缺失数据处理方法的比较研究*)

刘 鹏 雷 蕾 张雪凤
(上海财经大学经济信息管理系 上海200433)

摘 要 数据挖掘已被广泛用于医疗领域,而大多数医疗数据集都存在缺失值。本文介绍了一些缺失值估计算法。建立了5种模型来提高预测的有效性,它们是保留缺失模型、直接丢弃模型、贝叶斯补缺模型、贝叶斯重叠补缺模型和基于信息增益的贝叶斯重叠补缺模型。这些模型在Clinics数据集上进行了处理和分析。用C4.5决策树和10叠交叉确认法来检验这些模型的性能,结果表明根据信息增益递减顺序排序,用朴素贝叶斯分类器来预测缺失值是有效的。

关键词 数据挖掘,缺失值,朴素贝叶斯分类器,信息增益

A Comparison Study of Missing Value Processing Methods

LIU Peng LEI Lei ZHANG Xue-Feng
(Department of Information Systems, Shanghai University of Finance and Economics, Shanghai 200433)

Abstract Data mining approaches have been applied widely in the field of healthcare and most healthcare datasets are full of missing values. Some missing value estimation methods are introduced in this paper. Five models are built to improve the efficiency of the prediction: Basic model; Delete straight model; Bayesian estimation model; Bayesian estimation iteration model and Bayesian estimation iteration model based on information gain. The models are conducted and analyzed on Clinics dataset. Decision tree C4.5 and 10-folds cross-validation are used to estimate the performances of each model, which shows that use naive Bayesian classifier to predict missing values iteratively in degressive order of information gain is effective.

Keywords Data mining, Missing value, Naive bayesian classifier, Information gain

1 引言

近年来,数据挖掘技术被广泛地应用于医疗领域。数据挖掘的过程包括问题理解、数据采集和理解、预处理、数据挖掘工具、模型评估和知识应用[1]。根据文[1]的研究,在数据挖掘过程中20%的时间用于目标识别,60%用于数据准备,数据挖掘和知识分析都为10%。为什么人们要将超过50%的精力放在数据预处理上呢?在现实世界的数据库中存在着严重的质量问题:1)数据不完整,2)数据冗余,3)数据不一致,4)噪音数据。这些严重的质量问题会降低挖掘算法的性能[2],因此,人们不得不将大量的时间和精力花在数据预处理上。在保证不减少数据所含信息的前提下,合理有效的数据预处理可以压缩数据量,改善数据质量,提高数据挖掘算法的性能,减少学习时间[2]。

缺失数据的处理问题是预处理中的一个重要问题,本文第2节介绍数据预处理过程中的几种常见缺失数据处理技术。第3节给出了基于贝叶斯原理和信息增益的缺失数据处理模型。第4节是各模型在Clinics数据集的应用和模型比较。最后为结束语。

2 缺失数据处理技术

数据挖掘算法中,有些算法要求数据集的所有值都是可用值,缺失数据会降低算法的性能,如决策树、K平均聚类法等[3]。缺失数据处理不当,就会累计大量错误,增加后续算法的运算时间和复杂度。人们对缺失数据的处理方法展开了研究,在预处理过程中使用的缺失数据处理方法有:

1. 直接丢弃含缺失数据的记录[4]。

2. 用全局变量或是属性的平均值来代替所有缺失数据,把全局变量或是平均值看作属性的一个新值[4]。

3. K-最近距离邻居法[3]。先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的K个样本,将这K个值加权平均来估计该样本的缺失数据。

4. 用预测模型来预测每一个缺失数据[4]。用已有数据作为训练样本来建立预测模型,预测缺失数据。该方法最大限度地利用已知的相关数据,是比较流行的缺失数据处理技术。

3 缺失数据的处理模型

3.1 模型的建立

针对缺失数据的处理方法,本文提出如下的处理模型:

1. 基模型—保留缺失模型

不对缺失数据做任何处理,直接使用含大量缺失数据的初始数据集,简称数据集0。将该模型作为基模型,与其他缺失数据处理模型进行比较分析。

2. 模型1—直接丢弃模型

直接删除数据集0中所有含缺失数据的记录,仅用剩余的无缺失数据的记录组成新数据集,简称数据集A。

3. 模型2—贝叶斯补缺模型

该模型采取补值的方法来处理缺失数据。在数据集0的基

*)上海财经大学"211工程"重点学科建设项目资助(2004[9])。刘 鹏 副教授,主要研究领域为数据挖掘与信息系统。雷 蕾 硕士研究生,研究领域为数据挖掘与知识管理。张雪凤 讲师,研究领域为数据挖掘与数据库。

础上，分别将具有缺失数据的属性作为预测目标属性，用朴素贝叶斯分类器来预测该目标属性，并将所有预测所得值填补到原数据集0中的缺失数据处，生成一个无缺失数据的新数据集，简称数据集B。

4. 模型3－贝叶斯重叠补缺模型

贝叶斯补缺模型用原数据集0来预测所有具有缺失数据的属性，并没有利用先预测出的缺失值。对数据挖掘模型来说，训练数据集的数据越多，建立的数据挖掘模型的精度就越高。于是，我们利用先预测出来的属性值来充实训练数据集，将其填补到数据集0中，产生新的数据集C_1。再在数据集C_1的基础上预测下一个属性，并将预测所得值填补到数据集C_1中，产生新的数据集C_2，……依次类推，最后产生一个完整的新数据集C。这种补缺的处理方式每次都是在前一次的基础上再预测，充分利用了已预测出来的缺失数据。

在贝叶斯重叠补缺模型中补值的顺序不同，产生的结果也就不同。因此随机产生10组不同的补值顺序，并对这10组不同的补值顺序所产生的模型进行研究。

5. 模型4－基于信息增益的贝叶斯重叠补缺模型

在贝叶斯补缺模型中，每次所用的训练数据集是不变的，都是数据集0。所以，属性预测的先后顺序对最终产生的数据集并没有影响。但是，补值过程中的属性顺序对于贝叶斯重叠补缺模型来说很重要。先预测属性值的准确率会影响到后预测属性值的准确率。如果用错误的数据来建立贝叶斯模型并预测，那么后续预测值的错误率会更高。也就说值补得越多，后补值的错误率也就越高，产生的新数据集的真实性、可靠性也就越差。因而，利用贝叶斯重叠补缺模型对所有属性的缺失数据集进行补值，所得数据集未必是最好的。

针对上述问题，我们利用信息论中熵的概念，以信息增益(information gain)为标准来选择属性。在信息论中，有许多属性的选择方法，如信息增益比(gain ratio)、距离度量(distance measure)、χ^2统计和相关度(relevance)等等[5]。其中，信息增益属性选择标准是一个发展成熟的标准，已经被学者们广泛地接受和采纳，如著名的C4.5决策树模型就是采用信息增益来选取属性节点的。

根据属性的信息增益递减顺序排列，按照贝叶斯重叠补缺模型依次预测并填补属性中的缺失数据，最后产生数据集D。

将上述各缺失数据处理模型汇总至表1。

表1　5个模型汇总

模型	缺失处理方式	补值顺序
基模型	保留	N/A
直接丢弃模型	丢弃	N/A
贝叶斯补缺模型	补值	随机
贝叶斯重叠补缺模型	补值	随机
基于信息增益的贝叶斯重叠补缺模型	补值	排序

3.2　模型的检验

本文用到的数据挖掘相关技术有朴素贝叶斯分类器、决策树C4.5和信息增益。这里采用检验模型的方式来检验缺失数据的处理效果，如果模型检验的结果好则说明缺失数据的处理效果好。用缺失数据处理后的数据集建立C4.5决策树预测模型，并用10叠交叉确认法来检验该决策树模型的预测准确率。其中还涉及到三个概念：模型预测准确率，类预测准确率以及补缺收益。分别定义如下：

模型预测准确率＝

$$\frac{\text{整个数据集中预测正确的记录个数}}{\text{整个数据集的记录总数}}\times 100\%$$

类预测准确率＝

$$\frac{\text{某个类中预测正确的记录个数}}{\text{该类中的记录总数}}\times 100\%$$

补缺收益＝

$$\frac{\text{模型预测的准确率}-\text{基模型预测准确率}}{\text{基模型预测准确率}}\times 100\%$$

其中，“整个数据集中预测正确的记录个数”和“某个类中预测正确的记录个数”是利用C4.5决策树预测模型在补缺模型产生的数据集上计算得到的，而“基模型预测准确率”是在数据集0上计算得到的。

4　模型应用

数据挖掘在医疗领域有着广泛的应用，如将数据挖掘技术应用于病人住院期的预测，可以帮助医院更好地配置医疗资源[6]。然而大多医疗数据集都存在着大量的缺失数据，通过数据预处理阶段对数据集的改进，能提高预测模型的准确率。

4.1　Clinics 数据集

Clinics 数据集是来自伦敦 St. George's 医院老年医学部的一个临床医疗系统1994年到1997年的数据[6]。一共有4722条病人记录，其内容包括病人的个人数据、入院时间、入院原因、BARTHEL 指数、治疗结果以及住院期等等。其中，住院期属性的均值为85天，中值为17天。将病人的住院期分为三类：短期0～14 天，中期15～60天，长期61天以上。根据文[6]的研究，形成 Clinics 处理数据集，简称数据集0。

Clinics 数据集中缺失数据的比重相当大，一共有3017条记录含有缺失数据，占总记录数的63.89%。住院期短、中、长三类中缺失记录所占比重分别为63.29%、61.81%和74.86%。有8个属性具有缺失数据。通过对缺失数据的处理，期望提高预测模型的准确率，尤其是住院期为长期的预测准确率。

4.2　模型应用与分析

将第3节所提出的缺失数据处理模型应用于 Clinics 数据集，得到模型及类预测准确率(表2)，以及模型补缺收益(表3)。

表2　预测准确率汇总表

模型	模型预测准确率	类预测准确率		
		短期	中期	长期
基模型	52.44%	44.40%	68.80%	10.00%
模型1	54.31%	48.20%	65.30%	14.00%
模型2	51.21%	47.40%	62.30%	17.00%
模型3	55.62%	46.44%	69.26%	28.82%
模型4	55.87%	47.50%	68.40%	31.10%

表3　补缺收益汇总表

模型	模型预测准确率	类预测准确率		
		短期	中期	长期
模型1	3.6%	8.6%	－5.1%	40.0%
模型2	－2.3%	6.8%	－9.4%	70.0%
模型3	6.1%	4.6%	0.7%	188.2%
模型4	6.5%	7.0%	－0.6%	211.0%

(下转第174页)

```
    〈Time type="Caption_Content"〉
      ……
    〈/Time〉
      ……
  〈/Page〉
〈/ContentPage〉
```

4.3 动态文档的实现

如前所述，由于数据和数据的用户接口的独立性，当文档的内部结构需要更改的时候，比如增加一个需求的提出者，系统只需要修改 XML 文件就可以达到要求。

这里需要对应修改三个节点，首先是在 DataInfo 和 DateTemplate 的子节点 ContentPage 下增加个名为 Announcer 的子节点。Announcer 节点的数据就是系统增加的需求提出者。其次是修改 UIInfo 节点，同样在 ContentPage 的 Page 节点下增加 Announcer 节点，并调整 Announcer 节点和其余节点的属性，使"提出者"这个数据位于页面的合适位置。

当文档的内部结构没有改变仅仅需要调整数据的显示属性时也可以通过修改 XML 文件来实现。这时就要修改 UIInfo 节点，UIInfo 节点包括系统的默认字体，每一个数据用什么方式显示，数据所在位置，数据的外部组织形式等等信息。比如，我们在"需求文档生成时间"的位置想去掉"时间："这个标识，我们可以把相应的节点改为：

```
〈Time type="Content"〉
  〈Content height="30" width="160" x="120" y="60"〉
    〈Action name="FocusListener" source="FXML-
    Doc.plugin.DateCheckListener"/〉
  〈/Content〉
〈/Time〉
```

4.4 动态行为的实现

由于文档行为和系统架构之间的独立性，文档自身定义自己的行为，而在应用程序目录\plugin 下的一个个 class 文件才是行为的真正实现。

文档的 Action 节点就是系统行为的定义。如 Time 的子节点：

```
〈Action name="FocusListener" source="FXML-
Doc.plugin.DateCheckListener"/〉
```

Action 节点包括两个属性 name 和 source，name 指定的是行为的类型名，在这里就是 FocusListener；source 指定的是实现具体行为的类，在这里就是 FXMLDoc.plugin.DateCheckListener。

系统的行为框架是由一个叫 ActionAssembler 类实现的。这个类负责解析 Action 节点并生成相应类的实例，并根据行为的名字绑定到指定的 UI 组件上。由于在这里用到的是 JComponent 的 addFocusListener 方法，相应的 DateCheckListener 是从 FocusListener 继承而来。

```
public static void assemble(JComponent talker,Node actionNode)
{
  //解析 Action 节点得到_actionName 和_actionClass
  parseActionNode(actionNode);
  if(_actionName.compareTo(FOCUS_LISTENER)==0)
  {
    try
    {
      //生成一个_actionClass 的类的对象
      Class newClass=Class.forName(_actionClass);
      //生成对象的实例，并和 talker 这个 UI 组件绑定起来
      talker.addFocusListener ((FocusListener) newClass.new-
      Instance());
    }
    catch …
  }
  else if …
}
```

结束语 对于文档管理系统，软件维护需要的工作量占软件生命周期的比重非常大。如何有效地降低维护费用，延长软件的使用寿命，本文基于 XML 技术通过提高软件各部分的独立性，做到软件模块高内聚、低耦合，在降低软件的维护费用方面做了一些工作。

参考文献

1 Bray T, Paoli J, Sperberg-McQueen C M. Extensible Markup Language 1.0, Feb. 1998
2 Aviram M H. XML and Java: A powerful combination
3 Elliotte Rusty Harold, XML Bible, June, 1999

（上接第156页）

在对计算结果的分析中，我们注意到运用基于信息增益的贝叶斯重叠补缺模型补到第三个属性时，形成数据集 D'，此时模型预测准确率和长期类预测准确率最高。如果需要对剩余的全部缺失数据进行补缺，那么在数据集 D' 的基础上，运用贝叶斯补缺模型预测剩余的具有缺失数据的属性，生成最终数据集 D，其结果为：模型预测准确率55.53 %，短、中、长三类的类预测准确率分别是45.4%、69.7%和29.9%，计算结果亦令人满意。

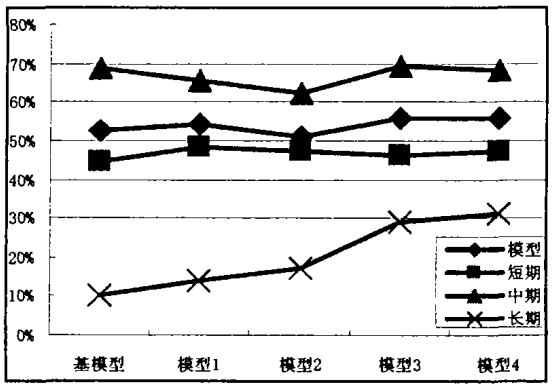

图1 模型及类预测准确率汇总图

根据表2可以画出模型及类预测准确率(图1)。可以看到，本文所提出的缺失数据处理模型不仅可以改善总体模型的预测准确率，而且能使长期类的预测率得到更大提高，其补缺收益为211%，效果十分明显。这说明对提高长期类预测准确率来说，缺失数据的处理起到了关键作用。

结束语 本文重点讨论了几种不同的缺失数据处理方法，建立了四个缺失数据处理模型，及其组合应用。通过将四种模型应用于医疗领域，预测病人住院期的长短，来说明本文所建立模型的有效性。

然而，除了信息增益之外，还有很多因素会影响补缺的性能，例如各属性中缺失数据所占比重及重要性。对本文所提出的模型进行改进，找出一个更加有效的属性补缺顺序的排列标准，还需要进一步的研究工作。

参考文献

1 Cios K J, Kurgan L A. Trends in Data Mining and Knowledge Discovery. In: Knowledge discovery in advanced information systems, Pal, N.R., Jain, L. C., Teodoresku N. eds. Springer, 2002
2 H Liu, Motoda H. Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic, Boston: MA, 1998
3 Troyanskaya O, et al. Missing value estimation methods for DNA, Bioinformatics, 2001. 520～525
4 Kantardzic M. Data Mining Concepts, Models, Methods and Algorithms, Wiley-IEEE Computer Society Pr, 2003
5 Ian H. Witten and Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, 2000
6 Marshel A H. Bayesian Belief Network Using Conditional Phase-type Distributions: [PhD Thesis]. University of Ulster, 2001