DOI:10.16515/j.cnki.32-1722/n.2018.02.0003

基于大间距思想的虚拟样本优化求解分类算法

陶玉婷[1,2],周 波[1],张 娜[1],张文月[1]

(1.金陵科技学院软件工程学院,江苏 南京 211169;2.南京大数据研究院,江苏 南京 211169)

摘 要:在训练样本中,某些类与其他类的一些样本混杂或距离很近,后者被称为边界异类。此时,类中心离边界异类也近,从而影响最小类中心分类器的识别率。基于大间距思想,提出了一种新的分类算法,旨在改进最小类中心分类器。新分类算法为每类求解一个虚拟样本,使其尽可能排斥异类样本,并让每类虚拟样本取代该类中心点做分类。与类中心相比,虚拟样本离边界异类更远,因此增强了分类的可靠性。在CENPARMI手写阿拉伯数字库和Yale B人脸数据库上的实验表明,新分类算法的分类效果优于最小类中心分类器。

关键词:最小类中心分类器;大间距;边界异类样本;虚拟样本

中图分类号:TP311 **文献标识码**:A **文章编号**:1672-755X(2018)02-0010-05

Classification Algorithm of Virtual Sample Optimization Based on the Large Margin Criterion

TAO Yu-ting[1,2], ZHOU Bo[1], ZHANG Na[1], ZHANG Wen-yue[1]

(1. Jinling Institute of Technology, Nanjing 211169, China; 2. Nanjing Institute of Big Data, Nanjing 211169, China)

Abstract: In the training data-set, some classes are located very close to, or mixed with the samples from other classes, therefore the latter are called marginal heterogeneous samples. In this case, the class center is close to these marginal heterogeneous ones, and impairs the classification accuracy. Based on the ideology of large margin, we propose a new classification algorithm, to modify the minimum class center classifier. Our proposed algorithm will work out one virtual sample for each class, which is supposed to be distant from marginal heterogeneous samples; finally the virtual sample is taken for classification instead of the class center. Therefore, the virtual sample is farther to the marginal heterogeneous samples than the class center, and the reliability of classification is enhanced. The experimental results on CENPARMI handwritten numerals database and Yale B face database show that our proposed classifier outperform the minimum class center classifier.

Key words: minimum class center classifier; large margin; marginal heterogeneous samples; virtual sample

图像分类是在图像样本集进行特征提取的基础上,将其进行类别区分的过程[1]。常见的分类器有K近邻分类器(K-Nearest Neighbor,KNN)[2]、支持向量机(Support Vector Machine,SVM)[3]和最小类中心分类器(Minimum Class Center Classifier)[4]等。当KNN的近邻参数$K=1$时,即退化成了最近邻分类

收稿日期:2018-05-11

基金项目:金陵科技学院博士科研启动基金(jit-b-201617);智能人机交互科技创新团队(金陵科技学院科技创新团队10186001)

通信作者:陶玉婷(1985—),女,江苏南京人,讲师,博士,主要从事模式识别、优化方法的研究。

器(Nearest Neighbor，NN)。最小类中心分类器，即最小距离分类器，计算测试样本到每类训练样本的中心点距离，将测试样本归到中心点距离最小的那一类。然而，当某类样本与异类样本混杂(或很近)时，类中心点可能离异类样本很近，称其为边界异类。此时，最小类中心分类器可能会将边界异类错分到该类中，从而造成误判，降低识别率。

近年来，最大间距准则成为模式识别领域的研究热点，其基本思想是寻找一个最优投影矢量，使得投影后不同类别样本之间的间距最大[5]。比如，程国提出了基于中间值的最大间距准则[6]，王振海提出了融合奇异值分解和最大间距准则的人脸识别方法[7]，均取得了较好的识别效果。

受到大间距思想的启发，本文提出了一种新的分类算法，旨在改进最小类中心分类器。对于每个类，该算法构建关于虚拟样本的二次目标函数，采用求一阶导数的优化方法[8]，求解排斥边界异类同时靠近类内样本的虚拟样本。最后，让每个类的虚拟样本取代该类中心点，来做分类。与类中心点相比，虚拟样本在优化求解后会离边界异类更远。本文的实验分别选用国际上通用的模式识别领域 CENPARMI 手写阿拉伯数字库[9-10] 和 Yale B 人脸库[11] 中的数据，先采取主成分分析(Principle Component Analysis，PCA)[12]方法进行特征提取，再将本文提出的新分类算法与其他分类器的识别率进行实验对比。结果表明，新分类算法的识别率要普遍高于最小类中心分类器。

1　最小类中心分类器

最小类中心分类器是一种最基本的分类方法[4]，它先计算测试样本 x_{test} 到训练样本各类别均值(类中心点)的距离，然后将 x_{test} 归到距离中最小的那一类，即：假设有 C 个类，每类中心点(即均值)$\{m_1, m_2, \cdots, m_c\}$；计算测试样本 x_{test} 到各类中心的欧氏距离 $d_i = \| x_{\text{test}} - m_i \|_2^2$。若 $d_z = \min\{d_1, \cdots, d_c\}(z = 1, \cdots, C)$，则 x_{test} 被判为第 z 类。

2　虚拟样本优化求解分类算法

2.1　算法简介

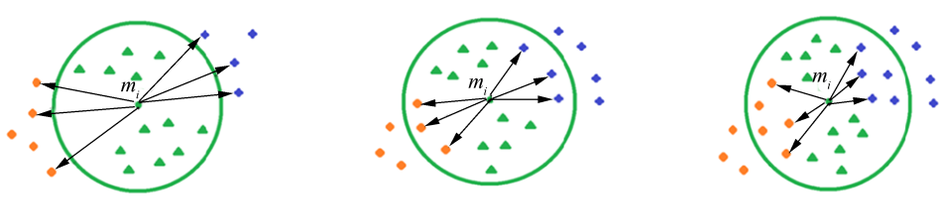
该算法构建关于虚拟样本的二次目标函数，采用求一阶导数的优化方法，求解排斥边界异类同时靠近类内样本的虚拟样本。最后，把每个类的虚拟样本取代各自类的中心点，来做分类。

2.2　详细步骤

假设总共有 N 个训练样本，C 个类，每类有 N_i 个样本，第 i 类中心点(即均值)为 m_i。对于第 i 类($i = 1, \cdots, C$)，有如下方法：

2.2.1　*最大半径 R_i 构造*　训练样本 $X_i = (x_{i1}, x_{i2}, \cdots, x_{iN_i})$。计算出该类每个训练样本 x_{ij} ($j = 1, 2, \cdots, N_i$)到 m_i 的欧氏距离，即 $d_{ij} = \| x_{ij} - m_i \|_2^2$。在此 N_i 个距离中取最大值作为该类的最大半径，即 $R_i = \max\{d_{i1}, \cdots, d_{iN_i}\}$。

2.2.2　*寻找以 R_i 为半径圆域内的异类样本*　在其他所有 $C-1$ 个类(都是异类)的 $N-N_i$ 个训练样本里，找出与 m_i 距离小于 R_i 的样本。基于该步骤，将会出现三种情况：1)在 R_i 范围内(圆域内)没有异类样本，即没有混杂，如图 1(a)；2)在 R_i 范围内(圆域内)异类样本少量混杂，如图 1(b)；3)在 R_i 范围内(圆域内)异类样本较多，即混杂厉害，如图 1(c)。



(a) R_i 范围内无混杂　　(b) R_i 范围内少量混杂　　(c) R_i 范围内混杂厉害

图 1　R_i 范围内异类样本混杂的三种情况

2.2.3　*边界异类的选取及异类近邻关系的构造*　事先给定阈值 K(该参数可调节)，根据 2.2.2 的三种不

同混杂情况，分别做不同的边界异类选取(边界异类为 T_i 个)：1)无混杂。则从圆域外寻找 K 个离 m_i 最近的异类样本，即令 $T_i=K$。2)少量混杂。则把这些异类样本个数记为 T_i，此时 $0<T_i<K$。3)混杂厉害。则取前 K 个离 m_i 最近的异类样本，即令 $T_i=K$。

如果把单个边界异类样本记作 y，则第 i 类的边界异类样本集合为 $Y_i=[y_1,\cdots,y_{T_i}]$。

2.2.4 *构建目标函数* 式(1)构建了以虚拟样本 S_i 为变量的二次目标函数 f_i，旨在使 S_i 到第 i 类的类内样本总体距离最小化(即等号右边第一项)，同时到边界异类样本的距离最大化(即等号右边第二项)，以达到函数值 f_i 最小化的目标。在函数 f_i 中，$\boldsymbol{e}_i=(1\quad\cdots\quad 1)^{\mathrm{T}}$($\boldsymbol{e}_i$ 中元素个数为 N_i)，$\boldsymbol{p}_i=(1\quad\cdots\quad 1)^{\mathrm{T}}$($\boldsymbol{p}_i$ 中元素个数为 T_i)。

$$f_i=\min_{S_i}\|X_i-S_i\boldsymbol{e}_i^{\mathrm{T}}\|_F^2-\|Y_i-S_i\boldsymbol{p}_i^{\mathrm{T}}\|_F^2 \tag{1}$$

式(1)中，函数 f_i 是关于变量 S_i 的二次非线性函数，将 f_i 对变量 S_i 求导[8]，即：

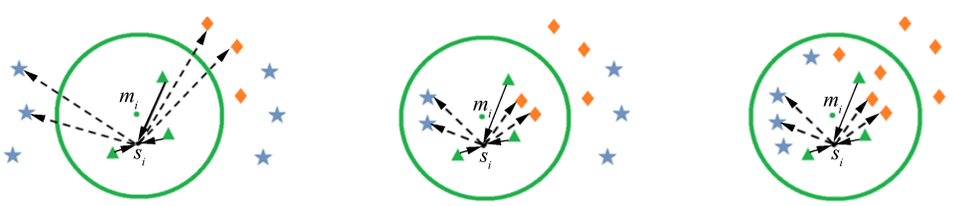
$$\frac{\partial f_i}{\partial S_i}=-2X_i\boldsymbol{e}_i+2S_i\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{e}_i+2Y_i\boldsymbol{p}_i-2S_i\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i \tag{2}$$

令 $\frac{\partial f_i}{\partial S_i}=0$，整理得：

$$S_i=\frac{1}{N_i-T_i}(X_i\boldsymbol{e}_i-Y_i\boldsymbol{p}_i) \tag{3}$$

式(1)中 S_i 的取值，能使函数 f_i 达到最小值。因此，式(3)的 S_i 是优化求解后第 i 类的虚拟样本。

我们对每类训练样本都构建如式(1)的二次目标函数，通过优化求解，得到如式(3)的虚拟样本 S_i 的表达式。图 2 是图 1 中三种混杂情况下 S_i 的示意图。其中，箭头向外的虚线表示 S_i 把边界异类样本排斥出去，即距离最大化；箭头向内的实线表示 S_i 把类内样本向里聚拢，即距离最小化。图 2 中 S_i 在保持与类内样本近距离的前提下，和类中心点 m_i 相比，与边界异类样本的距离更远了。



(a) R_i 范围内无混杂　(b) R_i 范围内少量混杂　(c) R_i 范围内混杂厉害

图 2　三种混杂情况下 S_i 的示意

2.2.5 *判断类别* 计算测试样本 x_{test} 与每个类的虚拟样本 S_i 的欧氏距离 $d_i=\|x_{\mathrm{test}}-S_i\|_2^2$。若 $d_z=\min\{d_1\quad\cdots\quad d_c\}(z=1,\cdots,C)$，则 x_{test} 被判为第 z 类。

3　实验及性能分析

3.1　数据库介绍

1)CENPARMI 手写阿拉伯数字库。该库包含 0～9 共 10 类手写体阿拉伯数字样本，每类有 600 个样本(图 3)。本文分别采用该库的 256 维 Gabor 变换特征[9]和 121 维 Legendre 矩特征[10]来做实验。

2)Yale B 人脸数据库。该库[10]包含 38 人，每个人 9 种姿态，64 种光照条件，尺寸 168 * 192。先对图像下采样变成 42 * 48(即 2016 维特征)，再做直方图均衡化，如图 4 所示。

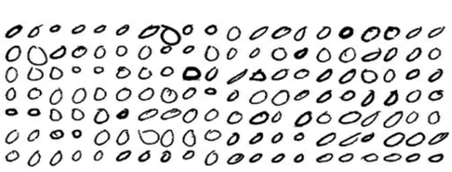


图 3　CENPARMI 手写阿拉伯数据库中的部分样本

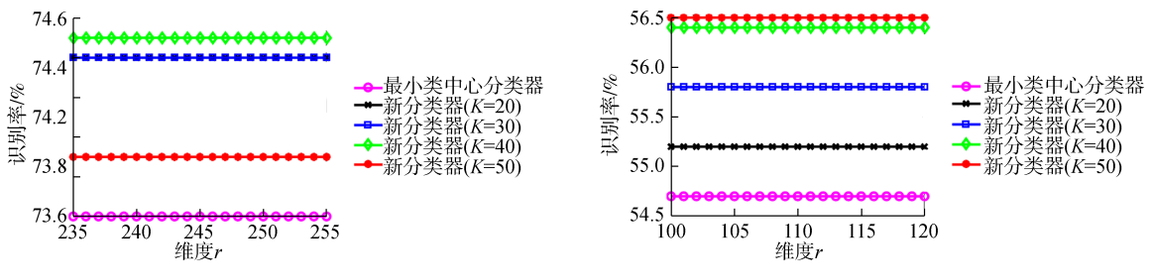
图 4　Yale B 人脸库中一个人的示例样本

3.2　实验结果与分析

在分类之前，先采用主成分分析(Principle Component Analysis，PCA)[12]降维来进行特征提取。如

果原始样本 X 的特征维度为 p，样本总个数为 N，则所提取的特征维度总共是 $\min\{p,N-1\}$（去掉一个总体均值）。

1）在 CENPARMI 手写阿拉伯数据库上的实验。每类选取前 200 个样本作为训练样本，后 400 个样本作为测试样本。新分类算法中，设定 $K=\{20,30,40,50\}$。Gabor 特征取 PCA 降维后的 $r\in[235,255]$；而 Legendre 特征取 $r\in[100,120]$。最小类中心分类器和新分类算法的识别率随维度 r 的变化情况分别如图 5(a)和(b)。可以看出，随着维度 r 的增加，同一 K 值的识别效果稳定。



注：$K=20$ 和 $K=30$ 时的线段重合。

（a）Gabor 特征的识别率　　（b）Legendre 特征的识别率

图 5　CENPARMI 手写阿拉伯数字库的识别率随特征维度的变化

当取全部维度（Gabor 取 256 维，Legendre 取 121 维）时，把新分类算法与最小类中心分类器、K 近邻分类器（$K=20,30,40,50$）、最近邻分类器以及支持向量机的识别率进行实验对比，结果如表 1。总体上看，SVM 识别率最优，其次是 NN。在 Gabor 特征下，KNN 识别效果不如新分类算法；而在 Legendre 特征下却相反。综合图 5 和表 1，不论在 Gabor 特征还是 Legendre 特征，当 K 的取值从 20 递增到 40 时，新分类算法的识别率越来越高，且都高于最小类中心分类器。

表 1　CENPARMI 手写阿拉伯数字库下的各个分类器识别率

分类器		分类器在不同特征下的正确识别率/%	
		Gabor 特征	Legendre 特征
虚拟样本优化求解分类算法	$K=20$	74.4	55.2
	$K=30$	74.4	55.8
	$K=40$	74.5	56.4
	$K=50$	73.9	56.5
K 近邻分类器（KNN）	$K=20$	71.7	83.6
	$K=30$	68.9	81.1
	$K=40$	66.7	78.8
	$K=50$	64.8	77.6
最小类中心分类器		73.6	54.7
最近邻分类器（NN）		79.5	88.2
线性支持向量机（linear-SVM）		85.1	89.5
非线性支持向量机（nonlinear-SVM）		87.1	44.9

值得注意的是，新分类算法里，当 $K=50$ 时，识别效果反而不如 K 取 20、30、40 的情况。对照图 2，可能的原因是在这 50 个边界异类样本（按照与类中心距离从近到远排列）中，前 40 个在圆域的同一侧，后 10 个在另一侧。同时最大化虚拟样本与两侧异类的距离，会使虚拟样本 S_i 偏向类中心点 m_i。

2）在 Yale B 人脸数据库上的实验。该库有 38 人，即 38 个类。每类选取前 16 个作为训练样本，后 48 个作为测试样本。新分类算法中，设定 $K=\{2,3,4,5\}$，PCA 降维后特征维度取 $r\in[587,607]$。最小类中心分类器和新分类算法的识别率随维度 r 的变化情况如图 6。随着维度 r 的增加，同一线条的识别效果稳定。

最后，取 PCA 降维后的全部维度（即 608 维），把新分类算法与其他方法的识别率进行实验对比，如表 2 所示。线性 SVM 识别率最优，其次是 NN，然后是 KNN；而非线性 SVM 却不如新分类算法中 $K=5$ 的情况。综合图 6 和表 2，当 K 的取值从 2 递增到 5 时，新分类算法的识别率越来越高，且都高于最小类中

心分类器。当 $K=5$ 时，新分类器识别率最优。

表 2 Yale B 人脸数据库下的各个分类器识别率

分类器		分类器的正确识别率/%
虚拟样本优化求解分类算法	$K=2$	47.6
	$K=3$	51.0
	$K=4$	51.4
	$K=5$	53.3
K 近邻分类器(KNN)	$K=2$	58.4
	$K=3$	57.7
	$K=4$	56.0
	$K=5$	57.5
最小类中心分类器		42.9
最近邻分类器(NN)		66.4
线性支持向量机(linear-SVM)		83.9
非线性支持向量机(nonlinear-SVM)		51.5

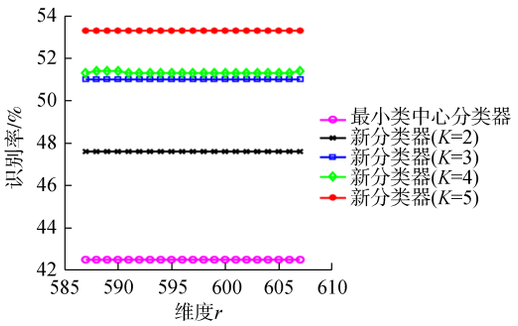


图 6 Yale B 人脸库识别率随特征维度的变化

4 结 语

新分类算法用一个固定的虚拟样本代表整个类；而 KNN 和 NN 针对不同的测试样本，都从每类训练集中选出最近的一个或多个样本来比较距离。所以前者的搜索空间和灵活度不及后两者。SVM 选出边界少量的样本作为支持向量，而离类边界较远的正负类样本没有参与最大边距的决策。相比之下，新分类算法只考虑了类边界上的异类样本，故无法排除离类边界较远的类内样本对边界信息的负面影响。总之，新分类算法不如 KNN、NN 和 SVM，根本原因在于分类思想和算法设计都不同，不具备可比性。新分类算法只是结合了类间大间距的思想，旨在改进最小类中心分类器。理论上，与类中心点相比，虚拟样本离边界异类的距离更远。本文的实验也证明了新分类算法的识别效果确实普遍高于最小类中心分类器。

参考文献：

[1] 徐彩云. 图像识别技术研究综述[J]. 电脑知识与技术，2013(10)：2446－2447

[2] 闭小梅，闭瑞华. KNN 算法综述[J]. 科技创新导报，2009(14)：31

[3] CSIE. 最新的 3.22 版本支持向量机 matlab 代码工具包[EB/OL]. [2017－11－20]. http://www.csie.ntu.edu.tw/～cjlin/libsvm/

[4] 任靖，李春平. 最小距离分类器的改进算法——加权最小距离分类器[J]. 计算机应用，2005，25(5)：992－994

[5] 陈才扣，杨静宇. Fisher 大间距线性分类器[J]. 中国图象图形学报，2007(12)：2143－1247

[6] 程国. 基于中间值的最大间距准则特征提取方法[J]. 甘肃科学学报，2014，26(4)：21－24

[7] 王振海. 融合奇异值分解和最大间距准则的人脸识别方法[J]. 计算机工程与应用，2011，47(8)：164－166

[8] Petersen K B，Pedersen M S. The matrix cookbook [EB/OL]. [2018－01－05]. http://matrixcookbook.com

[9] Hamamota Y，Uchimura S，Watanabe M，et. al. Recognition of handwritten numerals using Gabor features[J]. International Conference on Pattern Recognition (ICPR)，2002(3)：250－253

[10] Liao S X，Pawlak M. On image analysis by moments[J]. Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence，1996，18(3)：254－266

[11] Georghiades A S，Belhumeur P N，Kriegman D J. From few to many：illumination cone models for face recognition under variable lighting and pose [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2002，23(6)：643－660

[12] Kirby M，Sirovich L. Application of the Karhunen-Loeve procedure for the characterization of human faces [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1990，12(1)：103－108

（责任编辑：湛 江）