(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 109034392 A
(43)申请公布日 2018.12.18

(21)申请号 201811146505.3

(22)申请日 2018.09.29

(71)申请人 广西壮族自治区水产科学研究院
地址 530021 广西壮族自治区南宁市青山路8号

(72)发明人 肖俊 罗永巨 郭忠宝 杨弘 于凡 钟欢 周毅 梁军能 唐瞻杨 严欣 雷燕

(74)专利代理机构 重庆市信立达专利代理事务所(普通合伙) 50230
代理人 包晓静

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/06(2006.01)
G06Q 50/02(2012.01)
G06F 17/30(2006.01)
A01K 61/10(2017.01)

权利要求书5页 说明书12页 附图1页

(54)发明名称

一种罗非鱼杂交配套系的选育方法及系统

(57)摘要

本发明属于信息处理技术领域,公开了一种罗非鱼杂交配套系的选育方法及系统,包括:收集罗非鱼杂交配套系的相关数据;处理搜集的相关数据,进行统计分析,建立数据挖掘平台;搭建人机交互系统。本发明通过对采集的罗非鱼的相关信息进行数据分析,建立人机交互系统,推算出罗非鱼杂交配套系的可靠的选育方法,有效的节约了数据信息采集分析的时间,提高了罗非鱼杂交配套系的选育方法的工作效率,保证了选育的优良性与可靠性。

CN 109034392 A

1.一种罗非鱼杂交配套系的选育方法，其特征在于，所述罗非鱼杂交配套系的选育方法包括：

处理搜集的相关数据，进行统计分析，建立数据挖掘平台；处理搜集的相关数据采用神经网络训练算法，具体包括：

1)在前向阶段，输入层获取到输入信号并将其传递到隐藏层中的每个神经元；然后，隐藏层处理这些信号并将处理结果传递到输出层；对于一个输入向量义X＝(X1,X2,...Xm)，隐藏层中每个神经元的输入和输出信号标记为uj和hj，这两个信号分别可W通过公式计算；

$$u_j = \sum_{i=1}^{m} W_{ij} x_i + \theta_j \ (j = 1,2,\ldots) \quad ;$$

$$h_j = f(u_j) = \frac{1}{1 + \exp(u_j)} \ (j = 1,2,q) \quad ;$$

其中Wij是输入层神经元i和隐藏层神经元j之间的权重，θj是偏置；

输出层从隐藏层获取到信号之后同样需要进行后续处理；输出层神经元的输入信号lk和输出信号ck分别由公式计算得出；

$$l_k = \sum_{j=1}^{q} V_{jk} h_j + \gamma_k \ (k = 1,2,\ldots n) \quad ;$$

$$c_k = f(l_k) = \frac{1}{1 + \exp(-l_k)} \ (k = 1,2,n) \quad ;$$

其中Vij是输入层神经元j和隐藏层神经元jk之间的权重，γk是偏置；在前向过程中，神经网络模型权重W,V和偏置θ，γ并不发生变化；如果前向处理得出的神经网络最终输出信号与真实信号一致，那么下一个输入向量将被输入到该神经网络并开始新一轮的前向过程；否则，该算法将进入后向过程；这里，将神经网络的最终输出信号和真实信号之间的差值称为偏差；

(2)后向阶段

在后向过程，首先将采用公式计算出每个输出层神经元的偏差，然后进一步地利用公式计算出每个隐藏层神经元ei的偏差；

dk＝(yk-ck)ck(1-ck)(k＝1,2,)；

$$e_j = (\sum_{k=1}^{n} d_k V_j) h_j (1 - h_j) \ (j = 1,2,) \quad ;$$

偏差从输出层反向回馈到隐藏层；通过送种偏差后向传播方式，利用公式更新输出层和隐藏层的连接权重；利用公式更新隐藏层与输入层之间的连接权重；

$$\begin{aligned} V_{jk}(N+1) &= V_{jk}(N) + a_1 d_k(N) h_j \\ \gamma_k(N+1) &= \gamma_k(N) + a_1 d_k(N) \end{aligned} \quad ;$$

$$\begin{aligned} W_{ij}(N+1) &= W_{ij}(N) + a_2 e_j(N) \\ \theta_j(N+1) &= \theta_j(N) + a_2 d_j(N) \end{aligned} \quad ;$$

i＝1,2,…m；j＝1,2,…q；k＝1,2,..n；a1和a2是取值范围在0到1的学习率；N表示当前

训练轮数的编号。

2.如权利要求1所述的罗非鱼杂交配套系的选育方法,其特征在于,所述罗非鱼杂交配套系的选育方法,具体包括:

通过数据集成模块,将互相关联的分布式异构数据源集成到一起,使用户以透明的方式访问罗非鱼杂交配套系的相关数据源;

通过数据管理模块利用计算机硬件和软件技术对罗非鱼杂交配套系的相关数据进行有效的收集、存储、处理和应用;实现数据有效管理;具体有:收集罗非鱼杂交配套系的相关数据处理搜集的相关数据,进行统计分析,建立数据挖掘平台;搭建人机交互系统;

通过数据储存模块将罗非鱼杂交配套系的相关数据储存在计算机中;

通过数据挖掘模通过对大量的罗非鱼杂交配套系的相关数据进行分析,提取隐含的信息和知识。

3.如权利要求2所述的罗非鱼杂交配套系的选育方法,其特征在于,

收集罗非鱼杂交配套系的相关数据中,相关数据包括:

(1)罗非鱼的种类、数量、生存年份、形态标准、生存环境类数据信息;

(2)选育方案;

(3)基础群组建;

(4)关于专门化品系的选育法;

(5)数个杂交组合的比较试验,筛选组合。

4.如权利要求2所述的罗非鱼杂交配套系的选育方法,其特征在于,数据挖掘模块中基于关联规则映射的罗非鱼生物信息多维数据挖掘算法,进行分析、提取隐含的信息和知识;具体包括:

假设子空间的维度为d,先挖掘处于不同子空间的不同数据集,子空间用矩阵M表示,为:

$$M=\begin{pmatrix} M^1 \\ \vdots \\ M^d \end{pmatrix} \quad d\leqslant n \tag{1}$$

假设两个数据集V_i和V_k分别位于两个不同的子空间M^i(i≤d)和M^k(k≤d),其中这两个子空间的欧几里德距离为D(i,k),两个数据集的欧几里德距离为d(i,k),则对于不同子空间的两个数据集的挖掘公式为:

$$W(M^i,M^k)=\frac{\sigma}{2}\begin{pmatrix} M^1 \\ \vdots \\ M^d \end{pmatrix} P(V_i)\,P(V_k)\times\log_2\sqrt{D(i,k)^2+d(i,k)^2} \tag{2}$$

其中:σ表示子空间挖掘因子,P(V_i)、P(V_k),分别表示数据集V_i和数据集V_k的挖掘频率;

对于同一子空间的不同数据集的挖掘,通过不同数据集之间的关联程度进行区分,先通过式

$$K_1\begin{pmatrix}\alpha_{ik}\\ \beta_{ik}\\ \theta_{ik}\end{pmatrix}=\begin{pmatrix}\alpha_{i1} & \cdots & \alpha_{1k}\\ \vdots & & \vdots\\ \alpha_{1k} & \cdots & \alpha_{i1}\end{pmatrix}\begin{pmatrix}\beta_{i1} & \cdots & \beta_{1k}\\ \vdots & & \vdots\\ \beta_{1k} & \cdots & \beta_{i1}\end{pmatrix}\begin{pmatrix}\theta_{i1} & \cdots & \theta_{1k}\\ \vdots & & \vdots\\ \theta_{1k} & \cdots & \theta_{i1}\end{pmatrix};$$

$$K_2\begin{pmatrix}\frac{1}{\alpha_{ik}}\\ \frac{1}{\beta_{ik}}\\ \frac{1}{\theta_{ik}}\end{pmatrix}=\begin{pmatrix}\frac{1}{\alpha_{i1}} & \cdots & \frac{1}{\alpha_{1k}}\\ \vdots & & \vdots\\ \frac{1}{\alpha_{1k}} & \cdots & \frac{1}{\alpha_{i1}}\end{pmatrix}\begin{pmatrix}\frac{1}{\beta_{i1}} & \cdots & \frac{1}{\beta_{1k}}\\ \vdots & & \vdots\\ \frac{1}{\beta_{1k}} & \cdots & \frac{1}{\beta_{i1}}\end{pmatrix}\begin{pmatrix}\frac{1}{\theta_{i1}} & \cdots & \frac{1}{\theta_{1k}}\\ \vdots & & \vdots\\ \frac{1}{\theta_{1k}} & \cdots & \frac{1}{\theta_{i1}}\end{pmatrix};$$

求得K_1和K_2然后求得在同一空间下数据集V_i和V_k的关联因子:

$$g(i,k)=\sqrt{\frac{K_1}{K_2}\begin{pmatrix}x_{1i} & \cdots & x_{mi}\\ \vdots & & \vdots\\ x_{mi} & \cdots & x_{1i}\end{pmatrix}d(i,k)}-\ln 2\frac{1}{m}\left(\frac{K_2}{K_1}\begin{pmatrix}x_{1k} & \cdots & x_{mk}\\ \vdots & & \vdots\\ x_{mk} & \cdots & x_{1k}\end{pmatrix}\right)\quad(3);$$

得到数据集V_i和V_k的关联因子g(i,k)之后,得到相同子空间下这两个数据集的挖掘公式为

$$W(V^i,V^k)=(P(V_i)-P(V_k))\,g(i,k)\,d(i,k)\times\frac{e^{g(i,k)}}{\begin{pmatrix}x_{1k} & \cdots & x_{mk}\\ \vdots & & \vdots\\ x_{mk} & \cdots & x_{1k}\end{pmatrix}\begin{pmatrix}x_{1i} & \cdots & x_{mi}\\ \vdots & & \vdots\\ x_{mi} & \cdots & x_{1i}\end{pmatrix}}\quad(4)$$

假设在同一空间M^i下数据集之间关联程度限定阈值T(V),当数据集之间的关联因子g(i,k)大于T(V)时,则这两个数据集具有强相关性,则两个数据集的区分公式写成

$$f_{M^i}(V_i)=\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)-\ln\left(\sqrt{\sum_{j=1}^{n}(P(V_i)-P(V_j))}\right)\quad(5)$$

$$f_{M^i}(V_k)=\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)-\ln\left(\sqrt{\sum_{j=1}^{n}(P(V_k)-P(V_j))}\right)\quad(6)$$

当数据集之间的关联因子g(i,k)小于T(V)时,则这两个数据集具有弱相关性,则两个数据集的区分公式写成

$$f_{M^i}(V_i)=\frac{\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)}{(\pi e^2-1)}+\frac{e}{2}\sum_{j=1}^{n}(P(V_i)-P(V_j))$$

$$f_{M^i}(V_k)=\frac{\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)}{(\pi e^2-1)}+\frac{e}{2}\sum_{j=1}^{n}(P(V_k)-P(V_j))$$

。

5.如权利要求2所述的罗非鱼杂交配套系的选育方法,其特征在于,数据储存模块中采用朴素贝叶斯分类算法将罗非鱼杂交配套系的相关数据储存在计算机中,具体包括:

设D是训练对象与其相关联的类标号的集合,每个对象用一个n维属性向量X＝{x_1,x_2…x_n}表示,描述n维属性向量X＝{x_1,x_2…x_n}表示,描述n个属性A_1,A_2…,A_n的值,假定原始集合基于n维属性共划分为m个类C1,C2…Cm,计算每个类对X的后验概率,并将对象X归属于具有最高后验概率的类,后验概率P(C_i|X)的计算公式为:

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

;

由于P(C_i|X)的计算开销较大,进行类条件独立的假定,给定向量的类标号,并假定属性值有条件的相互独立,P(X|C_i)的计算公式为:

$$P(X|C_i)=\prod_{k=1}^{n}P(x_k|C_i)=P(x_1|C_i)P(x_2|C_i)\cdots P(x_n|C_i)$$

;

其中,P(x_1|C_i)P(x_2|C_i)…P(x_n|C_i)容易地由训练对象求算,X_k表示X在属性A_k上的值,对每个类别C_i计算P(X|C_i)P(C_i),当P(X|C_i)P(C_i)＞P(X|C_j)P(C_j),1≤j≤m,j≠i成立时,X属于类C_i。

6.如权利要求2所述的罗非鱼杂交配套系的选育方法,其特征在于,数据集成模块中采用不完备混合数据的集成聚类算法将互相关联的分布式异构数据源集成到一起,使用户以透明的方式访问罗非鱼杂交配套系的相关数据源;具体包括:

输入:带有缺失值的数据集D、聚类个数k;

输出:最终聚类结果π*(D);

步骤一,对数据集D分别运用平均值填充法、KNN填充法、SKNN填充法填充得到完备数据集D_1,D_2,D_3;

步骤二,对D_i(1≤i≤3)分别执行M_i次K-Prototypes聚类算法,得到基聚类结果集Π(D);

步骤三,根据式

$$SM(x_p,x_q)=\frac{1}{\sum_{i=1}^{3}M_i}\sum_{i=1}^{3}\sum_{j=1}^{M_i}SM_i^j(x_p,x_q)$$

;

计算样本与样本之间的相似度矩阵$SM_{n\times n}$;

步骤四,基于相似度矩阵$SM_{n\times n}$,分别根据以下式:

单链(single link)方法.由2个类中相似度最大的2个样本决定

$$sim(C,C')=\max_{x\in C,x'\in C'}sim(x,x')$$

;

全链(complete link)方法,由2个类中相似度最小的2个样本决定

$$sim(C,C')=\min_{x\in C,x'\in C'}sim(x,x')$$

;

组平均(average link)方法,由2个类中所有样本点相似度的平均值决定

$$sim(C,C')=\frac{1}{|C||C'|}\sum_{x\in C}\sum_{x'\in C'}sim(x,x')$$

;

式中:样本之间的相似度sim(x,x′)为相似度矩阵$SM_{n\times n}$中的对应元素值;

运行层次聚类算法得到最终的聚类结果$\pi^*(D)$。

7.一种罗非鱼杂交配套系的选育计算机程序,其特征在于,所述罗非鱼杂交配套系的选育计算机程序实现权利要求1～6任意一项所述的罗非鱼杂交配套系的选育方法。

8.一种终端,其特征在于,所述终端至少搭载实现权利要求1～6任意一项所述的罗非鱼杂交配套系的选育方法的控制器。

9.一种计算机可读存储介质,包括指令,当其在计算机上运行时,使得计算机执行如权利要求1-6任意一项所述的罗非鱼杂交配套系的选育方法。

10.一种实施权利要求1所述罗非鱼杂交配套系的选育方法的罗非鱼杂交配套系的选育系统,其特征在于,所述罗非鱼杂交配套系的选育系统包括数据集成模块、数据储存模块、数据管理模块、数据挖掘模块;

数据集成模块,用于将互相关联的分布式异构数据源集成到一起,使用户以透明的方式访问罗非鱼杂交配套系的相关数据源;

数据管理模块,用于利用计算机硬件和软件技术对罗非鱼杂交配套系的相关数据进行有效的收集、存储、处理和应用;实现数据有效管理;

数据储存模块,用于将罗非鱼杂交配套系的相关数据储存在计算机中;

数据挖掘模块,用于通过对大量的罗非鱼杂交配套系的相关数据进行分析,提取隐含的信息和知识。

一种罗非鱼杂交配套系的选育方法及系统

技术领域

[0001] 本发明属于信息处理技术领域，尤其涉及一种罗非鱼杂交配套系的选育方法及系统。

背景技术

[0002] 目前，我国通过全国水产原种和良种审定委员会审定通过的罗非鱼品种有奥尼罗非鱼、吉富品系尼罗罗非鱼、新吉富罗非鱼、“夏奥1号”奥利亚罗非鱼等。但面对我国庞大的罗非鱼产业，选育良种还远不能满足产业快速发展的需求。如何充分发挥我国尼罗罗非鱼引进资源群体多、遗传背景丰富的特点，开展这些引进后养殖群体种质资源的综合利用，将是我国罗非鱼良种选育和产业化开发的重要方向之一。配套系育种是应用具有某种(些)经济性状的“专门化”品系进行杂交，生产具有显著“杂种优势”的配套组合的一种育种方式，在农作物、家禽和家畜育种中应用普遍，并取得很好的效果，在水产业，鱼类的配套系育种近年已有所开展。

[0003] 现代罗非鱼杂交配套选育技术体系复杂，需要多个学科交叉和多种技术支撑，缺乏有效数据组织和管理。

[0004] 我国选育技术相关数据量很大，但分散，未有效组织。目前育种者在育种过程中利用的数据主要为自身内部数据，而公开的文献和基因组相关数据等其他数据很少利用或无法利用。导致大量内部数据成为“数据孤岛”，同时大量公开的育种相关数据(如基因组数据)成为“数据海洋”，无从下手。上述问题极大地限制了育种相关数据的利用和育种效率的提高。

[0005] 大数据有5大特征，即所谓5V:数量巨大(volume)，类型多样(variety)，处理速度快(velocity)，价值密度低(value)，真实性(veracity)。在这5V中，数量巨大、类型多样指数据量大而形式多样，同时要求处理速度要快，而其中价值密度低则指的是数据信息存在垃圾多、污染重以及利用难的问题，然而就是在这样的低密度中却实实在在蕴涵着巨大的价值。可以说，大数据时代的到来将对研究方式、思维方式乃至于生活方式和生产方式都产生革命性变化。

[0006] 综上所述，现有技术存在的问题是：

[0007] (1)现代罗非鱼杂交配套选育技术体系复杂，需要多个学科交叉和多种技术支撑，缺乏有效数据组织和管理；且选育技术相关数据量很大，分散、未有效组织，极大地限制了相关数据的利用和育种效率的提高。

[0008] (2)在生物信息网络中对复杂和大规模的数据集进行挖掘时所出现的算法挖掘精度低、运行速度慢、内存占用大等问题

[0009] (3)实际应用中面临的数据往往是兼具数值属性和分类属性共同描述的混合型数据，而且通常带有缺失值。

[0010] (4)分类算法不能能个处理多分类任务，不适合增量式训练，算法复杂，不具有稳定的分类效率。

发明内容

[0011] 针对现有技术存在的问题，本发明提供了一种罗非鱼杂交配套系的选育方法及系统。

[0012] 本发明是这样实现的，一种罗非鱼杂交配套系的选育方法，包括：

[0013] 处理搜集的相关数据，进行统计分析，建立数据挖掘平台；处理搜集的相关数据采用神经网络训练算法，具体包括：

[0014] 1)在前向阶段，输入层获取到输入信号并将其传递到隐藏层中的每个神经元；然后，隐藏层处理这些信号并将处理结果传递到输出层；对于一个输入向量义$X=(X_1,X_2,...X_m)$，隐藏层中每个神经元的输入和输出信号标记为u_j和h_j，这两个信号分别可W通过公式计算；

[0015] $$u_j=\sum_{i=1}^{m}W_{ij}x_i+\theta_j(j=1,2,...)\quad;$$

[0016] $$h_j=f(u_j)=\frac{1}{1+\exp(u_j)}(j=1,2,q)\quad;$$

[0017] 其中W_{ij}是输入层神经元i和隐藏层神经元j之间的权重，θ_j是偏置；

[0018] 输出层从隐藏层获取到信号之后同样需要进行后续处理；输出层神经元的输入信号l_k和输出信号c_k分别由公式计算得出；

[0019] $$l_k=\sum_{j=1}^{q}V_{jk}h_j+\gamma_k(k=1,2,...n)\quad;$$

[0020] $$c_k=f(l_k)=\frac{1}{1+\exp(-l_k)}(k=1,2,n)\quad;$$

[0021] 其中V_{ij}是输入层神经元j和隐藏层神经元jk之间的权重，γ_k是偏置；在前向过程中，神经网络模型权重W,V和偏置θ，γ并不发生变化；如果前向处理得出的神经网络最终输出信号与真实信号一致，那么下一个输入向量将被输入到该神经网络并开始新一轮的前向过程；否则，该算法将进入后向过程；这里，将神经网络的最终输出信号和真实信号之间的差值称为偏差；

[0022] (2)后向阶段

[0023] 在后向过程，首先将采用公式计算出每个输出层神经元的偏差，然后进一步地利用公式计算出每个隐藏层神经元e_i的偏差；

[0024] $d_k=(y_k-c_k)c_k(1-c_k)(k=1,2,)$；

[0025] $$e_j=(\sum_{k=1}^{n}d_kV_j)h_j(1-h_j)(j=1,2,)\quad;$$

[0026] 偏差从输出层反向回馈到隐藏层；通过送种偏差后向传播方式，利用公式更新输出层和隐藏层的连接权重；利用公式更新隐藏层与输入层之间的连接权重；

[0027] $$\begin{aligned}V_{jk}(N+1)&=V_{jk}(N)+a_1d_k(N)h_j\\\gamma_k(N+1)&=\gamma_k(N)+a_1d_k(N)\end{aligned}\quad;$$

[0028] $$W_{ij}(N+1)=W_{ij}(N)+a_2 e_j(N)$$ $$\theta_j(N+1)=\theta_j(N)+a_2 d_j(N)$$ ；

[0029] i＝1,2,…m;j＝1,2,…q;k＝1,2,..n;a_1和a_2是取值范围在0到1的学习率;N表示当前训练轮数的编号。

[0030] 进一步,所述罗非鱼杂交配套系的选育方法,具体包括：

[0031] 通过数据集成模块,将互相关联的分布式异构数据源集成到一起,使用户以透明的方式访问罗非鱼杂交配套系的相关数据源；

[0032] 通过数据管理模块利用计算机硬件和软件技术对罗非鱼杂交配套系的相关数据进行有效的收集、存储、处理和应用;实现数据有效管理;具体有:收集罗非鱼杂交配套系的相关数据处理搜集的相关数据,进行统计分析,建立数据挖掘平台;搭建人机交互系统；

[0033] 通过数据储存模块将罗非鱼杂交配套系的相关数据储存在计算机中；

[0034] 通过数据挖掘模通过对大量的罗非鱼杂交配套系的相关数据进行分析,提取隐含的信息和知识。

[0035] 3、如权利要求2所述的罗非鱼杂交配套系的选育方法,其特征在于,

[0036] 收集罗非鱼杂交配套系的相关数据中,相关数据包括：

[0037] (1)罗非鱼的种类、数量、生存年份、形态标准、生存环境类数据信息；

[0038] (2)选育方案；

[0039] (3)基础群组建；

[0040] (4)关于专门化品系的选育法；

[0041] (5)数个杂交组合的比较试验,筛选组合。

[0042] 进一步,数据挖掘模块中基于关联规则映射的罗非鱼生物信息多维数据挖掘算法,进行分析、提取隐含的信息和知识;具体包括：

[0043] 假设子空间的维度为d,先挖掘处于不同子空间的不同数据集,子空间用矩阵M表示,为:

[0044] $$M=\begin{pmatrix} M^1 \\ \vdots \\ M^d \end{pmatrix} \quad d\leqslant n \qquad (1)$$

[0045] 假设两个数据集V_i和V_k分别位于两个不同的子空间M^i(i≤d)和M^k(k≤d),其中这两个子空间的欧几里德距离为D(i,k),两个数据集的欧几里德距离为d(i,k),则对于不同子空间的两个数据集的挖掘公式为：

[0046] $$W(M^i,M^k)=\frac{\sigma}{2}\begin{pmatrix} M^1 \\ \vdots \\ M^d \end{pmatrix} P(V_i)\,P(V_k)\times\log_2\sqrt{D(i,k)^2+d(i,k)^2} \qquad (2)$$

[0047] 其中:σ表示子空间挖掘因子,P(V_i)、P(V_k),分别表示数据集V_i和数据集V_k的挖掘频率；

[0048] 对于同一子空间的不同数据集的挖掘,通过不同数据集之间的关联程度进行区

分,先通过式

[0049] $$K_1\begin{pmatrix}\alpha_{ik}\\ \beta_{ik}\\ \theta_{ik}\end{pmatrix}=\begin{pmatrix}\alpha_{i1} & \cdots & \alpha_{1k}\\ \vdots & & \vdots\\ \alpha_{1k} & \cdots & \alpha_{i1}\end{pmatrix}\begin{pmatrix}\beta_{i1} & \cdots & \beta_{1k}\\ \vdots & & \vdots\\ \beta_{1k} & \cdots & \beta_{i1}\end{pmatrix}\begin{pmatrix}\theta_{i1} & \cdots & \theta_{1k}\\ \vdots & & \vdots\\ \theta_{1k} & \cdots & \theta_{i1}\end{pmatrix};$$

[0050] $$K_2\begin{pmatrix}\frac{1}{\alpha_{ik}}\\ \frac{1}{\beta_{ik}}\\ \frac{1}{\theta_{ik}}\end{pmatrix}=\begin{pmatrix}\frac{1}{\alpha_{i1}} & \cdots & \frac{1}{\alpha_{1k}}\\ \vdots & & \vdots\\ \frac{1}{\alpha_{1k}} & \cdots & \frac{1}{\alpha_{i1}}\end{pmatrix}\begin{pmatrix}\frac{1}{\beta_{i1}} & \cdots & \frac{1}{\beta_{1k}}\\ \vdots & & \vdots\\ \frac{1}{\beta_{1k}} & \cdots & \frac{1}{\beta_{i1}}\end{pmatrix}\begin{pmatrix}\frac{1}{\theta_{i1}} & \cdots & \frac{1}{\theta_{1k}}\\ \vdots & & \vdots\\ \frac{1}{\theta_{1k}} & \cdots & \frac{1}{\theta_{i1}}\end{pmatrix};$$

[0051] 求得K_1和K_2然后求得在同一空间下数据集V_i和V_k的关联因子:

[0052] $$g(i,k)=\sqrt{\frac{K_1}{K_2}\begin{pmatrix}x_{1i} & \cdots & x_{mi}\\ \vdots & & \vdots\\ x_{mi} & \cdots & x_{1i}\end{pmatrix}d(i,k)}-\ln 2\frac{1}{m}\left(\frac{K_2}{K_1}\begin{pmatrix}x_{1k} & \cdots & x_{mk}\\ \vdots & & \vdots\\ x_{mk} & \cdots & x_{1k}\end{pmatrix}\right)\quad(3);$$

[0053] 得到数据集V_i和V_k的关联因子g(i,k)之后,得到相同子空间下这两个数据集的挖掘公式为

[0054] $$W(V^i,V^k)=(P(V_i)-P(V_k))\,g(i,k)\,d(i,k)\times\frac{e^{g(i,k)}}{\begin{pmatrix}x_{1k} & \cdots & x_{mk}\\ \vdots & & \vdots\\ x_{mk} & \cdots & x_{1k}\end{pmatrix}\begin{pmatrix}x_{1i} & \cdots & x_{mi}\\ \vdots & & \vdots\\ x_{mi} & \cdots & x_{1i}\end{pmatrix}}\quad(4)$$

[0055] 假设在同一空间M^i下数据集之间关联程度限定阈值T(V),当数据集之间的关联因子g(i,k)大于T(V)时,则这两个数据集具有强相关性,则两个数据集的区分公式写成

[0056] $$f_{M^i}(V_i)=\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)-\ln\left(\sqrt{\sum_{j=1}^{n}(P(V_i)-P(V_j))}\right)\quad(5)$$

[0057] $$f_{M^i}(V_k)=\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)-\ln\left(\sqrt{\sum_{j=1}^{n}(P(V_k)-P(V_j))}\right)\quad(6)$$

[0058] 当数据集之间的关联因子g(i,k)小于T(V)时,则这两个数据集具有弱相关性,则两个数据集的区分公式写成

[0059] $$f_{M^i}(V_i)=\frac{\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)}{(\pi e^2-1)}+\frac{e}{2}\sum_{j=1}^{n}(P(V_i)-P(V_j))$$

[0060] $$f_{M^i}(V_k)=\frac{\begin{pmatrix}V_1 & \cdots & V_n\\ \vdots & & \vdots\\ V_n & \cdots & V_1\end{pmatrix}W(V^i,V^k)}{(\pi e^2-1)}+\frac{e}{2}\sum_{j=1}^{n}(P(V_k)-P(V_j))$$ 。

[0061] 进一步，数据储存模块中采用朴素贝叶斯分类算法将罗非鱼杂交配套系的相关数据储存在计算机中，具体包括：

[0062] 设D是训练对象与其相关联的类标号的集合，每个对象用一个n维属性向量X＝{x_1，x_2…x_n}表示，描述n维属性向量X＝{x_1，x_2…x_n}表示，描述n个属性A_1，A_2…A_n的值，假定原始集合基于n维属性共划分为m个类C1，C2…Cm，计算每个类对X的后验概率，并将对象X归属于具有最高后验概率的类，后验概率P(Ci|X)的计算公式为：

[0063] $$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$ ；

[0064] 由于P(Ci|X)的计算开销较大，进行类条件独立的假定，给定向量的类标号，并假定属性值有条件的相互独立，P(X|Ci)的计算公式为：

[0065] $$P(X|C_i)=\prod_{k=1}^{n}P(x_k|C_i)=P(x_1|C_i)P(x_2|C_i)\cdots P(x_n|C_i)$$ ；

[0066] 其中，P(x_1|Ci)P(x_2|Ci)…P(x_n|Ci)容易地由训练对象求算，X_k表示X在属性A_k上的值，对每个类别Ci计算P(X|Ci)P(Ci)，当P(X|Ci)P(Ci)＞P(X|Cj)P(Cj)，1≤j≤m，j≠i成立时，X属于类Ci。

[0067] 进一步，数据集成模块中采用不完备混合数据的集成聚类算法将互相关联的分布式异构数据源集成到一起，使用户以透明的方式访问罗非鱼杂交配套系的相关数据源；具体包括：

[0068] 输入：带有缺失值的数据集D、聚类个数k；

[0069] 输出：最终聚类结果$\pi^*(D)$；

[0070] 步骤一，对数据集D分别运用平均值填充法、KNN填充法、SKNN填充法填充得到完备数据集D_1，D_2，D_3；

[0071] 步骤二，对Di(1≤i≤3)分别执行M_i次K-Prototypes聚类算法，得到基聚类结果集Π(D)；

[0072] 步骤三，根据式

[0073] $$SM(\boldsymbol{x}_p,\boldsymbol{x}_q)=\frac{1}{\sum_{i=1}^{3}M_i}\sum_{i=1}^{3}\sum_{j=1}^{M_i}SM_i^j(\boldsymbol{x}_p,\boldsymbol{x}_q)$$;

[0074] 计算样本与样本之间的相似度矩阵$SM_{n\times n}$;

[0075] 步骤四,基于相似度矩阵$SM_{n\times n}$,分别根据以下式:

[0076] 单链(single link)方法.由2个类中相似度最大的2个样本决定

[0077] $$sim(C,C')=\max_{x\in C,x'\in C'}sim(\boldsymbol{x},\boldsymbol{x}')$$;

[0078] 全链(complete link)方法,由2个类中相似度最小的2个样本决定

[0079] $$sim(C,C')=\min_{x\in C,x'\in C'}sim(\boldsymbol{x},\boldsymbol{x}')$$;

[0080] 组平均(average link)方法,由2个类中所有样本点相似度的平均值决定

[0081] $$sim(C,C')=\frac{1}{|C||C'|}\sum_{x\in C}\sum_{x'\in C'}sim(\boldsymbol{x},\boldsymbol{x}')$$;

[0082] 式中:样本之间的相似度sim(x,x′)为相似度矩阵$SM_{n\times n}$中的对应元素值;

[0083] 运行层次聚类算法得到最终的聚类结果$\pi^*(D)$。

[0084] 本发明的另一目的在于提供一种罗非鱼杂交配套系的选育计算机程序,其特征在于,所述罗非鱼杂交配套系的选育计算机程序实现所述的罗非鱼杂交配套系的选育方法。

[0085] 本发明的另一目的在于提供一种终端,所述终端至少搭载实现所述的罗非鱼杂交配套系的选育方法的控制器。

[0086] 本发明的另一目的在于提供一种计算机可读存储介质,包括指令,当其在计算机上运行时,使得计算机执行所述的罗非鱼杂交配套系的选育方法。

[0087] 本发明的另一目的在于提供一种罗非鱼杂交配套系的选育系统包括数据集成模块、数据储存模块、数据管理模块、数据挖掘模块;

[0088] 数据集成模块,用于将互相关联的分布式异构数据源集成到一起,使用户以透明的方式访问罗非鱼杂交配套系的相关数据源;

[0089] 数据管理模块,用于利用计算机硬件和软件技术对罗非鱼杂交配套系的相关数据进行有效的收集、存储、处理和应用;实现数据有效管理;

[0090] 数据储存模块,用于将罗非鱼杂交配套系的相关数据储存在计算机中;

[0091] 数据挖掘模块,用于通过对大量的罗非鱼杂交配套系的相关数据进行分析,提取隐含的信息和知识

[0092] 本发明的优点及积极效果为:

[0093] 本发明通过对采集的罗非鱼的相关信息进行数据分析,建立人机交互系统,推算出罗非鱼杂交配套系的可靠的选育方法;将大数据的处理方法与罗非鱼杂交配套系的选育方法相结合,利用大数据处理的特点,有效的对罗非鱼的前期采集数据进行采集,节省了大量的人力物力,提高了选育技术的效率;同时利用神经网络训练算法对采集的数据进行处理,在隐藏层神经元数目足够的情况下,可对任意精度近似逼近任何连续的非线性函数,提

高了数据分析的准确性，提高了罗非鱼杂交配套系的选育的优良性与可靠性；本发明在提高罗非鱼杂交配套系的选育的工作效率的同时，提高了选育的准确性与优良性。在生物信息网络中对复杂和大规模的数据集进行挖掘时，采用基于关联规则映射的罗非鱼生物信息多维数据挖掘算法挖掘精度高、运行速度快、内存占用小等问题；实际应用中面临的数据往往是兼具数值属性和分类属性共同描述的混合型数据，采用不完备混合数据的集成聚类算，可以可负带有缺失值的缺点带来的问题；采用朴素贝叶斯分类算法能个处理多分类任务，适合增量式训练，算法简单，具有稳定的分类效率。

附图说明

[0094] 图1是本发明实施例提供的罗非鱼杂交配套系的选育方法的数据管理模块管理方法流程图。

[0095] 图2是本发明实施例提供的非鱼杂交配套系的选育系统示意图。

[0096] 图中：1、数据集成模块；2、数据储存模块；3、数据管理模块；4、数据挖掘模块。

具体实施方式

[0097] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0098] 本发明实施例提供的罗非鱼杂交配套系的选育方法，具体包括：

[0099] 通过数据集成模块，将互相关联的分布式异构数据源集成到一起，使用户以透明的方式访问罗非鱼杂交配套系的相关数据源；

[0100] 通过数据管理模块利用计算机硬件和软件技术对罗非鱼杂交配套系的相关数据进行有效的收集、存储、处理和应用；实现数据有效管理；具体有：收集罗非鱼杂交配套系的相关数据处理搜集的相关数据，进行统计分析，建立数据挖掘平台；搭建人机交互系统；

[0101] 通过数据储存模块将罗非鱼杂交配套系的相关数据储存在计算机中；

[0102] 通过数据挖掘模通过对大量的罗非鱼杂交配套系的相关数据进行分析，提取隐含的信息和知识。

[0103] 下面结合附图及具体实施例对本发明的应用原理作进一步描述。

[0104] 如图1所示，本发明实施例提供的数据管理模块管理方法包括以下步骤：

[0105] S101：收集罗非鱼杂交配套系的相关数据；

[0106] S102：处理搜集的相关数据，进行统计分析，建立数据挖掘平台；

[0107] S103：搭建人机交互系统。

[0108] 作为本发明的优选实施例，所述S101的相关数据包括：

[0109] (1)罗非鱼的种类、数量、生存年份、形态标准、生存环境类数据信息；

[0110] (2)选育方案；

[0111] (3)基础群组建；

[0112] (4)关于专门化品系的选育法；

[0113] (5)数个杂交组合的比较试验，筛选“最佳”组合。

[0114] 作为本发明的优选实施例，所述S102的处理搜集的相关数据采用神经网络训练算

法,如下:

[0115] (1)前向阶段

[0116] 在前向阶段,输入层获取到输入信号并将其传递到隐藏层中的每个神经元。然后,隐藏层处理这些信号并将处理结果传递到输出层。对于一个输入向量义X=(X1,X2,...Xm),隐藏层中每个神经元的输入和输出信号标记为u_j和h_j,这两个信号分别可W通过公式(1)和公式(2)算出;

[0117] $$u_j = \sum_{i=1}^{m} W_{ij} x_i + \theta_j (j = 1,2,...) \quad (1)$$

[0118]

$$h_j = f(u_j) = \frac{1}{1 + \exp(u_j)} (j = 1,2,q) \quad (2)$$

[0119] 其中W_{ij}是输入层神经元i和隐藏层神经元j之间的权重,θ_j是偏置。

[0120] 输出层从隐藏层获取到信号之后同样需要进行后续处理。输出层神经元的输入信号l_k和输出信号c_k分别由公式(3)和公式(4)计算得出。

[0121] $$l_k = \sum_{j=1}^{q} V_{jk} h_j + \gamma_k (k = 1,2,...n) \quad (3)$$

[0122] $$c_k = f(l_k) = \frac{1}{1 + \exp(-l_k)} (k = 1,2,n) \quad (4)$$

[0123] 其中V_{ij}是输入层神经元j和隐藏层神经元jk之间的权重,γ_k是偏置。

[0124] 至此,前向过程的信息处理流程结束。在前向过程中,神经网络模型权重W,V和偏置θ,γ并不发生变化。如果前向处理得出的神经网络最终输出信号与真实信号一致,那么下一个输入向量将被输入到该神经网络并开始新一轮的前向过程。否则,该算法将进入后向过程。这里,将神经网络的最终输出信号和真实信号之间的差值称为偏差(Error)。

[0125] (2)后向阶段

[0126] 在后向过程,首先将采用公式(5)计算出每个输出层神经元的偏差,然后进一步地利用公式(6)计算出每个隐藏层神经元e_i的偏差。

[0127] $d_k=(y_k-c_k)c_k(1-c_k)$ (k=1,2,) (5)

[0128]

$$e_j = (\sum_{k=1}^{n} d_k V_j) h_j (1 - h_j) (j = 1,2,) \quad (6)$$

[0129] 偏差从输出层反向回馈到隐藏层。通过送种偏差后向传播方式,利用公式(7)更新输出层和隐藏层的连接权重。进一步地,利用公式(8)更新隐藏层与输入层之间的连接权重。

[0130]
$$V_{jk}(N+1) = V_{jk}(N) + a_1 d_k(N) h_j$$
$$\gamma_k(N+1) = \gamma_k(N) + a_1 d_k(N) \quad (7)$$

[0131]
$$W_{ij}(N+1) = W_{ij}(N) + a_2 e_j(N)$$
$$\theta_j(N+1) = \theta_j(N) + a_2 d_j(N) \quad (8)$$

[0132] 在上述的公式中，i＝1,2,…m;j＝1,2,…q;k＝1,2,..n。a_1和a_2是取值范围在0到1的学习率。N表示当前训练轮数的编号。

[0133] 如图2,本发明实施例提供的非鱼杂交配套系的选育系统包括:数据集成模块1、数据储存模块2、数据管理模块3、数据挖掘模块4;

[0134] 数据集成模块1,数据集成是要将互相关联的分布式异构数据源集成到一起,使用户能够以透明的方式访问罗非鱼杂交配套系的相关数据源;

[0135] 数据管理模块3,利用计算机硬件和软件技术对罗非鱼杂交配套系的相关数据进行有效的收集、存储、处理和应用的过程;其目的在于充分有效地发挥数据的作用,实现数据有效管理的关键是数据组织;

[0136] 数据储存模块2,将罗非鱼杂交配套系的相关数据储存在计算机中;

[0137] 数据挖掘模块4,通过对大量的数据进行分析,以发现和提取隐含在其中的具有价值的信息和知识的过程。

[0138] 所述数据挖掘模块中基于关联规则映射的罗非鱼生物信息多维数据挖掘算法为:

[0139] 假设子空间的维度为d,先挖掘处于不同子空间的不同数据集,其中子空间用矩阵M表示,定义为

[0140] $$M=\begin{pmatrix} M^1 \\ \vdots \\ M^d \end{pmatrix} \quad d\leqslant n \tag{1}$$

[0141] 假设两个数据集V_i和V_k分别位于两个不同的子空间M^i(i≤d)和M^k(k≤d),其中这两个子空间的欧几里德距离为D(i,k),两个数据集的欧几里德距离为d(i,k),则对于不同子空间的两个数据集的挖掘公式为:

[0142] $$W(M^i,M^k)=\frac{\sigma}{2}\begin{pmatrix} M^1 \\ \vdots \\ M^d \end{pmatrix}P(V_i)\,P(V_k)\times\log_2\sqrt{D(i,k)^2+d(i,k)^2} \tag{2}$$

[0143] 其中:σ表示子空间挖掘因子,P(V_i)、P(V_k),分别表示数据集V_i和数据集V_k的挖掘频率;

[0144] 对于同一子空间的不同数据集的挖掘,通过不同数据集之间的关联程度进行区分,先通过式

[0145] $$K_1\begin{pmatrix} \alpha_{ik} \\ \beta_{ik} \\ \theta_{ik} \end{pmatrix}=\begin{pmatrix} \alpha_{i1} & \cdots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{1k} & \cdots & \alpha_{i1} \end{pmatrix}\begin{pmatrix} \beta_{i1} & \cdots & \beta_{1k} \\ \vdots & & \vdots \\ \beta_{1k} & \cdots & \beta_{i1} \end{pmatrix}\begin{pmatrix} \theta_{i1} & \cdots & \theta_{1k} \\ \vdots & & \vdots \\ \theta_{1k} & \cdots & \theta_{i1} \end{pmatrix};$$

[0146] $$K_2\begin{pmatrix} \frac{1}{\alpha_{ik}} \\ \frac{1}{\beta_{ik}} \\ \frac{1}{\theta_{ik}} \end{pmatrix}=\begin{pmatrix} \frac{1}{\alpha_{i1}} & \cdots & \frac{1}{\alpha_{1k}} \\ \vdots & & \vdots \\ \frac{1}{\alpha_{1k}} & \cdots & \frac{1}{\alpha_{i1}} \end{pmatrix}\begin{pmatrix} \frac{1}{\beta_{i1}} & \cdots & \frac{1}{\beta_{1k}} \\ \vdots & & \vdots \\ \frac{1}{\beta_{1k}} & \cdots & \frac{1}{\beta_{i1}} \end{pmatrix}\begin{pmatrix} \frac{1}{\theta_{i1}} & \cdots & \frac{1}{\theta_{1k}} \\ \vdots & & \vdots \\ \frac{1}{\theta_{1k}} & \cdots & \frac{1}{\theta_{i1}} \end{pmatrix};$$

[0147] 求得K_1和K_2然后求得在同一空间下数据集V_i和V_k的关联因子:

[0148] $$g(i,k)=\sqrt{\frac{K_1}{K_2}\begin{pmatrix} x_{1i} & \cdots & x_{mi} \\ \vdots & & \vdots \\ x_{mi} & \cdots & x_{1i} \end{pmatrix}d(i,k)}-\ln 2\frac{1}{m}\left(\frac{K_2}{K_1}\begin{pmatrix} x_{1k} & \cdots & x_{mk} \\ \vdots & & \vdots \\ x_{mk} & \cdots & x_{1k} \end{pmatrix}\right) \quad (3);$$

[0149] 得到数据集V_i和V_k的关联因子g(i,k)之后,可以得到相同子空间下这两个数据集的挖掘公式为

[0150] $$W(V^i,V^k)=(P(V_i)-P(V_k))\,g(i,k)\,d(i,k)\times \frac{e^{g(i,k)}}{\begin{pmatrix} x_{1k} & \cdots & x_{mk} \\ \vdots & & \vdots \\ x_{mk} & \cdots & x_{1k} \end{pmatrix}\begin{pmatrix} x_{1i} & \cdots & x_{mi} \\ \vdots & & \vdots \\ x_{mi} & \cdots & x_{1i} \end{pmatrix}} \quad (4)$$

[0151] 假设在同一空间M^i下数据集之间关联程度限定阈值T(V),当数据集之间的关联因子g(i,k)大于T(V)时,则这两个数据集具有强相关性,则两个数据集的区分公式写成

[0152] $$f_{M^i}(V_i)=\begin{pmatrix} V_1 & \cdots & V_n \\ \vdots & & \vdots \\ V_n & \cdots & V_1 \end{pmatrix}W(V^i,V^k)-\ln\left(\sqrt{\sum_{j=1}^{n}(P(V_i)-P(V_j))}\right) \quad (5)$$

[0153] $$f_{M^i}(V_k)=\begin{pmatrix} V_1 & \cdots & V_n \\ \vdots & & \vdots \\ V_n & \cdots & V_1 \end{pmatrix}W(V^i,V^k)-\ln\left(\sqrt{\sum_{j=1}^{n}(P(V_k)-P(V_j))}\right) \quad (6)$$

[0154] 当数据集之间的关联因子g(i,k)小于T(V)时,则这两个数据集具有弱相关性,则两个数据集的区分公式写成

[0155] $$f_{M^i}(V_i)=\frac{\begin{pmatrix} V_1 & \cdots & V_n \\ \vdots & & \vdots \\ V_n & \cdots & V_1 \end{pmatrix}W(V^i,V^k)}{(\pi e^2-1)}+\frac{e}{2}\sum_{j=1}^{n}(P(V_i)-P(V_j))$$

[0156] $$f_{M^i}(V_k)=\frac{\begin{pmatrix} V_1 & \cdots & V_n \\ \vdots & & \vdots \\ V_n & \cdots & V_1 \end{pmatrix}W(V^i,V^k)}{(\pi e^2-1)}+\frac{e}{2}\sum_{j=1}^{n}(P(V_k)-P(V_j))$$ 。

[0157] 作为本发明的优选实施例,所述数据储存模块中采用朴素贝叶斯分类算法为:

[0158] 设D是训练对象与其相关联的类标号的集合,每个对象用一个n维属性向量X={x_1,x_2…x_n}表示,描述n维属性向量X={x_1,x_2…x_n}表示,描述n个属性A_1,A_2…A_n的值,假定

原始集合基于n维属性共划分为m个类C1,C2…Cm,计算每个类对X的后验概率,并将对象X归属于具有最高后验概率的类,后验概率P(Ci|X)的计算公式为:

[0159] $$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$;

[0160] 由于P(Ci|X)的计算开销较大,进行类条件独立的假定,给定向量的类标号,并假定属性值有条件的相互独立,P(X|Ci)的计算公式为:

[0161] $$P(X|C_i)=\prod_{k=1}^{n}P(x_k|C_i)=P(x_1|C_i)P(x_2|C_i)\cdots P(x_n|C_i)$$;

[0162] 其中,P(x1|Ci)P(x2|Ci)…P(xn|Ci)可以容易地由训练对象求算,Xk表示X在属性Ak上的值,对每个类别Ci计算P(X|Ci)P(Ci),当P(X|Ci)P(Ci)>P(X|Cj)P(Cj),1≤j≤m,j≠i成立时,X属于类Ci。

[0163] 作为本发明的优选实施例,所述数据集成模块中采用不完备混合数据的集成聚类算法:

[0164] 输入:带有缺失值的数据集D、聚类个数k;

[0165] 输出:最终聚类结果π*(D);

[0166] 步骤一,对数据集D分别运用平均值填充法、KNN填充法、SKNN填充法填充得到完备数据集D1,D2,D3;

[0167] 步骤二,对Di(1≤i≤3)分别执行Mi次K-Prototypes聚类算法,得到基聚类结果集Π(D);

[0168] 步骤三,根据式

[0169] $$SM(\boldsymbol{x}_p,\boldsymbol{x}_q)=\frac{1}{\sum_{i=1}^{3}M_i}\sum_{i=1}^{3}\sum_{j=1}^{M_i}SM_i^j(\boldsymbol{x}_p,\boldsymbol{x}_q)$$;

[0170] 计算样本与样本之间的相似度矩阵SMn×n;

[0171] 步骤四,基于相似度矩阵SMn×n,分别根据以下式:

[0172] 单链(single link)方法.由2个类中相似度最大的2个样本决定

[0173] $$sim(C,C')=\max_{x\in C,x'\in C'}sim(\boldsymbol{x},\boldsymbol{x}')$$;

[0174] 全链(complete link)方法,由2个类中相似度最小的2个样本决定

[0175] $$sim(C,C')=\min_{x\in C,x'\in C'}sim(\boldsymbol{x},\boldsymbol{x}')$$;

[0176] 组平均(average link)方法,由2个类中所有样本点相似度的平均值决定

[0177] $$sim(C,C')=\frac{1}{|C||C'|}\sum_{x\in C}\sum_{x'\in C'}sim(\boldsymbol{x},\boldsymbol{x}')$$;

[0178] 式中:样本之间的相似度sim(x,x′)为相似度矩阵SMn×n中的对应元素值;

[0179] 运行层次聚类算法得到最终的聚类结果π*(D)。

[0180] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用全部或部分地以计算机程序产品的形式实现，所述计算机程序产品包括一个或多个计算机指令。在计算机上加载或执行所述计算机程序指令时，全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL)或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输)。所述计算机可读取存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘Solid State Disk(SSD))等。

[0181] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

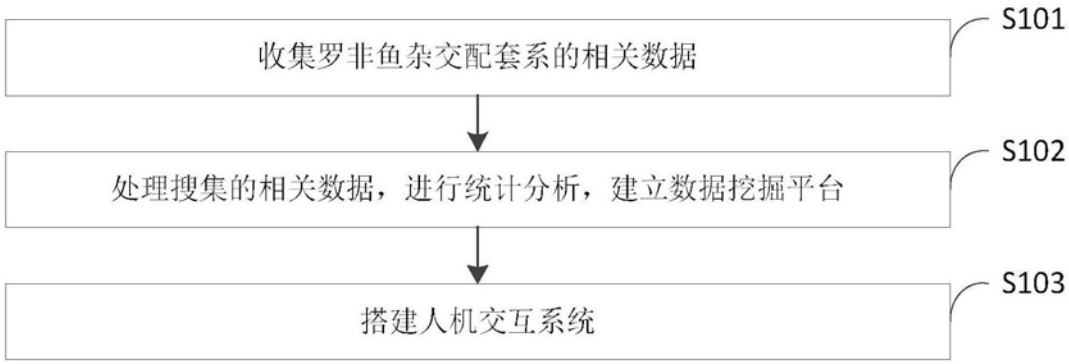
收集罗非鱼杂交配套系的相关数据
S101
处理搜集的相关数据，进行统计分析，建立数据挖掘平台
S102
搭建人机交互系统
S103

图1

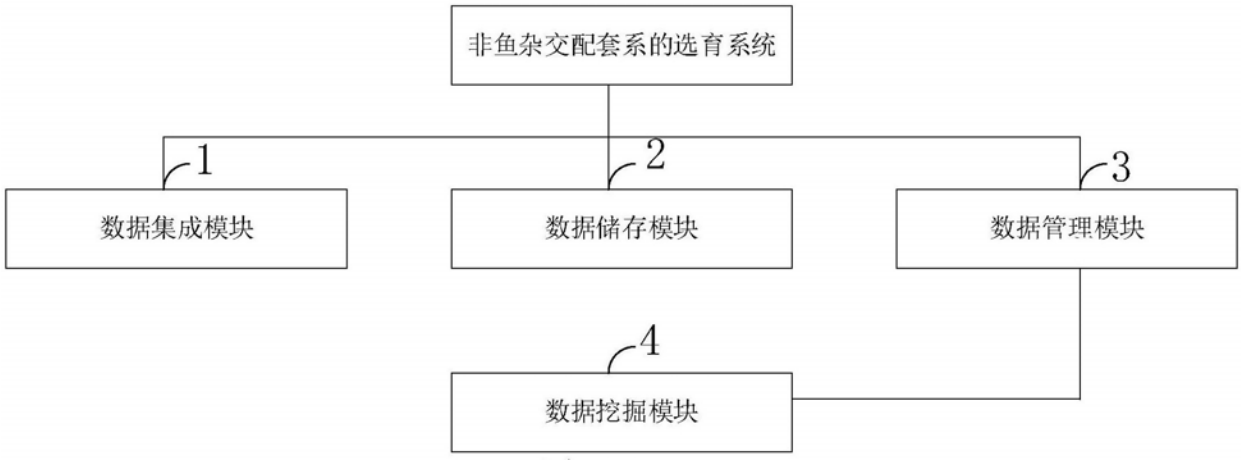
非鱼杂交配套系的选育系统
1
数据集成模块
2
数据储存模块
3
数据管理模块
4
数据挖掘模块

图2