第34卷 第15期 计 算 机 工 程 2008年8月
Vol.34 No.15 Computer Engineering August 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)15—0208—03 文献标识码：A 中图分类号：TP393.08

# 粗糙集理论和DT_SVM在Web信息过滤中的应用

衣治安，刘 杨
(大庆石油学院计算机与信息技术学院，大庆 163318)

**摘 要：**针对Web信息过滤问题，提出一种将粗糙集理论和决策树SVM(DT_SVM)相结合进行数据分类、过滤的新方法。该方法运用改进的启发式相对属性约简算法消除冗余、降低样本空间维数，通过聚类和DT_SVM相结合来训练SVM，将多分类问题转化为二值分类问题，提高了训练速度及过滤精度。实验表明，该算法得到了较高的查全率、查准率，体现了将粗糙集理论与DT_SVM算法结合的优越性。
**关键词：**Web信息过滤；粗糙集理论；DT_ SVM算法；属性约简；聚类

# Application of Rough Set Theory and DT_SVM in Web Information Filtering

YI Zhi-an, LIU Yang
(College of Computer and Information Technology, Daqing Petroleum Institute, Daqing 163318)

**【Abstract】** This paper advances a new data classification and filtering method based on rough set theory and Decision Tree SVM (DT_SVM) in allusion to the problem of Web information filtering. This method utilizes an improved heuristic algorithm of relative attribute reduction to eliminate redundancy, debase the spacial dimension of sample data, and train SVM by clustering integrated with DT_SVM, it can change multiclass problem into binary classification, and improve the training speed and the filtering precision. Experimental results demonstrate that the new algorithm gains a higher filtering recall and precision, manifests the algorithm's advantage of rough set theory integrated with DT_SVM.
**【Key words】** Web information filtering; rough set theory; DT_SVM; attribute reduction; clustering

## 1 概述

随着Internet的高速发展，网络资源如潮水般涌来，用户要找到自己感兴趣的信息，无异于大海捞针。为了更有效、更准确地为用户提供有用信息，满足用户的个性化需要，信息过滤技术应运而生。它是一种从动态的信息流中提取符合用户个性化需求的系统化数据处理方法。近年来，国内外学者取得了一些进展，重点是如何获取用户相对稳定的个性兴趣、中英文信息自动分类等，研究主要集中在特征抽取、学习算法和过滤算法上。由于信息内容的多样化、语言逻辑的复杂化以及更新速度快，因此信息自动过滤方面的研究进展十分缓慢。

针对这一问题，本文提出了一种将RS理论[1]和SVM[2]相结合的Web信息过滤新方法，由于SVM训练在处理数据量较大的模式分类问题时，会增加SVM分类器的复杂度、导致SVM训练时间较长，而粗糙集理论能够在处理不确定知识、消除冗余信息和发现样本数据属性关系上具有突出的优势，因此将两者结合可在分类前对数据进行属性约简，以压缩数据空间，降低分类的维数；同时RS理论方法在数据分类过程中对噪声比较敏感，而SVM方法则具有较好的抑制噪声干扰的能力和良好的泛化能力。因此，将无噪声的训练样本学习应用于有噪声的环境中时效果很理想。

信息过滤的实质是一种信息分类技术，信息训练方法和分类决策算法是信息过滤的核心。本文将粗糙集理论和信息熵相结合，从条件属性之间的关系出发，提出一种启发式相对属性约简算法，以RS处理后的数据作为样本数据，采用基于核函数的聚类DT_SVM算法，确定每个决策节点的最优超平面，根据用户的个性化需求有效的进行信息过滤。

## 2 基于RS理论的属性约简和DT_SVM的分类算法

### 2.1 改进的启发式相对属性约简

决策表 $S=<U,C\cup D,V,f>$ 中，若存在一个子集 $P\subseteq C$，使 $POS_R(U,D)=POS_C(U,D)$，并且 $P$ 是最小集，即不存在 $Q\subset P$，使 $POS_Q(U,D)=POS_C(U,D)$，称 $P$ 是 $C$ 关于 $D$ 的相对约改进算法。是以属性的核作为初始候选约简集 $R$，以属性约简的贡献度 $CON_{C-s}(s)$、条件熵作为启发信息。先计算 $CON_{C-s}(s)$，选择 $CON_{C-s}(s)$ 大的属性逐次加入到 $R$ 中，当最终属性约简集与决策属性集的条件熵等于初始条件属性集与决策属性集的条件熵( $H(D/red(C)=H(D/C))$ )时算法结束，这样就得到一个相对约简。

算法描述：

输入 决策系统 $S=<U,C\cup D,V,f>$。其中，$U$ 为论域；$C, D$ 分别为条件属性集和决策属性集。

输出 该决策系统的一个相对约简。

**Step1** 计算条件属性的核 $C_0=Core_D(C)$，作为初始属性约简集，计算 $C, D$ 的条件熵 $H(D/C)$，$red(C)$ 为最小约简，$U_1-POS_R(U,D)\Rightarrow U_1$。

**Step2** 将条件熵中贡献最大的属性逐次加入到约简集中，对每个属性 $s\in C$，计算 $CON_{C-s}(s)$，选出条件熵中最大

**基金项目：**黑龙江省研究生创新科研基金资金项目(YJSCX2006-38 HLJ)
**作者简介：**衣治安(1964－)，男，教授、在职博士研究生，主研方向：计算机网络与通信，人工智能；刘 杨，硕士研究生
**收稿日期：**2007-08-21 **E-mail：**nice_xiaoxiao@163.com

的属性，$C_0 = C_0 + \{s\}$，比较 $H(D/red(C))$ 和 $H(D/C)$ 的大小，转为 Step5。

**Step3** 令 $p \in C - red(C)$，$p$ 为部分属性约简后剩余的元素，计算

$CON_{red(C)Up}(D/p)$，

$CON_{red(C)Up}(D \mid p) = H(D \mid red(C)) - H(D \mid red(C) \cup \{p\})$，

$\max\{CON_{red(C)U\,p}(D|p)\}$，$red(C) = red(C) + \{p\}$

**Step4** 进行反向剔除测试，令 $R - C_0 \Rightarrow B$

(1)若 $B = \Phi$ 转(4)；

(2)$R$ 中任意取出一个属性 $q$，如果

$POS_{R-\{q\}}(U,D) = POS_C(U,D)$

那么

(3) $R \Leftarrow R - \{q\}$ ；

(4) $B \Leftarrow B - \{q\}$ (1)。

**Step5** 计算 $H(D/red(C))$，若 $H(D/red(C) = H(D/C))$，则 $R = red(C)$ 为最小约简，输出最小约简 $red(C)$，算法结束；若 $H(D/red(C) > H(D/C))$，则 $j$=$j$+1，转 Step3。

**Step6** 最后得到的 $red(C) = \{a_1, a_2, \cdots, a_n\}$ 就是 $C$ 相对于 $D$ 的一个约简。

**Step7** 根据各条件属性的贡献度，构造条件属性贡献度矩阵 $\boldsymbol{M}$，即 $\boldsymbol{M}_{CON} = \begin{pmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{pmatrix}$，$a_i = CON_{p-c_i}(D/c_i)$，$i$=1, 2,⋯,$n$。

算法改进后，属性的贡献度矩阵同时生成，避免了计算各属性组合与决策属性之间的互信息，从而减少了计算量并提高了属性约简的速度。

### 2.2 决策树支持向量机(DT_SVM)

SVM 分类算法最初是针对两分类问题提出的，而很多 Web 信息数据为多类数据，因此进行 Web 信息过滤必须对数据进行多分类。目前性能较好的多分类算法有 1-$v$-$r$ $SVM_S$ [3]，1-1-1 $SVM_S$ [4]和 DT_SVM[5]，前 2 类算法训练 SVM 时会存在大量不可分区域，且训练时间和测试时间较长，而 DT_SVM 可将多类分解成一系列二值分类问题，克服不可分现象，同时对于 $C$ 类分类问题只需构造 $C$-1 个分类决策函数，从而可节省测试时间。

然而在构建决策树时，易产生错分累积的现象，即决策点一旦被上层分类器错误分类，这种错误会被下层分类器传递下去，导致决策点离正确类别越来越远。如一个属于类别 $C_2$ 的决策点在 $SVM_0$ 处发生错分，这个错分被 $SVM_2$ 积累并传递给 $SVM_3$，使原本属于类别 $C_2$ 的测试点最终被错误分类。错分累积是影响其分类精度的重要因素，当各决策节点的 2 个子类交叠较为严重时，分类器的分类精度不高。

为了降低错分累积的程度，使 DT_SVM 具有最优的分类性能，本文采用 DT_SVM 与基于核函数的聚类算法相结合的思想。即通过聚类，将 $K$ 类样本聚为 2 类，使每类样本的聚类中心距离最大，类中的样本数据相差最小，从而确定每个决策节点的最优超平面。在算法中，分类器超平面：$\Phi(\omega,\varepsilon) - \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\varepsilon_i$，核函数：$K(x,x_i) = \exp\{-\frac{|x - x_i|^2}{\sigma^2}\}$。

最短距离法：把类 $S_p$ 与类 $S_q$ 中 2 个最近样本向量之间的欧氏距离作为 $S_p$ 与类 $S_q$ 之间的距离，即

$d_{p.q} = \min\{\|x_i - x_j\| \mid x_i \in S_p, x_j \in S_q\}$

算法如下：

**Step1** 初始状态只含有决策节点 $x$，将 $x$ 作为决策数的根，即 $x$ 为全部训练样本集。

**Step2** 采用核函数的聚类算法，将 $X$ 聚类成 2 个子集，$X'_{+1}$ 和 $X'_{-1}$；分别计算测试网页到当前类训练聚类中心 $O'_{+1}$ 和 $O'_{-1}$ 的最大相近度，以最短类距离法计算 $x$ 到中心的距离：

$d_x^+ = \max_{j=1}^{m} d(x, O'_{ij})$，$d_x^- = \max_{k=1}^{n} d(x, O'_{ik})$

若 $d_x^+ > d_x^-$，则 $x \in X'_{+1}$；否则，$x \in X'_{-1}$ ($O'_{ij}$, $O'_{ik}$ 分别为聚类中心；$m$, $n$ 分别为聚类中心的簇数)。

**Step3** 根据生成的决策树，调用 SVM 的学习算法构建决策树内各节点的最优超平面；分类函数为

$f(x) = \sum_{i=1}^{n} a_i^* y_i (x_i.x) + b^*$

将 $f(x)$ 的最大值放在 max 中，若 $f(x) > \max$，则 $\max \leftarrow f(x)$，$n_i \leftarrow i$。

**Step4** $i \leftarrow i+1$，若 $i < m+1$，返回 Step2。

对于测试网页 $x$，依次计算分类器的决策函数 $f(x)$，根据类序号 $n_1, n_2, \cdots, n_k$ 排序。由于测试网页和聚类中心的相似度之差的计算效率较高，且训练集合具有层次结构，每增加一个新类，子 SVM 分类器不必重新训练，只要在树根处增加子 SVM 分类器即可减少了训练样本的支持向量数，因此该分类算法具有较高的网页过滤速度。

## 3 本文算法在 Web 信息过滤中的应用

通过对文档数据的预处理，如 Web 信息向量化，中、英文分词，滤掉稀疏词条，删除禁用词，最后生成用户兴趣模版，更能体现文档主体，提高文档过滤的精度。

将粗糙集属性约简和决策 SVM 分类算法应用于 Web 信息过滤中，总体结构框架如图 1 所示。

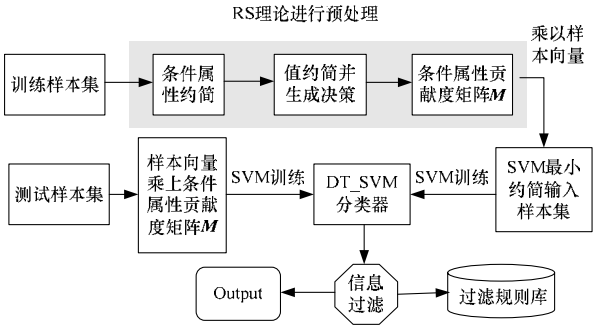


**图 1 粗糙集与决策 SVM 结合框架**

## 4 仿真实验

为了验证本文方法的可行性和有效性，从 Internet 上下载了 800 篇网页进行仿真实验，网页分为 6 个类别，将 600 篇作为训练样本，200 篇作为测试样本。

RS 属性约简通过 VC++语言编程实现，SVM 核函数选择 RBF 函数，采用交叉确认法来训练数据，训练结果如表 1 和表 2 所示，错分样本的惩罚因子 $C$ 越大表示惩罚越大，$\gamma = 1/\sigma^2$ 参数是核函数的控制因子，越大表示模型复杂度越低。

由表 1 和表 2 可以看出，效果最佳的一组数据为：参数 $C$=15 000, $\gamma$ =0.09，支持向量数为 1 070 个，过滤精度为 91.85%。

**表 1　$C$ 取不同值时的训练和测试结果($\gamma$ =0.02)**

| $C$ | 支持向量个数 | 过滤精度/(%) |
|---|---|---|
| 150 | 1 800 | 91.581 5 |
| 400 | 1 419 | 91.584 8 |
| 1 000 | 1 325 | 91.593 3 |
| 4 000 | 1 248 | 91.596 4 |
| 12 000 | 1 134 | 91.747 6 |
| 15 000 | 1 123 | 91.749 3 |
| 18 000 | 1 129 | 91.748 7 |

**表 2　$\gamma$ 取不同值时的训练和测试结果($C$=15 000)**

| $\gamma$ | 支持向量个数 | 过滤精度/(%) |
|---|---|---|
| 0.03 | 1 045 | 91.843 1 |
| 0.07 | 1 051 | 91.844 2 |
| 0.09 | 1 070 | 91.852 1 |
| 0.10 | 1 063 | 91.848 6 |
| 0.15 | 1 065 | 91.839 8 |
| 0.21 | 1 057 | 91.798 4 |

本实验采用查全率和查准率来评价信息过滤的性能。设样本集有 $N$ 个样本，实际上 $D_0$(感兴趣), $D_1$(不感兴趣)的样本个数分别为 $F_0$, $F_1$，采用过滤识别后，归入 $D_0$, $D_1$ 的样本数为 $R_0$, $R_1$，未识别数为 $N_R$，则 $R_0+ R_1+N_R=N$。在归入 $D_0$, $D_1$ 类的 $R_0$, $R_1$ 个样本中，实际上属于 $D_0$, $D_1$ 类的为 $P_0$, $P_1$ 个，即 $R_0$ 个分类为 $D_0$ 类的样本中有 $P_0$ 个分类是正确的，设 $i$ 表示分类，则：$D_i$ 类的查全率($Recall$)=$P_i/F_i$，查准率($Precision$)=$P_i/R_i$，实验结果如图 2 所示。

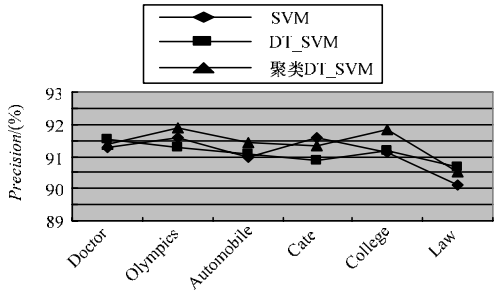


**图 2　SVM, DT_SVM, 聚类 DT_SVM 3 种不同算法的过滤精确率**

实验表明，经 RS 约简后的各类算法比较，聚类 DT_SVM 具有较高的过滤精度和分类速度，可见将聚类与 DT_SVM 结合减少了每个 SVM 的分类的训练样本总数，从而提高了信息过滤的效率。针对多分类问题，分别采用一对一、一对多的聚类 DT_SVM 算法进行实验，结果如表 3 所示。

**表 3　DT_SVM 算法与其他多分类算法的过滤精度比较**

| $C$ | $\gamma$ | 1-$r$-1/(%) | | 1-$r$-$n$/(%) | | 聚类+DT_SVM/(%) | |
|---|---|---|---|---|---|---|---|
| | | SVM | SVM+AR | SVM | SVM+AR | SVM | SVM+AR |
| 1 000 | 0.02 | 87.96 | 89.93 | 89.72 | 91.03 | 89.77 | 91.59 |
| 1 000 | 0.06 | 90.28 | 90.64 | 88.43 | 89.95 | 90.89 | 91.61 |
| 12 000 | 0.07 | 89.82 | 90.15 | 89.89 | 91.02 | 90.50 | 91.74 |
| 15 000 | 0.09 | 90.03 | 91.17 | 90.18 | 91.33 | 91.26 | 91.85 |
| 18 000 | 0.08 | 88.71 | 90.36 | 90.15 | 90.98 | 90.37 | 91.76 |

实验表明，经 RS 约简后的各类算法比较，本文算法降低了训练模型的复杂度，具有较高的过滤精度和分类速度，从而在一定程度上避免了模型的过拟合现象，并提高了 SVM 的推广能力和训练速度，取得了较好的过滤效果。

## 5　结束语

本文提出了基于粗糙集理论和 DT_SVM 的 Web 信息过滤方法，通过改进的启发式相对属性约简，对样本数据进行属性约简，再通过聚类和 DT_SVM 相结合，提高了模型的过滤精度。经实验验证，算法具有较高的查全率、查准率，缩短了信息过滤的时间，本文只进行了网页信息过滤的仿真实验，后续工作可以考虑将算法应用于其他网络信息过滤领域。

## 参考文献


[1] Han Jiawei, Kamber M. Data Mining Concepts and Techniques[M]. 2nd ed. Beijing: China Machine Press, 2006.

[2] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 2000.

[3] Knerr S, Personnaz L, Dreyfus G. Single-layer Learning Revisited: Stepwise Procedure for Building and Training a Neural Network[M]. New York: Springer-Verlag, 1990: 13.

[4] Bottou L, Cortes C, Denker J S. Comparison of Classifier Method: A Case Study in Handwrittern Digitrecognition[C]//Proc. of the 12th International Conference on Pattern Recognition. [S. l.]: IEEE Press, 1994: 77-87.

[5] 孟媛媛, 刘希玉. 一种新的基于二叉树的 SVM 多分类方法[J]. 计算机应用, 2005, 25(11): 195-196, 199.


（上接第 169 页）

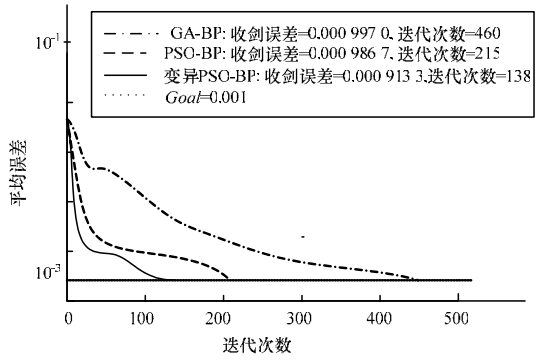


**图 3　变异 PSO, PSO, GA 算法训练 BP 网络性能曲线**

基于 BP 网络的各种算法的仿真结果比较如表 1 所示。

**表 1　仿真结果比较**

| 算法 | 训练时间/s | 迭代次数 | 收敛误差 | 相对误差/(%) | 平均准确率/(%) |
|---|---|---|---|---|---|
| BP | 1 496 | 89 582 | 0.000 999 9 | 0.351 | 93.5 |
| GA | 235 | 460 | 0.000 997 0 | 0.350 | 95.5 |
| PSO | 204 | 215 | 0.000 986 7 | 0.346 | 95.0 |
| 变异 PSO | 150 | 138 | 0.000 913 3 | 0.320 | 96.5 |

从图 2、图 3 及表 1 可知，在相同的环境下，变异 PSO 算法在用时、迭代次数、相对误差等方面都相对较小，且有较高的测试精确率，所以基于粒子群优化的 BP 网络在检测网络入侵数据时，在收敛速度和算法的性能上都优越于 PSO 算法、GA 算法和传统 BP 算法。

## 5　结束语

通过以上分析表明，变异粒子群算法在 BP 网络学习中，相对于 PSO 算法、GA 算法和传统 BP 算法，不仅速度快，算法简单，而且测试精确率高。特别是最后网络入侵数据的仿真实验的比较，进一步证明了基于变异 PSO 算法的 BP 网络学习算法的优越性和实用性。

## 参考文献


[1] 周开利, 康耀红. 神经网络模型及其 MATLAB 仿真程序设计[M]. 北京: 清华大学出版社, 2006.

[2] 余炳辉, 袁晓辉. 随机摄动粒子群优化算法[J]. 计算机工程, 2006, 32(12): 189-190

[3] 张丽平. 粒子群优化算法的理论与实践[D]. 杭州: 浙江大学, 2005.