基于支持向量数据描述的无标签数据多类分类

朱帮助[1], 林 健[2]

(1. 五邑大学 系统科学与技术研究所,广东 江门 529020; 2. 北京航空航天大学 经济管理学院,北京 100083)

摘 要:为解决支持向量机(SVM)在处理无标签数据多类分类上的难题,提出了一种基于支持向量数据描述(SVDD)的无标签数据多类分类算法.该方法只需要建立一个分类模型就可以实现多类聚类分类.首先采用主成分分析作数据预处理,提取输入数据的统计特征值,得到主成分特征指标输入到 SVDD分类器进行多类聚类分类.以珠三角地区物流中心城市分类评价为研究对象,实证结果表明,采用主成分分析降低了数据维度,有效浓缩了评估信息,SVDD分类器很好地区分了各中心城市,实现了多类分类的目的.

关键词:多类分类;无标签数据;支持向量数据描述;主成分分析

中图分类号: TP18 **文献标识码**: A **文章编号**: 1673-4785(2009)02-0131-06

Multi-class classification algorithm for unlabeled data using SVDD

ZHU Bang-zhu[1], LIN Jian[2]

(1. Institute of System Science and Technology, Wuyi University, Jiangmen 529020, China; 2. School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: Support vector machines (SVM) may encounter problems in dealing with multi-class classification of unlabeled data. So we suggested a new multi-class classification algorithm based on support vector data description (SVDD) in this paper. Compared with other multi-class classification algorithms, the proposed algorithm only needed one classifier to complete the multi-class clustering classification. With this method, principal component analysis (PCA) was used to preprocess original data to extract statistically characteristic values; inputting these values into an SVDD classifier completed multi-class clustering classification. Taking nine cities in the Pearl River delta area as an example, an evaluation was made of the developmental levels of the logistics of these cities. The test results showed that data dimensions were reduced by using principal component analysis, and the evaluated information was effectively concentrated by adopting feature extraction with PCA. Moreover, the SVDD classifier could distinguish the central cities very well, so it can be used as an effective approach for multi-class classification of unlabeled data.

Keywords: multi-class classification; unlabeled data; support vector data description; principle component analysis

多类分类问题是目前模式识别领域中的一个热点与难点课题.基于统计学习理论[1]的支持向量机(SVM)从新的角度有效地解决了两类分类问题.在此基础上,一些学者开展了多类分类问题研究,提出了一些具体的实现方法,代表性的有一对多法、一对一法、决策树法、Weston法等[2-3].但这些方法通常需要构造多个两类分类器,算法计算复杂度较高;此外,这些方法无一例外都是有监督学习方法,需要为每个样本附上类别标签.但在实际应用(如区域物流中心城市分类评价)中广泛存在着大量的无标签数据,上述要求常常很难得以满足[4],因此在一定程度上降低了这些方法的实用价值.

收稿日期: 2008-07-12.
基金项目:国家自然科学基金资助项目(70471074).
通信作者:朱帮助. E-mail: wpzbz@126.com.

支持向量数据描述(SVDD)是由 Tax和 Duin于1999年提出的一种一类分类方法[5],其理论源于 SVM. 目前,SVDD已在故障诊断、语音识别、图像识别等领域得到应用[6-9]. 与 SVM 寻求最优超平面不同的是,SVDD的出发点是寻求一个包容目标样本数据的最小容量超球体,将这种基于一类分类的分类思想引入多类分类可望解决 SVM 在处理无标签数据多类分类问题上存在的难题. 为此,提出了一种基于支持向量数据描述的无标签数据多类分类方法,并将其用于区域物流中心城市分类评价. 该方法采用主成分分析作数据预处理,提取输入数据的统计特征值,得到主成分特征指标输入到 SVDD 分类器进行多类聚类分类. 实证分析也验证了该方法的有效性和可行性.

1 主成分分析和支持向量机

1.1 主成分分析原理

主成分分析(PCA)是一种统计维数压缩方法[10]. 给定一个数据集 X:

$$X = (x_{ij}),\ i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, p$$

1) 将原始数据进行标准化处理: $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sqrt{\mathrm{var}(x_j)}}$.

式中:$\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, $\mathrm{var}(x_j) = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$.

2) 计算相关矩阵: $R = (r_{ij})_{p\times p}$, $r_{ij} = \dfrac{S_{ij}}{\sqrt{S_{ii}\cdot S_{jj}}}$,

$S_{ij} = \dfrac{1}{n-1}\sum_{k=1}^{n}(z_{ki} - \bar{z}_i)(z_{kj} - \bar{z}_j)$, $\bar{z}_i = \sum_{k=1}^{n} z_{ki}/n$, $\bar{z}_j = \sum_{k=1}^{n} z_{kj}/n$

3) 令 $|R - \lambda| = 0$,求解相关矩阵 R 的特征根 λ_j 且使得 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$,其对应的特征向量为 u_j,得到主成分: $y_j = u\sum_{j=1}^{p} x_j$.

4) 计算各主成分方差贡献率: $e_j = \lambda_j / \sum_{k=1}^{p}\lambda_k$ 和累计方差贡献率: $E = \sum_{j=1}^{m} e_j$.

当 $E \geq \theta$(θ通常取 80%)时 n的最小整数作为 m 的值,即主成分的个数为 m.

1.2 支持向量机算法

对于两类问题,假定样本集 $\{(x_i, y_i), x_i \in R^l, y_i \in \{+1, -1\}\}$ 能够被超平面 $(w \cdot x) + b = 0$ 分类. 优化超平面的求解问题为

$$\begin{aligned} &\min \phi(w) = \frac{1}{2}(w \cdot w) + C\sum_{i=1}^{n}\xi_i,\\ &\text{s.t. } y_i[(w \cdot x_i) + b] \geq 1 - \xi_i, \qquad (1)\\ &\xi_i \geq 0. \end{aligned}$$

这是一个二次规划问题,根据 KKT定理,最优解为其拉格朗日函数的鞍点:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(w \cdot x_i + b) + \xi_i - 1) - \sum_{i=1}^{n}\beta_i\xi_i. \qquad (2)$$

即

$$\begin{aligned} &\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0,\\ &\alpha_i[y_i(w \cdot x_i + b_i) - 1 + \xi_i] = 0, \qquad (3)\\ &\alpha_i, \beta_i, \xi_i \geq 0,\\ &\beta_i\xi_i = 0. \end{aligned}$$

得

$$\begin{aligned} &w = \sum_{i=1}^{n}\alpha_i y_i x_i,\\ &\sum_{i=1}^{n} y_i\alpha_i = 0, \qquad (4)\\ &C = \alpha_i + \beta_i. \end{aligned}$$

根据式(4)重构式(2),得到其对偶二次规划问题:

$$\begin{aligned} &\max Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i, x_j) + \sum_{i=1}^{n}\alpha_i,\\ &\text{s.t. } \alpha_i \geq 0, \qquad (5)\\ &\sum_{i=1}^{n} y_i\alpha_i = 0. \end{aligned}$$

根据 KKT定理,对于大多数的样本,$\alpha_i = 0$. 对应 $\alpha_i \neq 0$ 的样本为支持向量[11]. 由此可见,由支持向量决定的分类面和由全体样本集决定的分类面是等价的.

对于线性不可分问题,依据统计学习理论可知,如果选用适当的核函数,将低维的输入空间数据通过核函数映射到高维特征空间,输入空间线性不可分问题在特征空间将转化为线性可分问题. 满足 Mercer条件的对称函数都可以作为核函数.

引入核函数 $K(x_i, x_j)$ 代替式(5)中向量的内积 (x_i, x_j),得

$$\max Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\cdot$$

$$\alpha_i\alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{n}\alpha_i,$$
$$\text{s t}\quad \alpha_i \geq 0, \tag{6}$$
$$\sum_{i=1}^{n} y_i\alpha_i = 0.$$

决策函数为

$$f(x) = \operatorname{sgn}[\sum_{i=1}^{n}\alpha_i y_i K(x_i, x) + b]. \tag{7}$$

任选一支持向量 x_i,计算出 b:

$$y_i[\sum_{i=1}^{b}\alpha_i y_i K(x_i, x) + b] = 1. \tag{8}$$

2　支持向量数据描述多类分类

2.1　原　理

假定一个包含 l个目标样本的无标签数据 $\{x_i, i=1, \cdots, l\}$, $x_i \in R^d$. 支持向量数据描述(SVDD)的基本思想是寻求一个最小容积的超球体,使所有的(或绝大多数的)目标样本都包含在该球体内. 即设法找一个以 a为中心,以 R为半径的能够包含所有样本点的最小超球体.

$$\min F(R, a, \xi_i) = R^2 + C\sum_{i=1}^{l}\xi_i,$$
$$\text{s t}\quad (x_i - a)^{\mathrm{T}}(x_i - a) \leq R^2 + \xi_i, \tag{9}$$
$$\xi_i \geq 0.$$

这也是一个二次优化问题,可以构造出拉格朗日函数:

$$\min F(R, a, \xi_i) = R^2 + C\sum_{i=1}^{l}\xi_i -$$
$$\sum_{i=1}^{l}\alpha_i\{R^2 + \xi_i - (x_i - a)^{\mathrm{T}}(x_i - a)\} - \sum_{i=1}^{l}\gamma_i\zeta_i,$$
$$\text{s t}\quad \alpha_i \geq 0, \tag{10}$$
$$\gamma_i \geq 0.$$

求解得

$$\sum_{i=1}^{l}\alpha_i = 1,$$
$$a = \sum_{i=1}^{l}\alpha_i x_i, \tag{11}$$
$$C - \alpha_i - \gamma_i = 0.$$

根据式(11)重构式(9),得到其对偶二次规划问题:

$$\max\sum_{i=1}^{l}\alpha_i(x_i \cdot x_i) - \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j(x_i \cdot x_j),$$
$$\text{s t}\quad \sum_{i=1}^{l}\alpha_i = 1,\ 0 \leq \alpha_i \leq C. \tag{12}$$

根据统计学习理论,只要核函数满足 Mercer条件,它就对应某一变换空间 $\phi(x_i)$中的内积[1]. 因此,在最优分类面中采用适当的内积核函数就可以实现从低维向高维空间的映射,从而实现某一低维空间的非线性问题向高维特征空间的线性问题转换,而计算复杂度并没有增加. 于是,式(12)问题转换为求最优解:

$$\max\sum_{i=1}^{l}\alpha_i K(x_i, x_i) - \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i, x_j), \tag{13}$$
$$\text{s t}\quad \sum_{i=1}^{l}\alpha_i = 1,$$
$$0 \leq \alpha_i \leq C.$$

求解优化问题(13)可以得到 α_i的值,通常大部分 α_i将为 0,不为 0的 α_i所对应的样本称为支持向量[1]. 支持向量体现在超球体的边界上. 因此,超球体的半径由支持向量到球心的距离决定. 即对应于 $0 < \alpha_i < C$的样本满足

$$R^2 - (K(x_i, x_i) -$$
$$2\sum_{j=1}^{l}\alpha_j K(x_j, x_i) + a^2) = 0. \tag{14}$$

式中: $a = \sum_{i=1}^{l}\alpha_i\phi(x_i)$. 因此,任取一支持向量,根据式(14)可求出 R的值.

对于新样本 z,令

$$f(z) = (\phi(z) - a)(\phi(z) - a)^{\mathrm{T}} =$$
$$K(z, z) - 2\sum_{i=1}^{l}\alpha_i K(z, x_i) +$$
$$\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i, x_j). \tag{15}$$

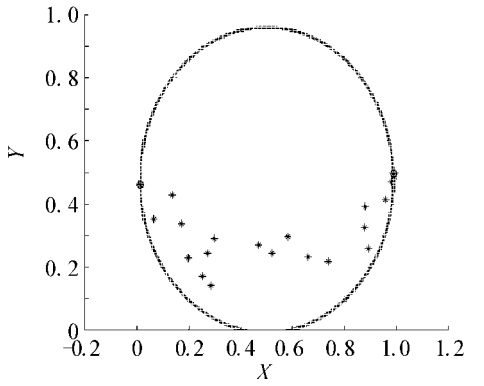
则可以判断 z是否在边界内:

$$\begin{cases} f(z) \leq R^2, & z\text{在边界内;} \\ f(z) > R^2, & z\text{在边界外.} \end{cases} \tag{16}$$

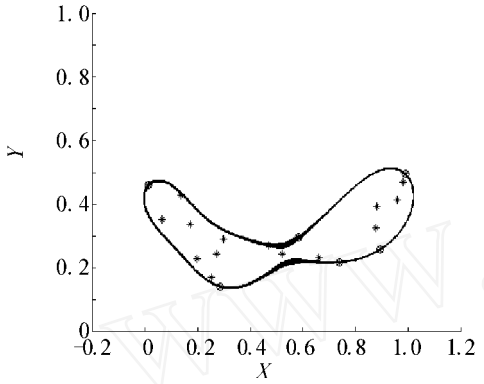
2.2　算例分析

利用 SVDD对无标签数据进行多类聚类分类,通过选择合适的核函数以及与之对应的参数,可以获得比较理想的分类边界.

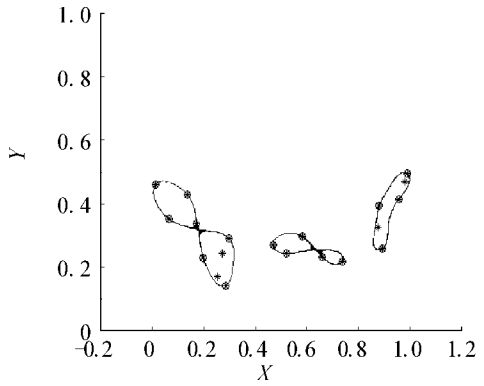
考虑如图 1所示的两维样本分布,采用 SVDD进行一类分类所得结果如图 1(a),通过调整高斯径向基核函数中参数 σ值的大小,来比较所取得的多类分类效果,如图 1(b)～(d)所示.



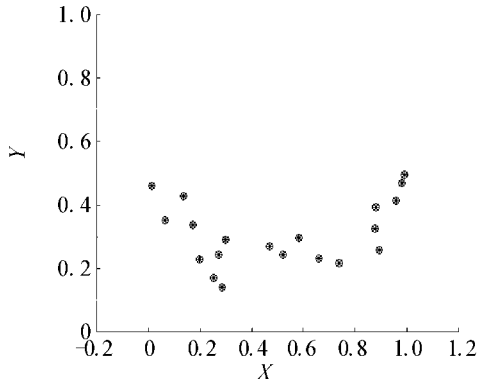
(a)σ = 1



(b)σ = 0.3



(c)σ = 0.1



(d)σ = 0.02

图 1 参数 σ对分类效果的影响

Fig 1 Multi-class classification outcomes of different σ

由此可见,随着 σ的减小,正常域范围不断缩小,支持对象数目则不断增加. 当 σ小到一定的时候,分类区域被分成若干个互不相通、相互孤立的子区域,每个子区域代表一类. 随着 σ的进一步减少,最终会形成每一个样本点对应一个子区域,即每一个样本点都是支持向量.

3 实证分析

为说明基于支持向量数据描述的无标签数据多类分类方法的有效性和可行性,以珠三角地区 9个物流中心城市分类评价作为研究对象开展实证研究. 20世纪 90年代中期以来,随着现代物流理念在我国的普及,我国部分省市已开始制定相应的物流规划,尤其是物流中心的规划与建设. 物流中心城市的确定是当前物流规划中的首要工作,其实质上是一个分类评价问题[11]. 依据综合性、客观性、可得性、可比性等原则,在对有关文献综合分析和征询有关专家意见的基础上,从 4个方面选取指标建立城市物流发展水平综合评价指标体系,这些指标从不同角度反映了城市物流的发展特征: 1)社会经济发展类:综合反映了城市物流发展的社会经济基础,包括 GDP(x_1,万元)、人均 GDP (x_2,元/人)、GDP增长率 (x_3, %); 2)生产消费流通类:分别从生产、消费等角度反映了城市物流服务的需求状况和需求规模,包括社会消费品零售总额 (x_4,万元)、工业总产值 (x_5,万元)、批发零售贸易业总额 (x_6,万元); 3)人力资源类:反映了物流发展的人力资源状况,包括物流从业人员比例,等于交通运输、仓储和邮政业人员数/总从业人员数 (x_7, %); 4)交通运输类:反映了城市物流发展的物质基础,包括运输量 (x_8,万吨)、港口吞吐量 (x_9,万吨)、运输网密度 (x_{10},千公里/万平方公里). 各指标数据均来自《广东省统计年鉴(2005)》.

采用 Matlab 7.0编程进行主成分分析,得特征值、贡献率和累积贡献率,如表 1.由于第一、二 2个主成分的特征值大于 1,且累积贡献率达到 84.09%,因此可以取前 2个主成分,作为特征提取的目的指标,如表 2.

表 1 特征值、贡献率和累积贡献率

Table 1 Eigenvalue, contribution and accumulative contribution

主成分	特征值	贡献率/%	累积贡献率/%
1	6.18	61.84	61.84
2	2.23	22.25	84.09

表 2　第一、二主成分

Table 2　The fisrt and second principle components

序号	y_1	y_2
1	- 0.66	- 1.01
2	2.21	- 0.51
3	1.05	0.10
4	0.04	1.33
5	- 0.56	0.99
6	- 0.47	- 0.27
7	- 0.27	1.10
8	- 0.77	- 0.11
9	- 0.58	- 1.62

将前 2个主成分值作为 SVDD分类器的输入向量进行多类分类.分类器核函数采用高斯径向基核函数.调节核函数的参数 σ等于 0.5,即

$$K(x_i, x_j) = \exp(- \| x_i - x \|^2 / 0.5^2).$$

SVDD的分类结果为 3类:

$$X_1 = \{2, 3\}; X_2 = \{4, 5, 6, 7, 8\}; X_3 = \{1, 9\}.$$

SVDD的分类效果如图 2所示.从图 2可以发现,基于 SVDD的无标签数据多类分类方法可以获得较为理想的分类边界.

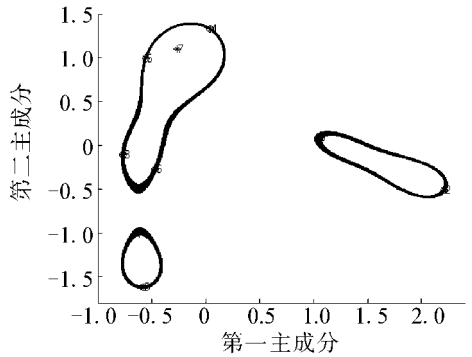


图 2　分类效果

Fig. 2　Classification outcome

为了对比,本文还采用 K-means法进行了 3类分类,分类结果与 SVDD分类结果一致,从而验证了本文所提出方法的有效性与可行性.

依据 SVDD分类结果,珠三角 9个中心城市的物流发展水平大致可分为 3个层次:第 1层次是第 1类城市,包括广州和深圳,由于在经济实力、消费能力、人力资源、物流运输能力等方面均很雄厚,导致物流发展综合水平很高,遥遥领先于其他城市;第 2层次是第 2类城市,包括佛山、中山、珠海、东莞和惠州,这些城市在经济增长、工业总产值及交通运输网络密度上具有一定的相对优势,致使物流发展综合水平较高;第 3层次是第 3类城市,包括江门和肇庆,这 2个城市位于珠三角西部,经济实力不强,物流人才短缺,交通基础设施较落后,尚未真正形成交通运输网络,造成物流发展综合水平偏低.该分类评价结果基本符合目前珠三角地区物流发展的实际情况,也得到了有关政府和物流专家的认可.

4　结　论

本文结合一类分类方法的最新成果,提出了基于支持向量数据描述的无标签数据多类分类算法,并将该方法应用于珠三角物流中心城市分类评价中,分类边界明确,分类结果合理,从而验证了该方法的有效性与可行性.

与已有的多类分类算法相比,本文提出的基于主成分分析的支持向量数据描述无标签数据多类分类方法具有如下几点优势:

1)采用主成分分析作数据预处理,提取主成分特征指标作为 SVDD分类器的输入,大大降低了数据维度,有效浓缩了评估信息,实现了多类分类的可视化;

2)将一类分类分类思想引入多类分类中,有效地解决了传统 SVM在处理无标签数据多类分类问题上存在的困难,较大程度上简化了多类分类过程中的计算复杂度,提高了模型的实用价值;

3)多类分类问题十分普遍,虽然本文是以珠三角地区物流中心城市分类评价为应用对象验证了所提出方法的有效性;但该方法具有较强的通用性,稍加变化就可以用于解决其他领域的多类分类问题.当然,在实际应用中,如何合理确定核函数的参数大小,将是未来进一步研究的问题之一.

参考文献:

[1]VAPNIK V. Statistical learning theory[M]. New York: Wiley, 1998: 59-64.

[2]PLATT J C, CRISTIANNINI, SHAWE T J. Large margin DAGs for multiclass classification[C]//Advances in Neural Information Processing Systems 12. Cambridge, Mass: MIT Press: 547-553.

[3]WESTON J, WATKINS C. Multi-class support vector machines[R]. CSD-TR-98-04. London: Royal Holloway University, 1998.

[4]ZHOU Z H, LI M. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[5]TAX D M J, DUIN R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11/13):

1191-1199.

[6] TAX D, DUIN R. Outlier detection using classifier instability[J]. Advances in Pattern Recognition, Lecture notes in Computer Science, 1998, 1451: 593-601.

[7] JIANG Shengyi, SONG Xiaoyu, WANG Hui, et al. A clustering-based method for unsupervised intrusion detections [J]. Pattern Recognition Letters, 2006, 27(7): 802-810.

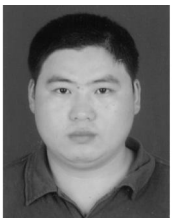
[8] BANERJEE A. A support vector method for anomaly detection in hyperspectral imagery[J]. IEEE Trans on Geoscience and Remote Sensing, 2006, 44(8): 2282-2290.

[9] 赵学风,段晨东,刘义艳,等. 一种基于支持向量数据描述的损伤诊断方法[J]. 系统仿真学报, 2008, 20(6): 1570-1573.
ZHAO Xuefeng, DUAN Chendong, LIU Yiyan, et al. Structure damage diagnosis method based on SVDD technique [J]. Journal of System Simulation, 2008, 20(6): 1570-1573.

[10] KIM K I, JUNG K, KIM H J. Face recognition using kernel principal component analysis[J]. IEEE Signal Procession Letters, 2002, 9(2): 40-42.

[11] 赵 闯,刘 凯,李电生. SOFM神经网络在物流中心城市分类评价中的应用[J]. 中国公路学报, 2004, 17(4): 119-122.
ZHAO Chuang, LIU Kai, LI Diansheng. Application of SOFM neural network for classification and evaluation of logistics center city[J]. China Journal of Highway and Transport, 2004, 17(4): 119-122.

作者简介:



朱帮助,男, 1979年生,讲师,博士,主要研究方向为复杂系统分析与建模、智能信息处理,发表学术论文近 20篇,其中多篇被 SCI、EI、ISTP收录.

林 健,男, 1958年生,教授,博士生导师,博士,主要研究方向为复杂系统建模与仿真、信息管理与信息系统,主持多项国家自然科学基金项目和省部级科研项目,发表学术论文 150余篇,其中多篇被 SCI、EI、ISTP收录.

2009中国智能自动化会议
Chinese Intelligent Automation Conference

2009中国智能自动化会议 (2009 CIAC) 将于 2009年 9月 27～29日在江苏南京举行. 会议由中国自动化学会智能自动化专业委员会和江苏省自动化学会主办,东南大学承办. 本次征文的内容包含以下 29大类:人工神经网络、模糊系统、进化计算、计算智能及软计算、智能控制、先进控制方法和技术、机器人技术与系统、多智能体系统、人工认知系统、生物信息学、离散事件系统与混合系统、无线传感器网络、智能信息处理、混沌、分形与小波、智能管理与决策、智能建模与仿真、智能故障诊断、数据挖掘与知识发现、智能技术在通信与网络中的应用、智能人机交互技术、虚拟现实及多媒体技术、计算机视觉、模式识别与图像处理、智能测量及多传感器信息融合、智能自动化装置、智能交通系统、人工生命系统及其应用、智能设计与制造、其它等.

2009年 4月 30日前通过 http://cms.amss.ac.cn/提交论文; 2009年 6月 15日前发出会议论文录用结果的通知;录用的论文将在《中国科学 F:信息科学》专刊 (SCI源)、《控制理论与应用》专刊 (EI源)、《东南大学学报》增刊 (EI源)、《南京理工大学学报》增刊 (EI源)、《中南大学学报》增刊 (EI源)或有出版号的论文集上发表. 会议网址为: http://www.iacaa.org/ciac/zh_CN/index.html 联系人:清华大学计算机系钱宗华,电话: 010 - 62788939, E-mail: qianzh@tsinghua.edu.cn.