

虚拟样本生成技术研究

于 旭 杨 静 谢志强

(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

摘 要 虚拟样本生成技术主要研究如何利用待研究领域的先验知识并结合已有的训练样本构造辅助样本,扩充训练样本集,提高学习器的泛化能力。作为一种在机器学习中引入先验知识的方法,虚拟样本生成技术已经成为提高小样本学习问题泛化能力的主要手段之一,受到了国内外学者广泛研究。首先介绍了虚拟样本的概念,给出了衡量虚拟样本生成技术性能的两个指标,讨论了虚拟样本生成技术对学习器泛化能力的影响。然后根据虚拟样本生成技术的本质将其划分为 3 类,并针对每一类讨论了几种典型的虚拟样本生成技术,进而指出了现有虚拟样本生成技术存在的一些不足。最后进行总结并对虚拟样本生成技术的进一步发展提出了自己的看法。

关键词 虚拟样本,先验知识,泛化能力,机器学习,小样本学习问题

中图法分类号 TP181 **文献标识码** A

Research on Virtual Sample Generation Technology

YU Xu YANG Jing XIE Zhi-qiang

(Institute of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract Virtual sample generation technology mainly makes research on how to combine the priori knowledge of a domain with the existing training samples to create additional training samples, enlarging the training dataset, and improving the generalization ability of learning machine. As a method of bringing priori knowledge to machine learning, virtual sample generation technology has become a primary means to improve the generalization ability on small sample learning problem, and has been studied widely by scholars home and abroad. In this paper, firstly the concept of virtual sample was introduced, two indexes of evaluating the capability of virtual sample generation technology were given, and the influence of virtual sample generation technology on the generalization ability of learning machine was discussed. Then the current virtual sample generation technologies were divided into 3 classes by their essence, and for each class several typical virtual sample generation technologies were discussed. Moreover the deficiencies of the current virtual sample generation technology were pointed out. Finally, made a summary and proposed our own viewpoint on the further development of virtual sample generation technology.

Keywords Virtual samples, Prior knowledge, Generalization ability, Machine learning, Small sample learning problem

1 引言

充足的训练样本是提高传统有监督学习泛化能力的重要保障。然而,获取充足的训练样本通常需要耗费大量的人力物力^[1],因此如何在少量训练数据下,提高传统学习器的泛化能力,就成了一个值得研究的课题。训练样本数据很少的情况下的机器学习问题被称为小样本学习问题^[2]。为了提高小样本问题上的学习能力,近些年产生了许多优秀的方法,如半监督学习^[3-5]、主动学习^[6,7]以及直推式支持向量机^[8]。这些方法的一个共同特点是需要大量的未标记样本作为学习的辅助样本,而这个要求在某些时候也是很难满足的。因此,1992 年 Poggio 和 Vetter 首次提出了虚拟样本的思想^[9]。虚拟样本是指在未知样本概率分布函数的情况下,利用所研究的领

域的先验知识,结合已有的训练样本产生待研究问题的样本空间中的部分合理样本。研究虚拟生成技术的主要目的是生成合理的虚拟样本,从而可以将它们添加到原始训练样本集,以扩充训练样本集,有效提高分类器的泛化能力。所以虚拟样本生成技术就成了一种提高小样本分类问题分类精度的有效手段。

自从 Poggio 和 Vetter 提出虚拟样本概念以来,虚拟样本生成技术引起了广大学者的关注,并在很多机器学习领域,尤其是小样本学习领域得到了广泛的应用。纵观国内外虚拟样本生成技术领域的研究情况,至今还未见有这方面的综述性文章。目前对于虚拟样本生成技术的研究主要强调在某一特定领域内如何利用先验知识构造合理的虚拟样本,而没有对虚拟样本生成技术有一个整体性的、本质的分析。因此,我们

到稿日期:2010-04-09 返修日期:2010-08-24 本文受国家自然科学基金(60873037,61073041,61073043)和黑龙江省自然科学基金(F200901)资助。

于 旭(1982—),男,博士生,主要研究方向为数据挖掘、机器学习,E-mail:yuxub09@hrbeu.edu.cn;杨 静(1962—),女,教授,博士生导师,主要研究方向为数据库与数据挖掘;谢志强(1962—),男,教授,主要研究方向为 CIMS 及数据挖掘。

认为有必要对该领域作一个全面的介绍,并对一些经典的虚拟样本生成技术在合理性及适应性上进行分析与总结。

本文作为一篇综述性文章,旨在向读者介绍国内外各领域的虚拟样本生成技术,使研究人员尽可能详细地总结该领域的研究现状。本文第2节介绍了虚拟样本的概念,给出了衡量虚拟样本生成技术性能的两个指标,讨论了虚拟样本生成技术对学习器泛化能力的影响;第3节通过深刻分析现有虚拟样本生成技术的核心思想将其划分为3类,并针对每一类讨论了几种典型的虚拟样本生成技术,进而指出了现有虚拟样本生成技术存在的一些不足;最后对全文进行总结并对虚拟样本生成技术的进一步发展提出了自己的看法。

2 虚拟样本分析

2.1 虚拟样本定义与实质

虚拟样本通常被译为 Virtual Sample,又可译为 Synthetic Sample 或 Artificial Sample,1992年由 Poggio 和 Vetter 首次提出。通常认为虚拟样本是指通过利用待研究领域的先验知识,结合已知的训练样本产生的辅助样本。由于 Poggio 和 Vetter 并没有明确给出虚拟样本的定义,因此本文对虚拟样本的定义总结如下。

虚拟样本:对于给定训练样本 (x, y) ,若由变换 T 得到的样本 $(Tx, y_T f(x))$ 也是一个合理的样本,那么就称新得到的样本 $(Tx, y_T f(x))$ 是由变换 T 生成的虚拟样本。其中变换 T 是根据所研究领域的先验知识得到的,其实就是一种虚拟样本生成方法。

因此给定由 n 个样本组成的训练样本集 $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$,通过某种合适的变换 T 可以生成虚拟样本集合 $D' = \{(x_1', y_1'), \dots, (x_n', y_n')\}$,其中 $x_i' = Tx_i, y_i' = y_T f(x_i)$ 。

类似于其它在学习过程中加入先验知识的方式,如利用先验知识选取特征,将先验知识嵌入到学习算法中,以及结合先验知识作为提示等^[10],虚拟样本实质也是一种在学习过程中加入先验知识的有效方式。

2.2 评价标准及分析

自从虚拟样本思想产生以来,虚拟样本生成技术在各机器学习领域得到了广泛的研究。为了更好地评价这些虚拟样本生成技术,本文给出合理性标准和适应性标准两个评价指标,具体含义如下。

定义1(合理性标准) 通过虚拟样本生成技术产生的虚拟样本是真实特征空间的样本的概率期望。

定义2(适应性标准) 虚拟样本生成技术适应的领域的广泛性。

当然合理性与适应性往往不能很好地兼顾。若要提高虚拟样本生成技术的合理性,最根本的办法是加深对所研究领域的把握,有效提炼该领域的先验知识,构造出合理的虚拟样本生成技术。而这样却很容易导致提出的虚拟样本生成技术适应性降低。反之,如果提出一种对于较多领域都具有普遍意义的虚拟样本生成技术,使得适应性得到了提高,往往会导致生成技术在合理性上有一定的下降。

2.3 虚拟样本提高泛化能力的原因

Niyogi 等在文献[10]中通过将分类器限制在神经网络分

类器和高斯径向基函数网络分类器,讨论了影响学习器泛化能力的因素。指出泛化误差可以被分解为近似误差和估计误差,并说明如果利用未受限制的假设集,那么就需要大量的样本才能保证达到低的估计误差。另一方面,如果用受限制的假设集,那么近似误差就会很大。

本文根据统计学习理论^[11]将文献[10]中对分类器的限制进行了推广,对基于训练样本的学习问题的泛化能力进行了综合的分析。根据统计学习理论,学习机器的实际风险 R 是由两部分组成的:一是经验风险 R_{emp} ;另一部分称作置信范围 Φ ,它和学习机器的VC维数及训练样本数 n 有关。可以简单地表示为: $R \leq R_{\text{emp}} + \Phi(h/n)$,其中 Φ 随着 h/n 的增加而单调增加, R_{emp} 随着 h 增加而减小,因此可以很容易地得出和 Niyogi 等同样的结论。

通过上述对分类器泛化能力的综合分析可知,为有效处理近似误差与估计误差之间的矛盾,唯一的方法就是使目标函数集缩小。因此,如果能够得到关于目标函数的更多的先验知识,就可以缩小目标函数集,泛化能力也就会得到相应的提高。没有免费的午餐定理也说明了结合先验知识或许是学习器从有限的样本学习中获得较强泛化能力的唯一途径。虚拟样本作为一种利用先验知识的方法,可以提高泛化能力。Niyogi 等在文献[10]中进一步说明虚拟样本在数学上等价于结合先验知识作为正则化项^[10]。

3 虚拟样本生成技术实质分析与评价

自从1992年 Poggio 和 Vetter 提出虚拟样本思想以来,虚拟样本技术在很多领域得到了广泛的应用,目前已经被成功应用到图像识别^[10,12]、语音识别^[10]、手写体数字识别^[13]、文本识别^[14]和噪声源识别^[15]等多个研究方向。由于虚拟样本应用的领域非常广泛,不同的领域甚至同一领域虚拟样本生成技术也有着很大的差异,没有普遍意义下的虚拟样本生成算法。因此本文不打算具体对每个虚拟样本生成算法进行详细的分析,而主要目的是通过综合分析虚拟样本生成技术,对它们进行归纳总结,挖掘出这些虚拟样本生成技术的实质,为进一步构造虚拟样本生成技术提供一定的指导。

通过对目前虚拟样本生成技术的分析,发现这些虚拟样本生成技术按照生成思想主要可以归纳为如下3类:

- (1) 基于研究领域具体的先验知识构造虚拟样本。
- (2) 基于扰动的思想构造虚拟样本。
- (3) 基于研究领域的分布函数构造虚拟样本。

按照生成思想(1)生成虚拟样本,要求所研究领域先验知识非常明显(例如知道分布函数为奇函数或偶函数后,可以容易地构造虚拟样本)或者研究者对研究领域有深刻认识,能够很好地把握先验知识,构造出合理的虚拟样本。

生成思想(2)、(3)主要是从训练样本出发,通过对训练样本进行扰动或者从训练样本中学习分布函数,尽可能挖掘出研究领域的部分先验知识,因此比较适应于研究领域先验知识不明显或难以把握的情况。

下面对这几类虚拟样本生成技术进行深入的分析。

3.1 基于研究领域具体的先验知识

基于研究领域具体的先验知识构造虚拟样本的重要工作首先来自于 Poggio。Poggio 是机器视觉方面的专家,对图像识别有着很深刻的理解。因此在提出虚拟样本思想后, Pog-

gio 首先在图像识别领域给出了一种虚拟样本生成技术,提出了基于几何变换的虚拟样本生成技术。Poggio 虚拟样本生成的主要思想是对于给定的一个物体的三维视角,通过几何变换从任何其它的角度生成虚拟样本。

在 Poggio 提出虚拟样本概念并利用几何变换在图像识别领域得到实现之后,很多专家学者也在图像识别领域进行了很多研究。国内学者在文献[12]中提出了基于特征不变法的虚拟样本生成方法。所谓特征不变法是指对于需扩充的样本,保持其表情特征不变,通过改变其姿势或旋转一定角度,创建虚拟样本。其主要原理是把人脸矢量化,将人脸图像看作一个二维图像,分别用图像的形状和纹理矢量描述该图像的几何特征,利用现有原型脸计算人脸角度变化后其特征矢量的变化关系,然后用原型脸的特征矢量来表示对于需创建的虚拟样本的人脸,利用与原型脸的特征矢量变换的对应关系构造出新样本的特征矢量。

基于生成思想(1)的虚拟样本生成技术除了在图像识别领域得到了成功的应用,在文本识别和噪声源识别方面也得到了成功的应用。王晓东等在文献[14]中提出了一种基于本体驱动文本虚拟样本构造方法。在确保类别不变性的前提下,该方法依据领域相关本体所明晰表达的领域知识,基于本体树的点、边、子树,从同义、父子、语义同构的多个词义关系角度实现了文本虚拟样本的构造。徐荣武等在文献[15]中利用虚拟样本技术对双层圆柱壳体结构噪声源识别这一实际问题进行了研究,提出了基于频响函数和傅里叶变换的两种虚拟样本生成算法。

3.2 基于扰动思想

上面所讨论的基于生成思想(1)的几种虚拟样本生成技术,在各自领域都获得了成功的应用。主要是因为研究者对该领域有很深刻的认识,很好地把握了该领域的先验知识,从而合理地生成了虚拟样本。然而很多时候某些领域先验知识不是很明显,或者说研究者对于该领域不是很熟悉,因此很难通过把握研究领域具体的先验知识来构造虚拟样本生成技术。在这种情况下,部分研究者从训练样本出发,利用扰动的思想生成虚拟样本。主要工作如下。

Bishop 在文献[16]中指出,样本输入中添加噪声也可以提高学习方法的泛化能力,并且在噪声标准差较小时,Bishop, An 等人已经证明,在输入样本中添加噪声,等价于神经网络结构设计的正则化方法^[17],而正则化系数则与噪声标准差有关^[16,18]。

王卫东在文献[19]中为解决小样本二次判别分析的各类协方差矩阵通常是奇异的问题,提出了对训练样本进行扰动的方法,以此产生新的虚拟样本。也就是在每个训练样本的各维数据上增加一个扰动 Δ ,因此对于每个 p 维训练样本可以生成 p 个虚拟样本,达到了扩充样本集的目的,有效解决了协方差矩阵奇异的问题。

台湾学者 Li Der-chiang, Fang Yao-hwei 在文献[20]中利用组发现技术(group discovery technology)和超球特征方程提出了非线性虚拟样本生成技术(NVSG)。该方法主要步骤如下:首先找到两类训练样本间的最小距离,利用组发现技术将非线性可分训练数据划分成组,对划分后的每个分组,利用超球特征方程在超球范围内或者超球边界上均匀生成随机虚拟样本,实现了样本的扩充。其虚拟样本生成技术的实质是

对训练样本进行细化,对细化后的每个分组生成虚拟样本,这其实就是基于扰动的思想。

张莉、陈恭和在文献[21]中提出了一种基于 K 近邻算法的虚拟样本生成技术。该方法首先利用 K 近邻算法将训练数据划分成 P 组,然后对每一组中的训练样本两两求平均值。其中,若属性的取值是离散类型的,则取一组中出现次数最多的值;若属性的取值是连续的,则取算数平均。该虚拟样本生成技术实质上也是基于扰动的思想。

3.3 基于研究领域的分布函数

与基于扰动思想生成虚拟样本类似的是基于研究领域分布函数生成虚拟样本。该方法抛开领域具体的先验知识,从训练样本出发,首先确定样本的分布函数,然后构造虚拟样本。具体的研究工作如下。

台湾学者 Li Der-chang, Lin Yao-San 在文献[22]中提出了一种基于间隔核密度估计的虚拟样本生成方法。该方法主要步骤如下:首先利用间隔核密度估计方法(intervalized kernel density estimation)根据已知的小样本训练数据,估计出样本数据服从的概率密度函数。然后,根据估计得到的概率密度函数生成充足的虚拟样本。

大陆学者周志华在文献[23]中提出了 NEC4.5 算法,该算法有效克服了神经网络算法可理解性差的缺陷,并且在精确度上较决策树算法 C4.5 有很大的提高。NEC4.5 算法的主要过程是:首先利用小样本训练数据集训练神经网络集成。接着用得到的神经网络集成判断原始训练样本的新类标,用新类标代替原始类标产生新的训练集。此外,用训练得到的神经网络集成产生一些额外的训练样本,并把它们加入到新的训练集中。最后利用 C4.5 决策树算法在新的训练集上学习。

NEC4.5 的核心思想是,利用神经网络集成算法推广能力强、分类精度高的优势,得到一个分类决策函数。利用决策函数生成足够多的虚拟样本,从而有效地扩充了训练样本集,为 C4.5 算法在新样本集上达到良好的学习效果创造了条件。由于 NEC4.5 通过决策函数生成虚拟样本,因此本文将 NEC4.5 算法中的虚拟样本生成技术也归为基于分布函数思想的虚拟样本生成技术。

3.4 虚拟样本生成技术评价

对上述主要虚拟样本生成技术按照合理性与适应性进行总结,如表 1 所列。

表 1 各种虚拟样本生成技术合理性与适应性总结		
虚拟样本生成技术	合理性	适应性
Poggio 在 3D 图像识别中的虚拟样本生成技术	较高	较低
基于特征不变法的虚拟样本生成方法	较高	较低
基于本体驱动的文本虚拟样本构造方法	较高	较低
基于频响函数和傅里叶变换虚拟样本生成算法	较高	较低
输入训练数据上加噪声	一般	较高
基于数据扰动的虚拟样本生成技术	一般	较高
基于组发现技术和超球特征方程的非线性虚拟样本生成技术	一般	较高
基于 K 近邻算法的虚拟样本生成技术	一般	较高
基于间隔核密度估计	一般	较高
NEC4.5 算法中虚拟样本生成技术	一般	较高

如表 1 所列,基于生成思想(1)的几种虚拟样本生成技术,由于研究领域先验知识比较明显,或者研究者对研究领域有深刻认识,能够很好地把握先验知识,因此普遍合理性较

高,适应性较低。而基于生成思想(2)、(3)的几种虚拟样本生成技术从训练样本出发,抛开了领域具体的先验知识,虽然不能保证每个虚拟样本都是真实的,也就是说在合理性上较基于生成思想(1)的虚拟样本生成技术有所降低,但是适应的范围却比较大,因此适应性却有了一定的提高。所以对于虚拟样本生成技术而言,合理性与适应性是密切联系的。设计出合理性与适应性均高的虚拟样本生成技术,是未来虚拟样本研究的一个重要内容。

3.5 虚拟样本生成技术存在的一些不足

虚拟样本提出以来,在机器学习领域获得了广泛的应用,也取得了较好的效果,但其合理性与适应性的矛盾也一直存在。目前还没有出现较好解决这一问题的方法。

另外,虚拟样本的实质是利用领域的先验知识,根据已有的训练样本生成虚拟样本。这就必然造成生成的虚拟样本分布与原始样本分布不一致。而传统学习方法成立的前提是样本分布一致,因此传统的学习方法难以在新的样本集上获得良好的学习效果。

结束语 本文首先简要介绍了虚拟样本思想产生的背景和应用的领域,并指出了该研究的必要性。然后总结了虚拟样本的定义,并给出了虚拟样本生成技术的两个评价指标,讨论了虚拟样本技术对泛化能力的影响。在第3节,详细讨论了当前的虚拟样本生成技术的思想,并将它们按照生成思想分为3部分。按照适应性和合理性对各种虚拟样本生成技术进行了分析,并对这两个指标之间的关系做了一定的探讨。最后指出了目前虚拟样本生成技术存在的一些不足。

本文3.4、3.5小节已经指出,目前的虚拟样本生成技术在合理性与适应性上存在一定的矛盾。因此,设计出合理性与适应性均高的虚拟样本生成技术,是未来虚拟样本研究的一个重要内容。

另外,虚拟样本分布与原始样本分布不一致,影响了传统学习器的学习效果,因此虚拟样本思想在小样本学习问题上的优势在一定程度上受到了影响。因此,针对异分布数据设计合理的分类方法,使虚拟样本的优势得以充分的发挥,是需要进一步研究的问题。借鉴迁移学习思想^[24]从虚拟样本中自适应地选取最合适的虚拟样本,或许是解决该问题的一个可行方向。

参 考 文 献

- [1] Zhou Zhi-hua. Learning with unlabeled data and its application to image retrieval[C]//Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence, Guilin, China, 2006;5-10
- [2] Raudys S. Trainable fusion rules. II. Small sample-size effects[J]. Neural Networks, 2006, 19(10):1517-1527
- [3] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI, 1998;92-100
- [4] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data[C]//Proceedings of the 17th International Conference on Machine Learning, San Francisco, CA, 2000;327-334
- [5] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labeled and unlabeled data[A]// Mozer M, Jordan M I, Petsche T, eds. Advances in Neural Information Processing Systems 9[C]. Cambridge, MA: MIT Press, 1997; 571-577
- [6] Abe N, Mamitsuka H. Query learning strategies using boosting and bagging[C]//Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 1998;1-9
- [7] Seung H, Opper M, Sompolinsky H. Query by committee[C]//Proceedings of the 5th ACM Workshop on Computational Learning Theory, Pittsburgh, PA, 1992;287-294
- [8] Joachims T. Transductive inference for text classification using support vector machines[C]//Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, 1999; 200-209
- [9] Poggio T, Vetter T. Recognition and structure from one 2 D model view: observations on prototypes, object classes, and symmetries[R]. A. I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992
- [10] Niyogi P, Girosi F, Poggio T. Incorporating prior information on machine learning by creating virtual examples[C]//Proceedings of the IEEE, 1998;86;2196-2209
- [11] Vapnik V N. Statistical Learning Theory[M]. New York, Wiley, 1998
- [12] 温津伟, 罗四维, 赵嘉莉, 等. 通过创建虚拟样本的小样本人脸识别统计学习方法[J]. 计算机研究与发展, 2002, 39(7):814-818
- [13] Scholkopf B, Simard P, Smola A, et al. Prior knowledge in support vector kernels[C]//Advances in Neural Information Processing Systems, MIT Press, 1998
- [14] 王晓东, 郭雷, 方俊. 本体驱动的文本文档虚拟样本构造方法研究[J]. 计算机科学, 2008, 35(3):142-144
- [15] 徐荣武, 何琳, 章林柯, 等. 基于虚拟样本的双层圆柱壳体结构噪声源识别研究[J]. 振动与冲击, 2008, 27(5):32-35
- [16] Bishop C M. Training with noise is equivalent to Tikhonov regularization[J]. Neural Computation, 1995, 7(1):108-116
- [17] Tikhonov A T. Solution of ill-posed problems[M]. New York: Winston-wiley, 1977
- [18] An G. The effect of adding noise during back propagation training on a generalization performance[J]. Neural Computation, 1996(8):643-671
- [19] 王卫东, 杨靖宇. 采用虚拟训练样本的二次判别分析方法[J]. 自动化学报, 2008, 34(4):400-407
- [20] Li Der-chiang, Fang Yao-hwei. A non-linearly virtual sample generation technique using group discovery and parametric equations of hypersphere[J]. Expert Systems with Applications, 2009, 36(1):844-851
- [21] 张莉, 陈恭和. 入侵检测系统中训练样本的构造方法[J]. 计算机工程与应用, 2006, 42(28):145-146, 180
- [22] Li D C, Lin Y S. Using virtual sample generation to build up management knowledge in the early manufacturing stages[J]. European Journal of Operational Research, 2006, 175(1):413-434
- [23] Zhou Zhi hua, Jiang Yuan. Nec4. 5; neural ensemble based C4. 5[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(6):770-773
- [24] Pan S J, Yang Q. A survey on transfer learning[R]. HKUST-CS08-08. Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China, 2008