模糊多类 SVM 模型

李昆仑[1,2], 黄厚宽[1], 田盛丰[1]
(1. 北京交通大学计算机与信息技术学院, 北京 100044; 2. 河北大学计算机学院, 河北保定 071002)

摘 要: 利用 SVM 处理多类分类问题, 是当前的研究热点之一. 本文提出了一种模糊多类支持向量机模型, 即 FMSVM. 该方法是在 Weston 等人提出的多类 SVM 模型中引入模糊成员函数, 针对每个输入数据对分类结果的不同影响, 该模糊成员函数得到相应的值, 由此得到不同的惩罚值. 从而在构造分类超平面时, 可以忽略那些对分类结果影响很小的数据. 理论分析与数值实验都表明, 该算法具有良好的鲁棒性.

关键词: 多类分类; 支持向量机 (SVM); 模糊成员函数
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2004) 05-0830-03

Fuzzy Support Vector Machine for Multi-Class Classification

LI Kun-lun[1,2], HUANG Hou-kuan[1], TIAN Sheng-feng[1]
(1. School of Computer & Information Technology, Beijing Jiaotong University, Beijing 100044, China;
2. School of Computer Science, Hebei University, Baoding, Hebei 071002, China)

Abstract: How to process multi-class problem with SVM is one of the present research focuses. We propose a fuzzy multi-class SVM model referred as FMSVM. It is constructed by introducing a fuzzy membership function to the penalty in the quadratic problem of Weston and Watkins, the membership function acquire different values for each input data according to their different affects on the classification results. Hence, we can ignore the data, which affect the classification result a little. Therefore different input points can make different contributions to the learning of the decision surface, i. e., the optimal separating hyper-plane. Both theoretical analysis and digital experiment results show that the model proposed here works very well on benchmark data sets and also has the property of robustness.

Key words: multi-class classification; support vector machine (SVM); fuzzy membership function

1 引言

支持向量机(Support vector machine, SVM)是统计机器学习理论(Statistical learning theory, SLT)的核心内容, 它基于 VC 维理论和结构风险最小化原理[1,2]. 在很大程度上克服了传统机器学习中维数灾难、局部极小点以及过学习等难以克服的困难. 它具有良好的泛化能力, 并拥有传统的机器学习方法无法比拟的性能. 这一理论已经成功应用于数据挖掘和模式识别, 例如手写体数字的识别、物体的识别、语音识别等.

SVM 的研究中有两类问题亟待解决. 其一, 是如何将两类问题的解决办法有效地推广至多类问题. 目前, 已经有一些卓有成效的方法, 如 1-v-1、1-v-r 以及 DAGSVM 等方法. 上述的方法都是基于解决多个两类分类问题的. 换言之, SVM 的多类分类器一般是从二类分类出发设计的. 其二, SVM 对孤立点和噪音数据是非常敏感的, 如何克服这一困难.

与传统的多类 SVM 不同, FMSVM 首先在 Weston 和 Watkins 所提出的多类 SVM 分类器直接构造方法[5]的惩罚项中引入模糊成员函数. 在处理训练数据时, 根据它们在训练过程中重要程度的不同, 区别对待. 也就是说, 将目标函数中的惩罚项模糊化, 重构优化问题及其约束条件, 然后重构其 Lagrangian 公式, 使得原公式(模型)所对应的最优分类超平面的解即为其对偶形式的解.

2 多类 SVM 分类方法简介

下面简单介绍几种常见的多类 SVM 分类算法.

2.1 “one-versus-rest”(一对多, 1-v-r)方法[2]

对于 k-($k>2$)类分类问题, 构造 k 个 2 类 SVM 分类器, 每一类对应其中的一个. 在构造第 i 个 2 类分类器时, 把属于第 i 类的样本点标记为正, 不属于这一类的样本点标记为负. 也就是说, 第 i 个二类 SVM 分类器所构造的分类超平面(separating hyperplane), 把第 i 类与其它的($i-1$)类分割开. 测试时, 对测试样本分别计算各个 2 类分类器的决策函数值, 选择最大的函数值所对应的类别为测试数据的类别.

2.2 “one-versus-one”(一对一, 1-v-1)方法[6]

对于 k-类分类问题, 任意两个不同的类构造一个 2 类 SVM 分类器. 这时共需要构造 $N=k(k-1)/2$ 个 2 类 SVM

收稿日期: 2003-07-20; 修回日期: 2003-09-16
基金项目: 国家“十五”重点科技攻关项目(No. 2002BA407B); 河北省自然科学基金(No. 603137)

分类器. 在构造第 i 类和第j 类的 2 类 SVM 分类器时, 分别选取属于第 i 类和第j 类的样本数据作为训练样本, 将属于第 i 类的样本标记为正, 将属于第 j 类的样本标记为负. 测试时, 将数据对所有的 2 类分类器分别进行测试, 并累计各类的得分, 选择得分最高的类为测试数据的类别.

2.3　构造 k-类 SVM 分类器的直接方法

Weston 和 Watkins 提出了解决 k-类问题的一种直接方法[5], 实际上它是标准 SVM 中二次优化问题的一种自然的推广:

$$\min \Phi(w,\xi)=\frac{1}{2}\sum_{m=1}^{k}(w_m\cdot w_m)+C\sum_{i=1}^{l}\sum_{m\neq y_i}\xi_i^m \tag{1}$$

其约束条件为:

$$(w_{y_i}\cdot x_i)+b_{y_i}\geq(w_m\cdot x_i)+b_m+2-\xi_i^m,$$
$$\xi_i^m\geq 0,\ i=1,\dots,l,\ m,y_i\in\{1,\dots,k\},\ m\neq y_i \tag{2}$$

由此, 得到下面的 k-类 SVM 分类器的决策函数:

$$f(x)=\arg\max_k[(w_i\cdot x)+b_i],\ i=1,\dots,k \tag{3}$$

本文将对上述 SVM 多类分类方法进行改进, 在惩罚项中引入模糊成员函数. 其目的在于根据每个数据点对分类结果的不同影响, 模糊成员函数为其赋予不同的值. 因而, 可以降低噪音对分类结果的影响.

3　引入模糊成员函数的多类 SVM——FMSVM

3.1　输入数据的模糊性

在传统的 SVM 理论中, 训练过程对于那些远离它们所属类的训练点是十分敏感的. 如何设置惩罚项中自由参数 C 的值是非常重要的. 较大的 C 意味着为错误项指定了较大的惩罚值, 可以减少错误分类点. 另一方面, 较小的 C 的值, 意味着忽略了一些“微不足道”的误分类点, 因而可得到较大的分类间隔(margin). 无论 C 的值是大还是小, 在SVM 的训练过程中这个参数的值始终是固定的. 也就是说, 在 SVM 的训练过程中, 所有的训练点都是被平等对待的. 这样就导致了 SVM 对某些特殊情形的过分敏感, 例如孤立点与噪声. 这种情形即是所谓的“过学习”(overfitting) 现象.

在很多实际应用问题中, 不同的训练点对分类结果的影响是不同的. 一般来说, 训练集中存在某些点对分类结果的影响很大, 同时也存在一些点对分类结果的影响很小甚至是微不足道的. 因此在处理分类问题时, 必须将那些“重要的点”正确分类, 并且可以忽略那些“微不足道”的点, 比如说带有“噪声”的点和距离类中心很远的孤立点.

3.2　引入模糊成员函数的多类 SVM 分类器

假设给出了一个带有类别标号和模糊成员函数的训练集 $S:(x_1,y_1,s_1),\dots,(x_l,y_l,s_l)$, 每一个训练样本 $x_i\in R^N$ 都给出了与其对应的类别标号 $y_i\in\{1,2,\dots,k\}$, 以及模糊成员函数 s_i, $\varepsilon\leq s_i\leq 1$, 其中 $i=1,\dots,l$, ε 是一个充分小的正数. 令 $z=\varphi(x)$ 表示特征空间中的向量, 而 φ 是由输入空间 R^N 到特征空间 Z 的映射.

最优分类超平面可由下面二次优化问题的解给出:

$$\min\frac{1}{2}\sum_{m=1}^{k}(w_m\cdot w_m)+C\sum_{i=1}^{l}\sum_{m\neq y_i}s_i\xi_i^m \tag{4}$$

s.t.　$(w_{y_i}\cdot x_i)+b_{y_i}\geq(w_m\cdot x_i)+b_m+2-\xi_i^m,$

$$\xi_i^m\geq 0,\ i=1,\dots,l,\ m,y_i\in\{1,\dots,k\},\ m\neq k \tag{5}$$

其中 C 是惩罚常数, ξ 是松弛变量, s 是所引入的模糊成员函数. 对应的 Lagrangian 公式为:

$$\begin{aligned}L(w,b,\xi,\alpha,\beta)=&\frac{1}{2}\sum_{m=1}^{k}(w_m\cdot w_m)+C\sum_{i=1}^{l}\sum_{m=1}^{k}s_i^m\xi_i^m\\&-\sum_{i=1}^{l}\sum_{m=1}^{k}\alpha_i^m[((w_{y_i}-w_m)\cdot x_i)\\&+b_{y_i}-b_m-2+\xi_i^m]-\sum_{i=1}^{l}\sum_{m=1}^{k}\beta_i^m\xi_i^m\end{aligned} \tag{6}$$

其中, α, β 均为 Lagrange 因子, 它们满足下面的条件:

$$\alpha_i^{y_i}=0,\ \beta_i^{y_i}=0,\ \xi_i^{y_i}=2,\ s_i^{y_i}=1,\ i=1,\dots,l$$

以及约束条件:

$$\alpha_i^m\geq 0,\ \beta_i^m\geq 0,\ \xi_i^m\geq 0,\ \varepsilon\leq s_i^m\leq 1$$

分别求关于 w_n, b_n 的偏导数, 则在鞍点处应满足:

$$\frac{\partial L}{\partial w_n}=0,\Rightarrow w_n=\sum_{i=1}^{l}(A_ic_i^n-\alpha_i^n)x_i \tag{7}$$

$$\frac{\partial L}{\partial b_n}=0,\Rightarrow\sum_{i=1}^{l}A_ic_i^n=\sum_{i=1}^{l}\alpha_i^n \tag{8}$$

$$\frac{\partial L}{\partial \xi_j^n}=0,\Rightarrow Cs_j^n=\alpha_j^n+\beta_j^n\ 和\ 0\leq\alpha_j^n\leq Cs_j^n \tag{9}$$

其中, 引入了符号 c_i^n 和 A_i, 其意义如下:

$$c_i^n=\begin{cases}1, & 当\ y_i=n\ 时\\0, & 当\ y_i\neq n\ 时\end{cases},\quad A_i=\sum_{m=1}^{k}\alpha_i^m \tag{10}$$

将上述结果代入式(6), 可以得到下面的:

$$\begin{aligned}W=&\frac{1}{2}\sum_{m=1}^{k}\sum_{i=1}^{l}\sum_{j=1}^{l}(c_i^mA_i-\alpha_i^m)(c_j^mA_j-\alpha_j^m)(x_i\cdot x_j)\\&-\sum_{m=1}^{k}\sum_{i=1}^{l}\alpha_i^m[\sum_{j=1}^{l}(c_j^{y_i}A_j-\alpha_j^{y_i})(x_i\cdot x_j)\\&-\sum_{j=1}^{l}(c_j^mA_j-\alpha_j^m)(x_i\cdot x_j)+b_{y_i}+b_m-2]\\&-\sum_{m=1}^{k}\sum_{i=1}^{l}\alpha_i^m\xi_i^m+C\sum_{m=1}^{k}\sum_{i=1}^{l}s_i^m\xi_i^m-\sum_{m=1}^{k}\sum_{i=1}^{l}\beta_i^m\xi_i^m\end{aligned} \tag{11}$$

将其进一步化简为:

$$\min W=2\sum_{i,m}\alpha_i^m-\frac{1}{2}\sum_{i,j,m}[c_j^{y_i}A_iA_j-2\alpha_i^m\alpha_j^{y_i}+\alpha_i^m\alpha_j^m](x_i\cdot x_j) \tag{12}$$

约束条件为:

$$\sum_{i=1}^{l}\alpha_i^n=\sum_{i=1}^{l}c_i^nA_i,\quad n=1,\dots,k;$$
$$0\leq\alpha_i^m\leq Cs_i^m;\ \alpha_i^{y_i}=0;\ i=1,\dots,l;\ m,y_i\in\{1,\dots,k\} \tag{13}$$

因而, 可以得到下面的决策函数:

$$f(x,\alpha)=\arg\max_n[\sum_{i=1}^{l}(c_i^nA_i-\alpha_i^n)(x_i\cdot x)+b_n] \tag{14}$$

若给定了一个模糊训练集, 根据上述方法构造的模糊多类 SVM, 即 FMSVM 仍采用分类间隔的最大化以及分类错误的最小化, 使得分类器具有较好的泛化能力. 与传统 SVM 所不同的是, 在 FMSVM 中将惩罚项模糊化, 以降低不太重要的数据对分类结果的影响.

4　数值实验

首先在UCI机器学习数据库[7]中选择了7个数据集,

分别是 iris, wine, glass, soy 和 vowel, 以及 Blood2cell 和 Thyroid. 为了验证 FMSVM 对噪音数据的可行性, 利用 NDC 数据生成器[8] 生成了一个数据集 FTRAIN. 其中, 含有1000 个样本点, 每个样本数据的维数为 8, 样本集的类别数为 3, 该样本集服从正态分布, 在此基础上加入了 50 个噪音数据(均值为 0, 谱密度为一个正常数的白噪声). 在实验中, 取 $s_i = y_i / k$, 其中 k 为类别数, y_i 为类别标号(y_i= 1, , , k).

为了使得实验更具有说服力, 对每一个算法都选取了不同的 C 值, 并使用了两个核函数: 多项式函数 $K(x, x_i) = [(x \# x_i) + 1]^d$ 与径向基函数 $K(x, y) = \exp(-R|x - y|^2)$. 每个算法在实现过程中, 核函数的参数, 比如多项式核函数中的次数 d, 以及径向基函数中的跨度系数 R 的取值使用了几个值进行比较. 采用 10 重交叉验证法估计分类器的正确性, 表 1 和表 2 中给出了平均正确率以及标准偏差.

对实验结果的分析可以看出, FMSVM 算法在所选定的数据集上的性能是较好的. 在/ glass0 ,/ soy0 和/ vow2el0 上的分类结果优于其他算法; 但在/ iris0 和/ wine0 上, 分类结果比 12v21 差. 在/ blood cell0 上, 分类正确率高于其它算法; 但在/ thyroid0 上, 分类正确率低于 12v21 算法, 高于其它算法. 在表 1 和表 2 中, Data 指数据集, Kernel 指核函数, Parm 指参数, J&C 表示文[5] 中所提出的多类 SVM 算法.

把所 FMSVM 算法与 12v2r, 12v21 以及 Weston 等人提出多类 SVM 算法进行了比较. 由实验结果可以看出, FMSVM 与其它几种分类器相比, 对由数据生成器所生成的带有噪声的 FTRAIN 来说, 分类精度大大提高了. 由此可见该算法对处理噪音数据是非常有效的. 并且, 在对实验结论进行的统计显著性分析来看, 该算法的分类正确率明显高于其他三种算法.

表 1　在第一组数据上的实验结论

Data	Kernel	Parm	12v2r(%)	12v21(%)	J&C(%)	FMSVM(%)
iris	Poly	4	961 00? 0127	961 29? 0125	96124? 0130	961 34? 0125
		5	961 89? 0130	961 94? 0125	96186? 0135	961 86? 0127
		6	961 87? 0187	961 89? 0175	96132? 0163	961 33? 0133
	RBF	011	971 02? 0152	971 02? 0169	97102? 0145	971 08? 0125
		1	961 81? 1130	961 89? 1112	96156? 0110	961 65? 1110
wine	Poly	3	971 46? 0156	971 56? 0145	97134? 0155	971 78? 0150
		4	971 57? 1155	971 53? 1140	97157? 1136	971 65? 1136
	RBF	3	971 76? 0127	971 77? 0127	97176? 0127	971 68? 0138
		012	971 56? 0167	971 64? 0165	97170? 0155	971 68? 0152
glass	Poly	3	821 12? 1107	821 25? 1102	81125? 0185	831 30? 0182
		5	821 12? 0176	821 45? 0172	82133? 0170	851 30? 0187
	RBF	013	811 56? 0166	831 35? 0157	82146? 0132	851 50? 0167
		1	821 56? 0128	831 75? 0130	82156? 0198	841 50? 1102
soy	Poly	015	961 66? 1103	971 35? 0192	97156? 0187	971 00? 0156
		2	971 66? 0165	971 35? 0152	97126? 0177	971 50? 0166
	RBF	4	951 66? 0156	961 35? 0166	97122? 0158	971 50? 0151
		5	951 68? 0145	961 33? 0153	97122? 0168	971 40? 0185
vowel	Poly	012	961 66? 0136	971 35? 0155	97146? 0162	971 50? 0187
		2	961 66? 0153	971 35? 0166	97146? 0137	971 50? 0125
	RBF	5	961 65? 0186	971 33? 0192	97145? 0176	971 55? 0195
		6	961 65? 0156	971 36? 0168	97145? 0146	971 55? 1102

表 2　在第二组数据 blood2cell 和 thyroid 上的实验结果

Blood cell	Poly	4	901 25? 1105	921 10? 1114	90135? 1120	911 33? 1118
		5	911 03? 1134	911 90? 1125	90120? 1152	921 10? 1131
		6	911 25? 0198	911 58? 1105	91124? 0187	921 12? 1168
	RBF	10	911 52? 0164	911 58? 0138	91124? 0158	921 12? 0166
Thyroid	Poly	4	921 27? 2115	961 56? 1167	96162? 1137	931 54? 1152
	RBF	10	921 30? 1182	951 75? 1146	95116? 1103	951 46? 1124
FTRAIN	RBF	8	751 03? 3125	761 36? 2187	73123? 3126	801 25? 1157
		10	761 30? 3174	761 35? 2167	75100? 3133	811 02? 1123

5　结论

在本文中, 提出了一种 FMSVM 模型, 即模糊多类 SVM 模型. 它是在 Weston 等人所提出的算法的基础上, 引入一个模糊成员函数. 对每一个数据, 模糊成员函数都给出一个值以确定该数据属于某一类的隶属度, 从而可以确定该点对分类结果的影响程度. 由此可见, 依据此模型得到的算法可以减少噪音数据或孤立点对分类结果的影响, 增强算法的鲁棒性. 理论分析与实验结果表明, 我们所提出的模型是可行的和有效的.

参考文献:

[1] J C Burges. A tutorial on support vector machines for pattern recogni2tion[J]. Data Mining and Knowledge Discovery, 1998, 2: 121- 167.

[2] V Vapnik. Statistical Learning Theory[M]. Wiley2Interscience, Publi2cation, 1998.

[3] C W Hsu, C J Lin. A comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks. 2002, 13(2): 415- 425.

[4] Li Kun2lun, Huang Hou2kuan, Tian Sheng2feng. A novel multi2class SVM classifier based on DDAG[A]. Proc. of IEEE ICMLC. 02[C]. China: IEEE, 2002. 1203- 1207.

[5] J Weston, C Watkins. Multi2class Support Vector Machines [R]. Tech2nical Report, CSD2TR29804, Department of Computer Science, Royal Holloway University of London, England, May 1998.

[6] Ulrich Kressel. Pairwise classification and support vector machines [A]. In B SchÊlkopf, C J C Burges, A J Smola, editors, Advances in Kernel Methods2Support Vector Learning[C]. Cambridge, MA, MIT Press, 1998. 255- 268.

[7] http: / / www. ics. uci. edu/ ~ mlearn/ MLRepository. html[DB/ OL].

[8] http: / / www. cs. wisc. edu/ musicant/ data/ ndc[DB/ OL]. 1998.

作者简介:

李昆仑　男, 1962 年出生于河北保定, 现为北京交大计算机与信息技术学院博士生, 主要研究方向为机器学习、数据挖掘.

黄厚宽　男, 1940 年出生于四川遂宁, 现为北京交大计算机与信息技术学院教授, 博导, 主要研究方向为机器学习、数据挖掘、智能网络安全和多 Agent 等.