文章编号：2095-2163(2021)02-0084-04 中图分类号：TP391 文献标志码：A

基于 K-medoids 聚类的贝叶斯集成算法

盛静文，于艳丽，江开忠
（上海工程技术大学 数理与统计学院，上海 201620）

摘 要： 朴素贝叶斯分类算法由于其计算高效在生活中应用广泛。本文根据集成算法的差异性特征，聚类算法聚类点的选择方式的可变性，提出了基于 K-medoids 聚类技术的贝叶斯集成算法，朴素贝叶斯的泛化性能得到了提升。首先，通过样本集训练出多个朴素贝叶斯基分类器模型；然后，为了增大基分类器之间的差异性，利用 K-medoids 算法对基分类器在验证集上的预测结果进行聚类；最后，从每个聚类簇中选择泛化性能最佳的基分类器进行集成学习，最终结果由简单投票法得出。将该算法应用于 UCI 数据集，并与其他类似算法进行比较可得，本文提出的基于 K-medoids 聚类的贝叶斯集成算法（NBKME）提高了数据集的分类准确率。

关键词： 朴素贝叶斯；分类；K-medoids 聚类；集成算法

Research on Bayesian ensemble algorithm based on K-medoids clustering

SHENG Jingwen, YU Yanli, JIANG Kaizhong
(School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 The Naive Bayes classification algorithm is widely used in life due to its computational efficiency. Based on the difference characteristics of the ensemble algorithm and the variability of the clustering point selection method of the clustering algorithm, this paper proposes a Bayesian ensemble algorithm based on K-medoids clustering technology, and the generalization performance of Naive Bayes has been improved. Firstly, multiple Naive Bayesian classifier models are trained through the sample set; then, in order to increase the difference between the base classifiers, the K-medoids algorithm is used to gather the prediction results of the base classifiers on the validation set; Finally, the base classifier with the best generalization performance is selected from each cluster for ensemble learning, and the final result is obtained by a simple voting method. The algorithm is applied to UCI data set and compared with other similar algorithms. The Bayesian ensemble algorithm based on K-medoids clustering (NBKME) proposed in this paper improves the classification accuracy of the data set.

【Key words】 Naive Bayes; classification; K-medoids clustering; ensemble algorithm

0 引 言

分类问题在机器学习理论中占据重要地位，在监督学习中也是一个核心问题。在日常生活中，可见到很多分类问题，比如在医学方面，将病人的检查结果分为阳性和阴性，这就是一个二分类问题；在电子邮箱中，邮件的种类有 3 种，可能被分为垃圾邮件、广告邮件和正常邮件，这就是一个多分类问题；图像中的图像分割，最简单的方法就是对每一个像素进行分类。因此，分类问题是机器学习的基础，研究分类算法对于社会发展具有不可替代的作用。

在众多分类算法中，朴素贝叶斯方法[1]是一种基于概率的简单有效的机器学习分类方法，现已广泛应用在数据挖掘、计算机视觉、自然语言处理、模式识别、生物特征识别等领域。在很多的场合中，贝叶斯分类算法的性能与神经网络和决策树等算法比较接近，甚至要更好。贝叶斯算法不仅简单，而且当数据集的规模很大时，分类的正确率也很高。同时，还可以精确地用数学公式表示出来。集成学习[2]，是指结合多个学习器进行学习任务的一种机器学习方法，也称为分类器的集成。该方法可以对线性回归、决策树、支持向量机等基学习器进行集成训练，令性能较单一学习器有较大的提升。集成学习主要分为 2 类。其中一类是由 Breiman[3]提出的一种叫 Bagging 的集成方法。该方法的主要思想是通过自助法（Bootstrap）从训练集中抽取 N 个训练子集，然后对这 N 个训练子集进行训练可以生成 N 个基学习器，最终结果由这 N 个基学习器投票或平均的方式得出，如此一来不仅提高了模型学习的精度，而且也降低了过拟合的风险。集成学习[4]通常适用于不稳定的学习算法，如决策树算法、BP 神经网络算法等，而朴素贝叶斯是一种稳定的学习方法，因此本

作者简介：盛静文（1995-），女，硕士研究生，主要研究方向：机器学习、大数据；于艳丽（1993-），女，硕士研究生，主要研究方向：不平衡数据；江开忠（1965-），男，博士，副教授，主要研究方向：大数据。
通讯作者：盛静文 Email：sjw103606@163.com

收稿日期：2020-09-30

文主要研究如何破坏朴素贝叶斯算法的稳定性并提高基分类器之间的差异性来进行集成，从而提高泛化性能。

针对贝叶斯集成算法，已经有很多学者做了相关的研究。文献[5]在房价评估模型中引入集成学习和贝叶斯优化，模型精度提升了 3.15 个百分点；文献[6]提出了一种深度集成朴素贝叶斯模型，缓解了朴素贝叶斯特征独立性假设的缺点；文献[7]建立动态随机树贝叶斯回归模型，并通过集成（平均）来提高回归模型的泛化能力；文献[8]提出一种基于贝叶斯集成的 SAR 目标识别方法，有效解决了同一类目标数据可能会因为不同的表现形式产生错误的识别的问题。

本文在钟熙等人[9]研发的基于 Kmeans++聚类的朴素贝叶斯分类器集成方法的基础上提出了基于 K-medoids 的贝叶斯集成算法。首先，通过对初始训练样本集自助采样形成样本子集，对样本子集随机抽取属性形成属性子集来训练基分类器，把每个基分类器在验证集上的预测结果作为聚类数据；然后，使用 K-medoids 算法对结果数据进行聚类，从而得到若干个基分类器簇，统计每个簇中的所有基分类器在验证样本集上的泛化性能，并选择每个簇中泛化性能最好的基分类器代表该簇；最后，对选择出来的基分类器进行集成，通过简单投票法得到最终的预测结果。对于本文提出的算法的有效性，在 UCI 数据集上得到了验证。

1　相关理论

1.1　朴素贝叶斯分类模型

朴素贝叶斯分类器[10]是一种以概率统计为基础的学习方法，基于假设特征之间的强独立性，需符合贝叶斯假设。假设有类别 $C(c_1,c_2,\cdots,c_n)$，特征属性为 $X(x_1,x_2,\cdots,x_n)$，朴素贝叶斯是指在 X 相互独立的条件下，对于需要分类的项目，在该事件发生的条件下，哪个类别出现的概率大，被认为属于哪个类别。

朴素贝叶斯的数学公式为：

$$p(c\mid x)=\frac{p(x\mid c)p(c)}{p(x)}=\frac{p(c)}{p(x)}\prod_{i=1}^{d}p(x_i\mid c)\ . \tag{1}$$

其中，d 表示条件属性的数量；x 表示条件属性；c 表示类别。

研究假设朴素贝叶斯各个条件是相互独立的，所以 $p(x\mid c)$ 可以变成 $\prod_{i=1}^{d}p(x_i\mid c)$，在 x 属性下，概率值最高的 c 就是 x 所属类别。

1.2　集成算法

Bagging[3,11]是并行式集成学习方法最著名的代表。通过有放回的独立重复抽样得到 Bootstrap 样本，通过训练得到多个不同的基分类器，再将这些基分类器进行结合得到最终的分类器。在对预测输出进行结合时，对于分类任务，Bagging 使用简单投票法；对于回归任务，Bagging 使用简单平均法。

Bagging 的算法流程如下：

输入：样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$

输出：最终的分类器 $f(x)$

（1）对于 $t=1,2,\cdots,T$：

① 对训练集进行第 t 次随机采样，共采集 m 次，得到包含 m 个样本的采样集 D_t；

② 用采样集 D_t 训练第 t 个弱学习器 $G_t(x)$。

（2）如果是分类算法，则根据简单投票法得出最终类别。如果是回归算法，最终的模型输出结果为多个弱学习器的回归结果的平均值。

1.3　K-medoids 聚类算法

K-medoids 聚类算法[12-13]是 K-means 聚类算法的一种改进，K-means 算法是在一组给定的数据中随机选取 k 个数据作为初始聚类中心，接着计算所有数据与初始聚类中心的相似度，将数据放入与其最相似的聚类中心所代表的类；接下来，根据每个类内的数据的均值更新聚类中心点并重新划分数据集中的数据，不断迭代上述过程直至数据集中的数据所属的类不再发生变化或者直至算法收敛为止[14]。K-medoids 聚类算法与 K-means 不同的是在算法迭代过程中，K-medoids 聚类算法选择与类簇内数据均值距离最近的数据对象作为新的聚类中心点。

K-medoids 的具体流程如下：

（1）任意选取 K 个对象作为初始聚类中心点。

（2）将余下的对象分到各个类中去（根据与中心点最相近的原则）。

（3）在每个类中，顺序选取一个，计算可用于代替的总代价。选择最小的来代替。这样 K 个 medodis 就改变了。

（4）重复（2）、（3），直到 K 个 medoids 固定下来。

2 基于聚类的选择性朴素贝叶斯集成

差异性是集成算法的基础,一般源于基分类器的训练样本,这样同一算法在不同训练样本下即可训练出不同的基分类器。经典的集成方法 Bagging 和 Adaboost 都是同一算法在不同设置下的集成,其差异性源于训练样本的不同。将一个大的训练样本集根据相似度分成多个组,在每组进行分类训练,这样就在分类器训练前显性地增加了差异性,从而不用在训练过程中考虑如何平衡差异性和分类正确率的问题,因为在训练样本划分时分类器训练前即对差异性做了保证,而分类正确率即成为在分类器训练过程中唯一需要考虑的因素[15-17]。

基于 K-medoids 聚类的贝叶斯集成算法主要分为如下方面:

(1)使用 Bagging 集成算法,进行自主采样获得样本子集,再对样本子集随机抽取一定比例的属性获得属性子集,这就使得朴素贝叶斯变得不稳定,进而基分类器差异性较大。

(2)本文在 K-means 的基础上改进采用 K-medoids 算法对测试集的结果进行聚类,以达到朴素贝叶斯分类模型聚类的目的,从而获得差异性更大的基分类器。

(3)进行选择性集成。选择性集成不仅可以提高算法性能,还能加快预测速度。从聚类产生的每个簇中选出泛化性能最佳的基分类器进行集成,并利用简单投票法得出最终的预测结果。

算法的主要步骤如下:

(1)输入样本集 $D = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,其中 x_n 为样本属性,y_n 为样本标签,将数据集分为训练集与测试集。

(2)假定基分类器个数为 T,簇的个数为 k。

(3)对样本集中的数据进行自主采样得到 T 个样本子集,再对样本子集随机抽取 1/2 个属性子集,从而获得基分类器的训练集 D_t。

(4)利用训练集 D_t 训练基分类器 p_t。

(5)对测试集进行分类,并根据分类结果计算准确率。

(6)利用 K-medoids 算法对所有基分类器的分类结果进行聚类,选出每个聚类簇中分类精确度最高的基分类器进行集成。

(7)输出。研究推得的数学公式可写为:

$$P(x) = \underset{y \in Y}{\operatorname{argmax}} \sum_{1}^{T} \| p_t(x) = y \| . \qquad (2)$$

3 实 验

3.1 实验数据

本文采用 UCI 数据库中的部分数据集,算法过程通过 Python 实现。为了验证本文提出的基于 K-medoids 聚类的贝叶斯集成算法(NBKME)的优势,将其与朴素贝叶斯算法(NB)、NBFS(即用朴素贝叶斯做基分类器,训练样本集是从样本子集中随机抽取属性形成的属性子集的 Bagging 算法)、NBK(基于 K-means 聚类的贝叶斯集成算法)、NBKM(基于 K-means++聚类的贝叶斯集成算法)进行比较。数据集见表 1。

表 1 数据集

Tab. 1 Dataset

数据集	样本数	特征数	分类数
german	1 000	24	2
wine	178	13	3
blood	748	4	2
CMC	1 473	9	3
Tic-Tac-Toe	958	9	2
heart	270	13	2
diabetes	768	8	2
cancer	683	9	2

3.2 实验结果及分析

本文选用的基分类器为朴素贝叶斯分类器,基分类器的数目为 100,聚类算法簇的个数为 10,在每个数据集上进行 10 次实验,实验结果见表 2。

表 2 分类精度对比

Tab. 2 Comparison of classification accuracy %

数据集	NB	NBFS	NBKM (kmeans++)	NBK (kmeans)	NBKME (k-medoids)
german	0.752	0.718	0.768	0.757	0.780
wine	1	0.978	0.981	0.988	0.985
blood	0.732	0.755	0.786	0.777	0.793
CMC	0.474	0.476	0.508	0.497	0.512
Tic-Tac-Toe	0.717	0.713	0.756	0.742	0.760
heart	0.823	0.850	0.876	0.875	0.882
diabetes	0.818	0.757	0.773	0.775	0.795
cancer	0.959	0.962	0.968	0.963	0.988

由表 2 可以看出 NBFS 的分类准确度高于 NB,这就说明在 Bagging 算法中随机抽取属性子集可以提高分类准确率,集成算法的泛化性能也提高了;而 NBK 算法和 NBKM 算法对所有基分类器的结果进行聚类,有效增强了基分类器间的差异性,分类准确

率也有所提高。而本文提出的基于 NBKME 算法将 NBK 和 NBKM 算法中的聚类方式变换为 K-medoids。根据表 2 中的结果可见，聚类中心选择方式的改变也进一步提高了分类准确率。

4　结束语

建立科学有效的信用评估模型，能够为研究人员提供重要的决策支持，减少损失，意义十分重大。针对传统的朴素贝叶斯分类算法的改进，本文提出的基于 K-medoids 聚类的贝叶斯集成算法，通过属性子集的选取，朴素贝叶斯的聚类和聚类中心点选取方式的改变提高了分类器的泛化性能，分类准确率更高。下一步的工作可以围绕分类器的运行效率和样本子集的选取方式展开。

参考文献

[1] 杜选. 基于加权补集的朴素贝叶斯文本分类算法研究[J]. 计算机应用与软件，2014，31(9)：253-255.

[2] TANG Jialin，SU Qinglang，SU Binghua，et al. Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition [J]. Computer Methods and Programs in Biomedicine，2020，197：105622.

[3] BREIMAN L. Bagging predicators [J]. Machine Learning，1996，24(1)：123-140.

[4] 刘振刚. 集成学习在三维模型分类中的应用[J]. 计算机产品与流通，2020(4)：261.

[5] 顾桐，许国良，李万林，等. 基于集成 LightGBM 和贝叶斯优化策略的房价智能评估模型[J]. 计算机应用，2020，40(9)：2762-2767.

[6] 吴皋，李明，周稻祥，等. 基于深度集成朴素贝叶斯模型的文本分类[J]. 济南大学学报(自然科学版)，2020，34(5)：436-442.

[7] 王双成，郑飞，唐晓清. 动态随机树贝叶斯集成回归模型研究[J]. 小型微型计算机系统，2019，40(4)：715-720.

[8] 陈博，王珺琳，刘长清，等. 基于贝叶斯集成算法的仿真 SAR 目标识别方法[J]. 中国电子科学研究院学报，2017，12(1)：73-77.

[9] 钟熙，孙祥娥. 基于 Kmeans++聚类的朴素贝叶斯集成方法研究[J]. 计算机科学，2019，46(S1)：439-441，451.

[10] 漆原，乔宇. 针对朴素贝叶斯文本分类方法的改进[J]. 电子科学技术，2017，4(5)：114-116，129.

[11] 孟小燕. 基于属性权重的 Bagging 回归算法研究[J]. 现代电子技术，2017，40(1)：95-98，103.

[12] 韩冰，姜合. 基于相似度计算公式改进的 K-中心点算法[J]. 计算机与现代化，2019(5)：113-117.

[13] YU Donghua，LIU Guojun，GUO Maozu，et al. An improved K-medoids algorithm based on step increasing and optimizing medoids[J]. Expert Systems With Applications，2018，92：464-473.

[14] XU Xingbang，CHI Nan. The phase estimation of geometric shaping 8-QAM modulations based on K-means clustering in underwater visible light communication [J]. Optics Communications，2019，444：147-153.

[15] 徐浩然，许波，徐可文. 机器学习在股票预测中的应用综述[J]. 计算机工程与应用，2020，56(12)：19-24.

[16] 武建军，李昌兵. 基于互信息的加权朴素贝叶斯文本分类算法[J]. 计算机系统应用，2017，26(7)：178-182.

[17] MURALIDHARAN V，SUGUMARAN V. A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis [J]. Applied Soft Computing，2012，12(8)：2023-2029.

(上接第 83 页)

detection approach based on shadow feature with multichannel high-resolution synthetic aperture radar [J]. IEEE Geoence & Remote Sensing Letters，2017，13(10)：1572-1576.

[5] REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：Unified，real-time object detection [J]. arXiv preprint arXiv：1506.02640，2015.

[6] REDMON J，FARHADI A. YOLO9000：Better，Faster，Stronger [C] // IEEE Conference on Computer Vision & Pattern Recognition. Honolulu，Hawaii：IEEE，2017：6517-6525.

[7] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single shot multiBox detector [C] // European Conference on Computer Vision. Springer International Publishing. Cham：Springer，2016：21-37.

[8] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//2014 IEEE Conference on Computer Vision and Pattern Recogntion. Ohio，USA：IEEE，2014：580-587.

[9] GIRSHICK R. Fast R-CNN [C] // Proceedings of the IEEE International Conference on Computer Vision. Santiago，Chile：IEEE，2015：1440-1448.

[10] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards real-Time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39(6)：1137-1149.

[11] LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature pyramid networks for object detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI：IEEE，2017：936-944.

[12] REDMON J，FARHADI A. YOLOv3：An incremental improvement [J]. arXiv preprint arXiv：1804.02767，2018.

[13] HOWARD A G，ZHU M，CHEN B，et al. MobileNets：Efficient Convolutional Neural Networks for mobile vision applications [J]. arXiv preprint arXiv：1704.04861，2017.

[14] 李汉冰，徐春阳，胡超超. 基于 YOLOV3 改进的实时车辆检测方法[J]. 激光与光电子学进展，2020，57(10)：332-338.

[15] 刘寒迪，赵德群，陈星辉，等. 基于改进 SSD 的航拍施工车辆检测识别系统设计[J]. 国外电子测量技术，2020，39(7)：127-132.

[16] IANDOLA F N，HAN S，MOSKEWICZ M W，et al. SqueezeNet：AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size [J]. arXiv preprint arXiv：1602.07360，2016.

[17] 李双峰. TensorFlow Lite：端侧机器学习框架[J]. 计算机研究与发展，2020，57(9)：1839-1853.