

生物缺失数据处理的贝叶斯模型研究

沈 奇, 王池社

(金陵科技学院 信息技术学院, 江苏 南京 211169)

摘 要: 文章针对生物信息实验中的分类预测问题, 以属性缺失数据为对象, 结合朴素贝叶斯算法的特点, 设计了一种基于改进 EM 算法的缺失数据朴素贝叶斯填充模型, 并应用于蛋白质作用位点的定位研究中. 实验结果表明, 通过算法进行生物缺失数据的处理, 在准确率、精度、召回率、ROC 方面均获得了比其他方法更好的效果.

关键词: 缺失数据; 朴素贝叶斯; 分类

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2011)07-0110-03

Research of Bayesian Model Based on Missing Data of Biology

SHEN Qi, WANG Chi-she

(School of Information Technology, Jinling Institute of Technology, Nanjing 211169, China)

Abstract: In order to solve the classification prediction problem in biological information experiments, this article designed a new Bayesian classification model based on attribute missing data which combined with the characteristics of Naïve Bayesian algorithm. This model was applied to the research of protein-protein interaction site localized. Experimental results show that the biological missing data processing through this model has obtained better results than other methods in the fields of accuracy, precision, recall and ROC.

Key words: missing data; Naïve Bayesian; classification

1 引言

现实世界的数据库(例如,商业数据库和医院数据库)中的数据很少是完全的:丢失的数据、观测不到的数据、隐藏的数据、录入过程中发生错误的数据等等是现实世界数据库的一个常见特征.为了能在现实世界中有效地应用知识发现的方法和系统,就必须面对数据不完全的挑战^[1].大部分学习算法在处理此类问题时采用的是直接删除的方法:将数据缺失严重的属性或元组直接删除,但这样必然会造成信息的丢失.一些传统的机器学习算法则是采用特殊值填充方法进行处理:利用数据集中已知信息对缺失数据进行估计,并用估计值替代相应的缺失数据.而另外一种常用的方法就是参数估计法:对所有的观测数据进行全局分析,对相应的缺失数据项进行参数估计,并用估计值替代相应的缺失数据.

贝叶斯网络是数据挖掘的有效工具之一,它提供了一种自然地表示因果信息的方法,用来发现数据间的潜在关系,并以图形的方法描述出来^[2],它在处理缺失数据时具有其它机器学习所没有的优点:在一个训练实例中,如果一个属性值缺失,它便不被包括在频率计算中,概率的计算取决于真正出现属性值的训练实例个数,而不是训练实例的总数.因此,贝叶斯方法可以有效的处理缺失数据^[3],但其缺点也是显然的,造成缺失数据是随机丢失的.

基于上述原因,设计了基于RBC(Robust Bayes Classifier)^[4]的缺失数据的分类器来解决MAR假设问题^[5].文中在改进的EM算法缺失数据填充基础上,结合蛋白质作用位点定位研究的生物信息数据缺失的特点,设计了一个针对属性值缺失的贝叶斯模型,并将此模型应用到我们的前期研究中^[6].实验结果证明,此方法可以有效处理生物数据的缺失

收稿日期: 2010-10-25; 修回日期: 2010-11-21

基金项目: 国家自然科学基金项目(61070131);金陵科技学院博士科研启动基金

问题.

2 主要理论

2.1 最大似然估计方法

最大似然参数估计方法是在贝叶斯网络结构已知情况下的一种参数学习方法,其基本思想是利用似然函数来度量模型参数与数据样本的拟合程度:

设有随机变量 $X_1, X_2, \dots, X_n$. 假设 X_i 具有 r_i 个可能的取值: $1, 2, \dots, r_i$. 在所给的贝叶斯网络结构 m 下, X_i 的父结点集 $\pi(X_i)$ 的所有可能的配置为 $1, 2, \dots, q_i$, 定义下式为最大似然估计:

$$\theta_{ijk} = \begin{cases} \frac{m_{ijk}}{r_i}, & \text{若 } \sum_{k=1}^{r_i} m_{ijk} > 0 \\ \frac{1}{r_i}, & \text{若否} \end{cases}$$

其中 m_{ijk} 是数据中满足 $X_i = k$ 和 $\pi(X_i) = j$ 的样本的数量.

2.2 缺失数据类型

缺失数据的产生是多种原因的,统计学家根据数据集中所出现的缺失数据的情况,将缺失数据分为下述三类:

完全随机缺失:数据的缺失是随机的,数据的缺失不依赖于任何不完全变量或完全变量.

$$P(Y_{\text{missing}} | X, Y) = P(Y_{\text{Missing}})$$

随机缺失:数据的缺失不是完全随机的,即该数据的缺失依赖于其他完全变量.

$$P(Y_{\text{missing}} | X, Y) = P(Y_{\text{Missing}} | X)$$

非随机缺失:除上述两种情况外的其它情况.

3 处理缺失数据的贝叶斯算法模型

算法的指导思想是利用数据集中已知信息对缺失数据进行估计,并用估计值替代相应的缺失数据;在填充后形成的完整数据集上,采用朴素贝叶斯方法构建模型.该方法的目标就是充分利用观测数据的全局信息,识别出数据集中数据值间的关联,运用这种关联帮助估计缺失值.

3.1 改进的 EM 算法进行缺失值填充

生物缺失数据的分类模型在蛋白质相互作用的研究中具有重要的应用价值,基于此,我们仅考虑基于朴素贝叶斯分类结构下属性信息的丢失情况,通过对 Robust 方法处理缺失数据的理论进行研究,结合 EM 算法与生物缺失数据的特点,变换 E 步骤与 M 步骤的公式,进行条件概率与先验概率区间

的计算:

考虑一个由 n 个变量 $X = \{X_1, X_2, \dots, X_n\}$ 组成的朴素贝叶斯结构,样本数据集记为 D ,并假定各属性满足离散属性(连续数据可以预先离散化);节点 X_i 共有 r_i 个取值 $1, 2, \dots, r_i$,类节点 C_j 共有 q_j 个取值 $1, 2, \dots, q_j$;记 $n(?, c_j)$ 为属性 A_i 的值缺失,类值为 j 的实例的个数.

改进的 EM 算法步骤如下:

(1) 利用下式计算条件概率 $P(x_i^k | C_j)$:

$$P(x_i^k | C_j) = \frac{1 + n(x_i^k, C_j)}{\sum_{m=1}^{r_i} [1 + n(x_i^m, C_j)] + n(?, c_j)}$$

将此概率值作为 EM 算法的参数 θ .

(2) E-步骤:计算不完整数据集的充分统计量的期望值

$$m_{ijk} = \sum_{l=1}^N P(x_i^k, C_j = j | y_l, \theta^t)$$

其中, y_l 是数据集中的第 l 个实例,可能含有缺失值. x_i^k 表示属性 X_i 取第 k 个值.

(3) M-步骤:基于修补后的完整数据计算 θ 的最大似然估计值.

$$\theta_{ijk} = \frac{m_{ijk}}{\sum_{k=1}^{r_i} m_{ijk}}$$

迭代第(2)(3)步,收敛到一个给定的阈值,得到的 θ_{ijk} 即可以作为朴素贝叶斯分类器的输入.

3.2 基于缺失数据的朴素贝叶斯模型(RNBC)

在上述数值填充算法的基础上,我们建立基于缺失数据的朴素贝叶斯模型如下:

(1) 对所有的缺失的样本实例采用上述方法进行缺失数据填充.

(2) 在得到的完整数据集上,计算先验概率 $P(C_j)$ 和条件概率 $P(x_k | C_j) = \theta_{ijk}$.

(3) 利用经典的朴素贝叶斯分类器进行分类计算如下:

$$C_{NB} = \arg \max P(C_i) \prod_{k=1}^n P(x_k | C_j)$$

4 实验结果

4.1 实验数据的选择

在生物信息学中,蛋白质间的作用位点预测研究是一个重要的研究内容,它的研究对于蛋白质的功能预测、蛋白质复合物的结构预测具有重要的研究意义.我们针对生物信息数据易缺失的特点,选择蛋白质作用位点定位研究的数据为研究对象.为了

比较方便,数据来源仍选择我们前期研究的数据,主要步骤简述如下:

(1) 选择包括了 53 个蛋白质复合物的 81 条蛋白质链。

(2) 选择进化保守性的残基保守性打分及溶剂可及表面积 ASA 两个主要的生化特征。

(3) 选择残基的 9 序列邻近信息。

通过对所有 81 条蛋白质链的相关特征信息的提取,发现基本上每一链的残基保守性打分都会出现缺失,少则几个原子,多的达到 56 个之多。将这些数据用三种不同的方法来处理,分别得到了三类样本:第一类是直接将缺失数据删除,得到了 2314 个界面残基和 5555 个非界面残基。第二类是将所有缺失数据直接用“?”替代而不加删除,得到 2659 个界面残基和 6190 个非界面残基。第三类是将所有缺失数据采用本文的算法进行缺失数据填充,仍然为 2659 个界面残基和 6190 个非界面残基。

4.2 实验结果分析

针对上述三类样本,文中采用朴素贝叶斯分类器与支持向量机(libSVM)对第一、二类样本进行学习;采用基于缺失数据的朴素贝叶斯模型(RNBC)进行第三类样本数据的学习。学习的结果采用 10-fold 交叉验证方法进行结果验证。所有的运行环境基于 WEKA 的源码开放机器学习算法,方法预测结果如图 1 所示。

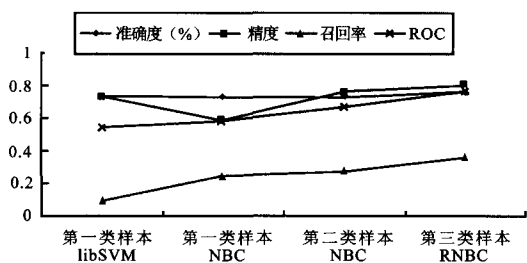


图 1 三类不同样本的实验结果图

由上图可以看出,在相同样本数据集上,NBC 由于具有缺失数据的处理能力,其运算效果比 libSVM 方法具有较好的效果。更重要的是,对样本

数据集中的缺失数据采用文中提出的改进 EM 算法进行缺失值填充处理后运用 NBC 得到的预测效果,比直接删除缺失数据或用“?”代替缺失数据都要好,其预测的准确率、精度、召回率以及 ROC 指标值均比前两种缺失数据处理方法有显著提高。

5 结束语

在当前信息数据量迅速发展的情况下,如何有效的进行缺失数据的处理越来越重要。尤其是在生物信息学的应用中,由于目前的实验手段等原因,缺失数据是普遍存在的,有的时候这样的缺失数据的信息量甚至是很大的。如何根据数据信息的特征进行有效的缺失数据算法设计是一项具有重要意义的研究方向。

参考文献:

- [1] 王洪春. 缺失数据的主曲线恢复方法[J]. 微电子学与计算机, 2008, 25(11):160-161.
- [2] 左春荣,余本宫,李娜,等. 贝叶斯网络在大规模医疗数据上的应用研究[J]. 微电子学与计算机, 2008, 25(6): 112-115.
- [3] Witten I H, Frank E. 数据挖掘实用机器学习技术[M]. 北京: 机械工业出版社, 2006.
- [4] Ramoni M, Sebastiani P. Robust bayes classifiers[J]. Artificial Intelligence, 2001, 125(1-2):209-226.
- [5] Chen J, Huang H, Tian F, et al. A selective Bayes classifier for classifying incomplete data based on gain ratio[J]. Knowledge-Based Systems, 2008, 21(7):530-534.
- [6] Wang C, Song J, Li F, et al. Identifying protein-protein interaction sites using adapted bayesian classifier [J]. CCCM, 2009:153-156.

作者简介:

沈奇女, (1975-), 硕士, 讲师. 研究方向为数据挖掘、特征处理。

王池社男, (1974-), 博士, 副教授. 研究方向为数据挖掘、智能计算、生物信息学。