基于改进的 SVM分类器的医学图像分类新方法*

蒋 芸[1,2], 李战怀[2]

(1.西北师范大学 数学与信息学院 计算机系, 兰州 730070; 2.西北工业大学 计算机学院, 西安 710072)

摘 要: 支持向量机(SVM)是一种准确度高的分类器,具有很好的容错和归纳能力;粗糙集理论方法在处理大数据量、消除冗余信息等方面具有优势。将两者相结合提出一种改进的 SVM分类算法 ISVM,并将其应用于乳腺 X光图像分类。实验结果表明, ISVM的分类精确度可达到 96.56%,比 SVM的分类精确度(92.94%)要高 3.42%,同时错误分辨率也平均接近 100%。

关键词: 改进的支持向量机方法; 粗糙集; 乳腺 X光图像

中图分类号: TP31 **文献标志码**: A **文章编号**: 1001-3695(2008)01-0053-03

New medical image classify approach based on improved SVM classifier

JIANG Yun[1,2], LI Zhan-huai[2]

(1. Dept. of Computer, College of Mathematics & Information Science, Northwestern Normal University, Lanzhou 730070, China; 2. College of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Support vector machine(SVM) has high classify accuracy and good capabilities of fault-tolerance and generalization. The rough sets theory approach has the advantages on dealing with great data and eliminating redundant information. This paper joined the SVM classifier with rough sets theory which called the improved SVM (ISVM) to classify digital mammography. The experimental results show that the improved SVM classifier can get 96.56% accuracy which is higher about 3.42% than 92.94% using SVM, and the error recognition rates are closed to 100% averagely.

Key words: ISVM; rough sets theory; mammography

0 引言

支持向量机(SVM)是一种建立在统计学习理论基础之上的机器学习方法,其最大的特点是根据 Vapnik[1]结构风险最小化原则,尽量提高学习机的泛化能力,即由有限的训练集样本得到小的误差仍然能够保证对独立的测试集保持小的误差。另外,由于支持向量算法是一个凸优化问题,所以局部最优解一定是全局最优解,这是其他学习算法所不及的[2]。SVM已被广泛应用于模式匹配、分类、聚类、回归估计等领域,但它仍存在一些缺点。经典的 SVM算法建立在二次规划基础之上,它无法区分训练集样本属性的重要性;同时,对于大数据量的模式分类和时间序列预测等问题,如何提高它的数据处理的实时性、缩短训练样本的时间、减少大训练样本集所占用的空间等方面仍是亟待解决的问题。目前已有几种技术用来降低 SVM的复杂度[3],主要是通过最小化核展开式来表示 SVM的解,因为在执行这种预处理技术之前要先计算 SVM的解,这些方法对降低训练阶段的复杂度并不合适。由波兰科学家 Pawlak于 1982年提出的粗糙集理论(rough sets theory)[4],在知识约简、消除冗余信息、处理不确定和不完整知识等方面具有巨大的优势: a)粗糙集仅利用数据本身提供的信息,不需要任何先验知识; b)粗糙集能够表达和处理不完备信息,能在保留关键信息的前提下对数据进行约简并求得知识的最小表达; c)能够识别和评估数据之间的依赖关系,揭示出概念简单的模式,同时能从经验数据中获取易于证实的规则知识。

本文将粗糙集理论和支持向量机相结合,提出了 ISVM算法。利用粗糙集理论处理大数据量、消除冗余信息等方面的优势,减少 SVM的训练数据,不但提高了 SVM的分类能力,而且增强了 SVM的分辨率。最后在乳腺 X光图像标准数据集 MIAS[5]上做实验,与单独使用 SVM方法相比较,比 SVM的分类精确度 92.94%高 3.42%,同时分辨率也平均接近 100%,这更加有利于医学诊断。

1 SVM的基本原理

SVM方法是在统计学习理论之上的一种机器学习方法,它建立在 VC理论和结构风险最小化原理基础上,根据有限样本信息在模型的复杂性和学习能力之间寻求最佳折中,以期获得更好的泛化能力。用 SVM算法来估计回归函数时,其基本思想就是通过一个非线性映射 φ,把输入空间的数据 x映射到一个高维特征空间中,然后在这一高维空间中作线性回归[2]。

给定数据点集 $G=\{(x_i, y_i)\}_{i=1}^{n}$。其中, $x\in R^n$是输入向量; y是期望值,对二类问题, $y\in\{-1, 1\}$; n是数据点的总数。SVM采用式(1)来估计函数:

$$y=f(x)=w\varphi(x)+b \tag{1}$$

其中: $\varphi(x)$是从输入空间到高维特征空间的非线性映射;系数 w和 b由最小化式(2)来估计:

$$R_{SVM}(c)=(c/n)\sum_{i=1}^{n}L_{\epsilon}[y_i, w\varphi(x_i)+b]+\|w\|^2/2 \tag{2}$$

收稿日期: 2006-11-24; **修回日期**: 2007-02-19 **基金项目**: 国家自然科学基金资助项目(60573096); 甘肃省自然科学基金资助项目(3ZS051-A25-042)

作者简介: 蒋芸(1970-),女,浙江绍兴人,副教授,博士研究生,主要研究方向为数据挖掘技术、粗糙集理论及应用(jiangyun@mail.nwpu.edu.cn); 李战怀(1961-),男,陕西旬邑人,教授,博导,主要研究方向为数据库理论与技术.

为了寻找系数 w和 b就需要引入松弛变量 ξ_i和 ξ_i^*，使式(3)最小化：

$$R_{SVM}(w, \xi^*) = \| w \|^2/2 + c\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (3)$$

$$w\varphi(x_i) + b_i - y_i \leqslant \varepsilon + \xi_i^*$$

$$y_i - w\varphi(x_i) - b_i \leqslant \varepsilon + \xi_i,\ \xi_i^* \geqslant 0,\ \xi_i \geqslant 0 \quad (4)$$

其中式(4)是约束条件。最后引入拉格朗日因子 α_i和 α_i^*，由式(1)给出的决策函数就变成下面的精确形式：$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x, x_i) + b$。对任何 i=1, …, n都有等式 $\alpha_i \times \alpha_i^* = 0$, $\alpha_i \geqslant 0$, $\alpha_i^* \geqslant 0$成立。要在条件(4)下最小化式(3)，在引入拉格朗日因子后，就可以把一凸优化问题简化为对一个二次优化问题寻找向量 w的问题。在这种情况下，要找到所求的向量 $w = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i$，必须找到最大化二次型，如式(5)所示：

$$R(\alpha_i, \alpha_i^*) = -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) - \sum_{i}^{n}\sum_{j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j)/2 \quad (5)$$

参数 α_i和 α_i^* (i=1, …, n)的约束条件是：$\sum_{i=1}^{n}\alpha_i^* = \sum_{i=1}^{n}\alpha_i$，$0 \leqslant \alpha_i \leqslant c$，$0 \leqslant \alpha_i^* \leqslant c$(i=1, …, n)。通过在二次优化方法中控制 c和 ε两个参数就可以控制(即使在高维空间中)SVM的泛化能力。根据二次规划中的库恩—塔克条件，在式(5)中系数 $(\alpha_i - \alpha_i^*)$只有一部分数目是非零值，它们所对应的数据点就是支持向量。这些数据点位于决策函数的 ε边界上或边界外。在式(5)中由于其他数据点的系数 $(\alpha_i - \alpha_i^*)$都等于零，从而证实了在所有的数据点中只有支持向量能够决定决策函数。一般来说，ε值越大，支持向量数目就越少，因而解的表达就越稀疏。然而大的 ε值也能降低数据点的逼近精度，从该意义上讲，ε也是解的表达稀疏程度与数据点的密度之间的平衡因子。式(5)中 $K(x_i, x_j)$称为核函数，核函数的值等于两个向量 x_i和 x_j在其特征空间 $\varphi(x_i)$和 $\varphi(x_j)$中的内积，即 $K(x_i, x_j) = \varphi(x_i) \times \varphi(x_j)$。任何函数只要满足 Mercer条件都可用做核函数，采用不同的函数作为核函数，可以构造实现输入空间中不同类型的非线性决策面的学习机器。

2 改进的 SVM 分类算法 ISVM

2.1 粗糙集理论[6]

决策系统 S=(U, A, V, f)。其中：U是全域，是一个非空有限集；A=C∪D, C和 D分别为条件和决策属性集；V是属性的值域集，$V = \bigcup_{a \in A} V_a$，$V_a$是属性 a的值域；f是信息函数 $f: U \times A \to V$，对 $\forall x \in U$, $a \in A$，存在 $f(x, a) \in V_a$。$\forall B \subseteq A$是条件属性集合的一个子集，称二元关系 Ind(B)为 S的不可区分关系：$Ind(B) = \{(x, y) \in U \times U \mid \forall a \in B, f(x, a) = f(y, a)\}$，它表示对象 x和 y关于属性集 A的子集 B是不可区分的。给定 $X \subseteq U$, $B(x_i)$是按等价关系 Ind(B)得到的包含 x_i的等价类。子集 X的下近似集 $\underline{B(X)}$和上近似集 $\overline{B(X)}$分别定义如下：

$$\underline{B(X)} = \{x_i \in U \mid B(x_i) \subseteq X\}$$

$$\overline{B(X)} = \{x_i \in U \mid B(x_i) \cap X \neq \varnothing\}$$

如果 $\overline{B(X)} - \underline{B(X)} = \varnothing$，则集合 X为 B上的可定义集合；否则称 X为 B上的粗糙集。X的 B正域是所有根据知识 B能确定地划入集合 X的 U中对象的集合，即 $POS_B(X) = \underline{B(X)}$。根据正域的概念，决策属性 D和条件属性 C的依赖度定义为：$\gamma(C, D) = card(POS_C(D))/card(U)$。其中，card(X)表示集合 X的基数；$\gamma(C, D) \in [0, 1]$。

粗糙集属性约简是在不损失信息的前提下删除冗余的属性，属性约简集的集合 R可以表示为 $R = \{R : R \subseteq C, \gamma(R, D) = \gamma(C, D)\}$，因此属性依赖度相等可以作为迭代运算的终止条件。

2.2 属性约简算法

设有一决策信息表 $S' = (U, C \cup D, V, f)$，集合 $C' \subseteq C$是 C的一个最小约简，如果 C′满足下条件：$POS_{C'}(\gamma) = POS_{C''}(\gamma)$；不存在 $C'' \subset C'$，使得 $POS_{C''}(\gamma) = POS_{C'}(\gamma)$。根据属性依赖度的定义，任意属性 $a \in C - R$的重要度可以定义为：$\theta(a, R, D) = \gamma(R \cup \{a\}, D) - \gamma(R, D)$。当 $R = \varnothing$时，$\theta(a, D) = \gamma(\{a\}, D)$。根据以上定义，设计以下属性约简算法。

算法 1 属性约简 reduce(S′, R)

输入：决策信息表 $S' = (U, C \cup D, V, f)$

输出：决策表 S′的一个属性约简集 R

a) $R = \varnothing$;

b)对每个属性 $a_i \in C - R$，计算其属性重要度 $\theta(a_i, R, D)$;

c)选择使 $\theta(a_i, R, D)$最大的属性 a_i，$R \leftarrow R \cup \{a_i\}$;

d) if $\gamma(R, D) = \gamma(C, D)$ then转 e), else转 b);

e) return (R); /*返回约简后的属性集 R

显然，此算法的计算复杂度为 $O(m^2)$。其中 m为决策表 S′中条件属性的个数。

2.3 ISVM 算法

ISVM算法主要由两部分组成。首先用 2.2节中提出的约简算法对数据集进行约简；然后再将约简后的数据集作为输入，用 SVM方法作分类。

算法 2 ISVM算法

输入：决策信息表 $S = (U, C \cup D, V, f)$

输出：分类结果 Y

a) discrete(S); /*离散化决策信息表 S

b) reduce(S, R);

/* 调用约简算法，得到约简后的条件属性集 R* /

c) $S = R \cup D$; /*约简条件属性后的新的决策信息表 S

d) SVM(S, Y); /* 调用 SVM分类器，新表 S作为它的输入，Y是分类结果* /

e) return(Y); /*返回分类结果

3 实验

3.1 数据集说明

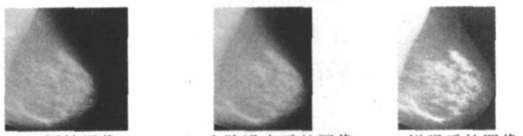
本文用于实验的数据集来自于 MIAS[5]，它是研究乳腺 X光图像的标准数据集。在 MIAS数据集中包含 322幅乳腺 X光图像，所有图像都是乳腺侧面图，它们分属于三类：正常、良性和恶性，后两类又统称为非正常。其中，属正常的图像 208幅，非正常 114幅(良性 63幅，恶性 51幅)。所有非正常的图像都包含出现异常的位置等信息，如肿瘤的圆区域、它的半径、乳房位置(左、右)、乳房组织的类型(密度、多脂的、多脂含腺的)以及是否存在肿瘤等。

3.2 数据集的预处理和特征提取

MIAS数据集中图像的典型尺寸是 1 024×1 024，由于这些图像是在不同外部条件下获取的，一些图像的亮度很高而另一

些图像却太暗,其中 50%的图像在背景中含有大量噪声。

去除噪声首先是用剪切操作来修剪图像;然后是图像增强。本文去除了几乎所有的背景信息和大多数噪声。图 1(a)是 MIAS中的一幅原图;(b)是经过剪切和去除噪声以及背景信息后的图像。由于图像的大小不同,在做剪切操作时横纵坐标的 x和 y的取值范围规定为(0, 255)。本文用垂直剪切的方法去除了多余的部分;然后用直方图均衡法增强图像避免图像过亮或过暗影响分类的效果,图 1(c)就是经过增强后的图像效果。



(a)原始图像　(b)去除噪声后的图像　(c)增强后的图像

图 1　图像对比

预处理之后,将提取的特征数据放入数据库中,并加入一些 MIAS数据集中已经存在的有关图像的信息,构成用于作数据分类的特征库。本文提取的特征是四个统计参数:均值(mean)、方差(variance)、偏斜度(skewness)和峰度(kurtosis)。这四个参数的计算公式分别如下[7]:

$$\text{mean:}\quad \mu=\sum_{k=1}^{N} f_k p_f(f_k) \tag{6}$$

$$\text{variance:}\quad \sigma^2=\sum_{k=1}^{N}(f_k-\mu)^2 p_f(f_k) \tag{7}$$

$$\text{skewness:}\quad \mu_3=\sum_{k=1}^{N}[(f_k-\mu)^3 p_f(f_k)]/\sigma^3 \tag{8}$$

$$\text{kurtosis:}\quad \mu_4=\sum_{k=1}^{N}[(f_k-\mu)^4 p_f(f_k)]/\sigma^4 \tag{9}$$

首先将图 1(c)中的图像均分成 4块,再将其中的每一块均分成 4块,最终将该图像均分成 16块,在每一块中分别提取 4个统计参数,一共获得了 64个统计特征。

3.3　实验结果及分析

本文用 10层交叉的方法在特征库上作分类测试,将特征库随机分成 10份,选择其中 90%作训练,其余 10%作测试,分别记录 SVM和 ISVM的分类精确度。特征库是由从 MIAS的每幅图中抽取的 64个统计参数和已存在的一些数据组成,共 69个属性,所有连续值属性都用算法 DBCHI2[8]进行了离散化处理,其中,SVM算法程序来自 LIBSVM[9]。表 1是实验结果。

表 1　ISVM与 SVM算法在数据集 MIAS上的实验结果比较

10次划分	SVM	ISVM	
	分类精确度/%	所选属性数量	分类精确度/%
1	93.56	21	96.42
2	90.21	16	97.12
3	92.19	18	97.56
4	93.88	15	96.87
5	93.47	23	96.06
6	94.66	20	96.44
7	92.25	13	95.15
8	90.83	26	94.96
9	93.64	19	97.34
10	94.75	15	97.69
平均值	92.94	18.6	96.56

其中:第一列是对数据集的 10次随机划分;第二列是 SVM的 10次分类精确度;第三、四列分别给出了 ISVM经约简后的条件属性数量以及 10次分类精确度。表 1的最后一行是相应列的平均值。从表 1中可以看出,虽然 SVM的平均分类精确度也达到了 92.94%,但仍比 ISVM的平均分类精确度 96.56%低 3.42%。同时由于先使用粗糙集原理对原数据集进行了约简,最终输入 SVM作分类的数据集的平均条件属性数量只有 18.6个,远远小于提取的 69个特征属性,从而简化了后继 SVM的处理过程。

笔者还通过实验对两种算法在 MIAS数据集上的小样本的错误分辨率以及训练所需时间作了比较。图 2是训练样本数从 20到 100个的错误分辨率比较,图 3是训练样本数从 10到 50个的训练所需时间的比较。从图 2可以看到,在 MIAS数据集上,ISVM的错误分辨率明显高于 SVM,平均都接近 100%;而 SVM的错误分辨率变化比较大,尤其是在小样本阶段,样本数小于 50时错误分辨率达不到 90%。实验结果说明使用 ISVM分类器将非正常乳腺 X光图像错误分类的可能性很小,这正是医学专家所期望的。从图 3可以看到,SVM和 ISVM训练所需时间很接近,ISVM花费的时间要比 SVM略多一些,主要原因是 ISVM算法首先要对数据集作约简。因为时间是以秒计,实际应用中时间的差别非常小。

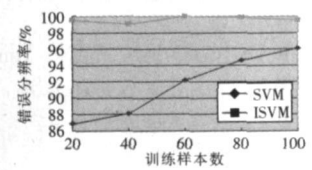


图 2　SVM和 ISVM的错误分辨率比较

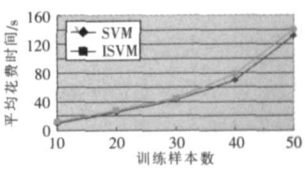


图 3　SVM和 ISVM的平均训练时间比较

4　结束语

本文将粗糙集属性约简原理与 SVM相结合,构造了改进的 SVM分类器 ISVM。首先用粗糙集的属性约简原理将数据集中不确定的、冗余的信息去除掉,然后将数据集中确定的部分交给 SVM作分类,从而增强了 SVM的分类能力。本文将 ISVM应用于乳腺 X光图像标准数据集 MIAS的分类。实验结果表明,ISVM在 MIAS数据集上的分类效果优于 SVM。这种分类方法还可以应用于其他领域。

参考文献:

[1] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer Verlag, 1995: 4-80.

[2] WANG L P. Support vector machine: theory and application[M]. New York: Springer Verlag, 2005: 1-66.

[3] SCHÖLKOPF B, SMOLA A J. Learning with kernels[M]. Cambridge: MIT Press, 2002: 54-62.

[4] PAWLAK Z W. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341-356.

[5] The mammographic image analysis society[DB/OL]. [2006-09]. http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html

[6] PAWLAK Z W. Rough sets and intelligent data analysis[J]. Information Sciences, 2002, 147(1-4): 1-12.

[7] ANTONIE M L, ZAIANE O R, COMAN A. Application of data mining techniques for medical image classification[C]//Proc of the 2nd International Workshop on Multimedia Data Mining. San Francisco: [s.n.], 2001: 94-101.

[8] HU Xiao-hua, CERCONE N. Data mining via generalization, discretization and rough set feature selection[J]. Knowledge and Information Systems: An International Journal, 1999, 1(1): 135-149.

[9] CHANG C, LIN C. LIBSVM[DB/OL]. [2006-09]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.