Journal of Computer Applications
计算机应用,2013,33(11):3300-3304
ISSN 1001-9081
CODEN JYIIDU
2013-11-01
http://www.joca.cn

文章编号:1001-9081(2013)11-3300-05
doi:10.11772/j.issn.1001-9081.2013.11.3300

基于超球体多类支持向量数据描述的医学图像分类新方法

谢国城*,蒋　芸,陈　娜
(西北师范大学 计算机科学与工程学院,兰州 730070)
(*通信作者电子邮箱 xieguocheng@126.com)

摘　要:针对乳腺X光医学图像多分类问题中训练速度比较慢的问题,提出超球体多分类支持向量数据描述(HSMC-SVDD)分类算法,即把超球体单分类支持向量数据描述直接扩展到超球体多分类支持向量数据描述。通过对乳腺X光图像提取灰度共生矩阵特征;然后用核主成分分析(KPCA)对数据进行降维;最后用超球体多分类支持向量数据描述分类器进行分类。由于每一类样本只参与构造一个超球体的训练,因此训练速度明显提高。实验结果表明,这种超球体多分类支持向量数据描述分类器的平均训练时间为21.369 s,训练时间比Wei等(WEI L Y, YANG Y Y, NISHIKAWA R M, *et al.* A study on several machine-learning methods for classification of malignant and benign clustered micro-calcifications. IEEE Transactions on Medical Imaging, 2005, 24(3): 371-380)提出的组合分类器(平均训练时间40.2 s)减少了10~20 s,分类精度最高达76.6929%,适合解决类别数较多的分类问题。

关键词:乳腺X光图像;多类支持向量数据描述;灰度共生矩阵;核主成分分析
中图分类号:TP391.413　　**文献标志码**:A

New medical image classification approach based on hypersphere multi-class support vector data description

XIE Guocheng*, JIANG Yun, CHEN Na
(*College of Computer Science and Engineering, Northwest Normal University, Lanzhou Gansu* 730070, *China*)

Abstract: Concerning the low training speed of mammography multi-classification, the Hypersphere Multi-Class Support Vector Data Description (HSMC-SVDD) algorithm was proposed. The Hypersphere One-Class SVDD (HSOC-SVDD) was extended to a HSMC-SVDD as a kind of immediate multi-classification. Through extracting gray-level co-occurrence matrix features of mammography, then Kernel Principle Component Analysis (KPCA) was used to reduce dimension, finally HSMC-SVDD was used for classification. As each category trained only one HSOC-SVDD, its training speed was higher than that of the present multi-class classifiers. The experimental results show that compared with the combined classifier, in which the average train time is 40.2 seconds, proposed by Wei (WEI L Y, YANG Y Y, NISHIKAWA R M, *et al.* A study on several machine-learning methods for classification of malignant and benign clustered micro-calcifications. IEEE Transactions on Medical Imaging, 2005, 24(3): 371-380), the training time of HSMC-SVDD classifier is 21.369 seconds, the accuracy is up to 76.6929% and it is suitable for solving classification problems of many categories.

Key words: mammograph; multi-class Support Vector Data Description (SVDD); Gray-Level Co-occurrence Matrix (GLCM); Kernel Principle Component Analysis (KPCA)

0　引言

乳腺癌是女性常见的肿瘤疾病之一,由于乳腺癌的病发机理还未完全弄清楚,因此早期诊断对防治乳腺癌十分重要[1]。乳腺癌的临床诊断方法包括触摸式诊断、组织学诊断、细胞学诊断和影像学诊断四大类,其中影像诊断是最适合适龄女性乳腺癌诊断方法,乳腺X光摄影技术是最常见乳腺癌早期诊断方法[2]。随着计算机技术的不断发展,医学图像上的计算机辅助诊断技术也得到了迅猛的发展,其中常用的方法有关联规则、决策树、遗传算法、人工神经网络神经网络、贝叶斯、粗糙集、模糊聚类和支持向量机(Support Vector Machine,SVM)等。支持向量机在近些年发展比较迅速,它是建立在统计学习理论中的VC理论和结构化风险最小原理基础上实现的一种机器学习方法[3],能较好地解决小样本、非线性、高维数和局部极小点等实际问题。Tax等[4]在1999年提出支持向量数据描述(Support Vector Data Description, SVDD),它是一种源于统计学习理论和SVM的全新的数据描述方法,与SVM寻求最优超平面不同,SVDD包容所有目标样本数据的最小超球体。并且相比SVM,SVDD有着复杂性低、易移植和训练速度快等优点,在信用卡欺诈检测、入侵检测、人脸识别领域等有着广泛的应用,同时SVDD在解决不平衡数据分类和多示例分类中已成为一种新方法。SVDD在单分类问题中分类效果比较好,其中超球体单分类支持向量机[5]表现的效果更好,但是超球体单分类SVM缺乏有效的训

收稿日期:2013-06-03;**修回日期**:2013-07-14。　**基金项目**:国家自然科学基金资助项目(61163036, 61263036);甘肃省自然科学基金资助项目(1010RJZA022,1107RJZA112);2012年度甘肃省高校基本科研业务费专项;甘肃省高校研究生导师项目(1201-16);西北师范大学第三期知识与创新工程科研骨干项目(nwnu-kjcxgc-03-67)。

作者简介:谢国城(1987-),男,江西瑞金人,硕士研究生,主要研究方向:数据挖掘、粗糙集;　蒋芸(1970-),女,浙江绍兴人,副教授,博士,主要研究方向:数据挖掘、粗糙集、模式识别、机器学习;　陈娜(1987-),女,山东泰安人,硕士研究生,主要研究方向:数据挖掘、粗糙集。

练算法,所以其在应用中受到限制。在实际问题的解决中,很多分类问题都是包含多个类别的多分类问题,而目前大多数多分类器基本上都是由二分类器组合而成的,当分类类别数达到一定的数量时,这种经二分类器组合而成的多分类器将会遇到诸如样本训练阶段速度较慢的问题。例如在医学图像识别问题中,从图像中提取出来的信息特征量往往比较大,如果直接用二分类器组合而成的多分类器,训练速度相对来说较慢,而文献[6-7]中分别提到的关于多球体支持向量数据描述和多分类支持向量机的基本思想和实现对进一步研究多分类问题提供了帮助。

Wei 等[8]提到用级联 AdaBoost 对标准医学图像数据集进行分类,分类精度达到 80.3%,但是训练时间比较长,平均训练时间为 40.2 s。文献[9]中提到对标准医学图像数据集进行分类的方法,如果只使用 ID3 进行分类,分类精度为 43.3%,只使用 K 最近邻(K-Nearest Neighborhood,KNN)分类法进行分类,分类精度为 40.3%,而使用 ID3 和 KNN 的组合分类器进行分类,分类精度为 47.6%。文献[10]中提到对标准医学图像数据集进行分类的方法,用主成分分析(Principal Component Analysis,PCA)和基于规则的粗糙集进行分类,分类精度为 69.27%;用 BP(Back Propagation)神经网络进行分类,分类精度为 51.51%;用学习向量量化(Learning Vector Quantization, LVQ)神经网络分类器进行分类,分类精度为 63.63%。从文献[8]中可以看出分类精度虽然达到 80.3%,但是训练速度比较慢;而文献[9-10]中的分类精度比较低。

针对上述问题,本文提出基于超球体多分类支持向量数据描述方法。一些冗余的特征信息不仅会增大分类算法在构建分类模型时的数据量,而且还会影响分类器的分类效果;所以本文提出的超球体多分类支持向量数据描述算法优点在于分类模型建立前期先运用核主成分分析(Kernel Principal Component Analysis,KPCA)来有效地对数据进行降维,然后在构建分类模型时使每一类样本只参与构造一个超球体的训练,以此来直接构造多个 SVDD 超球体的多分类器,在保证分类精度的基础上有效地提高了训练速度。乳腺 X 光医学图像标准数据集 MIAS(Mammographic Image Analysis Society)[11]分类实验效果表明:与文献[8]相比,超球体多分类 SVDD 分类器的训练时间减少了 10~20 s,而且分类精度最高达到 76.6929%。

1　KPCA 和 SVDD 的基本原理

1.1　核主成分分析

与传统的主成分分析法(PCA)相比,核主成分分析(Kernel Principal Component Analysis, KPCA)[12]引入核函数方法,通过非线性函数把输入空间映射到高维空间,在特征空间中对数据进行处理,把非线性变换后的特征空间内积运算转换为原始空间的核函数计算,从而大大简化了计算量。

核主成分分析是一种将原始数据通过非线性变换映射到高维特征空间 $\boldsymbol{F}$ 的非线性方法,因此在特征空间 $\boldsymbol{\Phi}(\boldsymbol{x}_i)(i=1,2,\cdots,l)$ 中存在:

$$\lambda \boldsymbol{V} = \overline{\boldsymbol{C}}\boldsymbol{V} \tag{1}$$

对协方差矩阵进行求特征值 $\lambda \geqslant 0$ 和特征向量 $\boldsymbol{V} \in \boldsymbol{F}\backslash\{0\}$,其中 l 为样本数。

考虑到所有的特征向量均可表示为 $\boldsymbol{\Phi}(\boldsymbol{x}_1),\boldsymbol{\Phi}(\boldsymbol{x}_2),\cdots,\boldsymbol{\Phi}(\boldsymbol{x}_l)$ 的线性张成:

$$\boldsymbol{V} = \sum_{i=1}^{l} \alpha_i \boldsymbol{\Phi}(\boldsymbol{x}_i) \tag{2}$$

通过定义一个 $l \times l$ 的矩阵 $\boldsymbol{K}$,其中 $\boldsymbol{K}_{\mu v} := (\boldsymbol{\Phi}(\boldsymbol{x}_\mu) \cdot \boldsymbol{\Phi}(\boldsymbol{x}_v))$,可以得到:

$$l\lambda \boldsymbol{\alpha} = \boldsymbol{K}\boldsymbol{\alpha} \tag{3}$$

求解式(3)就能得到特征值 $\lambda \geqslant 0$ 和特征向量,对于测试样本在特征向量空间 $\boldsymbol{V}^k$ 的投影,并将内积用核函数替换则有:

$$(\boldsymbol{v}^k \cdot \boldsymbol{\Phi}(\boldsymbol{x})) = \sum_{i=1}^{l} (\alpha_i)^k \boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}) \tag{4}$$

则核矩阵可修正为:

$$\widetilde{\boldsymbol{K}}_{ij} = \boldsymbol{K} - 1_l \boldsymbol{K} - \boldsymbol{K}1_l + 1_l \boldsymbol{K}1_l \tag{5}$$

其中 $(1_l)_{ij} := 1/l$。

1.2　支持向量数据描述

支持向量数据描述是将所要描述的样本作为一个整体,建立一个封闭紧凑的超球体,获得包含单类数据样本的最小球形边界,使目标对象尽可能多或全部被包括在球体内部,而非该类样本将没有或者尽可能少地包含在球体内部。

设 $\{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathbf{R}^d, i=1,2,\cdots,n\}$($d$ 为样本的维数)为输入空间上的训练数据集。为了构造最小超球体,在特征空间上求解以后的优化问题为:

$$\min_R \left(R^2 + C\sum_i \xi_i\right) \tag{6}$$

$$\text{s.t.} \quad \| \boldsymbol{x}_i - \boldsymbol{a} \|^2 \leqslant R^2 + \xi_i; \quad \xi_i \geqslant 0$$

运用核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 来代替内积运算,将低维空间的非线性问题转化为高维空间的线性问题,引入核函数后该问题的对偶形式为:

$$\max_R L = \sum_i \alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_i) - \sum_{i,j} \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{7}$$

$$\text{s.t.} \quad \sum_i \alpha_i = 1; \quad 0 \leqslant \alpha_i \leqslant C, i = 1,2,\cdots,n$$

在式(7)中对违反 KKT 条件的拉格朗日乘子 α_i 进行优化,直到所有的拉格朗日乘子均满足 KKT 条件为止。

假设 $\boldsymbol{\alpha}^* = [\alpha_1^*,\cdots,\alpha_N^*]^{\mathrm{T}}$ 为问题的最优解,且 $\alpha_i^* > 0$, $i=1,2,\cdots,n$,则称相应的 $\boldsymbol{x}_i$ 为支持向量。

则判断一个测试样本点 $\boldsymbol{z}$ 属于目标类的不等式:

$$\| \boldsymbol{z} - \boldsymbol{a} \|^2 = K(\boldsymbol{z},\boldsymbol{z}) - 2\sum_i \alpha_i K(\boldsymbol{z},\boldsymbol{x}_i) + \sum_{i,j} \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \leqslant R^2 \tag{8}$$

2　本文方法

2.1　超球体多分类支持向量数据描述算法

本文对单类支持向量数据描述(Hypersphere One-Class SVDD, HSOC-SVDD)进行改进,提出超球体多分类支持向量数据描述(Hypersphere Multi-Class Support Vector Data Description, HSMC-SVDD)算法,以克服多分类问题中存在的训练速度低和分类精度偏低等缺陷。

将核主成分分析数据降维的思想引入超球体多分类支持向量数据描述分类算法中能够有效地降低训练样本的数据维数,不仅能使训练速度明显提高,而且也能提高分类精度。

设 $\boldsymbol{D}$ 是训练元组和相关联的类标号的集合,每个元组可以用一个 n 维属性向量 $\boldsymbol{X}$ 可以表示为 $\boldsymbol{X} = \{x_1,x_2,\cdots,x_n\}$,其

中类标号属性中具有 T 类，每一类 n 维向量空间元素集合可以表示为 $X_t(t=1,2,\cdots,T)$，而且每类空间元素样本数为 L_t，其中第 t 类的第 i 个空间样本点可以表示为 $x_{t,i}(i=1,2,\cdots,L_t)$。每类元素 X_t 训练时只参与构造一个封闭而紧凑的超球体，使属于 X_t 的样本点全部或尽可能多地被包含在该球体内。其目标函数为：

$$\min_{R_t}\left(R_t^2+C_m\sum_{i=1}^{L_t}\xi_{t,i}\right) \tag{9}$$

$$\text{s.t.}\ \|\boldsymbol{x}_{t,i}-\boldsymbol{a}_t\|^2\leqslant R_t^2+\xi_{t,i};\ \xi_{t,i}\geqslant 0$$

其中：R_t 为第 t 类超球体的半径，$\boldsymbol{a}_t$ 第 t 类超球体的球心，$\xi_{t,j}$ 为松弛变量，C_t 为正则化参数。运用核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 来代替内积运算，引入核函数后该问题的对偶形式为：

$$\max_{R_t}L=\sum_i\alpha_{t,i}K(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,i})-\sum_{i,j}\alpha_{t,i}\alpha_{t,j}K(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j}) \tag{10}$$

$$\text{s.t.}\ \sum_i^{L_t}\alpha_{t,i}=1;0\leqslant\alpha_{t,i}\leqslant C_t,i=1,2,\cdots,L_t$$

对式(10)二次规划问题进行求解，来获得 T 个超球体。

则判断给定点 z 属于目标类的不等式：

$$d_{t,i}^2=\|\boldsymbol{z}_{t,i}-\boldsymbol{a}_i\|^2=K(\boldsymbol{z}_{t,i},\boldsymbol{z}_{t,i})-2\sum_i\alpha_{t,i}K(\boldsymbol{z}_{t,i},\boldsymbol{x}_{t,i})+\sum_{i,j}\alpha_{t,i}\alpha_{t,j}K(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j}) \tag{11}$$

$$\min_{\alpha_{t,i}}D_t=d_{t,i}^2-R_t^2 \tag{12}$$

即计算测试样本点到球 t 的球面距离的平方的最小值，则该样本点就属于 t 类。其中第 t 类球心和半径分别为：

$$a_t^2=\sum_i\sum_j\alpha_{t,i}\alpha_{t,j}K(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j}) \tag{13}$$

$$R_t^2=K(\boldsymbol{x}_{t,k},\boldsymbol{x}_{t,k})-2\sum_i\alpha_{t,i}K(\boldsymbol{x}_{t,k},\boldsymbol{x}_{t,i})+\sum_{i,j}\alpha_{t,i}\alpha_{t,j}K(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j}) \tag{14}$$

其中 $\forall\ \boldsymbol{x}_{t,k}\in SV_t<C_t$。

2.2 HSMC-SVDD 算法描述

算法描述如下。

输入　数据样本集 $\boldsymbol{D}$。

输出　分类结果 Y。

步骤1　X_t = separate($\boldsymbol{D}$);/* 将总的数据集 $\boldsymbol{D}$ 分成 T 个 n 维向量空间元素集合 X_t，每个集合中有 L_t 个样本点 */

步骤2　$train\_X_t\_unclean$ = randomGetTrainData(X_t);/* 对 T 个集合 X_t 用随机抽样抽取一定比例的数据作为构造该类球体的输入 */

步骤3　$test\_unclean$ = randomGetTestData($\boldsymbol{D}-\boldsymbol{X}$);/* 在总的样本集 $\boldsymbol{D}$ 抽取完训练样本后，在剩下的数据集中抽取测试集 */

步骤4　$train\_X_t$ = xgcDataNormalize($train\_X_t\_unclean$);/* 进行训练前的数据预处理，包括数据清理和数据标准化 */

步骤5　$testMatrix$ = xgcDataNormalize($test\_unclean$);/* 对测试集进行数据预处理 */

步骤6　$train\_X_t$ = Dimension_Reduction($train\_X_t$,KPCA); $KtestMatrix$ = Dimension_Reduction($testMatrix$,KPCA);/* 用KPCA进行数据集的降维 */

步骤7　$Ktrain\_X_t$ = ConstructKernelMatrix($train\_X_t$); $testMatrix$ = ConstructKernelMatrix($testMatrix$)/* 构造训练集和测试集的核矩阵 */

步骤8　[sv_num_t,sv_t,$alph\_sv_t$] = HSMC-SVDDProcess($Ktrain\_X_t$)/* 核心训练分类模型算法：

1）第 t 类球体圆心初值 $a_t=0$，半径 $R_t=0$，$\alpha_{t,i}=1/L_t$，计算 $d_{t,i}$。

2）在边界内和边界上找到违反 KKT 条件的第一个点 $x_{t,1}$，找到后选取 $\max|d_{t,1}^2-d_{t,2}^2|$ 的 $x_{t,2}$，然后再优化 $\alpha_{t,1}$，$\alpha_{t,2}$；否则搜索结束。

3）根据优化后的值更新式(11)、式(13)和式(14)$d_{t,1}$、$d_{t,2}$、球体圆心 a_t^2 和半径 R_t^2。返回2）继续搜索边界内和边界上的点。

4）搜索结束，输出支持向量信息。*/;

步骤9　Y = Predict($KtestMatrix$,sv_t,sv_num_t,$alph\_sv_t$);/* 测试阶段，用式(11)计算第 i 个测试样本点到第 t 个球心的距离的平方，并取式(12)的最小值，则该测试样本点属于第 t 类 */

3 实验分析

实验采用的是威廉康星乳腺癌数据集(MIAS)，它是研究乳腺 X 光图像的标准数据集，图像均为 1 024×1 024 像素的灰度图。数据集中包含 322 幅乳腺 X 光图像，其中共有 3 类：正常图像 208 幅、良性图像 63 幅和恶性图像 51 幅。

3.1 图像预处理

在乳腺癌辅助诊断中，许多客观因素都会导致图像受到噪声的污染，从而造成图像信息的不完整和不一致，因此图像预处理是图像分类中不可缺少的一步。

图像预处理的第一步就是对图像进行去噪，而目标物和背景的边界与噪声有共同的跃变特征，均值滤波器和中值滤波器在处理噪声点的同时也对边界点进行了处理。所以在处理噪声时，应该加入边界点与非边界点的确定，因此实验中使用的是 K 近邻(KNN)平滑滤波器，它是一种使用边界保持类滤波器。KNN 平滑滤波器的基本原理：以待处理的像素作为中心，取一个 $m\times m$(m 的值为3)的模板，在模板中选择 k(k 的值为5)个与待处理像素的值最接近的像素，用这 k 个像素的均值替换原来的像素值。第二步就是对去噪后的图像进行图像的增强，实验采用直方图均衡化法增强图像，使图像不会出现因亮度不均的问题影响分类效果。如图1为原始图像，图2～3分别为经过去噪和增强后的图像。

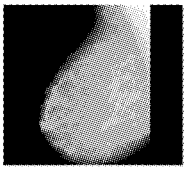

图1　原始图像

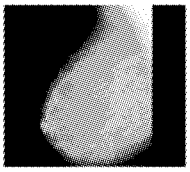

图2　去噪后的图像

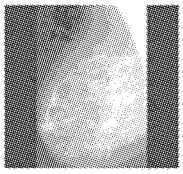

图3　增强后的图像

3.2 特征提取

图像中包含的特征千差万别，特征提取是图像分析的一个重要环节，特征提取是否恰当将直接影响后期图像的分类效果。而测度纹理的数学方法很多，本文采用的是其中比较有效的灰度共生矩阵法。因为灰度共生矩阵是建立在统计方法二阶组合条件概率密度基础上的纹理分析方法，由于纹理是灰度在空间分布位置上反复交替变化而形成的，所以图像中

两个像素一定存在灰度关系。灰度共生矩阵不仅反映了亮度分布特性,同时也反映了具有同样亮度或接近亮度的像素之间的位置分布特性,是有关图像亮度变化的二阶统计特征。灰度共生矩阵的定义:设图像某一区域有 N 个灰度值,则对应该区域的灰度共生矩阵是一个 $N \times N$ 阶矩阵,在矩阵中的位置 $(i,j)(1,\cdots,i,\cdots,N;1,\cdots,j,\cdots,N)$ 处元素是从灰度 i 的像元了离开某个固定的位置关系 $\delta = (DX,DY)$ 处像元灰度为 j 出现的概率,δ 称为位移量。本文提取了灰度共生矩阵中的 4 种统计量:角二阶矩、熵、对比度、相关系数;然后再求出相应的均值和方差。

角二阶矩:

$$f1 = \sum_{i=1}^{N}\sum_{j=1}^{N} P_\delta(i,j)^2 \tag{15}$$

熵:

$$f2 = \sum_{i=1}^{N}\sum_{j=1}^{N} P_\delta(i,j) \cdot \log P_\delta(i,j) \tag{16}$$

对比度:

$$f3 = \sum_{i=1}^{N}\sum_{j=1}^{N} P_\delta(i,j) * (i-j)^2 \tag{17}$$

相关性:

$$f4 = \sum_{i=1}^{N}\sum_{j=1}^{N} P_\delta(i,j) * (i-\mu_i) * (j-\mu_j)/\sigma_i\sigma_j \tag{18}$$

其中:

$$\begin{cases}\mu_i = \sum\limits_{i,j} i * P_\delta(i,j) \\ \mu_j = \sum\limits_{i,j} j * P_\delta(i,j)\end{cases} \tag{19}$$

$$\begin{cases}\sigma_i = \sum\limits_{i,j} (i-\mu_i) * P_\delta(i,j) \\ \sigma_j = \sum\limits_{i,j} (j-\mu_j) * P_\delta(i,j)\end{cases} \tag{20}$$

把每幅图像分成 4 块,分别每块子图像提取上述 8 个特征作为原始样本数据,共获得 32 个统计特征。

3.3　实验结果分析与比较

为了对本文提出的超球体多分类支持向量数据描述算法(HSMC-SVDD)的分类精度和训练速度进行定量分析,在事先抽取的总数据集中对正常、良性和恶性 3 种样本集进行 10 次随机抽样,分别进行 10 组仿真实验,对 3 种类别的数据集进行多分类实验,其中:训练样本集中样本数为 192,测试样本集中样本数为 130。因为训练样本和测试样本中均包含 3 类样本,并且 3 类样本同时进行测试,所以表 1 ~ 2 中的分类精度属于所有测试样本中正确分类的比例,即分类精度 = 正确分类样本数/总样本数。实验环境在 Windows XP 2 GB RAM CPU 主频 2.20 GHz,Matlab 7.0 平台下进行,在核主成分分析和构造核矩阵时所用的核函数均为高斯径向核函数。

1)实验总体结果分析,如表 1 所示。无论是训练速度还是分类精度,总体上都获得了良好的效果。其中训练时间最少的只需 14.156 s,分类精度最高达到 76.692 9%。实验中有些参数选取得好坏将影响 HSMC-SVDD 算法的分类效果,如高斯径向核函数中的参数 σ;经过实验对比得知,当 σ 的取值越大,训练的时间就越长,所以本文选取 $\sigma = 0.5$ 为最合适的值。其中的实验样本数据集经过数据预处理后直接作为 HSMC-SVDD 分类器的输入,但是由于经过预处理后的数据维数比较大,而且存在冗余数据,所以在 HSMC-SVDD 分类器中集成了核主成分分析方法。去除这些冗余数据后再进行分类模型的构建不仅在一定程度上可以提高分类精度,而且更能提高分类器的训练速度。

2)针对乳腺 X 光医学图像多分类实验,本文 HSMC-SVDD 算法与其他文献算法在训练时间和分类精度上进行比较:

①在训练时间上的比较:如表 2 所示,与文献[8]中的组合分类器相比,文献[8]中的组合分类器的平均训练时间为 40.2 s,本文的多分类器的平均训练时间为 21.368 8 s,训练时间减少了 10 ~ 20 s,如图 4 所示实验的训练时间在 14.156 ~ 44.563 s,平均训练时间为 21.368 8 s,平均测试时间为 0.428 1 s。

②在分类精度上的比较:如表 3 所示对本文方法与各种文献中的方法进行了比较,其中在分类精度上与文献[9 - 10]的比较,只利用 ID3 进行分类,分类精度为 40.3%,只利用 *K*NN 进行分类,分类精度只有 43.3%,利用 *K*NN 和 ID3 的组合分类器进行分类,分类精度为 47.6%,用主成分分析和基于规则的粗糙集进行分类,分类精度为 69.27%;用神经网络进行分类,分类精度为 51.51%;用学习向量量化神经网络分类器进行分类,分类精度为 63.63%。而采用本文方法进行分类,平均分类精度达到 67.76%,最高分类精度达 76.692 9%。

表 1　HSMC-SVDD 算法在乳腺 X 光医学图像 3 分类中的分类效果

组号	训练时间/s	测试时间/s	分类精度/%	组号	训练时间/s	测试时间/s	分类精度/%
1	16.844	0.469	64.330 7	6	16.546	0.422	72.094 5
2	18.218	0.422	64.730 7	7	14.156	0.422	65.905 5
3	21.187	0.453	56.692 9	8	17.766	0.422	68.818 9
4	44.563	0.422	66.456 7	9	21.735	0.421	76.692 9
5	20.862	0.422	64.881 9	10	21.813	0.406	67.007 9

表 2　训练速度上与其他方法的比较

分类方法	分类精度/%	平均训练时间/s
级联 AdaBoost[8]	80.340 0	40.200 0
本文方法	76.692 9	21.368 8

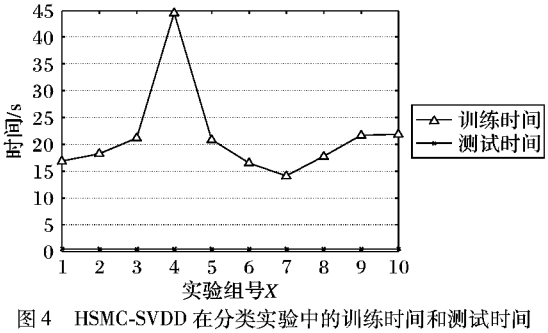

图 4　HSMC-SVDD 在分类实验中的训练时间和测试时间

SVDD 在单分类中的效果很好,在 HSMC-SVDD 中对每类样本构造一个超球体,在对各类样本进行分类时,也能达到良好的效果,在 10 组实验中分类精度最好的达到 76.692 9%。分类精度的好坏与前期提取的特征信息有很大关系,经过分析在一些分类精度好的实验组中抽取的样本数据较好,数据的相关性较好,而且属性信息具有区分性,所以分类精度比有些组实验高。但是在一些实验组中的分类精度不算太高,原因可能是提取的特征少,使得大多特征信息差别较小。如果

能进一步改进特征提取算法,将会取得更好的分类效果。

表3 各种方法的分类精度上对比

分类方法	分类精度/%
GLCM + ID3 和 KNN 的组合分类器[9]	47.6000
PCA + 基于关联规则的分类器[10]	69.2700
BP 神经网络分类器[10]	51.5100
LVQ 神经网络分类器[10]	63.6300
本文方法	76.6929

影响分类精度的因素比较多,如特征提取是实验的一个非常重要的阶段,选取好的特征也是高分类精度的一个前提,特征优化也可以作为提高分类精度的一个有效步骤。特征提取在近些年研究的也比较多,如 Xu 等[13]提到使用核方法进行癌症分类特征提取、轮廓特征提取和基于兴趣点特征提取等。

经过实验测试,在多类别中 HSMC-SVDD 对每一类样本构造一个超球体,与其他多分类器相比在计算代价上降低了很多,使训练速度有了明显的提高。

4 结语

本文把 HSOC-SVDD 分类器直接扩展到 HSMC-SVDD 分类器,在多类别中 HSMC-SVDD 对每一类样本构造一个超球体;而且在 HSMC-SVDD 中还引入了核主成分分析法,将数据集中不确定的、冗余的信息剔除,然后将处理后的数据集作为分类器的输入,因此在算法计算代价上明显降低。如果能进一步改进特征提取算法,将会取得更好的分类效果。但是理论分析和仿真实验表明,HSMC-SVDD 算法简单,在许多领域的多分类问题中,HSMC-SVDD 也可以作为一个新的思路加以应用。

参考文献:

[1] NISHIKAWA R M. Current status and future directions of computer-aided diagnosis in mammography [J]. Computerized Medical Imaging and Graphics, 2007, 31(4/5): 224 - 235.

[2] 张超,蒋宏传. 舒怡乳腺诊断仪在乳腺癌诊断中的应用[J]. 中华肿瘤防治杂志,2010,17(19):1600 - 1604.

[3] WANG L P. Support vector machine: theory and application [M]. Berlin: Springer-Verlag, 2005: 1 - 66.

[4] TAX D M J, DUIN R P W. Support vector data description [J]. Machine Learning, 2004, 54(1): 45 - 66.

[5] 徐图,罗瑜,何大可. 超球体单类支持向量机的 SMO 训练算法[J]. 计算机科学,2008,35(6):178 - 180.

[6] LE T, TRAN D, MA W, *et al*. A theoretical framework for multi-sphere support vector data description [C]// Proceedings of the 17th International Conference on Neural Information Processing: Models and Applications. Berlin: Springer, 2010: 132 - 142.

[7] LAUER F, GUERMEUR Y. MSVMpack: a multi-class support vector machine package [J]. Journal of Machine Learning Research 2011, 12: 2293 - 2296.

[8] WEI L Y, YANG Y Y, NISHIKAWA R M, *et al*. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications [J]. IEEE Transactions on Medical Imaging, 2005, 24(3): 371 - 380.

[9] OLIVER A, FREIXENET J, ZWIGGELAAR R. Automatic classification of breast density [C]// Proceedings of the 2005 IEEE International Conference on Image Processing. Washington, DC: IEEE Computer Society, 2005: 1258 - 1261.

[10] SWINIARSKI R, LIM H K. Independent component analysis, principal component analysis and rough sets in hybrid mammogram classification [C]// Proceedings of the 2006 International Conference on Image Processing. Washington, DC: IEEE Computer Society, 2006: 1121 - 1126.

[11] The mammography image analysis society [EB/OL]. [2013-03-09]. http://www.wiau.man.ac.uk/services/MIAS/MIAS web.html.

[12] SCHÖLKOPF B, SMOLA A, MÜLLER K-R. Kernel principal component analysis [M]// Advances in Kernel Methods. Cambridge: MIT Press, 1999: 327 - 352.

[13] XU Y, ZHANG D, YANG J, *et al*. Evaluate dissimilarity of samples in feature space for improving KPCA [J]. International Journal of Information Technology and Decision Making, 2011, 10(3): 479 - 495.

(上接第3299页)

意在每次迭代后对粒子位置进行判别,避免给模型参数带来大的偏差,同时,需要给定待估参数一个恰当的取值范围,否则难以快速收敛至最优解。通过改进引力搜索算法使得在任意大范围内能够快速搜索到发酵模型最优参数是今后的工作重点。

参考文献:

[1] 史仲平,潘丰. 发酵过程解析、控制与检测技术[M]. 北京:化学工业出版社,2005:144 - 153.

[2] 王景杨,范明哲. 基于最小二乘参数辨识的非线性机理模型研究[J]. 沈阳理工大学学报,2008,27(3):48 - 51.

[3] 薛尧予,王建林,于涛,等. 基于改进 PSO 算法的发酵过程模型参数估计[J]. 仪器仪表学报,2010,31(1):178 - 182.

[4] NIU B, LI L. A novel PSO-DE-based hybrid algorithm for global optimization [C]// Proceedings of the 4th International Conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence. Berlin: Springer-Verlag, 2008: 156 - 163.

[5] 王东阳,王健,陈宁. 谷氨酸发酵动力学参数估计[J]. 生物技术通讯,2005,16(4):407 - 408.

[6] RASHEDI E, NEZAMABADI-POUR H, SARYAZDI S. GSA: a gravitational search algorithm [J]. Information Sciences, 2009, 179(13): 2232 - 2248.

[7] ABBAS B, NEZAMABADI-POUR H, BAHROLOLOUM H, *et al*. A prototype classifier based on gravitational search algorithm [J]. Applied Soft Computing, 2012, 12(2): 819 - 825.

[8] BHATTACHARYA A, ROY P K. Solution of multi-objective optimal power flow using gravitational search algorithm [J]. IET Generation Transmission & Distribution, 2012, 6(8): 751 - 763.

[9] 叶勤. 发酵过程原理[M]. 北京:化学工业出版社,2005:4 - 7.

[10] 耿俊. 青霉素发酵过程的模型化研究[D]. 上海:上海交通大学,2009.

[11] 徐遥,王士同. 引力搜索算法的改进[J]. 计算机工程与应用,2011,47(35):188 - 192.

[12] RASHEDI E, NEZAMABADI-POUR H, SARYAZDI S. BGSA: binary gravitational search algorithm [J]. Natural Computing, 2010, 9(3): 727 - 745.

[13] BIROL G, ÜNDEY C, CINAR A. A modular simulation package for fed-batch fermentation: penicillin production [J]. Computers and Chemical Engineering, 2002, 26(11): 1553 - 1565.