doi:10.16104/j.issn.1673-1891.2017.02.006

# 基于数据挖掘的客户行为分析
## ——以中小餐饮企业为例

王召义
(安徽商贸职业技术学院经济贸易系,安徽 芜湖 241002)

**摘要**:随着“夜经济”日益成为商业活动的消费亮点,众多商家一直致力于通过满足消费者多样化的需求和体验来促使销售额的稳步提升。但是,很多店铺的夜间客流量并不理想,店铺的整体收益难以提高,是商家长期以来所面临的难题。借助多元线性回归和支持向量机对客户行为进行研究,发掘关键影响因素,预判客户消费行为。实证研究表明,分析客户行为可以为企业及时调整营销策略提供支持。

**关键词**:多元线性回归;支持向量机;LIBSVM;解释变量
**中图分类号**:F713.55 **文献标志码**:A **文章编号**:1673-1891(2017)02-0017-04

# Customer Behavior Analysis Based on Data Mining: a Case Study on Small and Medium Catering Enterprises

WANG Zhao-yi
(Department of Economics and Trade, Anhui Business College, Wuhu, Anhui 241002, China)

**Abstract**: At present, the "night economy" has become a highlight activity, and all businesses are working to meet the diversified market needs and improve the user's experience. While, the results are not satisfactory, and the total income is hard to increase. In this paper, multiple linear regression and support vector machines are used to study customer behavior, identify key influencing factors and predict customer behavior. The empirical research shows that, with the help of data mining technology, it can help enterprises to timely adjust advertising and marketing strategies to provide targeted service.

**Keywords**: multiple linear regression; support vector machine; LIBSVM; explanatory variables

## 1 问题由来

夜生活的经济效益(简称夜经济)是指人们在夜间所从事的生产性活动及消费性经济活动,目前已成为许多城市新的经济增长点[1]。因此,中小餐饮企业竭尽全力,变着花样玩促销,期望通过满足消费者多样化的需求和体验来促使销售额的稳步提升。但是,未经过市场调研和数据分析盲目延长夜间营业时间,不仅会增加营业成本,还会带来负面效果[2]。经过调查,有不少店铺的夜间营业额都不理想且难以提供有针对性的客户服务。因此,借助数据挖掘技术来发掘客户行为的关键影响因素并预判客户消费行为,为企业解决以上难题提供方法,并为企业制定营销服务策略提供决策支持。本文借助多元线性回归和支持向量机对上述问题进行分析探讨,以期对商家夜间竞争力的提升提供指导。

## 2 研究基础

### 2.1 多元线性回归

多元回归分析可以用于考察输出结果与多个解释变量之间存在的关联性,其数学模型如下:

设$x=(x_1,x_2,\cdots,x_p)$是解释变量,$y$是输出结果,如果$y$与$X$是线性的,那么进行$n$次试验后,可得$n$组数据:

$(y_i,x_{i1},x_{i2},\cdots,x_{ip}),(i=1,2,\cdots,n)$

此刻,$y$与$x$有如下线性关系:

**收稿日期**:2017-02-20
**基金项目**:2017年安徽省高校优秀青年人才支持计划重点项目(gxyqZD2017110);大学生创客实验室建设计划“Big Data& Analytics Hub 创客实验室”(2016ckjh088);安徽省高校自然科学研究重点项目“基于改进RFM模型的电子商务协同过滤推荐算法研究”(KJ2016A253);安徽省教学研究项目:基于校企合作的电子商务高素质技能型人才培养模式研究(2015jyxm751);2017年“三平台两基地”应用研究项目:基于SVM的网络创业过程性评价研究(2017ZDF05)。
**作者简介**:王召义(1983— ),男,安徽宿州人,讲师,硕士,研究方向:电子商务。

$y_1=b_0+b_1x_{11}+b_2x_{12}+\cdots+b_px_{1p}+\varepsilon_1$

$y_2=b_0+b_1x_{21}+b_2x_{22}+\cdots+b_px_{2p}+\varepsilon_2$

……

$y_n=b_0+b_1x_{n1}+b_2x_{n2}+\cdots+b_px_{np}+\varepsilon_n$

其中，$b_0,b_1,b_2,\cdots,b_p$是$x$的相应系数，$\varepsilon_i$是相应误差。

### 2.2 支持向量机

支持向量机（Support Vector Machine，SVM）是Corinna Cortes和Vapnik等于1995年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多独特优势，并能够推广应用到函数拟合等其他机器学习问题中[3]。

C-SVC模型是SVM的二分类模型，原理如下：

（1）设 $T=\{(x_1,y_1),\cdots,(x_i,y_i)\}\in(X\times Y)^l$

其中$x_i\in X=R^n$，$y_i\in Y=\{1,-1\}(i=1,2,\cdots,l)$；$x_i$为特征向量。

（2）参数寻优并求解下式最优解：

$$\min\frac{1}{2}\sum_{i=1}^{j}\sum_{i=1}^{i}y_iy_j\alpha_i\alpha_jK(x_i,x_j)-\sum_{j=1}^{l}\alpha_j$$

$$s.t.\sum_{i=1}^{l}y_i\alpha_i=0,$$

$$0\le\alpha_i\le C,i=1,2,\cdots,l$$

得到最优解：$\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_i^*)^T$。

（3）计算阀值$b^*$

$$b^*=y_i-\sum_{i=1}^{i}y_ia_i^*K(x_i-x_j)\qquad(0<\alpha_j^*<C)$$

（4）构造决策函数：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}a_i^*y_iK(x,x_i)+b^*)$$

$f(x)$是取值为1和-1的函数。sgn（*）为非负数时，$f(x)=1$，为负数时，$f(x)=-1$。

## 3 研究模型

基于数据挖掘的客户行为分析，核心是使用数据挖掘技术发掘关键影响因素和实现预判消费行为功能。研究模型主要分为3个部分：数据准备、数据分析和结论，具体流程如图1所示。

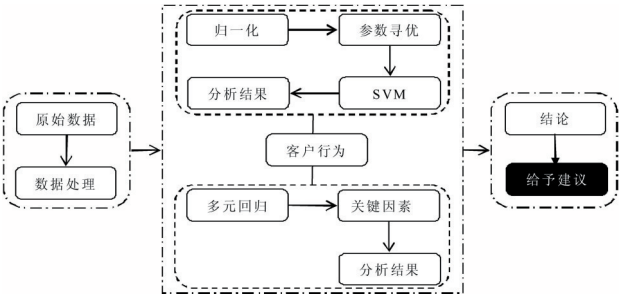


图1 研究模型

1）数据准备。根据企业的营销数据或调查结果，确定输出结果、分析单位和解释变量。根据数据挖掘技术对输入数据的格式要求，对原始数据做预处理。

2）数据分析。数据分析分为2个阶段展开工作：第一阶段是对预处理后的数据做多元线性回归分析，发掘关键影响因素；第二阶段是用支持向量机对客户消费行为进行预判。

3）结论。以分析结果为参考，为企业调整营销策略提供建议。

## 4 实证研究

有一家连锁餐厅A位于市区步行街某商业广场内，其工作日的夜间客流量很不理想，店铺的整体收益也难以提高。为此，A开展了一次市场调查活动，调查对象仅限于“过去3个月间（不限时间段）至少光顾过一次A店”这一设定给出肯定回答的客户。调查问卷包括9个问题：年龄、性别、婚姻情况、广告印象、光顾频率、消费金额、菜品种类等。

表1 解释变量

| 代号 | 解释变量 | 数值 |
|---|---|---|
| $x1$ | 调查ID | 1，2，3，……，1000 |
| $x2$ | 年龄 | |
| $x3$ | 性别 | 1代表男性，0代表女性 |
| $x4$ | 婚姻情况 | 0代表未婚，1代表已婚 |
| $x5$ | 广告印象 | 1代表印象很差；2代表没什么好印象；3代表没看到广告，不知道；4代表印象不错；5代表印象非常好 |
| $x6$ | 光顾次数 | 夜间（18:00—23:00）光顾次数 |
| $x7$ | 光顾人数 | 最常有的状况是几位一起夜间光顾，数值（0,1,2,3,4,5）代表有几人同行 |
| $x8$ | 消费金额 | 夜间平均消费金额 |
| $x9$ | 套餐 | 夜间光顾时，你自己点过的菜品；1代表点过，0代表没有点过。 |
| $x10$ | 面类 | |
| $x11$ | 盖浇饭 | |
| $x12$ | 甜品 | |
| $x13$ | 其他小食 | |
| $x14$ | 饮料 | |
| $x15$ | 酒类 | |
| $x16$ | 总消费金额 | 光顾次数与消费金额的乘积 |

表2 调查结果

| $x1$ | $x2$ | $x3$ | $x4$ | $x5$ | $x6$ | $x7$ | $x8$ | $x9$ | $x10$ | $x11$ | $x12$ | $x13$ | $x14$ | $x15$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 35 | 0 | 1 | 5 | 1 | 1 | 1700 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| 2 | 57 | 0 | 1 | 3 | 3 | 3 | 3 600 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 3 | 42 | 1 | 0 | 3 | 2 | 0 | 700 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 4 | 29 | 1 | 0 | 5 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| 5 | 54 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |

为此,A希望解决以下2个问题:

(1)采取什么措施来提高夜间营业额;

(2)预判客户是否会夜间光顾店铺。

根据调查问卷及调查结果,可整理出如表1所示的解释变量。

调查结果如表2所示。

### 4.1 多元线性回归分析

经过分析讨论,输出结果确定为"过去3个月夜间时段总消费金额",即光顾次数($x6$)与消费金额($x8$)之乘积。相较于光顾次数或人数、消费金额更能切实关系到营业额的走向;同时,将"光顾次数与消费金额"进行乘法运算能解决A"夜间客流量不理想"的难题。

解释变量要根据输出结果做出适当调整,删除无意义的解释变量,增加输出结果"总消费金额($x16$)"。多元回归分析选择以下解释变量进行分析:年龄($x2$)、性别($x3$)、婚姻状况($x4$)、广告印象($x5$)、套餐($x9$)、面类($x10$)、盖浇饭($x11$)、甜品($x12$)、其它小食($x13$)、饮料($x14$)、酒类($x15$)。以"总消费金额($x16$)",用SPSS22.0对数据进行多元回归分析,分析结果如表3。

**表3　多元回归分析结果**

| 变量 | 非标准化系数 | | $t$ | 显著性 |
|---|---|---|---|---|
| | $B$ | 标准错误 | | |
| (常量) | 679.536 | 1 209.978 | 0.562 | 0.575 |
| 年龄 | 32.740 | 14.679 | 2.230 | 0.026 |
| 性别 | 1 097.214 | 313.604 | 3.499 | 0.000 |
| 婚姻 | 279.135 | 331.669 | 0.842 | 0.400 |
| 广告印象 | 864.516 | 256.217 | 3.374 | 0.001 |
| 套餐 | -3 561.647 | 708.029 | -5.030 | 0.000 |
| 面类 | -569.993 | 329.919 | -1.728 | 0.084 |
| 盖浇饭 | -16.063 | 306.705 | -.052 | 0.958 |
| 甜品 | 348.929 | 578.807 | 0.603 | 0.547 |
| 其他小食 | 476.736 | 473.175 | 1.008 | 0.314 |
| 饮料 | -113.887 | 408.879 | -0.279 | 0.781 |
| 酒类 | 1 065.740 | 410.277 | 2.598 | 0.010 |

由表3的"显著性"列数值可知,年龄、性别、广告印象、套餐和酒类的显著性值均小于0.05,其值分别为0.026、0.000、0.001、0.000、0.010。也即,这5个变量对结果的影响是合理的、可接受的。

年龄、性别、广告印象、套餐和酒类等变量对应的回归系数分别为32.740、1 097.214、864.516、-3 561.647、1 065.740,既有正值也有负值,含义自然也有所不同。除"套餐"变量外,其他解释变量的回归系数都为正值,即倾向于"解释变量每增加1,输出结果总消费金额会随之有相应增长"。就"广告印象"变量来说,每增加1点印象值,总消费金额平均增加864.516。年龄、性别、酒类等解释变量也是如此。

需要注意的是"套餐"变量。这一解释变量的回归系数为-3 561.647,这意味着点套餐的顾客的总消费金额存在大幅度降低的倾向。这或许是点套餐的顾客工作日夜间不光顾或基本不消费的缘故。

从分析结果来看,可能会对连锁餐厅A夜间营业额的增长有所贡献的顾客,不论婚姻情况如何,主要是年龄较大的男性。其中,从没有点过套餐、不管时间早晚点酒类,即喜欢在这家店里就着下酒小菜小酌的顾客在夜间的总消费金额比较高。

为此,连锁餐厅A可采取以下措施来提高夜间营业额:(1)在年龄层偏高的男性常看的杂志或媒体上打广告;(2)将夜间的营业模式转型为"适合晚间小酌的店"。

具体来说,可以对酒水单进行修改调整,增加较符合四五十岁男性口味的白酒和啤酒的种类,在小食菜单里也要充实一些可以作为下酒菜的品种。可以将A形象定性为"比小酒馆安静、实惠,适宜喜欢夜间小酌的顾客的店"。遵循这一方针,重新制定宣传策略,将有可能进一步提高广告效果。

不过,仅凭此次调查结果就大刀阔斧地进行战略调整的风险太大。应该首先在小范围内进行验证,在看到效果之后再考虑总体战略上的调整。

### 4.2 支持向量机分析

SVM是一种典型的2类分类器,即它只回答属于正类还是负类的问题。把"夜间至少光顾过一次"做为输出结果,用1代表光顾过,-1代表没有光顾过。同时,对原始数据进行调整,将没有分析意义的列排除,例如调查ID。另外,把与输出结果存在绝对关系的列如总消费金额、光顾次数、消费金额等也一并排除。

目前,应用最广泛的SVM工具是台湾林智仁(Chih-Jen Lin)教授等开发设计的一个简单、易于使用和快速有效的SVM模式识别与回归的LibSVM软件包[4]。本文选择在MATLAB下使用LibSVM,完成相应的数据分析工作。

#### 4.2.1 数据预处理

LibSVM的数据格式为:label index1:value1 index2:value2 ... indexn:value$n$。其中label为目标值,index1……index$n$为从1到$n$的自然数,value为对应的特征值,数据之间用空格隔开[5]。处理后的数据格式如图2所示。选择1 000条数据作为研究

对象，把前700条数据作为训练数据集，后300条数据作为测试数据集。

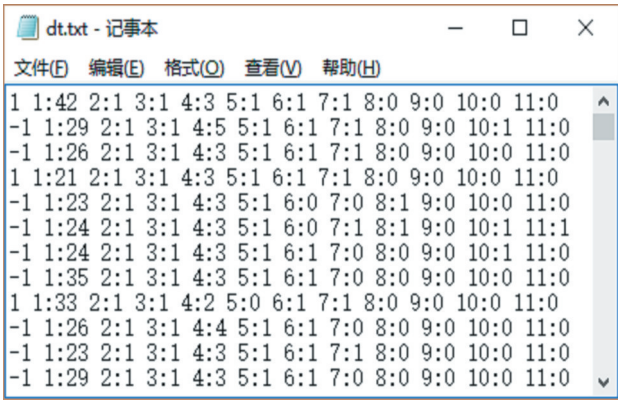


图2 数据格式

为了避免一些特征值范围过大而另一些特征值范围过小或避免在训练时为了计算核函数而计算内积的时候引起数值计算的困难[6]。通常需要将数据缩放到[-1,1]或者是[0,1]之间。在MATLAB中，Mapminmax函数可以实现数据归一化处理。注意，需要将训练数据和测试数据一起归一化。

### 4.2.2 参数寻优

对于采用RBF核函数的支持向量机的主要参数是惩罚系数$c$和核函数宽度$g$[7]。目前常用的优化选取方法就是让$c$和$g$在一定的范围内取值，对于取定的$c$和$g$，把训练集作为原始数据集并利用K-fold Cross Validation（记为K-CV）方法得到在此组$c$和$g$下训练集验证分类准确率，最后取使训练集验证分类准确率最高的那组$c$和$g$作为最佳的参数[8]。但有一个问题就是可能会有多组的$c$和$g$对应于最高的验证分类准确率，这种情况怎么处理？这里采用的手段是选取能够达到最高验证分类准确率中参数$c$最小的那组$c$和$g$作为最佳的参数，如果对应最小的$c$有多组$g$，就选取搜索到的第一组$c$和$g$作为最佳参数[9]。这样做的理由是：过高的$c$会导致过学习状态发生，即训练集分类准确率很高而测试集分类准确率很低（分类器的泛华能力降低），所以在能够达到最高验证分类准确率中的所有的成对的$c$和$g$中认为较小的惩罚参数$c$是更佳的选择对象[10]。

函数SVMcgForClass[11]可以实现以上选择$c$和$g$的算法，详细内容可参阅文献[12]，其用法如下：

```
[bestacc,bestc,bestg]
=SVMcgForClass(train_label,train,cmin,cmax,gmin,gmax,v,cstep,gstep,accstep)
```

使用函数SVMcgForClass选择$c$和$g$的代码及运行结果如下。

```
[bestacc,bestc,bestg]
=SVMcgForClass(train_label,train_data_scale,-10,10,-10,10,5,0.5,0.5,4.5)
```

运行结果为：bestacc =72.714 3%，bestc =512，bestg =0.007 8。

图2展示了此次参数选择结果3D视图。

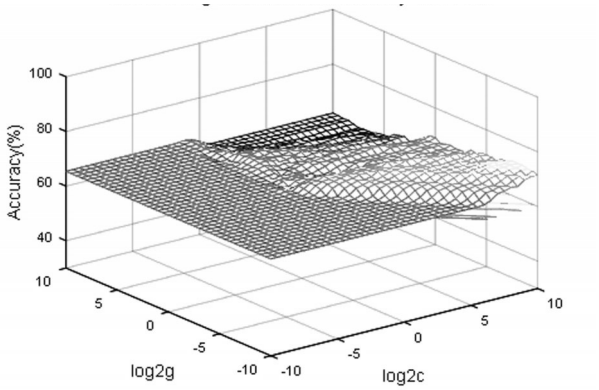


图2 SVC参数选择结果3D视图

### 4.2.4 训练与预测

建立分类训练模型代码如下：

```
model=svmtrain(train_label,train_data_scale,'-s 0 -t 2 -c 512 -g 0.0078')
```

运行结果如下：

```
model =
    Parameters: [5x1 double]      %结构体变量，依次保存的是 -s -t -d -g -r 等参数
      nr_class: 2                 %分类的个数
       totalSV: 499               %总的支持向量个数
           rho: 0.2698            %b=-model.rho
         Label: [2x1 double]
   sv_indices: [499x1 double]
         ProbA: []
         ProbB: []
           nSV: [2x1 double]      %每一类的支持向量的个数
     sv_coef: [499x1 double]      %支持向量的系数
           SVs: [499x11 double]   %具体的支持向量，以稀疏矩阵的形式存储
```

即$w*x+b=0$，其中$w=model.SVs'*model.sv_coef$，$b=-model.rho$。$w$是高维空间中分类超平面的法向量，$b$是常数项。

建立预测模型代码如下：

```
[predictlabel,acc,decision_values]=svmpredict(test_label,test_data_scale,model)
```

运行结果如下：

```
Accuracy = 77% (231/300) (classification)
```

得到预测的准确率为77%。

分析预测结果，发现广告印象为3，即回答“没看到广告，不知道”的调查对象中夜间光顾过本店的人只占19.35%。与之相对，广告印象不为3的调查对象中，夜间光顾过本店的人占到了55.77%。对于这一结果，我们可以做2种解释：一种是“没有看过广告，或者对其没印象的人基本不会光顾本店”；另一种解释是“原本就对这家店没有兴趣，所以即便看到过广告，

（下转第32页）

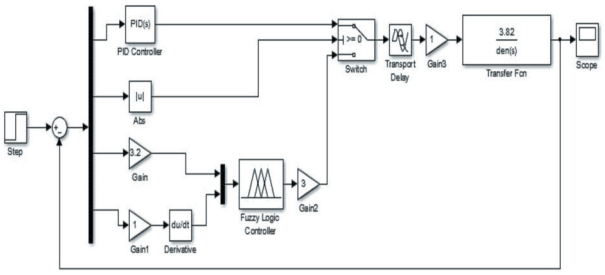


图4 控制系统仿真框图

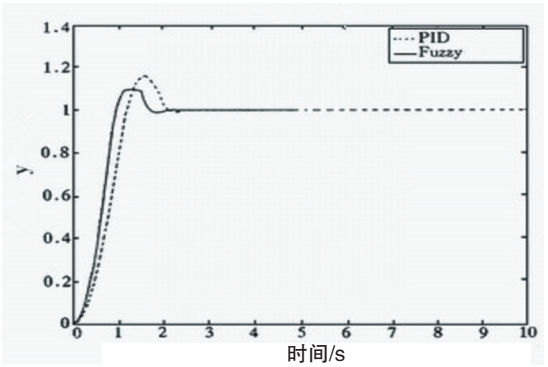


图5 模糊PID阶跃响应曲线

## 5 结语

本文从快速锻造油压机的结构和控制系统出发，深入探讨了自适应模糊PID技术在锻压机上的应用。建立了基于自适应模糊PID控制器的锻压机液压系统模型并进行了仿真。从仿真结果可以看出，自适应模糊PID算法把PID控制简便性、可靠性与模糊控制的智能型、灵活性融为一体，发挥了传统PID控制与自适应模糊控制的各自长处。

**参考文献：**

[1] 郭会娟.锻压机组控制系统[D].天津：天津工业大学,2005,20-23.

[2] 李士勇.模糊控制和智能控制理论与应用[M].哈尔滨：哈尔滨工业大学出版社,1990,40-42.

[3] 陈燕庆.工程智能控制[M].西安：西北工业大学出版社, 2003.

[4] 席爱民.模糊控制技术[M].西安：西安电子科技大学出版社,2005.

[5] 贾维宏.模糊PID在锻压机液压系统中的仿真研究[D].太原：太原理工大学,2012.

[6] 李毅波.重型模锻压机多学科集成建模与低速稳定性研究[J].长沙：中南大学,13-17,2013.

[7] 李艳杰,崔天宇,王海,等.比例阀控液压缸位置PID闭环控制的PLC软件实现 [J].沈阳理工大学学报,2013,32(4)：37-40.

[8] 窦丽娟.65MN自由锻压机液压系统的设计与仿真[D].秦皇岛：燕山大学,2011.

（上接第20页）

也完全没留下一点儿印象”。而至于哪种假设正确，就要根据在制作并发布容易给每个人留下深刻印象的广告的情况下，客流量的增长程度来判断了。

至此，连锁餐厅A可以按照此模型去分类预测客户，以判别其是否会夜间光顾。虽然准确率不是太高，但是与“还不如瞎猜的判别”相比，准确率要高很多，而且具有一定的可靠性和可信度。

## 5 结语

本文建立了一种基于多元线性回归和支持向量机的客户行为分析模式，并进行实证研究。研究表明，该模型可以挖掘客户消费习惯、预测客户消费行为，并在提高投资回报率、降低运营成本等方面提高企业的核心竞争力。

**参考文献：**

[1] 李经龙,张小林,马海波.夜生活与夜经济:一个不容忽视的生产力[J].生产力研究,2008(1):60-61+157.

[2] 卫志民. 解读“夜经济”[J].经济导刊,2004(9):88-90.

[3] 逄淑卉.基于支持向量机的纺织品服装安全风险评价模型研究[D].上海：东华大学,2011.

[4] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines, 2001[EB/OL].[2016-10-30]. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] 徐晟皓,杨楠堃,易梦乔.基于支持向量机的消费者行为分类方法[J].价值工程,2015(4):19-21.

[6] 邓乃扬,田英杰.支持向量机:理论算法与拓展[M].北京：科学出版社, 2009.

[7] 刘琰.支持向量机核函数的研究[D].西安:电子科技大学,2012.

[8] 杨雪梅,李书琴,杨会君.基于PCA和M-SVMs的化学物质生态危害预测应用研究[J].环境科学与技术.2012(10):195-200.

[9] 王升杰,李宝树,徐建云,等.基于多分辨率奇异谱熵和支持向量机的断路器机械故障诊断方法研究[J]. 电力科学与工程,2012(7):30-35.

[10] 王小川,史峰,郁磊,等. MATLAB神经网络43个案例分析[M].北京：北京航空航天大学出版社,2013.

[11] FARUTO,lIYANG, LIBSVM-farutoUltimateVersion a Toolbox with Implements for Support Vector Machines Based on Libsvm, 2011[EB/OL].[2016-10-30]. http://www.matlabsky.com.