

基于属性重要性的不完备数据填充算法

陈志奎, 吕爱玲, 张清辰

(大连理工大学 软件学院, 辽宁 大连 116621)

摘 要: 现有的不完备数据填充算法对所有缺失数据采用统一方式填充, 没有考虑数据的重要性, 效率低, 实时性差. 因此, 本文提出一种基于属性重要性的不完全数据填充算法. 通过差分矩阵求得属性约简, 根据约简区分重要属性和非重要属性, 对于重要属性数据填充采用改进的马氏距离填充方法, 而不重要属性数据填充采用相似度概率填充方法, 保证了数据精确度的同时, 提高了实时性, 具有实用性. 最后, 实验部分采用数据家庭系统数据和 UCI 标准数据集分别对算法性能进行了分析, 验证了该算法的优越性.

关键词: 不完备系统, 数据填充, 马氏距离, 属性约简

中图分类号: TP315

文献标识码: A

文章编号: 1000-7180(2013)07-0167-06

A New Algorithm for Imputing Missing Data Based on Distinguishing the Importance of Attributes

CHEN Zhi-kui, LV Ai-ling, ZHANG Qing-chen

(School of Software, Dalian University of Technology, Dalian 116621, China)

Abstract: Existing incomplete data filling algorithm are all use the same method to fill all the missing values, and did not consider the importance of each value, thus, makes all algorithms low efficiency and poor real-time. Therefore, this paper proposes a new data filling algorithm based on distinguishing the importance of attributes, it uses attribute reduction to distinguish important attributes and unimportant attributes, then, uses the improved mahalanobis-based algorithm to imputing the missing value that belong to the important attributes, and unimportant missing values according to the similarity-probabilistic method, thus, ensure that the accuracy of data, at the same time, make sure the real-time and practicality. at last, the experimental part using the Digital-home system and the UCI standard datasets to analysis the algorithm performance, verifying the superiority of the algorithm.

Key words: incomplete system; data filling; mahalanobis distance; attribute reduction

1 引言

随着信息技术的不断发展, 物联网应运而生. 物联网的出现实现了物物之间的“互联互通”, 从而极大地提高了人类生活的舒适性和娱乐性. 为了了解物体在工作状态, 就不得不利用传感器来采集状态信息, 以方便检测和控制. 然而, 由于传感器节点失效、网络不稳定、数据传输拥塞等一些因素, 造成了物联网系统中常常包含大量缺失数据. 为了保证数

据分析和梳理的正确性和有效性, 确保服务质量, 对缺失数据进行正确快速地填充将是一个非常有必要的预处理过程. 例如, 在数字家庭系统中, 安防子系统是必不可少的模块, 它是用户家庭的实时安全保护屏障, 因此, 必须要保证该子系统数据的完全准确性和实时性; 然而, 对于光照、湿度等信息, 并不需要完全的准确, 允许数据有小幅误差. 因此, 在数字家庭系统中, 就需要对缺失数据进行区分填充, 对安防采集的数据需要精确填充, 而对于湿度、光照等信息

收稿日期: 2012-09-02; **修回日期:** 2012-10-11

基金项目: 大连市科技局科技计划项目(2011A17GX076)

就可以相对粗略的填充方法. 这样才能在确保准确性的前提下, 提高数据处理效率, 满足实时性要求.

目前存在许多不完备数据填充算法, 主要包括: 基于马氏距离的不完备数据填充算法、基于决策树的不完整数据的数据填充算法^[1-2]、基于 EM 和贝叶斯网络的丢失数据填充算法等^[3-5]. 其中, 最邻近方法的填充效果最好^[6], 基于马氏距离的不完备数据填充就是一种最邻近方法的精确数据填充算法. 然而, 该算法对于所有数据采用统一的填充策略, 计算量大, 实时性差. 因此, 针对不完备信息系统中的缺失值填充问题, 本文提出了一种基于属性重要性的不完全数据填充算法. 首先通过差别矩阵求得属性约简, 根据约简区分重要属性和非重要属性, 对于重要属性数据填充采用改进的马氏距离填充方法填充, 而不重要属性数据填充采用基于相似度概率方法填充, 有效地保证了数据填充精确性的同时, 提高了数据填充效率.

2 相关理论

定义 1 一个决策表 $S = \langle U, C, D, V, f \rangle$, 其中, U 表示对象的非空有限集合, 称为论域; C 表示条件属性的非空有限集, D 表示决策属性的非空有限集, $C \cap D = \emptyset$; $V = \bigcup_{a \in C \cup D} V_a$, V_a 是属性 a 的值域; $f: U \times (C \cup D) \rightarrow V$ 是一个信息函数, 它为每个对象赋予一个信息值, 即 $\forall a \in C \cup D, x \in U$, 有 $f(x, a) \in V_a$. 若条件属性取值有遗漏或缺失, 记遗漏属性值为 *, 称信息系统 S 为不完备信息系统, 否则称为完备信息系统^[7].

定义 2 设 $|U| = n$, 则决策表 S 的差别矩阵 M^C , 其任一元素为:

$$m_{ij} = \begin{cases} \{c_l \in C: f(x_i, c_l) \neq f(x_j, c_l)\}, f(x_i, D) \neq f(x_j, D) \\ \emptyset, \text{ 其他} \end{cases}$$

定义 3 决策表 $S = \langle U, C, D, V, f \rangle$, 设 A 是 U 上的等价关系, 则对属性集 R 上的不可区分关系定义为: $IND(R) = \{(x, y) \in U \times U \mid \forall a \in R, f(x, a) = f(y, a)\}$, 若 $r \in R$, 并且 $IND(R) = IND(R - \{r\})$, 则称 r 是 R 中不必要的; 否则称 r 是必要的. 如果每一个 $r \in R$ 都是 R 中必要的, 则称 R 是独立的, 否则称 R 为依赖的. 设 $P \subseteq R$, 若 P 是独立的, 且 $IND(P) = IND(R)$, 则称 P 为 R 的一个约简.

定义 4 对于一个均值为 $\mu = (\mu_1, \mu_2, \dots, \mu_p)$, 协方差矩阵为 $\sum$ 的多变量向量 $x = (x_1, x_2, \dots,$

$x_p)$, 其马氏距离为^[3]:

$$D_M(x) = \sqrt{(x - \mu)^T \sum^{-1} (x - \mu)}$$

马氏距离也可以定义为两个服从同一分布, 并且其协方差矩阵为 $\sum$ 的随机变量 $\vec{x}$ 与 $\vec{y}$ 的差异程度:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \sum^{-1} (\vec{x} - \vec{y})}$$

定义 5 传统的基于马氏距离的数据填充算法步骤如下:

(1) 计算目标对象 g 到其他对象 x_i 的马氏距离: $d_i = \sqrt{(g - x_i)^T \sum^{-1} (g - x_i)^T}$;

(2) 马氏距离单位化: $p_i = \frac{d_i}{\sum_{i=1}^k d_i}$, $i = 1, 2, \dots, k$;

(3) 计算第 i 个邻居对象的熵值: $h_i = -(Ink)^{-1} p_i \ln p_i$;

(4) 计算每个邻居的变异程度系数: $v_i = 1 - h_i$;

(5) 计算加权系数: $\omega_i = \frac{1}{k-1} \left[1 - \frac{v_i}{\sum_{i=1}^k v_i} \right]$;

(6) 计算预测值: $\tilde{g} = \sum_{i=1}^k \omega_i \times x_i$, $i = 1, 2, \dots, k$.

定义 6 $\forall \emptyset \subset B \subseteq C$, 确定了 U 上的一个二元关系

$$SIM(B) = \{(x, y) \in U \times U \mid \forall a \in B, f(x, a) = f(y, a) \vee f(x, a) = * \vee f(y, a) = *\}$$

记 $S_B(x) = \{y \in U \mid (x, y) \in SIM(B)\}$, 称 $S_B(x)$ 为 x 在 B 上的相容类.

定义 7 基于相似度概率的数据填充规则定义为: 设 $a \in C, x, y, y_1 \in U, f(x, a) = *, y \in SIM(x)$, 则 $f(x, a)$ 的取值为相容类中出现频率最大的那个值(不包含 * 值情况下), 公式表示如下:

$$f(x, a) = \{f(y_1, a): f(y_1, a) \neq * \wedge P_{f(y_1, a)} = \max(P_{f(y, a)})\}$$

其中, $P_{f(y, a)}$ 表示 $f(y, a)$ 在属性 a 上出现的次数, 即频率值.

3 基于属性重要性的不完备数据填充算法

传统的不完备数据填充算法使用统一方法对缺省数据进行填充, 没有考虑数据的重要程度. 若所有数据都精确填充, 则时间复杂度高; 若所有数据都粗

略填充,则会影响数据的精确性.针对以上问题,本文提出了一种基于属性重要性的不完备数据填充算法,首先通过差别矩阵求得属性约简,根据属性约简区分重要属性和非重要属性,然后,对于重要数据采用基于马氏距离不完备数据填充方法,能够实现精确填充;对于非重要数据采用定义6给出的基于相似度最大出现概率的数据填充方法,实现快速填充.

3.1 区分重要属性和非重要属性

该算法用于区分信息系统中的重要属性和非重要属性.首先通过差别矩阵求得属性约简,选取未知值最少的那个约简为最终求得的最小约简,则最小约简所包含的所有属性均为重要属性,其余的属性为非重要属性.算法如下:

输入:不完备决策系统 $S = \langle U, C, D, V, f \rangle$

输出:重要属性集 IMP , 非重要属性集 $UNIMP$

算法:

(1) 初始化数组 $IMP, UNIMP$ 为 $NULL$, 数组 $COND$ 初始化为所有条件属性数组;

(2) 对系统 S 中的缺省值“*”进行粗略填充,对每一个缺省值采用该属性列中所有精确值的平均

值进行粗略填充, $f(x, a) = \frac{\sum_{i=r}^{i=t} f(x_i, a)}{t - r + 1}$, $f(x_i, a) \neq *$;

(3) 根据定义2构造差别矩阵 M^C ;

(4) 求得区分函数 $f = \prod_{(x_i, x_j) \in U \times U} \sum m_{ij}$

(5) 化简区分矩阵,变成所有属性合式的析取式,并计算每个合式中属性所含缺省值的个数,取缺省值个数最小的合式的每个属性到数组 IMP 中;同时,将在 $COND$ 中但不在 IMP 中的所有元素添加到 $UNIMP$;

(6) 输出 IMP 和 $UNIMP$, 算法结束.

3.2 基于改进的马氏距离数据填充算法

基于马氏距离的数据填充算法,是一种最邻近数据填充算法,通过马氏距离来定义对象间的相近程度,然后采用信息论中的熵值概念计算最邻近距离的加权系数,即每个对象对于预测值的影响因子;最后根据相关系数来确定预测值.信息熵是信息论中用于度量信息量的一个概念,是描述系统混乱的量,熵越大说明系统越混乱,所携带的信息量就越多;熵越小说明系统越有序,携带的信息也就越多.和其他数据填充算法比较,基于马氏距离的数据填

充算法能更好地描述对象间的相似性,更准确的填补缺失值.

文献[8]中的“熵”反映了对象之间的变异程度,即熵值的大小与变异程度相反原则;由 $\omega_i = \frac{1}{k-1} \left[1 - \frac{v_i}{\sum_{i=1}^k v_i} \right]$ 可知,当对象的变异程度越大,其加权系数越小;由此可得,熵值越大的对象,其加权系数越大,反之则越小;同时,根据最近邻方法可知,马氏距离越小的对象,其加权系数应该越大,这样得到的预测值才越准确.因此,当某个对象和目标对象的马氏距离越小时,熵值越大,加权系数就越大,从而预测值就越准确,即马氏距离和熵值大小应该存在单一的递减关系.然而文献[8]中的熵值计算公式 $h_i = -(Ink)^{-1} p_i \ln p_i$ 中,马氏距离和熵值大小并不存在单一的递减关系,当 $p_i \geq \frac{1}{e}$ ($e \approx 2.7183$)

时,熵值计算是马氏距离的增函数,即马氏距离越大的对象,其熵值也越大,从而相关系数也越大,这就导致了不相关数据对预测数据准确性的严重影响.如表1的不完备信息系统 U ,使用传统的基于马氏距离的数据填充算法,计算 x_4 到各对象的马氏距离排序为: $d_{45} < d_{43} < d_{42} < d_{41} < d_{46}$, 然后计算权值系数排序为: $\omega_{45} < \omega_{43} < \omega_{42} < \omega_{41} < \omega_{46}$, 使得距离最远的对象的权值系数最大,从而严重影响了最后的预测结果.

基于以上问题,本文提出了一种改进的基于马氏距离的数据填充算法,在传统的马氏距离数据填充算法的基础上,改进了熵值的计算方法,使得熵值大小和马氏距离之间存在单一的递减关系,这样减小了不相关数据对预测值的影响.算法具体步骤如下:

(1) 计算目标对象 g 到其他对象 x_i 的马氏距离:

$$d_i = \sqrt{(g - x_i) \sum^{-1} (g - x_i)^T};$$

(2) 马氏距离单位化:

$$p_i = \frac{d_i}{\sum_{i=1}^k d_i}, \quad i = 1, 2, \dots, k;$$

(3) 根据单位化的马氏距离,计算第 i 个邻居对象的熵值:

$$h_i = -(Ink)^{-1} \ln \left(p_i + \frac{1}{k} \right);$$

(4) 计算加权系数:

$$\omega_i = \frac{1}{k^{-1}} \left[1 - \frac{1 - h_i}{\sum_{i=1}^k (1 - h_i)} \right];$$

(5) 计算预测值:

$$\tilde{g} = \sum_{i=1}^k \omega_i x_i, \quad i = 1, 2, \dots, k;$$

(6) 预测值域化. 由于步骤 5 得到的预测值 $\tilde{g}$ 有可能不在对应的值域 V_a 内, 因此需要对预测值进行域化: $f(g, a) = \min\{|x - \tilde{g}|\}, x \in V_a$, 即选择值域 V_a 中离 $\tilde{g}$ 最近的即为最终的填充值.

根据传统的基于马氏距离的数据填充算法, 首先计算目标对象到其他对象间的马氏距离, 并将其单位化, 然后根据波尔兹曼的熵的统计学定义, 改进的熵值计算公式为 $h_i = -(Ink)^{-1} \ln(p_i + \frac{1}{k})$, h_i 为 p_i 的严格单调递增函数, 同时满足 $0 < h_i < 1$. 这样, 当对象间的马氏距离越小, 则计算的熵值 h_i 越大, 表明其所携带的知识越丰富, 其加权系数 ω_i 也越大, 预测值越准确.

3.3 基于相似度概率的缺失值填充算法

概率填充通过数据间的相似度定义来填充缺失数据, 具体算法步骤如下:

(1) 计算每个缺省数据 $x(x \in U)$ 的相容类 $S_C(x)$: $S_C(x) = \{y \in U \mid (x, y) \in SIM(C)\}$;

(2) 根据相似度概率填充定义, 计算每个缺省数据的值. 若计算结果有多个概率最大的值, 则随机选取一个作为填充值; 若相容类中的所有值均为缺省值, 则取值为: $f(x, a) = \{f(y_1, a) : \forall y \in U, f(y_1, a) \neq * \wedge P_{f(y_1, a)} = \max(P_{f(y, a)})\}$.

4 实验分析

4.1 与传统马氏距离填充算法比较

表 1 为数字家庭系统不完备数据样本 U , Rm * 代表传感器节点, C 为条件属性, D 为决策属性, * 代表缺失数据. 其中, temp 为采集的温度信息, humi 为湿度信息, lumi 为光照强度信息, power 为节点剩余能量值, sensitivity 为决策值.

表 1 数据家庭不完备数据样本

U	C				D
	t(temp)	h(humi)	l(lumi)	p(power)	s(sensitivity)
Rm1	28	41	200	31.4	3
Rm2	28	41	200	87.1	3
Rm3	29	42	*	43.5	2
Rm4	27.5	*	170	29.4	2
Rm5	29	42	*	40.1	1
Rm6	29	41	190	31.4	3
Rm7	28	42	170	29.4	2
Rm8	29	41	190	43.5	2
Rm9	28	42	180	32.7	1
Rm10	29	41	180	32.7	1

4.1.1 基于传统的马氏距离数据填充

(1) 首先对不完备信息系统 U 进行完备化, 对缺失值“*”进行均值填充, 则 $f(Rm_3, l) = f(Rm_5, l) = 185$, $f(Rm_4, h) = 41$;

(2) 计算属性间的协方差矩阵 Σ ;

(3) 对 $f(Rm_3, l)$ 进行填充: 根据 Σ 计算 Rm_i 和 Rm_3 的马氏距离分别为: $d_{1,3} = 3.9678$, $d_{2,3} = 3.3172$, $d_{4,3} = 3.3846$, $d_{5,3} = 0.2452$, $d_{6,3} = 2.4053$, $d_{7,3} = 1.8936$, $d_{8,3} = 2.0201$, $d_{9,3} = 2.1070$, $d_{10,3} = 2.2238$; 根据 $p_i = \frac{d_{i,3}}{\sum_{i=1}^9 d_{i,3}}$, $h_i = -(In9)^{-1} p_i \ln p_i$, ω_i

$$= \frac{1}{8} \left[1 - \frac{1 - h_i}{\sum_{i=1}^9 (1 - h_i)} \right], \text{ 得到 } \omega_1 = 0.1117, \omega_2 = 0.1115, \omega_3 = 0.1115, \omega_4 = 0.1098, \omega_5 = 0.1112, \omega_6 = 0.1110, \omega_7 = 0.1110, \omega_8 = 0.1111, \omega_9 = 0.1111; \text{ 最后得到预算值 } \tilde{g} = \sum_{i=1}^9 \omega_i \times f(Rm_i, l) = 184.4614,$$

因为 $f(Rm_i, l) \in \{170, 180, 190, 200\}$, 所以 $f(Rm_3, l) = \tilde{g} \approx 180$;

(4) 按照步骤 3, 对 $f(Rm_4, h)$ 和 $f(Rm_5, l)$ 进行填充. 填充结果为: $f(Rm_4, h) = \tilde{g} = 41.4443 \approx 41$, $f(Rm_5, l) = \tilde{g} = 184.4531 \approx 180$;

(5) 最终填充后的完备数字家庭信息系统如表 2 所示.

表 2 基于传统马氏距离填充后的完备信息系统

U	C				D
	t(temp)	h(humi)	l(lumi)	p(power)	s(sensitivity)
Rm1	28	41	200	31.4	3
Rm2	28	41	200	87.1	3
Rm3	29	42	180	43.5	2
Rm4	27.5	41	170	29.4	2
Rm5	29	42	180	40.1	1
Rm6	29	41	190	31.4	3
Rm7	28	42	170	29.4	2
Rm8	29	41	190	43.5	2
Rm9	28	42	180	32.7	1
Rm10	29	41	180	32.7	1

4.1.2 基于属性重要性的数据填充

(1) 首先对不完备信息系统 U 进行完备化, 对缺失值“*”进行均值填充, 则 $f(Rm_3, l) = f(Rm_5, l) = 185$, $f(Rm_4, h) = 41$;

$l) = 185, f(Rm_4, h) = 41$;

(2) 然后根据差别矩阵求属性约简,得到重要属性和非重要属性,完备信息系统 U' 构造差别矩阵如表 3,由差别函数 $f = p \wedge (t \vee l) \wedge (h \vee l) \wedge (t \vee h)$ 可得, U' 的约简有 $\{t, h, l\}$ 、 $\{t, h, p\}$ 和 $\{t, l, p\}$,由于 $\{t, h, p\}$ 中包含的缺失值最少,因此,选择数据样本 U 的重要属性为 t, h, p ,非重要属性为 l ;

表 3 差别矩阵

	Rm1	Rm2	Rm3	Rm4	Rm5	Rm6	Rm7	Rm8	Rm9	Rm10
Rm1	∅									
Rm2	D	∅								
Rm3	Abcd	abcd	∅							
Rm4	Ac	acd	abcd	∅						
Rm5	Abcd	abcd	d	abcd	∅					
Rm6	Ac	acd	bcd	acd	bcd	∅				
Rm7	Bcd	bcd	acd	ab	acd	abcd	∅			
Rm8	Ac	acd	bc	acd	bcd	d	abcd	∅		
Rm9	Bcd	bcd	acd	abcd	acd	abcd	cd	abcd	∅	
Rm10	Ac	acd	bcd	acd	bcd	cd	abcd	cd	ab	∅

(3) 对 $f(Rm_4, h)$ 精确填充:计算属性间的协方差矩阵 $\sum$ 为:

$$\sum = \begin{bmatrix} 0.3583 & 0.0222 & 1.3889 & -0.4489 \\ 0.0222 & 0.2667 & -2.2222 & -1.6422 \\ 1.3889 & -2.2222 & 111.1111 & 104.7778 \\ -0.4489 & -1.6422 & 104.7778 & 302.4440 \end{bmatrix}$$

根据 $\sum$ 计算 Rm_i 和 Rm_4 的马氏距离分别为:
 $d_{1,4} = 3.6980, d_{2,4} = 3.7198, d_{3,4} = 3.5825, d_{5,4} = 3.5758, d_{6,4} = 3.0076, d_{7,4} = 2.1750, d_{8,4} = 2.8993, d_{9,4} = 2.7830, d_{10,4} = 2.5396$; 根据 $p_i = \frac{d_{i,4}}{\sum_{i=1}^9 d_{i,4}}, h_i = -(\ln 9)^{-1} \ln(p_i + \frac{1}{9}), \omega_i = \frac{1}{8} \left[1 - \frac{1-h_i}{\sum_{i=1}^9 (1-h_i)} \right]$, 得到 $\omega_1 = 0.1092, \omega_2 = 0.1091, \omega_3 = \omega_4 = 0.1096, \omega_5 = 0.1114, \omega_6 = 0.1143, \omega_7 = 0.1117, \omega_8 = 0.1121, \omega_9 = 0.1130$; 最后得到预算值 $\tilde{g} = \sum_{i=1}^9 \omega_i \times f(Rm_i, h) = 41.4456$, 因为 $f(Rm_i, h) \in \{41, 42\}$, 所以 $f(Rm_4, h) = \tilde{g} \approx 41$;

(4) 非重要数据 $f(Rm_3, l)$ 和 $f(Rm_5, l)$ 的数据填充:首先根据定义 5 求得, $S_c(Rm_3) = S_c(Rm_5) = \emptyset$,由定义 6 可知 $f(Rm_3, l) = f(Rm_5, l) = 180$;

(5) 数据填充之后的完备信息系统如表 4 所示.

表 4 基于属性重要性的数据填充算法

U	C				D
	t(temp)	h(humi)	l(lumi)	p(power)	s(sensitivity)
Rm1	28	41	200	31.4	3
Rm2	28	41	200	87.1	3
Rm3	29	42	180	43.5	2
Rm4	27.5	41	170	29.4	2
Rm5	29	42	180	40.1	1
Rm6	29	41	190	31.4	3
Rm7	28	42	170	29.4	2
Rm8	29	41	190	43.5	2
Rm9	28	42	180	32.7	1
Rm10	29	41	180	32.7	1

由表 2 和表 4 对比可知,两种算法对于缺失值的填充结果都相同.不完备数据样本 U 使用单一基于马氏距离的数据填充算法填充需要进行 3 轮 (3×9 次) 马氏距离计算,计算复杂,效率低;而基于属性重要性的数据填充算法,单纯对于重要属性数据使用马氏距离数据填充算法,实现了重要属性的精确数据填充;而非重要数据使用基于相似度的概率填充算法,只需要计算 1 轮马氏距离,提高了数据填充效率.因此,基于属性重要性的数据填充算法在保证数据填充精确性的同时,提高了效率,具有实用性.

4.1.3 算法复杂度分析

设不完备决策系统 $S = \langle U, C, D, V, f \rangle$, 缺失数据个数为 n ,其中重要数据 $n_1 (0 \leq n_1 \leq n)$ 个.

(1) 基于马氏距离的数据填充算法:根据文献 [9],进行 n 各缺失数据的填充,需要对 S 进行 n 轮马氏距离填充,每轮数据填充中,计算协方差矩阵复杂度为 $O(|U| |C|)$,预测值计算为 $O(|U| |D|)$,因此总的时间复杂度为 $O(n * |U| |C|)$.

(2) 基于属性重要性的数据填充算法中,应用改进的马氏距离进行数据填充的复杂度为 $O(n_1 * |U| |C|)$,采用相似度填充的算法复杂度为 $O((n - n_1) * |U|)$,所以总的时间复杂度为 $O((n - n_1 + n_1 |C|) * |U|), O((n - n_1 + n_1 |C|) * |U|) \leq O((n - n_1 |C| + n_1 |C|) * |U|) =$

$O(n * |U|) \leq O(n * |U| * |C|)$, 当 n_1 足够小于 n 时, 该算法能有效地提高算法运行效率.

令 $\eta = \frac{n}{n_1}$, 则马氏距离填充算法和基于属性重要性的填充算法的算法复杂度对比曲线如图 1 所示.

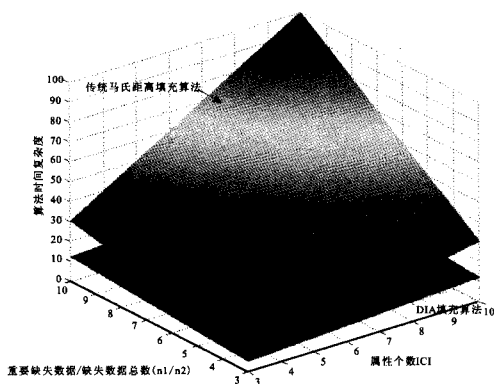


图 1 算法复杂度对比图

从图 1 可知, DIA 算法的时间复杂度明显低于马氏距离填充算法, 当属性个数逐渐增加时, 基于马氏距离的数据填充算法运行时间呈指数增长.

4.2 与经典填充算法性能比较

该实验采用 UCI 标准库中无缺失数据的 Iris 和 diabetes 数据集. 其中, Iris 数据集有 150 个实例, 4 个属性和 1 个决策类; diabetes 数据集有 768 个实例, 8 个条件属性和 1 个决策属性. 实验采用 Matlab 和 Weka 平台相结合, 具体实验步骤如下:

(1) 首先使用 MCAR (Missing Completely At Random) 得到 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50% 缺失率的缺失数据集;

(2) 在 matlab 上, 就每个缺失数据集分别使用 DIA (基于属性重要性的数据填充算法) 算法、Mean、KNN 进行数据填充处理;

(3) 使用 weka 平台上的 C4.5 分类器上对每个测试集进行分类性能预测;

(4) 重复上述步骤 50 次, 最后求取平均值. 实验结果如图 2、图 3 所示.

由图 2 和图 3 曲线图可知, DIA 算法在三种填充算法中具有最好的分类性能, 同时, 当属性集个数越多时, DIA 算法的填充精度越高, 分类性能越好.

5 结束语

针对物联网系统中的缺失值问题, 本文提出了一种属性重要性的数据填充算法. 首先, 通过属性约简求得重要属性和非重要属性, 对于重要属性数据

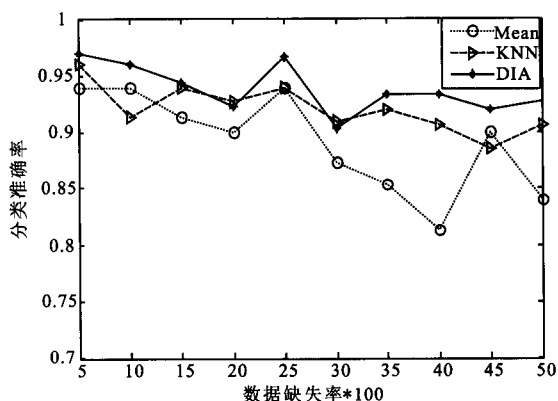


图 2 Iris 数据集上的分类性能比较

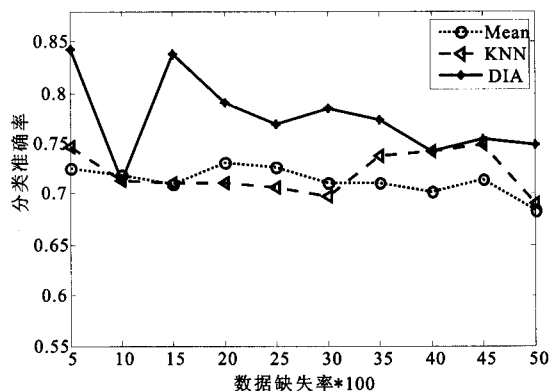


图 3 Diabetes 数据集上的分类性能比较

采用了一种基于改进的马氏距离的精确数填充算法, 在考虑了对象间关系的同时, 也考虑了属性之间的制约关系, 实现了数据的精确填充; 对于非重要属性数据提出了基于概率的填充算法, 提高了填充效率. 实验中针对数字家庭系统数据与传统的马氏距离算法进行了比较, 并且针对 UCI 数据集与 Mean、KNN 填充算法进行了对比分析. 实验结果表明, DIA 算法能够快速、精确地对缺失数据进行填充, 同时在对测试集的分类准确率上均优于其他算法, 并且当数据集的属性个数较多时, 越能体现 DIA 算法的优越性.

参考文献:

- [1] Lamis Hawarah, Ana Simonet, Michel Simonet TIMC-IMAG. Dealing with missing values in a probabilistic decision tree during classification[C]// proceedings of the 6th IEEE International Conference on Data Unining-workshops. washington, DC USA; IEEE, 2006: 325-329.
- [2] Zhang S C. Missing is useful; missing values in cost-sensitive decision trees[J]. IEEE transactions knowledge and data Engineering. 2008, 9(1): 32-38.

(下转第 176 页)

检率较低,证明了算法的有效性.

表 5 入侵检测比较表

数据集	PUDCE		Wenke Lee	
	DR	FPR	DR	FPR
A	91.72%	1.51%	82.37%	——
B	88.96%	1.73%	81.35%	——

4 结束语

本文针对 PU 学习分类算法在正例样本比例低时分类精度不高的问题,提出一种动态的集成 PU 学习数据流分类的入侵检测方法 PUDCE. 在只标注入侵样本(正例)条件下,在正例和未标注样本中学习得到分类器,可进行入侵行为识别. 在人工数据集和真实数据集上的实验表明,PUDCE 算法效果优于单个 PU 分类器,并在正例样本极低的条件下具有良好的分类效果,更加适合网络入侵检测中实际应用.

参考文献:

- [1] 康世瑜. 基于数据挖掘和特征选择的入侵检测模型[J]. 微电子学与计算机,2011(8):52-56.
- [2] Jiazhen He, Yang Zhang, Xue Li, et al. Bayesian classifiers for positive unlabeled learning[C]//Proceedings of the 12th international conference on Web-age information management, Berlin, Heidelberg: Springer-

Verlay,2011;81-93.

- [3] Yi Wang, Yang Zhang. Mining data streams with skewed distribution by static classifier ensemble[J]. Studies in Computational Intelligence, 2009(214):65-71.
- [4] Calvo B, Larranaga P, Lozano J A. Learning bayesian classifiers from positive and unlabeled examples[J]. Pattern Recognition Letters,2007, 28(16):2375-2384.
- [5] Zhengzheng Xing, Jian Pei, Eamonn Keogh. A brief survey on sequence classification[J]. SIGKDD Explorations Newsletter, 2010,12(1):40-48.
- [6] 田俊峰,刘仙跃. 入侵检测数据分类模型—PCANN[J]. 微电子学与计算机,2007(9):35-39.
- [7] 尹向东. 基于正态云模型的网络入侵检测[J]. 微电子学与计算机,2010(10):21-25.
- [8] Wenke Lee,Stolfo J. A data mining framework for constructing features and models for intrusion detection systems[D]. York: Columbia University New ,1999, 135-160.

作者简介:

宋 群 男,(1984—),博士研究生. 研究方向为数据挖掘、系统工程.

张 骏 男,(1964—),博士,教授,博士生导师. 研究方向为控制理论、系统工程等.

智永锋 男,(1978—),博士,讲师. 研究方向为信号与系统.

(上接第 172 页)

- [3] 李宏,阿玛尼,李平,等. 基于 EM 和贝叶斯网络的丢失数据填充算法[J]. 计算机工程与应用,2010,46(5):123-125.
- [4] JinlongSHI,ZhigangLUO. Missing value estimation for DNA microarray gene expression data with principal curves;2010[C]//International Conference on Bioinformatics and Biomedical Technology. Haesun Park: IEEE,2010;262-265.
- [5] XiaofengZhu, ShichaoZhang. Missing value estimation for mixed-attribute data sets[J]. IEEE transactions on knowledge and data engineering,2011,23(1):110-121.
- [6] 刘星毅,农国才. 几种不同缺失值填充方法的比较[J]. 南宁师范高等专科学校学报,2007,24(3):148-150.
- [7] 吕萍,钱进,王波,等. 一种快速差别矩阵属性约简算法

[J]. 计算机工程与应用,2010,46(20):164-167.

- [8] 刘星毅,檀大耀,曾春华,等. 基于马氏距离的缺失数据填充算法[J]. 微计算机信息,2010(9):225-226.

- [9] 杨涛,罗嘉伟,王艳. 基于马氏距离的缺失值填充算法[J]. 计算机应用,2005,25(12):2838-2871.

作者简介:

陈志奎 男,(1968—),博士,教授. 研究方向为无线认识网络、路由算法、数据挖掘.

吕爱玲 女,(1989—),硕士研究生. 研究方向为大规模异构网络不完全信息填充.

张清辰 男,(1988—),博士研究生. 研究方向为大规模数据处理、数据挖掘.