DOI:10.11721/cqnuj20180616

基于 k-means++的多分类器选择分类研究*

熊 霖,唐万梅
(重庆师范大学 计算机与信息科学学院,重庆 401331)

摘要:【目的】机器学习中不同算法适用于具有不同分布特征的数据集。在用整个训练集上训练得到的单个分类器预测新样本类别时,由于缺少对局部区域样本的针对性,可能导致分类器对某一区域数据的预测能力较差而产生错误分类。为了解决这个问题,提出基于 k-means++的多分类器选择算法。【方法】首先用 3 种分类综合性能较好的算法——AdaBoost、SVM、随机森林(RF)在训练集上分别训练得到 3 个分类器作为候选基分类器,然后利用 k-means++算法将训练数据集分为 k 个簇,用 3 个候选分类器分别对每个簇进行分类测试,选择对这一簇中数据分类精度最高的分类器作为与它的数据相似数据的分类器。在对新样本进行类别预测时,首先判定样本属于哪个簇,然后用它的分类器进行分类预测。【结果】实验结果表明,新算法在 9 个 UCI 数据集上优于单个分类算法。【结论】基于局部区域动态选择最优分类器可以提高模型分类准确性。

关键词:局部区域;AdaBoost;k-means++;随机森林;SVM
中图分类号:TP181 **文献标志码:**A **文章编号:**1672-6693(2018)06-0088-09

分类是模式识别、机器学习和数据挖掘等领域的一个重要研究方向。所谓分类是通过对已有的数据进行学习,提取数据中蕴涵的规律,然后利用所获得的规律(分类器)对新数据进行类别预测。一个好的分类器不仅仅能够对当前数据有很好的分类效果,还要对新的数据有很好的分类能力,即泛化能力。不同的数据集分布各异,有些呈现高斯分布、而有些呈现多个簇的分布等。研究表明[1],对于某些分类算法难分的数据集,采用其他的分类算法,却能得到很好的分类效果。也就是说,不同的分类算法对于不同的数据集具有不同的适应性。当对呈现多个簇分布的数据采用在整个数据集上学习的单个分类器进行分类时,由于缺少局部的针对性,可能会导致分类器对于不同的簇有不同的适应性:有些簇的分类效果很好,而对于有些簇的分类效果却很差。

1 相关研究

由于各个分类方法具有自己的局限性和优势,文献[2-3]指出:在解决复杂问题时使用多个不同的分类器构建多分类器系统,可以获得比单一分类器更好的预测效果。多分类器系统和传统单个分类方法不同,多分类器系统并非得到一个最优分类器,而是通过组合多个分类器得到最优的分类模型。但多个分类器会使模型过于庞大,不但影响预测速度而且会占用较大的存储空间。Zhou 等人[4]证明了可以用较少的分类器集成获得相同或更好的效果。针对分类器和数据适应性问题,可以采用动态选择分类器的方法来解决。窦鹏[5]用不同的分类算法训练多个基分类器,提出了基于识别性能矩阵投票的集成、基于类别权重的加权投票和基于全信息相关度的多分类投票 3 种投票集成算法,在遥感图像分类上获得了很好的分类效果。饶川等人[6]针对高速列车故障样本数据分布不均匀等情况,提出了基于选择性集成的 SVM 多分类器融合算法。这一算法选取测试样本最邻近的 k 个训练样本训练多个 SVM 分类器,然后对分类效果好的 SVM 分类器进行融合,能够获得比单个算法(如普通 SVM、KNN、Bayes)更好的分类效果。但这种方法在训练分类器时没有在整个数据集上进行学习,这会导致部分信息丢失,所以在某些时候分类效果不如普通的单个算法。为了使分类器得到较好的分类能力,训练集应尽

* **收稿日期:**2018-02-28 **修回日期:**2018-09-06 **网络出版时间:**2018-10-25 10:41
资助项目:重庆市自然科学基金(No. CSCT2015JCYJA00005);重庆市研究生教育教学改革研究重点项目(No. yjg152001);重庆市教改项目(No. yjg123040);重庆师范大学校级项目(No. xyjg16005;No. 201625)
第一作者简介:熊霖,男,研究方向为机器学习与数据挖掘,E-mail:921147163@qq.com;**通信作者:**唐万梅,女,教授,博士,E-mail:cqtwm@163.com
网络出版地址:http://kns.cnki.net/kcms/detail/50.1165.N.20181025.1041.010.html

可能的大，从而需要分类器在整个训练集上进行学习。王文哲等人[7]先将数据特征空间分为粗糙区域、确定区域和不确定 3 个区域，识别阶段用粗糙区域已知的数据训练 AdaBoost 分类器用以识别预测未知数据类别。文献[8]提出了基于动态权重的 AdaBoost 改进算法：先将训练集样本进行聚类，分析簇与学习得到的一系列基分类器的适应性，从而确定各个基分类器在每个簇上的权重，最后根据测试数据与簇之间的相似性来计算各个基分类器用于分类测试样本所占的权重，这种方法得到了比 AdaBoost 更好的分类效果。李凯等人[9]针对分类器差异性提出了基于 k-means 聚类技术的神经网络分类器集成方法，即先用神经网络在训练集上学习得到多个神经网络分类器，在测试集上用每个分类器的分类结果作为聚类对象，用 k-means 算法对这些分类结果进行聚类以度量基分类器之间的差异性。在每个簇中选择一个代表分类器，构成集成模型的分类器成员，这样可以去除冗余的分类器，得到一个更精简的集成模型。

集成模型多数时候可以提升分类性能，然而对于某些数据，只有少量的分类器能将其正确分类，大多数分类器会分错，这时候集成会起负作用。文献[10]中论证了多分类系统中是将多分类器组合集成起来还是选择单个好的分类器问题，得出的结论是根据不同的分类器在不同数据区域识别能力的好坏，直接选择最适应于该区域数据的分类器进行分类。Didaci 等人[11-12]提出了一种基于 KNN 的局部精度的动态分类器选择方法，选用的分类算法是线性贝叶斯、二次贝叶斯和 KNN 作为基分类器，但选用的基分类算法精度并不高。随机森林(RF)在决策树的基础上，通过随机选择属性得到多个具有多样性的决策树分类器，这些分类器多样性不仅来自于属性扰动，还来自于样本扰动，从而展现了强大的性能。文献[13]用多个指标对比了有监督学习中的多个分类算法，结果显示 RF 最好。文献[14]用包括 17 个类别的 179 个分类器在 UCI 上的 121 个数据集上进行分类实验，得出的结论是 RF 性能最好，其次是基于高斯核函数的 SVM。多分类器系统中分类器的精度和分类器之间的差异性是构建多分类器系统的关键。基于以上分析，本文在选取基分类器算法时从精度和多样性考虑，选择了有监督学习算法中综合性能较好的 3 个分类算法：AdaBoost、SVM 和 RF。

针对不同分类器适用于不同分布特性的数据，本文提出了基于 k-means＋＋的多分类器选择算法：KMDCS (k-means＋＋ dynamic classifier selection)，该算法针对不同分类算法对具有不同区域特性数据的适应性不同，采用 k-means＋＋算法将数据集聚为 k 个簇，在每个簇上分别选取适应该簇数据的分类器作为该簇和与该簇类似数据的分类器。实验结果表明，这一算法大多数情况下能获得比单个分类算法更好的分类效果。

2 KMDCS 算法

若采用传统的多数投票法，多数的分类器由于适应性不好很有可能会错分类边界上的样本，从而导致集成起负作用，比较好的选择是用对局部区域数据分类精确度最高的分类器进行分类。判断哪个分类器是最佳的选择就显得尤为重要。KMDCS 算法基于这样的假设：如果分类器对与测试样本相似的数据具有很好的分类能力，那么这一分类器对测试样本也有很好的分类效果[8]。KMDCS 算法采用计算样本间的欧式距离来度量样本间的相似性，所以对数据进行预处理显得尤为重要，数据预处理阶段包括数据标准化和特征选择。

2.1 特征选择

在机器学习中，样本数据包含许多与预测目标不相关或者冗余的特征。不相关或冗余特征不但会影响模型的解释性和精确度，而且模型的训练阶段还会花费更多的时间，甚至会造成模型的过拟合导致模型泛化能力差。当某个数据集中特征数量比样本数量还要多时，学习算法将会更容易达到局部最优而停止。这个问题可以通过特征选择的方法来解决，选择最具有预测性的特征来简化高维数据[15]。

由于 KMDCS 算法需要用 k-means＋＋算法将原始数据集聚为不同局部区域的数据，通过计算样本之间的欧式距离来度量样本之间的相似性，无关联或冗余特征会影响数据聚类效果。特征选择结果的好坏会直接影响分类器的泛化能力[16]。特征选择方法很多，常见的有单变量选择、递归消除、稳定性选择和基于 L1(lasso 正则化)的选择。单变量选择选择出的特征子集可能会包含冗余特征，递归消除虽然很有效，但实际上它是一种贪婪算法，对某些选择有一定的偏好。稳定性选择默认使用 L1 正则化，通过二次采样不断改变特征选择结果，它通过随机选择训练集中的部分特征，多次进行正则化计算，最终排除那些特征影响系数经常为 0 的变量。实验发现采用稳定性选择与基于 L1 的选择比其他特征选择的效果更好。

2.2 数据标准化

如不对数据进行标准化处理，那么数量级大的特征变量将会对目标预测起着决定性作用，数量级小的特征

变量对分类的影响就微乎其微。数据标准化是将数据按比例缩放,使数据规范到一个小而特定的区间,去除数据单位,将有量纲的数字转化为无量纲的纯数值,以便不同量级或单位的指标能够比较和加权。数据标准化有很多方法,最具代表的是数据的归一化处理。文献[17-18]指出,样本数据归一化可以加速学习训练速度并提高分类精度。本文采用(1)式进行归一化计算,即将原始数据值转换到[0,1]区间。

$$\chi'=\frac{\chi-\chi_{\min}}{\chi_{\max}-\chi_{\min}}, \tag{1}$$

其中 $\chi_{\max}$ 为数据集中某一属性样本出现的最大值,$\chi_{\min}$ 为该属性的最小值。

2.3 用 k-means++对数据进行聚类

KMDCS算法是基于局部区域数据选择最好的分类器,首先需要将原始数据集聚成多个簇,同一簇内数据相似度高,不同簇的数据之间相似度低。k-means是最为常见的聚类算法,但它需要用户随机选择初始质心,初始质心选择的随机性会导致聚类的不稳定。k-means++算法[19]则可以克服这个问题,其主要思想是:初始聚类中心点之间的距离要尽可能的远,在选取第1个聚类中心点($n=1$)时随机选择,在第 n 个初始聚类中心($1<n<k$)的选择时,距离前 $n-1$ 个聚类中心越远的数据点会有更高的概率被选中。计算样本与聚类中心点距离为:

$$d(x)=\sqrt{\sum_{i=1}^{m}(x_i-c_i)^2}, \tag{2}$$

计算 x 点被选择为聚类中心的概率为:

$$P(x)=\frac{d(x)^2}{\sum_{x\in D}d(x)^2}。 \tag{3}$$

k-means++算法流程如下所示:

算法 1 输入:训练集 D,簇数量 k。

输出:k 个簇数据的集合。

过程:1.从 D 中随机选择一个点作为第一个聚类中心;

2.用(2)式计算 D 中每个点 x 与最近聚类中心的距离;

3.用(3)式计算样本点 x 作为下一个初始质心的概率;

4.重复2步和3步直到 k 达到指定值;

5.用 k 个初始聚类中心进行标准k-means聚类。

2.4 基于局部精度选择最合适的分类器

针对不同算法与不同分布特性的数据适应性问题,选出对局部数据分类准确率高的分类算法。KMDCS算法在候选分类器的训练阶段,将训练集进行特征选择和数据标准化后,用AdaBoost、SVM和RF 3种算法在预处理后的数据集上分别进行学习得到3个候选分类器,它的训练过程与传统训练过程相同。然后将训练集聚为 k 个簇,根据每个簇与3个候选分类器的适应性从中选择一个分类准确率最高的分类器作为簇数据的分类器。如图1所示。

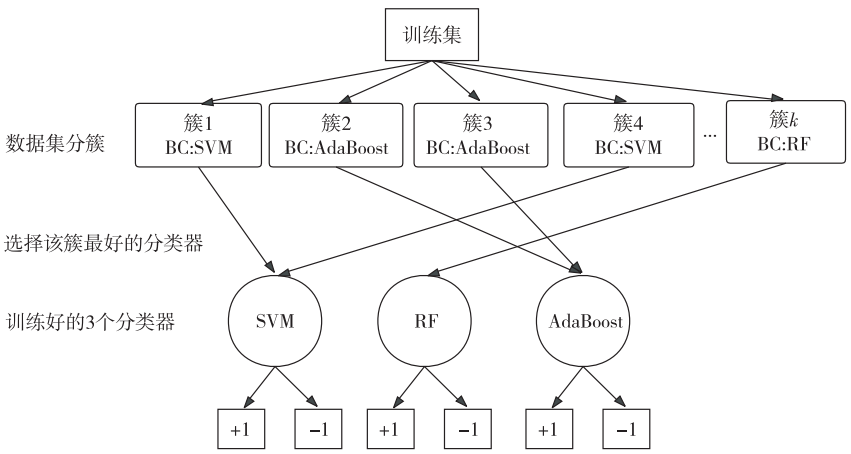

图1 基于局部区域选择最合适的分类器

Fig.1 Select the best classifier based on local region

2.5 KMDCS算法实现

KMDCS算法在分类器的训练阶段,首先对原始数据进行数据归一化和特征选择后,用AdaBoost、SVM和RF 3种算法分别在预处理后的数据集上进行训练得到3个候选分类器(AdaBoost分类器、SVM分类器、RF分类器)。然后用k-means++算法将训练集聚为 k 个簇,分别用3个候选分类器对每一个簇中的数据进行分类,对于某一个簇,3个候选分类器中哪一个分类器的分类精确度高,则选取最高的分类器作为簇数据或与簇数据相

似样本的分类器(有可能 3 个分类器分类精度一样，则随机选择一个)。在新样本的预测阶段，计算样本与 k 个簇质心的欧式距离，将样本判定为距离某簇的质心最近的那个簇，并用它对应的分类器进行类别判定。KMDCS 算法分类器选择过程伪代码描述如下。

算法 2　训练阶段：输入：训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中分簇的数量为 k。

输出：k 个簇与该簇对应的最佳分类器：$H=\{C_1:h_1,C_2:h_2,\cdots,C_k:h_k\}$。

过程：1. 对训练集 D 预处理(标准化、特征选择)；

2. 将处理后的训练数据集用 AdaBoost、SVM 和 RF 3 种算法分别训练得到 3 个候选分类器；

3. 用 k-means++将训练集聚为 k 个簇；

4. 对每个簇 j 从 1 到 k：用 3 个分类器分别测试 j 簇的分类误差，选择分类误差最小的分类器 h_j；

预测阶段：1. 对每个簇 j 从 1 到 k：计算新样本与簇 C_j 的质心的欧式距离，将该样本判定为距离某簇质心最近的簇 C_j；

2. 采用与该簇对应的分类器 h_j 进行类别预测。

3 实验结果与分析

本文从 UCI Machine Learning Repository 中选择了 9 个常用数据集进行仿真实验，实验在 HP EliteOne800 PC 上进行，代码用 Python 2.7 实现，运行于 windows 7 操作系统之上。仿真实验共有 2 部分，第 1 部分的仿真表明数据集通过特征选择后训练的模型的泛化能力更好；第 2 部分用原始数据集预处理过后的数据集进行仿真实验，来验证本文提出的 KMDCS 算法的有效性。

3.1 实验数据集

从 UCI 数据库中选择的 9 个常用的类别比较均衡的二分类数据集，分别是 liver-disorders，colic，statlog，pima，Credit，svmguide3，Dota2，phishing，magic04。它们包括了 7 个领域的真实数据，特征数均匀分布在 5～116 之间，样本数量分布在 145～19 020 之间。将数据集中每条数据类别标签统一成正类和负类(+1，−1)，由于 Dota2 数据集规模过于庞大，Dota2S 仅选择了 Dota2 数据集的前 7 760 条数据。实验数据集描述如表 1 所示，按照数据集样本数量递增排序。实验过程中，随机选择各个数据集中 2/3 的数据作为训练数据用于分类器的训练，余下的 1/3 数据作为测试数据，用于测试各分类器的泛化能力。

表 1　UCI 数据集信息描述

Tab. 1　Properties of UCI data sets

数据集	数据集样本数	特征数量	类别数
Liver-disorders	145	5	2
colic	366	21	2
statlog	690	14	2
pima	768	8	2
Credit	1 000	12	2
Svmguide3	1 243	23	2
Dota2S	7 760	116	2
phishing	11 055	30	2
Magic04	19 020	10	2

3.2 各分类算法主要参数设置

本文选择的 3 个基本分类算法分别是 AdaBoost 算法[21]、SVM 算法[22]和 RF[23]。

本文采用了基于“SAMME.R”实现的 AdaBoost 算法即 Real AdaBoost，不同于普通的“SAMME”弱学习器权重的度量方式，Real AdaBoost 中弱分类器输出不再是正类和负类，而是样本属于某类的概率即[−1，1]之间的实数，多数情况下这种方式比“SAMME”实现的更加精确[20]。实验过程中，分别测试了 AdaBoost 算法迭代次数为 1，5，10，50，100，200，500，1 000 时的分类情况，实验发现如果将 AdaBoost 算法的迭代次数设置过大(大于 50)，会造成很多簇内的数据 AdaBoost 分类器分类效果好于另外两个分类器，但是已经出现了过拟合问题。将迭代次数设置为 10 时，AdaBoost 算法的分类精度与 SVM 算法和 RF 算法分类精度大体相当，出现过拟合问题的概率非常小。选用机器学习库 Scikit-learn 所实现的 SVM 算法和 RF 算法作为基分类算法，在实验过程中参数都是 Scikit-learn 库所默认的参数。

3.3 特征选择前后训练模型泛化能力对比

特征选择作为 KMDCS 算法的前期工作，为了验证特征选择可以提高模型泛化精度这一理论，分别在 9 个

实验数据集上进行实验。将9个数据集数据归一化后采用稳定性特征选择进行特征提取,设置抽样次数为300(如果数据维数较大,相应的抽样次数需要适当的增加)。图2中灰色条表示数据没经过特征选择训练的AdaBoost分类器的分类精度,黑色条为数据经过稳定特征选择后训练的AdaBoost分类器的分类精度。同理,图3和图4分别显示9个数据集上特征选择前训练的SVM分类器、RF分类器分类精度与特征选择后分类器精度对比。

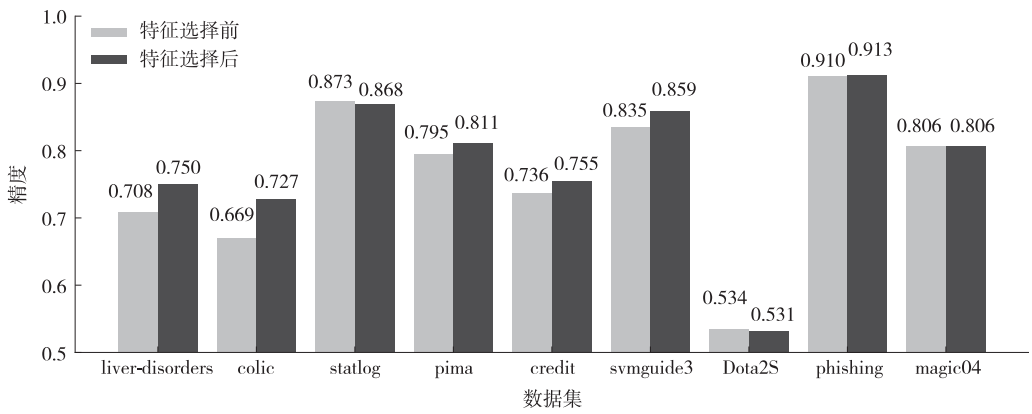

图2 特征选择前后数据集训练的AdaBoost分类器分类精度对比

Fig. 2 Accuracy comparison of the AdaBoost classifier trained before and after feature selection

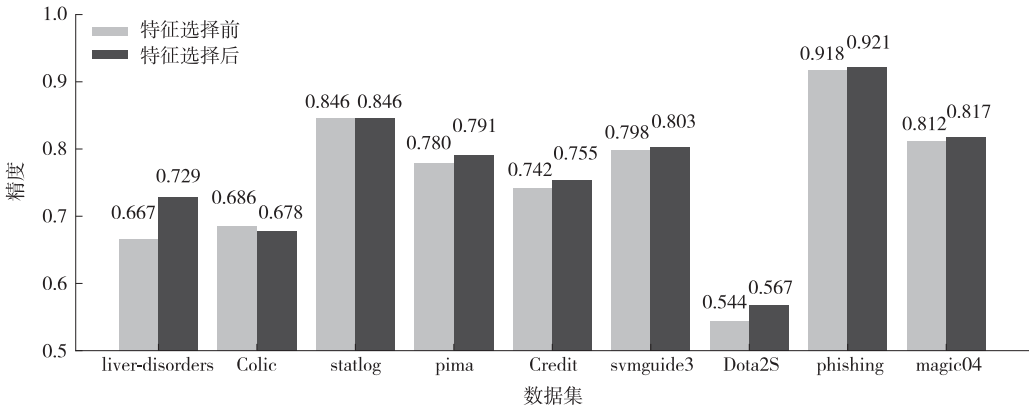

图3 特征选择前后数据集训练的SVM分类器分类精度对比

Fig. 3 Accuracy comparison of the SVM classifier trained before and after feature selection

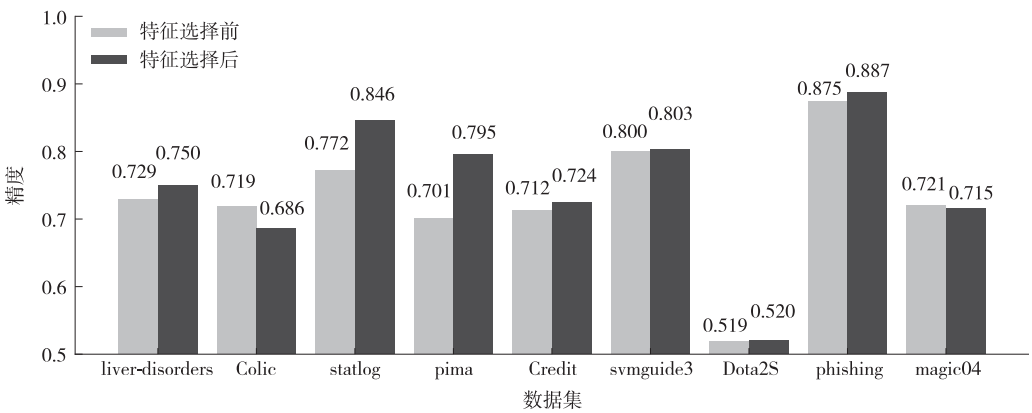

图4 特征选择前后数据集训练的随机森林分类器分类精度对比

Fig. 4 Accuracy comparison of the RF classifier trained before and after feature selection

由图2实验数据可以看出,在基于L1正则化的稳定性特征选择后,在statlog和Dota2S两个数据集上进行特征选择后训练的AdaBoost分类器较特征选择前训练的AdaBoost分类器泛化能力有少量的下降、在magic04上持平、在其余的6个数据集上特征选择后训练的模型的泛化能力有较大提升。由图3实验结果可知,只有在colic数据上经过特征选择后训练的SVM分类器模型泛化能力有所下降,在其余8个数据集上,经过特征选择后的数据训练的SVM分类器泛化能力更好。图4对比了数据经过特征选择前后所训练的RF分类器的泛化能力,发现在colic和magic04数据集上经特征选择后训练的RF分类器分类精度有所降低,其原因可能是RF在构建基本决策树分类器时随机选择某些属性,而这些属性在特征选择时被排除,从而影响了基分类器的多样性而导致的。在其余数据集上经过特征选择后的数据训练的模型泛化能力更好。总的来说,有很少的数据集通过特征选择过后,训练的模型泛化误差会上升,但是上升并不明显。而在多数数据集上经过特征选择后的数据训练的模型泛化能力更好。

3.4 KMDCS 算法实验结果与分析

这里对比了 3 个基本算法和本文中提出的动态选择分类器算法（KMDCS）在 9 个数据集上的分类精度。将经过预处理（数据归一化、特征选择）后的数据按训练集与测试集 2∶1 的比例进行分配，用 2/3 的数据分别训练 AdaBoost、SVM、RF 等算法得到 3 个候选分类器，KMDCS 分类器由 2.5 描述的算法选择得到，然后用后余下 1/3 数据测试 4 个分类器的分类误差。

本文提出的 KMDCS 算法中 k 值的不同，KMDCS 分类精度也不同，这里选用一个中等规模的数据集 Credit 进行分析，其中用于训练的样本有 670 个，实验发现有以下规律，如图 5 所示。

从图 5 实验数据可得：随着 k 值的增加，KMDCS 算法在数据集上的分类精确率先上升后下降，这是因为随着 k 值增加，训练集数据被划分越来越细，选出的局部最优分类器更有针对性，算法分类精度上升。由于真实的数据包含噪声，过度分簇使得分到一个簇的数据越来越少，导致选择的分类器不稳定，算法分类精度下降。在 Credit 数据集上簇数达到 400 左右时分类精度相对较好，对于其他数据集，因为数据分布特征与规模的不同，达到最好分类效果的 k 值也不相同。

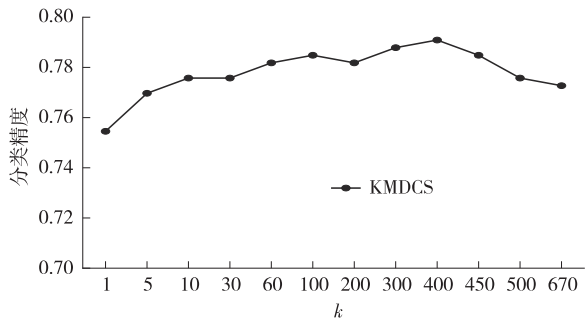

图 5　k 取不同值时 KMDCS 算法在 Credit 上的分类精度

Fig. 5　The classification accuracy of KMDCS algorithm on Credit when k takes different values

下面的实验测试不同 k 值的 KMDCS 算法与 AdaBoost 算法、SVM 算法和 RF 算法在各个数据集上训练得到的分类器的分类精确度。将原数据集分成 3 个簇（$k=3$）时，3 个基本分类器和 KMDCS 分类算法的分类准确率如表 2 所示。

观察表 2 中实验数据可以发现，$k=3$ 时，在数据集 Liver-disorders、colic、statlog、magic04 上，KMDCS 算法得到的分类效果比单个分类算法都好。在 pima、Credit、svmguide3、Dota2 数据集上，KMDCS 算法和 3 个分类算法（AdaBoost、SVM、RF）中分类精度最高的算法的分类精度相同。KMDCS 算法在 phishing 数据集上获得的分类精确度没有 SVM 算法分类精确度高，但要好于 RF 和 AdaBoost 算法。

表 2　KMDCS 算法 $k=3$ 时与 3 个基本分类算法在实验数据集上分类精度对比

Tab. 2　Comparison of accuracy between KMDCS algorithm $k=3$ and 3 basic classification algorithms on experimental data sets

数据集	精确度/%			
	AdaBoost	SVM	RF	KMDCS
Liver-disorders	75.000	72.917	75.000	77.083
colic	72.727	67.768	68.595	75.206
statlog	86.842	84.649	84.649	87.281
pima	81.102	79.133	79.527	81.102
Credit	75.454	75.454	72.424	75.454
Svmguide3	85.888	80.291	80.291	85.888
Dota2S	53.143	56.735	51.971	56.735
phishing	91.257	92.134	88.681	91.888
Magic04	80.611	81.519	72.295	81.742

增大 KMDCS 算法中 k 的取值，如果将各训练集数据分成 30 个簇（$k=30$）时，各种算法在 9 个实验数据集上分类精度如表 3 所示。

由表 3 的仿真结果可知，KMDCS 算法在 5 个数据集上比单个分类算法的分类效果更好，在 colic，Svmguide3 数据集上的分类效果次于 AdaBoost 算法，但是要比 SVM 和随机森林算法好。在 Dota2S 数据集上仅次于 SVM 算法。

增加 KMDCS 算法中的 k 值，当 $k=120$ 时，因为 Liver-disorders 的训练集只有 97 个样本，簇数量大于训练样本数，会产生空簇，所以表 4 中没有 Liver-disorders 数据集的测试数据。

从表 4 可以看出，KMDCS 算法在多数数据集上分类精确率比单个分类算法好，可以获得更加稳定可靠的分类效果。综合观察表 2、表 3、表 4 的仿真结果发现总体呈现的一个趋势是：随着 k 值增大，KMDCS 算法在样本数量较少的数据集上精度有所降低而在样本数量多的数据集上算法精度有所上升。KMDCS 算法分类精度的提升是因为将训练集中的数据从开始的一个整体聚成不同区域的数据之后，基于局部精度选择出最优分类器可以使得不同区域的数据采用更加适合于该分布特性的分类器进行分类，从而达到更加精确的分类效果。随着 k 值

的继续增大,算法在样本数量较少的数据集上精度下降的原因可能是训练集中数据少,导致数据分裂太细,以至于某些簇中只有一个样本,在选择最优分类器的时候出现了过拟合现象。

继续增加 k 值,$k=900$ 时,这里舍去了训练样本数量不足900的5个数据集:liver-Liver-disorders,colic,statlog,Credit,svmguide3。4种分类算法在3个数据集上的分类精度如表5所示。

综合观察表2、表3、表4、表5的实验结果可以发现,KMDCS算法对于样本数量小的数据集减少分簇数量、数据样本数量大时增大分簇数量可以得到很好的分类精确度。不同数据集由于数据量和数据分布的不同,不同的 k 值可以使得KMDCS算法的分类效果不同。从数据集角度上看,KMDCS算法在Dota2S上在 $k=3$ 时分类精度可以和SVM分类精度相当,随着 k 值的增大,KMDCS算法分类精度较SVM分类精度有所不足,这可能是因为该数据集数据本身是集中分布的,不太适合用KMDCS算法进行分类,但是该分类方法仍然好于AdaBoost算法和随机森林算法。在大多数数据集上KMDCS算法比单个分类算法有更好的分类效果。比如数据集Magic04,可以发现在该数据集上KMDCS算法比3个单个分类算法更好,这是因为数据集数据本身呈离散分布,本文提出的KMDCS算法在这类数据集上取得比原单个分类算法更好的分类准确性。对于数据呈现集中分布的数据集,本文提出的算法与3个分类算法最好的那个算法精确率相当。最差的情况是KMDCS算法可能会比3个基本分类算法中最好的那个分类算法略差,但是好于另外两种分类算法。由于训练阶段会针对3种不同算法训练分类器,预测阶段增加了动态选择的过程,相对于单个分类算法会花费约3倍的时间,但在可接受的范围之内。以上的仿真结果及分析表明了本文提出的KMDCS算法的有效性。

表3 KMDCS算法 $k=30$ 时与3个基本分类算法在实验数据集上分类精度对比

Tab.3 Comparison of accuracy between KMDCS algorithm $k=30$ and 3 basic classification algorithms on experimental data sets

数据集	精确度/%			
	AdaBoost	SVM	RF	KMDCS
Liver-disorders	75.000	72.916	75.000	79.166
colic	72.727	67.768	68.595	71.901
statlog	86.842	84.649	84.649	86.842
pima	81.102	79.133	79.527	82.677
Credit	75.454	75.454	72.424	77.576
Svmguide3	85.888	80.291	80.291	85.645
Dota2S	53.143	56.735	51.971	56.462
phishing	91.257	92.134	88.681	92.217
Magic04	80.611	81.519	72.295	82.173

表4 KMDCS算法 $k=120$ 时与3个基本分类算法在实验数据集上分类精度对比

Tab.4 Comparison of accuracy between KMDCS algorithm $k=120$ and 3 basic classification algorithms on experimental data sets

数据集	精确度/%			
	AdaBoost	SVM	RF	KMDCS
colic	72.727	67.768	68.595	74.380
statlog	86.842	84.649	84.649	86.842
pima	81.102	79.133	79.527	81.496
Credit	75.454	75.454	72.424	78.182
Svmguide3	85.888	80.291	80.291	85.645
Dota2S	53.143	56.735	51.971	55.369
phishing	91.257	92.134	88.681	92.134
Magic04	80.611	81.519	72.295	82.444

表5 KMDCS算法 $k=900$ 时与3个基本分类算法在实验数据集上分类精度对比

Tab.5 Comparison of accuracy between KMDCS algorithm $k=900$ and 3 basic classification algorithms on experimental data sets

数据集	精确度/%			
	AdaBoost	SVM	RF	KMDCS
Dota2S	53.143	56.735	51.971	54.002
phishing	91.257	92.134	88.681	92.244
Magic04	80.611	81.519	72.295	82.332

4 结语

本文提出了基于k-means++的动态选择分类器的KMDCS算法,通过选择对局部适应性好的分类器来预

测与该区域相似数据的类别，解决了单个分类算法缺少对不同局部区域样本的针对性问题。通过比较 3 个基本分类算法（AdaBoost，SVM，RF）和 KMDCS 算法在 9 个数据集上的泛化精度，实验发现 KMDCS 算法多数情况下能提升分类准确率。在实验中 KMDCS 算法还有许多参数没有调整到最优的组合，但可以看到分类精度比单个分类算法要好。在今后的研究中可以通过调整参数组合，使模型达到更好的分类效果。本文尝试了不同 k 值的情况，从将数据全部归为一个簇到每个样本单独作为一个簇，绝大多数情况下本文提出的算法都比单个分类算法有更高的分类准确性。对于不同的数据集，没有找到达到最佳效果的统一的 k 值，这也是接下来的工作需要研究的。本文在多分类器训练时仅选择了 3 个分类算法，在今后的研究中也可以加入更多的分类算法，进行更具有针对性的基于局部精度的多分类器动态选择的研究。

参考文献：

[1] CARUANA R，NICULESU M A. An empirical comparison of supervised learning algorithms using different performance metrics[J]. International Journal of Intelligent Information Processing，2005，1(4)：161-168.

[2] DIDACI L，FUMERA G，ROLI F. Diversity in classifier ensembles：fertile concept or dead end? [J]. Lecture Notes in Computer Science，2013，7872：37-48.

[3] BAR A，ROKACH L，SHANI G，et al. Improving simple collaborative filtering models using ensemble methods [M]//Multiple Classifier Systems. Berlin，Heidelberg：Springer，2013：1-12.

[4] ZHOU Z H，WU J X，TANG W. Ensembling neural networks：many could be better than all[J]. Artificial Intelligence，2002，137(1)：239-263.

[5] 窦鹏. 基于投票法的多分类器集成遥感影像分类技术[D]. 兰州：兰州交通大学，2014.
DOU P. Multi classifier integrated remote sensing image classification technology based on voting method [D]. Lanzhou：Lanzhou Jiaotong University，2014.

[6] 饶川，苟先太，金炜东. 基于选择性集成学习的高速列车故障识别研究[J]. 计算机应用研究，2018，35(5)：1-2.
RAO C，GOU X T，JIN W D. Study on recognition of high speed rail malfunction based on selective ensemble learning [J]. Application Research of Computers，2018，35(5)：1-2.

[7] 王文哲，吴华. 粗糙 k-means 和 AdaBoost 结合的雷达辐射源快速识别算法[J]. 空军工程大学学报（自然科学版），2016，17(1)：51-55.
WANG W Z，WU H. A fast radar emitter recognition algorithm based on rough k-means combined with AdaBoost [J]. Journal of Air Force Engineering University (Natural Science Edition)，2016，17(1)：51-55.

[8] 张亮，李智星，王进. 基于动态权重的 AdaBoost 算法研究[J]. 计算机应用研究，2017，34(11)：3233-3236.
ZHANG L，LI Z X，WANG J. Research on dynamic weights based on AdaBoost[J]. Application Reacher of Computers，2017，34(11)：3233-3236.

[9] 李凯，常圣领. 基于 k-means 聚类的神经网络分类器集成方法研究[J]. 计算机工程与应用，2009，45(22)：120-122.
LI K，CHANG S L. Study of ensemble method of classifiers for neural networks based on k-means clustering[J]. Computer Engineering and Applications，2009，45(22)：120-122.

[10] GIACINTO G，ROLI F. Adaptive selection of image classifiers[C]//International Conference on Image Analysis and Processing. Berlin：Springer-Verlag，1997：38-45.

[11] DIDACI L，GIACINTO G. Dynamic classifier selection by adaptive k-nearest-neighbourhood rule[J]. Lecture Notes in Computer Science，2004，3077：174-183.

[12] DIDACI L，GIACINTO G，ROLI F，et al. A study on the performances of dynamic classifier selection based on local accuracy estimation [J]. Pattern Recognition，2005，38(11)：2188-2191.

[13] CARUAUA R，KARAMPTIZIKIS N，YESENALINA A. An empirical evaluation of supervised learning in high dimensions [C]//International Conference on Machine Learning. [S. l.]：ACM，2008：96-103.

[14] 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.

[15] KOLLER D，SAHAMI M. Toward optimal feature selection[C]//ACM. Proc of Int Conf on Machine Learning. Bari Itali：ACM，1996：284-292.

[16] 姚旭，王晓丹，张玉玺. 特征选择方法综述[J]. 控制与决策，2012，27(2)：161-166.
YAO X，WANG X D，ZHANG Y X. Summary of feature selection algorithms[J]. Control and Decision，2012，27(2)：161-166.

[17] 柳小桐. BP 神经网络输入层数据归一化研究[J]. 机械工程与自动化，2010，3：122-123.
LIU X T. BP neural network input layer data normalization research [J]. Mechanical Engineering and Automation，2010，3：122-123.

[18] 王新志，陈伟，祝明坤. 样本数据归一化方式对 GPS 高程转换的影响[J]. 测绘科学，2013，38(6)：162-165.
WANG X Z，CHEN W，ZHU M K. Influence of sample data normalization ways on GPS elevation transformation

[J]. Science of Surveying and Mapping, 2013, 38(6): 162-165.

[19] ARTHUR D, VASSILVITSKII S. K-means++: the advantages of careful seeding[C]//ACM. Eighteenth Acm-Siam Symposium on Discrete Algorithms. [S. l.]: Society for Industrial and Applied Mathematics, 2007, 11(6): 1027-1035.

[20] 付忠良. 关于 Real AdaBoost 算法的分析与改进[J]. 电子科技大学学报, 2012, 41(4): 545-551.
FU Z L. Analysis and Improvement on Real AdaBoost Algorithm [J]. Journal of University of Electronic Science and Technology of China, 2012, 41(4): 545-551.

[21] HASTIE T, ROSSET S. Multi-class AdaBoost[J]. Statistics & Its Interface, 2009, 2(3): 349-360.

[22] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. Journal of ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[23] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

Multiple Classifiers Selection Classification Based on k-means++

XIONG Lin, TANG Wanmei

(College of Computer Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: [**Purposes**] Different algorithms in machine learning are suitable for data sets with different distribution characteristics. One algorithm may be better than other algorithms on data sets with some distribution characteristics. Classifier trained on the whole training set is used to predict the new sample class, because lack of pertinence to local region samples, it may lead to a wrong classification of a classifier with poor prediction ability in local regions. To solve this problem, a multi-classifier selection algorithm based on k-means++ is proposed. [**Methods**] The process of this algorithm is first used 3 kinds of classification algorithms: AdaBoost, SVM, Random Forests on the training set were respectively trained one classifiers as candidate classifier, and then use k-means++ algorithm divides the training set into k clusters, with 3 classifiers separately for each cluster classification, selection of the highest classification accuracy of cluster classifier as best classifier. When classifying the new sample, first determines which cluster is the sample belongs to, and then uses the best classifier for classification prediction. [**Findings**] The experimental results show that the algorithm improves accuracy rate of classification and recognition on 9 UCI data sets compare with a single algorithm. [**Conclusions**] The accuracy of model classification can be improved by selecting the best classifier dynamically based on local regions.

Keywords: local region; AdaBoost; k-means++; random forest; SVM

(责任编辑 许 甲)