计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
44(8):1324~1330, 2007

用于不完整数据的选择性贝叶斯分类器

陈景年[1,2] 黄厚宽[1] 田凤占[1] 付树军[1]
[1](北京交通大学计算机与信息技术学院 北京 100044)
[2](山东财政学院信息与计算科学系 济南 250014)
(jnchen06@163.com)

Selective Bayes Classifiers for Incomplete Data

Chen Jingnian[1,2], Huang Houkuan[1], Tian Fengzhan[1], and Fu Shujun[1]
[1](*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)
[2](*Department of Information and Computing Science, Shandong University of Finance, Jinan* 250014)

Abstract Selective classifiers have been proved to be a kind of algorithms that can effectively improve the accuracy and efficiency of classification by deleting irrelevant or redundant attributes of a data set. Though some selective 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 and have many redundant or irrelevant attributes because of various kinds of reason. Similar to the case of complete data, irrelevant or redundant attributes of an incomplete data set can also sharply reduce the accuracy of a classifier established on this data set. So constructing selective classifiers for incomplete data is an important problem. With the analysis of main methods of processing incomplete data for classification, two selective Bayes classifiers for incomplete data, which are denoted as SRBC and CBSRBC respectively, are presented. While SRBC is constructed by using the robust Bayes classifiers, CBSRBC is based on SRBC and chi-squared statistics. Experiments on twelve benchmark incomplete data sets show that these two algorithms can not only enormously reduce the number of attributes, but also greatly improve the accuracy and stability of classification as well. On the whole, CBSRBC is more efficient than SRBC and its classification accuracy is higher than that of SRBC. But some thresholds necessary to CBSRBC can be avoided by SRBC.

Key words Bayesian method; classification; feature selection; incomplete data; chi-squared statistics

摘 要 选择性分类器通过删除数据集中的无关属性和冗余属性可以有效地提高分类精度和效率.因此,一些选择性分类器应运而生.然而,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.由于各种原因,现实中的数据通常是不完整的并且包含许多冗余属性或无关属性.如同完整数据的情形一样,不完整数据集中的冗余属性或无关属性也会使分类性能大幅下降.因此,对用于不完整数据的选择性分类器的研究是一项重要的研究课题.通过分析以往在分类过程中对不完整数据的处理方法,提出了两种用于不完整数据的选择性贝叶斯分类器:SRBC和CBSRBC.SRBC是基于一种鲁棒贝叶斯分类器构建的,而CBSRBC则是在SRBC基础上利用χ^2统计量构建的.在12个标准的不完整数据集上的实验结果表明,这两种方法在大幅度减少属性数目的同时,能显著提高分类准确率和稳定性.从总体上来讲,CBSRBC在分类精度、运行效率等方面都优于SRBC算法,而SRBC需要预先指定的阈值要少一些.

关键词 贝叶斯方法;分类;特征选择;不完整数据;χ^2统计量

中图法分类号 TP181

收稿日期:2007-01-14;**修回日期**:2007-05-11
基金项目:国家自然科学基金项目(60503017,60673089)

通过删除数据集中的无关属性和冗余属性,在精选出的属性子集上构建的选择性分类器可以有效地提高分类精度和效率.因此,一些有效的选择性分类器应运而生.

由于处理不完整数据的复杂性,以往的选择性分类器[1-3]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并且包含许多无关属性和冗余属性.因此,对用于不完整数据的选择性分类器的研究是一项重要的研究课题.

为了在不完整数据集上构造选择性分类器,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,或者针对不同的变量分别设置某个特定的取值[4].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[5],从而容易引起大的估计偏差.

Friedman 等人[6]建议采用 EM 算法[7],梯度下降算法[8]或者 Gibbs 采样算法[9]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[10].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[11].

为避免 MAR 假设,Ramoni 与 Sebastiani 提出了一种 RBC(robust Bayes classifier)分类器[12],该方法可以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,并且这种方法具有很高的分类效率.但是,与朴素贝叶斯分类器相似,RBC 也是假定在给定类变量时,各个属性变量之间是相互独立的.当这一前提假设不成立时,分类精度往往会显著下降.

通过上述关于不完整数据分类方法的分析,本文首先基于 RBC 和包装法(wrapper)[13]构建了一种用于不完整数据的选择性贝叶斯分类器 SRBC.然后,为提高 SRBC 的效率,在 SRBC 基础上利用 χ^2 统计量属性评价算法[14]给出了一种基于混合特征选择方法的选择性贝叶斯分类器 CBSRBC.此后,通过在 12 个标准的不完整数据集上的实验,对提出的两种算法与 RBC 算法进行了比较和分析.最后对本文的工作进行了总结.

1 SRBC 分类器

SRBC(selective robust Bayes classifiers)是在 RBC 基础上利用包装法构建的用于不完整数据的选择性贝叶斯分类器.

为什么选取 RBC 来构建选择性分类器呢?

通过前面对用于不完整数据的分类算法的分析,可以将构建 SRBC 的原因和依据归结如下:

首先,在主要的用于不完整数据的分类算法中,例如 EM 算法、Gibs 抽样法等,虽然有时候能得到较好的分类效果,但它们都基于缺值数据满足 MAR 假设,而且计算复杂度一般都很高.因此,难以利用它们构造用于不完整数据的选择性分类器.再就是朴素贝叶斯分类器,虽然也可对不完整数据进行分类,而且分类效率很高,但它对不完整数据采取的两种简单处理方法:删除包含缺失数据的实例和给缺失数据指定一个虚拟值,都可能会引起大的估计偏差.RBC 分类器既不需要缺值数据满足任何假设,也不会引起大的估计偏差.其计算复杂度不高,而且分类准确率一般要比朴素贝叶斯方法的要高,完全克服了上述方法的不足.因此,对构建基于包装方法的选择性分类器来说,RBC 是一个理想的选择.

其次,前面已经提及,RBC 是以各个属性变量条件独立为前提的,而这一假定在实际中多数情况下不成立,这往往会导致分类准确率降低.因此,通过构建 SRBC,不仅可以避免 MAR 假设,而且属性变量条件独立假设也可避免.这正是构建 SRBC 的依据.

在给出算法 SRBC 之前,首先对 RBC 作进一步介绍.

Ramoni 与 Sebastiani 提出的 RBC 是一种在不完整数据集上构建的贝叶斯分类器.RBC 的训练过程是先在所给的不完整数据集上计算有关的不完整实例的频数,然后利用这些频数计算出各个属性变量的类条件概率分布的估计区间以及类变量的边缘分布的估计区间.在计算各个估计区间的过程中,不需要缺值数据满足 MAR 假设.在分类过程中,首先利用上述估计区间求出在给定新实例的条件下,类变量后验概率的估计区间.然后,通过给区间打分,将新实例分到最高分值关联的类中.在计算类变量后验概率的估计区间时,假定各个属性变量之间是类条件独立的.这种方法具有很高的分类效率.与朴素贝叶斯分类器相比,RBC 一般具有更高的分

类精度,分类性能更稳定.

在利用包装法构建 SRBC 的过程中,我们采用了搜索效果好而复杂度相对较低的最优优先(best first search)前向搜索方法[15]对属性空间进行搜索.

记 $A=\{a_1,a_2,\cdots,a_N\}$ 为整个属性集合,N 为 A 中属性的个数. Q 为一个队列,用来存放曾经是最优的属性子集及其对应的分类精度. S_b 为当前最优属性子集,$f(S)$ 表示 RBC 在属性子集 S 上的分类精度. 阈值 T 为用来控制搜索过程是否停止的参数,即如果连续 T 次对 Q 的头结点进行扩展都没有使当前最高分类精度改善,则搜索过程结束.

算法 SRBC 可描述如下:

① 初始化. 设置参数 T,令整数 $t=0$. 令属性 $a_s=\arg\max_{1\leqslant i\leqslant N}\{f(\{a_i\})\}$,当前最高分类精度 $f_{\max}=f(\{a_C s\})$. 将属性子集 $\{a_s\}$ 作为一个结点加入到队列 Q 中.

② 当 t<T 时执行步骤③④和⑤,否则,执行步骤⑥:

③ 取出 Q 的头结点 S_h(为一属性子集),令 added = *false* (added 用来标志在对 Q 的头结点的扩展中,是否向 Q 中加入了新的结点). 对每一属性 $a\in A-S_h$,如果 $S_h\cup\{a\}$ 没有被评价过,而且 $f(S_h\cup\{a\})>f_{max}$,那么,令 added = *true*,$S_b=S_h\cup\{a\}$,$f_{max}=f(S_h\cup\{a\})$,以及 t = 0,并且将 S_b 作为一个新结点加入到队列 Q 中.

④ 如果 added = *false*,那么 t←t+1.

⑤ 转到步骤②继续执行.

⑥ 在最终的属性子集 S_b 上构建 *RBC* 分类器.

在构建 SRBC 的过程中,每评价一个属性子集,就要构建一个 RBC 分类器. 因此 SRBC 的计算复杂度一般很高,尤其当属性个数较多时更是如此. 接下来的第 2 节通过利用 χ^2 统计量给出了一个更加高效的选择性分类器.

2 CBSRBC 分类器

在构建 CBSRBC (chi-square-based selective robust Bayes classifiers)过程中,利用 χ^2 统计量属性评价算法[14]来评价一个属性变量与类变量之间的相关程度. 一个属性变量与类变量之间的 χ^2 统计量越大,表明该属性变量与类变量越相关. 根据 χ^2 统计量的大小可以删除那些与类变量相关程度较低的属性.

χ^2 统计量计算复杂度相对较低. 其时间主要花费在构建每个属性与类变量之间的列联表上. 而要构建所有的列联表,只需对数据集扫描一次. 因此,该算法对于实例数目和属性数目都具有很好的扩展性. 在这方面要比包装方法(其时间开销随属性数目的增长呈指数增长趋势)好得多. 因此,χ^2 统计量属性评价算法能够用于具有大量属性或实例的大型数据集. 这也正是我们选择它来构造选择性分类器的原因之一.

在构建属性 A 与类变量 C 之间的列联表时,χ^2 统计量属性评价算法对 A 和 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 3 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,该算法能够充分利用观测值的信息处理缺失数据.

然而,χ^2 统计量属性评价算法不能够很好地发现属性集中的冗余属性. 当该算法直接用来构造选择性分类器时,一般难以获得理想的效果. 与此相反,SRBC 一般能得到较好的选择效果,而复杂度较高. 将 χ^2 统计量属性评价算法与 SRBC 结合,既可以利用前者高的计算效率,又可以利用后者好的选择效果;既可以通过前者选择出相关的属性,又可以通过后者去除冗余属性,从而达到在保持 SRBC 的分类精度不下降的前提下提高其计算效率的目的. 这正是我们构造 CBSRBC 的依据. 第 3 节的实验结果表明了 CBSRBC 的有效性.

CBSRBC 算法可描述如下:

1) 设置利用 χ^2 统计量属性评价算法选择的属性数目 T_c.

2) 扫描数据集 D,统计各有关频数(包括缺失数据的频数).

3) 对每个属性 A,构造 A 与类变量 C 之间的列联表 M. 在构造表 M 时,将 A 和 C 的缺失值的频数分别根据 A 和 C 的各个观察值的频数按比例地分配到 M 中相应元素表示的频数中.

4) 对于每一属性 A,计算 A 与类变量 C 之间的 χ^2 统计量. 假设数据集 D 包含 n 个实例,并假设 A 与 C 之间的列联表 M 为 m 行 c 列,其中 m 为属性 A 所有可能的取值个数,c 为类变量 C 的取值个数. M 中第 i 行第 j 列上的元素 f_{ij} 为数据集 D 中 A 取其第 i 个值并且 C 取其第 j 个值的所有实例的个数,即频数. A 与 C 之间的 χ^2 统计量可按下列步骤计算:

① 计算 M 的每一行之和 r_i 与每一列之和 s_j:

$$r_i=\sum_{j=1}^{c}f_{ij},i=1,2,\cdots,m;$$

$$s_j = \sum_{i=1}^{m} f_{ij}, j = 1,2,\cdots,c.$$

② 对每一对 i,j,计算 A 取其第 i 个值并且 C 取其第 j 个值的期望频数 e_{ij}:

$$e_{ij} = \frac{r_i \times s_j}{n}, i = 1,2,\cdots,m, j = 1,2,\cdots,c.$$

③ 计算 A 与 C 之间的 χ^2 统计量 $Chi(A,C)$:

$$Chi(A,C) = \sum_{i=1}^{m}\sum_{j=1}^{c} \frac{(e_{ij} - f_{ij})^2}{e_{ij}}.$$

5) 取前 T_c 个 χ^2 统计量最大的属性,并记它们构成的集合为 S_t.

6) 在属性子集 S_t 上执行算法 SRBC.

需要指出的是,在上述 CBSRBC 算法中,可以用其他的参数,比如 χ^2 统计量的大小来取代参数 T_c,只是设置选择的属性数目 T_c 更简便而已. 另外,RBC,SRBC,以及 CBSRBC 算法只考虑有限状态属性变量. 当有数值型属性变量时,需要先进行离散化处理.

3 实验结果及分析

3.1 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[16].

表 1 对这 12 个数据集进行了描述,从上到下按照数据集合中实例个数从大到小顺序依次排列. 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.

Table 1 Data Sets Used in the Experiments
表 1 实验中用到的数据集

No.	Names	Instances	Classes	Attributes
1	Mushroom	8124	2	22
2	Annealing	798	5	38
3	B.cancer	699	2	10
4	Credit	690	2	15
5	Cylinder	512	2	39
6	Arrhythmia	452	16	279
7	Vote	435	2	16
8	Horse-colic	368	2	27
9	Audiology	200	2	70
10	Echocardiogram	132	2	12
11	Bridges	108	6	12
12	L.cancer	32	3	56

3.2 实验结果与分析

所有实验是在 weka 系统[17]环境下,在内存为 1GB,主频为 2.93GHz 的 Pentium Ⅳ PC 机上运行的. 在实验过程中,我们令 $\alpha = 1$(α 在 RBC 算法中用来确定先验信息,详情参见文献[12]). 在执行算法 SRBC 时,参数 T 取 weka 系统中的默认值 $T = 5$. 对每个属性子集进行评价时,采用 weka 系统中默认的 5 重交叉验证. 在执行算法 CBSRBC 时,为方便操作,我们令 $T_c = k/4 + 5$,其中 k 为数据集中包含的属性数目. 当然,这样选取的参数 T_c 未必能使算法 CBSRBC 的性能达到最优,即使如此,CBSRBC 也能得到比较理想的分类结果. 对数值型属性,使用“weka.filters.supervised.attribute.Discretize”进行离散化.

表 2 列出了 RBC,SRBC 以及 CBSRBC 在每一个数据集上的 10 次 10 重交叉验证的平均准确率及相应的标准离差,并在表的底部给出了它们在 12 个数据集上的准确率的平均值和标准离差的平均值. 在每一个数据集上的较高的准确率,以粗体表示.

Table 2 Average Accuracy of the Three Classifiers
表 2 三个分类器的平均准确率

Data sets	RBC	SRBC	SRBCUC
Mushroom	95.96±0.02	**99.68±0.04**	**99.68±0.04**
Annealing	95.96±0.31	91.59±0.12	**96.31±0.33**
B.cancer	97.11±0.16	**97.31±0.11**	**97.31±0.11**
Credit	86.18±0.40	86.65±0.30	**87.04±0.27**
Cylinder	71.36±0.48	**76.02±0.55**	76.00±0.54
Arrhythmia	72.77±0.89	**75.01±0.62**	74.63±0.65
Vote	90.25±0.19	**96.31±0.00**	95.85±0.00
Horse-colic	85.20±0.59	88.09±0.39	**88.47±0.13**
Audiology	67.99±0.79	**76.53±0.41**	74.17±0.72
Echocardiogram	**98.36±0.87**	97.26±0.00	98.22±0.92
Bridges	61.62±2.20	**66.10±1.02**	**66.10±1.02**
L.cancer	56.13±1.67	80.32±3.86	**86.45±2.96**
Average	81.57±0.71	85.91±0.62	**86.69±0.64**

从表 2 可以看出,SRBC 在所有实验数据集中的 10 个数据集上,其分类准确率明显高于 RBC 的分类准确率. SRBC 在 12 个数据集上的平均准确率也比 RBC 高出 4.34%. 尤其是在数据集 L.cancer 上,SRBC 的分类准确率比 RBC 的分类准确率高出 24.19%.

之所以在数据集 L.cancer 上分类准确率会提高这么多,除了算法 SRBC 本身的作用外也与数据集 L.cancer 本身的特点有关. L.cancer 总共有 32 个实例,而属性个数却有 56 个之多. 一般情况下,当实例个数与属性个数的比例太小时,对各个属性

变量的类条件概率估计以及对类变量的概率估计都会变得非常不精确,利用这些估计得到的分类结果也会变得很不精确. 当通过 SRBC 使属性个数减少时,相对来讲,就相当于增加了实例的个数,从而使对上述概率的估计变得较精确,也就使得分类准确率可能会有较大提高. 这时,SRBC 的性能会更加显著.

通过考察表 2 还可以发现,CBSRBC 的分类性能比 SRBC 更加显著. 在所有实验数据集中的 11 个数据上,CBSRBC 的分类准确率明显高于 RBC 的分类准确率. 尤其在 L.cancer 数据集上,CBSRBC 的分类准确率比 RBC 的高出 30.32%,也比 SRBC 的分类准确率高出 6.13%. 只是在数据集 Echocardiogram 上 CBSRBC 的分类准确率略低于 RBC 的分类准确率. 与 SRBC 相比,在 5 个数据集上,CBSRBC 的分类准确率明显高于 SRBC 的分类准确率;在 3 个数据集上,它们的分类准确率相同;在其余 4 个数据集上,CBSRBC 的分类准确率略低于 SRBC 的分类准确率. CBSRBC 在 12 个数据集上的平均准确率也比 SRBC 高出 0.78%.

另外,通过比较 3 个分类器在每个数据集上分类准确率的标准离差可以发现,在大多数数据集上 CBSRBC 和 SRBC 的标准离差都比 RBC 的低. 在 12 个数据集上它们的标准离差的平均值也明显比 RBC 的低. 这说明 CBSRBC 和 SRBC 的分类性能比 SRBC 更加稳定.

为进一步对 CBSRBC 和 SRBC 在运行时间以及选择的属性数目等方面进行比较,表 3 给出了它们在每个数据集上选择的属性个数以及运行 10 次的平均时间.

Table 3 Runtime and Selected Attributes of SRBCUC and SRBC

表 3 SRBCUC 与 SRBC 的运行时间和选择的属性数

Data Sets	Total Attr.	Selected Attr.		Runtime(s)	
		SRBC	CBSRBC	SRBC	SRBCUC
Mushroom	22	3	3	110.81	41.11
Annealing	38	8	11	69.91	19.3
B.cancer	10	9	8	8.70	6.35
Credit	15	10	5	15.77	6.70
Cylinder	39	8	7	61.44	15.75
Arrhythmia	279	11	14	676.75	287.34
Vote	16	3	2	3.11	1.59
Horse-colic	27	5	6	13.59	6.45
Audiology	70	12	8	269.66	35.67
Echocardiogram	12	3	3	2.33	1.30
Bridges	12	6	6	3.02	1.56
L.cancer	56	5	10	4.16	2.52
Summation	602	83	83	1239.3	425.64

从表 3 可以看出,在 12 个数据集合中的每个数据集上,CBSRBC 和 SRBC 都能大幅度减少属性的数目,因此,可以在很大程度上对数据集和分类器进行简化. 尤其是在包含 279 个属性的数据集 Arrhythmia 上,CBSRBC 只选择了 14 个属性,而 SRBC 只选择了 11 个属性. 从总体上来看,12 个数据集包含的总的属性数为 602 个,而由 SRBC 和 CBSRBC 选择的总的属性数都是 83 个. 因此,在选择的属性数目上,SRBC 和 CBSRBC 没有明显差别.

但是,从运行时间上来看,CBSRBC 与 SRBC 却有着相当大的差别. 从表 3 可以看出,在所有 12 个数据集上,算法 CBSRBC 的运行时间明显少于 SRBC 的运行时间. 这充分说明 CBSRBC 的运行效率显著高于 SRBC 的运行效率.

对于算法 CBSRBC,需要指出的是其分类准确率和运行时间与参数 T_c 有密切的关系,对此,在 L.cancer 上进行了实验. 表 4 列出了在 L.cancer 上,CBSRBC 在 T_c 取各个不同值时的分类准确率和运行时间(单位:s).

Table 4 Performance of CBSRBC on L.Cancer with T_c Taking Various Values

表 4 在 T_c 的各种取值下 CBSRBC 在 L.cancer 上的运行结果

N	Accuracy(%)	Runtime(s)	N	Accuracy(%)	Runtime(s)
5	81.29	0.53	27	84.84	2.36
10	82.90	1.00	30	84.84	2.41
15	86.45	1.73	35	84.19	4.38
20	86.45	2.42	40	84.19	5.17
21	87.74	2.47	45	84.19	5.84
25	87.74	2.83	50	82.58	4.48
26	87.74	2.86	56	82.58	4.64

从表 4 可以看出,T_c 从 5 逐渐增大时,CBSRBC 的分类准确率也随之增高,当 T_c 增大到一定程度,分类准确率达到峰值(由表 4 知,T_c 在 21~26 之间取值时,分类准确率最高为 87.74%). 之后,随着 T_c 的增大,分类准确率将逐渐下降.

另外,从表 4 可以看出,CBSRBC 的运行时间总体上是随 T_c 的增加而增加. 当然,由于 CBSRBC 的总的运行时间还要包括执行 SRBC 所花费的时间,这样,可能会出现 T_c 增大时,反而 CBSRBC 的运行时间有所降低的现象. 例如,$T_c = 50$ 时,运行时间为 4.48s,比 $T_c = 45$ 时的运行时间(5.84s)还少. 这是由于在 $T_c = 50$ 时(SRBC 只选择了 5 个属性)

SRBC 的运行时间比 $T_c=45$ 时(SRBC 选择了 10 个属性)SRBC 的运行时间少. 因此,选择合适的 T_c 能够在很大程度上提高 CBSRBC 的分类精度和效率. 一般情况下,使算法 CBSRBC 能够达到理想的运行效果的 T_c 值分布在一个不太狭小的范围内. 因此,合适的 T_c 可以凭借经验比较容易地获取.

4 结 论

通过删除数据集中的无关属性和冗余属性,在精选出的属性子集上构建选择性分类器是提高分类精度和效率的一种非常有效的途径. 由于处理不完整数据的复杂性,以往的选择性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并且包含许多冗余属性和无关属性. 因此,对用于不完整数据的选择性分类器的研究是一项重要的研究课题.

本文通过分析已有的处理不完整数据的方法,提出了两种基于 RBC 算法的用于不完整数据的选择性贝叶斯分类模型:SRBC 和 CBSRBC. 在 12 个标准的不完整数据集上的实验结果表明,这两种方法在大幅度减少属性数目的同时,能显著提高分类准确率和稳定性. 从总体上来讲,CBSRBC 在分类精度和运行效率等方面都优于 SRBC 算法. 而 SRBC 算法需要预先指定的阈值要少一些. 因此,SRBC 算法的执行会更简便一些.

参 考 文 献

[1] P Langley, S Sage. Induction of selective Bayesian classifiers [C]. In: Proc of the 10th Conf on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1994. 399-406

[2] M Singh, G M Provan. Efficient learning of selective Bayesian network classifiers [C]. In: Proc of the 13th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufman, 1996

[3] Shang Wenqian, Huang Houkuan, Liu Yuling, *et al*. Research on the algorithm of feature selection based on Gini index for text categorization [J]. Journal of Computer Research and Development, 2006, 43(10): 1668-1694 (in Chinese)
(尚文倩, 黄厚宽, 刘玉玲, 等. 文本分类中基于基尼指数的特征选择算法研究[J]. 计算机研究与发展, 2006, 43(10): 1668-1694)

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

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

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

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

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

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

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

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

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

[13] R Kohavi, G H John. Wrappers for feature subset selection [J]. Artificial Intelligence, 1997, 97(1-2): 273-324

[14] I H Witten, E Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition [M]. San Francisco: Morgan Kaufmann, 2005

[15] P H Winston. Artificial Intelligence [M]. MA: Addison-Wesley, 1992

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

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

Chen Jingnian, born in 1970. Associate professor. He is currently Ph. D. candidate of Beijing Jiaotong University, Beijing, China. He is a student member of CCF. His main research interests include pattern recognition, machine learning and data mining.

陈景年,1970 年生,博士研究生,副教授,中国计算机学会学生会员,主要研究方向为模式识别、机器学习、数据挖掘.

Huang Houkuan, born in 1940. Professor and Ph. D. supervisor. He has published more than 100 journal and conference papers. He is a senior member of CCF. His main research fields include artificial intelligence, pattern recognition, data warehousing, data mining and multi-agent system, *etc*.

黄厚宽,1940 年生,教授,博士生导师,在期刊及会议上共发表了 100 多篇论文,中国计算机学会高级会员,主要研究方向为人工智能、模式识别、数据仓库、数据挖掘以及多智能体系统.

Tian Fengzhan, born in 1972. Ph. D. and associate professor. He is a member of CCF. His main research interests include Bayesian networks, machine learning and data mining.

田凤占,1972 年生,博士,副教授,中国计算机学会会员,主要研究方向为贝叶斯网络、机器学习、数据挖掘.

Fu Shujun, born in 1968. Associate professor. He is currently a Ph. D. candidate of Beijing Jiaotong University, Beijing, China. His main research fields include artificial intelligence, image processing, *etc*.

付树军,1968 年生,博士研究生,副教授,主要研究方向为人工智能、图像处理.

Research Background

Selective classifiers have proved to be a kind of algorithms that can effectively improve the accuracy and efficiency of classification by deleting irrelevant or redundant attributes of a data set. Though some efficient selective classifiers have been proposed, most of them deal with complete data. Yet actual data are often incomplete and have many redundant or irrelevant attributes because of various kinds of reason. So constructing selective classifiers for incomplete data is an important problem. With the analysis of main methods of processing incomplete data for classification, two selective Bayes classifiers for incomplete data are presented, which are denoted as SRBC and CBSRBC respectively. While SRBC is constructed by using the robust Bayes classifiers, CBSRBC is based on SRBC and chi-squared statistics. Experiments on twelve benchmark incomplete data sets show that these two algorithms can not only enormously reduce the number of attributes, but also greatly improve the accuracy and stability of classification as well. On the whole, CBSRBC is more efficient than SRBC and its classification accuracy is higher than that of SRBC. But some thresholds necessary to CBSRBC can be avoided by the SRBC. This work is supported by the National Natural Science Foundation of China under grant No. 60503017 and No. 60673089.

第 3 届中国高性能计算研讨会会议通知

http://www.atip.org/node/94

2007 年 11 月 11 日, 美国 内华达州 里诺市

由美国自然科学基金会(NSF)协办、亚洲科技资讯公司(ATIP)和超级计算大会组委会(SC07)共同主办的第 3 届中国高性能计算研讨会将于 2007 年 11 月 11 日在 2007 超级计算大会(SC07)之前举行. 本次研讨会还得到了许多企业和机构的赞助,如 Sun, Microsoft, SGI, Clearspeed, IDC, 曙光, 上海超算中心等. 中国高性能计算研讨会旨在加强中西方在该研究领域的理解,促进合作,已分别于 2003 年和 2004 年超级计算大会期间成功举办了两届. 欢迎中国从事高性能计算及相关研究和应用的专家、学者、用户、及厂商踊跃参加并投稿. 研讨会主要内容:

① 政府计划及项目;
② 超算相关基础设施(超算中心、网络等);
③ 高性能计算相关研究(算法、体系结构等);
④ 高性能计算应用 (科学计算、工程计算、企业应用等);
⑤ 高性能计算产业(市场情况、中外厂商);
⑥ 小组座谈:中国高性能计算现状;
⑦ 海报.

有意参会者可提出自己愿意在哪一方面做口头报告或海报展示. 所有被本次研讨会采用的口头报告和海报都将被 ACM 数字图书馆收录.

对于使用高性能计算机或对高性能计算感兴趣的人来说,SC07 (http://sc07.supercomp.org) 是全球最大和最重要的活动. 今年估计将有超过 5000 人参会. 参会者不仅有机会与国际高性能计算领域的重要人物进行详细交流,还可以参观来自世界各地的上百个展位(包括厂商、工业用户及研究中心等). SC07 的学术活动包括与高性计算有关的研究论文、小组座谈和讨论会等.

详细情况,请咨询 ATIP 北京代表处 陈道碧 女士

Tel:010-62136752, 13601192280

Email:dchen@atip.org.cn or debbiechen@vip.sina.com