基于局部加权的 Citation-kNN 算法

黄剑华 丁建睿* 刘家锋 张英涛
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要：Citation-kNN 算法对传统的 kNN 算法进行了改进，使其可以应用于多示例学习问题，但其 0-1 决策方式具有一定的局限性，没有充分考虑样本的分布情况。为解决该问题，该文提出局部加权的 Citation-kNN 算法，综合考虑样本的分布情况，提出基于样本距离加权、基于样本离散度加权的方法，并对各种组合情况进行了实验。在标准数据集 MUSK 和乳腺超声图像数据库上的实验结果表明，该文提出的方法与 Citation-kNN 相比，性能有明显提高，并具有良好的适应性。

关键词：图像识别；多示例学习；Citation-kNN；样本分布；局部加权
中图分类号：TP391.41 **文献标识码**： A **文章编号**：1009-5896(2013)03-0627-06
DOI: 10.3724/SP.J.1146.2012.00016

Citation-kNN Algorithm Based on Locally-weighting

Huang Jian-hua Ding Jian-rui Liu Jia-feng Zhang Ying-tao
(*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: The Citation-kNN algorithm improves traditional kNN algorithm and can be applied to solve multi-instance learning issue. But its 0-1 decision strategy has some limitations. To overcome this issue, the locally-weighted Citation-kNN algorithm is presented in this paper. Considering distribution of the samples, the distance-based weighted method and the scatter-based weighted method are proposed. And their combinations are discussed. The method is applied to the standard database MUSK and the breast ultrasound image database. The results confirm that the method has higher accuracy comparing with that by using Citation-kNN algorithm.

Key words: Image recognition; Multi-Instance Learning (MIL); Citation-kNN; Distribution of samples; Locally weighted

1 引言

1997 年，Dietterich 等人[1]在对药物活性预测问题的研究中，提出了多示例学习(Multi-Instance Learning, MIL)的概念。在传统学习框架中，一个样本代表一个示例，即样本和示例是一一对应的关系，同时示例的标签全部已知或者全部未知；而在多示例学习中，一个样本被定义为一个包，其中包含了多个示例，即样本和示例是一对多的对应关系，同时样本(包)的标签已知但是示例的标签未知。所以多示例学习中的训练样本的歧义性与传统学习中样本的歧义性完全不同，这使得传统学习方法难以解决多示例问题[2]。

文献[1]提出了 APR (Axis- Parallel Rectangle) 学习算法来解决多示例药物活性预测问题；Maron 等人[3]提出了多样性密度(Diverse Density, DD)算法；Wang 等人[4]提出了 Bayesian-kNN 和 Citation-kNN 两种算法；Zhang 等人[5]将 DD 算法与 EM 算法相结合提出了 EM-DD 算法；Andrews 等人[6]对支持向量机进行了扩展，得到了多示例算法 Bag-SVM 和 Inst-SVM。

在多示例学习概念和算法提出以后，已经在多个领域得到了成功应用，如：药物预测[1]；图像分类、标注[7]、图像分割[8]；视频中人的识别[9]，股票选择[10]；索引网页推荐[11]；显微镜下的肺癌细胞图像识别[12]等。

Citation-kNN 算法通过结合惰性学习(lazy learning)和 Hausdorff 距离，对 k-近邻算法进行了扩展，使其能够处理多示例学习问题，并在标准测试数据集 MUSK 上取得了良好的结果。但其所采用的 0-1 投票决策方式在实际应用中存在着一定的缺陷。因为在实际分类系统中样本库并不是分布在连续的空间中，并且样本也并不能均匀分布在包空间的任何一个角落，所以实际样本库缺乏足够的样本来表示完整的包空间。在这种不完全库所能表达的

2012-01-06 收到，2012-12-31 改回
国家自然科学基金(61073128, 61100097)资助课题
*通信作者：丁建睿 jrding@hit.edu.cn

特征空间中，样本分布会表现出密集与稀疏，散乱与聚集等特征，同时样本也可能在一个聚集的中心位置或不同聚集的交界处，而 Citation-kNN 的决策方式无法很好地解决这些特殊情况。

本文针对 Citation-kNN 算法的不足进行了改进，提出了基于局部加权的 Citation-kNN 算法(Locally Weighted Citation-kNN, LWCitation-kNN)。实验证明，该方法在标准库 MUSK1 上的分类效果较 Citation-kNN 算法有明显提高，同时，该方法在超声乳腺肿瘤良恶性分类问题中也得到了很好的效果。

2 局部加权的 Citation-kNN(LW Citation-kNN)算法

本文提出的局部加权的 Citation-kNN 算法(LWCitation-kNN)主要针对 Citation-kNN 算法中决策部分的不足进行了改进，将包特征空间的分布形式进行细分，得到两种分布特征：距离分布和散乱分布。算法根据以上分布特征的具体情况设置相应的权值和决策准则。

LWCitation-kNN 和 Citation-kNN 的算法流程如图 1 所示。

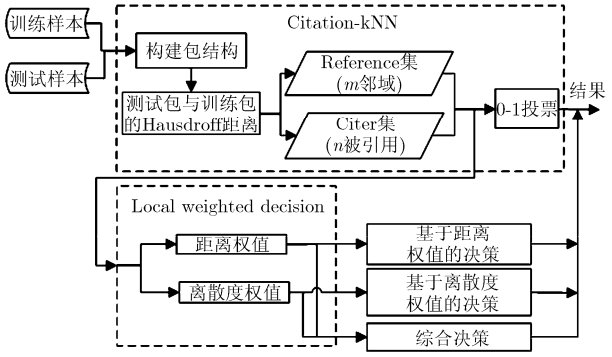


图 1 LWCitation-kNN 和 Citation-kNN 算法流程图

不同于 Citation-kNN 中的 0-1 决策方式，在 LWCitation-kNN 中，根据样本的分布情况，分别提出了基于距离权值、基于离散度权值以及综合这两种分布的决策方式。

2.1 基于距离权值的算法改进

图 2 是一种可能存在的包分布情况。

图中中心位置的白色方块为测试样本，周围的圆点为训练样本，颜色代表其类别，数量分别为 n_b 和 n_w。在图中所示的分布情况中，n_b/n_w=1，若采用 Citation-kNN 的 0-1 投票方式，则很难确定测试样本的类别。

基于距离加权的类别决策函数定义为

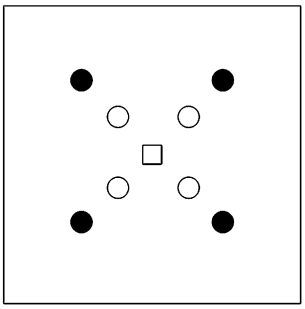


图 2 距离权值解决的情况

$$\left.\begin{aligned} f_d(X) &= \sum_{T_i \in \text{voter}} c_i \cdot W_i\left(d(T_i, X)\right) \\ L(X) &= \begin{cases} +1, & f_d(X) \geq 0 \\ -1, & f_d(X) < 0 \end{cases} \end{aligned}\right\} \tag{1}$$

本文所采用的距离加权函数为

$$W\left(d(T_i, X)\right) = \frac{d_{\max} - d(T_i, X)}{d_{\max} - d_{\min}} \tag{2}$$

其中 $d_{\max}$ 和 $d_{\min}$ 分别为训练样本到测试样本的最大和最小距离，该距离可以细分为局部形式和全局形式，采用局部形式时，$d_{\max}$ 和 $d_{\min}$ 为投票集合中的最大和最小距离；采用全局形式时，$d_{\max}$ 和 $d_{\min}$ 为所有训练样本到测试样本的最大和最小距离。

2.2 基于离散度权值的算法改进

离散度是用于描述特征空间中局部范围内的不同标签类别的样本点相互掺杂程度，离散度反映了样本局部范围内的可分程度。若在一个区域中心的样本点的可分性很差，则说明这个区域的散乱程度高，同时这个区域的样本点作为训练样本(投票者)的投票可信度就小，即权值要小。

图 3 是一种可能存在的包分布情况。

图 3 中的测试样本难以用 Citation-kNN 和基于距离权值的方法正确分类，但从离散度来看，上方白色训练样本的离散度更低，其权值应该更大，即测试样本判为白色更加合理。

基于离散度加权的决策函数为

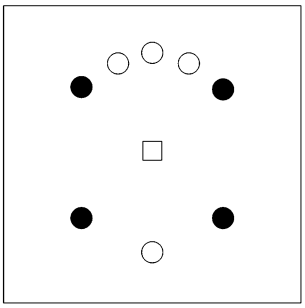


图 3 离散度权值所解决的情况

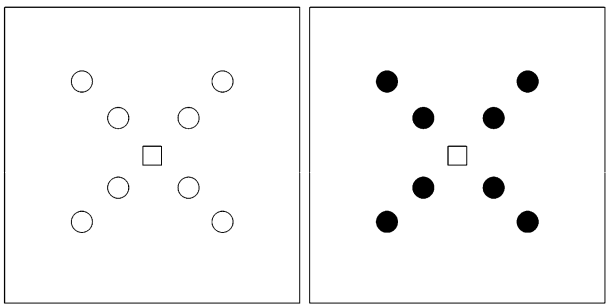
$$\left.\begin{aligned} f_s(X) &= \sum_{T_i\in\text{voter}} c_i\cdot S(T_i) \\ L(X) &= \begin{cases} +1, & f_s(X)\ge 0 \\ -1, & f_s(X)<0 \end{cases} \end{aligned}\right\} \tag{3}$$

其中 X 为测试示例包，T_i 为 X 投票集合中的训练示例包，其标签 c_i 用+1 或-1 来表示，$S(T_i)$ 为训练示例包的离散度权值，其定义为

$$S(T_i)=\left|\sum_{k\in\text{voter}}\left(c_k\cdot W\left(d(T_k,T_i)\right)\right)\right| \tag{4}$$

其中 T_i 为训练示例包，将 T_i 作为待测示例包，可以得到它的投票集合，T_k 为 T_i 投票集合中的训练示例包，其标签 c_k 用+1 或-1 来表示，$W(\cdot)$ 为根据式(2)所得到的 T_i 投票集合中的训练示例包的距离加权值。

训练样本中可能存在噪声，在图 4 所示的情况中，采用式(4)计算的离散度是相同的，但在图 4(b)中心的训练样本修正为黑色更为合理。



(a)正常情况　　(b)噪声情况

图 4 训练样本中存在噪声的情况

为解决这类问题，可以采用带有类别信息的离散度权值，对于不合理的训练样本可以采用式(5)进行修正：

$$\left.\begin{aligned} S(T_i) &= \sum_{T_k\in\text{voter}}\left(c_k\cdot W\left(d(T_k,T_i)\right)\right) \\ C(T_i) &= \begin{cases} 1, & S(T_i)>0 \\ -1, & S(T_i)<0 \end{cases} \end{aligned}\right\} \tag{5}$$

2.3 综合加权算法

综合考虑上面两种分布情况，将基于距离权值和基于离散度权值的方法相结合，得到基于综合加权的 Citation-kNN 改进算法，算法描述如表 1 所示。

3 实验结果

本文的实验样本库分别采用标准数据集 MUSK1 和 MUSK2 和由哈尔滨医科大学第二附属医院提供的乳腺超声图像库。实验分为两部分，第 1 部分是针对 MUSK1 和 MUSK2 标准数据集进行实验，主要比较本文方法与标准 Citation-kNN 算法在该数据集上的分类效果；第 2 部分是针对超声乳腺肿瘤图像库进行实验，分别对比分析不同的加权方法下的分类效果，同时选择适合于此类库的最佳参数组合。

表 1 综合加权算法描述

算法：基于综合加权的 Citation-kNN；
输入：训练样本集 $\{X_1, X_2, \cdots, X_L\}$，测试样本 X；
输出：$L(X)$；
(1)计算测试包 X 与训练包 X_i 的 hausdroff 距离；
(2)选出 m 个 reference 以及 n 个 citer，得到 X 的投票集合 voterset；
(3)基于 mindist 和 maxdist 求 voter 的投票者距离权值 W_d；
(4)对于训练样本集 $\{X_1, X_2, \cdots, X_L\}$ 中的每个训练样本 X_i
(4.1)选出 X_i 的 m 个 reference 以及 n 个 citer，得到 X_i 的投票集合 voterset；
(4.2)求 X_i 的离散度权值 S；
end
(5)voter 综合权值= W_d*S；
(6)最大化类别决策函数，得出测试样本的类别 $L(X)$。

实验中采用交叉验证的方法，将数据集随机分为 10 组，依次将其中 1 组作为测试样本，其他 9 组作为训练样本，在该训练样本上选取参数 c_n, r_n，并在测试样本上进行测试，最终结果取 10 次测试的平均值。实验采用准确率(ACC)来评价不同算法在数据集上的分类性能，其中参数 TP(True Positive)和 FN(False Negative)是被正确和错误判别的正例样本数，而 TN(True Negative)和 FP(False Positive)是被正确以及错误判别的反例样本数，准确率定义为式(6)：

$$\text{ACC}=(\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{TN}+\text{FP}) \tag{6}$$

通过对前面局部加权方法进行组合，可以得到 8 种加权方法，如表 2 所示。

3.1 MUSK 库的实验结果

MUSK 库包含 MUSK1 和 MUSK2 两个分子构

表 2 加权方式的组合

加权	加权方式
W_1	局部距离加权
W_2	全局距离加权
W_3	离散度加权
W_4	具有修正功能的离散度加权
W_5	局部距离加权+离散度加权
W_6	局部距离加权+具有修正功能的离散度加权
W_7	全局距离加权+离散度加权
W_8	全局距离加权+具有修正功能的离散度加权

造数据集，每个数据集中包含多个示例包，每个示例包包含多个示例，每个示例代表一种分子构造方法，其特点如表 3 所示。

表 3 MUSK 库

图像库属性	MUSK1	MUSK2
正负包比率	47:45	39:63
包平均示例数	5.2	64.7
正包平均示例数	4.4	26.1
负包平均示例数	6.0	88.6

实验中选择 reference 集合大小取值范围为 2～10, citer 集合大小取值范围为 0～10，在 MUSK1 和 MUSK2 上的实验结果如表 4 所示。

表 4 MUSK1 和 MUSK2 上的实验准确率(%)

加权	MUSK1 上的准确率	MUSK2 上的准确率
W_1	89.55	83.05
W_2	89.25	83.21
W_3	89.10	83.65
W_4	92.00	83.65
W_5	90.05	81.32
W_6	93.48	85.45
W_7	92.39	85.22
W_8	95.30	86.30

本文所采用的方法与其他多示例学习算法在 MUSK1 和 MUSK2 库上的性能比较结果如表 5 所示。

表 5 与其他算法的准确率比较(%)

方法	MUSK1	MUSK2
本文(W_8)	95.3	86.3
Citation-kNN	92.4	86.3
iterated-discrim APR	92.4	89.2
Bayesian-kNN	90.2	82.4
Diverse Density	88.9	82.5

结果表明本文的权值设置方法在 MUSK 库上有较好的效果，在综合加权方式下(W_8)达到最佳性能。本文方法同样具有较好的表现，准确率要高于 Bayesian-kNN 和 Diverse Density 算法，但略低于 iterated-discrim APR 算法，由于该算法针对 MUSK 数据集进行了优化[13]，因此并不具有代表性；另外一个原因是 MUSK 数据集符合多示例学习的标准定义，即：正例包中的正例数大于 1，负例包中所有示例均为负例，而 Citation-kNN 算法以及本文改进的算法，属于示例包级的算法，并没有考虑包中示例的正负问题，因此针对 MUSK 数据集的性能提高不大。

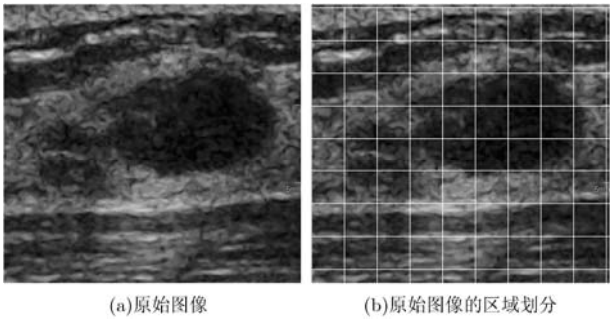
3.2 超声乳腺肿瘤良恶性分类实验

本文主要针对 116 幅乳腺超声图像(58 例良性，58 例恶性)进行分类，用以验证本文方法的效果。这些乳腺超声图像来源于哈尔滨医科大学第二附属医院的乳腺超声图像库，图像通过 GE VIVID7 超声影像系统和 5.6–14 MHz, 38 mm 线性探头采集。诊断结果均得到临床手术证实。

在传统超声乳腺肿瘤分类系统中，每个图像的特征信息需要在肿瘤感兴趣区域(Region Of Interest, ROI)部分提取，为了避免非肿瘤 ROI 区域信息对最终分类的干扰，应尽量提取出精准的 ROI 区域，但是现有超声肿瘤分割仍然是医疗图像领域的一个难题，并没有得到很好的解决[14]。同时在一些分类以及分割算法中，需要有医生手动提取的训练样本作为系统的初始条件，手画 ROI 边界在训练样本大量存在的时候是一项繁琐的工作。为避免 ROI 区域的分割误差对分类性能的影响，减轻医生的工作负担，本文将超声图像的分类问题转为多示例学习问题来进行处理。

将每幅超声图像划分为等大小的子区域，每个子区域作为示例，整幅图像作为示例包，提取每个子区域的纹理特征(灰度共生矩阵)[15]作为示例特征，如图 5 所示。

区域大小反映了所提取特征的分辨率，区域越小，则特征分辨率越高，不同区域之间的特征差异越小；区域越大，则特征分辨率越低，极端情况下，当区域大小为图像大小时，则多示例学习问题被还原为传统有监督学习问题，在区域大小的选择上，存在着既能够反映图像的局部特征，又能够使得区



(a)原始图像 (b)原始图像的区域划分

图 5 示例包的构建

域之间的可分性较好的一些划分，对于不同的图像和不同的样本库，该划分不尽相同。

实验中选择的子区域大小从 50×50～155×155 (步长为 15)，reference 集合大小取值范围为 2～10, citer 集合大小取值范围为 0～10。表 6 中给出了对图像进行不同划分时，采用不同加权方式时的分类准确率与 Citation-kNN, iterated-discrim APR 方法准确率的比较。

从实验结果来看，本文所采用的方法在不同的包构建方式下要普遍高于 Citation-kNN 和 iterated-discrim APR 的分类效果，在采用 W_8(全局距离加权+具有修正功能的离散度加权)加权方式时，效果最佳，同时也说明 iterated-discrim APR 算法在其他数据集上表现并不理想。

通过对实验结果进行分析，发现被错误分类的样本大部分集中在恶性样本集合内，这说明库中恶性样本之间的差异性较良性样本之间较大，使得恶性样本不如良性样本聚集，分布散乱，造成分类错误。

4　结束语

本文针对 Citation-kNN 算法进行了改进，充分考虑了样本的分布特征，提出了基于样本距离加权、基于样本离散度加权的方法，并对多种组合方式进行了实验。在标准数据集和乳腺超声图像数据集上的实验结果表明，该方法要明显优于 Citation-kNN 算法，具有良好的适应性。同时，在超声图像库上的实验结果表明，在医学图像分类中可以采用多示例学习方法，减少对感兴趣区域(ROI)准确定位的依赖，避免由于分割不准确所带来的特征提取误差。在今后的工作中，将进一步针对医学图像的特点，提取更为有效的特征，并探讨扩展传统的多示例学习方法，使之符合医学图像结构复杂，良恶性特征相互重叠等特性。

表 6 不同加权方式时的分类准确率(%)

加权	区域							
	50	65	80	95	110	125	140	155
W_1	68.35	80.17	78.95	72.41	**81.50**	75.86	77.50	77.59
W_2	70.30	81.03	77.50	74.14	**81.45**	78.45	76.35	80.17
W_3	77.12	83.62	81.45	76.72	**82.80**	78.45	80.55	81.03
W_4	81.49	87.93	85.70	88.79	**88.25**	87.07	87.00	88.79
W_5	80.15	85.34	80.30	74.14	**87.90**	77.59	79.50	82.76
W_6	83.40	89.66	85.45	84.48	**88.15**	86.21	85.90	88.79
W_7	74.55	83.62	80.30	75.86	**83.00**	81.90	75.43	81.03
W_8	**83.05**	**91.20**	**85.88**	**87.90**	**92.00**	**85.65**	**92.00**	**92.00**
Citation-kNN	75.20	81.90	75.50	75.00	81.60	77.59	80.00	78.45
iterated-discrim APR	56.00	52.30	50.00	51.05	53.45	53.45	51.05	52.66

参 考 文 献

[1] Dietterich T G, Lathrop R H, and Lozano-Pérez T. Solving the multiple-instance problem with axis-parallel rectangles [J]. *Artificial Intelligence*, 1997, 89(1–2): 31–71.

[2] Foulds J and Frank E. A review of multi-instance learning assumptions[J]. *Knowledge Engineering Review*, 2010, 25(1): 1–25.

[3] Maron O and Lozano-Pérez T. A framework for multiple-instance learning[J]. *Advances in Neural Information Processing Systems*, 1998, (10): 570–576.

[4] Wang J and Zucker J D. Solving the multiple-instance problem: a lazy learning approach[C]. Proceedings of the 17th International Conference on Machine Learning, San Francisco, CA, 2000: 1119–1125.

[5] Zhang Q and Goldman S A. EM-DD: an improved multiple-instance learning technique[J]. *Advances in Neural Information Processing Systems*, 2002, (14): 1073–1080.

[6] Andrews S, Hofmann T, and Tsochantaridis I. Multiple instance learning with generalized support vector machines [C]. Proceedings of the National Conference on Artificial Intelligence, Edmonton, Alta., Canada, 2002: 943–944.

[7] Katsamanis A, Gibson J, Black M P, *et al.*. Multiple instance learning for classification of human behavior observations [C]. Proceedings of Affective Computing and Intelligent Interaction, Memphis, TN, USA, 2011: 145–154.

[8] Vezhnevets A and Buhmann J. Towards weakly supervised semantic segmentation by means of multiple instance and

multitask learning[C]. Proceedings of Computer Vision and Pattern Recognition, San Francisco, CA, United States, 2010: 3249–3256.

[9] Guillaumin M, Verbeek J, and Schmid C. Multiple instance metric learning from automatically labeled bags of faces[C]. Proceedings of European Conference on Computer Vision, Heraklion, Crete, Greece, 2010: 634–647.

[10] Tsoumakas G, Katakis I, and Vlahavas I. Mining Multi-Label Data, Data Mining and Knowledge Discovery Handbook[M]. Berlin: Springer, 2010: 667–686.

[11] Zhou Z H, Jiang K, and Li M. Multi-instance learning based web mining[J]. *Applied Intelligence*, 2005, 22(2): 135–147.

[12] Zhu Liang, Zhao Bo, and Gao Yang. Multi-class multi-instance learning for lung cancer image classification based on bag feature selection[C]. Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD, Jinan, China, 2008, 2: 487–492.

[13] Zhou Z H. Multi-instance learning from supervised view[J]. *Journal Computer Science and Technology*, 2006, 21(5): 800–809.

[14] Cheng H D, Shan Juan, Ju Wen, *et al*.. Automated breast cancer detection and classification using ultrasound images: a survey[J]. *Pattern Recognition*, 2010, 43(1): 299–317.

[15] Liu B, Cheng H D, Huang J, *et al*.. Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images[J]. *Pattern Recognition*, 2010, 43(1): 280–298.

黄剑华： 男，1967 年生，教授，主要研究方向为医学图像处理、模式识别、人工智能.

丁建睿： 男，1973 年生，副教授，主要研究方向为医学图像处理、模式识别、人工智能.

刘家锋： 男，1968 年生，副教授，主要研究方向为模式识别、计算机视觉、字符识别技术、人体运动分析.

张英涛： 女，1975 年生，讲师，主要研究方向为医学图像处理、模式识别、人工智能.