

部分缺失数据的三次参数样条函数修补方法

张春生, 赵凌琪, 庄丽艳

(内蒙古民族大学 计算机科学与技术学院, 内蒙古 通辽 028043)

摘 要: 首先讨论了完整数据对数据挖掘的重要性, 分析了目前解决缺失数据的一些方法的局限性, 提出了利用三次参数样条函数对缺失数据进行修补的方法. 该方法简单易行, 不需要进行大量的样本训练, 并且与主成分分析方法和列平均值方法进行了比较, 确定三次参数样条函数方法与主成分分析方法接近. 若进一步对三次参数样条函数的构造进行完善, 将取得更好的效果.

关键词: 三次参数样条函数; 缺失数据; 修补; 主成分分析方法

中图分类号: TP392

文献标识码: A

文章编号: 1000-7180(2010)05-0196-03

Mending Method for Partially Missing Data Using Cubic Spline Parameter Functions

ZHANG Chun-sheng, ZHAO Ling-qi, ZHUANG Li-yan

(School of Computer Science and Technology of Inner Mongolia
University for the Nationalities, Tongliao 028043, China)

Abstract: This article first discusses the importance of complete data for data mining and analyzes the limits of some methods of solving missing data at present, and proposes the mending method for missing data using cubic parameter spline function, which is easy to use and do not need a lot of sample training. The method is compared with the principal component analytical method and the column average method, and it is ensured that cubic parameter spline function is close to the principal component analytical method. If the structure of cubic parameter spline function is further perfected, it will obtain the better effect.

Key words: cubic parameter spline function; missing data; mending; the principal component analytical method

1 引言

目前的数据挖掘算法一般都是基于完整数据而言的, 但实际情况下, 数据库中的数据不一定是完整的, 存在一定的数据缺失, 数据的不完整性严重影响了数据挖掘的效果, 目前的关联规则发现算法^[1]在数据库中含有缺失的数据时就得不到令人满意的结果^[2]. 所以缺失数据的修补至关重要.

Kononenko 等人^[3]采用贝叶斯方法确定缺失值; Quinlan 等人^[4]用其他已知属性的值和分类信息来预测一个数据的缺失值; 也有不少研究者研究了基于粗糙集理论的不完全系统中的知识发现方

法; Lakshminarayan 等人在已有的机器学习系统 AutoClass 的基础上开发了一个可处理缺失数据的知识发现系统^[5]. Ragel 在文献^[2]中提出了一种在含有缺失数据的数据库中发现关联规则的方法.

还有人应用小波分析、神经网络等算法对缺失的数据进行预测, 但必须进行大量样本的训练, 过程相对复杂^[6], 文献^[7]应用主成分分析方法实现缺失数据的修补.

由于三次参数样条曲线对曲线的拟合有较好的效果, 特别是算法简单易行, 复杂度低, 是缺失数据修补的一种有效方法, 而主成分分析方法本质上是一种线性分析方法, 如果对集合的归纳不是直线, 而

收稿日期: 2009-04-10; 修回日期: 2009-07-25

基金项目: 内蒙古教育科研项目(NJZY07140)

是曲线或曲面,单纯的主成分分析就不一定有效.

文中提出用三次参数样条曲线来拟合具有代表性的特征点,然后利用插值的方法实现缺失数据的修补.

2 三次参数样条函数原理

三次参数样条函数是根据一些有效特征点进行曲线拟合以达到对未知特征进行插值预测,以实现

对缺失数据修补的目的.
端点条件:设曲线始、末两端点的二阶导数为零.

三次参数样条函数公式为:

$$P(t) = B_1 + B_2t + B_3t^2 + B_4t^3 \tag{1}$$

可以看出,求得 P 的关键问题是确定 B_1 、 B_2 、 B_3 、 B_4 四个系数.

$$B_1 = P_i \tag{2}$$

$$B_2 = P'_i \tag{3}$$

$$B_3 = \frac{3(P_{i+1} - P_i)}{t_{i+1}^2} - \frac{2P'_i + P'_{i+1}}{t_{i+1}} \tag{4}$$

$$B_4 = \frac{2(P_i - P_{i+1})}{t_{i+1}^3} + \frac{P'_i + P'_{i+1}}{t_{i+1}^2} \tag{5}$$

其中, P_i 是已知的特征点. 由于 P_i 已知,关键在于求 P'_i ,在前面假设的端点条件下,可由以下方程求得:

$$\begin{bmatrix} 1 & 0.5 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & & \vdots & \vdots & \ddots & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & 1 & 2 \end{bmatrix} \begin{bmatrix} P'_1 \\ P'_2 \\ P'_3 \\ \vdots \\ P'_{n-1} \\ P'_n \end{bmatrix} = \begin{bmatrix} \frac{3(P_2 - P_1)}{2t_2} \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ \frac{3(P_n - P_{n-1})}{t_n} \end{bmatrix} \tag{6}$$

式中:

$$a_i = t_{i+1}$$

$$b_i = 2(t_{i+1} + t_i)$$

$$c_i = t_i$$

$$d_i = 3[\frac{t_i}{t_{i+1}}(P_{i+1} - P_i) + \frac{t_{i+1}}{t_i}(P_i - P_{i-1})]$$

$$i = 1, 2, 3, \dots, n - 1$$

求得 P'_i 则得到 B_1 、 B_2 、 B_3 、 B_4 四个系数,进而得到三次参数样条函数 $P(t)$,然后可利用变量 t_i 实现插值,以达到缺失数据修补的目的.

3 实例

文中采用文献[7]提供的测试数据以及用主成

分分析方法、列平均值法二种方法得到的测试结果,采用三次参数样条函数方法对测试数据进行测试,由于三次参数样条函数不适合于端点测试,所以,只对中间点进行测试,结果如表 1~表 4 所示.

表 1 文献[7]原始测试数据

Rec #	X1	X2	X3	X4
1	149.3	4.2	108.1	15.9
5	180.8	1.1	132.1	18.8
7	202.1	2.1	146.0	22.7
9	226.1	5.0	162.3	28.1
11	239.0	0.7	167.6	26.3

表 2 文献[7]主成份分析方法测试结果

Rec #	X1	X2	X3	X4
1	155.18	4.2	108.1	15.9
5	180.8	0.33	132.1	18.8
7	202.1	2.1	146.0	30.16
9	226.1	5.0	155.38	33.19
11	239.0	0.7	167.6	26.3

表 3 文献[7]列平均值方法测试结果

Rec #	X1	X2	X3	X4
1	194.59	4.2	108.1	15.9
5	180.8	3.3	132.1	18.8
7	202.1	2.1	146.0	21.89
9	226.1	5.0	139.74	21.89
11	239.0	0.7	167.6	26.3

表 4 三次参数样条函数方法测试结果

Rec #	X1	X2	X3	X4
1	188.1	4.2	108.1	15.9
5	180.8	2.36	132.1	18.8
7	202.1	2.1	146.0	20.4
9	226.1	5.0	152.3	22.3
11	239.0	0.7	167.6	26.3

4 结束语

采用 $RMS = \sqrt{\frac{1}{N \times M} \sum_{i=1}^N \sum_{j=1}^M (X'_{ij} - X_{ij})^2}$ 作为误差分析工具, N 和 M 是测试数据的行和列数, X'_{ij} 为测试值, X_{ij} 为精确值, RMS 为误差. 实例的误差比较结果如表 5 所示.

表 5 三种方法 RMS 比较表

方法	主成分方法	列平均值法	三次参数样条方法
误差	2.54	5.26	2.64

从表 5 可以看出,采用三次参数样条函数方法优于列平均值方法,接近主成分方法,三次参数样条函数方法与其他方法相比,算法简单、容易实现,特别是在数据点呈曲线变化时三次参数样条函数的效果更佳,由于本例中,数据点基本上呈直线变化,故主成分分析方法与三次参数样条方法相似。

对于利用小波分析、傅立叶函数、神经网络等技术进行缺失数据修补的方法这里没有应用,因为他们都要进行大量的样本训练,而三次参数样条函数方法不进行样本训练,所以我们采用了主成份分析方法和列平均值方法进行了比较。

文中算法可用于修补缺失数据的中间点,但对于端点的缺失数据修补应进一步研究,若采用文献[8]的方法进一步去构建精确的三次参数样条函数并采用 5 个定位点,将取得更好的效果。

参考文献:

- [1] Grawa A I, Rannila M H, Srikant R, et al. Fast discovery of association rules[C]// Advance in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1995:307-328.
- [2] Ragel A, Cremonilleux B. Treatment of missing values for

association[C]// Research and Development in Knowledge Discovery and Data Mining. New York: Springer-Verlag, 1998:258-269.

- [3] Kononenko I, Bratko I. Experiment in automatic learning of medical diagnostic rules[R]. Jozef Stefan Institute, 1984.
- [4] Quinlan J R. C4. 5: Programs for machine learning[M]. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1993.
- [5] Lakshminarayanan K, Harp S, Goldman R, et al. Imputation of missing data using machine learning techniques[C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1996:140-145.
- [6] 刘峰,翟俊.基于聚类分析和神经网络的时间序列预测方法[J].微电子学与计算机,2006,23(9):85-87.
- [7] 王清毅,蔡智,邹翔.部分数据缺失环境下的发现方法[J].软件学报,2001,12(10):1516-1524.
- [8] 张彩明.高精度三次参数样条曲线的构造[J].计算机学报,2002,25(3):262-268.

作者简介:

张春生 男,(1965-),硕士,教授.研究方向为数据库技术、数据挖掘、应用软件开发.

(上接第 195 页)

参考文献:

- [1] Chen Lei-Da. A model of consumer acceptance of mobile payment [J]. International Journal of Mobile Communications, 2008, 6(1):32-52.
- [2] 许芸.在电子商务数据管理中分布式数据库的安全性策略[J].微电子学与计算机,2002,19(9):13-15.
- [3] 王潇雨.移动支付安全交易平台的研究与开发[J].计算机工程与设计,2006(11):213-216.
- [4] 施小华,王志坚.移动支付的安全方案研究[J].计算机技术与发展,2007(4):114-118.
- [5] 张慧,章丽玲.移动支付及其安全技术研究[J].湖北第二师范学院学报,2008(8):52-55.

- [6] 湛玉婕,张必春.C2C 电子商务诚信问题改良初探[J].经济师,2009(1):254-255.

作者简介:

张晓杰 男,(1982-),硕士研究生.研究方向为管理信息系统与计算机网络.

卫红春 男,(1956-),教授.研究方向为管理信息系统与软件工程.

王义海 男,(1958-),高工.研究方向为信息系统安全技术.

雷卓琳 女,(1975-),工程师.研究方向信息系统工程.