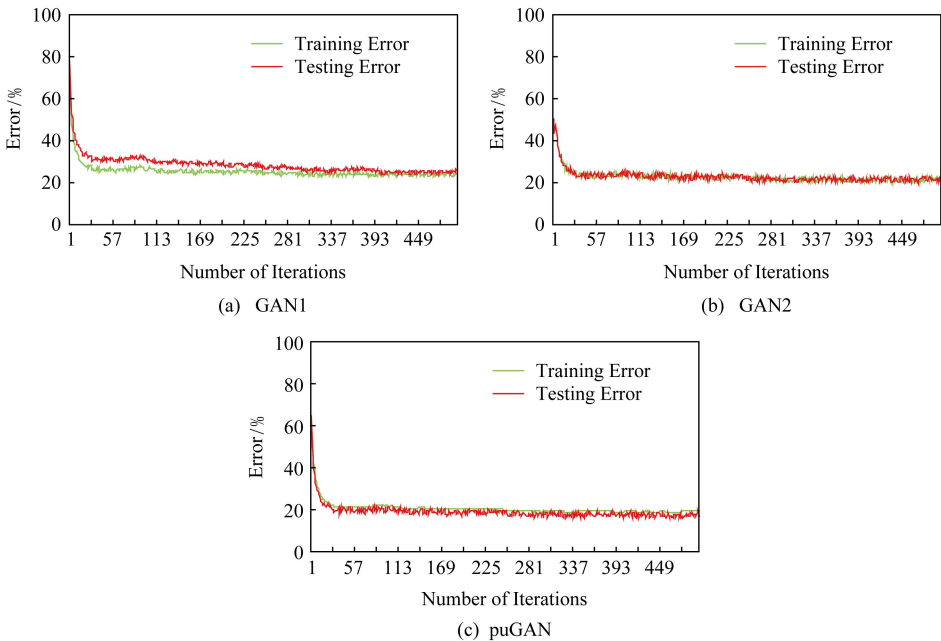


Fig. 7 Training error and Testing error among GAN1, GAN2 and puGAN

图 7 GAN1,GAN2 和 puGAN 训练误差和测试误差分析对比图

综上所述,在 GAN 的 3 类模型中,puGAN 模型在 ROC 面积以及稳定性上表现更为突出.

4 结 论

本文讨论了 puGAN 模型在 POI 定位中的应用.实验表明,该模型通过对正样本和无标签样本进行建模和生成,提升了 POI 定位的准确率和召回率,从而有效解决了数据样本不足以及半监督学习

中噪声过大的问题.

实验还表明,该模型能够避免复杂模型所带来的过拟合问题,且具有更好的稳定性.此研究为现有的 POI 定位系统提供了新的指导意义.

虽然 puGAN 模型在准确率、召回率和稳定性上表现优秀,但是却存在迭代速度较慢的缺点.在大数据的背景下,提升模型训练速度也早已成为了一个重要的课题.在今后的研究中,我们会改进训练过程中最优化的方法来提升模型训练的效率.

参 考 文 献

[1] Gao Rong, Li Jing, Du Bo, et al. A synthetic recommendation model for point-of-interest on location-based social networks: Exploiting contextual information and review [J]. Journal of Computer Research and Development, 2016, 53(4): 752-763 (in Chinese)
(高榕, 李晶, 杜博, 等. 一种融合情景和评论信息的位置社交网络兴趣点推荐模型[J]. 计算机研究与发展, 2016, 53(4): 752-763)

[2] Chen Cheng, Yang Haiqin, Lyu M R, et al. Where you like to go next: Successive point-of-interest recommendation [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 2605-2611

[3] Gao Huiji, Tang Jiliang, Hu Xia, et al. Content-aware point of interest recommendation on location-based social networks [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 1721-1727

[4] Yin Hongzhi, Zhou Xiaofang, Cui Bin, et al. Adapting to user interest drift for POI recommendation [J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(10): 2566-2581

[5] Wen Yutian, Tian Xiaohua, Wang Xinbing, et al. Fundamental limits of RSS fingerprinting based indoor localization [C] //Proc of the IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2015: 2479-2487

[6] Zhou Mu, Zhang Qiao, Xu Kunjie, et al. PRIMAL: Page rank-based indoor mapping and localization using gene-sequenced unlabeled WLAN received signal strength [J]. Sensors, 2015, 15(10): 24791-24817

[7] Zhang Wei, Hua Xianghong, Yu Kegen, et al. Radius based domain clustering for WiFi indoor positioning [J]. Sensor Review, 2017, 37(1): 54-60

[8] Matos P P D, Afonso A P, Carmo M B. Point of interest awareness using indoor positioning with a mobile phone [C/OL] //Proc of the 1st Int Conf on Pervasive and Embedded Computing and Communication Systems, 2011 [2018-12-19]. https://www.researchgate.net/publication/221156937_Point_of_Interest_Awareness_using_Indoor_Positioning_with_a_Mobile_Phone

[9] Moghtadaiee V, Dempster A G. WiFi fingerprinting signal strength error modeling for short distances [C] //Proc of the Int Conf on Indoor Positioning and Indoor Navigation. Piscataway, NJ: IEEE, 2013: 1-6

[10] Elkan C, Noto K. Learning classifiers from only positive and unlabeled data [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 213-220

[11] Du Plessis M C, Niu Gang, Sugiyama M. Analysis of learning from positive and unlabeled data [C] //Proc of the 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 703-711

[12] Chapelle O, Lkopf B S, Zien A. Semi-Supervised Learning [M]. Cambridge, MA: MIT Press, 2006

[13] Ren Yafeng, Ji Donghong, Zhang Hongbin, et al. Deceptive reviews detection based on positive and unlabeled learning [J]. Journal of Computer Research and Development, 2015, 52(3): 639-648 (in Chinese)
(任亚峰, 姬东鸿, 张红斌, 等. 基于 PU 学习算法的虚假评论识别研究[J]. 计算机研究与发展, 2015, 52(3): 639-648)

[14] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc of the 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680

[15] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans [J]. arXiv preprint, arXiv: 1606.03498, 2016

[16] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data [J]. arXiv preprint, arXiv: 1610.05755, 2016

[17] Odena A. Semi-supervised learning with generative adversarial networks [J]. arXiv preprint, arXiv: 1606.01583, 2016

[18] Arjovsky M, Chintala S, Bottou L. Wasserstein gan [J]. arXiv preprint, arXiv: 1701.07875, 2017

[19] Mordelet F, Vert J P. A bagging SVM to learn from positive and unlabeled examples [J]. Pattern Recognition Letters, 2010, 37(1): 201-209

[20] Singh G, Sachan M. Multi-layer perceptron (MLP) neural network technique for offline handwritten gurmukhi character recognition [C] //Proc of the IEEE Int Conf on Computational Intelligence and Computing Research. Piscataway, NJ: IEEE, 2015: 1-5



Tian Jiwei, born in 1993. MSc candidate. Student member of CCF. Her main research interests include data mining, wireless network application.



Wang Jinsong, born in 1970. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include network and information security, computer network.



Shi Kai, born in 1981. PhD, associate professor, MSc supervisor. Member of CCF. His main research interests include network and information security. (shikai0229@163.com)