基于粗集和SVM的客户抵押贷款违约评估

王 波[1],刘勇奎[1],郝艳友[2]
WANG Bo[1],LIU Yong-kui[1],HAO Yan-you[2]

1.大连民族学院 计算机科学与工程学院,辽宁 大连 116605
2.中国建设银行 大连分行,辽宁 大连 116001
1.College of Computer Science and Engineering,Dalian Nationalities University,Dalian,Liaoning 116605,China
2.Dalian Branch of China Construction Bank,Dalian,Liaoning 116001,China
E-mail:wangb@dlnu.edu.cn

WANG Bo,LIU Yong-kui,HAO Yan-you.Defaults assessment of mortgage loan with RS and SVM.Computer Engineering and Applications,2008,44(9):229-231.

Abstract: Credit risk is the primary source of risk to financial institutions.Support Vector Machine is a new machine learning method based on the idea of VC dimension and Statistical Learning Theory.It is a good classifier to solve binary classification problem and the learning results possess stronger robustness.We use Grid-search method adjusts these penalty parameters to achieve better generalization performances in our application.In this paper the attribute reduction of rough set has been applied as preprocessor so that we can delete redundant attributes,then default prediction model of the housing mortgage loan is established by using SVM.Classification performance is better than other classification algorithm.
Key words: credit rating;Support Vector Machine;attribute reduction;grid-search

摘 要:信贷风险是金融机构风险的主要来源。支持向量机是基于VC维和统计学习理论理念的一种新的机器学习方法。它在解决两类问题时是一种较好的分类方法,同时学习结果模型有较强的稳定性。在实际应用中,采用Grid-search方法调整支持向量机的惩罚参数,达到了更好的推广能力和预测结果。采用粗集对数据集进行预处理,属性约简,删除了多余的属性,然后再用支持向量机进行分类建立了住房抵押贷款信用风险评估模型,并与其他算法进行了比较,取得了良好的分类效果。
关键词:信用评估;支持向量机;属性约简;Grid-search
文章编号:1002-8331(2008)09-0229-03 **文献标识码:**A **中图分类号:**TP391.4

1 导言

信用风险是金融机构面对的主要风险。信用风险一般定义为借贷者因违约而不偿还贷款造成的风险。信用风险模型的基础是违约概率[1,2]。住房抵押贷款是银行贷款的重要组成部分,回收违约客户贷款的本金和利息对银行来说也存在巨大的信用风险。信用评级系统为客户抵押贷款应用产生一个内部评级,它考虑多方面的数量,如客户的年龄、家庭收入、利率等。经过评级系统分析所有信用记录的信息,最后通过总结一系列的计算得出一个客户的信用积分。Fair,Isaac & Co.开发了一种数学方法,在信用记录里找出影响客户偿还贷款能力和意愿的属性(http://www.fairisaac.com)。该方法的问题显然是主观方面的预测,很难做出一致的评估。信用评分问题将预测违约概率转化为分类问题。最近的研究表明人工智能方法比传统的统计方法能达到更好的效果[3,4]。

目前基于支持向量机(SVM)的信用评估方面的研究,大多数采用UCI中德国、澳大利亚等国外的信用数据,采用国内信用数据的研究较少[5-7]。而国内银行数据和国外银行数据有较大差异,本文采集抽取国内某银行房贷的真实数据进行了研究,对国内银行的个人房贷信用评估进行了初步探索。

本文采用SVM来评估抵押贷款的客户信用风险,得到的预测模型有较好的泛化能力。SVM中核函数的参数γ和上界C控制SVM的推广能力。Grid-search方法通过搜索来寻找SVM核函数参数γ和上界C的最佳取值[8]。我们扩充了该方法来调整这些参数使得在应用中取得更好的效果。

结合粗糙集(RS)的属性约简和SVM的分类机理,提出了一种混合算法,应用RS的属性约简过程作为预处理器,对数据集进行属性约简,可以把冗余的属性从决策表中删去,但不损失任何有效信息;然后基于SVM进行分类建模和预测,这样可大大降低数据维数,降低SVM分类过程中的复杂度,但分类性能并不会降低,最后的实验结果表明该方法的有效性,提高

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60675008)。
作者简介:王波(1981-),男,助教,主要研究领域为数据挖掘,生物信息;刘勇奎(1961-),男,博士,教授,主要研究领域为计算机图形学和图像处理;郝艳友(1971-),男,博士,工程师,主要研究领域为数据仓库和数据挖掘。
收稿日期:2007-07-09 **修回日期:**2007-10-19

了预测准确率和预测速度。

2 粗糙集和支持向量

2.1 粗糙集理论的属性约简概念

1982 年,Z.Pawlak 发表了一篇经典论文——Rough Sets,粗糙集理论[9]从此诞生。去除冗余属性,进行属性的约简,降低属性维数,是粗糙集理论的重要应用之一。

设信息系统为 IS=(U,A)[10]。其中:U 是全域(对象的有限集,$U=\{x_1,x_2,\cdots,x_m\}$),A 是属性集(特征变量)。每个属性 $a\in A$。定义信息函数为 $f_a:U\to V_a$,其中 V_a 是 a 值的集,称为属性 a 的域。检验属性集是否独立可以看属性一个个被去掉后,是否会增加信息系统中基本集的数目。如果 $ind(A)=ind(A-a_i)$,那么属性 a_i 称为冗余的;否则,属性 a_i 对 A 来说是必不可少的。若属性集不是独立的,则能找到所有可能的最小属性子集,这样就得到了相同数目的整个属性集的基本集(约简),并找到所有不可缺少的属性集(核)。

属性约简是粗糙集理论处理信息系统的重要手段。属性约简,是通过删除知识库中多余的部分(等价关系)或多余基本属性集来保留知识库中的重要属性,从而删除知识库中不必要的知识保留真正有用的部分。本文将这一理论应用到分类的训练前阶段,用粗糙集的属性约简算法实现属性约简,然后结合 SVM 分类方法对数据进行分类。

2.2 支持向量机和参数选择

SVM 是由 Vapnik 提出的基于结构风险最小化原理的一种新的机器学习方法[11]。先考虑两类问题,设输入变量为 $x_i(i=1,\cdots,l)$,相应的类标签为 $y_i=\{-1,1\}$。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优化问题:

$$\min\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \tag{1}$$

$$\text{Subject to: } y_i(w\cdot x_i+b)\geqslant 1 \tag{2}$$

图 1 是一个线性可分的例子。实心点和空心点分别代表两类样本,H 为分类超平面,H_1、H_2 为过最近样本点的两个超平面。分类间隔(margin)是两个分类超平面 H_1 和 H_2 之间的垂直距离。

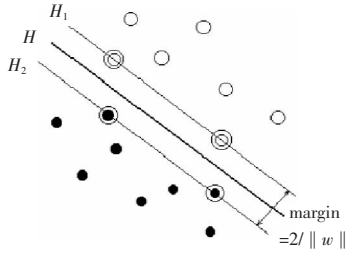

图 1 最优分类面

当分类问题是线性不可分的,这是上述的最优分类面是不存在的,无法使得所有样本都被正确分类。若仍采用上述思想,就必须"软化"对间隔的要求,允许有部分样本点被错分,引入松弛变量 ξ_i 和惩罚参数 C,将最优问题表示为:

$$\min\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{w}+\boldsymbol{C}\sum_{i=1}^{l}\xi_i \tag{3}$$

$$\text{Subject to: } y_i(w\cdot x_i+b)\geqslant 1-\xi_i \quad i=1,\cdots,l;\xi_i\geqslant 0 \tag{4}$$

$C\sum_{i=1}^{l}\xi_i$ 是控制错误分类样本的数量。C 越大,对错分的惩罚程度越大,则分类间隔越小,反之容许更多的被错分样本,则分类间隔越大[12]。

引入 Lagrange 乘子 α_i,可把问题转化为:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(x_i\cdot x_j)-\sum_{j=1}^{l}\alpha_j \tag{5}$$

$$\text{Subject to: } \sum_{i=1}^{l}y_i\alpha_i,0\leqslant\alpha_i\leqslant C,i=1,\cdots,l \tag{6}$$

此时的决策函数可以表示为:

$$f(x)=sign(w\cdot x+b)=sign\left(\sum_{i=1}^{l}\alpha_iy_i(x_i\cdot x)+b\right) \tag{7}$$

对于非线性情况,可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将数据映射到特征空间中是可分的。下面是三种最常用的核函数:

$$K_{poly}(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\gamma\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j+r)^d,\gamma>0 \tag{8}$$

$$K_{rbf}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\gamma\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2),\gamma>0 \tag{9}$$

$$K_{sig}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\tanh(\gamma\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j+r) \tag{10}$$

这里,γ,r和 d 是核参数。每一个核都有其优缺点[13]。通过核类型、核参数和 C 来控制支持向量机的泛化能力。

3 方法论

3.1 属性约简算法

目前属性约简的算法有利用辨识矩阵构造区分函数,基于属性重要性的启发式算法和复合系统的约简等[14]。几种算法各有其优缺点,本文采用的属性约简算法为基于属性重要性的约简算法。该算法通过所给决策表求出分辨矩阵以得到核,即通过函数计算属性的重要度,对数据集进行属性约简。从最重要的属性开始取,直到取出的属性构成一个约简为止,从而保证了求得的属性约简一定是最小约简。该算法的优点是简单直观,能够处理较大的数据集合。

具体的算法实现如下:

输入:一个决策表 $S(U,R)$,$R=C\cup D$,C 为条件属性集,D 为决策属性集;

输出:决策表 S 的决策规则。

步骤 1 消去重复的行,在二维数据表中,行表示不同的对象,当两个对象的属性相同,就消去其中的一行;

步骤 2 当两列的属性相同,就消去其中一列;

步骤 3 生成 S 的可分辨矩阵 $\boldsymbol{M}$:

```
FOR(i=1;i<=n;i++)
    FOR(j=1;j<i;j++)
        M=(C_ij);
```

即 $\boldsymbol{M}$ 为一个主对角线上无值的下三角矩阵;

步骤 4 由 $\boldsymbol{M}$ 计算 C 相对于 D 的核 $Cored(C)$;

设 $C_0=Cored(C)$;

步骤 5 令 $B=C_0$,且核以外的各条件属性为 AR,则 $AR=C-B$;对每一属性 $a_i\in AR$,计算其属性的重要性 $f(a_i)$,并按其值对 AR 进行降序排序;

步骤 6 分别计算 $r_B(D)$,$r_c(D)$;

步骤 7 WHILE($r_B(D)!=r_c(D)$) AND (AR! =ϕ)

$\{f(a_k)=\mathrm{MAX}(f(a_i))\}$;

$B=B\cup\{a_k\}$;$AR=AR-\{a_k\}$;}

步骤 8　最后 B 就是 C 相对于 D 的一个约简。

其中预处理过程设定为已经完成，已经得到了一个决策表,并且表的第一列定为决策属性,并且表中的属性值都已转化为数字形式。

3.2　SMO 算法

SMO 算法是 SVM 中较常用的训练算法,下面是算法的简单描述:

(1)遍历整个训练集,寻找违反 KKT 条件的拉格朗日乘子 a_1。若找到 a_1,转第 2 步;若找不到,则遍历界内支持向量对应的训练点,若找到 a_1,转第 2 步。在遍历"整个训练集"和"界内的支持向量对应的训练点"之间切换,直到整个训练集都满足条件为止。

(2)从非界集寻找拉格朗日乘子 a_2,取$|E_1-E_2|$值最大的对应点为 a_2。若两样本相同,则丢弃 a_2转第 3 步。否则为 a_2计算 L 和 H。若 $L=H$,则优化没有改进,丢弃 a_2转第 3 步。否则计算 η 值,若值为负计算新的 a_2,否则计算目标函数在 L 和 H 点的值,取较大值的点作为新的 a_2。$|a_2^{new}-a_2^{old}|$小于 ε 值,则丢弃 a_2,转第 3 步,否则转第 4 步。

(3)随机遍历非界集合直到找到在第 2 步中有优化改进的 a_2。若没有则随机遍历整个训练集直到找到在第 2 步中有优化改进的 a_2。若在这两个遍历中都没有找到 a_2,那么就舍弃 a_1的值返回第 1 步重新寻找新的违反 KKT 条件的 a_1。

(4)计算新的 a_1的值。更新阈值 b、误差 E_i和保存新的 a_1、a_2值。返回第 1 步。

3.3　寻找最优参数

抵押贷款风险评估是一个两类任务。准确率对于两类分类问题是一种典型的性能指标。SVM 有多个性能指标,SVM 参数最优选取也是一个多目标规划问题。用 RBF 核函数映射原始空间到高维特征空间,且要用到惩罚因子 C。

Grid-search 方法使用交叉确认 C 和 γ。每一个基本对(C,γ)都被尝试,然后选择交叉确认中正确率最好的。虽然 Grid-search 寻找最好参数计算时间比避免搜索全部近似或启发式方法多一点,但寻找参数要可靠。而且由于只有两个参数,每对(C,γ)都是相互相对独立的,所以 Grid-search 可以很容易地并行。Grid-search 方法在文献[8]中只被用于训练数据集中,将其扩展用在训练集和测试集中。在训练集和测试集中都寻找最优参数。

过程描述:将数据转换为程序的格式;将数据集随机分为训练集和测试集;考虑 RBF 核函数,给出 C 和 γ 的范围;使用这 3 个参数来训练整个训练集和预测测试集;在实验结果中找出最优参数。

3.4　粗集和支持向量机的结合

结合粗糙集的属性约简和支持向量机的分类机理,提出了一种混合算法(称为 R_SVM),应用粗糙集理论的基于属性重要性约简的算法对数据集进行属性约简,把冗余的属性从决策表中删去,但不损失任何有效信息;然后基于支持向量机的序列最小最优化算法(SMO)算法进行分类建模和预测,并用 Grid-search 方法选择最优参数 C 和 γ,然后对测试集进行预测。

4　实验及分析

数据集取自国内某商业银行实际的住房抵押贷款数据。这个数据集中包括从 1998 年的 1 月到 2004 年 12 月间的数据,从中抽取了 4 000 个客户样本进行研究。把客户定义为两类:"好"和"坏"客户。"坏"客户指借贷者至少有一次违约(超过 3 个月没有按期偿还分期付款)。"好"客户指借贷者按时偿还贷款。该数据集是指已经得到抵押贷款的客户的样本。其中信用为好的有 3 520 个样本,信用为差的有 480 个样本。本文对银行的数据进行了一些列初始化处理。将数据均分为两个子数据集,一半为训练集,另一半为测试集。训练集包含 2 000 样本点,其中 241 个"坏"客户和 1 759 个"好"客户。测试集包含 2 000 样本点,其中 240 个"坏"客户和 1 760 个"好"客户。"好"客户被标为"1","坏"客户被标为"-1"。

所有的客户是已经申请了抵押贷款的银行客户。数据集中包括客户自然信息、社会信息和贷款申请表中填写的其他信息。数据集中总结可用的 21 个应用属性为:年龄,性别,民族,婚姻状况,教育程度,职业,所在行业,工作时间,家庭收入,房屋单价,房屋面积,房屋总价,月供还款,首付比例,合同金额,贷款期限,户籍,有无职业资格,联系方式,是否本地人,贷款余额。

实验步骤:

(1)直接用 SVM(采用 SMO 算法)对数据集进行分类训练预测,然后用 Grid-search 方法选择最优参数 C 和 γ,并在相同的训练集和测试集上分别用 C4.5 和 BP 算法进行训练预测。

(2)用 R_SVM 对数据集进行分类训练预测,即先用粗集进行属性约简，然后用 SVM 对数据集进行分类训练预测，并用 Grid-search 方法选择最优参数 C 和 γ。

采用 3.1 章节的属性重要性约简算法，进行属性约简,结果属性由原来的 21 个属性约简为 16，约简掉的属性为 5,户籍,民族,有无职业资格,联系方式,是否本地人。对于属性约简的结果也是与现实相符合的,户籍所在地字段中数据分布很不均匀,属于本市的占大多数;是否本地人字段的区分度不高,由于在填写贷款申请时该项不是必填字段,缺省取值为 1—表示是本地人,且大部分人属于本地人;有无职业资格字段,该项不是必填字段,缺省取值为没有执业资格,而且很多申请人并不从事技术工作,所以没有职业资格;民族,大部分人民族是汉族;联系方式被离散化后,表示的意义不明显。故上面分析的字段与根据属性重要性分析的结果是一样的,证明了属性约简的正确性。

在 3 台 PC 服务器上同时并行作实验。Grid-search 中给出 C 和 γ 的范围为:10~80,2^{-10}~2^{10}。最后寻找的最优参数为 C=36,$\gamma=2^{-4}$。

个人信用评估的准确度对于降低银行贷款风险起着非常重要的作用。本文采用粗集和 SVM 结合的算法,对个人信用评估进行了研究。通过实际的银行数据的试验结果见表 1,分析

表 1　最终实验结果

分类算法	正确识别样本	错误识别样本	预测总正确率	属性数
Classic SVM	1 717	283	85.85%	21
BP	1 648	352	82.40%	21
C4.5	1 661	339	83.05%	21
R_SVM	1 764	236	88.20%	16

(下转 248 页)