ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 2,Mar. 2018,pp. 317－322
DOI:10.16337/j.1004-9037.2018.02.014

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

# 基于样本邻域保持的代价敏感特征选择

余胜龙　赵　红

(闽南师范大学粒计算及其应用重点实验室,漳州,363000)

**摘　要**:特征选择是机器学习和数据挖据中一个重要的预处理步骤,而类别不均衡数据的特征选择是机器学习和模式识别中的一个热点研究问题。多数传统的特征选择分类算法追求高精度,并假设数据没有误分类代价或者有同样的代价。在现实应用中,不同的误分类往往会产生不同的误分类代价。为了得到最小误分类代价下的特征子集,本文提出一种基于样本邻域保持的代价敏感特征选择算法。该算法的核心思想是把样本邻域引入现有的代价敏感特征选择框架。在8个真实数据集上的实验结果表明了该算法的优越性。

**关键词**:特征选择;邻域保持;有监督学习;代价敏感

**中图分类号**:TP391.7　　**文献标志码**:A

## Cost-Sensitive Feature Selection Based on Sample Neighborhood Preserving

Yu Shenglong, Zhao Hong

(Fujian Key Laboratory of Granular Computing and Application, Minnan Normal University, Zhangzhou, 363000, China)

**Abstract**: Feature selection is an important preprocessing step in machine learning and data mining. Feature selection of class-imbalanced dataset is a hot topic of machine learning and pattern recognition. Most traditional feature selection classification algorithms pursue high precision, and assume that the data have no misclassification costs or have the same costs. However, in real applications, different misclassifications always tend to produce different misclassification costs. To get the feature subset with minimum misclassification cost, a supervised cost-sensitive feature selection algorithm based on sample neighborhood preserving is proposed, whose main idea is to introduce the sample neighborhood into the cost-sensitive feature selection framework. The experimental results on eight real-life data sets demonstrate the superiority of the proposed algorithm.

**Key words**: feature selection; neighborhood preserving; supervised learning; cost-sensitive

## 引　　言

在当前的数据时代,无论在数量还是在维度上,数据的产生都是多种多样的,而且数据也越来越大。如在医学、天文和文本等领域,高维数据给数据分析带来了诸多挑战。因此,数据降维[1]成为机器学习中一个必不可少的过程,而特征选择是主要的降维技术之一。特征选择是通过选取能表示所有特征的

**基金项目**:国家自然科学基金(61703196)资助项目;福建省教育厅科技项目(JAT160305)资助项目。
**收稿日期**:2016-06-07;**修订日期**:2016-06-29

一个较小的特征子集。从数据是否带有标签来看,特征选择可以分为有监督和无监督两种。有监督的特征选择是利用标签信息评价特征的重要度来进行选择,运用比较广泛的有监督特征选择算法有 Fisher Score[2],Relief 和 ReliefF[3]。无监督的特征选择算法[4]是根据自己数据特征的关系来进行特征选择。

一部分有监督特征选择算法以追求高精度为目的,却忽略了误分类代价或默认误分类代价相同。然而,在现实应用中,不同的误分类通常会导致不同的误分类代价。例如,在交易过程中存在两种类型的误分类。类型Ⅰ定义为将正常交易误分为欺骗交易;类型Ⅱ定义为将欺骗交易误分为正常交易。类型Ⅰ的错误代价是工作人员重新审查交易;类型Ⅱ的错误代价是造成了巨大的金钱损失。显然,类型Ⅰ导致的错误代价要小于类型Ⅱ导致的错误代价。另一部分有监督特征选择算法以选择有益的特征为目标,假设不同类别的样本具有相等的权重,这种认为数据本身有均衡样本类别的想法会导致在数据具有不均衡的样本类别时,有监督特征选择算法的效果大打折扣。因此在现实应用中必须考虑数据类别不均衡问题。

20 世纪 90 年代,代价信息被考虑到算法中,因而提出了代价敏感学习,它是机器学习领域十大研究热点之一[5]。至今,众多研究人员提出了许多代价敏感学习算法[6-8]来解决代价敏感问题和类不均衡问题,并在不同研究领域证实了算法的有效性[9-11]。现实应用中的代价主要分为两类:误分类代价和测试代价。其中,误分类代价可分为:基于样本的误分类代价[12]和基于类别的误分类代价[13]。本文的算法主要是在基于类别的误分类代价基础上再引入邻域实现的[14-16]。

邻域在特征选择中有着广泛的应用。文献[17]提出一种基于邻域粗糙集的测试代价属性约简,但其代价设置未考虑与邻域大小的关系。本文利用邻域可以保持样本局部结构的性质,同时引入代价敏感,提出一种新的基于样本邻域保持的代价敏感特征选择算法(Cost sensitive feature selection based on sample neighborhood preserving,CSFN-SNP)。首先,根据邻域保持局部结构的性质得出邻域矩阵[18];其次,引入代价矩阵和样本重要度,在邻域矩阵上计算每个特征的分数,并对每个特征分数使用排序算法进行排序,从而返回特征排序。实验结果表明,提出的代价敏感特征选择算法具有很好的性能。

## 1 相关工作

在现实应用中,不同的误分类通常会导致不同的误分类代价。现假设有 $c$ 类数据样本,并且,假设将第 $i$ 类 ($i \in \{1,2,\cdots,c-1\}$) 数据样本误分类为第 $c$ 类数据样本造成的代价要高于将第 $c$ 类数据样本误分类为第 $i$ 类数据样本的代价。基于该假设,将第 1 类到第 $c-1$ 类设为"组内"类,将第 $c$ 类设为"组外"类。根据文献[9],这样的误分类代价可分为 3 类:

(1) 误识别代价 $C_{\mathrm{II}}$:将"组内"类中某一类的数据样本误分类为"组内"类中另一类数据样本产生的代价;

(2) 误接受代价 $C_{\mathrm{OI}}$:将"组外"类数据样本误分类"组内"类数据样本所产生的代价;

(3) 误拒绝代价 $C_{\mathrm{IO}}$:将"组内"类数据样本误分类"组外"类数据样本所产生的代价。

表 1 代价矩阵

Tab. 1 Cost matrix

| | $I_1$ | $\cdots$ | $I_{c-1}$ | $O$ |
|---|---|---|---|---|
| $I_1$ | 0 | $\cdots$ | $C_{\mathrm{II}}$ | $C_{\mathrm{IO}}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $I_{c-1}$ | $C_{\mathrm{II}}$ | $\cdots$ | 0 | $C_{\mathrm{IO}}$ |
| $O$ | $C_{\mathrm{OI}}$ | $\cdots$ | $C_{\mathrm{OI}}$ | 0 |

由常识可知,代价 $C_{\mathrm{IO}}$,$C_{\mathrm{II}}$ 和 $C_{\mathrm{OI}}$ 的值一般不相等。令 $\mathrm{Cost}(i,j)$ ($i,j \in \{1,2,\cdots,c\}$) 为将第 $i$ 类样本误分为第 $j$ 类样本的代价,可以构建如表 1 所示的代价矩阵 $\boldsymbol{C}$。正确预测则没有代价,即对角元素全是 0。但在现实应用中,代价矩阵通常由用户或该领域的专家给出。根据代价矩阵,定义函数 $f(\alpha)$ 来衡量第 $\alpha(1 \leqslant \alpha \leqslant c)$ 类样本的重要度,即

$$f(\alpha)=\begin{cases}(c-2)C_{\mathrm{II}}+C_{\mathrm{IO}} & \alpha=1,\cdots,c-1\\ (c-1)C_{\mathrm{OI}} & \text{其他}\end{cases} \tag{1}$$

## 2 基于邻域的代价敏感特征选择

本节提出了一种基于样本邻域保持的代价敏感特征选择算法 CSFN-SNP,该算法是在代价敏感信息的基础上引入邻域,使得每个样本的邻域内存在 $k$ 个节点的邻接矩阵,并且每个样本的每个特征都在其自己的邻域上讨论,保持了样本的局部结构,可以得到较优的特征子集。

给定一个数据集 $\boldsymbol{X}^{n\times m}$,$n$ 和 $m$ 分别表示样本数和特征数。使用 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 来表示 $n$ 个样本,其中 $\boldsymbol{x}_i\in\mathbf{R}^{1\times m}$。同样,使用 $\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m$ 来表示 $m$ 个特征,其中 $\boldsymbol{f}_i\in\mathbf{R}^{n\times 1}$。定义标签 $\boldsymbol{Y}=[y_1;y_2;\cdots;y_n]$,其中 $y_i\in\mathbf{R}$。令 $\boldsymbol{x}_i$ 为数据集中第 $i$ 个样本,$f_{ri}$ 表示样本 $\boldsymbol{x}_i$ 的第 $r$ 个特征值($r=1,2,\cdots,m$)。而对于本文,先找出特征的一个子集$\{\boldsymbol{f}'_1,\boldsymbol{f}'_2,\cdots,\boldsymbol{f}'_d\}$,其中 $d<m$,使得样本分类的误分类代价尽可能小。

首先对数据集进行正规化处理,其次构造邻域矩阵 $\boldsymbol{G}=\{\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_n\}$,其中 $\boldsymbol{G}_i$ 表示第 $i$ 个样本。构造邻域矩阵 $\boldsymbol{G}$ 有两种方法:(1) $K$ 近邻($K$ nearest neighbor,KNN):如果 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 是 $K$ 近邻。(2) $\varepsilon$ 近邻($\varepsilon$ neighborhood):如果第 $i$ 个样本和第 $j$ 个样本之间满足 $\|\boldsymbol{x}_i-\boldsymbol{x}_j\|\leqslant\varepsilon$。

在实际应用中,第 2 种方法很少用,因为很难找到一个好的 $\varepsilon$。本文使用第 1 种方法 KNN 来构造邻域矩阵 $\boldsymbol{G}$。在此基础上再引入代价,第 $r$ 个特征的 CSFN 得分 $C_r$($C_r$ 越小越好),即有

$$S_r=\sum_{\boldsymbol{G}_i}\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M}f(y_i)(f_{ri}-f_{rj})^2 \tag{2}$$

$$H_r=\lambda\sum_{\boldsymbol{G}_i}\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in C}\mathrm{Cost}(y_i,y_j)(f_{ri}-f_{rj})^2 \tag{3}$$

$$C_r=S_r-H_r \tag{4}$$

式中:定义 $M=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid\boldsymbol{x}_i,\boldsymbol{x}_j$ 属于同一类别$\}$,$C=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid\boldsymbol{x}_i,\boldsymbol{x}_j$ 属于不同类别$\}$;$\lambda$ 为一权重参数,用以调节样本误分类代价矩阵的权重。

本文提出的 CSFN-SNP 算法的时间复杂度为 $O(nm^2+nc)$,其算法步骤表述如下。

**算法 1** 基于样本邻域保持的代价敏感特征选择

输入:数据集 $\boldsymbol{X}^{n\times m}$,标签 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_n]$,参数 $k$ 和 $\lambda$

输出:特征子集$\{\boldsymbol{f}'_1,\boldsymbol{f}'_2,\cdots,\boldsymbol{f}'_d\}$

(1) 计算每个样本之间的相互距离,根据邻域的性质,找到每个样本的 $k$ 近邻(包括样本自己),得到每个样本的邻域矩阵 $\boldsymbol{G}_i$ ($1\leqslant i\leqslant n$);

(2) 根据表 1 和式(1)分别得出代价矩阵 $\boldsymbol{C}$ 和样本重要度 $f(\alpha)$;

(3) 在每个邻域矩阵 $\boldsymbol{G}_i$ 的基础上,根据式(4)逐一计算出每个特征的得分 $C_r$($1\leqslant r\leqslant m$);

(4) 对得到的 $C_r$ 进行排序,并返回特征排序,选取前 $d$ 个特征。

## 3 算法性能对比实验

### 3.1 实验设置

为了验证本文算法的有效性,将其和现有的代价敏感特征选择算法在 UCI 数据集上做了相关的对比实验。这 8 个 UCI 数据集分别是 Heart,Australian,German,Wpbc,Vehicle,Glass,Landsat 和 Segment。为了准确反映 CSFN 算法在类不均衡情况下的性能,Vehicle,Glass,Landsat 和 Segment 数据集的类别数是不均衡的。表 2 给出了数据集的详细信息。所有实验均在 MATLAB 平台上实现。选取的现有代价敏感特征选择算法包括

(1) Baseline:选择所有的特征。

(2) 基于最大方差的代价敏感特征选择算法(Cost-sensitive variance score, CSVS)[19]:目标是找到组内样本距离样本中心尽可能比组外样本距离样本距离中心近的特征。

(3) 基于约束保持的代价敏感特征选择算法(Cost-sensitive constraint score, CSCS)[19]:利用同一类别样本间距小于不同类别样本间距的特征进行逐一打分。

对于算法 CSFN-SNP 和 CSCS,用网络搜索策略来调节参数 $\lambda$,设定参数集合为{1 000,100,10,1,0.1,0.01,0.001}。CSFN-SNP 算法的近邻参数 $k$ 设置为 5。设置 $C_{\mathrm{II}}$,$C_{\mathrm{OI}}$ 和 $C_{\mathrm{IO}}$ 的值分别为 1,2 和 20。对选择的特征子集使用 1-$NN$ 来分类,同时将得到的误分类代价和分类精度作为参考指标。实验结果为 5 次 5 折交叉验证结果的平均值。结合特征选择数和参数,将每个算法的最佳实验结果分别在表 3 和表 4 中列出,表 3 列出了最小误分类代价和所对应的特征数,表 4 列出了分类精度,并且在表 3 和表 4 中最佳结果使用粗体来标出,次好的结果用下划线标出,代价旁边括号内表示对应最小代价的特征数。图 1 显示了 4 种算法在 8 个数据集上随特征数变化而得到的误分类代价的对比结果。

**表 2 数据集信息**

**Tab. 2 Information of data sets**

| 数据集 | 样本 | 特征 | 类别 |
|---|---|---|---|
| Heart | 270 | 13 | 2 |
| Australian | 690 | 14 | 2 |
| German | 1 000 | 24 | 2 |
| Wpbc | 198 | 33 | 2 |
| Vehicle | 846 | 18 | 4 |
| Glass | 214 | 9 | 6 |
| Landsat | 2 000 | 36 | 6 |
| Segment | 2 310 | 18 | 7 |

**表 3 不同特征选择算法的误分类代价**

**Tab. 3 Misclassification cost of different feature selection algorithms**

| Algorithm | CSFN-SNP | CSCS | CSVS | Baseline |
|---|---|---|---|---|
| Heart | **65.8(1)** | 242.0(13) | 148.9(3) | 243.7 |
| Australian | **152.5(2)** | 285.2(12) | 288.9(12) | 315.8 |
| German | **588.8(6)** | 713.7(23) | 711.5(21) | 759.6 |
| Wpbc | **77.0(1)** | 128.8(33) | 132.8(26) | 135.7 |
| Vehicle | **105.7(13)** | 146.0(14) | 173.0(16) | 177.6 |
| Glass | **34.2(1)** | 54.2(1) | 55.3(8) | 55.6 |
| Landsat | **279.0(33)** | 287.5(32) | 314.6(36) | 320.0 |
| Segment | **16.2(9)** | 17.5(16) | 17.2(17) | 18.0 |

**表 4 不同特征选择算法的分类精度**

**Tab. 4 Accuracy of different feature selection algorithms** %

| Algorithm | CSFN-SNP | CSCS | CSVS | Baseline |
|---|---|---|---|---|
| Heart | 57.0 | 57.8 | **72.7** | 57.6 |
| Australian | 77.3 | 80.1 | **80.3** | 79.2 |
| German | 62.4 | **66.6** | **66.6** | 65.2 |
| Wpbc | 70.9 | **71.0** | 68.6 | 69.4 |
| Vehicle | **71.3** | 69.8 | 69.5 | 69.3 |
| Glass | **32.5** | 26.0 | 30.1 | 29.5 |
| Landsat | 88.4 | **88.9** | 88.3 | 88.3 |
| Segment | **96.5** | 96.3 | 96.2 | 96.2 |

## 3.2 结果分析

从表 3 可以看出,算法 CSCS 和 CSVS 在特征选择时,有较小的误分类代价和与之对应的特征数。而本文提出的 CSFN-SNP 算法使得每个样本保持局部的结构,可以比 CSCS 和 CSVS 算法在 8 个数据集上获得更小的误分类代价,且 CSFN-SNP 算法在获得最小误分类代价时所需要的特征数大部分小于

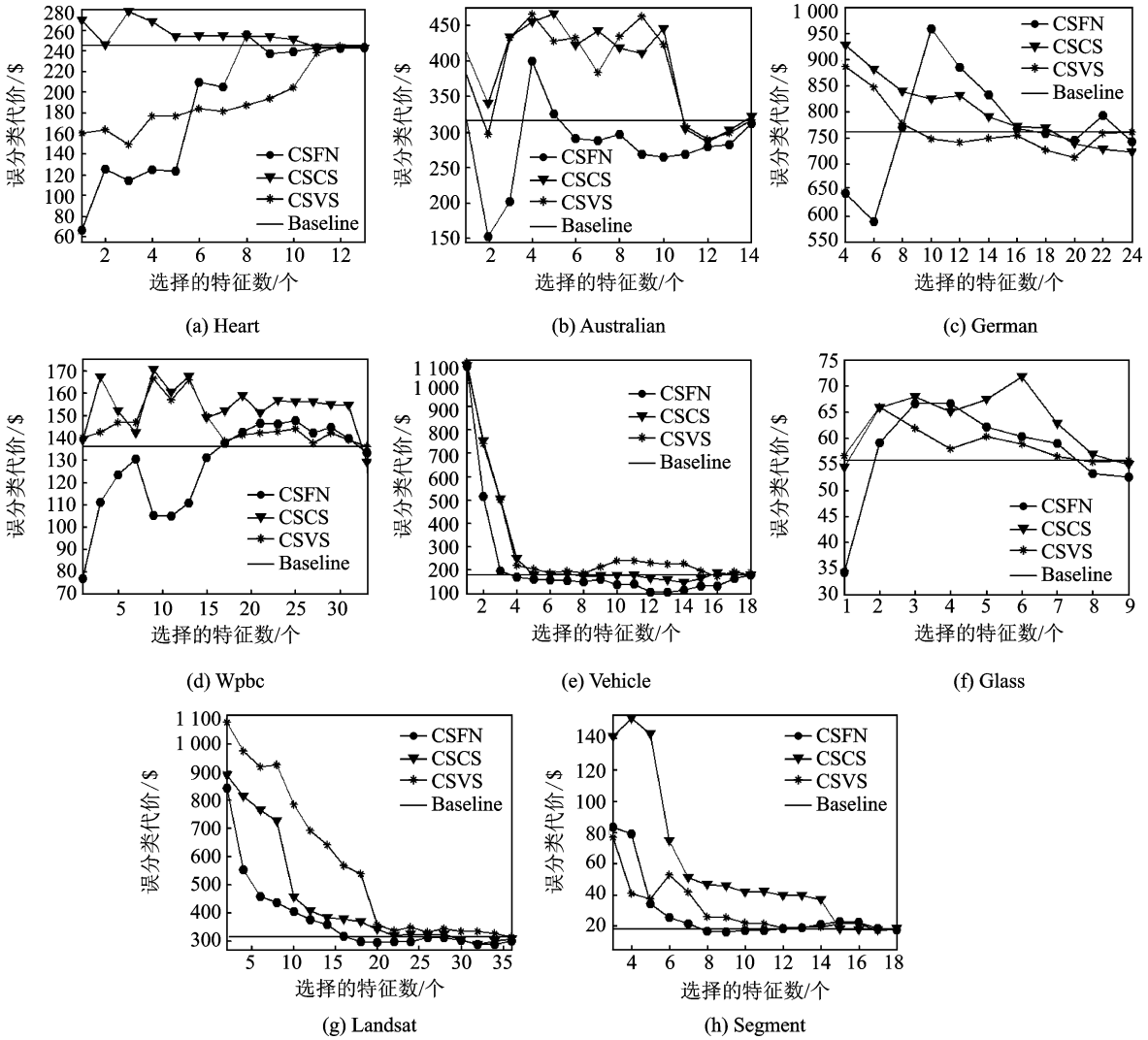


图 1 4 种算法在 8 个数据集上的误分差代价对比

Fig. 1 Misclassification cost contrast for four algorithoms on eight datasets

CSCS 和 CSVS 算法在达到最优性能时所需的特征个数。结合表 3 和表 4,可以得出误分类代价和分类精度有些关联,但并不是误分类代价越小,分类精度就会越高。说明本文的算法是使误分类代价尽可能小的情况下而得出尽可能高的分类精度,也说明了误分类代价是更重要的评价指标。从表 3 和图 1 可以得出,CSFN-SNP 与对比算法能较快地选出误分类代价最低时的特征数。

## 4 结束语

在现实世界中误分类代价一般是不相同的。本文在考虑到误分类代价的同时引入邻域,并且通过邻域保持局部几何结构的性质,提出了基于样本邻域保持的代价敏感特征选择算法 CSFN-SNP,引入的邻域矩阵在降维时既能用来保持原始数据的局部几何结构,又能充分利用其类别信息。为了验证该算法的有效性,将其与已有的代价敏感特征选择算法与 CSCS,CSVS 和 Baseline 作对比,实验结果表明 CSFN-SNP 具有更优的性能。在下一步的工作中,我们将同时考虑测试代价和误分类代价,以建立更好的模型来应

用于实际生活中。

**参考文献：**


[1] Liu H, Motoda H. Feature selection for knowledge discovery and data mining[M]. Netherlands: Springer Science & Business Media, 2012.

[2] Duda R O, Hart P E, Stork D G. Pattern classification[M]. New Jersey: John Wiley & Sons, 2012: 117-120.

[3] Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of Relief and ReliefF[J]. Machine Learning, 2003, 53(1/2): 23-69.

[4] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24(3): 301-312.

[5] Saitta L, Geibel P. Machine learning: A technological roadmap[M]. Netherlands: Department of Social Science Informatics, University of Amsterdam, 2001.

[6] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[7] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: One-sided selection[C]// Proceedings of the Fourteenth International Conference on Machine Learning. Nashville, Tennessee, USA: Morgan Kaufmann, 1997: 253-259.

[8] Ting K M. An empirical study of MetaCost using boosting algorithms[J]. Lecture Notes in Computer Science, 2000, 1810: 413-425.

[9] Zhang Y, Zhou Z H. Cost-sensitive face recognition[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(10): 1758-1769.

[10] Zhou Zhihua, Liu Xuying. On multi-class cost-sensitive learning[J]. Computational Intelligence, 2010, 26(3): 232-257.

[11] Liu Xuying, Zhou Zhihua. Learning with cost intervals[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington: ACM, 2010: 403-412.

[12] Zadrozny B, Elkan C. Learning and making decisions when costs and probabilities are both unknown[C]// Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. America San Francisco: ACM, 2001: 204-213.

[13] Domingos P. MetaCost: A general method for making classifiers cost-sensitive[C]// Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: ACM, 1999: 155-164.

[14] Wang Fei, Zhang Changshui. Label propagation through linear neighborhoods[J]. Knowledge and Data Engineering, IEEE Transactions on, 2008, 20(1): 55-67.

[15] 徐久成，李涛，孙林，等. 基于信噪比与邻域粗糙集的特征基因选择方法[J]. 数据采集与处理，2015，30(5)：973-981.
Xu Jiucheng, Li Tao, Sun Lin, et al. Feature gene selection based on SNR and neighborhood rough set[J]. Journal of Data Acquisition and Processing, 2015, 30(5): 973-981.

[16] 叶鑫晶，李洁，王颖，等. 基于边缘邻域的乳腺肿块特征提取算法[J]. 数据采集与处理，2015，30(5)：993-1002.
Ye Xinjing, Li Jie, Wang Ying, et al. Mammographic mass feature extraction algorithm based on edge of neighborhood[J]. Journal of Data Acquisition and Processing, 2015, 30(5): 993-1002.

[17] Zhao Hong, Min Fan, Zhu William. Test-cost-sensitive attribute reduction based on neighborhood rough set[C]// Granular Computing (GrC), 2011 IEEE International Conference on. Tsukuba, Japan: IEEE, 2011: 802-806.

[18] Yang Yi, Xu Dong, Nie Feiping, et al. Image clustering using local discriminant models and global integration[J]. Image Processing, IEEE Transactions on, 2010, 19(10): 2761-2773.

[19] Miao Linsong, Liu Mingxia, Zhang Daoqiang. Cost-sensitive feature selection with application in software defect prediction[C]// Pattern Recognition (ICPR), 2012 21st International Conference on. Kaohsiung, China: IEEE, 2012: 967-970.


**作者简介：**


**余胜龙**(1992-)，男，硕士研究生，研究方向：机器学习、数据挖掘，E-mail：Fishslyu@163.com。

**赵红**(1979-)，女，副教授，研究方向：粒计算、代价敏感学习、分层分类学习，E-mail：Hongzhaocn@163.com。


（编辑：王静）