计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(4):266−270 [doi: 10.15888/j.cnki.csa.007914]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于改进 SVM 的互联网用户分类①

尚　晖

(浙江工贸职业技术学院, 温州 325002)
通讯作者: 尚　晖, E-mail: 50703066@qq.com

**摘　要**: 由于传统模型大量约束样本, 导致其学习能力下降, 因此设计一个基于改进支持向量机 (Support Vector Machine, SVM) 的互联网用户分类模型. 该模型通过构造样本数据, 模拟互联网用户的浏览轨迹; 根据用户偏好, 制定全新的用户分类策略; 基于改进支持向量机, 实现对互联网用户的分类. 性能测试: 3 次实验下, 此次设计的模型分类准确率平均值为 98.56%, 超出了预设的期望值, 具备分类能力. 对比测试: 与两组传统用户分类模型相比, 此次设计的模型, 面对不断增加的样本数据, 同样能保持高水平的学习能力.

**关键词**: 改进 SVM; 互联网用户; 分类模型; 学习能力

引用格式: 尚晖.基于改进 SVM 的互联网用户分类.计算机系统应用,2021,30(4):266–270. http://www.c-s-a.org.cn/1003-3254/7914.html

## Internet User Classification Based on Improved SVM

SHANG Hui

(Zhejiang Industry & Trade Vocational College, Wenzhou 325002, China)

**Abstract**: The learning ability of traditional models is reduced by copious constrained samples, so an Internet user classification model based on improved Support Vector Machine (SVM) is designed, which simulates the browsing trajectories of Internet users by constructing sample data. A brand-new user classification strategy according to user preferences is formulated. Then, Internet users are classified based on improved SVM. According to the three performance tests, the model has satisfying classification ability because its average accuracy is 98.56%, higher than the expected value. Seen from the comparative tests with two traditional user classification models, this model can maintain a high level of learning ability in the face of increasing sample data.

**Key words**: improved SVM; Internet users; classification model; learning ability

各类 APP 依靠互联网扩大影响, 为提高自身在同行业中的竞争优势, 采用传统用户分类模型, 对注册用户进行分类以便提供更好的服务. 互联网具有强大的通讯和社交功能, 互联网企业以互联网为依托, 开发具有企业特色的 APP 软件, 吸引使用者利用 APP 浏览网上信息. 但随着信息化时代到来, 企业发现互联网带来的丰厚利益, 越来越多的企业投身到互联网行业中, 竞争变得越来越激烈, 因此为了提高自身的竞争优势, 提出利用一种分类手段, 将网站中的互联网用户进行分类, 相关学者对此进行了仔细研究. 欧阳晔等[1] 提出一个基于机器学习算法的分类模型, 旨在利用该算法, 对网络用户浏览偏好进行分类; 王嘉祺等[2] 提出用户分类系统在不同的社交网络中发挥着重要的作用, 例如恶意账号检测, 高影响力用户发现及会员用户发现. 引入深度学习技术来解决用户分类问题, 且使用了陌陌的真实数据进行评估, 对于不同的分类目标,均可取得较好的效果, 但是分类准确度较低; 蒲杰方, 卢荧玲[3] 筛选了 14 个关键变量作为影响客户是否购买定期存款

① 收稿时间: 2020-07-13; 修改时间: 2020-10-09, 2020-10-21; 采用时间: 2020-10-28; csa 在线出版时间: 2021-03-30

的影响因素, 并对重要特征进行初步分析; 根据数据特征利用 k-means 聚类算法对银行的客户群进行分类, 从而得出三类最有可能购买定期存款的客户群, 剖析每一类客户群的特征, 从而有针对性地为其提供差别化的分类, 但是分类用时较长. 这些传统模型的使用效果没有达到预期, 因此研究一个全新的互联网用户分类模型.

支持向量机简称为 SVM, 是将风险控制在最小的一个机器学习算法, 通过 SVM 的计算, 得到全局最优解, 同时将计算难度降至最低, 减小以往学习算法的计算误差. 支持向量机解决了局部极小值的问题, 且具有较好的推广能力, 对于数据检测、数据挖掘以及数据处理等研究领域, 有不错的使用效果. 为进一步提高支持向量机的使用性能, 以原有支持向量机为依据, 对 SVM 改进, 得到全新的 TWSVM 和 NPSVM. 改进后的 SVM 数据拟合性更好、求解数据的能力更强, 因此在互联网用户分类研究中, 引入改进的 SVM 进一步完善互联网用户分类方法.

## 1 基于改进 SVM 的互联网用户分类模型

### 1.1 构造样本数据

假设互联网用户浏览网络信息的时间序列为$u(t)$, 其中$t\in(1,N)$; 令嵌入维数为$n$, 时间延迟为$\lambda$, 则$N'=N-(n-1)\lambda$, 表示重构后的相空间矢量长度, 重构后获得$n$维相空间相点$U_m$, $m\in(1,N')$, 表示 $U_m$的每一个分量都有$n$个元素, 即维数[4]. 以$u(t)$中的$u(m)$为起点, 每隔$\lambda$个互联网用户信息, 重构相空间相点在相空间的轨迹, 公式为:

$$\begin{cases} U_1=(u_1,u_{1+\lambda},\cdots,u_{1+(n+1)\lambda})^{\mathrm{T}} \\ U_2=(u_2,u_{2+\lambda},\cdots,u_{2+(n+1)\lambda})^{\mathrm{T}} \\ \vdots \\ U_{N'}=(u_{N-(n-1)\lambda},u_{N-(n-2)\lambda},\cdots,u_N)^{\mathrm{T}} \end{cases} \tag{1}$$

模型设置合适的嵌入维数, 则重构的相空间可以准确模拟互联网用户的浏览轨迹. 根据混沌理论可知, 嵌入维数$n$的值太小, $c=1,2,\cdots,n$, 重构空间中的用户信息, 会因吸引子的作用, 而产生扭结和重叠现象, 此时的信息距离过于接近, 数据之间交融, 难以进行分类. 同时噪声的维数是无穷大的, 若嵌入维数$n$的值太大则$n-c$空间将被舍入误差完全覆盖, 因此在设置嵌入维数$n$时, 采用误差最小算法设置嵌入维数[5].

获得网络用户的时间序列数据$\{u_m\}_{m=1}^{M}$, 其中$u_m=u(t_0+m\Delta t)$, $M$表示样本数据个数; $t_0$表示用户浏览网页的初始时间; $\Delta t$表示样本时间间隔. 根据同样的假设条件, 则其在$n$维空间$D^n$中形成的新向量$U_m$可被定义为:

$$\begin{cases} U_m=(u_m,u_{m+1},\cdots,u_{m+(n-1)}) \\ m=1,2,\cdots,N_n \\ N_n=N-(n-1)\lambda \end{cases} \tag{2}$$

根据式 (2) 的计算结果, 在$D^n$中定义$U_i$到$U_j$的距离, 公式为:

$$\left\|U_i-U_j\right\|=\left\{\sum_{t=1}^{n-1}\left(u_{i+sr}-u_{j+sr}\right)^2\right\}^{1/2} \tag{3}$$

式中,$s$表示信息长度; $r$表示空间所占范围比[6]. 根据嵌入定理, 令最佳延迟时间为$\lambda$, 则$n$为最佳嵌入维数时的映射关系为$f:D\to D^n$, 其中$f$表示关系参数, $D$表示网络空间中的用户信息. 则存在公式:

$$N_{n+1}=f(N_n) \tag{4}$$

利用映射$f$的连续性, 当$U_i$靠近$U_j$时, $u_{i+n}$与$u_{j+n}$之间也应靠近. 记$U_i$的最邻近点是$U_i^*$, 则:

$$\left\|U_i^*-U_i\right\|=\min_{j=1,2,\cdots,N_n}\left\|U_j-U_i\right\| \tag{5}$$

计算平均一步误差, 结果为:

$$q(n,\lambda)=\frac{1}{N-N_n}\sum_{i=1}^{N_n-1}\left|u_{i+n}-U_{i+n}^*\right| \tag{6}$$

当$n$比最佳嵌入维数小时, 误差$q(n,\lambda)$较大; 当$n$达到最小嵌入维数时, 因为映射$f$所以$q(n,\lambda)$减少. 当$n$继续增大时, $q(n,\lambda)$随之变化, 当$q(n,\lambda)$为最小时得到的最佳嵌入维数$n$, 可以作为最佳结果[7]. 将该结果带入式 (1), 重构的相空间可以反映互联网用户的浏览轨迹, 完成对样本数据的构造.

### 1.2 制定用户分类策略

根据互联网用户在浏览网页信息时浏览轨迹, 计算用户属性偏好度, 将分值作为用户分类的依据. 根据物联网客户的浏览轨迹, 设置用户标签, 包括: 财经、科技、数码、社交、交通、天气、新闻、法律、品牌、美食以及保险等. 利用数学算法, 计算用户浏览轨迹中, 存在的逻辑、类似偏好等, 从而形成分类定义[8].

对第 1.1 节构造的样本进行统计, 合理转化统计结果拟合出函数图像, 根据图像中正负样本的差异指标重新清洗用户信息, 再次通过转化得到拟合函数图像, 若图中的样本数据分布分散, 说明提取的构造样本存

在问题, 需要重新执行上述操作; 若函数分布差异性明显, 说明维度有效. 用户偏好 $B$ 的变化控制样本在相空间的变化. 假设用户偏好存在$w$个, 则有$B_1,B_2,B_3,\cdots,B_w$, 数学算法的计算结果为:

$$Z_i \sim D\left(\tau_i,\varphi^2\right), i=1,2,\cdots,w \tag{7}$$

式中, $Z_i$表示构造的样本数据集合; $\tau_i$表示受偏好$B$变化影响的标签偏移阈值; $\varphi$表示偏好差异[9]. 将显著性问题转化为偏好$B$在$D$空间内是否影响网页浏览选择行为, 即检验$G_0:\tau_1=\tau_2=\cdots=\tau_w$是否成立. 给出下列方程, 其中各项参数为验证所需的指标.

$$\begin{cases} \bar{z}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{w}\sum_{j=1}^{n_i}z_{ij} \\ S_1^2=\displaystyle\sum_{i=1}^{w}\sum_{j=1}^{n_i}\left(z_{ij}-\bar{z}\right)^2 \\ S_2^2=\displaystyle\sum_{i=1}^{w}\sum_{j=1}^{n_i}\left(z_{ij}-\bar{z}_i\right)^2 \\ S_3^2=\displaystyle\sum_{i=1}^{w}\sum_{j=1}^{n_i}\left(z_i-\bar{z}\right)^2 \end{cases} \tag{8}$$

上述公式中$n$表示结果总数; $\bar{z}$表示总均值; $S_1^2$表示总方差平方和; $S_2^2$表示组内平方和; $S_3^2$表示组间平方和[10]. 根据上述指标, 得到$G_0$的拒绝域为:

$$V=\left\{\frac{(n-w)S_3^2}{(w-1)S_2^2}>G(w-1,n-w)\right\} \tag{9}$$

得到的检验结果可分为 4 种情况: 高度显著、显著、有一定影响、无显著影响, 根据该结果得到用户偏好 $B$ 变化下对于互联网信息选择的影响程度建立一个互联网需求客户分类数据表, 如表 1 所示[11].

表 1 互联网需求客户分类数据表

| 高端商务人群 | 白领用户人群 | 校园用户人群 | 农村用户人群 | 其他用户人群 |
|---|---|---|---|---|
| 理财 | 购物 | 家居 | 医疗 | 通讯 |
| 国家 | 影视 | 社交 | 房产 | 网络 |
| 财经 | 微博 | 数码 | 手机 | 养生 |
| 时尚 | 游戏 | 房产 | 科技 | 音乐 |
| 阅读 | 饮食 | 服饰 | 教育 | 资讯 |

按照上述分解结果, 制定一个详细的用户分类策略, 加强模型的分类效果.

### 1.3 基于改进的 SVM 设计分类模型

根据制定的分类策略, 利用改进的 SVM 设计分类模型对互联网用户进行分类. 用户的非线性可分情形下假设两个用户的选择向量分别为$x$和$y$, 则经过改进 SVM 的非线性函数 $F$ 的分类模型过程如下:

步骤 1. 计算待分类样本与训练集之间的距离, 计算方法主要有欧氏距离;

步骤 2. 按距离递增次序排序;

步骤 3. 选取与当前点距离最小的$k$个互联网用户;

步骤 4. 统计前$k$个互联网用户所在类别出现的频率;

步骤 5. 返回前$k$个互联网用户出现频率最高的类别作为互联网用户划分目标.

将用户选择向量映射到特征空间 $K$ 内, 则两个向量的欧氏距离为:

$$l^K(x,y)=\sqrt{H(*)-2H(x,y)+H(y,y)} \tag{10}$$

式中, $H(*)$代表核函数[12,13], 那么特征空间样本的中心向量$C$为:

$$C_F=\frac{1}{n}\sum_{i=1}^{n}F(x_i) \tag{11}$$

根据上述公式计算类中心, 再计算两类中心的距离, 公式为:

$$L=\left|C^+-C^-\right| \tag{12}$$

式中, $C^+$表示正类中心; $C^-$表示负类中心. 计算两类样本与其他用户样本信息之间的距离, 当该距离小于公式 (12) 的计算结果时, 将样本作为有效候选支持向量, 即:

$$L'=|x_i-C| \tag{13}$$

图 1 为保留满足$L'<L$时, 样本作为有效候选支持向量的示意图[14].

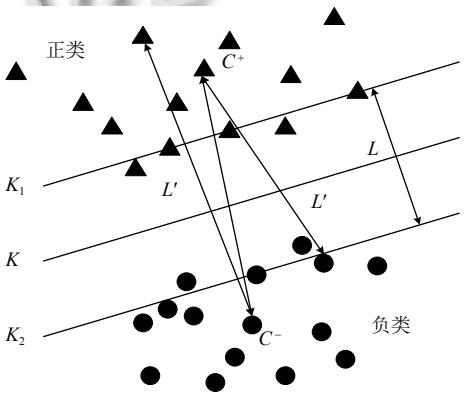



图 1 预选有效的候选支持向量

根据图 1 的示意图可知, 根据$L'$和$L$对特征空间中互联网用户选择进行划分, 以此将选择偏好相同的用户归集到一个数据集合中得到如表 1 所示的分类结果, 至此实现基于改进 SVM 的互联网用户分类[15].

## 2 实验研究

### 2.1 性能测试

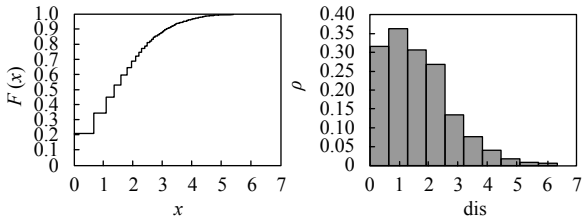
以互联网上某一期间的新闻作为实验测试基本条件, 利用设计的分类模型分别统计该期间的新闻展现量$P$和点击量click, 其中得到的新闻展现量统计分析结果如表 2 所示.

表 2 新闻展现量统计分析表

| 展现次数 | 新闻数量(个) | 展现次数比重 | 累积分布比值 |
|---|---|---|---|
| 1 | 30498 | 0.2041 | 0.0254 |
| 2 | 20156 | 0.1349 | 0.3411 |
| 3 | 15579 | 0.1051 | 0.4459 |
| 4 | 12304 | 0.0814 | 0.5290 |
| 5 | 9812 | 0.0661 | 0.5944 |
| 6 | 7597 | 0.0499 | 0.6458 |
| 7 | 6188 | 0.0332 | 0.6872 |
| 8 | 4872 | 0.0268 | 0.7213 |
| 9 | 4035 | 0.0217 | 0.7485 |
| 10 | 3301 | 0.0195 | 0.7806 |
| 11 | 2857 | 0.0159 | 0.7895 |
| 12 | 2388 | 0.0142 | 0.8101 |
| 13 | 2136 | 0.0133 | 0.8221 |
| 14 | 1915 | 0.0111 | 0.8334 |
| 15 | 1599 | 0.0112 | 0.8435 |
| 16 | 1440 | 0.0077 | 0.8552 |
| 17 | 1255 | 0.0082 | 0.8609 |
| 18 | 1198 | 0.0075 | 0.8788 |
| 19 | 1106 | 0.0069 | 0.8868 |
| 20 | 997 | 0.0068 | 0.8942 |

表 2 中, 展现次数为浏览过某条新闻的用户数量. 已知此次展现次数的最小值为 1, 最大值为 645, 均值为 11, 其中展现次数为 50 的新闻, 所占比例为 0.0009, 表 1 是 20 次以内展现次数的统计结果. 根据表中数据可知, 展现次数小于 10 的累积分布率约为 78.06%, 展现次数小于 20 的累积分布率约为 89.42%. 分类模型取新闻展现量 $P$ 的对数, 得到下图 2 所示的新闻展现量 $P$ 的分布图.

根据图中显示数据可知, 得到的分布是一个长尾的幂律分布, 大部分点集中分布在较小展现量处. 新闻作为网民了解国情、社会事件的重要媒介, 更新速度十分迅速. 用户根据自身偏好, 只浏览自身感兴趣的新闻类型. 因此该模型推断出大量用户浏览新闻的时间较为零散, 专门定点浏览新闻的用户数量较少. 因此该分类模型根据这一分析, 以用户偏好作为参考进行互联网用户分类. 为了实验测试的严谨性, 对该模型进行 3 次性能测试, 并计算该分类模型的分类准确率, 当该模型的分类准确率在 95% 以上时, 证明该模型成立且具有使用价值. 表 3 为模型分类准确性计算结果.




(a) 新闻展现量对数概率累积分布图 (b) 新闻展现量对数概率密度分布图

图 2 分类模型得到的新闻展现量分布图

表 3 分类模型分类准确性测试结果

| 测试组 | SVM的参数选项 | 分类准确率(%) |
|---|---|---|
| 第1组 | $c$=2, $m$=0.1 | 98.67 |
| 第2组 | $c$=2, $m$=0.1 | 98.33 |
| 第3组 | $c$=2, $m$=0.1 | 98.67 |

根据表中的数据计算结果可知, 3 次测试下基于改进 SVM 的互联网用户分类结果, 其分类平均准确率为 98.56%, 满足预期, 因此进行下一步对比实验.

### 2.2 对比测试

实验测试环境和测试条件不变, 分别利用 3 个模型对浏览新闻的用户进行分类, 对照组 1 是基于 SVM 的互联网用户分类模型, 对照组 2 是文献 [3] 模型, 实验组为基于改进 SVM 的互联网用户分类模型, 对比 3 种模型. 实验结果如图 3 所示.

根据图中数据可知, 实验组模型的学习样本数量与模型自身提供的样本数量几乎一致. 对照 1 组模型的学习样本数量, 在模型自身提供的样本数量达到 1000 时其学习数量迅速下降且难以恢复. 对照 2 组模型的学习样本数量, 比其自身提供的样本数量少了近 1 倍. 相比较而言, 此次设计的模型性能更好. 表 4 为模型性能比较分析结果.

根据表中分析结果可知, 3 组模型虽然都是根据用户偏好特征进行分类, 但获取偏好特征的方式不同, 再加之模型自身约束了选择的样本, 导致模型学习性能下降. 可见此次设计的分类模型, 解决了模型学习能力不足的问题.

## 3 结束语

传统的分类模型与此次设计的分类模型都将用户偏好作为详细分类的依据, 改进的 SVM 充分发挥其强大的学习能力, 对分类后的样本数据进行学习, 当该模型获取到入网用户信息后, 根据其浏览内容迅速判断用户类型, 提醒软件推送用户感兴趣的各类信息. 此次研究受时间的限制没有介绍 SVM 的改进内容, 而是直

接将改进后的 SVM 投入使用, 在今后的研究项目将对改进过程、改进内容加以描述.

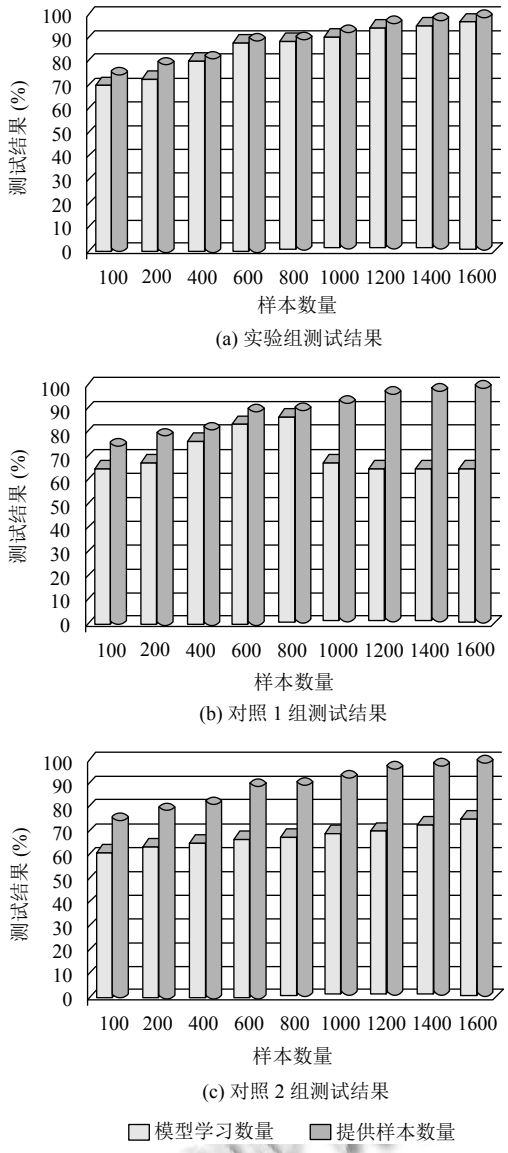



图 3 模型学习能力比较结果

表 4 模型性能比较

| 性能指标 | 实验组 | 对照1组 | 对照2组 |
|---|---|---|---|
| k值邻近 | 是 | 是 | 是 |
| 用户偏好特征 | 分析特征 | 统计特征 | 统计特征 |
| 样本约减 | 否 | 是 | 是 |
| 样本均衡 | 否 | 是 | 否 |

## 参考文献


1 欧阳晔, 杨爱东, 孟凡语. 一种博弈论辅助的机器学习算法检测用户流失行为. 电信科学, 2020, 36(6): 79–89.

2 王嘉祺, 何新磊, 汪天一, 等. 基于深度学习的移动社交网络用户分类研究. 计算机应用与软件, 2018, 35(12): 42–48.

3 蒲杰方, 卢荧玲. 基于聚类算法和神经网络的客户分类模型构建. 软件, 2018, 39(4): 130–136. [doi: 10.3969/j.issn.1003-6970.2018.04.028]

4 洪翠, 付宇泽, 郭谋发, 等. 改进多分类支持向量机的配电网故障识别方法. 电子测量与仪器学报, 2019, 33(1): 7–15.

5 彭姣, 朱建青. 时间尺度上相空间中非完整系统相对运动动力学的 Lie 对称性. 云南大学学报 (自然科学版), 2020, 42(3): 492–498.

6 李慧, 陈湘萍. 基于相空间重构和长短期记忆网络的风电预测. 新型工业化, 2020, 10(3): 1–6.

7 戎子睿, 林湘宁, 金能, 等. 基于相空间轨迹识别和多判据融合的变压器保护新策略. 中国电机工程学报, 2020, 40(6): 1924–1937.

8 胡晓丽, 张会兵, 董俊超, 等. 基于 CNN-LSTM 的用户购买行为预测模型. 计算机应用与软件, 2020, 37(6): 59–64. [doi: 10.3969/j.issn.1000-386x.2020.06.012]

9 武慧娟, 尚冰琦, 孙鸿飞, 等. 微阅读用户持续使用行为影响因素及作用路径研究. 情报科学, 2020, 38(6): 76–82, 102.

10 乔兴媚, 杨娟. 学习风格用户模型分类及其自适应策略. 现代教育技术, 2019, 29(1): 100–106. [doi: 10.3969/j.issn.1009-8097.2019.01.015]

11 吴要毛, 陈昊, 王龙, 等. 一种面向用户特征的配电网投资决策分析方法. 湖北大学学报 (自然科学版), 2019, 41(4): 391–398.

12 程永锋, 汉京善, 刘彬, 等. 基于 Bagging 算法构造强分类器的 one class SVM 导线舞动预测应用. 振动与冲击, 2020, 39(9): 152–158.

13 叶明珠, 赵治栋. 基于迁移学习和支持向量机的胎心率分类方法. 杭州电子科技大学学报 (自然科学版), 2020, 40(3): 14–18, 43.

14 王福斌, 潘兴辰, 王宜文. 基于 SVM 的多核学习飞秒激光烧蚀光斑图像分类. 激光杂志, 2020, 41(4): 86–91.

15 张菊, 杨勇. 基于 SVM 算法的高考语文中现代文阅读材料体裁自动分类研究. 自动化技术与应用, 2020, 39(4): 162–164. [doi: 10.3969/j.issn.1003-7241.2020.04.036]