基于 ODR-ADASYN-SVM 的极端金融风险预警研究①

林　宇[1]，黄　迅[1]，淳伟德[1]，黄登仕[2]
（1. 成都理工大学商学院，成都 610059；2. 西南交通大学经济管理学院，成都 610031）

摘要： 针对合成少数类过采样（synthetic minority over-sampling technique，SMOTE）方法在提升支持向量机（support vector machine，SVM）的非均衡样本学习能力中出现的过拟合（over fitting），引入自适应合成抽样方法（adaptive synthetic sampling approach，ADASYN）和逐级优化递减欠采样方法（optimization of decreasing reduction，ODR）分别克服 SMOTE 在生成新样本中的盲目性和在处理对象上的局限性，进而与 SVM 相结合，构造出改进 SVM，即 ODR-ADASYN-SVM 模型来预测中国极端金融风险；最后运用 T 检验对各模型预测精度的差异性进行显著性检验以及对各模型的预测稳定性进行评价. 实证结果表明，ODR-ADASYN-SVM 模型不仅能够显著地提升 SVM 的非均衡样本学习能力，同时也能够有效地克服 SMOTE 的过拟合，从而展示出优越的极端金融风险预测性能.

关键词： ODR；ADASYN；支持向量机；极端金融风险；预警模型
中图分类号： F832.5　　**文献标识码：** A　　**文章编号：** 1007－9807(2016)05－0087－15

0　引　言

近年来爆发的次贷危机、欧债危机等金融风险危机事件，不仅给各国经济带来了沉重打击，也使金融风险预警面临着更为严峻的挑战. 因此，建立科学的风险预警模型，准确地识别并有效地防范和控制风险，以维护金融经济安全，促进经济社会和谐稳定，既是新形势下金融经济管理部门面临的重要任务，也是学术界关注的热点问题.

然而，无论是对金融经济管理部门还是对投资者来说，都更关注金融市场极端下滑（extreme downside）所引发的极端金融风险（extremely financial risk）[1]. 这是因为，极端金融风险尽管发生概率极小，但一旦发生，就不仅会给投资者带来巨大损失，或是灭顶之灾，甚至还可能引起社会动荡、政权解体等严重后果. 因此，探讨极端金融风险的有效预警方法具有至关重要的现实意义.

在这里需要指出的是，自从 1997 年东南亚金融危机发生以来，极端金融风险预警就一直都是学术界关注的焦点与热点. 学者们已运用概率比（probit）回归、横截面回归（STV）、信号法（KLR 模型）、逻辑（Logit）回归、神经网络（neural network，NN）等模型对金融风险危机预警展开了研究[2-6]，虽然在预测极端金融风险上都取得了较好的效果，但却存在前提条件过于苛刻、过学习、欠学习以及局部极小等问题[7]. 而 SVM[8] 运用计算机智能技术自动确定模型结构，不仅能有效克服上述模型所存在的问题，而且也具有优越的智能学习与智能预测能力，因而被广泛运用于风险预警中且取得了良好的研究效果[9]. 然而，SVM 的良好预测性能却需要依赖数量相同的不同类别样本进行学习才能获得，事实上，现实生活中大量存在的样本却往往属于非均衡的[10]，如金融市场

① 收稿日期：2013－06－07；修订日期：2015－07－13.
基金项目：国家自然科学基金资助项目（71171025）；国家社会科学基金资助项目（12BGL024）；四川省软科学研究计划资助项目（2014ZR0093）；成都理工大学“金融与投资”优秀创新团队计划资助项目（KYTD201303）.
作者简介：林　宇（1973—），男，四川仪陇人，博士，教授. Email：linyuphd@126.com

中的极端与非极端金融危机事件所分别代表的少数和多数类样本就正好构成一类典型的非均衡样本. 如果仍然运用 SVM 对这样一类非均衡样本进行学习,就很可能会造成 SVM 的分类超平面(separating hyperplane)向极端金融危机事件所代表的少数类样本偏移,从而导致该类样本被大量错分,最终降低 SVM 的预测性能[11]. 由此可见,非均衡样本使得 SVM 学习面临严峻的挑战,因而就必须改进 SVM 模型,以提升 SVM 的非均衡样本学习能力.

尽管学者们已研究出多种方法,尤其是研究出 SMOTE 方法[12]来提升分类模型的非均衡样本学习能力. 但在 SMOTE 方法中,每个少数类样本都盲目地生成相同数量的人造样本,却忽略了其邻近样本的分布特点,因而容易造成少数类样本的相互重叠. 同时,SMOTE 也仅仅处理少数类样本,无法实现对多数类样本的处理. 因此,SMOTE 方法在生成少数类样本中的盲目性以及在处理对象上的局限性,使得分类模型不仅容易错分少数类样本,而且也无法有效删除多数类样本中的噪声信息,从而出现不能准确预测少数类样本的过拟合问题[13]. 而 ADASYN[14]却是专门针对 SMOTE 的盲目性而设计的解决方法,它利用少数类样本的密度分布来计算少数类样本生成的数量,使学习困难的少数类样本生成更多的人造样本,以增强分类模型的学习能力. 同时,ODR[15]利用 K 最邻近规则(K-nearest neighbor,KNN)来计算多数类样本对邻近样本的影响程度,进而依次删除对分类模型的学习能力有负面影响或者影响不大的多数类样本,以实现对多数类样本有目的的筛选,从而有效克服 SMOTE 在处理对象上存在的局限性. 由此可见,将 ADASYN 和 ODR 相结合,既能有效地克服 SMOTE 在生成少数类样本中的盲目性,又能有效地克服 SMOTE 在处理对象上的局限性,因而在提升非均衡样本学习能力上具有十分显著的优势. 但令人遗憾的是,就本文所掌握的文献而言,尚未发现有学者将 ADASYN 与 ODR 相结合以提升分类模型的非均衡样本学习能力.

值得注意的是,中国金融市场成立时间较短,缺乏应对金融风险危机的成熟经验,在极端金融风险的控制与防范上还存在诸多薄弱性[16-17];与此同时,随着经济全球化的深入发展,中国金融市场与国外金融市场的联系日益密切,因而极易受到国外极端金融风险的传染与冲击,从而承受着更为艰巨的风险考验,面临着更为严峻的风险挑战[18-19]. 那么,能否将 ODR、ADASYN 与 SVM 相结合,构建出适合中国金融市场的极端金融风险预警方法? 如果能,那么在极端金融风险的预测上,构建的预警模型又是否较 SVM、SMOTE-SVM、ODR-SVM 以及 ADASYN-SVM 预警模型具有更为明显的优势呢?

基于以上分析,本文以沪深 300 指数(China securities index 300, CSI300)为研究对象,既引入 ADASYN 来克服 SMOTE 在生成新样本中的盲目性. 又针对 SMOTE 在处理对象上的局限性,引入 ODR 来删除非极端金融风险样本中的噪声信息;并与 SVM 相结合,提出改进 SVM,即 ODR-ADASYN-SVM 预警模型来预测中国极端金融风险. 进而运用性能评估指标对 SVM、SMOTE-SVM、ODR-SVM、ADASYN-SVM 以及 ODR-ADASYN-SVM 模型的预测精度进行评估,并运用 T 检验对各模型预测精度的差异性进行显著性检验并对各模型的预测稳定性进行评价,以充分展示出 ODR-ADASYN-SVM 模型在预测性能上的显著优越性,从而为金融经济管理部门及时控制与防范极端金融风险以及投资者制定合理的投资策略提供良好的决策借鉴.

目前,国内外已有部分学者运用 SVM 对金融市场风险预警展开了研究. Ahn 等[20]以韩国金融市场为研究对象,建立了 SVM 金融市场风险预警系统; Groth 和 Muntermann[21]又以文本信息为研究对象,运用支持向量机(SVM)、神经网络(NN)、贝叶斯(Bayes)等方法,探索了金融市场风险危机管理; 徐国祥和杨振建[22]将主成分分析方法(principle component analysis, PCA)、遗传算法(genetic algorithm, GA)和 SVM 相结合,构建了 PCA-GA-SVM 预警模型,并对沪深 300 指数和大盘股走势进行了预测研究; 李云飞和惠晓峰[23]运用 SVM 建立了股票投资价值分类模型,并与 BP 和 RBF 神经网络的预测性能进行了比较.

与本文所掌握的研究文献相比,本文的差异性显而易见. 1)尚未发现有文献集成 ODR 和 ADASYN 来克服 SMOTE 所存在的过拟合,尤其

是没有发现有文献将 SVM 与该技术相结合,对中国新兴金融市场(本文用 CSI300 作为中国新兴金融市场的代表)的极端金融风险展开预警研究; 2)除运用常用指标来评价非均衡样本分类模型的性能之外,本文尤其是运用 T 检验对模型预测精度的差异性进行显著性检验,以及对各模型的预测稳定性进行评价; 3)在提取能够显著刻画中国新兴金融市场极端金融风险的特征指标上,不仅运用统计假设检验和逐步判别分析法来提取市场内部特征指标,还运用 Copula 方法专门对市场外部特征指标进行了提取,从而使得提取出的特征指标能够更为全面而显著地刻画中国新兴金融市场的极端金融风险.

1　研究方法

1.1　极端金融风险 SVM 预警方法

就金融市场 i 而言,每个时刻都存在着发生极端金融风险与未发生极端金融风险两种情况,分别用状态指标变量 $y^{(i)} \in \{+1,-1\}$ 表示. 其中,"+1"代表发生极端金融风险,"-1"代表未发生极端金融风险. 而每个时刻还包含 n 维特征指标变量 $x_d^{(i)} = (x_1^{(i)},x_2^{(i)},\cdots,x_n^{(i)})$, $d = 1,2,\cdots,n$,它的每个分量都代表市场 i 的极端金融风险特征指标. 于是,特征指标变量与状态指标变量就构成样本点 $(x_d^{(i)},y^{(i)})$.

由于本文研究的是极端金融风险预警方法,即通过当前特征指标变量来预测出下一个时刻的状态指标变量,则由 $t-1$ 个不同时刻的样本点就构成了 1 个样本数据集 $(x_{k,d}^{(i)},y_h^{(i)})$, $k = 1,2,\cdots,t-1$,代表了 t 时刻前的某个时刻, $h = 2,3,\cdots,t$,表示对应每个 k 时刻的下 1 个时刻.

在构造出样本数据集的基础上,将样本数据集中 1 个时间长度为 m 的样本数据 $(x_{j,d}^{(i)},y_p^{(i)})$ 作为训练集, $j = 1,2,\cdots,m$; $p = 2,3,\cdots,m+1$, $m < t-1$. 另一部分数据 $(x_{k-j,d}^{(i)},y_{h-p}^{(i)})$ 作为测试集. 然后,运用 SVM 进行训练,就得到极端金融风险预警模型

$$y^{(i)} = \operatorname{sgn}(f(x)) \tag{1}$$

式中 $\operatorname{sgn}(f(x))$ 是符号函数; $f(x)$ 是以特征指标变量为自变量的决策函数. 当金融市场处于极端金融风险状态时, $y^{(i)} = \operatorname{sgn}(f(x)) = +1$; 否则,就为 -1.

运用 SVM 方法在训练集上寻找最优分类函数(即式(1)代表的极端金融风险预警模型),就需要将寻找最优分类函数转化为求解以下最优问题

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{p=2}^{m+1}\xi_p \tag{2}$$

$$\text{s.t. } y_p^{(i)}\left((\boldsymbol{w}^{\mathrm{T}}x_{j,d}^{(i)}) + b\right) + \xi_p \geqslant 1 \tag{3}$$

式中 $\boldsymbol{w}$ 是可调权值向量; b 是分类阈值(threshold); ξ_p 是非负的松弛变量(slack variable),其作用是适当地"软化"约束条件,允许样本误判的存在; C 是惩罚参数(penalty parameter),其作用是权衡分类间隔 $2/\|\boldsymbol{w}\|$ 和错划程度 $\sum_{p=2}^{m+1}\xi_p$,目的是使得 $2/\|\boldsymbol{w}\|$ 尽可能大而 $\sum_{p=2}^{m+1}\xi_p$ 尽可能小.

为了求解上述最优问题,就要利用拉格朗日乘子(Lagrange multiplier)法将上述最优问题转化为对偶(dual)问题

$$\max_{\alpha} -\frac{1}{2}\sum_{p=2}^{m+1}\sum_{q=2}^{m+1}\sum_{j=1}^{m}\sum_{s=1}^{m}y_p^{(i)}y_q^{(i)}\alpha_j\alpha_s \times \exp\left(-\gamma\left|x_{j,d}^{(i)} - x_{s,d}^{(i)}\right|^2\right) + \sum_{s=1}^{m}\alpha_s \tag{4}$$

$$\text{s.t. } \sum_{p=2}^{m+1}\sum_{j=1}^{m}y_p^{(i)}\alpha_j = 0 \tag{5}$$

$$0 \leqslant \alpha_j \leqslant C,\ j = 1,2,\cdots,m \tag{6}$$

其中 α 为拉格朗日乘子, $(x_{s,d}^{(i)},y_p^{(i)})$ 为 α 的一个正分量 α_s 所对应的样本点, $\exp(-\gamma|x_{j,d}^{(i)} - x^{(i)}|^2)$ 是引入的径向基核函数(radical basis function, RBF), γ 是 RBF 核函数的参数. 由于大量研究文献均已证实 RBF 核函数下 SVM 的预测性能较其余核函数下 SVM 的预测性能更为优异[24-26]. 因此,本文选用 RBF 作为 SVM 的核函数.

于是,通过求解上述对偶问题,就得到以下最优分类函数

$$f(x) = \operatorname{sgn}(\boldsymbol{w}^* x + b^*) = \operatorname{sgn}\left(\sum_{j=1}^{m}\sum_{p=2}^{m+1}\alpha_j^* y_p^{(i)}\exp(-\gamma|x_{j,d}^{(i)} - x^{(i)}|^2) + b^*\right) \tag{7}$$

其中

$$b^* = y_q^{(i)} - \sum_{j=1}^{m}\sum_{p=2}^{m+1}\sum_{s=1}^{m} y_p^{(i)}\alpha_j^* \exp(-\gamma |x_{j,d}^{(i)} - x_{s,d}^{(i)}|^2) \quad (8)$$

至此，极端金融风险预警的 SVM 方法就已构建完毕. 然而，正如前文所分析的那样，由极端与非极端金融风险样本所构成的非均衡训练样本会大大削弱 SVM 的学习能力，最终降低 SVM 对极端金融风险样本预测的精度. 因此，下文将重点探讨能够提升 SVM 的非均衡样本学习能力的有效方法，以降低数据的过度偏斜对 SVM 预测极端金融风险样本所造成的影响.

1.2 提升 SVM 预测性能的 ODR-ADASYN 方法

为克服 SMOTE 在生成新样本中的盲目性以及在处理对象上的局限性，本文运用 ODR 和 ADASYN 方法来改进 SMOTE，以提升 SVM 的极端金融风险预测性能.

假定训练样本集为 S，其中非极端金融风险样本集为 S_{maj}，极端金融风险样本集为 $S_{\min}$，x 属于 S_{maj} 中的样本. 定义样本 x 的关联集是指 S_{maj} 中除 x 以外的其他非极端金融风险样本的 K 个最邻近样本中包含 x 的样本集，简称为关联集 x. 则 ODR-ADASYN 的计算机实验步骤如下:

步骤 1 根据式(9)，计算出需要删除的非极端金融风险样本数量 Num

$$Num = (|S_{\text{maj}}| - |S_{\min}|)\alpha \quad (9)$$

其中 $\alpha \in [0,1]$，表示删除指定的非极端金融风险样本后期望得到的样本非均衡水平，若 $\alpha = 1$，则表示删除指定的非极端金融风险样本后，两类样本数量达到完全均衡;

步骤 2 计算出非极端金融风险样本集中每个样本的关联集;

步骤 3 就样本 x 而言，运用 KNN 算法对关联集 x 中的所有样本进行分类，能够被正确分类的个数记录为 $withx$;

步骤 4 从关联集 x 包含的所有样本的最邻近样本中将样本 x 删除，并将第 $K+1$ 个最邻近样本加入，再运用 KNN 算法对关联集 x 中的所有样本进行分类，能够被正确分类的个数记录为 $withoutx$;

步骤 5 根据式(10)，计算 $withx$ 与 $withoutx$ 的差值 $diff$

$$diff = withx - withoutx \quad (10)$$

并通过下式判断 $diff$ 与 0 的大小关系来定义样本 x 的属性

$$x \in \begin{cases} N & diff < 0 \\ B & diff = 0 \\ S & diff > 0 \end{cases} \quad (11)$$

其中 N 代表噪声样本，B 代表边界样本，S 代表安全样本;

步骤 6 根据步骤 3 至步骤 5，定义出 S_{maj} 中除样本 x 外的其余每个非极端金融风险样本的属性;

步骤 7 根据步骤 6 定义出的样本属性，保留 S_{maj} 中的边界样本且删除噪声样本，并针对安全样本作下一步处理;

步骤 8 对安全样本而言，需要计算它们与其最近极端金融风险类样本之间的欧式距离，用 d 表示，再按照 d 值从大到小的顺序依次将每个 d 所对应的每个安全样本进行排序，并从第 1 个安全样本开始删除，直到非极端金融风险样本的数量递减到指定的数量，即 $|S_{\text{maj}}| - Num$ 为止. 此时，剩余的非极端金融风险样本集记为 S_{newmaj}，且满足下式

$$|S_{\text{newmaj}}| = |S_{\text{maj}}| - Num \quad (12)$$

步骤 9 根据步骤 8 得到的新的非极端金融风险样本 S_{newmaj} 的数量 $|S_{\text{newmaj}}|$ 以及原极端金融风险样本 $S_{\min}$ 的数量 $|S_{\min}|$，计算出新生成的极端金融风险样本数量 Num_{new}

$$Num_{\text{new}} = |S_{\text{newmaj}}| - |S_{\min}| \quad (13)$$

从式(13)可知，当合成数量为 Num_{new} 的人造极端金融风险样本后，两类样本数量就达到了完全均衡;

步骤 10 对每一个极端金融风险样本 x_i 而言，从所有样本中找出其 K 个最邻近样本，并根据下式计算出最邻近样本中非极端金融风险样本的占比 r_i

$$r_i = \frac{\Delta_i}{K} \quad (14)$$

其中 $\Delta_i \in [0,1]$，指 K 个最邻近样本中属于非极端金融风险样本的个数;

步骤 11 根据下式归一化 r_i，得到分布函数 $\hat{r}_i$

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{|S_{\min}|} r_i} \quad (15)$$

且 $\hat{r}_i$ 应满足下式

$$\sum_{i=1}^{|S_{\min}|} \hat{r}_i = 1 \tag{16}$$

步骤 12　计算出每一个极端金融风险样本需要合成的人造样本个数 g_i

$$g_i = \hat{r}_i Num_{\text{new}} \tag{17}$$

步骤 13　对每一个极端金融风险样本 x_i 而言,都运用 SMOTE 方法合成 g_i 个极端金融风险样本,此时两类样本数量达到一致.

需要说明的是,步骤 1 至步骤 8 是 ODR 的实验步骤,步骤 9 至步骤 13 是 ADASYN 的实验步骤. 于是,就能够基于上述 ODR-ADASYN 的实验步骤平衡非均衡的训练样本,进而运用 SVM 对得到的均衡训练样本进行训练,最后再基于测试样本对通过训练得到的 SVM 模型进行性能测试与评价.

1.3　极端金融风险预警模型的性能评估指标

基于均衡样本构建预警模型时,通常以分类准确率为评估指标来评价模型性能. 然而,当样本非均衡时,该指标仅能刻画模型的整体预测性能,却无法有效地评估模型对于极端金融风险样本的预测性能[27]. 基于此,本文运用目前针对非均衡样本分类的常用评估指标——几何平均正确率 G_{mean}、少数类的 F_{measure} 和 AUC(area under ROC curve)值,对本文所提出的极端金融风险预警模型进行性能评估(以下用 G 和 F 分别代替 G_{mean} 和 F_{measure}). 3 种评估指标的构建过程如下:

假定 $|FS_{\min}|$ 和 $|FS_{\text{maj}}|$ 分别为将非极端金融风险样本错划为极端金融风险样本和将极端金融风险样本错划为非极端金融风险样本的数量,$|TS_{\min}|$ 和 $|TS_{\text{maj}}|$ 分别为极端和非极端金融风险样本被正确划分的数量,通常采用混淆矩阵(confusion natrix)来表示(见表 1).

表 1　二分类数据集的混淆矩阵

Table 1 Confusion matrix of data sets of binary classification

	判断为非极端金融风险样本	判断为极端金融风险样本
实际为非极端金融风险样本	$\|TS_{\text{maj}}\|$	$\|FS_{\min}\|$
实际为极端金融风险样本	$\|FS_{\text{maj}}\|$	$\|TS_{\min}\|$

根据混淆矩阵并通过下面 3 个式子就可以得到灵敏度 SE、特异度 SP 和极端金融风险样本查准率 P 3 类基本指标

$$SE = \frac{|TS_{\min}|}{|TS_{\min}| + |FS_{\text{maj}}|} \tag{18}$$

$$SP = \frac{|TS_{\text{maj}}|}{|TS_{\text{maj}}| + |FS_{\min}|} \tag{19}$$

$$P = \frac{|TS_{\min}|}{|TS_{\min}| + |FS_{\min}|} \tag{20}$$

于是,再根据上述 3 类基本指标并通过下面两式就得到几何平均正确率 G 以及少数类的 F

$$G = \sqrt{SE \times SP} \tag{21}$$

$$F = \frac{2 \times SE \times P}{SE + P} \tag{22}$$

其中几何平均正确率 G 综合考察了模型对于两类样本的预测性能,若 G 较大,则说明模型预测两类样本的精度都较高,反之亦然; 而少数类样本的 F 主要考察了模型对极端金融风险样本的预测性能,若 F 较大,就说明模型预测极端金融风险样本的性能较优越,反之亦然. 此外,AUC 代表受试者工作特征曲线(receiver operating characteristic, ROC)下方的面积. 面积越大,说明模型对两类样本的综合预测性能越优异.

2　实证结果与分析

2.1　样本选择

本文选取沪深 300 指数(CSI300) 2005 - 04 - 08 ~ 2012 - 12 - 31 的样本作为研究对象. 需要说明的是,之所以选择 CSI300 进行研究,原因在于 CSI300 包含了沪深股市 60% 左右的市值,被视为反映沪深两市整体走势的“晴雨表”,具有良好的市场代表性[22,28-29],而之所以选择从 CSI300 创建以来至 2012 年末这段时间为研究区间,是因为沪深股市在这段时间内经历了暴涨暴跌的全过程,从而使沪深股市整体走势的预测结果具有更强的说服力.

2.2　状态指标变量的确定

要确定出所选每个样本的状态指标变量,即确定出所选的每个样本究竟属于极端金融风险样本还是非极端金融风险样本,关键在于选择极端金融风险门槛值. 只有选择出门槛值,才能以该门槛值为标准,将超过门槛值的样本认

定为极端金融风险样本，未超过门槛值的样本认定为非极端金融风险样本，以确定出所有样本的状态指标变量，从而才能进一步开展预警模型的构建工作.

需要指出的是，目前还没有确定门槛值的最优数学理论模型. Hill [30] 尝试运用 Hill 图法来确定门槛值; Stelios 和 Dimitris [31] 则运用了超额均值函数(mean excess function，MEF) 图法来确定门槛值; 而 Neftci [32] 把 1.65σ 当作门槛值，超过 1.65σ 的值被作为极端金融风险样本; 另外，Du-Mouchel [33] 选择了 10% 左右的数据作为极端金融风险样本进行研究，取得了较好的研究效果. 因此，本文选取占总体样本 9% 的收益率最低的样本作为极端金融风险样本，即极值尾部. 由于根据极值理论(extreme value theory，EVT) ，对于充分高的门槛值，超过门槛值的样本都近似服从广义帕累托分布(generalized Pareto distribution，GPD) 簇，用下式表示

$$\lim_{k\to u}\sup_{0\leqslant y\leqslant x-u}|F_k(x)-G_{\xi,\beta}(x)|=0 \qquad (23)$$

其中 k 表示超过门槛值的极端金融风险样本的个数; $F_k(x)$ 为极值分布函数; $G_{\xi,\beta}(x)$ 为 GPD 簇分布函数. 因此，本文还对这些极值尾部样本用准极大似然估计方法(quasi maximum likelihood estimation，QMLE) 估计 GPD 簇分布函数的参数得到尾部 GPD 分布曲线与经验分布的拟合效果图(见图 1) . 从图 1 可以直观地看出，GPD 与经验分布具有较好的拟合效果，说明设定的极端金融风险样本门槛值是科学合理的. 此时，极端与非极端金融风险样本的非均衡比例达到 1∶10.

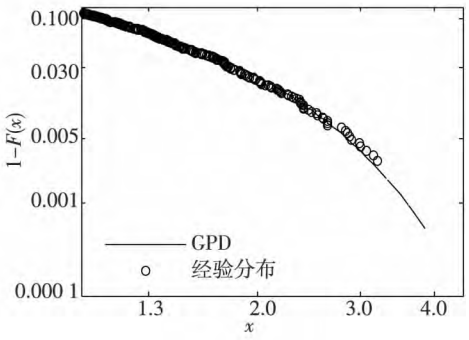

图 1 CSI 300 标准收益尾部经验分布与 GPD 分布拟合效果图

Fig. 1 Empirical distribution of standard return of CSI 300 and GPD plot

2.3 特征指标变量的选择与提取

金融风险管理的目标在于建立有效的风险预警系统. 而风险预警的重点又在于准确地提取出诱发金融市场爆发极端风险的特征指标. 但极端风险的爆发，不仅是市场自身风险的累积，也有外部极端风险危机的传导作用. 因此，本文通过借鉴相关文献 [22,34] 选择出由 8 项股指基本指标构成的内部特征指标. 同时，由于相关文献 [35-36] 将具有代表性的国际金融市场作为了中国金融市场的外部风险危机因素，又考虑到收益率是反映市场综合信息的重要指标，因此，本文将 8 个具有代表性的国际金融市场的股指日收益率作为了外部特征指标(见表 2) .

表 2 极端金融风险预警模型的特征指标

Table 2 Characteristic indexes of the early warning model for extremely financial risk

变量	外部特征指标名称	变量	内部特征指标名称
$X1$	恒生指数日收益率(HSI)	$X9$	开盘价
$X2$	韩国股指日收益率(KOSPI)	$X10$	收盘价
$X3$	台湾加权指数日收益率(TWII)	$X11$	最高价
$X4$	标准普尔指数日收益率(GSPC)	$X12$	最低价
$X5$	纳斯达克指数日收益率(NASDAQ)	$X13$	涨跌额
$X6$	日经 225 指数日收益率(Nikkei 225)	$X14$	涨跌幅
$X7$	道琼斯工业平均指数日收益率(DJIA)	$X15$	成交量
$X8$	英国金融时报指数日收益率(FTSE 100)	$X16$	成交额

注: 样本的特征指标数据来源于雅虎财经(http: //finance. yahoo. com) .

然而，最终用于预警的指标不仅是可以量化的，而且还需要存在显著差异，这样才可以起到预警效果. 同时，宋新平和丁永生 [26] 也指出，预警是个多指标的复杂系统，提取合适的特征指标可以

有效地提升模型的性能. 因此,为了从上述指标中进一步提取出能够显著刻画中国新兴金融市场极端风险的市场内外部特征指标,本文还将对表 2 中的国外股市指标和基本指标分别作如下处理.

一方面,为了提取出对中国极端风险的爆发具有显著传导作用的外部特征指标,本文运用 Clayton Copula 对 CSI 300 日收益率和表 2 中的前 8 项国外股指日收益率的下尾相关性进行分析. 这是因为下尾相关性能够反映两个市场同时暴跌的概率,这正好能够刻画外部极端金融风险对中国极端金融风险的影响程度[37]. 而 Clayton Copula 又恰好是专门刻画下尾相关性的 Copula 模型. 由此,计算出 CSI 300 与各国外股指间的下尾相依系数(见表 3).

表 3　CSI 300 与其余收益率之间的下尾相依系数

Table 3 Dependence coefficient of lower tail between different return rates

特征指标	HSI	KOSPI	TWII	GSPC	NASDAQ	DJIA	Nikkei 225	FTSE 100
CSI 300	0.238 7	0.112 5	0.112 3	$2.437\ 7\times10^{-4}$	$1.043\ 5\times10^{-4}$	$1.612\ 0\times10^{-4}$	0.054 1	0.009 9

从表 3 可以看出,CSI 300 与 HSI、KOSPI 和 TWII 的下尾相依系数大于 0.1,与 Nikkei 225 和 FTSE 100 的下尾相依系数小于 0.1,而与美国股市 3 大指数(GSPC,NASDAQ 和 DJIA)的下尾相依系数几乎为 0. 从而说明,CSI300 与 HSI、KOSPI 和 TWII 的下尾相关性较高,与 Nikkei 225 和 FTSE 100 的下尾相关性较低,而与美国股市 3 大指数(GSPC,NASDAQ 和 DJIA)几乎不存在下尾相关性. 基于此,本文选择与 CSI 300 下尾相关性较高的 HSI、KOSPI 和 TWII 的日收益率作为最终的外部特征指标.

另外,为了检验上述提出的 Clayton Copula 模型是否充分描述了 CSI 300 与国外股指之间的下尾相依关系,本文还将进行卡方检验,结果见表 4. 从表 4 可见,在 5% 的显著性水平下,各卡方检验值都通过了拟合优度检验,由此证明 CSI 300 与国外股指存在的相依结构科学合理. 另外,由于 CSI300 与 HSI、KOSPI 和 TWII 在 2005 - 04 - 08 ~ 2012 - 12 - 31 这段区间内的样本不匹配而出现缺失数据,因此,针对这些缺失数据,本文的处理方式是假如某个市场因节假日而没有进行交易,则其余市场的当日样本数据就被删除. 最终选择出 1 767 个样本进行研究.

另一方面,为提取出能够显著区分风险状态的内部特征指标,本文还借鉴相关文献[26,38]提出的统计假设检验方法和逐步判别分析法对表 1 中后 8 项基本指标进行提取. 首先对这 8 项基本指标进行正态性检验,进而对符合与不符合正态分布的基本指标分别做单样本 T 检验和 K-S 检验,从中提取出若干基本指标; 然后,运用逐步判别分析法进一步提取基本指标; 最终将这 8 项基本指标约简为开盘价、收盘价、成交量和成交额这 4 项基本指标. 并与通过 Clayton Copula 提取出的 HSI、KOSPI 和 TWII 日收益率这 3 项国外股市指标共同构成最终的能够显著刻画中国新兴金融市场极端金融风险的市场内外部特征指标.

表 4　卡方检验统计量

Table 4 Chi-square test statistics

特征指标	HSI	KOSPI	TWII	GSPC	NASDAQ	DJIA	Nikkei 225	FTSE 100
CSI 300	31.921 3**	15.718 7**	4.267 5**	0.092 5**	0.403 8**	0.003 9**	14.912 2**	0.285 5**

注: ** 代表各相依系数在 5% 的显著水平上显著.

2.4　最优 ODR-ADASYN-SVM 预警模型的确立

在将 SVM 的核函数设为 RBF 核函数、惩罚参数 C 设为 0.5、核函数参数 γ 设为 0.5 的基础上,本文采用 10 折交叉验证法(cross validation, CV)对 ODR 与 ADASYN 的最邻近参数 K 以及 ODR 中的参数 α 进行讨论,从而选择出性能最优的 ODR-ADASYN-SVM 模型. 需要强调的是,之所以将惩罚参数 C 值与核函数参数 γ 值都设为 0.5,是因为本文经过反复实验,最后确定模型在这两个值上的预测效果最为理想. 本文主要使用 Matlab 2011b 进行编程分析. 实验结果如图 2 和图 3 所示.

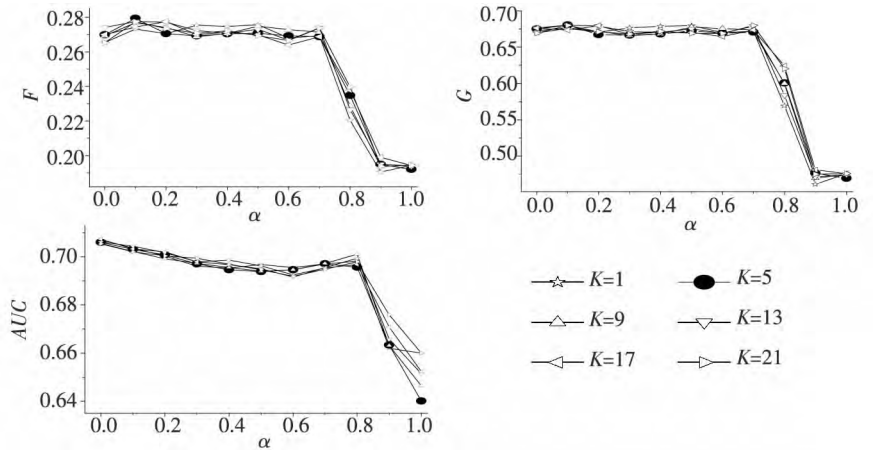

图 2 不同最邻近参数 K 下，参数 α 对模型预测精度的影响

Fig. 2 Influence of parameter alpha on prediction accuracy in different nearest neighbor K

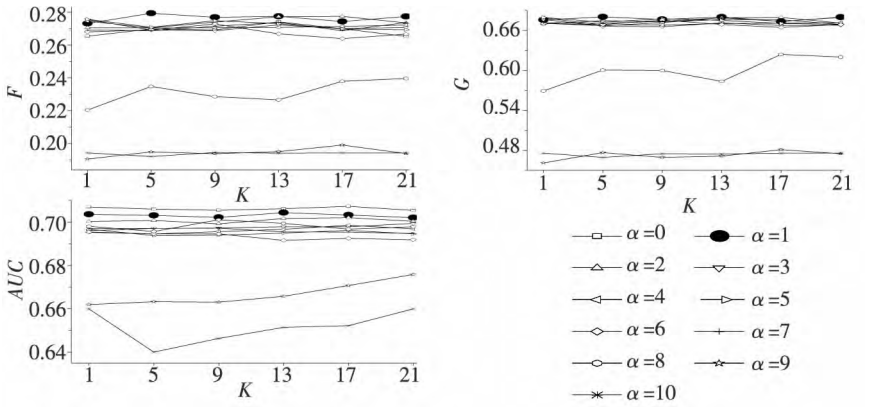

图 3 不同参数 α 下，最邻近参数 K 对模型预测精度的影响

Fig. 3 The influence of nearest neighbor K on prediction accuracy in different parameter alpha

从图 2 可以看出，不论 K 取何值，随着参数 α 的增大，特别是当 $\alpha > 0.7$ 时，模型的 F 和 G 值都同时呈现出明显的下降趋势，但当 $\alpha \leqslant 0.7$ 时，模型的 F 和 G 值却无明显的波动. 此外，当 $\alpha > 0.8$ 时，模型的 AUC 值同样也下降明显，但当 $\alpha \leqslant 0.8$ 时，模型的 AUC 值却无明显波动. 由此说明，尽管 ODR 能够有效地删除非极端金融风险样本的噪声信息，但过大的 α 会使得 ODR 过多地删除有效的非极端金融风险样本，最终降低模型的预测精度.

而从图 3 可以看出，在绝大部分的 α 值下，随着 K 的增大，模型的 G、F 和 AUC 值都无明显波动，从而说明 K 对模型的预测精度影响不大. 分析原因，主要在于当样本的非均衡水平较高时，极端金融风险样本的数量十分稀少，这使得每个样本的 K 个最邻近样本中属于极端金融风险类的样本较少且数量较为固定，并不会随着 K 的增大而增加，从而使得模型的预测精度较为稳定.

基于上述分析，本文认为，选择合适的 α，对于构建精度最优的 ODR-ADASYN-SVM 预警模型起着至关重要的作用. 因而，再次分析图 2 和图 3 可知，较其余绝大部分 α 下的 ODR-ADASYN-SVM，α 为 0.1 下的 ODR-ADASYN-SVM 具有最高的预测精度，特别是 α 为 0.1、K 为 5 下 ODR-ADASYN-SVM 模型的 AUC 值达到 0.703 1，G 值达到 0.680 4，F 值达到 0.279 3，相比其他 α 与 K 值下的 ODR-ADASYN-SVM 模型，其预测精度最为优异.

2.5 最优 ODR-ADASYN-SVM 模型与其余模型的预测性能对比评估

尽管通过上述实验，本文选择出最优的 ODR-ADASYN-SVM 预警模型，然而，该模型是否有效地提升了 SVM 的非均衡样本学习能力并成功地克服了 SMOTE 的过拟合，这是本文接下来需要

继续深入探讨的内容.

在将 SVM 的核函数设为 RBF 核函数、惩罚参数 C 设为 0.5、核函数参数 γ 设为 0.5、最邻近参数 K 设为 5、ODR-ADASYN-SVM 中 ODR 的参数 α 设为 0.1、ODR-SVM 中 ODR 的参数 α 设为 1 的基础上,本文仍然采用 10 折交叉验证法对 SVM、SMOTE-SVM、ODR-SVM、ADASYN-SVM 和 ODR-ADASYN-SVM 模型的预测性能进行对比研究. 实验结果如表 5 所示.

表 5　模型预测精度的对比结果

Table 5 Results of the comparison among prediction accuracy of each model

模型	F	G	AUC
SVM	NAN	0.000 0	0.568 0
SMOTE-SVM	0.275 8	0.675 5	0.704 4
ODR-SVM	0.192 1	0.469 5	0.640 1
ADASYN-SVM	0.277 0	0.676 1	0.706 1
ODR-ADASYN-SVM	0.279 3	0.680 4	0.703 5

注: 产生 NAN 的原因在于 SVM 将 CSI300 的所有极端金融风险样本判断错误,从而导致 $|TS_{\min}|$ 值为 0.

从表 5 可以看出,SVM 模型的预测性能最不理想,尤其是它的 G 值为 0 而 F 值为 NAN 的实证结果充分说明 SVM 模型将所有的极端金融风险样本都错误地判定为非极端金融风险样本,由此证明 SVM 模型无法有效地预测极端金融风险. 从表 5 还可以看出,ODR-ADASYN-SVM 模型在 G、F 值上都高于 SMOTE-SVM、ODR-SVM 和 ADASYN-SVM 模型,而在 AUC 值上略低于 SMOTE-SVM 和 ADASYN-SVM 模型,该结果能够说明,从预测精度的角度,ODR-ADASYN-SVM 模型的预测性能明显优于 ODR-SVM 模型且略优于 SMOTE-SVM 和 ADASYN-SVM 模型.

然而,仅根据各评估指标值就判断模型的性能还缺少类似于统计学检验所具有的科学性与客观性. 因此,本文还将对预测结果进行配对样本 T 检验,从而检验出各模型的预测性能差异是否显著. 检验结果如表 6 所示.

表 6　模型预测性能的配对样本 T 检验结果

Table 6 Results of paired T-test of prediction performance of each model

评估指标	模型	SMOTE-SVM	ODR-SVM	ADASYN-SVM	ODR-ADASYN-SVM
F	SVM	0.000 0***	0.000 0***	0.000 0***	0.000 0***
	SMOTE-SVM	–	0.001 0***	**0.826 0**	**0.557 0**
	ODR-SVM			0.001 0***	0.000 0***
	ADASYN-SVM				**0.629 0**
G	SVM	0.000 0***	0.000 0***	0.000 0***	0.000 0***
	SMOTE-SVM	–	0.000 0***	**0.938 0**	**0.630 0**
	ODR-SVM			0.000 0***	0.000 0***
	ADASYN-SVM				**0.595 0**
AUC	SVM	0.005 0***	**0.130 0**	0.005 0***	0.007 0***
	SMOTE-SVM	–	0.011 0**	**0.374 0**	**0.844 0**
	ODR-SVM			0.009 0***	0.009 0***
	ADASYN-SVM				**0.508 0**

注: *** 表示 $p<0.01$, ** 表示 $p<0.05$, 加粗值表示 $p>0.1$.

从表 6 可以看出,SVM 与其余模型在 G、F 和 AUC 值上的绝大部分配对样本 T 检验统计量在 1% 的显著水平下拒绝零假设(null hypothesis),即 SVM 与其余模型的预测性能差异显著. 由此表明,基于预测精度的角度,SVM 模型的预测性能显著低于其余模型,从而证明了 SMOTE、ODR、ADASYN 以及 ODR 与 ADASYN 相结合的 ODR-ADASYN 能够提升 SVM 的非均衡样本学习能力. 此外,从表 6 还可以看出,ODR-SVM 与 SMOTE-SVM、ADASYN-SVM、ODR-ADASYN-SVM 模型的预测性能也存在显著差异,表明 ODR-SVM 模型的预测性能显著低于 SMOTE-SVM、ADASYN-SVM 和 ODR-ADASYN-SVM 模型的预测性能,由此可见,仅仅运用 ODR 来删除非极端金融风险样

本中的噪声信息，以克服 SMOTE 在处理对象上存在的局限性，而不结合 ADASYN 来克服 SMOTE 在生成极端金融风险样本中的盲目性，并不能十分有效地提升 SVM 模型的预测性能. 令人遗憾的是，从表 6 中无法观察到 ODR-ADASYN-SVM 模型与 SMOTE-SVM、ADASYN-SVM 模型的预测性能存在显著差异，因而也就无法充分证明 ODR 与 ADASYN 相结合在克服 SMOTE 的过拟合问题上具有显著优势. 基于此，本文还将从预测稳定性的角度来继续探讨 ODR-ADASYN-SVM 模型较 SMOTE-SVM 和 ADASYN-SVM 模型的显著优越性.

为了对比 ODR-ADASYN-SVM 模型与 SMOTE-SVM、ADASYN-SVM 模型的预测稳定性，本文在设置与上述实验相同参数的基础上，通过修改极端金融风险样本的门槛值，将极端金融风险样本与非极端金融风险样本的比例分别设置为 1∶5、1∶10、1∶15、1∶20、1∶25、1∶30、1∶35、1∶40、1∶45、1∶50 这 10 个水平，进而对这 10 个水平下 3 类模型的预测精度变化情况进行分析，从而判定 3 类模型的预测稳定性，并最终挖掘出 3 类模型在预测性能上的显著差异. 实验结果如图 4 和表 7 所示.

从图 4 可以看出，在各种非均衡比例下，ADASYN-SVM 的 G 和 AUC 值几乎都高于 SMOTE-SVM，并且 ADASYN-SVM 的波动幅度也小于 SMOTE-SVM，而从表 7 也可以看出，不同非均衡比例下，ADASYN-SVM 的 G 和 AUC 值的均值都大于 SMOTE-SVM，并且 ADASYN-SVM 的 G 和 AUC 值的标准差也明显小于 SMOTE-SVM，证明 ADASYN-SVM 的波动幅度明显小于 SMOTE-SVM，从而说明在不同的非均衡比例下，ADASYN-SVM 模型对两类样本共同预测的稳定性明显优于 SMOTE-SVM 模型.

更为重要是，图 4 还展示出 ODR-ADASYN-SVM 的 G 和 AUC 值几乎都高于 ADASYN-SVM 和 SMOTE-SVM，并且 ODR-ADASYN-SVM 的波动幅度也最小，同时，从表 7 也可以看出，ODR-ADASYN-SVM 的 G 和 AUC 值的均值都大于 ADASYN-SVM 和 SMOTE-SVM，并且 ODR-ADASYN-SVM 的 G 和 AUC 值的标准差都低于 0.05，明显小于 ADASYN-SVM 和 SMOTE-SVM，表明 ODR-ADASYN-SVM 的波动幅度明显小于 ADASYN-SVM 和 SMOTE-SVM，从而说明在不同的非均衡比例下，ODR-ADASYN-SVM 模型在对两类样本的共同预测上具有最优的稳定性.

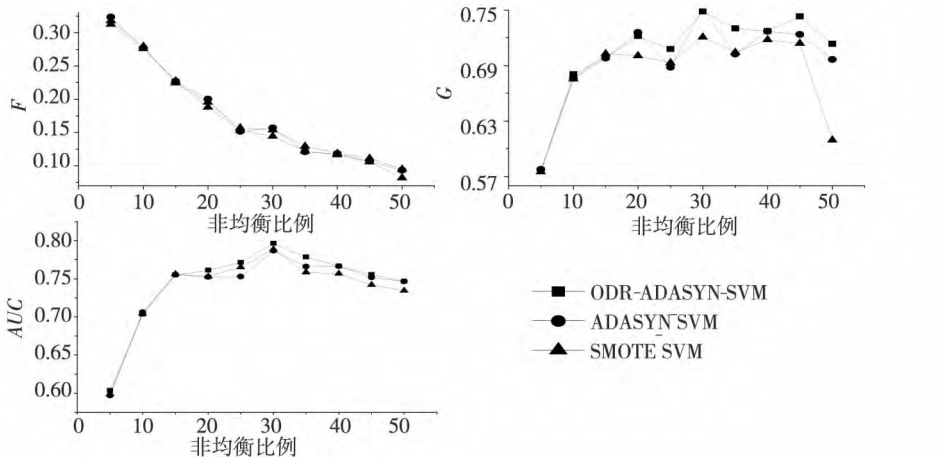

图 4 不同非均衡比例下 ODR-ADASYN-SVM、ADASYN-SVM 和 SMOTE-SVM 的 G、F 和 AUC 值比较

Fig. 4 Comparison of G、F and AUC among ODR-ADASYN-SVM, ADASYN-SVM and SMOTE-SVM in different imbalance ratios

表 7 不同非均衡比例下 3 类模型预测值的均值与标准差

Table 7 Means and standard deviations of three models' prediction accuracy in different imbalance ratios

模型	ODR-ADASYN-SVM	ADASYN-SVM	SMOTE-SVM
F	0.178 1(0.077 3)	0.177 4(0.077 5)	0.172 7(0.077 6)
G	0.704 7(0.047 5)	0.696 7(0.051 2)	0.681 3(0.055 6)
AUC	0.743 8(0.048 1)	0.738 0(0.050 3)	0.735 8(0.053 6)

注：不带括号外的值表示均值，带括号的值表示标准差.

然而,从图 4 却无法看出 3 类模型的 F 值在不同非均衡比例下的大小差异,也看不出各模型 F 值的波动幅度的差异. 尽管通过表 7 能够看出在不同非均衡比例下,ODR-ADASYN-SVM 的 F 值的均值大于 ADASYN-SVM 和 SMOTE-SVM,ADASYN-SVM 的 F 值的均值大于 SMOTE-SVM,以及 ODR-ADASYN-SVM 的 F 值的标准差小于 ADASYN-SVM 和 SMOTE-SVM,ADASYN-SVM 的 F 值的标准差小于 SMOTE-SVM,但大小差异都很不明显,因而无法判定 3 类模型在极端金融风险预测上稳定性的显著差异,也就更加无法展现 3 类模型的极端金融风险预测性能的显著差异. 为了展现出 3 类模型的极端金融风险预测性能的显著差异性,本文还将对在不同非均衡比例下 3 类模型的 F 值进行独立样本 T 检验. 实验结果见表 8.

表 8　不同非均衡比例下 3 类模型 F 值的独立样本 T 检验结果

Table 8 Results of the independent T-test of F of three models in different imbalance ratios

比例	10	15	20	25	30	35	40	45	50
5	**0.169 0** **0.161 0** (**0.161 0**)	0.001 0*** 0.001 0*** (0.001 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)
10	–	0.037 0** 0.093 0* (0.063 0***)	0.001 0*** 0.008 0*** (0.001 0***)	0.000 0*** 0.002 0*** (0.000 0***)	0.000 0*** 0.001 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)
15		–	0.094 0* 0.099 0* (0.038 0**)	0.003 0*** 0.001 0*** (0.001 0***)	0.001 0*** 0.011 0** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)
20			–	0.022 0** 0.009 0*** (0.046 0**)	0.006 0*** 0.002 0*** (0.002 0***)	0.000 0*** 0.001 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)	0.000 0*** 0.000 0*** (0.000 0***)
25				–	**0.834 0** **0.774 0** (**0.528 0**)	**0.101 0** **0.118 0** (0.054 0*)	0.045 0** 0.089 0* (0.047 0**)	0.012 0** 0.003 0*** (0.007 0***)	0.001 0*** 0.015 0** (0.001 0***)
30					–	**0.110 0** 0.079 0* (0.056 0*)	0.048 0** 0.075 0* (0.057 0*)	0.011 0** 0.006 0*** (0.004 0***)	0.001 0*** 0.006 0*** (0.001 0***)
35						–	**0.489 0** **0.858 0** (**0.706 0**)	**0.206 0** **0.269 0** (**0.220 0**)	0.016 0** 0.056 0* (0.021 0**)
40							–	**0.666 0** **0.492 0** (**0.494 0**)	**0.143 0** **0.112 0** (0.070 0*)
45								–	**0.238 0** **0.440 0** (**0.162 0**)

注: 未加括号的值表示 ODR-ADASYN-SVM 的 p 值,加下划线的值表示 ADASYN-SVM 的 p 值,而加有括号的值表示 SMOTE-SVM 的 p 值,* 表示 $p<0.1$,** 表示 $p<0.05$,*** 表示 $p<0.01$,加粗值表示 $p>0.1$.

分析表 8 可知,ODR-ADASYN-SVM 模型加粗的 p 值达到 9 个,带“***”的 p 值只有 28 个; ADASYN-SVM 模型加粗的 p 值有 8 个,带“***”的 p 值有 29 个; 而 SMOTE-SVM 模型加粗的 p 值只有 6 个,带“***”的 p 值却达到 31 个,说明在 10% 的显著水平下,ODR-ADASYN-SVM 模型的 F 值在所有非均衡比例下接受零假设 9 次,ADASYN-SVM 模型接受 8 次,SMOTE-SVM 模型仅接受 6 次,而在 1% 的显著水平下,ODR-ADASYN-SVM 模型的 F 值在所有非均衡比例下拒绝零假

设仅 28 次，ADASYN-SVM 拒绝 29 次，SMOTE-SVM 模型却拒绝 31 次，从这个角度来说，ODR-ADASYN-SVM 模型在各非均衡比例下的预测精度差异最不显著，其次是 ADASYN-SVM 模型，而最为显著的是 SMOTE-SVM 模型，从而说明，在不同的非均衡比例下，ODR-ADASYN-SVM 模型对于极端金融风险样本预测的稳定性优于 SMOTE-SVM 模型和 ADASYN-SVM 模型，而 ADASYN-SVM 模型又优于 SMOTE-SVM 模型.

基于图 4、表 7 与表 8 的分析，本文认为，尽管从预测精度分析，ODR-ADASYN-SVM 的预测性能并不显著优于 SMOTE-SVM 和 ADASYN-SVM，但从预测稳定性分析，ODR- ADASYN-SVM 的预测性能显著优于 SMOTE-SVM 和 ADASYN-SVM，而 ADASYN-SVM 却又显著优于 SMOTE-SVM，从而证明了 ADASYN 能够有效地克服 SMOTE 在生成极端金融风险样本中的盲目性，但仅仅运用 ADASYN 并不能最为有效地提升 SVM 模型的预测性能，只有将 ODR 与 ADASYN 相结合来克服 SMOTE 存在的过拟合，才能促使 SVM 模型的性能提升达到最为优异的效果.

综上所述，参数 K 对 ODR-ADASYN-SVM 模型的预测性能影响不大，而参数 α 却对 ODR-ADASYN-SVM 模型的预测性能有着较大的影响. 因此，通过对参数 α 的调整，就能够得到最优的 ODR-ADASYN-SVM 预警模型. 并且无论是从预测精度分析，该模型的预测性能显著优于 SVM 和 ODR-SVM 模型，还是从预测稳定性分析，该模型也显著优于 SMOTE-SVM 和 ADASYN-SVM 模型，从而充分证明了将 ODR 与 ADASYN 相结合不仅能够提升 SVM 的非均衡样本学习能力，而且比单独运用 ODR 和 ADASYN 来克服 SMOTE 的过拟合能够取得更加优异的效果. 从上述分析可知，由 ODR、ADASYN 和 SVM 结合成 ODR- ADASYN-SVM 模型，能够较好地预测中国极端金融风险，从而为金融经济管理部门和投资者应对与防范极端金融风险提供了有效的操作工具.

3 结束语

本文以 CSI300 指数为研究对象，运用 EVT 理论确定出极端与非极端金融风险样本，并运用统计假设检验和逐步判别分析法提取市场内部特征指标以及运用 Clayton Copula 方法提取市场外部特征指标，进而引入 ADASYN 来克服 SMOTE 在生成新样本中的盲目性; 针对 SMOTE 在处理对象上的局限性，还引入 ODR 来删除非极端金融风险样本中的噪声信息; 并与 SVM 相结合，构造了改进 SVM，即 ODR-ADASYN-SVM 模型，并对该模型的最邻近参数 K 和 ODR 的参数 α 进行讨论，进而运用最优的 ODR-ADASYN-SVM 模型来预测中国极端金融风险，然后再运用性能评估指标对 SVM、SMOTE-SVM、ODR-SVM、ADASYN-SVM 以及 ODR-ADASYN-SVM 模型的预测精度进行评估，并最终运用 T 检验对各模型预测精度的差异性进行显著性检验以及对各模型的预测稳定性进行评价. 实证结果表明，无论是从预测精度分析，ODR-ADASYN-SVM 模型的预测性能显著优于 SVM 和 ODR-SVM 模型，还是从预测稳定性分析，ODR-ADASYN- SVM 模型也显著优于 SMOTE-SVM 和 ADASYN-SVM 模型，从而说明 ODR-ADASYN 方法能够成功地克服 SMOTE 的过拟合，并有效地提升 SVM 的非均衡样本学习能力，同时，ODR-ADASYN-SVM 预警模型在中国极端金融风险预测上具有最为优越的预测性能.

本文的研究能够为金融经济管理部门和投资者应对与防范极端金融风险提供可操作性的应用工具与方法. 对于金融经济管理部门而言，能够运用 ODR-ADASYN-SVM 模型对未来一段时间内的金融市场极端风险进行准确预测，及时制定并实施应对金融风险危机的相关宏观经济政策，从而构建金融市场极端风险危机的“防火墙”，积极地化解与防范金融风险危机，以维护金融市场稳定，促进经济持续健康发展; 对于投资者而言，能够运用 ODR-ADASYN-SVM 模型提前捕获金融市场风险危机信号，进而即时调整金融资产投资策略，以优化金融资产投资组合，从而更为有效地管理金融资产，保证金融资产的保值甚至增值.

本文提出的风险预警方法，虽然丰富了市场主体对极端金融风险预警的研究手段，但仍需要指出的是，金融市场样本除存在极端与非极端两种分类状态外，也存在多分类状态，运用本文提出

的风险预警模型将无法进行多分类研究. 因此,如何改进 SVM 方法,并结合非均衡样本处理方法进行多分类状态下的极端金融风险预警研究,将是下一步的主要研究方向.

参 考 文 献:

[1] 魏　宇. 股票市场的极值风险测度及后验分析研究 [J]. 管理科学学报, 2008, 11(1): 78 - 88.
Wei Yu. EVT risk measures and its backtesting in stock markets [J]. Journal of Management Sciences in China, 2008, 11(1): 78 - 88. (in Chinese)

[2] Frankel J A, Rose A K. Currency crashes in emerging markets: An empirical treatment [J]. Journal of International Economics, 1996, 41(3/4): 351 - 366.

[3] Sachs J, Tornell A, Velasco A. Financial crises in emerging markets: The lessons from 1995 [J]. Brookings Papers on Economic Activity, 1996, 27(1): 147 - 215.

[4] Kaminsky G, Lizondo S, Reinhart C M. Leading Indicators of Currency Crises [R]. IMF Staff Paper, 1998, 45(1): 1 - 48.

[5] 陈守东, 杨　莹, 马　辉. 中国金融风险预警研究 [J]. 数量经济技术经济研究, 2006, (7): 36 - 48.
Chen Shoudong, Yang Ying, Ma Hui. The study of early-warning on Chinese financial risk [J]. The Journal of Quantitative & Technical Economics, 2006, (7): 36 - 48. (in Chinese)

[6] Kim T Y, Oh K J, Sohn I, et al. Usefulness of artificial neural networks for early warning system of economic crisis [J]. Expert Systems with Applications, 2004, 26(4): 583 - 590.

[7] 周　敏, 王新宇. 基于模糊优选和神经网络的企业财务危机预警 [J]. 管理科学学报, 2002, 5(3): 86 - 90.
Zhou Min, Wang Xinyu. Pre-warning systems of enterprise financial crisis based on fuzzy selection and artificial neural networks [J]. Journal of Management Sciences in China, 2002, 5(3): 86 - 90. (in Chinese)

[8] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 1995.

[9] Li H, Sun J. Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples-evidence from the Chinese hotel industry [J]. Tourism Management, 2012, 33(3): 622 - 634.

[10] Maratea A, Petrosino A, Mario M. Adjusted F-measure and kernel scaling for imbalanced data learning [J]. Information Sciences, 2014, 257(1): 331 - 341.

[11] Sundarkumar G G, Ravi V. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance [J]. Engineering Applications of Artificial Intelligence, 2015, 37(1): 368 - 377.

[12] Chawla N, Bowyer K, Hall L, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, (16): 321 - 357.

[13] Wang B X, Japkowicz N. Imbalanced data set learning with synthetic samples [C] // Proc. IRIS Machine Learning Workshop, 2004: 1 - 25.

[14] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning [C] // Proc. Int'l J. Conf. Neural Networks, 2008: 1322 - 1328.

[15] 陶新民, 童智靖, 刘　玉, 等. 基于 ODR 和 BSMOTE 结合的不均衡数据 SVM 分类算法 [J]. 控制与决策, 2011, 26(10): 1535 - 1541.
Tao Xinmin, Tong Zhijing, Liu Yu, et al. SVM classifier for unbalanced data based on combination of ODR and BSMOTE [J]. Control and Decision, 2011, 26(10): 1535 - 1541. (in Chinese)

[16] 戴文华, 夏　峰. 关于中国证券市场 20 年发展的基本分析与思考 [J]. 证券市场导报, 2014, (1): 4 - 11.
Dai Wenhua, Xia Feng. The basic analysis and reflections on the 20 years of development of the securities market [J]. Securities Market Herald, 2014, (1): 4 - 11. (in Chinese)

[17] 黄金波, 李仲飞, 姚海洋. 基于 CVaR 核估计量的风险管理 [J]. 管理科学学报, 2014, 17(3): 49 - 59.
Huang Jinbo, Li Zhongfei, Yao Haiyang. Risk management based on the CVaR kernel estimator [J]. Journal of Management Sciences in China, 2014, 17(3): 49 - 59. (in Chinese)

[18] 熊 熊，张 珂，周 欣. 国际市场对我国股票市场系统性风险的影响分析 [J]. 证券市场导报，2015，(1)：54－58.
Xiong Xiong，Zhang Ke，Zhou Xin. Analysis of impact of international market on China stock market systemic risk [J]. Securities Market Herald，2015，(1)：54－58. (in Chinese)

[19] 吴勇民，纪玉山，吕永刚. 技术进步与金融结构的协同演化研究——来自中国经验证据 [J]. 现代财经(天津财经大学学报)，2014，(7)：33－44.
Wu Yongmin，Ji Yushan，Lü Yonggang. Study on the co-evolution of technological progress and financial structure：The empirical evidence from China [J]. Modern Finance and Economics (Journal of Tianjin University of Finance and Economics)，2014，(7)：33－44. (in Chinese)

[20] Ahn J J，Oh K J，Kim D H. Usefulness of support vector machine to develop an early warning system for financial crisis [J]. Expert Systems with Applications，2011，38(4)：2966－2973.

[21] Groth S S，Muntermann J. An intraday market risk management approach based on textual analysis [J]. Decision Support Systems，2011，50(4)：680－691.

[22] 徐国祥，杨振建. PCA-GA-SVM 模型的构建及应用研究——沪深 300 指数预测精度实证分析 [J]. 数量经济技术经济研究，2011，(2)：135－147.
Xu Guoxiang，Yang Zhenjian. Research for construction and application of PCA-GA-SVM model [J]. The Journal of Quantitative & Technical Economics，2011，(2)：135－147. (in Chinese)

[23] 李云飞，惠晓峰. 基于支持向量机的股票投资价值分类模型研究 [J]. 中国软科学，2008，(1)：135－140.
Li Yunfei，Hui Xiaofeng. The classification model for stock investment value based on SVM [J]. China Soft Science Magazine，2008，(1)：135－140. (in Chinese)

[24] Farquad M A H，Indranil B. Preprocessing unbalanced data using support vector machine [J]. Decision Support Systems，2012，53(1)：226－233.

[25] Huang Z，Chen H C，Hsu C J，et al. Credit rating analysis with support vector machines and neural networks：A market comparative study [J]. Decision Support Systems，2004，37 (4)：543－558.

[26] 宋新平，丁永生. 基于最优支持向量机模型的经营失败预警研究 [J]. 管理科学，2008，21(1)：115－121.
Song Xinping，Ding Yongsheng. Study on business failure prediction based on an optimized support vector machine model [J]. Journal of Management Sciences，2008，21(1)：115－121. (in Chinese)

[27] 邹 鹏，李一军，郝媛媛. 基于代价敏感性学习的客户价值细分 [J]. 管理科学学报，2009，12(1)：48－56.
Zou Peng，Li Yijun，Hao Yuanyuan. Customer value segmentation based on cost-sensitive learning [J]. Journal of Management Sciences in China，2009，12(1)：48－56. (in Chinese)

[28] 魏 宇，赖晓东，余 江. 沪深 300 股指期货日内避险模型及效率研究 [J]. 管理科学学报，2013，16(3)：29－40.
Wei Yu，Lai Xiaodong，Yu Jiang. Intra-day hedging models and hedging effectiveness of CSI300 index futures [J]. Journal of Management Sciences in China，2013，16(3)：29－40. (in Chinese)

[29] 陈 莹，武志伟，王 杨. 沪深 300 指数衍生证券的多市场交易与价格发现 [J]. 管理科学学报，2014，17(12)：75－84.
Chen Ying，Wu Zhiwei，Wang Yang. Multi-market trading of HS300 index derivatives and price discovery of stock market index [J]. Journal of Management Sciences in China，2014，17(12)：75－84. (in Chinese)

[30] Hill B M. A simple general approach to inference about the tail of a distribution [J]. The Annals of Statistics，1975，3 (5)：1163－1174.

[31] Stelios D B，Dimitris A G. Estimation of Value-at-Risk by extreme value and conventional methods：A comparative evaluation of their predictive performance [J]. Int. Fin. Markets，Inst. and Money，2005，15 (3)：209－228.

[32] Neftci S. Value at risk calculations，extreme events and tail estimation [J]. The Journal of Derivations，2000(Spring)，7 (3)：23－37.

[33] DuMouchel W H. Estimating the stable index α in order to measure tail thickness：A critique [J]. Annals of Statistics，1983，11(4)：1019－1031.

[34] Kimoto T，Asakawa K. Stock market prediction system with modular neural networks [C] // Proc. Int'l J. Conf. Neural

Network，1990：1 - 6.

[35] 叶五一，缪柏其. 基于 Copula 变点检测的美国次级债金融危机传染分析 [J]. 中国管理科学，2009，17(3)：1 - 7.
Ye Wuyi，Miao Baiqi. Analysis of sub-prime loan crisis contagion based on change point testing method of Copula [J]. Chinese Journal of Management Science，2009，17(3)：1 - 7. (in Chinese)

[36] 王永巧，刘诗文. 基于时变 Copula 的金融开放与风险传染 [J]. 系统工程理论与实践，2011，31(4)：778 - 784.
Wang Yongqiao，Liu Shiwen. Financial market openness and risk contagion：A time-varying Copula approach [J]. Systems Engineering-Theory & Practice，2011，31(4)：778 - 784. (in Chinese)

[37] 叶五一，韦　伟，缪柏其. 基于非参数时变 Copula 模型的美国次贷危机传染分析 [J]. 管理科学学报，2014，17(11)：151 - 158.
Ye Wuyi，Wei Wei，Miao Baiqi. Analysis of sub-prime loan crisis contagion based on non-parametric time-varying Copula [J]. Journal of Management Sciences in China，2014，17(11)：151 - 158. (in Chinese)

[38] 宋新平，丁永生，曾月明. 基于遗传算法的同步优化方法在财务困境预警中的应用 [J]. 预测，2009，28(1)：48 - 55.
Song Xinping，Ding Yongsheng，Zeng Yueming. Application of a GA-based simultaneous optimization method in financial crisis prediction [J]. Forecasting，2009，28(1)：48 - 55. (in Chinese)

Early warning for extremely financial risks based on ODR-ADASYN-SVM

LIN Yu[1]，*HUANG Xun*[1]，*CHUN Wei-de*[1]，*HUANG Deng-shi*[2]

1. Business School，Chengdu University of Technology，Chengdu 610059，China；
2. School of Economics and Management，Southwest Jiaotong University，Chengdu 610031，China

Abstract: Synthetic minority over-sampling technique (SMOTE) has the problem of over fitting in improving the imbalanced samples' learning ability of support vector machine (SVM). In this paper，adaptive synthetic sampling approach (ADASYN) and optimization of decreasing reduction approach (ODR) are assembled into an ODR-ADASYN to overcome the blindness in generating new samples and the limitations in processing the object. Combining SVM with ODR-ADASYN，an improved SVM，named ODR-ADASYN-SVM，is put forward to predict extremely financial risks；T-test is also applied to the significance test of the difference of the prediction accuracy of all models and to the evaluation of the prediction stability of all models. The result illustrates that the ODR-ADASYN-SVM can not only significantly improve the imbalanced samples' learning ability of SVM，but also overcome the problem of over fitting for SMOTE effectively. Hence，the ODR-ADASYN-SVM has a superior ability to predict extremely financial risks.

Key words: ODR；ADASYN；SVM；extremely financial risk；early warning model