DOI:10.3969/j.issn.1673-4785.201307010
网络出版地址:http://www.cnki.net/kcms/doi/CNKI:23-1538/TP.20131105.1201.004.html

不完备信息系统中测试代价敏感的可变精度分类粗糙集

鞠恒荣[1],马兴斌[1],杨习贝[1,2],祁云嵩[1],杨静宇[2]
(1.江苏科技大学 计算机科学与工程学院,江苏 镇江 212003; 2. 南京理工大学 计算机科学与技术学院,江苏 南京 210094)

摘　要:在不完备信息系统中,可变精度分类关系是限制容差关系的改进形式,但其并未考虑数据集中属性的测试代价。为解决这一问题,提出了基于测试代价敏感的可变精度分类粗糙集模型。进一步地,通过分析传统启发式算法没有考虑测试代价以及回溯算法的时间消耗等因素,提出一种新的属性重要度测量,并在此基础上设计了一种新的启发式算法。通过实验对比分析,说明了新提出算法的有效性。

关键词:属性约简;不完备信息系统;测试代价敏感;变精度分类粗糙集
中图分类号: TP18　**文献标志码**:A　**文章编号**:1673-4785(2014)02-0219-05

中文引用格式:鞠恒荣,马兴斌,杨习贝,等. 不完备信息系统中测试代价敏感的可变精度分类粗糙集[J]. 智能系统学报, 2014, 9(2): 219-223.
英文引用格式:JU Hengrong, MA Xingbin, YANG Xibei, et al. Test-cost-sensitive based variable precision classification rough set in incomplete information system[J]. CAAI Transactions on Intelligent Systems, 2014, 9(2): 219-223.

Test-cost-sensitive based variable precision classification rough set in incomplete information system

JU Hengrong[1], MA Xingbin[1], YANG Xibei[1,2], QI Yunsong[1], YANG Jingyu[2]
(1. School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 2. School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: In an incomplete information system, the precision-variable classification relation is an improvement of the limited tolerance relation. However, the test costs of the data concentration attributes are not taken into account. To solve this problem, a test-cost-sensitive-based precision-variable precision classification rough set is proposed. Furthermore, the traditional heuristic algorithm does not take the importance of the test costs of the attributes into account, and backtracking algorithm is very time-consuming. Therefore, not only was a new importance of the attribute proposed, but a new heuristic algorithm was also presented for obtaining reduction with minor test costs. The experimental results show the effectiveness of the new algorithm by comparing it with the other algorithms.

Keywords: attribute reduction; incomplete information system; test cost sensitive; variable precision classification rough set

作为一种处理不精确、不确定性问题的数学工具,粗糙集理论[1](rough set)由波兰学者 Pawlak 提出后便受到了广泛关注[2-4]。然而由于数据测量的误差、数据获取的限制等原因,导致了所面临的信息系统往往是不完备的。为处理这类问题,王国胤[5]提出了限制容差关系。进一步,杨习贝[6]提出了一种新的基于可变精度分类的拓展粗糙集模型,对限制容差关系进行了改进。然而,在实际工程应用中,数据的获取是需要付出一些成本或代价的,称其为测试代价。针对该问题,Min 等[7-11]率先将测试代价引入到

收稿日期:2013-07-05.　**网络出版日期**:2013-11-05.
基金项目:国家自然科学基金资助项目(61100116, 61203024);江苏省自然科学基金资助项目(BK2011492, BK2012700);江苏省高校自然科学基金资助项目(11KJB520004, 13KJB520003);高维信息智能感知与系统教育部重点实验室(南京理工大学)基金资助项目(30920130122005);江苏省普通高校研究生科研创新计划项目资助项目(CXLX13_707).
通信作者:杨习贝. E-mail:yangxibei@hotmail.com.

粗糙集的约简问题中,终究未能将测试代价引入到不完备信息系统环境下粗糙集本身的近似模型上。

1 基本概念

形式化地, 信息系统可表示为四元组 $I_S=\langle U, A_T, V, f\rangle$, 其中 $U=\{x_1, x_2, \cdots, x_m\}$ 为研究对象的有限集合, 称为论域; $A_T=\{a_1, a_2, \cdots, a_n\}$ 为描述对象的全部属性所组成的集合; $V=\cup_{a\in A_T}V_a$ 为属性集合 A_T 的值域, 其中 V_a 为属性 a 的值域; $f: U\times A_T\rightarrow V$ 为信息函数, 表示对每一个 $x\in U$, $a\in A_T$, $f(x, a)\in V_a$。特别地, 当信息系统中属性集 $A=A_T\cup D$ 且 $A_T\cap D=\varnothing$(其中 A_T 为条件属性集合, D 为决策属性集合)时, 信息系统也被称为决策系统。

定义 1[6] 设 S 为不完备信息系统, $\forall A\subseteq A_T$, 由 A 决定的可变精度分类关系记为 V_A^{α},且

$$V_A^{\alpha}=\{(x, y)\in U^2: \forall a\in P_A(x)\cap P_A(y), f(x, a)=f(y, a)\wedge\frac{|P_A(x)\cap P_A(y)|}{|P_A(x)|}\geqslant\alpha\}\cup \boldsymbol{I}_U \tag{1}$$

式中: $P_A(x)=\{a\in A: f(x, a)$已知$\}$, $\alpha\in[0, 1]$, $|X|$ 表示集合 X 的基数, $\boldsymbol{I}_U$ 为恒等关系且 $\boldsymbol{I}_U=\{(x, x): x\in U\}$。

定义 2[6] 设 S 为不完备信息系统, $\forall A\subseteq A_T$, 对于任意的 $X\subseteq U$, X 基于可变精度分类关系 V_A^{α} 的下、上近似集合分别记为 $\underline{A}_V^{\alpha}(X)$ 和 $\bar{A}_V^{\alpha}(X)$:

$$\underline{A}_V^{\alpha}(X)=\{x\in U: V_A^{\alpha}(x)\subseteq X\}$$

$$\bar{A}_V^{\alpha}(X)=\{x\in U: V_A^{\alpha}(x)\cap X\neq\varnothing\}$$

式中: $V_A^{\alpha}(x)=\{y\in U: (x, y)\in V_A^{\alpha}\}$ 表示对象 x 的可变精度容差类。

2 测试代价与可变精度分类粗糙集

不完备信息系统环境下的粗糙集模型未考虑数据的代价问题, Min 等[8]将测试代价引入到信息系统中, 具体的描述见定义 3。

定义 3[8] 一个测试代价敏感的不完备决策系统 TCSIIDS 为一个五元组 $\langle U, A_T\cup D, V, f, c^*\rangle$, U, $A_T\cup D$, V 和 f 的含义与第 1 节所示相同, $c^*: A_T=\{a_1, a_2, \cdots, a_n\}\rightarrow\mathbf{R}^+\cup\{0\}$ 为测试代价函数($\mathbf{R}^+$表示正整数集), 即 $c^*(A_T)=\sum_{i=1}^{|A_T|}c^*(\{a_i\})$, 其中 $c^*(\{a_i\})$ 表示单个属性 a_i 的测试代价。本文假设所有属性的测试代价均为正整数, 即 $c^*(\{a_i\})>0$, $a_i\in A_T$。

定义 4 令 TCSIIDS 为测试代价敏感不完备决策系统, 其中 $A\subseteq A_T$, $\forall x, y\in U$, $\forall a\in A$, 定义特征函数如下所示:

$$\delta_a(x)=\begin{cases}1: P_a(x)\neq\varnothing\\ 0: P_a(x)=\varnothing\end{cases}$$

$$\varphi_a(x, y)=\begin{cases}1: P_a(x)\cap P_a(y)\neq\varnothing\\ 0: P_a(x)\cap P_a(y)=\varnothing\end{cases}$$

式中: $P_a(x)=\{a\in A: f(x, a)$已知$\}$。

定义 5 设 TCSIIDS 为测试代价敏感不完备决策系统, 其中 $A\subseteq A_T$, 由 A 决定的可变精度分类关系记为 V_A^{α, c^*} 且

$$V_A^{\alpha, c^*}=\{(x, y)\in U^2: \forall a\in P_a(x)\cap P_a(y), f(x, a)=f(y, a)\wedge\frac{\sum_{a\in A}(c^*(a)\times\varphi_a(x, y))}{\sum_{a\in A}(c^*(a)\times\delta_a(x))}\geqslant\alpha\}\cup\boldsymbol{I}_U$$

其中: $\alpha\in[0, 1]$。

定义 6 设 TCSIIDS 为测试代价敏感不完备决策系统, $A\subseteq A_T$, 对于 $\forall X\subseteq U$, X 基于可变精度分类关系 V_A^{α, c^*} 的下、上近似集分别记为 $\underline{A}_V^{\alpha, c^*}(X)$ 和 $\bar{A}_V^{\alpha, c^*}(X)$

$$\begin{cases}\underline{A}_V^{\alpha, c^*}(X)=\{x\in U: V_A^{\alpha, c^*}(x)\subseteq X\}\\ \bar{A}_V^{\alpha, c^*}(X)=\{x\in U: V_A^{\alpha, c^*}(x)\cap X\neq\varnothing\}\end{cases}$$

式中: $V_A^{\alpha, c^*}(x)=\{y\in U: (x, y)\in V_A^{\alpha, c^*}\}$ 表示在测试代价敏感不完备决策系统中对象 x 的可变精度容差类。

3 属性约简

属性约简是粗糙理论的主要研究内容之一。然而, 寻找决策表的最小约简已被证明是一个 NP-hard 问题, 在处理大规模数据时计算时间代价很大, 针对这一问题, 许多学者提出了许多高效的约简算法, 启发式搜索方法就是其中的一个典型代表。

定义 7 设 TCSIIDS 为测试代价敏感不完备决策系统, $\alpha\in[0, 1]$, $U/\mathrm{IND}(D)=\{X_1, \ldots, X_n\}$ 表示根据决策属性得到的所有决策类的合集, 那么 $U/\mathrm{IND}(D)$ 的近似质量可定义为

$$\gamma(A, \alpha, D)=|\cup\{\underline{A}_V^{\alpha, c^*}(X_j): 1\leqslant j\leqslant m\}|/|U|$$

定理 1 令 TCSIIDS 为测试代价敏感不完备决策系统, $\forall A\subseteq AT$, 若 $0<\alpha_1<\alpha_2<1$, 则有

$$\gamma(A, \alpha_1, D)\leqslant\gamma(A, \alpha_2, D)$$

证明 根据定义 6, 定理 1 易证。

定义 8 令 TCSIIDS 为测试代价敏感不完备决策系统, $\alpha\in[0, 1]$, $A\subseteq AT$ 为一个约简当且仅当 $\gamma(A, \alpha, D)=\gamma(AT, \alpha, D)$ 且 $\forall B\subset A$, $\gamma(B, \alpha, D)\neq\gamma(A, \alpha, D)$。

令 TCSIIDS 为测试代价敏感决策系统, $\alpha \in [0, 1]$, $\forall A \subseteq AT$, $\forall a_i \in A$, a_i的重要度为

$$\mathrm{LSig}_{\mathrm{in}}(a_i, A, D) = \gamma(A, \alpha, D) - \gamma(A - a_i, \alpha, D)$$

可以看出, $\mathrm{LSig}_{\mathrm{in}}(a_i, B, D)$反映了 a_i从当前条件属性集 A 中删除后近似质量的变化, 相应地也可定义

$$\mathrm{LSig}_{\mathrm{out}}(a_i, A, D) = \gamma(A \cup a_i, \alpha, D) - \gamma(A, \alpha, D)$$

式中:$a_i \in A_T - A$, $\mathrm{LSig}_{\mathrm{out}}(a_i, A, D)$用以度量向属性集 A 增加属性 a_i后近似质量的变化。根据上述属性的重要度可以设计启发式属性约简算法。Min 等在文献[11]中设计了传统的启发式算法(记为算法 1)。其算法复杂度为 $O(|A_T||U| + \sum_{i=1}^{|A_T|} |U|(|A_T| - i + 1))$。

Min 等从获取约简的测试代价最小出发设计出新的约简算法, 即回溯算法(记为算法 2)。其算法复杂度为 $O(2^{|A_T|} - 1)$。详细算法见文献[11]。

3.1 考虑属性测试代价的启发式算法

由上文可知回溯算法的时间复杂度为 $O(2^{|A_T|} - 1)$。考虑到现实生活中存在着大量高维属性的数据, 这样一种机制将会大大制约属性约简的时间。为解决这一问题, 本文依然从启发式算法的角度出发, 将属性的测试代价考虑到属性重要度定义中。为此, 给出如下的属性融合重要度定义。

$$\mathrm{TCSLSig}_{\mathrm{in}}(a_i, A_T, D) = 0.1 \times \mathrm{LSig}_{\mathrm{in}}(a_i, A_T, D) + 0.9 \times (c^*(a_i))^{\theta}$$

$$\mathrm{TCSLSig}_{\mathrm{out}}(a_i, A_T, D) = 0.1 \times \mathrm{LSig}_{\mathrm{out}}(a_i, A_T, D) + 0.9 \times (c^*(a_i))^{\theta}$$

式中:$\theta \leqslant 0$。可以看出, $\mathrm{TCSLSig}_{\mathrm{in}}(a_i, A_T, D)$是 $\mathrm{LSig}_{\mathrm{in}}(a_i, A_T, D)$与$(c^*(a_i))^{\theta}$之和。当 $\theta = 0$ 时, 无论属性的测试代价取何值, 都有 $\mathrm{TCSLSig}_{\mathrm{in}}(a_i, A_T, D) = 0.1 \times \mathrm{LSig}_{\mathrm{in}}(a_i, A_T, D) + 1$, 这说明 $\mathrm{TCSLSig}_{\mathrm{in}}(a_i, A_T, D)$的大小与属性测试代价的大小无关, 仅与 $\mathrm{LSig}_{\mathrm{in}}(a_i, A_T, D)$的值有关。随着 θ 值的不断减小, $(c^*(a_i))^{\theta}$的值也不断减小(本文随机产生测试代价在[10, 100]), 此时$(c^*(a_i))^{\theta}$在 $\mathrm{TCSLSig}_{\mathrm{in}}(a_i, A_T, D)$中所占的比重也越来越小, 表明属性测试代价在属性的重要度中的影响作用越来越小。根据新的属性融合重要度, 不难设计出综合考虑了测试代价的启发式算法, 具体算法流程见算法 1。

算法 1　基于测试代价敏感不完备决策系统 TCSIIDS 综合考虑测试代价的启发式算法

输入:测试代价敏感不完备决策系统 TCSIIDS, α;

输出:约简 red 及约简的测试代价 $c^*(\mathrm{red})$。

1) 计算 $\gamma(A_T, \alpha, D)$; 令 $\theta = 0$, $c^*(\mathrm{red}) = c^*(A_T)$;

2) $\mathrm{red} \leftarrow \varnothing$;

3) $\forall a_i \in A_T$, 计算属性 a_i的重要度 $\mathrm{TCSLSig}_{\mathrm{in}}(a_i, A_T, D)$;

4) 若 $\mathrm{TCSLSig}_{\mathrm{in}}(a_j, A_T, D) = \max\{\mathrm{TCSLSig}_{\mathrm{in}}(a_i, A_T, D): a_i \in A_T\}$, 则 $\mathrm{red} \leftarrow a_j$,计算 $\gamma(\mathrm{red}, \alpha, D)$;

5) 若 $\gamma(\mathrm{red}, \alpha, D) \neq \gamma(A_T, \alpha, D)$, 则重复以下循环, 否则转 6);

①$\forall a_i \in A_T - \mathrm{red}$,计算 $\mathrm{TCSLSig}_{\mathrm{out}}(a_i, \mathrm{red}, D)$;

②若 $TCSLSig_{out}(a_i, B, D) = \max\{\mathrm{TCSLSig}_{\mathrm{out}}(a_i, B, D): a_i \in A_T - \mathrm{red}\}$,则 $\mathrm{red} \leftarrow a_i$;

6) $\forall a_i \in \mathrm{red}$, 若 $\gamma(\mathrm{red} - a_i, \alpha, D) = \gamma(A_T, \alpha, D)$, 则 $\mathrm{red} = \mathrm{red} - a_i$, $\mathrm{tmp} = c^*(\mathrm{red})$;

7) $c^*(\mathrm{red}) = \min\{c^*(\mathrm{red}), \mathrm{tmp}\}$;

8) 若 θ 大于给定阈值, 则 $\theta = \theta - \delta$(此处 δ 为步长,$\delta > 0$)且重复 2)~7),否则转 9);

9) 输出 red 及 $c^*(\mathrm{red})$。

3.2　实验分析

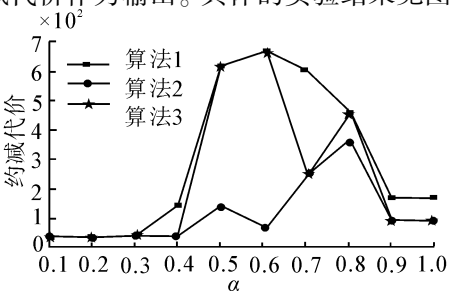
本节将通过实验, 对比算法 1、算法 2 和算法 3。表 1 列出了实验中使用的 4 组测试数据的基本信息, 所有数据集均下载于 UCI 数据集。由于 UCI 数据集中的大部分数据不含有测试代价, 所以在本组实验中为每个数据集的属性随机增加了取值在[10, 100]之间的测试代价。

表 1　**实验数据基本信息**

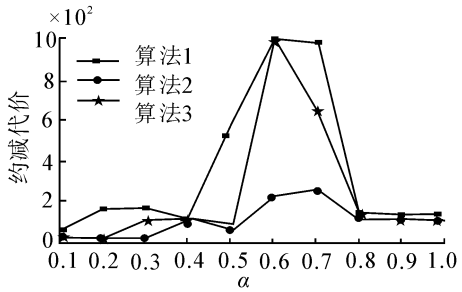
Table 1　**Data sets descriptions**

Data ID	Data sets	Samples	Attributes	Decision Classes
1	Bridges	108	12	7
2	Credit Approval	263	15	3
3	Heart-Disease	303	14	5
4	Hepatitis	155	19	2

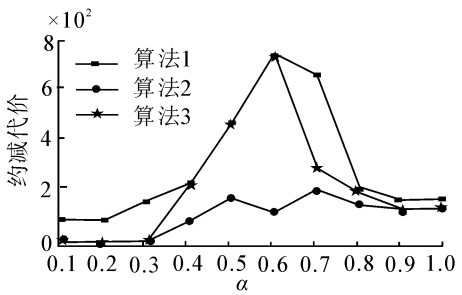
由于基于测试代价敏感的可变精度分类粗糙集模型的下、上近似集由阈值 α 控制, 因此, 在实验中选取了 10 组不同的 α 值分别对比 3 个算法的测试代价, 在算法 3 的第 8 步中, 给定阈值设为-5, 步长 δ 为 0.5, 即重复求得 10 次约简, 取其中的最小测试代价作为输出。具体的实验结果见图 1。



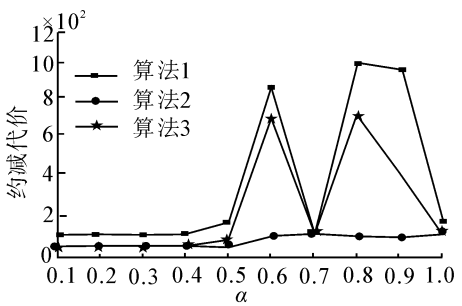
(a) Bridges



(b) Credit approval



(c) Heart-disease



(d) Hepatitis

图 1　3 种约简算法所求得的测试代价对比

Fig.1　Comparisons among test costs obtained by three algorithms

由图 1 的实验结果，可以得到：1)传统的启发式算法所获取的约简的测试代价最大，回溯算法所约简的测试代价最小，而综合考虑测试代价的改进的启发式算法得到约简的测试代价则是基于两者之间。2)从图 1 的 4 个子图可以发现，3 种算法的测试代价随 α 值的不断增加呈现先增加达到一定峰值后再下降的大致趋势，从实验的角度可看出 α 值在这 3 个算法中发挥着调节正域的作用。

4　结束语

本文将测试代价引入不完备信息系统中，提出了基于测试代价敏感的可变精度分类粗糙集模型。进一步地，通过分析传统启发式约简算法未考虑测试代价以及回溯约简算法为获取最优测试需要消耗大量时间的不足，本文对传统属性重要度测量进行了改进，并根据新的属性重要度测量设计了一种新的启发式算法用以获取测试代价次优的约简。实验表明，总体而言，改进的启发式算法是寻找约简测试代价次优的合适方法。

参考文献：

[1]PAWLAK Z. Rough sets-theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic, 1991.

[2]HU Q H, CHE X J, ZHANG L, et al. Rank entropy based decision trees for monotonic classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(11): 2052-2064.

[3]HU Q H, PAN W W, ZHANG L, et al. Feature selection for monotonic classification[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(1): 69-81.

[4]LUO G Z, YANG X B. Limited dominance-based rough set model and knowledge reductions in incomplete decision system[J]. Journal of Information Science and Engineering, 2010, 26(6): 2199-2211.

[5]王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展, 2002, 39(10): 1238-1243.
WANG Guoyin. Extension of rough set under incomplete information systems[J]. Journal of Computer Research and Development, 2002, 39(10): 1238-1243.

[6]杨习贝, 杨静宇, 於东军, 等. 不完备信息系统中的可变精度分类粗糙集模型[J]. 系统工程理论与实践, 2008, 28(5):116-121.
YANG Xibei, YANG Jingyu, YU Dongjun, et al. Rough set model based on variable parameter classification in incomplete information systems[J]. System Engineering-Theory and Practice, 2008, 28(5): 116-121.

[7]MIN F, HE H P, QIAN Y H, et al. Test-cost-sensitive attribute reduction[J]. Information Sciences, 2011, 181(22): 4928-4942.

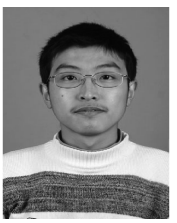
[8]MIN F, LIU Q H. A hierarchical model for test-cost-sensitive decision systems[J]. Information Sciences, 2009, 179(14): 2442-2452.

[9]MIN F, ZHU W. Test-cost-sensitive attribute reduction based on neighborhood rough set[C]//2011 IEEE International Conference on Granular Computing. Kaohsiung, China, 2011: 802-806.

[10]MIN F, ZHU W. Attribute reduction of data with error ranges and test costs[J]. Information Sciences, 2012, 211: 48-67.

[11]MIN F, HE H P, QIAN Y H, et al. Test-cost-sensitive attribute reduction[J]. Information Sciences, 2011, 181: 4928-4942.

作者简介:



鞠恒荣, 男, 1989 年生,硕士研究生,主要研究方向为粗糙集, 主持江苏省普通高校研究生科研创新计划项目一项。

马兴斌, 男, 1992 年生, 硕士研究生, 主要研究方向为粗糙集。

杨习贝, 男, 1980 年生, 副教授, 博士(后), 江苏省青蓝工程优秀青年骨干教师。主要研究方向为粗糙集、粒计算、知识工程、人工智能等。近年来, 发表学术论文 60 余篇, 出版英文学术专著一部。现主持国家自然科学基金、江苏省自然科学基金等多项科研项目。

第 11 届全球智能控制与自动化大会(WCICA2014)
The 11th World Congress on Intelligent Control and Automation(WCICA2014)

The 11th World Congress on Intelligent Control and Automation (WCICA 2014) will be held in Shenyang, China, from June 27 to 30, 2014. WCICA 2014 is technically sponsored by IEEE Robotics and Automation Society, IEEE Control Systems Society, National Natural Science Foundation of China, the Chinese Association of Automation, and the Chinese Association of Artificial Intelligence.

WCICA 2014 features plenary keynotes and plenary panel discussion sessions by the world leading researchers as well as awards to honor outstanding papers presented at this Congress. The awards include Best Paper on Theory, Best Paper on Applications, Best Student Paper, Best Poster Paper, Best Paper on Biomedical & Biosystem Related Areas, SUPCON Best Paper on Industrial Automation, and AIAG Best Paper on Supply Chain Related Topics.

Moreover, we are in the process to arrange post-conference publication of a selected group of accepted papers at WCICA 2014 in more than ten leading international journals and top Chinese journals as special issues, among them are IEEE/ASME Transactions on Mechatronics, IEEE Transactions on Robotics, IEEE Transactions on Control Systems Technology, IEEE Transactions on Industrial Electronics, IEEE Transactions on Systems, Man, and Cybernetics, Part B, and Acta Automation Sinica.

It is our great pleasure to invite you to submit your original research papers to the Congress. WCICA 2014 also welcomes proposals for organizing Focused Theme Sessions on the conference topics, Tutorials and Workshops on emerging topics. You are invited to submit focused theme session proposals to Prof. Simon X. Yang (syang@uoguelph.ca) before Dec. 1, 2013. Please refer to the Congress website for details.

Website: http://2014.wcica.info/index.shtml.
Tel: 02483681047-8013.
E-mail: wcica2014@gmail.com.