基于构造映射的支持向量分类机

刘琼荪,杜会锋
LIU Qiong-sun,DU Hui-feng

重庆大学 数理学院,重庆 400030
College of Mathematics and Physics,Chongqing University,Chongqing 400030,China

LIU Qiong-sun,DU Hui-feng.Support vector machine classifier based on construction map.Computer Engineering and Applications,2009,45(27):130-132.

Abstract: A distance map and conditional probability map based on mahalanobis distance and copula function are constructed, the samples in high-dimensional space are transformed into new samples in two-dimensional space.A divided Support Vector Machine Classifier can be constructed by easily processing the new samples.It is simple and easy to implement.Experiments show that it has made better classification results.

Key words: map;two-dimensional;space;divided support vector machine classifier

摘 要: 构造了基于马氏距离和 Cupula 函数的距离映射和条件概率映射,将高维空间中的样本转化为二维空间中的新样本,并对新样本进行简易处理,构建了可分支持向量分类机,其特点是简单,易于实现。实验表明取得了较好的分类效果。

关键词: 映射;二维空间;可分支持向量机
DOI: 10.3778/j.issn.1002-8331.2009.27.039 **文章编号:** 1002-8331(2009)27-0130-03 **文献标识码:** A **中图分类号:** TP181

1 引言

支持向量机(SVM)是根据 Vapnik 提出的统计学习理论发展而来,它以坚实的理论基础和良好的泛化性能被广泛应用于模式识别诸多领域。为了使 SVM 适应多指标、大数据集的应用,并提高分类的精度和算法的运行速度,许多学者做了大量的工作。大致可归纳为两个方面:一是基于距离核函数的除噪和减样方法,或从训练集合中选择出最有可能成为支持向量的样本进行再训练,以此提高训练速度和分类精度[1-2]。另一方面,根据样本集构造新的核函数和选择较优的核函数[3],以保证分类精度。首先基于降维的思想,将高维空间中的样本通过马氏距离或 copula 连接函数转化为二维空间中的样本, 然后构造新的核函数,得到可分支持向量机,不仅算法易于实现,而且有效地提高了分类精度。

2 支持向量机分类的基本原理

支持向量机是基于统计学习理论的一种新型机器学习方法。主要有三个特点:(1)基于结构风险最小化原则,最大化分类间隔以得到较好的推广能力;(2)算法设计为凸二次规划;(3)引入核函数将样本映射到其他的特征空间,构建出非线性支持向量机。设 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$ 用于分类的训练集,$x_i\in R^p$ 为样本特征,$y_i\in Y=\{1,-1\}$,$i=1,\cdots,l$ 表示相应的分类结果,即正类和负类[4]。

当训练集线性可分时,SVM 的目标就是构造线性最优分类超平面 $w\cdot x+b=0$,要求将两类样本完全正确地分开,并使分类间隔最大,其优化问题如下:

$$\min_{w,b}\frac{1}{2}\|w\|^2 \tag{1}$$
$$\text{s.t. } y_i((w\cdot x_i)+b)\geqslant 1,i=1,\cdots l$$

它的对偶问题为:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i$$
$$\text{s.t. }\sum_{i=1}^{l}y_i\alpha_i=0 \tag{2}$$
$$\alpha_i\geqslant 0,i=1,\cdots,l$$

其中$(x_i\cdot x_j)$表示两个向量$\boldsymbol{x}_i,\boldsymbol{x}_j$的内积。

当训练集线性不可分时,一般做法是采用非线性映射的方法,将原始空间的样本映射到高维特征空间中,使样本在此特征空间中线性可分,即寻找非线性变换 ϕ,有映射 $\phi:x\in R^p\rightarrow\phi(x)\in R^m$,$m>p$。将式(2)转化为如下问题:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j))-\sum_{i=1}^{l}\alpha_i$$
$$\text{s.t. }\sum_{i=1}^{l}y_i\alpha_i=0 \tag{3}$$
$$\alpha_i\geqslant 0,i=1,\cdots,l$$

由于内积运算$(\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j))$是在相对的高维空间中进行,容易引起维数灾难。为此引入核函数 $K(\cdot)$,使之满足 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j))$。实际上可用原空间中的函数来实现,不必知道

作者简介: 刘琼荪(1956-),女,教授,主要研究方向:智能计算,数据挖掘和应用统计;杜会锋(1983-),男,硕士,主要研究方向:数据分析和数据挖掘。
收稿日期: 2009-02-12 **修回日期:** 2009-04-22

变换 ϕ 的具体形式。常见的核函数有高斯径向基核、多项式核、Sigmoid 核。将式(3)转化为:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jK(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i$$

$$\text{s.t.}\ \sum_{i=1}^{l}y_i\alpha_i=0 \qquad (4)$$

$$\alpha_i\geqslant 0,i=1,\cdots,l$$

经过求解对偶问题(式(4)),得最优解 $\boldsymbol{\alpha}^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_l^*)^{\mathrm{T}}$,再计算 $w^*=\sum_{i=1}^{l}y_i\alpha_i^*x_i$,选择 α^* 的正分量 α_j^*,以此计算 $b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^*K(x_i\cdot x_j)$,由此,构造出线性最优分类超平面 $w^*\cdot x+b^*=0$。由此可知,核函数类型和参数的选择对样本的正确分类尤为重要。

首先对原样本进行特征提取,引入距离映射或条件概率映射 ϕ,将高维空间的样本 $x\in R^p$ 转化为二维空间的样本 $\boldsymbol{\phi}(x)=(z_1,z_2)^{\mathrm{T}}\in R^2$;其次引入核函数 $K(\phi(x),\phi(x'))=(\phi(x)\cdot\phi(x'))$。

3　距离映射

设有两个总体(正类和负类)分别记为 G_1 和 G_2,设来自总体 $G_i(i=1,2)$ 的训练样本为 $\boldsymbol{x}_t^{(i)}=(x_{t1}^{(i)},x_{t2}^{(i)},\cdots,x_{tp}^{(i)})^{\mathrm{T}}$,$(i=1,2;t=1,2,\cdots,n_i)$,其中 n_i 是总体 G_i 的样本个数,则样本 G_i 的均值向量为 $\boldsymbol{\mu}_i$,它的估计量为:

$$\hat{\boldsymbol{\mu}}_i=\left(\frac{1}{n_i}\sum_{t=1}^{n_i}x_{t1}^{(i)},\frac{1}{n_i}\sum_{t=1}^{n_i}x_{t2}^{(i)},\cdots,\frac{1}{n_i}\sum_{t=1}^{n_i}x_{tp}^{(i)}\right)^{\mathrm{T}},(i=1,2) \qquad (5)$$

总体 $G_i(i=1,2)$ 的协方差矩阵 $\boldsymbol{\Sigma}_i$ 的估计量为:

$$\hat{\boldsymbol{\Sigma}}_i=\frac{1}{n_i-1}\sum_{t=1}^{n_i}(x_t^{(i)}-\hat{\mu}_i)(x_t^{(i)}-\hat{\mu}_i)^{\mathrm{T}},(i=1,2) \qquad (6)$$

其中 $\hat{\boldsymbol{\Sigma}}_i(i=1,2)$ 为 p 阶方阵。则任意的样本 $\boldsymbol{x}=(x_1,\cdots,x_p)^{\mathrm{T}}\in\boldsymbol{R}^p$ 到总体 G_1 的马氏距离[5]的平方为:

$$d_1^2(x)=(x-\hat{\mu}_1)^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_1^{-1}(x-\hat{\mu}_1) \qquad (7)$$

同理,样本 $x\in R^p$ 到总体 G_2 的马氏距离的平方为:

$$d_2^2(x)=(x-\hat{\mu}_2)^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_2^{-1}(x-\hat{\mu}_2) \qquad (8)$$

定义距离映射 $\phi:X\subset R^p\to R^2$,即 $\phi:x\to(d_1^2(x),d_2^2(x))^{\mathrm{T}}$。

显然映射 ϕ 就是将 p 维空间的样本点 x 通过马氏距离映射到二维空间上的点 $(d_1^2(x),d_2^2(x))^{\mathrm{T}}$ 与之对应。

由此定义核函数:

$$K(x,x')=(\phi(x)\cdot\phi(x'))=d_1^2(x)d_1^2(x')+d_2^2(x)d_2^2(x'),x,x'\in R^p \qquad (9)$$

4　条件概率映射

基于贝叶斯分类器的思想,对样本进行特征提取。设有两个类别 G_1,G_2,给定一个具体的样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$,其中 x_i 表示第 i 个属性指标。判断样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$ 属于类别 $G_k(k=1,2)$ 的概率可由贝叶斯公式计算。需要计算:$P(G_k|x)=\frac{P(G_k)P(x|G_k)}{P(x)}\propto P(G_k)P(x|G_k)$,显然欲计算 $P(G_k|x)$ 关键取决于计算 $P(x|G_k)$ 和 $P(G_k)(k=1,2)$。为了计算 $P(x/G_k)$,引入 Copula 函数的概念。

4.1　Copula 函数的定义及 Sklar 定理

定义[6]　一个 n 维函数 $C(u_1,u_2,\cdots,u_n)$ 满足如下属性:

(1)定义域为 $I^n=[0,1]^n$;

(2)$0\leqslant C(u_1,u_2,\cdots,u_n)\leqslant 1,\forall a,b\in I^n$ 且 $a\leqslant b,V_c([a,b])\geqslant 0$;

(3)对任意的 $u_i=0,C(u_1,u_2,\cdots,u_n)=0$,当所有的 $u_j=1(j\neq i)$ 时,$C(u_1,u_2,\cdots,u_n)=u_i$;

则称函数 $C(u_1,u_2,\cdots,u_n)$ 为 copula 函数。

Skarl **定理**[6]　设 H 是随机变量 $X_1,X_2,\cdots,X_n$ 具有边缘分布 $F_1,F_2,\cdots,F_n$ 的联合分布函数,则存在一个 n 维的 Copula 函数 $C(\cdot)$ 满足:

$$H(x_1,x_2,\cdots,x_n)=C(F(x_1),F(x_2),\cdots,F(x_n))$$

特别地,如果 $F(x_1),F(x_2),\cdots,F(x_n)$ 是连续的,那么函数 $C(\cdot)$ 是唯一的。

4.2　构造映射

设负、正类总体分别为 G_1 和 G_2,设样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$。$P(G_k)$ 被称为先验概率,可通过 $P(G_k)=d_k/d$ 计算,其中 d_k 是属于类 G_k 的训练样本数,d 是训练样本总数。边缘概率 $P(x_i|G_k)=d_{ik}/d_k(i=1,2,\cdots,p)$,其中 d_{ik} 是属于类 G_k 的训练样本中属性 x_i 的个数。由 Copula 函数来确定条件概率 $P(x|G_k)$,选择适当的 Copula 函数 $C'(\cdot)$ 得到联合概率:

$$P(x|G_k)=C'(P(x_1|G_k),P(x_2|G_k),\cdots,P(x_p|G_k)),k=1,2 \qquad (10)$$

特别地,当选取 $C'(u_1,u_2,\cdots,u_p)=\prod_{i=1}^{p}u_i$ 时,则 $P(x|G_k)=\prod_{i=1}^{p}P(x_i|G_k)$。

定义映射 $\phi:X\subset H^p\to R^2$,即 $\phi:\boldsymbol{x}\to(P(G_1)P(x|G_1),P(G_2)\times P(x|G_2))^{\mathrm{T}}$。

由此定义核函数为:

$$K(x,x')=(\phi(x)\cdot\phi(x'))=P(G_1)^2P(x|G_1)P(x'|G_1)+P(G_2)^2P(x|G_2)P(x'|G_2) \qquad (11)$$

其中 $x,x'\in H^p$。

5　可分支持向量分类机

通过构造的距离映射和条件概率映射,将原来的训练集映射到二维空间上,得到新的训练集 $T=\{(x'_1,y_1),(x'_2,y_2),\cdots(x'_l,y_l)\}=\{(\phi(x_1),y_1),(\phi(x_2),y_2),\cdots,(\phi(x_l),y_l)\}$。其中 $x'_i\in R^2,y_i\in Y=\{1,-1\},i=1,\cdots,l$。在二维空间中求解原始最优化问题:

$$\min_{w,b}\frac{1}{2}\|w\|^2$$

$$\text{s.t.}\ y_i((w\cdot x'_i)+b)\geqslant 1,i=1,\cdots l \qquad (12)$$

式(12)中的决策变量只有三个 w_1,w_2,b,则问题大大简化,且易于求解。

在构建可分支持向量机时,需要对映射后的新样本做简易

处理。设映射后的训练样本 $\phi(x_i)=x'_i=(x'_{i1},x'_{i2})$,对不满足条件 $\mathrm{sgn}(x'_{i2}-x'_{i1})=y_i(i=1,\cdots,n)$的样本予以剔除,用剩余的样本构建可分支持向量分类机,剩余样本平面分布如图 1 所示。

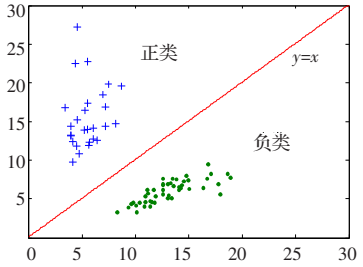

图 1 特征提取后的剩余样本分布示意图

通过求解原始最优化问题(式(12)),得到最优解(w_1^*,w_2^*,b^*),然后构建出决策函数:

$$f(x)=\mathrm{sgn}((\phi(x)\cdot w^*)+b^*)=\mathrm{sgn}(\alpha w_1^* x'_1+w_2^* x'_2+b^*) \quad (13)$$

其中 α 为修正参数,经实验发现 $\alpha\in(0,2)$。

算法步骤:

步骤 1 选择距离映射或条件概率映射,将高维空间的样本映射到二维空间 R^2,映射后的样本记为:

$$T=\{(x'_1,y_1),(x'_2,y_2),\cdots(x'_l,y_l)\}=$$
$$\{(\phi(x_1),y_1),\phi(x_2),y_2),\cdots,\phi(x_l),y_l)\}$$

步骤 2 对经过映射变换后的样本进行处理,对不满足条件 $\mathrm{sgn}(x'_{i2}-x'_{i1})=y_i,(i=1,\cdots,n)$的样本予以剔除。

步骤 3 经处理过的样本构建可分支持向量分类机,求解原问题(式(12)),得到最优解(w_1^*,w_2^*,b^*)。

步骤 4 确定式(13)中的 $\alpha\in(0,2)$。记 α_m 对应的分类正确率为 r_m,从 $\alpha_0=1$ 以步长 0.1 向左搜索,如果 $r_k<r_{k-1}(k<10)$,则停止搜索,再从 $\alpha_0=1$ 以步长 0.1 向右搜索,如果 $r_j<r_{j-1}(j<10)$,则停止搜索,令 r_{k-1} 与 r_{j-1} 较大者对应的 α 为 α',然后从 $\alpha_0=\alpha'$ 以步长为 0.01 按上面的方式进行搜索,进而找到分类效果最好的 α^*。

步骤 5 确定决策函数:$\mathrm{sgn}(\alpha^* w_1^* x'_1+w_2^* x'_2+b^*)$。

6 实验及结果分析

为了验证该算法(GSVM)的分类效果,从 UCI 数据库[8]中选取实验数据集,利用 MATLAB 软件编程实现,并与采用径向基核函数的支持向量分类机(SVM)进行分类效果比较。对未给出测试集的数据集进行随机干扰,打乱数据的排列顺序,然后采取 10 折交叉验证的方式,取 10 次的平均值作为实验的测试结果。表 1 是对算法进行分类正确率比较结果。

表 1 采用标准支持向量机与该文算法分类结果的比较

数据集	数据规模	SVM/(%)	GSVM/(%)
Pima Indians Diabetes	768	70.61	76.54
Wisconsin Breast Cancer	699	92.63	97.91
Spect heart	267	84.35	91.99
Harberman'Survival	306	69.74	75.17
Tic-Tac-Toe	958	–	75.46

实验结果表明,设计的 GSVM 算法比 SVM 算法的分类效果有明显提高,对于数据集 Tic-Tac-Toe,由于其样本属性不是数值型的,SVM 算法不能直接进行分类,而该文的算法可以直接进行分类,并且 GSVM 算法简单、易于实现,采用条件概率映射可以处理非数据型样本的分类问题。

7 结束语

构造的距离映射和条件概率映射,将高维空间的样本转化为二维空间上的样本,再建立可分支持向量分类机,其特点是简单,易于实现,且取得了较好的分类效果。并且建立的可分支持向量分类机,不仅可以处理数据型样本,也可以处理非数据型样本。今后的发展方向还可将马氏距离推广为选用其他距离以及条件概率计算推广为选择其他形式的 copula 函数,从而构造出各种映射和可分的支持向量分类机。

参考文献:

[1] 刘万里,刘三阳,薛贞霞.基于距离核函数的除噪和减样方法[J].系统工程理论与实践,2008(7):160-164.

[2] 孙发圣,肖怀铁.基于 K 最近邻的支持向量机快速训练算法[J].电光与控制,2008(6):44-47.

[3] 朱树先,张仁杰.支持向量机核函数选择的研究[J].科学技术与工程,2008(8):4513-4516.

[4] 邓乃杨,田英杰.数据挖据中的新方法:支持向量机[M].北京:科学出版社,2004:187-189.

[5] Marques deSa J P.模式识别:原理、方法及应用[M].吴逸飞,译.北京:清华大学出版社,2002:80-83.

[6] Nelsen R B.An introduction to copulas[M].[S.l.]:Springer,1999:38-42.

[7] 周涛,张艳宁.基于改进粒子算法的支持向量机[J].计算机工程与应用,2007,43(15):44-46.

[8] UCI Repository of Machine Learning Databases[EB/OL].http://archive.ics.uci.edu/ml/datasets.html.

(上接 129 页)

[3] Agrawal R,Srikant R.Fast algorithms for mining association rules in large databases[C]//Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94),Chile,1994:487-499.

[4] Han J,Pei J,Yin Y.Mining frequent patterns without candidate generation [C]//Proceedings of the 2000 ACM-SIGMOD International Conference on Management of Data(SIGMOD'00),Dallas,Texas,2000:1-12.

[5] Carvalho J P,Tomé J A B.Rule based fuzzy cognitive maps and fuzzy cognitive maps-a comparative study[C]//Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society,New York,USA,1999:115-119.

[6] Carvalho J P,Tomé J A B.Rule Based Fuzzy Cognitive Maps-qualitative systems dynamics[C]//Proc 19th Internat Conf of the North American Fuzzy Information Processing Society,2000:407-411.

[7] Aguilar J.A survey about fuzzy cognitive maps papers (Invited Paper) [J].International Journal of Computational Cognition,2005,3(2):27-33.

[8] Stylios C D,Gmumpca P P.Fuzzy Cognitive Maps:A Soft Computing Technique for Intelligent Contro1[C]//Proc of the IEEE International Symposium on Intelligent Control,Patras,2000:97-102.