一种基于成对约束的特征选择改进算法

杨 杨[1],刘会东[2]

(1. 南京师范大学 强化培养学院,江苏 南京 210046;2. 南京师范大学 计算机科学与技术学院,江苏 南京 210046)

[摘要] 基于成对约束的特征选择算法通过度量单个特征的重要性得到一个特征序列,但由单个重要特征构成的特征子集未必是最有效的. 为此,提出了一种基于成对约束的特征选择改进算法,该算法采用对特征子集进行度量的策略,逐步选择使新的特征子集最有效的特征,从而得到一个有效的特征序列. 实验表明新提出的算法是有效可行的.

[关键词] 机器学习,特征选择,成对约束,分类

[中图分类号] TP311 [文献标识码] A [文章编号]1672-1292(2011)01-0056-06

An Improved Algorithm for Feature Selection Based on Pairwise Constraint

Yang Yang[1], Liu Huidong[2]

(1. Intensification Culture School, Nanjing Normal University, Nanjing 210046, China;
2. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210046, China)

Abstract: Feature selection is key issue in machine learning field. As compared with unsupervised feature selection methods, supervised feature selection approaches have more better performances. However, most of the existing supervised feature selection algorithms mainly aim at the cases using the labels as supervised information, here these methods are not applied to the cases with pairwise constraints. In the real application, it is more easier to get the pairwise constraints as comparing with getting labels. So the researchers proposed a feature selection based on pairwise constraint, the algorithm obtains a feature sequence by measuring the significance of each single feature, but in fact the feature subset combining by those more important features may be not an effective feature subset. Therefore, in this paper, we introduce an improved feature selection algorithm based on pairwise constraint, the newly developed algorithm focuses on evaluating the importance of a feature subset but not a single feature, that is, it uses the empty feature subset as starting point, and then gradually extends this feature subset by adding a most effective feature in every round, in this way an effective ranking feature list is obtained. Experimental results show that the newly proposed algorithm is flexible.

Key words: machine learning, feature selection, pairwise constraint, classification

特征选择是机器学习、数据挖掘及模式识别等领域的主要研究内容之一[1],其研究受到众多研究者的关注[2-9]. 特征选择方法大致可分为 Filter[2] 和 Wrapper[4] 两种模型. Filter 模型通过某个适应函数(fitness function)的值来估计某个特征子集的有效性,与具体的分类器无关. Wrapper 模型是用某个特定分类器的性能作为特征子集选择的准则,这种直接优化分类器的策略可改进分类器的泛化性,但计算代价相对高,且不具有通用性.

从是否使用监督信息角度看,特征选择大体上可分为无监督特征选择[5,6]、监督特征选择[7,8]和半监督特征选择[9]等方法. 与无监督特征选择方法相比,监督特征选择算法具有更好的性能. 但大多数监督特征选择算法中使用的监督信息是类标号,很少考虑成对约束等其他监督信息. 本文的主要目标即为如何设计基于成对约束的特征选择改进算法.

实际应用中,相对于类标记监督信息,成对约束监督信息更容易得到[10]. 文献[11]提出了一种基于成

收稿日期: 2010-10-18.
基金项目: 南京师范大学 2010 年学生科学基金.
通讯联系人: 刘会东,硕士,研究方向:稀疏化和多任务学习. E-mail: huidong_liu@163.com

对约束的特征选择模型—Constraint Score,并由该模型诱导出两种算法(Constraint Score-1, Constraint Score-2) ,这两种算法均通过度量单个特征的重要性得到一个特征序列,但事实上由单个重要特征构成的特征子集未必是最有效的.

为克服文献[11]仅采用对单个特征进行度量来确定其重要性并得到特征序列的不足,本文提出了一种基于成对约束的特征选择改进算法,该算法采用对特征子集进行度量的策略,逐步选择使新的特征子集最有效的特征,从而得到一个有效的特征序列. 该算法从一个空集开始,首先得到一个最重要的特征并加入到该集合中,然后逐个加入特征到该集合并进行度量以确定加入的特征是否有效,直到得到一个有效的特征序列. 类似于文献[11],本文也采用 Filter 模型.

1　基于成对约束的特征选择算法—Constraint Score

相对于无监督特征选择算法[5,6],监督特征选择算法[4,7,8]具有更好的性能. 但这些监督特征选择算法仅适用于以类标记为监督信息的情况,不适用于其他监督类型的情况. 为此,文献[11]提出了一种基于成对约束的特征选择算法—Constraint Score.

对给定的样本集 $X=\{x_1,x_2,\cdots x_m\}$,成对约束分 Must-link 约束和 Cannot-link 约束. Must-link 约束集定义为 $M=\{(x_i,x_j)\mid x_i$ 与 x_j 同类$\}$; Cannot-link 约束集定义为 $C=\{(x_i,x_j)\mid x_i$ 与 x_j 异类$\}$. Constraint Score 模型中对第 r 个特征的评价函数为:

$$C_r^1=\frac{\sum\limits_{(x_i,x_j)\in M}(f_{ri}-f_{rj})^2}{\sum\limits_{(x_i,x_j)\in C}(f_{ri}-f_{rj})^2}, \tag{1}$$

$$C_r^2=\sum_{(x_i,x_j)\in M}(f_{ri}-f_{rj})^2-\lambda\sum_{(x_i,x_j)\in C}(f_{ri}-f_{rj})^2. \tag{2}$$

式中, f_{ri} 为第 i 个样本的第 r 个特征.

Constraint Score 算法通过选择那些使评价函数值较小的特征,即使同类样本分布更加紧凑,而不同类样本分布较分散的特征. Constraint Score 算法的描述过程如下:

算法 1　Constraint Score;

输入: 样本集 X,成对约束集 M 和 C,λ(对 Constraint Score-2) ;

输出: 特征序列.

步骤:

1. 对于样本集中每一个特征, 利用式(1) 或式(2) 计算 Constraint Score 值;

2. 根据 Constraint Score 值升序排列得到一个特征序列.

从 Constraint Score 算法可以看出,Constraint Score-1(简记为 CS1) 和 Constraint Score-2(简记为 CS2) 均通过度量一个特征来确定特征序列,以此通过组合那些重要的特征构成一个特征子集,但某些情况下单个重要特征构成的特征子集未必是有效的.

2　基于成对约束的特征选择改进算法—Improved Constraint-Score

为克服 Constraint Score 模型的不足,本文采用对特征子集进行类似 Constraint Score 模型的度量. 对给定的样本集 $X=\{x_1,x_2,\cdots x_m\}$,其 $x_i=(f_{1i},f_{2i},\cdots,f_{ni})$. 对任意的特征子集(其索引为 A) ,基于成对约束的特征选择改进模型(Improved Constraint Score) 对特征子集的评价函数为:

$$IC_A^1=\frac{\sum\limits_{(x_i,x_j)\in M}\|f_{Ai}-f_{Aj}\|_2^2}{\sum\limits_{(x_i,x_j)\in C}\|f_{Ai}-f_{Aj}\|_2^2}, \tag{3}$$

$$IC_A^2=\sum_{(x_i,x_j)\in M}\|f_{Ai}-f_{Aj}\|_2^2-\lambda\sum_{(x_i,x_j)\in C}\|f_{Ai}-f_{Aj}\|_2^2. \tag{4}$$

式中, f_{Ai} 为特征子集 A 在第 i 个样本上的投影.

$$S_{ij}^{M} = \begin{cases} 1 & \text{if } (x_i, x_j) \in M \text{ or } (x_j, x_i) \in M, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$$S_{ij}^{C} = \begin{cases} 1 & \text{if } (x_i, x_j) \in C \text{ or } (x_j, x_i) \in C, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

假设 $\boldsymbol{F}_r = (f_{r1}, f_{r2}, \cdots, f_{rm})^{\mathrm{T}}$,其中 f_{ri} 表示第 i 个样本的第 r 个特征. 另设 $\boldsymbol{D}^M$ 和 $\boldsymbol{D}^C$ 分别为对角矩阵,其中对角线元素分别为 $D_{ii}^M = \sum_j S_{ij}^M$ 和 $D_{ii}^C = \sum_j S_{ij}^C$. 令 $\boldsymbol{L}^M = \boldsymbol{D}^M - \boldsymbol{S}^M$, $\boldsymbol{L}^C = \boldsymbol{D}^C - \boldsymbol{S}^C$. 由式(5)和(6),可得:

$$\sum_{(x_i,x_j)\in M} \| \boldsymbol{f}_{Ai} - \boldsymbol{f}_{Aj} \|_2^2 = \sum_{i,j} \| \boldsymbol{f}_{Ai} - \boldsymbol{f}_{Aj} \|_2^2 S_{ij}^M = \sum_{i,j} (\boldsymbol{f}_{Ai}\boldsymbol{f}_{Ai}^{\mathrm{T}} + \boldsymbol{f}_{Aj}\boldsymbol{f}_{Aj}^{\mathrm{T}} - 2\boldsymbol{f}_{Ai}\boldsymbol{f}_{Aj}^{\mathrm{T}}) S_{ij}^M =$$

$$\sum_{i,j,r\in A} (f_{ri}^2 + f_{rj}^2 - 2f_{ri}f_{rj}) S_{ij}^M = 2\sum_{r\in A} \boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^M \boldsymbol{F}_r.$$

类似可得, $\sum_{(x_i,x_j)\in C} \| \boldsymbol{f}_{Ai} - \boldsymbol{f}_{Aj} \|_2^2 = 2\sum_{r\in A} \boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^C \boldsymbol{F}_r$. 于是,式(3)和(4)可定义如下:

$$IC_A^1 = \frac{\sum_{r\in A} \boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^M \boldsymbol{F}_r}{\sum_{r\in A} \boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^C \boldsymbol{F}_r}, \tag{7}$$

$$IC_A^2 = \sum_{r\in A} \boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^M \boldsymbol{F}_r - \lambda \sum_{r\in A} \boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^C \boldsymbol{F}_r. \tag{8}$$

而由式(1)和(2)可得[11]:

$$C_r^1 = \frac{\boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^M \boldsymbol{F}_r}{\boldsymbol{F}_r^{\mathrm{T}} \boldsymbol{L}^C \boldsymbol{F}}, \tag{9}$$

$$C_r^2 = \boldsymbol{F}_r{}^{\mathrm{T}} \boldsymbol{L}^M \boldsymbol{F} - \lambda \boldsymbol{F}_r{}^{\mathrm{T}} \boldsymbol{L}^C \boldsymbol{F}_r. \tag{10}$$

依据式(7)和(8),Improved Constraint Score 算法的主要思路为:(1)让 A 为空集, $F = \{1,2,\cdots,n\}$;(2)计算 $r = \min_{r\in F-A} IC_{f_{\{r\}\cup A}}^1$,令 $A = \{r\} \cup A$, $F = F - \{r\}$;(3)反复重复第2步直到 F 为空,进而得到一个特征序列. 依据上述思路,Improved Constraint Score 算法的描述过程如下:

算法2 Improved Constraint Score;

输入:样本集 X,成对约束集 M 和 C, λ(对 Improved Constraint Score-2);

输出:特征序列.

步骤:

1. 设 $A = \{\}$, $F = \{1,2,\cdots,n\}$;
2. 利用式(7)或(8),计算 $r = \min_{r\in F-A} IC_{f_{\{r\}\cup A}}^1$,令 $A = \{r\} \cup A$, $F = F - \{r\}$;
3. 重复步骤2直到 F 为空,并得到一个特征序列.

类似于 Constraint Score,由 Improved Constraint Score 算法诱导出两个算法 Improved Constraint Score-1(简记为ICS1)和 Improved Constraint Score-2(ICS2),其中 ICS1 和 ICS2 分别由式(7)和(8)得到.

定理1 对 Constraint Score 和 Improved Constraint Score 算法,由式(2)和(4)诱导出的 CS2 和 ICS2 算法得到同样的特征序列.

证明 不妨设由 CS2 得到的特征序列为 $f_{i_1}, f_{i_2}, \cdots, f_{i_n}$,其中 $i_1, i_2, \cdots, i_n$ 为 $1,2,\cdots,n$ 的一个排列. 由式(10)和 CS2 算法知, $C_{f_{i_1}}^1 \leqslant C_{f_{i_2}}^1 \leqslant \cdots \leqslant C_{f_{i_n}}^1$. 由 $r = \min_{r\in F-A} IC_{f_{\{r\}\cup A}}^1$ 和式(8)知,依次扩展的特征分别为 f_{i_1}, $f_{i_2}, \cdots, f_{i_n}$,故定理1成立. 证毕.

由定理1可见,Constraint Score 和 Improved Constraint Score 算法对应的 CS2 和 ICS2 得到相同的特征序列. 这进一步表明 Constraint Score-2 本质上相当于通过特征子集的度量来确定特征序列的,这与文献[11]指出的 Constraint Score-2 优于 Constraint Score-1 是一致的. 由于 Constraint Score-2 和 Improved Constraint Score-2 得到相同的特征序列,为节约篇幅,本文实验主要对 Constraint Score-1(CS1)和 Improved Constraint Score-1(ICS1)进行比较.

3　实验结果

3.1　数据集描述

为进一步验证算法的性能,本文采用网上(http://www.ics.uci.edu) 提供的 UCI 数据集,共有 6 个数据集,各数据集的描述见表 1.

表 1　实验中所采用的数据集描述

Table 1　Description of data sets in experiments

序号	数据集	样本数	类别数	特征数
1	Ionosphere	351	2	34
2	Liverdata	345	2	6
3	Sonar	208	2	60
4	Soybean	47	4	35
5	Vehicle	846	4	18

3.2　实验分析

为方便计,将 Constraint Score-1 和 Improved Constraint Score-1 分别简记为 CS1 和 ICS1. 为测试 ICS1 和 CS1 算法的性能,用表 1 中的数据集来测试由特征子集诱导出的分类器精度. 在实验中,随机选择 60% 作为训练集,其余 40% 作为测试集,重复 10 次实验,并以 10 次的平均测试精度作为分类器精度. 为有效验证两种算法得到的特征序列的有效性,分别用 3NN 和 C4.5(简记为 C45) 两个不同的分类器来评价得到特征子集的有效性. 这里分类器 3NN 和 C4.5 采用 Weka(Version 3-5) 软件[12] 的缺省参数.

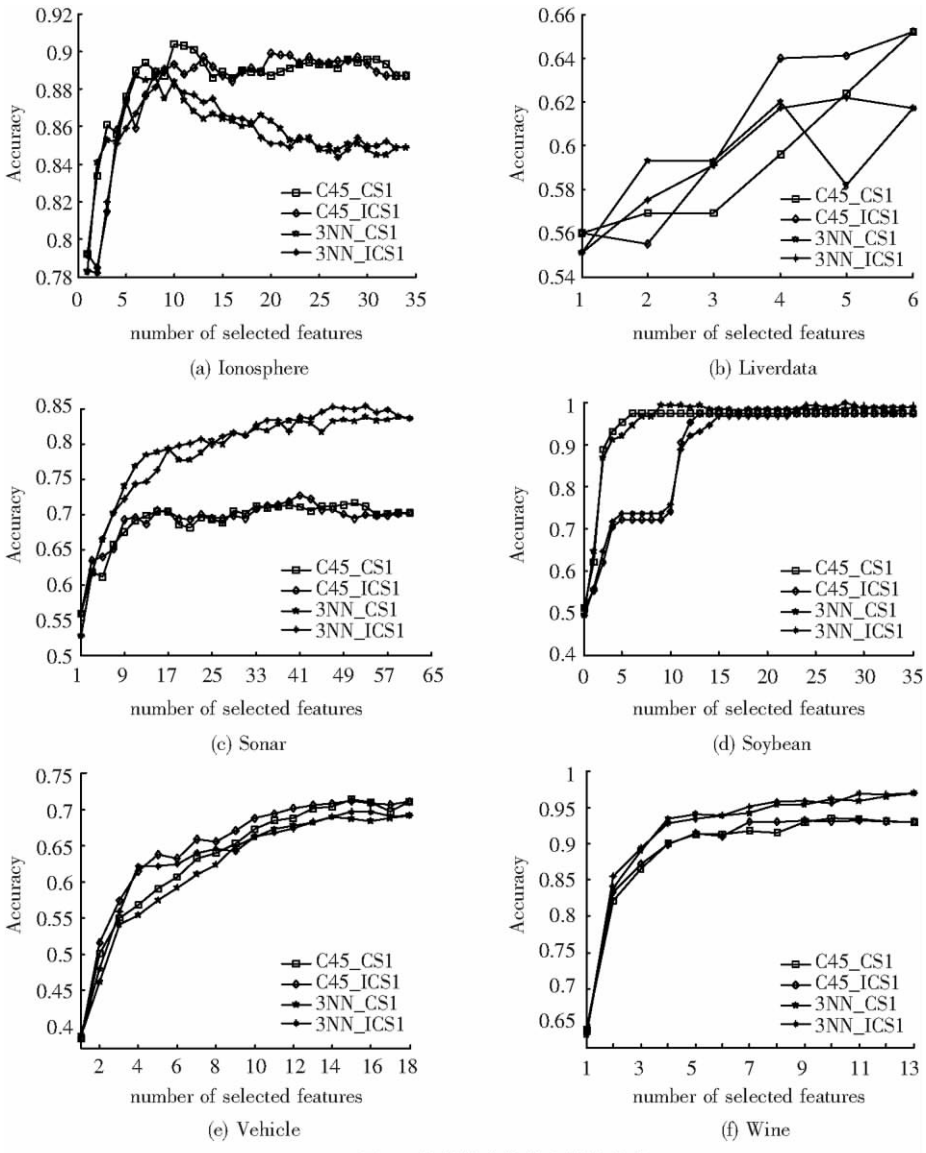

图 1　分类精度随特征数的变化

Fig.1　Accuracy vs. number of features

为有效测试 CS1 和 ICS1 的性能,首先进行第 1 组实验,即取 Must-link 和 Cannot-link 约束数均为 5,并对全部 6 个数据集用两个不同的分类器 3NN 和 C45 对由 CS1 和 ICS1 得到的特征序列进行测试,侧重测试分类精度随特征数的变化情况,实验结果如图 1 所示. 进一步,为测试不同成对约束数目对特征序列有效性的影响,取 Must-link 和 Cannot-link 约束数均为 10 和 20 两种情况,并对 Ionosphere 数据集用 C45 进行实验,以测试分类精度随成对约束数的变化情况,实验结果如图 2 所示.

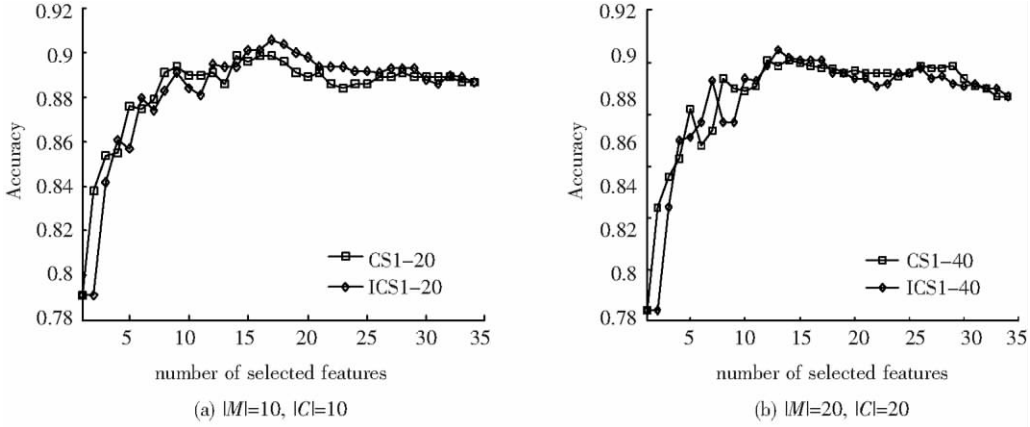

图 2　不同成对约束数下分类器性能比较(Ionosphere)

Fig.2　Comparison of classification performance under different pairwise constraints(Ionosphere)

从图 1 可以看出,CS1 和 ICS1 都是非常有效的特征选择算法. 但与 CS1 相比,ICS1 具有更好的性能,如: 对 Vehicle 数据集,ICS1 明显优于 CS1. 对 Wine 数据集,当特征数取 8 左右时,ICS1 优于 CS1,且此时分类器基本上处于最优. 对 Sonar 数据集,在 3NN 上 ICS1 的性能优于 CS1; 对 C45,当特征数取 41 左右时,ICS1 的性能优于 CS1. 对 Soybean 数据集,CS1 优于 ICS1 的性能,即 CS1 取较小的特征子集即可得到较高的分类精度. 对数据 Ionosphere, 在 3NN 上 ICS1 的性能优于 CS1; 在 C45 上,当特征数取 10 左右时,CS1 的性能优于 ICS1,而当特征数取 21 左右时,ICS1 的性能优于 CS1. 对 Liverdata 数据集,当特征数为 4 和 5 时,ICS1 的性能优于 CS1. 总体上看, ICS1 在 4 个数据集上的性能优于或略优于 CS1,且在另外 2 个数据集上的性能与 CS1 是可以比较的.

由图 2 可以看出,ICS1 的性能优于 CS1, 如: 对图 2(a) ,当特征数在 [15,25] 中取时,ICS1 明显优于 CS1,且此时分类器基本上达到最优. 此外,还可看出,分类器的性能并不随着约束数的增加而增加,即当约束数取得一定数后,得到的特征序列的有效性反而有所下降,这与文献 [11] 的结论是一致的.

综上分析,通过对特征子集的度量,ICS1 可有效改进特征序列的有效性,进而改进分类器的性能.

4　结语

本文提出了一种基于成对约束的特征选择改进算法——Improved Constraint Score,该算法采用对特征子集进行度量的策略,逐步选择使新的特征子集最有效的特征,进而得到一个有效的特征序列. 通过在 6 个 UCI 数据集上的实验表明,新提出的算法 Improved Constraint Score 是基于成对约束的特征选择算法 Constraint Score 的改进.

[参考文献](References)

[1] Liu H, Motoda H. Feature selection for knowledge discovery and data mining [M]. Boston: Kluwer,1998.

[2] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution [C] // Proceedings of the 20th International Conferences on Machine Learning. Washington DC, 2003: 856-863.

[3] Kohavi R, John G. Wrappers for feature subset selection [J]. Artificial Intelligence, 1997,19(1/2) : 273-324.

[4] 毛勇,周晓波,夏铮,等. 特征选择算法研究综述 [J]. 模式识别与人工智能,2007,20(2) : 211-218.
Mao Yong, Zhou Xiaobo, Xia Zheng, et al. A survey for study of feature selection algorithms [J]. Pattern Recognition & Artificial Intelligence, 2007,20(2) : 211-218. (in Chinese)

[5] 朱颢东,李红婵,钟勇. 新颖的无监督特征选择方法[J]. 电子科技大学学报,2010,39(3):412-415.
Zhu Haodong, Li Hongchan, Zhong Yong. New unsupervised feature selection method[J]. Journal of University of Electronic Science and Technology of China,2010,39(3):412-415. (in Chinese)

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

[7] Bishop C M. Neural Networks for Pattern Recognition[M]. Oxford: Oxford University Press, 1995.

[8] 王博,黄九鸣,贾焰,等. 适用于多种监督模型的特征选择方法研究[J]. 计算机研究与发展,2010,47(9):1 548-1 557.
Wang Bo, Huang Jiuming, Jia Yan, et al. Research on a common feature selection method for multiple supervised models[J]. Journal of Computer Research and Development, 2010,47(9):1 548-1 557. (in Chinese)

[9] Zhao Z, Liu H. Semi-supervised feature selection via spectral analysis[C]// Proceedings of the 7th SIAM International Conference on Data Mining. Minneapolis: MN,2007: 641-646.

[10] Xing E P, Ng A Y, Jordan M I, et al. Distance metric learning, with application to clustering with side-information[C]// Proceedings of the Conference on Advances in Neural Information Processing Systems(NIPS). 2002: 505-512.

[11] Zhang D, Chen S, Zhou Z H. Constraint Score: A new filter method for feature selection with pairwise constraints[J]. Pattern Recognition, 2008,41(5):1 440-1 451.

[12] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques[M]. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[责任编辑: 严海琳]