基于排序选择度量的自适应集成方法

沈先宝 宋余庆 刘 哲
(江苏大学计算机科学与通信工程学院 江苏 镇江 212013)

摘 要 针对集成过程中基分类器的集成优先性缺少精确化度量而导致的模型选择严谨性不高、系统精简性设计难以实现的问题,文中提出了一种基于排序选择度量方式、自适应权重设置的集成分类方法。首先,利用 K 折交叉验证及设计的误差熵与分类器互补性相结合的组合指数度量方法,选出集成优先性最高的两个分类器;然后,通过构造的以组合指数为基础的整体组合指数度量方法,实现对不同模型的优先性排序选择;最后,通过设置自适应权重的方式为不同模型找到最佳权重进行集成分类。在 UCI 数据集上的实验结果表明,所提方法与其他分类模型相比,各项分类评价指标均有提高,验证了该集成方法的可行性。该方法通过设计的模型选择定量性依据及自适应权重设置机制,使得整个集成分类系统具有模型选择分层性、可自适应精简化的特点。

关键词 排序选择,误差熵,互补性,组合指数,自适应权重
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/jsjkx.181102173

Adaptive Integrated Method Based on Sorting Selection Metrics

SHEN Xian-bao SONG Yu-qing LIU Zhe
(Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China)

Abstract Aiming at the problem that the rigor of model selection is not high and the system simplification design is difficult to achieve due to the lack of accurate measurement of the integration priority of the base classifier in the integration process, an integrated method based on sorting selection metrics and adaptive weighting setting was proposed. Firstly, the K-fold cross-validation and the combined index metric method constructed by combining the error entropy of the design and the complementarity of the classifier are utilized to select two classi-fiers with the highest integration priority. Then, considering the integration influence between the remaining candidate classifiers and the selected classifier subsets, the overall combination index metric based on combination index is constructed to realize the prioritization of different models. Finally, the best weights are found for different models for integration classification by adaptive weight method. The experimental results on the UCI dataset show that compared with other classification models, the classification evaluation indicators of the proposed method are improved, proving the feasibility of the integration method. This method selects quantitative basis of design model and adaptive weight setting mechanism through the designed model, making the whole integrated classification system have the stratification for model selection and the characteristics of adaptive simplification system.

Keywords Sorting selection, Error entropy, Complementarity, Combination index, Adaptive weight

1 引言

集成学习作为一个有效的机器学习范式,不仅能够提高分类精度,还能够提高系统的泛化性能和鲁棒性[1-2]。设计合理的分类器选择与融合方案已成为集成学习研究的关键点[3-7]。

文献[8]讨论了分类器组合和分数级融合的研究现状,并提出了一种将分类器输出转换为分数,以便分数级别的组合过程更有效工作的方法。

文献[9]通过采用 K-均值聚类结合动态选择循环框架生成混合模型的方法,可提高模型在多标签数据集上的性能。相比聚类,KNN 可以更精确地估计局部区域[10-11],因此其被用于估计测试样本的近邻集合,但使用 KNN 会涉及更高的计算成本。因此,有研究者考虑了一种自适应 KNN[12],其把竞争区域从类边界转移到类中心,以尽量减弱噪声样本的影响。也有学者采用遗传算法对每个基分类器进行权值优

到稿日期:2018-09-26 返修日期:2018-12-15 本文受国家自然科学基金项目(61572239,61772242),国家自然科学基金青年基金项目(61402204)资助。

沈先宝(1993-),男,硕士,主要研究方向为机器学习,E-mail:1633602665@qq.com;宋余庆(1959-),男,博士,教授,博士生导师,主要研究方向为机器学习、数据挖掘、医学图像处理与分析等,E-mail:yqsong@ujs.edu.cn(通信作者);刘 哲(1982-),女,博士,副教授,主要研究方向为数据挖掘、图像处理等。

化[13-14]，然后用设置好的阈值选择相应的分类器进行集成。还有基于基分类器的个体表现的集成，用于测量基分类器能力的标准有准确率、概率、数据复杂性[15]和元学习[16-19]等。上述研究方法先将分类器选择问题转化为最优化问题，采用各种优化算法找出合适的分类器子集进行集成，使得分类性能达到最佳。但不足的是，这类方法对基分类器的集成优先性缺少精确化度量，无法对后续集成研究提供参考优化基础。而且以上方法实现的待集成分类器集合确定后，结构固定，无二次更新系统的设计存在。若是对集成系统有精简、高效的要求，这些方法都有所欠缺。

基于上述原因，本文提出了一种基于排序选择度量方式、自适应权重设置的集成方法(Integrated method based on Sorting Selection Metrics and Adaptive Weight Settings，SSAW)。该方法以结合基分类器性能差异影响及互补性为设计基础，可对基分类器的集成优先性进行精确化度量，同时结合自适应权重设置方式，建立系统可精简化机制。

2 本文方法

本文提出的基于排序选择度量的自适应集成方法的整体思路为：首先进行 K 折交叉验证，并将设计的误差熵与分类器互补性相结合，构造出组合指数度量法，选择出集成优先性最高的两个分类器；然后以组合指数为基础，设计整体组合指数度量法，实现对余下基分类器的优先性排序选择；最后以自适应权重设置的方式为不同模型找到最佳权重并赋予集成。

2.1 排序选择度量

分类器的集成合理程度直观地反映了不同模型组合后得到的数据处理效果的好坏，根据选择优先性排序，可为后续集成及进一步研究提供参考。因此需要解决分类器选择的集成合理程度的度量问题。

K 折交叉验证可避免过拟合及欠拟合状态的发生，且能得到具有说服性的结果，可依此选择出第一个分类器。为首个分类器匹配集成效果最佳的双分类器组合，本文基于分类器个体差异性影响及互补配合程度，提出了一种将误差熵与互补性相结合、用于描述分类器组合效果的度量方法——组合指数。其主要思想如下：设有非空样本集 D，$N(D)$为样本数，待选择分类器 E_i 和 E_j 在样本集上的 K 次交叉验证得分列表分别为 y_{cvi} 和 y_{cvj}，y_{ik} 和 y_{jk} 分别表示 E_i 和 E_j 第 k 次交叉验证的得分，两分类器在整个训练集上的错分样本集合分别为 D_i 和 D_j。$\max(y_{cvi},y_{cvj})$表示 E_i 和 E_j 这 K 次交叉验证的最大值，d_{ik} 表示分类器 E_i 第 k 次交叉验证得分与最大值的误差绝对值，如式(1)所示：

$$d_{ik}=|y_{ik}-\max(y_{cvi},y_{cvj})| \tag{1}$$

根据上式求出误差和，可将各次误差值与该误差和进行比值计算，以代表差异程度出现的概率；再将该概率值代入信息熵中进行计算，结合信息熵具有的描述信息不确定性的特性，进而评判分类器对系统带来的差异性影响程度。计算分类器 E_i 的差异性影响熵值 M_i，如式(2)所示：

$$M_i=-\sum_{k=1}^{K}\frac{d_{ik}}{\sum_{k=1}^{K}d_{ik}}\ln\frac{d_{ik}}{\sum_{k=1}^{K}d_{ik}} \tag{2}$$

根据熵的极值性，分类器各次交叉验证差异程度出现的概率越接近，误差熵就越大，给系统带来的无序程度就越高，不同分类器之间的差异性就越大。

选择互补性强的分类器进行集成可以进一步提高分类效果。假设分类器 E_i 和 E_j 的误分类样本总数为 $N(D_i\cup D_j)$，均错误分类的样本数为 $N(D_i\cap D_j)$，二者差集的占比表示两个分类器的互补程度。占比值越大，两分类器的互补性就越强，反之，互补性越弱。

CI_{ij} 表示两个分类器 E_i 和 E_j 的组合指数，将分类器差异性影响熵与互补性相结合设计度量指标——组合指数，如式(3)所示：

$$CI_{ij}=\frac{N(D_i\cup D_j)-N(D_i\cap D_j)}{N(D)}*\frac{M_i+M_j}{2} \tag{3}$$

在上述定义中，CI_{ij} 体现的是两个分类器在组合效果上的综合衡量，值越大，说明这两个分类器集成组合性越好，反之则代表这两个分类器的组合效果欠佳，不宜被选择。由此，可选出组合性最佳的两个分类器。

除了考虑两个分类器间的组合性，还要考虑待选分类器与已有分类器集合间的集成性，从而进一步设计能够度量该情况的方法——整体组合指数。

假设已有分类器集合的分类器个数为 n，待选分类器为 E_k，则这 $n+1$ 个分类器的两两组合数为 C_{n+1}^2，则整体组合指数 OCI 如式(4)所示：

$$OCI=(\sum_{i=1,i\neq j}^{n+1}CI_{ij})/(2*C_{n+1}^2) \tag{4}$$

本文设计的排序选择度量算法的核心思想是根据组合指数和整体组合指数，对分类器进行排序选择。其过程为：首先利用 K 折交叉验证，从分类器集合中选择得分最高的分类器，放于排序集合的第一位；然后依据设计的组合指数，将与第一个分类器组合性最优的分类器排在第二位；最后依据整体组合指数，从剩余分类器集合里不断选择分类器，从而构造出排序集合。

依据此算法度量模型的集成优先性，若要求分类器集成个数为、，则选择序列中的前 n 个分类器组合即可，因此该算法实现的效果具有模型选择分层性的特点。

2.2 自适应权重设置

在分类器排序选择过程完成后，可以设计某种准则函数，采用最优化方法为模型找到最佳权重进行集成。

假设样本数为 N；标签类别数为 M；排序集合含有 m 个分类器；y_{ij} 表示若第 i 个样本属于第 j 类则取值为 1，否则取值为 0；p_{kij} 表示分类器排序集合中索引为 k 的分类器在第 i 个样本上被预测为第 j 类的概率；w_k 表示分类器排序集合中索引为 k 的分类器的权重。损失函数为模型在训练集上的预测值与标签值之间的平均损失值，定义如式(5)所示：

$$loss=-\frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}y_{ij}*\log(\sum_{k=0}^{m-1}w_k*p_{kij}) \tag{5}$$

排序集合中分类器的次序可代表其在初始阶段的权重地位，初始权值用统一差值递减设置，如式(6)所示：

$$1,\frac{m-1}{m},\cdots,\frac{2}{m},\frac{1}{m} \tag{6}$$

权重在向损失函数值不断减小的方向变化的过程中，其数值被归一化处理，表现不佳的分类器的权重将下降，若下降为 0 或接近 0，则该分类器就可以被舍弃，相当于进一步精简集成系统。本文的优化方法在程序编写中直接调用较为成熟

且高效的序列二次规划法库，因为该法库在许多实例测试中都能快速且准确地找到最优解。

3　实验分析

3.1　实验数据

本文在 UCI 数据集中依据数据量规模选择了 Breast_cancer，Financial_anti-fraud 和 Bank_customer_churn 3 个数据集进行实验，数据集的基本信息如表 1 所列。

表 1　数据集信息

Table 1　Information of data set

数据集	样本总数	特征维数	类别数目
Breast_cancer	683	11	2
Financial anti-fraud	95210	88	2
Bank_customer_churn	17240	179	2

为了验证本文方法的可行性，将主流模型如 KNN，C4.5，LR(Logistic regression)，SVM，RF(Random Forest)，ET(Extra-Trees)，GBDT 及其目前的先进版模型 XGB(XGBoost)作为本文方法框架下的基分类器列表进行建模，并将它们作为对比算法，在相同的数据集上进行实验对比。

3.2　评价指标

本文采用 5 个评价指标来测试本文方法的性能，各项指标的定义如下：

$$Accuracy=\frac{TP+TN}{TP+FN+FP+TN} \tag{7}$$

$$Precision=\frac{TP}{TP+FP} \tag{8}$$

$$Recall=\frac{TP}{TP+FN} \tag{9}$$

$$F1=2*Precision*\frac{Recall}{Precision+Recall} \tag{10}$$

3.3　实验结果统计与分析

AUC 值指 ROC 曲线下方的面积，范围介于 0.5 到 1 之间，值越大，分类模型的性能就越好。

其中，*TP* 表示正确预测的正例数，*FN* 表示被预测为负类的正例数，*TN* 表示正确预测的反例数，*FP* 表示被预测为正类的反例数。

(1)训练阶段

在训练集上，本文采用十折交叉验证，以获得较好的误差估计。各个模型在不同训练集上的交叉验证准确率均值如表 2 所列。

表 2　各个模型在不同训练集上交叉验证的准确率均值

Table 2　Average accuracy of cross-validation of each model on different training sets

(单位：%)

Model	Breast_cancer	Financial_anti-fraud	Bank_customer_churn
KNN	97.05	93.05	92.72
C4.5	93.37	89.67	98.63
LR	95.69	93.09	95.64
SVM	96.09	93.11	89.80
RF	95.49	93.12	98.87
ET	96.26	93.14	93.87
GBDT	96.46	93.05	98.94
XGB	97.06	93.15	98.76

由表 2 可知，各个模型在较小数据集上的准确率均值普遍高于较大数据集上的准确率均值，这是因为模型对小样本数据更容易充分学习，而模型本身的设计局限性无法保证其能更充分地学习大量数据的特征，从而导致准确率有所下降，这也说明了数据量对模型性能存在一定的影响。

各个模型在整个训练数据集上进行预测，得分数值与错分样本统计如表 3、表 4 所列。可以看出，得分越高，错分样本的个数越少。

表 3　各个模型在不同训练集上交叉验证后的得分

Table 3　Score of cross-validation of each model on different training sets

Model	Breast_cancer	Financial_anti-fraud	Bank_customer_churn
KNN	97.07	93.05	92.72
C4.5	93.94	89.65	98.65
LR	95.70	93.09	95.63
SVM	96.09	93.10	89.81
RF	96.48	93.11	98.92
ET	96.09	93.13	93.64
GBDT	96.48	93.06	98.93
XGB	97.07	93.15	98.74

表 4　各个模型在不同训练集上的错分样本数

Table 4　Number of misclassified samples of each model on different training sets

Model	Breast_cancer	Financial_anti-fraud	Bank_customer_churn
KNN	15	4633	627
C4.5	31	6895	104
LR	22	4605	375
SVM	20	4593	879
RF	18	4590	83
ET	20	4578	548
GBDT	18	4628	80
XGB	14	4562	96

(2)排序选择度量阶段的相关结果

依据表 2—表 4，选取得分最高的分类器作为集成选择的第一个分类器，将剩余的分类器与第一个分类器计算组合指数，结果如表 5 所列，表中的 * 表示该分类器已被选择。

表 5　不同训练集上最优模型与待考查模型的组合指数

Table 5　Combination index of optimal model and model to be investigated on different training sets

Model	Breast_cancer	Financial_anti-fraud	Bank_customer_churn
KNN	0.0229	0.0020	0.0834
C4.5	0.0886	0.0859	0.1052
LR	0.0349	0.0011	0.1160
SVM	0.0626	0.0346	0.0182
RF	0.0441	0.0009	0.0686
ET	0.0273	0.0005	0.0023
GBDT	0.0122	0.0021	*
XGB	*	*	0.1318

依据表 5，只需在同一数据集下找到组合指数值最大的分类器，其即为与最优分类器组合性最好的分类器。计算剩余模型的整体组合指数，并建立分类器排序集合，结果如表 6、表 7 所列。

表 6 不同训练集上已选模型子集与待考查模型的整体组合指数

Table 6 Overall combination index of selected model subsets and models to be investigated on different training sets

Model	Breast_cancer	Financial_anti-fraud	Bank_customer_churn
KNN	0.0314	0.0290	0.0407
C4.5	*	*	0.0306
LR	0.0368	0.0127	0.0467
SVM	0.0440	0.0213	0.0386
RF	0.0411	0.0179	0.0390
ET	0.0337	0.0151	0.0318
GBDT	0.0291	0.0221	*
XGB	*	*	*

表 7 分类器排序选择结果

Table 7 Sorting selection results of classifier

优先次序	Breast_cancer	Financial_anti-fraud	Bank_customer_churn
1	XGB	XGB	GBDT
2	C4.5	C4.5	XGB
3	SVM	KNN	LR
4	RF	GBDT	KNN
5	LR	SVM	RF
6	ET	RF	SVM
7	KNN	ET	ET
8	GBDT	LR	C4.5

依据表 6 和表 7，被优先选择的分类器整体组合指数较大，而后被选择的分类器的值会越来越小，因为整体组合指数的设计考量了当前待考查分类器与已选分类器集合的集成效果。随着被选择分类器的增多，已选分类器集合的子规模会增大，这部分集合相比单个模型在集成系统中的作用也会越来越大，因此，后面待考查分类器的作用将显得逐渐弱化，反映在其数值上是逐渐减小。将排序选择结果显示出来，可清楚地观察到在特定数据集上更适宜集成的模型，方便进一步深入研究。

(3)自适应集成

在排序集合基础上，对各个基分类器权重进行分配以找到最佳权重，如表 8 所列。

表 8 不同数据集上模型的权重

Table 8 Weights of models on different datasets

Model	Breast_cancer	Financial_anti-fraud	Bank_customer_churn
KNN	0.1251	0.1434	0.3987
C4.5	0.1252	0.1429	0.0019
LR	0.1253	0.1406	0.0021
SVM	0.1252	0.1343	0.0003
RF	0.1245	0.1400	0.0016
ET	0.1252	0.1401	0.0028
GBDT	0.1240	0.1354	0.2615
XGB	0.1254	0.1633	0.3012

观察数据集 Breast_cancer 和 Financial_anti-fraud 上的权重分布，各个模型权重虽有差异但大致接近，说明为了达到性能最佳，各模型需要全部参与集成。而在数据集 Bank_customer_churn 上，各模型的权重比例出现差异，部分的分类器权重十分接近 0，这些分类器可以被舍弃，最终仅保留 KNN，GBDT 和 XGB 3 个分类器进行集成，从而精简集成系统。

(4)评价指标

依据表 9－表 11，在较小数据集上，各个模型的分类性能普遍较好，而本文设计的基于排序选择度量方式、自适应权重设置的集成方法进一步将各项指标大约提高了 2%，但因数据量小，其效果还不能完全体现。当数据增多时，各个模型的分类性能都有所下降，而本文设计的方法仍保证了各项指标的提高，从而验证了该方法的可行性。

表 9 Breast_cancer 测试集上的评价指标

Table 9 Evaluation index on Breast_cancer test set

Model	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
KNN	0.98	0.99	0.99	0.99	0.98
C4.5	0.95	0.96	0.96	0.96	0.95
LR	0.98	0.99	0.99	0.99	0.98
SVM	0.97	0.98	0.98	0.98	0.98
RF	0.98	0.98	0.98	0.98	0.97
ET	0.98	0.98	0.98	0.98	0.98
GBDT	0.97	0.97	0.97	0.97	0.98
XGB	0.98	0.98	0.98	0.98	0.98
SSAW	0.99	0.99	0.99	0.99	0.98

表 10 Financial_anti-fraud 测试集上的评价指标

Table 10 Evaluation index on Financial_anti-fraud test set

Model	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
KNN	0.94	0.92	0.94	0.93	0.62
C4.5	0.91	0.92	0.91	0.91	0.61
LR	0.94	0.92	0.94	0.93	0.67
SVM	0.94	0.92	0.93	0.92	0.69
RF	0.94	0.93	0.94	0.93	0.65
ET	0.94	0.94	0.94	0.94	0.67
GBDT	0.94	0.93	0.94	0.93	0.76
XGB	0.94	0.94	0.94	0.94	0.78
SSAW	0.98	0.97	0.98	0.97	0.86

表 11 Bank_customer_churn 测试集上的评价指标

Table 11 Evaluation index on Bank_customer_churn test set

Model	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
KNN	0.92	0.91	0.92	0.91	0.87
C4.5	0.93	0.92	0.92	0.92	0.89
LR	0.94	0.95	0.95	0.94	0.93
SVM	0.90	0.81	0.90	0.85	0.78
RF	0.95	0.94	0.94	0.94	0.94
ET	0.94	0.93	0.93	0.93	0.92
GBDT	0.96	0.95	0.94	0.94	0.94
XGB	0.96	0.96	0.95	0.95	0.95
SSAW	0.98	0.97	0.96	0.96	0.96

结束语 针对集成过程中基分类器的集成优先性缺少精确化度量、系统精简性设计难以实现等问题，本文提出了一种基于排序选择度量方式、自适应权重设置的集成方法，该方法能兼顾利用不同基分类器的个体差异性影响及组合性两方面因素，构造出对目标有较高识别率的分类器排序集合，为后续的集成提供参考依据。同时，在此基础上，能利用自适应方式为模型找到最佳权重，使集成分类器系统达到性能最佳，进而得到目标对象的分类结果。实验结果表明，与其他模型相比，本文方法的各项指标均有提高，验证了其可行性。但本文方

法主要是针对二分类问题提出的，未来将对如何提高多分类问题的准确率进行进一步研究。

参 考 文 献

[1] FUMERA G，FABIO R，ALESSANDRA S. A theoretical analysis of bagging as a linear combination of classifiers[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2008，30(7)：1293-1311.

[2] FARID D M，AL-MAMUN M A，MANDERICK B，et al. An adaptive rule-based classifier for mining big biological data[J]. Expert Systems with Applications，2016，64(C)：305-316.

[3] KUNCHEVA L I. A theoretical study on six classifier fusion strategies[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2002，24(2)：281-286.

[4] CERNADAS E，AMORIM D. Do we need hundreds of classifiers to solve real world classification problems? [J]. Journal of Machine Learning Research，2014，15(1)：3133-3181.

[5] POLIKAR R. Ensemble based systems in decision making[J]. IEEE Circuits & Systems Magazine，2006，6(3)：21-45.

[6] JR A S B，SABOURIN R，OLIVEIRA L E S. Dynamic selection of classifiers — A comprehensive review[J]. Pattern Recognition，2014，47(11)：3665-3680.

[7] RAYHAN F，AHMED S，SHATABDA S，et al. iDTI-ESBoost：Identification of Drug Target Interaction Using Evolutionary and Structural Features with Boosting[J]. Scientific Reports，2017，7(1)：17731.

[8] CHITROUB S. Classifier combination and score level fusion：concepts and practical aspects[J]. International Journal of Image and Data Fusion，2010，1(2)：113-135.

[9] LIN C，CHEN W，QIU C，et al. LibD3C：Ensemble classifiers with a clustering and dynamic selection strategy[J]. Neurocomputing，2014，123(1)：424-435.

[10] CRUZ R M O，SABOURIN R，CAVALCANTI G D C. Analyzing different prototype selection techniques for dynamic classifier and ensemble selection[C]// International Joint Conference on Neural Networks. Anchorage，AK，USA：IEEE，2017：3959-3966.

[11] DE SOUTO M C P，SOARES R G F，SANTANA A，et al. Empirical comparison of Dynamic Classifier Selection methods based on diversity and accuracy for building ensembles[C]// IEEE International Joint Conference on Neural Networks. Hong Kong，China：IEEE，2008：1480-1487.

[12] CRUZ R M O，SABOURIN R，CAVALCANTI G D C. Prototype selection for dynamic classifier and ensemble selection[J]. Neural Computing & Applications，2016，29(2)：1-11.

[13] FARID D M，ZHANG L，RAHMAN C M，et al. Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks[J]. Expert Systems with Applications An International Journal，2014，41(4)：1937-1946.

[14] SUN Z，SONG Q，ZHU X，et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition，2015，48(5)：1623-1637.

[15] BRUN A L，JR A S B，OLIVEIRA L S，et al. Contribution of Data Complexity Features on Dynamic Classifier Selection[C]// International Joint Conference on Neural Networks. Vancouver，BC，Canada：IEEE，2016：4396-4403.

[16] CRUZ R M O，SABOURIN R，CAVALCANTI G D C. On Meta-learning for Dynamic Ensemble Selection[C]// International Conference on Pattern Recognition. Stockholm，Sweden：IEEE Computer Society，2014：1230-1235.

[17] PINTO F，SOARES C，MENDES-MOREIRA J. CHADE：Metalearning with Classifier Chains for Dynamic Combination of Classifiers[M]// Machine Learning and Knowledge Discovery in Databases. Springer International Publishing，2016：410-425.

[18] CRUZ R M O，SABOURIN R，CAVALCANTI G D C. META-DES. Oracle：Meta-learning and feature selection for dynamic ensemble selection[J]. Information Fusion，2017，38(1)：84-103.

[19] XIAO Z H，YU H，WANG Y C. Unsupervised Machine Learning Based on the Sun of Squares of the Dynamic Deviations[J]. Journal of Chongqing University of Technology(Natural Science)，2018，32(11)：134-139，186.(in Chinses)
肖枝洪，于浩，王一超. 基于动态离差平方和准则的无监督机器学习[J]. 重庆理工大学学报(自然科学)，2018，32(11)：134-139，186.