一种有效的不完整数据分类器*)

陈景年[1,2] 黄厚宽[1] 田凤占[1] 邱桃荣[1]

(北京交通大学计算机与信息技术学院 北京 100044)[1] (山东财政学院信息与计算科学系 济南 250014)[2]

摘 要 在模式识别、机器学习以及数据挖掘中,分类是一个基本而又重要的问题。虽有大量的分类器应运而生,但由于处理不完整数据的复杂性,它们大都是针对完整数据的。然而,由于各种原因,现实中的数据通常是不完整的。因此,对不完整数据分类器的研究具有重要意义。通过分析以往在分类过程中对不完整数据的处理方法,提出了一种不完整数据分类器:DBCI。在DBCI的训练过程中,将缺失值的频数按比例地分配到其它观测值的频数中。因此,不完整数据集所包含的信息可以得到充分利用。在12个标准的不完整数据集上的实验结果表明,与分类效果显著的不完整数据分类器RBC相比,DBCI具有更高的分类效率和更稳定的性能,并且它的分类准确率可以与RBC相媲美。

关键词 分类,贝叶斯方法,不完整数据

Effective Classifier for Incomplete Data

CHEN Jing-nian[1,2] HUANG Hou-kuan[1] TIAN Feng-zhan[1] QIU Tao-rong[1]

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)[1]

(Department of Information and Computing Science, Shandong University of Finance, Jinan 250014, China)[2]

Abstract Classification is an elementary and important problem in pattern recognition, machine learning and data mining. Though many classifiers have been proposed, most of them deal with complete data, which is due to the complexity of dealing with incomplete data. Yet actual data sets are often incomplete because of various kinds of reason. So the study of classifiers for incomplete data is of great significance. With the analysis of main methods of processing incomplete data for classification, a new classifier for incomplete data denoted as DBCI is presented. In the training process of DBCI, frequencies of missing values are distributed proportionally across frequencies of other observed values. So the information contained in incomplete datasets can be sufficiently utilized. Experiments are carried out on twelve benchmark incomplete data sets. Compared to the remarkable Robust Bayes Classifier (RBC) that is very effective for classifying incomplete data, DBCI is more efficient and more stable and its classification accuracy is comparable to that of RBC.

Keywords Classification, Bayesian method, Incomplete data

分类是模式识别、机器学习以及数据挖掘中一个基本而又重要的问题。因此,一些有效的分类算法应运而生。

由于处理不完整数据的复杂性,以往的分类器大都是针对完整数据的。然而,由于各种原因,实际中的数据通常是不完整的。因此,对用于不完整数据的分类器的研究具有重要的意义。

为了在不完整数据集上构造分类器,下面先简单回顾一下在分类过程中对不完整数据的主要的处理方法。

以往有关不完整数据分类方法的文献并不多见。已有的能够处理不完整数据的分类器,例如朴素贝叶斯分类器和C4.5决策树,在遇到不完整数据时通常采用的简单方法是丢弃包含缺值的数据项,或者针对不同的属性分别设置某个特定的取值[1]。简单丢弃法浪费了部分数据样本中的信息,在数据样本较少或者代价昂贵时不宜采用;而设置特定数值的方法容易产生数据的偏斜[2],从而容易引起大的估计偏差。

Friedman等人[3]建议采用EM算法[4]、梯度下降算法[5]或者Gibbs采样算法[6]对缺值数据进行修补,之后在得到的完整数据集上构建分类器。但是,上述方法都基于缺值数据满足MAR (Missing at Random) 假设[7]。当不满足这个假设时,这些数据修补方法的精度会显著下降,由此构造的分类器的精度也会下降[8]。

为避免MAR假设,Ramoni与Sebastiani提出了一种RBC(Robust Bayes Classifier)分类器[9],该方法可直接从不完整数据构造贝叶斯分类器,也不要求缺值数据满足MAR假设,并且这种方法具有较高的分类效率和精度。

但是,RBC分类器对缺失数据是采用区间来估计的,这使得对缺失数据的估计信息显得粗糙且不直观,而且复杂的估计公式,增加了计算的复杂度。

针对上述问题,本文采用一种较为直观有效的不完整数据处理方法构造了不完整数据分类器DBCI (Distribution-based Bayes Classifiers for Incomplete data)。然后,通过在12个标准的不完整数据集上的实验,对提出的DBCI算法和分类效果显著的不完整数据分类器RBC进行了比较和分析。最后对本文的工作进行了总结。

*)本文受国家自然科学基金(60503017和60673089)资助。陈景年 博士研究生,副教授,CCF会员,主要研究方向为模式识别、机器学习、数据挖掘;田凤占 博士,副教授,CCF会员,主要研究领域为贝叶斯网络、机器学习、数据挖掘;黄厚宽 教授,博士生导师,CCF会员,主要研究领域为人工智能、模式识别、数据仓库、数据挖掘以及多智能体系统;邱桃荣 博士研究生,副教授,主要研究领域为人工智能、数据挖掘、数据库。

1 DBCI 分类器

为避免上文提到的朴素贝叶斯分类器和 RBC 分类器中的问题，在接下来的部分给出了 DBCI 分类器。DBCI 分类器将每个属性 A 与类变量 C 的缺失值的频数(包括 A 缺失，C 不缺失；A 不缺失，C 缺失；A 和 C 都缺失三种情况)进行统计，并且将这些频数根据 A 和 C 的各个观察值的频数按比例地分配到各有关的频数中。因此，DBCI 能够充分利用观测值的信息对缺失数据进行处理，并且具有相对较低的计算复杂度。

DBCI 的训练过程可描述如下：

输入：不完整数据集 D 作为训练集，设 D 有 n 个属性。

输出：分类所需要的任一属性 A 的类条件概率估计和类变量 C 的概率估计。

过程：

(1)扫描数据集 D，统计各有关频数。设类变量 C 的所有可能的取值为 l 个(即有 l 个类)。统计 C 取每个值 c_j 的频数 $f(c_j)$，$(j=1,2,\cdots,l+1)$，其中 $f(c_{l+1})$ 表示类值缺失的实例数。对每一属性 A，设 A 的所有可能的取值为 m 个。统计 A 取第 i 个值 a_i 且 C 取第 j 个值 c_j 的实例的频数 f_{ij}，$i=1,2,\cdots,m+1$，$j=1,2,\cdots,l+1$。其中 $f_{(m+1)j}$，$f_{i(l+1)}$ 及 $f_{(m+1)(l+1)}$ 分别表示 A 的值缺失而 C 取第 j 个值的实例的频数；A 取第 i 个值而 C 的值缺失的实例的频数以及 A 和 C 的值都缺失的实例的频数。

(2)将类值缺失的实例的频数根据各个类值的频数按比例地分配到各相应类值的频数中：

首先计算和式 $sc=\sum_{j=1}^{l} f(c_j)$，然后令：

$$f'(c_j)\leftarrow f(c_j)+f(c_{l+1})\times\frac{f(c_j)}{sc}。$$

(3)将 $f_{(m+1)j}$，$f_{i(l+1)}$ 及 $f_{(m+1)(l+1)}$ 根据 A 和 C 的各个观察值的频数按比例地分配到各有关频数中：

首先，对于 $i=1,2,\cdots,m$，$j=1,2,\cdots,l$，计算和式：

$$r_i=\sum_{j=1}^{l} f_{ij}\ ,s_j=\sum_{i=1}^{m} f_{ij}\ 以及\ s=\sum_{i=1}^{m} r_i$$

然后 $f'_{ij}\leftarrow f_{ij}+f_{i(l+1)}\times\frac{s_j}{s}+f_{(m+1)j}\times\frac{r_i}{s}+f_{(m+1)(l+1)}\times\frac{f_{ij}}{s}$。

(4)利用 $f'(c_j)$ 和 f'_{ij} 计算类变量 C 的概率估计 $P(C=c_j)$(简记为 $P(c_j)$)和任一属性 A 的类条件概率估计 $P(A=a_i|C=c_j)$(简记为 $P(a_i|c_j)$)：

$$P(c_j)=\frac{f'(c_j)+1}{\sum_{k=1}^{l} f'(c_k)+l}\ ,P(a_i|c_j)=\frac{f'_{ij}+1}{\sum_{i=1}^{m} f'_{ij}+m}。$$

设有未分类实例 $e=(e_1,e_2,\cdots,e_n)$，对 e 的分类过程为：

对每一 $j=1,2,\cdots,l$，由于 $P(c_j|e)\propto P(c_j,e)$，由条件独立假设，$P(c_j,e)=P(c_j)\prod_{k=1}^{n} P(e_k|c_j)$。令 $c=\arg\max_{1\leqslant j\leqslant l}\{P(c_j,e)\}$，则 c 为 e 的类别。

需要指出的是 DBCI 只考虑字符型属性变量，对于包含数值型属性的数据集要先进行离散化。

2 实验结果及分析

2.1 实验数据集

为了验证所提出的算法的有效性，我们在 12 个包含缺失数据的数据集合上进行了实验。这 12 个数据集合均来自 UCI 机器学习知识库[10]。表 1 对这 12 个数据集进行了描述。数据集中实例个数从最多 8124 到最少 32 个，属性个数从最多 279 个到最少 10 个，分别分布在一个很宽的范围内。

表 1 实验中用到的数据集

Data sets	Instances	Classes	Attributes
Annealing	798	5	38
Arrhythmia	452	16	279
Audiology	200	2	70
B. cancer	699	2	10
Bridges	108	6	12
Credit	690	2	15
Cylinder	512	2	39
Echocardiogram	132	2	12
Horse-colic	368	2	27
L. cancer	32	3	56
Mushroom	8124	2	22
Vote	435	2	16

2.2 实验结果与分析

所有实验是在 weka 环境[11]下，在内存为 1GB，主频为 2.93 GHz 的 Pentium IV PC 机上进行的。为验证提出的不完整数据分类器 DBCI 的有效性，将它与分类效果显著的不完整数据分类器 RBC 进行比较。对数值型属性，使用“weka. filters. supervised. attribute. Discretize”进行离散化。

表 2 列出了 RBC 和 DBCI 在每个数据集上的 10 次 10 重交叉验证的平均准确率及相应的标准离差。通过显著程度为 95%的双尾 t 测试来检验两种分类器在每个数据集上的分类准确率有无显著差别。在每个数据集上显著高的准确率以粗体表示。如果在一个数据集上两种分类器的分类准确率无显著差别，则都以粗体表示。另外，为了比较这两种算法的运行效率，还列出了它们在每个数据集上的运行时间。在表 2 的底部列出了这两种分类器在 12 个数据集上的分类准确率的平均值、标准离差的平均值及运行时间的平均值。

表 2 RBC 与 DBCI 的分类准确率和运行时间

Data sets	Classification accuracy		Runtime(second)	
	RBC	DBCI	RBC	DBCI
Annealing	95.96±0.31	92.74±0.30	2.92	2.83
Arrhythmia	72.77±0.89	73.13±0.70	30.64	26.86
Audiology	67.99±0.79	68.64±0.83	4.09	1.22
B. cancer	97.11±0.16	97.02±0.06	1.47	1.49
Bridges	61.62±2.20	64.00±1.61	0.56	0.39
Credit	86.18 0.40	85.70±0.34	1.55	1.45
Cylinder	71.36±0.48	75.37±1.19	2.95	2.50
Echocardiogram	98.36±0.87	97.26±0.00	0.75	0.69
Horse-colic	85.20±0.59	83.71±0.54	1.39	1.39
L. cancer	56.13±1.67	56.13±1.67	0.45	0.33
Mushroom	95.96±0.02	95.93±0.05	8.72	8.80
Vote	90.25±0.19	90.18±0.25	0.58	0.50
Average	81.57±0.71	81.65±0.63	4.67	4.04

从表 2 可以看出如下几点：

(1) DBCI 与 RBC 相比，在所有实验数据集中的 3 个数据集上，其分类准确率明显高于 RBC 的分类准确率。尤其是在数据集 Cylinder 上，DBCI 的分类准确率比 RBC 的分类准确率高出 4.01%，在 4 个数据集上它们的分类准确率没有明显差别，在其余 5 个数据集上 DBCI 的分类准确率低于 RBC 的分类准确率，而 DBCI 在 12 个数据集上的准确率的平均值比 RBC 的要高。因此，总体来看 DBCI 的分类准确率与 RBC 的分类准确率大致相当。

(2) 比较 DBCI 与 RBC 在 12 个数据集上的标准离差可以发现,在 7 个数据集上 DBC 的标准离差低于 RBC,在 1 个数据集上二者的标准离差相同,在其余 4 个数据集上,DBC 的标准离差高于 RBC,而 DBCI 在 12 个数据集上的标准离差的平均值也明显比 RBC 的要低。因此,总体来说 DBCI 的分类性能比 RBC 的更加稳定。

(3)从 DBCI 和 RBC 的运行时间来看,在所有实验数据集中的 9 个数据集上,DBCI 的运行时间少于 RBC 的运行时间,在 1 个数据集上二者的运行时间相同,只在其余 2 个数据集上 DBCI 的运行时间略多于 RBC 的运行时间。而 DBCI 在 12 个数据集上的平均运行时间也明显比 RBC 的要少。因此,从总体来看 DBCI 的运行效率明显高于 RBC。

结束语 分类是模式识别、机器学习以及数据挖掘中一个基本而又重要的问题。虽然大量有效的分类算法应运而生,但由于处理不完整数据的复杂性,以往的分类器大都是针对完整数据的。然而,实际中的数据通常是不完整的。因此,对不完整数据分类器的研究具有重要的意义。

本文通过分析已有的在分类过程中处理不完整数据的方法,提出了一种不完整数据分类器:DBCI。为验证 DBCI 的有效性,将它与分类效果显著的不完整数据分类器 RBC 进行了实验比较。在 12 个标准的不完整数据集上的实验结果显示,DBCI 的分类准确率与 RBC 的分类准确率大致相当。但 DBCI 的运行效率明显高于 RBC,而且其分类性能比 RBC 更加稳定。

参 考 文 献

[1] Quinlan J R. C4. 5:Programs for Machine Learning [M]. San Francisco:Morgan Kaufmann,1993

[2] Kohavi R, Becker B, Sommerfield D. Improving simple Bayes [C]// M. van Someren, G. Widmer, eds. Poster Papers of the ECML-97. Prague:Charles University,1997:78-87

[3] Friedman N,Geiger D,Goldszmidt M. Bayesian network classifiers [J]. Machine Learning,1997,29 (2/3):131-163

[4] Lauritzen S L. The EM algorithm for graphical association models with missing data [J]. Computational Statistics and Data Analysis,1995,19(2):191-201

[5] Russell S,Binder J,Koller D,et al. Local learning in probabilistic networks with hidden variables[C]// Proc. of IIJCAI-95. San Francisco:Morgan Kaufmann,1995:1146-1151

[6] Geman S, Geman D. Stochastic relaxation, Gibbs distributions andthe Bayesian restoration of images [J]. IEEE Trans on Pattern Analysis and Machine Intelligence,1984,6(6):721-741

[7] Little R J A,Rubin D B. Statistical Analysis with Missing Data [M]. New York:Wiley,1987

[8] Spiegelhalter D J,Cowell R G. Learning in probabilistic expert systems[C]// J Bernardo, J Berger, A PDawid, eds. Bayesian Statistics 4,Oxford:Oxford University Press,1992:447- 466

[9] Ramoni M,Sebastiani P. Robust Bayes classifiers [J]. Artificial Intelligence,2001,125(1/2):209-226

[10] Blake C,Keogh E,Merz C J. UCI Repository of machine learning databases [OL]. Department of Information and Computer Sciences,University of California,Irvine. http:// www. ics. uci. edu/~mlearn /MLRepository. html],1998

[11] Weka: Data Mining Software in Java [OL]. http://www. cs. waikato. ac. nz/ ml/weak,2007

(上接第 152 页)

从 FC-tree 中删除,这样可以以 2 的指数级减少搜索节点。如本文图 1 所示,如果节点{2}从树中删除,则其后继节点{2,3},{2,5},{2,6},{2,3,5},{2,3,6},{2,5,6},{2,3,5,6}也必不满足最小频繁闭项目集约束条件,从而应被剪枝。

为了进一步验证算法 Max-FCIA 的优越性,我们用 VC++ 6.0 在内存 512M、操作系统为 Windows XP 的机器上实现了 Max-FCIA 算法和 Max-Mine 算法。Max-Mine 算法是一种基于树型结构和 Apriori 算法思想的最大频繁项目集挖掘算法,它比 Apriori 算法有很大优势。本实验使用了 chess 数据集中的 3196 条数据,实验结果如图 4 所示,图 4 表明 Max-FCIA 算法是有效的。

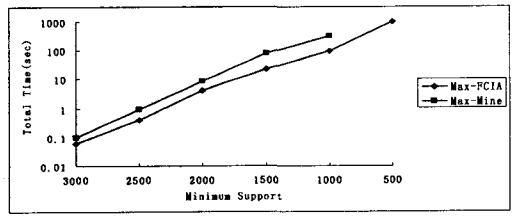

图 4 算法执行时间比较

结束语 频繁闭项目集提供了事务数据库的一个最小描述,其数量介于最大频繁项目集和频繁项目集之间,同时记录了所有频繁项目集的支持度,因此发现频繁闭项目集对数据挖掘具有十分重要的意义。本文系统地给出了频繁闭项目集及其存储结构 FC-tree 和挖掘算法 Max-FCIA,并对一些关键技术进行了改进,实验表明该算法是有效可行的。今后研究工作的重点将放在频繁闭项目集挖掘算法改进和增量式更新方面。

参 考 文 献

[1] Pasquier N, Bastide Y, Taouil R, et al. Discovering frequent closed itemsets for association rules// Beeri C, et al. , eds. Proc. of the 7th Int'l. Conf. on Database Theory. Jerusalem: Springer-Verlag,1999:398-416

[2] Agrawal R, Srikant R. Fast algorithms for mining association rules// Beeri C, et al. , eds. Proc. of the 20th Int'l. Conf. on Very Large Databases. Santiago: Morgan Kaufmann Publishers, 1994:487-499

[3] Pei J, Han J, Mao R. CLOSET: An efficient algorithm for mining frequent closed itemsets// Gunopulos D, et al. , eds. Proc. of the 2000 ACM SIGMOD Int' l. Workshop on Data Mining and Knowledge Discovery. Dallas: ACM Press,2000:21-30

[4] Burdick D, Calimlim M, Gehrke J. MAFIA: A maximal frequent itemset algorithm for transactional databases// Georgakopoulos D, et al. , eds. Proc. of the 17th Int'l. Conf. on Data Engineering. Heidelberg: IEEE Press,2001:443-452

[5] Zaki M J, Hsiao C J. CHARM: An efficient algorithm for closed itemset mining// Grossman R, et al. , eds. Proc. of the 2nd SIAM Int'l. Conf. on Data Mining. Arlington: SIAM,2002:12-28

[6] 朱玉全,杨鹤标,孙蕾. 数据挖掘技术[M]. 南京:东南大学出版社,2006

[7] 朱玉全,宋余庆. 频繁闭项目集挖掘算法研究[J]. 计算机研究与发展,2007,44(7):1177-1183