Advances in Applied Mathematics 应用数学进展, 2020, 9(5), 622-629
Published Online May 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.95074

Hans汉斯

An Ensemble Imbalanced Data Classification Algorithm Based on Random *k*-Rank Nearest Neighbor Rules

Yixin Shen[1], Subhash C. Bagui[2], Shuangge Ma[1*]

[1]College of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi
[2]Department of Mathematics and Statistics, The University of West Florida, Pensacola FL
Email: *986177822@qq.com

Received: Apr. 7th, 2020; accepted: Apr. 29th, 2020; published: May 6th, 2020

Abstract

In this article, a random ensemble *k*-RNN algorithm called REKRNN is proposed to deal with the imbalanced data classification. The algorithm incorporates the *k*-rank nearest neighbor classifier into the frame of Bagging algorithm. At the same time, resampling techniques and random feature method are applied to deal with the imbalanced issue. We observe that the proposed method performed remarkably well on different imbalanced dataset. The random ensemble *k*-RNN algorithm can be considered as a promising tool for imbalanced classification.

Keywords

Imbalanced Data Classification, *k*-Rank Nearest Neighbor Rule, Bagging, Resampling Techniques

基于随机秩次*k*近邻规则的不平衡数据分类算法

沈怡欣[1]，Subhash C. Bagui[2]，马双鸽[1*]

[1]太原理工大学数学学院，山西 晋中
[2]西佛罗里达大学数学与统计系，佛罗里达 彭萨科拉
Email: *986177822@qq.com

收稿日期：2020年4月7日；录用日期：2020年4月29日；发布日期：2020年5月6日

*通讯作者。

文章引用: 沈怡欣, Subhash C. Bagui, 马双鸽. 基于随机秩次 *k* 近邻规则的不平衡数据分类算法[J]. 应用数学进展, 2020, 9(5): 622-629. DOI: 10.12677/aam.2020.95074

摘　要

针对不平衡数据分类问题，为提高二分类任务中少数类样本分类准确率低的问题，本文提出一种随机秩次*k*近邻集成学习算法——REKRNN。该方法将秩次*k*近邻算法应用于Bagging集成学习框架中，同时采用混合重采样和随机子空间法平衡训练集，增加基学习器差异性。仿真实验证明，该算法在处理不平衡数据分类任务时性能良好。

关键词

不平衡数据分类，秩次*k*近邻，集成学习，重采样，随机子空间法

1. 引言

分类算法是机器学习的一个重要分支。传统的分类算法大多基于平衡的数据集，即假设不同类别间样例数目相对平衡，并且以总体分类准确率为分类器效能评价指标。然而，在诸多实践中，我们能够获取的数据集中不同类别的样例数往往是不平衡的。例如在病例诊断[1]中，多数样例为健康，仅有少数为患病；又如在信用评估[2]中，也仅有极少数的样例信用记录不良。当面对不平衡分类任务时，传统的算法会倾向于多数类，假设一个数据集中多数类与少数类的比例为 99:1，即使将全部样例均预测为多数类，总体准确率依然可以高达 99%。然而在不平衡分类中，更为重要的往往是识别少数类的能力，这使得接近于 1 的总体分类准确率并无意义。因此，提出针对不平衡数据的有效分类算法尤为重要，这也是近年分类算法研究的一个趋势。

本文从集成学习的角度出发，构建了一个基于 Bagging 框架的随机秩次 *k*-近邻不平衡数据分类算法(Random Ensemble *k*-rank nearest neighbor Algorithm, REKRNN)。它将秩次 *k*-近邻分类器(*k*-rank nearest neighbor，简称 *k*-RNN)应用于集成学习框架中，使 *k*-RNN 学习能力强、算法复杂度低的优势，与 Bagging 的高泛化性能相结合，将重采样和分类算法嵌入推进，最终获得了良好的分类性能。

2. 相关研究

2.1. 不平衡数据分类的集成学习算法

近年来的国内外研究中，提高不平衡数据中少数类分类准确率问题的方法主要分为三个类别：数据层面，算法层面和集成学习。数据层面是指在训练模型之前，对原始数据集进行不同方式的重采样，调整多数类和少数类样本的个数，从而降低不平衡度。例如，通过上采样重复选取少数类样本[3]，或者通过降采样减少多数类样本[4]，来重构数据集使其中不同类别样本数目相对平衡。然而，预处理后用于最终训练分类器的单个数据集，会损失样例特征信息或重复采样多次而造成大量的信息浪费，因此数据层面的方法在不平衡率较高时不适用。算法层面是指针对不平衡数据的特点，提出新的分类算法或者对已有算法做出改变。例如代价敏感学习，通过为不同类别样本设置不同错分代价提高分类准确率。

集成学习是数据和算法的混合方法，它的主要思想是创建一系列弱学习器，再以一定的策略将其结

合做出最终决策，利用弱学习器之间的差异性得到更全面的强分类模型。集成学习能够有效地将采样方法和分类算法嵌入推进，是效能更高的不平衡数据分类算法。其中 Boosting 类通过引入代价敏感学习更新权重，改变多数类和少数类样本的分布，串行生成分类器。它通常基于整个特征空间构建算法，因此整体复杂度较高，不利于高维数据处理。与之相比，Bagging 类并行生成基分类器，算法结构更为简单，具有较高的泛化性能[5] [6]。因此本文提出的 REKRNN 算法采用 Bagging 框架，构建以 *k*-RNN 为弱学习器的不平衡数据分类算法。

现有的基于 Bagging 的集成算法中，多数采用决策树作为弱学习器，例如随机森林。近年来，研究者将更多的机器学习算法引入其中，例如神经网络，支持向量机等，使得单个学习器的性能通过集成得到提升[7]。然而，在特征数较多时，这些算法在学习过程中容易因为复杂度高而产生过拟合现象，使得模型整体泛化性能较差。同时，计算过程时间和资源消耗巨大。

为了降低运算成本，本文首次采用了模型复杂度更低的秩次 *k* 近邻算法(k-rank nearest neighbor classifier)作为集成算法的弱学习器。单变量的*k*-RNN 规则首先由 Anderson 等人于 1996 年提出，随后 Bagui 和 Pal 进一步将之拓展为适用于多变量的分类算法[8] [9]。与最为经典的 *k*-NN 算法相同，它的主要思想是如果一个样本在样本空间中的大多数“邻居”都属于一个类别，那这个样本也属于这个类别[10]。在分类决策时，只依据这些邻居的类别来做出预测。所不同的是，在寻找最近邻样本时，*k*-NN 基于每两个样本间的距离，而 *k*-RNN 基于对各个类别样本混合排序后的次序，即秩次。试验证明，*k*-RNN 在拥有与 *k*-NN 相当的分类性能的同时，算法的复杂度更低。

2.2. *k*-RNN 分类器

原始的 *k*-RNN 规则适用于单变量总体，即总体中样本仅有一个特征，它规则和主要思想如下：

设 $\{X_1, X_2, \cdots, X_{n_1}\}$ 和 $\{Y_1, Y_2, \cdots, Y_{n_2}\}$ 是分别来自两个给定总体 π_1 和 π_2 的训练样本集，测试样例 Z 可被分类至总体 π_1 或 π_2。将 X,Y 和 Z 全部样本(n_1+n_2+1个)按照降序(或升序)排列，得到混合样本的秩次；然后，取 Z 左边的样本和右边的样本各 k 个作为秩次最近邻，样例 Z 的预测类别将由这 $2k$ 个最近邻中的出现次数较多的类别标记为预测结果。

尽管原始的 *k*-RNN 规则有很好的分类能力，但由于“按照降序(或升序)排列”这一概念无法自然地扩展到多变量总体，它只能用于单一特征的总体分类任务。为了将 *k*-RNN 规则能够适用于多特征总体分类，并使其在大多数现实问题中得到应用，Bagui 扩展了该算法，将 Randlest 的多维样本排序法引入其中，提出了基于排序的多元 *k*-RNN 分类器。该排序法综合了不同总体间的均值和协方差矩阵的差异，更加具有统计学意义。多元 *k*-RNN 分类器的主要思想如下：

设 $\{X_1, X_2, \cdots, X_{n_1}\}$ 和 $\{Y_1, Y_2, \cdots, Y_{n_2}\}$ 是分别来自两个给定总体 π_1 和 π_2 的训练样本集。其中 $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip}) \in R^p$，总体均值为 μ_1，协方差矩阵为 Σ_1；$Y_i = (y_{i1}, y_{i2}, \cdots, y_{ip}) \in R^p$，总体均值为 μ_2，协方差矩阵为 Σ_2。测试样例 $Z = (z_1, z_2, \cdots, z_p) \in R^p$ 可被分类至总体 π_1 或 π_2。基于原始 *k*-RNN 规则，测试样例 Z 在多维样本 X,Y 和 Z 中的混合秩次由以下得分函数获得：

$$R(Z;\mu_1,\mu_2,\Sigma_1,\Sigma_2) = \left(\mu_1^{\mathrm{T}}\Sigma_1^{-1} - \mu_2^{\mathrm{T}}\Sigma_2^{-1}\right)Z - \frac{1}{2}Z^{\mathrm{T}}\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)Z \tag{1}$$

其中，μ_i^{T} 表示均值向量 μ_i 的转置，Σ_i^{-1} 表示协方差矩阵 Σ_i 的逆矩阵。

在大多数的应用环境中，参数 μ_1, μ_2, Σ_1 和 Σ_2 均为未知。Johnson 等人证明，可以由它们的无偏估计 $\bar{X}, \bar{Y}, S_1$ 和 S_2 代替得到得分函数的无偏估计：

$$\hat{R}(Z;\bar{X},\bar{Y},S_1,S_2) = \left(\bar{X}^{\mathrm{T}}S_1^{-1} - \bar{Y}^{\mathrm{T}}S_2^{-1}\right)Z - \frac{1}{2}Z^{\mathrm{T}}\left(S_1^{-1} - S_2^{-1}\right) \tag{2}$$

其中，

$$\bar{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i,\ \bar{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i \tag{3}$$

$$S_1=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)^{\mathrm{T}},\ S_2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)^{\mathrm{T}} \tag{4}$$

更进一步地，Bagui 和 Pal 利用大数定律证明了 $\hat{R}(.)\xrightarrow{P}R(.)$。即得分函数的估计值 $\hat{R}(.)$ 依概率收敛到 $R(.)$ [8]。

由此可知，得分函数 $R(.)$ 或者 $\hat{R}(.)$ 是一个从 R^p 到 R^1，从 p 维到 1 维的映射，它的得分值代表了任一样例在 X,Y 样本集中的混合秩次。通过得分函数计算 X,Y 中每个样例的分值，就能够得到混合样本中全部样例的秩次。原始 k-RNN 规则中排序得到秩次的思想通过得分函数得以解决，使之可以成功应用于多特征分类任务中，依据秩次选择双侧最近邻进行样例的分类预测。

多元 k-RNN 分类器的训练过程如下表 1：

Table 1. k-RNN classifier algorithm
表 1. k-RNN 分类器训练过程

输入：训练集 $S=\{x_i,y_i\},i=1,2,\cdots,m$，$y_i\in Y,Y=\{0,1\}$，其中 0 和 1 为相应样例类别标签；测试集
1.分别计算 1-类别和 0-类别样例的均值向量和协方差矩阵 2.通过得分函数 $\hat{R}(.)$ 计算训练集中所有样本的分值作为其秩次 3.依秩次将全部训练样本排序 4.选择参数值 k 5.对于测试集中的任一样例： -通过得分函数 $\hat{R}(.)$ 计算分值作为其在混合样本中的秩次 -依据秩次将测试样例插入排序后的训练样本中 -选择其左右方向各 k 个样例作为最近邻 -将 $2k$ 个最近邻中出现次数较多的类别标记为该样例的预测结果 -结束循环
输出：测试集样本预测分类

3. 随机秩次 k-近邻集成学习算法

3.1. Bagging 算法

Bagging 是一种典型的集成学习算法，是当前机器学习算法研究的热点之一，在分类和回归任务中都有广泛的应用和良好的效果。它的实现步骤为：每次从训练集中有放回地抽取 n 个样本，重复抽取 T 次；由这 T 个样本各训练一个弱学习器；由 T 个弱学习器各对测试样本进行预测，按照投票取众数的方法得到最终预测结果。其中重要的步骤有放回抽样，即为 Bootstrap 抽样，Bagging 名称的来源。它的主要思想是通过弱分类器之间的差异性提高模型泛化性能。针对不平衡数据分类的任务，REKRNN 算法对 Bagging 框架做出如下改进：1) 采用混合重采样法对初始训练集进行采样以获得较为平衡的训练集；2) 引入随机特征子集降低维度和计算复杂度。

3.2. 混合重采样

弱分类器间的差异首先通过重采样得到不同的训练子集获得。在 Bagging 中，通常由 Bootstrap 实现重采样过程，即从初始训练集(样本容量为 N)中有放回地随机抽取同等数量的样本(N 个)形成训练子集。可以证明初始训练集中约 63.2%的样本会被选入训练集中，且其中一些样本重复出现。

而针对不平衡数据集，Bootstrap 并不会对多数类和少数类样本有任何偏好，依然均衡地进行有放回抽样。因此重采样后的样本子集仍然为不平衡数据集，且当总样本量小或不平衡率很高时，有极大可能 Bootstrap 子集中没有少数类样本或样本过少影响分类算法的学习。同时，由于 REKRNN 中弱分类器的算法k-RNN 仍然以类别间相对平衡为前提，因此仅仅对少数类获取 Bootstrap 样本，而对多数类采用降采样随机抽取与少数类相同数目的样本，由此得到少数类数目相对稳定且类别间相对平衡的数据集。

3.3. 随机特征子集

第二个获得弱分类器差异的方法为随机特征子集，即创建若分类器时仅以特征空间的一个子集为分类依据。例如在随机森林中，每棵决策树的节点不是在整个特征空间中进行搜索，而是在一个随机产生的特征子集中选择最优分裂[11]。与此类似，REKRNN 算法在每一个重采样的样本子集上，随机选择不同的特征子集训练弱分类器，保证弱分类器的多样性，增强整体算法的泛化性能。另一方面，随机选择子集可以降低特征维度，在高维数据分类中能够有效提高算法效率。同时，多个子集集成也可以弥补降维带来的潜在精度损失。随机特征子集的选择过程如下表 2：

Table 2. Random feature space algorithm
表 2. 随机特征子空间算法

输入：任一降采样后的训练子集 l；特征空间 $\mathcal{L}$ 由特征向量 $P=(p_1,p_2,p_3,\cdots,p_d)$ 构成
1.随机选择数字 $n(n<d)$ 作为子集 l 选用的特征数 2.从特征空间中 $\mathcal{L}$ 随机选择 n 个特征形成特征子集 3.依据选出的特征子集，在训练子集 l 的基础上生成训练子集 l^*
输出：随机特征训练子集 l^*

3.4. 训练及测试过程

随机秩次 k-近邻集成学习算法(REKRNN)将秩次 k-近邻分类器 k-RNN 应用于 Bagging 集成学习框架中，将 k-RNN 分类器学习能力强、算法复杂度低的优势，与 Bagging 算法随机样本空间、随机特征空间分割集成的优势相结合。同时，对该框架加以改进，针对多数类样本采用降采样，针对少数类样本采用 Bootstrap 重复采样，以获得更加平衡的训练集，有效地提高数据集不平衡时的分类性能。

REKRNN 算法的实现主要有四个过程，如图 1 所示：

1) 按一定比例将数据集划分为测试集和训练集，利用混合重采样法，对少数类样本取 Bootstrap 样本，多数类进行降采样，从训练集中抽取样本生成 r 个训练子集；

2) 针对每个训练子集，随机选择将采用的特征数，在特征空间中随机抽取特征子集，生成最终训练样本；

3) 在每个训练样本子集上建立 k-RNN 弱分类器；

4) 由弱分类器预测测试样本类别，使用“投票法”，将多数弱分类器的预测结果标记为最终分类结果。

4. 实验与结果分析

4.1. 实验数据集

为了验证 REKRNN 算法对于不平衡数据集的分类性能，以及对少数类样本的识别能力，本文选取 KEEL[1] 中 Abalone9-18 数据集进行仿真实验。该数据集含有 731 个样本的 8 个特征，其中多数类样例 684 个，少数类样例仅有 42 个，不平衡率(多数类/少数类数目)高达 16.68，是一个典型的不平衡分类数据集。

[1] http://www.keel.es/

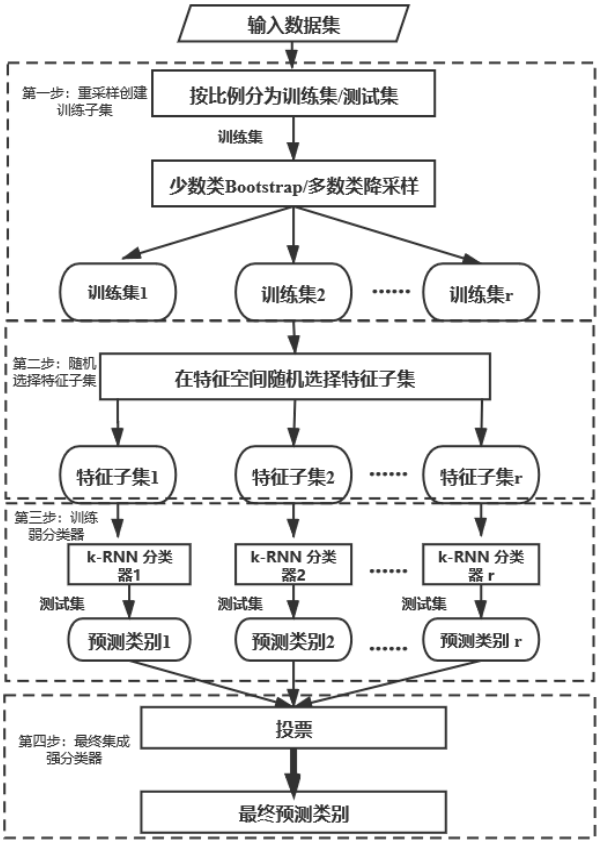

Figure 1. REKRNN algorithm flow chart
图 1. REKRNN 算法流程图

4.2. 评价指标

定义实验数据集中少数类样本数为 P，多数类样本为 N，相对应的混淆矩如表 3 所示：

Table 3. Confusion matrix
表 3. 混淆矩阵

类别	预测少数类	预测多数类
少数类	*TP*	*FN*
多数类	*FP*	*TN*

依据混淆矩阵，总体精确度(Overall accuracy)为分类正确的样例数占总样例数的比例：

$$\text{Overall accuracy} = \frac{TP + FP}{P + N} \tag{5}$$

敏感度(Sensitivity)，或查全率(Recall)为被预测为少数类的样例中真正少数类的比例：

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6}$$

特异性(Specificity)，或查准率为实际类别为多数类的样例中也被预测为多数类的比例：

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{7}$$

F 值(F-measure)，是对敏感度和特异性的折中：

$$F\text{-measure} = \frac{2}{1/\text{Sensitivity} + 1/\text{Specificity}} \tag{8}$$

ROC 曲线是对敏感度和特异性的综合图示，AUC (area under ROC curve)，即 ROC 曲线下的面积，经常作为一种不平衡分类的评价指标。

4.3. 实验过程与结果分析

为了验证 REKRNN 算法的分类性能，以及过程中使用的混合重采样，随即子空间两种技术对于提升学习器 k-RNN 算法分类性能的效果，分别使用单个基础 k-RNN 算法，重采样后平衡数据集的 k-RNN 算法，以及最终形成的 REKRNN 算法进行仿真实验。同时，以 KNN 和 Adaboost 两种模型为参照，比较 REKRNN 算法的分类效果。

本文使用 Python 进行仿真实验。k-RNN 和 KNN 中参数 k，Adaboost 中决策树的数目均由十折交叉验证获得。对于数据集的划分均采用训练集 70%，测试集 30%的分类比例。以上文提到的五种评价指标为分类效果准则，结果均用百分数表示。实验结果见表 4。

Table 4. Experimental output
表 4. 实验结果

Model	Accuracy	Sensitivity	Specificity	*F*-measure	AUC
Single Multivariate *k*-RNN	96.36%	37.5%	98.58%	54.33%	68.04%
***k*-RNN based on balanced dataset**	92%	92.86%	83.33%	87.84%	88.10%
REKRNN	92.27%	92.86%	92.23%	92.54%	92.55%
Adaboost	96.17%	28.6%	98.86%	44.37%	50%
KNN	93.89%	14.29%	97.16%	24.92%	55.72%

由上表，我们可以看出，在单个 k-RNN 分类的基础上，使用混合重采样得到平衡数据集后算法的查全率，F 值和 AUC 分别提高了 55.36%，33.51%和 20.06%，这表明重采样使得算法对于少数类样本的分类能力大大提升；在此基础上，加入 Bagging 和随机子空间法后，使用 REKRNN 模型后查准率，F 值和 AUC 又分别提高了 8.9%，4.7%，4.45%，这表明通过集成和随机特征选择又进一步提高了模型的泛化能力。同时，与 Adaboost 和 KNN 算法相比，本文提出的 REKRNN 算法在各个指标下均表现更优，表明此算法在不平衡数据分类中具有良好的表现。

5. 结论与展望

针对不平衡分类任务中少数类样本分类准确率低的问题，本文提出基于随机秩次 K 近邻规则的不平衡数据分类算法——REKRNN。k-RNN 分类能力强，算法复杂度低，Bagging 集成框架提高算法整体泛化性能，混合重采样使得子训练集相对平衡，随即子空间算法降低维度的同时增大基学习器差异性。从仿真实验结果来看，这四个元素均对少数类分类准确率的提高有很大贡献，使得算法能够有效识别少数类样例。与两种传统分类算法相比，REKRNN 也体现了它在不平衡数据分类中的优势。在下一步的研究中，可以考虑将代价敏感学习引入算法中，通过为少数类样例设置较大错分代价提高其分类准确率，进一步提高算法分类性能。

参考文献

[1] He, H. and Garcia, E.A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263-1284. https://doi.org/10.1109/TKDE.2008.239

[2] Zakaryazad, A. and Duman, E. (2016) A Profit-Driven Artificial Neural Network (ANN) with Applications to Fraud Detection and Direct Marketing. *Neurocomputing*, **175**, 121-131. https://doi.org/10.1016/j.neucom.2015.10.042

[3] Liu, G., Yang, Y. and Li, B. (2018) Fuzzy Rule-Based Oversampling Technique for Imbalanced and Incomplete Data Learning. *Knowledge-Based Systems*, **158**, 154-174. https://doi.org/10.1016/j.knosys.2018.05.044

[4] Lin, W.C., Tsai, C.F., Hu, Y.H., *et al*. (2017) Clustering-Based Undersampling in Class-Imbalanced Data. *Information Sciences*, **409-410**, 17-26. https://doi.org/10.1016/j.ins.2017.05.008

[5] 沈学华, 周志华, 吴建鑫, 等. Boosting 和 Bagging 综述[J]. 计算机工程与应用, 2000, 36(12): 31-32, 40.

[6] 张翔, 周明全, 耿国华, 等. Bagging 算法在中文文本分类中的应用[J]. 计算机工程与应用, 2009, 45(5): 135-137, 179.

[7] 毛国君, 段立娟. 数据挖掘原理与算法[M]. 第 3 版. 北京: 清华大学出版社, 2016.

[8] Bagui, S.C., Bagui, S., Pal, K. and Pal, N.R. (2003) Breast Cancer Detection Using Rank Nearest Neighbor Classification Rules. *Pattern Recognition*, **36**, 25-34. https://doi.org/10.1016/S0031-3203(02)00044-4

[9] Bagui, S.C. and Vaughn, B. (1998) Statistical Classification Based on k-Rank Nearest Neighbor Rule. *Statistical Decisions*, **16**, 181-189. https://doi.org/10.1524/strm.1998.16.2.181

[10] Gul, A., Perperoglou, A., Khan, Z., *et al*. (2018) Ensemble of a Subset of KNN Classifiers. *Advanced Data Analysis and Classification*, **12**, 827-840. https://doi.org/10.1007/s11634-015-0227-5

[11] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013(4): 1190-1197.