

基于改进遗传算法的不完整大数据填充挖掘算法

郑 斌

(福建江夏学院 电子信息科学学院, 福建 福州 350108)

摘 要: 传统的大数据填充挖掘过程存在耗时高、效率低的缺陷, 对此提出基于改进遗传算法的不完整大数据填充挖掘算法. 该算法先融合于 BP 人工神经网络算法随机产生不完整大数据初始网络权值, 并对该组权值进行训练, 设计一个不完整大数据基因矩阵, 利用遗传算法以该基因矩阵为依据计算出不完整大数据适应值函数, 并同时同时进行编码、杂交、变异等遗传操作, 在此基础上进行不完整大数据信息的全局搜索, 以其搜索的结果为核心将大数据划分为完整与不完整数据, 采用信息论中熵值的概念利用同一类完整数据信息的属性值对缺失值进行填充. 实验仿真证明, 基于改进遗传算法的不完整大数据填充挖掘方法能实现对不完整大数据的挖掘, 提高缺失数据的填充精度.

关键词: 不完整大数据; 数据填充; 遗传神经网络

中图分类号: TP311. 13

文献标识码: A

文章编号: 1000 - 7180(2016)02 - 0096 - 04

Incomplete Data Filling Mining Algorithm Based on the Improved Genetic Algorithm

ZHENG Bin

(College of Electronics and Information Science, Fujian Jiangxia University, Fuzhou 350108, China)

Abstract: Traditional big data high filling mining process is time-consuming, the defects of low efficiency. Based on improved genetic algorithm of incomplete data filling mining algorithm. This algorithm fusion in the BP artificial neural network algorithm first randomly generated incomplete data the initial network weights, and to train the weights of the group to design a incomplete data gene matrix, by using the genetic algorithm based on the gene matrix to calculate the incomplete data adaptive value function, and at the same time coding, hybridization and mutation genetic operation, on the basis of incomplete big data information of global search, with its search results as the core will be divided into complete and incomplete data, using the concept of entropy in information theory, and the same kind of complete data information of attribute values to fill the missing value. The experimental simulation proves that the incomplete big data based on improved genetic algorithm of incomplete filling mining method can realize data mining, improve the filling precision of the missing data.

Key words: incomplete data; Data filled; Genetic neural network

1 引言

在国内, 由于网络科技的飞速发展, 数据信息呈现了大规模的增长趋势, 此时迫切地需要一种将大量数据转换成相应的信息和知识的数据挖掘技术^[1-3]. 然而在实际的操作过程中, 由于信息无法被获取的原因造成了现实中的数据大量缺失, 致使挖

掘过程产生混乱^[4-5]. 在这种情况下, 不完整大数据填充挖掘方法可以实现对不完整大数据的挖掘, 提高了对缺失数据的填充, 是解决上述问题的根本途径, 引起了很多专家与学者的重视^[6]. 由于不完整大数据填充挖掘方法具有深远的发展意义, 因此也成为了业内人士研究的焦点课题, 受到了广泛的关注.

目前已有不少的解决方法.文献[7]采用了基于改进朴素贝叶斯算法的不完整大数据填充挖掘方法.该方法先利用朴素贝叶斯算法对源数据进行分类,在以分类的结果为基础利用EM算法实现对不完整大数据填充挖掘.该方法较为简单,但是存在挖掘误差大的问题.文献[8]重点研究了基于改进粗糙集算法的不完整大数据填充挖掘方法.该方法先利用粗糙集理论对不完整大数据的特征信息进行缺陷识别,融合于谱系聚类重心算法对不完整大数据进行填充,有效地完成了对不完整大数据填充挖掘.该方法时间复杂度较低,但是存在对大数据无法进行精确分类的问题.文献[9]重点研究了基于近邻传播算法的不完整大数据填充挖掘方法.该方法利用AP聚类算法对不完整的大数据进行聚类,融合于云计算技术算法对聚类的结果进行填充,进而实现了对不完整大数据填充挖掘.该方法鲁棒性较强,但是采用当前的算法进行不完整大数据填充挖掘时,缺失的数据难以被确定,导致了不完整大数据填充挖掘过程较为混乱.

针对上述问题的产生,文中提出了一种基于改进遗传算法的不完整大数据填充挖掘方法.该方法先利用BP人工神经网络算法建立不完整大数据基因矩阵,利用改进遗传算法对该矩阵进行复制、杂交、变异等遗传操作,并进行不完整大数据全局搜索,将大数据划分为完整与不完整数据,采用信息熵的数据填充理论思想利用同一类完整数据信息的属性值对缺失值进行填充,有效完成了对不完整大数据的填充挖掘.

2 不完整大数据填充挖掘优化原理

2.1 不完整大数据适应度函数值的计算

在对不完整大数据填充优化挖掘时,先融合于BP人工神经网络算法随机产生不完整大数据初始网络权值,对该组权值进行训练,设计一个不完整大数据基因矩阵,利用遗传算法以该基因矩阵为依据计算出不完整大数据适应值函数,并同时编码、杂交、变异等遗传操作,在此基础上进行不完整大数据信息的全局搜索.具体的步骤如下所示.

随机选取 N 组不完整大数据的初始网络权值,将其取值设定在 $(-1, 1)$ 区间内,对 N 组权值分别进行训练,依据训练好的 N 组优质的权值所相对应的上下限再次形成 N 组新的权值,与经过训练的权值在一起,设计出一个由 $2 * N$ 代表的不完整大数据基因矩阵,利用下式表示:

$$2 * N = \frac{N \otimes (-1, 1)}{N} \quad (1)$$

利用遗传算法以该基因矩阵为依据计算出不完整大数据的适应值函数,利用下式表述:

$$\text{Fithess} = \frac{\sum_{i=1}^N (Y_i^2 - \overset{\circ}{Y}_i)^2}{2 * N} \quad (2)$$

式中, Fithess 代不完整大数据适应值, $\overset{\circ}{Y}_i$ 代表第 i 个不完整大数据样本的实际输出, Y_i 代表第 i 个不完整大数据样本的期望输出, N 代表对于不完整大数据的学习样本.

在对不完整大数据填充挖掘优化过程中,利用浮点数直接表示编码问题,依据杂交算子建立具有不完整大数据的数值特点的向量的线性组合,假设由 S_a, S_b 分别代表两个杂交个体,则可以得出下式:

$$S_a' = \frac{uS_a + (1 - u)S_b}{\text{Fithess}} \quad (3)$$

$$S_a' = \frac{uS_b + (1 - u)S_a'}{\text{Fithess}} \quad (4)$$

式中, u 的代表浮点数,取值设定在 $(0, 1)$ 区间内, S_a', S_b' 代表杂交个体的不完整大数据数值特点的向量的线性组合.

在对不完整大数据填充优化挖掘过程中,设由 $X = (V_1, V_2, \dots, V_n)$ 代表不完整大数据的分量变异个体,如果每次分量以一定概率进行变异,则由 $(V'_1, V'_2, \dots, V'_n)$ 代表一次变异后的结果.则可以得出下式:

$$V'_k = \begin{cases} v_k = \frac{\Delta(t, \text{UB} - v_k)}{S'} \\ v_k = \frac{\Delta(t, v_k - \text{LB})}{S'_b} \end{cases} \quad (5)$$

式中, LB, UB 分别代表第 k 个不完整大数据变量的左、右邻居, v_k 代表变异点的位置, t 代表变异算子, V'_k 代表一次变异后的结果.

在对不完整大数据填充优化挖掘过程中,利用 $\Delta(t, y)$ 代表的不完整大数据目标函数进行不完整大数据信息的全局搜索,利用下式可以表述

$$\Delta(t, y)''' = \frac{y' * (1 - r^{(1-t/T^b)})}{\Delta(t, y) V'_k} \quad (6)$$

式中, $\Delta(t, y)'''$ 代表进行不完整大数据信息全局搜索的结果, r 代表 $[0, 1]$ 范围内的随机数, T 代表预制的最大代数, y' 代表不完整大数据的基因.

由以上可知,在对不完整大数据填充优化挖掘

过程中,先融合于BP人工神经网络算法随机产生不完整大数据初始网络权值,对该组权值进行训练,设计一个不完整大数据基因矩阵,利用遗传算法以该基因矩阵为依据计算出不完整大数据适应值函数,并同时进行编码、杂交、变异等遗传操作,在此基础上进行不完整大数据信息的全局搜索,为实现不完整大数据填充优化挖掘奠定了基础。

2.2 不完整大数据填充挖掘的实现

以上式获取的不完整大数据信息的全局搜索结果为核心将大数据划分为完整与不完整数据集,采用信息论中熵值的概念利用同一类完整数据信息的属性值对缺失值进行填充。具体的步骤如下所示。

假设由 O 代表源大数据集依据上式获取的不完整大数据信息的全局搜索结果为核心将大数据集划分为 I 代表的不完整大数据集与 C 代表的完整大数据集,利用下式表示:

$$O = \Delta(t, y)^m \times (I, C) \quad (7)$$

采用信息论中熵值的概念,设由 o 代表的任意缺失数据属性信息,将其位于同一区间内的完整数据属性表述为 $\{o_1, o_2, o_m \dots\}$, 并计算 o 与 $\{o_1, o_2, o_m \dots\}$ 的相似度,将其结果表述为 $S_1, S_2, \dots, S_m$, 利用下式表示:

$$\text{similarity}\{S_1, S_2, \dots, S_m\} = (I, C) \times \frac{O}{\{o_1, o_2, o_m \dots\}} \quad (8)$$

将 $S_1, S_2, \dots, S_m$ 依据下式进行单位化:

$$p_i = \frac{\{S_1, S_2, \dots, S_m\}}{s_i / \sum_{i=1}^m s_i} \quad (9)$$

式中, s_i 代表不完整大数据的隶属概率, m 代表对不完整大数据的估计值。

依据下式计算各个完整数据信息的熵值:

$$h_i = \frac{-p_i}{\ln p_i} \quad (10)$$

式中, h_i 代表各个完整数据信息的熵值。

各个完整数据信息的权重利用下式表述:

$$w_i = (1 - h) / (m - \sum h_i) \quad (11)$$

式中, w_i 代表各个完整数据信息的权重。

在对不完整大数据填充优化挖掘过程中,依据下式计算 o 的缺失属性值 f :

$$f = \frac{\sum_{i=1}^m w_i \times x_i | o_i |}{o} \quad (12)$$

式中, x_i 代表不完整数据 o_i 的对应属性值。

由以上可知,不完整大数据填充挖掘优化的原理,这能够有效地完成了对不完整大数据填充优化挖掘。

3 实验结果与分析

为了证明基于改进遗传算法的不完整大数据填充挖掘方法的有效性需要进行一次实验,利用 Matlab 平台搭建不完整大数据填充挖掘的仿真虚拟实验平台。

实验 1: 任意从由 p 代表的源数据集中选取 3%、6%、9%、10%、15%、10%、26% 的数据信息,删除这些数据信息中的部分属性值,利用改进算法、文献[8]算法、文献[9]算法进行不完整大数据填充挖掘实验,将 2 种算法的挖掘精度进行对比,对比结果见图 1。

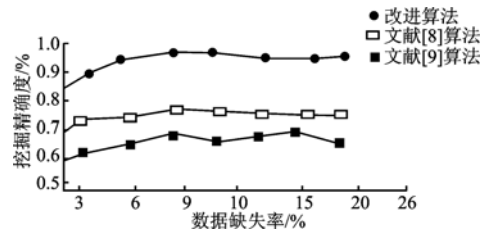


图 1 不同算法数据挖掘精度对比

从图 1 中可以看出,改进算法对不完全数据的挖掘精度要优于传统算法,这主要是因为改进算法先融合于BP人工神经网络算法随机产生不完整大数据初始网络权值,对该组权值进行训练,设计一个不完整大数据基因矩阵,利用遗传算法以该基因矩阵为依据计算出不完整大数据适应值函数,并同时进行编码、杂交、变异等遗传操作,在此基础上进行不完整大数据信息的全局搜索,以其搜索的结果为核心将大数据划分为完整与不完整数据,从而保障了改进算法对不完整数据挖掘的精度。

实验 2: 任意从由 p 代表的源数据集中选取 3%、6%、9%、10%、15%、10%、26% 的数据信息,删除这些数据信息中的部分属性值,模拟不完整数据集 o , 分别利用改进算法、文献[8]算法、文献[9]算法进行不完整大数据填充,对比结果见图 2。

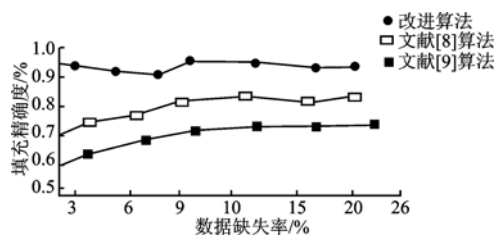


图2 不同算法填充精度比较

从图2可以说明,改进算法对缺失大数据填充精确度要优于传统算法,这是因为改进算法采用信息论中熵值的概念利用同一类完整数据信息的属性值对缺失值进行填充,从而提升了改进算法对不完整数据填充的精确性。

4 结束语

针对采用当前的算法进行不完整大数据填充挖掘时,缺失的数据难以被确定,导致了不完整大数据填充挖掘过程较为混乱的问题,提出了一种基于改进遗传算法的不完整大数据填充挖掘算法。该算法先融合于BP人工神经网络法随机产生不完整大数据初始网络权值,对该组权值进行训练,设计一个不完整大数据基因矩阵,利用遗传算法以该基因矩阵为依据计算出不完整大数据适应值函数,并同时编码、杂交、变异等遗传操作,在此基础上进行不完整大数据信息的全局搜索,以其搜索的结果为核心将大数据划分为完整与不完整数据,采用信息论中熵值的概念利用同一类完整数据信息的属性值对缺失值进行填充。实验仿真证明,基于改进遗传算

法的不完整大数据填充挖掘方法能实现对不完整大数据的挖掘,提高缺失数据的填充精度。

参考文献:

- [1] 朱彦君,吴向阳. 基于张量分解的多维数据填充算法[J]. 计算机工程, 2014, 40(5):45-48.
- [2] 陈志奎,吕爱玲,张清辰. 基于属性重要性的不完备数据填充算法[J]. 微电子学与计算机, 2013,30(7):167-172.
- [3] 陈志奎,杨英达,张清辰,等. 基于属性约简的物联网不完全数据填充算法[J]. 计算机工程与设计, 2013, 34(2):418-422.
- [4] 冷泳林,陈志奎,张清辰,等. 不完整大数据的分布式聚类填充算法[J]. 计算机工程, 2015, 41(5):19-25.
- [5] 梁肖肖,常家树,董沁怡,等. 基于遗传算法的PID参数整定与优化[J]. 福建电脑, 2014, 29(2):31-32.
- [6] 李家成,苏一丹,覃华,等. 基于遗传算法的K调和均值聚类算法[J]. 计算机技术与发展, 2013, 22(9):55-58.
- [7] 王梦雪. 数据挖掘综述[J]. 软件导刊, 2013, 12(10):135-137.
- [8] 王慧,张翠羽. 基于改进遗传算法的网络差异数据挖掘算法[J]. 计算机仿真, 2015, 32(5):311-314.
- [9] 杨海霞. 基于改进数据过滤算法的异构数据库优化挖掘技术[J]. 科技通报, 2013, 29(6):125-127.

作者简介:

郑斌女,(1977-),硕士,讲师.研究方向为信息处理和数据分析. E-mail:frostzb@163.com

(上接第95页)

- [9] 杨跃东,王莉莉,郝爱民. 运动串:一种用于行为分割的运动捕获数据表示方法[J]. 计算机研究与发展, 2008, 45(3):527-534.
- [10] 柴桦,邹北骥. 基于条件随机场的连续运动识别技术[J]. 计算机工程 and 科学, 2009, 31(5):53-56.
- [11] 田苗,冯进良,文琴,等. 基于Linux系统的指纹识别门禁系统的研究[J]. 电子设计工程, 2014, 22(2):134-137.

作者简介:

刘凌霄女,(1977-),硕士研究生,副教授.研究方向为图形图像处理、智能信息控制.

宋强男,(1971-),博士研究生,副教授.研究方向为人工智能、图形图像处理.

E-mail:feixiang_883366@163.com