网络出版时间：2018-11-15 10:12:14
网络出版地址：http://kns.cnki.net/kcms/detail/61.1450.TP.20181114.1554.014.html

基于支持向量机的英文字符识别研究[1]

郑宇晨

（四川师范大学 数学与软件科学学院，四川 成都 610068）

摘要：图像识别是“大数据”时代的热门研究领域之一，而英文字符识别是图像识别领域重要的研究方向。对于手写数据的辨认在移动智能、刑侦、医学、考古学等诸多领域有广泛的应用，同时，国内在该领域的建模探索相对匮乏。本文使用机器学习领域的经典手写字符数据集，基于统计机器学习理论，建立英文字符识别的支持向量机（Support Vector Machine，SVM）模型。鉴于国内外对于参数选择至今没有公认的方法，本文依据支持向量的个数、训练误差、测试误差作为评价指标，对惩罚参数 C 的选取进行探索并给出了在字符识别领域的推荐值。实证结果表明，对“变体”英文字母的识别准确率很高，且非常稳健，没有“过拟合”现象，说明支持向量机适用于处理字符识别问题。本质上，相比经典的二分类问题，本文是多分类支持向量机（Multi-class Classification Support Vector Machine，MCSVM）应用的研究与探索。

关键词：手写英文字符识别；数据挖掘；高斯径向基核函数；多分类支持向量机（MCSVM）；统计机器学习；惩罚参数 C

Research of English Character Recognition Based on Support Vector Machine

Yuchen Zheng

(School of Mathematics and Software Sciences, Sichuan Normal University, Chengdu 610000, China)

Abstract: Pattern Recognition is one of the hottest research fields in the “Age of Big Data”, and English character recognition is considered a significant research orientation of Pattern Recognition. Recognizing handwritten data is widely used in hundreds of fields, such as mobile intelligent, criminal investigation, medicine, and archaeology. However, there is rare domestic modeling research of English character recognition. Based on the statistical machine learning theory, this article makes use of a classic handwritten data set in the field of machine learning in order to build a support vector machine (SVM) model of English character recognition. It is well-known that there isn’t any widely accepted method to select parameters for SVM even in foreign articles. According to the fact, research on how to select penalty parameter C is implemented based on the index like, the number of support vector, training error, and test error. More than that, a recommended penalty parameter C of letter recognition is also proposed. The experimental results indicate that this model has high accuracy and robustness without overfitting to recognize variation English character. So SVM is a favorable choice to handle with character recognition problem. Essentially, this article aims at applied research and exploration of the Multi-class Classification Support Vector Machine (MCSVM) compared with classical binary Support Vector Machine (SVM).

Keywords: handwritten English character recognition； data mining；radial basis function(RBF)；MCSVM；Statistical machine learning；penalty parameter C

1 引言

伴随着大数据时代的来临，各种社会、生活现象背后的大数据“内核”逐渐被人们关注、重视。图像与语音识别是最为人熟知并受到高度关注的八个领域之一。本文研究图像识别中较简单、实用的英文字符识别。限于 PC 的运算能力，这里研究样本容量为 20000，参数空间为 16 维（加上 1 维的响应值，是 17 维）的英文文本数据。

研究目的：针对“变体”的 26 个英文字母，建立模型进行准确识别。事实上，这是个（多）分类问题。“变体”包括：多种不同的随机重塑、扭曲或印刷颜色（黑体或白体）的改变。见图 1 举例。

支持向量机是统计学习理论中最具代表性的模型，也是数据挖掘、机器学习领域的重要模型之一。本文选择文本（图像）识别中的英文字符识别为研究对象，以国外前沿的科研成果为灵感来源，旨在促进国内相关领域的研究再前进“一小步”。

---

[1]基金项目：教育部人文社科规划项目（12YJA630197）；安徽省质量工程项目（2016jyxm0017）。
作者简介：郑宇晨（1993 年生），男，安徽蚌埠人，硕士，研究方向为统计机器学习。

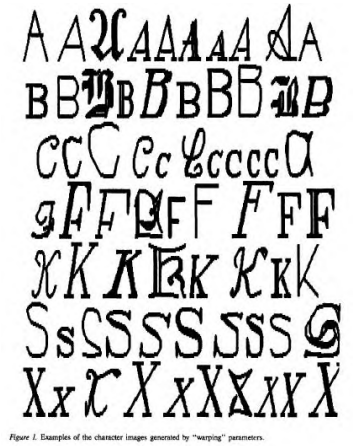

Figure 1. Examples of the character images generated by "warping" parameters.

图 1 英文字母“变体”示例[1]

2 相关工作

无论是支持向量机的建模还是实证研究，相比国外的高速发展，国内的研究主要集中在部分领域的应用实证分析，前人研究文本识别，尤其（英文）字符识别不够充分。

国外相关研究：Corinna Cortes、Vladimir（1995）[2]建立支持向量网（Support Vector Networks）用于二分类特征识别，通过引入核函数做高维映射的方式，提供了处理低维空间线性不可分数据的“全新思路”。Hamed Pirsiavash、Deva Ramanan、Charless Fowlkes（2008）[3]提出双线性支持向量机用于可视化数据的识别，取得了较好的分类结果。M.Nazir、Muhammad Ishtiaq、Anab Batool、M.Arfan Jaffar、Anwar M.Mirza（2012）[4]基于离散余弦变换（Discrete Cosine Transform，DCT）、K-最近邻分类器建立了分类器，用于图像的性别识别，取得了 99.3%的分类准确率。Luo Luo、Yubo Xie、Zhihua Zhang、Wu-Jun Li（2015）[5]提出支持矩阵机（Support Matrix Machine，SMM）模型，捕捉了输入矩阵（型数据）行、列指标间的关系，使支持向量机理论取得突破，并且模型预测精度高、稳健性好且计算效率大大提高。

国内相关研究：吴一全、朱丽、周怀春（2014）[6]构造火焰图像的特征向量，建立基于 Krawtchouk 支持向量机（SVM）模型，进行火焰熄灭与否的状态识别，实现了 95.56%的二分类识别率。薛浩然、张珂荇、李斌、彭晨辉（2015）[7]运用 SVM 适用于解决小样本、非线性特征的特点，建立多分类模型，进行变压器故障类型判断，实现了 95%的小样本识别率。此外，在电子故障检测与诊断领域，万鹏、王红军、徐小力（2012）[8]，史丽萍、王攀攀、胡泳军、韩丽（2014）[9]也进行了一些实证研究。

3 支持向量机理论

支持向量机（Support Vector Machine,SVM）[2]是一类由 Vapnik 及其同事提出的图像分类算法。其训练算法的思想基于“结构风险最小化”而非“经验风险最小化”原理，并且为训练多项式、神经网络和径向基函数（RBF）分类器提供了新的方法。[10]SVM 已经被证明对于许多分类任务是卓为有效的。[11][12][13]

经典的 SVM 是一个二分类模型。假设样本点由如下带分类标签的数据集构成 $D=\{x_i,y_i\}_{i=1}^{m}$，其中 $y_i\in\{-1,1\}$，即观测值是二元取值。目标是通过有限数量的线性分类器将数据分开，使得泛化误差最小，或至少使其上界最小。因此，要寻找使得两类数据“边界”的距离最大化的超平面。判别超平面可以按照如下方式定义：

$$f(x)=\omega^T x+b \tag{1}$$

其中，w 是一个正交于超平面的向量。最优分类超平面的计算是一个约束优化问题，可以用二次规划的技术技巧解决。SVM 的优

化问题可以用如下方式呈现：

$$\min_{\omega,b,\xi} \ \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to } \ y_i(\omega^T x_i + b) \geq 1 - \xi_i \qquad (2)$$

当有新的样本点 x 进入分类器时，将根据其与决策边界（decision boundary）的关系对其进行分类，相应的决策函数为

$$g(x) = sign(\omega^T x + b) \qquad (3)$$

4 建模及实证分析

试验环境：本文所有的试验在 Intel Genuine 的 CPU（型号 T2050,1.60GHz，1.99GB 的 RAM）以及 32bit 的 Windows Server XP 系统下完成。通过 R3.3.3 实现支持向量机。

假设：1.文件已经分割成矩形区域，2.每个区域包含一个单一字符，3.文件中只包含英文字母字符。

4.1 收集数据

数据来源：使用由 W.Frey 和 D.J.Slate 捐赠给 UCI 机器学习数据仓库（UCI Machine Learning Data Repository）（http://archive.ics.uci.edu/ml）的一个数据集。该数据集包含了 26 个大小写英文字母的 20000 个案例，使用多种不同的随机重塑和扭曲的黑色和白色字体印刷。

4.2 探索和准备数据

Frey 和 Slate 提供的文件，当图像字符被扫描到计算机中，它们将被转换成像素，并且拥有 16 个统计属性，这就是本文的建模指标。详见附录 1：建模指标附表。可以预期，响应值 letter 有 26 个水平（或者 26 类），因为英文字母有 26 个。

下面进入机器学习的训练和测试阶段。通常情况下，需要随机地将数据集划分为训练数据集和测试数据集。然而，Frey 和 Slate 已经将数据随机化，并建议使用前 16000 个记录（80%）来建立模型，即作为训练样本；使用后 4000 条记录（20%）来进行测试，即作为测试样本。

到此为止，数据准备完毕。

4.3 基于数据训练模型

使用 R 软件自带的 kernlab 添加包中的 ksvm()函数建立模型，其具体语法如下。其中，核函数的选择对于 SVM 模型的表现至关重要。相关研究表明，在机器学习领域的研究中，高斯径向基核函数（rbfdot）是首选。[14] [15] [16] [17] [18] 主要建模结果如下：

```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)

parameter : cost C = 1



Gaussian Radial Basis kernel function.

Hyperparameter : sigma =  0.0473840758601753

Number of Support Vectors : 8679



Training error : 0.051688
```

图 2 支持向量机建模结果

由训练误差为 0.051688 可见，模型对训练样本拥有较高的分类精度 94.83%。但因为数据挖掘模型可能出现“过拟合”（overfit）的问题，所以还需要检验模型的泛化能力。

4.4 评估模型的性能

根据测试数据集研究模型的性能，将测试数据集中的预测值与真实值进行比较，从而判断它能否很好的推广到未知的数据，即建立的模型能否在现实世界中较好的应用。使用 R 软件中的 Table（）函数来实现，结果如下：

letter_predictions	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	144	0	0	0	0	0	0	0	0	1	0	0	1	2	2	0	5
B	0	121	0	5	2	0	1	2	0	0	1	0	1	0	0	2	2
C	0	0	120	0	4	0	10	2	2	0	1	3	0	0	2	0	0
D	2	2	0	156	0	1	3	10	4	3	4	3	0	5	5	3	1
E	0	0	5	0	127	3	1	1	0	0	3	4	0	0	0	0	2
F	0	0	0	0	0	138	2	2	6	0	0	0	0	0	0	16	0
G	1	1	2	1	9	2	123	2	0	0	1	2	1	0	1	2	8
H	0	0	0	1	0	1	0	102	0	2	3	2	3	4	20	0	2
I	0	1	0	0	0	1	0	0	141	8	0	0	0	0	0	1	0
J	0	1	0	0	0	1	0	2	5	128	0	0	0	0	1	1	3
K	1	1	9	0	0	0	2	5	0	0	118	0	0	2	0	1	0
L	0	0	0	0	2	0	1	1	0	0	0	133	0	0	0	0	1
M	0	0	1	1	0	0	1	1	0	0	0	0	135	4	0	0	0
N	0	0	0	0	0	1	0	1	0	0	0	0	0	145	0	0	0
O	1	0	2	1	0	0	1	2	0	1	0	0	0	1	99	3	3
P	0	0	0	1	0	2	1	0	0	0	0	0	0	0	2	130	0
Q	0	0	0	0	0	0	8	2	0	0	0	3	0	0	3	1	124
R	0	7	0	0	1	0	3	8	0	0	13	0	0	1	1	1	0
S	1	1	0	0	1	0	3	0	1	1	0	1	0	0	0	0	14
T	0	0	0	0	3	2	0	0	0	0	1	0	0	0	0	0	0
U	1	0	3	1	0	0	2	0	0	0	0	0	0	0	1	0	0
V	0	0	0	0	0	1	3	4	0	0	0	0	1	2	1	0	3
W	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0
X	0	1	0	0	2	0	0	1	3	0	1	6	0	0	1	0	0
Y	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7	0
Z	2	0	0	0	1	0	0	0	3	4	0	0	0	0	0	0	0

letter_predictions	R	S	T	U	V	W	X	Y	Z
A	0	1	1	1	0	1	0	0	1
B	3	5	0	0	2	0	1	0	0
C	0	0	0	0	0	0	0	0	0
D	4	0	0	0	0	0	3	3	1
E	0	10	0	0	0	0	2	0	3
F	0	3	0	0	1	0	1	2	0
G	2	4	3	0	0	0	1	0	0
H	3	0	3	0	2	0	0	1	0
I	0	3	0	0	0	0	5	1	1
J	0	2	0	0	0	0	1	0	6
K	7	0	1	3	0	0	5	0	0
L	0	5	0	0	0	0	0	0	1
M	0	0	0	3	0	8	0	0	0
N	3	0	0	1	0	2	0	0	0
O	0	0	0	3	0	0	0	0	0
P	0	0	0	0	0	0	0	1	0
Q	0	5	0	0	0	0	0	2	0
R	138	0	1	0	1	0	0	0	0
S	0	101	3	0	0	0	2	0	10
T	0	3	133	1	0	0	0	2	2
U	0	0	0	152	0	0	1	1	0
V	1	0	0	0	126	1	0	4	0
W	0	0	0	4	4	127	0	0	0
X	0	1	0	0	0	0	137	1	1
Y	0	0	3	0	0	0	0	127	0
Z	0	18	3	0	0	0	0	0	132

图 3 模型分类结果

结果分析：这是一个矩阵，在 R 软件中分成两部分显示，因为响应值矩阵的维数较高，有 26 维。对角线的值 144、121、120、156 和 127 表示的预测值与真实值相匹配的总数。非对角线的值表示预测出现错误。例如：位于行 B 和列 D 的值 5 表示有 5 种情况将字母 D 误认为字母 B。

最终，在 4000 个测试样本点中，分类器正确识别的字母有 3357 个。以百分比计算，预测精度为 83.925%，接近 84%。

到此为止，支持向量机分类模型建立完毕，有令人满意的训练误差（5.1688%）；并且经过检验，模型的泛化能力较好，对未知的数据有较高的预测精度（83.925%），对实践的指导大有裨益。

5 模型提升

83.925%的预测精度虽然不低，但仍然有明显的提升空间，故需进行模型优化。这里采用如下方式实现：调整惩罚参数 C 的取值。这是一种主流的探索方式。建模和测试过程与“三、四”中的方法相同，这里直接展示主要返回结果，代码参见附录。

C	支持向量的个数	前 6 个预测样本	训练误差	测试误差
1	8679	U N V X N H	0.052	0.161
2	7661	U N V I N H	0.033	0.053
5	6692	U N V I N H	0.020	0.038
20	6096	U N V I N H	0.007	0.032
50	5956	U N V I N H	0.002	0.030
100	5918	U N V I N H	0.001	0.030
500	5870	U N V I N H	0	0.029
5000	5874	U N V I N H	0	0.029

表 惩罚参数 C 的探索

表的结果如果进行可视化，会更加直观，限于篇幅，不再展示。分析如下：

（1）支持向量的个数基本随着 C 的增大而减小，说明对样本点的划分越严格（或对错分的“惩罚”越严厉），处于“边界位置”的点就越少。

（2）训练误差随着 C 的增大而减小，直至为 0，说明对样本点的划分越严格（或对错分的“惩罚”越严厉），即制定越严格的分类准则，模型对训练数据的拟合优度越高。这是数据挖掘算法的普遍特点。

（3）测试误差随着 C 的增大而持续减小，直至达到一个非常低的水平（0.029<0.03）；当 C=50 时，测试误差为 0.03025，往后基本就稳定在 0.03 左右了。说明对样本点的划分越严格（或对错分的“惩罚”越严厉），模型对测试数据的拟合优度越高，即对未知样本的分类准确度越高。体现了支持向量机分类模型的优越性，即强大的泛化能力，相比深度学习的核心算法——神经网络，不容易过拟合。

（4）“前 6 个样本的预测结果”是特别增加的一个细节。从预测结果来看，C 在取默认值为 1 时，第四个样本的预测结果为 X；而 C 在其它取值的情形下，第四个样本的预测结果为 I。结合（2）、（3）的结果可知，适当大的惩罚参数 C 得到更优的模型，所以有理由相信第四个样本的预测结果为 I，而非 X。仅仅取 6 个样本，就已经产生了预测结果的差异，所以，参数的调整对于模型优化来说是很有必要的。

总之，随着惩罚常数 C 的不断增加，测试样本的预测精度始终不下降（事实上，还在不断上升），最终错分率小于 3%，没有出现过拟合。反映本文建立的基于支持向量机的英文字符识别模型较为优越，不仅分类精度高，而且非常稳健。所以，可以为现实提供有价值的指导。比如：针对不同书写者的手癖、使用的书写工具，可以有效地对英文文字加以识别，并为电子方式的呈现提供基础。

当然，从理论上而言，随着惩罚参数 C 的增加，模型的稳健性是在下降的，几何上反映为两分类超平面间距离缩小。所以实践中，C 的取值需要“合适”，而不宜过大。就本文建立的模型而言，C 可以取 20，此时测试误差为 0.0315。

6 总结与展望

本文建立了一个多分类支持向量机（SVM）模型，操作简单，容易实现。样本容量充足且指标维数“适中”，所以结论也是较有说服力的。实证结果表明，对于各种“变体”的手写英文字符，此模型可以实现很高的预测精度和令人满意的稳健性。在此基础上，本文对惩罚参数 C 的选择做了一定的探索，而关于 SVM 选参，在国内外学术界始终没有形成具有共识的理论。

相比只有 26 种分类的英文字母，如何进行更为复杂的中文字符识别是未来的研究方向。有许多问题需要面对，包括：核函数的选择、惩罚常数 C 的选择、训练样本及测试样本量的选择、非平衡样本、非数值型特征等。此外，对于有污染或字符缺损的文本数据，本文建立的模型是否依然有良好的表现，也有待求证。

参考文献：

[1]P.W.Frey, D.J.Slate. Letter recognition using Holland-style adaptive classifiers[J]. Machine Learning，Vol.6，page 161-182，1991.

[2]C.Cortes, V.Vapnik. Support Vector Networks[J]. Machine Learning, 20(3), page 273-297, 1995.

[3]H.Pirsiavash, D.Ramanan, C.Fowlkes. Bilinear classifiers for visual recognition[J]. Neural Information Processing Systems（NIPS），volume 09，page 1482-1490，2009.

[4]M.Nazir, M.Ishtiaq, A.Batool, M.A.Jaffar, A.M.Mirza. Feature Selection for Efficient Gender Classification.Wseas International Conference on Nural Networks & Wseas International Conference on Evolutionary Computing & Wseas International Conference on Fuzzy Systems，Athens ， Valencia，Spain，volume 11，page 70-75，2013.

[5]L.Luo, Y.B.Xie, Z.H. Zhang, W.J. Li. Support Matrix Machines[J]. International Conference on Machine Learning, volume 37, page 938-947, 2015.

[6]吴一全、朱丽、周怀春.基于 Krawtchouk 矩和支持向量机的火焰状态识别[J].中国电机工程学报，2014，34（5），734-740.

[7]薛浩然、张珂荇、李斌、彭晨辉.基于布谷鸟算法和支持向量机的变压器诊断[J].电力系统保护与控制，2015，43（8），8-13.

[8]万鹏、王红军、徐小力.局部切空间排列和支持向量机的故障诊断模型[J].仪器仪表学报，2012，33（12）,2789-2795.

[9]史丽萍、王攀攀、胡泳军、韩丽.基于骨干微粒群算法和支持向量机的电机转子断条故障诊断[J].电工技术学报，2014，29（1）,147-154.

[10]G.Blanchard, O.Bousquet, P.Massart. Statistical Performance of Support Vector Machines[J].The Annals of Statistics， 36(2)，page489-531，2008.

[11]R.Ronfard, C.Schmid, B.Triggs. Learning to parse pictures of people[J]. Lecture Notes in Computer Science，2353（6），page 700-714，2002.

[12]B.B.Li, A.Artemiou, L.X.Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction[J].The Annals of Statistics，39(6)，page3182-3210，2011.

[13]B.Kalantar, B.Pradhan, S.A.Naghibi, A.Motevalli, S.Mansor. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine(SVM), logistic regression(LR) and artificial neural networks(ANN)[J].Geomatics,Natural Hazards and Risk，9(1)，page49-69，2018.

[14]I.Steinwart, M.Anghel. Consistency of Support Vector Machines for Forecasting the Evolution of an Unknown Ergodic Dynamical

System from Observations with Unknown Noise[J].The Annals of Statistics, 37(2)，page841-875，2009.

[15]I.Steinwart, C.Scovel. Fast Rates for Support Vector Machines Using Gaussian Kernels[J].The Annals of Statistics， 35(2)，page575-607，2007

[16]苑利、赵锐、谭孝元、苟先太. 基于红外成像技术的零值绝缘子检测[J].高压电器，2018，54（2），97-102.

[17]周爱红、倪莹莹、尹超、孙武. 一种盾构施工引起的地面沉降预测方法[J].测绘科学，2018，43（3），167-172.