混合式采样方法之朴素贝叶斯决策树算法应用于心电图数据

许召召[1]，李京华[1]，陈同林[1]，李昕洁*[1+]

1 云南大学 软件学院，云南省昆明市 650091

+ 通讯作者：邮箱：camhero@hotmail.com

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而又准确地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、不平衡和有噪声的，基于此，本文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——用于支持临床医疗决策。特别的，本文提出的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。将本文的算法应用于 ECG 临床医疗决策中，并取得了很好的效果。

关键词：数据平衡；wrapper 特征选择；朴素贝叶斯；决策树

The Naive Bayesian Decision Tree Algorithm for Hybrid Sampling Method Applied to ECG Data

Abstract： It is still a very serious problem that how efficiently and accurately dig out the medical data generated by the Internet-based wisdom medical system with "Industrial 4.0". However, the medical data is often high-dimensional, unbalanced and noisy, on account of which this paper proposes a new data processing method combing SMOTE method with Filter-Wrapper feature selection algorithm to support clinical decision-making. In particular, the method proposed in this paper not only overcomes the situation of bad prediction result of the independent assumptions in the practical attribute application of Naive Bayesian, but also avoid over-fitting problem caused by constructing the model of C4.5 decision tree. What's more, the algorithm proposed is applied to ECG clinical decision-making, which has achieved very good results.

Key words: data balance, wrapper feature selection, naive bayesian, decision tree

0. 引言

2013 年，德国政府率先提出“工业 4.0”，通过将信息技术与制造业融合，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病人资料和医疗数据，这些数据包含着十分重要的信息，如何提取出这些信息，成为现阶段医疗数据挖掘中的热点。医疗数据往往是不完整、含有噪声和不平衡的[3][4]，因此要进行准确的医疗决策，不仅要选择合适的分类算法，而且也必须对物联网智能设备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，以及 Filter 特征选择算法选择的特征与后续算法偏差较大的缺点。与此同时，选择合适的分类算法也是本文研究方向之一。通过数据挖掘算法提取医疗数据中隐藏的有用的信息，主要通过数据挖掘算法分析训练数据样本，给出高效而又准确的临床诊断。分类算法又是数据挖掘领域中的一个研究热点，常用的分类算法主要有神经网络、决策树、朴素贝叶斯等[5][6][7]。

朴素贝叶斯是一种具有高效而又容易理解的分类算法[8]，但其条件独立性和属性重要性相等的假设并不适合实际应用，影响了分类器的性能。同样的，决策树算法在构造时，由于数据往往含有噪声，会造成过度拟合并且导致分类精度下降。因此本文通过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，因此能否高效的构建分类模型是本文研究的主要方向之一。

将经过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，这些都将会在实验中得到验证。将本文所提算法用于心电图数据集中，可以得到高效而又准确的病人病症分类结果，在心电图治疗之中起到了辅助作用。

1. 相关工作

1.1. 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其中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。 基于本文对心电图医疗数据诊断的背景，在对医疗数据进行预测时，需要对原始数据集进行平衡处理，以确保医疗诊断的准确性。

基于数据抽样方法主要分为欠采样和过采样、欠采样和过采样相结合的方法。过采样中，最为常用的是 Chawla 等人[9]提出的 SMOTE 算法，该算法通过合成少数类样本，得到了广泛的应用；类似的还有 Borderline-SMOTE 算法[10]，这是一种改进的 SMOTE 算法，只对数据的少数类的边界样本进行抽样处理，使得增加的样本实有价值的。欠采样中 S.J.Yen 等人[11]提出一种基于聚类的欠抽样方法，该方法通过聚类后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。

其中 SMOTE[12]算法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。然后按照公式（1）将新和成的样本加入到数据中。

$$Pnew = x + rand(0,1) \times (y_i - x) \qquad (1)$$

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

1.2. 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]，根据是否独立于算法模型分为过滤式 (Filter)和封装式 (Wrapper)[14]两种。其中，滤波式(Filter)特征选择算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。而嵌入式(Wrapper)评价策略的特征选择算法，其原理是使用后续学习模型的算法的预测性能来评价特征子集的好坏。目前，此类方法是特征领域的研究热点。Hsu 等人[17]提出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。类似的还有使用遗传算法结合人工神经网络来选择特征子集[18]，也有提出用启发式搜索策略(SBS，SFS，FSFS)和分类器性能评价准则相结合来评价特征子集[19]。

虽然这两种方法都有各自的优点，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又高，花费的代价又低的特征选择方法是我们本文研究的一个重点之一。

1.3. 贝叶斯网络

贝叶斯分类算法基于贝叶斯定理，使用概率的形式表示样本的不确定性，通过改变时间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个简单的概率分类器，通过将贝叶斯定理应用于数据的朴素独立假设。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]，本文选择朴素贝叶斯算法与决策树算法相融合。

朴素贝叶斯算法的主要思想如下：

设 n 维特征向量，并代表 n 个属性的值，即

$A = \{a_1, a_2, ..., a_n\}$，给定一个未知数据集，得到目标值为

$$V_{map} = \arg\max P(V_j \mid a_1, a_2, ..., a_n) \qquad (2)$$

其中，$V_j \in V$ 。

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

$$P(V_I \mid X) = \frac{P(X \mid V_i)P(V_i)}{P(X)} \qquad (3)$$

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

由于朴素贝叶斯假设各个属性之间相互独立，即：

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

其中先验概率（ $p(x_k \mid V_i), k \in 1,2,...,n$ ，）可以从训练数据中求得。

如上所述，对于一个未知数据集 X，计算属于类别 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。

1.4. 决策树

决策树是通过一系列规则对未知数据进行分类的过程。因其规则简易、计算量小以及具有较高的分类准确率等优点得到广泛的应用。决策树的构造过程与人做决策的原理相似，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 提出 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，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。为了克服信息增益来选择属性时偏向选择值多的属性的不足。Quinlan 提出 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，重复的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造成决策树深度增加，划分数值型属性时则需要大量的排序和计算。

2. 算法设计与分析

本文提出的方法是基于 NB-C4.5 算法的基础，在整个流程结构中，首先通过 SMOTE 方法，可以有效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，最后使用 NB-C4.5 算法进行模型训练。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，获得分类预测结果。

2.1. 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函数，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集。

SMOTE-Filter-Wrapper 算法步骤如下：

Input: 数据集 D，评价学习器 NB-C4.5，

1. 对于数据集 D，若抽样率为 m，对于每个负类样本点 x_i，找出它的 k 个负类近邻点。从中任选 m 个近邻点 $y_{ij}(j = 1, 2, \ldots, m)$。按照公式（1），合成 N 个新的负类样本；

2. 将步骤 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，通过设置不同的阈值(根据属性的信息增益，设置阈值)，并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；

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'。

Output： 新数据集 D'

2.2. 朴素贝叶斯决策树算法构造

据上文所述朴素贝叶斯算法的思想，以下给出一种朴素贝叶斯和 C4.5 决策树融合的 NB-C4.5 分类算法的设计思路：

（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，即 BN 的值为 0.

其中 C4.5 决策树的信息增益率计算方法[26]如下所述：

假设训练样本 S，样本中有 n 个类。那么 S 的熵（信息增益）可以表示为等式（5）：

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

在等式（5）中，p_i 表示属于类 i 的训练样本概率。如果 A 是具有 v 个不同子集的数据对象属性 $\{S_1, S_2, ..., S_V\}$，其中 S_j 由 S 的属性 A 中的值为 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如下：

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + L + S_{nj}}{S} I(S_{1j}, L, S_{nj}) \quad (6)$$

其中 $\frac{sij+...+snj}{s}$ 是第 j 个子集的权重，对于给定的子集，有：

$$I(S_{1j}, L, S_{nj}) = -\sum_{i=1}^{n} p_{ij} log_2 p_{ig} \qquad (7)$$

表示属于类的样本的概率。那么属性 A 的信息增益可以描述如下：

$$(Gain(A)) : Gain(A) = I(S) - E(A) \qquad (8)$$

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，因此在这里使用信息增益率来选择分支属性。

$$Ratio(S, A) = Gain(S, A) / Split(S, A) \quad (9)$$

其中

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

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。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。

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，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那一类作为该训练样本的类别[27]。

2.3. 算法分析

本文所使用的算法模型同决策树一样，具有规则简易、分类性能较高等优点。

实验主要分为三步: 1）首先使用 SMOTE 方法平衡数据集，2）对 1 中处理过数据，应用 Filter-Wrapper 特征选择算法，筛选出最优的特征子集，3）最后使用 NB-C4.5 算法对数据集进行建模预测。

本文的方法具有的优点如下:

（1）具有更好的预测能力。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不同的属性但是相同的信息增益率，那么就会造成属性二义性，对于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去处数据二义性，从而提高了模型的预测能力。

（2）具有更强的分类鲁棒性。数据挖掘一般是不完整、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，提高了 NB-C4.5 分类器的鲁棒性。

（3）提高了算法的性能。对于医疗数据而言，往往是不平衡的，通过本文的方法不仅降低了医疗数据的不平衡性，还提高了特征选择的效率，提高了整体算法的性能。

初步的处理。在此基础上，对数据结果部分贴上标签，将心电图结果全部划分为正常与不正常两类，因此数据变成一个二分类问题。数据的总数有 222594 条实例，13 个特征属性和 1 个类别标签属性。所有实验数据特征描述如下如表 1 所示。

表 1 UCI 标准数据集和 ECG 数据集的特征描述

3. 实验结果与分析

3.1. 数据集和性能指标

本节主要通过实验研究基于 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[28]中下载的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，并由该医院的医生对数据进行

为了验证本文算法的性能，选取了一些常用的评价指标：准确率（ *Accuracy* ）、*MCC*

$$Accuracy = \frac{P_T}{P_T + P_F} \quad (11)$$

其中 P_T 是正确分类的实例， P_F 是错误分类的实例， N_T 是属于该类但未分配给该类的实例。另外，MCC 是另一个有效的不均衡数据分类性能评价手段，对于一个给定的两分类问题，MCC 的计算方法为：

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

3.2. 实验设计

仿真实验中，采用 10 层交叉（10-fold Cross Validation）的方法在数据集上进行实验测试，分别记录其分类准确率、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，然后比较传统特征选择算法与本文的算法，最后将本文的算法应用于临床医疗决策中。

3.3. 实验结果

3.3.1. 基于 SMOTE 数据重抽样方法的 Filter-Wrapper 特征选择算法

对本文提出的方法进行验证，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，选择降维最优的阈值，相关的阈值选择如下图所示：

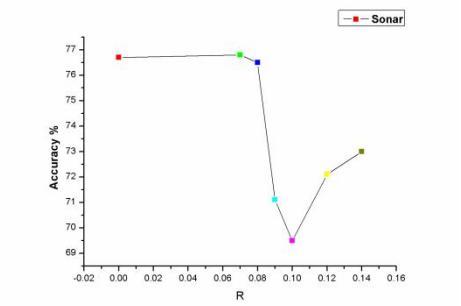

图 1 通过设置不同的阈值所对应的算法性能 (Sonar)

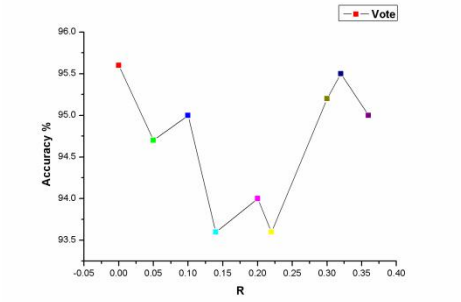

图 2 通过设置不同的阈值所对应的算法性能 (Vote)

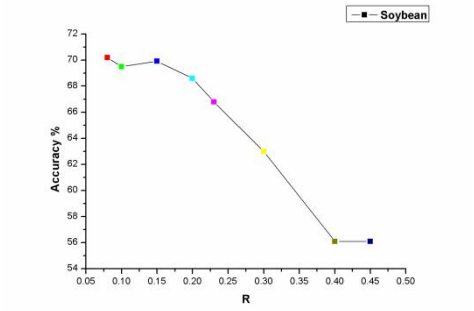

图 3 通过设置不同的阈值所对应的算法性能 (Soybean)

由图 1,2,3 可知，通过设置不同的阈值（即根据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 Sonar 数据集，当 R=0 时，可以得到最高的准确率。

数据集	样本数	属性数	类别数
Sick	3772	29	2
Spice	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

表 2 原始 UCI 数据集与经过处理后的数据集之间的对比

数据集	原始数据集		SMOTE 方法后		本文的方法	
	Accurac%	MCC%	Accurac%	MCC%	Accurac%	MCC%
Waveform	79.5	69.1	83.4	74.5	85.9	76.7
Sick	98.0	82.7	97.7	88.8	98.3	90.1
Splice	95.3	92.5	95.8	93.4	97.6	95.5
Segment	95.5	94.7	96.3	95.7	97.4	96.3
Anneal	98.0	94.8	98.2	95.5	98.6	97.2
Soybean	91.6	90.3	93.3	92.6	95.3	94.1
Vote	95.6	90.8	96.4	92.6	97.8	94.3
Sonar	75.5	50.9	83.9	65.2	89.5	64.9

由表 2 可知：

1. 根据 NB-C4.5 算法的 MCC 可以得出，经过 SMOTE 处理后数据的不平衡性得到了很好的改善，并且本文方法的 MCC 具有更好的效果。

2. 本文提出的方法同原始数据集以及经过 SMOTE 方法处理后的结果相比，都有很好的效果。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。

由表 3 可知 Filter 特征选择算法筛选后的特征子集数最多，并且准确率最低。Wrapper 特征选择算法筛选的特征子集较少，有较好的效果，但是其工作效率有最低，通过比较两种传统的方法，可以得出本文提出的方法具高的准确率，并且工作效率介于 Filter 特征选择算法与 Wrapper 特征选择算法之间。

表 3 原始 UCI 数据集、经过 Wrapper 特征选择方法筛选过的 UCI 数据集与经过本文的方法的数据集（基于 SMOTE 方法的 Filter-Wrapper 特征选择算法）

数据集	Filter 特征选择算法		Wrapper 特征选择算法		本文的方法	
	Attribute	Accuracy%	Attribute	Accuracy%	Attribute	Accuracy%
Waveform	40	79.5	10	83.1	14	85.9
Sick	29	98.0	11	98.1	6	98.3
Spice	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	91.5

3.4. 基于 UCI 数据集的其他方法与本文的对比

将本文所提出的算法与其他方法所对比，对比方法中，不止包括以朴素贝叶斯为核心的方法，还包括决策、神经网络、支持向量机等方法。对比结果如下表：

表 4 不同研究方法与本文方法的对比

数据集	Model	Accuracy %
Waveform	LinearSVM-BGA[27]	87.12
	V-CELMC1[29]	87.88
	EITL[30]	85.80
	本文的方法	**85.90**
Sonar	Boosting-NN-Ada[31]	87.50
	IGLB-NB[30]	87.62
	He-Bagging[32]	82.54
	本文的方法	**89.5**
Segment	HSICmkl[33]	96.43
	Libsvm[34]	97.10
	C4.5[36]	96.79
	本文的方法	**97.40**
Anneal	Boosting C4.5[36]	95.27
	Bagging C4.5[36]	93.75
	本文的方法	**98.60**
Soybean	CN2[36]	82.70
	SVM_{mkl}[33]	97.00
	HAC[37]	89.80
	本文的方法	**98.10**
Splice	AB_{NN}-AB_{SVM}[38]	96.40
	SVM-ABSVM[38]	97.80
	BoostC4.5[35]	95.70
	本文的方法	**97.60**

由表 4 可知

1. 本文提出的方法不仅优于最新提出的朴素贝叶斯算法，同样的和近期提出的决策树相比，本文的方法也有着很好的分类效果，如 IGLB-NB、BoostC4.5。
2. 与其他不同的方法相比，本文的方法也有着非常的效果，不仅仅局限于决策树和贝叶斯之间对比。

常规方法与本文的方法对 ECG 数据进行分类的准确率比较如表 5 示，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　常规方法与本文的方法对 ECG 数据进行分类的准确率比较

模型	Accuracy %
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
本文的方法	**90.1**

如表 5 所示，本文所提出的方法比其他方法获得更好的精度。因此，将本文的方法应用于 ECG 数据集中，可以获得更好的准确率，这意味着所提出的方法可以有效的应用于 ECG 医疗诊断中。

4. 总结

在很多现实领域中，数据集通常是不平衡的、高维的并且数据量较大，如本文所研究的 ECG 数据集，因此如何有效的处理 ECG 医疗数据是近期研究的热点。本文提出 SMOTE-Filter-Wrapper 方法可以有效处理 ECG 医疗数据集，提高了算法的预测能力。此外，本文的方法不仅降低了数据集的不平衡性，而且极大的提高了 Wrapper 方法的工作效率。更重要的是提高了 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，因此本文的方法可以有效的解决临床医疗决策的关键问题。

Reference

[1] Hermann M, Pentek T and Otto B. Design Principles for Industrie 4.0 Scenarios: A Literature Review. Working Paper No. 01.2015, Technische Universitat Dortmund, 2015.

[2] Kagerman H., Wahlster W. and Helbig J. Recommendations for implementing the strategic initiative INDUSTRIE 4.0. Final report of the Industrie 4.0 Working Group, Federal Ministry of Education and Research, 2013.

[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]. Proc 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] Quilan J R. Learning efficient classification procedures and their application to chess end games [J]. **Machine Learning: An Artificial Intelligence Approach**, 1983, 1.

[7] Michalski R S，Carbonell J G，Mitchell T M. Machine learning: an artificialintelligence approach，CA: Morgan Kaufmann，1983: 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**, 2002, **16**: 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] Mazurowski M A, Habas P A, Zurada J M.et al. Training neural network classifiers for medical decision making :The effects of imbalanced datasets on classidication performance [J]**.Neural networks**,2008,**21**(2):427-436.

[12] Chawla, N. V., Bowyer, K. W., Hall, L. O. et al. SMOTE: Synthetic Minority Over -sampling Technique. JAIR 16 (2002), 321–357.

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

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

[15] ZhouXiaobo,WangXiaodong, DoughertyER. Nonlinear-Probit Gene Classification Using Mutual Information and WaveIet-Based Feature Selection. 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. IEEE Trans on Neural Networks,2004,15(4):937 一 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.Bioinformatics,2001,17(12):1131 一 1142.

[19] Inza l, Larranaga P, Blanco E R, et al. Filter Versus Wrapper Gene Selection Approaches in DNA Microarray Domains. **Artificial Intelligence in Medicine**,2004,**31** （ 2 ）：91-103.

[20] 张依杨，向 阳，蒋锐权，等. 朴素贝叶斯算法的 Map R educe 并行化分析与实现[J]. 计算机技术与发展，2013，23(3) : 23 – 26.

[21] P Domingos, and M J Pazzani. On The Optimality of The Simple Bayesian Classifier under Zero-One Loss [J]. **Machine Learning**, 1997.**29**(2-3), pp. 103-130.

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

[23] Shannon C E.A note on the concept of entropy [J].**Bell System Tech.I,** 1948, **27**, pp:379-423.

[24] Quinlan J R.C4.5:Programming for machine learning[M].Morgan Kauffmann,[s.n.],1993.

[25] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees[M].[s.l.], CRC Press,1984

[26] 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] Aksakalli V, Malekipirbazari M. Feature selection via binary simultaneous perturbation stochastic approximation [J]. **Pattern Recognition Letters**, 2016, **75**: 41-47.

[28] 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, 2010.

[29] Singh R G, Pandey A. The Impact of Randomization on Circular-Complex Extreme Learning Machine for Real Valued Classification Problems[J]. I**nternational Journal of Computer Applications**, 2014, **103**(2).

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

[31] Maclin R, Opitz D. An empirical evaluation of bagging and boosting [J]. **AAAI/IAAI,** 1997,: 546-551.

[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]//AAAI/IAAI, Vol. 1. 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]//Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on. IEEE, 2008, 1: 393-400.

[38] De Keyser R. Engineering Applications in Artificial Intelligence [J]. 2005.