灵敏度分析:分类器中的缺失数据

雷 蕾　吴乃君　刘 鹏　刘兰娟
(上海财经大学信息管理与工程学院)

摘要:在数据挖掘技术中,分类预测具有十分广泛的应用。由于数据集中总是存在着不同程度的数据缺失,降低了分类模型的预测准确率。主要通过灵敏度分析来研究缺失数据对分类算法的影响。对 6 种分类器进行实验,结果显示,当数据集中缺失数据超过 20 %,会对分类模型的预测准确率产生很大的不利影响,而且对于不同特征的数据集影响也不同。在这 6 种分类器中,朴素贝叶斯分类器对缺失数据最不敏感。对于目前流行缺失数据的处理方法——利用预测模型来预测并填补缺失数据,朴素贝叶斯分类器将是一个不错的选择。

关键词:分类器;缺失数据;灵敏度分析;数据挖掘
中图分类号:TP181　　**文献标识码**:A　　**文章编号**:1672 - 884 X(2005) S2 - 0153 - 05

Sensitivity Analysis : Missing Data in Classifiers

Lei Lei　Wu Naijun　Liu Peng　Liu Lanjuan
(Shanghai University of Finance & Economics , Shanghai ,China)

Abstract :Among all the technologies of data mining , predictive classification has a wide range of application. There are always missing data , which would affect the accuracy of classification models in the datasets. In this paper the influence of missing data on classifiers is investigated. First , basic knowledge of predictive classification is introduced briefly. Then , the sensitivity of six representative classifiers to missing data is studied by sensitivity experiments. The results indicated that when the proportion of missing data exceeds 20 % in the datasets , they do have a huge adverse effect on the prediction accuracy of the model. Moreover , missing data have different effects on different datasets , depending on their characteristics. Among the six classifiers , the Naive Bayesian classifier is the least sensitive to missing data.

Key words :classifier ; missing data ; sensitivity analysis ; data mining

数据挖掘是从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘有趣知识的过程[1]。数据挖掘的方法很多,常用的有描述、关联分析、分类预测、聚类分析等。其中,分类预测具有十分广泛的应用。然而,现实数据集中存在着许多数据质量问题,如数据不完整、数据冗余、数据不一致、噪音数据等。这些严重的数据质量问题都会降低数据挖掘算法的性能[2]。缺失数据就是一个非常普遍的数据质量问题。那么,缺失数据到底会对分类算法产生什么样的影响呢?各个分类器对缺失数据会产生什么样的反应呢?

对于现实数据集中的大量缺失数据,研究者们已经提出了许多缺失数据的处理方法。例如,样本丢弃法、均值替代法,模型预测法等。其中,模型预测法的基本思想就是利用已知数据建立预测模型,来预测数据集中的缺失数据并用预测值填补缺失值。缺失处理的效果取决于该预测模型。

本文将通过灵敏度分析来研究缺失数据对分类模型的影响程度。

1　分类预测

1.1　分类器

"分类"是指在已有数据的基础上学会一个分类函数或构造出一个分类模型,即通常所说的分类器[3],该函数或模型能够把数据库中的数据记录映射到给定类别中的某一个,从而可以在给定

收稿日期:2005 - 07 - 08

的条件下对人们感兴趣的未知变量做出预测[3]。许多因素都会影响分类器的预测准确率,但主要有 4 个[3]:

(1) 训练集的记录数量　分类器要利用训练集进行学习,因而训练集越大,分类器越可靠。相对地,构造分类器的时间也越长。

(2) 数据的质量　诸如噪音数据、缺失数据、不一致数据等的存在会引入大量的错误信息,从而无法创建出令人信服的分类器。

(3) 属性的质量　属性对于分类目标提供不同的信息。一般,一个属性是无法提供足够的信息来正确地预测分类的,如试图根据某人眼睛的颜色来决定他的收入。而加入其他的属性,如职业、每周工作小时数和年龄等,则可以提高预测准确率。但是,属性数量的增加意味着要计算更多的属性组合,这将大大增加分类器的生成难度和时间,因此,要选择对于分类目标最有价值的属性。

(4) 待预测记录的特性　如果待预测记录相异于训练集中的记录分布,那么错误率有可能会很高。显而易见,从轿车数据中构造出的分类器,用来对跑车记录进行分类就没有意义。

1.2　用分类器处理缺失数据

在目前最流行的缺失数据处理方法中,利用分类器建立预测模型来预测并填补缺失数据的方法是发展很快的一系列方法。在众多分类器中,选择哪一个分类器,以及如何利用这些分类器是该方法尚待研究的问题。目前已经提出的有:

(1) K-最邻近模型(KNN)　该模型使用 KNN 预测并替换缺失数据[4]。在数据集中某条记录只要少量属性出现缺失数据,此记录的基本特征、与其他记录的相似度等就会发生变化,从而导致分类错误的上升。

(2) C4.5 内置模型(C4.5)　该模型使用 C4.5 决策树预测并替换缺失数据。C4.5 决策树是最为广泛接受的一种分类器,嵌入式的缺失数据处理是其优点之一[5]。

(3) 基于朴素贝叶斯分类器的缺失数据处理模型 (NB)　该模型主要利用朴素贝叶斯分类器对缺失数据进行预测并替换[6],朴素贝叶斯分类器也是一种非常流行的分类器。

也有人使用神经网络来预测缺失数据。然而,不论采用何种分类算法,它们都是在含有缺失数据的数据集上进行训练并建立分类器,缺失数据对其知识发现过程都会产生或大或小的影响,其影响程度随具体的分类算法而定[4]。缺失数据对某分类算法的影响小,也就是说,在含有缺失数据的数据集上,该分类算法的预测准确率较高,其补缺的效果也会比较好。缺失数据对分类算法的影响程度可以是选择分类器的一个指标。

2　实验分析

2.1　灵敏度分析

计算机模型已被广泛地使用于实际问题的解决,一般模型都由许多参数组成,参数的变动会导致模型输出结果的变化。灵敏度分析(sensitivity analysis, SA) 就是研究一个或多个不确定性输入参数的变化对模型结果产生的影响,即模型对某个参数或参数组合变化的灵敏程度[7]。如果输入参数的微小变化导致了输出结果很大的变化量,则模型对该参数的变化非常灵敏。通过灵敏度分析可以识别对模型结果起决定性作用的输入参数,提高模型的可信度或预测准确性[8,9]。灵敏度分析的方法也有很多,不同学科领域的灵敏度分析各有自己的特点,总体上可以分为数学方法、统计方法和图形法[8]。

在本实验中,数据集含有的缺失数据比例就是影响分类模型输出结果的参数。本实验将检测在缺失比例有微小变化的情况下,分类器的预测准确率将会产生怎样的变动。

2.2　实验设计

本文选取了 6 种具有代表性的分类算法:朴素贝叶斯分类器(NB)、逻辑回归(LR)、神经网络向后传播分类(NN)、K-最邻近分类(KNN)、决策树 C4.5 以及逻辑分类树(LMT) 来研究缺失数据对分类模型的影响。实验中用到的 10 个数据集都来自 UCI[10],如表 1 所示。

表 1　数据集介绍

No.	Datasets	Records	Attr.	Classes
1	Breast	699	9	2
2	Bupa	345	6	2
3	Nursery	12 960	8	5
4	German	1000	24	2
5	Crx	690	15	2
6	Pima	768	8	2
7	Vehicle	846	18	4
8	Cmc	1473	9	3
9	Ionosphere	351	34	2
10	Segment	2310	19	7

相关的评测指标有 3 个:预测准确率、模型相对预测损失率以及预测收益[11],分别定义如下:

预测准确率 =

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

相对预测损失率 =

$$\frac{AC - \text{某缺失率下的预测准确率}}{AC} \times 100\%,$$

$$\text{预测收益} = \frac{\text{模型预测的准确率} - MD}{MD} \times 100\%,$$

式中,AC 为无缺失情况下模型的预测准确率;MD 为原始数据分布中所占比例最大的类的百分比[11],也就是说,如果不用任何预测模型,而将所有记录都归为该类的话,整个数据集的预测准确率为 MD。

预测准确率是最常用的评价分类器效果的指标。预测收益是文献[11]所提出的一个用来评价分类器效果的指标。这个指标比较了不使用分类器和使用分类器两种情况,显示了分类器的使用效果。该指标也可以显示缺失数据对分类器的实用性的影响。相对预测损失率是一个非常直接的显示指标,将无缺失情况与不同比例缺失情况进行比较。同时,除以 AC 可以去除量纲,使得不同数据集的实验结果可以综合分析。

在实验过程中,首先,随机地把每一个数据集分成两部分:2/3 的记录作为训练数据集,其余 1/3 作为测试数据集。此过程重复 10 次,生成 10 组训练集和测试集。然后,按缺失比率 10 %、20 %、…、90 %依次对训练数据集插入缺失数据。插入缺失数据的过程是完全随机的。最后,将上述 6 个分类算法应用于各个训练集建立分类预测模型,并用不含缺失数据的测试集进行检验,记录不同缺失率下的预测准确率。实验所得结果的均值显示于图 1 图 6。由于篇幅限制,只显示了部分实验结果。

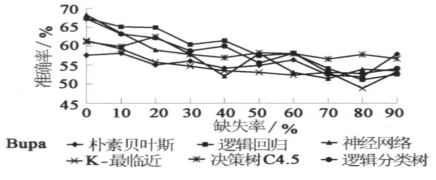

图 1 数据集 Bupa 的比较结果

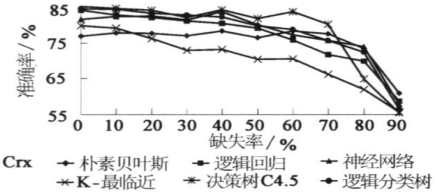

图 2 数据集 Crx 的比较结果

2.3 结果与分析

从图中可以看到,随着缺失比率的增大,所有分类预测算法的预测准确率都有明显的下降趋势。总体上来说,当缺失数据在数据集中所占比例少于 10 %的时候,缺失数据对分类器的影响很小。对于这 6 种分类器,由缺失数据所造成的预测损失率平均在 2 %左右(见表 2 和图 7)。若缺失数据的比例为 10 %20 %,缺失数据对分类器的影响是不容忽视的,6 种分类器的平均预测损失率上升到 5 %左右。但是,通过简单的处理,比如用近似值替代等,仍能很好地减少数据缺失对分类器所造成的不利影响。当缺失比例超过 20 %的时候,预测准确率有很明显的下降,这就需要谨慎处之。需要选择适当的缺失处理方法来消除缺失数据带来的不利影响,以提高分类器的性能。然而,当数据缺失率超过 50 %以后,6 种分类器的平均预测损失率超过了 10 %,数据集 Vehicle 还接近 25 %,显然,数据缺失所造成的预测准确率的损失是巨大的。这时,缺失数据处理技术所能挽回的损失是非常有限的。从图 7 中还可以看到,随着缺失比例的上升,预测损失率是加速上升的。

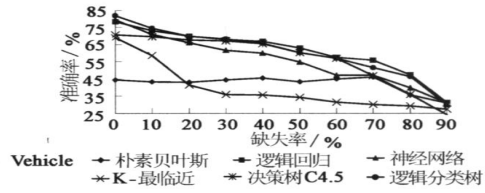

图 3 数据集 Vehicle 的比较结果

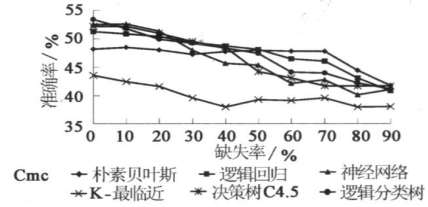

图 4 数据集 Cmc 的比较结果

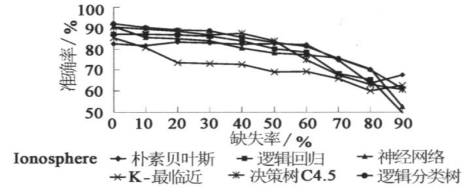

图 5 数据集 Ionosphere 的比较结果

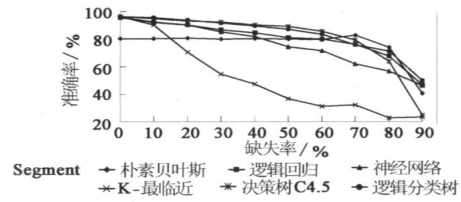

图 6 数据集 Segment 的比较结果

这说明,随着缺失数据数量的增加,缺失比例较小的上升,会引起预测准确率越来越大的下降,缺失数据对分类器的影响以指数形式上升。

表 2　6 种分类器的平均预测损失率

Missing / %	NB	LR	NN	KNN	C4.5	LMT	Average
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.08	1.93	3.27	4.29	0.87	2.53	2.16
20	0.59	2.85	5.2	13.22	1.6	4.27	4.63
30	0.47	4.65	8.00	16.45	3.19	5.65	6.40
50	1.24	8.41	12.45	22.45	6.38	9.57	10.08
70	2.70	14.11	18.25	24.95	13.29	15.37	14.78
80	9.73	19.76	20.93	29.03	23.37	18.55	20.23

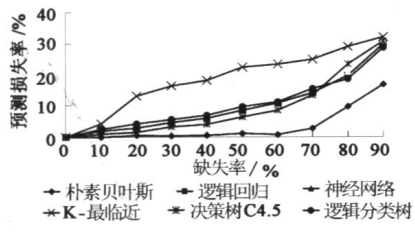

图 7　6 种分类器的平均预测损失率

数据集本身的特征也使得缺失数据对分类模型的影响不同,结构越是复杂的数据集对数据质量的要求越高。数据集预测类的类型越多,预测准确率下降的幅度也越大。在上述 10 个数据集中,6 个数据集的分类只有 2 类,数据集‘Cmc’分为 3 类,数据集‘Nursery’、‘Vehicle’、‘Segment’的分类均在 4 个以上。随着缺失比例的上升,除朴素贝叶斯分类器以外,各分类器对数据集‘Nursery’、‘Vehicle’、‘Segment’的预测准确率的下降幅度都比较大。当缺失比例达到 30 %的时候,这 3 个数据集的平均预测损失率超过 10 %,而其余数据集的平均预测损失率只有 4 %左右。

缺失数据对不同分类器影响程度也不同。相比较而言,在这 6 种分类器中,朴素贝叶斯分类器对缺失数据最不敏感,仅次于它的是决策树 C4.5;对缺失数据最敏感的是 K-最邻近分类,仅次于它的是神经网络算法。随着数据集中缺失比率的增加,朴素贝叶斯分类器的预测准确率一直接近无缺失值的情况,且比较稳定,只有当缺失比例超过 70 %时,才呈现明显的下降。从图 7 可以看到,朴素贝叶斯分类器的预测损失率一直是最低的,而 K-最邻近分类是最高的。而且 K-最邻近分类的上升速度也是最快的,在缺失率为 10 %的时候,其预测损失率就出现明显的上升。可以说 K-最邻近分类对缺失数据是非常敏感的,少量的数据缺失就会对该分类器造成很大的不利影响。从图 8 也可以看到,随着缺失比例的增加,各分类器的预测收益一直在下降,其中,朴素贝叶斯分类器的下降幅度最小,走势最平缓,而其他分类器都有明显的下降趋势。K-最邻近分类的预测收益曲线走势最为陡峭。总体来说,朴素贝叶斯分类器对缺失数据不敏感,即使训练集含有大量的缺失数据,仍能很好地运作。它最大化地利用了所有已知数据,且运算效率非常高。在众多分类器中,虽然,朴素贝叶斯分类器本身的预测准确率(即无缺失值时)并不是最高的(这个问题可以通过其他方法来改善),但它对缺失数据的适应性是最好的。因此,当数据集中有大量数据缺失的时候,朴素贝叶斯分类器是一个不错的选择。同时,在众多分类器中,选择用朴素贝叶斯分类器来处理缺失数据也能得到比其他分类器更好的结果。

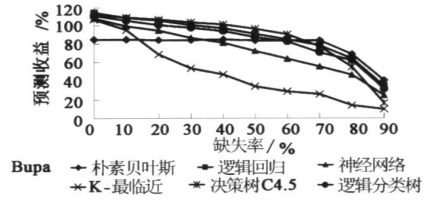

图 8　6 种分类器的预测收益

3　结论

数据集中总是存在着一些数据质量问题,这将影响分类模型的预测准确率。缺失数据就是一种非常典型的数据质量问题,可以说几乎所有的数据集都存在不同程度的缺失问题。目前,人们比较多地利用分类器来预测并填补缺失数据,如何选择合适的分类器便是一个问题。本文主要研究缺失数据对分类算法的影响,分析了缺失数据对 6 种分类器的影响程度。这 6 个分类算法分别是朴素贝叶斯分类器、逻辑回归、神经网络向后传播分类、K-最邻近分类、决策树 C4.5 以及逻辑分类树。实验显示,随着缺失比率的增大,所有分类预测算法的预测准确率都有明显的下降趋势。当数据集中缺失数据超过 20 %的时候,预测准确率有很明显的下降,这就要求谨慎地选择适当的缺失处理方法来提高分类器的预测准确率。在这 6 个分类器中,朴素贝叶斯分类器对缺失数据最不敏感,K-最邻近算法对缺失数据最敏感。综上所述,对于含有大量缺失的数据集,选择朴素贝叶斯分类器来处理缺失数据能得到令人满意的效果。在笔者的另一篇论文[12]中有对基于朴素贝叶斯分类器的缺失数据处理方法的详细讨论以及与其他分类器的比较。如果不对缺失数据进行处理,

直接运用分类器来完成分类预测的任务,朴素贝叶斯分类器应是一个不错的选择。

参 考 文 献

[1] Han J, Kamber M. Data Mining Concepts and Technique[M]. USA:Morgan Kaufmann Publishers, 2000.

[2] Cios K J, Kurgan L. Trends in Data Mining and Knowledge Discovery[R]. In N. R. Pal, L. C. Jain, 2002.

[3] Tian Jinlan, Li Ben. Tools for Data Mining: Classifiers[M]. Beijing:China Computerworld Corporation,Department of Computer Science, China, Tsinghua University. Computer World, 20th Periodical,1999.

[4] Marvin L Brown, John F Kros. Chapter VII-The Impact of Missing Data on Data Mining[M]. Data Mining: Opportunities and Challenges, Idea Group Publishing, USA, 2003.

[5] Quinlan J R. C4.5 Programs for Machine Learning[M]. Morgan Kaufmann, CA, USA, 1993.

[6] Liu P, Lei L, Zhang X F. A Comparison Study of Missing Value Processing Methods[J]. Computer Science, 2004,31(10): 155156

[7] Yao J T. Sensitivity Analysis for Data Mining[A]. Proceedings of The 22nd International Conference of NAFIPS[C]. July 24 - 26, Chicago, USA, 2003:272277

[8] Frey H C, Patil S. Identification and Review of Sensitivity Analysis methods[J]. Risk Analysis,2002,22(3):553578

[9] Joao W Cangussu, Raymond A DeCarlo, Aditya P Mathur. Using Sensitivity Analysis to Validate a State Variable Model of the Software Test Process[J]. IEEE Trans, Software Eng, 2003, 29(5):430443

[10] Merz C J, Murphy P M. UCI Repository of Machine Learning Datasets[EB]. http://www.ics.uci.edu/ mlearn/MLRepository.html,1998

[11] Peng Liu, Elia El-Darzi et al. Comparative analysis of Data Mining Algorithms for Predicting Inpatient Length of Stay[A]. Proceedings of the Eighth Pacific-Asia Conference on Information Systems[C]. 2004.

[12] Liu P,Lei L. An Analysis of Missing Data Treatment Methods and the Application of Naive Bayesian Classifier[A]. China: CSCA, 2005.

作者简介:雷蕾(1982～),女,汉族,上海人。上海财经大学(上海市 200433)信息管理与工程学院硕士研究生。研究方向为数据挖掘。