控制与决策
Control and Decision

基于随机投影的快速凸包分类器

顾晓清， 张聪， 倪彤光

引用本文:
顾晓清, 张聪, 倪彤光. 基于随机投影的快速凸包分类器[J]. 控制与决策, 2020, 35(5): 1151−1158.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.1266

文章编号: 1001-0920(2020)05-1151-08 DOI: 10.13195/j.kzyjc.2018.1266

基于随机投影的快速凸包分类器

顾晓清, 张 聪, 倪彤光†

(常州大学 信息科学与工程学院，江苏 常州 213164)

摘 要: 传统的基于核函数的分类方法中核矩阵运算复杂度较高,无法满足大规模数据分类的要求. 针对这一问题,提出基于随机投影的快速凸包分类器(FCHC-RP). 首先,使用随机投影的方法将样本投影到多个二维子空间,并将子空间数据映射到特征空间;其次,根据数据分布的几何特征得到凸包候选集;再次,基于凸包的定义计算出特征空间中的凸包向量;最后,使用与凸包向量对应的原始样本及其权值训练支持向量机. 此外,FCHC-RP还适用于不平衡数据的分类问题,根据两类样本的不平衡程度选择不同的参数,可以得到规模相当的两类样本的凸包集,实现训练数据的类别平衡. 理论分析和实验结果验证了FCHC-RP在分类性能和训练时间上的优势.

关键词: 大规模数据；凸包；随机投影；核方法；分类；快速

中图分类号: TP273 **文献标志码:** A

Fast convex hull classifier based on random projection

GU Xiao-qing, ZHANG Cong, NI Tong-guang†

(School of Information Science and Technology，Changzhou University，Changzhou 213164，China)

Abstract: Due to the high computational complexity of kernel matrixes, traditional kernel-based methods can not satisfy the requirement of large-scale data classification. To solve this problem, a fast convex hull classifier based on random projection (FCHC-RP) is proposed. In the FCHC-RP, the samples in the original space are projected into several two-dimensional subspaces, and then the data in subspaces are mapped into the kernel space. Then, the convex hull candidate set is computed according to the geometric characteristics of data distribution in the kernel space. Based on the definition of the convex hull, the convex hull vectors in the kernel space are computed. Finally, the support vector machine is trained by the convex hull vectors and their weights. In addition, the FCHC-RP is also suitable for imbalanced classification problems. The FCHC-RP adopts classifier parameters according to the degree of class imbalance between two classes, so that the size of convex hull sets belonging to two class samples is similar. Thus, the training data in two classes are comparative. Theoretical analysis and experimental results verify the advantages of the FCHC-RP in classification performance and training time.

Keywords: large-scale data；convex hull；random projection；kernel method；classification；fast

0 引 言

基于核函数的方法在机器学习和模式识别领域取得了广泛应用,如文本识别、图像处理、基因数据识别、人脸识别以及语音识别等领域[1-5]. 基于核函数方法的主要思想是将数据通过特征映射(非线性变换)转换到希尔伯特空间(特征空间),使得原空间的线性不可分的问题转变成特征空间的线性可分的问题,这些方法具有较好的分类性能. 但是,核矩阵的存储和计算使得基于核函数分类方法的时间复杂度至少为$O(N^2)$[6],其中N是训练集样本数. 显然,随着数据量的增加,训练时间也会大量增加,因此,这些分类方法不适合大规模数据的分类问题.

以支持向量机(support vector machine, SVM)[7-8]为代表的核分类方法在构建分类面时旨在寻找不同类别样本的最优分类超平面,使得其两侧的样本点到分类面的最小距离最大化. 目前,越来越多的研究者从考虑样本分布的几何结构入手,筛选出能影响分类面的核心样本以减少训练集的规模,提高分类器的训练效率. 如:核心集向量机(generalized core vector machines, GCVM)及其变种[9-10]寻找一个最小体积超球,使得数据最大可能或全部包含在内;快速核密度估计(fast density estimator, FastKDE)[11]从

收稿日期: 2018-09-16；**修回日期:** 2018-11-27.

基金项目: 国家自然科学基金项目(61806026, 61572085);江苏省自然科学基金项目(BK20160187, BK20180956);江苏省教育科学“十三五”规划2018年度课题项目(B-a/2018/01/41).

†通讯作者. E-mail: hbxtntg-12@163.com.

理论上建立了核密度估计与二次规划问题间的联系,并使用简单采样的方法建立训练集. 但GCVM和FastKDE只适用于平方型hinge损失函数的SVM. 随机逼近凸包(randomized approximation convex hull, ApproxHull)[12]使用遗传算法求得样本集在线性空间的凸包集并用于分类问题,但该方法无法扩展至特征空间. 快速凸包方法(fast convex-hull vector machine, CHVM)[13]使用高斯核和可加性核得到特征空间的凸包集并应用于大规模ncRNA数据的分类,但实际应用时需将数据作降维处理. 凸包顶点选择(convex hull vertices selection, CHVS)[14]通过计算样本与线性子空间的距离选择凸包,但该方法的时间复杂度与样本维数的4次方成正比,随着样本维数的增加,CHVM的时间复杂度会急剧增加.

任意点集的凸包是指能包含点集中所有样本的最小凸多边形. 因此,SVM最优分类面可等价为寻找最接近两类样本凸包且具有最大距离的超平面. 基于这一思想,本文提出一种基于随机投影的快速凸包分类器(fast convex hull classifier based on random projection, FCHC-RP). 首先,FCHC-RP通过随机投影策略将样本投影到多个二维子空间,并将二维子空间数据映射至特征空间;其次,FCHC-RP在特征空间求得凸包候选集;再次,在凸包候选集的基础上进一步得到凸包集;最后,FCHC-RP将凸包向量还原至原始样本,连同其权值一起用于分类器的训练. FCHC-RP能有效减少训练集的规模,能够处理大规模数据的分类问题. 另外,现实中的数据还常常面临数据不平衡的问题. 对此,本文在不同类别的样本中调整算法参数,得到规模相当的凸包集,将FCHC-RP拓展至不平衡数据的分类问题. FCHC-RP的优势在于,能充分利用样本的几何结构,分类效果几乎不受训练样本减少的影响,同时分类器的训练时间有效减少.

1 相关工作

1.1 凸 包

定义1(凸包)[15] 设$X \subset R^d, X = \{x_1, x_2, \ldots, x_N\}$由$N$个点组成,样本集$X$的凸包为

$$\mathrm{CH}(X) = \Big\{ \sum_{x_i \in X} \alpha_i x_i \Big| \sum_{i=1}^{N} \alpha_i = 1,\ 0 \leqslant \alpha_i \leqslant 1 \Big\}. \quad (1)$$

样本集X的凸包是指能包含X中所有点的最小凸集. 式(1)给出了线性空间中凸包的定义,如果使用特征映射$\phi : x \to \phi(x)$将样本x映射到特征空间,特征空间中X的映像表示为$\phi(X) = \{\phi(x_1), \phi(x_2), \ldots, \phi(x_N)\}$,则样本集$X$在特征空间中的凸包可定义为

$$\mathrm{KerCH}(X) = \Big\{ \sum_{x_i \in X} \alpha_i \phi(x_i) \Big| \sum_{i=1}^{k} \alpha_i = 1,\ 0 \leqslant \alpha_i \leqslant 1 \Big\}. \quad (2)$$

1.2 随机投影

随机投影(random projections, RP)[16]使用随机矩阵R将向量投影到一个低维空间,随机矩阵独立于训练数据,无需按照某种准则通过训练数据产生,能够快速有效地解决高维数据的降维问题. 定义$R^d \to R^p$是随机投影映射,对于给定的数据集X,使用随机矩阵R将d维向量投影到一个p维空间($p \ll d$),得到低维向量集$\{z_1, z_2, \ldots, z_N\}$,其中

$$z_i = Rx_i. \quad (3)$$

R是标准正交矩阵,各行都是单位化的零均值正态变量,满足独立同分布. 随机投影矩阵能保持样本间成对的距离,与SVM同属于距离学习方法. 文献[16]从理论上证明了基于随机投影的SVM可以得到与原问题相近的分类误差,投影后的数据可以保持特征空间的几何性质.

2 基于随机投影的快速凸包分类器

2.1 基于随机投影的凸包学习方法

定理1[17] 假设X_A是X的凸包集(A是X中的样本标号),将X投影至任意低维子空间,得到低维子空间的凸包集$Y_{\bar{A}}$($\bar{A}$是投影空间中的样本标号),则得到$\overline{A} \subseteq A$.

根据定理1,样本在投影低维空间时仍能保持高维空间下的几何结构.如果样本是投影低维空间的凸包向量,则它在原空间中一定也是凸包向量. 当高维空间的凸包难以计算时,将数据多次投影到低维子空间,低维子空间下的凸包集组合可近似等价于高维空间下的凸包集.基于这一想法,本文提出随机投影策略下凸包学习方法. 本算法由以下几个阶段组成:

1) 随机投影阶段. 使用高斯随机投影矩阵将数据集X投影至L个二维子空间,得到子空间数据$P_j(X)(j = 1, 2, \ldots, L)$,然后使用特征函数将各子空间的数据$P_j(X)$映射至特征空间,得到特征空间数据$\phi(P_j(X))$. 以特征空间的原点为中心将$\phi(P_j(X))$划分为$k$组对称等分区域,第$i$组对称区域分别表示为$s^+_{j,i}$和$s^-_{j,i}(i = 0, 1, \ldots, k-1)$,即

$$s^+_{j,i} = \{\phi(P_j(X)) : \mathrm{atan}(P_j(X)) \in [\alpha i, \alpha(i+1))\},$$
$$s^-_{j,i} = \{\phi(P_j(X)) : \mathrm{atan}(P_j(X)) \in [\pi + \alpha i, \pi + \alpha(i+1))\}, \quad (4)$$

其中$\alpha = \pi/k$.

2) 凸包候选集计算阶段. 定义对称区域$s_{j,i}^{+}$和$s_{j,i}^{-}$的单位向量keru_i^{+}和keru_i^{-},即

$$\mathrm{keru}_i^{+}=\phi\Big(\cos\Big(\alpha i+\frac{\alpha}{2}\Big),\ \sin\Big(\alpha i+\frac{\alpha}{2}\Big)\Big),$$
$$\mathrm{keru}_i^{-}=\phi\Big(-\cos\Big(\alpha i+\frac{\alpha}{2}\Big),\ -\sin\Big(\alpha i+\frac{\alpha}{2}\Big)\Big). \tag{5}$$

在第i组对称区域$s_{j,i}^{+}$和$s_{j,i}^{-}$中,计算$\phi(P_j(X))$与单位向量keru_i^{+}和keru_i^{-}内积的最大值$m_{j,i}^{+}$和$m_{j,i}^{-}$,有

$$m_{j,i}^{+}=\max_{\phi(x)\in s_{j,i}^{+}}(\mathrm{keru}_i^{+\mathrm{T}}\phi(x)),$$
$$m_{j,i}^{-}=\max_{\phi(x)\in s_{j,i}^{-}}(\mathrm{keru}_i^{-\mathrm{T}}\phi(x)). \tag{6}$$

在区域$s_{j,i}^{+}$中求得与keru_i^{+}内积值为$m_{j,i}^{+}$的向量$M_{j,i}^{+}$,在区域$s_{j,i}^{-}$中求得与keru_i^{-}内积值为$m_{j,i}^{-}$的向量$M_{j,i}^{-}$,有

$$M_{j,i}^{+}=\{\phi(x)\in s_{j,i}^{+}:\mathrm{keru}_i^{+\mathrm{T}}\phi(x)=m_{j,i}^{+}\},$$
$$M_{j,i}^{-}=\{\phi(x)\in s_{j,i}^{-}:\mathrm{keru}_i^{-\mathrm{T}}\phi(x)=m_{j,i}^{-}\}. \tag{7}$$

在区域$s_{j,i}^{+}$和$s_{j,i}^{-}$中,求得凸包候选向量$v_{j,i}^{+}$和$v_{j,i}^{-}$,有

$$v_{j,i}^{+}=\max(\max_{\phi(x)\in(s_{j,i}^{-}-M_{j,i}^{-})}(\mathrm{keru}_i^{+\mathrm{T}}\phi(x)),M_{j,i}^{+}),$$
$$v_{j,i}^{-}=\max(\max_{\phi(x)\in(s_{j,i}^{+}-M_{j,i}^{+})}(\mathrm{keru}_i^{-\mathrm{T}}\phi(x)),M_{j,i}^{-}). \tag{8}$$

值得说明的是: 如果特征空间的原点在$\phi(P_j(X))$几何结构的内部,则凸包候选向量$v_{j,i}^{+}$和$v_{j,i}^{-}$分别等于向量$M_{j,i}^{+}$和$M_{j,i}^{-}$;但如果特征空间的原点不在$\phi(P_j(X))$几何结构的内部,则由式(8)可以计算得到$\phi(P_j(X))$的全部凸包候选向量. 然后,使用每一组对称区域中的凸包候选向量$v_{j,i}^{+}$和$v_{j,i}^{-}$构建凸包候选集V,即

$$V=\bigcup_{j=1}^{L}\{\{v_{j,i}^{+}:\|v_{j,i}^{+}\|\neq\infty\}\bigcup\{v_{j,i}^{-}:\|v_{j,i}^{-}\|\neq\infty\}\}. \tag{9}$$

最后,将凸包候选集V中的特征向量还原至原始数据中的d维样本,并统计各样本出现的频次.

3) 凸包集构建阶段. 根据凸包的定义,设V^*是V的任意子集,即$V^*\subseteq V$. 对于V中的任意样本v_i,建立函数$f(v_i,V^*)$,即

$$f(v_i,V^*)=\min_{\mu}\Big\|v_i-\sum_{v_t\in V^*}\mu_{i,t}v_t\Big\|^2;$$
$$\text{s.t. } 0\leqslant\mu_{i,t}\leqslant 1,\ \sum_{v_t\in V^*}\mu_{i,t}=1. \tag{10}$$

如果满足$\max\limits_{v_i\in V}f(v_i,V^*)$小于阈值$\varepsilon$,则定义$V^*$为$V$在特征空间中的凸包集. 在这种情况下,$V$中的任意样本$v_i$都可以写成与$V^*$中向量的线性组合关系,即

$$v_i=\sum_{v_t\in V^*}\gamma_{i,t}v_t+\delta_i. \tag{11}$$

其中: $\|\delta_i\|^2\leqslant\varepsilon$;且

$$\gamma_{i,t}=\begin{cases}\mu_{i,t},\ v_t\in V^*,\ v_i\in V\text{ and }v_i\notin V^*;\\ 0,\ \text{otherwise.}\end{cases}$$

需要指出的是,计算凸包集V^*时,根据凸包候选集V中各样本出现的频次,设定出现频次最高的样本为凸包集V^*的初始值. 然后,按照各样本频次的降序依次将各样本代入式(10). 如果$\max\limits_{v_i\in V}f(v_i,V^*)>\varepsilon$,说明$v_i$不能用当前凸包集$V^*$线性表示,则将$v_i$加入到凸包集$V^*$中,更新$V^*$得到$V^*=V^*\bigcup v_i$. 对式(10)进行求解时,展开各项并忽略常数项,式(10)可以写成如下形式:

$$\min_{\mu}2{v_i}^{\mathrm{T}}V^*\mu+\mu^{\mathrm{T}}{V^*}^{\mathrm{T}}V^*\mu;$$
$$\text{s.t. }\sum_{v_t\in V^*}\mu_{i,t}=1,\ 0\leqslant\mu_{i,t}\leqslant 1. \tag{12}$$

式(12)是一个标准的凸二次规划问题,本文使用序贯最小优化(sequential minimal optimization, SMO)[18]算法来求解.

设β_t是凸包集V^*中每个凸包向量x_t的权值,用来体现每个凸包向量在凸包集中的“重要程度”. β_t的值可以通过下式计算得到:

$$\beta_t=\sum_{i=1}^{N}\mu_{i,t}. \tag{13}$$

2.2　基于随机投影的快速凸包分类器

不失一般性,本文讨论二元分类问题. 假设数据集X在正负两类样本X^+和X^-上分别得到特征空间凸包集KerCH(X^+)和KerCH(X^-). KerCH(X^+)和KerCH(X^-)对应的原始样本连同其权值一起参与SVM分类器的训练. 因此,基于随机投影的快速凸包分类器FCHC-RP的无约束原始问题表示为

$$\min_{w,b}\frac{1}{2}\|w\|^2+\frac{C}{N}\sum_{t=1}^{|V^*|}\beta_t l(w,b,\phi(x_t)). \tag{14}$$

其中: $l(w,b,\phi(x_t))$为损失函数,$|V^*|$为凸包集V^*的容量. FCHC-RP能使用不同类型的损失函数,鉴于hinge损失函数在SVM中应用最广泛,将hinge损失函数代入式(14),得到

$$\min_{w,b}\frac{1}{2}\|w\|^2+\frac{C}{N}\sum_{t=1}^{|V^*|}\beta_t\xi_t;$$
$$\text{s.t. }y_i(w^{\mathrm{T}}\phi(x_i)+b)\geqslant 1-\xi_i,$$
$$\xi_i\geqslant 0,\ i=1,\ldots,|V^*|. \tag{15}$$

引入拉格朗日乘子α,式(15)可写成如下的对偶形式:

$$\max_{\alpha}\sum_{i=1}^{|V^*|}\alpha_i-\frac{1}{2}\sum_{i=1}^{|V^*|}\sum_{j=1}^{|V^*|}\alpha_i\alpha_j y_i y_j\phi(x_i)^{\mathrm{T}}\phi(x_j);$$

$$\text{s.t.}\ \sum_{i=1}^{|V^*|}y_i\alpha_i=0,\ 0\leqslant\alpha_i\leqslant\frac{C}{N}\beta_t,\ i=1,\ldots,|V^*|. \tag{16}$$

与传统SVM方法类似,求解出超平面法向量w和偏移量b,可得FCHC-RP的分类决策函数

$$f(x)=w^{\mathrm{T}}\phi(x)+b. \tag{17}$$

2.3 FCHC-RP算法描述

根据以上分析,本节给出FCHC-RP算法的描述.

算法1 FCHC-RP算法.

step 1: 使用式(3)将X投影至L个二维空间;

for $j=1$ to L

step 2: 使用式(4)建立k个等分区域对$\{s_{j,i}^{+},s_{j,i}^{-}\}$;

step 3: 使用式(5)定义$s_{j,i}^{+}$和$s_{j,i}^{-}$的单位向量keru_i^{+}和keru_i^{-};

step 4: 使用式(6)~(8)计算$s_{j,i}^{+}$和$s_{j,i}^{-}$中的凸包候选点;

endfor

step 5: 使用式(9)~(12)建立凸包候选集V;

step 6: 还原V至原始空间,统计样本出现的频次并删除重复样本;

step 7: 使用式(12)计算凸包集V^*;

step 8: 使用式(13)计算权重β_t;

step 9: 将V^*对应的原始样本和β代入式(15)和(16)求解参数(w,b);

step 10: 将(w,b)代入式(17), 得到决策函数$f(x)$.

3 问题讨论

3.1 FCHC-RP精度误差分析

定义L1-SVM的无约束最优化问题$F_1(w,b)$为

$$\min_{w,b}F_1(w,b)=\frac{1}{2}\|w\|^2+\frac{C}{N}\sum_{t=1}^{N}l(w,b,\phi(x_t)),$$

FCHC-RP的无约束最优化问题$F_2(w,b)$为

$$\min_{w,b}F_2(w,b)=\frac{1}{2}\|w\|^2+\frac{C}{N}\sum_{t=1}^{|V^*|}\beta_t l(w,b,\phi(x_t)),$$

其中

$$l(w,b,\phi(x_t))=\max\{0,1-y_t(w^{\mathrm{T}}\phi(x_t)+b)\}.$$

定理2 设L1-SVM的最优解是(w_1^*,b_1^*),FCHC-RP的最优解是(w_2^*,b_2^*),则

$$F_1(w_1^*,b_1^*)-F_2(w_2^*,b_2^*)\leqslant C\sqrt{C\varepsilon}.$$

证明 定义$F_3(w,b)$为

$$\min_{w,b}F_3(w,b)=\frac{1}{2}\|w\|^2+\frac{C}{N}\sum_{t=1}^{N}l(w,b,u_t),$$

其中$u_i=\sum_{t=1}^{|V^*|}\mu_{i,t}\phi(x_t),\phi(x_t)\in V^*$,则

$$L_3(w,b,V^*)=$$

$$\frac{C}{N}\sum_{i=1}^{N}\max\Big[0,\Big\{1-y_i\Big(w^{\mathrm{T}}\sum_{t=1}^{|V^*|}\mu_{i,t}\phi(x_t)+b\Big)\Big\}\Big]\leqslant$$

$$\frac{C}{N}\sum_{i=1}^{N}\sum_{t=1}^{|V^*|}\max[0,\mu_{i,t}\{1-y_t(w^{\mathrm{T}}\phi(x_t)+b)\}]=$$

$$\frac{C}{N}\sum_{t=1}^{|V^*|}\max[0,1-y_t(w^{\mathrm{T}}\phi(x_t)+b)]\sum_{i=1}^{N}\mu_{i,t}=$$

$$L_2(w,b,V^*).$$

等式的两端加上$(1/2)\|w\|^2$可得$F_3(w,b)\leqslant F_2(w,b)$. 由式(10)同理可得

$$-\frac{C}{N}\sum_{i=1}^{N}\max\{0,y_i w^{\mathrm{T}}\delta_i\}\leqslant$$

$$F_1(w,b)-F_3(w,b)\frac{C}{N}\sum_{i=1}^{N}\max\{0,-y_i w^{\mathrm{T}}\delta_i\}.$$

由此可得

$$F_1(w_1^*,\ b_1^*)-F_3(w_2^*,b_2^*)\leqslant$$

$$\frac{C}{N}\sum_{i=1}^{N}\max\{0,-y_i {w_2^*}^{\mathrm{T}}\delta_i\}\leqslant\frac{C}{N}\sum_{i=1}^{N}\|w_2^*\|\|\delta_i\|.$$

由文献[19]可得$\|w_2^*\|\leqslant\sqrt{C}$,因此

$$F_1(w_1^*,\ b_1^*)-F_3(w_2^*,\ b_2^*)\leqslant\frac{C}{N}\sum_{i=1}^{N}\sqrt{C\varepsilon}=C\sqrt{C\varepsilon}.\ \square$$

定理2证明了FCHC-RP与传统使用hinge损失函数的L1-SVM在分类精度上是相当的.

3.2 FCHC-RP精度误差分析

FCHC-RP方法可以分为随机投影、计算凸包候选集、计算凸包集和建立分类器4个阶段. 给定规模为N的d维数据集,FCHC-RP前2个阶段的时间复杂度分别是$O(L(2dN))$和$O((N+2k)/2)$. 其中:L是执行随机投影的次数,k是等分区域对的个数. 第3阶段以渐近方式使用SMO算法求解式(10),时间复杂度是$O\Big(\sum_{i=1}^{|V|}n_i^2\Big)$,其中$|V|$和$n_i$分别是凸包候选集和当前凸包集的容量. 第4阶段仍然使用SMO算法训练分类器,时间复杂度为$O(|V^*|^2)$,其中$|V^*|$是FCHC-RP在第3阶段结束时得到的凸包集的容量. 因此,FCHC-RP分类器的时间复杂度是

$$O\Big(L(2dN+(N+2k)/2)+\sum_{i=1}^{|V|}n_i^2+|V^*|^2\Big).$$

需要说明的是:第1和第2阶段在并行环境下执行时间复杂度会有效降低;而凸包集V^*的规模远小

于样本总体规模N. 因此,FCHC-RP时间复杂度远小于传统SVM高达$O(N^3)$的时间复杂度.

3.3 FCHC-RP不平衡数据的分类

一般认为,两分类样本规模的比例低于1:2时,数据集具有不平衡特征. SVM在处理不平衡数据的分类问题时,分类超平面易向少数类样本偏移,导致少数类样本的误判率较高. 对此,FCHC-RP可通过等分区域参数k或凸包阈值ε来调节凸包集的规模. 当k值较小时,等分区域数量就少,FCHC-RP最终得到的凸包集的规模也较小. 当ε值较小时,凸包候选集中不满足$\max\limits_{v_i\in V} f(v_i,V^*)\leqslant\varepsilon$的向量数目就多,最终得到的凸包集的规模就越大;反之,当ε值较大时,得到的凸包集的规模较小. 实际应用中,在不平衡比例不高的大规模数据分类问题中(两类比例小于1:30),可在k或ε设定的搜索范围内,在少数类样本中使用较小的k或ε值,同时在多数类样本中使用较大的k或ε值,使得两类样本得到的凸包集规模相当. 当大规模数据不平衡比例较高时(正负类比例大于1:30),可仅在多数类样本中计算凸包集,保留全部的少数类样本,此时式(14)中少数类样本对应的β值均为1.

4 实验分析

4.1 实验设置

为验证本文所提出FCHC-RP分类器的有效性,本节将在10个不同类型的数据集[20](如表1所示)上设计多个学习任务: 1) FCHC-RP与基线分类器L1-SVM[21]以及几种适用于大规模数据分类器进行比较,包括CVM[22]、CHVM[14]和FastKDE[11]; 2) FCHC-RP与几种适用于不平衡数据分类方法进行比较,包括IM-FastKDE[11]、CS-SVM[23]和MLWSVM[24].

实验参数设置如下: FCHC-RP等分区域参数k的取值范围是$\{4,6,9,18\}$,凸包阈值ε的取值范围是$\{10^{-3},10^{-2},10^{-1}\}$. 随机投影矩阵有多种形式,本文使用高斯随机投影矩阵. 根据大量的实验,当样本的维数小于20维时,随机投影迭代次数等于样本维数×2;当样本的维数大于20维时,随机投影迭代次数等于样本维数×1.2. FastKDE按15%训练集的样本数进行采样;IM-FastKDE基于FastKDE在两类样本上按不同比例进行采样,使得两类样本比例为1:1. CVM的逼近精度是10^{-5};L1-SVM使用libsvm工具箱[21]实现;CS-SVM和MLWSVM中两类样本的正则化参数比例与两类样本的容量成反比. 所有SVM算法中的核函数均采用高斯核,核参数和正则化参数的取值范围均为$\{10^{-3},10^{-2},\ldots,10^{3}\}$. 所有参数均采取5重交叉验证法来选取最优值.

表1　10个真实数据集的基本信息

数据集	规模	维数	两类样本比例
a9a	35 000	123	1:2
buzz	140 707	77	4:5
dosvsnormal	250 000	41	1:1
forest	581 012	54	1:1
ijcnn1	33 895	22	2:3
kdd99	300 000	41	1:1
letter26	20 000	16	1:1
multi-ncRNA	448 601	8	1:2
normalvsprb	123 000	41	1:2
nursery	12 960	8	2:3

大规模数据分类方法采用分类精度、G-mean[25]和训练时间3种评价指标;不平衡数据分类方法采用G-mean、F-measure[25]和训练时间3种评价指标. 其中: 分类精度用来评价数据集整体分类性能,G-mean用来评价在保持正负类分类精度平衡下最大化两类的精度,F-measure用来评价分类器对少数类的分类性能. 实验在2.53-GHz quad-core CPU, 8-GB RAM, Windows 7系统下执行,所有分类器均在Matlab 2016b环境下实现.

4.2 大规模数据集的分类

本节在10个数据集上评价FCHC-RP方法的性能: 表2比较了所有方法的分类精度及其方差(括号内数值为方差),表3比较了所有方法的G-mean及其方差,表4比较了所有方法的训练时间及其方差.

表2　不同方法在10个真实数据集上的分类精度及其方差的比较

数据集	L1-SVM	CVM	CHVS	FastKDE	FCHC-RP
a9a	82.42(0.28)	80.78(0.32)	81.58(0.30)	80.22(0.36)	82.42(0.28)
buzz	–	92.39(0.33)	93.56(0.32)	92.23(0.35)	94.53(0.32)
dosvsnormal	–	94.23(0.17)	96.25(0.17)	94.39(0.37)	96.60(0.15)
forest	–	93.97(0.20)	97.40(0.10)	95.40(0.28)	97.88(0.10)
ijcnn1	91.02(0.38)	90.48(0.43)	90.12(0.37)	88.05(0.49)	90.74(0.36)
kdd99	–	91.00(0.33)	91.95(0.30)	90.51(0.42)	92.43(0.30)
letter26	97.60(0.10)	97.42(0.13)	97.25(0.12)	97.23(0.15)	97.64(0.11)
multi-ncRNA	–	93.49(0.28)	95.41(0.24)	92.37(0.31)	95.66(0.21)
normalvsprb	–	92.67(0.25)	92.74(0.19)	92.05(0.27)	93.14(0.19)
nursery	90.78(0.27)	90.02(0.34)	90.51(0.28)	90.00(0.39)	90.57(0.27)

表 3 不同方法在10个真实数据集上的G-mean及其方差的比较

数据集	L1-SVM	CVM	CHVS	FastKDE	FCHC-RP
a9a	82.20(0.28)	80.70(0.34)	81.36(0.32)	79.65(0.35)	82.21(0.28)
buzz	–	92.51(0.32)	93.82(0.32)	92.38(0.34)	94.61(0.31)
dosvsnormal	–	94.17(0.19)	96.09(0.19)	94.18(0.23)	96.63(0.16)
forest	–	93.90(0.16)	97.35(0.15)	95.26(0.26)	97.34(0.12)
ijcnn1	90.41(0.40)	90.20(0.48)	90.03(0.43)	87.27(0.52)	90.19(0.42)
kdd99	–	91.02(0.42)	91.37(0.40)	90.54(0.50)	92.35(0.40)
letter26	97.63(0.11)	97.49(0.12)	97.16(0.13)	97.12(0.12)	97.63(0.11)
multi-ncRNA	–	93.10(0.24)	95.38(0.27)	92.04(0.32)	95.62(0.23)
normalvsprb	–	92.61(0.23)	92.56(0.21)	92.31(0.24)	93.40(0.19)
nursery	90.54(0.30)	89.99(0.32)	90.04(0.30)	89.87(0.40)	90.46(0.31)

表 4 不同方法在10个真实数据集上的训练时间(单位: s)及其方差的比较

数据集	L1-SVM	CVM	CHVS	FastKDE	FCHC-RP
a9a	568.85(5.67)	84.75(0.80)	67.85(0.67)	11.56(0.16)	17.47(0.23)
buzz	–	2 865.85(20.45)	248.71(2.62)	497.21(5.00)	92.39(0.87)
dosvsnormal	–	12 586.48(109.17)	313.39(3.48)	215.38(3.40)	200.15(3.00)
forest	–	40 344.37(220.69)	466.60(3.16)	2 196.58(22.52)	325.98(3.33)
ijcnn1	480.45(4.55)	70.75(0.47)	14.72(0.15)	10.49(0.19)	10.48(0.12)
kdd99	–	3 300.90(40.15)	480.36(4.22)	350.78(4.18)	329.12(3.60)
letter26	198.42(3.01)	20.53(0.42)	4.97(0.07)	2.47(0.03)	3.73(0.05)
multi-ncRNA	–	3 160.18(30.63)	155.89(0.79)	750.83(8.92)	83.70(0.71)
normalvsprb	–	5 268.66(24.63)	107.54(1.92)	80.43(0.55)	78.42(0.50)
nursery	150.82(1.08)	25.48(0.22)	12.74(0.17)	9.36 (0.08)	9.59 (0.10)

1) 从表2中可以看出,由于L1-SVM使用Libsvm工具箱实现,其在multi-ncRNA、kdd99和forest等6个数据集上的训练时间超过5个小时,实验没有记录其结果.但在有记录的a9a、ijcnn1和nursery这3个数据集上L1-SVM取得了最高分类精度,这是因为所有的训练样本均参与到L1-SVM分类器的训练中. FCHC-RP则在其余的7个数据集上取得了最佳分类精度,因为FCHC-RP中的训练样本是能够表示数据集在特征空间几何结构的凸包向量. CHVS和CVM也取得了较高的分类精度. 而FastKDE的分类精度相对较低,其原因在于FastKDE使用随机采样的策略来获得训练样本,而随机采样的不确定性容易导致得到的样本不能表示数据集的几何结构.

2) 从表3中可以看出,G-mean评价指标上的结果与表2的分类精度保持了一致性. 由于10个数据集都属于类别平衡的数据集,少数类和多数类的样本规模大致相当,所有分类器在不同类别样本上的准确率均较接近,可以看到,FCHC-RP取得了令人满意的结果,在10个数据集上胜出了7次.

3) 从表4中可以看出,除a9a、nursery和letter26数据集外,FCHC-RP在7个真实数据集上的训练时间最短. FCHC-RP计算凸包候选集的时间复杂度是近似线性的,大多数数据集上获得的凸包候选集仅占训练集规模的不到10 %,随后FCHC-RP以渐近的方式计算凸包集,因此,FCHC-RP可以实现分类器的快速训练. 随着训练数据规模的增加,CVM的训练时间增加幅度较大,因为CVM的时间复杂度在理论上与训练数据的规模无关,但在大规模数据的分类问题中,CVM往往需要较长时间找到近似最小包含球. CHVS在维数较高的数据集上训练时间较长,CHVS的时间复杂度是

$$O\Big(Nd^4+\sum_{i=1}^{l}n\bar{m}_{sv}\Big).$$

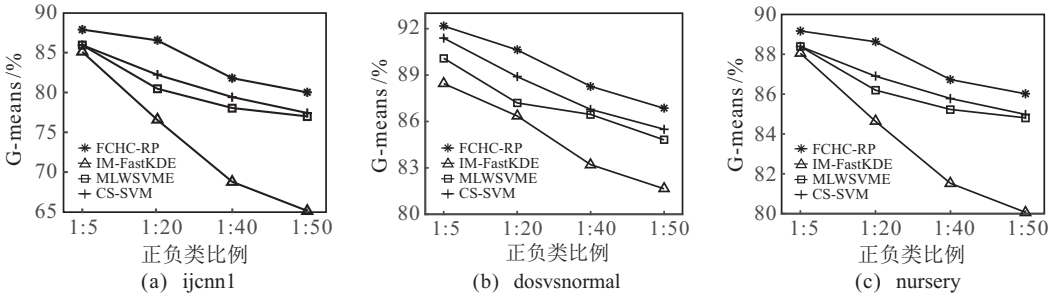
其中:d是样本的维数,n和$\bar{m}_{sv}$分别是凸包向量的个数和每次迭代中的支持向量数,l是迭代的次数. 因此,CHVS在高维大规模数据上训练时间较长. FastKDE在10万规模以下的训练集上具有一定的优势,在a9a、letter26和nursery数据集上取得了最优结果.

4.3 不平衡数据集的分类

为了更好地呈现样本的不平衡性对算法性能的影响,本节对数据集ijcnn1、dosvsnormal和Nursery进行改造. 依照不平衡数据分类问题中常用的设定方法,少数类为正类,多数类为负类. ijcnn1数据集分别随机取正类样本4 000、1 000、500和400,负类样本取20 000; dosvsnormal数据集分别取正类样本2 500、625、312和250,负类样本取125 000; Nursery数据集分别取正类样本1 555、389、195和156,负类样本取7 776. 正负类样本比例分别为1: 5、1: 20、1: 40、1: 50.

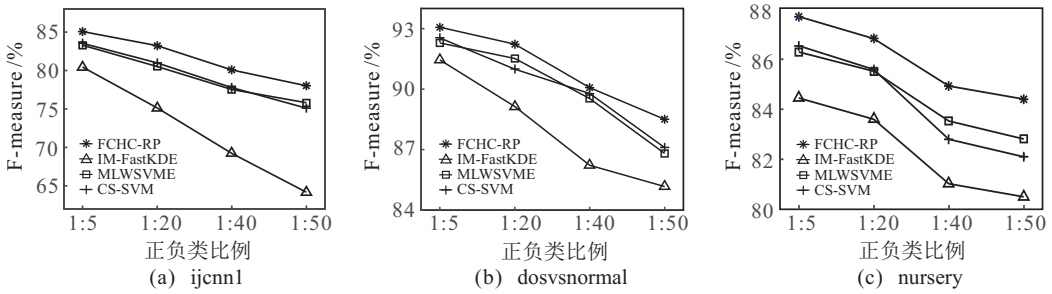
如前文所述,FCHC-RP在处理不平衡数据分类问题时,若正负类比例小于1: 30,则在正负类样本中使用不同的k来计算两类样本的凸包集;若正负类比

例达到或大于1:30,则仅在多数类样本中计算凸包集,保留全部的少数类样本. 图1比较了所有方法的G-mean及其方差,图2比较了所有方法的F-measure及其方差,图3比较了所有方法的训练时间及其方差.



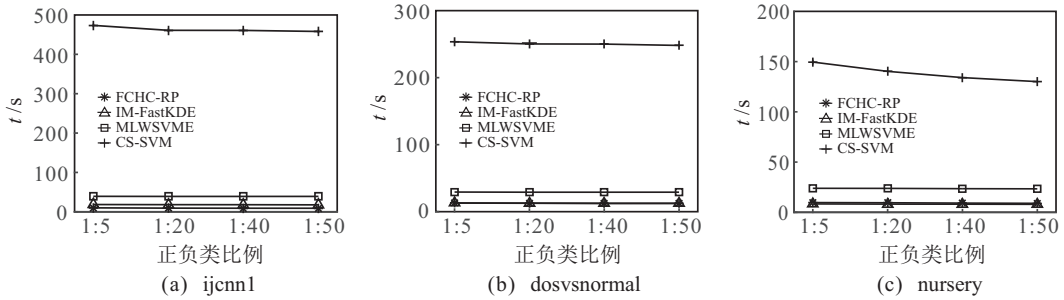
(a) ijcnn1　(b) dosvsnormal　(c) nursery

图 1　不同方法的G-mean比较



(a) ijcnn1　(b) dosvsnormal　(c) nursery

图 2　不同方法的F-measure比较



(a) ijcnn1　(b) dosvsnormal　(c) nursery

图 3　不同方法的训练时间比较

1) 从图1中可以看出,随着数据集不平衡比例的提高,IM-FastKDE、CS-SVM、MLWSVM和FCHC-RP的G-mean值呈现出不同程度的下降,这是因为数据的不平衡性是影响分类效果的因素之一,此时分类超平面向正类样本发生偏移,正类样本的分类精度下降,G-mean值下降. 但同时也注意到,由于充分考虑不同类别样本分布结构的同时保持类别比例的平衡,FCHC-RP在所有对比方法中G-mean值最优. CS-SVM和MLWSVM通过给正负类样本设置不同的分类代价来提高正类样本的分类精度,但会牺牲一部分负类样本的精度,因此,这两种方法的实验结果逊于FCHC-RP. IM-FastKDE的简单采样策略则容易造成分类器过拟合的现象.

2) 从图2中可以看出,与图1结果相似,各方法的F-measure值随着不平衡性的加剧而下降,但FCHC-RP在所有对比算法中F-measure下降的幅度最小. 由于简单采样的不稳定性,IM-FastKDE的方差在全部分类器中是最大的.

3) 由于基于SVM的分类器的训练时间与训练样本的规模成正比,训练样本的规模越大,训练分类器花费的时间就越多. 从图3中可以看出: CS-SVM使用全部样本训练分类器,训练时间最长; MLWSVM使用KNN算法找出样本的代表点,训练时间较CS-SVM短; IM-FastKDE和FCHC-RP在3个数据集上的训练时间相当.

5　结　论

FCHC-RP首先通过随机投影和核函数将原始样本映射到多个特征子空间,选择能表示样本集在特征空间分布特征的凸包候选集; 然后进一步计算出凸包向量; 最后,以凸包向量对应的原始样本和其权值训练支持向量机分类器. 理论分析和真实数据集的仿真实验表明了本文方法优良的分类性能和较高的执行效率.

另外,FCHC-RP还适用于不平衡数据的分类问

题. 应当指出,本文对能否有效解决大规模噪声数据分类问题没有进行讨论,若训练样本中包含噪声数据时,需要剔除噪声数据再计算样本的凸包集. 此外,本文的另一研究方向是将FCHC-RP拓展至在线学习方法,根据凸包的定义更新训练样本以有效控制在线学习中训练样本的规模.

参考文献(References)

[1] Ni T G, Gu X Q, Wang J, et al. Scalable transfer support vector machine with group probabilities[J]. Neurocomputing, 2018, 273(17): 570-582.

[2] Unar S, Wang X Y, Zhang C. Visual and textual information fusion using kernel method for content based image retrieval[J]. Information Fusion, 2018, 44(11): 176-187.

[3] Xu X Z, Deng J, Cummins N, et al. A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(7): 1436-1449.

[4] Yan J J, Zheng W M, Xu Q Y, et al. Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech[J]. IEEE Transactions Multimedia, 2016, 18(7): 1319-1329.

[5] Cao X C, Ren W Q, Zuo W M, et al. Scene text deblurring using text-specific multiscale dictionaries[J]. IEEE Transactions on Image Processing, 2015, 24(4): 1302-1314.

[6] 史荧中, 王士同, 王骏, 等. 基于最小包含球的非静态大数据集的快速分类算法[J]. 控制与决策, 2013, 28(7): 1065-1072.
(Shi Y Z, Wang S T, Wang J, et al. Fast classification for nonstationary large scale data sets using minimal enclosing ball[J]. Control and Decision, 2013, 28(7): 1065-1072.)

[7] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 955-974.

[8] Fan R E, Chang K W, Hsieh C. LIBLINEAR: A library for large linear classification[J]. Journal of Machine Learning Research Applied, 2018, 9(6): 1871-1874.

[9] Tsang I W, Kwok J T, Zurada J M. Generalized core vector machines[J]. IEEE Transactions on Neural Networks, 2006, 17(5): 1126-1140.

[10] Tsang I, Kwok A, Kwok J. Simpler core vector machines with enclosing balls[C]. Proceedings of the 32nd International Conference on Machine Learning. Corvallis, 2007: 911-918.

[11] Wang S T, Wang J, Chung F. Kernel density estimation, kernel methods, and fast learning in large data sets[J]. IEEE Transactions on Cybernetics, 2014, 44(1): 1-20.

[12] Zhao J, Fernandes V B, Jiao L, et al. Multiobjective optimization of classifiers by means of 3D convex-hull-based evolutionary algorithms[J]. Information Sciences, 2016, 367(3): 80-104.

[13] Gu X Q, Chung F L, Wang S T. Fast convex-hull vector machine for training on large-scale ncRNA data classification tasks[J]. Knowledge-Based Systems, 2018, 151(1): 149-164.

[14] Ding S, Nie X, Hong Q, et al. A fast algorithm of convex hull vertices selection for online classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(4): 792-806.

[15] Wang D, Qiao H, Zhang B, et al. Online support vector machine based on convex hull vertices selection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(4): 593-609.

[16] Paul S, Boutsidis C, Magdon-Ismail M, et al. Random rrojections for linear support vector machines[J]. Acm Transactions on Knowledge Discovery from Data, 2014, 8(4): 1-25.

[17] Zhou T Y, Bian W, Tao D C. Divide-and-conquer anchoring for nNear-separable nonnegative matrix factorization and completion in high dimensions[C]. Proceedings of IEEE 13th International Conference on Data Mining. Dallas: IEEE, 2013: 917-926.

[18] Takahashi N, Nishi T. Rigorous proof of termination of SMO algorithm for support vector machines[J]. IEEE Transactions on Neural Network, 2005, 16(3): 774-776.

[19] Talathi S S, Hwang D U, Spano M L, et al. Non-parametric early seizure detection in an animal model of temporal lobe epilepsy[J]. Journal of Neural Engineering, 2008, 5(1): 85-98.

[20] Bache K, Lichman M. UCI database[EB/OL]. [2018-02-28]. http://www.ics.uci.edu/.

[21] Chang C, Lin C. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[22] Tsang I W, Kwok J T, Cheung P M. Core vector machines: Fast SVM training on very large data sets[J]. Journal of Machine Learning Research, 2005, 6(12): 363-392.

[23] Shirazi H M, Vasconcelos N, Iranmehr A. Cost-sensitive support vector machines[J]. Journal of Machine Learning Research, 2012, 6(5): 387-389.

[24] Talayeh R, Oleg R, Ilya S, et al. Multilevel weighted support vector machine for classification on healthcare data with missing values[J]. Plos One, 2016, 11(5): e0155119.

[25] 顾晓清, 蒋亦樟, 王士同. 用于不平衡数据分类的0阶TSK型模糊系统[J]. 自动化学报, 2017, 43(10): 1773-1788.
(Gu X Q, Jiang Y Z, Wang S T. Zero-order TSK-type fuzzy system for imbalanced data classification[J]. Acta Automatica Sinica, 2017, 43(10): 1773-1788.)

作者简介

顾晓清(1981−), 女, 副教授, 博士, 从事机器学习的研究, E-mail: czxqgu@163.com;

张聪(1996−), 男, 硕士生, 从事模糊识别与人工智能的研究, E-mail: 1203638098@qq.com;

倪彤光(1978−), 男, 副教授, 博士, 从事智能计算与机器学习等研究, E-mail: hbxtntg-12@163.com.

(责任编辑：孙艺红)