计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(7):121−126 [doi: 10.15888/j.cnki.csa.007001]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

实例加权类依赖 Relief①

邱海峰, 何振峰

(福州大学 数学与计算机科学学院, 福州 350116)
通讯作者: 邱海峰, E-mail: 1397750932@qq.com

摘　要: Relief 算法是一个过滤式特征选择算法, 通过一种贪心的方式最大化最近邻居分类器中的实例边距, 结合局部权重方法有作者提出了为每个类别分别训练一个特征权重的类依赖 Relief 算法 (Class Dependent RELIEF algorithm: CDRELIEF). 该方法更能反映特征相关性, 但是其训练出的特征权重仅仅对于衡量特征对于某一个类的相关性很有效, 在实际分类中分类精度不够高. 为了将 CDRELIEF 算法应用于分类过程, 本文改变权重更新过程, 并给训练集中的每个实例赋予一个实例权重值, 通过将实例权重值结合到权重更新公式中从而排除远离分类边界的数据点和离群点对权重更新的影响, 进而提高分类准确率. 本文提出的实例加权类依赖 RELIEF(IWCDRELIEF) 在多个 UCI 二类数据集上, 与 CDRELIEF 进行测试比较. 实验结果表明本文提出的算法相比 CDRELIEF 算法有明显的提高.

关键词: Relief 算法; 特征加权; 实例加权; 局部权重; 分类

引用格式: 邱海峰,何振峰.实例加权类依赖 Relief.计算机系统应用,2019,28(7):121–126. http://www.c-s-a.org.cn/1003-3254/7001.html

Instance Weighted Class Dependent Relief

QIU Hai-Feng, HE Zhen-Feng

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

Abstract: The Relief algorithm is a filtering feature selection algorithm that maximizes the instance margins in the nearest neighbor classifier in a greedy manner. Combined with the local weight method, the authors proposed a Class Dependent RELIEF (CDRELIEF) algorithm that trains one feature weight for each category. This method can better reflect the correlation of features. However, feature weight vector are only effective for measuring the correlation of features to a certain class, and classifying them in actual classification. In the actual classification, the classification accuracy is not high enough. In order to apply the CDRELIEF algorithm to the classification process, this study changes the weight update process, and assigns an instance weight to each instance in the training set. By combining the instance weight value into the weight updating formula, the influence of data points far from the classification boundary and outliers on weight updating is excluded, thereby improving the classification accuracy. The Instance Weighted CDRELIEF (IWCDRELIEF) algorithm proposed in this study is compared with CDRELIEF algorithm on multiple UCI 2-class datasets. Experimental results show that the algorithm proposed in this study has significantly improved the CDRELIEF algorithm.

Key words: Relief algorithm; feature weighting; instance weighting; local weight; classification

① 基金项目: 福建省自然科学基金 (2018J01794)
Foundation item: Natural Science Foundation of Fujian Province (2018J01794)
收稿时间: 2019-01-19; 修改时间: 2019-02-19; 采用时间: 2019-03-01; csa 在线出版时间: 2019-07-01

© 中国科学院软件研究所 http://www.c-s-a.org.cn

1 引言

作为一种重要的降维技术, 特征选择是一个热门的研究课题, 现有的特征选择方法可以分为两大类:过滤法和封装法. 过滤法先对数据集进行特征选择, 然后再训练学习器, 特征选择过程与后续学习器无关. 与过滤法不同, 封装法直接把最终要使用的学习器的性能作为特征子集的评价准则. 换言之, 封装法的目的就是为给定的学习器选择最有利于其性能的特征子集. 封装法的性能常依赖于具体的分类器, 而过滤法的性能通常无此依赖性, 由于过滤法的较好适应性, 相比封装法, 过滤法得到了更多的关注.

Relief 是一种广泛应用的过滤型方法, 在文献[1]中首次被提出用于二类数据的特征选择, 虽然 Relief 算法比较简单, 运行效率高, 并且结果也比较令人满意, 但是其局限性在于只能处理二类数据, Kononenko 将其扩展到多类情况, 提出 ReliefF 算法, 并在文献[2]中对 ReliefF 算法做了深入探讨. 虽然 Relief 已经得到较广泛的应用, 但它依然存在一些不足之处[3], 例如, 该类算法的数学形式依然没有得到很好的定义, 故它的特点和性质还难以得到深入的研究, 此外, 它依然缺乏强大的处理异常数据点的机制, 以及需要提高在噪音环境下的鲁棒性. 目前, 已有许多改进 Relief 算法的文献, 如迭代 Relief 算法 I-RELIEF[4], I-RELIEF 算法基于间隔最大化构造优化目标函数, 并以类 EM 算法的迭代策略来导出权重向量的学习规则. 另外, 文献[5]中提出了类依赖特征权重 Relief 算法, 由于不同类别数据点的各个特征重要性可能存在很大不同, 类依赖特征权重 Relief 算法为每个类别数据点单独训练一个权重, 以克服使用全局权重时不同类别数据点间特征重要性不同带来的影响.

另外, 已有许多结合实例选择和特征选择的研究. 有研究通过进化计算同时进行实例和特征选择以及加权[6], 提出了组合这四项任务的一般框架, 并对 15 种可能的组合的有用性进行了全面研究. 还有基于动态不完整数据粗糙集的增量特征选择[7], 提出了一种增量的特征选择方法, 可以加速动态不完整数据中的特征选择过程. 还有研究提出结合实例选择的三种策略进行基于实例的学习[8], 首先, 它使用 CHC 遗传算法的框架. 其次, 它包含了多次选择每个实例的可能性. 最后, 它使用的本地 k 值取决于每个测试实例的最近邻居, 这三种组合策略能够比以前的方法实现更好的减少, 同时保持与 k 近邻规则相同的分类性能. 目前已经有多个实例加权方案用于改进 Relief 算法的准确率, 如 Iterative Relief, I-RELIEF, 和 SWRF, 这些方法应用不同的实例加权方案并且有不错的效果.

为了克服类依赖特征权重的不足, 提高类依赖特征权重 Relief 算法准确率. 本文从局部特征权重数据分类的角度修改权重训练过程并引入实例权重来提高对边界点的敏感性. 本文第 2 部分先介绍 Relief 和类依赖 Relief, 并分析类依赖 Relief 的不足之处, 第 3 部分提出本文算法, 第 4 部分采用 8 个 UCI 数据集进行实验. 第 5 部分对文章内容进行总结.

2 Relief 和类依赖 Relief

Relief 算法中使用全局权重, 但是因为全局距离度量使用的特征权重没有区别不同的类别, 所以当一些特征对于不同的类表现得不同时会导致分类性能不佳. 相比全局权重, 局部特征权重更能反映不同类中相同特征的不同重要性, 因此, CDRELIEF 通过学习局部权重来提高权重关于类别的相关性, 目前, 已有许多方法[9,10]用于在局部区域上学习距离度量, 也有局部和全局相结合的距离度量[11]. 对于不同的类别来说特征权重是不一样的. 最有代表性的方法是类依赖加权距离度量 (CDW), 该距离与原型的类标签相关:

$$d_{CDW}(x,y)=\sqrt{\sum_{j=1}^{D}w_{c,j}^2(x_j-y_j)^2} \tag{1}$$

式中$d_{CDW}(x,y)$是点 x 和点 y 的类依赖加权距离, D 表示数据维度, c 是点 x 的类标签, $w_{c,j}$表示类别 c 第 j 个特征的权重.

2.1 Relief

Relief 特征加权[1]的核心思想是根据每一个特征区分不同类实例的能力来估计特征权值及其重要性, 给定一个包含 N 个实例的二类数据集 X, C 是类标签集合, x 是 X 中的一个实例, 每个实例$x=(x_1,x_2,\cdots,x_D)$是一个维度为 D 的实值向量. Relief 进行如下迭代学习: 随机的选取一个实例 x, 然后寻找同类最近实例$NH(x)$和异类最近实例$NM(x)$, 接着利用如下规则更新权值:

$$w_j=w_j+\frac{1}{T}\cdot|x_j-NM(x)_j|-\frac{1}{T}|x_j-NH(x)_j| \tag{2}$$

其中, $|x_j-y_j|$是用来计算两个实例第 j 维特征值的差异

© 中国科学院软件研究所 http://www.c-s-a.org.cn

程度, 即特征差值向量的绝对值向量. 具体地, Relief 算法如算法 1 所示.

算法 1. Relief 算法

① 给定一个包含 N 个实例和 D 个特征的二类数据集 X, 设置初始权值 $w_j=0\ (1\le j\le D)$以及最大迭代次数 T, 并且设置迭代初始值 t=1.

② 从数据集 X 中随机选取一个实例 x 并计算该实例的同类最近实例$NH(x)$和异类最近实例$NM(x)$.

③ 对于每一维权值, 利用式 (2) 更新权值.

④ 若 t=T, 算法结束, 否则 t=t+1 返回步骤②.

⑤ 输出更新以后的权值向量 w.

从最近邻居 Relief 发展出了考虑 K 个邻居的变体, 它的权重更新公式为:

$$w_j=w_j+\sum_{z\in KNN(x,l),l\neq c}|x_j-z_j|/T-\sum_{z\in KNN(x,c)}|x_j-z_j|/T \tag{3}$$

$KNN(x,c)$是 x 在X_c中通过欧氏距离求得的 K 个最近邻居的集合.

2.2 类依赖 Relief

Elena Marchiori[5]研究将 Relief 分解为类依赖特征权重, 并表示使用全局特征权重时将同一特征在不同类中的权重相加会抵消彼此关于单个类别的相关性, 导致特征关于单个类别的相关性可能不会被检测到, 因此他们提出将原来的所有数据共用一个特征权重改为一个类别一个特征权重, 类 c 的特征权重为w_c, 这样可以保留特征关于单个类别的相关性. 计算类别权重w_c时只选取类别为 c 的实例 x, 然后找该实例邻居, 对类别权重进行更新. 权重更新公式为:

$$w_c=\sum_{x\in X_c}\sum_{z\in KNN(x,c)}-|x-z|/T+\sum_{z\in \underset{l\neq c}{KNN}(x,l)}|x-z|/T \tag{4}$$

w_c被看做类别 c 的特征权重, X_c是类别为 c 的数据点集合, $KNN(x,c)$是 x 的同标签 k 近邻, $\underset{l\neq c}{KNN}(x,l)$是 x 的标签不为 c 的 k 近邻. 根据式 (4) 可以为数据集中每个类别数据求得一个特征权重.

3 实例加权类依赖 Relief

然而, 存在如下问题:在训练权重w_c过程中, 对属于类 c 的数据点x^1和不属于类 c 的数据点x^2, 目的是使x^1和x^2在w_c下的加权距离比x^1和同属于类 c 的数据点x^3在w_c下的加权距离要大. 即$\|x^1-x^2\|_{w_c}\ge\|x^1-x^3\|_{w_c}$.

但是在分类过程中, 与权重训练过程中使不同类数据点在同一个权重下比较距离大小的思想不同, 现有一个属于类别 c 的数据点x^1, 一个属于类别 l 的数据点x^2. 要正确分类一个属于类 c 的数据点 y, 需要满足条件: $\|y-x^2\|_{w_l}\ge\|y-x^1\|_{w_c}$, 即点 y 与点x^2在w_l下的加权距离要比 y 与点x^1在w_c下的加权距离要大. 点 y 和类 c 数据点x^1间的距离用w_c计算, $d(y,x^1)=\|y-x^1\|_{w_c}$和类 l 数据点x^2的距离用w_l计算, $d(y,x^2)=\|y-x^2\|_{w_l}$. 另外, 为了提高训练出的特征权重的分类精度, 本文将参与权重训练的实例限制在分类边界附近的点.

3.1 实例权重

本文中设置实例权重是一方面由于难分类的点是位于类边界的点, 那些远离类边界的点不容易分类错误. 当类边界处的点能够正确分类时远离类边界的点也能分类正确. 另一方面由于远离类边界的点在参与特征权重更新公式中对特征权重值造成的变化量较大, 而类边界处点对特征权重值造成的变化量较小, 因此远离类边界点的参与容易使得训练出的分类边界不能够正确分类类边界点. 因此只需要选取类边界附近的点参与分类边界的确定, 从而避免了远离类边界的点对特征权重的影响, 进而提高了分类准确率.

在权重更新过程中通过令远离类边界的数据点实例权重值为 0, 来排除远离类边界的数据点对特征权重更新的影响, 同时也排除了离群点的影响, 进而提高训练出的特征权重具有更高的分类精度. 实例权重公式如下:

$$IW(x)=\begin{cases}1, & \text{if}(\min(d_1/d_2,d_2/d_1)>threshold)\\ 0, & \text{if}(\min(d_1/d_2,d_2/d_1)<threshold)\end{cases} \tag{5}$$

其中, $threshold$ 是设定的阈值, 取值为 0 到 1 之间的值. d_1 是 x 到 k 个同类邻居的距离和, d_2 是 x 到 k 个异类邻居的距离和, 如果当前实例到同类邻居的距离之和 d_1 与到异类邻居的距离之和 d_2 的比值$d_1/d_2<threshold$说明当前实例点远离类边界, 实例点权重设为 0, 从而不影响特征权重更新. 另一方面, 当$d_2/d_1<threshold$时, 该实例点是离群点, 权重值也应该为 0, 从而排除了离群点对特征权重的影响.

3.2 新的特征权重更新公式

本文结合实例权重提出新的类依赖特征权重更新过程如下:

输入: 最大迭代次数 T, 以及一个包含 N 个实例的 D 维二类数据集: $X=\{(x^n,l(x^n))\}_{n=1}^N$, C 是数据的类别标

© 中国科学院软件研究所 http://www.c-s-a.org.cn

签集合, 因为算法用于二类数据集分类, 所以 C 只包含两个元素.

Step1. 为每个类别的特征权重设置初始权值 $w_{c,j}=0(c\in C,1\le j\le D)$.

Step2. 从集合 C 中取出一个类标签 c.

Step3. 从数据集 X 中随机选取一个类别为 c 的实例 x. 根据如下过程更新权重:

Step3.1. 找出 x 的 k 个同类最近邻居集合 $KNN(x,c)$, 还有 k 个异类最近邻居集合 $\underset{l\neq c}{KNN}(x,l)$, 以及到 $KNN(x,c)$ 中 k 个点的距离之和 d_1. 到 $\underset{l\neq c}{KNN}(x,l)$ 的 k 个点距离之和 d_2.

Step3.2. 将 d_1, d_2 代入式 (5) 计算 x 的实例权重 $IW(x)$.

Step3.3. c 为 x 的类标签, l 为不同于 c 的类标签, 即集合 C 中的另一个类. 对两个类别的特征权重 $w_{c,j}(j\in D)$, $w_{l,j}(j\in D)$ 进行更新:

$$\begin{aligned}&\text{If}\quad \sum_{z\in KNN(x,l),l\neq c}\|x-z\|-\sum_{z\in KNN(x,c)}\|x-z\|>0:\\&\quad w_{c,j}=w_{c,j}+IW(x)*\left(\sum_{z\in KNN(x,l),l\neq c}|x_j-z_j|-\sum_{z\in KNN(x,c),}|x_j-z_j|\right)/(T*k)\\&\text{else}:\\&\quad w_{l,j}=w_{l,j}+IW(x)*\left(\sum_{z\in KNN(x,l),l\neq c}|x_j-z_j|-\sum_{z\in KNN(x,c),}|x_j-z_j|\right)/(T*k)\end{aligned}\tag{6}$$

$\|x-z\|$表示点 x 和点 z 的欧式距离.

Step4. $t=T$, 则执行 Step5, $t<T$ 则返回 Step3.

Step5. 若 C 中所有值都取出, 算法结束, 输出 $w_c(c\in C)$ 否则返回 Step2.

本文提出的新特征权重更新公式中由于引入了实例权重避免远离类边界的点大幅度影响特征权重值而导致分类边界不能正确分类类边界点. 另一方面从局部权重分类的角度出发修改特征权重更新过程:当异类邻居的特征差值小于与同类邻居的特征差值时减小同类特征权重值, 当异类邻居的特征差值大于与同类邻居的特征差值时增大异类特征权重值.

4 实验与分析

实验中采用了 8 个二类 UCI 数据集 (见表 1). 所有数据都用 z-score 标准化进行预处理. 对每个数据集都进行了 10 折交叉验证, 取 10 折交叉准确率的平均值作为最后的准确率. 实验中阈值 *threshold* 取值范围从 0.1 到 0.9, 以 0.1 为间隔一共 9 个取值, 对每个数据集选择效果最好的那个. 为了验证本文方法的实际效果. 实验中取 k=5, 对比了本文提出的算法和类依赖 Relief 的准确率, 表 2 显示了两个算法的平均准确度以及标准差. 可以看到本文提出的算法对数据集的分类准确率有很明显的提高, 并且从图一可以看出相比 CDRELIEF, 当 k 取不同值时分类准确率更加稳定且明显高于 CDRELIEF.

表 1 数据集相关信息

数据集	实例个数	属性数目
HEART	270	13
B.CANCER	277	9
DIABETES	768	8
SPLICE	3175	60
THYROID	215	5
BREAST-W	683	9
BUPA	345	6
PIMA	768	8

表 2 CDRELIEF 和 IWCDRELIEF 算法准确率对比 (%)

数据集	CDRELIEF 算法 (STD)			IWCDRELIEF 算法 (STD)		
	k=3	k=5	k=7	k=3	k=5	k=7
HEART	76.90(0.096)	72.96(0.083)	76.30(0.089)	81.11(0.075)	80.74(0.086)	80.74(0.052)
B.CANCER	52.75(0.202)	61.82(0.160)	45.89(0.173)	75.10(0.072)	75.46(0.055)	69.23 (0.098)
DIABETES	55.21(0.043)	49.47(0.067)	51.69(0.038)	71.49(0.048)	70.19(0.034)	68.10(0.058)
SPLICE	52.16(0.006)	51.88(5.896)	52.07(0.004)	89.15(0.014)	93.57(0.009)	94.51(0.013)
THYROID	93.48(0.067)	93.10(0.067)	93.05(0.071)	96.28(0.046)	94.89(0.049)	93.51(0.055)
BREAST-W	96.04(0.021)	96.48(0.018)	96.92(0.016)	97.51 (0.015)	97.51(0.013)	97.07(0.009)
BUPA	45.29(0.094)	48.17(0.110)	42.93(0.065)	64.38(0.036)	61.45(0.067)	59.44(0.048)
PIMA	79.54(0.054)	81.90(0.039)	82.29(0.051)	88.42(0.026)	86.07(0.037)	86.47(0.025)

© 中国科学院软件研究所 http://www.c-s-a.org.cn

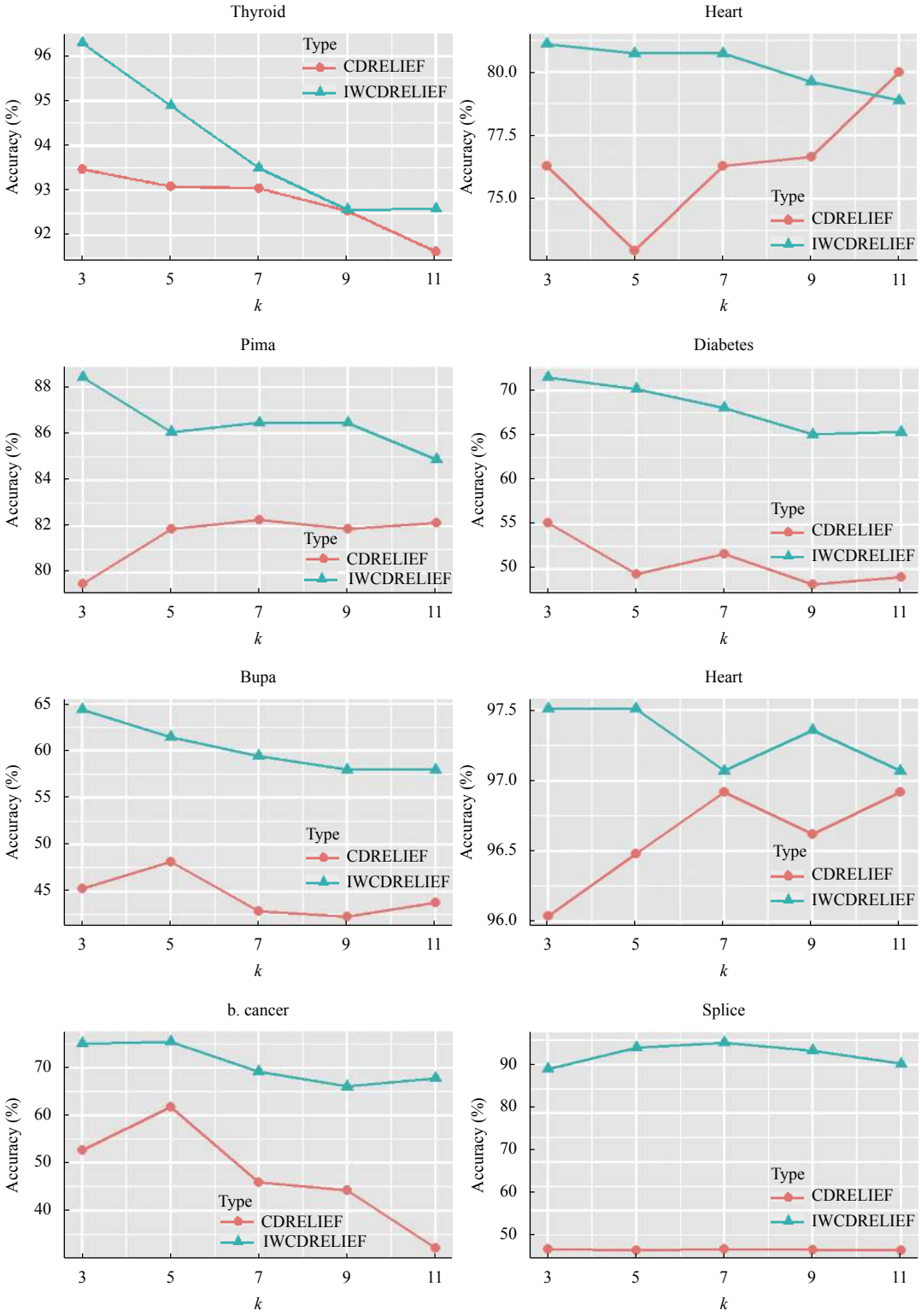

图 1 CDRELIEF 和 IWCDRELIEF 对实验数据集在不同 k 值下分类准确率的对比 (%)

5 结语

本文通过应用实例权重到类依赖 Relief 特征权重更新公式中, 提出了具有更好鲁棒性的实例加权类依赖 Relief 算法, 提出的新算法在 8 个二类 UCI 数据集

© 中国科学院软件研究所 http://www.c-s-a.org.cn

上验证了其有效性. 未来的工作中, 研究如何进一步提出更精确有效的实例加权方案以及如何结合快速学习理论加快算法执行速度, 减小算法时间复杂度是重点方向.

参考文献

1 Kira K, Rendell LA. A practical approach to feature selection. Proceedings of the 9th International Workshop on Machine Learning. Aberdeen, Scotland, UK, 1992: 249–256.

2 Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning, 2003, 53(1-2): 23–69.

3 Urbanowicz RJ, Meeker M, La Cava W, *et al*. Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics, 2018, 85: 189–203. [doi: 10.1016/j.jbi.2018.07.014]

4 Sun YJ. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligenc, 2007, 29(6): 1035–1051. [doi: 10.1109/TPAMI.2007.1093]

5 Marchiori E. Class dependent feature weighting and K-nearest neighbor classification. Proceedings of the 8th IAPR International Conference on Pattern Recognition in Bioinformatics. Nice, France, 2013: 69–78.

6 Pérez-Rodríguez J, Arroyo-Peña AG, García-Pedrajas N. Simultaneous instance and feature selection and weighting using evolutionary computation: Proposal and study. Applied Soft Computing, 2015, 37: 416–443. [doi: 10.1016/j.asoc.2015.07.046]

7 Shu WH, Shen H. Incremental feature selection based on rough set in dynamic incomplete data. Pattern Recognition, 2014, 47(12): 3890–3906. [doi: 10.1016/j.patcog.2014.06.002]

8 de Haro-García A, Pérez-Rodríguez J, García-Pedrajas N. Combining three strategies for evolutionary instance selection for instance-based learning. Swarm and Evolutionary Computation, 2018, 42: 160–172. [doi: 10.1016/j.swevo.2018.02.022]

9 Zhang H, Yu J, Wang M, *et al*. Semi-supervised distance metric learning based on local linear regression for data clustering. Neurocomputing, 2012, 93: 100–105. [doi: 10.1016/j.neucom.2012.03.007]

10 Jiao LM, Pan Q, Feng XX, *et al*. An evidential K-nearest neighbor classification method with weighted attributes. Proceedings of the 16th International Conference on Information Fusion. Istanbul, Turkey, 2013: 145–150.

11 Wang W, Hu BG, Wang ZF. Globality and locality incorporation in distance metric learning. Neurocomputing, 2014, 129: 185–198. [doi: 10.1016/j.neucom.2013.09.041]

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn