动态粒度 SVM 学习算法*

程凤伟[1] 王文剑[1,2] 郭虎升[1]

[1](山西大学 计算机与信息技术学院 太原 030006)
[2](山西大学 计算智能与中文信息处理教育部重点实验室 太原 030006)

摘 要 粒度支持向量机(GSVM)在处理分布均匀的数据集时较有效,但现实生活中数据集的分布往往是不可预测的,且分布不均匀.文中提出一种动态粒度支持向量机(DGSVM)学习算法,根据粒的不同分布自动粒划分,使SVM可在不同层次的粒上训练.标准数据集上的实验表明,与GSVM相比,DGSVM具有更好的分类性能.

关键词 粒度支持向量机(GSVM),不均匀数据集,分布,动态粒度支持向量机(DGSVM)
中图法分类号 TP 181

Dynamic Granular Support Vector Machine Learning Algorithm

CHENG Feng-Wei[1], WANG Wen-Jian[1,2], GUO Hu-Sheng[1]
[1](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education*, *Shanxi University*, *Taiyuan* 030006)

ABSTRACT

Granular support vector machine (GSVM) is effective when dealing with distribution uniform datasets. However, the distribution of the dataset in the real world is unpredictable, and the density is uneven. In this paper, a dynamic granular support vector machine learning algorithm (DGSVM) is proposed. According to the different distribution of the granules, some granules are divided automatically and SVM training is performed on different levels of granule space. The experimental results on benchmark datasets demonstrate that DGSVM algorithm obtains better classification performance compared with GSVM.

Key Words Granular Support Vector Machine (GSVM), Uneven Dataset, Distribution, Dynamic Granular Support Vector Machine(DGSVM)

* 国家自然科学基金项目(No. 60975035,61273291)、山西省回国留学人员科研项目(No. 2012-008)、山西省研究生教育创新项目(No. 2013-3001)资助
收稿日期:2013-04-12;修回日期:2013-08-30
作者简介 **程凤伟**,女,1988 年生,硕士研究生,主要研究方向为机器学习. E-mail:867964783@qq.com. **王文剑**(通讯作者),女,1968 年生,教授,博士生导师,主要研究方向为机器学习、计算智能、数据挖掘等. E-mail:wjwang@sxu.edu.cn. **郭虎升**,男,1986 年生,博士研究生,主要研究方向为机器学习、计算智能等.

1　引　言

在许多数据分类应用问题中,由于数据的收集可能从大量、模糊、随机的环境中获得,因此常会遇到数据分布不均匀的情况. 对于这类样本分布不均匀的数据,获得的样本无法反映整个空间的数据分布. 考虑到算法的容错性,应关注密度较大的区域,以保证这部分区域的错分率较低. 数据集分布不均匀也是医疗诊断、天气预报等分类应用中经常遇到的问题. 目前绝大多数分类方法都是假设数据集的分布是均匀的,所以在处理不均匀的数据集时效果不理想,特别是在支持向量机的基础上,几乎没有算法是针对密度分布不均匀数据集提出的,如何提高对这类数据集的分类准确率,成为一个亟待解决的问题.

粒度计算[1-3]是目前智能信息处理领域中模拟人类思考问题和解决大规模复杂问题的方法,而支持向量机在解决小样本、非线性及高维模式识别中表现出特有的优势,将二者有效融合,可在不降低 SVM 泛化性能的条件下较大幅度提高学习效率. 在粒度意义下探讨支持向量机学习机制,并明确粒度支持向量机(Granular SVM, GSVM)的含义是由 Tang[4]于 2004 年提出的. 其主要思想是首先构建粒度空间获得一系列信息粒,然后在每个信息粒上进行 SVM 学习,最后通过聚合信息粒上的信息获得最终的决策函数.

GSVM 用有效信息粒代替传统的数据点训练,可大幅度提高支持向量机的训练效率,同时获得较令人满意的泛化能力[5-6]. 典型的 GSVM 主要有:基于关联规则的 GSVM[7],基于聚类的 GSVM[8-9]. 此外也有不少的学者研究基于粗糙集[10]、决策树[11]、商空间[12]及神经网络[13]的 GSVM .

上述 GSVM 在解决实际问题时取得较好效果[14-15],尤其对分布均匀的数据集较有效. 有别于研究时的理想状态,真实数据集中的数据分布是难以预料的,往往会出现分布不均匀的情况. 本文结合粒度计算理论和层次分类思想,提出一种动态粒度支持向量机(Dynamic GSVM, DGSVM)学习算法. 该算法通过构建层次粒划分模型,设计高效的动态粒度划分机制,细化密度大的粒,因此不同分布的粒可能落在不同的粒度层次上,选取动态粒划分后各个粒的代表点加入训练集,进行 SVM 训练. 从而在不改变原数据集分布的情况下有效提高 SVM 的训练效率,同时保持较高的泛化能力.

2　动态粒度支持向量机

传统的 GSVM 的粒划分机制是静态的,即每次训练过程中粒划分的个数由用户给定的初始参数决定,粒划分后,取部分代表点(如粒中心) 加入训练集训练. 传统 GSVM 实现起来较简单,易于操作,处理密度均匀的数据集效果较好. 但它对初始粒划参数较敏感,而且在处理密度分布不均匀的数据集时,对密度大的区域会丢失一些重要的分类信息,对密度小的区域又会出现选取的分类信息冗余.

DGSVM 考虑到不同粒度之间的分布密度差异对最终分类结果的影响程度不同,在选择重要分类信息时采取动态粒划分机制. 密度较大且离超平面较近的粒对分类正确率的影响也相对较大,因此对这些粒再次粒划分,提取更多的代表点加入训练集. 对于密度小且远离超平面的粒,只抽取少量代表点加入训练集,这种动态粒划分方法较好地保持原有数据集的分布. 有些粒可能多次粒划分,而有些粒也可能只一次粒划分,尽管它们不在同一层次上,但从各粒中选取的代表信息都是重要数据,因此仍能获得较好的泛化能力.

本文通过动态粒划分的方法对密度不均匀的数据集提取重要样本. 首先根据数据集的大小,用户给定一个初始粒划参数 K,对数据集初次粒划分. 其次,取每个粒的粒中心进行 SVM 训练,得到一个分类超平面,综合考虑粒密度和粒半径两个因素,计算出那些靠近边界且粒密度较大的粒,对这些粒再次粒划分(再次粒划分参数称动态粒化因子,它由粒密度和粒半径共同决定),得到一组新的粒,取新粒的粒中心,加入训练集进行 SVM 训练. 重复此过程,直到满足停止条件为止.

图 1 给出 DGSVM 动态粒划分的示意图,假设经过初始的粒划分得到图中的 7 个粒,分别取其粒中心加入训练集进行 SVM 训练,得到初始超平面,黑圈代表密度较大且离超平面较近的粒,这些粒对超平面的影响较大,需再次粒划分. 对 E_s 再次粒划分,可得到一系列新粒,取新粒的粒中心代替 E_s 加入训练集进行 SVM 训练,获得新的超平面. 若新粒中 E_1 密度较大且离超平面较近,将对其再次粒划分,仍取新粒的粒中心加入训练集训练,直到不需继续粒划分为止. 在同一层次上的粒划分过程可并行执行,进一步提高算法效率.

本文根据动态粒划因子进行粒度自动划分. 假

设数据集 X 经过初次粒划分,得到一组粒

$$E = \{E_i\}_{i=1}^{K},$$

其中,若 E_i 含有 n_i 个 m 维的数据点 $\{\boldsymbol{x}_j\}_{j=1}^{n_i}$,粒 E_i 的中心 $\boldsymbol{c}_i$ 定义为

$$\boldsymbol{c}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j. \tag{1}$$

粒 E_i 的半径反映一个粒的大小,其定义如下:

$$r_i = \max\{\sqrt{|\boldsymbol{x}_j - \boldsymbol{c}_i|^2}\}. \tag{2}$$

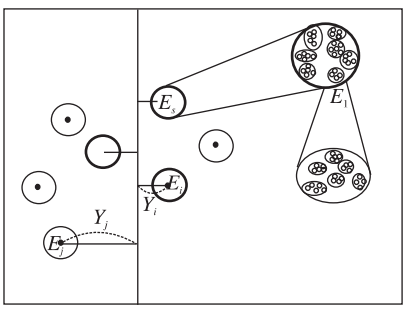

图 1 动态粒划分过程

Fig. 1 Process of dynamic granular division

粒密度是用来衡量粒中样本的稠密程度,因为样本点集的密度与各个样本点到粒中心的平均距离成反比,与样本点集所在区域内样本的个数成正比. 因此计算出粒中各个样本点到粒中心的平均距离,再用粒中样本总个数除以样本到粒中心的平均距离,如此计算出的粒密度能反映一个粒所包含样本点的稠密程度. 其定义如下:

$$d_i = \frac{n_i^2}{\sum_{j=1}^{n_i}\sqrt{|\boldsymbol{x}_j - \boldsymbol{c}_i|^2}}. \tag{3}$$

动态粒划因子 k_i 用于确定一个粒需再次粒划分的个数,希望得到密度较大且含有样本个数较多的粒,再次粒划分时划分出粒的数目较多,这样定义的优点是能从密度较大且相对重要的区域提取出更细更多的分类信息参与训练,而从密度较小相对不重要的区域中只抽取少量代表点加入训练,以便获得更优的超平面. 它由粒密度、半径和一个调和参数 *para* 共同决定,*para* 决定一个粒再次划分粒时的程度大小,它由网格搜索得出. k_i 定义如下:

$$k_i = \left\lceil \frac{d_i r_i}{para} \right\rceil, \tag{4}$$

满足 $Y_i < 1 + r_i$, $k_i > 2$ 的粒需再次粒划分.

若所有粒都不满足此条件则算法达到停止条件. 本文选择密度较大且离分类超平面较近的粒再次粒划分,因为这些粒有可能成为潜在的支持向量,且密度较大对分类结果的影响较大. $Y_i < 1 + r_i$ 用来选取离超平面较近的粒,其中,Y_i 表示粒中心到超平面的距离,r_i 表示粒的半径. $k_i > 2$ 用来选取密度较大的粒,其中 k_i 与粒密度成正比.

综上所述,DGSVM 的主要步骤如下.

算法 动态粒度支持向量机

step 1 根据初始粒划参数 K 对数据集初次粒划分,得到一系列 $E = \{E_i\}_{i=1}^{K}$. 由式(1) ~ 式(3) 计算出每个粒的中心、半径和密度.

step 2 将每个粒的粒中心加入训练集,训练得到一个分类超平面,找出那些离超平面较近且密度较大的粒,由式(4) 可得每个粒的动态粒划因子 k_i,根据动态粒划因子 k_i 再次粒划分.

step 3 用划分之后的粒代替原来的粒,重复 step 2, 直到满足停止条件,不需继续粒划分为止,此时得到的超平面即为最优超平面,得到的优化模型为

$$f(x) = \mathrm{sgn}(W\phi(\boldsymbol{x}) + b),$$

算法结束.

DGSVM 在传统 GSVM 的基础上,考虑到样本的分布密度,主要根据数据集中样本分布的稀疏程度和离超平面的远近来提取重要分类信息,对于密度较小且离超平面较远的粒,只取粒中心加入训练集,而对于密度较大且离超平面近的粒,再次粒划分,将更细的粒的中心加入训练集. 因此它在处理密度分布不均匀的数据集时,可表现出更好的特性.

3 实验及结果分析

将 DGSVM 分别与 GSVM、经典 SVM 对比,实验采用的标准数据集如表 1 所示.

表 1 实验数据集

Table 1 Datasets used in the experiment

数据集	训练集	测试集	维数
Thyroid	2800	1500	5
Diabetis	4680	3000	8
Breast_cancer	2000	770	9
Heart	4250	2500	13
Image	6500	5050	18
German	3500	1500	20

实验中采用高斯核函数,核参数取 1.0,其中正则参数取 1 000. 由于本文的重点在于如何结合粒度计算理论和层次分类模型来设计具有较好泛化能力的 DGSVM,以克服密度不均匀的数据分布对分类器泛化性能产生的负面影响,所以有关如何选择核参

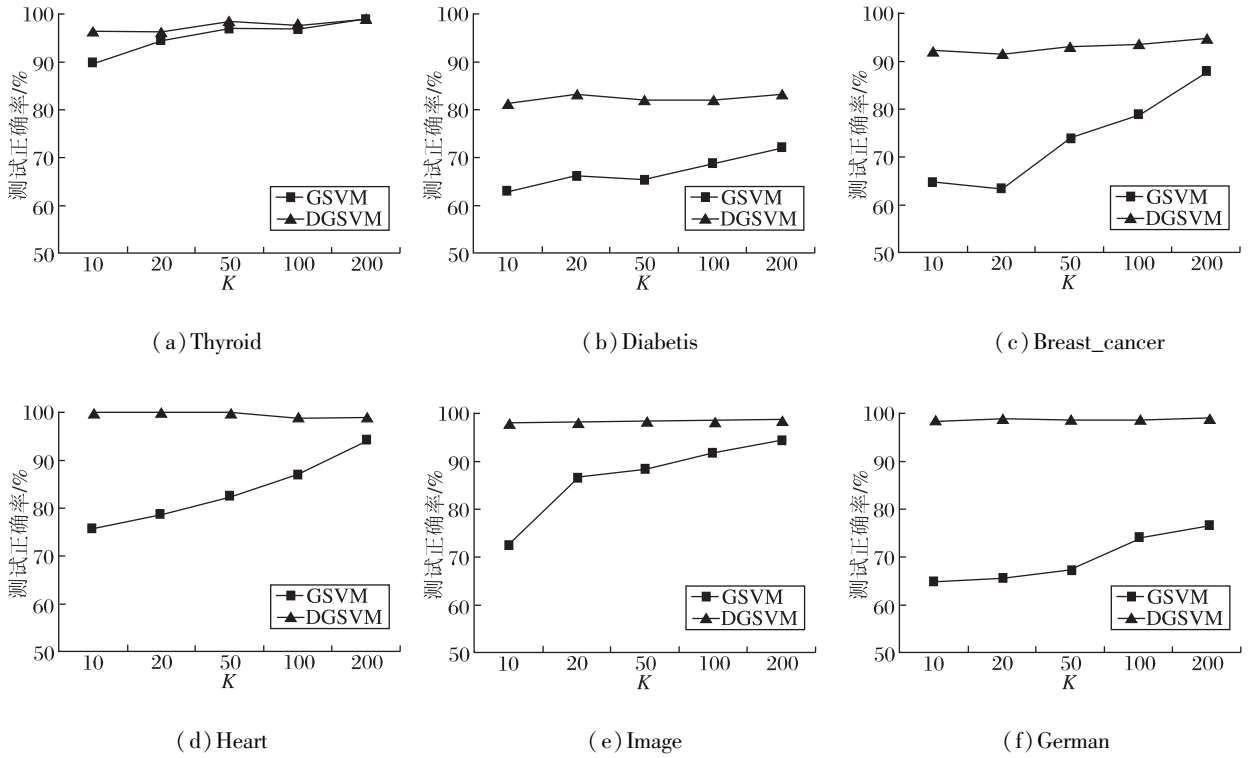

图 2　GSVM 与 DGSVM 的测试结果对比

Fig. 2　Result comparison between GSVM and DGSVM

数及正则化参数将不在文中具体讨论,相关内容可参考文献[16] ~ 文献[21]. 调和参数 *para* 取 1.0, *para* 直接影响动态粒划因子 K 的大小,*para* 越大,K 值越小;*para* 越小,K 值越大. 根据实验所需粒划分的程度来确定 *para* 的取值. 初始粒划参数 K 根据数据集的大小而定.

为测试初始粒划参数 K 对 GSVM 和 DGSVM 的影响,在每个数据集上分析初始参数 K 对算法的影响. 图 2 给出两种算法测试正确率对比.

从图 2 中可看出, 对于 GSVM,在 $K=10$ 时,正确率在 6 个数据集上都处在最低值,随着 K 值的增加,GSVM 正确率基本处于上升趋势,在 $K=200$ 时,正确率均达到最大值. 整体上看,GSVM 随 K 值的变化,正确率有大幅度的变化. 对于 DGSVM,在所有的 K 值下,正确率都高于 GSVM,且正确率一直很平稳,几乎不受初始粒划参数的影响.

实验还统计 DGSVM 中的最终划分粒的个数,图 3 给出最终划分粒的个数在不同初始粒划参数下的统计结果.

从图 3 可看出,对于 DGSVM,除数据集 Thyroid 外,最终划分粒的个数几乎不受 K 值的影响,保持平稳状态. 而对于 GSVM,最终粒划参数始终等于用户给定的初始参数,尽管 DGSVM 中最终粒的个数多于 GSVM,但 DGSVM 采用的动态粒划分机制对初始参数不敏感,结果稳定,表现出较好的鲁棒性.

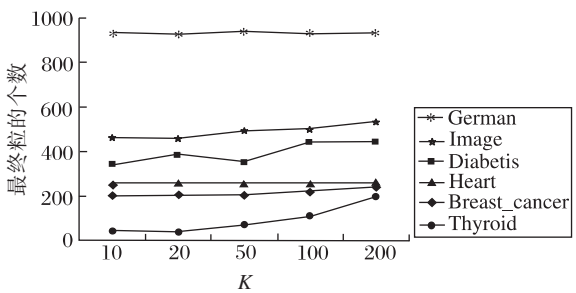

图 3　DGSVM 中最终划分粒的个数

Fig. 3　Final number of granular division in DGSVM

表 2 是 DGSVM、GSVM 和传统 SVM 测试结果对比,从实验结果中可看出,在 6 个数据集上 DGSVM 和 GSVM 的训练效率比传统 SVM 有上千倍的提高. DGSVM 与 SVM 相比,分类正确率虽有所下降,但仍在可接受范围之内. 与 GSVM 相比,DGSVM 具有更好的泛化性能,特别是在数据集 Thyroid、Heart 和 German 上,正确率都在 95% 以上,最高达 99.6%. 上述实验结果表明,DGSVM在经过粒划分

表 2 3 种方法测试结果对比

Table 2 Comparison of experimental results by 3 algorithms

数据集		$K=10$		$K=20$		$K=30$		$K=100$		$K=200$		SVM
		DGSVM	GSVM	DGSVM	GSVM	DGSVM	GSVM	DGSVM	GSVM	DGSVM	GSVM	
Thyroid	训练时间/s	0.0938	0.001	0.0625	0.016	0.2343	0.078	0.6875	0.344	2.6719	1.66	5569.578
	分类正确率/%	96.2	89.6	96.1	94.4	98.1	97.0	97.5	96.8	98.9	98.9	100.0
Diabetis	训练时间/s	10.4219	0.0166	12.8438	0.0156	10.1250	0.0781	19.2656	0.036	19.375	1.72	60008.39
	分类正确率/%	81.3	62.7	83.0	66.1	81.9	65.2	81.8	68.7	83.0	71.7	100.0
Breast_ cancer	训练时间/s	2.6885	0.016	2.7343	0.016	2.7656	0.078	2.3594	0.344	4.2813	1.75	2047.25
	分类正确率/%	92.1	64.4	91.4	63.2	92.7	73.6	93.33	78.7	94.8	87.1	97.5
Heart	训练时间/s	3.4788	0.001	3.3384	0.016	3.3696	0.078	3.1980	0.359	3.1980	1.75	389.35
	分类正确率/%	99.6	75.5	99.6	78.3	99.6	82.1	98.7	86.8	98.7	93.8	100.0
Image	训练时间/s	27.678	0.016	14.414	0.016	17.486	0.078	20.360	0.359	22.050	1.78	17256.7
	分类正确率/%	97.8	72.4	97.7	86.4	98.0	88.2	98.2	91.6	98.4	94.1	98.2
German	训练时间/s	111.322	0.001	247.3	0.016	172.125	0.078	201.656	0.0385	172.719	1.86	19330.906
	分类正确率/%	98.1	64.7	98.6	65.4	98.5	67.10	98.3	73.8	98.7	76.4	100.0

压缩后的数据集上训练,在正确率几乎没有太大变化的情况下,速度有较大提高,而且采用动态粒划分机制,使得实验结果较稳定. 因为算法是根据数据集样本分布信息动态粒划分,对于很多“不重要”的粒,只保留粒中心加入训练集,这在很大程度上减少训练样本,缩短训练时间. 同时,对那些对最终结果“有影响”的粒细化,让更多的潜在支持向量加入训练集进行训练,所以分类正确率仍保持较高水平.

4 结 束 语

本文提出一种动态粒度支持向量机学习算法,通过在不同层次的粒上,有效提取重要的分类信息进行 SVM 训练,获得较好的泛化性能,同时提高训练效率. 本文对二分类问题进行实验验证. 在未来的工作中,考虑将算法扩展到多类分布不均匀数据的分类问题中. 另外,可将本文算法应用于网页分类、疾病监测等大规模分布不均匀的实际问题中.

参 考 文 献

[1] Yao Y Y. Perspectives of Granular Computing // Proc of the IEEE International Conference on Granular Computing. Beijing, China, 2005, I: 85-90

[2] Xu C F, Wang J L. An Efficient Incremental Algorithm for Frequent Itemsets Mining in Distorted Databases with Granular Computing // Proc of the International Conference on Web Intelligence. Hong Kong, China, 2006: 913-918

[3] Yao Y Y. Granular Computing for Web Intelligence and Brain Informatics // Proc of the International Conference on Web Intelligence. Silicon Valley, USA, 2007: 1-4

[4] Tang Y C, Jin B, Zhang Y Q. Granular Support Vector Machines for Medical Binary Classification Problems // Proc of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. La Jolla, USA, 2004: 73-78

[5] Wang W J, Guo H S, Jia Y F, *et al*. Granular Support Vector Machine Based on Mixed Measure. Neurocomputing, 2013, 101(4): 116-128

[6] Guo H S, Wang W J, Men C Q. A Novel Learning Model-Kernel Granular Support Vector Machine // Proc of the International Conference on Machine Learning and Cybernetics, Baoding, China. 2009, II: 930-935

[7] Tang Y C, Jin B, Zhang Y Q. Granular Support Vector Machines with Association Rules Mining for Protein Homology Prediction. Artificial Intelligence in Medicine, 2005, 35(1): 121-134

[8] Yu H, Yang J, Han J W. Classifying Large Data Sets Using SVMs with Hierarchical Clusters // Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2003: 306-315

[9] Wang W J, Xu Z B. A Heuristic Training in Support Vector Regression[EB/OL]. [2013-02-10]. http://www. sciencedirect. com/science/article/pii/S0925231203005307?via=ihub

[10] Chen R C, Cheng K F, Chen Y H, *et al*. Using Rough Set and Support Vector Machine for Network Intrusion Detection System // Proc of the 1st Asian Conference on Intelligent Information and Database Systems. Dong Hoi, Vietnam, 2009: 465-470

[11] Kumar A M, Gopal M. A Hybrid SVM Based Decision Tree. Pattern Recognition, 2010, 43(12): 3977-3987

[12] Cui H X, Zhang L B, Kang R Y, *et al*. Research on Fault Diagnosis for Reciprocating Compressor Valve Using Information Entropy and SVM Method. Journal of Loss Prevention in the Process Industries, 2009, 22(6): 864-867

[13] Boni A D. Improved Neural Network for SVM Learning. IEEE Trans on Neural Networks, 2002, 13(5): 1243-1244

[14] Yu H, Yang J, Han J W, *et al*. Making SVMs Scalable to Large Data Sets Using Hierarchical Cluster Indexing. Data Mining and Knowledge Discovery, 2005, 11(3): 295-321

[15] Awad M, Khan L, Bastani F, *et al*. An Effective Support Vector Machines (SVMs) Performance Using Hierarchical Clustering // Proc of the 16th IEEE International Conference on Tools with Arti-

ficial Intelligence. Boca Raton, USA, 2004: 663-667
[16] Liu X D, Luo B, Chen Z Q. Optimal Model Selection for Support Vector Machines. Journal of Computer Research and Development, 2005, 42(4): 576-581 (in Chinese)
(刘向东,骆 斌,陈兆乾. 支持向量机最优模型选择的研究. 计算机研究与发展, 2005, 42(4): 576-581)
[17] Shawkat A, Smith-Miles K A. A Meta-Learning Approach to Automatic Kernel Selection for Support Vector Machines. Neurocomputing, 2006, 70(1/2/3): 173-186
[18] Wang W J, Xu Z B, Lu W Z. *et al*. Determination of the Spread Parameter in the Gaussian Kernel for Classification and Regression. Neurocomputing, 2003, 55(3/4): 643-663
[19] Wang W J, Guo J L, Men C Q. An Approach for Kernel Selection Based on Data Distribution // Proc of the 3rd International Conference on Rough Sets and Knowledge Technology. Chengdu, China, 2008: 596-603
[20] Liao S Z, Jia L. Constructing a New Spherical Kernel Function. Journal of Computer Research and Development, 2007, 44(z2): 398-402 (in Chinese)
(廖士中,贾 磊. 一类新的球面核函数的构造. 计算机研究与发展, 2007, 44(z2): 398-402)
[21] Wu T, He H G, He M K. Interpolation Based Kernel Function´s Construction. Chinese Journal of Computers, 2003, 26(8): 990-996 (in Chinese)
(吴 涛,贺汉根,何明科. 基于插值的核函数构造. 计算机学报, 2003, 26(8): 990-996)

* *

(上接第 362 页)

二、投稿须知

Springer 论文集中的稿件应为原创,并且从未发表或提交其他期刊或会议。

所有论文将通过在线系统提交,在投稿之前必须先注册作者帐号。文章格式必须采用规定的模板,每篇文章至多 10 页,包含文字、图表和参考文献。所有投稿论文必须采用英文书写,会上报告可用中文讲解。会议论文集将在 Springer 出版社的 CCIS 系列出版,并提交 EI 和 ISTP 审核后收录。

论文录用后,每篇投稿必须有至少一名作者注册,并保证在会上报告论文。缺席报告的论文将从会后的最终论文集中删除,不提交数据库检索。

请通过此网站提交文章:https://www.easychair.org/conferences/?conf=ccpr2014

论文将会被专家审阅,最终结果将会以邮件的形式发给作者。被录用的论文需准备一份讲稿或展板在会上报告,可使用中文。

三、稿件要求

每篇文章须包含以下 4 种形式的文件。

1)使用 Springer LNCS 模板编写的 word 或 latex 稿件源文件(请于 http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0 下载论文模版文件)。

2)对应的 pdf 文件。

3)论文中每幅图片的源文件。位图/线条图分辨率大于等于 400dpi 即可,矢量图无此要求。

4)版权转让页(模版可从论文版权表下载,如果论文有多个作者,允许其中一个作者作为代表签署该篇论文的版权转让页)。

具体注意事项请参见会议网站。

四、重要日期

论文投稿截止日期: 2014 年 5 月 30 日
论文录用通知日期: 2014 年 7 月 30 日
最终论文提交日期: 2014 年 8 月 25 日

会议网站:http://eeit.hnu.cn/ccpr2014/index.html