doi:10.11887/j.cn.202203010 http://journal.nudt.edu.cn

众包标签的双重置信度推断算法*

张 琳,姜高霞,王文剑
(山西大学 计算机与信息技术学院,山西 太原 030006)

摘 要:标记者的知识水平、评价标准等均具有显著差异,导致收集到的标签质量参差不齐,提高标签和学习模型的质量对众包标签中学习起着关键作用。针对众包标签推断问题,提出了一种双重置信度推断算法,分别从数据分布特征及标签信息两方面计算得到标记者置信度,再通过此置信度推断数据集的集成标签,以此提高集成标签的质量。实验结果表明,与其他仅使用标签信息的推断算法相比,所提算法可以得到更优结果。

关键词:众包;标签集成;聚类;双重置信度
中图分类号:TP391 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2022)03-077-08

听语音 聊科研 与作者互动

Crowdsourced label inference algorithm using double-confidence

ZHANG Lin, *JIANG Gaoxia*, *WANG Wenjian*
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract: Since the workers have significant differences in the knowledge level and evaluation criteria, the quality of the collected labels varies a lot. It's of key importance to improve the quality of labels and learning models in crowdsourced label learning. A novel double-confidence inference algorithm was proposed to solve the problem of crowdsourced label inference. The workers' confidence was obtained via the data distribution characteristics and label information, and then the label was inferred by this confidence so as to improve the quality of the integrated label. The experimental results show that the proposed algorithm outperforms other ground truth inference algorithms only based on label information.

Keywords: crowdsourced; integrated label; clustering; double-confidence

机器学习是现代计算机快速发展的领域之一,涉及各个领域的应用。目前,研究人员在深度学习方面取得了很多成果,但其可解释性依然是难以解决的问题。因此,人在机器学习中的作用依然不可忽视。众所周知,传统的数据注释依赖于领域专家,成本高且耗时长,其限制了标签数据集的获得。而随着众包系统的快速发展,标签数据的获取变得比较容易,但由于收集的标签中存在噪声,众包数据训练的学习模型质量通常低于专家标记数据训练的模型。直观上,学习模型的质量与标签的质量密切相关,因此提高标签质量是提高学习模型质量的一个直接途径。

在众包平台获取带噪声的标签数据集后,通过设计算法,从多个标签集合中归纳出一个完整的标签,这些算法被称为真相推理算法(ground truth inference algorithms)。在不知道某实例真实标签的情况下,一般使用真相推理算法推断其真实标签,但单纯的推理算法的学习是相对困难的。因此,利用实例数据分布信息帮助推理算法推断真实标签十分重要,且希望每个实例的集成标签为其真实标签。

使用重复标签来提高标签质量可以追溯到20多年前,Smyth等在1994年使用极大似然估计算法和重复标记去解决金星图像标记的不确定性[1]。多人投票机制(majority vote, MV)是一种最简单高效的方法,除了MV算法,近几年还提出了一些其他的推理算法。这些推理算法可以根据其数学方法分为两类:基于机器学习的算法和基于线性代数的方法。基于机器学习的方法是当前研究的主流,早在1979年,Dawid等提出了一种

* **收稿日期**:2021-08-30
基金项目:国家自然科学基金资助项目(62076154,61906113,U1805263);山西省国际合作重点研发计划资助项目(201903D421050);山西省高等学校科技创新资助项目(2020L0007);中央引导地方科技创新资助项目(YDZX20201400001224)
作者简介:张琳(1997—),女,山西吕梁人,硕士研究生,E-mail:15735166294@163.com;
王文剑(通信作者),女,教授,博士,博士生导师,E-mail:wjwang@sxu.edu.cn

基于最大似然估计的真相推理算法(Dawid-Skene model,DS)[2],除了为每个例子推断出集成标签外,DS 还为每个贴标签者估计了混淆矩阵;Demartini 等提出的 ZenCrowd[3]算法通过对标记者质量建模来确定每个样本属于特定类的概率,但其没有关注任务难度;而 Whitehill 等提出的 GLAD[4]算法则在此基础上,对任务难度建模,将此结果应用到对标记者质量的建模中,使推断结果具有更高的准确性,但导致其代码运行迭代过程十分缓慢。Sheng 等和 Ipeirotis 等分别于 2008 年和 2014 年在研究了 MV 算法后,提出了简单的概率模型来描述单样本的标签质量[5-6],但其假设每个标签的质量均相同。Jung 等在 2011 年针对实例,提出了一种利用评分机制和权重提高 MV 精度的方法[7],其中推理模型通常是基于概率图形模型,在概率图形模型中进行推理的一种重要的主流通用方法是期望最大化(expectation-maximization,EM)算法。2014 年 Li 等在 DS 模型下推导了具有任意有限标记者和项目数的一般类型聚合规则的错误率边界[8],可用于设计最优加权多数投票。Khetan 等在 2016 年也分析了广义 DS 模型下众包的可靠性[9]。在基于 DS 改进的模型中,当类的数量较大而标签的数量较少时,会造成混淆矩阵十分稀疏。

众包通常基于两个基本的共同假设:贴标签者有不同的可靠性,并且独立做决定。一些推理算法只遵循这些假设,例如 DS 和 ZenCrowd。而 Raykar 等[10]还基于标签标记者在提供标签时的偏见,IEThresh[11]则是基于了学习者的经验,正标签阈值(positive label frequency threshold,PLAT)[12]方法假设标签者对消极和积极的例子有不同的校正率,LC-ME[13]算法假设每个项都属于一个潜在类,标签程序对同一类的项具有一致的视图,但对不同类的项具有不一致的视图。

由于仅根据带噪音的标签推断实例的真实标签比较困难,因此除了一般的真相推理算法外,近几年一些学者提出了根据实例特征和标记者影响等因素去提高标签集成效果的方法。Tian 等在 2018 年提出了 M^3V 算法[14],Ruiz 等在 2018 年提出了一种通过变分高斯过程从众包中学习正确标签的方法[15]。还有一些学者提出不学习真相推理算法,而直接根据带噪音的标签集学习分类算法。另一些学者利用多任务学习方法去处理众包标签噪音,也取得了不错的效果。目前来看,利用其他信息提高众包标签集的集成算法效果是可行的。但现有的算法在多分类问题中的表现相对较差,且在众包系统中获取标签时,标记者越少代价越小,因此希望可以在获取的众包标签数较少的情况下得到较优的标签推断结果。由此考虑使用聚类算法研究众包噪音标签过滤问题。本文提出的算法主要通过对实例特征聚类及分析标记者相似性两部分来提高标签推断的准确率。

1 单标记众包标签的双重置信度算法

1.1 问题描述

众包标签推断问题是基于样本特征及多标签数据等信息推断实例的真实标签。众包数据 $D=\{\langle x_i,\hat{y}_i,y_i\rangle\}_{i=1}^{n}$,标签数据集 $label=\{l_i\}_{i=1}^{k}$。其中,n 表示样本总数,x_i 表示样本特征,$\hat{y}_i$ 表示实例的真实标签,y_i 表示标记者提供的数据标签,$y_i\in label$,k 表示类别数。标记者数据集 $worker=\{w_1,w_2,\cdots,w_m\}$,其中 m 表示标记者总数。

1.2 双重置信度计算方法

由于提高集成标签的质量面临巨大的挑战,许多研究人员试图在标签聚合过程中引入更多的先验信息,这些方法违反了通用标签聚合的不可知前提,即除了收集的噪声标签外,不能使用任何先验知识。实例的特征携带有价值的信息,如果可以应用适当的机器学习方法,这些特征可以帮助识别和分类这些项目。因此,在标签推断过程中完全忽略实例的特征是不明智的,且应用特征并不违反不可知论的先决条件。

本文算法主要通过设置标记者置信度推断实例的集成标签。标记者置信度由两部分确定:第一部分是通过聚类算法将实例的特征部分聚类,再根据聚类结果和标记者提供的标签,确定每位标记者的正确率,从而得到标记者置信度的第一部分基于数据分布特征的置信度;第二部分通过计算标记者提供的标签间的相似度得到基于标记信息的置信度,若某位标记者与其他标记者提供的结果越相似,则其置信度越高。将两部分置信度合并,由此得到完整的标记者置信度。

1.2.1 基于数据分布特征的置信度

通过对数据特征执行聚类算法得到基于数据分布特征的置信度。通常数据集中大多数的数据标记问题是较为简单的,标记者可以较容易地为这些问题打出正确的标签,而对于较为困难的一部分数据,标记者所提供的标签很有可能是错误的,这时可以通过聚类结果与某一组标签的一致性来确定某个标记者的置信度,从而提高标签推断结果的准确率。

图 1 通过一个简单的例子来说明聚类置信度的可行性。图中圆形、正方形、三角形分别表示 3 类数据,图 1(a)表示数据的真实分布情况,图 2(b)表示标记者 w_1 提供的数据划分结果,图 1(c)表示另一标记者 w_2 提供的数据划分结果,图 1(d)表示将原始数据进行聚类得到的分类结果。

图 1 中两位标记者在为众包数据提供标签时均存在错误标签,而此时有较为可靠的聚类结果,则可利用这一结果帮助确定标记者的置信度。

由于采用了聚类算法,因此聚类结果与众包标记结果的相似性采用常见的聚类评价指标 Rand Index 系数表示,将其作为标记者的第一部分置信度:

$$v_1 = R = \frac{T_p + T_n}{T_p + F_p + T_n + F_n} = \frac{T_p + T_n}{C_n^2} \tag{1}$$

式中:T_p 表示在聚类结果和标记者提供的标签中均属于同一类的实例数;T_n 表示在聚类结果和标记者提供的标签中均不属于同一类的实例数;F_p 表示聚类结果属于一类,而标记者提供的标签不属于一类的实例数;F_n 表示聚类结果不属于一类,而标记者提供的标签属于一类的实例数;n 表示实例数。

1.2.2　基于标签信息的置信度

在众包数据中,某位标记者与其他标记者所提供的标签信息越相似,则这位标记者提供的标签越可信。因此,将标记者之间的相似性作为标记者置信度的一部分,用以计算完整的置信度。

用 p_{ab} 表示标记者 a 与标记者 b 所提供的标签信息间的相似度,p_a 表示标记者 a 与其他标记者的平均相似度,$a,b \in (1,m)$。

根据某位标记者为某个实例提供的标签 y_i^w 计算其与其他标记者提供标签的相似度 p_{ab},再将每个标记者与其他标记者的相似度相加,根据式(3)求得平均相似度 p_a,即为第二部分置信度 v_2。

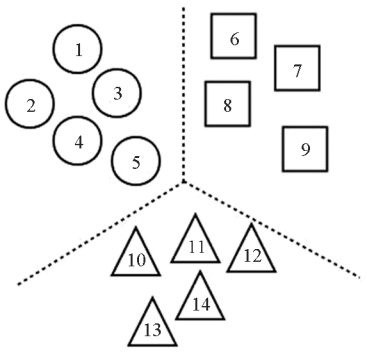
$$p_{ab} = \frac{1}{n} \cdot \sum_{i=1}^{n} I(y_i^{w_a} = y_i^{w_b}) \tag{2}$$

$$v_2 = p_a = \frac{1}{m-1} \cdot \sum_{b=1, b \neq a}^{m} p_{ab} \tag{3}$$

根据上文分别得到基于数据分布特征和标签信息的置信度 v_1 和 v_2,$v_1, v_2 \in (0,1)$,计算两者的几何平均值 V,即标记者置信度,其中 $V \in (0,1)$。

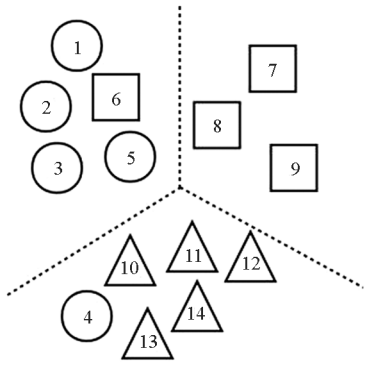
$$V = \sqrt{v_1 \cdot v_2} \tag{4}$$

图 2 所示为标记者置信度 V 随 v_1、v_2 变化的趋势图。由图可知,v_1、v_2 均对 V 有整体的影响。



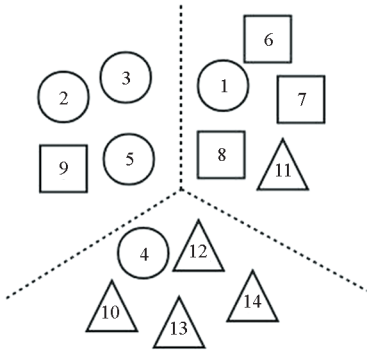
(a) 原始数据的分布

(a) Original data distribution



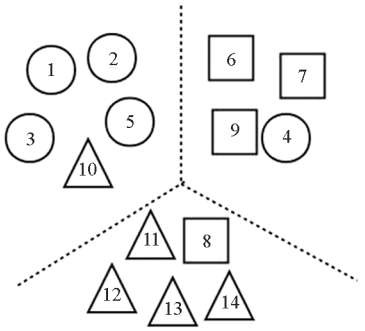
(b) w_1 的数据划分结果

(b) Data partitioning results of w_1



(c) w_2 的数据划分结果

(c) Data partitioning results of w_2



(d) 数据聚类结果

(d) Data clustering results

图 1　数据标记

Fig. 1　Data labeling

若使用算数平均值,则当其中一个变量不变时,另一变量对其整体的影响有限,不能充分表现变量较小时对整体权重的影响。

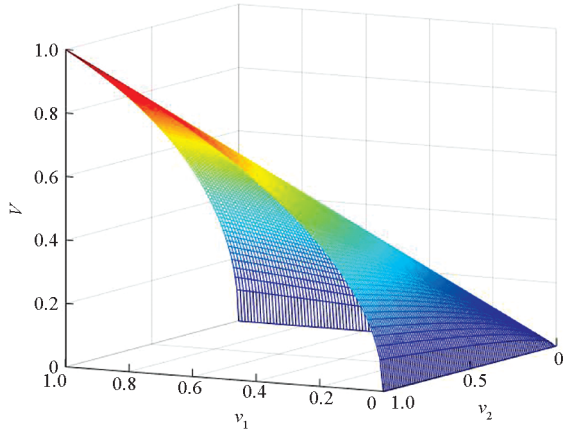


图 2 置信度的变化情况

Fig. 2 Variation of confidence

根据得到的标记者置信度推断得出实例的真实标签:

$$l_i = \arg\max_{i=1,2,\cdots,k} p(l_i) \tag{5}$$

式中,$p(l_i)$ 表示实例推断标签为 l_i 的可能性,$i \in (1,k)$。

双重置信度推断(double confidence inference, DC)算法的步骤如算法 1 所示。

算法 1 双重置信度推断算法

Alg. 1 Double confidence inference algorithm

输入:多标签数据集
输出:实例预测标签

1. 对实例特征部分执行 DBSCAN 聚类算法;
2. 将聚类算法结果与每个标记者提供的标签数据集作对比,分别计算 T_p 及 T_n,根据式(1)计算 Rand Index 系数作为每个标记者的第一部分置信度;
3. 根据式(2)计算标记者间的相似性,再由式(3)计算每个标记者的第二部分置信度;
4. 根据式(4)计算标记者置信度;
5. 根据得到的标记者置信度,计算真实标签等于每种标签的可能性,由式(5)选取可能性最大的标签作为该实例的预测标签

本文算法在每个数据集上的每种标记数下的计算复杂度 $O(n) = n \cdot p \cdot l$,其主要由算法 1 的第 5 步推断实例的真实标签过程产生。其中:n 为数据集实例数;p 表示标记数,即为该实例提供标签的人数;l 为数据集的类数。

2 实验与分析

本节对表 1 所示的 10 个真实世界数据集进行实验,并与 4 种经典的方法 MV、DS、GLAD、ZC 算法在标签推断准确率和时间效率两方面进行了比较。

表 1 实验数据集

Tab. 1 Experimental data set

数据集	样本数	类别数	特征数
Svmguide2	391	3	20
Svmguide4	612	6	10
Vehicle	846	4	18
Phishing	1 353	3	10
Steel Plates Faults	1 941	7	27
Segment	2 310	7	19
Satimage	4 435	6	36
Letter	20 000	26	16
Australian	690	2	14
Breast_cancer	699	2	10

2.1 实验设计

使用的数据集均为 UCI 数据集,参照文献[16]中所示方法模拟生成众包数据。首先为每位标记者随机生成标记准确率 t,$t \in (0.3, 0.9)$;再根据每位标记者的准确率对原数据集随机抽样,样本数据按照准确率 t 被赋予正确标签,其他数据均被随机赋予错误标签,由此生成标签数据集。实验为每个数据集生成了标记者数分别为 1 ~ 10 的标签数据集,为保证实验结果可信,每次的标签生成过程重复 5 次。

考虑数据分布的多样性,实验中采用的是经典的噪声环境下基于密度的聚类(density-based spatial clustering of applications with noise, DBSCAN)算法。DBSCAN 算法包含 3 个参数,分别是 X、ε、m_p。其中:X 为数据集的特征;ε 表示样本间的最小距离,即若两样本间距离小于 ε,则样本互为邻域;m_p 表示形成簇类所需的最小样本个数,将 m_p 设定为特征数的 2 倍。ε 和 m_p 的计算方法分别为:

$$\varepsilon = \boldsymbol{d}(0.005l(\boldsymbol{d})) \tag{6}$$

$$m_p = 2d_x \tag{7}$$

其中,$\boldsymbol{d}$ 表示排序后的数据特征的距离向量,ε 为该向量千分之五处的值,$l(\cdot)$ 表示计算向量长度的函数,d_x 表示数据集的特征数。

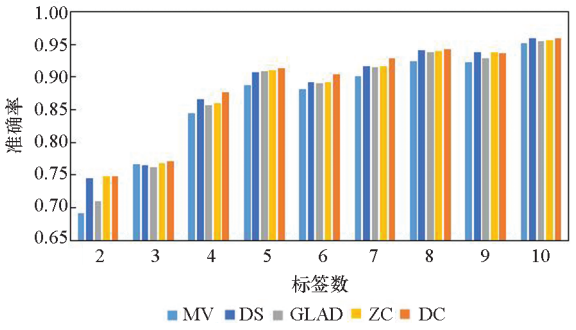
2.2　实验结果分析

为验证本文算法的有效性,从时间效率及标签推测准确性两方面考虑方法效率。时间效率为各个算法在每个数据集上的 5 次实验的平均运行时间,标签准确率如式(8)所示。

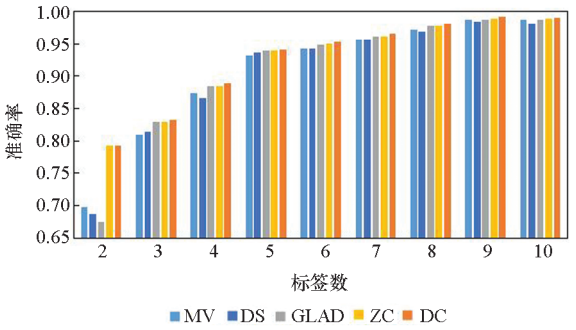
$$c = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i = y_i) \tag{8}$$

2.2.1　不同标签数对准确率的影响

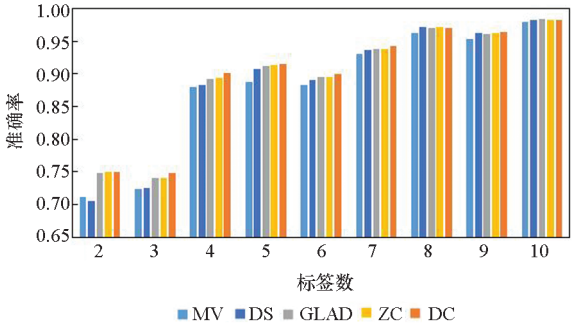
图 3 所示为 5 种算法在 10 个数据集上实例被标记次数为 2 ~ 10 的准确率。由图 3 可知,5 种方法的实验准确率均随标签数的增多而提高。标签数为 1 时的实验结果参考性不大,因此图中未表示出来。在标签数为 2 时,MV 算法与 GLAD 算法的实验效果均较差,DS 算法的表现也差强人意。在大部分数据集上,当标签数多于 2 时,各个算法的准确率均随着标签数的增多快速提升,当每个实例的标签数多于 7 时,各个算法的准确率



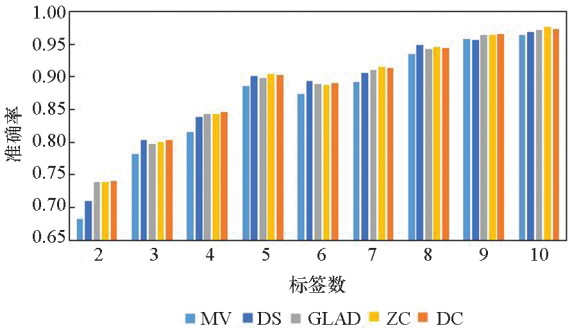
(a) Svmguide2



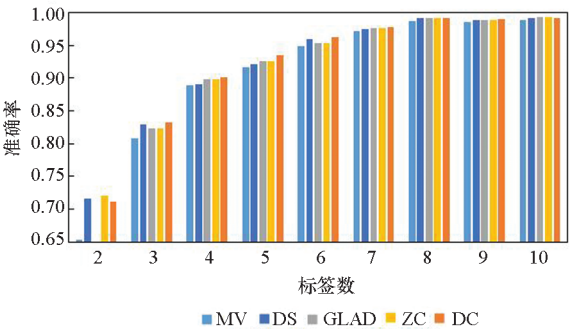
(b) Svmguide4



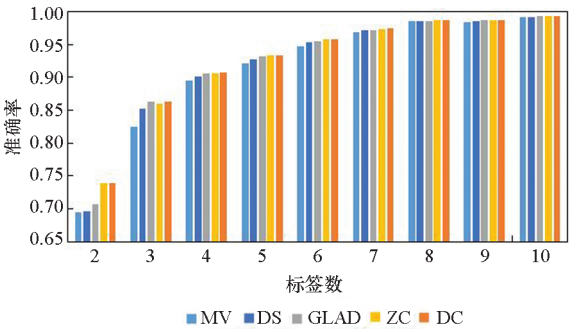
(c) Vehicle



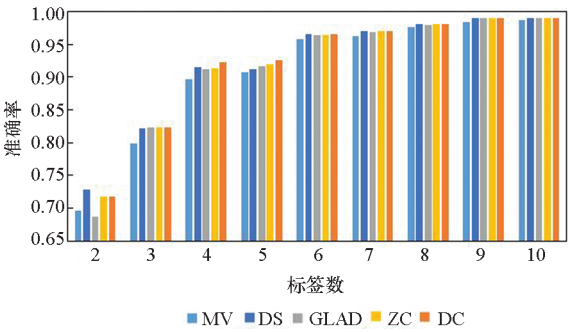
(d) Phishing



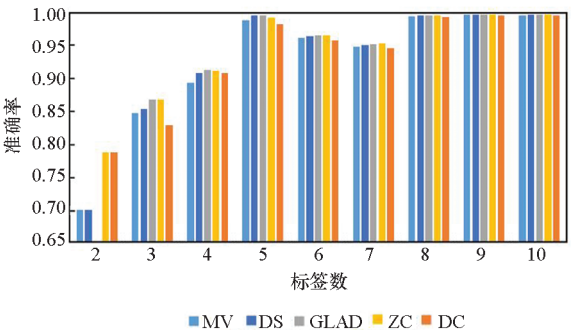
(e) Steel Plates Faults



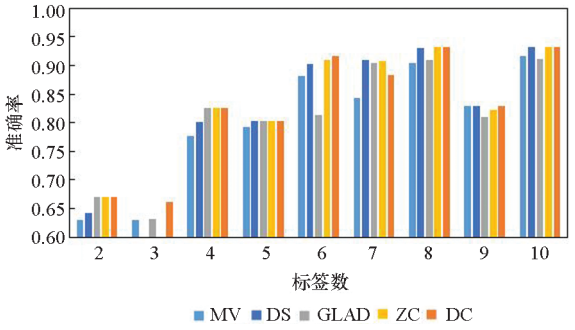
(f) Segment



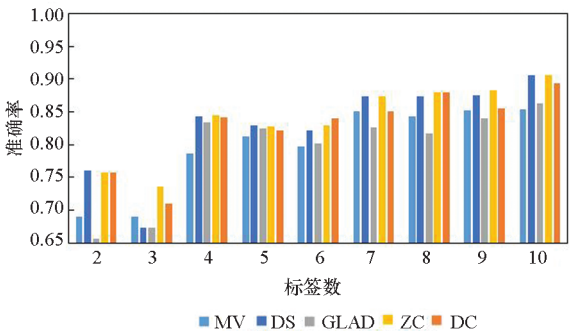
(g) Satimage



(h) Letter



（i）Australian



（j）Breast_cancer

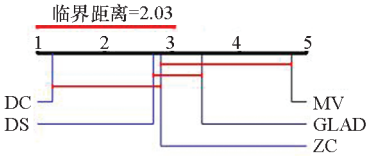
图 3　不同标记者数目在各数据集下的算法准确率

Fig. 3　Accuracy of the algorithm for the number of different markers in each data set

均趋于 1，效果相差较小，且几乎不再随着标签数的增多而有所提高。

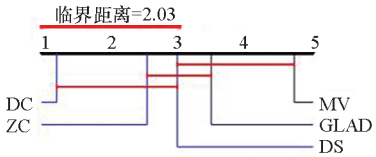
在真实众包数据中，每个样本的标签数均大于 7 的可能性较小，因此主要分析标签数为 4 ~ 7 的实验结果。显而易见，在大部分数据集上，当标签数为 4 ~ 7 时，DC 算法的实验效果略优于其他 4 种算法，其他情况时与另外 4 种算法的实验效果相近，而在实际的标注中，得到的标签数据往往没有很多，因此 DC 算法相较其他方法有一定优势。DS 算法及 ZC 算法在二分类问题中的实验效果比 MV、GLAD、DC 算法好，而在多分类问题中 GLAD、DC 算法的效果较其在二分类中的效果有明显的提升。

图 4 给出了各算法在标签数分别为 4 ~ 7 时准确率的临界差异（critical difference，CD）图。CD 图是基于统计显著性差异的算法对比模式，可以给出不同算法的排名。图 4 中算法排名越小表示算法效果越好。在标签数为 4 ~ 7 时，DC 算



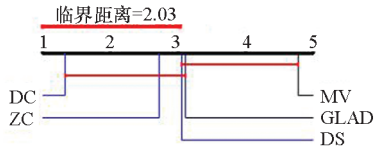
（a）标签数为 4

（a）Number of labels is 4



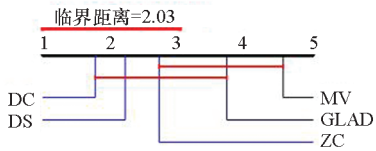
（b）标签数为 5

（b）Number of labels is 5



（c）标签数为 6

（c）Number of labels is 6



（d）标签数为 7

（d）Number of labels is 7

图 4　临界差异图

Fig. 4　Critical difference diagram

法准确率均高于其他 4 种算法。图中算法的平均排名基于 10 个数据集和 10 种标签数量。

2.2.2　与其他算法的比较分析

表 2 所示为 5 种方法分别在标签数为 1 ~ 10 时的实验数据均值的比较。由表 2 所示可知，DC 算法在 8 个数据集上的表现最优，相比最差的实验数据分别提高了 1.9%、2.23%、1.52%、2.14%、1.62%、1.45%、3.86%、5.39%；相比其他 4 种算法平均提高了 1.12%、1.52%、0.65%、1.11%、1.06%、0.72%、1.68%、2.26%。在 Segment 和 Satimage 数据集上的表现也仅次于最好的算法结果，且分别相差 1.03%、0.79%。这是由于数据集的类别数增多，噪音分散严重。

在各个数据集中，由于标记者的准确率随机给出，且绝大部分标记者的准确率高于 0.5，因此 MV 算法的准确率随着标签数的增多而提高，而在真实数据集中，标记者的准确率并没有随机给出的准确率高，导致 MV 算法的准确率其实并没有实验所示的那么高。又因为标签数据中的噪音是根据给定的标记者的准确率随机给出，所以多分类噪音没有二分类问题中的噪音结果集中，而是被分散到了所有可能的标签上，这样就导致 MV 算法在多分类问题中的准确率更高，其他4 种算法也受此影响。

当标记者准确率较低时，GLAD 算法相较其

他算法在各个数据集中的实验效果较优,但该算法的迭代时间相较其他算法过长;DS 算法使用的基本算法是 EM 算法,其实验效果受初值影响较大,DC 算法结果为运行效果最好的结果;ZC 算法在各数据集上的表现相对 MV、DS、GLAD 算法是最好的。DC 算法在大部分数据集上的实验效果相较其他 4 种算法均为最优的。

表 2　各算法在各数据集中的平均准确率

Tab. 2　Average accuracy of each algorithm in each data set

%

数据集	算法				
	MV	DS	GLAD	ZC	DC
Svmguide2	86.76 ±0.93	87.32 ±1.83	87.75 ±1.28	88.32 ±2.39	**88.66 ±1.56**
Svmguide4	90.64 ±1.73	90.38 ±1.84	90.99 ±3.94	92.35 ±2.66	**92.61 ±2.71**
Vehicle	87.91 ±2.37	88.49 ±2.50	89.33 ±2.98	89.40 ±3.14	**89.43 ±2.92**
Phishing	86.52 ±1.58	86.72 ±1.70	88.36 ±1.05	88.59 ±1.55	**88.66 ±1.34**
Steel Plates Faults	90.53 ±1.82	91.26 ±1.32	90.48 ±2.06	91.89 ±2.00	**92.10 ±1.90**
Segment	91.25 ±1.49	91.82 ±1.56	92.25 ±1.83	92.60 ±1.90	**92.70 ±1.42**
Satimage	90.67 ±2.30	92.05 ±1.76	91.30 ±2.60	**92.06 ±2.53**	91.03 ±3.59
Letter	92.55 ±1.90	92.92 ±1.25	92.25 ±2.00	**94.13 ±1.35**	93.34 ±1.47
Australian	79.57 ±0.81	82.07 ±1.24	81.94 ±1.04	83.43 ±1.57	**83.43 ±1.93**
Breast_cancer	82.45 ±2.91	84.38 ±1.10	79.58 ±0.30	84.43 ±1.35	**84.97 ±2.30**

图 5 为 5 种算法在 10 个数据集上的平均运行时间,由于 GLAD 算法的运行时间相较其他 4 种算法长很多,因此图中所示的 GLAD 算法的运行时间为真实运行时间的 1%。5 种算法的运行时间均随数据集数目的增大而变长,其中:MV 算法的运行时间最短;ZC 算法与 DS 算法耗时相近;DC 算法实验过程中要运行聚类算法,因此耗时较 DS、ZC、MV 算法略长。

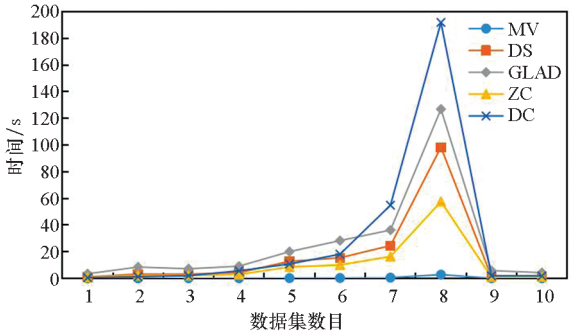


图 5　算法运行时间

Fig. 5　Running time of algorithms

3　结论

本文考虑了数据分布和标签信息两方面的置信度,首先根据 DBSCAN 聚类算法将样本数据聚类,由此结果得到标记者置信度的第一部分,再由标记者相似性计算得到置信度的第二部分,最后根据标记者置信度推断实例的真实标签。实验结果表明,DC 算法在标签数处于 3 ~7 的情况时,效果优于其他算法。

由于 DC 算法使用了聚类方法,考虑到聚类结果的准确性可能对实验结果有一定影响,因此未来工作将针对聚类算法对实验结果的影响以及纠正聚类结果的准确性这两部分展开,以提高标签推断算法的准确性。

参考文献(References)

[1] SMYTH P, BURL M C, FAYYAD U M, et al. Knowledge discovery in large image databases: dealing with uncertainties in ground truth[C]//Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1994: 109 - 120.

[2] DAWID A P, SKENE A M. Maximum likelihood estimation of observer error-rates using the EM algorithm[J]. Applied Statistics, 1979, 28(1): 20 - 28.

[3] DEMARTINI G, DIFALLAH D E, CUDRÉ-MAUROUX P. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking[C]// Proceedings of the 21st International Conference on World Wide Web, 2012: 469 - 478.

[4] WHITEHILL J, RUVOLO P, WU T F, et al. Whose vote should count more: optimal integration of labels from labelers

of unknown expertise [C]//Proceedings of the 22nd International Conference on Neural Information Processing Systems, 2009: 2035 – 2043.

[5] SHENG V S, PROVOST F, IPEIROTIS P G. Get another label? improving data quality and data mining using multiple, noisy labelers [C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008: 614 – 622.

[6] IPEIROTIS P G, PROVOST F, SHENG V S, et al. Repeated labeling using multiple noisy labelers [J]. Data Mining and Knowledge Discovery, 2014, 28(2): 402 – 441.

[7] JUNG H J, LEASE M. Improving consensus accuracy via Z-score and weighted voting [C]//Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011.

[8] LI H W, YU B. Error rate bounds and iterative weighted majority voting for crowdsourcing [EB/OL]. (2014 – 11 – 15) [2021 – 05 – 10]. https://arxiv.org/abs/1411.4086.

[9] KHETAN A, OH S. Reliable crowdsourcing under the generalized dawid-skene model [EB/OL]. (2016 – 02 – 10) [2021 – 05 – 10]. https://arxiv.org/abs/1602.03481v1.

[10] RAYKAR V C, YU S, ZHAO L H, et al. Learning from crowds [J]. Journal of Machine Learning Research, 2010, 11(4): 1297 – 1322.

[11] DONMEZ P, CARBONELL J, SCHNEIDER J. A probabilistic framework to learn from multiple annotators with time-varying accuracy [C]//Proceedings of the 2010 SIAM International Conference on Data Mining, 2010: 826 – 837.

[12] ZHANG J, WU X D, SHENG V S. Imbalanced multiple noisy labeling [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(2): 489 – 503.

[13] TIAN T, ZHU J. Uncovering the latent structures of crowd labeling [C]//Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2015: 392 – 404.

[14] TIAN T, JUN Z, YOU Q B. Max-margin majority voting for learning from crowds [C]//Proceedings of Advances in Neural Information Processing Systems, 2015.

[15] RUIZ P, MORALES-ÁLVAREZ P, MOLINA R, et al. Learning from crowds with variational Gaussian processes [J]. Pattern Recognition, 2019, 88: 298 – 311.

[16] TAO F N, JIANG L X, LI C Q. Label similarity-based weighted soft majority voting and pairing for crowdsourcing [J]. Knowledge and Information Systems, 2020, 62(7): 2521 – 2538.