

## 基于显露模式的出生缺陷判别算法

吴保华<sup>1</sup>, 段磊<sup>1</sup>, 于中华<sup>1</sup>, 唐常杰<sup>1</sup>, 朱军<sup>2</sup>

(1. 四川大学 计算机学院, 成都 610065; 2. 中国出生缺陷监测中心, 成都 610064)

(baohuawu@foxmail.com; leidian@scu.edu.cn; yuzhonghua@scu.edu.cn)

**摘要:** 出生缺陷是目前世界各国关注的公共卫生问题, 采用数据挖掘技术提高出生缺陷的诊断水平是当前数字医学的热点研究方向。为此, 提出了适合出生缺陷特征提取的两种显露模式: 有缺陷相比于无缺陷的显露模式和无缺陷相比于有缺陷的显露模式。将新模式与决策树 C4.5 算法结合, 实现了基于显露模式的出生缺陷判别 (BDD-EP) 算法。实验结果表明 BDD-EP 算法判别准确率高达 90.1%, 判别正常类的  $F$  度量值为 93.9%, 判别缺陷类的  $F$  度量值为 74.1%, 均高于其他几种著名的分类算法的判别效果。

**关键词:** 显露模式; 决策树; 特征提取; 出生缺陷

**中图分类号:** TP311.13 **文献标志码:** A

## Birth defects detection algorithm based on emerging patterns

WU Bao-hua<sup>1</sup>, DUAN Lei<sup>1</sup>, YU Zhong-hua<sup>1</sup>, TANG Chang-jie<sup>1</sup>, ZHU Jun<sup>2</sup>

(1. College of Computer Science, Sichuan University, Chengdu Sichuan 610065, China;

2. National Center for Birth Defects Monitoring, Chengdu Sichuan 610064, China)

**Abstract:** The problem of birth defects is one of the most important public health problems in the world, and the application of data mining method to improve the diagnostic accuracy for birth defects is a hot medical research issue. The authors proposed two emerging patterns for birth defects feature extraction: the defection contrast to normal and the normal contrast to defection. The Birth Defects Detection based on Emerging Patterns (BDD-EP) algorithm was implemented through combining the proposed patterns with C4.5 decision tree. The extensive experimental results show that the detection accuracy of BDD-EP is as high as 90.1%, the  $F$ -measure of normal samples is 93.9%, and the  $F$ -measure of defect samples is 74.1%. Compared with other famous classical classification algorithms, BDD-EP algorithm can get better results.

**Key words:** Emerging Pattern (EP); decision tree; feature extraction; birth defect

## 0 引言

出生缺陷包括先天畸形、智力障碍和代谢性疾病等<sup>[1]</sup>, 已成为婴儿死亡和残疾的主要原因之一<sup>[2]</sup>。目前, 新生儿的缺陷病诊断滞后于出生<sup>[3]</sup>, 提高诊断水平, 尽可能早地发现缺陷婴儿及干预婴儿缺陷的发生, 已成为世界各国所面临的共同挑战。

显露模式 (Emerging Pattern, EP)<sup>[4-5]</sup> 旨在描述一个数据集到另一个数据集支持度发生显著变化的项集, 能够捕获目标类和非目标类上多组属性之间的差异。文献[6-9]的实验表明, 基于 EP 的分类算法有较好的区分能力与分类性能。EP 已广泛应用在医学、生物信息学及其他各种分类预测中<sup>[10-12]</sup>。

本文提出并实现了一种基于 EP 和决策树 C4.5 相结合的出生缺陷判别算法 BDD-EP。对从出生监测数据库中随机抽取的含有 8 个属性的 15 096 条样本, 按照 3:2 的比例随机分成训练集和测试集分别进行训练、测试, 实验结果表明准确率达到了 90.1% 分类精度。

本文主要工作包括:

- 1) 介绍了数据挖掘技术在出生缺陷挖掘中的研究进展, 分析了临床中出生缺陷判别的难点;
- 2) 针对分类特征不足而导致的判别误差, 提出了新的特征提取方法, 并对这些特征采用  $strength$  进行度量筛选, 从而减少维数增长引起的计算压力;
- 3) 实现了基于 EP 的出生缺陷判别算法 BDD-EP;
- 4) 在真实出生缺陷数据集上做了详尽的实验, 结果表明: BDD-EP 算法判别准确率高达 90.1%, 判别正常类的  $F$  度量值为 93.9%, 判别缺陷类的  $F$  度量值为 74.1%, 高于其他分类算法的判别效果。

## 1 相关工作

采用数据挖掘技术从已积累的临床资料中提炼出蕴含的信息进行临床诊断及科学研究, 已成为数字医学研究的重要课题。文献[13]采用贝叶斯方法来估计出生缺陷监测数据中缺陷发生率, 并得出了采用贝叶斯方法进行估计的置信度平均要高出采用非贝叶斯方法 29%; 文献[14]采用贝叶斯方

收稿日期: 2010-08-24; 修回日期: 2010-10-11。

**基金项目:** 国家自然科学基金资助项目 (60773169); 国家“十一五”科技支撑计划项目 (2006BAI05A01); 高等学校博士学科点专项科研基金资助项目 (20100181120029); 四川大学青年教师科研启动基金资助项目 (2009SCU11030)。

**作者简介:** 吴保华 (1985-), 男, 河南项城人, 硕士研究生, 主要研究方向: 数据挖掘、自然语言处理; 段磊 (1981-), 男, 四川成都人, 讲师, 博士, 主要研究方向: 数据挖掘; 于中华 (1967-), 男, 黑龙江齐齐哈尔人, 副教授, 博士, 主要研究方向: 数据挖掘、自然语言处理; 唐常杰 (1946-), 男, 重庆人, 教授, 博士生导师, 博士, CCF 高级会员, 主要研究方向: 数据库、知识工程; 朱军 (1964-), 女, 四川成都人, 研究员, 主要研究方向: 出生缺陷干预。

法和空间统计方法对空间范围内的出生缺陷环境致畸因子识别进行了探索与分析;文献[15]采用决策树 C4.5 算法进行出生缺陷判别并实现了一个出生缺陷监测系统;文献[16]以山西省和顺县出生的婴儿为数据集,采用粗糙集理论确定影响出生缺陷变化因素,进而发现了空间因素和婴儿发生神经管缺陷之间的一个关系;文献[17]采用贝叶斯信念网络算法来确定新生儿神经管缺陷发生的概率,并且在数据集上准确率达到了 95% 的置信度;文献[18-19]对出生缺陷数据中的朴素干预规则进行了相关的研究。

文献[4]首次研究了 EP 挖掘算法;文献[20]提出了一种基于零压缩二元决策图 (Zero-suppressed Binary Decision Diagrams, ZBDD) 的 EP 挖掘算法,其实验表明,该算法在高维属性的复杂数据集上挖掘对比模式比较有效;文献[5]采用 Equivalence Class 概念,结合经过修改后的 FP-tree 算法,提出了一种高效的 EP 挖掘算法,实验表明该算法挖掘 EP 比较准确全面,且在时间效率上高于若干经典挖掘算法。EP 具有较强的分类能力,许多基于 EP 的分类算法相继被提出,如: CAEP<sup>[6]</sup>、DeEP<sup>[7]</sup>、JEP<sup>[8]</sup>、CEEP<sup>[9]</sup>等,实验结果表明这些分类器在其实验数据集上具有较强的分类性能;EP 在医学、生物信息学等其他领域也表现出了较强的应用价值。如文献[10]基于 EP 设计了一种针对急性淋巴细胞白血病 (Acute Lymphoblastic Leukemia, ALL) 判别的分类器 PCL,其实验结果表明该分类器的判别错误率比 C4.5 分类器低 71%、比朴素贝叶斯分类器低 50%、比 K 近邻 (K Nearest Neighbor, KNN) 分类器低 43%;文献[11]基于 EP 实现了一种挖掘基因表达谱中具有较强诊断能力的基因组的算法,在 ALL、AML 和 colon tumor 数据集上其准确率达到了 90% 以上。

## 2 BDD-EP 算法的设计

本文的基本出发点是:EP 能够捕获目标类和非目标类上多组属性之间的差异,基于 EP 的分类器改进了传统分类器因没有考虑多组属性之间的差异而引起的分类准确度不高的缺点。在属性个数已确定的出生缺陷监测数据集上,为了能更准确地对出生婴儿进行缺陷诊断,本文首先在训练集上分别提取缺陷类和正常类的对比特征——EP;然后把这些特征应用决策树 C4.5 算法进行分类,以达到缺陷判别的效果。

### 2.1 出生缺陷数据预处理

在出生缺陷监测数据中存在一些数值型属性(如:身高、体重、妊周等),EP 方法要求所有的属性必须是非数值的范畴型。本文沿用基于熵的方法对数值型属性进行了离散化<sup>[21]</sup>,其大致思路如下。

设出生缺陷训练数据集  $D$  中包含  $k$  个类,  $D_i$  为  $D$  中基于某数值属性  $A$  的第  $i$  个划分区间所对应的子集,  $m_i$  为  $D_i$  中属性  $A$  所具有的属性值数量,  $m_{ij}$  为  $D_i$  中类  $j$  的婴儿数据在属性  $A$  上所具有的属性值的数量,  $p_{ij} = m_{ij}/m_i$  为类  $j$  的婴儿数据在  $D_i$  中关于属性  $A$  所占的比例;则  $D_i$  的信息熵可以用式(1)来表示:

$$Entropy(D_i) = - \sum_{j=1}^k p_{ij} \lg(p_{ij}) \quad (1)$$

又设整个出生缺陷训练数据集  $D$  中属性  $A$  的属性值数量为  $m$ , 基于属性  $A$  共产生了  $n$  个划分区间, 则基于属性  $A$  对  $n$  个划分区间的期望信息需求可由式(2)表示:

$$Info_A(D) = \sum_{i=1}^n \frac{m_i}{m} Entropy(D_i) \quad (2)$$

在选择属性  $A$  的分裂点时,选择产生最小期望信息需求的属性值;并且此过程递归地用于所得到的每个划分,直到满足某个终止标准。其中,测试数据集基于训练数据集中的离散化区间进行离散。

### 2.2 显露模式 EP 的挖掘

显露模式在出生缺陷监测数据中其定义描述如下。

定义1 给定正常类(或缺陷类)数据集  $D_1$  和缺陷类(或正常类)数据集  $D_2$ , 项集  $X$  在  $D_1$  和  $D_2$  中的支持度分别为  $sup_1(X)$ 、 $sup_2(X)$ , 则项集  $X$  从  $D_1$  到  $D_2$  的增长率  $GR(X)$  可由式(3)表示:

$$GR(X) = \begin{cases} 0, & sup_1(X) = 0 \text{ 且 } sup_2(X) = 0 \\ \infty, & sup_1(X) = 0 \text{ 且 } sup_2(X) \neq 0 \\ \frac{sup_2(X)}{sup_1(X)}, & \text{其他} \end{cases} \quad (3)$$

给定某一增长率阈值  $\rho > 1$ , 如果项集  $X$  从  $D_1$  到  $D_2$  的增长率  $GR(X) \geq \rho$ , 则称  $X$  是从  $D_1$  到  $D_2$  的  $\rho$ -显露模式, 简称  $X$  是  $D_2$  的 EP; 如果  $GR(X) = \infty$ , 则称  $X$  为  $D_2$  的 JEP (Jumping EP)<sup>[8]</sup>。

例1 给定 100 例出生缺陷监测数据, 其中正常婴儿和缺陷婴儿分别为 76 例和 24 例, 设增长率阈值为 2.5, 下面为两个典型的 EP:

$$X = \{(\text{体重} = 3400 \sim 4000\text{g}), (\text{身长} = 49 \sim 54\text{cm})\}$$

$$Y = \{(\text{体重} = 700 \sim 1900\text{g}), (\text{身长} = 9 \sim 32\text{cm})\}$$

表1 例1中EP的支持度和增长率

EP	正常类中的支持度/%	缺陷类中的支持度/%	增长率
X	73.7	25.8	2.9
Y	1.3	37.5	28.8

则在该数据集中特征  $X$  为正常婴儿的 EP, 特征  $Y$  为缺陷婴儿的 EP。

为建立基于 EP 的分类器, 首先需要挖掘出出生缺陷监测数据中的 EP。文献[5]中给出了被学术界广为应用的 EP 挖掘算法, 其大致思路如下: 对于正常类(或缺陷类)数据集  $D_1$  和缺陷类(或正常类)数据集  $D_2$ , 通过设置:

$$ms = (|D_1| * \alpha\%) / (|D_1| + |D_2|) \quad (4)$$

$$\text{和 } \delta = |D_2| * \beta\% \quad (5)$$

挖掘出的频繁模式即为在  $D_1$  中支持度大于  $\alpha\%$  且在  $D_2$  中的支持度小于  $\beta\%$  的最短的 EP, 这些 EP 相比普通 EP 非冗余, 拥有更强的对比表达功能<sup>[20]</sup>, 且具有较好的相对危险度和让步比<sup>[5]</sup>。

### 2.3 EP 的筛选

从出生缺陷监测训练集中挖掘出这些最短的 EP 后, 假如使用挖掘出来的所有的 EP 进行分类将会带来一系列的问题, 首先维数太大给计算带来非常大的压力, 存储空间大、处理速度慢; 其次一些 EP 的支持度及增长率相对并不高, 甚至会影响分类效果。为此, 本文采用了 *strength* 对挖掘出的 EP 进行了度量: 一个 EP 的 *strength* 越大, 它对分类的贡献就越大, 则它就应优先选择用来分类。strength 的定义如式(6):

$$strength(X) = \frac{GR(X)}{GR(X) + 1} * sup_2(X) \quad (6)$$

有关 EP 筛选的具体描述如算法 1 所示。其中算法 1 的关键描述是 1) ~ 5), 该步骤通过为每一条 EP 遍历训练集中所有婴儿数据, 计算其在正常类和缺陷类子集上的支持度, 以计算该条 EP 的 *strength* 值。该算法的时间复杂度为  $O(n^2)$ 。

算法 1 基于 *strength* 的 EP 筛选算法。

输入: 未筛选的正常类(或缺陷类)的 EP 集合 *OEP*; 缺陷婴儿训练子集 *DTR*; 正常婴儿训练子集 *NTR*; 经过筛选后输出的 EP 数量 *t*。

输出: 筛选后的正常类(或缺陷类)的 EP 集合 *SEP*。

```

1) for each EP  $e \in OEP$ 
2)  $sup_1 = support(e, DTR)$ ;
   // 该 EP 在缺陷婴儿(或正常婴儿)子集中的支持度
3)  $sup_2 = support(e, NTR)$ ;
   // 该 EP 在正常婴儿(或缺陷婴儿)子集中的支持度
4)  $calculate\ gr = sup_2 / sup_1$ ;
   // 计算正常类(或缺陷类)EP 的增长率
5)  $calculate\ str = gr / (gr + 1) \times sup_2$ ;
   // 计算正常类(或缺陷类)EP 的 strength
6) end
7) sort(OEP); // 所有 EP 按 strength 从大到小的顺序排序
8)  $SEP = select(t, OEP)$ ; // 选择前 t 个 EP 输出
9) return SEP;
```

## 2.4 BDD-EP 挖掘算法

出生缺陷的判别可看做采用分类器对监测数据集中的婴儿数据进行分类的过程。算法 BDD-EP 的核心思想是: 把从数据集中抽取的特征——缺陷类和正常类的 EP, 应用于 C4.5 算法中。即 C4.5 算法中的节点由已选择好的具有较强辨别力的 EP 组成, 分裂规则采用基于 EP 增益率的选择度量。具体地, 该分裂规则是基于信息熵(*Info*)、信息增益(*Gain*)、分裂信息(*SplitInfo*)对 C4.5 算法中的增益率(*GainRatio*)进行了扩展; 以某一  $EP_j$  为例, 定义如下:

$$Info(D) = - \sum_{i=1}^k p_i \lg(p_i) \quad (7)$$

$$Gain(EP_j) = Info(D) - \left[ \frac{|D_j|}{|D|} * Info(D_j) + \left( 1 - \frac{|D_j|}{|D|} \right) * Info(D_{-j}) \right] \quad (8)$$

$$SplitInfo_{EP_j}(D) = - \frac{|D_j|}{|D|} * \lg\left(\frac{|D_j|}{|D|}\right) - \left( 1 - \frac{|D_j|}{|D|} \right) * \lg\left( 1 - \frac{|D_j|}{|D|} \right) \quad (9)$$

$$GainRatio(EP_j) = \frac{Gain(EP_j)}{SplitInfo(EP_j)} \quad (10)$$

其中:  $k$  表示出生缺陷数据集中不同的类别数,  $p_i$  表示任意元组属于类  $C_i$  的概率,  $D_j$  是  $D$  中能覆盖  $EP_j$  的所有元组组成的数据集,  $D_{-j}$  是  $D$  中不能覆盖  $EP_j$  的所有元组组成的数据集, 即  $D = D_j \cup D_{-j}$ 。最终, 选择增益率最大的 EP 作为分裂 EP。

BDD-EP 算法由五个部分组成:

- 1) 把出生缺陷数据库中连续型的数值属性转化为非数值的范畴型;
- 2) 挖掘出生缺陷数据库中训练集中的 EP;
- 3) 对挖掘出的 EP 进行筛选;
- 4) 对训练集和测试集中每一条婴儿数据以筛选后的 EP 为属性进行转化;

5) 采用 C4.5 分类器分别对转化后的训练集和测试集进行训练与测试。

第 1) ~ 3) 部分的实现方法, 已经在本文第 2.1 ~ 2.3 节说明; 对于第 5) 部分, 可以直接采用 C4.5 分类器实现, 其原理已经在第 2.4 节说明; 第 4) 部分即采用 EP 对数据集的转化是 BDD-EP 算法的核心问题。通过 2.3 节方法, 从挖掘出的缺陷类的 EP 集和正常类的 EP 集中分别选择 *strength* 最大的前  $k_1$  和  $k_2$  个 EP 作为分类器采用的特征, 然后使每一出生婴儿数据采用由这些 EP 组成的向量  $\langle V_1, V_2, \dots, V_{k_1}, V_1', V_2', \dots, V_{k_2}' \rangle$  表示。设  $m \in [1, k_1], n \in [1, k_2]$ , 则  $V_m$  为出生婴儿监测数据集中 *strength* 较大的缺陷类的  $EP_m$ ,  $V_n'$  为出生婴儿监测数据集中 *strength* 较大的正常类的  $EP_n$ 。假如某一婴儿数据包含  $EP_m$ , 则  $V_m$  的值为 1, 否则为 0 ( $V_n'$  的取值与之类似)。具体算法描述如下所示。算法 2 的关键步骤是 4) ~ 8), 该步骤以 EP 数据集中的每一条 EP 为属性建立关系数据表, 如果该婴儿数据包含该 EP 则该婴儿数据在该 EP 下的属性值为 1, 否则为 0。该算法的时间复杂度为  $O(n^2)$ 。

算法 2 基于 EP 的数据集转化算法。

输入: 已经离散化而待转化的数据集 *DS*; 经过 *strength* 筛选后的 EP 数据集 *SEP*。

输出: 转化后的数据集 *EDS*。

```

1)  $EDS = \emptyset$ ;
2) for each infant  $I$  in DS
3)  $I' = \emptyset$  //  $I'$  为新的婴儿数据
4) for each EP  $e$  in SEP
5) if  $e \subset I$  then // 若该婴儿数据能覆盖该 EP
6)  $I'$  add 1; // 该 EP 在婴儿数据中的属性值为 1
7) else  $I'$  add 0; // 否则属性值为 0
8) end;
9)  $EDS$  add  $I'$ ; // 此条婴儿数据转化结束, 加入 EDS 中
10) end;
11) return EDS;
```

例 2 对于例 1 数据集中婴儿采用 BDD-EP 算法进行缺陷判别, 首先在该数据集中挖掘并筛选出正常婴儿 EP 的数量为 40 个 ( $N_1, N_2, \dots, N_{40}$ ), 缺陷婴儿 EP 的数量为 22 个 ( $B_1, B_2, \dots, B_{22}$ ), 把此 62 个 EP 看做出生缺陷监测数据集中每一婴儿数据的属性来转化原数据集; 如果某一婴儿数据包含某一 EP, 则该婴儿数据在该 EP 属性上的属性值为 1, 否则为 0。然后再把转化好的数据集采用决策树 C4.5 算法进行分类训练与测试。

## 3 实验及性能分析

### 3.1 实验平台与数据预处理

实验平台为 Athlon 3.2 GHz 和 2 GB 内存的 Windows XP 平台, JDK 6.0 编译环境。

数据来自于中国出生缺陷监测中心监测数据库中 1986—1988—1988—12 出生监测的数据, 从中随机抽取了 15 096 条病历, 按照 3:2 的比例随机分成训练集和测试集。为了减少 EP 挖掘过程中的计算复杂度以提高 EP 挖掘的效率, 并尽量减少因信息不足引起的诊断错误, 本实验仅采用了已由医学专家选好的 8 个属性作为数据集的属性, 如表 2 所示。其中, 体重、身长、母亲年龄、父亲年龄等数值型属性, 实验中采用 2.1 节介绍的方法对其进行了离散化, 母亲职业属性直接采用了原数据库中已经定义好的名词属性值。

### 3.2 实验结果及分析

出生缺陷判别的性能评价通常借用分类的相关指标。具



体地,设  $N_{\text{defect}}$  为测试集中缺陷样本的总数,  $N_{\text{normal}}$  为测试集中正常样本的总数,  $t_{\text{defect}}$  是被正确分类的缺陷样本数,  $f_{\text{defect}}$  是被错误分类的缺陷样本数,  $t_{\text{normal}}$  是正确分类的正常样本数,  $f_{\text{normal}}$  是被错误分类的正常样本数, 则有以下几个评价指标用来衡量出生监测数据中不同的判别器的性能 (只列举缺陷样本的评价指标, 正常样本的评价指标与之类似)。

1) 召回率 (recall): 即缺陷 (或正常) 样本的查全率。

$$\text{recall}_{\text{defect}} = t_{\text{defect}} / N_{\text{defect}} \quad (11)$$

召回率反映了判别器发现缺陷 (或正常) 样本的能力, 召回率越高其“漏网”的缺陷 (或正常) 样本就越少。

2) 精确度 (precision): 即缺陷 (或正常) 样本的查准率。

$$\text{precision}_{\text{defect}} = \frac{t_{\text{defect}}}{t_{\text{defect}} + f_{\text{defect}}} \quad (12)$$

精确度反映了判别器“找对”缺陷 (或正常) 样本的能力, 精确度越高将缺陷 (或正常) 样本判断为正常 (或缺陷) 的可能性就越小。

3)  $F$  度量值 ( $F$ -measure)。

$$F_{\text{defect}} = 2 \times \frac{\text{recall}_{\text{defect}} * \text{precision}_{\text{defect}}}{\text{recall}_{\text{defect}} + \text{precision}_{\text{defect}}} \quad (13)$$

$F$  度量值实际上反映的是召回率和精确度的调和平均值。

4) 准确度 (accuracy)。

$$\text{accuracy} = \frac{t_{\text{defect}} + t_{\text{normal}}}{N_{\text{defect}} + N_{\text{normal}}} \quad (14)$$

准确度反映的是对所有样本 (缺陷和正常) 的判对率。

表 2 出生缺陷监测数据属性

属性	描述
TZ	婴儿出生时的身体重量 (g)
SC	婴儿出生时身体长度 (cm)
CC	母亲的生产次数 (times)
YC	母亲的怀孕次数 (times)
RZ	母亲的妊娠周数 (weeks)
MNL	母亲的实足年龄 (years)
FNL	父亲的实足年龄 (years)
MZY	母亲的职业: 0, 未填; 1, 工人; 2, 农民; 3, 商人; 4, 军人; 5, 干部; 6, 科技人员; 7, 医药工作者; 8, 其他

在挖掘正常类 (或缺陷类) EP 的过程中, 为了能够挖掘出较多的 EP, 同时又使挖掘出的 EP 能够被较多的目标类的婴儿数据包含, 且被较多的非目标类婴儿数据不能包含, 本文把式 (4) 中的  $\alpha$  设置为 0.8 以上, 通过改变不同的  $\alpha, \beta$  值并分别把挖掘出的正常类和缺陷类的全部 EP 作为特征用于 C4.5 分类器, 用判别的准确度作为度量, 度量结果如图 1、2 所示。

其中, 挖掘正常类 EP 时, 当  $\beta < 0.27$  时 EP 的数量为 0。挖掘缺陷类 EP 时, 当  $\alpha$  为 0.8 和 0.83,  $\beta < 0.29$  时 EP 的数量为 0 ( $\beta = 0.29$  时判别准确度分别为 86.2% 和 86.8%, 为便于作图没有显示在图 2 中); 当  $\alpha$  为 0.86 和 0.9,  $\beta < 0.31$  时 EP 的数量为 0。为了更能准确地判别出婴儿是否患有缺陷, 本文选择了正常类和缺陷类各自判别准确度最高时的那些 EP 做下一步的实验, 此时的  $\alpha, \beta$  值如表 3 所示。

这样, 在训练集中共挖掘出 961 条缺陷类的 EP 和 2 600 条正常类的 EP。通过采用 2.3 节所论述的筛选方法, 分别在缺陷类与正常类 EP 集中筛选不同的 EP 作为分类器所采用的特征 EP, 并采用判别的准确度进行度量, 实验结果如图 3

所示。其中, 在使用挖掘出的全部缺陷类 EP, 正常类 EP 的数量为 100 时分类的准确度达到最高 90.1%。

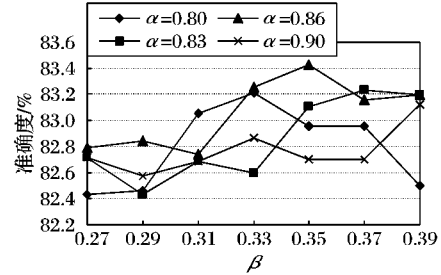


图 1  $\alpha, \beta$  不同值时正常类 EP 判别准确度

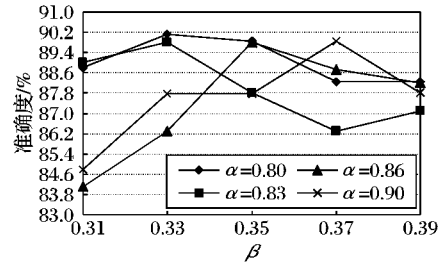


图 2  $\alpha, \beta$  不同值时缺陷类 EP 判别准确度

表 3 两类 EP 各自判别准确度最高时的  $\alpha, \beta$  值

类别	$\alpha$	$\beta$
缺陷类	0.80	0.33
正常类	0.86	0.35

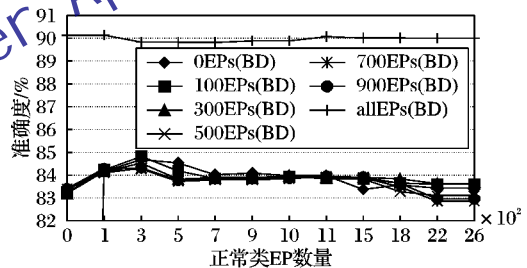


图 3 不同 EP 数量时的 BDD-EP 的判别准确度

为了验证 BDD-EP 算法对出生缺陷判别的有效性, 在本实验中分别采用了在出生缺陷研究中广泛采用的朴素贝叶斯 (NB)、贝叶斯网络 (BN)、C4.5 及 SMO、KNN 等其他几种知名的分类器与之对比, 判别的实验结果如表 4 所示。

通过对比分析可知, BDD-EP 无论在准确度、召回率、精确度、 $F$  度量值方面都要高于其他几种常见算法的判别效果, 可见本文提出的算法对出生缺陷的判别是可行与高效的, 对出生缺陷的诊断将具有一定的指导意义。

## 4 结语

出生缺陷是严重的社会公共卫生问题, 提高出生缺陷诊断力是当前数字医学的重要方向。本文研究了出生缺陷诊断方法, 提出了一种新的特征提取方法——挖掘数据集中有缺陷相比于无缺陷的 EP 和无缺陷相比于有缺陷的 EP, 并将这些特征应用于决策树 C4.5 算法中, 实现了一种基于 EP 的出生缺陷判别算法 BDD-EP。实验表明此算法的判别准确率达到了 90.1% 的较好效果, 对出生缺陷的诊断具有一定的指导意义。

由于训练数据集的限制, BDD-EP 只能处理两类数据的问题: 缺陷类和正常类。在下一步工作中, 将扩大缺陷类别的数据集以拓展 BDD-EP 用于其他特定缺陷疾病的诊断能力。

表4 不同的分类器的判别结果对比

分类器	建立分类器 模型时间/s	准确度/%	缺陷类/%			正常类/%		
			召回率	精确度	F 度量值	召回率	精确度	F 度量值
NB	0.11	82.7	55.4	66.1	60.3	91.2	86.8	89.0
BN	0.24	84.5	58.8	70.8	64.3	92.5	87.9	90.1
SMO	22.81	84.0	79.1	44.1	55.6	86.4	84.8	90.2
KNN	0.00	84.6	44.6	82.4	57.8	97.0	84.9	90.6
C4.5	1.05	84.3	52.9	73.3	61.5	94.0	86.6	90.1
BDD-EP	6.38	90.1	59.7	97.8	74.1	99.6	88.9	93.9

## 参考文献:

- [1] 钟南山. 浅谈我国出生缺陷研究的现况与展望[J]. 中国优生优育, 2007, 13(1): 10-11.
- [2] CARMONA R H. The global challenges of birth defects and disabilities [J]. Lancet, 2005, 366(9492): 1142-1144.
- [3] EPTEIN C J. Genetic disorders and birth defects [M]// RUDOLPH A M, HOFFMAN J, RUDOLPH C D. Rudolph's Pediatrics, 19th ed. Norwalk, CT: Appleton & Lange, 1991: 265-269.
- [4] DONG G Z, LI J Y. Efficient mining of emerging patterns: Discovering trends and differences [C]// Proceedings of the 5th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Diego: ACM, 1999: 43-52.
- [5] LI J, LIU G, WONG L. Mining statistically important equivalence classes and delta-discriminative emerging patterns [C]// Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Jose: ACM, 2007: 430-439.
- [6] DONG G Z, ZHANG X Z, WONG L, et al. CAEP: Classification by aggregating emerging patterns [C]// Proceedings of the 19th International Conference on Discovery Science, LNCS 1721. Berlin: Springer, 1999: 30-42.
- [7] LI J, DONG G, RAMAMOCHANARAO K, et al. DoEP: A new instance-based lazy discovery and classification system [J]. Machine Learning, 2004, 54(2): 99-124.
- [8] LI J, DONG G, RAMAMOCHANARAO K. Making use of the most expressive jumping emerging patterns for classification [C]// Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNCS 1805. Berlin: Springer, 2000: 220-232.
- [9] 李曼, 范明. 一种新颖的基于最有效的跳跃显露模式的分类法[J]. 计算机科学, 2002, 29(8): 73-76.
- [10] LI J, LIU H, DOWNING J R, et al. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients [J]. Bioinformatics, 2003, 19(1): 71-78.
- [11] LI J, WONG L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns [J]. Bioinformatics, 2002, 18(5): 725-734.
- [12] MAO S, DONG G. Discovery of highly differentiative gene groups from microarray gene expression data using the gene club approach [J]. Bioinformatics and Computational Biology, 2005, 3(6): 1263-1280.
- [13] HEMMI I. Bayesian estimation of the incidence rate in birth defects monitoring [J]. Congenit Anom, 1988, 28(2): 103-109.
- [14] WU J, WANG J, MENG B, et al. Exploratory spatial data analysis for the identification of risk factors to birth defects [J]. BMC Public Health, 2004, 23(4): 23-33.
- [15] 杨峰. 基于决策树的出生缺陷预警系统研究与实现[D]. 长春: 东北师范大学, 2006.
- [16] BAI H, GE Y, WANG J, et al. Using rough set theory to identify changes affected by birth defects: the example of Heshun, Shanxi, China [J]. International Journal of Geographical Information Science, 2010, 24(4): 559-576.
- [17] LIAO Y, WANG J, GUO Y, et al. Risk assessment of human neural tube defects using a Bayesian belief network [J]. Stochastic Environmental Research and Risk Assessment, 2009, 24(1): 93-100.
- [18] 张悦, 唐常杰, 李川, 等. 出生缺陷监测数据中的朴素干预规则挖掘[J]. 计算机科学与探索, 2009, 3(2): 188-197.
- [19] 唐常杰, 段磊, 王悦, 等. 干预规则挖掘的任务分类和三项技术进展[J]. 计算机应用, 2010, 30(1): 10-14.
- [20] LOEKITO E, BAILEY J. Fast mining of high dimensional expressive contrast patterns using zero-suppressed binary decision diagrams [C]// Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Philadelphia: ACM, 2006: 307-316.
- [21] FAYYAD U, IRANI K. Multi-interval discretization of continuous-valued attributes for classification learning [C]// Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chamberg: Morgan Kaufmann, 1993: 1022-1027.

## 全国抗恶劣环境计算机第二十一届学术年会征文通知

全国抗恶劣环境计算机第二十一届学术年会将于2011年8月23—26日在山东青岛召开。本次会议由中国计算机学会主办, 中国计算机学会抗恶劣环境计算机专委会、中国船舶重工集团公司第716研究所承办。会议将通过学术报告、专题讨论等多种方式, 充分交流我国抗恶劣环境计算机科研成果, 介绍抗恶劣环境计算机技术与产品, 展望国内外抗恶劣环境计算机发展趋势和前景。会议将邀请著名专家学者到会做专题报告。

## 一、征文范围(包括但不限于)

- 抗恶劣环境计算机发展现状与趋势
- 抗恶劣环境计算机需求分析
- 用于自主可控信息系统的计算机技术
- 复杂电磁环境下的计算机技术
- 安全防护与可信计算技术
- 物联网的计算机技术

## 二、联系信息

投稿邮箱: songly706@sina.com(请注明“2011 抗恶劣征文”)

联系电话: 宋凌云(010-68387785); 张淑萍(010-68388715)

征文投稿截止日期: 2011年5月30日 录用通知发出日期: 2011年7月10日 学术年会召开日期: 2011年8月23—26日  
详情请见: <http://www.ccf.org.cn/sites/ccf/nry.jsp?contentId=2600536795851>