基于朴素贝叶斯的 EM 缺失数据填充算法

邹 薇,王会进
(暨南大学 信息科学技术学院,广东 广州 510632)

摘 要:实际应用中大量的不完整的数据集,造成了数据中信息的丢失和分析的不方便,所以对缺失数据的处理已经成为目前分类领域研究的热点。由于 EM 方法随机选取初始代表簇中心会导致聚类不稳定,本文使用朴素贝叶斯算法的分类结果作为 EM 算法的初始使用范围,然后按 E 步 M 步反复求精,利用得到的最大化值填充缺失数据。实验结果表明,本文的算法加强了聚类的稳定性,具有更好的数据填充效果。

关键词:数据填充;EM 算法;朴素贝叶斯算法
中图分类号:TP311 **文献标识码**:A **文章编号**:1674-7720(2011)16-0075-03

EM algorithm to implement missing values based on Naïve Bayesian

Zou Wei, Wang Huijin
(College of Information Science and Technology, Jinan University, Guangzhou 510632, China)

Abstract: Dataset with missing values is quite common in real applications. It is a big problem of data pretreatment, and handling missing values has become a research hot issue. EM chooses the center of cluster randomly leading to cluster irregularly, so this paper uses the result of Naïve Bayesian as the initial range of EM, then refines the value reduplicative, finally gets the excepted maximize value. The research result suggests that this algorithm improved the level of cluster and had a better data make-up result.

Key words: missing values implement; EM algorithm; Naïve Bayesian algorithm

在数据泛滥的今天,迫切地需要一种将数据转换成有用的信息和知识的数据挖掘技术。然而,由于信息无法获取或者在操作过程中被遗漏等原因,现实中的数据往往存在大量的缺失[1]。数据缺失对数据挖掘的过程和结果有严重的影响:首先,系统丢失了大量有用的信息;其次,系统中所表现出的不确定性更加显著,系统中蕴涵的确定性成分更难把握[2];第三,包含空值的数据会使挖掘过程陷入混乱,导致不可靠的输出;第四,可能直接影响到数据挖掘模式发现的准确性和运行性能,甚至导致错误的挖掘模型[3]。因此,在数据预处理过程中,缺失数据的处理是一个重要的环节。

目前,国外对数据缺失问题的研究取得了很多成果,提出了最近似值替换方法、随机回归填补法、神经网络、贝叶斯网络等理论来解决缺失数据填充问题。国内对填充缺失数据的研究还处在一个开始的阶段,只有银行、保险业等在针对其自身具体的应用进行了缺失数据处理的研究。

总体上说,对缺失值的处理分为三大类:删除元组、数据填充和不处理[4]。其中,处理数据缺失最简单的方法是删除元组,当缺少类标号时通常这样做(假定挖掘任务设计分类),但是当每个属性缺少值的百分比变化很大时,该方法性能特别差[5]。处理数据缺失的有效方法是使用最可能的值填充缺失值,可以用回归、贝叶斯形式化的基于推理的工具或决策树归纳确定[6]。近年来,学术界提出了很多数据填充算法。宫义山提出了基于贝叶斯网络的缺失数据处理方法[7],彭红毅针对数据之间存在相关性且为非高斯分布这种情况提出了 ICA-MDH 数据估计方法[8],Hruschkaetal.使用贝叶斯算法对实例中的缺失值进行估计[9]。

在众多算法中,EM 算法能通过稳定、上升的步骤可靠地找到全局最优值,算法适应性更强。尽管 Gibbs 抽样(Gibbs samplig)[10]、GEM(Generalized EM)算法、Monte Carlo

EM 算法都改进了 EM 算法，但 EM 算法收敛速度慢的缺点仍然没有得到很好的解决。基于此，本文提出结合朴素贝叶斯分类改进传统 EM 算法的方法填充缺失数据的新算法。给 EM 初始值界定了范围，提高了 EM 算法的收敛速度和算法的稳定性，克服了边缘值造成 EM 算法结果偏差大的缺点，实现了良好的缺失数据填充效果。

1 朴素贝叶斯分类的 EM 数据填充算法及其改进

1.1 符号定义

首先对算法中使用到的符号进行定义，如表 1。

表 1 符号定义一览表

符号	定义说明
$X=\{x_1,x_2,x_3,\cdots,x_n\}$	N 维的元组
C	簇。有 $C_1,C_2\cdots C_k$，总 k 个簇
$P(C_k\|X_i)$	每个对象 X 属于 C_k 簇的概率
m	每个对象通过 EM 算法计算得到的最大化值
D	训练元组和相关联的类标号的集合
L	类别，指定有 w 个类：$L_1,L_2,\cdots,L_w$
S_d	总集合 D 元组的总个数
S_{ci}	集合 D 中属于类 C_i 的个数
ξ	概率函数的母体
θ	估计值
X_{L_i}	属于类 L_i 的元组
C_{LiK}	属于类 L_i 中的簇 K

1.2 传统 EM 算法介绍

EM(期望最大化)算法是一种流行的迭代求精算法，它的每一步迭代都由一个期望步(expectation step)和一个最大化步(maximization step)组成。其基本思想是，首先估计出缺失数据初值，计算出模型参数的值，然后再不断迭代执行 E 步和 M 步，对估计出的缺失数据值进行更新，直到收敛。EM 算法的具体描述如下：

(1)随机选择 K 个对象代表簇的中心，以此猜测其他的参数；

(2)反复执行 E 步和 M 步对参数进行求精，直到收敛。

①E(期望)步：用概率 $P(X_i\in C_k)$ 将每个对象 X 指派到簇 C_k。

$$P(X_i\in C_k)=P(C_k|X_i)=P(C_k)\times P(X_i|C_k)/P(X_i)=P(C_k)\times P(X_i|C_k)/\sum_{i=1}^{k}P(C_k)\times P(X_i|C_k)$$

其中，$P(X_i|C_k)$ 表示簇 C_k 中 X_i 的概率，是对象 X_i 的簇隶属概率。

②M(最大化)步：利用前面得到的概率估计重新计算模型参数，$m_k=1/n\times\sum_{i=1}^{m}X_i\times P(X_i\in C_k)/\sum_{i=1}^{j}P(X_i\in C_j)$

1.3 EM 算法改进

EM 算法随机选择对象作为簇的中心，会导致 EM 算法聚类结果的不稳定性，以及边缘数据对整个算法影响过大，使得填充数据正确率偏低。本文提出了基于朴素贝叶斯的 EM 缺失数据填充算法。本算法使用朴素贝叶斯算法对源数据进行分类，将分类结果作为 EM 算法使用范围，在每个类中反复执行 E 步 M 步直至收敛，充分利用了 EM 算法容易达到局部最优的优点，使得 EM 算法更好地聚类，更快地收敛，从而得到更准确的数据填充值。本文算法的具体描述如下：

(1)利用朴素贝叶斯算法对源数据进行分类；

$$P(L_i/X)=P(X/L_i)\times P(L_i)/P(X)=P(X/L_i)\times P(L_i)/\sum^{w}P(X/L_i)\times P(L_i)$$

其中，$P(L_i)$ 为先验概率，等于 S_{Ci}/S_d。

$P(X/L_i)$ 为 L_i 条件下 X 的条件概率密度函数。假定 X/L_i 为一整体 T，该概率密度函数母体 ξ 是离散型，则 $L(\theta\wedge;T_1,T_2,\cdots,T_n)=\sup_{\theta\in\Theta}L(\theta;T_1,T_2,\cdots,T_n)$，满足这个式子的 $\theta\wedge(T_1,T_2,\cdots,T_n)$ 就有可能产生 $T_1,T_2,\cdots,T_n$ 的参数 θ 的值，其相应的统计量 $\theta\wedge(\xi_1,\xi_2,\cdots,\xi_n)$ 称作 θ 的极大似然估计量。如果该概率密度函数母体 ξ 是连续型，则只需求出使得 $L(\theta\wedge;T_1,T_2,\cdots,T_n)=\prod f(T_i;\theta)$ 达到极大的 $\theta\wedge(T_1,T_2,\cdots,T_n)$，便可得到极大似然估计，即 $\ln L(\theta\wedge;T_1,T_2,\cdots,T_n)=\sup_{\theta\in\Theta}L(\theta;T_1,T_2,\cdots,T_n)$。

计算出 $P(L_i/X)$，分类法将预测 X 属于具有最高后验概率(条件 X 下)的类。即朴素贝叶斯分类预测 X 属于类 C_i，当且仅当 $P(C_i/X)>P(C_j/X)\quad 1\leqslant j<\mathrm{m},j\neq\mathrm{i}$。

这样就得出了每个数据元组 X 所属的类，分类完成。

(2)利用(1)分类的结果分别作为新的数据集，在这些数据集中分别使用 EM 算法计算期望最大化值。

在类 $L_1,L_2,\cdots,L_w$ 这 W 个分类中，分别选出 K 个对象代表簇的均值，再反复执行 E 步和 M 步对参数进行求精，直到收敛。

E(期望)步：用概率 $P(X_{Li}\in C_{LiK})$ 分别将类 L_i 中的每个对象 X_{Li} 指派到簇 C_{LiK} 中。

M(最大化)步：利用前面得到的概率估计重新计算模型参数，$m_{Lik}=1/n\times\sum_{i=1}^{m}X_i\times P(X_{Li}\in C_{LiK}))/\sum_{i=1}^{j}P(X_{Li}\in C_{Lij}))$

算法收敛后，用计算得到的最大化值 m_{Lik} 作为类 L_i 中簇 k 的最大化值，并使用这个值填充缺失数字。

1.4 算法伪代码实现

上节描述的算法由程序实现，具体的算法伪代码如下：

输入：使用完全随机缺失方法剔除部分数据的数据集；
输出：使用填充算法处理过的数据集；
读入：使用完全随机缺失方法剔除部分数据的数据集。

(1)使用朴素贝叶斯算法对数据集进行分类：

```
for(int  m=0;m<=MaxItemNo;m++){ //预测分类
  for(n=0;n<=MaxGradeNo;n++){
    for(int  k=0;k<=Max;k++){
```

```
        if(Max [k])
s[n]*=(gm->PF[n][k][Item[m][k].DiscrValue]/gi->GF[n]);
        else{
          if(Item[m][k].Val<=gm->SP[k])
s[n]*=(gm->PF[n][k][1]/gi->GF[n]);
        else
        s[n]*=(gm->PF [n][k][2]/gi->GF[n]);
                }
            }
    s[n]*=(gi->GradeFreq[n]/(gi->MaxItemNo+1));
    if(s[n]>MaxProb){//MaxProb 为最大的概率
    MaxProb=s[n];
    BestI=n;
            }   }
```

(2)将朴素贝叶斯算法的分类结果分别作为 EM 的初始范围。分别在每个类中使用 EM 算法,计算出期望最大值。

```
for(u = 0 ; u < R ; u++){
  if(vaule.update_value){
    for (m = d ; m< dK ; m++){
      bd[i - d] = beta[m]*dens[m];
                    }
    for(m= 0 ; m < d ; m++){
id= m*d;
    for(j = 0 ; j < d ; j++){
      v  = j + id;
      alphabd[v]  = 0.0;
    for (k = 0 ; k < K - 1 ; k++){
      kd  = k*d;
    alphabd[v] += alpha[j + kd]*bds[i + kd];
                              }
            }  }  }  }
```

2 实验结果及分析

从 UCI 机器学习数据库中,选取 4 个没有数据缺失的完整数据集,表 2 列出了它们的详细信息。

表 2 论文中使用的数据集

数据集	实例个数	属性个数	是否缺失
Iris	150	4	N
Vehicle	846	18	N
Heart	270	13	N
Glass	214	10	N

实验设计具体步骤如下:

(1) 将原始数据集准备二份,一份作为原始集,一份作为测试集。用 MCAR(missing completely at random,完全随机缺失)方法随机去掉测试集的不同比率的属性值,并剔除原有类标;

(2) 使用本文算法对(1)后的测试集的属性值和类标进行预测,填充缺失值和类标志;

(3) 反复进行试验 20 次;

(4) 本文使用填充数据与真实数据的平均绝对离差(MAD)和标准平均离差($RMSD$)作为比较标准。其中 $MAD=\sum_{1}^{20}|Y_{填充值\ i}-Y_{真实数据}|/20\times n$,$RMSD=[\sum_{1}^{20}(Y_{填充值\ i}-Y_{真实数据})^2/n]^{1/2}/20$。其中 $Y_{填充值\ i}$ 表示第 i 次填充的数据,$Y_{真实数据}$ 是真值,n 表示缺失个数。

对于不同缺失率的数据集,分别使用 EM 算法和本文算法进行填充,比较结果如表 3~表 5 所示。

表 3 缺失率 15%下 MAD、RMSD 比较结果

		EM 算法	本文算法
Iris	MAD	2.061 4	1.984
	RMSD	2.841 7	2.642 9
Vehicle	MAD	2.061 3	1.985
	RMSD	2.841 8	2.642 4
Heart	MAD	2.060 9	1.983
	RMSD	2.842 0	2.642 8
Glass	MAD	2.061 2	1.984
	RMSD	2.841 8	2.642 7

表 4 缺失率 30%下 MAD、RMSD 比较结果

		EM 算法	本文算法
Iris	MAD	1.914 4	1.746 2
	RMSD	2.552 7	2.246 8
Vehicle	MAD	1.913 9	1.754 2
	RMSD	2.553 4	2.287 6
Heart	MAD	1.914 5	1.748 9
	RMSD	2.553 6	2.249 3
Glass	MAD	1.914 8	1.751 2
	RMSD	2.552 9	2.286 4

表 5 缺失率 50%下 MAD、RMSD 比较结果

		EM 算法	本文算法
Iris	MAD	3.268 4	2.942 3
	RMSD	4.558 1	4.264 9
Vehicle	MAD	3.271 6	2.964 2
	RMSD	4.557 6	4.265 7
Heart	MAD	3.282 1	2.946 3
	RMSD	4.549 4	4.259 1
Glass	MAD	3.274 9	2.954 65
	RMSD	4.551 3	4.258 6

由上述三表可以看出,在缺失率不同的情况下与经典 EM 算法相比,本文算法稳定,且减少了与真实数值的偏差,这样使得实际运用中的填充数据值更真实地反映数据信息。EM 算法提出较早,GEM 算法、Monte Carlo EM 算法和界定折叠法等都改进了 EM 算法,相比较于这些算法,本文充分利用了 EM 算法容易实现局部最优的特点,将 EM 初始范围界定在一个类内,使得 EM 算法很好地聚类和收敛,使得填充值更接近于真实数值。

数据缺失是数据预处理中亟须解决的问题,本文为填充缺失数据提出了基于朴素贝叶斯的 EM 数据填充

算法。该算法使用朴素贝叶斯分类算法的结果作为 EM 算法的初始范围，然后按 E 步 M 步反复求精，利用得到的最大化值填充缺失数据。该算法充分利用了 EM 算法容易实现局部最优的特点，使得 EM 算法更好地聚类，更快地收敛，从而得到更准确的数据填充值。实验结果表明，该算法得到了预期的效果。由于本论文主要是针对数值型属性进行分析，下一步的研究是考虑非数值型属性缺失问题。

参考文献

[1] GRZYMALA-BUSSE J W. Rough set approach to incomplete data. In:LNAI 3070,2004:50~55.

[2] (加)Han Jiawei, KAMBER M. 数据挖掘概念与设计[M]. 北京：机械工业出版社,2008.

[3] LAKSHMINARAYAN K,(1999).Imputation of missing data in industrial databases[J],Applied Intelligence 11：259-275.

[4] HUANG X L.A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets[J]. Pattern Recognition Letters，2002(23)：1613-1622.

[5] GRZYMALA-BUSSE J W,FU M,(2000).A comparison of several approaches to missing attribute values in data mining[C].In:Proc of the 2nd Int'Conf on Rough Sets and Current Trends in Computing.Berlin:Springer-Verlag, 2000：378-385.

[6] ZHANG S C,QIN Y S,ZHU X F,et al.Optimized parameters for missing data imputation.PRICAI06,2006:1010-1016.

[7] 宫义山，董晨.基于贝叶斯网络的缺失数据处理[J].沈阳工业大学学报，2010,32(1)：79-83.

[8] 彭红毅，朱思铭，蒋春福.数据挖掘中基于 ICA 的缺失数据值的估计[J].计算机科学，2005，32(12)：203-205.

[9] HRUSCHKA E R,EBECKEN N F F.Missing values prediction with K2[J].Intelligent Data Analysis,2002,6(6):557-566.

[10] GEMAN S，GEMAN D.Stochastic relaxation,Gibbs distribution and the Bayesian restoration of images[J].IEEE Trans onPattern Analysis and Machine Intelligence, 1984(6):721.

(收稿日期：2011-05-19)

作者简介：

邹薇，女，1987 年生，硕士在读，主要研究方向：数据库应用。

王会进，男，1965 年生，硕士，副教授，主要研究方向：计算机网络和数据库系统应用与开发。