基于智能手机大数据的交通出行方式识别研究*

李 喆[1], 孙 健[2a,2b], 倪训友[2b]

(1. 上海市政交通设计研究院有限公司综合交通所, 上海 200030; 2. 上海交通大学 船舶海洋与建筑工程学院 a. 海洋工程国家重点实验室; b. 交通研究中心, 上海 200240)

摘 要: 智能手机时代所产生的大数据能够为交通研究者带来大量信息,基于智能手机采集交通出行大数据,再利用基于粒子群的支持向量机模型进行交通出行方式识别研究。在分析数据特点的基础上提出用于建模的特征变量,之后使用粒子群算法优化支持向量机参数,并基于成都市的实证数据进行模型的训练与出行方式识别研究。研究结果表明,该模型识别正确率为 95.1%,高于决策树、BP 神经网络、基于网格搜索的支持向量机模型,且该模型在时间效率方面具有明显的优越性,因而在出行方式识别方面具有良好的现实意义。

关键词: 粒子群; 支持向量机; 出行方式识别; 智能手机大数据; 模式识别

中图分类号: U116.2 **文献标志码**: A **文章编号**: 1001-3695(2016)12-3527-03

doi:10.3969/j.issn.1001-3695.2016.12.002

Travel mode recognition based on smart phone big data

Li Zhe[1], Sun Jian[2a,2b], Ni Xunyou[2b]

(1. *Shanghai Municipal Transportation Design Institute Co., Ltd., Shanghai* 200030, *China*; 2. *a. State Key Laboratory of Ocean Engineering, b. Transportation Research Center, School of Naval Architecture, Ocean & Civil Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*)

Abstract: The big data generated by smart phones can bring a lot of information for traffic investigators, this paper proposed a model based on particle swarm optimization and support vector machine to recognize different travel modes based on the smart-phone data. After analyzing the characteristics of data collected by smartphones, it proposed several feature variables for mode-ling. Further on, it used particle swarm optimization for optimizing the support vector machine model, which was trained and tested for travel mode recognition based on the empirical data in Chengdu, Sichuan province. The results indicate that, the recognition accuracy of the proposed model attains 95.1%, is higher than that of the decision trees, back propagation neural network model and the support vector machine based on grid search optimization. The time efficiency of the proposed model has good performance in urban travel mode recognition.

Key words: particle swarm; support vector machine; travel mode recognition; smart phone big data; pattern recognition

0 引言

随着智能手机的广泛普及以及移动应用程序的爆炸式增长,如何利用智能手机应用程序进行交通数据的采集以及利用采集的数据进行交通规划与管理成为重要的研究方向。以 IOS 和 Android 系统为代表的智能手机迅速发展,市场研究机构 DisplaySearch 预测 2016 年全球智能手机出货量将达 14.5 亿部[1]。根据 Gartner 统计数据,预计到 2016 年,全球手机应用程序下载次数将达 3 000 亿次[2]。传统交通信息采集模式落后、系统服务质量偏低,如何利用手机应用程序采集交通出行数据,尤其是对个体某次出行方式的识别是当前研究的热点与难点。利用手机数据或者手机应用程序采集交通数据获得出行参数是最近兴起的研究课题。Zheng 等人[3]基于 GPS 设备采集原始 GPS 定位数据,并进行交通出行方式的识别研究,分别比较决策树(decision trees)、贝叶斯网络、支持向量机(SVM)等方法进行交通出行方式识别,各类方法选取的输入变量包括路段长度、平均速度、均值及协方差、三个最大速度、三个最大加速度等。

李海峰等人[4]建立了基于神经网络的交通方式选择模型,选用性别、年龄、收入、职业、出行目的、出发地点、到达地点、出发时间、到达时间九个变量作为模型的输入,利用江苏省徐州市大规模城市人口出行调查中交通抽样调查的实测数据为实例,对步行、自行车和公交车出行方式进行识别。但该研究识别缺乏出租车、地铁等出行方式,而这些都是交通出行方式的重要组成部分。

张治华[5]采集了 222 个交通出行,共有近 580 000 个 GPS 记录,且每个记录包括速度、加速度、行驶方向变化特征。在此基础上,利用速度的 75 分位数、速度的离差、矢量加速度的均值、信号质量等作为模型输入,分别构造了多层感知器神经网络、贝叶斯网、决策树等模型,结果表明多层感知器神经网络和决策树具有良好的识别精度。

闫彭[6]提出基于 AGPS 手机的交通方式识别方法,通过手机 GPS 数据记录软件采集得到 GPS 数据,并选择瞬时速度、加速度、平均的视野内卫星数量、平均 HDOP 值(HDOP 是描述水平坐标精度的误差程度)作为交通方式识别的特征,并利用 BP 神经网络(back propagation neural network)进行交通出行方式识

收稿日期: 2015-09-06; **修回日期**: 2015-11-09 **基金项目**: 国家自然科学基金资助项目(71101109);上海市"科技创新行动计划"软科学研究重点资助项目(15692105400)

作者简介:李喆(1992-),男,河南信阳人,硕士研究生,主要研究方向为交通运输规划与管理(lizhe199266@mail.sjtu.edu.cn);孙健(1977-),男,教授,博导,主要研究方向为公交系统与智能交通系统;倪训友(1985-),男,博士研究生,主要研究方向为智能交通、停车诱导.

别。研究结果表明,BP 神经网络对出行方式的识别有很好效果。

汪磊等人[7]采用 SVM 模型对出行方式进行识别,结果表明 SVM 对出行方式识别具有良好的应用效果。但其对于 SVM 参数并未进行优化,导致虽采用了 SVM 模型,但由于 SVM 参数非最优组合而引起最终识别精度不高。

上述研究采用了多种识别方法对交通出行方式进行识别研究,考虑到 SVM 具有较强学习能力,且在交通领域已有较多应用,比如行程时间预测[8]、短时交通流预测[9]、公交车到站时间预测[10]等。本文将其引入到出行方式的识别研究,在此基础上引入粒子群算法对支持向量机模型进行参数寻优,对基于智能手机应用的实证数据进行研究,并将结果与决策树、BP 神经网络和基于网格搜索(grid search,GS)的支持向量机模型进行比对。

1 模型原理

1.1 支持向量机

SVM 模型能非常有效地解决小样本、非线性以及高维度的模式识别和预测问题,具有在各类函数集中构造函数的通用性,文献[11]给出了详细 SVM 模型的推导过程。SVM 不要求具体的函数形式,可捕获非线性系统输出变量(如出行方式)和输入变量(如速度均值、加速度均值、速度方差、加速度方差、速度 75 分位数等)之间的复杂关系[12,13]。

本研究基于已有研究[14,15],选择径向基核函数进行分类分析。由于径向基核函数被选用,为获得 SVM 最佳出行方式识别结果,需要对惩罚参数和核函数参数进行最佳参数组合寻优。

1.2 粒子群算法

1995 年 Kennedy 等人[16]提出粒子群算法,也称做粒子群优化算法(particle swarm optimization,PSO),该算法能有效解决复杂空间的优化问题,相关原理与过程参见文献[17,18]。粒子群算法作为一种高效的优化工具,能在系统辨识、神经网络训练等方面发挥重要作用。粒子群算法既有优异的全局搜索能力且收敛速度快,因此被用于进行 SVM 最优参数的搜索。

1.3 PSO-SVM 模型

支持向量机的最大优点在于能解决非线性划分、小样本量、高维数的学习问题,虽然人工神经网络模型也善于非线性拟合,但容易获得局部最优。训练良好的支持向量机能高效映射模型输入与输出之间的非线性关系,是有效的分类技术。

基于粒子群算法的支持向量机(PSO-SVM)模型如图 1 所示,其进行出行方式识别流程如图 2 所示。

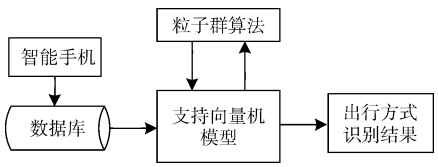

图 1 PSO-SVM 模型

如图 1 所示,PSO-SVM 模型进行出行方式识别,首先利用智能手机应用程序(APP)进行出行数据采集并实时上传到数据库;之后,利用支持向量机对数据进行训练;训练过程中有两个待定参数,其取值利用粒子群算法进行寻优。在此基础上采用最优的支持向量机进行出行方式识别研究,并得到结果。

图 2 给出了 PSO-SVM 模型进行出行方式识别的具体流程。其中需要强调的是粒子群算法被应用于优化支持向量机的参数,如果 PSO 算法寻找到最优的参数和,则根据最优参数和对应支持向量机进行出行方式识别;如果没有寻找到最优参数和,则继续寻优,直到满足为止。为了比较该模型对出行方式的识别精度高低,将该模型结果与决策树、BP 神经网络、基于网格搜索的支持向量机(GS-SVM)结果进行对比分析。

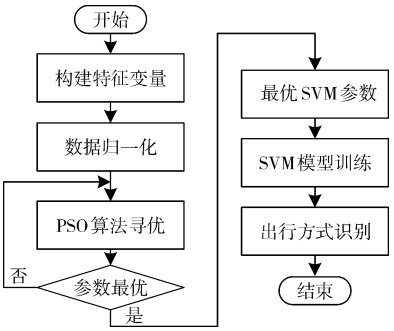

图 2 PSO-SVM 模型进行出行方式识别流程

2 出行方式特征分析与特征变量选择

基于智能手机应用程序采集居民出行数据,本章依据智能 APP 采集数据的特点对于步行、自行车、公交车、出租车、地铁等出行方式进行特征分析。基于带有 A-GPS 功能的智能手机是数据采集的重要载体,实现了借助 GPS、基站等进行高精度的混合定位,手机应用程序能够实时记录出行者的经度坐标、纬度坐标、定位时刻、定位精度、定位数据类型、瞬时速度、定位卫星数等数据,实时计算、记录出行者的瞬时加速度等数据。

智能手机应用程序定位优先采用高精度定位方式的原则。首先,智能手机 GPS 模块与 GPRS 模块应同时开启,双方信号均可获得时优先采用 GPS 定位,仅在 GPS 无法定位时采用基站定位。根据不同定位方式可得不同数据项,如表 1 所示。

表 1 不同定方式获得数据项

定位方式	数据项							
	标志符	速度	加速度	经度	纬度	时刻	精度	卫星数
GPS	61	有	有	有	有	有	有	有
基站	161	无	无	有	有	有	有	无

其中,标志符为 61(GPS 定位)和 161(基站定位)的数据为有效定位数据,其他均为无效数据。

2.1 不同定位方式百分比

对步行、自行车、公交车、出租车、地铁五种出行方式中不同定位方式所占百分比进行统计,结果如表 2 所示。

表 2 不同定位方式百分比 /%

定位方式	步行	自行车	公交车	出租车	地铁
GPS	90.7	89.0	98.3	92.3	0.8
基站	8.6	9.6	1.6	5.3	82.4
无效	0.7	1.4	0.1	2.4	16.8

由表 2 可知,基于步行、自行车、公交车、出租车出行方式一般在无遮挡空间,GPS 信号良好,大多以 GPS 定位方式为主;而地铁一般运行于地下空间,GPS 信号较差,多以基站定位为主,且在隧道内部还会存在无信号的情况,导致出现较多无效定位数据(占地铁总定位数据的 16.8%)。

根据不同定位方式百分比差异,地铁以基站定位为主、无效定位数据较多的特征很容易将其与其他出行方式区分开。

2.2 速度特征

由于地铁主要是以基站定位为主,根据表 2,基于基站定

位得到的数据项中并无速度信息,且根据不同定位方式百分比能够将其识别出来,故仅分析步行、自行车、公交车、出租车四种出行方式的速度特征。

对不同出行方式速度进行累计频率分布,得到表 3。

表 3　不同出行方式速度累计频率(%)分布

出行方式	速度/km/h								
	0	1	5	6	10	15	40	60	80
步行	19	20	86	95	99	99	100	100	100
自行车	12	12	22	24	44	93	100	100	100
公交车	16	19	30	34	45	55	96	99	100
出租车	23	24	29	30	34	43	71	90	99

由表 3 可得步行、自行车、公交车、出租车的 0 速度百分比较高,均在 10% 以上。其中,步行速度在区间[1,6]比例占 75%,非 0 速度百分比达 93%;自行车速度在区间[5,15]占 71%,非 0 速度百分比达 81%;公交车速度在区间[10,40]比例占 51%,非 0 速度百分比达 60%;出租车速度在区间[10,60]比例占 56%,非 0 速度百分比达 72%。不同出行方式有各自主要的速度分布区间,因此对非 0 速度[1,6]百分比、非 0 速度[5,15]百分比、非 0 速度[10,40]百分比、非 0 速度[10,60]百分比等进行统计。

2.3　加速度特征

对步行、自行车、公交车、出租车四种出行方式的加速度特征进行统计分析,结果如表 4 所示。由表 4 可知,不同出行方式的加速度大小与离散程度不同,其中出租车的各项指标均为最高,而步行各项指标均为最低,在一定程度上可表现出不同出行方式的加速度特征,故需要对加速度均值、加速度标准差、加速度中位数、加速度 85 分位数进行统计。

表 4　不同出行方式加速度(m/s^2)统计结果

出行方式	均值	标准差	中位数	85 分位数	最大值
步行	0.030	0.050	0.022	0.05	0.125
自行车	0.041	0.071	0.025	0.075	0.195
公交车	0.173	0.197	0.104	0.375	0.73
出租车	0.205	0.261	0.122	0.417	1.041

2.4　特征变量

根据对出行特征的分析,选取 GPS 定位方式百分比、基站定位方式百分比、无效定位百分比、速度均值、非 0 速度均值、速度的 85 位分位数、速度标准差、非 0 速度[1,6]百分比、非 0 速度[5,15]百分比、非 0 速度[10,40]百分比、非 0 速度[10,60]百分比、加速度均值、加速度标准差、加速度中位数、加速度 85 分位数等作为特征变量,将这些特征变量作为模型的输入,将不同出行方式作为模型的输出。

3　实证研究

3.1　数据来源与统计

在成都市进行基于智能手机应用程序的数据采集工作,本测试由 14 位学生志愿者组成,共采集到 433 个出行样本,其中步行样本 96 个、自行车样本 78 个、公交车样本 131 个、出租车样本 71 个、地铁样本 57 个。利用 K-CV 方法进行交叉验证,将数据均分为三组,将每个子集数据分别作为验证集,其余两组子集作为训练集,获得三个结果,用这三个预测结果准确率的平均数作为此 K-CV 下分类器的性能指标。

3.2　识别精度

本研究利用 K-CV 交叉验证方法,基于成都市实证数据进行出行方式识别,同时采用决策树和 BP 神经网络方法对同样数据进行训练和出行方式识别,其交叉验证准确率结果如图 3 所示。

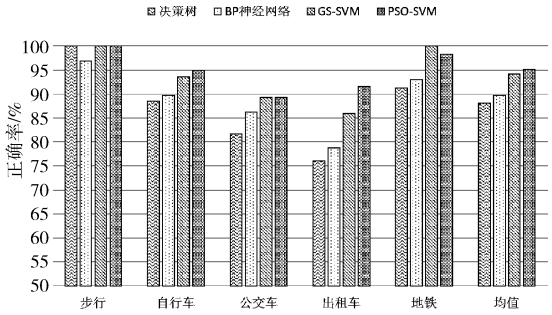

图 3　出行方式识别的交叉验证准确率结果

由图 3 可知,PSO-SVM 模型在总体上对出行方式识别的精度最高,达到 95.1%;在出行方式识别精度上,GS-SVM 模型次之,再次是 BP 神经网络,决策树模型对出行方式识别精度最低。

对于不同出行方式,PSO-SVM 模型在步行、自行车、出租车三种交通方式识别中精度均最高,只是在公交车、地铁出行方式识别上略微差于 GS-SVM 模型。PSO-SVM 模型对于步行识别正确率达到 100%,对自行车、出租车、地铁的识别正确率达到 90% 以上,对公交车的识别正确率也高达 89.29%(略低于 GS-SVM 模型的 89.32%)。

总体而言,PSO-SVM 模型在出行方式识别精度方面对 BP 神经网络、决策树两种模型具有较大优势,同时整体上也优于 GS-SVM 模型。

3.3　时间效率

为进一步了解决策树、BP 神经网络、GS-SVM、PSO-SVM 四种算法执行的时间效率,对其执行时间进行统计,结果如表 5 所示。

表 5　四种算法执行的平均时间　　/s

	决策树	BP 神经网络	GS-SVM	PSO-SVM
时间均值	73.33	112.33	236.67	79.67

由表 5 可见,PSO-SVM 的执行时间虽不是最短,为 79.67 s,仅高于决策树的 73.33 s,但考虑其识别精度为 95.1%,为四种方法中最高,而决策树的识别精度为 88.11%,为四种方法中最低,可见 PSO-SVM 算法的时间效率优良,仍旧为最合适的算法。

4　结束语

本文探讨了基于智能手机应用程序采集数据,利用粒子群算法优化支持向量机参数进行出行方式识别。首先介绍了国内外的相关研究,并对支持向量机与粒子群算法进行了简要介绍,采用粒子群算法优化支持向量机参数模型进行分析;之后,对不同出行方式的数据特征进行分析,得出基于定位方式、速度、加速度等特征变量用于模型的输入。在基于成都市实证数据的基础上,采用同样训练与测试数据对决策树、BP 神经网络、基于网格搜索的支持向量机模型、基于粒子群的支持向量机模型进行交叉验证。结果表明,基于粒子群支持向量机模型的交通出行方式识别精度最高,达到 95.1%,且时间效率优良,该模型能有效应用于出行方式的识别研究。利用智能手机数据对出行方式进行自动识别不但能为交通规划提供大量数据基础,还能为交通实时控制与诱导提供依据[19]。然而,交通出行行为本身非常复杂,出行背景、距离等属性都是影响出行方式识别的重要因素,将来在该方向还有大量工作可开展。

(下转第 3558 页)

实验 1　依次增加节点数目运行并行 FCM 聚类集成算法，节点数目由 1 个增加至 9 个，分别聚类数据样本 1、样本 2 和样本 3，记录每次运行的时间，每组实验分别运行 10 次并计算平均值。运行情况如图 5 所示。随着节点数目的增加，三种大小的样本加速比都呈现出增长的趋势，而随着样本数据量的增加，这种增长趋势越明显，表明了该算法有较强的处理大规模数据的能力，加速比性能良好。

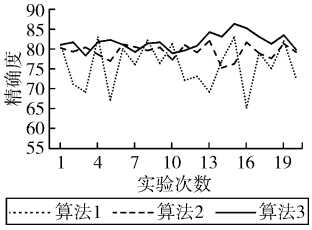

图 4　精确度对比实验

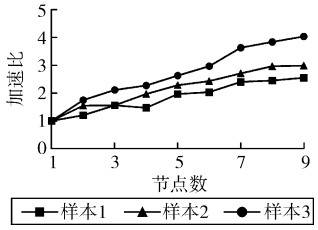

图 5　加速比实验结果

实验 2　将三种大小的样本数据分别在 3 个节点、6 个节点和 9 个节点上进行聚类集成运算，记录每次实验的运行时间，每组实验重复运行 10 次并计算其平均值。运行情况如图 6 所示。随着节点数目的增加，三种样本的算法运行时间都呈现降低趋势，数据量越大，这种减小趋势越明显。从图中可以看出，样本 3 减少的趋势最明显，9 个节点的运行时间基本只有 3 个节点运行时间的三分之一，而其他两种样本只降低了一半左右，所以该算法在处理大规模数据时具有良好的扩展性。

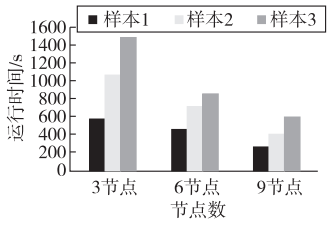

图 6　扩展性实验结果

4　结束语

本文基于 MapReduce 计算模型对 FCM 聚类集成算法进行了并行化的改进，将集成算法拆解成三个步骤，分别由在个 job 进行并行化的实现，经过实验验证，算法在精确度上分别优于单一的并行 FCM 算法和并行 K-means 聚类集成算法，并且具有良好的加速比和扩展性，证明了算法具有较好的并行能力，可以处理较大规模的数据集。但是，该算法需要预先固定 K 值，这是制约算法性能的一个缺陷，将在后续的研究中进行改进。

参考文献：

[1] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61.
[2] Zhou Zhihua, Tang Wei. Clusterer ensemble [J]. Knowledge-Based Systems, 2006, 19(1):77-83.
[3] 罗会兰，孔繁胜，李一啸. 聚类集成中的差异性度量研究[J]. 计算机学报，2007，30(8)：1315-1324.
[4] 孟小峰. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，50(1)：146-169.
[5] Hall L O. Clustering with a genetically optimized approach[J]. IEEE Trans on Evolutionary Computation, 1993, 3(2): 103-112.
[6] Yu Qianqian. Parallel fuzzy C-means algorithm based on MapReduce [J]. Computer Engineering and Applications, 2013, 49(14): 133-137.
[7] Deam J, Ghemawat S. MapReduce: simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.
[8] Zhao Weizhong, Ma Huifang, He Qing. Parallel K-means clustering based on MapReduce[C]// Proc of the 1st International Conference on Cloud Computing and Big Data. 2009: 674-679.
[9] Ngazimbi M. Data clustering using MapReduce [D]. Idaho: Boise State University, 2009.
[10] 冀素琴，石洪波. 基于 MapReduce 的 K-means 聚类集成[J]. 计算机工程，2013，39(9)：84-87.
[11] 武小红. 可能性模糊 C-均值聚类新算法[J]. 电子学报，2008，36(10)：1996-2000.
[12] 王永贵. MapReduce 模型下的模糊 C-均值算法研究[J]. 计算机工程，2014，40(10)：47-51.
[13] 刘秉义. 聚类集成算法及应用研究[D]. 南京：南京理工大学，2012.
[14] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3(5): 583-617.

(上接第 3529 页)

参考文献：

[1] 黄龙中. App 市场使用数据分析[EB/OL]. (2013-06-27) [2015-08-25]. http://mobile.51cto.com/data-401049.htm.
[2] 21N 科技. App 改变世界：2016 年下载量将达 3000 亿次[EB/OL]. (2013-06-03) [2014-12-20]. http://it.21cn.com/mi/a/2013/0603/09/21996474.shtml.
[3] Zheng Yu, Liu Like, Wang Longhao, *et al*. Learning transportation mode from raw GPS data for geographic applications on the Web [C]//Proc of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 247-256.
[4] 李海峰，王炜. 基于神经网络的交通方式选择模型[J]. 公路交通科技，2007，24(7)：132-136.
[5] 张治华. 基于 GPS 轨迹的出行信息提取研究[D]. 上海：华东师范大学，2010.
[6] 闫彭. 基于 AGPS 手机的交通方式识别研究[D]. 北京：北京交通大学，2012.
[7] 汪磊，左忠义，傅军豪. 基于 SVM 的出行方式特征分析和识别研究[J]. 交通运输系统工程与信息，2013，14(3)：70-75.
[8] 陈旭梅，龚辉波，王景楠. 基于 SVM 和 Kalman 滤波的 BRT 行程时间预测模型研究[J]. 交通运输系统工程与信息，2012，12(4)：29-34.
[9] 朱征宇，刘琳，崔明. 一种结合 SVM 与卡尔曼滤波的短时交通流预测模型[J]. 计算机科学，2013，40(10)：248-251.
[10] Yu Bin, Lam W H K, Tam M L. Bus arrival time prediction at bus stop with multiple routes[J]. Transportation Research Part C: Emerging Technologies, 2011, 19(6): 1157-1170.
[11] Hearst M A, Dumais S T, Osman E, *et al*. Support vector machines[J]. Intelligent Systems and Their Applications, 1998, 13(4): 18-28.
[12] 李玉鑑，李玉雄，冷强奎. 基于 LASVM-NC 和 TF. RF 的文本分类方法[J]. 计算机工程与应用，2014，50(10)：136-140.
[13] 王修信，秦丽梅，罗玲，等. 遥感图像森林林型 SVM 分类的多特征选择[J]. 计算机工程与应用，2013，49(20)：256-262.
[14] Wu C, Ho J M, Lee D T. Travel-time prediction with support vector regression [J]. IEEE Trans on Intelligent Transportation Systems, 2004, 5(4): 276-281.
[15] Thissen U, Van Brakel R, De Weijer A P, *et al*. Using support vector machines for time series prediction [J]. Chemometrics and Intelligent Laboratory Systems, 2003, 69(1-2): 35-49.
[16] Kennedy J, Kennedy J F, Eberhart R C. Swarm intelligence[M]. San Francisco: Morgan Kaufman Publishers, 2001.
[17] 张树团，张晓赋，雷涛，等. 基于粒子群算法和支持向量机的故障诊断研究[J]. 计算机测量与控制，2009，16(11)：1573-1574.
[18] 胡坤，余健明. 基于粒子群优化 SVM 的电能质量复合扰动分类的研究[J]. 西安理工大学学报，2012，28(3)：352-355.
[19] Sun D J, Liu Xiaofeng, Ni Anning, *et al*. Traffic congestion evaluation method for urban arterials: case study of Changzhou, China[J]. Journal of the Transportation Research Board, 2014, 2461(1): 9-15.