SVM中不平衡数据的分离超平面的校正方法

刘万里,刘三阳
LIU Wan-li,LIU San-yang

洛阳师范学院 数学科学学院,河南 洛阳 471022
Department of Mathematics,Luoyang Normal College,Luoyang,Henan 471022,China
E-mail:lwanli@lynu.edu.cn

LIU Wan-li,LIU San-yang.Revising method for separation hyperplane of imbalanced data in SVM.Computer Engineering and Applications,2008,44(19):169-171.

Abstract: A balance method for the offset of separation hyperplane of biclassification imbalanced data is proposed.Firstly,the principal eigenvalues are found respectively of the two classes of samples in feature space by using Kernel Principal Component Analysis(KPCA).Secondly,one penalty proportion is given based on the information provided by the sizes of the two sample data and their eigenvalues.Finally,a new separation hyperplane is generated by the optimization training.The hyperplane revises the error of the standard Support Vector Machines.Experiments show the efficiency of proposed method,i.e.comparing with standard Support Vector Machines the classification error can be balanced and be also decreased.

Key words: imbalanced data;Kernel Principal Component Analysis(KPCA);Support Vector Machines(SVM);offset

摘 要:针对两类不平衡数据的分离超平面的偏移问题提出一种平衡方法。首先,对两类样本数据在核空间中进行核主成分分析,分别求出两类样本数据的在特征空间中的主要特征值;然后,根据两样本容量以及各自的特征值所提供的信息,对两类数据给出惩罚因子比例;最后,通过优化训练,产生一个新的分离超平面。该分类面校正了标准的支持向量机的分类误差。实验显示了所提出方法的有效性,即与标准的支持向量机相比,不仅平衡了错分率而且还能减少错分率。

关键词:不平衡数据;核主成分分析;支持向量机;偏移
DOI:10.3778/j.issn.1002-8331.2008.19.051 **文章编号**:1002-8331(2008)19-0169-03 **文献标识码**:A **中图分类号**:TP181

1 引言

由Vapnik等人创立的支持向量机(SVM)[1]已经在许多领域得到很成功的应用。标准的支持向量机是在假设类分布平衡,样本数据大致相当的前提下使用时,具有较好的精度。然而对于不平衡数据标准的支持向量机的性能大大下降。近几年来,关于不平衡数据的分类问题的研究成为关注的热点[2-9]。针对不平衡数据的挖掘,现有的研究包括两方面的内容:其一是实验研究类分布对各种传统分类算法结果的影响,验证有偏性的存在[2,7];其二是采用适当的方法重构训练样本集,提高分类性能[3-6]。为了解决不平衡问题,文献[5]提出了重新增加正类样本数量(样本数量较少的类称为正类,另一类称为负类)用来弥补与负类的差距,达到平衡作用。该方法的优点是增加了原有信息,但是重新增抽的样本难以保证与原来样本同分布,整体的随机性也不好保持,运作的时间及重新抽样的条件在实际中不一定能满足,不仅增加了运算量,而且过学习情况很可能发生。文献[6]提出减少负类样本数量来达到平衡。这种作法实际上是把相邻的边界点去掉一些,这自然会失去一些有用的信息,随机性也难以保证。认为不平衡数据主要是以下三种情况:(1)两类数据数量差别很大,比如特殊疾病的诊断等;(2)两类数据数目相当,但是类分布差别较大,一类比较集中,另一类比较分散;(3)两类数据数目和类分布都差别很大。这三种情况使用标准的支持向量机都不合适。而从有关的参考文献来看,绝大多数的研究都是针对第(1)种情况来考虑的,即数据数目比例失衡的情况。关于类分布差异的研究较少[8]。文献[9]提出一种加权支持向量机算法,没能给出确定类权重和向量权重的确切方法。将针对两类不平衡数据给出一种方法,称为分布平衡法(DBM)。其步骤是首先对两类样本数据在核空间中进行核主成分分析,然后根据两类样本数据的在特征空间中特征值所提供的信息,对两类数据给出惩罚因子比例;然后通过优化训练,产生一个新的分离超平面。本方法避免了对原有训练样本进行修改,只是将原有样本集中具有的信息充分提取出来,通过自身信息进行调整,达到平衡目的。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60574075,No.60674108)。
作者简介:刘万里(1964-),男,副教授,博士生,研究方向为:机器学习、最优化方法及应用;刘三阳(1959-),男,教授,博士生导师,研究方向为:最优化理论、方法及应用。
收稿日期:2007-09-28 **修回日期**:2008-01-21

2 支持向量机(SVM)简介[1]

设给定样本集$\{x_i\}\in R^d, y_i\in\{-1,1\}$为相应的类标,其中$i=1,2,\cdots,n$。通过引入非线性映射$\varphi: x_i\to\varphi(x_i)$,将$x_i$映射到高维特征空间中。选取适当的核函数$k$,使得$k(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$。通过引入松弛变量$\xi_1,\xi_2,\cdots,\xi_n$,及惩罚因子$C$,求如下规划问题:

$$\min\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t. } y_i[w\cdot\varphi(x_i)+b]>1-\xi_i,\xi_i\geqslant 0,i=1,2,\cdots,n,C>0 \tag{1}$$

其对偶规划为:

$$\max\ w(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)$$

$$\text{s.t. }\sum_{i=1}^{n}\alpha_i y_i=0,0\leqslant\alpha_i\leqslant C \tag{2}$$

求得超平面的法向量为:

$$\boldsymbol{w}^*=\sum_{i=1}^{n}\alpha_i y_i\phi(x_i) \tag{3}$$

选取某个$0<\alpha_j^*<C$,代入下式:

$$b^*=y_j-\sum_{i=1}^{n}y_i\alpha_i^*k(x_i\cdot x_j) \tag{4}$$

求得判别函数为:

$$f(x)=\sum_{i=1}^{n}\alpha_i y_i k(x_i\cdot x)+b^* \tag{5}$$

3 分布平衡方法(DBM)

在使用 SVM 对不平衡数据分类时,偏移性存在的原因除了两类样本数差异外,还与它们的分散程度有关。因为当正类样本数与负类样本数大致相当,如果正类分散程度小于负类分散程度,根据概率论知识可以知道,分离超平面会向分散程度大的类负类偏移,从而使负类的错分率增加。因此,仅由样本数的差异来平衡不平衡数据的偏移性是不全面的。下面将分布因素考虑进去,结合样本容量差异,给出一种算法,称为分布平衡法(DBM)。该方法需要三步进行:(1)使用核主成分分析提取每类数据前p个主要特征值;(2)利用两类样本容量差异以及特征值所提供的信息来确定两类数据的惩罚因子比例;(3)建立模型求解。

3.1 平衡因子的确定

设两类样本集非线性可分:$\{x_i\in R^d\}, y_i=\begin{cases}1, i=1,\cdots,n_1\\-1, i=n_1+1,\cdots,n\end{cases}$,其中$n=n_1+n_2$。引入映射$\phi: x_i\to\phi(x_i)$将其映射到高维特征空间上,并通过选取适当的核函数$k$,对$\forall x_i,x_j$有:$k(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$,其核矩阵可记为:$\boldsymbol{K}=(k(x_i,x_j))_{n\times n}=(k_{ij})_{n\times n}$。

设正类和负类在核空间内的协方差矩阵分别为:$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2$。使用核主成分求得的$\Sigma_1$前$p$个主要特征值为$\lambda_1,\lambda_2,\cdots,\lambda_p$,$\boldsymbol{\Sigma}_2$的前$p$个主要特征值为$\mu_1,\mu_2,\cdots,\mu_p$。分别用下列式子代表两类数据的分散程度:

$$|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}\approx\sqrt{\lambda_1,\lambda_2,\cdots,\lambda_p},\ |\boldsymbol{\Sigma}_2|^{\frac{1}{2}}\approx\sqrt{\mu_1,\mu_2,\cdots,\mu_p} \tag{6}$$

许多研究[3-8]表明:使用 C-SVM 时,分离超平面会向样本数少的类即正类偏移,为了纠正偏移性,采用不同惩罚因子C^+,C^-的比例$\frac{C^+}{C^-}\propto\frac{n_2}{n_1}$[9]。由上述分析,两类惩罚因子比例应有$\frac{C^+}{C^-}\propto\frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}$。因此,综上所述,因子$C^+/C^-\propto\frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{n_1}/\frac{|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}{n_2}$,所以确定$C^+,C^-$如下:

$$C^+=\lambda C,C^-=(1-\lambda)C \tag{7}$$

其中$\lambda=\frac{|\Sigma_1|^{\frac{1}{2}}}{n_1}/(\frac{|\Sigma_1|^{\frac{1}{2}}}{n_1}+\frac{|\Sigma_2|^{\frac{1}{2}}}{n_2})$,$C$为某一正常数。

3.2 模型建立

所求超平面的法向量$\boldsymbol{w}$应满足如下规划:

$$\min\frac{1}{2}\|\boldsymbol{w}\|^2+C^+\sum_{i=1}^{n_1}\xi_i+C^-\sum_{i=n_1+1}^{n}\xi_i$$

$$\text{s.t. }\ y_i[\boldsymbol{w}\cdot\varphi(x_i)+b]>1-\xi_i,\xi_i\geqslant 0,i=1,2,\cdots,n \tag{8}$$

其中C为某一正常数,$C^+=\lambda C,C^-=(1-\lambda)C$。上述规划的对偶规划为:

$$\max\ w(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)$$

$$\text{s.t. }\sum_{i=1}^{n}\alpha_i y_i=0,0\leqslant\alpha_i\leqslant\lambda C,y_i=1,i=1,\cdots,n_1$$

$$0\leqslant\alpha_i\leqslant(1-\lambda)C,y_i=-1,i=n_1+1,\cdots,n \tag{9}$$

由此可重新获得一个新的分离超平面(在应用时,核参数及常数C需要调节,但是只要C被确定,C^+,C^-就被确定)。

4 实验

为了检验 DBM 的性能,在 7 个基准分类问题上将其与 C-SVM 比较。它们来自 UCI 机器学习库。对每个分类问题,将数据标度到区间[-1,1]上。除了数据 Ann-Throid 外,随机选择数据集的 2/3 来训练,1/3 来测试。对 Ann-Throid 取第 2 类和第 3 类训练,测试集取第 2 类的全部测试集和第 3 类测试集中的前 1 000 个作为测试集。所有实验均采用 Gauss 核函数,一律取核参数σ=0.5。惩罚参数C利用 5 倍交叉验证来选取。实验是在奔腾 1.73 G,512 M 内存的 PC 机上安装的 Matlab7.0 软件上实现的,SVM 是利用 http://theoval.sys.uea.ac.uk/svm/toolbox/提供的 Matlab SVM 软件工具包来实现的,DBM 是编制的 Matlab 程序。表 1 给出了所选数据集的特征。"N^+"和"N^-"分别代表正类和负类的个数。表 2 分别给出了支持向量机(C-SVM)和 DBM 在这 7 个基准问题上的实验结果(20 次实验的平均值),C^+代表正类的惩罚参数,C^-代表负类的惩罚参数。表 2 中 C-SVM 的参数C和 DBM 中C^+,C^-的C都是在 1,10,100,1 000,10 000 中利用验证取最有利的数据。在确定每一类主特征值个数时,在大于 0.001 的范围内选取,并且以较小个数为准保持两类具相同的个数。

注意:(1)对多类分类转化为两类问题,本文将其中一类看作正类,它类看作负类;(2)本文采取平均错分率来衡量精度,因为对于不平衡测试数据用总错分率表达精度是不合适[4-6]。其中$Perr(\%)$,$Nerr(\%)$分别代表正类和负类的错分率。用几何平均错分率来表达精度[4-6]:$Merr(\%)=1-\sqrt{(1-Perr)(1-Nerr)}$。

表 1 分类基准问题的数据特征

问题	样本规模	训练集(N^+,N^-)	测试集(N^+,N^-)	类别数	属性
Wisconsin breast cancer	569	380(142,238)	189(70,119)	2	31
Glass2v	214	143(52,91)	71(25,46)	6	10
Breast cancer	699	467(161,306)	232(80,152)	2	10
Bupa liver disorder	345	231(97,134)	114(48,66)	2	6
Iris3v12	150	100(33,67)	50(17,33)	3	4
Wine3v12	178	119(32,87)	59(16,43)	3	13
Ann-Throid2v3	4 838	3 579(191,3 488)	1 177(177,1 000)	3	21

表 2 SVM 与 DBM 实验结果比较

问题	C-SVM				DBM				
	C	*Perr*/%	*Nerr*/%	*Merr*/%	C^+	C^-	*Perr*/%	*Nerr*/%	*Merr*/%
Wisconsin breast cancer	100	4.29	4.20	4.25	2.449	7.550	4.30	2.52	3.40
Glass2v	10	0	0	0	83.170	16.830	0	0	0
Breast cancer	10	12.50	4.61	8.56	9.468	0.532	5	5.21	5.10
Bupa liver disorder	100	33.30	28.80	31.10	6.149	3.251	22.91	33.94	28.60
Iris3v12	100	0	0	0	68.342	31.658	0	0	0
Wine3v12	10	12.50	0	3.39	99.770	0.023	0	0	0
Ann-Throid2v3	100	39.00	1.40	20.20	97.050	2.950	14.20	9.30	11.80

5 实验的结果及分析

由表 2 可知前 6 个实验数据中有 2 个数据 Glass 和 Iris,使用 C-SVM 方法与 DBM 方法具有相同的效果,因为它们的错分率都是 0;其它 5 个数据使用 DBM 方法不仅平衡了正类和负类的错分率而且降低了错分率。这就说明提出的方法不仅能保持 C-SVM 好的分类效果而且能够平衡和减少两类的错分率。对于第 7 数据看到:不仅较好地平衡两类错分率,而且使两类的平均错分率减少 8.4%,这很符合本文目的。因为对于不平衡程度较大的大样本来说,该方法更能体现其价值。

6 结论

本文提出的调整方法有如下特点:(1)使用核主成分分析求出两类核空间数据的协方差矩阵的特征值,根据两类样本数的差异及特征值所提供的信息确定出不同类惩罚因子的比例,减少了以往方法的盲目性;(2)不需要修改原始样本信息,仅把原始样本点所存在的信息提取出来,能平衡两类错分率,甚至能减少错分率;(3)本方法适合于任何形式的不平衡数据,包括样本容量差异和分散差异。

今后研究应考虑把本方法推广到多类分类问题以及实际应用上。

参考文献:

[1] Cristianini N,Shawe-Taylor J.An introduction to support vector machines and other kernel based learning methods[M].Cambridge:Cambridge University Press,2000.

[2] Japkowicz N,Stephen S.The class imbalance problem:a systematic Study[J].Intelligent Data Analysis,2002,6(5):429-449.

[3] Chawla N V,Bowyer K W,Hall L O,et al.Smote:synthetic minority over-sampling technique[J].Journal of Artificial Intelligence Research,2002,16(3):321-357.

[4] Ling C,Li C.Data mining for direct marketing problems and solutions[C]//Proceedings of the Fourth International Conference on Knowledge Discovery and Data Ming.New York:AAAI Press,1998:73-79.

[5] Miroslav K,Stan M.Addressing the curse of imbalanced datasets:one-sided sampling[C]//Proceedings of the Fourteenth International Conference on Machine Learning,Nashville,Tennessee,1997:178-186.

[6] Rehan A,Stephen K,Nathalie J.Applying support vector machines to imbalanced datasets[C]//Fifteenth European Conference on Machines Learning.Berlin:Springer-Verlag,2004:39-50.

[7] 郑恩辉,李平,宋执环.不平衡数据挖掘:类分布对支持向量机的影响[J].信息与控制,2005,34(6):703-708.

[8] Lin I,Lee Y,Wahba G.Support vector machines for classification in nonstandard situations[J].Machine Learning,2002,46(2):191-202.

[9] 贾银山,贾传荧.一种加权支持向量机分类算法[J].计算机工程,2005,31(12):23-25.