粗糙支持向量机

梁宏霞 闫德勤
(辽宁师范大学计算机与信息技术学院 大连 116029)

摘 要 支持向量机(SVM)是一种重要的数据机器学习工具,其有效性依赖于对数据信息获取的准确性。以往的支持向量机模型都没有考虑到数据等价类信息。为此,基于粗糙集理论和支持向量机思想,提出了一种新的支持向量机模型——粗糙支持向量机(RSVM)。采用UCI机器学习数据库中的数据做对比实验,结果表明RSVM比传统支持向量机(SVM)和模糊支持向量机(FSVM)都有较高的测试精度。

关键词 支持向量机,等价类,粗糙支持向量机
中图法分类号 TP391 **文献标识码** A

Rough Support Vector Machine

LIANG Hong-xia YAN De-qin
(School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract As a powerful machine learning tool, support vector machine (SVM) is now widely used and studied. However, one kind model of SVM is not absolutely suitable for all kind of data. The effectiveness of a SVM depends on correctness of acquiring the information of data. With consideration of acquiring the equivalence information among data, based on the theory of rough sets and SVM, a new model called rough support vector machine (RSVM) was proposed. The data used in experiments are selected from UCI machine learning data base. For comparison, three kinds of support vector machine, traditional support vector machine (SVM), fuzzy support vector machine (FSVM) and RSVM are used in experiments. The results of experiments show that compared with SVM and FSVM, RSVM has a remarkable predictive accuracy.

Keywords SVM, Equvalence class, RSVM

1 引言

支持向量机(SVM)是20世纪90年代中期在统计学习理论的基础上由Vapnik提出的一种新的机器学习方法[1],它基于VC维和结构风险最小化理论[2](SRM),在很大程度上解决了传统机器学习中的维数灾难及局部极小等问题[3]。由于SVM具有很好的分类能力而成为机器学习的重要工具。为了使SVM有更广的适应性,人们对其结构进行了多种改进。如模糊支持向量机(FSVM)[4]针对每个输入数据对分类结果的不同影响,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,提高了支持向量机的抗噪性;概率支持向量机(PSVM)[5],分配概率值为每一个样本,体现样本间的概率分布特性。

对于SVM而言,其有效性在很大程度上依赖于对数据信息获取的准确性。在很多情况下,数据不仅含有聚类(基于距离的)信息,同时含有(基于特定性质的)等价类信息。由于以往的支持向量机模型都没有考虑到数据等价类信息,面对处理存在等价类属性的数据时,人们一般的做法是先利用粗糙集的属性约简功能将数据样本进行属性约简和规则提取,之后送入支持向量机进行训练和测试。这样,支持向量机要承担属性约简算法本身产生的条件限制和信息损失。

为更好地获取数据中所隐含的信息,本文基于粗糙集理论和支持向量机思想,提出了一种新的支持向量机模型——粗糙支持向量机(RSVM)。应用这个新的模型,数据的等价信息的获取可以在RSVM中明确地体现出来。采用UCI机器学习数据库中的数据做对比实验,结果表明RSVM比传统支持向量机(SVM)和模糊支持向量机(FSVM)都有较高的测试精度。

2 基本概念

2.1 SVM[1]

设给定的训练数据集为:$\{x_i, y_i\}$,其中$i=1,\cdots,N$,相应的类标签为$y_i=\{-1,+1\}$。在线性可分的情况下,SVM能找到一个超平面来使两类的分类间隔最大。这等价于解决下面的规划问题:

$$\min \frac{1}{2}w^T w$$
$$\text{s.t: } y_i(w \cdot x_i + b) \geqslant 1 \quad (1)$$

到稿日期:2008-05-23 本文受国家自然科学基金(No. 60372071),中国科学院自动化研究所复杂系统与智能科学重点实验室开放课题基金(20070101),辽宁省教育厅高等学校科学研究基金(2004C031)资助。

梁宏霞 硕士研究生,主要从事数据挖掘、图像检索方面的研究,E-mail:lhx-19822003@163.com;闫德勤 博士,教授,主要从事模式识别、数据挖掘、密码学和图像检索方面的研究。

决策函数可以表示为

$$f(x)=sign(w\cdot x+b)=sign(\sum_{i=1}^{N}\alpha_i y_i(x_i\cdot x)+b) \quad (2)$$

对于非线性的情况，我们可以通过非线性映射将原始空间映射到高维特征空间。可以找到一个合适的核函数 $K(x_i\cdot x)$将数据映射到特征空间。这样式(2)变为

$$f(x)=sign(w\cdot x+b)=sign(\sum_{i=1}^{N}\alpha_i y_i K(x_i\cdot x)+b) \quad (3)$$

支持向量机是在最小化错分，同时最大化分类间隔的情况下获取最优分类面。

2.2 粗糙集[7]

设 $U=\{x_1,x_2,\cdots,x_n\}$是一有限集，称为论域。R 是 U 上的一个等价关系，U/R 表示在 U 上导出的所有等价类；$[x]_R$ 表示包含元素 x 的 R 的等价类，$x\in U$。

Pawlak 粗糙集模型基于传统的粗糙集定义方法[1,5]：

对任一集合 $X\subseteq U$

$R_-(X)=\{x\in U|[x]_R\subseteq X\}$

$R^-(X)=\{x\in U|[x]_R\cap X\neq\varnothing\}$ ($\varnothing$ 为空集)

分别称 $R_-(X)$与 $R^-(X)$为 X 的 R 下近似和 X 的 R 上近似。

3 粗糙支持向量机

3.1 RSVM 理论模型

设给定的训练数据集为$\{x_i,y_i,r_i\}$，其中 $i=1,\cdots,N$，相应的类标签为 $y_i=\{-1,+1\}$，r_i 为等价类系数，其目标函数可以表示为：

$$\min\frac{1}{2}w^Tw+\sum_{i=1}^{N}[C\xi_i+r_i\eta_i]$$

$$\text{s.t.:}\begin{cases}y_i(wx_i+b)\geqslant 1-\xi_i\\ y_i(wx_i+b)\geqslant 1-\eta_i\end{cases}$$

其中 $\xi_i\geqslant 0,\eta_i\geqslant 0$ (4)

式(4)的对偶问题为：

$$\min Q(\alpha,\beta)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j(\alpha_i+\beta_i)(\alpha_j+\beta_j)K(x_i\cdot x_j)-\sum_{i=1}^{N}(\alpha_i+\beta_i)$$

$$\text{s.t.:}\begin{cases}\sum_{i=1}^{N}(\alpha_i+\beta_i)y_i=0\\ 0\leqslant\alpha_i\leqslant C\\ 0\leqslant\beta_i\leqslant Cr_i\end{cases} \quad (5)$$

其中 $C\sum_{i=1}^{N}\xi_i$ 是为了控制错误分类样本的数量；r_i 是等价类系数因子，用来控制样本的等价性，刻画样本的等价类信息。此时最优判别函数为：

$$f(x)=sign(w\cdot x+b)$$

$$sign(\sum_{i=1}^{N}(\alpha_i+\beta_i)y_iK(x_i\cdot x)+b)$$

$$\text{s.t.:}\begin{cases}0\leqslant\alpha_i\leqslant C\\ 0\leqslant\beta_i\leqslant Cr_i\end{cases} \quad (6)$$

3.2 分配等价类系数

提出基于粗糙集的支持向量机，重要的是如何根据样本数据的特点来分配它的等价性系数。这个系数刻画了样本数据之间的有效联系，特别是对那些数据之间具有很强等价性的样本进行了很好的刻画及描述，然后将这些样本送入粗糙支持向量机模型中训练学习。

在具体实现的过程中，我们依据数据的最普遍联系，考虑其属性的相同程度为其等价性比例。

定义等价类系数

$r_i=C(a)/N,i=1,2,\cdots,N$

其中，N 为样本总数；$a=\{a_1,a_2,\cdots,a_j,\cdots,a_d\}$表示条件属性集合，$d$ 表示条件属性个数。当 $\frac{j}{d}=1$，表示全精度等价；$\frac{j}{d}=90\%$，表示 90%等价。依此类推。$C(a)$表示所有样本中条件属性 a 相同的个数。

这样做的好处是可以最广泛、最大众地刻画数据之间的有效联系，将等价性强的数据可以更好地进行分类识别，提高支持向量机的泛化能力。

4 实验设计及结果

对于粗糙支持向量机模型(RSVM)，我们通过大量的实验来验证它的有效性与可行性。实验一共分为两部分来进行：(1)两类样本实验；(2)多类样本实验。其中两类样本实验和多类样本实验的数据源都是从 UCI 机器学习数据库中选取的。所选取的数据集类别 2～10 不等，样本数 150～846 不等，属性数 4～18 不等，如见表 1 所列。

表 1 数据集说明

Data set	Attribute	Class	Number
breast	9	2	683
pima	8	2	768
heart	13	2	296
bupa	6	2	345
iris	4	3	150
auto	7	3	392
wine	13	3	178
vehicle	18	4	846
glass	9	7	214
machine	7	8	209

测试精度是 SVM 中评价一个模型优劣的主要衡量标准。测试精度也是预测精度，即用已知类别的样本去测试这个模型。具体可以用以下式子来表示：$p=\frac{n}{N}$，$n\in N$，其中 n 是预测正确的样本数，N 是测试总样本数。支持向量个数的多少也对模型优劣有一定的影响，主要体现在存储空间和运算时间上。因此，我们将从这两个评判标准来进行我们的实验。

在实验中我们比较了 3 种不同的支持向量机模型的测试，从每个数据集中随机选取 3/4 的数据作为训练集，其余的数据作为测试集进行实验。在考虑等价性的情况下，我们不仅考虑了全精度的等价，而且考虑了变精度下的等价关系，分别有 90%，80%，70%的等价。10 次平均测试精度如下(括号中为支持向量个数)：SVM 表示原始的 SVM 模型测试；FSVM 表示模糊支持向量机模型测试；RSVM 表示粗糙支持向量机模型测试。下面是各个部分的实验结果及分析情况。

(1) 两类样本实验

表 2 是核函数为 polynomial 条件下两类的分类测试精度，3 种核函数对比结果将在后面的泛化实验中体现。从实验的结果可以看出：pima 和 heart 数据的测试精度明显得到

了提高，并且在测试精度相同的情况下，RSVM 的支持向量会相对较少(如 breast 数据)。这样，对于大数据来说，减少了存储要求，提高了运算速度。RSVM 一般情况比 FSVM 的测试情况要好。

(2)多类样本实验

提出粗糙支持向量机模型主要是考虑了数据样本之间的等价性关系，这同样适应于多类样本的分类。在多类实验中，我们分别用多类分类中的 1-v-1 和 1-v-a 方法进行了测试，核函数同样选取了 3 种核函数来进行实验。

在多类样本实验中，不论是 1-v-1 方法还是 1-v-a 方法，iris，auto，vehicle 3 种数据在 3 种核函数的测试下，我们所提的粗糙支持向量机(RSVM)模型相对于原始 SVM 和 FSVM 的测试效果都有了明显的提高；machine 数据在 polynomial 和 RBF 核函数下、glass 数据在 RBF 测试下，RSVM 的测试结果也较 SVM 和 FSVM 有所提高。

(3)平均测试精度

我们将测试数据样本所得的测试精度进行平均，结果如图 1—图 3 所示。

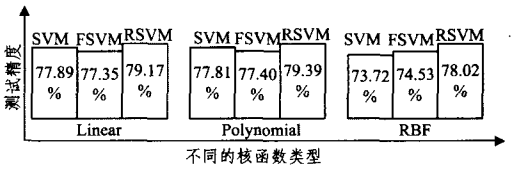

图 1　两类实验对比结果

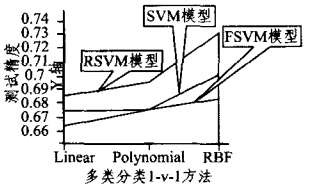

图 2　多类实验对比结果(1-v-1)

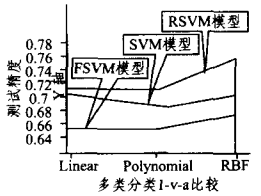

图 3　多类实验对比结果(1-v-a)

从上面各图可以看出，将数据结果进行平均以后，不论是两类实验结果还是多类实验结果，粗糙支持向量机(RSVM)的测试精度始终比 SVM 和 FSVM 的测试精度高。

表 2　核函数为 polynomial 时的测试精度(两类)

	Polynomial					
Data	SVM	FSVM	全精度等价性(100%)	变精度等价性(90%)	变精度等价性(80%)	变精度等价性(70%)
breast	0.9474(26)	0.9649(25)	RSVM:0.9649(24)	RSVM:0.9474(25)	RSVM:0.9474(25)	RSVM:0.9474(24)
pima	0.7865(302)	0.7760(292)	RSVM:0.7865(300)	RSVM:0.7865(304)	RSVM:0.7917(304)	RSVM:0.7917(305)
heart	0.7432(89)	0.7432(87)	RSVM:0.7432(89)	RSVM:0.7838(88)	RSVM:0.7432(87)	RSVM:0.7432(90)
bupa	0.6353(198)	0.6118(175)	RSVM:0.6353(198)	RSVM:0.6353(197)	RSVM:0.6353(196)	RSVM:0.6353(196)

结束语　支持向量机能否很好地分类，一个重要方面就是其模型能否有效提取数据的分布与关联信息。粗糙集理论方法的成功应用使人们对数据的等价类信息给予关注，很多研究者试图把粗糙集理论方法与支持向量机结合起来。但所用的结合方法都是分别使用粗糙集和支持向量机处理数据。本文首次把粗糙集信息利用结合到支持向量机模型中，使得该模型有更强的信息提取和利用能力。所做的实验证明了我们所提的 RSVM 模型优于传统的 SVM 和 FSVM 模型，大大提高了测试精度。

参 考 文 献

[1] Cortes C, Vapnik V. Support Vector Networks. Machine learning, 1995, 20(3): 273-297

[2] Vapnik V N. Estimation of Dependencies Based on Empirical Data. Berlin: Springer-Verlag, 1982

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

[4] Huang H P, Liu Y H. Fuzzy support vector machines for pattern recognition and data mining. Int'l Journal of Fuzzy Systems, 2002, 4(3): 826-835

[5] Lee Ki-Young, Kim Dae-Won. Possibilistic support vector machines. Pattern Recognition, 2005, 38(8): 1325-1327

[6] Bo Liefeng, Jiao Licheng, Wang Ling. Working Set Selection Using Functional Gain for LS-SVM. IEEE Transactions on Neural Networks, 2007, 18(5): 1541-154

[7] Pawlak Z. Rough set. International Journal of Information and Computer Science, 1982, 11(5): 341-356

(上接第 207 页)

[6] Jarillo G, Succi G, Pedrycz W, et al. Analysis of Software Engineering Data Using Computational Intelligence Techniques // Proceedings of the 7th International Conference on Object-Oriented Information Systems. Calgary, AB, 2001

[7] Kim Kyung-Sup, Han Ingoo. The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. Expert Systems with Applications, 2001, 21(3): 147-156

[8] Changchiena S W, Linb Ming-Chin. Design and implementation of a case-based reasoning system for marketing plans, 2005, 28(1): 43-53

[9] Vapnik V N. The Nature of Statistical Learning Theory. Springer, 1995

[10] Min J H, Lee Young-Chan. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications, 2005, 28: 603-614

[11] Tam K Y, Kiang M Y. Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. Management Science, 1992, 38: 926-947

[12] Dutta S, Shekhar S. Bond Rating: A Non-Conservative Application of Neural Networks // Proceedings of the IEEE International Conference on Neural Networks. San Diego, New York: IEEE Press, 1989: 443-450