文章编号:1000 - 582X(2003)07 - 0047 - 05

一种 Web 文档在线自适应分类新方法*

黎　昱,黄 席 樾,周　欣,刘　涛
(重庆大学 自动化学院,重庆　400044)

摘　要:Web 文档自动分类是 Web 挖掘中的重要研究内容。文档向量空间模型(VSM)是实现文档自动分类的基础,但如何排除冗余属性并降低向量空间的维数是一个难点。文中运用粗集理论对由样本文档集合构成的信息系统进行数据泛化,并求取文档的最优规约属性集,大大降低了文档的特征空间的维数,减少了冗余属性对文档分类的干扰,提高了分类效率。运用 Fuzzy ARTMAP(adaptive resonance theory mapping)神经网络,利用其自适应分类和增量学习的优良特性,实现了对 Web 文档的在线自适应分类。

关键词:网页分类;粗集;属性规约;在线自适应分类
中图分类号:TP391　　**文献标识码**:A

Internet 信息的有效利用有赖于对 Web 文档有效的组织和管理,恰当的分类可以帮助人们准确、迅速地定位所需的信息和分流信息,因而对 Web 信息采掘具有重要意义。因此,网页的自动分类正逐渐受到研究者们的重视并成为越来越重要的研究领域。

一个 Web 信息采掘系统,其通常的工作方式是:通过多个 Robot(或称 Spider)进程在 Internet 上进行漫游,下载遇到的网页并交由分类器进行分析,然后将其链接归入恰当的分类。在实际工作中,预先人为划定的分类不太可能覆盖整个文档特征空间,新的模式的出现要求及时产生新的分类,并且对实时下载的网页也需要进行实时的分类。因此,具备在线的、增量的、自适应学习能力的自动分类器是非常必要的。

许多文献中已提出了一些 Web 文档自动分类的方法。较多的是采用贝叶斯分类器来实现自动分类[1]。贝叶斯分类方法虽然理论上可达到最小经验风险,但在实际应用中一般很难获得先验概率和类条件概率,尤其对于高维的情况,影响了该方法的推广应用;支持向量机是在统计学习理论基础上发展起来的机器学习方法,一些研究者提出采用支持向量机进行网页分类[2],但由于网页数据规模非常大,支持向量机的训练过程过于漫长,且至今还没有一种自适应的学习算法,难以适应网页分类中在线学习的要求;文献[3]采用一种近邻算法进行分类,文献[4]运用决策树技术来产生分类器。这两种方法,都未能解决在线自适应分类问题,而后者由于文档的属性空间维数通常很高,决策树的构造还存在一定困难。

笔者针对以上问题,在已有研究成果的基础上,提出了一种基于粗集理论和 Fuzzy ARTMAP 神经网络的在线自适应分类方法。

1　文档特征空间的建立及属性规约

1.1　建立 Web 文档的向量空间

在对 Web 文档进行矢量化之前需要对文档进行预处理。主要包括英文文档的 Stemming 处理,对中文文档需进行分词处理。与传统的基于关键字(词)的布尔模型不同,文档向量空间模型(Vector Space Model,VSM)能较为全面地反映文档的特征[5]。在理想情况下,每一文档都映射为一组规范化正交词条向量($T_1, T_2, \cdots, T_n$)张成的向量空间 S 中的点,($T_1, T_2, \cdots, T_n$)构成 S 的坐标系(基)。Salton[5] 提出了一种文档向量表示方法,文档 D_i 可被表示为一个 n 维向量 $W_i = (w_{i1}, w_{i2}, \cdots, w_{in})$;对于 HTML 文档,其分量还与该词条出现的 HTML 标记域有关[6]。通常的作法是给不同的标记域 F 赋以不同的权重 R_F,则 w_{ik} 计算如下:

$$w_{ik} = \log(N/n_k) \cdot \sum_F R_F \cdot tf_{ik}^F \Big/ \sqrt{\sum_l \left(\log(N/n_k) \cdot \sum_F R_F \cdot tf_{lk}^F\right)^2} \qquad (1)$$

其中 tf_{ik}^F 表示特征词条 k 在文档 i 内的标记域 F 中的出现次数,tf_{lk}^F 的含义类似,l 表示文档编号,n_k 为出现该特征词条的文档在训练样本集中的数量,N 为训练样本数。

在上述模型中特征词条的选择至关重要,如果把

* 收稿日期:2003 - 02 - 30
作者简介:黎昱(1974 -),男,四川宜宾人,重庆大学博士研究生,主要研究领域为数据挖掘、模式识别、神经网络。

这些特征词条看作文档的属性集，那么提高属性集的完备性以及降低属性间的相关性是提高文本向量空间中模式的可分性以及分类效率的重点。遗憾的是，相关的文献大多忽略了这一重要工作。

1.2 基于粗糙集的属性规约及文档特征空间降维

粗集(RS)理论于1982年由波兰学者Pawlak Z首先提出[7]，目前有多种粗集模型。作为一种数据推理的新方法，它能够在缺少关于数据的先验知识的情况下，仅仅以对观测数据的分类为基础，解决模糊或不确定性数据的分析和处理。粗集理论以信息系统的形式表示数据，对象以属性表征，其描述形式类似于关系型数据库，基于RS理论的数据归约的基本原理是通过求属性重要性并排序，在泛化关系中找出与原始数据具有相同决策或分辨能力的相关属性的最小集合，实现信息约简[8-9]。

从粗集理论的观点看，特征词条 $T_1, T_2, \cdots, T_n$ 构成了属性集 $\{T_1, T_2, \cdots, T_n\}$。定义信息系统：

$$K = < U, Q, V, f > \quad (2)$$

其中 $U = \{d_1, d_2, \cdots, d_L\}$ 为所有训练样本文档构成的论域；$Q = \{C, D\}$ 为属性集合，其中 $C = \{T_1, T_2, \cdots, T_n\}$ 称为条件属性，$D = \{dp\}$ 为结果属性(或决策属性)，dp 的取值为文档所属的分类；$V = U_{q \in Q} V_q$ 为属性值的集合；f 为信息函数，对 $\forall d \in U, q \in Q$，有 $f(d, q) \in V_q$。各属性的取值范围为[0,1]连续取值，为了便于粗集处理，需先进行数据泛化，将区间[0,1]离散化为 M_i 份。

1.2.1 样本文档集的数据泛化

数据泛化就是进行元组规约，其目的在于离散化属性的值域，并裁减冗余元组，从而提高属性规约的效率。因此在进行属性规约之前需对数据进行元组规约处理。

数据泛化常用的方法是对每一个属性 T_i，根据其数据分布的特征建立概念树 H_i(实际上是一种数字层次结构)，然后采用概念树提升方法对属性进行泛化。

Han J和Fu Y提出了一种自动生成数字层次结构的算法(AGHF算法)[9]，其基本思想是：首先构造属性值集合 V_{Ti} 的统计直方图，然后用等宽区间产生叶节点，再用等频率区间法产生高层节点。该方法的算法时间复杂度为 $O(Z)$(Z 为直方图中矩形条数)。由于生成的概念树是用于元组规约，还需引入次序约束，并限制一定的扇出系数 F。前者保证划分的区间为连续的区间；后者使每一节点的分枝数小于或等于 F，这样避免了生成的概念树层次太少，有利于在概念树提升操作中保证数据的精度。

图1是一个实际计算中的一个属性“政治”(特征词条之一)生成的概念树的例子(最大扇出系数限制为3，统计直方图的区间宽度为0.02，文档样本数为2 625)，因篇幅所限不能完整列出：

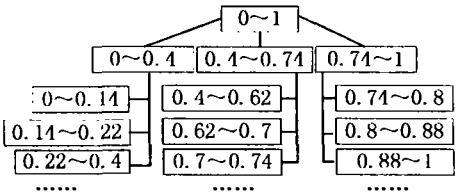

图1 概念树示例

现指定泛化率 $p(0 < p \leq 1)$ 和属性值集合基数的域值 $m_i < M_i$，并设概念树集合为 H，其元素 H_i 对应于相应的属性 T_i 的概念层次，样本文档集合 U，条件属性集 C，描述算法如下：

1) 令 $d_m = p \times card(U)$；$U' = U$；$C' = C$；

2) 计算每个属性值集合的基数 $d_i = card(V_{Ti})$；

3) 如果 $card(U') > d_m$ 并且有 $d_i > m_i$ 转下一步，否则结束，输出泛化的信息系统；

4) 选择使 d_i/m_i 最大的属性 $T_i \in C'$，如果 T_i 为可泛化，则在 Hi 中提升到上一级概念，并用该级概念的区间编号替代 V_{Ti} 中的值，否则，$C' = C' - \{T_i\}$；

5) 从 U' 中删除所有重复元组，然后转第2步；

以上是数据泛化的整个过程，各属性的值域变为：$V_{Ti} = \{1, 2, \cdots, M_i\}$；信息系统 K 转化为泛化的信息系统：

$$K' = < U, Q, V', f > \quad (3)$$

不难证明，这一算法的时间复杂度为 $O(L \times n)$，其中 L 为样本文档数，n 为条件属性集的基数。

1.2.2 文档属性规约

1) 属性的依赖度

对两个属性集合 $R \in C$ 和 D，则 R 在 $U/IND(D)$ 中的正域(或 D 的 R 正域)定义为：

$$POS_R(D) = \bigcup_{X \in U/\Phi} R\_(x) \quad (4)$$

其中 $U/IND(D)$ 表示论域 U 在属性集 D 的不可分辨关系 $IND(D)$ 上的划分；$R\_(X)$ 称为集合 X 的 R 下近似：

$$R\_(X) = \cup \{Y \in U/IND(R) : Y \subseteq X\} \quad (5)$$

下近似 $R\_(X)$ 是那些由 R 定义的知识能确定地归入集合 X 的对象的集合。由于 D 中只有一个决策属性 dp，则式(5)可化为：

$$POS_R(D) = \bigcup_{X \in U/\Phi} R\_(X) \quad (6)$$

其中 U/dp 为 dp 的所有等价类族，具体来说，就是样本文档集 U 的所有分类。

现在，可以定义属性集 D 对属性集 R 的依赖度：

$$\gamma_R(D) card(POS_R(D))/card(U) \quad (7)$$

这里 $card(\cdot)$ 表示集合的基数，对给定的文档训练样本集，$card(U)$ 为一常数。

2) 属性的重要性

属性的重要性反映了条件属性中不同属性对条件属性集与决策属性集间的依赖程度的贡献。设属性集 $R \subseteq C$，条件属性 $a \in R$；则属性 a 在 R 中对 $U/IND(D)$ 的重要程度定义为：

$$G_{R,D(a)} = \gamma_R(D) - \gamma_{R-|a|}(D) \quad (8)$$

3）条件属性集的相对核

核是计算属性规约集的基础，在粗集理论中，核反映了相对于属性集 D，属性集 C 所有规约属性集的交集。对属性 $r \in C$，称属性 r 为 D 不可省略的，当且仅当 $POS_C(D) = POS_{C-\{r\}}(D)$。$C$ 中所有 D 不可省略属性的集合称 C 的 D 核，记为：$CORE_D(C)$。

核可通过分辨矩阵来计算。定义关于属性集 C 的分辨矩阵 $M(C)$：$M(C) = (m_{i,j})_{n\times n}$

$$(m_{i,j}) = \begin{cases} \Phi & d_i, d_j \in D \text{ 的同一等价类} \\ \{T \in C : f(d_i,T) \neq f(d_j,T) \neq f(d_j,T)\} & ; d_i, d_j \in D \text{ 的不同等价类} \end{cases}$$

由于 D 中只有一个属性元素 ds，表示训练样本文档的分类编号，则上式可简化为：

$$(m_{i,j}) = \begin{cases} \Phi & f(d_i,dp) = f(d_j,dp) \\ \{T \in C : f(d_i,T) \neq f(d_j,T)\} & \\ f(d_i,S) \neq f(d_j,dp) & \end{cases} \quad (9)$$

由分辨矩阵可计算：

$$CORE_D(C) = \{T \in C : m_{i,j}\} = \{T\}, 1 \leqslant j < i \leqslant n\} \quad (10)$$

即核中的元素为分辨矩阵中只含有单个元素的单元中的元素构成的集合。

至此，可以计算条件属性集 C 的最佳规约集。算法步骤如下：

1）建立信息系统 K 中关于条件属性集 C 的分辨矩阵 $M(C)$；

2）计算属性集 C 的相对核 $CORE_D(C)$；

3）令 $R = C, E = CORE_D(C)$；$R' = R - E$；

4）对每个属性 $T_i \in R'$，按式(7)计算属性的重要性 $G_{R',D}(T_i)$，并按大小排序；计算属性集 D 对 E 和 R 的依赖程度 $\gamma_E(D)$，$\gamma_R(D)$；

5）如果 $\gamma_E(D) = \gamma_R(D)$，或 $R' = \Phi$ 则转 8 步；

6）选择 $G_{R',D}(T_j)_{max}$ 对应的属性 $T_j, T_j \in R'$。注意可能有多个属性的重要性等于最大重要性，则在其中选择属性值最少的属性（即 $card(V_j)$ 最小）；

7）$E = E \cup \{T_j\}$，$R' = R' - \{T_j\}$，重新计算 D 对 E 的依赖程度：$\gamma_E(D)$，并转第 5 步；

8）令 $n' = card(E)$，$\Gamma = \gamma_E(D)$，下面是一个反馈过程，以伪代码的形式表述如下；

```
9)for (int i = 0;i < n';i ++) {
    if (T_i ∉ CORE_D(C) )E = E - {T_i} ;// 从E逐个中去掉不属于相对核的属性
    计算 γ_E(D);
    if (γ_E(D) ≠ Γ)   E = E ∪ {T_i} ; // 如依赖度改变，则恢复该属性
}
```

10）输出最终得到的最佳属性规约集 E，$card(E) = m, m \leqslant n$。

至此，可得到文档的最佳属性规约集 E。在测试中，初始属性集的属性数量为 378 个，用该算法求得的最优属性集包含的属性数为 205，显示这一算法对文档特征空间的降维是行之有效的。

3　基于神经网络的文档在线自适应分类

自适应共振理论（ART）是由美国 Boston 大学的 Grossberg 和 Carpenter 两人提出的，能较好地模拟人类的认知行为，较好地解决了神经网络的稳定性 - 可塑性的两难问题，并能在动态环境下进行实时在线学习。此后又发展出了 Fuzzy ARTMAP[10]，它能处理离线的、有示教的学习。该网络由两个 Fuzzy ART 模块级联构成，分别称 ART_a 和 ART_b。ART_a 接受 m 维输入模式向量 a，在离线有监督训练阶段，ART_b 的输入为输入模式 a 的期望输出 b，表示其正确分类。将网络用于文档分类时，ART_b 的示教输入为样本文档的确切分类，表示为向量：

$$\boldsymbol{b} = (b_1, b_2, \cdots, b_n) \quad (11)$$

其中：$b_i = \begin{cases} 1 & i = k(k \text{ 为样本文档 } a \text{ 所属分类}) \\ 0 & \text{其他} \end{cases}$

n 为样本文档总的分类数。因此直接将 $\boldsymbol{b}$ 加载到映射场，得到如图 2 所示的简化的 Fuzzy ARTMAP。

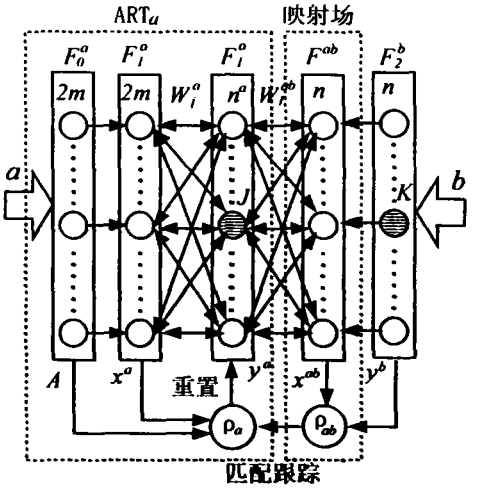

图 2　简化的 Fuzzy ARTMAP 结构

在 F_0^a 层将 a 转换为补码形式，即：$A = (a, a^c)$，其中，$a_i^c = 1 - a_i (0 \leqslant a_i \leqslant 1)$。这样处理实际上是规格化输入向量，使 $|A|$ 为常数。网络中映射场的神经元数目与 F_2^b 相同，且为一一对应关系。令 x^a，y^a 分别为 F_1^a，F_2^a 的输出，W_j^a 为连接 F_2^a 中第 j 个神经元的 $2m$ 维连接权向量；y^b 为 F_2^b 的输出。ART_a 中 F_2^a 的第 j 个神经元与映射场神经元间的连接权向量为 W_j^{ab}，x_{ab} 为匹配域 F^{ab} 的输出向量。

Fuzzy ARTMAP 的监督训练算法如下：

初始化网络。置 x^a, y^a, y^b, x^{ab} 为零向量，W_j^{ab}、W_j^a 中的权值分量均置为 1。警戒参数 ρ_a 置为基准值 $0 < \rho_a < 1$。

输入样本 (a^i, b^i)，a^i 通过 F_0^a 变换为 A。f_2^a、F_2^b 的输出 y^a, y^b 如下：

$$\left.\begin{aligned} y_j^a &= |A \wedge W_j^a| / (\alpha + |W_j^a|) \\ y_j^b &= b_j \end{aligned}\right\} \quad (11)$$

其中 α 为选择参数，一般取为很小的正数；$\wedge$ 为模糊与，定义为：$(p \wedge q)_i \equiv \min(p_i, q_i)$，这一式子反映了输入向量 A 与长期记忆权阵的匹配程度。F_2^a 通过竞争，获胜单元产生输出。设 F_2^a 的获胜单元为 J，则映射场激活。映射场的 F^{ab} 层的输出为：

$$x_{ab} = \begin{cases} y^b \wedge W_J^{ab} & F_2^b \text{ 的第 } J \text{ 节点与 } F_2^b \text{ 均激活} \\ W_J^{ab} & F_2^a \text{ 的第 } J \text{ 节点激活而 } F_2^b \text{ 未激活} \\ y^b & F_2^a \text{ 未激活而 } F_2^b \text{ 激活} \\ 0 & F_2^a, F_2^b \text{ 均未激活} \end{cases} \quad (12)$$

如果 y^b 与 W_J^{ab} 失配，ART_a 开始匹配跟踪过程以搜索一个更好的类别节点。

匹配跟踪。若 $|x_{ab}| < \rho_{ab}|y^b|$，则 ART_a 的预测分类与实际分类失配，此时应修改 ART_a 的警戒参数 ρ：

$$\rho_a = |A \wedge W_J^a| / |A| + \delta \quad (13)$$

其中 δ 为略大与 0 的数，使 F_1^a 层输出 x^a 满足下式：

$$|x^a| = |A \wedge W_J^a| < \rho_a |A| \quad (14)$$

则 ART_a 中原有的谐振条件被破坏。ART_a 开始选择其他 F_2^a 节点 K，使其同时满足：

$$|x^a| \geqslant \rho_a |A| \text{ 和 } |x^{ab}| \geqslant \rho_{ab} |y^b| = \rho_{ab}$$

如满足上两式，则进行"慢学习"，更新其连接权向量

$$W_{K(\text{new})}^a = \beta(A \wedge W_{K(\text{old})}^a) + (1-\beta) W_{K(\text{old})}^a \quad (15)$$

$$W_{K(\text{new})}^{ab} = \beta(y^b \wedge W_{K(\text{old})}^{ab}) + (1-\beta) W_{K(\text{old})}^{ab} \quad (16)$$

其中 $0 \leqslant \beta \leqslant 1$，称为学习速率。

如果所有的 F_2^a 节点都不满足，则产生一个新的 F_2^a 节点 N，初始化其连接权（快学习）：

$$W_{N(\text{init})}^a = A \quad (17)$$

在映射场，假定这个新节点与 F_2^b 的第 S 类别节点对应，则该节点与映射场间的连接权向量为：

$$W_{NS}^{ab} = 1 \text{，其他 } W_{Nk}^{ab} = 0 (k \neq S) \quad (18)$$

当网络工作于在线分类状态时，不再对网络提供示教输入，其工作方式与普通的 Fuzzy ART 相同，对于新的文档于监督训练时产生的模式分类均不匹配时，Fuzzy ARTMAP 在 F_2^a 层中产生新的类别节点以存储这一新的模式。

4 实验结果

整个系统的工作流程如图 3 所示，系统基于 Windows2000 平台，采用 VC ++6.0 开发实现。其中的自动分词模块采用了东北大学中文信息处理实验室研制的 CipSegSDK 动态链接库，谨此表示感谢。

在测试中，使用 Google 搜索引擎，以"新闻"为主题搜索中文网页，然后在返回的结果中进一步分别以"政治"、"财经"、"娱乐"、"体育"共 5 个主题进行检索，从中共挑选网页 2 625 个，其中选择 1 475 个网页用于训练

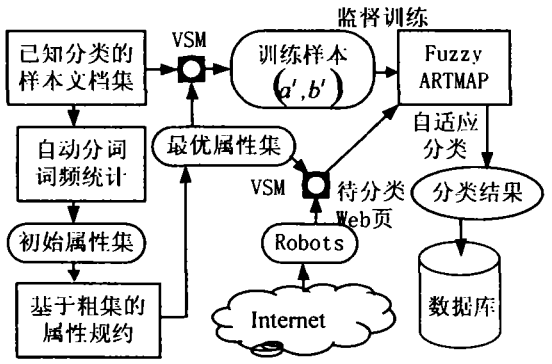

图 3 系统工作流程图

（通过数据规约后，余下 1 087 个），其余 1 150 个网页用于测试。

为减少数据泛化和属性规约的计算规模，并避开与主题无关的词条，从人民日报标注语料库中挑选了 563 个常用词条，并加入近期新闻中较常出现的地名、人名等实词共计 614 个，再通过对样本文档进行词频统计，从中选择了与上述 5 类网页相关的词频较高的前 378 个词条，得到初始属性集；然后采用前文所述的基于粗集理论的属性规约算法，得到最优属性集，其中共包含属性（特征词条）205 个。进而用 VSM 模型将所有网页表示为 205 维向量。

分别采用传统的支持向量机（SVM）方法和 K－近邻法作为对比。测试分为两部分，一是将余下的 1 150 个网页用作测试集，用于对比本文的方法和这两种方法的分类准确率。二是通过 Robot 程序在网上实时下载网页。作为一个试验，先由 Robot 程序依据超链接是否包含特征词条，对网页进行初步过滤，通过过滤后的网页（共计 1 200 个）再交由分类器分类。

由于单一的支持向量机用于多类分类问题比较困难，笔者构造了一种二叉树型的分类器，在每个节点用支持向量机分类，如图 4 所示。由于在训练阶段，SVM 算法的时间开销较大，并且需计算 3 个 SVM，故耗时最多。

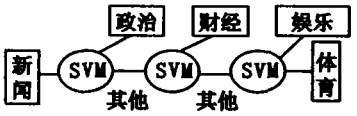

图 4 组合的 SVM 分类器

K－近邻算法的时间复杂度较高。每次输入待分类样本时，判别其类别均需要计算与所有样本的距离，并选择最近的 K 个样本，因此在给定特征属性集的前提下，每判别一个样本的时间复杂度为 $O(N)$，N 为样本文档数。

在第一部分测试中，每种方法均分别采用了初始属性集和规约后的属性集构造输入模式向量作为对比，测试结果如表 1 所示（其中 R 表示采用规约属性集，UR 表示采用初始属性集）。结果显示，SVM 算法在第一部分测试中的平均精度最高，Fuzzy ARTMAP 的精度略低于 SVM。测试数据还显示"娱乐"和"体育"

两类的识别率较低，分析分类后的网页发现，这两类网页发生混淆的样本较多。

表 1　第 1 部分测试结果　　%

类别	政治		财经		娱乐		体育	
	UR	*R*	*UR*	*R*	*UR*	*R*	*UR*	*R*
SVM	86.0	98.4	86.2	97.8	83.2	94.7	80.2	91.6
K－近邻法	96.5	88.7	77.6	89.7	75.5	86.4	72.3	84.5
FARTMAP	84.7	97.2	84.0	961	82.3	92.5	79.1	93.3

第 2 部分测试采用在线方式，Fuzzy ARTMAP 的 F_2^a 层节点被限制为 12，但最后使用了 10 个类别节点（即产生了 6 个新的类别节点）。由于这一部分测试的样本为直接从网上获取，有一定的随机性；并且，大量的新闻类网页并不能简单地归入已有的类别，如同一篇网页中可能既包含娱乐信息，也有财经报道等。因此，在第 2 部分测试中，SVM 和 K－近邻法的分类效果均不理想，前者有约 42% 的文档被正确归类，后者仅为 35.3%，其余大量的网页被归入不能识别类。而 Fuzzy ARTMAP 分类器不但以较高的精度归类属于已有 4 个类别的文档，并且产生了新的类别以记录包含混合信息的新的文档模式，平均准确率达 92.8%，充分体现了自适应分类器的优越性。

5　结　论

采用基于粗集理论的数据规约，有效地解决了文档特征空间坐标系的生成以及特征空间维数过高的问题，从理论上给出了减少冗余属性的方法，有助于文档判别正确率的提高并减少了计算规模，使 3 种算法在第 1 部分测试中分类精度均有显著提高。

通过比较可以看出，将 Fuzzy ARTMAP 用于在线网页分类时，其自适应聚类的能力具有明显的优势，虽然在离线测试中，其平均精度略小于支持向量机，但它所具有的自适应能力弥补了这一不足。因此，这一基于自适应神经网络的分类算法在网页自动在线分类中具有较好的应用前景，其研究对面向 Internet 的信息管理和知识发现具有重要意义。

参考文献：

[1] KONT KANEN P, MYLLYMAKI P, SILANDER T, et al. BYDA: software for Bayesian classification and feature selection[A]. AGRAWAL R, STOLORZ P E, PIATETSKY－SHAPIRO G, eds. Processdings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD '98) [C]. Menlo Park: AAAI Press, 1998,254－258.

[2] 李晓黎，刘继敏，史忠植，等. 基于支持向量机与无监督聚类相结合的中文网页分类器[J]. 计算机学报，2001，24(1)：62－68.

[3] YANG Y. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval [A]. Proc·Seventeenth International ACM SIGIR Conference on Research and Developmentin Information Retrieval [C]. Dublin, 1994.

[4] APTE C, DAMERAU F, WEISS S. Automated learning of decision rules for text categorization[J]. ACM Transactions on Information System, 1994, 12 (3): 233－251.

[5] SALTON G, WONG YAND C S. A Vector space model for automatic indexing[J]. Communications of ADC, 1975, 18 (11): 613－620.

[6] SALTON G. Introduction to Modern Information Retrieval [M]. New York: Mc Graw－Hill Book Company, 1983.

[7] PAWLAK Z. Rough Sets－Theoretical Aspects of Reasoning About Data[M]. Kluwer Academic Pub, 1991.

[8] 刘同明. 数据挖掘技术及其应用[M]. 北京：国防工业出版社，2001.

[9] HAN J, FU Y. Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases[A]. Proc. AAAI'94 Workshop on Knowledge Discovery in Database (KDD'94) [C]. 1994,157－168.

[10] CARPENTER G A. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps[J]. IEEE Trans. Neural networks, 1992,3 (5):698－713.

A New Method of Online Adaptive Classification of Web Pages

LI Yu, HUANG Xi-yue, ZHOU Xin, LIU Tao

(College of Automation, Chongqing University, Chongqing 400044, China)

Abstract: The web documents classification is an important research content of web mining. Document vector space model is the foundation of automatic classification of documents, while it is difficult to eliminate redundant attributes and reduce the dimension of the vector space. The Rough Sets Theory is applied to generalize the information system comprised by document samples set, and to compute the best reducing properties set. So dimension of document feature space is reduced greatly, and disturbance to document classification is decreased too, which improve the efficiency of classification. In addition, using the advantage of adaptive classification and incremental learning of Fuzzy ARTMAP neural network, the online adaptive classification of web document is achieved.

Key words: web pages classification; rough sets; attributes reduction; online adaptive classification

（编辑　吕赛英）