DOI：10.13334/j.0258-8013.pcsee.2015.01.005 文章编号：0258-8013 (2015) 01-0037-06 中图分类号：TM 715

海量数据下的电力负荷短期预测

张素香[1]，赵丙镇[1]，王风雨[2]，张东[3]

(1. 国家电网公司信息通信分公司，北京市 西城区 100761；2. 北京国电通网络技术有限公司，北京市 丰台区 100070；3. 国家电网公司农电工作部，北京市 西城区 100031)

Short-term Power Load Forecasting Based on Big Data

ZHANG Suxiang[1], ZHAO Bingzhen[1], WANG Fengyu[2], ZHANG Dong[3]

(1. State Grid Information & Telecommunication branch, Xicheng District, Beijing 100761, China;
2. Beijing Guodiantong Networks Technology Co., Ltd., Fengtai District, Beijing 100070, China;
3. State Grid Corporation of China, Xicheng District, Beijing 100031, China)

ABSTRACT: The short-term power load forecasting method had been researched based on the big data. And combined the local weighted linear regression and cloud computing platform, the parallel local weighted linear regression model was established. In order to eliminate the bad data, bad data classification model was built based on the maximum entropy algorithm to ensure the effectiveness of the historical data. The experimental data come from a smart industry park of Gansu province. Experimental results show that the proposed parallel local weighted linear regression model for short-term power load forecasting is feasible; and the average root mean square error is 3.01% and fully suitable for the requirements of load forecasting, moreover, it can greatly reduce compute time of load forecasting, and improve the prediction accuracy.

KEY WORDS: big data; cloud computing; load forecasting; local weighted linear regression

摘要：该文研究海量数据下的短期电力负荷预测方法，基于局部加权线性回归和云计算平台，建立并行局部加权线性回归模型。同时，为剔除坏数据，采用最大熵建立坏数据分类模型，保证历史数据的有效性。实验数据来自已建的甘肃某智能园区。实验结果表明，提出的并行局部加权模型用于短期电力负荷预测是可行的，平均均方根误差为 3.01%，完全满足负荷预测的要求，并极大地减少了负荷预测时间，提高预测精度。

关键词：大数据；云计算；负荷预测；局部加权线性回归

0 引言

电力负荷预测在保证电力系统规划与可靠、经济运行方面具有十分重要的意义。在我国经济高速发展的今天，解决电力负荷预测问题已成为重要而艰巨的任务。高质量的负荷预测需要准确的数学模型，随着现代技术的不断进步和智能用电的深入[1]，负荷预测理论与技术得到很大发展，理论研究逐步深入[2-3]。多年来，电力负荷预测理论和方法不断涌现，神经网络[4-7]、时间序列[8-9]、贝叶斯[10]、模糊理论[11]、小波分析[12]、回归分析[13-14]、支持向量机[15]等技术为电力负荷预测提供了有力的工具。但目前已有的方法仍具有局限性。神经网络方法：一是无法避免在训练过程中产生的学习不足或者是过拟合现象；二是收敛速度慢且易陷入局部极小。时间序列法：对历史数据准确性要求高，短期电力负荷预测时对天气因素不敏感，难以解决因气象因素造成的短期负荷预测不准确问题。回归分析方法是在统计平均意义下定量地描述所观察变量之间的数量关系，往往对数据量有所限制。

随着智能用电海量数据的涌现，必须要寻找一种新的方法满足海量用电大数据分析的要求。目前已有的预测算法无法满足预测速度和预测精度的要求，传统的局部加权线性回归预测用于小数据预测时，具有训练速度快、预测误差率小等优点。但是当数据量非常大时，由于该算法需要为每个测试点寻找近邻，运算量很大，单机运算的时间会达到几个小时或者几天。因此，解决海量数据基础上的预测问题显得十分重要。

本文以智能工业园区海量数据为基础，将局部加权线性回归预测算法和云计算 Mapreduce 模型相

基金项目：国家 863 高技术基金项目(2011AA05A116)。
The National High Technology Research and Development of China 863 Program (2011AA05A116).

结合展开短期电力负荷预测方法研究。该方法首先将海量数据分割成多个数据子块，然后通过云平台将各子块的数据同时进行分析和处理，最后将结果进行归并，该处理过程降低了海量数据的时间处理开销。同时，本文对枚举型数据也进行了处理，并将其加入到距离计算中，提高了预测的准确率。

1 基于云计算的局部加权线性回归模型

1.1 传统局部加权线性回归模型

局部加权线性回归(locally weighted linear regression，LWLR)模型以局部数据为基础拟合多项式回归曲线，观察数据在局部展现出来的规律和趋势。确定预测点周围最邻近的数据点，常用的确定局部数据点的方法为 K 最邻近(k-Nearest Neighbor，KNN)[16-17]算法，其主要思想为计算预测点到特征空间中所有数据点的距离，从中找出距离预测点最近的 k 个点的集合。

设任意 1 个实例用 $X=\{s_1,s_2,\cdots,s_n\}$描述，2 个实例 X_1 和 X_2 之间的距离可以用式(1)得到：

$$d(X_1,X_2)=\sqrt{\sum_{i=1}^{n}(s_i-s_j)^2} \tag{1}$$

建立回归公式：

$$\hat{f}(x)=\omega_0+\omega_1a_1(x)+\omega_2a_2(x)+\cdots+\omega_na_n(x) \tag{2}$$

式中 ω_i 代表根据距离公式(1)计算出的权重大小，其计算公式为

$$\omega_i=\frac{1}{d(x_q,x_i)^2} \tag{3}$$

式中：x_q 为预测点；x_i 为 x_q 的临近点；两者之间距离的倒数为权重的大小。

在式(2)中，ω_0 为回归常数项，ω_1、ω_2、…、ω_n 为回归系数，$\hat{f}(x)$ 为回归预测值。$\alpha_i(x)$表示实例 x 的第 i 个属性值。在拟合以上形式的线性函数到给定的训练集合时，通常采用梯度下降方法，找到使误差最小化的系数 ω_1、ω_2、…、ω_n，即满足：

$$E(x)\equiv\frac{1}{2}\sum_{x\in\text{最近点}}(f(x)-\hat{f}(x))^2 \tag{4}$$

通过满足误差准则满足局部逼近，得到梯度下降训练法则：

$$\Delta\omega_j\equiv\eta\sum_{x\in x\text{ 的 }k\text{ 个最近点}}K(d(x_q,x))(f(x)-\hat{f}(x))a_j(x) \tag{5}$$

式中 η 为学习速率。

1.2 基于云计算的局部加权线性回归算法实现

1.2.1 系统结构

从 1.1 节描述可以看出传统局部加权线性回归算法存在严重缺陷，即当待回归数据增多时，从海量数据中确定近邻数据点集合而产生的计算量是非常巨大的。本文结合云计算技术，将 LWLR 算法和 MapReduce 模型框架相结合，实现电力负荷并行预测。

MapReduce 是一种处理海量数据的并行编程模型和计算框架，它采用一种“分而治之”的思想。因此，本文的并行局部加权线性回归模型包括 3 个阶段：map 阶段、合并阶段、reduce 阶段，每个阶段的数据将以<键, 值>的方式进行交换。系统框架如图 1 所示。

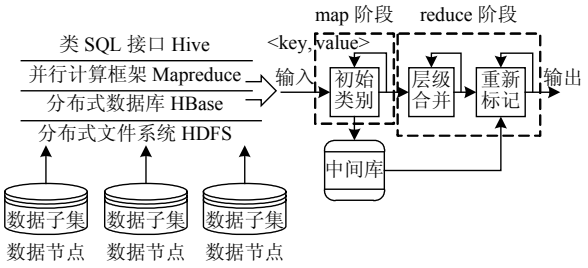


图 1 并行局部加权线性回归系统框架

Fig. 1 Framework of the parallel locally weighted linear regression

1）map 阶段。

首先将输入的数据集合分为若干个数据子集，数据用<key, value>表示。key 是当前数据相对的偏移量，value 值解析成当前数据各个维度的坐标值。基于局部最小距离算法计算出测试点与数据子集的最近 k 个中心点的距离，其运算中间结果将被放入中间库中。

2）合并阶段。

该阶段的任务即将处理完后的数据进行本地层级合并。将中间键值对集合重新排序产生一个新的二元组，相同的键值将被归为一类。

3）reduce 阶段。

reduce 函数首先解析样本个数和相应节点各个维度累加的坐标值，计算出各个数据子集中离预测点最近的 k 个点，并基于混合高斯模型计算出各属性的加权值，该结果将被更新到分布式文件系统中并进行下一次迭代直至算法收敛。

1.2.2 数据来源与处理

1）数据采集网络架构。

如图 2 所示的数据采集网络通过在用能设备信息计量点上部署计量设备，利用工业总线将数据进行集中到采集点，并与不同通信网络对接。采集的数据类型包括用电设备的电能基本参数和电能质量信息等，同时还包括温度、流量等其他能源数据的采集，实现了多能源、全覆盖的数据信息采集。

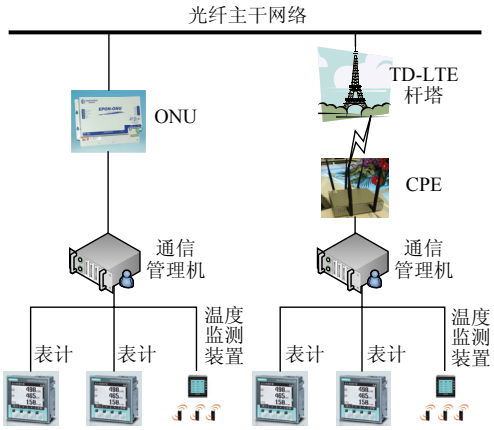


图 2　数据采集网络架构图

Fig. 2　Architecture of data collection network

2）基于最大熵的坏数据分类模型。

由于人为因素或某些特殊原因存在，通常采样得到的异常数据将影响预测结果的精确度及可靠性。本文首先对样本历史数据进行了预处理，基于最大熵算法建立了坏数据分类模型，如图 3 所示。

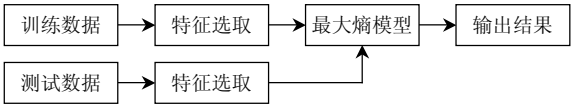


图 3　基于最大熵的坏数据分类模型

Fig. 3　Bad data classification model based on maximum entropy

最大熵原理是在 1950 年由 E. T. Jaynes 提出的，其主要思想是：在用有限知识预测未知假设时，应该选取符合已知假设条件但熵值最大的概率分布。即在已知部分知识的前提下，关于未知分布最合理的推断就是符合已知知识的最不确定或最随机的推断。

在最大熵模型中，信息以特征的方式表达，其中特征为二值特征 $f_i(x,y)$，若 f_i 对模型有用，则构建一个能生成训练样本 $\tilde{p}(x,y)$ 的约束模型(模型期望=经验期望)：

$$P=\{p \mid E_p(f_i)=E_{\tilde{p}}(f_i),1\le i\le k\} \tag{6}$$

最大熵算法提出在与约束集合一致的模型中，选择具有最大熵的 p_*。在有用特征 f_i 的基础上进行推论，它能产生最优化和唯一无偏估计值 p_*。

$$p_*=\arg\max_{p\in C} H(p) \tag{7}$$

式中 $H(p)$为模型 p_*下的熵。

本文数据为时间序列数据，因此，首先进行归一化处理，然后送入最大熵模型中迭代。所有元素按照公式(8)进行标准归一化处理。

$$r_i=\frac{x_i-\min(x_i)}{\max(x_i)-\min(x_i)} \tag{8}$$

3）枚举型数据。

本文将时间、温度等数据用于并行 LWLR 算法，但由于以上数据具有连续特点，本文将其转化为向量。如时间类数据：一周为 7 天，向量的维度为 7，则设时间向量为$\{t_1,t_2,\cdots,t_7\}$，如星期日被表示为向量$\{0,0,0,0,0,0,1\}$。

1.2.3　基于云计算的 LWLR 预测算法

基于 MapReduce 的 LWLR 预测算法如图 4 所示。在 LWLR 算法中，首先解决 Map 个数问题。通过读取数据源及其数据结构、并行度、增量字段、异常处理方式等多种参数信息，并根据增量字段当前的最大值对数据集进行划分和调整，确定 Map 个数。其次，利用 KNN，对每个 Map 所处理的数据块选择离预测点最近的 K 个点；最后将每个 Map 的 K 个点与预测点进行距离比较，筛选出最小的 K 个点，并基于混合高斯模型计算权重，确定参数，完成了模型建立任务。

Map 阶段：
$\exists x_q$
$\forall D=\{D_1,D_2,\cdots,D_n\}$
foreach D_i
$\{$foreach$\{x_i\in D_i\}$
$\{\mu_j\}\leftarrow\arg\min\{d(x,\ x_i),j=1,2,\cdots,K\}\}$
Reduce 阶段：
$\exists x_q$
$\forall$map$=\{\mu_1,\mu_2,\cdots,\mu_n\}$
foreach μ_i
$\{$foreach$\{y_i\in\mu_i\}$
$\{\beta_i\}\leftarrow\arg\min\{d(x_q,y_i),i=1,2,\cdots,K\},n=1,2,\cdots,K\}$
基于混合高斯模型计算 k 个点的加权值 ω_i
$\hat{f}(x)=\omega_0+\omega_1a_1(x)+\omega_2a_2(x)+\cdots+\omega_na_n(x)$

图 4　基于云计算的 LWLR 预测算法

Fig. 4　LWLR forecasting based on cloud computing

2　负荷预测实验及结果分析

2.1　负荷预测误差评测指标

由于电力负荷预测是通过历史数据对未来电

力负荷的估算，因此预测值与实际值存在差距，产生电力负荷预测误差。产生误差的原因很多，归纳起来主要包括：1）数学模型的简化和忽略各种因素的关系；2）历史数据不够完整；3）参数选取不当造成误差。

本文采用的评测指标如下：

设 $y(i)$和 $\hat{y}(i)$ 分别表示 i 时刻的实际负荷值和预测值，则有：

绝对误差：

$$E = y(i) - \hat{y}(\mathrm{i}) \tag{9}$$

相对误差：

$$e_1 = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y(i)-\hat{y}(i)}{y(i)}\right| \times 100\% \tag{10}$$

式中 e_1 为日平均误差。由于预测误差有正负，为了避免正负相抵消，计算其平均数的时候取误差的绝对值。

$$e_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y(i)-\hat{y}(i)}{y(i)}\right|^2} \times 100\% \tag{11}$$

式中 e_2 为均方根误差。均方根误差指标加强了数值大的误差的作用，提高了该指标的灵敏性。

2.2 实验数据

本文的数据来源为甘肃省某电网企业所采集的负荷数据和天气数据，训练数据范围为 2011 年 11 月 24 日至 2011 年 11 月 30 日的用电数据，每个设备的采样间隔周期为 15 min，如表 1 所示。预测 2011 年 12 月 1 日的电力负荷，如表 2 所示。同时考虑了温度、湿度、工作日、节假日、季节等负荷影响因素对电力用户负荷波动的影响，通过计算与负荷的关联强度，为建立更加精确的负荷预测模型提供依据。

本文的数据表达为：负荷时间序列为 $X_1, X_2, \cdots, X_n$；$x_{1,i}$ 为负荷数据；$x_{2,i}$ 为温度序列；$x_{3,i}$ 为湿度序列；以此类推。

表 1 训练数据

Tab. 1 Training data

日期	小时	分钟	最高温度	最低温度	电量/(kW·h)
24	00	15	7	−3	52 679.184 2
24	00	30	7	−3	58 495.673 0
24	00	45	7	−3	57 386.641 0
⋮	⋮	⋮	⋮	⋮	⋮
30	23	30	3	−5	84 704.784 1
30	23	45	3	−5	72 975.840 8

表 2 2011 年 12 月 1 日的实际负荷数据

Tab. 2 Actual load data on December 1，2011

小时	分钟	最高温度	最低温度	电量/(kW·h)
00	15	8	−1	53 100.606 1
01	30	8	−1	53 105.972 3
02	45	8	−1	54 000.864 3
⋮	⋮	⋮	⋮	⋮
23	30	8	−1	72 000.504 5
23	45	8	−1	71 100.308 5

2.3 实验结果

1）并行局部加权线性回归算法与传统算法对比

从图 5 所示的结果可以看出，在小数据样本时，两者之间的预测时间相差不大，相反，传统线性回归方法所需要的时间略优于并行局部线性加权算法，原因在于：并行局部线性加权算法在小样本集下仍将数据分成若干个子样本集，不同数据子集之间的通讯代价增高反而影响预测速度；但随着样本集的增大，预测算法所需的迭代时间有了明显不同，并行加权线性回归算法所需的时间要远远小于传统方法。

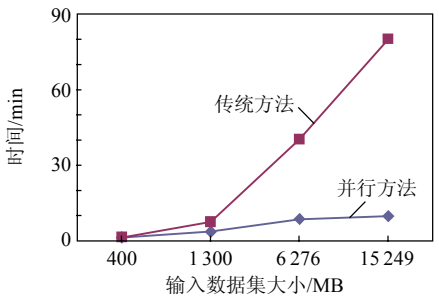


图 5 传统算法和并行算法所需时间对比

Fig. 5 Consume time contrast between the traditional algorithm and the parallel algorithm

2）本文实验结果

本文基于并行局部加权线性回归算法得到的负荷预测值与实际负荷值对比结果如表 3 所示。

如图 6、7 所示，预测值的曲线与实际值的曲线趋势相似，其均方根误差平均值为 3.01%。预测

表 3 对比结果

Tab. 3 Comparison results

序号	预测值	实际值	误差/%
1	50 976.581 8	53 100.606 1	−4.0
2	54 386.020 9	53 105.972 3	2.4
3	56 034.593 1	54 000.864 3	3.8
4	52 065.280 6	53 400.287 8	−2.5
⋮	⋮	⋮	⋮
96	73 517.718 9	71 100.308 5	3.4

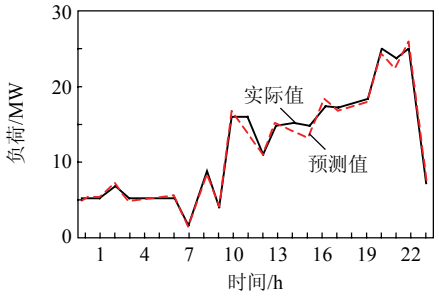


图 6　负荷预测对比图

Fig. 6　Load forecasting contrast curve

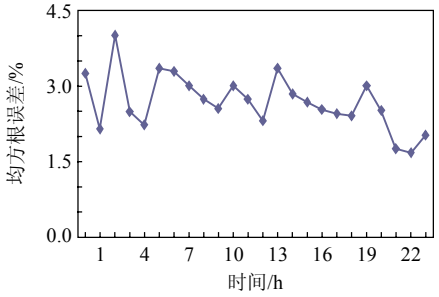


图 7　负荷预测的均方误差曲线

Fig. 7　Mean square error curve of load forecasting

结果的误差符合负荷预测的误差标准。证明基于云计算的局部加权线性回归方法是可行的，该系统软件一直在某智能园区运转正常，为电力企业管理该园区的电力负荷起到了很重要的作用。

3　结论

本文针对传统局部加权线性回归算法的严重缺陷，研究了海量数据电力负荷短期预测问题。通过最大熵剔除坏数据模型进行数据预处理后，将具有并行编程模型和计算框架的 Mapreduce 和局部加权线性回归算法相结合，提出了并行局部加权线性回归算法，解决了海量数据的计算量问题，预测所耗的时间大大缩短，同时还保证了预测精度满足负荷预测要求。

下一步，将围绕多模型相结合的方法解决负荷预测中的因素关联问题。

参考文献

[1] 中国电力大数据发展白皮书[M]. 北京：中国电力出版社，2013.
The white paper for the development of Chinese electric power big data[M]. Beijing: China Electric Power Press, 2013(in Chinese).

[2] 钟清，孙闻，余南华，等. 主动配电网规划中的负荷预测与发电预测[J]. 中国电机工程学报，2014，34(19)：3050-3055.
Zhong Qing，Sun Wen，Yu Nanhua，et al. Load and power forecasting in active distribution network planning[J]. Proceedings of the CSEE，2014，34(19)：3050-3055(in Chinese).

[3] 肖白，周潮，穆钢. 空间电力负荷预测方法综述与展望[J]. 中国电机工程学报，2013，33(25)，78-92.
Xiao Bai，Zhou Chao，Mu Gang. Review and prospect of the spatial load forecasting methods[J]. Proceedings of the CSEE，2013，33(25)：78-92(in Chinese).

[4] Liang Z S. The short term load forecast of power system based on adaptive neural network[J]. Journal of Northeast China Institute of Electric Power Engineering，1994，14(1)：27-35.

[5] 谢开贵，李春燕，周家启. 基于神经网络的负荷组合预测模型研究[J]. 中国电机工程学报，2002，22(7)：85-89.
Xie Kaigui，Li Chunyan，Zhou Jiaqi. Research of the combination forecasting model for load based on artificial neural network[J]. Proceedings of the CSEE，2002，22(7)：85-89(in Chinese).

[6] 刘玲，严登俊，龚灯才，等. 基于粒子群模糊神经网络的短期电力负荷预测[J]. 电力系统及其自动化学报，2006，18(3)：47-50.
Liu Ling，Yan Dengjun，Gong Dengcai，et al. New method for short term load forecasting based on particle swarm optimization and fuzzy neural network[J]. Proceedings of the CSU-EPSA，2006，18(3)：47-50(in Chinese).

[7] 傅忠云. 粒子群优化 BP 算法在电力系统短期负荷预测中的应用[J]. 重庆工学院学报(自然科学版)，2007，21(10)：93-96.
Fu Zhongyun. Application of PSO-BP algorithm in electric power system short-term load forecast[J]. Journal of Chongqing Institute of Technology (Natural Science Edition)，2007，21(10)：93-96(in Chinese).

[8] Hagan M T，Behr S M. The time series approach to short-term load forecasting[J]. IEEE Transactions on Power System，1987，2(3)：25-30.

[9] 赵宏伟，任震，黄雯莹. 考虑周周期性的短期负荷预测[J]. 中国电机工程学报，1997，17(3)：211-213.
Zhao Hongwei，Ren Zhen，Huang Wenying. Short-term load forecasting considering weekly period based on PAR[J]. Proceedings of the CSEE，1997，17(3)：211-213(in Chinese).

[10] 陶文斌，张粒子，潘弘，等. 基于双层贝叶斯分类的空间负荷预测[J]. 中国电机工程学报，2007，27(7)：13-17.
Tao Wenbin，Zhang Lizi，Pan Hong，et al. Spatial electric load forecasting based on double-level Bayesian classification[J]. Proceedings of the CSEE，2007，27(7)：13-17(in Chinese).

[11] Bakirtzis A G，Theocharis J B. Short term load forecasting using fuzzy neural networks[J]. IEEE Transactions on

Power System，1995，10(3)：1518-1524.

[12] 姚李孝，刘学琴.基于小波分析的月度负荷组合预测[J]. 电网技术，2007，31(19)：65-68.
Yao Lixiao，Liu Xueqin．A wavelet analysis based combined model for monthly forecasting[J]．Power System Technology，2007，31(19)：65-68(in Chinese).

[13] 雷绍兰，孙才新，周湶，等．电力短期负荷的多变量时间序列线性回归预测方法研究[J]．中国电机工程学报 2006，26(2)：25-29.
Lei Shaolan，Sun Caixin，Zhou Quan，，et al．The research of local linear model of short-term electrical load on multivariate time series[J]．Proceedings of the CSEE，2006，26(2)：25-29(in Chinese).

[14] 张伏生，汪鸿，韩悌，等．基于偏最小二乘回归分析的短期负荷预测[J]．电网技术，2003，27(3)：36-40.
Zhang Fusheng，Wang Hong，Han Ti，et al．Short-term load forecasting based on partial least-squares regression[J]．Power System Technology，2003，27(3)：36-40(in Chinese).

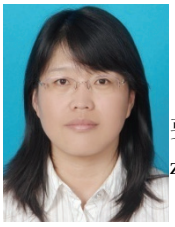
[15] 牛东晓，谷志红，邢棉，等．基于数据挖掘的 SVM 短期负荷预测方法研究[J]．中国电机工程学报，2006，26(18)：6-12.
Niu Dongxiao，Gu Zhihong，Xing Mian，et al．Study on forecasting approach to short-term load of SVM based on data mining[J]．Proceedings of the CSEE，2006，26(18)：6-12(in Chinese).

[16] Thomas M C，Peter E H．Nearest neighbor pattern classification[J]．IEEE Transaction Theory，1967，13(1)：21-27.

[17] Bremner D，Demaine E，Erickson J，et al．Output-sensitive algorithms for computing nearest-neighbor decision boundaries[J]．Discrete and Computational Geometry，2005，33(4)：593-604.

收稿日期：2014-09-07。

作者简介：



张素香

张素香(1973)，女，博士，副教授，主要研究方向为数据挖掘、智能用电，zsuxiang@163.com。

(编辑 李泽荣)