ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2008/02(2)-0192-06
DOI:10.3778/j.issn.1673-9418.2008.02.007

E-mail:fcst@public2.bta.net.cn
http://www.ceaj.org
Tel:+86-10-51616056

SVM 置信度在线评估以及决策改进*

凌 萍[1,2+],周春光[1]
LING Ping[1,2+], ZHOU Chunguang[1]

1. 吉林大学 计算机科学与技术学院 教育部符号计算与知识工程重点实验室,长春 130012
2. 徐州师范大学 计算机科学与技术学院,江苏 徐州 221116
1. Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science, Jilin University, Changchun 130012, China
2. College of Computer Science, Xuzhou Normal University, Xuzhou, Jiangsu 221116, China
+ Corresponding author: E-mail: lingicehan@yahoo.com.cn

LING Ping, ZHOU Chunguang. For SVM: confidence online evaluation & decision improvement. Journal of Frontiers of Computer Science and Technology, 2008,2(2):192-197.

Abstract: An algorithm of confidence evaluation for SVM is presented. Based on the evaluation, decision risk is specified in the context of multi-classification. Evaluation approach combines the theoretic analysis and empirical analysis. The decision with low confidence is refined by a local classifier that is formulated online. The local classifier works in query's neighborhood, and the neighborhood is developed according to a local metric. Experiments demonstrate the fine performance of the designed algorithm, and its improvement in classification accuracy over the state of the arts.
Key words: SVM; confidence evaluation; decision risk amount; local classifier; local metric

摘 要:设计了 SVM 置信度在线评估方案,以此确定 SVM 做多分类时结果的风险程度,对高风险决策结果进行修正。置信度评估采用理论估计和经验估计相结合的方式。多分类决策结果的修正由在线生成的局部分类器完成。局部分类器在待查询数据的邻域内工作,此邻域基于一个局部测度而生成。实验表明,所设计的算法呈现

* the Key Program of National Natural Science Foundation of China under Grant No.60433020,60673099,60773095 (国家自然科学基金重点项目); 985 Project, Technological Creation Support of Computation and Software Science (985 工程"计算与软件科学科技创新平台"项目); the National High-Tech Research and Development Plan of China No.2007AA04Z114 (国家高技术研究发展计划(863)); the Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China (教育部"符号计算与知识工程"重点实验室资助).
Received 2007-11, Accepted 2008-02.

了较好的分类能力,提高了传统同类方法的分类准确率。

关键词:SVM;置信度评估;决策风险值;局部分类器;局部测度
文献标识码:A　　中图分类号:TP301

1　介绍

SVM是近年来机器学习领域的研究热点，是借助最优化理论解决分类问题的有力工具[1]。它于20世纪90年代由Vapnik基于统计学意义提出,其理论研究和算法实现的基本框架已经形成，并在模式识别、回归分析、时间序列预测等方面发挥了重要作用。对于二分类问题,SVM良好的能力毋庸置疑,但在多分类问题上SVM则表现欠佳。目前对此的解决思路有:(1)一类对余类SVMs的构造(SVM_{1r})[2];(2)成对分类SVMs的构造(SVM_{11})[3];(3)纠错输出编码[4];(4)多类目标函数的SVM[5]。这些方法在实际应用中达到了多分类目的,但效果并不理想,且各种方法存在弊端:SVM_{1r}的基本SVM分类器建立在不对称的两类上，这影响了基本SVM以及总体分类器的决策结果;SVM_{11}须生成大数目的基本分类器，其决策机制缺乏理论依据;纠错编码方法仅适用于一些可合理分解为若干二分类问题的多分类问题,其实际输出常带有明显偏差；第四种方法使用含多个优化目标的函数,具体实现上有困难。

本文以SVM_{1r}为基础，对其基本SVM的决策结果的置信度进行在线评估,进而给出决策结果的风险度,并对高风险决策结果进行修正。置信度值的评估综合了理论分析和经验分析,决策结果的修正由一个局部分类器完成。局部分类器建立于查询数据的邻域之内,邻域则基于一个位查询数据定制的测度而生成。上述设计构成了一个改进的SVM_{1r},称之为$iSVM_{1r}$。

文章的第2、3章将详细讨论$iSVM_{1r}$的步骤细节,第4章给出了$iSVM_{1r}$的实验结果,最后是结论。在进入$iSVM_{1r}$之前,先简要回顾SVM的基本意义。

2　背景知识

SVM寻求具有最大间隔的分类超平面[1]。对训练集$\{(x_1,y_1)(x_2,y_2)\cdots(x_l,y_l)\}\in X\times Y$,其中$X=R^n$,$Y=\{1,-1\}$,SVM构造并求解如下的最优化函数:

$$\max_{\alpha}\sum_i\alpha_i-\frac{1}{2}\sum_i\sum_j\alpha_i\alpha_j y_i y_j K(x_i,x_j)$$
$$\text{s.t. } 0\le\alpha_i\le C_{svm},\ \sum_i\alpha_i y_i=0 \tag{1}$$

得最优解$\alpha^*=(\alpha_1^*,\cdots,\alpha_l^*)^T$。选择$\alpha_j^*>0$,计算$w^*=\sum_{i=1}^{l}y_i\alpha_i^*x_i$,$b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^*(x_i\cdot x_j)$。构造决策函数$f(x)=\mathrm{sgn}((w^*\cdot x)+b^*=0)$。根据各数据点在决策函数值的符号决定类别归属。那些对应了$0<\alpha_i<C_{svm}$的数据点称为支持向量,C_{svm}作为惩罚参数表示在准确性和松弛性之间所做的权衡。

3　$iSVM_{1r}$算法及SVM置信度评估

3.1　$iSVM_{1r}$算法步骤

$iSVM_{1r}$的思想是首先做正常SVM多分类，当决策有高风险时,在线生成局部分类器,进行决策修正。其算法步骤如下所述。

训练阶段:

生成M个一类对余类SVM,其决策函数分别为:$f_1,\cdots,f_M$。

测试阶段:

对查询数据Q，在线生成各基本SVM的置信度系数:$\lambda_1,\cdots,\lambda_M$。

$A1=\max\{\lambda_j\cdot f_j(Q)\}(j=1,\cdots,M)$

$I=\max_j\{\lambda_j\cdot f_j(Q)\}$

$A2=\mathrm{second\_max}\{\lambda_j\cdot f_j(Q)\}(j=1,\cdots,M)$

$J=\text{second\_max}_j\{\lambda_j \cdot f_j(Q)\}$

Q 的决策风险值：$risk(Q)=A1-A2$

If $risk(Q)>\varepsilon$ $label(Q)=I$；

Else

构造局部分类器，作决策修正→

> 构建局部 SVM，相应分界面函数为 g
> 基于 g 建立新测度，并生成 Q 的新邻域
> 在新邻域中用 kNN 重新决策

EndIf

3.2 SVM 置信度评估

设以欧氏测度建立的 Q 邻域为 N1，各基本 SVM 的信用度取决于其自身的分类能力和在 Q 的邻域之上生成决策的一致程度。本文中，前者为理论评估，以决策函数自身分界面的间隔大小表达：$(\|w_j\|)^{-1}$，$(j=1,\cdots,M)$。后者采用经验评估，以 SVM 对 Q 的邻居的决策正确率作为其可信度的一种度量。设 η_j 为 f_j 在 Q 的邻域 N1 中的分类正确率，则基本 f_j 的置信度定义为：

$$\lambda_j=\exp\left(-\frac{\|w_j\|}{\eta_j}\right) \tag{2}$$

其中 exp 机制的引入保证值的稳定性。

4 决策修正

决策修正由局部分类器完成，局部分类器的工作内容为：(1) 构建局部 SVM，相应分界面函数为 g；(2) 基于 g 建立新测度，并生成 Q 的新邻域；(3) 在新邻域中用 kNN 重新决策。

4.1 构建局部 SVM

高风险的决策结果意味着分类器在两个最有可能的类别—I，J—之间缺乏判断能力，局部 SVM 将基于这两个类别建立。即，训练数据包括了 I，J 的 SVs 以及 I，J 两类在 N1 中出现的成员。设 $N1_I=\{x|x\in N1, x\in I\}$，$N1_J=\{x|x\in N1, x\in J\}$，$SV_I=\{SV|SV\in I\}$，$SV_J=\{SV|SV\in J\}$，则局部 SVM 的训练数据为：$T=N1_I\cup N1_J\cup SV_I\cup SV_J$。局部 SVM 设计为线性版本，即其中使用的 Kernel 函数为：$K(x,y)=\langle x,y\rangle$。对应的优化目标为：

$$\max_{\alpha}\sum_i\beta_i-\frac{1}{2}\sum_i\sum_j\beta_i\beta_j y_i y_j\langle x_i,x_j\rangle$$
$$\text{s.t. } 0\le\beta_i\le C_{svm},\ \sum_i\beta_i y_i=0 \tag{3}$$

得最优解 $\beta^*=(\beta_1^*,\cdots,\beta_l^*)^T$。计算局部分界面函数 $g(x)=(w^*\cdot x)+b^*$，其中，$w^*=\sum y_i\beta_i^* x_i$，$b^*=y_j-\sum y_i\beta_i^*(x_i\cdot x_j)$，$\beta_j^*>0$。

4.2 新测度定义

利用 SVM 的分界面函数来定义测度，原因在于此函数蕴含了丰富的类别判断性信息。从数学角度看，SVM 的优化过程保证了分界面函数是在最小结构化风险原则之下的最优化分结果。从几何角度看，分界面函数方向表示了两类别之间的最佳区分方向。更具体地，对决策函数 f 上的任意一点 x，$f(x)=0$，其导数 $f'(x)$ 揭示了在 x 的邻域内数据可被最大程度正确分类的方向。用此作为定义新测度的依据，可有效帮助类别辨认。对局部 SVM 的分界面函数 g，寻求 Q 在 $g(x)=0$ 上的最近邻居 P：$\min\|Q-P\|$，s.t. $g(P)=0$。$g'(P)$ 指明了 P 邻域内的最佳分类方向。然而 P 的寻求产生额外计算耗费，此处直接用 Q 取代 P，记生成导数向量为：$V=g'(Q)=(V_1,\cdots,V_n)$。新测度定义为：

$$D(x,y)=\sqrt{(x-y)^T\mu(x-y)} \tag{4}$$

$$\mu_i=\frac{\exp(B\cdot|V_i|)}{\sum_{j=1}^{n}\exp(B\cdot|V_j|)} \tag{5}$$

式 (5) 中 exp 机制的引入保证了值的稳定性。参数 B 控制了各分量对测度的作用大小，本文定义为 $B=(|g(Q)|)^{-1}$。若 $|g(Q)|$ 小，Q 接近 g，$g'(Q)$ 的效果接近 $g'(P)$，则 $g'(Q)$ 的作用应加强；否则 Q 远离 g，$g'(Q)$ 的效果与 $g'(P)$ 有明显偏差，则 B 较小，μ_i 接近 1/n。

4.3 决策修正

使用 kNN 进行决策修正。kNN 工作于 Q 的新邻域 N2 中，N2 由测度 $D^*(\cdot)$ 生成。文章以 Q 周围自然密集区的大小作为 N2 大小的估计值。具体细节如下：

(1) 从小到大排列 Q 与其它数据点的新距离值 $\{D^*(Q,x_i)\}$;

(2) $|N2|=\max_j\{\frac{D^*(Q,x_j)-D^*(Q,x_{j-1})}{D^*(Q,x_{j-1})}\}$。

前文 N1 的大小也用此方式确定,只是那里使用欧氏测度代替 $D^*(\cdot)$。

5 实验分析

首先测试文中 $D^*(\cdot)$ 的质量。将 $D^*(\cdot)$ 和一些广泛使用的测度进行比较:Machete[6]、Scythe[6]、DANN[7]、Adamenn[8]以及欧氏测度。Machete 是通过寻求维之间的最大局部相关性来定义测度;Scythe 是通用型的 Machete 方法,它改进了 Machete 的贪心策略;DANN 对各个输入维赋予权值,权值的学习引入自适应思想;Adamenn 实际是一种加权的欧氏测度,其权值源于各输入维上的统计信息,其中也用到了自适应策略。另外 Gaussian Kernel 函数产生的相似度也视为一种测度参与比较,其尺度参数由 10 折交叉验证方式确定。这些测度生成的邻域作为 kNN 分类器的工作环境,用分类准确率表示它们的优劣。kNN 工作的邻域大小由文章 3.3 提及的方式确定。实验数据来自 UCI[9]:Musk1、Musk2、Waveform、Diabetes、Liver、Vote。随机抽取 20%的数据作为测试数据,实验结果见表1。

从表中结果可看出,$D^*(\cdot)$ 与 Adamenn 相对于欧氏、Kernel、Machete 以及 Scythe 的优势明显。$D^*(\cdot)$ 在 Musk1、Waveform、Diabetes、Vote 数据集上产生最佳结果,后者在 Musk1、Musk2 以及 Liver 上有最佳分类结果。但 Adamenn 在不能达到最佳分类结果的情况下,能够产生仅次于最优情况的结果,这说明 Adamenn 有较好的算法稳定性。$D^*(\cdot)$ 相对于 Adamenn 的算法稳定性稍差。DANN 稍逊于 Adamenn,它在 Liver 一个数据集上有最好的表现。欧氏测度定义实际是对各维信息的无差别利用,在非高斯分布或分布形状不规则的数据集上不能生成高质量邻域。Kernel 核函数定义隐含了一个从原输入空间到特征空间的映射,力求使新空间中数据的分布趋于规则,这在一定程度上提高了所生成邻域的质量,显示出比欧氏测度稍高的质量。但它对特征空间中各维信息的利用仍是无差别状态,因而结果仍不够理想。在四种专门的测度定义中 Adamenn 和 DANN 优势明显,这归功于它们在权衡各个维的重要程度时使用的自适应策略,可灵活应对各种数据集。Scythe 优于 Machete 则归功于它对贪心策略的改进。综上可知,$D^*(\cdot)$ 与 Adamenn 有相近的性能,这说明 $D^*(\cdot)$ 具有较好的数据特征捕捉能力。注意到 Adamenn 中有 6 个参数需要人为调整,计算耗费巨大,而 $D^*(\cdot)$ 的计算方便,是一个性价比较高的选择。

第二部分实验测定 $iSVM_{1r}$ 分类能力。进行比较的其他分类器有:SVM_{11}、SVM_{1r} 以及前文基于 Adamenn 的 kNN(命名为 AkNN)。另外考虑几种专门的邻域生成方法:iLSH[10]、bLSH[11]、VA-file[12]。这三种方法并不依赖于提出某种新的测度,而是在原有欧氏测度空间中设计方法来生成高质量的邻域。iLSH 和 bLSH 源于 Locality Sensitiveness Hashing(LSH)思想,

Table 1 Comparison of diverse metrics

表 1 不同测度的性能对比

数据集	Euclidean	Kernel	Machete	Scythe	DANN	Adamenn	$D^*(\cdot)$
Musk1	90.4	91.6	94.9	96.0	96.2	97.0	97.0
Musk2	63.8	64.5	66.1	67.8	70.2	71.8	70.0
Waveform	80.3	81.7	82.2	84.9	84.6	85.8	86.7
Diabetes	85.2	87.1	87.0	88.3	90.0	91.2	91.6
Liver	68.1	68.9	70.3	70.4	71.3	71.3	70.8
Vote	92.8	93.2	94.3	96.0	95.1	96.7	96.9

可归纳为利用随机映射寻找数据间的关系,从而建立邻域。它将映射过程视为哈希散列过程,所以经常用于建立数据索引机制的过程中。其中 iLSH 定义了从原输入空间到一维空间的哈希映射,并根据数据的一维射影值确定邻域。bLSH 定义了从原输入空间到某一低维空间的哈希映射,并以低维空间中的球状区域作为生成邻域的基本单位。VA-file 则将数据空间划分为 2^b 个超矩方体,为每一超矩方体建立标识,用这些标识判断相应的超矩方体是否参与构成邻域。实验中设 b=8,实验的邻域大小仍用 3.3 小节的方法确定。

实验数据为 News Group[13],该数据集中含有 20 000 个文档(电子邮件),被平均地分为 10 个类,这里记这些类别为:

NG1:alt.atheism;NG2:comp.graphics;NG3:comp.os.ms.windows.misc;NG4:comp.sys.ibm.pc.hardware;NG5:comp.sys.mac.hardware;NG6:comp.windows.x;NG7:misc.forsale;NG8:rec.autos;NG9:rec.motorcycles;NG10:rec.sport.baseball;NG11:rec.sport.hockey;NG12:sci.crypt;NG13:sci.electronics;NG14:sci.med;NG15:sci.space;NG16:soc.religion.christian;NG17:talk.politics.guns;NG18:talk.politics.mideast;NG19:talk.politics.misc;NG20:talk.religion.misc.

取用其中的若干类构成不同的实验数据:

(1) {NG1(200),NG2(350),NG7(200)}

(2) {NG6(200),NG7(100),NG8(200)}

(3) {NG7(100),NG8(150),NG12(100),NG16(150),NG17(100)}

(4) {NG2(200),NG3(200),NG4(200)}

(5) {NG4(300),NG5(350),NG6(300)}

(6) {NG12(100),NG13(150),NG14(300)}

这里使用 tf.idf 方法把每个文档表示为向量,并对文档向量做标准化。那些出现此数较少的词被删除。在各个数据集中,随机抽取 10%的数据作为测试数据。实验结果见表 2。

从实验结果可以看出,AkNN 和 iLSH 的性能接近,其原因在于 iLSH 是把数据映射到一维空间并用此一维射影值进行分析,而 AkNN 是用从数据的统计信息中提取的一维的测度值进行分析,二者均是基于一维的标量信息讨论数据距离。在高维数据环境下它们对数据间分布结构的刻画能力相近。iLSH、bLSH 和 VA-file 三种方法生成邻域的质量依次递增。正如前文所述,iLSH 仅用一维信息描述高维数据,因而在获取数据的分布结构能力上稍显薄弱,生成的邻域质量欠佳。VA-file 保持了数据的多维特征,可比 iLSH 获知更多的数据局部分布信息,所以提高了邻域质量。但其用于邻域生成的基本单元是具有固定形状的超矩方体,在数据分布不规则时,适应性稍差。bLSH 使用球状区域作为构成邻域的基本单位,球状与邻域的一般形状接近,形成的邻域更贴合数据分布轮廓。$iSVM_{lr}$ 优于传统 SVM_{l1} 和 SVM_{lr},$iSVM_{lr}$ 的实验表现接近并在一些数据集上超过 bLSH,特别是表 2 中数据集 4-6 中,数据类别界限模糊,类别分布有重叠,$iSVM_{lr}$ 凭借对高风险决策结果的修正过程,显示出良好的分类能力。前三种方法是特别为高维数据而设计

Table 2 Comparison of diverse classifiers

表 2 不同分类器性能对比

数据集	SVM_{l1}	SVM_{lr}	AkNN	iLSH	bLSH	VA-file	$iSVM_{lr}$
1	83.5	84.2	85.7	85.3	88.3	86.4	88.5
2	79.6	78.3	82.9	82.8	84.0	83.2	83.6
3	81.0	81.2	83.5	83.1	85.2	83.0	85.1
4	59.0	59.1	58.3	60.7	62.1	61.9	63.5
5	53.4	52.7	58.3	60.7	62.1	61.9	63.5
6	56.0	56.5	62.3	65.1	70.6	69.3	70.7

的,$iSVM_{lr}$与之对比的结果说明$iSVM_{lr}$在改进了原SVM_{lr}的分类准确率的基础上具备了良好的分类能力。

6 结束语

文章对基本SVM的决策能力做了置信度评估,并确定SVM多分类决策的风险值,设计了对高风险结果的修正方案。实验表明,所设计的算法有效提高了传统SVM多分类器的性能并达到了良好的分类效果。

References:

[1] Vapnik V. Statistical learning theory[M]. New York: John Wiley & Sons Publisher, 1998.

[2] Murino V, Bicego M, Rossi I A. Statistical classification of raw textile defects[C]//Proceedings of the 17th International Conference on Pattern Recognition, 2004,4:311-314.

[3] Hastie T J, Tibshirani R J. Classification by pairwise coupling[J]. Advances in Neural Information Processing Systems, 1998,10:507-513.

[4] Dietterich T G, Bakiri G. Solving multi-class learning problems via error-correcting output codes[J]. Journal of Artificial Intelligence Research, 1995,2:263-286.

[5] Weston J, Watkin C. Multi-class support vector machines[C]// Proceedings ESANN, Brussels, 1999.

[6] Friedman J H. Flexible metric nearest neighbor classification[R]//USA: Dept of Statistics, Stanford University, 1994.

[7] Hastie T, Tibshirani R. Discriminant adaptive nearest neighbor classification[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1996,18(6):607-615.

[8] Domeniconi C, Peng J, Gunopulos D. An adaptive metric machine for pattern classification[J]. Advances in Neural Information Processing Systems, 2000,13:458-464.

[9] http://www.uncc.edu/knowledgediscovery.

[10] Mayur D, Nicole I, Piotr I, et al. Locality-sensitive hashing scheme based on p-stable distributions[C]//Proceedings of the 20th Annual Symposium on Computational geometry, 2004:253-262.

[11] Alexandr A, Piotr I. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions[C]//Proceeding of 47th Annual IEEE Symposium on Foundations of Computer Science, 2006:459-468.

[12] Weber R, Schek H, Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces[C]//Proceedings of 24th International Conference on Very Large Data Bases, 1998:194-205.

[13] http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html.

LING Ping was born in 1979. She received the M.S. degree at Jilin University in 2006 and now she is a Ph.D. candidate at Jilin University. Her research interests include computation intelligence, support vector technique and relational data mining, etc. She published 7 papers that are indexed by SCI and EI.

凌萍(1979-),女,江苏徐州人,2006年于吉林大学攻读博士学位,主要研究领域为计算智能、支持向量技术、关系数据挖掘,发表论文被SCI、EI检索7篇。

ZHOU Chunguang was born 1947. He received the Ph.D. degree overseas. He is a professor and doctoral supervisor at Jilin University. His research interests include computation intelligence, machine gestation, neural network, fuzzy system, theories, models and algorithms of evolution computing. Recent years, he took charge of multi national and provincial projects, published more than 120 papers and an academic monograph.

周春光(1947-),男,吉林长春人,归国博士,教授,博士生导师,1982年毕业于吉林大学计算机系,主要研究方向为计算智能、机器味觉、神经网络、模糊系统和进化计算相关理论、模型和算法。近年来,承担了多项国家和省部级项目,发表学术论文120多篇,出版学术专著一部。