基于平衡采样的轻量级广告点击率预估方法

施梦圜[1], 顾津吉[2]
(1. 南京大学 软件新技术国家重点实验室, 南京 210093; 2. 百度中国有限公司 联盟研发部, 上海 210203)

摘 要: 类似 Google AdSense 这样的定向广告投放系统在过去十年得到了长足的发展和进步,在定向广告投放系统中,机器学习方法在广告点击率预估扮演着重要角色。目前,广告点击率预估模型中的训练数据逐渐呈指数级增长,越来越大的训练数据给模型的扩展性带来了极大的不便。很多有用的特征以及复杂的模型受限制于训练集规模而无法加入到模型之中。借鉴类别不平衡问题中的平衡采样策略,通过多次采样的负样本数据和集成学习,缩短训练时间,改善学习准确率。实验证明在采用了平衡采样之后,点击率预估效果和线上资源消耗都得到了优化。

关键词: 广告点击率; 机器学习; 计算广告学; 类别不平衡学习
中图分类号: TP181 **文献标志码**: A **文章编号**: 1001-3695(2014)01-0033-04
doi:10.3969/j.issn.1001-3695.2014.01.006

Balance-sampling based light-weighted advertisement CTR prediction method

SHI Meng-yuan[1], GU Jin-ji[2]
(1. *National Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*; 2. *Dept. of Union Development, Baidu Inc, Shanghai* 210203, *China*)

Abstract: Targeting advertisement system, such as Google AdSense, media.net, has grown dramatically in the past 10 years, and machine learning plays a significant role in advertisement click-through-rate(CTR) prediction. However, the size of training set grows exponentially as time passed by, which greatly impedes the extendibility of the model. Large number of powerful features and complex models couldn't be applied to CTR prediction due to the large scale of training set size. This paper adopted balanced sampling strategy to CTR prediction. After sampling small portion of negative examples multiple times and ensemble learning, it cut down the training time and improved prediction accuracy. Experiment tests show after balance sampling, prediction results is improved and computation resources are saved.

Key words: click-through rate(CTR); machine learning; computational advertising; class imbalance learning

0 引言

随着计算机的发展,百度、谷歌等大型公司会为具有一定访问量的网站发布或展示与网站内容相关的广告,从而将网站流量转换为收入,这种广告投放方式通常被称做定向广告联盟。近些年来,定向广告联盟逐渐成为大型公司的重要收入来源之一。2013 年,Google AdSense 的广告营收占广告总收入的 27%[1]。通常来说,联盟的收入主要通过下式进行计算:

$$收入 = 广告点击率 \times 广告主出价$$

从上式可以看到,整个联盟的收入主要由两个因素决定,一个是广告主出价,一个是广告本身的点击率。广告主的出价是在与广告主进行协商之后就确定的,无法进行优化。然而如果能准确地预估广告的点击率,就能发布点击率更高的广告,从而提高联盟的整体收入。因此广告点击率预估在整个联盟广告系统中扮演着举足轻重的作用。

随着机器学习技术的发展,这一技术被逐渐采用到广告点击率预估中,大量关于线上广告点击日志很大程度上反映了广告点击率信息,可以被当成是监督学习中的训练集;在训练集训练出模型之后,使用该模型对线上广告进行点击率预估。

然而,随着互联网的发展,联盟的广告点击日志也以一种令人吃惊的速度增长着。从表 1 中可以看到,Google 公司四个广告产品的点击日志(训练集)都达到了 10 TB 的数量级,越来越多的训练数据给广告点击率预估带来了诸多困难。首先模型的训练时间变长,让线上模型的更新周期变长;其次,训练集的增加使得模型训练会消耗大量的计算资源和存储资源,采用复杂模型以及新特征变得非常困难,直接影响点击率预估效果。因此如何缩短训练时间,删除无用的训练数据一直以来是学术界研究的课题[2~5]。

表 1 Google 广告产品的训练数据规模[6]

产品	压缩前大小/TB	训练数据/TB	压缩率
A	9.98	2.00	4.99x
B	2.67	0.71	3.78x
C	66.66	15.54	4.29x
D	61.93	17.24	3.59x

类别不平衡问题指的是训练样本的分布不均匀问题,具体地说就是某些类的样本远多于其他类的样本。通常定向广告联盟的广告点击日志中正样本(被点击的记录)和负样本(未被点击的记录)比例为 1∶1000,很显然这是一个类别不平衡问题。众多处理类别不平衡算法中,对训练数据进行下采样是一种被广泛接受和采纳的策略[7~9]。下采样能削减训练数据的大小,缩短训练时间,然而单纯的下采样无法全面利用大类样

收稿日期: 2013-05-14; **修回日期**: 2013-06-23

作者简介:施梦圜(1987-),江苏常州人,硕士研究生,主要研究方向为机器学习、信息检索等(shimy@lamda.nju.edu.cn);顾津吉(1987-),江苏苏州人,硕士研究生.

本提供的信息。平衡采样通过集成一组在下采样之后训练集上训练所得分类器,补回了下采样所造成的信息丢失。本文借鉴平衡采样思想,构建一组逻辑回归模型对进行点击率预估,称这一方法为平衡采样逻辑回归。

1 类别不平衡问题和下采样算法

类别不平衡问题是指训练样本的分布不均匀问题,如在人脸检测[10]中,照片中人脸的数据显然是大大少于背景数目的。类别不平衡问题在生活中非常常见,它会给机器学习的算法带来诸多困难。首先由于类别的不平衡性,稀有类的样本过少,很难真实地刻画稀有类的分布信息;同时由于正样本过于稀疏,很容易在特征空间中形成一些比较小的数据区块。Weiss[11]表明分类错误很多时候都会集中在一些很小的区块中间。其次,传统的机器学习算法在处理类别不平衡问题时都会有不同程度的下降。最后,以准确率为标准的评判标准通常会忽视稀有类的重要性和影响。以广告点击率为例,在一个千分之一的广告点击率下,即使认为所有的广告都不会被点击,它仍然有 99.9% 的准确率。这显然是不符合实际情况的。

通常解决类别不平衡问题的方法主要分为两类:a)从训练集角度出发,通过改变训练样本的分布,降低训练样本之间的不平衡性,提高分类效果;b)从算法的角度出发,针对算法在处理类别不平衡问题中所碰到的缺陷,有目的性地解决这些问题。基于训练集的算法最重要的策略就是上采样和下采样。上采样通过对稀有类增加数据来调整训练数据的类别不平衡性;下采样方法通过删除大类的数据从而改变训练数据的分布使得训练数据更加平衡。在点击率预估问题中,考虑到线上日志的数据量已经达到 TB 级别,上采样会给计算资源和存储资源带来更加沉重的消耗,因此本文不予考虑。最原始的下采样方法就是随机下采样[3],事实证明,这样的采样方法在一定程度上降低了类别不平衡性。在此基础上,很多启发式的下采样方法被提了出来,Hart[8]给出了 CNN 原则,提出了稳定子集的概念,并以此为采样原则对数据进行采样。Wilson[9]提出了 ENN 原则,如果一个大类样本最近的三个样本有两个是被标记为稀有样本的,则删除这一样本。在 ENN 的基础上,Laurikkala[12]提出了 NCL 方法,使用 ENN 在稀有类和大类同时进行样本的删除。Tomek[13]考虑到类别不平衡问题在决策边界中的不稳定性,使用样本之间距离来删除训练样本中决策边界的样本点来进行数据的下采样。

2 平衡采样逻辑回归

2.1 基于逻辑回归的点击率预估

逻辑回归是机器学习中被广泛使用的分类模型,它的输出值表示了某个样本被标记为正样本的概率,在定义上它和点击率(广告被点击的概率)是相吻合的。因此,在文献[14]中被提出后,它就被广泛采用到了各大联盟的点击率预估系统之中。由于联盟广告点击率系统中特征总数非常庞大,然而真正有用的特征较少,L1 范式的正则化项能有效地让模型变得稀疏,起到一个特征选择的作用。随后,微软提出的 OWLQN[15]算法有效地解决了 LBFGS[16]优化算法中 L1 范式不可微的问题,因此,L1 范式 + 逻辑回归的模型以及 OWLQN 算法的优化成了各大联盟系统的首选。本文也采用这样的模型进行训练和点击率预估,给定一个用户 u、网页 p、广告 a:

$$p(c|p,u,a)=\frac{1}{1+\exp(\sum_{i=1}^{d}w_i f_i(p,u,a))} \tag{1}$$

其中:$f_i(p,u,a)$表示对三元组(p,u,a)提取的第 i 维特征;w_i 表示在这一维度的权重;$p(c|p,u,a)$也就是用户 u 在网页 p 上点击广告 a 的概率,即广告点击率。

2.2 平衡随机森林和 easy ensemble

随机森林通过集成多个决策树在机器学习中得到了广泛的应用。它通过对原始数据的自采样、决策树训练过程中对特征空间的随机划分,使得模型训练过程中在样本和特征选择上都随机进行,让不同的决策树都有较高的准确率以及决策树之间较大的差异性,实验表明,它比普通单个决策树在泛化误差要好。然而与传统的机器学习算法一样,它在处理类别不平衡数据的时候很难。文献[17,18]发现,对于决策树这样的分类器,如果能人工地使用采样或者插值的方法将数据集调整到类别平衡时将会取得更好的分类效果,它也证明了下采样通常能取得比上采样更好的分类效果。根据上述理论,Chen 等人[19]提出了平衡随机森林方法,该方法对正类和反类分别进行不同力度重采样,使得采样后的数据类别平衡(这个过程称之为平衡采样)。最后在重采样多次后采用多数投票的方法进行集成学习,整个算法的流程如下所示。

算法 1 平衡随机森林算法

输入:稀有类样本 P,大类样本 N,其中$|P| << |N|$,随机森林中决策树的个数 T,每一轮采样的数据大小 n。

$i \leftarrow 0$

repeat:

1 $i \leftarrow i+1$;

2 从 P 中采样一批数据 p_i,从 N 中采样一批数据 n_i,使得$|p_i| = |n_i| = n$

3 使用随机森林中对某个特定决策树训练的方法训练一个决策树 $h_i(x)$

until $i = T$

输出:$H(x)=\begin{cases}\frac{1}{n}\sum_{i=1}^{T}h_i(x) & \text{回归}\\ \operatorname{sgn}(\sum_{i=1}^{T}h_i(x)) & \text{分类。}\end{cases}$

借鉴了平衡采样的策略,Liu 等人[20]提出了 easy ensemble 算法,该算法可以看成一种集成 + 集成学习思想,它使用 bagging 消除不同分类器之间的多样性,boosting 提高分类器的准确率,多次平衡采样弥补下采样所带来的负样本信息缺失,最终取得了比随机森林更加良好的表现。

2.3 平衡逻辑回归的点击率预估

上文介绍了平衡采样策略在处理类别不平衡问题时的广泛应用。它删除了大量的负样本数据,缩短训练时间;而训练多个分类器进行集成能有效地将下采样所带来的信息丢失补回。当将它应用到广告点击率预估问题中时,考虑到线上模型都是使用逻辑回归模型,本文提出了平衡逻辑回归算法进行点击率预估。平衡采样逻辑回归算法的流程如算法 2 所示,与平衡随机森林相比,两者之间有以下两点区别:

a)平衡逻辑回归在每一步采样之后,所采用的方法是逻辑回归,这一策略主要的考虑在于现有联盟所采用的基本都是逻辑回归算法,独立开发一套新的大规模并行机器学习算法需要消耗大量的时间。

b)平衡逻辑回归在采样过程中没有真正地做到类别平衡,每一轮都是以一定的采样力度 α 对负样本进行采样,使用这样的策略主要是出于运算和存储资源考虑。现有的定向广

告的点击率通常只有千分之一,这意味着正样本和负样本比值达到 1∶1 000,如果在每一轮都做到真正类别平衡,意味着要删除 99.9% 的负样本,这会删除大量的负样本,相应地,也需要训练大量的分类器进行集成以补回采样造成的信息丢失。这样的策略在现实点击率预估系统是不现实的,首先训练大量的分类器是一个非常消耗计算资源的过程,其次,线上加载大量模型会造成机器的 CPU 空闲时间大大减少,甚至造成系统的瘫痪。因此在平衡采样的逻辑回归算法中,通常只能加载若干(3 ~5 个)模型,因此也在每一轮采样过程中无法做到真正的平衡采样。

算法 2　平衡采样逻辑回归算法

输入:稀有类样本 P,大类样本 N,其中 $|P| \ll |N|$,逻辑回归分类器的个数 T,每一轮采样的数据大小 n,删除力度 $0 < \alpha < 1$。

$i \leftarrow 0$

repeat:

1 $i \leftarrow i + 1$;

2 从 N 中采样一批数据 n_i,使得 $|n_i| = (1 - \alpha)n$

3 使用采样的数据 n_i 和 P 作为训练集,使用逻辑回归训练,得到分类器(逻辑回归函数)$h_i(x)$

until $i = T$

输出:$H(x) = \begin{cases} \frac{1}{n}\sum_{i=1}^{T} h_i(x) & \text{回归} \\ \text{sgn}(\sum_{i=1}^{T} h_i(x)) & \text{分类。} \end{cases}$

3　实验结果

3.1　实验设置

实验采用国内最大的定向广告联盟——百度联盟的点击日志作为训练集训练逻辑回归模型,并使用这一模型在百度线上进行了多天线上实验。该数据集包含 150 多种特征,既有广告 ID、网站 ID 等 ID 类特征,也有广告物料等泛化类特征。由于所有类的特征都被离散化处理,因此每条记录的特征数目是非常庞大的。所有实验数据的准备和 OWLQN 算法的训练过程都在 Hadoop 集群上运行[21]。集群共有 4 000 个节点,每个节点由 8 核 ×2.4 GHz 的 CPU 以及 16 GB 内存和 12 TB 硬盘组成。每天产生原始的日志数据量达到 20 TB,在经过特征归一化、hash 索引后,压缩至 200 GB 的训练数据。

3.2　实验设计

实验主要从以下几个方面展开:a)作为一个广告点击率相关的策略,虽然它主要的目的在于缩短训练时间,减小线上模型,但是它不能在点击率预估的效果上有太大下降,要争取做到持平;b)在点击率预估效果没有太大下降的前提之下,考察这一算法在性能上的表现,包括计算资源和存储资源的节约;c)性能的提升可以带来模型的可扩展性,为了展示这一扩展性,一些过去无法被加入点击率预估的特征被采纳进来,并取得了良好的点击率预估效果。

在评判指标的选择中,采用了类别不平衡问题中最常见的评判标准 AUC。同时也在真实线上环境中做实验,观察策略对线上广告点击率造成的影响。

3.3　点击率预估效果实验

3.3.1　不同采样力度 α 对点击率预估效果影响

联盟广告的点击日志是一份非常冗余的日志,很多网页或广告很可能在广告点击日志中只有个位数的展现,无法反映广告点击率这样的数据占据了广告点击日志的很大部分,因此本文大胆地删除大量负样本数据。在实验中,考察四个删除力度 α(0%、90%、95%、99%),为了排除多个模型带来的影响,暂时只使用一个线上模型。

图 1 反映了在不同删除力度下,模型线下 AUC 的改变;表 2 是不同删除力度下和不删除数据相比线上广告点击率的变化。可以看到即使删除了 99% 的负样本,无论是线上的点击率还是模型线下评估的 AUC 都没有显著地下降。变化基本保持百分位,产生这一现象的主要原因还是广告点击日志中的数据有大量的冗余信息,很多低展现、无点击广告的信息不能提高模型的预估能力。因此,单边的平衡采样并不会对广告点击率造成太大的影响。出于保险的考虑,在下文的实验中都是用 90% 的删除力度。

表 2　不同删除力度下线上广告点击率

	删除力度 α		
	90%	95%	99%
点击率变化	-0.5%	-1.08%	-2.01%

3.3.2　模型个数对点击率预估效果的影响

在平衡采样的逻辑回归过程中,模型个数是非常重要的参数。模型个数过少则线上点击率预估效果会下降,模型个数过多则会消耗过多的计算、存储资源。图 2 是删除力度为 90% 的情况下,采用不同模型个数线下 AUC 的变化。可以看到,整体上 AUC 是随着的模型个数的增加而逐渐增加,但是这样的趋势逐渐放缓。模型个数为 7 和 8 的 AUC 基本持平,造成这样的现象主要在于逻辑回归本身是一种比较稳定的模型,虽然对样本的采样能造成一定的不稳定性,然而随着模型个数的增多,这样不稳定性逐渐被消除,最终趋于统一。表 3 反映了加载 1、3、5 个模型和不进行采样只使用一个模型的线上的点击率变化,可以看到,线上实验和线下的 AUC 评估效果基本一致。而且,在加载 5 个模型之后,数据删除所带来的信息丢失被多个模型的集成补了回来,线上广告点击率效果上已经基本与未删除样本和集成的策略基本持平,这也与本文算法中加载若干(3 ~5)模型的假设是一致的。

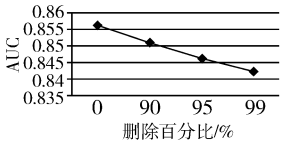

图 1　不同删除力度下 AUC 变化

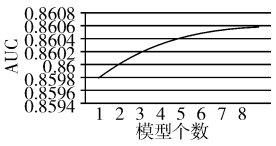

图 2　不同模型个数下的 AUC

表 3　不同删除力度下线上广告点击率

	模型个数		
	1	3	5
点击率变化	-0.5%	-0.28%	-0.098%

3.3.3　关于采样的讨论

在平衡采样的逻辑回归算法中,节约资源的主要原因在于采样,然而本文没有采用类别不平衡算法中启发式的 ENN、CNN 采样方法,而是用了最简单的随机采样,这主要从性能因素考虑。无论是 ENN、CNN 或者 NCL 这样的采样方法,都需要计算样本和样本之间的距离或者样本和决策边界的距离。由于百度联盟点击率日志规模庞大,训练样本数目较多(TB 级别),特征维度也非常庞大(几十万),计算样本之间距离过程是非常消耗计算资源的。以 ENN 为例,在每一轮样本采样的过程中,都需要计算每个样本的 3NN 样本,这个过程意味着要重新对整个训练集(TB 级样本)进行一次遍历,时间复杂度为 $O(n)$,显然这是非常消耗资源的。然而,随机采样则完全不同,决定每个样本是否删除只需要产生一个随机数就能完成,

时间复杂度为 $O(1)$。

在本算法过程中,本文采用了一种单边采样的策略,即只对负样本进行采样。为了证明这一思想的正确性,本文进行了一次实验,对比了对正样本和负样本同时采样和单边采样在不同模型个数下的线下 AUC。双边采样中对正样本采样 20%,负样本采样 90%;单边采样只对负样本采样 90%。实验结果(表4)表明,双边采样效果明显不如单边采样。

表4 双边采样和单边采样在不同模型个数下的 AUC

采样	模型个数			
	1	3	5	7
单边采样	0.859 6	0.860 2	0.860 3	0.860 6
双边采样	0.854 8	0.856 0	0.857 0	0.857 3

3.4 性能实验

本算法最主要的目的在于性能上的提升。表5 比较了未对数据进行采样和采样之后整个系统的性能参数。其中,删除力度 α 为 90%,使用三个模型进行投票。

表5 平衡采样逻辑回归算法带来的性能提升

性能	未采样	采样	对比/%
训练样本数据量(60/d)	11.4 TB	1.2 TB	-88
每轮迭代时间	600 s	100 s	-83.3
模型大小	1.2 GB	600 MB	-50
线上 CPU 空闲	75%	60%	-20
整体训练时间	8.5 h	1.5 h	-82.5

从表5 中可以看到,在进行数据删除之后,在主要的几个性能指标上都有了显著的性能提升。每个模型所需的训练数据大小缩减了 88%,这与删除 90% 负样本的预期基本一致,节约了大量的线上数据存储资源。由于 OWLQN 算法每一轮都要遍历所有的样本,因此在样本删除之后,算法每一轮所遍历样本数目减少,迭代的时间也大幅度缩减,从 600 s 下降到了 100 s。整体训练时间从过去的 8 h 下降到了 1.5 h,大大加快了模型的产出速度。与此同时,模型的大小也随之缩减了 50%。由于在线上进行点击率预估的时候需要同时加载三个模型,因此线上机器的 CPU 空闲时间下降了,然而这样的降低和整体性能的提升相比完全是可以接受的。

综合效果实验和性能实验的结果可以看到,平衡采样逻辑回归方法节约了大量的线下存储和计算资源,有效地提升了点击率预估系统的性能。更重要的是,这样的性能提升并没有造成点击率预估效果上的下降。主要原因在于两点,首先广告点击率预估的负样本中的冗余性非常高,即使删除大量的负样本也不会造成显著的预估效果下降;其次多次采样的集成多个模型的学习方法能有效地补回由于删除负样本所造成的信息丢失。

3.5 模型可扩展性实验

平衡采样逻辑回归帮助广告点击率预估节省了大量的计算和存储资源,这些节省下来的资源帮助后续的策略团队上线更多过去不可能被使用的特征。为了说明这一扩展性,本文简单地做了一个实验。网站 URL 和用户 cookie 是两个非常细粒度的特征,每天的点击日志中会有数以亿计的 URL 和 cookie。在离散化的逻辑回归算法下,如果把这两个特征加入到模型中会造成线上模型大小以及训练时间的大幅度增加,因此在过去几年,特征团队一直没有将其放入点击率预估模型。然而在进行平衡采样之后,计算和存储资源被节省了下来,这两个特征可以被尝试加入到广告点击率预估模型之中,首先看一下在加入特征之后系统整体的性能参数。

从表6 中看到,在加入了细粒度特征之后,训练的数据量和算法迭代时间及不加入细粒度特征之后有一定程度的增加,但是与未采样的数据相比仍然是有较大的改善。

表6 增加了 URL 和 cookie 特征之后系统性能参数

性能	未采样	采样	采样加特征
训练样本数据量	11.4 TB	1.2 TB	3 TB
每轮迭代时间	600 s	100 s	200 s
模型大小	1.2 GB	600 MB	1.1 GB
线上 CPU 空闲	75%	60%	55%

表7 反映了在加入 URL 和 cookie 特征之后,线下 AUC 以及线上点击率(与未进行采样不加入 cookie,URL 特征相比),可以很明显地看到,在加入了 URL 和 cookie 两个细粒度特征之后,无论是线下 AUC 还是线上广告点击都取得了长足的进步,尤其是线上广告的点击率。在广告点击率预估中增加 2% 的点击率是一个非常不错的效果,尤其是在百度联盟这样一个已经拥有 150 多种丰富特征和大量点击日志的定向广告联盟中。更重要的是,即使加入了这两个特征,资源消耗依然没有达到未采样之前的水平。

表7 增加了 URL 和 cookie 之后模型 AUC 和线上点击率效果对比

策略	线下 AUC 评估	线上点击率效果
加入 cookie、URL	0.868	+2.25%
未加入 cookie、URL	0.861	-0.28%

4 结束语

本文借鉴了在类别不平衡算法中平衡采样方法,并针对广告点击率问题中训练数据的特点提出了平衡采样逻辑回归算法。事实证明,该算法能在不牺牲模型点击率预估效果的前提下有效地提升整个系统的性能指标。然而,现有的平衡采样方法由于加载多个模型造成线上机器 CPU 空闲时间的下降,因此,是否存在其他方法不让系统同时加载多个模型、是否有更好的办法弥补数据删除所带来的信息丢失是下一步需要考虑的问题和工作。

参考文献:

[1] 搜狐 IT. 谷歌赚钱机器 AdSense 的诞生 10 年前收购功不可没[EB/OL]. (2013-03-16). http://it.sohu.com/20130316/n369051579.shtml

[2] LI Wei, WANG Xue-ri, ZHANG Ruo-fei, *et al*. Exploitation and exploration in a performance based contextual advertising system[C]// Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 27-36.

[3] KARIMZADEHGAN M, LI Wei, ZHANG Ruo-fei, *et al*. A stochastic learning-to-rank algorithm and its application to contextual advertising [C]//Proc of the 20th International Conference on World Wide Web. New York: ACM Press, 2011: 377-386.

[4] CHENG Hai-bin, Van ZWOL R, AZIMI J, *et al*. Multimedia features for click prediction of new ads in display advertising[C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 777-785.

[5] GRAEPEL T, CANDELA J Q, BORCHERT T, *et al*. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine[C]//Proc of the 27th International Conference on Machine Learning. 2010: 13-20.

[6] 李沐. 大数据:"人工特征工程 + 线性模型"的尽头[EB/OL]. (2013-03-25). http://qing.blog.sina.com.cn/tj/74733da9330036o7.html.

[7] KOTSIANTIS S B, PINTELAS P E. Mixture of expert agents for handling imbalanced data sets[J]. Annals of Mathematics, Computing &Teleinformatics, 2003, 1(1): 46-55. (下转第 39 页)

种 $W(U;c,d)=0$ 的情况,避免了因数据丢失给结果带来的不准确性。

根据计算出的结果,当 $c=3$ 时 VS(U,V)的值最大,划分效果最明显。与原 MPO(U,V)指标进行对比,表明加入距离临界值 L 的 VS(U,V)有效性函数不仅与原 MPO(U,V)函数一样,都能确定最优聚类数为 3,划分效果更明显,而且能避免因噪点的过量删除带来的数据信息丢失,并能更加准确地得到类与类之间的模糊点。

3.2　算例 2

对全国 31 个省、市、自治区和直辖市的科研课题能力的模糊聚类结果进行有效性评价。仿照算例 1 步骤,得到结果如表 3、4 所示。

表 3　VS(U,V)算例结果

c	com′(c)	sep′(c)	VS′(U,V)
3	3.916 4	0.440 6	8.888 8
4	2.610 7	1.485 4	1.757 6
5	2.577 8	2.348	1.097 9

表 4　MPO(U,V)算例结果

c	com(c)	sep(c)	VS(U,V)
3	15.050 6	1.706 6	8.019
4	13.230 4	3.953 5	3.346 5
5	16.165 7	7.310 6	2.211 3

当 $c=3$ 时,仅有上海第一类和第二类的隶属度差值为 0.01,小于等于 $T=0.01$,但是第一类和第二类的距离的差值小于 L_j。因此,在 $c=3$ 时没有被删除的噪点,所有值均取 $1-|u_{aj}-u_{bj}|$,代入公式,最终得到 sep′(c)的值为 0.4406。

当 $c=4$ 时,四川的 $u_{23,1}-u_{23,2}=0.0079$ 同时满足约束,$W(U;c,d)$取值为 0,其他值均取 $1-|u_{aj}-u_{bj}|$,最终得到 sep′(4)的值为 1.4854。

当 $c=5$ 时,满足$|u_{aj}-u_{bj}|\leqslant T$ 的点有 19 个,加入$|d_{aj}-d_{bj}|\leqslant L_j$ 约束后,最终 $W(U;c,d)$取值为 0 的点为 4 个。

通过计算得 VS(U,V)和 MPO(U,V)的值,确定最优聚类数为 3,再次证明了函数的有效性。

4　结束语

本文在分析了各种聚类有效性评价函数优缺点的基础上对其中的 MPO 评价函数进行改进,得到了新的聚类有效性函数。该函数通过距离和隶属度结合了几何结构域模糊划分两种理论,构建紧密性度量和分离性度量,在原有隶属度临界值 T 的基础上引入距离临界值 L,降低了类间噪点对聚类评价结果的影响。通过算例说明了该方法的合理性与有效性。

参考文献:

[1] 胡雅婷. 可能性聚类方法研究及应用[D]. 吉林:吉林大学,2012.

[2] 彭勇,吴友情. 一种新的聚类有效性函数[J]. 计算机工程与科学,2010,46(6):124-132.

[3] DUNN J C. Well-separated clusters and the optimal fuzzy partitions [J]. Journal of Cybernetics, 1974, 4(1):95-104.

[4] MAULIK U, BANDYOPADHYA S. Performance evaluation of some clustering algorithms and validity indices[J]. Pattern Analysis and Machine Intelligence, 2002, 24(12):1650-1654.

[5] CALINSIK T, HARABASZ J. A dendrite method for cluster analysis [J]. Communications in Statistics, 1974, 3(1):1-27.

[6] BEZDEK J C. Numerical taxonomy with fuzzy sets[J]. Journal of Mathematical Biology, 1974, 1(1):57-71.

[7] BEZDEK J C. Cluster validity with fuzzy sets[J]. Journal of Cybernetics, 1973, 3(3):58-73.

[8] WINDHAM M P. Cluster validity for fuzzy clustering algorithms[J]. Fuzzy Sets and Systems, 1981, 5(2):177-185.

[9] ZAHID N, LIMOURI M, ESSAID A. A new cluster-validity for fuzzy clustering[J]. Pattern Recognition, 1999, 32(7): 1089-1097.

[10] XIE Xuan-li, BENI G. A validity measure for fuzzy clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-847.

[11] 唐明会, 杨燕. 模糊聚类有效性的研究进展[J]. 计算机工程与科学, 2009, 31(9): 112-114.

[12] 刘万里,刘三阳,薛贞霞. 基于距离核函数的除噪和减样方法[J]. 系统工程理论与实践, 2008, 7(7): 160-163.

[13] HU Ya-ting, ZUO Chun-chang, YANG Yang, *et al*. A cluster validity index for fuzzy C-means clustering[C]//Proc of the 2nd International Conference on System Science, Engineering Design and Manufacturing Informatization. 2011: 263-266.

[14] ZALIK K R. Cluster validity index for estimation of fuzzy clusters of different sizes and densities[J]. Pattern Recognition, 2010, 43(10): 3374-3390.

[15] 谢中华. MATLAB 统计分析与应用: 40 个案例分析[M]. 北京: 北京航空航天大学出版社, 2010.

(上接第 36 页)

[8] HART P E. The condensed nearest neighbor rule[J]. IEEE Trans on Information Theory, 1968,14(3): 515-516.

[9] WILSON D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Trans on Systems, Man and Cybernetics, 1972,SMC-2(3): 408-421.

[10] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society,2001:511-518.

[11] WEISS G M. Learning with rare cases and small disjuncts[C]//Proc of the 12th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1995:558-565.

[12] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C]//Proc of the 8th Conference on AI in Medicine in Europe. London:Springer-Verlag, 2001:63-66.

[13] TOMEK I. Two modifications of CNN [J]. IEEE Trans on Systems, Man and Cybernetics, 1976,SMC-6(11):769-772.

[14] RICHARDSON M, DOMINOWSKA E,RAGNO R. Predicting clicks: estimating the click-through rate for new ads[C]//Proc of the 16th International Conference on World Wide Web. New York: ACM Press,2007:521-530.

[15] ANDREW G,GAO Jian-feng. Scalable training of L^1-regularized log-linear models[C]//Proc of the 24th International Conference on Machine Learning. New York:ACM Press,2007:33-40.

[16] BYRD R H, NOCEDAL J,SCHNABEL R B. Representations of quasi-Newton matrices and their use in limited memory methods[J]. Mathematical Programming,1994,63(1-3):129-156.

[17] LING C X,LI Cheng-hui. Data mining for direct marketing: problems and solutions [C]//Proc of the 4th International Conference on Knowledge Discovery and Data Mining. 1998:217-225.

[18] KUBAT M,MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//Proc of the 14th International Conference on Machine Learning. 1997:179-186.

[19] CHEN Chao, LIAW A,BREIMAN L. Using random forest to learn imbalanced data, 666[R]. Berkeley:Statistics Department, University of California, 2004.

[20] LIU Xu-ying, WU Jian-xin, ZHOU Zhi-hua. Exploratory under-sampling for class-imbalance learning[C]//Proc of the 6th International Conference on Data Mining. 2006:965-969.

[21] CHU C, KIM S K, LIN Y A, *et al*. Map-Reduce for machine learning on multicore [C]//Advances in Neural Information Processing Systems. Cambridge:MIT Press,2007:281-288.