

复杂系统多因素估计函数分析及其应用

陈红英¹, 王 涛¹, 毛革非²

(1 华南师范大学 计算机学院, 广东 广州 510631; 2 广州亿程交通信息有限公司, 广东 广州 510630)

摘 要: 复杂的系统通常有多因素与目标相关, 这些因素中哪些与目标最相关, 以及相关因素与目标之间的关系对复杂系统的分析很重要. 首先采用基于信息熵的方法求取最关键因素, 然后对训练集进行聚类, 最后采用拉格朗日多项式定义关键因素与目标之间的函数关系, 并将该方法应用于智能交通系统中的行车油量分析系统, 取得了较好的效果.

关键词: 信息熵; 聚类; 多项式插值

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2012)06-0068-03

Complicated System Factor's Function Analysis and Application

CHEN Hong-ying¹, WANG Tao¹, MAO Ge-fei²

(1 Institute of Computer, South China Normal University, Guangzhou 510631, China;

2 YiCheng Traffic Information Corporation, Guangzhou 510630, China)

Abstract: Complicated system usually has many factors that have relation with result. Which is the most important factor, and what is the relation between these factors with result. In this paper these most important factors are found by method of information entropy, then training set is clustered, finally Lagrange interpolation method is used to find the function between important factors and result. In the end of this paper, this method is used in vehicle oil quantity analysis system, and gets satisfying effect.

Key words: information entropy; cluster; lagrange interpolation method

1 引言

复杂的系统通常有多因素与目标相关, 如何发现这些因素中哪些与目标最相关, 相关因素与目标之间是否有函数关系, 如果有能否求得函数关系是什么, 这些问题对复杂系统的分析很重要. “信息熵”是 1948 年由香农提出的概念, 它用来度量信息量的多少. 香农指出, 任何信息都存在冗余, 冗余大小与信息中每个符号(数字、字母或单词)的出现概率或者说不确定性有关. 他借鉴了热力学的概念, 把信息中排除了冗余后的平均信息量称为“信息熵”, 并给出了计算信息熵的数学表达式. 文献[1]针对传统 K-means 算法存在的问题, 提出一种基于信息熵的精确属性赋权 K-means 聚类算法来提高聚类的精

确度. 文献[2]提出采用信息熵对连续属性离散化方法. 文献[3]采用信息熵定义知识量函数. 本文通过求取系统信息的信息熵, 找到信息量最大的因素. 为了求因素和目标之间的函数关系, 本文采用拉格朗日多项式插值, 文献[4]提出了拉格朗日乘子方法, 该方法可实现并具有全局收敛性. 为了降低插值多项式的阶数, 我们采用了层次聚类方法.

2 系统原理

本文首先采用基于信息熵的方法求取最关键因素, 然后对训练集进行聚类, 最后采用拉格朗日多项式定义关键因素与目标之间的函数关系, 最后将该方法应用于智能交通系统中行车油量分析系统, 实践证明取得了较好的效果.

收稿日期: 2011-09-29; **修回日期:** 2011-11-24

基金项目: 广东省自然科学基金项目(9451063101002238)

2.1 最关键因素的求取方法

2.1.1 训练数据标准化

我们将复杂系统各个因素用数据集定义如下:

定义 1 训练集中的因素定义如下: $(x_1, x_2, \dots, x_i, \dots, x_n, y)$, 其中 x_i 是第 i 个因素, y 为目标.

定义 2 设训练集有 S 个训练实例, 则训练实例集合定义为 $\{(x_1^{(1)}, x_2^{(1)}, \dots, x_i^{(1)}, \dots, x_n^{(1)}, y^{(1)}), (x_1^{(2)}, x_2^{(2)}, \dots, x_i^{(2)}, \dots, x_n^{(2)}, y^{(2)}), (x_1^{(3)}, x_2^{(3)}, \dots, x_i^{(3)}, \dots, x_n^{(3)}, y^{(3)}), (x_1^{(s)}, x_2^{(s)}, \dots, x_i^{(s)}, \dots, x_n^{(s)}, y^{(s)})\}$, 其中 $x_i^{(j)}$ 表示因素 x_i 的第 j 个实例的值.

定义 3 一个 X 值域为 $\{x_1, \dots, x_n\}$ 的随机变量的熵值 H 定义为

$$H(X) = E(I(X)) \quad (1)$$

式中, E 代表了期望函数, 而 $I(X)$ 是 X 的信息量. $I(X)$ 本身是个随机变量. 如果 p 代表了 X 的机率函数, 则信息熵的公式可以表示为

$$H(X) = \sum_{i=1}^n p(x_i) I(x_i) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (2)$$

根据定义 3, 本系统中训练实例集合中的 $x_i^{(j)}$ 大小不一, 需要映射到 $p(x_i)$ (其中 $p(x_i) \leq 1$, 且 $\sum_{i=1}^n p(x_i) = 1$), 因此, 需要将训练实例数据进行标准化.

具体步骤如下:

(1) 对训练集进行归一化

$$z_i^{(j)} = x_i^{(j)} / y^{(j)} \quad (j=1, \dots, s, i=1, \dots, n) \quad (3)$$

训练集变为 $\{(z_1^{(1)}, z_2^{(1)}, \dots, z_i^{(1)}, \dots, z_n^{(1)}, y^{(1)}), (z_1^{(2)}, z_2^{(2)}, \dots, z_i^{(2)}, \dots, z_n^{(2)}, y^{(2)}), (z_1^{(3)}, z_2^{(3)}, \dots, z_i^{(3)}, \dots, z_n^{(3)}, y^{(3)}), (z_1^{(s)}, z_2^{(s)}, \dots, z_i^{(s)}, \dots, z_n^{(s)}, y^{(s)})\}$, 其中 $y^{(1)} = y^{(2)} = \dots = y^{(s)} = 1$.

$$W_i = \sum_{j=1}^s z_i^{(j)} \quad (4)$$

(2) 训练集取

$$h_i^{(j)} = z_i^{(j)} / w_i \quad (5)$$

经过参数标准化后, $\sum_{j=1}^s h_i^{(j)} = 1$, 且 $y^{(1)} = y^{(2)} = \dots = y^{(s)} = 1$, 满足式(2)的要求.

2.1.2 利用信息熵计算因素重要度

最终求的属性(因素)信息熵 i 的信息量

$$H(h_i) = \sum_{j=1}^s h_i^{(j)} \log_2 h_i^{(j)} \quad (6)$$

定义 4 以式(5)计算的 k 个最大信息熵因素作为关键因素, 训练集标记为 $\{g_1^{(j)}, g_2^{(j)}, \dots, g_i^{(j)}, \dots,$

$g_k^{(j)}\} (j=1, \dots, s)$.

2.2 层次聚类

下面我们用层次聚类算法 single-linkage 对训练集进行聚类. single-linkage 每次将两个旧类合并成一个新类, 直到最终合并成一个类为止. 每合并一次, 则在距离矩阵中删除相对应的行与列.

定义 5 将待聚类的 s 个实例分别标号为 $0, 1, \dots, s, D=[d(i, j)]$ 表示对应的 $s \times s$ 距离矩阵. $L(k)$ 表示第 k 个类所处的层次, 由对象 m 构成的类记为 (m) , 类 (r) 与类 (s) 的距离记为 $d[(r), (s)]$.

定义 6 两个类的相似度用它们对应向量的夹角的余弦值来表示, 定义 4 求得的数据集中任意两个实例类 (i) 和实例类 (j) 的向量分别为 $(g_1^{(i)}, g_2^{(i)}, \dots, h_i^{(i)}, \dots, g_k^{(i)}, y^{(i)})$ 和 $(g_1^{(j)}, g_2^{(j)}, \dots, g_i^{(j)}, \dots, g_k^{(j)}, y^{(j)})$, 它们之间的距离如下:

$$d[(i), (j)] = \frac{\sum_{m=1}^k g_m^{(i)} \times g_m^{(j)}}{\sqrt{\left(\sum_{m=1}^k g_m^{(i)2}\right) \left(\sum_{m=1}^k g_m^{(j)2}\right)}} \quad (7)$$

single-linkage 聚类算法过程如下:

(1) 初始时共有 s 个类, 每个类由一个实例类成. 令顺序号 $t=0, L(t)=0$.

(2) 利用式(7)在 D 中寻找最小距离 $d[(r), (s)] = \min d[(i), (j)]$.

(3) 将两个类 (r) 和 (s) 合并成一个新类 (r, s) ; 令 $t=t+1, L(t)=d[(r), (s)]$.

(4) 更新距离矩阵 D : 将表示类 (r) 和类 (s) 的行列删除, 同时加入表示新类 (r, s) 的行列; 定义新类 (r, s) 与各旧类 (k) 的距离为 $d[(k), (r, s)] = \min d[(k), (r)], d[(k), (s)]$.

(5) 反复步骤(2)~(4), 直到所有对象合并成一个类为止.

2.3 多项式插值

根据定义 4, k 个最大信息熵因素作为最关键因素, 需要计算 k 个函数与之对应似的 $f_i(g_i^{(j)}) = y_i (j=1, \dots, s, i=1, \dots, k)$ 简化为 $f_i(x) = y_i, i \in \{0, 1, 2, \dots, k\}$.

定义 7 插值多项式 $f_i(x)$ 的形式为

$$f_i(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x + a_0 \quad (8)$$

定理 1 如果 $x_0, x_1, \dots, x_n$ 是互异的实数, 那么对于任意的实数 $y_0, y_1, \dots, y_n$, 都存在唯一的至多 n 次多项式 $f_i(x) = y_j (j=0, 1, \dots, n)$.

证明: $f_i(x)$ 是式(8)的形式, 就得到系数为 a_k

的线性方程系统,用矩阵向量形式表示为

$$\mathbf{X} = \begin{bmatrix} x_0^n & x_0^{n-1} & x_0^{n-2} & \cdots & x_0 & 1 \\ x_1^n & x_1^{n-1} & x_1^{n-2} & \cdots & x_1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^n & x_n^{n-1} & x_n^{n-2} & \cdots & x_n & 1 \end{bmatrix}_{(n+1) \times (n+1)},$$

$$\mathbf{a} = \begin{bmatrix} a_n \\ a_{n-1} \\ \cdots \\ a_0 \end{bmatrix}_{(n+1) \times 1}, \quad \mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \cdots \\ y_n \end{bmatrix}_{(n+1) \times 1} \quad (9)$$

为了构建插值多项式 $f_i(x)$,我们要解这个系统计算系数 a . 左侧的矩阵 $\mathbf{X}$ 通常叫作范德蒙矩阵, 由于 $x_0, x_1, \cdots, x_n$ 是互异的实数, 它的行列式 $\det(\mathbf{X})$ 不为零. 于是对于任意的实数 $y_0, y_1, \cdots, y_n$, 式(9)有唯一解 $\mathbf{a} = \mathbf{X}^{-1} \mathbf{y}$.

不失一般性求 $f_1(g_i^{(j)}) = y_i$ 的函数 $f_1(x) = y_i$, 其中 $x = g_i^{(j)} (j=1, \cdots, s)$.

由于实例个数为 s , 可知式(8)的阶数为 s , 通常实例个数数目巨大, 采用多项式插值不现实. 2.2 节中我们采用聚类的方法将类似的实例聚类到不同的类, 在带入式(8)时只需要从每个类中选择一个实例即可. 如果聚类后的类别个数为 m , 则多项式的阶数只要有 m 阶即可 (其中 $m \ll S$), 这种方法使得拉格朗日插值算法具有了可行性^[6].

3 系统应用

智能交通系统(ITS)是将先进的信息技术、通讯技术、传感技术、控制技术以及计算机技术等有效地集成运用于整个道路交通管理体系. 我们采用上述方法对系统中参数之间的关系进行分析, 可以分析的信息有很多种, 每种分析依赖的公用数据集为卫星定位系统采集的数据: 速度、加速度、时间、方向、经度、纬度、里程、用油量、刹车、ACC、发动机、温度、车门、其他 I/O 量. 除此以外, 不同的分析系统还有不同的额外数据集, 例如:

(1) 油量分析系统

额外数据集: 无;

分析目标: 分析油量与各个因素之间的关系.

(2) 路段分析

额外数据集: 车辆在高速、国道、省道、市区主干道、市区次干道的速度;

分析目标: 不同路段限速阈值、行驶速度.

(3) 行车路况分析

额外数据集: 坡道、颠簸程度、路面材质、道路拥

堵情况;

分析目标: 车辆保养、行驶速度、油量消耗.

(4) 行车风险分析

额外数据集: 超速、疲劳驾驶、急转弯、急刹车 (速度较高时刹车) 次数;

分析目标: 事故发生原因.

(5) 行为习惯分析

额外数据集: 启动预热不足够、急转弯、急刹车、停车空调用量过度、加油地点、减油地点;

分析目标: 车况、行车风险、油量消耗.

不失一般性, 进行油量分析, 得到与油量最相关的因素是加速度、速度、温度差 (空调温度和室外温度之差), 它们与油量之间的关系通过函数插值得到如图 1 所示的函数曲线.

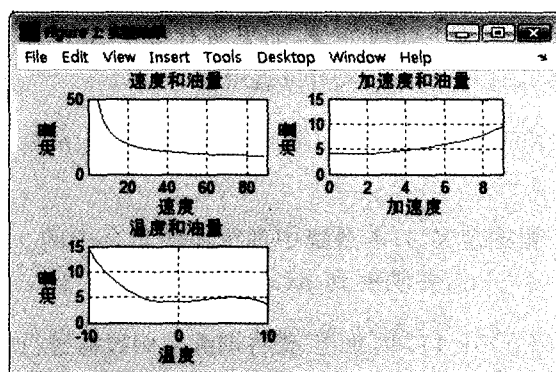


图 1 油量与加速度、速度、温度差函数

为了验证函数的正确性, 我们将一部分数据用作训练, 另一部分数据用于测试, 测试结果: 速度和油量函数匹配率为 85%~91%, 加速度和油量匹配率 88%~90%, 温度和油量 82%~88%, 具体匹配率与车辆有关, 同一种车辆的模型匹配率较高.

同理, 行车风险分析、路段分析、行车路况分析、行为习惯分析, 也可得到相关分析函数, 在此不一一例举.

4 结束语

本文对复杂系统的多因素估计函数^[6-7]进行了讨论, 主要有以下创新:

(1) 在采用信息熵作为关键因素的度量值时将训练数据标准化, 提供了将一般数据用于信息熵计算的方法.

(2) 拉格朗日插值算法通常存在阶数太高的问题, 限制了其实际应用, 本文在进行拉格朗日插值前, 利用聚类的方法降低了拉格朗日插值算法的阶数, 使其具有可行性.

(下转第 75 页)

不规则度相同的情况下,EFL算法都能够提供较好的定位服务.下一步的工作是研究运动模型和动态特性对算法健壮性、网络能耗等的影响,以使算法在更加真实的环境中表现出较好的定位效果.

参考文献:

- [1] 王福豹,史龙,任丰原.无线传感器网络中的自身定位系统和算法[J].软件学报,2005,16(5):857-868.
- [2] Isaac Amundson, Xenofon D. Koutsoukos. A survey on localization for mobile wireless sensor networks[C] // the 2nd International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments. FL, USA, 2009: 235-254.
- [3] 周彤.移动传感器网络节点部署及自定位技术研究[D].哈尔滨:哈尔滨工业大学,2008.
- [4] Lingxuan Hu, David Evans. Localization for mobile sensor networks[C] // Proceedings of ACM MobiCom. Philadelphia, Pennsylvania, USA, 2004: 45-57.
- [5] Martins M H T, Hongyang Chen, Sezaki K. OT-MCL: orientation tracking-based monte carlo localization for mobile sensor networks[C] // Proceedings of

Networked Sensing Systems (INSS). PA, USA, 2009: 1-8.

- [6] Guodong Teng, Kougen Zheng, Wei Dong. MA-MCL: mobile-assisted monte carlo localization for wireless sensor networks[J]. International Journal of Distributed Sensor Networks. 2011:671-814.
- [7] Savvides A, park H, Srivastava M B. The bits and flops of the n-hop multilateration primitive for node localization problem[C] // Proceedings of ACM WSNA. NY, USA, 2002:112-121.

作者简介:



杨雪女,(1986—),硕士研究生.研究方向为无线传感器网络节点定位技术.

王辉女,(1962—),博士,教授,硕士生导师.研究方向有光传输理论与系统、信号处理与控制技术.

(上接第70页)

(3) 利用了训练集中数据的特点,层次聚类算法中将关键因素的向量距离作为类之间的距离.

系统最终应用于智能交通中的行车油量分析系统,获得良好的效果.

参考文献:

- [1] 原福永,张晓彩,罗思标.基于信息熵的精确属性赋权 K-means 聚类算法[J].计算机应用,2011,31(6):1675-1677.
- [2] 高建国,崔业勤.基于信息熵理论的连续属性离散化方法[J].微电子学与计算机,2011,28(7):187-189.
- [3] 朱莉莉,严洪森,崔鹏飞,等.基于信息熵的知识网度量方法及应用[J].东南大学学报:自然科学版,2010,40(5):1097-1102.
- [4] 濮定国,金中.新的拉格朗日乘子方法[J].同济大学

学报(自然科学版),2010,38(8):1387-1391.

- [5] 毛韶阳,李肯立.一种基因数据的聚类并行算法研究[J].微电子学与计算机,2007,24(9):130-133.
- [6] 彭新俊,王翼飞.快速原空间孪生支持向量回归算法[J].模式识别与人工智能,2011,24(1):22-29.
- [7] 岳昊.三划分问题可多项式归约为唯一可达向量 Petri 网可达性问题[J].微电子学与计算机,2008,25(10):144-146.

作者简介:

陈红英女,(1969—),博士,副教授.研究方向 GIS、人工智能.

王涛男,(1985—),硕士研究生.研究方向语义网络、人工智能.

毛革非男,(1965—),高级工程师.