融合SMOTE与Filter-Wrapper的朴素贝叶斯决策树算法及其应用

许召召[1] 李京华[1] 陈同林[1] 李昕洁[1,2]

(云南大学软件学院 昆明 650091)[1] (云南省软件工程重点实验室 昆明 650091)[2]

摘 要 如何对以“工业4.0”为背景的物联网智慧医疗系统所产生的医疗数据进行高效且准确的挖掘仍然是一个十分严峻的问题。而医疗数据往往是高维的、不平衡的和有噪声的,因此提出一种新的数据处理方法——将SMOTE方法与Filter-Wrapper特征选择算法融合,并将其应用于支持临床医疗决策。特别地,所提方法不仅克服了朴素贝叶斯在属性实际应用中因属性独立假设而造成的预测不佳的情况,而且避免了C4.5决策树在构建模型时的过拟合问题。将所提算法应用于ECG临床医疗决策中,取得了很好的效果。

关键词 数据平衡,Wrapper特征选择,朴素贝叶斯,决策树

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.09.009

Naive Bayesian Decision Tree Algorithm Combining SMOTE and Filter-Wrapper and It's Application

XU Zhao-zhao[1] LI Ching-hwa[1] CHEN Tong-lin[1] LEE Shin-jye[1,2]

(School of Software, Yunnan University, Kunming 650091, China)[1]

(Key Laboratory in Software Engineering of Yunan Province, Kunming 650091, China)[2]

Abstract How to efficiently and accurately dig out the medical data generated by the Internet-based wisdom medical system with “Industrial 4.0” is still a very serious problem. However, the medical data is often high-dimensional, unbalanced and noisy, so this paper proposed a new data processing method combining SMOTE method with Filter-Wrapper feature selection algorithm to support clinical decision-making. In particular, the proposed method not only overcomes the situation of bad prediction result of the independent assumptions in the practical attribute application of Naive Bayesian, but also avoids over-fitting problem caused by constructing the model of C4.5 decision tree. What's more, when the proposed algorithm is applied to ECG clinical decision-making, good results can be obtained.

Keywords Data balance, Wrapper feature selection, Naive Bayesian, Decision tree

1 引言

2013年,德国政府率先提出“工业4.0”,其将信息技术与制造业融合并用于人工智能、智慧医疗、智慧交通等领域[1-2]。随着医疗信息化建设的发展,医疗信息系统中积累了大量的医疗数据,这些数据蕴含着十分重要的信息,如何提取出这些信息成为了现阶段医疗数据挖掘的热点。然而医疗数据往往是不完整、不平衡且含有噪声的[3-4],因此要进行准确的医疗决策,不仅需要选择合适的分类算法,而且必须对所获取的数据进行相关处理。基于此,本文提出一种新的数据处理方法——SMOTE-Filter-Wrapper。首先,通过SMOTE方法来降低原始医疗数据的不平衡性;然后,将Filter特征选择算法和Wrapper特征选择算法相融合,解决了Wrapper特征选择时效率低的问题,而且弥补了Filter特征选择算法选择的特征与后续算法偏差较大的不足。与此同时,选择合适的分类算法也是本文的研究方向之一。通过数据挖掘算法提取医疗数据中潜在的信息,主要通过数据挖掘算法训练数据集,给出高效且准确的临床诊断。

分类算法是数据挖掘领域的一个研究热点,常用的分类算法主要有神经网络、决策树、朴素贝叶斯等[5-7]。朴素贝叶斯是一种高效且容易理解的分类算法[8],但其属性之间同等重要性的假设在实际应用中并不适用,从而影响了分类器的性能。同样地,在构造决策树算法时,由于数据中往往含有噪声,从而造成过度拟合,导致分类精度下降。因此,本文将朴素贝叶斯与决策树融合,使用朴素贝叶斯的概率优化方法去除数据中的噪声实例,不仅削弱了属性之间的独立性,而且避免了决策树的过度拟合。虽然将朴素贝叶斯与决策树进行融合可以避免各自算法的缺点,但是该算法的模型构建仍然在内存中,因此能否高效地构建分类模型是本文研究的主要方向之一。

到稿日期:2017-10-09 返修日期:2018-01-14 本文受国家自然科学基金:云计算环境下双模型驱动面向软件动态演化的建模与分析(61379032)资助。

许召召(1991—),男,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:243642549@qq.com;**李京华**(1994—),女,硕士生,主要研究方向为模糊逻辑系统,E-mail:chinghwali@hotmail.com;**陈同林**(1992—),男,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:tonglinchen@hotmail.com;**李昕洁**(1974—),男,博士,教授,主要研究方向为机器学习、计算智慧与决策支援系统,E-mail:camhero@hotmail.com(通信作者)。

经实验得出,将本文提出的方法应用于朴素贝叶斯决策树分类器中进行模型训练与预测,可以取得很好的效果;将其应用于心电图数据集中,可以得到高效且准确的病人病症分类结果,在心电图治疗中起到了辅助作用。

2 相关工作

2.1 SMOTE 算法

非平衡数据分类是指分类中的数据集类别分布不均匀,某类样本数在整个数据集中占主要优势。这一情况多存在于医疗诊断、故障检测、信用卡诈骗等领域中。本文基于心电图医疗数据诊断的背景,在预测医疗数据时,先对原始数据集进行平衡处理,以确保医疗诊断的准确性。

基于数据抽样的方法主要分为欠采样、过采样、欠采样和过采样相结合的方法。过采样中,最常用的是 Chawla 等[9]提出的 SMOTE 算法,该算法能够合成少数类样本,得到了广泛的应用;类似的还有 Borderline-SMOTE 算法[10],它是一种改进的 SMOTE 算法,只对数据的少数类的边界样本进行抽样处理,使得增加的样本具有价值。欠采样中,Yen 等[11]提出了一种基于聚类的欠抽样方法,该方法通过聚类后的每个簇内的多数类和少数类样本数目的比例确定抽样比例参数,使被选择的多数类样本更具有代表性。也有学者提出将两种方法融合,如 SMOTE+ENN 和 SMOTE+Tomeklink 算法。

SMOTE 算法[12]是为克服随机过抽样不足而提出的再抽样算法。其基本原理是在近邻少数类样本之间插入新值,合成新的少数类样本。具体做法是:假设过采样倍数为 N,首先从每个少数类样本的 K 个同类最近邻中随机选择 N 个样本;然后按照式(1)将新合成的样本加入数据中。

$$Pnew=x+rand(0,1)\times(y_i-x) \tag{1}$$

其中,$i=1,2,\cdots,N$,$rand(0,1)$表示 0 到 1 之间的一个随机数。

2.2 Wrapper 型特征选择

特征选择是指从数据集的所有特征中挑选出最优的特征子集的过程[13]。根据其是否独立于后续学习算法,可被分为过滤式(Filter) 和封装式(Wrapper)[14]两种。其中,过滤式(Filter)方法一般依据评价准则来增强特征与类的相关性,弱化特征之间的相关性。目前,使用得最多的是概率距离和相关测量法[15]、类之间的距离测量法、信息熵法[16]等。而封装式(Wrapper)评价策略的特征选择算法的原理是使用后续学习算法的分类性能来评价特征子集的优势。目前,此类方法是该领域的研究热点。Hsu 等[17]提出了使用决策树来进行特征选择的 Wrapper 方法,通过遗传算法来搜索使得决策树分类错误率最小的特征子集。类似的方法还有将遗传算法与人工神经网络相结合来选择特征子集[18],以及将启发式搜索策略(SBS,SFS,FSFS)和分类器性能评价准则相结合来评价特征子集[19]。

虽然这两种方法均有各自的优点,但是 Filter 特征选择算法选择的特征子集与后续算法无关,偏差较大;而 Wrapper 效率低,并不适合大数据集。因此,如何构造一个准确率高且代价低的特征选择方法是本文研究的重点之一。

2.3 贝叶斯网络

贝叶斯分类算法源于贝叶斯定理,使用概率的形式表示数据样本的不确定性,通过改变事件的先验概率和后验概率,假设各个属性相互独立来预测分类结果[20]。朴素贝叶斯算法是一种简单的概率分类器,其将贝叶斯定理应用于数据的属性独立假设。虽然朴素贝叶斯简单,但是其往往比一些复杂的分类器拥有更高的效率[21],因此本文选择朴素贝叶斯算法与决策树算法相融合的算法。

朴素贝叶斯算法的主要思想如下:

设有 n 维特征向量,其代表 n 个属性的值,即 $A=\{a_1,a_2,\cdots,a_n\}$,给定一个未知数据集,得其目标值为:

$$V_{map}=\arg\max P(V_j|a_1,a_2,\cdots,a_n) \tag{2}$$

其中,$V_j\in V$。

假设有 m 个类别,分别用 $V_1,V_2,\cdots,V_m$ 表示。给定一个未知数据集 X(没有类别号),由贝叶斯定理可得:

$$P(V_I|X)=\frac{P(X|V_i)P(V_i)}{P(X)} \tag{3}$$

由于对于所有类,$P(X)$均为常数,因此最大后验概率 $P(V_i|X)$可以转化为最大先验概率 $P(X|V_i)P(V_i)$。

朴素贝叶斯假设各个属性之间相互独立,即:

$$P(X|V_i)=\prod_{k=1}^{n}p(x_k|V_i) \tag{4}$$

其中,先验概率 $p(x_k|V_i)$,$k\in1,2,\cdots,n$,可以从训练样本中求得。

综上所述,对于一个未知数据集 X,计算其属于类别 V_i 的概率 $P(X|V_i)P(V_i)$,选择其中概率值最大的类别作为分类类别。

2.4 决策树

决策树是通过一系列规则对未知数据分类的过程,因规则简易、计算量小且具有较高的分类准确率等优点而得到广泛应用。决策树的构造过程与人的决策行为的原理相似。给定一个数据集,通过统计方法选择数据集中的属性 A 作为根节点,计算属性 A 的信息增益等方法,将数据集 S 分为多个子集。在这些子集的基础上重复上述过程,直到满足一个特定的停止准则。

Quinlan 于 1986 年提出了 ID3 算法[22],通过信息熵[23]来选择属性。使用贪心算法,自顶向下进行搜索,对单个属性进行多叉划分,为属性的所有取值都建立一个分支,并采用信息增益作为评价标准。信息增益的缺点是倾向于选择取值较多的属性,这时就会面临多值偏向性的问题。为了克服信息增益选择属性时偏向选择值较多的属性的不足,Quinlan 提出了 C4.5 算法[24],根据信息增益率(Gain Ratio)的值来选择属性,克服了 ID3 算法的缺点。CART 算法[25]使用 GINI 不纯度作为度量准则,采用二分递归分割的方法,重复地将样本集分为两个子集,使得每个非叶子节点都有两个分支,最后产生一棵二叉决策树。与 ID3 算法相比,二叉划分的方法更为适用,可以很好地处理数值型的属性,但是使用这种方法划分离散值属性时会增加决策树深度,且划分数值型属性时也需要大量的排序和计算。

3 算法设计与分析

本文提出的方法是基于 NB-C4.5 算法的,在整个结构中,首先通过 SMOTE 方法来有效地降低数据集的不平衡性,然后使用 Filter-Wrapper 特征选择算法剔除负作用特征,最后将上述方法处理后得到的数据集应用于 NB-C4.5 算法中进行模型训练,最终获得分类预测结果。

3.1 基于 SMOTE 采样方法的 Filter-Wrapper 特征选择算法

Wrapper 方法使用后续算法的分类准确性作为评价指标，虽然 Wrapper 方法筛选的特征子集提高了后续分类器的预测能力，但泛化能力较差，效率较低。为了改善 Wrapper 特征选择方法处理大数据时效率较低的问题，本文提出将 Filter 与 Wrapper 算法相融合的方法。首先，使用 Filter 特征选择算法来降低高维数据的维度，并通过设置不同的阈值来确定特征筛选的幅度；然后，使用 Wrapper 方法筛选出更加精确的特征子集。SMOTE-Filter-Wrapper 算法的步骤如算法 1 所示。

算法 1　SMOTE-Filter-Wrapper 算法

输入：数据集 D，评价学习器 NB-C4.5

输出：新数据集 D′

Step 1　对于数据集 D，若抽样率为 m，对于每个负类样本点 x_i，找出其 k 个负类近邻点。从中任选 m 个近邻点 y_{ij}（j=1，2，…，m），根据式(1)合成 N 个新的负类样本。

Step 2　对于 Step 1 获得的数据集，首先使用 Filter 特征选择算法降低数据的维度，并通过设置不同的阈值（根据属性的信息增益来设置）将筛选后的数据集应用于 Wrapper 方法中，其中分类算法选择 NB-C4.5。根据分类器的预测性能进行更精确的筛选。

Step 3　根据分类器的预测性能选择分类效果最好的特征子集，并输出新的数据集 D′。

3.2 朴素贝叶斯决策树算法的构造

根据上文所述朴素贝叶斯算法的思想，提出一种朴素贝叶斯和 C4.5 决策树相融合的 NB-C4.5 分类算法，其设计思路如下。

1）如果通过 C4.5 决策树的信息增益率的值可以选择某个属性分支，则 BN 的值为 0。其中 C4.5 决策树的信息增益率的计算方法如下。

假设训练样本为 S，样本中有 n 个类，则 S 的熵（信息增益）可以表示为：

$$I(S)=-\sum_{i=1}^{m}p_i\log_2(p_i) \tag{5}$$

其中，p_i 表示属于类 i 的训练样本概率。设属性 A 是具有 v 个不同子集的数据对象属性$\{S_1,S_2,\cdots,S_V\}$，其中 S_j 由 S 的属性 A 中值为 a_j 的样本组成。假设 s_{ij} 是子集中的类 C_i 的样本数。

根据属性 A 划分的信息增益可以描述为：

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{nj}}{s}I(s_{1j},\cdots,s_{nj}) \tag{6}$$

其中，$\frac{s_{1j}+\cdots+s_{nj}}{s}$ 是第 j 个子集的权重。对于给定的子集，有：

$$I(s_{1j},\cdots,s_{nj})=-\sum_{i=1}^{n}p_{ij}\log_2 p_{ij} \tag{7}$$

上式表示属于类的样本的概率。此时，属性 A 的信息增益可以描述为：

$$(Gain(A)):Gain(A)=I(S)-E(A) \tag{8}$$

该过程的目的是选择具有最大信息增益的属性作为分支节点。为了避免通过信息增益选择时面临多值偏向性问题，使用信息增益率来选择分支属性。

$$Ratio(S,A)=Gain(S,A)/Split(S,A) \tag{9}$$

其中：

$$Split(S,A)=-\sum_{i=1}^{v}\frac{S_j}{S}\log_2\frac{S_i}{S} \tag{10}$$

C4.5 决策树分类器是根据上述属性选择方法自上而下形成的。其内部节点表示分支属性，叶节点表示类。在形成决策树分类器之后，以从根到叶节点的合并范式提取的方式形成分类规则。

2）当数据元素的类别无法确定时，BN 的取值为 f。其中，f 是朴素贝叶斯的概率公式，即根据实验结果得出，该数据对象先验概率值由 f 可以得出，将训练样本分到某一个类中，然后再计算出后验概率，最后选择后验概率最大的类作为该训练样本的类别[26]。

3.3 算法分析

本文所使用的算法模型与决策树相同，具有规则简易、分类性能较高等优点。

实验主要分为 3 步：1）使用 SMOTE 方法平衡数据集；2）对步骤 1）中处理过的数据，应用 Filter-Wrapper 特征选择算法筛选出最优的特征子集；3）使用 NB-C4.5 算法对数据集进行建模预测。

本文方法具有以下优点：

1）更好的预测能力。传统的 C4.5 决策树使用信息增益率来选择属性，假设在决策树构建过程中出现了属性不同但是信息增益率相同的情况，则会造成属性二义性，这对数据集的分类预测产生了不良影响。而 NB-C4.5 算法通过朴素贝叶斯的先验概率来处理数据二义性，能提高模型的预测能力。

2）更强的分类鲁棒性。数据挖掘一般是从不完整、含有噪声的数据集中提取出隐藏在数据中的信息，而朴素贝叶斯可以剔除数据集中的噪声，能提高 NB-C4.5 分类器的鲁棒性。

3）更高的算法性能。医疗数据往往是不平衡的，通过使用本文的方法不仅能降低医疗数据的不平衡性，还能提高特征选择的效率和整体算法的性能。

4 实验结果与分析

4.1 数据集和性能指标

本节主要通过实验来研究基于 Wrapper 特征选择的 NB-C4.5 算法的性能，为此，从 UCI 标准及其学习库[27]中下载基准数据集进行算法的训练与测试 。另外，本文选取的 ECG 数据是由玉溪人民医院提供的真实数据，该医院的医生对数据进行了初步处理。在此基础上，对部分数据结果贴上标签，将全部心电图结果划分为正常与不正常两类，因此该问题变成一个二分类问题。数据共有 222594 条实例，13 个特征属性和 1 个类别标签属性。所有实验数据的特征描述如表 1 所列。

表 1　UCI 标准数据集和 ECG 数据集的特征描述

Table 1　Feature description of UCI standard dataset and ECG dataset

数据集	样本数	属性数	类别数
Sick	3772	29	2
Splice	3190	61	3
Segment	2310	19	7
Anneal	898	38	5
Vehicle	846	18	4
Soybean	683	35	19
Vote	435	16	2
Sonar	208	60	2
ECG	222594	13	2

为了验证本文算法的性能，选取了一些常用的评价指标：

如准确率(Precision)和 MCC。

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

其中,TP 为正确分类的实例,FP 为错误分类的实例。MCC 是另一个有效的不均衡数据分类性能的评价指标,对于一个给定的两分类问题,MCC 的计算方法为:

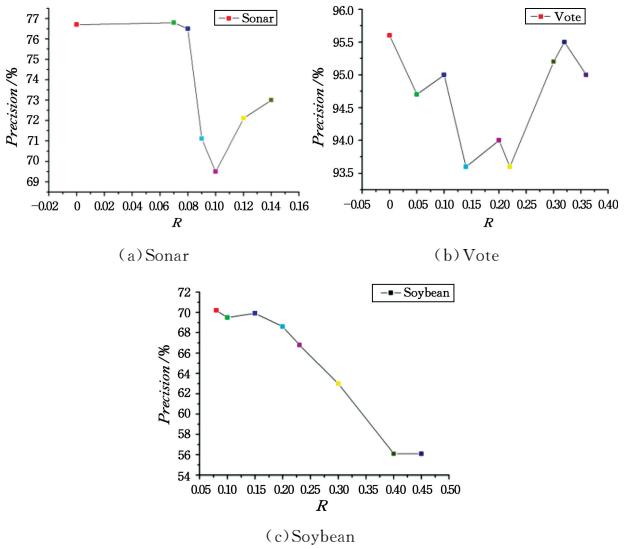
$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{12}$$

4.2 实验设计

本文采用 10 层交叉(10-fold Cross Validation)方法在数据集上进行实验测试,分别记录其分类准确率和 MCC,最后求得平均值,即算法的分类准确率。首先,对 SMOTE-Filter-Wrapper 算法进行验证,然后将传统特征选择算法与本文算法进行比较,最后将本文算法应用于临床医疗决策中。

4.3 实验结果

对本文方法进行验证。首先,使用 SMOTE 方法平衡数据集,然后使用 Filter-Wrapper 特征选择算法降低数据的属性维度。另外,对原始数据集进行 SMOTE 合成后,使用 Filter 特征选择算法,通过设置不同的阈值选择最佳的降维因子,并选择效果最好的阈值,不同阈值的算法性能如图 1 所示。



(a)Sonar (b)Vote

(c)Soybean

图 1 设置不同的阈值时所对应的算法性能

Fig. 1 Performance of algorithm when setting different thresholds

图 1 分别为对数据集 Sonar,Vote,Soybean 设置不同的阈值时对应的算法性能。由图 1 可知,通过设置不同的阈值(由信息增益的大小来选择特征子集)来选择特征子集,可以得出最优节点,如对于 Sonar 数据集,当 $R=0$ 时,可以得到最高的准确率。

不同方法在 8 个数据集上的 MCC 值对比如表 2 所列。

表 2 不同方法在 8 个数据集上的 MCC 值对比

Table 2 Comparison of MCC of different methods on 8 datasets

方法	MCC/%							
	Wave-form	Sick	Splice	Segment	Anneal	Soybean	Vote	Sonar
原始方法	69.1	82.7	92.5	94.7	94.8	90.3	90.8	50.9
SMOTE 方法	74.5	88.8	93.4	95.7	95.5	92.6	92.6	65.2
本文方法	76.7	90.1	95.5	96.3	97.2	94.1	94.3	64.9

由表 2 可知:1)从 NB-C4.5 算法的 MCC 可以看出,经过 SMOTE 处理后数据的不平衡性得到了很好的改善,且本文方法的 MCC 具有更好的效果;2)本文提出的方法同其他两种方法相比,具有更好的效果,经过 Filter-Wrapper 特征选择算法筛选后的特征子集具有更高的泛化能力。

3 种方法在 8 个数据集上所得特征子集数量与准确率对比如表 3 所列。

表 3 3 种方法在 8 个数据集上所得特征子集数量与准确率对比

Table 3 Comparison of attribute and precision of 3 methods on 8 datasets

数据集	Filter 特征选择算法		Wrapper 特征选择算法		本文方法	
	Attribute	*Precision*/%	*Attribute*	*Precision*/%	*Attribute*	*Precision*/%
Waveform	40	79.5	10	83.1	14	85.9
Sick	29	98.0	11	98.1	6	98.3
Splice	61	95.3	23	96.3	20	97.6
Segment	19	95.5	7	96.2	8	97.4
Anneal	38	98.0	9	98.9	11	98.6
Soybean	35	91.6	18	94.4	19	95.3
Vote	16	95.6	5	96.8	9	98.1
Sonar	60	75.5	4	75.0	8	89.5

由表 3 可知,Filter 特征选择算法筛选后的特征子集数量最多,且准确率最低。Wrapper 特征选择算法筛选的特征子集较少,具有较好的效果,但是其工作效率最低。通过与这两种传统方法进行比较可以得出,本文提出的方法具有较高的准确率,且工作效率介于 Filter 特征选择算法与 Wrapper 特征选择算法之间。

4.4 UCI 数据集上其他方法与本文方法的对比

将本文所提算法与其他方法进行对比,对比方法包括以朴素贝叶斯为核心的方法、决策方法、神经网络、支持向量机等。对比结果如表 4 所列。

表 4 不同方法与本文方法的对比

Table 4 Comparison of different methods and proposed method

(单位:%)

数据集	模型	*Precision*
Waveform	PSO-SWCM[28]	86.40
	V-CELMC1[29]	87.88
	EITL[30]	85.80
	本文方法	85.90
Sonar	TEC[31]	87.47
	IGLB-NB[30]	87.62
	He-Bagging[32]	82.54
	本文方法	89.50
Segment	HSICmkl[33]	96.43
	Libsvm[34]	97.10
	本文方法	97.40
Anneal	Boosting C4.5[36]	95.27
	Bagging C4.5[36]	93.75
	本文方法	98.60
Soybean	CN2[36]	82.70
	SVMmkl[33]	97.00
	HAC[37]	89.80
	本文方法	98.10
Splice	random forest[38]	96.40
	BoostC4.5[35]	95.70
	本文方法	97.60

由表 4 可知:1)本文方法不仅优于最新提出的朴素贝叶斯算法,与近期提出的决策树(如 IGLB-NB,BoostC4.5)相比,本文方法同样具有很好的分类效果;2)与其他不同的方法

相比，本文方法也具有非常好的效果，不仅仅局限于决策树和贝叶斯之间。

常规方法与本文方法对 ECG 数据进行分类的准确率结果如表 5 所列。对比方法包括决策树(C4.5)、朴素贝叶斯网络(NB)、朴素贝叶斯决策树(NB-C4.5)和基于 Wrapper 特征选择的包装学习(Wrapper-NB-C4.5)。

表 5　常规方法与本文方法对 ECG 数据进行分类的准确率比较

Table 5　Accuracy comparison of classification of ECG data between conventional methods and proposed method

(单位：%)

模型	*Precision*
C4.5	75.8
NB	73.4
NB-C4.5	75.1
KNN	75.6
Bagging C4.5	75.9
Wrapper-NB-C4.5	76.2
本文方法	90.1

如表 5 所列，本文所提方法相比其他方法获得了更高的精度，因此将本文方法应用于 ECG 数据集中可以获得更高的准确率，这意味着所提方法可以有效地应用于 ECG 医疗诊断中。

结束语　在很多现实领域中，数据集通常是不平衡、高维且数据量较大的，如本文研究的 ECG 数据集，因此，如何有效地处理 ECG 医疗数据是近期研究的热点。本文提出的 SMOTE-Filter-Wrapper 方法可以有效地处理 ECG 医疗数据集，提高了算法的预测能力。此外，本文方法不仅降低了数据集的不平衡性，而且极大地提高了 Wrapper 方法的工作效率。更重要的是，其提高了 NB-C4.5 分类器的性能，并能应用于 ECG 医疗数据集中，在临床医疗诊断中取得了很好的效果，可以有效地解决临床医疗决策的关键问题。

参 考 文 献

[1] CHENG Y Y, QU H B, ZHANG B L. Chinese medicine industry 4.0: advancing digital pharmaceutical manufacture toward intelligent pharmaceutical manufacture[J]. China Journal of Chinese Materia Medica, 2016, 41(1): 1.

[2] LI X, LI D, WAN J, et al. A review of industrial wireless networks in the context of Industry 4.0[J]. Wireless Networks, 2017, 23(1): 23-41.

[3] WILK S, SLOWINSKI R, MICHALOWSKI W, et al. Supporting triage of children with abdominal Pain in the emergency room[J]. European Journal of Operationl Research, 2005, 160(3): 696-709.

[4] CHEN J M, SUN Y X. Experiments study on a dynamic priority scheduling for wireless sensor networks[C]// Proceedings of Mobile Ad-hoc and Sensor Networks. Wuhan, 2005: 613-622.

[5] QUINLAN J R. Induction of decision tree[J]. Machine Learning, 1986, 1(1): 81-106.

[6] QUINLAN J R. Learning Efficient Classification Procedures and Their Application to Chess End Games[M]// Machine Learning. Springer Berlin Heidelberg, 1984.

[7] MICHALSKI R S, CARBONELL J G, MITCHELL T M. Machine learning: an artificial intelligence approach[M]. London: Morgan Kaufmann, 1984: 463-482.

[8] PALACIOS-ALONSO M A, BRIZUELA C A, SUCAR L E. Evolutionary learning of dynamic Nave Bayesian classifiers[J]. Journal of Automated Reasoning, 2010, 45(1): 21-37.

[9] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1): 321-357.

[10] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]// Proceedings of the 2005 International Conference on Intelligent Computing. Berlin: Springer Press, 2005: 878-887.

[11] YEN S J, LEE Y S. Cluster-based under-sampling approaches for imbalanced data distributions[J]. Expert Systems with Applications, 2009, 36(3): 5718-5727.

[12] BATISTA G, PRATI R C, MONARD M C. A study of the behaviour of several methods for balancing machine learning training data[J]. SIGKDD Explor, 2004, 6(1): 20-29.

[13] 边肇祺，张学工. 模式识别(第 2 版)[M]. 北京：清华大学出版社，2000.

[14] LANGLEY P. Selection of relevant features in machine learning[C]// Proceedings of the AAAI Fall Symposium on Relevance. New Orleans, 1994: 1-5.

[15] ZHOU X B, WANG X D, DOUGHERTY E R. Nonlinear-Probit Gene Classification Using Mutual Information and WaveIet-Based Feature Selection[J]. Biological Systems, 2004, 12(3): 371-386.

[16] SINDHWANI V, RAKSHIT S, DEODHARE D, et al. Feature Selection In MLPs and SVMs Based on Maximum Output Information[J]. IEEE Transactions on Neural Networks, 2004, 15(4): 937-948.

[17] HSU W H. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning [J]. Information Sciences, 2004, 163(17): 103-122.

[18] LI L, WEINBERG C R, DARDEN T A, et al. Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method [J]. Bioinformatics, 2001, 17(12): 1131-1142.

[19] INZA I, LARRANAGA P, BLANCO E R, et al. Filter Versus Wrapper Gene Selection Approaches in DNA Microarray Domains[J]. Artificial Intelligence in Medicine, 2004, 31(2): 91-103.

[20] ZHANG Y Y, XIANG Y, JIANG R Q, et al. Analysis and Implementation of Map Reduce Parallelization of Naive Bayes Algorithm [J]. Computer Technology and Development, 2013, 23(3): 23-26. (in Chinese)
张依杨，向阳，蒋锐权，等. 朴素贝叶斯算法的 MapReduce 并行化分析与实现[J]. 计算机技术与发展，2013，23(3)：23-26.

[21] DOMINGOS P, PAZZANI M J. On The Optimality of The Simple Bayesian Classifier under Zero-One Loss[J]. Machine Learning, 1997, 29(2/3): 103-130.

[22] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[23] SEGAL I E A. note on the concept of entropy[J]. Journal of Mathematics and Mechanics, 1960, 9(4): 623-629.

[24] QUINLAN J R. C4.5: Programming for machine learning[M]. London, Morgan Kauffmann, 1993.

[25] BREIMAN L, FRIEDMAN J H, STONE C J, et al. Classification and regression trees[M]. Chapman and Hall, 1984.

(下转第 74 页)

不能互相访问;同时也将保证某个核心上的分区发生错误时不会传播,如 Core1 上的分区在运行过程中发生错误时不能传播,受影响的只是 Core1,其他分区仍会正常运转,这正体现了分区机制的容错和安全特性。

结束语 本文就 seL4 的多核调度与分区隔离进行了研究,设计并实现了在单核 seL4 的基础上分别加入多核和分区隔离的支持。最后将这两部分工作结合起来,提出了多核平台的分区机制方案。未来可在 qemu 模拟器上运行该方案。

参 考 文 献

[1] KLEIN G,ANDRONICK J,ELPHINSTONE K,et al. seL4:formalverification of an operating system kernel[J]. Communications of the Acm,2010,53(6):107-115.

[2] PETERS S,DANIS A,ELPHINSTONE K,et al. For a Microkernel,a Big Lock Is Fine[C]// Proceedings of the 6th Asia-Pacific Workshop on Systems. Newyork. NY,USA:ACM,APSys'15,2015.

[3] ALVES-FOSS J,OMAN P W,TAYLOR C,et al. The MILS architecture for high-assurance embedded systems[J]. Internationl Journal of Embedded Systems,2006,2(3/4):239-247.

[4] PRISAZNUK P J. ARINC 653 role in integrated modular avionics(IMA)[C]// 2008 IEEE/AIAA 27th Digital Avionics Systems Conference. 2008.

[5] COMMITTEE A E E,et al. Avionics application software standard interface part 1required services[EB/OL]. https://standards. globalspec. com/std/280829/arinc-653p1.

[6] HAN S,JIN H W. Kernel-level ARINC653 partitioning for Linux[C]// Proceedings of the 27th Annual ACM Symposium on Applied Computing. 2012:1632-1637.

[7] DELANGE J,POK L L. An ARINC653-compliant operating system released under the BSD license[C]// 13th Real-Time Linux Workshop. 2011.

[8] ASBERG M,NOLTE T. Towardsauser-mode approach to partitioned scheduling in the seL4 microkernel[J]. ACM SIGBED Review,2013,10(3):15-22.

[9] FORD B,SUSARLA S. CPU inheritance scheduling[C]// Proceedings of the Second USENIX Symposium on Operating Systems Design and Implementation(OSDI'96). 1996:91-105.

[10] LACKORZYN′SKI A,WARG A,VÖLP M,et al. Flattening hierarchical scheduling[C]// Proceedings of the Tenth ACM International Conference on Embedded Software. 2012:93-102.

[11] HEISER A L G,LYONS A,VANGE M,et al. FlaRe:Efficient Capability Semantics for Timely Processor Access[EB/OL]. https://people. mpi-sws.org/~bbb/papers/pdf/preprint-FlaRe. pdf.

[12] SERGEY B,ALEXANDRA F. User-level schedulingon NUMA multicore systems under Linux[EB/OL]. http://citeseerx. ist. psu. edu /viewdoc/download?doi=10. 1. 1. 369. 9422&rep=rep1&type=pdf.

[13] MERCER C W,SAVAGE S,TOKUDA H. Processor capacity reserves:An abstraction for managing processor usage[C]// Fourth Workshop on Workstation Operating Systems. 1993:129-134.

[14] UHLIG R,NEIGER G,RODGERS D,et al. Intel virtualization technology[J]. Computer,2005,38(5):48-56.

[15] MERKEL D. Docker:lightweight linux containers for consistent development and deployment[J]. Linux Journal,2014,2014(239):2.

[16] NICTA. The seL4 microkernel[EB/OL]. https://github. com/seL4/seL4.

[17] KLEIN G,ANDRONICK J,ELPHINSTONE K,et al. Comprehensive formal verification of an OS micro-kernel[J]. ACM Transactions on Computer Systems(TOCS),2014,32(1):32-102.

(上接第 69 页)

[26] FAN J C,ZHANG W Y,LIANG Y Q. Decision tree classification algorithm based on Bayesian method[J]. Journal of Computer Applications,2005,25(12):2882-2884. (in Chinese)
樊建聪,张问银,梁永全. 基于贝叶斯方法的决策树分类算法[J]. 计算机应用,2005,25(12):2882-2884.

[27] FRANK A,ASUNCION A. UCI Machine Learning Repository[DB/OL]. http://archive. ics. uci. edu/ml/Irvine,CA:University of California,School of Information and Computer Science.

[28] YANG L Y,ZHANG J Y,WANG W J. Selecting and Combining Classifiers Simultaneously with Particle Swarm Optimization[J]. Information Technology Journal,2009,8(2):241-245.

[29] SINGH R G,PANDEY A. The Impact of Randomization on Circular-Complex Extreme Learning Machine for Real Valued Classification Problems[J]. International Journal of Computer Applications,2014,103(2):1-7.

[30] LIPITAKIS A D,ANTZOULATOS G S,KOTSIANTIS S,et al. Integrating global and local boosting[C]// 2015 6th International Conference on Information,Intelligence,Systems and Applications(IISA). IEEE,2015:1-6.

[31] RAHMAN A,VERMA B. A novel ensemble classifier approach using weak classifier learning on overlapping clusters[C]// International Joint Conference on Neural Networks. IEEE,2010:1-7.

[32] COELHO A L V,NASCIMENTO D S C. On the evolutionary design of heterogeneous bagging models[J]. Neuro Computing,2010,73(16):3319-3322.

[33] CHEN J,JI S,CERAN B,et al. Learning subspace kernels for classification[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2008:106-114.

[34] DO T N,POULET F. Enhancing svm with visualization[C]// International Conference on Discovery Science. Springer Berlin Heidelberg,2004:183-194.

[35] QUINLAN J R. Bagging,boosting,and C4. 5[C]// Association for the Advancement of Artificial Intelligence. 1996:725-730.

[36] CLARK P,BOSWELL R. Rule induction with CN2:Some recent improvements[C]// European Working Session on Learning. Springer Berlin Heidelberg,1991:151-163.

[37] JO H,NA Y,OH B,et al. Attribute value taxonomy generation through matrix based adaptive genetic algorithm[C]// 20th IEEE International Conference on Tools with Artificial Intelligence. IEEE,2008,1:393-400.

[38] SAEED A A,CAWLEY G C,BAGNALL A. Benchmarking the semi-supervised naïve Bayes classifier[C]// International Joint Conference on Neural Networks. IEEE,2015:558-561.