DOI: 10.11992/tis.201807023
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190409.0946.012.html

基于异构距离的集成分类算法研究

张燕, 杜红乐
(商洛学院 数学与计算机应用学院, 陕西 商洛 726000)

摘 要: 针对异构数据集下的不均衡分类问题, 从数据集重采样、集成学习算法和构建弱分类器 3 个角度出发, 提出一种针对异构不均衡数据集的分类方法——HVDM-Adaboost-KNN 算法 (heterogeneous value difference metric-Adaboost-KNN), 该算法首先通过聚类算法对数据集进行均衡处理, 获得多个均衡的数据子集, 并构建多个子分类器, 采用异构距离计算异构数据集中 2 个样本之间的距离, 提高 KNN 算法的分类准性能, 然后用 Adaboost 算法进行迭代获得最终分类器。用 8 组 UCI 数据集来评估算法在不均衡数据集下的分类性能, Adaboost 实验结果表明, 相比 Adaboost 等算法, F_1 值、AUC、G-mean 等指标在异构不均衡数据集上的分类性能都有相应的提高。

关键词: 异构数据; 不均衡数据; 异构距离; 集成学习; 过取样; 欠取样
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 1673−4785(2019)04−0733−10

中文引用格式: 张燕, 杜红乐. 基于异构距离的集成分类算法研究 [J]. 智能系统学报, 2019, 14(4): 733–742.
英文引用格式: ZHANG Yan, DU Hongle. Imbalanced heterogeneous data ensemble classification based on HVDM-KNN[J]. CAAI transactions on intelligent systems, 2019, 14(4): 733–742.

Imbalanced heterogeneous data ensemble classification based on HVDM-KNN

ZHANG Yan, DU Hongle
(School of Math and Computer Application, Shangluo University, Shangluo 726000, China)

Abstract: A novel classification method, the heterogeneous value difference metric-Adaboost-KNN (HVDM-Adaboost-KNN), is proposed to achieve data resampling, to obtain an ensemble learning algorithm, and to construct a weak classifier for addressing the imbalanced classification of a heterogeneous dataset. This algorithm initially equalizes the dataset using a clustering algorithm to obtain several equalized data subsets and constructs several sub-classifiers. Further, the heterogeneous distance is used to calculate the distance between two samples in the heterogeneous dataset to improve the classification accuracy of the KNN algorithm. Subsequently, the Adaboost algorithm is used to iteratively obtain the final classifier. Eight groups of UCI datasets are used to evaluate the classification performance of the algorithm in imbalanced datasets. The Adaboost experimental results denote that the classification performance of indices, such as the F1 value, AUC, and G-means, using the heterogeneous imbalanced datasets was better when compared with that exhibited by other algorithms.

Keywords: heterogeneous data; imbalanced data; heterogeneous value difference metric; ensemble learning; over sampling; undersampling

传统的算法多是面向均衡数据集, 有较好的分类性能, 而实际应用中的数据集多是不均衡、异构的。面向不均衡数据分类的研究是数据挖掘、机器学习等领域当前的研究热点之一[1-16], 主要集中在数据层面[1-8] 和算法层面[9-16]。

数据层面的方法[1-5], 又称为重采样法, 多采用减少多数类样本或增加少数类样本, 使得数据

收稿日期: 2018−07−22. **网络出版日期**: 2019−04−10.
基金项目: 陕西省自然科学基础研究计划项目 (2015JM6347); 陕西省教育厅科技计划项目 (15JK1218); 商洛学院科学与技术项目 (18sky014); 商洛学院科技创新团队建设项目 (18SCX002); 商洛学院重点学科建设项目, 学科名: 数学".
通信作者: 杜红乐. E-mail: dhl5597@163.com.

集均衡化，过采样[1-3]是依据少数类样本的空间特征，通过一定的方法增加少数类样本数量，该方法容易导致过拟合，为此许多研究者提出了解决方法，例如合成少数类样本过采样技术 (synthetic minority oversampling technique,SMOTE) 及对 SMOTE 的改进算法；欠采样[5-6]则是通过一定的方法删除多数类样本中信息重复或者包含信息量较少的样本，但由于计算方法的不同，会删除包含丰富信息的样本，导致欠学习，为此研究者结合集成学习和重采样思想，不删除样本，而是对多数类样本按照一定策略进行抽取，然后与少数类样本一起构成训练子集[7]。

算法层面的方法则是提出新方法或者改进已有算法，减少数据不均衡对分类性能的影响，主要包括代价敏感学习[8-9]、单类学习、集成学习[10-14]等。其中集成学习方法是通过迭代逐步把弱分类器提升为强分类器，能够较好的提高分类器的性能，也是解决不均衡分类问题的常用方法，在一些领域得到应用[15-16]。文献 [11] 首先将数据集划分为多个均衡的子集，训练各个子集获得多个分类器，然后把多个分类器按照一定的规则 (文中给出 5 种集成规则) 进行集成，从而提高分类性能，该方法中对数据集的划分方法对最终的分类器性能有较大的影响，为此，文献 [5] 中通过聚类对多数类样本进行欠取样，获得与少数类样本数量相同的样本，然后采用 Adaboost 算法获得最终分类器，该方法保证所选样本的空间分布，但不能对分类错误样本和正确样本进行区别对待，而文献 [6] 利用抽样概率进行抽样，通过迭代不断修正抽样概率，对于分类错误的样本加大抽样概率，而分类正确的样本减小抽样概率，目的是争取下轮迭代中能选中进行学习。因此，本文方法既要充分考虑样本的空间分布，又要考虑到正确分类和错误分类样本之间的区别，采用聚类和抽样概率的方式进行数据集的划分，获得多个均衡的数据子集。

KNN 算法是一种简单而有效的分类算法，通过计算与样本最近的 K 个样本的类别来判断样本的类别，计算样本的 K 近邻经常采用欧氏距离、相关距离等，而对于异构数据集下，这些距离不能准确的表达样本的相似程度，针对此问题，Wilson 等[18]提出了异构距离，可以更准确的度量异构数据下 2 个样本之间的相似度，因此，本文采用基于异构距离的 KNN 算法作为弱分类器。

基于以上分析，本文提出一种面向不均衡异构数据的集成学习算法 (imbalanced heterogeneous data ensemble classification based on HVDM-KNN, HK-Adaboost 算法)，提高异构不均衡数据下的分类性能。该算法首先用聚类算法把数据集划分为多个均衡的数据子集，对于每个子集采用基于异构距离的 KNN 算法，然后用 Adaboost 算法对弱分类器进行训练，然后依据一定的评价指标进行调整，获得最终的强分类器。

1　相关概念

1.1　异构距离

定义 1 异构不均衡数据 (heterogeneous data)：设数据集 $\boldsymbol{X}$ 上的每条记录共有 m 个属性，$k(0<k<m)$ 个属性取值为连续值，其余 m-k 个属性取值为离散值，则称该数据集为异构数据集，若该数据集中类样本数量有较大差异，则称数据集为异构不均衡数据集。

根据样本到类中心的距离判断样本的类别，其实质就是计算样本与类的相似度，然而欧氏距离以及其它距离都不能准确度量异构数据集中记录的相似度。为了有效度量异构数据之间的相似度，实现数据分类，Wilson 等[18]提出 HVDM(heterogeneous value difference metric) 距离函数，能够反映出不同属性对相似度的影响，有效度量数据之间的差异，其定义如下：

定义 2 异构距离：设 $x,y\in\boldsymbol{X}$，则 x, y 之间的异构距离 $\boldsymbol{H}$(x,y) 定义为

$$\boldsymbol{H}(x,y)=\sqrt{\sum_{i=1}^{m}\boldsymbol{d}_j^2(x_j,y_j)} \tag{1}$$

式中：

$$\boldsymbol{d}_j(x_j,y_j)=\begin{cases}1, & x_j \text{ or } y_j\\ d_{\text{vdm}}(x_j,y_j), & x_j,y_j\text{为离散属性}\\ d_{\text{diff}}(x_j,y_j), & x_j,y_j\text{为连续属性}\end{cases} \tag{2}$$

$$d_{\text{diff}}(x_j,y_j)=\frac{|x_j-y_j|}{4\sigma_j} \tag{3}$$

$$d_{\text{vdm}}(x_j,y_j)=\sum_{i=1}^{k}\left|\frac{N_{a,x,i}}{N_{a,x}}-\frac{N_{a,y,i}}{N_{a,y}}\right| \tag{4}$$

式中：σ_j 为数据集上第 j 个属性的方差；$N_{a,x}$ 是数据集 $\boldsymbol{X}$ 上第 a 个属性取值为 x_a 的记录数；$N_{a,x,i}$ 为数据集 $\boldsymbol{X}$ 上第 a 个属性取值为 x_a 且类别为 i 的记录数。可以看到，异构距离依据数据集上每个属性的统计信息，而不是简单的 2 个属性的差值，在异构数据集上能够更加准确的描述 2 个样本之间的差异。

1.2　KNN 算法

K-近邻算法通过取测试样本的 K 个近邻，然后依据 K 个近邻的类别进行投票，确定测试样本

的类别, 由于算法简单、易于实现等特点, 被广泛应用。KNN 是依据 K 个近邻的类别决定测试样本的类别, 因此 K 个近邻的选取将影响算法的性能, 与测试样本越相近实质就是与测试样本越相似, 而计算相似度可以采用距离、夹角余弦等方法, 基于距离相似度中常采用欧氏距离, 尤其是对连续属性的向量之间, 能较好的度量 2 个向量间的相似程度。而对于既有数值属性又有字符属性的异构数据, 采用欧氏距离不能准确描述 2 个向量间的相似程度, 而实际应用中的数据有相当部分属于这样的异构数据集, 进行训练、分类时多是采用简单的数字替换, 把数据集转换为数值型的向量, 例如 red、blue、yellow 3 种颜色, 若用 1、2、3 进行代替, 原来 red 与 blue 之间的差别与 red 与 yellow 之间的差别是相同的, 但是用数字替换后的距离计算中, $(1-2)^2$ 与 $(1-3)^2$ 间的差别是不相同的, 因此本文的 KNN 算法中采用文献 [18] 给出的异构距离作为度量选择 K 个近邻样本。

1.3　数据均衡化

在集成学习中, 对多个训练集进行训练获得分类器, 然后把分类器进行集成, Adaboost 算法是通过修改每个样本的权重, 改变原有的数据分布从而得到新的训练集, 但是该方法无法改变 2 类样本数量不成比例的问题, 为此, 文献 [11] 提出一种新的面向不均衡数据的集成方法, 把多数类样本划分为多个与少数类样本规模相当的子集, 然后与少数类样本一起构成多个均衡的子集, 该方法的关键是如何对多数类样本进行划分; 文献 [5] 采用 K 均值聚类, 产生与少数类样本数量相同的簇数, 用簇代表原来的多数类样本, 从而对数据进行均衡化, 该方法会导致丢掉较多的样本, 进而导致出现欠学习现象; 文献 [6] 中依据抽样概率从多数类样本中随机抽取与少数类样本数量相等的样本, 与少数类一起构成训练集, 这样同样会导致丢掉较多的样本, 为此, 文中采用迭代的方式多次抽取, 每次抽取都会修改样本的抽样概率, 一方面该方法仍然会有部分样本不被选中, 另一方面, 抽取的样本无法保持原有数据的空间分布, 为此本文采用先聚类再抽取的方式对多数类样本进行划分, 划分方法如算法 1。

该方法抽取的样本包含有对应簇的空间信息, 使得针对每个子集获得的分类器有较好的分类性能, 另外选取合理的 m 值, 几乎不会有样本不被抽取, 并且抽取的样本与多数类样本有相似的空间分布。

算法 1　数据划分

输入　数据集

$$\text{train\_data} = \{(x_i, y_i)\}, x_i \in \mathbf{R}^n, y_i \in \boldsymbol{Y} = \{-1, 1\}$$

1) 把数据集分为min_data和maj_data, 并计算 2 类样本数量min_num和maj_num;

2) 若maj_num < min_num, 终止, train_data为最终数据集; 否则调用 Cluster(maj_data,min_num), 获得min_num个簇;

3) 分别从 min_num 个簇中放回抽取一个样本, 与少数类样本一起构成数据子集 $\boldsymbol{B}_m$。

输出　获得 m 个均衡的子集 $\boldsymbol{B}_m = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$。

2　HK-Adaboost 算法

算法需要依据每个子分类器的分类性能计算每个子分类器的权重, 这里采用子分类器的分类错误率描述子分类器的权重, 子分类器的权重表示为

$$w_i^t = \frac{1}{2}\ln((1-e_i^t)/e_i^t) \tag{5}$$

式中: $e_i^t = \sum_{i=1}^{l} \boldsymbol{D}_{ti}[y \neq h_i^t(x)]$表示子分类器的分类错误率。

在第 t 轮迭代中需要更新样本的权重, 改变子分类器的分类性能, 新样本的权重依据式 (6) 进行更新:

$$\boldsymbol{D}_{(t+1)i}(j) = \boldsymbol{D}_{ti}\exp(-w_i^t y_j h_i^t(x_j))/z_{ti} \tag{6}$$

每轮迭代结束, 对分类器进行集成时, 要考虑上轮所获得的分类器和本轮分类器进行集成, 集成方法如下:

$$\boldsymbol{H}_i^t(x) = \boldsymbol{H}_{t-1}(x) + \sum_{t=1}^{\mathrm{T}} w_i^t h_i^t(x) \tag{7}$$

计算每轮迭代结束所获得分类器的分类性能提升情况, 并获得该轮迭代后的分类器:

$$\boldsymbol{H}_t(x) = \boldsymbol{H}_i^t(x), t = \arg\max(\sigma_{ti}) \tag{8}$$

对$\boldsymbol{H}_t(x)$的分类性能表示为

$$a_t = 1 - \frac{1}{2l}\sum_{i=1}^{l}|\boldsymbol{H}_t(x_i) - y_i| \tag{9}$$

迭代结束后获得最终分类器为

$$\boldsymbol{H}(x) = \frac{1}{Z}\sum_{i=1}^{t} a_i \boldsymbol{H}_i(x) \tag{10}$$

算法 2 的详细过程如下:

算法 2　HK-Adaboost 算法

输入　数据集$\text{train_data} = \{(x_i, y_i)\}, x_i \in \mathbf{R}^n, y_i \in \boldsymbol{Y} = \{-1, 1\}$, 迭代次数 T, 基础分类器 C。

输出　最终分类器为$\boldsymbol{H}(x) = \text{sign}(\frac{1}{Z}\sum_{i=1}^{t} a_i \boldsymbol{H}_i(x))$。

1) 用 K 均值聚类算法对数据集进行划分, 获得 m 个均衡的子集$\boldsymbol{B}_m = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$;

2) 初始化样本权重: $\boldsymbol{D}_{1m} = (d_{1m}, d_{2m}, \cdots, d_{lm}) = 1/l$,

m 表示第 m 个子集；

for t=1:T

3) for i=1:M

调用分类器$C(\boldsymbol{B}_i, \boldsymbol{D}_{ti})$训练并获得第 i 个子集上的子分类器h_i^t；

利用式 (5) 计算每个分类器权重；

利用式 (6) 对各子集中样本进行权重更新；

利用式 (7) 获得本轮结束时分类器$\boldsymbol{H}_i^t(x)$；

end for i

4) 计算所得各子分类器$\boldsymbol{H}_i^t(x)$与上轮所得分类器的提升效果：$\sigma_{ti} = \boldsymbol{H}_i^t(x) - \boldsymbol{H}_{t-1}(x)$；

若分类效果没有提升，则结束该子集上的迭代，若有提升，则依据公式 (8) 选择提升效果最好的分类器作为第 t 次迭代后的分类器；

利用公式 (9) 计算分类器的分类性能a_t；

end for t

5) 利用式 (10) 获得最终分类器$\boldsymbol{H}(x)$。

1) 中，利用 K 均值聚类算法对多数类样本进行聚类，K 值为少数类样本数，得到 K 个簇，然后采用有放回抽样，从每个簇中随机取出一个样本，与少数量样本一起构成一个均衡的训练子集，然后重复该步骤，产生 m 个均衡的训练子集。

2) 是对每个训练子集中的每个样本赋予权值，初始权值都相等。

3) 是第 t 次迭代时，每个训练子集上的训练过程，依据 Adaboost 算法思想，对每个子集上的每个样本的权重进行更新，当第 t 次迭代结束后获得的分类器为第 t-1 次迭代获得分类器与第 i 个子集上前 t-1 次迭代获得的分类器的加权和。每个分类器的分类性能评价指标可以是 F_1 值、G-mean、AUC 等，本文算法中采用与 Adaboost 一致的评价方式-分类错误率。

式 (5) 是依据分类器对样本的分类错误样本的权重之和计算分类器的权重，然后依据 Adaboost 算法中更新样本权重的思想，应用式 (6) 更新每个样本的权重，第 t 轮迭代结束。

4) 是计算第 t 次迭代结束后获得的分类器，如果分类效果比上次迭代好，则进行后面步骤，否则丢弃该次迭代产生分类器。这里从 m 个子分类器中选择提升效果最好的分类器作为第 t 次迭代后的分类器，提升效果的评价仍然可以采用 F_1 值、G-mean、AUC 等评价指标进行评价，本文为简化算法，仍采用准确率作为评价指标，获得本轮迭代的分类器。然后计算本轮所得分类器的分类性能，并计算本轮迭代所得分类器在最终分类器中的权重。

5) 获得最终分类器，是每轮迭代所得分类器的加权和，其中Z为归一化因子，这里取$Z = \sum_{i=1}^{t} a_t$。

3　实验分析

本文选择 8 组不同的数据集进行实验，8 组数据集来自 UCI 数据库，Car Evaluation、TIC-TAC-Toe Endgame、Liver Disorders、Breast Cancer、Haberman's Survival、Blood transfusion、Contraceptive Method Choice 和 Teaching Assistant Evaluation，所选实验数据集的详细信息如表 1 所示，可以看到数据集在一定程度上都是不均衡的，并且数据集的各个属性是不连续的，另外本文算法是针对 2 类分类的，因此把数据集都转化为 2 类的数据集[17,19]。

表 1　实验数据集

Table 1　dataset

序号	数据集	属性	属性类别	多数类	少数类	比例
1	Car	6	6离散	1 210	384	3.15
2	Tic-Tac-To	9	9离散	626	332	1.89
3	Liver	7	7整数	200	145	1.38
4	Breast	9	8离散, 1整数	201	85	2.36
5	Haberman	3	3整数	225	81	2.78
6	Blood	4	4整数	570	178	3.20
7	Contraceptive	9	2整数, 7离散	844	300	2.81
8	Teaching	5	5整数	102	49	2.08

3.1　实验评价指标

针对均衡数据的分类多采用分类精度作为评价指标，而对于不均衡数据，更多关注的是少数类样本的分类情况，这种基于相同错分代价的评价指标不能很好描述分类性能。针对不均衡数据分类的评价指标多采用 Recall、Precision、F-mean、

G-mean、ROC 曲线和 AUC 等, 这些性能指标是基于混淆矩阵来计算的, 对于二分类问题的混淆矩阵如表 2 所示。

表 2　混淆矩阵

Table 2　Obfuscation matrix

类别	预测正类	预测负类
正类	TP	FN
负类	FP	TN

依据混淆矩阵可以计算上面评价指标的计算公式:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

$$\text{F}_1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{13}$$

$$\text{G}-\text{mean} = \sqrt{\frac{\text{TN}}{\text{TN}+\text{FP}}\times\frac{\text{TP}}{\text{TP}+\text{FN}}} \tag{14}$$

Recall 表示正类的查全率; Precision 表示正类的查准率; F-mean 同时考虑查全率和查准率, 只有当两个都大时 F-mean 的值才较大, 可以较好的描述不均衡数据集下的分类性能; G-mean 综合考虑 2 类的准确率, 任何一类准确率较低时, G-mean 的值都会较小, 因此能够较好评价不均衡数据集下的分类性能。

ROC 曲线则是以正负类的召回率为坐标轴, 通过调整分类器的阈值而获得一系列值对应的曲线, 由于 ROC 曲线不能定量评价分类器的分类性能, 因此常采用 ROC 曲线下的面积 AUC 来评价分类器的分类性能, AUC 值越大代表分类器的分类性能越好, 本文实验主要从上面所列评价指标来对比算法的性能。

3.2　异构距离有效性验证

表 3 中 HDVM_KNN 是指采用 HDVM 距离的 K 近邻算法, KNN 指采用欧氏距离的 K 近邻算法, 其中 K 取值为 5 的实验结果, SVM 是依据动态错分代价的支持向量机算法, SVM 和 KNN 是对数据进行归一化操作后采用 matlab 中自带的支持向量机算法进行实验的结果, Car 数据集的实验是从数据集中隔一条记录取一条的方式选取训练集, 用全部样本作为测试集的结果, 其他数据集均是全部数据既是训练集又是测试集的结果。实验结果主要依据常见的性能指标样本准确率 ACC、Recall、Precision、F-mean、G-mean 和 AUC 验证算法的性能, 实验更加关注少数类样本的分类性能。

表 3　实验结果

Table 3　Experimental result

数据集	算法	性能指标					
		ACC	Recall	Precision	AUC	F_mean	G_mean
Car	HDVM_KNN	**0.895 9**	**0.746 6**	**0.907 4**	**0.882 8**	**0.799 0**	**0.843 6**
	KNN	0.886 4	0.706 7	0.881 0	0.875 2	0.793 1	0.826 4
	SVM	0.623 6	0.377 8	0.545 5	0.677 4	0.526 8	0.592 6
Tic-Tac-Toe	HDVM_KNN	**0.955 3**	**0.905 4**	**0.904 0**	**0.905 9**	**0.853 5**	**0.904 4**
	KNN	0.932 9	0.787 9	0.772 4	0.781 2	0.588 7	0.778 1
	SVM	0.597 2	0.429 5	0.587 7	0.597 1	0.492 9	0.558 4
Liver	HDVM_KNN	**0.918 8**	**0.903 4**	**0.930 0**	**0.931 1**	**0.903 4**	**0.916 6**
	KNN	0.785 5	0.803 4	0.885 0	0.789 8	0.717 6	0.789 8
	SVM	0.684 1	0.609 8	0.680 0	0.675 1	0.647 2	0.676 9
Breast	HDVM_KNN	0.779 7	**0.727 3**	**0.930 3**	**0.699 0**	**0.571 4**	**0.765 2**
	KNN	**0.790 2**	0.720 0	0.925 4	0.655 3	0.533 3	0.755 3
	SVM	0.716 8	0.520 8	0.771 1	0.500 9	0.552 5	0.651 8
Haberman	HDVM_KNN	**0.843 1**	**0.779 7**	**0.942 2**	**0.791 3**	**0.657 1**	**0.818 0**
	KNN	0.797 4	0.711 1	0.942 2	0.767 6	0.507 9	0.760 0
	SVM	0.738 6	0.507 9	0.862 2	0.679 1	0.444 4	0.636 8
Blood	HDVM_KNN	**0.847 6**	**0.775 9**	**0.954 4**	**0.871 2**	**0.612 2**	**0.817 2**
	KNN	0.819 5	0.693 7	0.940 4	0.849 6	0.532 9	0.764 0
	SVM	0.493 3	0.309 3	0.361 4	0.806 7	0.462 4	0.536 9
Contraceptive	HDVM_KNN	**0.863 7**	**0.779 7**	**0.771 2**	0.661 0	0.707 2	**0.773 1**
	KNN	0.753 6	0.698 6	0.759 0	0.756 3	**0.730 9**	0.753 3
	SVM	0.758 3	0.631 1	0.671 4	**0.774 3**	0.590 5	0.662 6
Teaching	HDVM_KNN	**0.892 2**	**0.796 3**	**0.887 4**	**0.886 8**	**0.835 0**	**0.864 3**
	KNN	0.803 9	0.565 2	0.715 2	0.666 7	0.547 4	0.664 4
	SVM	0.107 8	0.340 6	0.384 1	0.609 3	0.502 7	0.536 8

由表 3 的实验结果可以看出，基于异构距离的 KNN 实验结果除了 Breast 数据集的准确率和 Contraceptive 数据集的 AUC、F_1 值外，其他指标都优于采用欧氏距离的 KNN 算法的实验结果，说明对于异构数据集，异构距离比欧氏距离能更准确的描述 2 个样本之间的相似度。

3.3　与其他算法的性能对比

用每个数据集的全部数据作为训练集，同时用该数据作为测试集的实验结果，详细实验数据如表 4 所示，其中 KNN 是 K 近邻算法、Adaboost 用的是 matlab 中自带的算法、OK-Adaboost 是本文所提算法 (K 近邻算法采用欧氏距离进行计算)、HK-Adaboost 是本文所提算法 (K 近邻算法采用异构距离进行计算)，实验中中 K 近邻算法的 K 取值为 5，划分子集数 m 取值为 5，Adaboost 算法迭代 100 次、OK-Adaboost 和 HK-Adaboost 算法迭代 20 次。

表 4　算法性能对比 1

Table 4　Algorithm performance comparison 1

数据集	算法	性能指标					
		ACC	Recall	Precision	AUC	F_mean	G_mean
Car	KNN	0.886 4	0.706 7	0.881 0	0.875 2	0.793 1	0.826 4
	Adaboost	0.922 2	0.901 0	0.967 3	0.993 0	0.848 0	0.933 0
	OK-Adaboost	0.967 4	0.942 7	0.981 7	0.998 6	0.993 0	0.958 8
	HK-Adaboost	**0.997 5**	**1**	**1**	**1**	**0.994 8**	**0.998 3**
Tic-Tac-Toe	KNN	0.932 9	0.787 9	0.772 4	0.781 2	0.588 7	0.778 1
	Adaboost	0.816 3	0.596 4	0.813 4	0.556 4	0.692 3	0.745 9
	OK-Adaboost	0.841 3	0.566 3	0.811 0	0.559 0	0.712 1	0.747 7
	HK-Adaboost	**0.937 4**	**0.891 6**	**0.943 6**	**0.857 4**	**0.908 0**	**0.926 0**
Liver	KNN	0.785 5	0.803 4	0.885 0	0.789 8	0.717 6	0.789 8
	Adaboost	0.811 6	0.724 1	0.814 0	0.811 7	0.763 6	0.796 0
	OK-Adaboost	0.785 5	0.648 3	0.776 3	0.773 6	0.717 6	0.757 4
	HK-Adaboost	**1**	**1**	**1**	**1**	**1**	**1**
Breast	KNN	0.790 2	0.720 0	0.925 4	0.655 3	0.533 3	0.755 3
	Adaboost	0.765 7	0.470 6	0.799 1	0.419 1	0.544 2	0.647 4
	OK-Adaboost	0.870 6	0.717 6	0.886 8	0.671 2	0.767 3	0.819 3
	HK-Adaboost	**0.940 6**	**0.894 1**	**0.955 4**	**0.858 5**	**0.899 4**	**0.926 6**
Haberman	KNN	0.797 4	0.711 1	0.942 2	0.767 6	0.507 9	0.760 0
	Adaboost	0.781 0	0.407 4	0.811 0	0.652 8	0.496 2	0.610 7
	OK-Adaboost	0.875 8	0.691 4	0.894 5	0.777 6	0.746 7	0.807 1
	HK-Adaboost	**1**	**1**	**1**	**1**	**1**	**1**
Blood	KNN	0.819 5	0.693 7	0.940 4	0.849 6	0.532 9	0.764 0
	Adaboost	0.814 2	0.353 9	0.826 0	0.720 9	0.475 5	0.582 3
	OK-Adaboost	0.907 8	0.809 0	0.940 2	0.889 1	0.806 7	0.871 4
	HK-Adaboost	**0.963 9**	**0.927 0**	**0.977 2**	**0.972 9**	**0.924 4**	**0.950 9**
Contraceptive	KNN	0.753 6	0.698 6	0.759 0	0.756 3	0.730 9	0.753 3
	Adaboost	0.720 1	0.519 1	0.708 5	0.937 3	0.612 8	0.671 9
	OK-Adaboost	0.843 8	0.780 3	0.844 9	0.976 0	0.809 9	0.833 8
	HK-Adaboost	**0.957 9**	**0.949 0**	**0.962 2**	**0.998 2**	**0.950 6**	**0.956 7**
Teaching	KNN	0.803 9	0.565 2	0.715 2	0.666 7	0.547 4	0.664 4
	Adaboost	0.853 3	0.625 0	0.844 8	0.985 3	0.731 7	0.774 9
	OK-Adaboost	0.746 7	0.250 0	0.735 3	0.985 3	0.387 1	0.495 1
	HK-Adaboost	**0.933 3**	**0.958 3**	**0.979 2**	**0.996 7**	**0.902 0**	**0.939 8**

由于训练集和测试集相同，各个算法的各项指标都比较好，但是每项指标本文算法的性能都优于其他算法，KNN 算法和 Adaboost 算法没有考虑数据集不均衡的问题，算法 OK-Adaboost 中若

分离器采用的是欧氏距离, 不能准确描述样本间的相似程度。

表 5 给出的是各个数据集, 取一半作为训练集、一半作为测试集。从第一条开始隔一条去一条作为训练集, 剩余为测试集, 实验中取 $m = 5$, $K = 5$, $T = 20$, 详细的实验结果如表 5 所示。由于采用一半作为训练集, 一半作为测试集, 实验所得各项性能指标明显比表 4 要差, 尤其是在 Car 和 Contraceptive 数据集上的分类性能, 可以看到基于欧氏距离的算法要比基于异构距离的算法效果要好, 查看样本之间的异构距离, 发现他们之间的差异远远小于欧氏距离之间的差异。分析数据发现计算样本$\frac{N_{a,x,i}}{N_{a,x}}$的值之间差异很小, 即属性值在各个类中出现的频率相近, 导致$\frac{N_{a,x,i}}{N_{a,x}} - \frac{N_{a,y,i}}{N_{a,y}}$值接近 0, 即任意 2 个样本之间的相似度都很高, 也就不能很好区分 2 个样本之间的差异。由此可以看到, 异构距离在这样的数据集中, 同样不能很好的计算 2 个样本的相似度。图 1 和图 2 是 Haberman 和 Blood2 个数据集下采用本文算法和 Adaboost 算法下的 ROC曲线。

表 5　算法性能对比 2

Table 5　Algorithm performance comparison 2

数据集	算法	性能指标					
		ACC	Recall	Precision	AUC	F_mean	G_mean
Car	KNN	0.865 7	0.838 5	0.944 6	0.979 7	0.750 6	0.856 3
	Adaboost	**0.883 3**	0.875 0	0.957 1	0.985 7	0.778 5	0.880 5
	OK-Adaboost	0.849 4	**0.994 8**	**0.997 9**	**0.999 0**	0.761 0	0.893 9
	HK-Adaboost	0.865 7	0.979 2	0.992 1	0.996 5	**0.778 5**	**0.901 4**
Tic-Tac-Toe	KNN	0.772 4	0.469 9	0.768 4	0.438 4	0.588 7	0.662 1
	Adaboost	0.730 7	0.524 1	0.784 2	0.480 6	0.623 7	0.693 2
	OK-Adaboost	0.699 4	0.747 0	0.801 9	0.499 6	0.625 0	0.704 7
	HK-Adaboost	**0.780 8**	**0.771 1**	**0.843 3**	**0.539 4**	**0.657 8**	**0.734 4**
Liver	KNN	0.641 6	0.465 8	0.663 8	0.877 1	0.523 1	0.598 9
	Adaboost	0.739 9	0.520 5	0.720 0	0.921 6	0.628 1	0.684 5
	OK-Adaboost	0.669 4	**0.712 3**	0.766 7	0.910 8	0.666 7	0.701 1
	HK-Adaboost	**0.751 4**	0.698 6	**0.770 8**	**0.940 8**	**0.681 5**	**0.727 5**
Breast	KNN	0.720 3	0.309 5	0.756 3	0.275 8	0.393 9	0.525 2
	Adaboost	0.671 3	0.238 1	0.728 8	0.202 7	0.298 5	0.450 3
	OK-Adaboost	0.692 3	0.714 3	0.851 9	0.488 0	0.576 9	0.698 6
	HK-Adaboost	**0.734 3**	**0.761 9**	**0.879 5**	**0.550 7**	**0.627 5**	**0.742 1**
Haberman	KNN	0.758 2	0.236 8	0.786 8	0.591 8	0.327 3	0.469 4
	Adaboost	0.712 4	0.157 9	0.773 0	0.543 0	0.240 0	0.386 9
	OK-Adaboost	0.712 4	0.631 6	0.858 6	0.689 0	0.521 7	0.683 2
	HK-Adaboost	**0.761 6**	**0.710 5**	**0.881 7**	**0.728 4**	**0.551 0**	**0.718 8**
Blood	KNN	0.767 4	0.329 9	0.796 9	0.694 9	0.423 8	0.551 1
	Adaboost	0.663 1	0.226 8	0.780 0	0.682 7	0.346 5	0.469 3
	OK-Adaboost	0.703 2	0.773 2	**0.887 8**	0.724 8	0.539 4	0.692 1
	HK-Adaboost	**0.778 1**	**0.773 2**	0.887 2	**0.727 3**	**0.543 5**	**0.694 9**
Contraceptive	KNN	0.715 0	0.366 7	0.788 4	0.897 9	0.402 9	0.554 6
	Adaboost	**0.812 9**	0.433 3	0.824 7	**0.970 5**	0.548 5	0.640 9
	OK-Adaboost	0.702 8	**0.653 3**	**0.859 5**	0.914 6	**0.556 8**	**0.701 9**
	HK-Adaboost	0.725 5	0.366 7	0.791 2	0.907 0	0.412 0	0.553 9
Teaching	KNN	0.613 3	0.160 0	0.666 7	0.865 6	0.216 2	0.366 6
	Adaboost	**0.680 0**	0.480 0	0.750 0	0.865 6	0.216 2	0.611 9
	OK-Adaboost	**0.680 0**	**0.680 0**	0.741 9	**0.904 8**	0.492 8	0.559 3
	HK-Adaboost	0.666 7	**0.680 0**	**0.804 9**	0.891 2	**0.576 3**	**0.669 9**

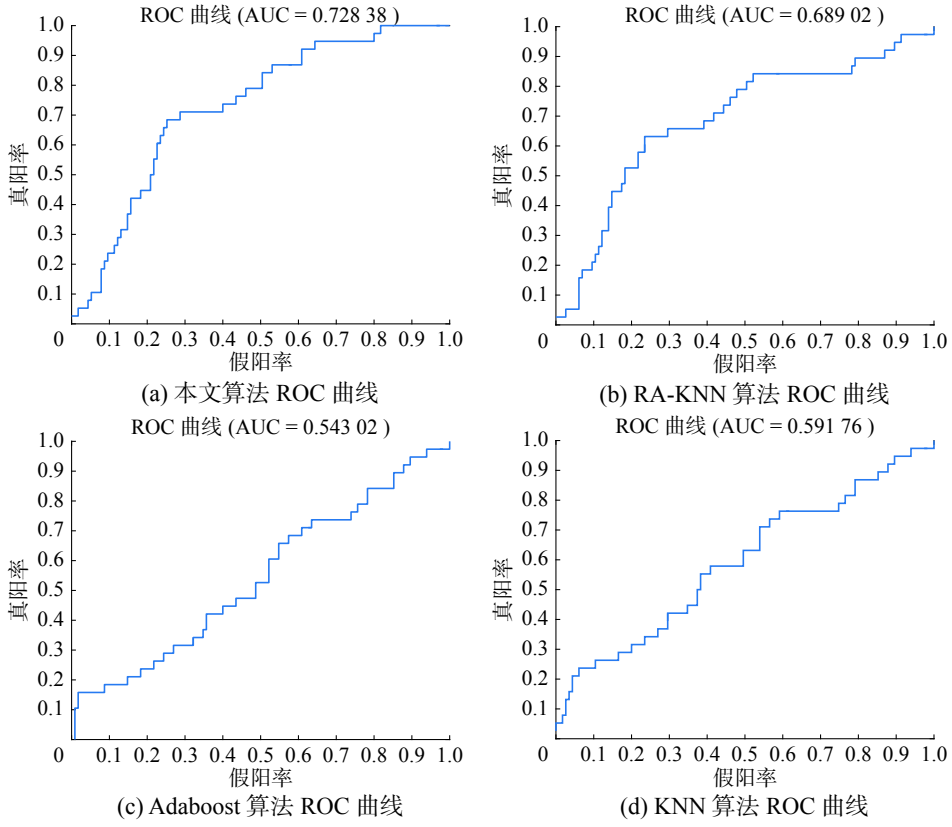

图 1　Haberman 数据集的 ROC 曲线对比

Fig. 1　RCO figure of Haberman dataset

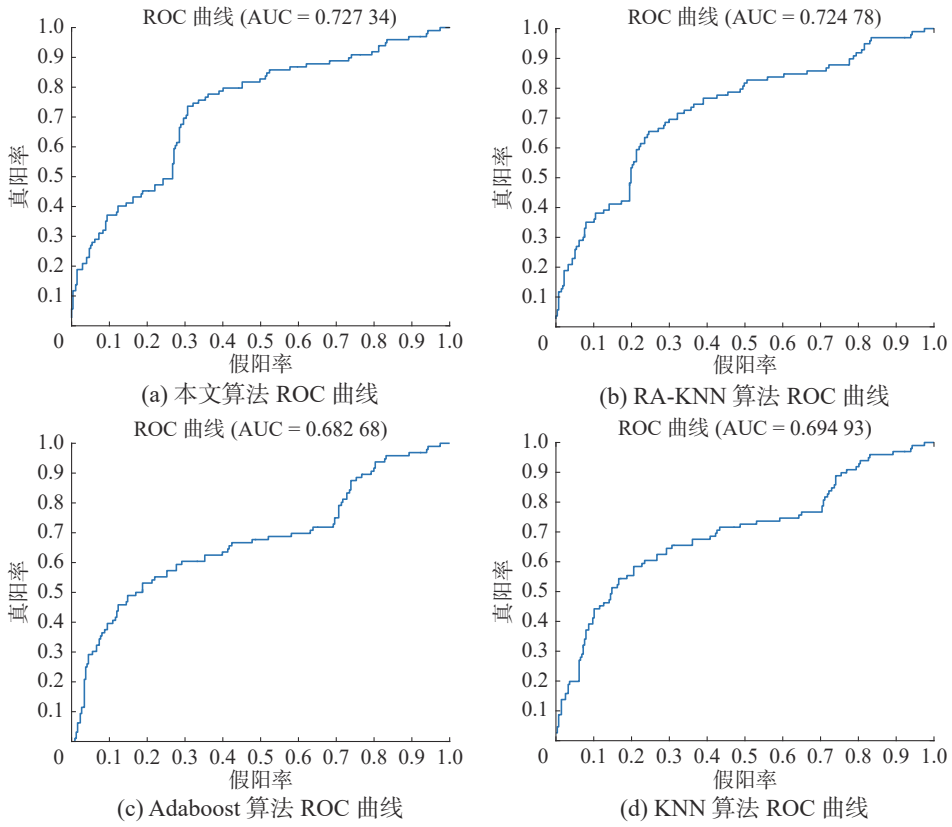

图 2　Blood 数据集 ROC 曲线对比

Fig. 2　RCO figure of Blood dataset

4 结束语

针对异构不均衡数据集下的分类问题, 本文提出一种面向不均衡异构数据的集成学习算法-HK-Adaboost 算法, 该算法从数据划分、基于异构距离的 KNN 及多个分类器的迭代集成等方面进行改进, 可以提高分类器在异构不均衡数据集下的分类性能, 通过在 8 组 UCI 异构数据集上进行实验, 验证了算法在异构不均衡数据上的分类性能。但实验中遇到一些问题, 如 Adaboost 算法对数据进行归一化后的分类性能会更差、异构距离计算时间复杂度高、数据划分子集的个数如何最优、如何扩展到多类问题等问题将是下阶段的主要工作。

参考文献:

[1] 胡峰, 王蕾, 周耀. 基于三支决策的不平衡数据过采样方法 [J]. 电子学报, 2018, 46(1): 136–144.
HU Feng, WANG Lei, ZHOU Yao. An oversampling method for imbalance data based on three-way decision model[J]. Acta electronica sinica, 2018, 46(1): 136–144.

[2] SÁEZ J A, LUENGO J, STEFANOWSKI J, et al. SMOTE–IPF: addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering[J]. Information sciences, 2015, 291: 184–203.

[3] PRUSTY M R, JAYANTHI T, VELUSAMY K. Weighted-SMOTE: a modification to SMOTE for event classification in sodium cooled fast reactors[J]. Progress in nuclear energy, 2017, 100: 355–364.

[4] MATHEW J, LUO Ming, PANG C K, et al. Kernel-based SMOTE for SVM classification of imbalanced datasets [C]//Proceedings of the 41st Annual Conference of the IEEE Industrial Electronics Society. Yokohama, Japan, 2015: 1127-1132.

[5] 武森, 刘露, 卢丹. 基于聚类欠采样的集成不均衡数据分类算法 [J]. 工程科学学报, 2017, 39(8): 1244–1253.
WU Sen, LIU Lu, LU Dan. Imbalanced data ensemble classification based on cluster-based under-sampling algorithm[J]. Chinese journal of engineering, 2017, 39(8): 1244–1253.

[6] 陈旭, 刘鹏鹤, 孙毓忠, 等. 面向不均衡医学数据集的疾病预测模型研究 [J]. 计算机学报, 2019, 42(3): 596–609.
CHEN Xu, LIU Penghe, SUN Yuzhong, et al. Research on disease prediction models based on imbalanced medical data sets[J]. Chinese journal of computers, 2019, 42(3): 596–609.

[7] JIAN Chuanxia, GAO Jian, AO Yinhui. A new sampling method for classifying imbalanced data based on support vector machine ensemble[J]. Neurocomputing, 2016, 193: 115–122.

[8] DU Hongle, TENG Shaohua, ZHANG Lin, et al. Support vector machine based on dynamic density equalization [C]//Proceedings of the Second International Conference on Human Centered Computing. Colombo, Sri Lanka, 2016: 58-69.

[9] ZHOU Yuhang, ZHOU Zhihua. Large margin distribution learning with cost interval and unlabeled data[J]. IEEE transactions on knowledge and data engineering, 2016, 28(7): 1749–1763.

[10] WANG Shuo, MINKU L L, YAO Xin. Resampling-based ensemble methods for online class imbalance learning[J]. IEEE transactions on knowledge and data engineering, 2015, 27(5): 1356–1368.

[11] SUN Zhongbin, SONG Qinbao, ZHU Xiaoyan, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern recognition, 2015, 48(5): 1623–1637.

[12] GUO Haixiang, LI Yijing, LI Yanan, et al. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification[J]. Engineering applications of artificial intelligence, 2016, 49: 176–193.

[13] WANG Qi, LUO Zhihao, HUANG Jincai, et al. A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM[J]. Computational intelligence and neuroscience, 2017, 2017: 1827016.

[14] POTHARAJU S P, SREEDEVI M. Ensembled rule based classification algorithms for predicting imbalanced kidney disease data[J]. Journal of engineering science and technology review, 2016, 9(5): 201–207.

[15] ZHAI Junhai, ZHANG Sufang, WANG Chenxi. The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers[J]. International journal of machine learning and cybernetics, 2017, 8(3): 1009–1017.

[16] YU Hualong, NI Jun. An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2014, 11(4): 657–666.

[17] HAQUE M N, NOMAN N, BERRETTA R, et al. Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification[J]. PLoS one, 2016, 11(1): e0146116.

[18] WILSON D R, MARTINEZ T R. Improved heterogeneous distance functions[J]. Journal of artificial intelligence research, 1997, 6(1): 1–34.

[19] ZHANG Yishi, YANG Anrong, XIONG Chan, et al. Feature selection using data envelopment analysis[J]. Knowledge-based systems, 2014, 64: 70–80.

作者简介：

张燕，女，1977 年生，讲师，主要研究方向为模式识别、机器学习。主持和参加省部级及企业合作项目 6 项。发表学术论文 10 余篇。

杜红乐，男，1979 年生，副教授，主要研究方向为数据挖掘、机器学习。主持或承担校级以上项目 12 项。发表学术论文 30 余篇，被 EI 检索 10 余篇。

第十七届中国机器学习会议（CCML 2019）

第十七届中国机器学习会议 (CCML 2019) 由中国人工智能学会和中国计算机学会联合主办，中国人工智能学会机器学习专业委员会和中国计算机学会人工智能与模式识别专业委员会协办，贵州大学承办。该系列会议每两年举行一次，现已成为国内机器学习界最主要的学术活动。此次会议将为机器学习及相关研究领域的学者交流最新研究成果、进行广泛的学术讨论提供便利，并且将邀请国内机器学习领域的著名学者做精彩报告。

重要时间

会议举办时间：2019 年 7 月 23 日—25 日

会议投稿截止时间：2019 年 3 月 25 日

会议录用通知时间：2019 年 5 月 10 日

征稿范围

机器学习的新理论、新技术与新应用；特征选择；聚类；人类学习的计算模型；流形学习与降维；异常检测；计算学习理论；信息检索；演化学习；监督学习；生物特征识别；符号学习；非监督学习；生物信息学；多 Agent 学习；半监督学习；距离度量学习；机器学习应用；强化学习；基于案例的推理；主动学习；多示例学习；增量学习与在线学习；多标记学习；神经网络；对复杂结构数据的学习；模式识别;集成学习；增强学习系统可理解性；大数据学习；多任务学习；数据挖掘与知识发现

投稿要求

论文必须未公开发表过，一般不超过 6 000 字；只接受中文稿；

论文应包括题目、作者姓名、作者单位、摘要、关键字、正文和参考文献；另附作者通讯；

地址、邮编、电话及 E-mail 地址；

学生 (不包括博士后和在职博士生) 第一作者的论文稿件请在首页脚注中注明，否则将不具有参选“优秀学生论文”的资格。

论文出版

会议录用论文将被推荐到《中国科学》、《软件学报》、《计算机科学与探索》、《模式识别与人工智能》、《智能系统学报》、《数据采集与处理》、《计算机科学》、《计算机应用》、《小型微型计算机系统》、《计算机工程与应用》、《南京大学学报(自然科学版)》、《山东大学学报(工学版)》、《南京师范大学学报》等期刊正刊发表。

投稿方式

本会议采用在线投稿方式，在线投稿网址为：https://easychair.org/conferences/?conf=ccml2019

咨询电话

16608512304(龙老师)、18586818471(徐老师)

会议邮箱：CCML2019@126.com

会议网址：http://bdiri.gzu.edu.cn/ccml2019