(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110674860 A
(43)申请公布日 2020.01.10

(21)申请号 201910884965.4

(22)申请日 2019.09.19

(71)申请人 南京邮电大学
地址 210023 江苏省南京市亚东新城区文苑路9号

(72)发明人 仇晨晔

(74)专利代理机构 南京苏科专利代理有限责任公司 32102
代理人 姚姣阳

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 20/00(2019.01)

权利要求书2页 说明书9页 附图3页

(54)发明名称

基于邻域搜索策略的特征选择方法、存储介质和终端

(57)摘要

一种基于邻域搜索策略的特征选择方法、存储介质和终端，所述方法包括：获取特征搜索空间；所述特征搜索空间包括多个特征子集；采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集。上述的方案，可以提高所选取的特征子集的准确性，进而可以提高采用所选取的特征子集中的特征所构建的模型的准确性。

CN 110674860 A

1.一种基于邻域搜索策略的特征选择方法,其特征在于,包括:

获取特征搜索空间;所述特征搜索空间包括多个特征子集;

采用基于邻域搜索策略的微粒群算法,从所述特征搜索空间中的特征子集中搜索出全局最优特征子集。

2.根据权利要求1所述的基于邻域搜索策略的特征选择方法,其特征在于,所述采用基于邻域搜索策略的微粒群算法,从所述特征搜索空间中的特征子集中搜索出全局最优特征子集,包括:

初始化所述特征搜索空间中的特征子集;

计算所述特征子集之间的相似度矩阵;

基于计算得到的相似度矩阵,找到每个特征子集最相似的预设数量个邻近特征子集;

从所述预设数量个邻近特征子集中找出适应度数值最大的特征子集,分别作为每个特征子集对应的局域导引;

基于对应的局域导引对每个特征子集的位置进行更新,得到每个特征子集更新后的新特征子集;

当确定新特征子集的适应度数值大于对应的特征子集的适应度数值时,采用新特征子集代替对应的特征子集;

从所述计算所述特征子集之间的相似度矩阵开始执行下一次迭代,直至迭代次数达到预设的次数阈值,得到全局最优特征子集。

3.根据权利要求2所述的基于邻域搜索策略的特征选择方法,其特征在于,采用如下的公式计算所述相似度矩阵中的特征子集之间的相似度:

$S_{ij}=||x_i-x_j||$;

其中,S_{ij}表示第i个特征子集x_i与第j个特征子集x_j之间的相似度,$||x_i-x_j||$表示第i个特征子集x_i与第j个特征子集x_j之间的欧式距离。

4.根据权利要求1所述的基于邻域搜索策略的特征选择方法,其特征在于,采用如下的公式计算所述特征子集的适应度数值:

$$f(x_i)=\frac{TP+TN}{TP+TN+FP+FN};$$

其中,$f(x_i)$表示第i个特征子集x_i的适应度数值,TP表示正确分类的正样本数目,FP表示错误分类的正样本数目,TN表示正确分类的负样本数目,FN表示错误分类的负样本数目。

5.根据权利要求1所述的基于邻域搜索策略的特征选择方法,其特征在于,基于对应的局域导引对每个特征子集的位置进行更新,包括:

$X_i^{t+1}=X_i^t+V_i^{t+1}$,且:

$$V_i^{t+1}=w\times V_i^t+c_1\times r_1\times(pbest_i-X_i^t)+c_2\times r_2\times(lbest_i^t-X_i^t);$$

其中,X_i^{t+1}表示执行第t次迭代得到的第i个特征子集X_i^t进行更新的新特征子集,V_i^t表示执行第t-1次迭代得到的第i个特征子集X_i^t的位置,w表示预设的惯性权重,$pbest_i$表示所记录的第i个特征子集X_i^t的全局最优,$lbest_i^t$表示第i个特征子集X_i^t的局域导引,c_1和c_2是分别表示个体认知和社会认知权重,r_1 r_2分别表示[0,1]之间的随机数。

6.根据权利要求2至5任一项所述的基于邻域搜索策略的特征选择方法,其特征在于,

在基于对应的局域导引对每个特征子集的位置进行更新之后，还包括：

按照预设的概率对更新后的新特征子集执行变异操作，得到变异后的新特征子集，并采用变异后的新特征子集替换所述更新后的新特征子集。

7.根据权利要求6所述的基于邻域搜索策略的特征选择方法，其特征在于，所述对更新后的新特征子集执行变异操作，包括：

$$x_{new,d}=\begin{cases}x_{r1,d}+F\bullet(x_{r2,d}-x_{r3,d}),若rand()<MR\\x_{id},\qquad\qquad\qquad\qquad\qquad\qquad 其它;\end{cases}$$

其中，$x_{new,d}$表示更新后的新特征子集变异后的位置，F表示缩放因子，MR表示变异概率，x_{id}表示更新后的新特征子集，$x_{r1,d}$、$x_{r2,d}$和$x_{r3,d}$表示种群中三个随机选取的特征子集。

8.一种计算机可读存储介质，其上存储有计算机指令，其特征在于，所述计算机指令运行时执行权利要求1至7任一项所述的基于邻域搜索策略的特征选择方法的步骤。

9.一种终端，其特征在于，包括存储器和处理器，所述存储器上储存有能够在所述处理器上运行的计算机指令，所述处理器运行所述计算机指令时执行权利要求1至7任一项所述的基于邻域搜索策略的特征选择方法的步骤。

基于邻域搜索策略的特征选择方法、存储介质和终端

技术领域

[0001] 本发明属于计算机技术领域，特别是涉及一种基于邻域搜索策略的特征选择方法、存储介质和终端。

背景技术

[0002] 在机器学习和数据挖掘问题中，经常遇到高维数据集。很显然并非所有特征在建模时都是有用的，其中存在一些无关或冗余特征。在高维数据集上建模会带来高昂的计算成本，同时会降低预测模型的准确性。特征选择的目的是从所有特征中辨别并选出那些有价值的特征，并利用这些特征来建立预测模型。特征选择可以有效地降低建模时间，提升模型准确率以及提供更好的模型解读能力。

[0003] 特征选择是一个非常复杂的组合优化问题。在高维度数据集中，搜索空间规模很大，且特征之间存在复杂的相互关系。假设一个数据集中含有n个特征，那么可能的特征子集就有2^n个。传统的穷举方法显然无法应用于高维数据集。进化计算技术是一种具备很强的全局搜索能力的优化算法，很适合用于高维空间的优化问题。其中，微粒群算法因其具备操作简单、收敛速度快等特点，已经被应用于从高维数据集中筛选出优秀的特征子集。

[0004] 但是现有的基于微粒群算法的特征选择方法，存在着所选取的特征子集准确性差的问题。

发明内容

[0005] 本发明解决的技术问题是如何提高所选取的特征子集的准确性。

[0006] 为了达到上述目的，本发明提供一种基于邻域搜索策略的特征选择方法，所述方法包括：

[0007] 获取特征搜索空间；所述特征搜索空间包括多个特征子集；

[0008] 采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集。

[0009] 可选地，所述采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集，包括：

[0010] 初始化所述特征搜索空间中的特征子集；

[0011] 计算所述特征子集之间的相似度矩阵；

[0012] 基于计算得到的相似度矩阵，找到每个特征子集最相似的预设数量个邻近特征子集；

[0013] 从所述预设数量个邻近特征子集中找出适应度数值最大的特征子集，分别作为每个特征子集对应的局域导引；

[0014] 基于对应的局域导引对每个特征子集的位置进行更新，得到每个特征子集更新后的新特征子集；

[0015] 当确定新特征子集的适应度数值大于对应的特征子集的适应度数值时，采用新特

征子集代替对应的特征子集；

[0016] 从所述计算所述特征子集之间的相似度矩阵开始执行下一次迭代，直至迭代次数达到预设的次数阈值，得到全局最优特征子集。

[0017] 可选地，采用如下的公式计算所述相似度矩阵中的特征子集之间的相似度：

[0018] $S_{ij}=||x_i-x_j||$；

[0019] 其中，Si_j表示第i个特征子集x_i与第j个特征子集x_j之间的相似度，$||x_i-x_j||$表示第i个特征子集x_i与第j个特征子集x_j之间的欧式距离。

[0020] 可选地，采用如下的公式计算所述特征子集的适应度数值：

[0021] $$f(x_i)=\frac{TP+TN}{TP+TN+FP+FN}$$；

[0022] 其中，$f(x_i)$表示第i个特征子集x_i的适应度数值，TP表示正确分类的正样本数目，FP表示错误分类的正样本数目，TN表示正确分类的负样本数目，FN表示错误分类的负样本数目。

[0023] 可选地，基于对应的局域导引对每个特征子集的位置进行更新，包括：

[0024] $X_i^{t+1}=X_i^t+V_i^{t+1}$，且：

[0025] $$V_i^{t+1}=w\times V_i^t+c_1\times r_1\times(pbest_i-X_i^t)+c_2\times r_2\times(lbest_i^t-X_i^t)$$；

[0026] 其中，X_i^{t+1}表示执行第t次迭代得到的第i个特征子集X_i^t进行更新的新特征子集，V_i^t表示执行第t-1次迭代得到的第i个特征子集X_i^t的位置，w表示预设的惯性权重，$pbest_i$表示所记录的第i个特征子集X_i^t的全局最优，$lbest_i^t$表示第i个特征子集X_i^t的局域导引，c_1和c_2是分别表示个体认知和社会认知权重，r_1r_2分别表示[0,1]之间的随机数。

[0027] 可选地，在基于对应的局域导引对每个特征子集的位置进行更新之后，所述方法还包括：

[0028] 按照预设的概率对更新后的新特征子集执行变异操作，得到变异后的新特征子集，并采用变异后的新特征子集替换所述更新后的新特征子集。

[0029] 可选地，所述对更新后的新特征子集执行变异操作，包括：

[0030] $$x_{new,d}=\begin{cases}x_{r1,d}+F\bullet(x_{r2,d}-x_{r3,d}),若rand()<MR\\x_{id},\quad 其它\end{cases}$$；

[0031] 其中，$x_{new,d}$表示更新后的新特征子集变异后的位置，F表示缩放因子，MR表示变异概率，x_{id}表示更新后的新特征子集，$x_{r1,d}$、$x_{r2,d}$和$x_{r3,d}$表示种群中三个随机选取的特征子集。

[0032] 本发明实施例还提供了一种计算机可读存储介质，其上存储有计算机指令，所述计算机指令运行时执行上述任一项所述的基于邻域搜索策略的特征选择方法的步骤。

[0033] 本发明实施例还提供了一种终端，包括存储器和处理器，所述存储器上储存有能够在所述处理器上运行的计算机指令，所述处理器运行所述计算机指令时执行上述任一项所述的基于邻域搜索策略的特征选择方法的步骤。

[0034] 与现有技术相比，本发明的有益效果为：

[0035] 上述的方案，通过获取包括多个特征子集的特征搜索空间，并采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集，可以

从多个特征子集中筛选出更佳的特征子集，提高所选取的特征的准确性，从而可以提高模型构建的准确性。

[0036] 进一步地，通过采用邻域搜索策略，每个个体在它的邻域范围内选择最优个体作为其学习对象，可以使得每个粒子在它的邻域范围内搜寻高质量的特征子集，因此可以搜索特征空间内更多区域，提高搜索的准确性。

[0037] 进一步地，按照预设的概率对更新后的新特征子集执行变异操作，可以提升微粒群算法的全局搜索能力，丰富算法的搜索行为，提升搜寻特征子集的准确性。

附图说明

[0038] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0039] 图1是本发明实施例的一种基于邻域搜索策略的特征选择方法的流程示意图；

[0040] 图2是本发明实施例的另一种基于邻域搜索策略的特征选择方法的流程示意图；

[0041] 图3是本发明实施例的一种基于邻域搜索策略的特征选择装置的结构示意图。

具体实施方式

[0042] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。本发明实施例中有关方向性指示(诸如上、下、左、右、前、后等)仅用于解释在某一特定姿态(如附图所示)下各部件之间的相对位置关系、运动情况等，如果该特定姿态发生改变时，则该方向性指示也相应地随之改变。

[0043] 如背景技术所述，现有技术中的基于微粒群算法的特征选择方法，每个个体通过学习它自身的个体最优(pbest)和整个种群内的全局最优(gbest)来更新自身的位置。在高维特征选择问题中，这种搜索策略可能会导致算法快速收敛到一个局部最优特征子集，无法找到真正的最佳特征子集。同时，微粒群算法在全局搜索能力方面逊色与其他一些常见的进化算法，这会导致它无法有效地遍历整个特征空间，尤其是在高维度的特征选择问题中。

[0044] 本发明的技术方案通过获取包括多个特征子集的特征搜索空间，并采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集，可以从多个特征子集中筛选出更佳的特征子集，提高所选取的特征的准确性，从而可以提高模型构建的准确性。

[0045] 为使本发明的上述目的、特征和有益效果能够更为明显易懂，下面结合附图对本发明的具体实施例做详细的说明。

[0046] 图1是本发明实施例的一种基于邻域搜索策略的特征选择方法的流程示意图。参见图1，一种基于邻域搜索策略的特征选择方法，具体可以包括如下的步骤：

[0047] 步骤S101：获取特征搜索空间。

[0048] 在具体实施中，所述特征搜索空间包括多个特征子集。

[0049] 步骤S102：采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集。

[0050] 在具体实施中，通过基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集，每个个体在它的邻域范围内选择最优个体作为其学习对象，可以使得每个粒子在它的邻域范围内搜寻高质量的特征子集，因此可以搜索特征空间内更多区域，提高搜索的准确性，从而可以提高模型构建的准确性，具体请参见图2。

[0051] 上述的方案，通过获取包括多个特征子集的特征搜索空间，并采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集，可以从多个特征子集中筛选出更佳的特征子集，提高所选取的特征的准确性。

[0052] 下面将结合图2对本发明实施例中的基于邻域搜索策略的特征选择方法做进一步详细的说明。

[0053] 步骤S201：初始化所述特征搜索空间中的特征子集。

[0054] 在具体实施中，初始化所述特征搜索空间中的特征子集，即执行种群初始化操作，其中通过这一步骤，种群中的每一个粒子即对应于一个候选的特征子集。对于种群中的每个粒子i，其位置可以表示为：$X_i=\{x_{i1},x_{i2},\cdots,x_{iD}\}$，其中，D代表问题的搜索空间维度，即总的候选特征的数目。粒子的位置采用实数编码，全都是[0,1]范围内的实数。

[0055] 种群中粒子的初始位置随机生成，并采用下面的公式将位置对应于特征子集：

[0056] $$A_{id}=\begin{cases}1, 若x_{id}>0.5\\0, 其它\end{cases} \tag{1}$$

[0057] 其中，$A_{id}=1$代表第d个特征被选中，否则，没有选中。通过上述的解码方式，可以将微粒群算法应用于特征选择问题。

[0058] 接着，对初始生成的候选特征子集分别进行适应度评价。在本发明一实施例中，采用K近邻分类模型(KNN)对计算每个候选特征子集的分类准确率。其中，K的数值可以根据实际的需要设置，如设为5等。每个特征子集的适应度数值采用如下的公式计算：

[0059] $$f(x_i)=\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

[0060] 其中，$f(x_i)$表示第i个特征子集x_i的适应度数值，TP表示正确分类的正样本数目，FP表示错误分类的正样本数目，TN表示正确分类的负样本数目，FN表示错误分类的负样本数目。

[0061] 采用上述的公式(2)计算得到的适应度数值越大，分类准确性越高，代表特征子集的质量越高。

[0062] 步骤S202：计算所述特征子集之间的相似度矩阵。

[0063] 在本发明一实施例中，假设种群中含有n个粒子，则特征子集之间的相似度矩阵为n×n的相似度矩阵。其中，矩阵中的每一个元素、采用如下的公式计算：

[0064] $S_{ij}=||x_i-x_j||$ (3)

[0065] 其中，S_{ij}表示第i个特征子集x_i与第j个特征子集x_j之间的相似度，$||x_i-x_j||$表示第i个特征子集x_i与第j个特征子集x_j之间的欧式距离。

[0066] 步骤S203:基于计算得到的相似度矩阵,找到每个特征子集最相似的预设数量个邻近特征子集。

[0067] 在具体实施中,当计算得到特征子集之间的相似度矩阵时,可以基于候选特征子集之间的相似度,找到每个候选特征子集的领域,即从候选特征子集之间的相似度中找出相似度数值最大的数个特征子集,作为每个特征子集最相似的预设数量个邻近特征子集。

[0068] 步骤S204:从所述预设数量个邻近特征子集中找出适应度数值最大的特征子集,分别作为每个特征子集对应的局域导引。

[0069] 在具体实施中,当每个特征子集最相似的预设数量个邻近特征子集时,计算每个邻近特征子集的适应度数值,并比较每个邻近特征子集的适应度数值,找到其中最大的一个适应度数值对应的邻近特征子集,即为该候选特征子集对应的局域导引。其中,每个邻近特征子集的适应度数值可以采用上述的公式(2)进行计算。

[0070] 步骤S205:基于对应的局域导引对每个特征子集的位置进行更新,得到每个特征子集更新后的新特征子集。

[0071] 在具体实施中,当得到每个特征子集对应的局域导引时,根据对应的局域导引对每个特征子集的位置进行更新。在本发明一实施例中,采用如下的公式基于对应的局域导引对每个特征子集的速度和位置进行更新:

[0072] $$V_i^{t+1} = w \times V_i^t + c_1 \times r_1 \times (pbest_i - X_i^t) + c_2 \times r_2 \times (lbest_i^t - X_i^t) \tag{4}$$

[0073] $$X_i^{t+1} = X_i^t + V_i^{t+1} \tag{5}$$

[0074] 其中,X_i^{t+1}表示执行第t次迭代得到的第i个特征子集X_i^t进行更新的新特征子集,V_i^t表示执行第t-1次迭代得到的第i个特征子集X_i^t的位置,w表示预设的惯性权重,$pbest_i$表示所记录的第i个特征子集X_i^t的全局最优,$lbest_i^t$表示第i个特征子集X_i^t的局域导引,c_1和c_2是分别表示个体认知和社会认知权重,r_1、r_2分别表示[0,1]之间的随机数。

[0075] 通过上述公式(4)和(5),可以计算得到每个候选特征子集的新位置,即更新后的新特征子集。通过采用这个邻域搜索策略,每个候选特征子集可以在其邻域范围内得到有价值的信息,并用于更新自身的位置,避免了种群多样性的快速流失。

[0076] 在本发明一实施例中,为了跳出局部最优,在更新位置之后,还包括:

[0077] 步骤S206:按照预设的概率对更新后的新特征子集执行变异操作,得到变异后的新特征子集,并采用变异后的新特征子集替换所述更新后的新特征子集。

[0078] 在本发明一实施例中,采用如下的公式对更新后的新特征子集执行变异操作:

[0079] $$x_{new,d} = \begin{cases} x_{r1,d} + F \bullet (x_{r2,d} - x_{r3,d}), & \text{若} rand() < MR \\ x_{id}, & \text{其它} \end{cases} \tag{6}$$

[0080] 其中,$x_{new,d}$表示更新后的新特征子集变异后的位置,F表示缩放因子,MR表示变异概率,x_{id}表示更新后的新特征子集,$x_{r1,d}$、$x_{r2,d}$和$x_{r3,d}$表示种群中三个随机选取的特征子集。

[0081] 如果有些粒子陷入了停滞状态,通过按照预设的概率对更新后的新特征子集执行变异操作,可以保住某些陷入停滞状态的粒子跳出局部最优,可以为算法带来更多的随机性,找到更好的特征子集。同时,该操作可以丰富算法的搜索行为,而且不会带来额外的算法评价次数。

[0082] 步骤S207:判断更新后的新特征子集的适应度数值是否大于对应的特征子集的适应度数值;当判断结果为是时,可以执行步骤S208;反之,则可以直接执行步骤S209。

[0083] 这里需要指出的是,该步骤中更新后的新特征子集可以为执行步骤S205得到的新的特征子集,也可以为执行步骤S206对更新后的新特征子集执行变异操作得到的新特征子集。

[0084] 骤S208:采用新特征子集代替对应的特征子集。

[0085] 在具体实施中,当得到新的特征子集后,重新评价新找到的特征子集,并当确定新特征子集的适应度数值大于对应的特征子集的适应度数值时,采用新特征子集代替之前的特征子集,即更新该粒子的个体最优。

[0086] 骤S209:判断迭代次数是否达到预设的次数阈值;当判断结果为是时,可以执行步骤S211;反之,则可以执行步骤S210。

[0087] 在具体实施中,所述预设的次数阈值可以根据实际的需要进行设置,在此不做限制。

[0088] 步骤S210:执行下一次迭代。

[0089] 在具体实施中,当迭代次数未达到预设的次数阈值时,可以接着执行下一次迭代,即从步骤S202重新开始执行,直至迭代次数达到所述次数阈值时止。

[0090] 步骤S211:输出全局最优特征子集。

[0091] 在具体实施中,所述全局最优特征子集为达到预设的次数阈值时,对所得到的每个位置上的特征子集的适应度数值进行计算,并将其中适应度数值最大的特征子集作为所述全局最优特征子集进行输出。

[0092] 为了验证本发明提出的特征选择算法的效果,选取了三个数据集来进行验证,和另外三种基于进化算法的特征选择模型进行对比。三个数据集分别为wine(包含13个特征,178个样本),ionosphere(包含34个特征,351个样本),musk1(包含166个特征,476个样本)。三个对比算法分别为:遗传算法(GA),微粒群算法(PSO),骨干粒子群算法(BBPSO)。每个数据集70%的样本用于训练,30%的样本用于测试。本发明提出的算法PSO-NS,种群数目设置为20,迭代次数为50次,c1和c2均为2,惯性权重w采用时变的惯性权重,初始值为0.9,逐步降低为0.4。其余对比算法的种群数目和迭代次数与PSO-NS一致。遗传算法的交叉概率设为0.8,变异概率为0.1。

[0093] 每种算法在每个数据集上都随机运行20次。表1中给出了实验结果,包括:特征子集的平均分类准确率和方差,选出的特征数目。

[0094] 表1

[0095]

数据集	指标	GA	PSO	BBPSO	PSO-NS
wine	准确率	96.29	97.04	97.04	98.87
	方差	2.33	1.66	1.29	1.44
	特征数目	6.35	8.6	7.8	7.44
ionosphere	准确率	83.62	84.91	84.72	87.32
	方差	3.12	2.04	1.64	1.91

[0096]

	特征数目	10.93	10.7	9.74	10.43
musk1	准确率	82.81	83.92	84.42	85.85
	方差	2.36	2.08	2.13	1.47
	特征数目	81.83	82.14	80.43	82.96

[0097] 从表1的结果中可以看出，本发明提出的特征选择算法在三个测试集中都获得了更高的分类准确率，方差也更小。这表明本发明中提出的邻域搜索策略和变异算子，提升了微粒群算法的种群多样性和全局搜索能力，因此使得算法能找到质量更高的特征子集。

[0098] 上述对本发明实施例中的基于邻域搜索策略的特征选择方法进行了描述，下面将对上述的方法对应的装置进行介绍。

[0099] 图3示出了本发明实施例中的一种基于邻域搜索策略微粒群算法的特征选择装置的结构示意图。参见图3，一种基于邻域搜索策略微粒群算法的特征选择装置30可以包括获取单元301和选取单元302，其中：

[0100] 所述获取单元301，适于获取特征搜索空间；所述特征搜索空间包括多个特征子集；

[0101] 所述选取单元302，适于采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集。

[0102] 在具体实施中，所述选取单元302，适于初始化所述特征搜索空间中的特征子集；计算所述特征子集之间的相似度矩阵；基于计算得到的相似度矩阵，找到每个特征子集最相似的预设数量个邻近特征子集；从所述预设数量个邻近特征子集中找出适应度数值最大的特征子集，分别作为每个特征子集对应的局域导引；基于对应的局域导引对每个特征子集的位置进行更新，得到每个特征子集更新后的新特征子集；当确定新特征子集的适应度数值大于对应的特征子集的适应度数值时，采用新特征子集代替对应的特征子集；从所述计算所述特征子集之间的相似度矩阵开始执行下一次迭代，直至迭代次数达到预设的次数阈值，得到全局最优特征子集。

[0103] 在本发明一实施例中，所述选取单元302，适于采用如下的公式计算所述相似度矩

阵中的特征子集之间的相似度：

[0104] $S_{ij}=||x_i-x_j||$；

[0105] 其中，S_{ij}表示第i个特征子集x_i与第j个特征子集x_j之间的相似度，$||x_i-x_j||$表示第i个特征子集x_i与第j个特征子集x_j之间的欧式距离。

[0106] 在本发明另一实施例中，所述选取单元302，适于采用如下的公式计算所述特征子集的适应度数值：

[0107] $$f(x_i)=\frac{TP+TN}{TP+TN+FP+FN}$$；

[0108] 其中，$f(x_i)$表示第i个特征子集x_i的适应度数值，TP表示正确分类的正样本数目，FP表示错误分类的正样本数目，TN表示正确分类的负样本数目，FN表示错误分类的负样本数目。

[0109] 在本发明又一实施例中，所述选取单元302，适于基于对应的局域导引对每个特征子集的位置进行更新，包括：

[0110] $X_i^{t+1}=X_i^t+V_i^{t+1}$，且：

[0111] $$V_i^{t+1}=w\times V_i^t+c_1\times r_1\times(pbest_i-X_i^t)+c_2\times r_2\times(lbest_i^t-X_i^t)$$；

[0112] 其中，X_i^{t+1}表示执行第t次迭代得到的第i个特征子集X_i^t进行更新的新特征子集，V_i^t表示执行第t-1次迭代得到的第i个特征子集X_i^t的位置，w表示预设的惯性权重，$pbest_i$表示，$lbest_i^t$表示第i个特征子集X_i^t的局域导引，c_1和c_2是分别表示个体认知和社会认知权重，r_1 r_2分别表示[0,1]之间的随机数。

[0113] 在具体实施例中，所述选取单元302，还适于在基于对应的局域导引对每个特征子集的位置进行更新之后，按照预设的概率对更新后的新特征子集执行变异操作，得到变异后的新特征子集，并采用变异后的新特征子集替换所述更新后的新特征子集。

[0114] 在本发明一实施例中，所述选取单元302，适于采用如下的公式所述对更新后的新特征子集执行变异操作：

[0115] $$x_{new,d}=\begin{cases}x_{r1,d}+F\bullet(x_{r2,d}-x_{r3,d}), & 若rand()<MR\\ x_{id}, & 其它\end{cases}$$；

[0116] 其中，$x_{new,d}$表示更新后的新特征子集变异后的位置，F表示缩放因子，MR表示变异概率，x_{id}表示更新后的新特征子集，$x_{r1,d}$、$x_{r2,d}$和$x_{r3,d}$表示种群中三个随机选取的特征子集。

[0117] 本发明实施例还提供了一种计算机可读存储介质，其上存储有计算机指令，所述计算机指令运行时执行所述的基于邻域搜索策略的特征选择方法的步骤。其中，所述基于邻域搜索策略的特征选择方法请参见前述部分的介绍，不再赘述。

[0118] 本发明实施例还提供了一种终端，包括存储器和处理器，所述存储器上储存有能够在所述处理器上运行的计算机指令，所述处理器运行所述计算机指令时执行所述的基于邻域搜索策略的特征选择方法的步骤。其中，所述基于邻域搜索策略的特征选择方法请参见前述部分的介绍，不再赘述。

[0119] 采用本发明实施例中的上述方案，通过获取包括多个特征子集的特征搜索空间，

并采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集，可以从多个特征子集中筛选出最相关最有价值的特征，故可以提高所选取的特征的准确性。

[0120] 进一步地，通过采用邻域搜索策略，每个个体在它的邻域范围内选择最优个体作为其学习对象，可以使得每个粒子在它的邻域范围内搜寻高质量的特征子集，因此可以搜索特征空间内更多区域，提高搜索的准确性。

[0121] 进一步地，按照预设的概率对更新后的新特征子集执行变异操作，可以提升微粒群算法的全局搜索能力，丰富算法的搜索行为，提升搜寻特征子集的准确性。

[0122] 以上显示和描述了本发明的基本原理、主要特征和本发明的优点。本行业的技术人员应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的只是说明本发明的原理，在不脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，本发明要求保护范围由所附的权利要求书、说明书及其等效物界定。

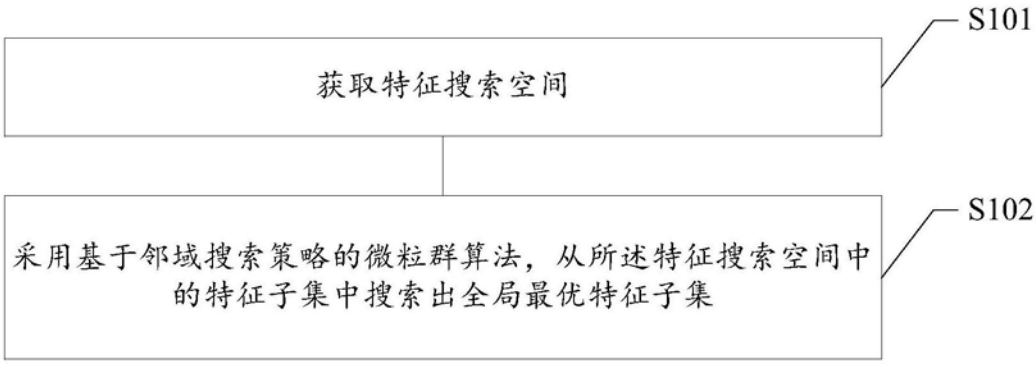
S101
获取特征搜索空间
S102
采用基于邻域搜索策略的微粒群算法，从所述特征搜索空间中的特征子集中搜索出全局最优特征子集

图1

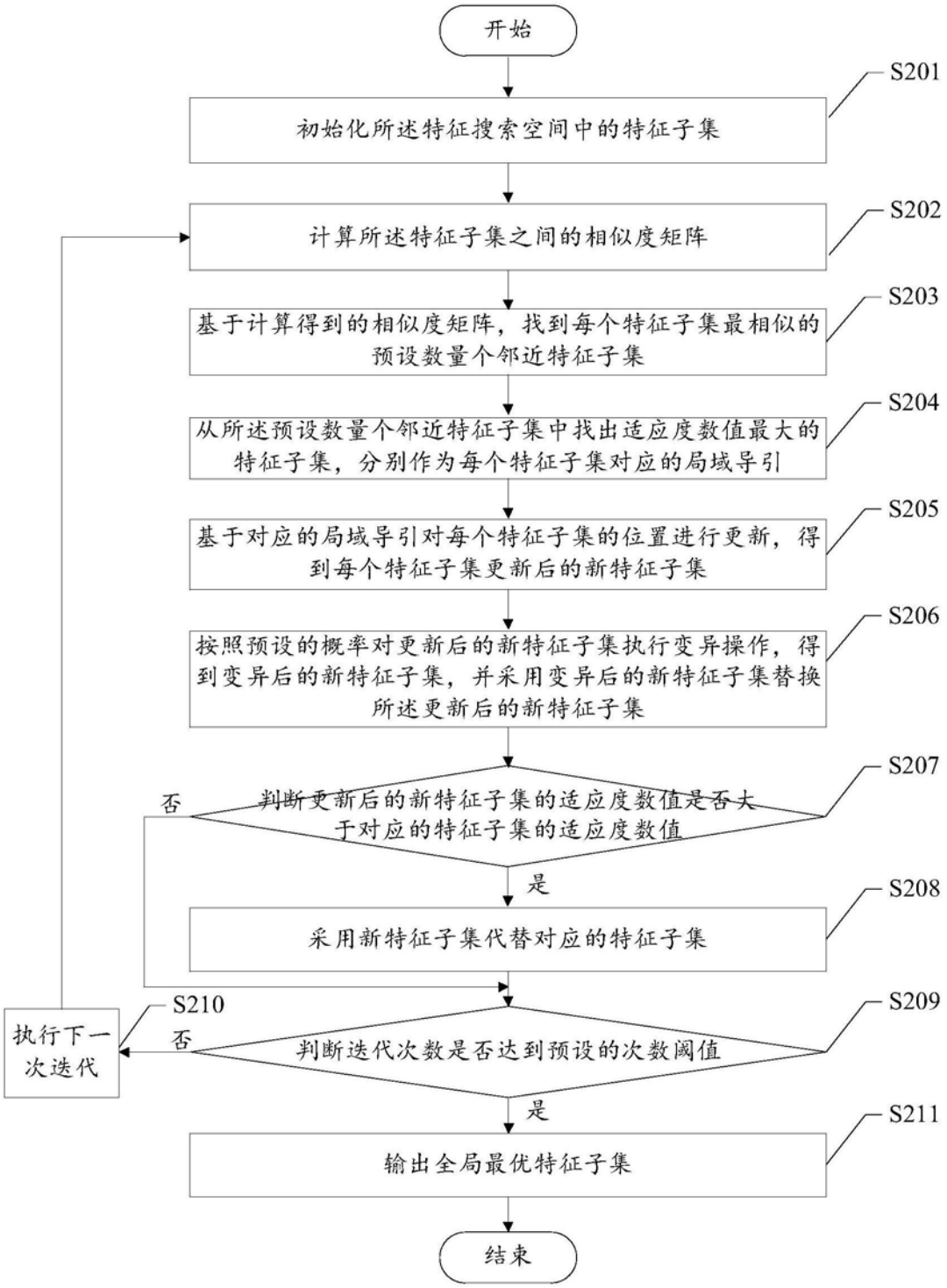
开始
S201
初始化所述特征搜索空间中的特征子集
S202
计算所述特征子集之间的相似度矩阵
S203
基于计算得到的相似度矩阵，找到每个特征子集最相似的预设数量个邻近特征子集
S204
从所述预设数量个邻近特征子集中找出适应度数值最大的特征子集，分别作为每个特征子集对应的局域导引
S205
基于对应的局域导引对每个特征子集的位置进行更新，得到每个特征子集更新后的新特征子集
S206
按照预设的概率对更新后的新特征子集执行变异操作，得到变异后的新特征子集，并采用变异后的新特征子集替换所述更新后的新特征子集
S207
判断更新后的新特征子集的适应度数值是否大于对应的特征子集的适应度数值
否
是
S208
采用新特征子集代替对应的特征子集
S209
判断迭代次数是否达到预设的次数阈值
否
S210
执行下一次迭代
是
S211
输出全局最优特征子集
结束

图2

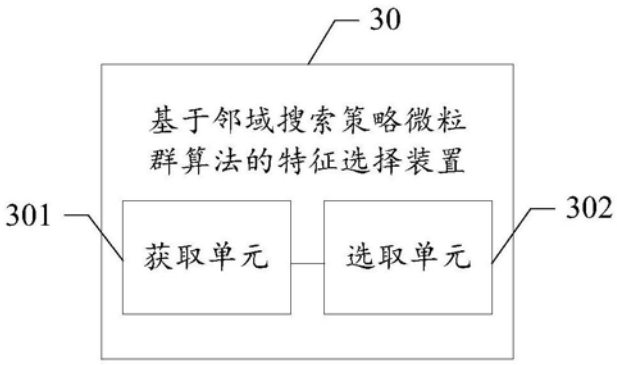

图3