文章编号：2095－0411（2013）04－0044－04

基于 PSO－SVM 算法的长微博贴图识别方法*

刘　平，叶施仁，杨长春，侯振杰，肖　飞
（常州大学 信息科学与工程学院，江苏 常州 213164）

摘要：微博由于字数的限制，当用户需要发较多内容时通常以附图的形式给出，识别包含文本内容贴图的长微博能够为微博研究提供更多有用的数据。在支持向量机（SVM）的基础上结合粒子群算法（PSO）提出了一种识别长微博贴图的 PSO－SVM 算法。该方法提取长微博贴图的颜色矩和灰度共生矩阵特征，然后利用 PSO 算法对 SVM 模型中的误差惩罚参数和核函数进行优化得到最佳分类模型，其最优参数将被用作长微博贴图和非长微博贴图进行分类。实验表明，与传统的基于网格搜索法优化的 SVM 算法相比，PSO－SVM 算法对长微博贴图识别具有更高的准确率和召回率。

关键词：长微博贴图；支持向量机；粒子群优化算法；最佳分类模型

中图分类号：TP 391　　**文献标识码**：A　　**doi**：10.3969/j.issn.2095－0411.2013.04.009

Identifying Images Representing Long Microblog Based on PSO-SVM Algorithm

LIU Ping，YE Shi-ren，YANG Chang-chun，HOU Zhen-jie，XIAO Fei
（School of Information Science and Engineering，Changzhou University，Changzhou 213164，China）

Abstract：Due to the length limitation of micro－blogs，users have to post images containing original text contents when they want to post long micro－blogs. If such images representing long micro－blog can be identified，it will provide more useful information for micro－blog research. An identification method based on support vector machine（SVM）and particle swarm optimization（PSO）is proposed in the paper. Firstly，the color moments and gray level concurrence matrix is extracted from image representing long micro－blog，then the error penalty parameter and kernel function of SVM are optimized by PSO algorithm to obtain the optimal classification model. The results show that，compared with traditional SVM based on grid searching method，the PSO－SVM algorithm has higher accuracy and recall rate of identifying images representing long microblog.

Key words：images representing long microblog；support vector machine；particle swarm optimization；best classification model

随着互联网应用的发展和移动终端服务的普及，以微博为代表的社会媒体发展迅猛，已经成为了人们交互和分享信息的重要手段。由于微博出现时采用短信息即时发布的方式，通常要求只能发布 140 个字以内的内容，即使有些中文微博放宽了限制，如网易微博能发布 163 个字以内的内容，但是

* **收稿日期**：2013－07－30
基金项目：国家自然科学基金项目资助（61272367）；江苏省科技厅项目资助（BZ2010021）
作者简介：刘平（1989－），男，江苏扬州人，硕士生；通讯联系人：叶施仁。

用户还是经常遇到超出字数限制的问题。为此很多微博就会提供长微博工具，将用户超出规定长度的内容或者是若干幅图片拼接用一幅贴图的形式呈现（称之为长微博贴图）。一般情况下这些长微博是用户花费更多精力撰写的，有时比随意书写的短微博更具有参考价值。微博贴图可以分为长微博贴图、多拼图、实时贴图、截屏贴图、相关贴图和非相关贴图 6 类。由于长微博是文本内容的图片化，如果能将长微博贴图与其他贴图先分开，后续利用 OCR（字符识别）工具将其中的文字识别出来，将对微博的内容研究具有重要的意义。

目前，支持向量机（SVM）是使用最广泛的图像分类算法[1]，但用户通常面临怎样选择合适的核函数的困难，需要实验数据对比、大范围搜索或采用网格搜索等优化算法进行寻优。本文针对长微博贴图分类识别的问题，探索一种利用 PSO 对 SVM 的参数进行优化的方法，该方法采用 PSO 算法对 SVM 模型中的误差惩罚参数和核函数进行优化得到最佳分类模型，有效地减少了 SVM 算法的训练次数，相对传统的网格搜索法该算法具有更高的学习精度和更好的适用性。

1 支持向量机

SVM 是 Vapnik 等人于 20 世纪 90 年代根据统计学习理论提出的一种新的机器学习理论方法，它是以结构风险最小化原则为理论基础，通过选择适当的函数子集及该函数子集中的判别函数来使学习机的期望风险达到最小，保证通过有限训练样本得到小误差分类器对独立测试集的测试误差比较小，从而得到一个具有最优分类能力和推广泛化能力的学习机[2]。SVM 能有效地解决有限样本的高维模型构造问题，而且所构造的模型有很好的预测能力。

支持向量机在解决小样本、非线性和高维模式识别问题中表现出很多优势，因此在很多分类问题中表现了良好的性能[3]。下面对支持向量机进行模式学习和分类的基本原理进行简单的介绍。

对于二分问题，原始问题为：

$$\min\left(\frac{1}{2}\omega^2+c\sum_{i=1}^{l}\xi_i\right)$$

$$\text{s.t. } y_i\ ((\omega\cdot x_i)+b)\geqslant(1-\xi_i);$$

$$\xi_i\geqslant 0,\ i=1,2\cdots l \qquad (1)$$

其中对偶问题为：

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

$$\text{s.t. }\sum_{i=1}^{l}y_j\alpha_j=0;\ 0\leqslant\alpha_j\leqslant c,\ i=1,2\cdots l \qquad (2)$$

SVM 利用核函数 $K(x_i, x_j)$ 将低维特征空间不可分的数据转换到高维线性可分的特征空间中，从而实现某一非线性分类转换后的线性可分。选取适合的核函数 $K(x_i, x_j)$ 和参数 c，求解式（2），得出构造决策函数为式（3）：

$$f(x)=\text{sgn}\left(\sum_{i=1}^{l}\alpha_i^*y_iK(x_i,x_j)+b^*\right) \qquad (3)$$

公式（3）中目前满足 Mercer 条件的内积核函数 $K(x_i, x_j)$ 主要有线性核函数（linear）、多项式核函数（polynomial）、径向基核函数（RBF）和 Sigmoid 核函数。其中，RBF 核函数在低维、高维、小样本和大样本等情况下都表现出很好的学习能力，受到了广泛的应用[4]。因此，本文选择 RBF 函数作为 SVM 的核函数，其中 RBF 核表示为：

$$K(x_i,x_j)=\exp(-g\|x_i-x_j\|^2),\ g>0$$

2 粒子群寻优算法

近年来，PSO 算法开始应用到 SVM 的核函数参数寻优中。PSO 算法是继蚁群算法、遗传算法后的又一种新的群智能寻优算法，该算法最初是模仿鸟群寻找食物的社会行为建立的[5]。

PSO 算法中每一个粒子都代表解空间的一个解，它根据自己的飞行经验及其它粒子的飞行经验来调整自己的飞行。粒子在飞行过程中经过的最好位置即为自身最优解（也叫个体极值 P_{best}），整个群体所经历过的最好位置为群体的最优解（全局极值 G_{best}）[5]。粒子通过跟踪局部极值和全局极值来调整自身的飞行，从而产生新一代粒子。

PSO 算法的基本原理如下[6]：假设在一个 D 维的搜索空间中，由 n 个粒子组成的种群 $\boldsymbol{X}=(\boldsymbol{X}_1, \boldsymbol{X}_2\cdots\boldsymbol{X}_n)$，其中第 i 个粒子表示为一个 D 维的向量 $\boldsymbol{X}_i=(X_{i1}, X_{i2}\cdots X_{iD})^{\mathrm{T}}$，代表第 i 个粒子在 D 维搜索空间中的位置，也代表一个问题的一个潜在解。第 i 个粒子的速度为 $\boldsymbol{V}_i=(V_{i1}, V_{i2}\cdots V_{iD})^{\mathrm{T}}$，其个体极值为 $\boldsymbol{P}_i=(P_{i1}, P_{i2}\cdots P_{iD})^{\mathrm{T}}$，全局极值为 $\boldsymbol{G}_i=(G_{i1}, G_{i2}\cdots G_{iD})^{\mathrm{T}}$。在每一次迭代过程中，粒子通过个体极值和全局极值更新自己的速度和位置，更新公式如下：

$$\boldsymbol{V}_{id}^{k+1}=\omega\boldsymbol{V}_{id}^{k}+c_1r_1(\boldsymbol{P}_{id}^{k}-\boldsymbol{X}_{id}^{k})+c_2r_2(\boldsymbol{G}_{id}^{k}-\boldsymbol{X}_{id}^{k}) \qquad (4)$$

$$\boldsymbol{X}_{id}^{k+1}=\boldsymbol{X}_{id}^{k}+\boldsymbol{V}_{id}^{k+1} \qquad (5)$$

其中 ω 为惯性权重；$d=1, 2\cdots D$；$i=1, 2\cdots n$；k 为当前迭代次数；$\boldsymbol{V}_{id}$ 是粒子的速度；c_1 与 c_2 为非负的常数，称为加速因子。r_1 与 r_2 为分布于 [0，1] 之间的随机数。第 d （$1\leqslant d\leqslant D$）维的位置变化范围为 $[-X_{\max d}, X_{\max d}]$，速度变化范围为 $[-V_{\max d}, V_{\max d}]$，在问题的搜索空间内，可设定 $V_{\max d}=k\cdot X_{\max d}$，$0.1\leqslant k\leqslant 1.0$。迭代中若出现粒子的位置或者速度超过边界范围则取边界值。

3 PSO-SVM 的长微博贴图识别算法

本文采用径向基函数作为 SVM 的核函数，利用粒子群算法的局部搜索能力和全局搜索能力，对支持向量机的参数进行合理寻优，得到参数的最优解。基于 PSO - SVM 的长微博贴图识别算法的具体步骤如下：首先，初始化一个种群规模为 n 的粒子群，并设置加速常数、最大进化代数、粒子最大速度，初始化 SVM 的误差惩罚参数 c 和核参数 g 的搜索空间范围，并在允许的范围内随机地设定每个粒子的初始位置和初始速度。接着根据适应度函数计算每个粒子的适应度值 $f(x_i)$，对于每个粒子，将其 $f(x_i)$ 和该粒子自身的最优适应度值 $f(P_{\text{best}})$ 相比较，若 $f(x_i)<f(P_{\text{best}})$，则调整 $f(P_{\text{best}})=f(x_i)$，并将粒子的当前位置作为该粒子的最优位置 P_{best}，同时将 $f(x_i)$ 与所有粒子的最优位置 $f(G_{\text{best}})$ 相比较，若 $f(x_i)<f(G_{\text{best}})$，则调整 $f(G_{\text{best}})=f(x_i)$，并将粒子的当前位置作为所有粒子的最优位置 $f(G_{\text{best}})$。根据式 (4) 和式 (5) 更新粒子的速度和位置，从而得到新的粒子位置，即可以得到新的 SVM 的参数值；最后判断得到的结果是否满足终止条件（达到最大迭代次数或者最小适应度阈值）。若是，则输出最优参数，否则，返回重新计算适应度函数 $f(x_i)$ 重复上述步骤。

4 实 验

本文以新浪微博中的长微博贴图作为实验数据，从其提供的开放平台 open. weibo. com 中下载相应的 Java 开发程序包，编写并调试微博采集的程序进行微博贴图信息的爬取工作，本文通过人工标注的方式对采集的博文进行标注，选取贴图 540 幅，其中长微博贴图 254 幅，非长微博贴图 286 幅，所有采集的贴图均为 JPG 格式，每幅长微博贴图的宽度集中在 440 像素左右。为了验证本文提出的方法，以 Matlab 2012a 和台湾大学林智仁教授团队开发的 LibSVM[7] 作为实验平台进行两组对比实验，其中长微博贴图和非长微博贴图各选择 2/3 作为训练集，剩余的 1/3 作为测试集。

4.1 长微博贴图的特征表示

图像特征提取是图像识别的关键步骤，图像特征提取的效果如何直接决定着图像识别的效果。在文献 [8] 中介绍了 4 种图像特征常用的提取方法，本文结合长微博贴图中文字和背景交叉相间的纹理特征，精心选择了颜色特征中的颜色矩和纹理特征中灰度共生矩阵表示长微博贴图。颜色矩是一种简单且有效的特征表示方式，它的数学基础是图像中任何颜色的分布都可以用它的矩来表示[9]。但是它是使用低阶矩来表示图像的，对图像的分辨能力比较低，为了弥补这样的不足，本文引入了纹理特征中比较常用的灰度共生矩阵来补充描述长微博贴图。文献 [10] 提出了 14 种基于灰度共生矩阵的纹理特征参数，本文采用描述能力较好的 6 个参数：角二阶矩、对比度、相关、逆差矩、熵和方差。其中，角二阶矩能够反映图像的均匀程度；对比度能够反映图像的清晰程度；相关能够反映纹理的主方向；逆差矩是反映图像纹理的同质性，度量图像纹理局部变化的多少；熵反映了图像具有的信息量，即图像中纹理的复杂程度或非均匀度；若纹理越复杂，熵具有较大值，若灰度越均匀，熵则越小；和方差可以反映纹理的周期。

4.2 对长微博贴图的 PSO - SVM 识别实验

本实验利用 RBF 作为核函数，利用 PSO 算法对 SVM 的模型参数进行优化，然后基于颜色矩 $\boldsymbol{F}_{\text{CM}}$、灰度共生矩阵 $\boldsymbol{F}_{\text{GLCM}}$ 及其组合分别作为 SVM 分类器的输入，利用得到的最优参数对长微博贴图和非长微博贴图进行分类。误差惩罚参数 c 和核函数 g 的搜索范围分别为 $2^{-4}\sim 2^{10}$ 和 $2^{-4}\sim 2^{6}$。PSO 算法参数的确定对识别的结果有很大的影响。本文针对新浪微博中长微博贴图的识别问题进行了精心的设计，最终确定了 PSO 算法的初始条件，即种群数量为 20，粒子维度为 2，加速度因子 c_1、c_2 均为 2，最大迭代次数设置为 1 000，分别采用单一特征和组合特征对长微博贴图进行识别，长微博贴图的准确率和召回率如表 1 所示。

从表 1 中可以看出，相对于使用两种单一的图像特征对两类图像进行分类，本算法分类效果较好，准确率和召回率分别达到 91.62% 和 90.48%，

明显高于采用其它两种特征的分类方法。这就说明在粒子群优化算法下，颜色矩与灰度共生矩阵之间具有优势互补的特性，更能表现出长微博贴图的特征。

表 1 采用单一特征及其组合对长微博贴图的 PSO - SVM 识别结果

Table 1 PSO - SVM results using diverse features

特征	c	g	准确率/%	召回率/%
$\boldsymbol{F}_{\mathrm{GLCM}}$	0.010	3.084	91.30	90.05
$\boldsymbol{F}_{\mathrm{CM}}$	3.705	0.010	85.47	84.52
$\boldsymbol{F}_{\mathrm{GLCM}}+\boldsymbol{F}_{\mathrm{CM}}$	28.572	0.010	91.62	90.48

4.3 对长微博贴图的 SVM 识别实验

为了衡量本文利用 PSO 对支持向量机参数优化的有效性，本实验采用传统支持向量机作为参比模型进行比较，采用相同数据集，其中参数采用 5 折交叉验证的网格搜索法进行优化。误差惩罚参数 c 和核函数 g 的步进大小均设为 0.5；同 PSO 算法，c 和 g 的搜索范围也分别设为 $2^{-4}\sim2^{10}$ 和 $2^{-4}\sim2^{6}$。分别采用单一特征和其组合特征对两类图像进行分类，识别长微博贴图的准确率和召回率如表 2 所示。

表 2 采用单一特征及其组合对长微博贴图的 SVM 识别结果

Table 2 SVM results using diverse features

特征	c	g	准确率/%	召回率/%
$\boldsymbol{F}_{\mathrm{GLCM}}$	1.380	97.251	89.26	90.05
$\boldsymbol{F}_{\mathrm{CM}}$	0.930	43.710	88.58	88.29
$\boldsymbol{F}_{\mathrm{GLCM}}+\boldsymbol{F}_{\mathrm{CM}}$	106.020	1.070	86.62	89.36

从表 2 中可看出，在网格搜索算法下，相对于使用单一的长微博特征对两类贴图进行分类对比时，采用颜色和纹理的组合特征时分类效果反而较差。对比表 1 和表 2 可以看出，PSO 算法对 SVM 的参数优化性能明显优于网格搜索法，并且具有较高的识别率，尤其是采用组合特征时，其优越性更加明显。

5 结 论

本文在 SVM 算法的基础上结合 PSO 算法提出了一种 PSO - SVM 算法来识别长微博贴图，该方法利用 PSO 算法对 SVM 模型参数中的误差惩罚参数和核函数进行优化得到最佳分类模型，并将本算法应用到支持向量机对新浪微博长微博贴图的识别中。实验结果表明，PSO - SVM 算法在基于颜色矩和灰度共生矩阵的组合特征对描述新浪微博中长微博贴图分类具有较理想的效果，有助于对长微博贴图进行识别。相对于基于网格搜索法优化的 SVM 算法，该算法具有较高的准确率和召回率，同时对社交媒体中其它的贴图分类方法具有一定的适用性和参考价值。

参考文献：

[1] 谢菲．图像纹理特征的提取和图像分类系统研究及实现［D］．成都：电子科技大学，2009.

[2] Cortes C，Vapnik V. Support－vector networks［J］. Machine Learning，1995，20（3）：273 - 297.

[3] Cristianini N，Shawe - Taylor J. An Introduction to Support Vector Machines and Other Kernel - Based Learning Methods［M］. London：Cambridge University Press，2000.

[4] 高锦．基于 SVM 的图像分类［D］．西安：西北大学，2010.

[5] 胡旺，李志蜀．一种更简化而有效的粒子群优化算法［J］．软件学报，2007，18（4）：860－868.

[6] 付燕，聂亚娜，靳玉萍，等．PSO - SVM 算法在肝脏 B 超图像识别中的应用［J］．计算机测量与控制，2012，20（9）：2491 - 2500.

[7] Osuna E，Freund R，Girosi F. Training support vector machines：An application to face detection［C］//Proceedings of CVPR′97 Puerto Rico. San Juan：IEEE Computer Society，1997.

[8] 翟俊海，赵文秀，王熙照．图像特征提取研究［J］．河北大学学报：自然科学版，2009，29（1）：106 - 112.

[9] 汤海缨，庄天戈．计算机色彩模型在图像显示与分割中的应用［J］．计算机学报，1999，22（41）：375 - 382.

[10] 苑丽红，付丽，杨勇，等．灰度共生矩阵提取纹理特征的实验结果分析［J］．计算机应用，2009，29（4）：1018 - 1021.