

基于SVM的多分类器构造算法的研究

成洪静,陈立潮,张英俊,李 静

(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘 要:在对传统的多类分类算法研究的基础上,针对基于二值分类器的多分类器构造算法存在的预测精度低、训练时间长的缺点,提出了一种基于SVM的组合回归机构造多类分类器的算法。该算法解决了二值分类器方法中存在的信息丢失问题,同时避免了由于参数调整而造成的计算代价过大的问题。实验结果表明:新的SVM多分类算法大大降低了计算代价,提高了运行效率和预测的精度,减少了运行时间。

关键词:支持向量机;回归机;多分类;分类器

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2008)12-0109-04

On Research of Algorithms about Structuring Multi-Classifier Based on SVM

CHENG Hong-jing, CHEN Li-chao, ZHANG Ying-jun, LI Jing

(Sch. of Computer Sci. and Techn., Taiyuan University of Sci. and Techn., Taiyuan 030024, China)

Abstract: By researching the foundation of the traditional multi-classification algorithms, the algorithm about structuring multi-classifier based on SVM is proposed by combining regression machines in this paper. The algorithm can resolve the problem of how to prevent the loss of information which occurs in the usual bi-classifier algorithms, and can avoid the problem of calculated price over great because of adjusting parameter. The results show that the new algorithm decreases the calculated price and the circulated time consumedly so that it enhances the running efficiency and the estimating accuracy.

Key words: support vector machine; regression; multi-classification; multi-classifier

0 引言

SVM最初是针对两分类问题提出的,而实际需要解决的问题大部分为多类分类问题,所以,研究用SVM处理多类分类问题具有重要的现实意义。现有的用SVM处理多类分类问题的算法分为两大类型:(1)解决 n 类问题的直接方法^[1](确定多类目标函数方法)。该算法对所有的样本使用同一个二次规划,只需要一次就可以决定分类。它的局限在于由于要一次处理所有数据,约束条件急剧增加,进行分类的二次规划相当庞大,即使是转化为线性规划,数据的规模依然受限,从而导致它的计算复杂度增加。(2)通过组合多个二值分类器来构造多类分类器。这种类型常见的构造方法有两种:一类对多余类和一对一方法^[2~4]。近年来的改进算法有:纠错输出编码方法(ECOC)^[5]、层

(树)分类方法^[6]、QP-MC-SV算法和LP-MC-SV算法^[7,8]。这些通过组合多个二值分类器构造多类分类器的算法存在一个共同的问题:在训练每个学习器时,仅考虑两类的数据,因此,输出差异很大并且忽视了剩余类的信息,必然会导致数据之间联系的弱化,进而影响了多类分类问题的处理。笔者针对这个问题,提出了一种通过组合回归机来构造多类分类器的算法,是分类和回归技术的有机结合。

1 二值多分类器

支持向量机通常以最大分类间隔作为执行原则。给定一训练集 $Z = \{(x_1, y_1), \dots, (x_n, y_n)\} = \{z_1, \dots, z_n\} \in (X \times Y)_n$, 其中 X 是输入空间, $Y = \{\theta_1, \theta_2\} = \{-1, +1\}$ 是输出空间。 $\Phi: X \rightarrow F$ 是一特征映射, F 为特征空间, 包含在 d 维的实向量集合中, $\Phi = (\Phi_1, \dots, \Phi_d)$ 。

两分类分类器:

$$f_w(x) = \langle \Phi(x), w \rangle + b = \langle X, w \rangle + b \quad (1)$$

$$\text{输出由 } h_w(x) = \text{sign}(f_w(x)) \quad (2)$$

收稿日期:2008-03-26

基金项目:山西省自然科学基金项目(20051044)

作者简介:成洪静(1978-),女,山东菏泽人,硕士研究生,研究方向为数据挖掘;陈立潮,博士,教授,研究方向为数据挖掘、智能化软件等。

得到,最大分类间隔:

$$w_{\text{svm}} = \arg \max_{w \in F} \frac{1}{\|W\|} \cdot \min_{z_i \in Z} \langle x_i, w \rangle \quad (3)$$

通常处理问题的方法是最小化 $\|w\|$ 在约束条件下,(3)式可转化为下列常见的形式:

$$\min_{w \in F} \frac{1}{2} \|w\|^2 \quad (4)$$

约束条件: $y_i \langle x_i, w \rangle \geq 1$, 其中 $z_i \in Z$

$$w_{\text{svm}} = \sum_i a_i y_i x_i \quad (5)$$

其解的形式为:

$$f_{w_{\text{svm}}}(x) = \sum_i a_i y_i k(x_i, x) \quad (6)$$

$k(x_i, x)$ 为核函数,只有少量的 a_i 非零,与它联系的向量为支持向量。

对于多分类情况,可能的类集用 $\{\theta_1, \dots, \theta_l\}$ 表示,其中 $l > 2$,子集 $z_k \in Z, k \in l, z_k = \{z_i = (x_i, y_i): y_i = \theta_k\}$, I_k 是与索引 i 属于同一类的类标识,则 $\bigcup_{i \in I_k} \{(x_i, y_i)\} = Z_k$ 。通常采用一对一的算法分解多分类,首先在分解阶段产生几个平行的学习器,其中每个学习器只对应两个类。其次,重构的思想允许通过合并来自分解阶段的输出计算出所有的输出。这样就需要训练 $l(l-1)/2$ 个两类分类器,产生超平面 $f_{kh}, 1 \leq k < h \leq l$,分隔两类 θ_k 和 θ_h 。如果 f_{kh} 无误差的分隔,则

$$\begin{aligned} \Theta(f_{kh}(x)) &= \\ \begin{cases} \theta_k, \text{sign}(f_{kh}(x)) = 1 \\ \theta_h, \text{sign}(f_{kh}(x)) = -1 \end{cases} \end{aligned} \quad (7)$$

在重构阶段中,用合并的思想分析由平行分解算法训练器产生的类分布。该方法的缺点主要是对于训练每个学习器仅考虑两类问题的数据,因此,输出差异很大且剩余类的信息被忽视。

2 基于 SVM 的组合回归机多分类器

如果超平面 f_{kh} 对 x_i 分类 $i \in I_k \cup I_h$,只有一种正确的解释就是 $f_{kh}(x_i) = 0$ 。很自然地可以想到,对每个不同于类 θ_k 和 θ_h 的训练输入包含在分隔超平面 $f_{kh}(x_i) = 0$ 。为解决剩余类的信息被忽视的问题,这里引入一个参数 $\delta (0 \leq \delta \leq 1)$,即剩余训练向量被压缩在沿着分隔超平面的 δ 管道内,参数 δ 允许覆盖剩余向量的超平面附近产生松弛区域。超平面需解决的优化问题就变成如下形式:

$$\min_{w \in F} \frac{1}{2} \|w\|^2 + C_1 \sum_i \zeta_i + C_2 \sum_j (\varphi_j + \varphi_j^*) \quad (8)$$

约束条件: $y_i \langle w, x_i \rangle \geq 1 - \zeta_i \quad z_i \in Z_{1,3}$

$$-\delta - \varphi_j^* \leq \langle w, x_i \rangle \leq \delta + \varphi_j \quad z_j \in Z_2$$

$$\zeta_i \geq 0 \quad z_i \in Z_{1,3}$$

$$\varphi_j, \varphi_j^* \geq 0 \quad z_j \in Z_2$$

其中, $Z_{1,3}$ 分别属于用 $\{-1, +1\}$ 标识的类,而 Z_2 表示用 0 标识的类。其解的形式与式(6)类似, a_i 是与优化问题相关的拉格朗日乘子,且 $\sum_i a_i = 0$ 。对于新输入 x ,学习器 $f_w(x)$ 的数值输出可以由以下公式解释:

$$\Theta(f_{kh}(x)) = \begin{cases} 1, & \text{若 } f_w(x) > \delta \\ -1, & \text{若 } f_w(x) < -\delta \\ 0, & \text{若 } |f_w(x)| \leq \delta \end{cases} \quad (9)$$

这种方法已经被证明具有较好的结果^[9],然而,通常情况下需要选择一些参数,如核函数 k 、区别两类时误差和的权重 C_1 、区别剩余类时误差和的权重 C_2 和不敏感参数 δ 。

在回归机中,只有 ϵ -支持向量回归机保持了 SVM 的稀疏性^[1],其中 ϵ -不敏感损失参数是这样定义的:设 $\epsilon > 0$,一个超平面 $y = (\omega \cdot x) + b$ 的 ϵ -带是指该超平面沿 y 轴依次上下平移 ϵ 所扫过的区域^[1]。在这里将 ϵ -支持向量回归机引入三值分类器中,由 0 标示的类存在 δ 管道内也可以说存在 ϵ -带区域内,即参数 δ 和参数 ϵ 具有相同的意义。取线性 ϵ -不敏感损失函数,超平面需解决的优化问题(式(8))转化为如下形式:

$$\min_{w \in F, b_j \in R} \frac{1}{2} \|w\|^2 + C \sum_i \sum_j (\zeta_i^j + \zeta_i^{*j+1}) \quad (10)$$

$$\begin{aligned} \text{其中: } & \langle x_i, w \rangle - b_j \leq -1 + \zeta_i^j \quad z_i \in Z_j \\ & \langle x_i, w \rangle - b_j \geq 1 - \zeta_i^{*j+1} \quad z_i \in Z_{j+1} \\ & \zeta_i^j, \zeta_i^{*j+1} \geq 0 \end{aligned}$$

式(10)引入拉格朗日函数:

$$\begin{aligned} L(b_j, \omega, \zeta_i^j, \zeta_i^{*j+1}, \alpha_{ij}^{(*)}, \beta_{ij}^{(*)}) &= \frac{1}{2} \|w\|^2 + \\ & C \sum_i \sum_j (\zeta_i^j + \zeta_i^{*j+1}) - \sum_i \sum_j \alpha_{ij} (-1 + \zeta_i^j - \langle x_i, w \rangle + b_j) - \\ & \sum_i \sum_j \alpha_{ij}^* (-1 + \zeta_i^{*j+1} + \langle x_i, w \rangle - b_j) - \\ & \sum_i \sum_j (\beta_{ij} \zeta_i^j + \beta_{ij}^* \zeta_i^{*j+1}) \end{aligned} \quad (11)$$

式(11)中的拉格朗日系数 $\alpha_{ij}^{(*)}, \beta_{ij}^{(*)} \geq 0$,分别对 $b_j, \omega, \zeta_i^j, \zeta_i^{*j+1}$ 求偏导数或梯度并令它们为 0,得到:

$$\frac{\partial L}{\partial \omega} = 0 \Rightarrow \omega - \sum_i \sum_j (\alpha_{ij}^* - \alpha_{ij}) x_{ij} = 0 \quad (12)$$

$$\frac{\partial L}{\partial b_j} = 0 \Rightarrow \sum_i \sum_j (\alpha_{ij} - \alpha_{ij}^*) = 0 \quad (13)$$

$$\frac{\partial L}{\partial \zeta_i^j} = 0 \Rightarrow C - \alpha_{ij} - \beta_{ij} = 0 \quad (14)$$

$$\frac{\partial L}{\partial \zeta_i^{*j+1}} = 0 \Rightarrow C - \alpha_{ij}^* - \beta_{ij}^* = 0 \quad (15)$$

将式(12~15)代入式(11)中,并对它关于 $\alpha_{ij}^{(*)}$ 求极大,得到如下对偶问题为:

$$\min_{w \in F, b_i \in R} \frac{1}{2} \sum_{i,k} \sum_{j,l} (\alpha_{ij}^* - \alpha_{ij}) (\alpha_{kl}^* - \alpha_{kl}) (x_{ij} \cdot x_{kl}) + \sum_i \sum_j (\alpha_{ij}^* - \alpha_{ij}) - \sum_i \sum_j y_i (\alpha_{ij}^* - \alpha_{ij}) \quad (16)$$

$$\text{其中: } \sum_i \sum_j (\alpha_{ij} - \alpha_{ij}^*) = 0$$

$$z_{i,k} \in Z_{1,3}, z_{j,l} \in Z_2$$

$$0 \leq \alpha_{ij}, \alpha_{ij}^* \leq C$$

其解函数为:

$$f(x) = \sum_i \sum_j (\bar{\alpha}_{ij}^* - \bar{\alpha}_{ij}) K(x_{ij}, x) \quad (17)$$

对于预测样本的输入 x , 其输出可以由式(9)来解释, 其中

$$\delta = \max_{z_i \in Z_1} |f_w(x_i)| = \max_{z_i \in Z_1} | \langle w, x_i \rangle |$$

在处理多于三分类的问题时, 分解阶段采用回归机方法, 把所有的样本分成 n 组, 对每组的样本继续划分成用 $\{-1, 0, 1\}$ 标识的三类, 而在重构阶段仍用 $1-v-1$ SVMs 方法中合并的思想分析 n 个训练器产生的类分布。

3 仿真实验

银行风险评估是一个典型的多分类问题, 文中针对中国银行某支行提供的风险评估数据进行了仿真实验。由于银行数据的保密性, 故详细数据不宜公开, 这里选取 198 个有效数据, 分成两部分即训练样本和测试样本。采用 MATLAB 实现。

算法的评价主要从时间复杂度和空间复杂度来考虑。算法的时间复杂度可以从训练速度的角度进行分析: $1-v-1$ SVMs 是在每两个类别之间构造一个分类面, 得到 $k(k-1)/2$ 个分类函数。假设每个类别的训练样本数相同, 那么每个 SVM 的训练中将有 $2l/k$ 个训练样本参加。整个训练时间为:

$$T_{1-v-1} = \frac{k(k-1)}{2} c \left(\frac{2l}{k} \right)^r \quad (18)$$

其中: l 为训练样本总数, k 为类别数, r 的大小与分解算法有关^[10]。文献[10]中已经证明 $1-v-1$ SVMs 在时间复杂度上优于其它算法, 如一对余类算法及层次分类算法等。而新的 SVMs 是在每三个类别之间构造分类面, 得到 $k(k-1)(k-2)/6$ 个分类函数。假设每个类别的训练样本数相同, 那么每个 SVM 的训练中将有 $3l/k$ 个训练样本参加。这种情况下整个训练时间为:

$$T_3 = \frac{k(k-1)(k-2)}{6} c \left(\frac{3l}{k} \right)^r \quad (19)$$

从计算公式来看, $1-v-1$ SVMs 的训练时间要高于新的 SVMs 的训练时间, 这里 $r=3, k=3, 4, 5, l=198$, 图 1 证实了这一点。算法的空间复杂度可以从

影响算法分类速度两个制约因素进行分析: 1) 对单个未知样本分类所需分类器的数量; 2) 分类器中支持向量的多少。从图 2 中可以看出在类别数较小的情况下, $1-v-1$ SVMs 分类器的数量明显高于新的 SVMs 的分类器的数量。而在分类数为 5 时, 分类器的数量是相等的。文献[10]中已经证明 $1-v-1$ SVMs 在空间复杂度上优于其它算法, 如一对余类算法及层次分类算法等。

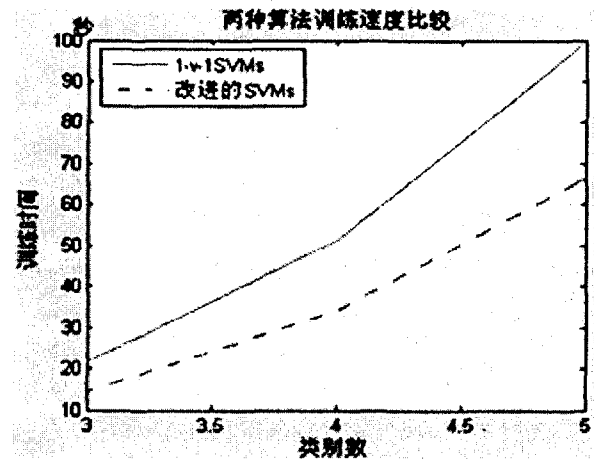


图 1 两种算法训练速度的比较

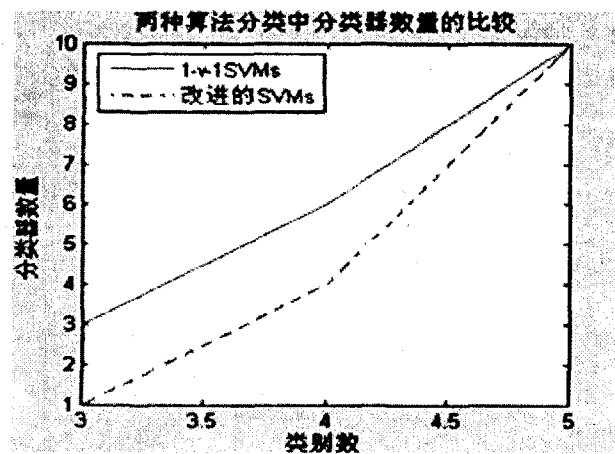


图 2 两种算法分类中分类器数量的比较

将银行的信贷风险分成四类, 分别用组合二值分类器和组合回归机的方法进行对比验证。利用上述计算分类器的公式可知, 需组合六个二值分类器, 而对于回归机只需组合四个就够用。实验结果见表 1, 从总体性能上比较分析说明: 由回归机作为三值分类器避免了信息的丢失和减少了调整参数, 其预测精度比组合二值分类器高出近 3%, 训练时间比组合二值分类器约减少 31%, 而预测时间则减少的更多, 比组合二值分类器减少了 70%, 明显小于组合二值分类器的预测时间。采用回归机这种三值分类器解决多分类问题时, 减少了优化问题的优化规模, 这一点可以从图 1 和图 2 中看出。这些都说明实验结果与上述理论分析是

一致的。

表 1 对于四分类问题分别采用两种算法的运行结果

类别	预测精度	训练时间	预测时间
二值分类器	83.87%	23.6 秒	0.7 秒
回归机	86.29%	16.4 秒	0.2 秒

根据上面的实验结果和分析可知,基于三值分类器的多分类算法无论在时间复杂度上还是在空间复杂度上都优于基于二值分类器的多分类算法。

4 结束语

在实际运用中,大都需要解决多类别的分类问题,该研究具有较强的实用价值。随着支持向量机在各个领域的广泛运用,如何有效地将该方法推广到多分类问题中已引起越来越多的研究者的重视。该文在对两类分类器分析的基础上,针对 $1-v-1$ SVMs 在分解阶段使用两类分类器造成的信息丢失问题,提出了用回归机作为三值分类器,并将其应用于该阶段。实验证明了该方法的有效性,为支持向量机在多分类问题中的应用起到一定的促进作用。

参考文献:

- [1] 邓乃扬,田英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2004.
- [2] Shashua A, Levin A. Taxonomy of large margin principle algorithms for ordinal regression problems[R/OL]. 2007. Leib-

niz Center for Research, School of Computer Science and Eng. The Hebrew University of Jerusalem. <http://citeseer.ist.psu.edu/shashua02taxonomy.html>.

- [3] Kreel U H G. Pairwise classification and support vector machines[C]//Advances in Kernel Methods: Support Vector Learning. Cambridge, MA, USA: MIT Press, 1999: 255 - 268.
- [4] Angulo C, Gonz lez L. 1 - v - 1 tri - class SV machine[C]//In: Proceedings of the 11th European Symposium on Artificial Neural Networks. Bruges, Belgium: [s. n.], 2003: 355 - 360.
- [5] 吴成东,杜崇峰,杨丽英.基于误差修正码的支持向量机大类别分类方法[J].沈阳建筑工程学院学报:自然科学版,2004,20(1):66 - 70.
- [6] Weston J, Watkins C. Support vector machines for multi - class pattern recognition[EB/OL]. 2007. In Proceedings of 7th European Symposium on Artificial Neural Networks, April 1999. <http://citeseer.ist.psu.edu/article/weston99support.html>.
- [7] 黄勇,郑春颖,宋忠虎.多类支持向量机算法综述[J].计算技术与自动化,2005,24(4):61 - 63.
- [8] 唐发明,王仲东,陈锦云.支持向量机多类分类算法研究[J].控制与决策,2005,20(7):746 - 749.
- [9] Angulo C. Learning with Kernel Machines into a Multi - Class Environment[D]. Spanish: Technical University of Catalonia, 2001.
- [10] 刘志刚,李德仁.支持向量机在多分类问题中的推广[J].计算机工程与应用,2004(7):10 - 13.

(上接第 108 页)

表 1 200 次对抗成功率对比

	Q 学习前手工编码	Q 学习后手工编码
DreamWing2006	38%	52%
DreamWing2007	32%	48%

可以看出,学习后的守门员防守能力相比学习前有了较大提高。本实验考虑的是一对一场,学习结果可以用来产生守门员决策,用于防守对方单刀球以及点球。

4 结束语

守门员防守策略问题是 RoboCup 中一个典型的子问题,但它同样是一个实时的 Agent 决策问题。根据我们的经验发现,守门员防守能力作为球队的最后一道防线,往往能在很大程度上影响一支队伍的水平。应用 Q 学习,解决了守门员防守策略问题中状态空间和动作的离散化,实现了 RoboCup 中的守门员防守策

略的优化,实验结果表明了 Q 学习在一定的训练周期结束后收敛,最终得到了优化的防守行为策略。研究结果对解决 Agent 智能决策问题具有普遍意义。

参考文献:

- [1] Stone P. Layered learning in Multi - Agent System[D]. Pittsburgh: Computer Science Department, Carnegie Mellon University, 1998.
- [2] Mihal B, Kay S, Jan W. Learning of kick in artificial soccer [C]//Robot Soccer World Cup IV. Berlin: [s. n.], 2000.
- [3] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence, 1996, 4: 237 - 285.
- [4] Sutton R S, Barto A G. Reinforcement Learning[M]. Cambridge, MA: The MIT Press, 1998.
- [5] 韩学东,洪炳熔,孟伟.强化学习在机器人足球比赛中的应用[J].计算机应用研究,2002(6):79 - 81.