基于 QBSO 算法的肝病检测

陆军华 李丹
四川大学锦城学院 四川 成都 610065

【摘 要】在机器学习和数据挖掘中，挑选最合适的特征对创建高效的数据模型应是非常关键的。但是目前对于优化特征选择还是比较困难的。本文提到的 QBSO 算法是在蜂群优化算法的基础上，用强化学习来执行其子集生成步骤中的搜索过程，并将之用于特征选择。本文将 QBSO 算法用于对印度肝病患者数据集的特征选择，再用 KNN 算法进行分类，来评估这些特征子集，并且与决策树相比较，选择出更合适的特征，提高对检测的准确率。

【关键词】特征选择；蜂群优化；强化学习；肝病患者

1 介绍

1.1 策树

特征选择作为一个数据预处理过程，在数据挖掘、模式识别和机器学习中有着重要地位。数据特征的多少会直接影响算法的准确度，当然，也并不是说数据的特征越多越好，因为数据的特征过多就会在训练分析模型的过程中增加时间开支，使分类器的性能下降。特征选择的目的就是为了减少过多与目标相关度不高的特征值，保留相关度高的特征值，并且我们选用的特征值在有良好的可靠性、尽可能高的权威性的前提下，数目尽可能少，另一方面损失的信息量尽可能小，可以减小疑难问题的复杂度，提升学习算法的预估精确度。[2]

特征选择的一般是从特征是否发散和特征与总体目标的相关性这两个角度展开。特征不发散是指样本在这个特征上几乎没有差异，即这个特征对我们进行分类的贡献度不高。特征与目标的相关度越高，这个特征对我们进行分类的贡献度就越高。特征选择一般有两种办法：过滤法，包装法其中，过滤法是指先设定一个阈值，再将特征与目标的相关度进行一个打分，相关度越高的，分数越高，就被保留下来，反之舍去；包装法是指每次选择若干个特征值，根据将要使用的目标函数对所选择特征进行预测效果的评分，分数高者保留。

QBSO 算法是在 BSO 算法的基础上，将 BSO 算法中的迭代搜索的过程用增强学习搜寻最佳途径的方法所取代，以此来展开简单的局部搜索。并将其用于特征选择，从而获得一组与目标相关度高的且分类错误率小的分类特征。[1]

2 算法

2.1 BSO 蜂群算法

BSO 是一种蜂群算法，其灵感来源于自然蜜蜂能够自我组织，自我适应环境和动态的任务分配的有趣的觅食行为，这个算法已经成功地应用于各种优化问题。蜜蜂的觅食行为是一个迭代的过程，其令人满意的表现可以解释为集约化和多样化之间的良好平衡，分别导致了对搜索空间的良好开发和探索。

蜂群算法是模仿蜜蜂觅食的行为，它的基本步骤是：第一步，先确定蜜蜂的搜索区域；第二步，蜜蜂搜索的迭代过程；第三步，选择合适的特征子集。在第一次的迭代过程中，开始先随机或通过启发式生成一个初始的解（就是蜜蜂的初始位置）叫做参考解 1，然后参考解都有序的存在 Tabu 表中；而搜索区域的 N 个解的集合是由参考解确定搜索区域中的 N 个解的集合，然后从搜索区域中分配一个解决方案给每一只蜜蜂，然后更新蜜蜂的位置，形成新的参考解；然后每一次迭代都是从参考解确定搜索区域中的 N 个解的集合，形成搜索区域，在搜索。然后每次按照相同的迭代过程，只是除了第一次迭代外的之后的迭代的初始值都为上一次新生成的参考解，最后达到给定的阈值或者得出来最优的特征子集。

BSO 蜂群算法还有两个重要的参数，分别是：flip 和 MaxChances。其中前者用于确定搜索区域的解集，这些解与参考解的距离与 flip 成反比。因此，必须仔细选择此参数的值，以确保搜索空间的良好覆盖率。后者是指在转移到另一个搜索区域之前所获得的机会数。它的作用是避免在局部最优值中得到存量。在选择参考方案时，被考虑在内，并允许通过明智地应用强化和多样化原则来保证开发和勘探之间的良好平衡。每当当前的解决方案得到改进，就会进行强化。然而，在迭代之后，如果没有发现任何改进，就会启动分散化。它包括从 Dance 表中选择距离存储在 Tabu 列表中的所有参考解决方案最远的解决方案。当找到最优解或达到最大迭代次

数时，算法停止。

2.2 BSO-FS

BSO 能够有效的搜寻到最优的解决方案，因此将 BSO 算法用于特征选择可以有效地去除特征中的无关特征，从而降低问题的复杂度，提高算法预测的精确度。当然 BSO 被应用于特征选择要对一般算法进行调整，使其适应问题的特性。（1）编码：将特征值按照原始特征长度的二进制向量表示的，在选择了某个特征之后就会将向量在对应的位置上，置为 1，反之，没有被选择的特征所在对应的位置上，置为 0；（2）拟合度：表示解决方案经过分类器之后的分类精度。（3）搜索区域：太小的值将有利于探索而不是利用搜索空间，而太高的值可能导致算法收敛到局部最优。

2.3 BSO 的优化算法：QBSO 算法

QBSO 算法是将 Q-learning 整合到 BSO 中，以便让蜜蜂在搜索过程中根据自己的经验进行学习。事实上，在 BSO 中，在蜜蜂被分配了一个解决方案后，它会执行一个经典的局部搜索，在这个过程中，它评估了它附近的所有解决方案，并返回最好的一个。因此，蜜蜂在这个过程中从不使用自己的经验。在这项工作中，我们提出用 Q-learning 算法代替蜜蜂进行的局部搜索，它在搜索过程中收集经验，并且可以获得其他蜜蜂的经验。组成坏境的状态应是蜜蜂附近的所有可能的特征子集。一个解决方案是一个布尔向量，表示哪些特征属于特征子集。

2.4 KNN 算法

KNN 算法是一种有监督学习的分类算法。KNN 的原理是选用在未知样本的必然范围内的 K 个样本，观察 K 个样本所属的类别，如果 K 个样本中大多数都同属于某一个类别，则未知样本被认为是该类的类别。[3]

算法流程：KNN（D，d，K 其中 D 为训练数据集，d 为不知道类别的测试样例）

（1）计算 d 和 D 中所有样例的距离；这是 KNN 算法最关键的过程。在距离（相似度）计算方法中，最常用的是“欧氏距离”，公式如下：

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

（2）选择 D 中离 d 最近的 K 个样本

（3）将这 K 个样本出现最多的类别赋予 d

KNN 算法最简单粗暴的就是将预测点与所有点距离进行计算，然后保存并排序，选出前面 K 个值看看哪些类别比较多。

2.5 决策树

决策树是一种树的结构，是一种分类算法。其中根节点和内部节点在决策树中代表了相应的测试条件，叶子节点代表最后的输出结果：（1）根节点：它是位于最上层，作为第一个测试条件，值得注意的是：一颗决策树有且只有一个这样的节点，没有入边，拥有零条或以上的出边；（2）内部节点：位于根节点之下，作为一种测试条件。这样的节点有一条入边，拥有两条或以上的出边；（3）叶子节点：是决策树最后的节点，它被用来判定最后的结果。这种节点只有一条入边而没有出边；从最上层的节点出发，到任意一个尾部的节点，都会形成一条规则。首先，将所有的数据特征看作是各个树的节点，遍历所有特征，其中每当遍历到其中某个特征时，对特征进行分割处理，并记录分割点的数据信息，作为划分子节点的纯度依据。其次，比较记录的数据特征以及判定最优特征，寻找最优划分方式，对样本数据集进行分割操作。最后，构建符合规则的决策树。值得注意的是，通过决策树所形成的规则应当是互斥且完备的，即对于任意一个样本数据，有且只有一条规则与其一一对应输出分类结果。

决策树算法的核心算法是 ID3 算法。ID3 算法融入了信息论中熵的概念，然后通过信息熵来推算信息增益，以此来衡量判断能力 [4]。其中信息增益的计算步骤：第一步：对给定的一个数据集，首先用熵计算数据集的混杂度。[5] 第二步：然后，把所有属性都计算一遍，找出用哪个属性来划分可以将数据集的混杂度减少最多。设属性 可以取 v 个值，假设我们用 来划分数据集，则我们可以将数据集划分成 v 个不相交的子集。第三步：计算属性的信息增益计算公式如下：

$$gain(D,\ A_i)=entropy(D)-entropy_{A_i}(D)$$

3 数据处理

3.1 数据集

人的各个器官分工协作，共同构建了一个完整健康的人，其中肝脏在人体中极其重要的作用，它一旦出现问题，就会引发多种肝脏疾病，例如：肝硬化、脂肪性肝病、自身免疫性肝病、药物性肝病，以及肝性脑病等，伴随着众多的身体危害。因此对肝病患者进行调查，记录患者身体的各项指标，统计出那些指标的改变可能会预示着肝病的发生，从而可以起到警示以及预判的作用。

本文用到了印度肝病患者数据集，该数据集收集于印度安得拉邦东北部（数据来源：http://archive.ics.uci.edu/）。这个数据集包含 10 个变量，分别是年龄、性别、总胆红素、直接胆红素、总蛋白、清蛋白、清蛋白 – 球蛋白比值、血清谷氨酸丙酮酸转氨酶、血清谷草转氨酶、

碱性磷酸酶。包含实例数：583 个，其中有 416 个肝脏患者记录和 167 个非肝脏患者记录；包含 441 条男性患者记录和 142 条女性患者记录 [6]。

属性信息：

（1）年龄（Age）：病人的年龄，其中凡是年龄超过 89 岁的患者都被列为 “90 岁”，如图 1 所示。

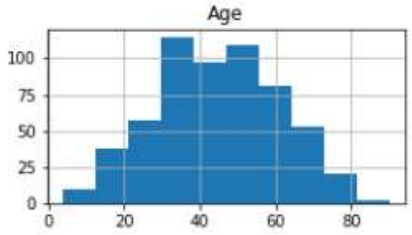

图 1 年龄（age）

（2）性别 (Gender)：患者的性别，如图 2 所示

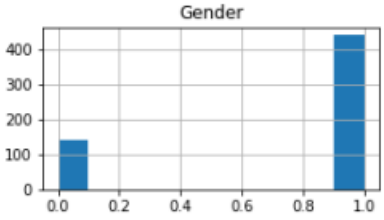

图 2 性别 (Gender)

（3）总胆红素 (TB)

（4）直接胆红素 (DB)

（5）碱性磷酸酶 (Alkphos)

（6）血清谷草转氨酶 (S_Alamine)

（7）血清谷氨酸丙酮酸转氨酶 (S_Aspartate)

（8）总蛋白 (TP)

（9）清蛋白 (ALB)

（10）清蛋白 – 球蛋白比值 (AG)(如图 3 所示)

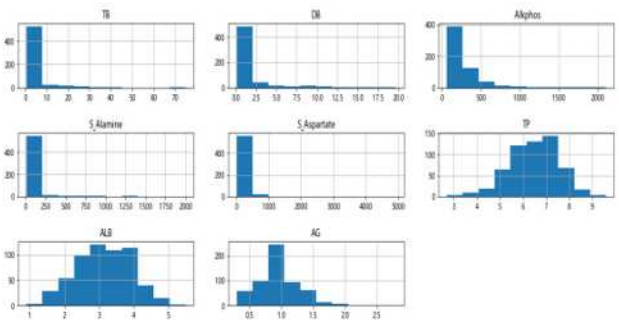

图 3 数据特征

3.2 数据处理

（1）将病人的年龄中凡是超过 89 岁的患者都被列为 “90 岁”；

（2）416 个肝脏患者记录（用 “1” 表示）和 167 个非肝脏患者记录（ “0” ）；

（3）441 条男性患者（用 “1” 表示）记录和 142 条女性患者（用 “0” 表示）记录；

（4）所使用的数据集是已经经过处理的，所以没有缺失值，就没有进行缺失值处理。

4 结果分析

本文将 QBS0 算法用在了印度肝病患者数据集上，对它的 583 条数据，其中有 416 个肝脏患者记录和 167 个非肝脏患者记录进行处理，并且用该算法对其中的 10 个变量：年龄、性别、总胆红素、直接胆红素、总蛋白、清蛋白、清蛋白 – 球蛋白比值、血清谷氨酸丙酮酸转氨酶、血清谷草转氨酶、碱性磷酸酶进行特征选择。在进行特征选择之后会发现，数据的特征值由原来的 10 个特征，选择出了 5 个相关度较高的数据特征，如图 4 所示。减少了特征的数量，但是最后分类的准确率提高了。增加了我们排除疑是患者，检测出真实患者的准确率。

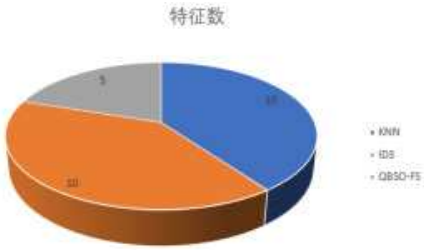

图 4 特征数比例

通过分类器对数据进行分类，在文中采用 9：1 的比例将原始数据集分成两类，分别为训练集和测试集，以此来分配样本数据，再进行 10 折交叉验证，得到了如表 1 所示的特征数据。

表 1 数据特征表

	QBSO–FS	KNN	ID3
特征数	5	10	10
样本数	583	583	583
分类数	2	2	2
准确率	0.74	0.6	0.62

分析发现：在使用 QBSO 算法对数据集进行特征选择后，对 10 个特征的编码为：1，1，1，1，0，0，0，0，0，1；其中在选择了某个特征之后就会将向量在对应的位置上置为 1，反之，没有被选择的特征所在对应的位置上置为 0；之后我们观察到准确率为 0.74，而没有进行特征选择，并用 KNN 算法和 ID3 算法对数据集进行分类，最后的准确率分别是 0.6 和 0.62。尽管在进行特征选择之后，特征数降低了，如图 5 所示，但是最后分类的准确率提高了，如图 6 所示。

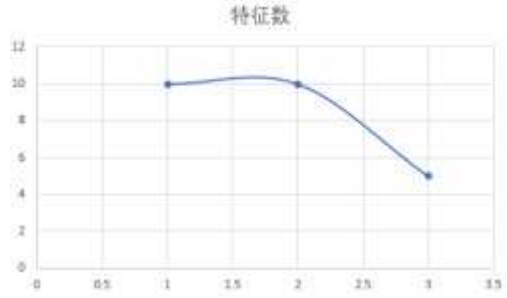

图 5 特征数

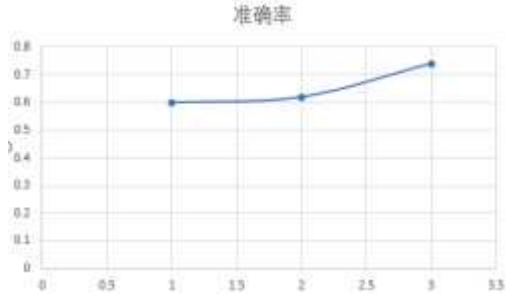

图 6 准确率

本文运用 KNN 和 ID3 两种模型分别对数据进行分析，可以看出在使用该算法对印度肝病患者数据集进行特征选择后能够有效的提高分类的正确率。运用该算法可以在一定程度上帮助我们提高对肝病的检测。

【参考文献】

[1] Sadeg S., Hamdad L., Remache A.R., Karech M.N., Benatchba K., Habbas Z. (2019) QBSO-FS: A Reinforcement Learning Based Bee Swarm Optimization Metaheuristic for Feature Selection. In: Rojas I., Joya G., Catala A. (eds) Advances in Computational Intelligence. IWANN 2019. Lecture Notes in Computer Science, vol 11507. Springer, Cham.

[2] 张俐 , 王枞 , 郭文明 . 利用近似马尔科夫毯的最大相关最小冗余特征选择算法 [J]. 西安交通大学学报 ,2018,52(10):141-145.

[3] 窦小凡 .KNN 算法综述 [J]. 通讯世界 ,2018(10):273-274.

[4] 谷雨 . 多源视觉场景下目标特征数据融合与识别技术的研究 [D]. 沈阳 : 沈阳理工大学 ,2017.

[5] 林志远 . 基于决策树算法的心脏病预测研究 [J]. 电子制作 ,2019(06):23-25.

[6] 李春生 , 焦海涛 , 刘澎 , 刘小刚 . 基于 C4.5 决策树分类算法的改进与应用 [J]. 计算机技术与发展 ,2020,30(05):185-189.

[7] 谢林瀚 . 基于数据挖掘探究肝脏疾病诊断模型 [J]. 中国科技投资 ,2019,(17):254-255.