基于 DTRS 模型的邮件过滤方法研究

赵春生[1] 冯 林[2] 何志勇[1]

[1](四川理工学院计算机学院 四川 自贡 643000)

[2](四川师范大学计算机科学学院 四川 成都 610101)

摘 要 邮件过滤是当前网络信息安全研究的一个热点。针对传统邮件过滤方法容错能力方面的不足,提出一种基于决策粗糙集模型 DTRS(Decision-Theoretic Rough Set)的邮件过滤方法。通过将无法明确判断的邮件用 DTRS 的边界域进行刻画,实现正常邮件、垃圾邮件和可疑邮件的三枝决策,确保总体决策的完备性。仿真实验结果表明文中方法是有效的,并且在控制邮件误分类上具有优势。

关键词 粗糙集 决策粗糙集 朴素贝叶斯 邮件过滤

中图分类号 TP181 **文献标识码** A DOI: 10.3969/j.issn.1000-386x.2013.05.043

ON DTRS MODEL-BASED EMAIL FILTERING

Zhao Chunsheng[1] Feng Lin[2] He Zhiyong[1]

[1](*School of Computer Science, Sichuan University of Science and Engineering, Zigong 643000, Sichuan, China*)

[2](*College of Computer Science, Sichuan Normal University, Chengdu 610101, Sichuan, China*)

Abstract Email filtering is a research focus in regard to current networks information security. Aiming at the weak fault-tolerant ability of traditional email filtering methods, we propose an email filtering approach which is based on DTRS (decision-theoretic rough set). A three-way decision for normal, spam and suspicious emails is realised by depicting those emails incapable of explicit judging with boundary region of DTRS, and this ensures the completeness of overall decision. The results of simulation experiment show that this approach is effective and has the advantage in controlling the email misclassification.

Keywords Rough set DTRS Naive Bayes Email filtering

0 引 言

随着电子邮件的广泛使用,垃圾邮件越来越严重地侵袭着人们的生活和工作。垃圾邮件过滤成为信息安全领域研究的一个热点,引起了国内外众多学者的重视[1-4]。邮件过滤是一个典型的不确定性分类问题,粗糙集[5]作为处理不确定、不精确问题的重要数学工具,其在邮件过滤中的应用受到越来越多学者的关注。如李志君等提出了基于粗糙集的邮件分类模型[6],邓维斌等研究了基于粗糙集的两阶段邮件过滤方法[7]。但经典 Pawlak 粗糙集的核心思想是基于等价关系的已知概念粒化和上下近似集合对未知概念的逼近[8,9],只有完全包含于目标概念的等价类才能判定为确定属于决策概念,而对于包含度介于 0 与 1 之间的对象集合则划分到边界域。因此,基于 Pawlak 经典粗糙集的邮件过滤模型将邮件过滤视为正常邮件与垃圾邮件的二枝决策问题,对现实中存在噪声的邮件过滤而言显得过于严格,缺乏对分类的容错能力,导致邮件误分类率较高。

针对 Pawlak 经典粗糙集模型容错能力的不足,Yao 将 Pawlak 粗糙集模型中的代数包含关系拓展为可调的概率包含关系,使得单个决策类的正域、边界域和负域由相应的概率包含范围确定,并基于最小风险 Bayes 决策方法确定概念边界,提出了决策粗糙集模型 DTRS[10,11]。DTRS 模型中单个决策类的负域不一定划分到互补的决策类的正域或边界域中,从而总体决策类的负域可以为非空,整个论域可以表示为决策类的正域、边界域和负域的并集,实现了总体决策类刻画的完备性。

本文将 DTRS 理论应用到邮件过滤中,通过 DTRS 三枝决策语义[12]和概念容错分析方法,改变传统非此即彼的二枝决策模式,将信息不足无法明确判断的邮件用 DTRS 模型的边界域进行刻画,允许在信息不足背景下的中间决策,从而实现垃圾邮件、正常邮件和可疑邮件的完整分类,有效减少邮件的误分类数量。

1 DTRS 基本概念

定义 1[13] 一个决策信息系统 S 是一个四元组: $S = <U, R, V, f>$。其中 U 是对象集合,也称为论域; $R = C \cup D$ 是属性集合,子集 C 和 D 分别称为条件属性和决策属性, $C \cap D = \phi$, $D \neq \phi$; V 是属性值的集合; $f: U \times R \rightarrow V$ 是一个信息函数,它指定 U 中每个对象 x 的属性值。

收稿日期: 2012-07-07。四川省科技厅项目(2009JY0134);自贡市科技局项目(2011G04)。赵春生,讲师,主研领域: 粗集理论及应用。冯林,副教授。何志勇,教授。

定义 2[13]　给定决策信息系统 $S=<U,R,V,f>$,对每个子集 $X\subseteq U$ 和不分明关系 B,X 的上近似集和下近似集定义如下:

$$B^{-}(X)=\{x\mid x\in U\wedge [x]_B\cap X\neq\phi\} \quad (1)$$

$$B_{-}(X)=\{x\mid (x\in U\wedge [x]_B\subseteq X)\} \quad (2)$$

定义 3[13]　U 为论域,集合 $POS(X)=B_{-}(X)$ 称为 X 的正域; 集合 $BND(X)=B^{-}(X)-B_{-}(X)$ 称为 X 的边界域; 集合 $NEG(X)=U-B^{-}(X)$ 称为 X 的负域。

在 DTRS 模型中,利用 2 个状态集和 3 个行动集来描述决策过程。状态集 $\Omega=(X,\neg X)$ 分别表示某事件属于 X 和不属于 X。行动集 $A=\{a_P,a_N,a_B\}$ 分别表示判定当前对象 x 属于 $POS(X)$,$NEG(X)$ 和 $BND(X)$ 三种动作。用 $\lambda_{PP},\lambda_{BP},\lambda_{NP}$ 分别表示当 x 属于 X 时采取行动 a_P,a_B 和 a_N 的损失; 用 $\lambda_{PN},\lambda_{BN}$ 和 λ_{NN} 分别表示当 x 不属于 X 时采取行动 a_P,a_B 和 a_N 的损失。根据最小风险 Bayes 决策方法,采取 a_P,a_B 和 a_N3 种行动的期望风险可分别表示为:

$$R(a_P\mid [x]_R)=\lambda_{PP}Pr(X\mid [x]_R)+\lambda_{PN}Pr(\neg X\mid [x]_R) \quad (3)$$

$$R(a_B\mid [x]_R)=\lambda_{BP}Pr(X\mid [x]_R)+\lambda_{BN}Pr(\neg X\mid [x]_R) \quad (4)$$

$$R(a_N\mid [x]_R)=\lambda_{PN}Pr(X\mid [x]_R)+\lambda_{NN}Pr(\neg X\mid [x]_R) \quad (5)$$

其中 $[x]_R$ 是用等价类形式表示的特征描述。

根据最小风险 Bayes 决策原则,可以得到如下形式的决策规则:

(P) 若 $R(a_P\mid [x]_R)\leqslant R(a_B\mid [x]_R)$ 且 $R(a_P\mid [x]_R)\leqslant R(a_N\mid [x]_R)$,则 $x\in POS(X)$

(B) 若 $R(a_B\mid [x]_R)\leqslant R(a_P\mid [x]_R)$ 且 $R(a_B\mid [x]_R)\leqslant R(a_N\mid [x]_R)$,则 $x\in BND(X)$

(N) 若 $R(a_N\mid [x]_R)\leqslant R(a_P\mid [x]_R)$ 且 $R(a_N\mid [x]_R)\leqslant R(a_B\mid [x]_R)$,则 $x\in NEG(X)$

由状态集 X 和 $\neg X$ 的互补关系可得 $P(X\mid [x]_R)+P(\neg X\mid [x]_R)=1$,同时,正确分类决策的损失值应不大于待定分类决策的损失值,小于错误分类决策的损失值,待分类决策的损失值应小于错误分类决策的损失值。因此,决策损失值的大小关系为 $\lambda_{PP}\leqslant\lambda_{BP}<\lambda_{NP}$,$\lambda_{NN}\leqslant\lambda_{BN}<\lambda_{PN}$。将期望风险的计算式带入决策规则(P),(B),(N)中,可得到如下含阈值形式的决策规则[13],即:

(P1) 若 $P(X\mid [x]_R)\geqslant\gamma$ 且 $P(X\mid [x]_R)\geqslant\alpha$,则 $x\in POS(X)$

(B1) 若 $P(X\mid [x]_R)\geqslant\beta$ 且 $P(X\mid [x]_R)\leqslant\alpha$,则 $x\in BND(X)$

(N1) 若 $P(X\mid [x]_R)\leqslant\beta$ 且 $P(X\mid [x]_R)\leqslant\gamma$,则 $x\in NEG(X)$

阈值 α,β,γ 分别由下列三式计算:

$$\alpha=\frac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})} \quad (6)$$

$$\beta=\frac{\lambda_{BN}-\lambda_{NN}}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})} \quad (7)$$

$$\gamma=\frac{\lambda_{PN}-\lambda_{NN}}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})} \quad (8)$$

由规则(B1)可得 $\alpha>\beta$,则:

$$\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}<\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$$

根据不等式 $\frac{b}{a}>\frac{d}{c}\Rightarrow\frac{b}{a}>\frac{b+d}{a+c}>\frac{d}{c}(a,b,c,d>0)$ 则 $\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}<\frac{\lambda_{NP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{NN}}<\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$,从而得到关系 $0\leqslant\beta<\gamma<\alpha\leqslant 1$。

三个决策规则可重写为:

(P2) 若 $Pr(X\mid [x]_R)\geqslant\alpha$,则 $x\in POS(X)$

(B2) 若 $\beta<Pr(X\mid [x]_R)<\alpha$,则 $x\in BND(X)$

(N2) 若 $Pr(X\mid [x]_R)\leqslant\beta$,则 $x\in NEG(X)$

DTRS 模型将传统正域、负域的二枝决策语义拓展为正域、边界域和负域的三枝决策语义,并根据最小风险 Bayes 决策给出了决策阈值的计算方法,为不确定决策提供了充分的理论依据。

2　基于 DTRS 的邮件过滤模型

按照 DTRS 相关理论,邮件过滤可看作一个决策系统 $S=<U,R=C\cup D,V,f>$,其中 U 是邮件的集合,C 为邮件特征集合,D 是邮件的状态特征,$D=\{0,1\}$,1 表示垃圾邮件,0 表示正常邮件。用行动集 $A=\{a_P,a_B,a_N\}$ 分别表示将邮件分类为垃圾邮件、可疑邮件和正常邮件三种判断。用 $\lambda_{PP},\lambda_{BP},\lambda_{NP}$ 分别表示当邮件为垃圾邮件时,采取 a_P,a_B 和 a_N 行动下的损失。用 $\lambda_{PN},\lambda_{BN}$ 和 λ_{NN} 表示当邮件为正常邮件时,采取 a_P,a_B 和 a_N 行动下的损失。根据式(1)、式(2)计算出阈值 α 和 β,根据 *Bayes* 公式从决策系统计算得出条件概率 $Pr(X\mid [x]_R)$,按照决策规则(P2),(B2),(N2)将邮件分类为垃圾邮件、可疑邮件和正常邮件。下面,我们给出基于 DTRS 模型的邮件过滤方法。

算法 1　基于 DTRS 模型的邮件过滤算法。

输入: 邮件决策系统 $S=<U,C\cup D,V,f>$ 和行动损失因子 $\lambda_{PP},\lambda_{BP},\lambda_{NP},\lambda_{PN},\lambda_{BN},\lambda_{NN}$。

输出: 邮件分类(垃圾邮件集 *Spam*、可疑邮件集 *Doubt* 和正常邮件集 *Normal*)。

步骤 1　预处理: 将邮件样本进行补齐并离散化,然后划分为训练集 *TE* 和测试集 *TR*。

步骤 2　初始化: 初始化邮件分类集 *Spam*、*Doubt* 和 *Normal* 均为 ϕ。

步骤 3　根据式(6)、式(7)计算阈值 α 和 β。

步骤 4　如果是测试任务,转步骤 6,如果是训练任务,则转步骤 5。

步骤 5　先验概率学习。

步骤 5.1　计算先验概率 $Pr(X)=\frac{|X|}{|TR|}$ 和 $Pr(\neg X)=\frac{|\neg X|}{|TR|}$,$|\cdot|$ 表示集合中元素的基数。

步骤 5.2　计算 $\forall c\in C$ 的先验概率 $Pr(c_{ik}\mid X)$,即在分类 X 中条件属性 c_i 的第 k 种取值的概率。

步骤 5.3　计算 $\forall c\in C$ 的先验概率 $Pr(c_{ik}\mid\neg X)$,即在分类 $\neg X$ 中条件属性 c_i 的第 k 种取值的概率

步骤 6　邮件过滤。

步骤 6.1　计算条件概率 $Pr(X\mid [x]_R)$。

$$Pr([x]_R)=Pr(v_1,v_2,\cdots,v_m)=\prod_{i=1}^{m}Pr(v_i) \quad (9)$$

$$Pr([x]_R\mid X)=Pr(v_1,v_2,\cdots,v_m\mid X)=\prod_{i=1}^{m}Pr(v_i\mid X) \quad (10)$$

$$Pr(X\mid [x]_R)=\frac{Pr(X)\,Pr([x]_R\mid X)}{Pr([x]_R)} \quad (11)$$

步骤 6.2　比较 $Pr(X\mid [x]_R)$ 与阈值 α,β 之间的关系,根据规则(P2),(B2),(N2)将邮件归入 *Spam* 集合,*Doubt* 集合和 *Normal* 集合。

3 仿真实验

为了验证基于 DTRS 模型的邮件过滤方法的有效性，本文采用 UCI 机器学习数据资料库的 spambase[14] 数据集为实验数据。该数据集中共包含 4601 个样本邮件，其中垃圾邮件 1813 封，正常邮件 2788 封。每个邮件样本包含 57 个条件属性和 1 个决策属性，决策属性取值为 1 表示该样本为垃圾邮件，取值为 0 表示为正常邮件。在 JDK6.0 下使用 Java 语言进行 DTRS 邮件过滤算法实现。实验过程如下:

步骤 1 使用 Rosetta[15] 分析平台中的 Entropy/MDL 算法对样本数据进行离散化。

步骤 2 通过实验不同行动损失因子 λ 对邮件分类效果的影响，设定阈值 $\alpha = 0.85$，$\beta = 0.15$，调用 DTRS 邮件过滤算法进行 5 折交叉验证，获取平均结果如表 1 所示。

步骤 3 从离散化后的样本数据中，随机抽取垃圾邮件和正常邮件各 1/5(垃圾邮件 362 封，正常邮件 557 封) 构成测试集，其余邮件构成训练集，使用朴素贝叶斯 NB[16] 分类算法进行过滤，结果如表 2 所示。

步骤 4 在 Rosetta 分析平台上，使用遗传算法对离散后的数据进行属性约简，并转化为 LIBSVM[17] 的输入格式，经过线性归一化后，随机抽取垃圾邮件和正常邮件各 1/5(垃圾邮件 362 封，正常邮件 557 封) 构成测试集，其余邮件构成训练集，进行基于粗糙集属性约简的支持向量机 SVM[18] 分类，采用 RBF 为 SVM 核函数，通过交叉验证参数寻优确定参数 $g = 0.03125$，$c = 128$，实验结果如表 3 所示。

表 1 基于 DTRS 的邮件分类结果

	Total	Spam	Doubt	Normal
垃圾邮件	361	304	43	14
正常邮件	556	14	86	456

表 2 基于朴素贝叶斯的邮件分类结果

	Total	Spam	Normal
垃圾邮件	362	320	42
正常邮件	557	27	530

表 3 基于粗糙集属性约简的 SVM 邮件分类结果

	Total	Spam	Normal
垃圾邮件	362	294	68
正常邮件	557	24	533

为了衡量邮件过滤效果，本文引入正确率 *Precision*、召回率 *Recall* 和误分率 *Misclassification* 三个指标对结果进行评估。用 N_L 表示测试集中正常邮件总数，N_S 表示测试集中垃圾邮件总数，$n_{L\to L}$ 表示正确分检出的正常邮件数，$n_{L\to S}$ 表示被误判为垃圾邮件的正常邮件数，$n_{S\to S}$ 表示正确分检出的垃圾邮件数，$n_{S\to L}$ 表示被误判为正常邮件的垃圾邮件数。以垃圾邮件为例，有如下的计算公式:

$$Precision = \frac{n_{S\to S}}{n_{S\to S} + n_{L\to S}} \tag{12}$$

$$Recall = \frac{n_{S\to S}}{n_{S\to S} + n_{S\to L}} \tag{13}$$

$$Misclassification = \frac{n_{S\to L}}{N_S} \tag{14}$$

基于朴素贝叶斯(NB)、粗糙集属性约简的 SVM 分类(RS_SVM) 和 DTRS 的三种邮件分类效果对比如表 4 所示。

表 4 不同分类模型的比较结果

	垃圾邮件			正常邮件		
	P	*R*	*Mis*	*P*	*R*	*Mis*
NB	91.7	88.5	11.5	93.1	95.1	4.9
RS_SVM	92.4	81.2	18.8	88.7	95.7	4.3
DTRS	95.6	84.2	3.9	97.0	82.0	2.5

注: *P* 表示分类正确率 Precision; *R* 表示召回率 Recall; *Mis* 表示误分率 Misclassification; 数值为百分比。

通过对实验结果数据的分析比较，可以看出:

(1) 基于 DTRS 模型的邮件过滤方法在分类正确率上比基于朴素贝叶斯和基于粗糙集属性约简的 SVM 分类都要高。由于将部分邮件分类为可疑邮件，使其在召回率上略有降低。但整体而言该方法是有效的。

(2) 基于 DTRS 的邮件过滤模型在控制邮件误分类方面具有明显优势。邮件误分类，特别是将一封正常邮件分类为垃圾邮件，极可能给用户造成严重损失。作为对比的两种分类模型分别将 24 封和 27 封正常邮件分类为垃圾邮件，而在基于 DTRS 的邮件过滤中被误分类的正常邮件数下降到 15 封，误分类率约 2.5%，误分类数量减少约 42%，而垃圾邮件的误分类数量减少超过 66%。

4 结 语

对垃圾邮件进行有效过滤是机器学习和网络信息安全领域研究的热点。本文将 DTRS 理论应用于垃圾邮件过滤，通过将无法明确判断的邮件用 DTRS 的边界域进行刻画，实现了正常邮件、垃圾邮件和可疑邮件的三枝决策，弥补了传统二枝决策在容错方面的不足，确保了总体决策的完备性。分类阈值通过行动损失因子计算得出，而行动损失的大小可以由行为学实验或专家的意见给出，是主观的，条件概率从决策系统中计算得出，是客观的，基于 DTRS 模型的邮件过滤方法体现了邮件过滤中主客观结合、人机合一的需求。仿真实验结果表明了该方法的有效性，并且与传统方法相比在控制邮件误分类方面效果明显，邮件误分类数量减少超过 40%。

参 考 文 献

[1] Thiago S Guzella, Walmir M C. A review of machine learning approaches to spam filtering [J]. Expert System with Applications, 2009, 36(7): 10206 - 10222.

[2] Lai C C. An empirical study of three machine learning methods for spam filtering [J]. Knowledge-Based System, 2007, 20(3): 249-254.

[3] 郑伟，沈文，张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究 [J]. 西北工业大学学报，2010，28(4)：622 - 627.

[4] 赵静，刘培玉，许明英. 邮件过滤中特征选择方法的性能评价与分析 [J]. 计算机应用研究，2012，29(2)：693 - 697.

[5] Pawlak Z. Rough set [J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341 - 356.

[6] 李志君，王国胤，吴渝. 基于 Rough Set 的电子邮件分类系统 [J]. 计算机科学，2004，31(3)：58 - 66.

[7] 邓维斌，洪智勇. 基于粗糙集的两阶段邮件过滤方法 [J]. 计算机应用，2010，30(8)：2006 - 2009.

(下转第 235 页)

边缘提取的是方法是计算图像的一阶或二阶导数，寻找灰度值突变的区域。OpenCV(EmguCV) 支持多种求导方式，有 Sobel 算子，Laplacian 算子和 Canny 算子等，前两种微分算子实现起来相对简单，适合一般情况下的边缘检测，分别具有主体与背景不能分离和对阶跃背景无法检测等缺点。Canny 算子的实现较为复杂，但其效果最为理想，本文选用 EmguCV 封装好的 cvCanny( ) 方法来对图像进行边缘提取:

```
CvInvoke. cvCanny( img, img, 100, 100, 5) ;
```

之后对其进行膨胀操作，滤去噪点:

```
CvInvoke. cvDilate( img, img, element, 1) ;
```

4) 计算相似度　相似度是图形匹配的结果表示，代表了图像与既定模板之间的相似关系，OpenCv(EmguCV) 为图形轮廓间的匹配提供了几种不同的方法: 轮廓树匹配、成对几何直方图匹配和 Hu 矩匹配等，前两种方法顾名思义，是用树的形式和直方图对比来进行轮廓的匹配; 而第三种方法是通过计算轮廓的 Hu 矩来计算轮廓间的相似度，Hu 矩对包括缩放、旋转和镜像映射在内的变化具有不变性，我们选用这种方法来计算图形与模板间的相似度。这里调用 CvInvoke. cvMatchShapes( ) 方法对输入图像和模板进行形状匹配，返回值是一个 double 类型的数，代表了两幅图的相似程度，返回值越接近 0 则说明两幅图的相似度越高，通过如下计算公式定义相应的相似度百分比，来表征图形与模板的相似度:

```
similaraty = Math. Ceiling( 100 *  Math. Pow( 0. 00067, mat_result) ) ;
                                                    //mat_result 为匹配返回值
```

5) 图形识别　通过所画图形与各个模板的匹配，得到一组相似度的数值，这些数值表示了图形与模板间的相似程度，对这组数值进行简单的排序，理论上将相似度最高的模板定义为最终匹配结果。

2.3.4　匹配实验与结果

图 2 所示为用户通过触摸显示屏绘制的动物“马”的原始图形，将此图形经过以上几个步骤的处理，同下面的模板进行相似度比对，模板如图 3 所示。

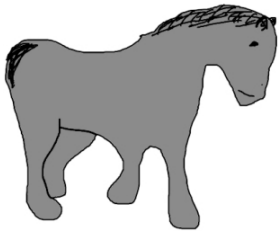

图 2　原始输入图形

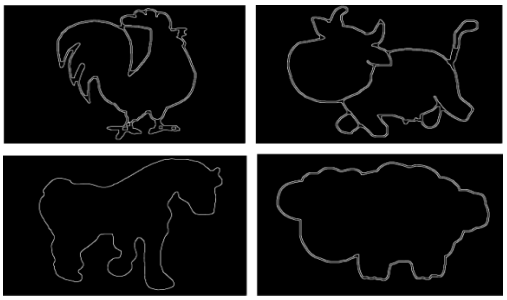

图 3　图形模板

经过比对，我们分别得到了以上图形模板的相似度值，分别为鸡: 26%，牛: 36%，马: 91%，羊: 53%，最终“马”匹配成功。

综上所述，实验结果比较理想。在软件实际使用过程中，匹配成功率也比较高，达到了 90% 以上。可见，通过在 VS2010 中为 WPF 配置加入 EmguCV，将 OpenCV 强大的图形处理功能加入到. net 平台中，我们并不需要将大量的时间花费在具体算法的实现和调试上，只需注重任务本身的要求，对症下药，设计合理的处理框架，选取合适的方法来调用，开发效率之高可见一斑。

3　结　语

本文简单介绍了新一代客户端设计系统 WPF 的特点与优势，并以一款游戏软件的设计为背景，说明了如何在 WPF 中配置 EmguCV 的过程，详细叙述了运用基于 C#语言的 EmguCV 图像处理库来完成具体的图像处理任务，主要包括图像的滤波，特征提取，相似度计算等内容，最后通过实验来验证处理的效果。

本文介绍的图形处理方法在 WPF 设计框架下具有广泛的用途，而在普遍使用面向对象语言的今天，基于 C#的 EmguCV 想必也能够延续 OpenCV 的辉煌。

参　考　文　献

[1] 黎松，平西建，丁益洪. 开放源代码的计算机视觉类库 OpenCV 的应用 [J]. 计算机应用与软件，2005，22(8) : 134 - 136.

[2] 黄建岗. 浅谈 WPF 设计模式 [J]. 中小企业管理与科技(上旬刊) ，2010(10) .

[3] 李学勇，路长厚，李国平. 基于二阶梯度图的 Canny 检测边缘修补方法 [J]. 光电子激光，2007，18(3) : 377 - 380.

[4] 秦小文，温志芳，乔维维. 基于 OpenCV 的图像处理 [J]. 电子测试，2011(7) .

[5] 方玫. OpenCV 技术在数字图像处理中的应用 [J]. 北京教育学院学报: 自然科学版，2011，6(1) : 7 - 11.

(上接第 154 页)

[8] 冯林，王国胤，李天瑞. 连续值属性决策表中的知识获取方法 [J]. 电子学报，2009，37(11) : 2432 - 2438.

[9] 冯林，原永乐，苟仕蓉，等. 一种实域粗糙集模型及属性约简方法 [J]. 控制与决策，2012，27(4) : 562 - 566.

[10] Yao Y Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences，2010，180(3) : 341 - 353.

[11] Yao Y Y，Zhao Y. Attribute reduction in decision-theoretic rough set models [J]. Information Sciences，2008，178(1) : 3356 - 3373.

[12] Yao Y Y. Three-way decision: an interpretation of rules in rough set theory [C] //The 4th International Conference on Rough Sets and Knowledge Technology，2009.

[13] 曾黄麟. 粗糙集理论及其应用 [M]. 重庆: 重庆大学出版社，1996.

[14] Frank A，Asuncion A. UCI Machine Learning Repository [EB/OL]. Irvine，CA: University of California，School of Information and Computer Science. (2010) . http: //archive. ics. uci. edu/ml.

[15] Øhrn A，Komorowski J. Rosetta: A rough Set Toolkit for Analysis of Data [C] //Third International Joint Conference on Information Sciences，Fifth International Workshop on Rough Sets and Soft Computing，Durham，NC，USA，RSSC'97(3) : 403 - 407.

[16] Nir Friedman，Dan Gelger. Bayesian Network Classifiers [J]. Machine Learning，1997，29: 131 - 163.

[17] Chang Chih-Chung，Lin Chih-Jen. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology，2011，2(3) : 1 - 27.

[18] Nelllo Cristianini，John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods [M]. England: Cambridge University Press，2000.