不平衡支持向量机的调整方法

刘万里[1,2] 刘三阳[2] 王金艳[3]
(洛阳师范学院数学系 洛阳 471022)[1] (西安电子科技大学应用数学系 西安 710071)[2]
(洛阳师范学院物理系 洛阳 471022)[3]

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调整方法。首先对两类样本数据进行初步训练,产生一个初步分离超平面的法向量,然后把高维样本投影到该法向量上得到一维数据,然后由一维投影数据所提供的信息利用非参数密度估计,根据误差率最小准则,得到一个修正的偏置。几个真实数据试验表明,该方法不仅能保持标准支持向量机的良好性能,而且能平衡或减少总的错分率。

关键词 不平衡数据,投影,支持向量机,Parzen窗密度,偏置
中图法分类号 TP181 **文献标识码** A

Adjusting Method for Imbalanced Support Vector Machines

LIU Wan-li[1,2] LIU San-Yang[2] WANG Jin-yan[3]
(Department of Mathematics, Luoyang Normal College, Luoyang 471022, China)[1]
(Department of Applied Mathematics, Xidian University, Xi'an 710071, China)[2]
(Department of Physics, Luoyang Normal College, Luoyang 471022, China)[3]

Abstract An adjusting method for the offset of separating hyperplane of binary-classification imbalanced data was proposed. Firstly, the two class of samples were trained preliminarily and a normal vector of the separating hyperplane was generated. Secondly one dimensional data were obtained by projecting the high dimensional sample data onto the normal vector. Finally a new algorithm was proposed by using the information provided by the estimation of the non-parameter density for the projection data. Thus the offset of the separating hyperplane of the standard support vector machines (SVM) was revised. Several real experiments show the proposed method can keep the good performance of standard SVM and balance or decrease the whole error ratio.

Keywords Imbalanced data, Projection, Support vector machines (SVM), Parzen-window densiry, Offset

1 引言

由Vapnik等人创立的支持向量机[1](Support Vector Machines, SVM)已经在许多领域得到很成功的应用。我们知道,标准的支持向量机在假设类分布平衡、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[2-10]。针对不平衡数据的挖掘,现有的研究包括两方面的内容:其一是试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[2,8];其二是采用适当的方法重构训练样本集,来提高分类性能[3-7]。为了解决不平衡问题,文献[4]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,的确有改善作用,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,而且很可能发生过学习情况。文献[7]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,在精度上还不如上一种方法好[7]。我们认为,不平衡数据主要是以下3种情况:(1)两类数据数量差别很大,像特殊疾病的诊断等;(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3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[9]。文献[10]提出一种加权支持向量机算法,该方法没能给出确定类权重和向量权重的确切方法。我们将给出一种方法——投影分布平衡法,其步骤是按照标准的支持向量机先进行初步训练,得到超平面的法向量 w 以后,把高维数据投影到 w 上得到一维数据,然后对两类一维投影数据所提供的信息使用非参数密度估计来调整分离超平面的偏置。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

到稿日期:2008-03-20 本文受国家自然科学基金项目(60574075)资助。
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,E-mail:lwanli@yahoo.cn;刘三阳(1959-),男,博士,教授,博士生导师,研究方向为最优化理论、方法及应用;王金艳(1963-),女,副教授,研究方向为实验数据处理。

有的信息充分提取出来，得到一个新的偏置。

2 支持向量机(SVM)模型及算法简介[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 $i=1,2,\cdots,n$。通过引入隐函数 $\phi: x_i \to \phi(x_i)$ 将 x_i 映射到高维特征空间中。选取适当的核函数使得 $k(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$。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：

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

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

其对偶规划为

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

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

求得超平面的法向量为

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

选取某个 x_j 满足 $0<\alpha_j^*<C$，代入

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

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 偏置 b 的估计

设两类样本集 $\{x_i\in R^d\}$，$y_i=\begin{cases}1, & i=1,\cdots,n_1\\ -1, & i=n_1+1,\cdots,n\end{cases}$ 非线性可分，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 $k(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$，其核矩阵可记为 $K=(k(x_i,x_j))_{n\times n}=(k_{ij})_{n\times n}$。

首先使用标准的核支持向量机求得参数 w 表达式 $w=\sum_{i=1}^{n}\alpha_i y_i\phi(x_i)$ 计算投影值，即

$$w\cdot\phi(x_j)=\sum_{i=1}^{n}\alpha_i y_i\phi(x_i)\cdot\phi(x_j)=\sum_{i=1}^{n}\alpha_i y_i k(x_i,x_j) \tag{6}$$

注意：为了方便，这里的内积值可理解为投影值(不必标准化)。

我们认为，利用软间隔法基本上可以消除野点的作用，所以初步训练得到的法向量 w 比较稳定。偏移问题主要是 b 值的偏向所致，所以重新修正一个 b 值。该方法称为投影密度法(PDM)。

假设分离超平面具有如下形式：

$$w\cdot\phi(x)=b \tag{7}$$

3.1 密度估计

我们知道，Parzen 窗密度估计法[11]计算简单，使用方便，对于模式分类是一种非常有效的方法。本文借用 Parzen 方法给出密度估计如下：

针对投影数据式(6)，假设正、负类样本所满足的分布密度分别为

$$P_1(x)=\frac{1}{n_1}\sum_{i=1}^{n_1}(x-w\cdot\phi(x_i))^2$$

$$P_2(x)=\frac{1}{n_2}\sum_{i=1}^{n_2}(x-w\cdot\phi(x_i))^2 \tag{8}$$

由 Parzen 窗密度估计法可知，式(8)中两个密度函数 $P_1(x)$，$P_2(x)$ 相当于窗宽取为 1，省略了标准化常数因子的 Parzen 窗密度。

注意：式(8)中两个密度函数 $P_1(x)$，$P_2(x)$ 不是严格意义上的密度，只是密度与这两个估计呈单调关系，因此可仅考虑式(8)中形式。

3.2 偏置的确定

如图 1 所示，b 点在两类投影数据中概率密度分别为

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(b-w\cdot\phi(x_i))^2,\frac{1}{n_2}\sum_{j=1}^{n_2}(b-w\cdot\phi(x_j))^2$$

由图 1 可知，b 的选取能使两类错分率达到最小，即：

$$\mathrm{Min}\,(Err)=\int_{-\infty}^{b}P_1(x)\mathrm{d}x+\int_{b}^{+\infty}P_2(x)\mathrm{d}x \tag{9}$$

且满足在两类中的密度值相等，即

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(b-w\cdot\phi(x_i))^2=\frac{1}{n_2}\sum_{j=1}^{n_2}(b-w\cdot\phi(x_j))^2 \tag{10}$$

解方程(9)，可得

$$b=[n_2\sum_{i=1}^{n_1}(w\cdot\phi(x_i))^2-n_1\sum_{j=1}^{n_2}(w\cdot\phi(x_j))^2]/2[n_2\sum_{i=1}^{n_1}w\cdot\phi(x_i)-n_1\sum_{j=1}^{n_2}w\cdot\phi(x_j)] \tag{11}$$

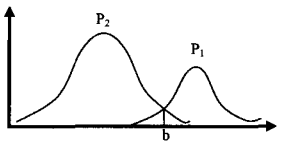

图 1

注意：$P_1(x)$，$P_2(x)$ 未必具有图 1 中那么标准的类似正态分布的密度曲线，但是它们一定有式(10)成立的 b 点存在。所以我们选取 b 点为偏置是合理的。

4 试验

为了检验 PDM 的性能，我们在 6 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[6]。对每个分类问题我们将数据标度到区间[−1,1]，随机选择数据集的 2/3 来训练，1/3 来测试。所有试验均采用 Gauss 核函数，一律取核参数 $\sigma=0.5$，惩罚参数 C 利用 5 倍交叉验证来选取。我们的试验是在奔腾 1.73G，512 内存的 PC 机上安装的 Matlab7.0 软件实现的。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，PDM 是我们编制的 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、表 3 分别给出了支持向量机(C-SVM)和 PDM 在这 6 个基准问题上的实验结果(10 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。注意：对多类分类转化为两类问题，本文将其中一类看作正类，其他类看作负类。表 2 中 Perr(%)，Nerr (%)分别代表正类和负类的错分率；Err(%)代表两类平均错分率。试验结果如表 2 所列。

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

(下转第 155 页)

进一步，为了比较 FAMC 与 AMC 和 FCM 的识别性能，取实验参数：$c=8$；模糊度 $m(1<m\leqslant 2)$。为了避免伪随机数、误差等因素对实验结果的影响，针对每个聚类参数和模糊度，3 种算法各计算 50 次，其实验结果如图 2 所示。

结束语 本文研究了基于伪 F 统计的 FAMC 算法，实验表明：该算法不仅有效地解决了参数的初始化问题；而且与 ACM 及 FCM 算法相比，FAMC 算法在稳健性和识别率上都有明显的改善。因此，本文算法对于处理基因芯片数据这类典型的高维小样本对象来说，是一种较为有效的分析工具。但是，目前关于基因表达数据分析方法的研究仍处于起步阶段，并且聚类结果的正确性在很大程度上仍旧依赖于基因表达数据的预处理。因此，今后我们应着重研究基因表达数据的标准化，使得标准化后的基因数据能够在同一个较小的范围内，从而可以在一定程度上提高聚类结果的正确性。

参 考 文 献

[1] Karayiannis N B. Weighted fuzzy learning vector quantization and weighted generalized fuzzy C-means algorithms. IEEE International Conference on Fuzzy Systems. 1996,2:773-779

[2] 程乾生. 属性均值聚类[J]. 系统工程理论与实践，1998,18(9):124-126

[3] 张媛祥. 一种稳健的聚类方法[J]. 数学的实践与认识，2003,33(8):8-10

[4] Dembele D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics, 2003:973-980

[5] 高新波，裴继红，谢维信. 模糊 C-均值聚类算法中加权指数 m 的研究[J]. 电子学报，2000,28 (4):80-83

[6] Vogel M A, Wong A C. PFS clustering method [J]. IEEE PAMI (S0001-0782),1979,3:237-245

[7] Culley T M, Wallace L E. Calculating F-Statistics [EB/OL]. (2001)[2004]. Http://ib. Berkeley. edu/courses/ibl60/hl3a.html

[8] Iyer, et al. Http://www. sciencemag. org/feature/data/984559.shl. 1999

(上接第 149 页)

]Breast cancer	699	467(161,306)	232(80,152)	2	10
Pima Indians Diabetics	768	513(179,334)	255(89,166)	2	31
Glass2v	214	143(52,91)	71(25,46)	6	10
Ionosphere	351	234(84,150)	117(42,75)	2	34
Iris3v12	150	100(33,67)	50(17,33)	3	4
Bupa liver disorder	345	231(97,134)	114(48,66)	2	6

表 2 SVM 与 PDM 试验结果比较

问题	C-SVM(b=0)				PDM			
	C	Perr (%)	Nerr (%)	Merr (%)	b_1	Perr (%)	Nerr (%)	Merr (%)
Breast cancer	10	12.5	4.61	8.56	0.1672	5	7.24	6.12
Pima Indians Diabetics	100	39.33	21.08	30.21	−0.4634	24.72	26.51	25.62
Glass2v	10	0	0	0	0.0253	0	0	0
Ionosphere	100	45.24	2.67	23.91	8.056	0	20	10
Iris3v12	100	0	0	0	−0.711	0	0	0
Bupa liver disorder	100	35.42	27.27	31.35	0.0389	33.33	27.27	30.3

5 结果及分析

表 2 中，参数 C 是在 1,10,100,1000,10000 中利用验证取得最有利的数据。由表 2 可知，6 个试验数据中有两个数据 Glass 和 Iris 使用 C-SVM 方法与 PDM 方法具有相同的效果，因为它们的错差率都是 0；其他 4 个数据 PDM 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。此结果并不是偶然的，因为毕竟考虑了密度因素，所以我们相信对于大样本不平衡该效果会更佳。

结束语 本文提出的调整方法有如下特点：

1)不需要修改原始样本信息，仅根据原始样本点投影特征分布密度的估计，纠正了由 C-SVM 训练的偏置 b。试验结果表明，本方法不仅能保持标准支持向量机的良好性能，而且能平衡甚至减少错分率。

2) 巧妙地把高维样本点的分类问题转化为一维来处理，利用 Parzen 窗估计法增强了理论根据，减少了以往方法的盲目性。

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程度差异。

本方法不足之处是本文仅在少数几个数据上验证了提出方法的有效性，没能应用到大样本的不平衡数据。今后的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] Nello C, John S T. An introduction to support vector machines and other kernel based learning methods. Cambridge: Cambridge University Press, 2000

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

[3] Chew H G, Bonger R E, Lim C C. Dnal nu-support vector machine with error rate and training size biasing // Proceedings of the 25th IEEE International Conference Acoustics, Speech, and Signal. Piscataway: USA, IEEE, 2001: 1269-1272

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

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

[6] Kubat M, Matwin S. Addressing the curse of imbalanced datasets // One-sided Sampling Proceedings of the Fourteenth International Conference on Machine Learning. Nashville: Tennessee, 1997: 178-186

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

[8] 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. 信息与控制，2005,34 (6):703-708

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

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

[11] 边肇祺，张学工. 模式识别. 北京：清华大学出版社，2000