一种基于混合差别矩阵的属性约简算法及其在入侵检测中的应用

邱玉祥, 杨 明

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] 针对入侵检测问题, 提出了构造混合辨别矩阵的方法, 并用 C4.5分类器测试选择子集的有效性. 实验表明分类器在新算法得到的特征子集上有较好的分类效果.

[关键词] 入侵检测, 粗糙集, 混合辨别矩阵, 属性约简
[中图分类号] TP311 [文献标识码] A [文章编号] 1672-1292(2008)03-0071-06

An Algorithm for Attribute Reduction in Rough Set and its Application in Intrusion Detection

Qiu Yuxiang, Yang Ming

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

Abstract A novel rough set-based method followed by establishing a mix discernibility matrix is introduced for intrusion detection, and choose C4.5 algorithm for testing the effectiveness of selected attribute subsets. Experimental results show that the classifiers developed using the selected attribute subsets have better performance than those generated by all attributes.

Key words intrusion detection, rough set, mix discernibility matrix, attribute reduction

入侵检测[1]发展至今已有 20多年的历史. 从机器学习角度看, 入侵检测实际上是一个分类问题. 在基于 kdd cup 99数据集的研究中, 入侵特征选择和分类算法研究是两大主要研究内容. 特征选择是根据某种性能判据, 从所有输入特征中选择重要特征, 去掉次要特征, 这往往有利于缩短检测时间、发现某类攻击的本质特征. 有监督的特征选择方法可被分为两类: filter模型和 wrapper模型[2,3]. filter模型评估依赖于数据集本身, 通常是选择和目标函数相关度大的特征或者特征子集, 一般认为相关度较大的特征或者特征子集会对应得到后续学习算法较高的准确率, 其评估方法主要有类间距离、信息增益以及不一致度等. filter型的特征选择因为通常运行效率较高而适用于大规模数据集, Relief算法[4]是其中公认效果比较好的一种算法.

粗糙集(Rough Set)理论[5]是 20世纪 80年代由波兰教授 Pawlak Z提出的, 目前已受到国内外学者的广泛关注. 在粗糙集理论中, 属性约简是重要研究内容之一, 也是知识获取的关键步骤, 因此属性约简研究备受粗集研究者的关注, 也取得了很大的进展[6-12]. 现有的属性约简大体上可分为基于差别矩阵或在此基础上的改进的属性约简算法[7-9]、基于正域的属性约简算法[10,11]及基于启发式的属性约简算法[12,13], 但这些算法只能够处理离散的数据, 不能直接处理 kdd cup 99这样的既有离散属性又有连续属性的数据. 因此使用基于粗集的特征选择算法, 需要先对连续属性离散化. 本文主要基于构造混合差别矩阵的方法, 无需离散化, 直接对 kdd cup 99进行属性约简.

1 Relief算法及其变种

Relief评估[4]最早由 Kira K提出, 其核心思想为: 好的特征应该使同类的样本接近, 而使不同类的样

收稿日期: 2008-01-07.
基金项目: 国家自然科学基金(40771163)资助项目.
通讯联系人: 杨 明, 教授, 博士, 研究方向: 数据挖掘、机器学习和粗糙集理论及应用研究. E-mail: myang@njnu.edu.cn

本之间远离. 算法从训练集 D 中随机选择一样 R, 然后从和其同类的样本中寻找最近邻样本 H, 称为 NearestHit; 从和其不同类的样本中寻找最近邻样本 M, 称为 NearestMiss. 然后对于每维特征, 如果 R 和 H 在其上的距离小于 R 和 M 上的距离, 则说明此维特征对区分同类和不同类的最近邻是有益的, 则增加该特征权值; 反之, 则说明此特征对区分同类和不同类的最近邻起反作用, 则降低该特征权值; 重复以上过程 m 次, 最后得到各特征的平均权值. 特征的权值越大, 表示该特征的分类能力越强, 反之, 表示该权值分类能力越弱.

Relief算法提出时仅限于处理类别数为两类的数据的分类问题, 后来 Kononenko[14]扩展了 Relief算法得到了 ReliefF算法. ReliefF可以解决多类问题以及回归问题, 在处理多类问题时, 不是从所有不同类样本集合中统一选择最近邻样本, 而是从每个不同类别的样本集合中选择 K 个最近邻样本.

2　传统的辨别矩阵的属性约简

一个信息系统或决策表 DT 是指由 U, Q, V, f 构成的一个四元组, $DT = < U, Q, V, f >$. 其中, U 为论域; $Q = C \cup D$ 是 U 上的一组属性集合, 子集 C 和 D 分别称为条件属性集和决策属性集, 并且 $C \cap D = \{\}$, 为方便记, 记 $D = \{d\}$ 为单个决策属性; $V = \bigcup_{a \in Q} V_a$, V 是属性 a 的值域, f 是 $U \times Q \rightarrow V$ 的映射, 它为每个对象的每个属性赋予一个信息值, 即 $a \in Q$, $x \in U$, $f(x, a) \in V_a$.

设 U/Q 表示由属性集 Q 形成的等价类的集合, 称 U/Q 为不可识别关系或不可辨关系. 若 $U/P = U/Q$, $P \subseteq Q$, 且对任意 $a \in P$ 有 $U/(P - \{a\}) \neq U/Q$, 则称 P 为属性集的 Q 的属性约简. Q 的约简的全体记为 $\mathrm{red}(Q)$. 若 $U/(Q - \{r\}) \neq U/Q$, $r \in Q$, 则称 r 为属性集的不可缺少属性, 这些不可缺少的属性的集合称为属性集的核.

可辨别矩阵 $\boldsymbol{M}(T)$[15]是一个 $|U| \times |U|$ 的对称矩阵, 矩阵中的每一个元素 m_{ij} 定义如下:

$$m_{ij} = \begin{cases} a \in C: f(x_i, a) \neq f(x_j, a) & \text{当 } f(x_i, d) \neq f(x_j, d); \\ \approx & \text{其它.} \end{cases} \quad (1)$$

根据可辨矩阵的定义可知: 当两个决策规则的决策属性取值相同时, 它们所对应的可辨矩阵中的元素的取值为 $\approx$; 当两个决策规则的决策属性值不同, 且可以通过某些条件属性的不同取值加以区分时, 它们所对应的可辩矩阵元素的取值为两个决策规则属性值不同的条件属性集合, 即可以区分这两个样本的条件属性集合.

由属性约简的定义可得到相应的基于辨别矩阵的求属性约简的等价命题: R 是 C 的一个属性约简, 当且仅当 $R \cap m_{ij} \neq \{\}$, 其中 $m_{ij} \in DM(m_{ij} \neq \{\})$, 并且对任意的 $S \subset R$, $\exists m_{ij} \in DM(m_{ij} \neq \{\})$ s.t. $S \cap m_{ij} = \{\}$. 根据辨别矩阵, Johnson提出了启发式的属性约简算法[16].

3　基于混合特征属性的属性约简

在可辨别矩阵的基础上前人提出了很多有效的属性约简的算法[8,9], 但是这些算法只适合那些决策表中的元素为离散值的情况. 文献[17]提出了基于连续属性的决策表的可辨别矩阵的构造算法. 该算法只针对那些决策表中的元素为连续属性的情况. 基本思想如下: 决策表 $DT = < U, Q, V, f >$, $Q = C \cup D$, $C = \{a_1, a_2, \cdots, a_m\}(V_{aj} \subset R, 1 \leqslant j \leqslant m)$, $D = \{d\}(V_d = \{d_1, d_2, \cdots d_s\})$, $U = \{x_1, x_2, \cdots x_n\}$, $x_i = \{x_{i1}, x_{i2}, \cdots, x_{im}, y_i\}$, $x_{ij} \in V_{aj}$, $y_i \in V_d$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$.

文献[17]按下列步骤构造辨别矩阵, 而不需要离散化决策表. 首先, 对决策表进行标准化, 把属性 a 的一个值 v 映射 v':

$$v' = (v - \min_a) \times (\mathrm{new\_max}_a - \mathrm{new\_min}_a)/(\max_a - \min_a) + \mathrm{new\_min}_a. \quad (2)$$

该算法中 $\mathrm{new\_min}_a$, $\mathrm{new\_max}_a$ 分别为 0和 1. 具有不同决策属性的两个对象 x_i, x_j 在给定的集合 C 上的距离定义为 $d_C(x_i, x_j) = \sum_{a_k \in C} |x_{ik} - x_{jk}|$, 即 $d_C(x_i, x_j) = \left(\sum_{a_k \in C} |f(x_i, a_k) - f(x_j, a_k)| \right)$. 其中, $1 \leqslant i, j \leqslant m$. $d_C(x_i, x_j)$ 代表了 x_i, x_j 在给定的属性集 C 上的相异性, 距离越大越相异. 并且设定一个阈值 δ, 当 $d_C(x_i, x_j) < \delta$ 时, 认为 x_i, x_j 是不一致的, 否则认为是一致的. 为了得到 C 中相对重要的属性, 需要把那些对 $d_C(x_i, x_j)$ 有贡献的属性进行排序, $|f(x_i, a_k) - f(x_j, a_k)|$ 或 $|x_{ik} - x_{jk}|$ 的值越大, 属性 a_k 越重要, 这里 a_k

$\in C$. 为了有效的控制相对重要属性子集的基数, 提出如下简明有效的策略:

$$\min \operatorname{card}(B)$$
$$\text{s.t}\ \frac{d_B(x_i, x_j)}{d_C(x_i, x_j)} \geqslant \varepsilon,\ B \subseteq C. \tag{3}$$

其中 $1 \leqslant i, j \leqslant n$, ε是一个阈值, 是一个可以调整的非负参数. 解决上面的优化问题, 就能够得到区分 x_i, x_j 的那些重要的属性. 为了简化, 把式 (3) 写成 $g_\varepsilon(x_i, x_j)$, 即 $g_\varepsilon(x_i, x_j) = B$, 其中 B 满足式 (3).

对于给定的阈值 ε, 假设 $|f(x_i, a_k) - f(x_j, a_k)| \geqslant |f(x_i, a_{k+1}) - f(x_j, a_{k+1})|$, 其中 $1 \leqslant k \leqslant m-1$, $g_\varepsilon(x_i, x_j) = \{a_1, a_2, \cdots a_t\}$ 可以由下面等价的公式得到:

$$\frac{\sum_{k=1}^{t} |x_{ik} - x_{jk}|}{\sum_{k=1}^{m} |x_{ik} - x_{jk}|} \geqslant \varepsilon \wedge \frac{\sum_{k=1}^{t-1} |x_{ik} - x_{jk}|}{\sum_{k=1}^{m} |x_{ik} - x_{jk}|} < \varepsilon \tag{4}$$

于是可得到模糊可辨别矩阵 $\boldsymbol{M}(T)$. $\boldsymbol{M}(T)$ 是一个 $|U| \times |U|$ 的对称矩阵, 矩阵中的每一个元素 C_{ij} 定义如下:

$$C_{ij} = \begin{cases} g_\varepsilon(x_i, x_j):\ d_C(x_i, x_j) \geqslant \delta \quad \text{当 } f(x_i, d) \neq f(x_j, d) \text{ 时},\ 1 \leqslant i, j \leqslant n; \\ \{\} \qquad\qquad\qquad\qquad\qquad\qquad \text{其它}. \end{cases} \tag{5}$$

综上所述, 结合文献 [15] 和 [17] 可以处理条件属性中既有离散属性又有连续属性的情况, 即把式 (1) 和式 (5) 有效结合. 决策表 $DT = <U, Q, V, f>$ 如上所述, 其中 $Q = C \cup D$, $C = \{a_1, a_2, \cdots, a_m, a_{m+1}, \cdots, a_r\}$, $C_1 = \{a_1, a_2, \cdots a_m\}$, $C_2 = \{a_{m+1}, a_{m+2}, a_r\}$, $C = C_1 \cup C_2$, $C_1 \cap C_2 = \{\}$, C_1 代表离散的条件属性的集合, C_2 代表连续的条件属性的集合. x_i, x_j 在给定的集合 C_2 上的距离定义为 $d_{C_2}(x_i, x_j) = \sum_{a_k \in C_2} |x_{ik} - x_{jk}|$, 即 $d_{C_2}(x_i, x_j) = \left(\sum_{a_k \in C_2} |f(x_i, a_k) - f(x_j, a_k)| \right)$, 其中 $1 \leqslant i, j \leqslant n$. 对于给定的阈值 ε, 假设 $|f(x_i, a_k) - f(x_j, a_k)| \geqslant |f(x_i, a_{k+1}) - f(x_j, a_{k+1})|$, 其中 $1 \leqslant k \leqslant m-1$, $a_k \in C_2$, $g_\varepsilon(x_i, x_j) = \{a_1, a_2, \cdots a_t\}$ 可由式 (4) 得到, 因此对于既有离散属性又有连续属性的决策表 (称为混合决策表), 辨别矩阵 (称为混合辨别矩阵) 可以定义为:

$$C_{ij} = \begin{cases} \{g_\varepsilon(x_i, x_j):\ d_{C_2}(x_i, x_j) \geqslant \delta\} \cup \{a \in C_1: f(x_i, a) \neq f(x_j, a)\} \quad \text{当 } f(x_i, d) \neq f(x_j, d) \text{ 时}; \\ \{\} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{其它}. \end{cases}$$

由上述分析, 混合辨别矩阵的算法 (Mix discernibility matrix, MDM) 可描述如下:

MDM$(C_1 \cup C_2, D, U)$

输入: C_1: 离散条件属性集, C_2: 连续条件属性集, D: 决策属性, U: 论域, ε, δ为阈值.

输出: 混合辨别矩阵.

步骤:

对决策表中的连续条件属性集进行标准化:

```
for(i = 1 to card(U) - 1){
  for(j = i + 1 to card(U)){
    if f(x_i, d) ≠ f(x_j, d)
      C_ij = {g_ε(x_i, x_j): d_C2(x_i, x_j) ≥ δ} ∪ {a ∈ C_1: f(x_i, a) ≠ f(x_j, a)}
    else
  C_ij = {}}}
```

输出: 混合辨别矩阵.

基于辨别矩阵的属性约简算法一般需要很大的存储空间, 以 Johnson提出的属性约简算法为例, 对 n行、m 列的决策表, 需要的空间复杂度为 $O(m * n^2)$. 文献 [18] 的作者提出了浓缩树 (C-tree) 进行辨别矩阵的压缩, 并且证明了 C-tree能够存储辨别矩阵的所有信息, 实验表明 C-tree能够大大减少程序的空间复杂度. 因此根据浓缩树的思想, 上述的 MDM 算法可以更改为混合辨别树 (Mix discernibility tree, MDT), 描述如下:

MDT($C_1 \cup C_2, D, U$)

输入: C_1: 离散条件属性集, C_2: 连续条件属性集, D: 决策属性, U: 论域, ε, δ为阈值.

输出: 混合辨别树.

步骤:

对决策表中的连续条件属性集进行标准化;

创建 MC-tree树的根结点

for(i = 1 to card(U) - 1){

for(j = i+ 1 to card(U)){

if$f(x_i, d) \neq f(x_j, d)$ {

$C_{temp} = \{g_\varepsilon(x_i, x_j): d_{C_2}(x_i, x_j) \geqslant \delta\} \cup \{a \in C_1: f(x_i, a) \neq f(x_j, a)\}$

把 C_{temp}压缩到 MC-tree上 }}}

输出混合辨别树 MC-tree

用 MC-tree的结果, 替换[18]中的 JreductBtree(C, D, U)中的 C-tree即可得到一个属性约简 $R \subseteq C$, 这样的算法能够直接对既有离散属性又有连续属性的数据之间进行属性约简, 而不必预先进行离散化.

4 实验结果

为有效验证算法在入侵检测中的应用效果, 将本文算法与 ReliefF进行了比较. 实验采用已由哥伦比亚大学完成数据预处理的 kdd cup 1999 data, 该数据集提供了从一个模拟的局域网上采集来的 9个星期的网络连接数据, 数据集中的每条记录包含了 41维特征. 其中第 2、3、4、7、12、21、22维是离散属性, 其它 34维特征是连续属性, 即 $C_1 = \{a_2, a_3, a_4, a_7, a_{12}, a_{21}, a_{22}\}$, $C_2 = C - C_1$. 数据的类别大致分为 5类, 分别为 Normal, Dos, Probe, R2L, U2R. 实验从 kdd cup 99中随机不放回抽取 2 989条作为训练, 另外抽取 2 989条作为测试. 数据的分布如表 1所示.

表 1 数据集

Table 1 Data collection

Kdd cup 99原始数据	训练数据	测试数据
Dos总数据 3 883 370	Dos(800)—train	Dos(800)—test
Normal总数据 972 782	Normal(800)—train	Normal(800)—test
Probe总数据 41 102	Probe(800)—train	Probe(800)—test
R2L总数据 1 126	R2L(563)—train	R2L(563)—test
U2R总数据 52	U2R(26)—train	U2R(26)—test

由于 $d_C(x_i, x_j) < \delta$时, 认为 x_i, x_j是不一致的数据, 又所有数据标准化在[0, 1]之间, 假设 x_i, x_j在每一维上的距离都小于 0.1例如 x_i, x_j在第 k维上的距离 $|f(x_i, a_k) - f(x_j, a_k)| < 0.1$, x_i, x_j在给定的集合 C上的距离定义为 $d_C(x_i, x_j) = \sum_{a_k \in C} |x_{ik} - x_{jk}| < 0.1 \times 41 \approx 4$, 故参数 δ设为 4; ε在 0.1-0.9之间调整. 本实验采用 C4.5算法对所得的属性约简进行测试, 以分类的精度作为测试结果, 结果如表 2所示.

表 2 实验结果

Table 2 Experimetal result

ε值	Normal	Dos	Probe	R2L	U2R	total	属性个数
0.1、0.2	0.963 75	0.986 25	0.982 23	0.998 75	0.500 00	0.978 59	21
0.3、0.4	1.000 00	0.986 25	0.985 79	0.998 75	0.769 23	0.991 30	19
0.5、0.6、0.7	1	0.986 25	0.980 46	0.998 75	0.65384	0.989 29	19
0.8	1	0.986 25	0.980 46	0.998 75	0461 54	0.987 62	21
0.9	1	0.986 25	0.975 13	0.998 75	0.576 92	0.987 29	23

从表 2可知, ε在 0.3和 0.4时的分类精度最高, 得到的属性约简集为: 3, 4, 5, 8, 11, 12, 13, 14, 16, 17, 19, 22, 25, 30, 31, 32, 33, 34, 35.

对属性约简前与属性约简后的数据进行测试, 比较结果如表 3所示.

表 3 属性约简前后比较

Table 3 Comparison before and after attributive reduction

	Normal	Dos	Probe	R2L	U2R	total	属性个数
属性约简前	0.986 25	0.983 75	0.976 91	0.998 75	0.615 38	0.983 94	41
属性约简后	1	0.986 25	0.985 79	0.998 75	0.769 23	0.991 30	19

从表 3可以看出, 分类器在特征子集上有较好的效果.

ReliefF得到的属性约简集为: 2, 3, 4, 12, 23, 24, 25, 26, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39.用 C4.5对 ReliefF与本文提出的算法得到的属性约简集进行测试, 结果如表 4所示.

表 4 实验结果比较

Table 4 Comparison of experimental result

	Normal	Dos	R2L	Probe	U2R	total	属性个数
本文算法	1	0.986 25	0.985 79	0.998 75	0.769 23	0.991 30	19
ReliefF	0.997 5	0.997 5	0.978 69	0.998 75	0.576 92	0.990 63	19

从表 4可以看出, 本文的算法得到的测试精度略好于 ReliefF算法, 在 U2R的识别上明显高于 ReliefF.

5 结语

本文在文献[17]提出的模糊辨别矩阵的基础上, 提出了新的混合辨别矩阵, 可以对既有离散属性又有连续属性的数据直接进行属性约简, 并在 kdd cup 99上对新的算法进行了测试. 实验表明分类器在新算法得到的特征子集上有较好的分类效果.

[参考文献] (References)

[1] Denning D E. An intrusion detection model[J]. IEEE Transactions on software Engineering, 1987, 13(2): 222-232.

[2] Kohavi R, John G. Wrapper for feature subset selection[J]. Artificial Intelligence, 1997, 97(1/2): 273-324.

[3] Ran Gilad-B, Amir N, Naftali T. Margin based feature selection-theory and algorithms[C] // Proc of the 21st Int Conf on Machine Learning. Banff, Canada, 2004: 43.

[4] Kira K, Rendell L A. A practical approach to feature selectionA[C] // Sleeman D, Edwards P, eds. Proceedings of the 9th International Workshop on Machine Learning C. San Francisco, CA: Morgan Kaufmann, 1992: 249-256.

[5] Pawlak Z. Rough sets[J]. Computer and Information Science, 1982, 11(2): 341-356.

[6] Hu X H, Cercone N. Learning in relational databases: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338.

[7] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for neural networks[J]. Computational Intelligence, 1995, 11(2): 339-347.

[8] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J]. 计算机学报, 2006, 29(3): 407-413.
Yang Ming. An incremental updating algorithm of the computation of a core based on the improved discernibility matrix[J]. Chinese Journal of Computers, 2006, 29(3): 407-413. (in Chinese)

[9] Wang Jue, Wang Ju. Reduction algorithm based on discernibility matrix: the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489-504.

[10] Liu Shaohui, Sheng Qiujian, Wu Bin, et al. Research on efficient algorithms for rough set methods[J]. Chinese Journal of Computers, 2003, 26(5): 524-529.

[11] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligences, 1998, 105(1/2): 77-103.

[12] 苗夺谦, 胡桂荣. 知识约简的一中启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681-684.
Miao Duoqian, Hu Guirong. A heuristic algorithm for reduction of knowledge[J]. Journal of COmputer Research and Development, 1999, 36(6): 681-684. (in Chinese)

[13] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报, 2007, 30(5): 815-822.
Yang Ming. An incremental updating algorithm for attribute reduction based on improved discernibility matrix[J]. Chinese Journal of Computer, 2007, 30(5): 815-822. (in Chinese)

[14] Kononenko I. Estimating attributes: analysis and extensions of Relief A[C] // De Raedt L, Bergadano F. Proceedings of

the 7th European Conference on Machine Learning C. Berlin: Springer, 1994: 171-182.
[15] Skowron A, Rauszer C. The Discernibility Matrices and Functions in Information Systems[C] // Slowinski R. Intelligent Decision Support: The Netherlands: Kluwer Academic Press, 1992: 331-362.
[16] Gupta K M, David W A, Philip M. Rough set feature selection algorithms for textual case-based classification[C] // Roth-Berghofer T R. ECCBR 2006, Lecture Notes in Artificial Intelligence, 2006, 4106: 166-181.
[17] Yang Ming, Chen Songcan, Yang Xubing. A novel approach of rough set-based attribute reduction using fuzzy discernibility matrix[C] // Proceedings of 4th International Conference on Fuzzy Systems and Knowledge Discovery. Washington, DC: IEEE Computer Society, 2007: 96-101.
[18] Yang M, Yang P. A novel condensing tree structure for rough set feature selection[J]. Neurocomputing, 2008, 71(4-6): 1092-1100.

[责任编辑: 严海琳]

(上接第 70页)

[5] Schapire R E, Singer Y. Boostexter: a boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2/3): 135-168.
[6] Lu Yuchang, Lu Mingyu, Li Fan. Analysis and construction of word weighing function in VSM[J]. Journal of Computer Research & Development, 2002, 39(10): 1205-1210.
[7] Xue Xiaobing, Zhou Zhihua. Distributional features for text categorization[C] // Proceedings of the 17th European Conference on Machine Learning (ECML'06). Berlin: LNAI 4212, 2006: 497-508.
[8] Lewis D D. Naive(Bayes) at forty: The independence assumption in information retrieval[C] // Proceedings of 10th European Conf on Machine Learning. Berlin: Springer, 1998: 4-15.
[9] Sauban M, Pfahringer B. Text categorization using document profiling[C] // Proceedings of PKDD-2003. Berlin: Springer-Verlag, 2003: 411-412.
[10] Craven M, DiPasquo D, Freitag D, et al. Learning to extract symbolic knowledge from the World Wide Web[C] // Proceedings of AAAI-98. Madison: WI, 1998: 509-516.

[责任编辑: 严海琳]