

DOI:10.3969/j.issn.1673-4785.201208002

网络出版地址: <http://www.cnki.net/kcms/detail/23.1538.TP.20130125.1528.012.html>

# 非平衡集成迁移学习模型及其在桥梁结构健康监测中的应用

于重重<sup>1,2</sup>, 吴子珺<sup>1</sup>, 谭励<sup>1</sup>, 涂序彦<sup>2</sup>, 田蕊<sup>1</sup>

(1. 北京工商大学 计算机与信息工程学院, 北京 100048; 2. 北京科技大学 计算机与通信工程学院, 北京 100083)

**摘要:** 在桥梁结构健康监测与状态评估过程中所获得的桥梁结构数据库常存在间断性异常或缺损, 且不同样本分类数据不均匀, 难以在信息缺失、数据分布失衡的情况下完成对桥梁结构健康的监测与状态评估。针对这一问题, 在改进相似性度量函数的 SOM 聚类算法和非平衡集成迁移学习算法的基础上, 提出了一种改进的迁移学习模型。通过对实际监测数据的分析, 该迁移学习模型分类精度随着目标数据集所占比例的不断增加而提高, 验证了该模型的有效性和科学性。

**关键词:** 非平衡集成迁移学习算法; SOM 算法; 迁移学习模型; 桥梁结构健康监测

**中图分类号:** TP391 **文献标志码:** A **文章编号:** 1673-4785(2013)01-0046-06

## Unbalanced integrated transfer learning model and its application to bridge structural health monitoring

YU Chongchong<sup>1,2</sup>, WU Zijun<sup>1</sup>, TAN Li<sup>1</sup>, TU Xuyan<sup>2</sup>, TIAN Rui<sup>1</sup>

(1. School of Computer & Information Engineering, Beijing Technology and Business University, Beijing 100048, China; 2. School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

**Abstract:** The examination of bridge structural data obtained in the bridge structural health monitoring and condition assessment process had the problem of intermittent abnormalities or defects in the past. However, the classification of different samples of data is seen to be uneven, thus, making it difficult to complete structural health monitoring and condition assessment of the bridge under the condition of the absence of information and data distribution imbalance. In order to solve the problem mentioned above, this paper proposes an improved transfer learning model based on self-organizing map (SOM) clustering algorithm to improve the similarity measure function and unbalanced integration transfer learning algorithm. According to the analysis of actual monitoring data, the classification accuracy of the proposed transfer learning model increased as the increasing of the proportion of the target data set, validating the efficiency and scientificity of the proposed model.

**Keywords:** unbalanced integrated transfer learning algorithm; self-organizing map algorithm; transfer learning model; bridge structural health monitoring

随着智能监测方法与评估技术的不断发展, 桥梁结构健康监测及状态评估的研究逐渐成为热点。为了对桥梁的安全和维护提供有力保障, 通常需要连续监测桥梁静应变、沉降等数据, 同时必须保证监

测数据的真实可靠。但通常存在一些干扰因素, 如传感器损坏、异常受力、天气条件恶劣等, 会造成监测数据的间断性异常或缺损, 从而无法对所获得的监测数据进行有效的数据分析<sup>[1-2]</sup>。因此, 利用相似的辅助测点数据, 实现对有效数据量少的目标测点的建模与分析, 是解决上述问题的重要思路。

迁移学习是一种强调在相似但不完全相同的领域、任务和分布之间实现知识转化的学习方式。本文

收稿日期: 2012-08-08 网络出版日期: 2013-01-25

基金项目: 国家自然科学基金资助项目(61070182); 北京市组织部优秀人才培养资助项目(2010D005003000008); 北京市学科建设项目(PXM2012\_014213\_0000\_74)。

通信作者: 于重重 E-mail: chongzhy@vip.sina.com

针对桥梁监测数据的间断性异常或缺损的特点,将迁移学习引入到桥梁结构健康监测领域,通过建模分析,解决目标测点实测有效数据过少时的数据分类和预测问题.文中提出了一个改进的迁移模型,首先通过对原始监测数据进行预处理,得到能够表征数据变化规律的训练数据集,然后再利用相似性度量函数,完成对具有相似数据变化规律的测点聚类,最后运用非平衡集成迁移学习算法(the unbalanced integrated transfer learning algorithm, UBITLA)<sup>[3]</sup>建立桥梁结构健康监测的分类模型,利用过往监测数据模型的迁移对新数据进行预测,以达到对监测数据的类别划分和评定桥梁结构损伤级别的目的.

1 迁移学习的基本理论

迁移学习作为机器学习和人工智能领域研究的新方向,近年越来越受到关注.迁移学习强调的是区域、任务、分布相似但不相同的知识的传递<sup>[4-5]</sup>,其突破了传统数据挖掘的两大假设:一是训练数据与测试数据必须同概率分布;二是两者必须同特征空间.然而,随着对迁移学习的研究,不同学者对迁移学习中存在的一些缺陷,提出了很多解决问题的算法,并得到了有效的应用.如运用基于 EM 的跨语言文本分类算法用以解决跨特征空间的迁移学习问题<sup>[6]</sup>,基于朴素贝叶斯分类器的支持跨领域文本分类的分类器实现了不同领域文本之间知识的迁移问题<sup>[7]</sup>,基于 TrAdboost 算法解决对称的二分类问题等<sup>[8-9]</sup>.这些成功的应用推动了迁移学习在处理新任务中标记数据量小的问题中的迅猛发展.

在桥梁结构健康监测领域,监测数据存在以下

特点:

- 1) 不同类别的样本分布极度不平衡,差异大;
- 2) 辅助数据中存在大量的与目标数据集不相似的冗余信息;
- 3) 桥梁实际监测的表征结构健康的正数据和表征结构损伤的负数据分布极度不平衡,负样本数量远远小于正样本,但负样本对于桥梁健康监测影响很大.因此,对于桥梁结构的损伤判定可以将其看作是一个非对称的二分类问题.

本文将迁移学习引入到桥梁结构健康监测领域,利用已有监测点数据,基于非平衡集成迁移学习算法(UBITLA),提出了一种改进的迁移学习模型,该方法能够有效地解决桥梁结构健康监测中存在的实际问题.

2 改进的迁移学习模型在桥梁结构健康监测中的应用

2.1 改进的迁移学习模型描述

本文提出的迁移学习模型总体思路如下:1) 将桥梁结构健康监测系统中监测到的原始数据进行预处理,其中包括信息筛选、去冗余、去噪声以及去除重复数据;2) 对预处理后的监测数据,根据测点间的相似度,对具有相似变化规律的测点进行聚类;3) 对聚类结果使用非平衡集成迁移学习算法,建立针对目标测点的数据分类模型,对目标数据集中的无标签数据进行预测分类,以实现对监测数据进行类别划分,评定桥梁各级损伤情况的目的.总体流程图如图 1 所示.

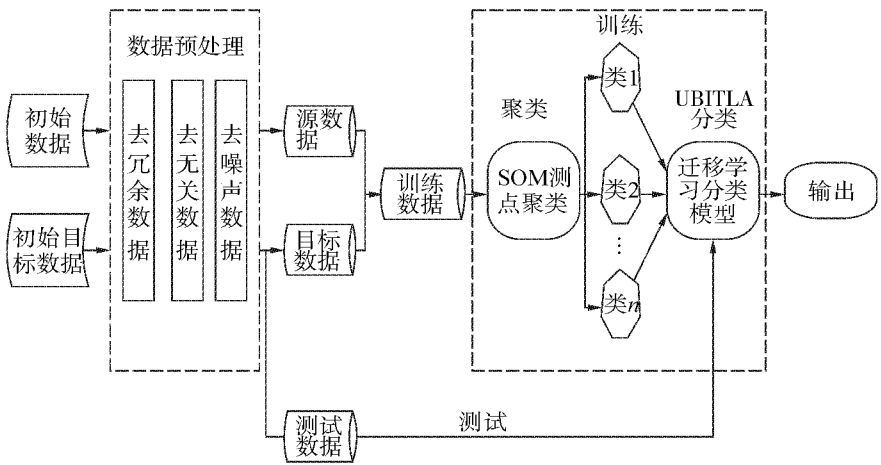


图 1 改进的迁移学习模型

Fig.1 Improved transfer learning model

2.2 基于相似测点的 SOM 聚类

相似度函数是聚类的前提条件,这是由于分类

是将相似的模式样本聚为一类.因为数据存在不同的类型,所以有不同的方法计算相似度,如欧氏距

离、切比雪夫距离、曼哈顿距离、明氏距离、加权的明氏距离、马氏距离、夹角余弦函数等<sup>[10]</sup>。在自组织映射(self-organizing map, SOM)算法中经常使用相似度函数作为分类的基础。SOM算法是一个由全连接的神经元阵列组成的无导师的自组织、自学习网络,是神经网络用于解决聚类问题的典型应用<sup>[11-12]</sup>。

由于桥梁结构健康监测系統所获得的监测数据受突发应力与异常负载等影响,经常会出现个别监测点数据存在异常,如果使用上述方法难以有效克服个别维度的干扰,且易受到高维特性的干扰,若存在2个数据的个别维度上的值差异过大或过小,这会影响整体度量效果,使数据间的相似信息被这少数维度的过大差异所淹没<sup>[13-14]</sup>。因此为减小这些异常数据对整体数据的影响,对传统的距离度量函数进行修改,从根本上解决传统的相似度量函数在高维数据空间中存在的不足,有效地缓解大差异个别维度对相似度的影响,具体如下。

**定义1** 设有 $d$ 维数据向量 $X = (x_1, x_2, \dots, x_d)$ 及 $Y = (y_1, y_2, \dots, y_d)$ ,则相似度量函数(similarity measurement points clustering, SMPC)如式(1)所示:

$$f_{\text{smpc}}(X, Y) = \frac{\sum_{i=1}^d e^{-|x_i - y_i|}}{d}. \quad (1)$$

改进后的相似度量函数通过引入指数 $e$ ,实现了在 $x_i$ 与 $y_i$ 大差异的情况下,分子数值变小的需求。通过对比计算发现,改进后的相似度量函数SMPC对比于传统的相似性度量函数,具有如下性质:

1) 新函数充分考虑了2个数据向量间的所有

维度,使大差异维度的贡献率降低;

2) 新函数的最大值为1,代表 $X$ 和 $Y$ 在所有维度上的值都相等,此时 $X$ 和 $Y$ 在 $d$ 维空间上是完全重合的,相似度最大。若最小值为0,则代表 $X$ 和 $Y$ 在每个维度上的差异均接近于无穷大,此时 $X$ 和 $Y$ 相似度最小。

通过基于SMPC相似度量函数的SOM网络,完成对桥梁监测点的聚类,其具体训练步骤如下:

1) 初始化SOM网络结构,对神经元权值向量随机赋初值;

2) 随机选择输入样本;

3) 计算各神经元之间的相似度,选择具有最大相似度的神经元作为获胜神经元;

4) 使用单调递减函数对获胜神经元及其邻域范围内的神经元的权值进行调整;

5) 对算法的收敛性进行判断,若不满足收敛条件则重复上述步骤3)直至算法达到收敛。

## 2.3 UBITLEA 算法

针对桥梁结构健康监测系統表征桥梁结构健康与否的正、负数据分布极度不均衡,若对其施行相同的权值调整策略会由于稀缺类样本权值的迅速减小而导致训练的失败问题,因此模型采用非平衡样本分类的集成迁移学习算法UBITLEA进行处理。算法在TrAdaboost算法的基础上,通过改变稀缺类样本权值的调整策略,使其不会迅速变小,从而保证这部分样本对模型建立的贡献率;另外该算法还引入冗余数据动态剔除策略,适时剔除辅助训练数据集中的冗余数据,以确保训练数据的高质。UBITLEA算法流程如图2所示。

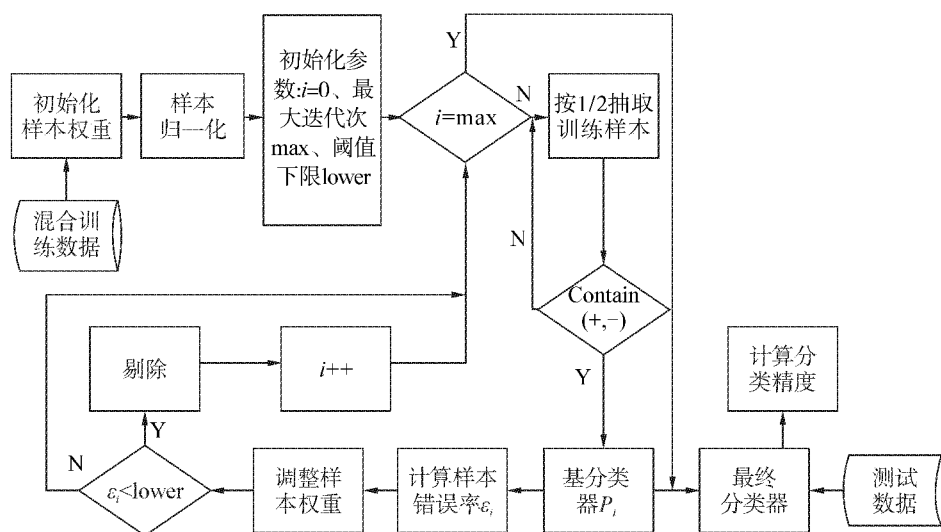


图2 UBITLEA 算法

Fig. 2 UBITLEA algorithm

UBITLA 算法的主体思想为:将迁移辅助数据集  $A$  与目标数据集  $O$  按比例抽取部分数据后混合成训练数据集  $C$ ,并通过初始化样本权重与归一化样本完成对原始样本的预处理;然后在每次的训练结果中选择误差最小的弱分类器  $h_t$ ,经  $T$  轮迭代后得到弱分类序列  $h_1, h_2, \dots, h_t$ ;最终将多个弱分类器叠加起来得到一个强分类器。

3 实验结果与分析

3.1 实验样本数据

实验采用已有 2 年监测历史的杭州湾大桥作为研究对象,选用某天 6:00—20:00 共 20 个测点所测得的相邻时段的数据差值作为实验数据,其中包含了早、中、晚 3 个高峰时段,以及上、下午低峰时段的数据变化情况,并分别从各数据类型中抽取一定数目的数据按比例组成初始数据集  $S_1$  和初始目标数据集  $S_2$ ,其具体组成如表 1 所示。

表 1 实验数据组成描述

Table 1 Description of the experimental data

数据类型	时间	比例/%	样本数	
			$S_1$	$S_2$
早高峰	7:00—9:00	30	1 500	600
上午低峰	9:00—11:00	5	250	100
午高峰	11:00—13:00	30	1 500	600
下午低峰	13:00—17:00	5	250	100
晚高峰	17:00—19:00	30	1 500	600

其中,正负样本比例为 5:1;输入数据为以 1 h 为监测周期,采集当日从 6:00—20:00 共 14 h 的静应变数据,并组成具有 14 维的属性向量(其包含监测数据在 1 d 内的变化情况)。在训练过程中从  $S_1$  内随机抽取 1/2 的数据,经数据预处理后作为源数据集  $Y_1$ ,同样从  $S_2$  内分别随机抽取 1/2 的数据,经数据预处理后作为目标数据集  $G_1$  与测试数据集  $T_1$ ,并通过由源数据集  $Y_1$  与目标数据集  $G_1$  所组成的训练数据完成对迁移学习模型的训练,实现对目标数据集  $T_1$  中无标签数据的预测分类,这样就可以通过监测数据的类别来评定桥梁的级别损伤。

3.2 结果与分析

3.2.1 不同距离函数

通过对训练数据采用不同的相似度距离函数的 SOM 网络完成对监测点的聚类,其聚类结果如表 2 所示,其中  $D^{(2)}(x, y)$  为欧式距离的相似性度量函数。

表 2 杭州湾大桥沉降数据监测点聚类

Table 2 Hangzhou Bay Bridge settlement data monitoring points clustering

类别	距离函数	
	$D^{(2)}(x, y)$	$f_{\text{smpe}}(x, y)$
聚类 1	测点 5、6、7、11、13、16、17、22	测点 5、6、7、11、13、16、17、20、22、24
聚类 2	测点 8、9、10、19、20、21、23、24、	测点 8、9、10、19、21、23
聚类 3	测点 12、14、15、18	测点 12、14、15、18

然后将上述 2 种不同的聚类结果,应用 UBITLA 迁移学习算法建立针对目标数据的分类模型,完成对目标数据中无标签数据的预测分类。采用欧式距离作为相似度距离函数与采用 SMPC 相似度距离函数的 SOM 网络聚类结果,对迁移学习模型分类精度的影响如图 3 所示。

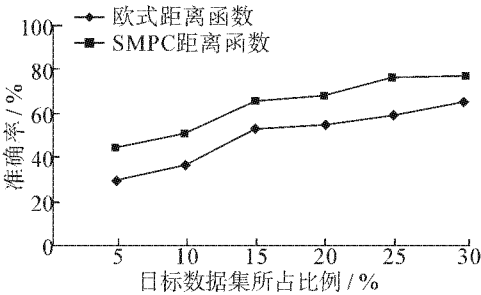


图 3 2 种不同的距离函数对迁移学习模型分类精度的影响

Fig. 3 Effects of two different distance functions on migration learning model classification accuracy

从上述的结果来看,2 种度量函数的总体聚类效果基本相近,但是对于个别测点,如杭州湾大桥的 20 号测点,在采用欧式距离作为相似度距离函数的 SOM 网络的聚类结果中将其分入了类别 2 中,而在采用 SMPC 相似度距离函数的 SOM 网络的聚类结果中将该测点分入了类别 1 中。通过对该测点和所在类别特点的进一步分析发现,该测点只是在某 2 个维度上的数据波动较大而在其余的 12 个维度的波动均比较小。由于传统的距离函数将所获得的 14 维数据向量都均等看待,放大了个别维度中异常数据所造成的影响,因此影响了整个聚类结果的合理性;但是本文提出的 SMPC 距离函数克服了传统度量函数的这一缺陷,使得聚类结果更接近实际。因

此,采用 SMPC 距离函数所获得的迁移学习模型分类精度高于传统的欧式距离函数.

### 3.2.2 辅助数据集的加入

选择杭州湾大桥测点 20 作为目标分类测点,分别选择同类辅助测点 5 和异类辅助测点 8 作为辅助训练测点,代入 UBITLA 算法进行训练. 辅助数据集对迁移学习模型分类精度的影响如图 4 和 5 所示.

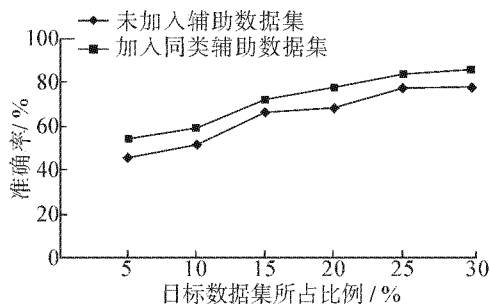


图4 同类辅助数据集的加入对迁移学习模型分类精度的影响

Fig.4 Effects of similar auxiliary data sets on migration learning model classification accuracy

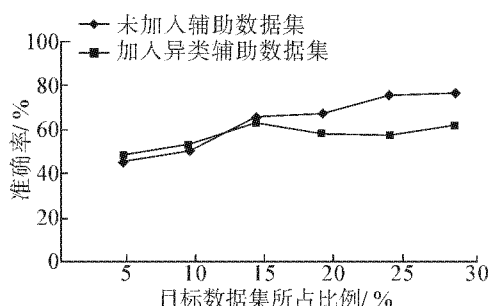


图5 异类辅助数据集的加入对迁移学习模型分类精度的影响

Fig.5 Effects of heterogeneous auxiliary data sets on migration learning model classification accuracy

实验结果表明,同类辅助数据集的加入可以有效地帮助目标数据的学习,随着目标数据集比例的增加,迁移学习模型分类精度不断提高;异类辅助数据集的添加,有时会对目标数据的学习起到促进作用,但是随着目标数据集所占比例的增加会出现相反的情况,这可能是由于异类辅助数据集数据与目标数据差异性较大,对迁移学习模型分类精度造成了一定的消极影响,导致分类精度下降. 总之,引入同类辅助数据集能够有效地在辅助数据与目标数据间进行知识迁移,提高对目标数据的学习效率和迁移学习模型分类的准确性.

## 4 结束语

本文围绕迁移学习的基本理论和在数据分类问题中的应用,针对桥梁健康监测中监测数据存在的问题,提出了一个改进的迁移学习模型,并且介绍了模型总体思路和关键技术. 通过对现役桥梁的监测数据的实验,证明了该模型的有效性. 随着技术的创新、方法的完善,迁移学习在不久的将来会应用在不同的领域上,期待其有更好的创新.

## 参考文献:

- [1] 刘永前. 大型桥梁结构健康监测技术研究与应用[D]. 北京: 北京交通大学, 2007: 1-12.  
LIU Yongqian. Research and application on structural health monitoring of long-span bridge[D]. Beijing: Beijing Jiaotong University, 2007: 1-12.
- [2] 李鹏飞, 吴太成. 桥梁健康监测技术研究综述[J]. 建筑监督检测与造价, 2010, 7(7): 28-31, 40.  
LI Pengfei, WU Taicheng. A review of health monitoring techniques of bridge[J]. Supervision Test and Cost of Construction, 2010, 7(7): 28-31, 40.
- [3] 于重重, 田蕊, 谭励, 等. 非平衡样本分类的集成迁移学习算法[J]. 电子学报, 2012, 40(7): 1358-1363.  
YU Chongchong, TIAN Rui, TAN Li, et al. Integrated transfer learning algorithmic for unbalanced samples classification[J]. Acta Electronica Sinica, 2012, 40(7): 1358-1363.
- [4] 许至杰. 迁移学习理论与算法研究[D]. 上海: 华东师范大学, 2012: 4-10.  
XU Zhijie. Research about the theories and algorithms of transfer learning[D]. Shanghai: East China Normal University, 2012: 4-10.
- [5] 戴文渊. 基于实例和特征的迁移学习算法研究[D]. 上海: 上海交通大学, 2008: 14-21.  
DAI Wenyuan. Instance-based and feature-based transfer learning[D]. Shanghai: Shanghai Jiao Tong University, 2008: 14-21.
- [6] UNKELBACH J, YI S, SCHMIDHUBER J. An EM based training algorithm for recurrent neural networks[J]. Proceedings of the 19th International Conference on Artificial Neural Networks. Berlin/Heidelberg: Springer-Verlag, 2009: 964-974.
- [7] DAI Wenyuan, XUE Guirong, YANG Qiang, et al. Transferring naive Bayes classifiers for text classification[C]// Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver, Canada, 2007: 540-545.
- [8] RAINA R, BATTLE A, LEE H, et al. Self-taught learn-

ing: transfer learning from unlabeled data[C]//The Twenty-fourth International Conference on Machine Learning. Corvallis, USA, 2007: 759-766.

- [9] 刘伟, 张化祥. 数据集动态重构的集成迁移学习[J]. 计算机工程与应用, 2010, 46(12): 126-128.

LIU Wei, ZHANG Huaxiang. Ensemble transfer learning algorithm based on dynamic dataset regroup[J]. Computer Engineering and Applications, 2010, 46(12): 126-128.

- [10] 邵昌昇, 楼巍, 严利民. 高维数据中的相似性度量算法的改进[J]. 计算机技术与发展, 2011, 2(2): 7-10.

SHAO Changsheng, LOU Wei, YAN Limin. Optimization of algorithm of similarity measurement in high-dimensional data[J]. Computer Technology and Development, 2011, 2(2): 7-10.

- [11] 杜俊卫, 李爱军. 一种基于聚类的文本迁移学习算法[J]. 计算机系统应用, 2010, 19(12): 238-241.

DU Junwei, LI Aijun. Transfer learning algorithm for text classification based on clustering[J]. Computer Systems & Applications, 2010, 19(12): 238-241.

- [12] 戴群, 陈松灿, 王喆. 一个基于自组织特征映射网络的混合神经网络结构[J]. 软件学报, 2009, 20(5): 1329-1336.

DAI Qun, CHEN Songcan, WANG Zhe. Hybrid neural network architecture based on self-organizing feature maps[J]. Journal of Software, 2009, 20(5): 1329-1336.

- [13] 王戒躁, 钟继卫, 王波. 大跨桥梁健康监测系统设计构成及其进展[J]. 桥梁建设, 2009(增刊2): 11-16.

WANG Jiezhao, ZHONG Jiwei, WANG Bo. Designed components and development of health monitoring systems for long span bridges[J]. Bridge Construction, 2009(S2): 11-16.

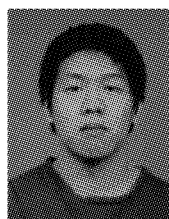
- [14] 于重重, 杨扬, 涂序彦, 等. DBSCAN 算法在桥梁健康监测预测模型中的应用[J]. 计算机工程与应用, 2008, 44(12): 224-227.

YU Chongchong, YANG Yang, TU Xuyan, et al. Application of DBSCAN algorithm in bridge-health monitoring prediction model[J]. Computer Engineering and Applications, 2008, 44(12): 224-227.

#### 作者简介:



于重重, 女, 1971 年生, 教授, 硕士生导师, 主要研究方向为智能信息处理与模式识别、复杂实时监测系统预测与评估等. 参与“十一五”重大科技支撑项目子课题、国家自然科学基金项目多项, 获部科技进步奖 1 项. 发表学术论文 40 余篇, 其中被 EI 检索 20 余篇, 合作出版著作 4 部.



吴子珏, 男, 1989 年生, 硕士研究生, 主要研究方向为机器学习、人工智能等.



谭励, 女, 1980 年生, 副教授, 博士, 主要研究方向为智能信息处理、机器学习、计算机视觉、无线传感器网络等. 作为项目负责人和主研人员参与国家自然科学基金项目 1 项、北京市优秀人才项目 1 项. 发表学术论文 20 余篇, 出版著作 1 部.



涂序彦, 男, 1935 年生, 教授, 博士生导师, 北京科技大学计算机与系统科学研究所所长, 中国人工智能学会荣誉理事长, 北京科技大学信息工程学院特聘教授. 创立多变量协调控制理论、最经济控制理论; 倡导人体控制论、大系统控制论新学科, 开拓人工智能、专家系统、智能控制、智能管理、人工生命新方法、新技术. 荣获国家科技攻关重大成果奖, 冶金部、机械部、电子部等科技进步奖多项. 发表学术论文 300 余篇, 出版专著 12 部.



田蕊, 女, 1986 年生, 硕士研究生, 主要研究方向为机器学习、人工智能等.