基于 EM 和贝叶斯网络的丢失数据填充算法

李 宏,阿玛尼,李 平,吴 敏
LI Hong,EMMANUEL Amani,LI Ping,WU Min

中南大学 信息科学与工程学院,长沙 410083
School of Information Science and Engineering,Central South University,Changsha 410083,China

LI Hong,EMMANUEL Amani,LI Ping,et al.Imputation algorithm of missing values based on EM and Bayesian network.Computer Engineering and Applications,2010,46(5):123-125.

Abstract: Dataset with missing values is quite common in real applications,and handling missing values has become a research hot issue in the classification field.This paper analyzes and compares several popular missing values imputation algorithms,and has proposed a novel imputation algorithm for missing values based on EM(Expectation Maximization) and Bayesian network.In this algorithm,the Naïve Bayesian is employed to estimate the initial values of EM algorithm,and the EM inspired approach for filling up missing values is incorporated to Bayesian network learning with the objective of ensuring the ultimate updater.As a result,the complete dataset is got after imputation.Experiment results demonstrate that the proposed algorithm enables much higher classification accuracy and lower cost when compared with other classical imputation algorithm.

Key words: missing values imputation;parameter updater;Expectation-Maximization(EM);Bayesian network

摘 要:实际应用中存在大量的丢失数据的数据集,对丢失数据的处理已成为目前分类领域的研究热点。分析和比较了几种通用的丢失数据填充算法,并提出一种新的基于 EM 和贝叶斯网络的丢失数据填充算法。算法利用朴素贝叶斯估计出 EM 算法初值,然后将 EM 和贝叶斯网络结合进行迭代确定最终更新器,同时得到填充后的完整数据集。实验结果表明,与经典填充算法相比,新算法具有更高的分类准确率,且节省了大量开销。

关键词:丢失数据填充;参数更新器;最大期望值算法(EM);贝叶斯网络
DOI:10.3778/j.issn.1002-8331.2010.05.037 **文章编号**:1002-8331(2010)05-0123-03 **文献标识码**:A **中图分类号**:TP301

1 引言

通常现实世界中存在大量的不完整数据集,如数据集中的属性或类标的丢失,对这些丢失数据的处理成为机器学习分类领域中的一个重要研究方向。在机器学习中,最简单的处理丢失数据的方法是摈弃包含丢失值的实例,这种粗糙处理方法当且仅当数据集包含少量具有缺失值的实例时才可行。传统的数据填充方法基于统计算法可分为两种:一种基于模型;一种基于准随机推理[1-3]。统计的方法涉及到诸如均值填充的简单数据驱动也涉及到执行参数估计的复杂模型。基于回归和似然函数是两种常见的填充算法模型。对于回归填充,缺失值由给定实例的未观测属性值基于观测属性值的回归模型来预测。对于似然填充,基于缺失数据的参数估计通过最大似然或最大后验步骤对缺失值进行填充[2]。Gustavo etal.[4]将 kNN 方法运用到丢失数据的填充当中并与 C4.5,均值方法和模式方法进行了比较。Huang etal.[5]采用了上述类似方法对没有缺失数据的训练集进行了测试。Hruschka etal.[6]使用贝叶斯算法对实例中的缺失值进行估计。对不完整数据集的填充算法进行深入研究和探讨,旨在提出一种更好的丢失数据填充方法改善数据集类标的预测准确性。

2 填充算法 EBN

这部分先对算法中将使用到的符号进行定义,然后阐述算法的主要思想并进行相应的算法分析,最后给出算法伪代码。

2.1 符号定义

表 1 符号定义一览表

符号	定义说明
D_h	数据集 D 中的第 h 个实例
A_{ik}	属性 A_i 的取值
A_{ijk}	属性为 A_i 类标为 j 的取值
θ_{ijk}	$P(A_{ik}\|C_j)$的估计值
C_j	实例属于第 j 类
N	数据集 D 中的实例总数
$N(A_{ik}\|C_j)$	D 中属性 A_i 的值为 A_{ik},且类标值为 j 的实例数目
$N(C_j)$	属于第 j 类的实例个数

2.2 算法描述及分析

经典算法 EM(Expectation-Maximization)通常能处理丢失数据的情形,但它必须假设数据源来自某种高斯参数模型或混合参数模型,所以它仅能处理数值属性的数据,而且 EM 算法

基金项目:国家杰出青年基金(the National Science Fund of China for Distinguished Young Scholar under Grant No.60425310)。
作者简介:李宏(1966-),男,博士,教授,主要研究领域为数据挖掘,图像处理;阿玛尼(1982-),男,硕士研究生,主要研究领域为数据挖掘;李平(1986-),男,硕士研究生,主要研究领域为数据挖掘;吴敏(1964-),男,教授,博导,主要研究领域为智能系统和机器学习。
收稿日期:2008-08-19 **修回日期**:2008-11-03

收敛进程相当慢[7]。为了改善 EM 算法的原有不足,提出一种新的基于 EM 和贝叶斯网络的丢失数据填充算法 EBN(EM and Bayes Network imputation)。该算法不要求参数模型的成立,而假设数据集中的各属性的取值一定程度上依赖于其他属性的取值。对于具有丢失数据的不完整数据集 D,各属性存在相互关联性,可以用式(1)和式(2)分别表示条件概率 $P(A_{ik}|C_j)$和类别概率 $P(C_j)$如下:

$$P(A_{ik}|C_j)=\frac{A_{ijk}+N(A_{ik}|C_j)}{A_{ijk}+N(C_j)} \tag{1}$$

$$P(C_j)=\frac{A_j+N(C_j)}{A_j+\sum_j N(C_j)} \tag{2}$$

分子分母中加 A_i 和 A_{ijk} 是为了防止出现零的情况,第 3 部分的设计实验中取 1。

EBN 算法对应各属性的初值通过分别计算实例各属性的条件概率和类别概率来获取,在计算实例属性的期望函数时,建立贝叶斯网络 w 作为更新器 U。

$$E_{P(A|D,\theta_w,w)}(N(A_{ik}|C_j))=\sum_{h=1}^{N}P(A_{ik}|D_h,\theta_w,w) \tag{3}$$

在参数最大化步,计算各实例相应属性的更新参数值 θ_{ijk} 如式(4)所示:

$$\theta_{ijk}=\frac{A_{ijk}+E_{P(A|D,\theta_w,w)}(N(A_{ik}|C_j))}{\sum_{k=1}^{N}(A_{ijk}+E_{P(A|D,\theta_w,w)}(N(A_{ik}|C_j)))} \tag{4}$$

输出 θ_{ijk} 是对所有概率 $P(A_{ik}|C_j)$的估计值,对于一个未分类的新实例,如同传统的 Naïve-Bayes 分类器,由 θ_{ijk} 可得到给定新实例条件下的类变量的后验概率值,新实例将被划分到最大后验概率值对应的类别中。

见表 2 的算法描述,由于给 EM 选择了合适的估计初值,其收敛速度大大提高,并且减小了 EM 算法陷入局部极值点的可能性。通过 k 次迭代最终得到较优的贝叶斯网络更新器 U_i 对丢失数据进行填充,使得填充数预测数据更接近原始值,间接提高了训练集在贝叶斯网络分类器上对测试集的分类性能。表 2 中详细给出了 EBN 丢失数据填充算法的执行伪代码。

表 2 EBN 填充算法伪代码

Pseudocode of EBN Imputation Algorithm
1:Input:D:The dataset within missing values
2:Initialization:Fill the missing values using RBE method
3: Compute P(A_{ik}|C_j)and P(C_j)according to Eqs.(3)(4),and regard them as the initial
4: estimation value of EM algorithm
5:Process(E-step and M-step):
6: k=0;//the number of iterations
7:For attribute A_i do
8: k=k+1;
9://E-step:
10: For attribute A_i do
11: Construct an updater U_i for A_i,according to Eq.(5)
12://M-step:
13: For attribute A_i do
14: Update missing values of attribute A_i through updater U_i,according to Eq.(6)
15:Repeat E-step and M-step until convergence
16:Output:
17: The iteration times k,and the ultimate updaters for imputing missing values of each
18: attribute A_i in the dataset D

3 实验设计与结果分析

这部分先对实验数据集预处理作说明,然后进行实验设计,最后列出实验结果并作分析。

3.1 数据集预处理

实验中使用的数据集全部来自 UCI 数据库,其中 4 个天然丢失数据的数据集和 8 个没有数据缺失的完整数据集,表 3 列出了 12 个数据集的实例数和属性个数。

表 3 实验中使用的数据集分析

No.	DataSets	Instances (Attributes)	Missing Rate/(%)	Numerical Attributes	Norminal Attributes
1	Audiology	226(69)	2.02	0	69
2	Mushroom	8 124(22)	1.38	0	22
3	Vote	435(16)	0	0	16
4	Breast	699(10)	0.22	10	0
5	Diabetes	768(8)	0	8	0
6	Glass	214(10)	0	10	0
7	Heart	270(13)	0	13	0
8	Ionosphere	351(34)	0	34	0
9	Iris	150(4)	0	4	0
10	Vehicle	846(18)	0	18	0
11	Hepatitis	155(19)	5.67	6	13
12	Lymph	148(18)	0	3	15

不完整数据处理的效果在很大程度上取决于不完整数据产生的机制, 大多数填充算法均假设丢失数据采用 MCAR(Missing Completely At Random) 也即完全随机缺失方法来处理,实验中同样作类似假设,分别对 UCI 中选取的非天然丢失数据的数据集使用 MCAR 方法进行不同数据丢失率的预处理。

3.2 实验设计

实验平台采用 MATLAB 与 WEKA 相结合。先考虑 UCI 数据集中 4 个天然丢失数据的数据集,分别使用新算法 EBN、Naïve-Bayes、kNN、平均值等填充方法在 C4.5 分类器上进行测试集分类性能的比较,实验结果如表 4 所示。随后分别选取了 8 个使用 MCAR 方法得到的缺失属性的数据集使用不同的填充方法 EBN、Mean、Naïve-Bayes、kNN 在相同分类器 C4.5 上进行测试集分类精度的比较,结果如表 5 所示。具体实验步骤如下:

(1)将原始数据集一分为二,变成数目相等的训练集和测试集,并用 MCAR 方法随机去掉不同比率 10%到 50%不等的属性值,同时原有实例类标剔除。

(2)分别在 MATLAB 中运行 EBN、Mean、Naïve-Bayes(N.B.)、kNN 等丢失数据填充算法对训练集和测试集的丢失数据进行填充,添加原有类标,从而分别得到完整的训练集和测试集。

(3)使用完整的训练集通过 WEKA 上的 C4.5 决策树(J48)进行建模分类,对测试集进行类标预测。

(4)根据测试集的预测类标与原有类标进行对比,得到算法分类准确率。

(5)重复以上步骤 20 次,得到用不同填充算法基于测试集的平均分类准确性。

3.3 实验结果及分析

对 UCI 中的 4 个天然丢失数据的数据集进行分类准确率的测试比较结果如表 4 所示。

表 5　不同数据丢失率下的各种填充算法分类性能比较表

Data Sets	Percentage of Missing Values											
	10%				30%				50%			
	EBN	Mean	N.B.	kNN	EBN	Mean	N.B.	kNN	EBN	Mean	N.B.	kNN
Diabetes	**78.04**	71.52	75.36	71.44	**76.79**	70.73	73.62	68.53	**72.97**	69.04	71.84	66.36
Glass	**69.21**	61.73	57.89	65.53	**64.33**	58.79	56.48	59.92	**58.27**	52.36	52.49	54.57
Heart	**85.19**	77.48	83.70	76.32	**82.46**	78.22	80.87	71.14	**77.36**	75.54	75.93	67.11
Ionosphere	**90.15**	87.12	81.77	86.04	**82.35**	79.64	73.79	79.23	**78.18**	74.07	72.13	75.87
Iris	93.68	93.21	**94.67**	92.44	91.55	86.67	**92.62**	87.71	**87.44**	82.00	86.00	79.93
Lymph	**83.61**	79.74	81.08	78.69	**80.76**	72.24	78.70	72.94	**76.35**	73.75	73.57	70.08
Vehicle	**71.17**	68.19	56.68	65.11	**67.04**	62.59	56.93	60.72	**62.42**	57.26	53.14	57.09
Vote	93.21	91.34	88.97	**93.42**	**91.77**	89.75	88.96	88.74	87.57	**89.41**	88.27	86.44

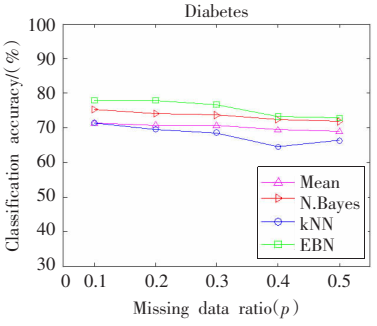

图 1　Diabetes 上的分类性能比较图

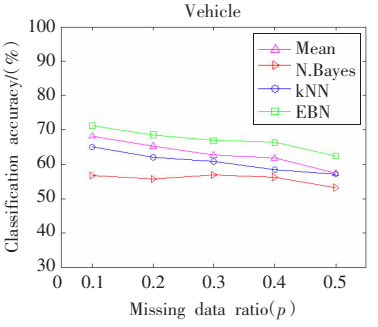

图 2　Vehicle 上的分类性能比较图

表 4　天然丢失的数据集填充算法分类性能比较表

Datasets	Instances (Attributes)	Missing Rate/(%)	EBN	Mean	N.B.	kNN
Audiology	226(69)	2.02	**77.91**	75.26	73.24	71.59
Breast	699(10)	0.22	**96.14**	93.42	93.77	95.64
Hepatitis	155(19)	5.67	**84.63**	81.79	83.87	82.58
Mushroom	8 124(22)	1.38	**85.72**	82.55	84.52	81.37

对 UCI 中的 8 个数据集分别进行 10%,20%,30%,40%,50%等 3 种丢失率的处理后分别用 EBN、Mean、Naïve-Bayes、kNN 进行数据填充处理。其中数据丢失率分别为 10%,30%,50%的测试集分类准确率记录如表 5 所示。由表 5 可知,相比其他 3 种算法,EBN 对于大多数数据集均具有更高分类准确率。

图 1 和图 2 分别描绘了数据集 Diabetes 和 Vehicle 在 5 种不同数据丢失率的情形下经过 4 种填充算法得到的测试集分类准确率比较曲线图。由曲线图可知,EBN 算法在 4 种填充算法中具有最好的分类性能。从两幅图的对比还可以发现,Naïve-Bayes 方法在数据集 Vehicle 上的分类性能优于 Mean 和 kNN;而在数据集 Diabetes 上 Naïve-Bayes 方法的分类性能最差,得知不同的算法在不同的数据集上的分类性能有所不同。

4　结束语

针对数据集中存在丢失数据的情形,提出了一种基于 EM 和贝叶斯网络的填充算法 EBN。算法通过朴素贝叶斯算出条件概率给 EM 中的期望函数赋予初始参数值,随后通过建立贝叶斯网络更新器对参数进行更新,迭代数次后到算法收敛,从而得到最后的填充数据,该算法收敛速度比传统 EM 快。实验中针对 UCI 数据集与 Mean、Naïve-Bayes、kNN 等填充算法进行了比较。实验结果表明,EBN 算法在对测试集的分类准确率上优于其他填充算法。下一步的研究工作将考虑把 EM 和 PCA (Principle Component Analysis)相结合[8],进一步提高丢失数据的填充准确率,从而改善数据集的分类性能。

参考文献:

[1] Lakshminarayan K,Harp S A,Samad T.Imputation of missing data in industrial databases[J].Applied Intelligence,1999,11:259-275.

[2] Li K H.Imputation using Markov chains[J].Journal of Statisticalt Comput Simul,1988,30:57-79.

[3] Little R J,Rubin D B.Statistical analysis with missing data[M].[S.l.]:John Wiley and Sons,1987.

[4] Gustavo E A,Batista P A,Monard M C.An analysis of four missing data treatment methods for supervised learning[J].Applied Artificial Intelligence,2003,17(5/6):519-533.

[5] Huang C,Lee H.A grey-based nearest neighbor approach for missing attribute value prediction[J].Applied Artificial Intelligence,2004,20(3):239-252.

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

[7] Petersen B,Winther O,Hansen L K.On the slow convergence of EM and VBEM in low-noise linear models[J].Neural Computation,2005,17(9):1921-1926.

[8] Stanimirova I,Daszykowski M,Walczak B.Dealing with missing values and outliers in principle component analysis[J].Tanlanta,2007,72:172-178.