一种改进的降噪自编码神经网络不平衡数据分类算法 *

张成刚 [1a, 1c]，宋佳智 [2]，姜静清 [1b*, 3]，裴志利 [1b]

(1. 内蒙古民族大学 a. 数学学院；b. 计算机科学与技术学院；c. 网络中心 内蒙古 通辽 028000；2. 东北师范大学 计算机科学与信息技术学院，长春 130000；3. 吉林大学 符号计算与知识工程教育部重点实验室，长春 130012)

摘 要：针对少数类样本合成过采样技术(synthetic minority over-sampling technique，SMOTE)在合成少数类新样本时会带来噪音问题，提出了一种改进降噪自编码神经网络不平衡数据分类算法(SMOTE-SDAE)。该算法首先通过 SMOTE 方法合成少数类新样本以均衡原始数据集，考虑到合成样本过程中会产生噪音的影响，利用降噪自编码神经网络算法的逐层无监督降噪学习和有监督微调过程，有效实现对过采样数据集的降噪处理与数据分类。在 UCI 不平衡数据集上实验结果表明，相比传统 SVM 算法，该算法显著提高了不平衡数据集中少数类的分类精度。

关键词：神经网络；过采样；不平衡数据；分类

中图分类号：TP183

Imbalanced data classification algorithm of improved de-noising auto-encoder neural network

Zhang Chenggang[1a, 1c], Song Jiazhi[2], Jiang Jingqing[1b*, 3], Pei Zhili[1b]

(1. a. College of Mathematics; b. College of Computer Science & Technology; c. Network Center, Inner Mongolia University for the Nationalities, Tongliao Inner Mongolia 028000, P. R. China; 2. College of Computer Science & Information Technology, Northeast Normal University, Changchun 130000, P. R. China; 3. Key Laboratory of Symbolic Computation & Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, P. R. China)

Abstract: Aiming at the noise problems of SMOTE algorithm when synthesizing new minority class samples, the paper proposed a Stacked De-noising Auto-Encoder neural network algorithm based on SMOTE, SMOTE-SDAE. The proposed algorithm balances the original data sets by using SMOTE to synthesize new minority class samples, and then effectively de-noises and classifies the oversampling data sets through the layer-by-layer unsupervised de-noise learning and supervised fine-tuning process of de-noising auto-encoder neural network given the impact of noise produced in the process of synthesizing samples. Experimental results on UCI imbalanced data sets indicate that compared with traditional SVM algorithms, SMOTE-SDAE algorithm significantly improves the minority class classification accuracy of the imbalanced data sets.

Key Words: neural network;over-sampling;imbalanced data;classification

0 引言

分类问题是数据挖掘和机器学习领域的一项重要研究内容，针对平衡数据分类问题现有的方法一般都能得到较好的效果，但是在实际的应用环境中存在和产生着大量的不平衡数据，例如网络入侵、文本分类、欺骗信用卡检测、医疗诊断等[1-4]，其中准确识别少数类信息有着重要意义。为弥补少数类在样本在数据分布方面不足的问题，Chawla 等人[5]提出的 SMOTE 算法不仅能有效的人工合成少数类样本，而且在一定程度上避免了过拟合问题，但由于人工生成了新的样本数据，所以会带来噪音等问题。近些年来，研究人员提出了许多 SMOTE 的改进算法。张永等人[6]提出的 ALSMOTE 算法将主动学习支持向量机作为分类器，采用基于距离的主动选择最佳样本的学习策略，用来改进 SMOTE 方法的不足。王超学等人[7]将遗传算法引入到 SMOTE 中提出 GA-SMOTE 算法，能够有区别的选择少数类样本，并且有效控制合成样本的质量。但上述算法均存在以下不足：(1)没有考虑到数据集内部的本质特征表示，所以对数据的泛化能力不高；(2)针对新增的噪音数据缺少必要的处理过程，导致算法的鲁棒性较差。

基于深度学习思想的自编码神经网络（Auto-Encoder neural

基金项目：国家自然科学基金资助项目（61163034；61373067）；内蒙古自然科学基金资助项目（2016MS0336）；内蒙古民族大学科学研究项目（NMD1231）；内蒙古自治区“草原英才工程”基金项目（2013）；内蒙古自治区“青年科技领军人才”基金项目（NJYT-14-A09）；内蒙古自治区“321 人才工程”二层次人选基金项目（2010）.

作者简介：张成刚（1986-），男，硕士研究生，主要研究方向为人工智能、机器学习；宋佳智（1993-），男，硕士研究生，主要研究方向为数据挖掘；姜静清（1968-），女（通信作者），教授，博士，硕导，主要研究方向为人工智能、机器学习（jiangjingqing@aliyun.com）；裴志利（1968-），男，教授，博士，主要研究方向为机器学习、文本挖掘.

network ,AE）在机器学习和数据挖掘领域已经取得巨大成功[8]，首先利用无监督学习方式预训练网络参数，依靠逐层调整参数来学习数据的内在特征并消除无关和冗余信息，然后使用有监督学习将重构误差反向优化参数。堆栈降噪自编码神经网络(stacked de-noising auto-encoder neural network ,SDAE)通过将原始数据加入噪声，可训练出对原始输入信息更加鲁棒的表达特征，从而提升自编码神经网络对输入数据的泛化能力[9]。本文提出的基于 SMOTE 方法降噪自编码神经网络算法(SMOTE-SDAE)，既减小了 SMOTE 方法带来的噪音问题，又对采样后的数据降噪并且进行分类，并通过实验验证了改进算法能够有效提高少数类样本的分类效果。

1　相关工作

1.1　SMOTE 算法

少数类样本合成过采样技术(synthetic minority over-sampling technique ,SMOTE)是 Chawla 等人在 2002 年提出的一种典型的过采样方法，由于产生新的少数类样本，它较传统随机采样算法能有效避免分类器的过拟合现象。其主要思想是在距离较近少数类样本之间按照公式(1)随机插入一个人工合成的少数类样本，达到平衡数据集的目的[7]。

$$x_{new} = x + rand(0,1) \times (y[i] - x) \tag{1}$$

这里，$i = 1,2,\ldots N$,样本的过采样率为 N ，$rand(0,1)$ 表示在区间 (0,1) 之间的一个随机生成数；x_{new} 表示新生成的少数类样本，x 表示原始的少数类样本，$y[i]$ 表示少数类 x 周围的第 i 近邻少数类样本。

1.2　自编码神经网络(auto-encoder neural networks, AE)

基于深度学习(deep learning)思想的自编码神经网络是一种尽可能重构输入数据的无监督学习神经网络[8]，利用逐层训练优化算法（如批梯度下降算法 BGD 或 L-BFGS 算法）初始化网络权重并使用反向传播算法（BP 算法）微调网络参数，优化整体性能。多次组合自编码神经网络，即把网络中当前层的输出作为下一层的输入，就构成堆栈自编码神经网络(stack auto-encoder neural networks, SAE)的深度结构。

1.3　降噪自编码神经网络（denoising auto-encoder neural networks ,SDAE)

降噪自编码神经网络是在自编码神经网络(AE)的基础上，对原始输入数据加入按一定概率分布的噪声，作为神经网络的输入数据，自编码神经网络尝试学习如何去除噪声，并且尽最大可能重构没有噪声扰乱过的输入，因此从含有噪声的输入中学习得到的特征更具鲁棒性，提升了自编码神经网络模型对输入数据的泛化能力[10-11]。多次重复降噪自编码神经网络，并同样把当前层的输出作为下一层的输入，就构成了堆栈降噪自编码神经网络(Stacked Auto-Encoder neural networks, SDAE)深度结构。堆栈降噪自编码神经网络如图 1 所示。

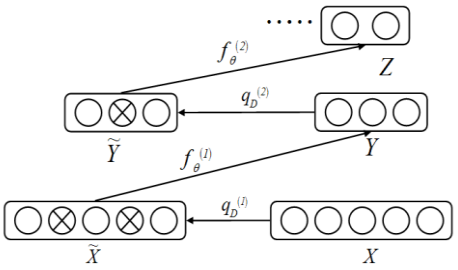


图 1 堆栈降噪自编码神经网络结构图

这里，将原始数据 x 中加入按一定概率分布 q_D 的噪音，扰乱生成 $\tilde{x}$ 作为自编码器的输入，然后通过激活函数 f_θ 计算出隐藏层中各个神经元的激活值，计算公式如下：

$$h_{w,b}(\tilde{x}) = f(\sum w\tilde{x} + b) \tag{2}$$

所以降噪自编码器的代价函数为：

$$J_{DAE}(w,b) = \frac{1}{m}\sum(\frac{1}{2}\left\|h_{w,b}(\tilde{x}) - x\right\|^2) \tag{3}$$

其中 w 神经元之间的权重，b 为偏置，m 为训练样本数，本文采用 $sigmoid$ 函数作为神经元的激活函数 $f(x) = 1/(1+e^{-x})$，其值域为 $[0,1]$ 。

2　提出的算法

结合 SMOTE 与 SDAE 的优点，本文提出一种基于 SMOTE 方法的降噪自编码神经网络不平衡数据分类算法(SMOTE-SDAE)：首先将 SMOTE 方法作用于原始数据集 *Dataset*，降低数据集内的不平衡度，针对 SMOTE 方法带来新的数据噪音问题，经过堆栈降噪自编码神经网络无监督逐层贪婪训练之后，模型抽象出的特征更具鲁棒性，增强了自编码神经网络模型对输入数据的泛化能力，从而提升了少数类及整体的分类效果。算法流程图如图 2 所示。

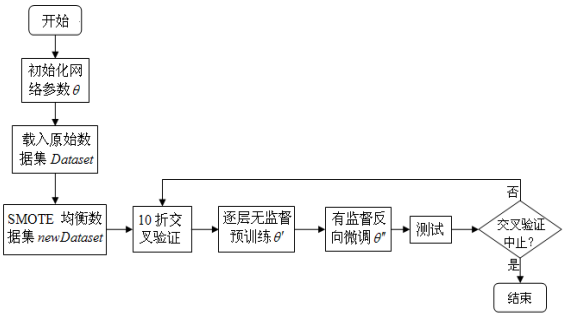


图 2 算法流程图

算法描述如下：

步骤 1　随机初始化参数 $\theta = \{W, b\}$，其中 $\boldsymbol{W}$ 为网络的权重，$\boldsymbol{b}$ 为偏置；可视层神经元个数 v，隐藏层神经元个数 $h1, h2$；q_D 为噪音系数，T 为数据集中少数类样本的总数，N 为采样率（样本合成比例），k 为选取近邻的个数，默认为 5；

步骤 2　载入原始数据集 *Dataset*；

步骤 3　过采样数据集 *newDataset = SMOTE(T, N, k)*;

步骤 4　对 *newDataset* 加入高斯噪音 q_D 生成新的训练集 *N-D*;

步骤 5 L-BFGS 优化算法逐层学习网络参数得到 θ'；

步骤 6　反向传播算法(BP)微调整个网络并更新参数得到 θ''；

步骤 7 测试并返回 AUC、F-value 和 G-mean 的结果。

3　实验

3.1　数据集描述

文中实验的数据是 UCI 机器学习数据库[12]中的 4 个常用的数据集。其中 ionosphere、german、wpbc 是二分类不平衡数据集；为了增大类别的不平衡程度，将 satimage 数据集中 Class 4 作为少数类，其他类共同作为多数类，构成了不平衡率较高的二分类数据集。具体描述如表 1 所示。

表 1　数据集描述

数据集	样本总数	属性数	少数类	多数类	不平衡率
ionosphere	351	34	126	225	1:1.8
german	1000	24	300	700	1:2.3
wpbc	198	33	47	151	1:3.2
satimage	6435	37	626	5809	1:9.3

3.2　基于混淆矩阵的评估指标

用于两分类问题的混淆矩阵(confusion matrix)定义如表 2 所示。

表 2　两分类问题的混淆矩阵

分类	实际少数类	实际多数类
预测少数类	TP	FP
预测多数类	FN	TN

这里，TP(true positive)代表实际少数类样本判定为少数类样本的数量；TN(true negative)代表实际多数类样本判定为多数类样本的数量；同理，FN(false negative)、FP(false positive)分别代表判定错误的情况下实际多数类样本和少数类样本的数量。

在传统的分类方法中，通常采用以整体正确率作为评价标准，由于分类器对少数类的不敏感，当对不平衡数据进行分类时，少数类在很大程度上被判为了多数类，导致少数类样本的识别率较低，目前出现了一些新的不平衡数据的分类评价指标，例如 AUC、F-value 和 G-mean 等方法[13]。

AUC(area under roc curve) 为其对应的 ROC(receiver operating characteristic)曲线以下的面积，是一种用来度量分类器好坏的一个标准，面积越大，则分类器的性能将被认为越好[14]。

F-value 是衡量准确率和召回率的分类评价指标，比较偏向对少数类的分类性能评价，定义如下：

$$F-value=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision} \tag{4}$$

其中，准确率 $precision=TP/(TP+FP)$，召回率 $recall=TP/(TP+FN)$，β 取值为 $[0,+\infty]$。本实验中取 $\beta=1$，此时 F-value 则表示召回率和准确率之间权重的平均。

G-mean 表示少数类分类精度和多数类分类精度的几何平均值，用来评价分类器整体的分类性能，其定义如下：

$$G-mean=\sqrt{\frac{TN}{TN+FP}\times\frac{TP}{TP+FN}} \tag{5}$$

G-mean 表示只有在少数类和多数类分类精度同时都高的情况下，此时 G-mean 的值最大[15-16]。

3.3　实验结果及分析

本文的对比算法为只有一个隐层的 SVM 算法，通过将输入样本映射到一个高维的特征空间，并在这个特征空间中构造出最优分类超平面，实现分类任务。而 SDAE 算法则有 4 层网络结构，其中包括 2 个隐层，SDAE 通过组合低层特征产生更加抽象的高层特征，以发现数据的内在特征表示[17]，所以本文的 SMOTE-SDAE 算法将合成新少数类样本作为训练集，利用无监督学习方式，预训练出较好的初始化参数，经过微调后，抽象出最能表达原始数据的本质特征，有助于提升分类能力。

3.3.1 实验中相关参数设置

本文实验环境为 WIN7 64 位，CPU 3.4GHz，RAM 4G，在 Matlab2012b 中实现提出的 SMOTE-SDAE、SVM、SMOTE-SVM、SDAE、SAE 算法，其中 SVM 算法通过调用 Libsvm-3.12 工具箱中的函数得以实现[18]，核函数使用 RBF 核函数，系数 $\sigma=0.5$，惩罚因子 $\cos t=2$；SMOTE-SVM 与 SMOTE-SDAE 算法中，近邻个数 $k=5$，采样率为 *N*；噪音系数 $q_D=0.15$；对于 SAE、SDAE 和 SMOTE-SDAE 算法均采用 4 层网络结构：输入层神经元数目 *v* 为各数据集的属性数；由于每个数据集的内部分布不同，经过实验凑试法，当第一隐藏层神经元数量 *h1* 在区间[25,35]时，第二隐藏层神经元数目 *h2* 在区间[6,10]时，结果最好；因为是二分类问题，所以输出层节点数目均为 1；在无监督训练和监督微调过程中，使用 L-BFGS 优化算法，迭代次数为 50；为避免数据重复计算，客观评价算法的性能，实验均采用 10 次 10 折交叉验证的平均值作为数值结果。

3.3.2 实验结果对比

表 3 与 4 为各算法(SVM、SMOTE-SVM、SAE、SDAE、SMOTE-SDAE)的 AUC、F-value 对比表，mean 为算法的平均结果，在 4 个不平衡数据集(ionosphere、german、wpbc、satimage)实验可知，SMOTE-SDAE 算法结果均大于传统算法取得结果。

表 3　不同算法的 *AUC* 对比(%)

AUC	ionosphere	german	wpbc	satimage	mean
SVM	89.91	82.95	66.18	85.75	81.20
SMOTE-SVM	91.85	87.36	71.09	84.15	83.61
SAE	85.67	74.12	57.84	80.73	74.59
SDAE	88.36	78.09	63.86	83.18	78.37

SMOTE-SDAE	**97.92**	**91.13**	**79.49**	**91.57**	**90.03**

表 4　不同算法的 *F-value* 对比　(%)

F-value	ionosphere	german	wpbc	satimage	mean
SVM	57.38	64.32	61.50	86.55	67.44
SMOTE-SVM	74.89	79.28	77.37	84.14	78.92
SAE	56.48	62.05	58.57	76.12	63.31
SDAE	52.36	71.18	52.53	79.97	64.01
SMOTE-SDAE	**78.42**	**81.92**	**83.71**	**89.48**	**83.38**

表 5 为不同算法的 G-mean 结果对比表，在 german 数据集上，SMOTE-SDAE 出现偏向于少数类样本识别而忽略多数类样本的现象，由于 G-mean 指标描述的是整体分类效果，导致 SMOTE-SDAE 算法关于 G-mean 指标表现略差，但在其他数据集上 SMOTE-SDAE 算法的 G-mean 指标以及总体表现优于其他算法。

表 5　不同算法的 *G-mean* 对比　(%)

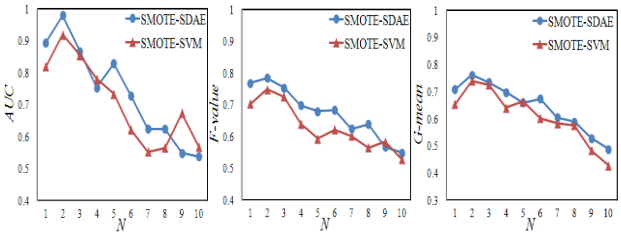
G-mean	ionosphere	german	wpbc	satimage	mean
SVM	60.22	81.28	48.30	84.62	68.61
SMOTE-SVM	73.96	**86.30**	52.08	85.74	74.52
SAE	68.12	74.65	46.77	77.22	66.69
SDAE	70.98	75.95	58.72	82.58	72.06
SMOTE-SDAE	**76.15**	**84.50**	**65.24**	**90.77**	**79.17**

此外，从以上三个指标评价表中可以得出以下结论：a)经过降噪处理的自编码神经网络的分类效果优于无降噪过程的自编码神经网络；b)经过 SMOTE 算法采样过的数据集，降噪自编码神经网络(SDAE)相对于其他算法明显提高少数类的分类效果。

3.3.3 实验结果与过采样率 N 的关系

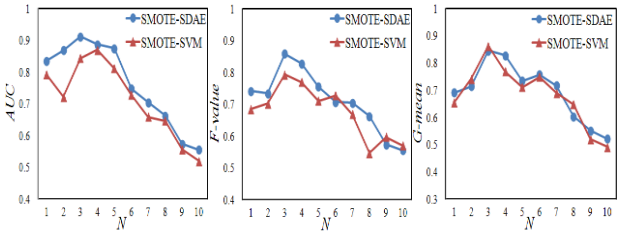
图 3~6 表示的是在 ionosphere、german、wpbc 和 satimage 不平衡数据集上实验结果关于采样率 N 的变化情况。对于 ionosphere 数据集，当采样率 $N=2$ 时，SMOTE-SDAE 算法达到最大值，此时数据集内部达到类别平衡的状态，而随着 N 不断增大，分类精度急剧下降，原因是数据集中的少数类成为了多数类，类别又出现不平衡现象；同理，对于 german 数据集，采样率 $N=3$，对于 wpbc 数据集，采样率 $N=4$，对于 satimage 数据集，采样率 $N=10$ 时，SMOTE-SDAE 算法达到最大值，并且整体表现优于 SMOTE-SVM 算法。

通过以上分析，还可以得出下面结论：a)当采样率 N 接近数据集中样本不平衡率时，数据集趋于平衡，此时分类效果最好；b) 由于采样后数据集内训练样本数量增多，更有利于 SMOTE-SDAE 算法通过无监督训练过程从中学习到数据内部特征。从总体看来 SMOTE-SDAE 算法对不平衡数据分类效果和稳定性都比较好。



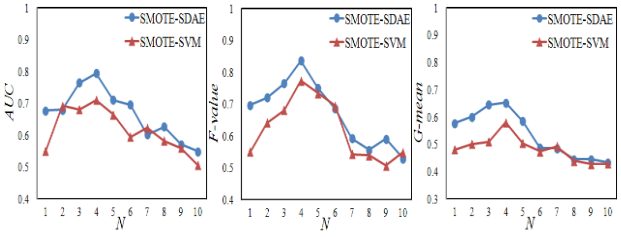
(a)AUC　(b) F-value　(c) G-mean

图 3　ionosphere 的结果关于采样率 N 的变化曲线



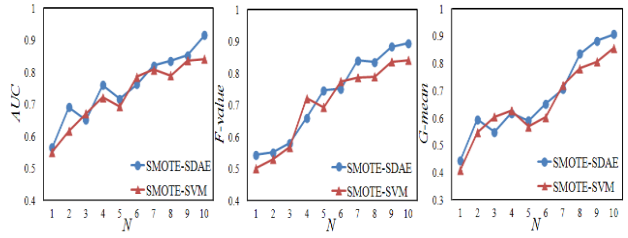
(a)AUC　(b) F-value　(c) G-mean

图 4　german 的结果关于采样率 N 的变化曲线



(a)AUC　(b) F-value　(c) G-mean

图 5　wpbc 的结果关于采样率 N 的变化曲线



(a)*AUC*　(b) *F-value*　(c) *G-mean*

图 6　*Satimage* 的结果关于采样率 *N* 的变化曲线

4　结束语

不平衡数据分类问题一直是数据挖掘和机器学习领域的研究重点，本文针对 SMOTE 算法合成少数类新样本带来的噪音等问题，将降噪自编码神经网络与之结合，提出了一种基于 SMOTE 方法的降噪自编码神经网络算法。相比传统降噪自编码神经网络与 SVM 等算法，本文算法提高了少数类的分类精度。实际应用中对一个结构未知的数据集，目前还没有比较有效的方法来确定合适的过采样率以达到较好的分类精度，根据经验通过多次实验，对比结果选定最优的过采样率。如何更有效地合成少数类样本、神经网络节点数目的选择以及对高维数据的处理等问题还有待于进一步研究。

参考文献:

[1]　Thomas C. Improving intrusion detection for imbalanced network traffic[J].

Security & Communication Networks, 2013, 6(6): 309–324.

[2] 骆自超, 金隼, 邱雪峰. 考虑类内不平衡的谱聚类过抽样方法[J]. 计算机工程与应用, 2014, 50(11): 120-125.

[3] 李伟, 赵亚欧, 陈月辉. 均衡数据法提高蛋白质二级结构预测[J]. 计算机工程与应用, 2009, 45(6): 219-220.

[4] Drosou K, Georgiou S, Koukouvinos C, *et al*. Support vector machines classification on class imbalanced data: a case study with real medical data[J]. Journal of Data Science, 2014, 12(4): 143-155.

[5] Chawla N V, Bowyer K W, Hall L O, *et al*. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321--357.

[6] 张永, 李卓然, 刘小丹. 基于主动学习 SMOTE 的非均衡数据分类[J]. 计算机应用与软件, 2012, 29(3): 91-93.

[7] 王超学, 张涛, 马春森. 面向不平衡数据集的改进型 SMOTE 算法[J]. 计算机科学与探索, 2014, 8(6): 727-734.

[8] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[9] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends® in Machine Learning, 2009, 2(1): 1-127.

Vincent P, Larochelle H, Bengio Y, *et al*. Extracting and composing robust features with denoising autoencoders[C]//Proc of the 25th international conference on Machine learning. ACM, 2008: 1096-1103.

Vincent P, Larochelle H, Lajoie I, *et al*. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. [J]. Journal of Machine Learning Research, 2010, 11(6): 3371-3408.

[10] Blake C, Merz C, Blake C, *et al*. UCI repository of machine learning databases[C]//Proc of International Conference on Document Analysis and Recognition. 2003.

[11] Provost F. Machine learning from imbalanced data sets 101[C]//Proc of the AAAI Workshop on Imbalanced Data Sets. 2000: 1-3.

[12] 程险峰, 李军, 李雄飞. 一种基于欠采样的不平衡数据分类算法[J]. 计算机工程, 2011, 37(13): 147-149.

[13] Menardi G, Torelli N. Training and assessing classification rules with imbalanced data[J]. Data Mining and Knowledge Discovery, 2014, 28(1): 92-122.

[14] Di Martino M, Fernández A, Iturralde P, *et al*. Novel classifier scheme for imbalanced problems[J]. Pattern Recognition Letters, 2013, 34(10): 1146-1151.

[15] 孙志军, 薛磊, 许阳明, 等. 深度学习研究综述[J]. 计算机应用研究, 2012, 29(8): 2806-2810.

[16] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Trans on Intelligent Systems and Technology, 2011, 2(3): 27.