基于 Matlab的支持向量机工具箱

郭小荟[1,2] 马小平[1]
[1](中国矿业大学信息与电气工程学院 江苏 徐州 221008)
[2](徐州师范大学计算机科学与技术学院 江苏 徐州 221116)

摘 要 介绍了基于 MATLAB的支持向量机工具箱,详细说明了工具箱中用于支持向量分类和支持向量回归的函数。并通过两个具体的实例来说明利用 SVM工具箱进行分类和回归方面的方法。

关键词 Matlab 支持向量机工具箱 分类 回归

SUPPORT VECTOR MACHINES TOOLBOX IN MATLAB ENVIRONMENT

Guo Xiaohui[1,2] Ma Xiaoping[1]
[1](College of Information and Electrical Engineering, CUMT, Xuzhou 221008, Jiangsu, China)
[2](College of Computer Science and Technology, XZNU, Xuzhou 221116, Jiangsu, China)

Abstract Support vector machines(SVM) toolbox in MATLAB environment is briefly introduced, and an extensively overview of the entire collection of toolbox functions used to support vector classification and support vector regression is given. And two examples are presented to illustrate how to solve classification and regression problems with the SVM toolbox.

Keywords Matlab Support vector machines Classification Regression

0 引 言

MATLAB已经成为国际上最流行的科学与工程计算的软件工具,现在的 MATLAB已经不仅仅是一个“矩阵实验室”了,它已经成为一种具有广泛应用前景的全新的计算机高级编程语言,有人称它为“第四代”计算机语言,MATLAB语言的功能越来越强大。它在国内外高校和研究部门正扮演着重要的角色,在科学运算、信号处理、自动控制与科学绘图等许多领域得到了广泛的应用。

V. Vapnik等人根据统计学习理论提出支持向量机(SVM)学习方法[1]。近年来受到国际学术界的广泛重视,并且已经广泛用于解决分类和回归问题。将支持向量机用于解决分类问题即支持向量回归(SVC),将支持向量机用于解决回归问题即支持向量回归(SVR)。Southampton大学 S.R.Gunn编写了 Matlab SVM Toolbox[2]。该工具箱运行在 MATLAB环境下,由许多用 m语言编写的脚本文件和函数组成,为 SVM技术的工程化、实用化提供了一个良好的平台。

本文以 MATLAB6.5为开发环境,详细讨论支持向量机工具箱及其相关函数,并通过两个具体的实例说明利用支持向量机工具箱进行分类和回归的方法。结果表明,SVM支持向量机工具箱可以很好地用于分类和回归的问题之中,对于支持向量机理论的推广应用具有很大的实用价值。

1 支持向量机工具箱函数及相关函数简介

SVM工具箱主要有两大功能:支持向量分类和支持向量回归。

1.1 支持向量分类的相关函数

1)支持向量机设计和训练函数 svc 该函数用来进行 SVM分类器的设计和对训练样本进行训练。该函数有四个参数,分别是训练样本的输入、训练样本的输出、核函数和惩罚因子。输出参数为支持向量个数、拉格朗日乘子及偏置量。其语法为:

[nsv alpha b0] = svc(X, Y, ker, C)

其中:

X是训练样本的输入;Y是训练样本的输出;ker是核函数;C是惩罚因子。

支持向量机工具箱支持如下几种核函数:

'linear', 'poly', 'rbf', 'sigmoid', 'spline', 'bspline', 'fourier', 'erbf', 'anova',除了'linear'和'spline'这两个核函数之外,其他的核函数还需要设定一些参数,如指定多项式核函数'poly'的阶数、径向基核函数'rbf'的宽度等,这些参数的设置在工具箱中的全局变量 p1, p2中设置。

nsv是 svc函数返回的训练样本中支持向量的个数;

alpha是 svc函数返回的每个训练样本对应的拉格朗日乘子,拉格朗日乘子不为零的向量即为支持向量。

b0是偏置量。

2)输出函数 svcoutput 该函数根据训练样本得到的最优分类面计算实际样本的输出。利用它还可以得到测试样本的分类情况,对最优分类面进行测试。

收稿日期:2006-01-20。江苏省自然科学基金项目(BK2003026)。
郭小荟,讲师,主研领域:软件工程,故障诊断,人工智能应用等。

3）支持向量机分类绘图函数 svcplot 该函数用来绘制出最优分类面，并标识出支持向量。

4）统计测试样本分类错误数量的函数 svcerror 该函数统计出利用已知的最优分类面对测试样本进行分类，发生错误分类的数目。

5）uiclass 该函数是一个具有简单图形用户界面的函数，可以用它方便地选择导入数据、选择核函数、显示最优分类面等功能。

1.2 支持向量回归的相关函数

1）支持向量机回归函数 svr 该函数根据训练样本设计最优回归函数，并找出支持向量。该函数有6个参数，分别是训练样本的输入、训练样本的输出、核函数、惩罚因子、损失函数和不敏感系数。输出参数为支持向量个数、拉格朗日乘子及偏置量。其语法为：

[nsv, beta, bias] = svr(X, Y, ker, C, loss, e)

X训练样本的输入	Y训练样本的输出
ker核函数	C惩罚因子
loss损失函数	e不敏感系数
nsv支持向量的个数	beta拉格朗日乘子
bias偏置量	

2）输出函数 svroutput 该函数利用 svr函数得到的最优回归函数来计算测试样本的输出，并返回。

3）svrplot 该函数用来绘制出最优回归函数曲线，并标识出支持向量。

4）svrerror 该函数用来显示根据最优回归函数计算的测试样本的拟合误差。

5）uiregress 该函数是一个具有简单图形用户界面的函数，可以用它方便地导入数据、选择损失函数、输入惩罚因子和ε不敏感系数、显示最优回归函数曲线。对于非线性回归，还有输入核函数宽度系数等功能。

1.3 SVM工具箱中的其它函数

1）数据归一化函数 svdatanorm 有的核函数对输入的数据有一定的要求。例如，当核函数是‘spline’，要求输入数据的上界为1，下界为0；当核函数是‘fourier’时，要求输入数据的上界为－pi/2，下界为 pi/2。这时，需要对输入数据进行归一化。

2）核函数计算函数 svkernel 该函数的主要功能是根据不同的核函数，进行相应的核函数的计算。

1.4 数据的导入方法

要利用 SVM工具箱进行样本分类或数据回归，必须准备训练样本和测试样本。对样本数据的获取，可以通过如下方式进行数据的获取。具体采用哪种方法，取决于数据的多少，数据文件的格式等。

用元素列表方式直接输入数据。

创建数据文件，通过 MATLAB提供的装载数据函数，从数据文件中读取。函数 load适合从 MAT文件、ASCII文件中读取数据；MATLAB I/O函数适合从其它应用中的数据文件中读取数据；

还可以通过数据输入向导（ ImportWizard）从文件或剪贴板中读取数据，单击 File菜单下的“ ImportData...”将出现“ ImportWizard”窗口，通过该窗口进行设置，该方法不适合从 M文件中读取数据。

2 SVM工具箱应用实例

2.1 支持向量机分类应用

利用 SVM工具箱进行数据样本的分类时，核函数的选择，惩罚因子的大小以及有关核函数的宽度参数对支持向量的个数和最优分类面的建立有很大影响。需要经过多次实验才能够确定使分类结果较好的参数。 SVM工具箱中的函数仅支持两分类问题。要解决多分类问题，可以通过组合多个二值子分类器来实现，具体的构造方法有一对一和一对多两种。

下面通过一个非线性分类的例子来说明 SVM支持向量分类的应用。在 MATLAB6.5中编写程序如下：

```
% a nonlinear separation example
load nlinesep; % nlinesep is a data file
ker='poly'; % kernel function
C=Inf; % chengfayinzi
[nsv, alpha, b0] = svc(trnX, trnY, ker, C); % design a classifier and obtain support vectors
svcplot(trnX, trnY, ker, alpha, b0); % draw the optimum separable plane
tstX=[1 2]; tstY=[1]; % test sample
predictedY= svcoutput(trnX, trnY, tstX, ker, alpha, bias); % output the separation result of test sample
err= svcerror(trnX, trnY, tstX, tstY, ker, alpha, b0)
```

在 matlab6.5中运行上述程序，就可以得到相应的结果。限于篇幅，这里不再列出。

上述程序中用到的数据样本如表1所示。其中，1到8号样本保存在数据文件 nlinesep中，9号样本做为测试样本。

表1 非线性可分数据表

序号	X		Y
1	1	1	－1
2	2	2	1
3	1	3	1
4	2	1	－1
5	2	2.5	1
6	3	2.5	－1
7	3	3	－1
8	1.5	1.5	1
9	1	2	1

如果用 uiclass对同样的数据文件 nlinesep中的数据进行分类，可以得到如图1所示的最优分类面。

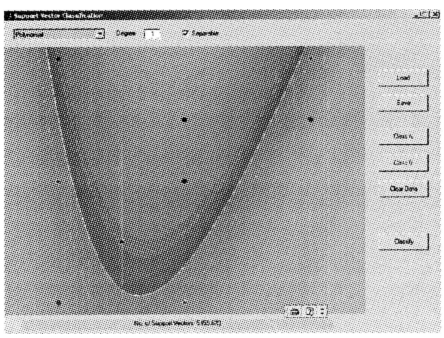

图1 最优分类面

2.2 支持向量回归应用

利用 SVM工具箱进行数据样本的回归时,不敏感系数、惩罚因子、核函数及其宽度参数对支持向量的个数和最优回归函数曲线的建立有很大影响。需要经过多次实验才能够确定使分类结果较好的参数。

下面通过一个非线性回归的例子来说明 SVM支持向量回归的应用。在 MATLAB6.5中编写程序对 sinc函数进行回归拟合。程序如下:

```
% a nonlinear regression example
load sinc; sinc is a data file
ker='rbf'; % kernel function
C=5; % upper bound
e=0.01; % insensitivity
loss='einsensitive'; % loss function
[nsv beta bias] = svr(trnX, trnY, ker, C, loss, e);
svrplot(trnX, trnY, ker, beta, bias, e);
tstX=0.1:tstY=sinc(tstX); % test sample
TstY=svroutput(trnX, tstX, ker, beta, bias); % output of the regression result of test sample
err=svrerror(trnX, tstX, tstY, ker, beta, bias, loss, e);
```

在 MATLAB6.5中运行上述程序,就可以得到相应的结果。最优回归函数曲线如图 2所示。

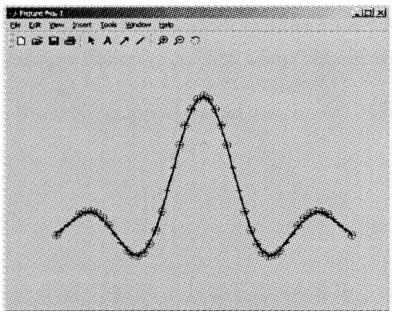

图 2　最优回归函数曲线

图中,标记为⊕的即为支持向量。从图 2可以得知模型数据与实际数据几乎完全逼近。

3 结　论

SVM在近十几年里已越来越受到人们的重视,如何结合实际问题实现分类和回归是支持向量机应用的一个重要方面。

S.R. Gunn编写的 Matlab SVM Toolbox则是一个实现支持向量机学习的有力工具。它充分利用了 MATLAB的优秀的数据可视化技术,它也具有开放性、可修改性、可扩展性,由于大部分的函数是用 m语言写的,阅读理解现有程序非常容易。在这个工具箱里有很好的说明文件和 DEMO程序,对如何使用该工具箱都做出详细的说明和示例,阅读完它们之后对于熟练使用该工具箱是没有多大困难的。如果能熟练使用该工具箱,并在此基础上针对具体问题进行二次开发,则分析设计基于支持向量机理论的应用系统就会"站在巨人的肩膀上"。

附:Matlab SVM Toolbox是免费软件,可以从下面网址自由下载使用:

http://www.isis.ecs.soton.ac.uk/resources/svminfo/

参 考 文 献

[1] Vlandimir N Vapnik.统计学习理论[M].许建华,张学工,译.北京:电子工业大学出版社,2004

[2] Steve R Gunn. Support Vector Machines for Classification and Regression[R]. Technical Report Southampton: University of Southampton, 1998: 1-28

(上接第 35页)

H_3;数据组:DG_1={手术病人表,编号,病人姓名,性别,年龄,手术原因,手术时间},DG_2={手术病人表,编号},DG_3={手术病人表};指派角色与安全级:$u_1 \rightarrow R_1$,$u_2 \rightarrow R_2$,$u_3 \rightarrow R_3$,$H(u_1)=H_3$,$H(u_2)=H_2$,$H(u_3)=H_3$,$H(u_4)=H_1$,$H(R_1)=H_3$,$H(R_2)=H_2$,$H(R_3)=H_3$,$H(DG_1)=H_3$,$H(DG_2)=H_2$,$H(DG_3)=H_3$;角色与数据组:$R_i \rightarrow DG_j$,$i=1,2$;$j=1,2,3$.

(2)分别执行下列查询并得到相应结果:

a 张医师(u_1)执行对数据组的查询。因为根据策略 2,$u_1 \rightarrow R_1$,$u_1 \in \{R_i \rightarrow DG_j\}$,$i=1,2$;$j=1,2,3$.所以 u_1可访问 DG_1,DG_2,DG_3.又因为 $H(u_1)=H_3 \geqslant H(R_1)=H_3 \geqslant H(DG_j)$,$i=1,2,3$.所以张医师可以访问 DG_1,DG_2,DG_3的全部数据。

b 李医师(u_2)执行对数据组的查询。根据策略 2,$u_2 \rightarrow R_2$,$u_2 \in \{R_i \rightarrow DG_j\}$,$i=1,2$;$j=1,2,3$.所以 u_2可以访问 DG_1,DG_2,DG_3.又因为 $H(u_2)=H_2$,$H(R_2)=H_2$,而 $H(DG_1)=H_3$,$H(DG_2)=H_2$,$H(DG_3)=H_1$。所以李医师只能访问 DG_2,DG_3。

c 王医师(u_3)执行对数据组的查询。根据策略 2,$u_3 \rightarrow R_3$,$u_3 \notin \{R_i \rightarrow DG_j\}$,$i=1,2$;$j=1,2,3$.所以王医师不能访问 DG_1,DG_2,DG_3.

d 赵医师(u_4)执行对数据组的查询。根据策略 1,$H(u_4)=H_1$,而 $H(DG_1)=H_3$,$H(DG_2)=H_2$,$H(DG_3)=H_1$,$H(u_4) \geqslant H(DG_3)$。所以赵医师只能访问 DG_3.

e 假设 u_5与 u_1具有共同访问对象 DG_1,DG_2,DG_3,但是,u_1仅负责 A公司的手术病人,u_5仅负责 B公司的手术病人,他们的查询经过 a后进入第二阶段,有选择的自主访问控制模块会根据他们的查询条件,经过选择和计算,清除既不是 A公司也不是 B公司的病员信息,最后分别显示满足条件的信息。如果 u_5企图用 u_1的查询条件了解 u_1范围的信息,仅由现有的数据库管理系统是不能控制的,而本文提出的两阶段的确认,可以方便地解决类似的访问控制问题。

4 结束语

对于数据仓库,存在许多影响安全的因素,本文策略适用于综合信息量很大,并且有安全要求的数据仓库,尤其对数据仓库中可变大小的数据组可实施灵活地访问控制,实用性很强。但查询速度对系统效率有所影响。接下来要对减小系统开销进行研究,并且在为数据仓库的访问控制制定新的策略时,会影响已有策略功能,因此需继续研究减少冗余策略的方法。

参 考 文 献

[1] Lindgreen R, Herschberg J. On the Validity of the Bell-LaPadula Model[J]. Computer & Security, 1994, 13: 317-338

[2] Sandhu R, Cope E, Feinstein H, Youman C. Role-based access control models[J]. IEEE Computer, 1996, 42(2): 38-47

[3] Min-A Jeong, Jung Ja Kim, Yonggwan Won. A Flexible Database Security System using Multiple Access Control Policies[J]. IEEE, 2003

[4] 徐兰芳,潘芸.数据仓库安全需求模型研究.华中科技大学学报,2005,33(7).