

基于层级规则树的跨平台基因表达数据分类

蔡瑞初¹, 侯永杰¹, 郝志峰^{1,2}

(1. 广东工业大学 计算机学院, 广州 510006; 2. 佛山科学技术学院 数学与大数据学院, 广东 佛山 528000)

摘 要: 基因检测技术运用至今已积累大量来自不同平台的数据, 针对传统数据分类模式难以在不同平台间进行有效迁移的问题, 提出一种基于层级规则树的基因表达数据分类算法 k-HRT。设计数据转换与规则预筛选策略, 实现算法的快速挖掘, 以解决由跨平台特性所带来的大规模数据问题。在真实基因表达数据集上的实验结果表明, 相对 k-TSP 算法、SVM-RFE 算法, k-HRT 算法能够有效提高分类精度。

关键词: 数据分类; 跨平台; 规则学习; 特征选择; 基因表达数据

中文引用格式: 蔡瑞初, 侯永杰, 郝志峰. 基于层级规则树的跨平台基因表达数据分类[J]. 计算机工程, 2019, 45(7): 26-31.

英文引用格式: CAI Ruichu, HOU Yongjie, HAO Zhifeng. Cross-platform gene expression data classification based on hierarchical rule tree[J]. Computer Engineering, 2019, 45(7): 26-31.

Cross-platform Gene Expression Data Classification Based on Hierarchical Rule Tree

CAI Ruichu¹, HOU Yongjie¹, HAO Zhifeng^{1,2}

(1. School of Computers, Guangdong University of Technology, Guangzhou 510006, China;

2. School of Mathematics and Big Data, Foshan University, Foshan, Guangdong 528000, China)

[Abstract] The application of genetic testing technology has accumulated a large amount of data from different platforms. To address the problem that it is difficult to migrate traditional data classification modes across different platforms, this paper proposes a gene expression data classification algorithm k-HRT based on Hierarchy Rule Tree (HRT). The strategy of data conversion and rule pre-screening is designed to realize the fast mining of the algorithm to solve the large-scale data problems caused by cross-platform characteristics. Experimental results on real gene expression datasets show that, compared with k-TSP algorithm and SVM-RFE algorithm, k-HRT algorithm can effectively improve classification accuracy.

[Key words] data classification; cross-platform; rule learning; feature selection; gene expression data

DOI: 10.19678/j.issn.1000-3428.0051856

0 概述

人的精神状况、行为偏好由先天基因组与外界刺激共同决定^[1-2]。基因检测技术的发展使得从全基因组表达水平定量检测基因转录产物 mRNA 得以实现。“基因→mRNA→蛋白质”为基因表达过程, 可以通过分析 mRNA 数据来研究发生变化的基因表达。运用机器学习方法对基因表达数据进行研究, 对医学临床诊断具有重要的参考意义。但不同于普通的分类问题, 由于伦理和法律方面的约束与潜在风险, 在医学诊断中对基因表达数据进行研究

的目标是得到精确且可解释的分类机制。机器学习领域主流的 SVM^[3]、神经网络^[4]等方法对数据具有很强的拟合能力, 但其结果的可解释性较差, 应用于医学诊断领域存在一定局限。

针对上述问题, 有学者将关联规则引入到分类问题中。文献[5]提出 CBA 算法, 其挖掘基因数据间的关联规则并用于分类。基于关联的分类规则具有直观的可解释性, 但其在高维基因数据上会产生大量的冗余规则, 且其规则挖掘与分类器构建为两阶段算法, 导致即使有巨量规则, 仍会有大量样本未

基金项目: 国家自然科学基金(61472089); NSFC-广东联合基金(U1501254); 广东省自然科学基金(2014A030306004, 2014A030308008); 广东省科技计划项目(2015B010108006, 2015B010131015); 广东特支计划(2015TQ01X140); 广州市科技计划项目(201604016075); 广州市珠江科技新星专项(201610010101)。

作者简介: 蔡瑞初(1983—), 男, 教授、博士, 主研方向为大数据分析、因果关系发现、机器学习; 侯永杰, 硕士研究生; 郝志峰, 教授。

收稿日期: 2018-06-19 **修回日期:** 2018-08-09 **E-mail:** cairuichu@gmail.com

被分类器所覆盖。文献[6]从样本覆盖的角度出发, 提出一种 Top-k 算法, 其以样本枚举的方式挖掘可以交叉覆盖到全部样本的多个规则组, 以进行分类器构建。以上算法在数据预处理阶段需将连续的基因表达水平值离散化, 即根据阈值的大小将基因表达水平值划分为激活态和抑制态。跨平台基因表达数据的特点是各平台的样本具有相同的特征和标签, 但由于平台的特性, 数据值之间存在尺度差异, 因此不同平台间的阈值也不同, 导致已有分类模式只适用于单一平台。

受 k-TSP 分类模式中“相对表达反转”概念^[7]的启发, 本文设计一种层级规则树(Hierarchy Rule Tree, HRT)分类模式。传统有监督学习算法推广到基因表达数据领域时, 一般通过基因选择和关联规则发现 2 种方式^[8-10]。HRT 分类模式将上述 2 种方式进行结合, 基于基因间相对表达反转的特性, 设计相应的数据转换和规则预筛选策略, 进行迭代式分类规则挖掘, 以提高跨平台基因表达数据挖掘的效率。

1 跨平台分类模式

1.1 最高得分对

文献[11]提出一种最高得分对(Top Scoring Pair, TSP)算法, 该算法是纯数据驱动的机器学习方法, 无需使用者设定参数, 避免了精确调整参数所带来的过拟合问题。TSP 分类器的出发点是不同基因表达水平之间的相对差异性, 即寻找事件 $\{V_a < V_b\}$ (V_a (V_b) 代表基因 a (b) 的表达值) 在不同类样本中发生概率显著不同的那些标志性基因对 (a, b) 。TSP 算法关注事件 $\{V_a < V_b\}$ 在每一类样本中发生的概率 $p_{ab}(C_m)$, 其公式定义如下:

$$p_{ab}(C_m) = \text{Prob}\{V_a < V_b | Y = C_m\}$$

其中, C_m 为样本类别, 此处仅考虑二分类问题, $m = \{0, 1\}$ 。用 Δ_{ab} 代表每一个基因对 (a, b) 的得分, $\Delta_{ab} = |p_{ab}(C_0) - p_{ab}(C_1)|$, 以得分最高的基因对作为分类器的判定依据。基因数据具有高维度特性, 上万个基因组合出的基因对的数量是亿级, 因此, 有可能出现得分一致的基因对。如果出现这种情况, TSP 算法会计算第 2 个评分指标, 即平均排序差, 以选取唯一的最高得分对。

TSP 算法理论上可以作为跨平台分类模式的基础, 但以单个基因对作为全部样本的分类模式过于简单。有研究者基于机器学习集成思想, 以 TSP 作为基分类器, 提出改进的 k-TSP 算法, 但仍难以拟合复杂的数据分布。

1.2 层级规则树

从规则构成的角度出发, 以基因对 (a, b) 之间表达值反转为基础的分类模式, 是由 2 个独立的分类规则 $\{< a, b > \rightarrow C_m\}$ 、 $\{\neg < a, b > \rightarrow C_m\}$ ($m = \{0, 1\}$) 所构成, 其中, $< a, b >$ 代表 $\{V_a < V_b\}$, $\neg < a, b >$ 代表 $\{V_a \geq V_b\}$ 。

本文提出的层级规则树以支持度(supp)和置信度(conf)作为规则兴趣度的度量。支持度和置信度是传统关联规则挖掘中的度量标准, 现已扩展到分类规则挖掘中。给定一个数据集 R , d 为数据集样本, X 和 Y 为样本属性, 对于分类规则 $X \rightarrow Y$, supp 和 conf 公式定义分别如下:

$$\text{supp}(X \rightarrow Y) = \frac{|\{d \in R; X \subseteq d, Y \subseteq d\}|}{|R|} \quad (1)$$

$$\text{conf}(X \rightarrow Y) = \frac{|\{d \in R; X \subseteq d, Y \subseteq d\}|}{|\{d \in R; X \subseteq d\}|} \quad (2)$$

以表 1 中的数据分布为例, 有:

$$\text{supp}(< a, b > \rightarrow C_0) = \frac{n_1}{n_1 + n_2 + n_3 + n_4}$$

$$\text{supp}(< a, b > \rightarrow C_1) = \frac{n_3}{n_1 + n_2 + n_3 + n_4}$$

$$\text{conf}(< a, b > \rightarrow C_0) = \frac{n_1}{n_1 + n_2}$$

$$\text{conf}(< a, b > \rightarrow C_1) = \frac{n_3}{n_3 + n_4}$$

表 1 数据分布示例

特征	标签	样本数量
< a, b >	C_0	n_1
	C_1	n_2
$\neg < a, b >$	C_0	n_3
	C_1	n_4

可以看出, $< a, b >$ 、 $\neg < a, b >$ 这 2 个分类规则基于完全不同的样本子集而得出, 两者之间并不存在关联关系。有可能其中一个规则具有很强的分类能力, 而另外一个规则没有参考价值。将两者相结合进行衡量, 会限制分类器对数据的拟合能力。基于此, 本文以其中的单个规则为基础, 建立一种层级规则树分类模式。

层级规则树首先对置信度最高的规则特征 $< a, b >$ 所覆盖的样本进行有效分类, 对于剩余样本(特征 $\neg < a, b >$ 所覆盖的样本), 寻找更精确的新规则(如 $< c, d >$)作为分类依据, 迭代挖掘直至样本集中的样本为空或小于设定阈值。后续规则形式可解释为 $\{\neg < a, b > \wedge < c, d > \rightarrow C_m\}$, $\{\neg < a, b > \wedge \neg < c, d > \wedge < e, f > \rightarrow C_m\}$, ...。如图 1 所示, 在预测样本类别时, 以挖掘出的顺序作为优先级依次使用多个分类规则, 构成层级规则树分类器 CL_{HRT} 。

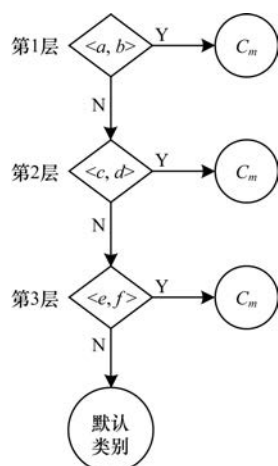


图1 层级规则树示例

上述过程是 HRT 分类模式在二分类问题中的应用,在解决多分类问题时,采用如下方式:

给定一个多分类标签 $C = \{C_0, C_1, \dots, C_{m-1}\}$, 将多分类问题分解为 m 个二分类问题, 即 $\{C_0, \sim C_0\}, \{C_1, \sim C_1\}, \dots, \{C_{m-1}, \sim C_{m-1}\}$, 由此得到 m 个 HRT 分类器。若 m 个分类器预测结果无冲突, 则输出此唯一类别; 若存在分类器冲突, 则选择缺省类(样本中的多数类)作为最终结果。

1.3 k-HRT 算法

集成学习是一种通用的机器学习策略, 与具体的算法模型无关, 其利用多个弱分类器进行集成以提升分类器的最终效果^[12]。常用的集成学习方法有 bagging、boosting, 以及由两者衍生出的随机森林、Adaboost、GBDT 等。为提升分类器在跨平台基因表达数据集上的泛化能力, 本文借鉴 bagging 的思想, 以 HRT 作为基分类器进行集成学习, 得到 k-HRT 算法。从训练集中随机去除小部分样本, 对多个基分类器 CL_i 进行 k 次重复训练, 构成分类器 h_{k-HRT} , 对 k 个基分类器的预测结果进行无权投票, 将得分最高的类别作为样本的最终预测类别并输出。设 x_{new} 是等待预测的样本, 待预测样本类别为 $C = \{C_0, C_1, \dots, C_{m-1}\}$, 则分类器 h_{k-HRT} 定义如下:

$$h_{k-HRT}(x_{new}) = \arg\max_C \sum_{i=1}^k I(CL_i(x_{new}), C) \quad (3)$$

$I(CL_i(x_{new}), C)$ 函数定义如下:

$$I(CL_i(x_{new}), C) = \begin{cases} 1, & CL_i(x_{new}) = C \\ 0, & \text{其他} \end{cases} \quad (4)$$

2 快速挖掘算法

在大规模基因表达数据集中, 快速挖掘算法基于相对偏移表(Relative Shifting Table, RST)实现规则预筛选, 以避免对规则空间进行暴力搜索。用 $N \times P$ 的矩阵 R 表示数据集, 基因表达数据样本数量为 N , 每个样本记录的基因(即特征)值的数量为 P 。算法首先对矩阵 R 进行数据转换, 然后构建 RST 进行规则预筛选, 从候选集中选择得分最高的分类规则, 删去该规则所命中的样本, 最后迭代构建层次规则树。完整

的层次规则树挖掘算法 BuildHRT 描述如下。

算法1 BuildHRT 算法

输入 数据集 R , 最小支持度 θ , 候选组数量 m

输出 分类器 CL

```

1. DataConvert( $R$ );
2. while  $|R| \geq N \times \theta$  do
3.  $RST = \text{RelativeShiftTable}(R)$ ;
4.  $Can = \text{GetCandidates}(T, m)$ ;
5.  $\gamma = \text{GetTop}(RST, Can, \theta)$ ;
6. if  $\gamma \neq \text{NULL}$  then
7. insert  $\gamma$  into  $CL$ ;
8. for each  $d \in R$  do
9. if  $d$  satisfies  $\gamma$  then
10. delete  $d$  from  $R$ ;
11. else
12. break
13. Set majority class of remain  $R$  as default class of  $CL$ .
```

BuildHRT 算法的时间复杂度和其在数据集 R 上的迭代效率有关。在最佳情况下, 迭代一次算法终止, 即 HRT 算法为一层, 算法时间复杂度为 $O(mN)$ 。在最差情况下, 迭代 $\frac{1}{\theta}$ 次算法终止, 算法时间复杂度

为 $O\left(\frac{1+\theta}{2\theta}mN\right)$, 此时最小支持度 θ 决定了时间复杂度, θ 的理论取值范围为 $\left[\frac{1}{N}, 1\right]$, θ 值越小, 算法复杂度越高, 取最小值 $\frac{1}{N}$ 时算法复杂度为 $O(mN^2)$ 。

2.1 数据转换

对于以基因对表达反转为基础的分类模式, 基因表达数据上有意义的是各基因之间表达值的大小关系, 而非单个基因的表达值。因此, 在 BuildHRT 算法的第 1 行, $DataConvert(R)$ 把记录基因表达值的矩阵转换为基因序列矩阵, 以此作为 HRT 分类模式挖掘算法的基础。具体地, 将矩阵 R 中每一个样本的基因表达值和基因名组合成一个二元组, 对每个样本内的 P 个二元组, 以基因表达值为键进行排序, 删去二元组中的基因表达值, 从而将样本转换为由基因名构成的有序序列, 该过程如图 2 所示。

基因	a	b	c	d
样本				
GPLx_sample1	957	642	821	203
GPLy_sample2	23	12	14	17
GPLz_sample3	6.6	3.7	2.3	4.2

(a) 转换前数据矩阵

下标	0	1	2	3
样本				
GPLx_sample1	a	c	b	d
GPLy_sample2	a	d	c	b
GPLz_sample3	a	d	b	c

(b) 转换后数据矩阵

图2 数据转换示例

2.2 相对偏移表

解决分类问题的关键是发现在不同类样本中表现差异最大的基因表达模式,并对其进行高效率的挖掘。为避免在高维度、大样本的跨平台基因表达数据集中进行暴力搜索,减小规则搜索空间,本文在数据转换的基础上设计一种相对偏移表 RST,以进行规则预筛选。具体过程如下:

1) 计算正、负类样本特征。根据式(5)计算样本的 P 个基因在矩阵 R 正类样本 C_0 中的总得分值 $Score_g^{C_0}$,根据总得分值对 P 个基因由高到低排序,将排序后得到的基因序列定义为正类样本特征 EP 。同理可以获取负类样本特征 EN 。

$$Score_g^{C_0} = \sum_{d \in C_0} I(d, g) \quad (5)$$

2) 根据式(6)计算每个基因 g 在正类样本特征 EP 上相对于负类样本特征 EN 中的相对偏移量 Dis_g 。

$$Dis_g = I(EP, g) - I(EN, g) \quad (6)$$

3) 对 P 个基因按照偏移量得分值进行排序,得到的基因序列即为相对偏移表 RST。假设在数据矩阵 R 上得到正类样本特征 $EP = daceb$ 、负类样本特征 $EN = dbeca$,则计算的相对偏移量 Dis_g 如表 2 所示,最终求得相对偏移表 $RST = acdeb$ 。

表 2 相对偏移量示例

基因	Dis_g
a	-3
b	3
c	-1
d	0
e	1

2.3 层级规则树挖掘

利用相对偏移表 RST 进行规则预筛选,目标是找到在不同类中相对表达反转的基因对 (a, b) 。令 x 为负类样本特征 EN 相对于正类样本特征 EP 向前偏移最多的基因, y 为 EN 相对于 EP 向后偏移最多的基因,则基因对 (x, y) 有极大概率在不同类样本上完成相对表达反转。向前、向后偏移最多的基因分别位于相对偏移表 T 的首、尾位置。因此,为了提高算法的稳定性,本文从相对偏移表首、尾各取出 m 个基因组合为候选基因对 Can (算法 1 中的第 4 行),并从中选取置信度最高的基因对作为目标分类规则(算法 1 中的第 5 行)。

在每一层寻找分类规则 γ 的具体过程如下:

- 1) 从相对偏移表 RST 的首、尾各随机取出 m 个基因,按照位置对应组合为 m 个候选基因对。
- 2) 每一个候选基因对 (a, b) 可以产生 4 个分

类规则: $\{ \langle a, b \rangle \rightarrow C_0 \}, \{ \langle a, b \rangle \rightarrow C_1 \}, \{ \neg \langle a, b \rangle \rightarrow C_0 \}, \{ \neg \langle a, b \rangle \rightarrow C_1 \}$ 。从中选取置信度得分最高的作为基因对 (a, b) 的唯一分类规则,共生成 m 个候选分类规则。

3) 从 m 个候选分类规则中选取置信度得分最高且满足最小支持度 θ 的规则,作为目标分类规则 γ 。

将分类规则 γ 加入层次规则树分类器中,从训练数据集上删去被分类规则 γ 前项所命中的样本,重复算法 1 的第 3 行~第 12 行,直至剩余样本数量小于最小支持度 θ ,将剩余样本中的多数类作为层次规则树分类器中的缺省类。

3 实验结果与分析

为验证本文分类模式的性能,从 NCBI 上获取多个平台的基因数据集作为原始实验数据,进行仿真对比与分析,多个数据集所共有的标签是性别与年龄。

3.1 数据预处理

原始基因表达数据样本的维度为 20 660,在这些属性上存在很多缺失值。由于这些样本来自不同的 GPL 平台,各样本上缺失值的数量与位置不同,因此需要找到缺失值最少的样本与维度。本文受数据库搜索方法 ripple join^[13]的启发,使用类似的方法进行缺失值数据的消除。首先将样本及其维度根据 NA 值的数量从高到低分别进行排序,然后记录维度映射表。初始化一个空矩阵,迭代扩展行和列,找到最大的无缺失样本矩阵。经过数据清洗后,最终可以得到无缺失值的 2 个样本集,数据分布如表 3 所示。其中,性别数据集来自 7 个平台: GPL10558, GPL6102, GPL6884, GPL4133, GPL6947, GPL6480, GPL570;年龄数据集来自 6 个平台: GPL10558, GPL6102, GPL6884, GPL6947, GPL6480, GPL570。

表 3 基因表达数据集信息

数据集	样本数量(正:负)	平台数量	特征数量
性别	5 729 (Male): 5 077 (Female)	7	13 876
年龄	9 261 (Young): 3 502 (Old)	6	12 763

3.2 实验设置

性别数据集的 2 个标签为 Male、Female,在数据预处理阶段,去除两性分别独有的基因特征,以男性、女性所共有的基因作为样本特征。年龄数据集的正、负类标签为 Young、Old,由于年龄是连续的数值,年龄数据集对 Old 标签的定义是年龄大于 60 岁的人群。

实验中的对比算法为本文 k-HRT 算法、k-TSP 算法^[7]、SVM-RFE 算法^[3]。实验分为 2 组:

实验 1 2 个数据集上的无平台信息实验。

实验 2 2 个数据集上的平台迁移实验。

本文自行从 NCBI 获取并整合实验数据,无训练集与测试集的分别。对于第 1 组无平台信息实验,训练集与测试集按 8:2 的比例进行无偏划分;对于第 2 组平台迁移实验,将源平台作为训练集、目标平台作为测试集。对于 BuildHRT 算法中的 2 个自定义参数,最小支持度 θ 设定为 $20/N$,候选基因对数量 m 为 50。用于集成的层级规则树数目 k 设定为 3,对比算法 k-TSP 中的 k 值也为 3。

3.3 结果分析

在无平台信息实验中,图 3、图 4 分别为性别、年龄 2 个数据集上各算法的准确度随样本量的变化情况。由图 3、图 4 可以看出,在类均衡的性别数据集上,k-HRT 算法与 k-TSP 算法均优于无法消除平台差异性的 SVM-RFE 算法;在类不均衡的年龄数据集上,k-HRT 算法的稳定性与准确性相比其他算法均具有较大优势。图 5、图 6 分别为性别、年龄 2 个数据集上各算法的 CPU 运行时间随样本量的变化情况。由图 5、图 6 可以看出,相比其他 2 种算法,k-TSP 算法效率较差,难以应对大规模数据集。表 4 所示为 2 个数据集上最大样本量时各算法的准确度对比。

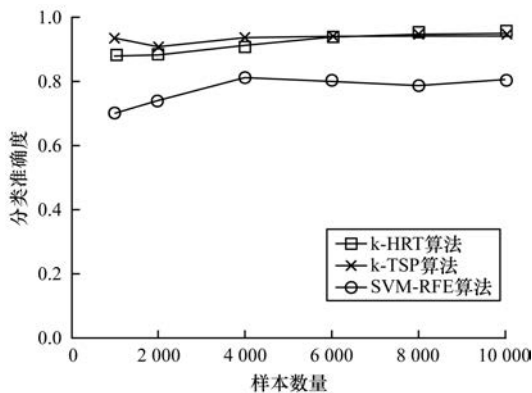


图 3 实验 1 中各算法在性别数据集上的准确度对比

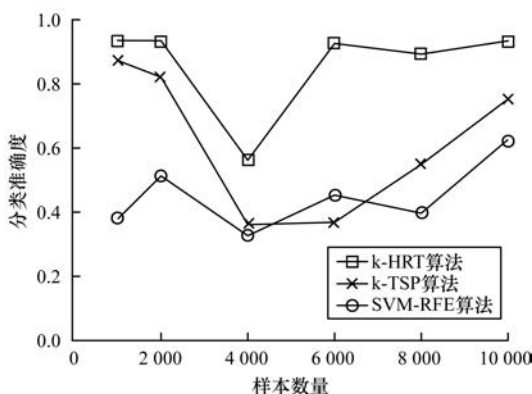


图 4 实验 1 中各算法在年龄数据集上的准确度对比

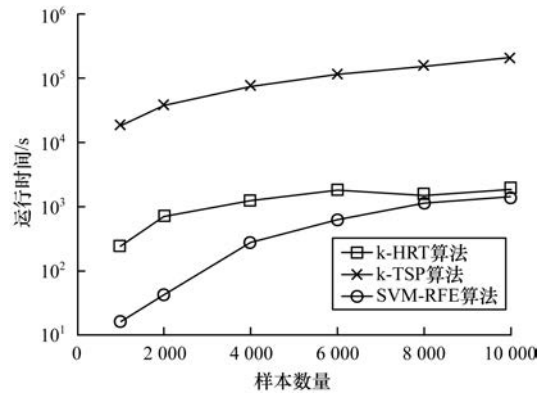


图 5 实验 1 中各算法在性别数据集上的运行时间对比

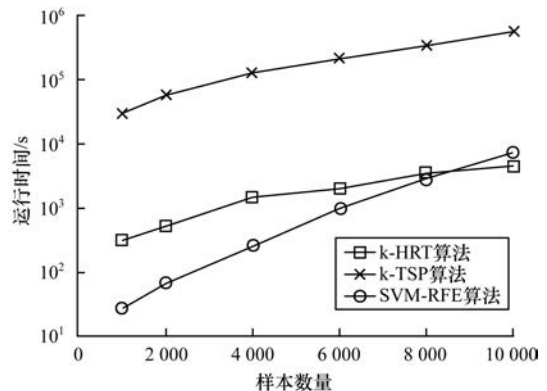


图 6 实验 1 中各算法在年龄数据集上的运行时间对比

表 4 实验 1 中各算法在最大样本量时的分类精度对比

数据集	分类精度		
	k-HRT	k-TSP	SVM-RFE
性别	0.950	0.941	0.806
年龄	0.934	0.753	0.627

对于平台迁移实验,表 5 所示为性别、年龄 2 个数据集上各算法的准确度对比,其中,每个数据集上准确度最高的值用加粗标出。由表 5 可以看出,该实验的结论与无平台信息实验基本一致。

表 5 实验 2 中各算法的分类精度对比

数据集	源平台	目标平台	分类精度		
			k-HRT	k-TSP	SVM-RFE
性别	GPL10558	GPL6884	0.843	0.892	0.799
	GPL6480	GPL4133	0.949	0.943	0.558
	GPL6884	GPL4133	0.910	0.925	0.575
年龄	GPL6947	GPL6480	0.727	0.568	0.618
	GPL10558	GPL6480	0.726	0.561	0.603
	GPL 6947	GPL10558	0.769	0.634	0.674

3.4 生物学解释

生物的衰老过程和某些基因表达水平的变化有关^[14]。本文对年龄组实验中 k-HRT 分类器上的如下分类规则进行生物学分析:

If IFNA17 < ALKBH1, then Young.

If BCORL1 < COX17, then Young.

If BCORL1 ≥ COX17 and MRAS < ZNF75D, then Old.

ALKBH1 与 DNA 烷基化损伤修复机制有关^[15]。文献[16]研究表明,ALKBH1 在胚胎干细胞转录网络中扮演着核心角色。IFNA17 由巨噬细胞产生,具有抗病毒活性,文献[17]发现其和细胞凋亡相关。文献[18]在 CEPH Utah 家族数据集上的研究结果也表明,IFNA17 的表达水平和年龄有关。COX17 则与生物的新陈代谢功能相关^[19]。BCORL1 编码的蛋白质是转录辅阻遏物,目前尚未发现 BCORL1 在生物衰老方面的相关特性。

4 结束语

传统分类模式在跨平台基因表达数据平台间难以有效迁移。为此,本文提出一种基于层级规则树的 k-HRT 算法。针对由跨平台特性所带来的大规模数据问题,设计与分类模式相切合的数据转换和规则预筛选策略。同步进行规则挖掘与分类器构建,以解决规则冗余与样本覆盖等问题。在真实基因表达数据集上的实验结果验证了 k-HRT 算法的可行性与高效性。本文所提出的数据转换与规则预筛选策略能够为相关的数据集挖掘工作提供一种新思路。下一步考虑将 k-HRT 算法应用到连续数值型跨平台数据集中,以进行数据分类。

参考文献

- [1] CLAES P, ROOSENBOOM J, WHITE J D, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape[J]. *Nature Genetics*, 2018, 50(3): 414-420.
- [2] CASPI A, SUGDEN K, MOFFITT T E, et al. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene[J]. *Science*, 2003, 301(5631): 386-389.
- [3] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. *Machine Learning*, 2002, 46(1-3): 389-422.
- [4] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. *Nature*, 2015, 521(7553): 436-440.
- [5] LIU Bing, HSU W, MA Yiming. Integrating classification and association rule mining [C]// *Proceedings of International Conference on Knowledge Discovery and Data Mining*. New York, USA: AAAI Press, 1998: 80-86.
- [6] CONG G, TAN K L, TUNG A K, et al. Mining top-k covering rule groups for gene expression data [C]// *Proceedings of 2005 ACM SIGMOD International Conference on Management of Data*. New York, USA: ACM Press, 2005: 670-681.
- [7] TAN A C, NAIMAN D Q, XU Lei, et al. Simple decision rules for classifying human cancers from gene expression profiles[J]. *Bioinformatics*, 2005, 21(20): 3896-3904.
- [8] CAI Ruichu, HAO Zhifeng, YANG Xiaowei, et al. An efficient gene selection algorithm based on mutual information[J]. *Neurocomputing*, 2009, 72(4-6): 991-999.
- [9] CAI Ruichu, TUNG A K, ZHANG Zhenjie, et al. What is unequal among the equals? Ranking equivalent rules from gene expression data[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2011, 23(11): 1735-1747.
- [10] 蔡瑞初, 王美华, 郝志峰, 等. 基于最大间隔的基因表达规则筛选[J]. *计算机工程与应用*, 2011, 47(26): 11-13.
- [11] GEMAN D, D'AVIGNON C, NAIMAN D Q, et al. Classifying gene expression profiles from pairwise mRNA comparisons [J]. *Statistical Applications in Genetics and Molecular Biology*, 2004, 3(1): 1-19.
- [12] 蔡毅, 朱秀芳, 孙章丽, 等. 半监督集成学习综述[J]. *计算机科学*, 2017, 44(增刊): 7-13.
- [13] HAAS P J, HELLERSTEIN J M. Ripple joins for online aggregation[J]. *ACM SIGMOD Record*, 1999, 28(2): 287-298.
- [14] BAHAR R, HARTMANN C H, RODRIGUEZ K A, et al. Increased cell-to-cell variation in gene expression in ageing mouse heart[J]. *Nature*, 2006, 441(7096): 1011-1012.
- [15] FEDELES B I, SINGH V, DELANEY J C, et al. The AlkB family of Fe(II)/ α -Ketoglutarate-dependent dioxygenases: repairing nucleic acid alkylation damage and beyond[J]. *Journal of Biological Chemistry*, 2015, 290(34): 20734-20742.
- [16] OUGLAND R, JONSON I, MOEN M N, et al. Role of ALKBH1 in the core transcriptional network of embryonic stem cells[J]. *Cellular Physiology and Biochemistry*, 2016, 38(1): 173-184.
- [17] FONSECA R R D, KOSIOL C, TOMÁ V, et al. Positive selection on apoptosis related genes [J]. *Febs Letters*, 2010, 584(3): 469-476.
- [18] TAN Qihua, ZHAO Jinghua, LI Shuxia, et al. Differential and correlation analyses of microarray gene expression data in the CEPH Utah families[J]. *Genomics*, 2008, 92(2): 94-100.
- [19] GLERUM D M, SHTANKO A, TZAGOLOFF A. Characterization of COX17, a yeast gene involved in copper metabolism and assembly of cytochrome oxidase[J]. *Journal of Biological Chemistry*, 1996, 271(24): 14504-14509.

编辑 吴云芳