

# 特征选择方法中三种度量的比较研究

宋智超<sup>1,2</sup>, 康健<sup>3</sup>, 孙广路<sup>1,2</sup>, 何勇军<sup>1,2</sup>

- (1. 哈尔滨理工大学 计算机科学与技术学院, 黑龙江 哈尔滨 150080;  
2. 哈尔滨理工大学 信息安全与智能技术研究中心, 黑龙江 哈尔滨 150080;  
3. 北京宇航系统工程研究所, 北京 100076)

**摘 要:** 不同类型数据中特征与类别以及特征与特征之间存在一定的线性和非线性相关性。针对基于不同度量的特征选择方法在不同类型数据集上选取的特征存在明显差别的问题, 本文选择线性相关系数、对称不确定性和互信息三种常用的线性或非线性的度量, 将它们应用于基于相关性的快速特征选择方法中, 对它们在基因微阵列和图像数据上的特征选择效果进行实验验证和比较。实验结果表明, 基于相关性的快速特征选择方法使用线性相关系数在基因数据集上选取的特征集往往具有较好分类准确率, 使用互信息在图像数据集上选取的特征集的分类效果较好, 使用对称不确定性在两种类型数据上选取特征的分类效果较为稳定。

**关键词:** 特征选择; 线性相关系数; 对称不确定性; 互信息; 基于相关性的快速特征选择方法

**DOI:** 10.15938/j.jhust.2018.01.020

**中图分类号:** TM391.1      **文献标志码:** A      **文章编号:** 1007-2683(2018)01-0111-06

## The Comparison of Three Measures in Feature Selection

SONG Zhi-chao<sup>1,2</sup>, KANG Jian<sup>3</sup>, SUN Guang-lu<sup>1,2</sup>, HE Yong-jun<sup>1,2</sup>

- (1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;  
2. Research Center of Information Security and Intelligent Technology, Harbin University of Science and Technology, Harbin 150080, China;  
3. Beijing Institute of Astronautical Systems Engineering, Beijing 100076, China)

**Abstract:** It has been known that either linear correlation or nonlinear correlation might exist between feature-to-feature and feature-to-class in datasets. In this paper, we study the differences of selected feature subset when different kinds of measures are applied with same feature selection method in different kinds of datasets. Three representative linear or nonlinear measures, linear correlation coefficient, symmetrical uncertainty, and mutual information are selected. By combining them with the fast correlation-based filter (FCBF) feature selection method, we make the comparison of selected feature subset from 8 gene microarray and image datasets. Experimental results indicate that the feature subsets selected by linear correlation coefficient based FCBF obtain better classification accuracy in gene microarray datasets than in image datasets, while mutual information and symmetrical uncertainty based FCBF tend to obtain better results in image datasets. Moreover, symmetrical uncertainty based FCBF is more robust in all datasets.

**Keywords:** feature selection; linear correlation coefficient; symmetrical uncertainty; mutual Information; fast correlation-based filter

**收稿日期:** 2016-05-17

**基金项目:** 黑龙江省普通高等院校新世纪优秀人才培养计划(1155-ncet-008); 黑龙江省教育科学规划课题(GBC1211062); 黑龙江省自然科学基金(QC2015084)。

**作者简介:** 宋智超(1990—), 男, 硕士研究生;

康健(1985—), 男, 博士, 工程师。

**通信作者:** 孙广路(1979—), 男, 博士, 教授, E-mail: guanglu\_sun@gmail.com.

## 0 引言

数据挖掘方法能够从数据中获取到潜在的有效信息,在金融预测、模式识别等多个领域得到了广泛应用。随着互联网和生物信息学技术的不断进步,数据朝着更大规模的方向发展,并带来了“维度灾难”等问题<sup>[1]</sup>。解决上述问题的有效方法之一是降低数据集中特征的维数。特征选择作为数据挖掘和机器学习中的重要研究内容,其通过删除数据集中的无关和冗余特征,达到有效的降低特征维数,提高分类的准确率和效率的目的,并且具有去噪、防止机器学习模型过拟合的作用<sup>[2]</sup>。

现有的特征选择方法主要可以分为过滤方法、封装方法和嵌入方法<sup>[3]</sup>。封装方法使用预先选定的机器学习方法作为评价特征集优劣的准则,存在时间复杂度高的问题。嵌入方法则将特征选择和机器学习算法的训练过程相结合。过滤方法不依赖特定的机器学习方法,具有运行效率高的特点,适用于解决高维数据中的特征选择问题。本文主要针对过滤方法进行研究。

搜索策略和度量的选取是过滤方法的两个重要研究内容。学者们提出了基于一致性、基于距离、基于信息论等多种度量,并据此提出了多种评价函数<sup>[4-6]</sup>。当前研究者们重点关注特征选择方法的设计,实验常用数据集主要有基因生物数据、图像数据和文本数据等<sup>[7]</sup>。据作者调研,目前尚无针对不同度量在不同类型数据上可能存在的效果差异性的研究。本文选取常用的三种度量——线性相关系数、对称不确定性和互信息,并结合经典的特征选择方法,对这3种度量应用到不同类型数据集上的效果进行研究。

基于相关性的快速特征选择方法是一种经典的特征选择方法,其在多种数据集上都具有较好的效果,并且对于高维数据具有较快的运行效率。本文将上述不同的度量应用于基于相关性的快速特征选择方法中,通过实验验证对不同度量在基因生物数据和图像数据上效果的差异,并对度量和数据类型之间的关系进行研究。

本文第二节为相关工作,对目前影响较大的特征选择方法和度量的应用进行介绍;第三节描述特征选择中的3种度量和基于相关性的快速特征选择方法;第四节是实验数据集和实验结果,第五节为总结。

## 1 相关工作

变量间的相关关系在机器学习和模式识别领域得到了广泛的研究。研究者们提出了多种度量对变量间的相关性进行挖掘,目前而言,变量之间的相关关系主要分为线性相关和非线性相关两类。早期的特征选择方法一般应用马氏距离、相关系数等线性度量<sup>[8]</sup>。文<sup>[9]</sup>使用相关系数、Wilcoxon 秩和检验两种度量对基因数据中的特征关系进行挖掘。文<sup>[10]</sup>提出了最小平方误差和最大信息压缩指数两种线性度量并应用于无监督的特征选择方法中,取得了较好的效果。然而,现实世界中的数据并不总是满足线性关系,对数据间线性关系的假设并不完备<sup>[11]</sup>。针对这种情况,学者们提出了多种非线性相关的度量,其中基于信息论的度量被认为是最有前景的度量,信息增益<sup>[12]</sup>、互信息<sup>[13]</sup>、归一化互信息<sup>[14]</sup>和条件互信息<sup>[15]</sup>等被应用到特征选择中,取得了不错的效果。

基于上述度量可以构建特征选择方法进行最优特征子集的选取。早期的特征选择方法只考虑特征与类别之间的相关性,如信息增益、Relief<sup>[16]</sup>和 Relieff<sup>[17]</sup>等。随着特征维数的增加,该类方法的时间复杂度呈线性增长并且能够适用于高维数据的特征选择。但是由于没有考虑冗余特征的影响,该类方法选取特征子集的分类效果往往不理想。

冗余特征的存在不仅增加了机器学习模型的时间复杂度,而且对最后的分类任务有干扰作用,也应该被去除。基于相关性的特征选择<sup>[18]</sup>、最小冗余最大相关<sup>[19]</sup>等方法可以对冗余特征进行处理,然而其使用的贪心序列搜索、最优搜索等搜索策略的时间复杂度为 $O(n^2)$ ,使得这些方法很难应用到高维数据的特征选择中。

针对上述问题,马尔科夫毯首次被 Koller 等人应用到特征选择中,取得了很好的效果<sup>[20]</sup>。随后的学者们对马尔科夫毯方法进行了广泛的研究<sup>[7,11,21]</sup>。其中,论文<sup>[11]</sup>提出一种基于相关性的快速特征选择方法,并对特征选择中的基本问题进行了定义。后续研究者在此基础上进行改进并应用到不同的特征选择任务中<sup>[7,22]</sup>。从算法效率和选取的特征子集的分类效果两方面来看,基于相关性的快速特征选择方法具有一定的优势。

## 2 特征选择中的度量和方法

首先对本文中使用的符号进行定义,给定特征

集  $F = \{f_1, f_2, \dots, f_m, c\}$ , 其中特征数为  $m$ , 样本类别为  $c$ , 包含该特征集定义特征的样本数为  $n$ 。

## 2.1 3 种度量

### 2.1.1 线性相关系数

线性相关系数又称相关系数或者 Pearson 系数, 是最常见的随机变量间相似性度量, 得到了广泛的应用。但是, 它只能对变量间的线性相关性进行度量。给定任意特征  $f_i$  和类别  $c$ , 相关系数具体定义如下。

$$\rho(f_i, c) = \frac{\text{cov}(f_i, c)}{\sqrt{\text{var}(f_i)} \sqrt{\text{var}(c)}} = \frac{\sum_{k=1}^n (f_{k,i} - \bar{f}_i) (c_k - \bar{c})}{\sqrt{\sum_{k=1}^n (f_{k,i} - \bar{f}_i)^2} \sqrt{\sum_{k=1}^n (c_k - \bar{c})^2}} \quad (1)$$

其中,  $f_{k,i}$  是第  $k$  个样本的第  $i$  维特征;  $\bar{f}_i$  是第  $i$  维特征在所有样本上的均值;  $c_k$  表示第  $k$  个样本的类别。若特征与类别间完全线性相关, 则  $\rho(f_i, c)$  值为 1 或者 -1; 若特征与类别间完全不相关, 则  $\rho(f_i, c)$  值为 0。故  $|\rho(f_i, c)|$  值在 0 ~ 1 间变化,  $|\rho(f_i, c)|$  值越大表示特征与类别间的相关性越强, 特征对分类的贡献越大, 且相关系数满足对称性  $|\rho(f_i, c)| = |\rho(c, f_i)|$ 。特征与特征间关系类似,  $|\rho(f_i, f_j)|$  值越大, 表示特征间的相似性越强。

### 2.1.2 对称不确定性

信息熵作为信息量的重要度量得到了广泛的应用。研究者们基于信息熵理论, 提出了很多衡量变量间线性和非线性关系的度量, 如信息增益、对称不确定性、互信息和条件互信息等。给定随机变量  $X$ , 其熵定义为:

$$H(X) = - \sum_i P(x_i) \log(P(x_i)) \quad (2)$$

其中,  $P(x_i)$  为变量  $X$  的先验概率。在获取观测值  $Y$  后, 变量  $X$  的条件熵定义为:

$$H(X|Y) = - \sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j)) \quad (3)$$

其中  $P(x_i|y_j)$  为变量的后验概率。给定上述定义, 信息增益定义为:

$$IG(X|Y) = H(X) - H(X|Y) \quad (4)$$

因为不均衡数据集中信息增益更倾向于选取高频特征, 所以需要使用时熵值对其进行归一化来修正。由此提出对称不确定性的定义:

$$SU(X, Y) = 2 \left[ \frac{IG(X|Y)}{H(X) + H(Y)} \right] \quad (5)$$

对称不确定性满足对称性, 即  $SU(X, Y) = SU$

$(Y, X)$ , 通过添加熵值的约束补偿了高频特征的信息增益值并且得到归一化的值。 $SU(X, Y)$  值为 1, 表示变量  $X$  与变量  $Y$  间完全相关; 值为 0, 表示两变量独立。

### 2.1.3 互信息

互信息也是一种基于信息论的度量, 可以对非线性关系进行度量。与上述两种度量类似, 其也满足对称性。给定随机变量  $X$  和  $Y$ , 若  $X, Y$  为离散型随机变量, 则其互信息定义为:

$$MI(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (6)$$

若  $X, Y$  为连续型随机变量, 则其互信息定义为:

$$MI(X, Y) = \iint_{X,Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad (7)$$

两变量间互信息值越大, 表示变量间相关性越强。 $MI(X, Y)$  为 0, 则表示两变量独立。

## 2.2 基于相关性的快速特征选择方法

首先对数据中特征与类别间的相关性以及特征与特征间的相关性进行定义。

定义 1 (特征 - 类别相关) 特征  $f_i \in F$  与类别  $c$  间的相关性定义为 FC 相关, 使用  $FC(f_i, c)$  表示。

定义 2 (特征 - 特征相关) 二维特征  $f_i$  和  $f_j$  之间相关性定义为 FF 相关, 使用  $FF(f_i, f_j)$  表示。

根据上述定义, 基于相关性的快速特征选择方法的特征选择过程如下: 第一阶段, 计算特征与类别之间的 FC 值, 并保留特征集中的相关特征; 第二阶段, 应用近似马尔科夫毯删除第一阶段保留的特征集中的冗余特征。算法描述如算法 1 所示。

本文分别使用线性相关系数、对称不确定性和互信息来定义 FC 相关和 FF 相关, 并应用于算法 1 中, 得到的特征选择方法分别称为 FSCC、FSSU 和 FSMI。

算法 1 中主元素为待选特征集中  $FC(f_i, c)$  值最大的特征, 每次迭代均选取  $FC(f_i, c)$  最大的特征添加到最优特征集中。然后利用该主元素对与主元素冗余的特征进行删除, 算法中阈值  $\theta$  根据预期选取特征的维数设定。

算法 1:

输入:  $F(f_1, f_2, \dots, f_m, c)$  // 训练集

$\theta$  // 相关性阈值

输出:  $F_{opt}$  // 选择的特征子集

// 第一阶段: 无关或者弱相关特征删除

1) for  $f_i \in F$

2) 计算特征和类别之间的  $FC(f_i, c)$  相关得分

- 3) if  $FC(f_i, c) > \theta$
- 4) 将特征  $f_i$  添加到特征子集  $F'$  中
- 5) end for
- //第二阶段: 冗余特征删除
- 6) 对  $F'$  中特征按照  $FC(f_i, c)$  值降序排序并将排序结果赋给  $F_{order}$
- 7) while  $F_{order} \neq \emptyset$
- 8) 查找  $F_{order}$  中的主元素  $f_i$ , 并添加到  $F_{opt}$  中
- 9) 将  $f_i$  从  $F_{order}$  中删除
- 10) 查找以  $f_i$  为近似马尔科夫毯的特征子集  $\{f_j\}$
- 11) 将冗余特征子集  $\{f_j\}$  中元素从  $F_{order}$  中删除
- 12) end while

算法1中使用近似马尔科夫毯定义冗余特征, 其基于 Pearl 等人提出的马尔科夫毯理论, 相关定义如下。

定义3 (马尔科夫毯) 给定特征  $f_i \in F$  和特征子集  $M_i \subset F (f_i \notin M_i)$ , 当且仅当公式(8)成立, 特征集  $M_i$  为  $f_i$  的马尔科夫毯。

$$p(F - M_i - \{f_i\}, c | f_i, M_i) = p(F - M_i - \{f_i\}, c | M_i) \quad (8)$$

由上述定义可知, 如果  $M_i$  为  $f_i$  的马尔科夫毯, 那么由于特征集  $M_i$  的存在,  $f_i$  对类别的区分性没有贡献, 即  $f_i$  的存在不能提高分类效果, 应该被删除。

通过马尔科夫毯条件来判断条件独立具有很高的计算复杂度, 因此算法1中使用单个特征来对某一特征  $f_i$  的马尔科夫毯条件进行近似, 并得到近似马尔科夫毯, 定义如下。

定义4 (近似马尔科夫毯) 给定特征集  $F$  中特征  $f_i$  和  $f_j$ , 如果  $FC(f_i, c) \geq FC(f_j, c) \wedge FF(f_i, f_j) > FC(f_j, c)$  成立, 则特征  $f_i$  为特征  $f_j$  的近似马尔科夫毯。

上述定义中假定, 如果某一特征  $FC$  值越大, 则该特征包含了类别越多的信息,  $FC(f_i, c)$  大于  $FC(f_j, c)$ , 则保留特征  $f_i$  可以保留对类别决策的更多信息。同时利用  $FC(f_j, c)$  来对两维特征间  $FF$  相关的程度进行限定。

### 3 实验结果与分析

#### 3.1 实验设置

为了验证本文提出的3种度量在基因和图像数据上选取特征的分类效果是否存在差异, 选取8个数据集进行实验研究。由于对选取的数据集无法事先得知最优特征子集, 同时为了增强实验的说服力、

避免实验结果的偏置, 在不同数据集上应用本文提出的3种特征选择方法 FSCC、FSSU 和 FSMI 分别选取 10, 20, 30, 40 维特征, 对3种不同特征选择方法选取特征差异性进行比较。由于对数据集我们没有先验知识, 当前特征选择工作一般使用分类器的准确率对最终选取的特征集优劣进行评价。本文使用常用的朴素贝叶斯(Naïve Bayes, NB)和支持向量机(Support Vector Machine, SVM)分类器, 统一使用 10-fold 交叉验证得到3种特征选择方法选取不同维数特征的分类准确率。

实验中将数据随机均等分成2份, 1份为训练集, 1份为测试集, 使用本文提出的3种特征选方法从训练集中选取预先设定维数的特征, 然后根据选取的特征子集重新构造测试集, 并应用 NB 和 SVM 分类器, 采用 10 折交叉验证得到不同特征选择方法选取的特征集在测试集上的分类准确率。为了使得最后的实验结果更具统计意义, 重复上述实验过程 10 次, 并对 10 次实验的结果取平均值得到最终的分

#### 3.2 数据集

实验中使用基因和图像两类数据集, 对3种度量的效果进行实验分析。每类数据选取四个不同的数据集, 有二分类也有多分类数据集, 特征维数从 280 维到 19993 维, 具体数据信息如表1所示。

表1 数据集

| 编号 | 数据集        | 特征总数  | 样本数 | 类别数 | 类型 |
|----|------------|-------|-----|-----|----|
| 1  | SMK-CAN    | 19993 | 187 | 2   | 基因 |
| 2  | TOX-171    | 5748  | 171 | 4   | 基因 |
| 3  | Arrhythmia | 279   | 452 | 16  | 基因 |
| 4  | Leukemia   | 7129  | 72  | 2   | 基因 |
| 5  | PIE10P     | 2420  | 210 | 10  | 图像 |
| 6  | ORL10P     | 10304 | 100 | 10  | 图像 |
| 7  | PIX10P     | 10000 | 100 | 10  | 图像 |
| 8  | AR10P      | 2420  | 130 | 10  | 图像 |

#### 3.3 实验处理和结果

本文算法1为两阶段特征选择方法, 算法第一阶段通过对 FSCC、FSSU 和 FSMI 3 种方法设定不同的阈值  $\theta$ , 选取预期维数的特征。表2为3种特征选择方法分别选取不同维数特征在 NB 分类器上的实验结果。表3为3种特征选择方法选取的特征在 SVM 分类器上的实验结果。

表 2 3 种特征选择方法选取的特征在 NB 分类器上的实验结果

%

| 特征维数       | FSCC  |       |       |       | FSSU  |       |       |       | FSMI  |       |       |       |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|            | 10    | 20    | 30    | 40    | 10    | 20    | 30    | 40    | 10    | 20    | 30    | 40    |
| SMK-CAN    | 67.38 | 70.21 | 71.99 | 71.99 | 65.7  | 65.1  | 66.02 | 65.1  | 59.56 | 65.95 | 60.63 | 58.86 |
| TOX-471    | 69.38 | 72.09 | 68.21 | 69.77 | 64.5  | 66.1  | 66.7  | 68    | 65.5  | 67.83 | 69.38 | 67.05 |
| Arrhythmia | 57.52 | 56.34 | 55.16 | 59.30 | 59.78 | 67.46 | 63.58 | 66.89 | 61.77 | 59.29 | 57.51 | 50.59 |
| Leukemia   | 93.52 | 96.29 | 96.29 | 96.29 | 94.5  | 95.31 | 95.86 | 96.11 | 95.37 | 92.59 | 94.45 | 94.44 |
| PIE10P     | 63.17 | 71.9  | 76.18 | 81.27 | 86.1  | 88.9  | 88.9  | 90.56 | 91.42 | 92.37 | 94.28 | 92.06 |
| ORL10P     | 60.67 | 57.33 | 64    | 61    | 78.9  | 76.5  | 75.46 | 73.88 | 84.67 | 81.33 | 81.33 | 79.33 |
| PIX10P     | 80    | 84    | 81.33 | 80.67 | 93.33 | 94.44 | 97.56 | 96.89 | 98    | 93.33 | 97.33 | 97.33 |
| AR10P      | 44.57 | 52.3  | 53.33 | 58.46 | 63.59 | 64.22 | 66.12 | 64.57 | 68.20 | 63.59 | 65.13 | 65.65 |

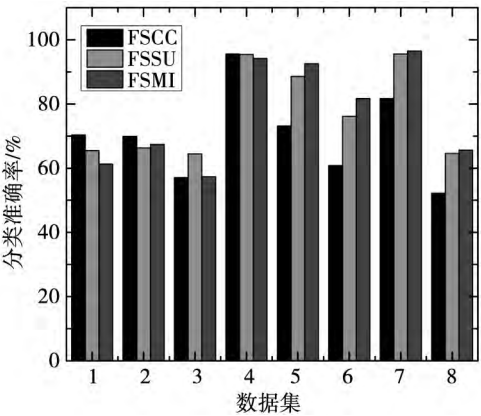


图 1 3 种特征选择方法在 NB 分类器上的实验结果

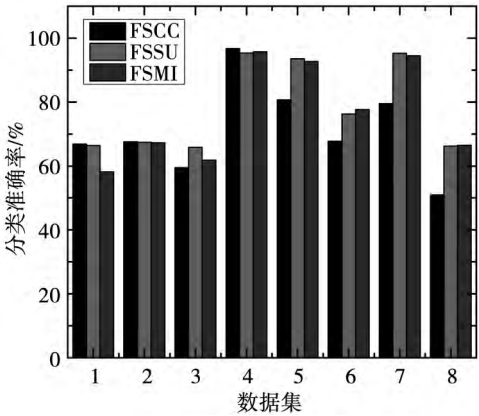


图 2 3 种特征选择方法在 SVM 分类器上的实验结果

图 1 和图 2 为 3 种特征选择方法选取不同维数特征在 NB 和 SVM 两个分类器上的准确率均值。从表 2 和图 1 中实验结果来看,FSCC 在 SMK-CAN、TOX-471 和 Leukemia 3 个基因类型数据集上的分类准确率最高,而在四个图像类型数据上的分类准确率较 FSSU 和 FSMI 方法有明显的差距。在 Arrhythmia 数据集上与 FSMI 分类准确率相近。FSMI 在 4 个图像数据上的分类效果最好,但在四个基因类型的数据集上的分类效果较差。FSSU 在 Arrhythmia 数据上的分类效果最好,在 TOX-471 数据集上的分类准确率最差,在其余的六个数据集上的效果与分类效果最好的方法效果相近。因此,从最终的分类结果来看,FSMI 效果最好,并且其更适合处理图像

类型数据。而 FSCC 更适合处理基因数据,并且 FSCC 在图像数据上的分类效果明显差于 FSMI 和 FSSU 方法。尽管 FSSU 方法只在 Arrhythmia 数据上的分类准确率最高,但是从所有八个数据集上的分类准确率来看,FSSU 选取的特征在不同数据集上的分类效果更加稳定。

图 2 和表 3 中在 SVM 分类器上的实验结果与图 1 和表 2 中实验结果类似,FSCC 方法仍然在 SMK-CAN、TOX-471 和 Leukemia 3 个基因数据上的分类准确率最高,FSSU 在 Arrhythmia、PIE10P 和 PIX10P 3 个数据集上的效果最好,FSMI 在其他两个图像数据上的效果最好。

表 3 3 种特征选择方法选取的特征在 SVM 分类器上的实验结果

%

| 特征维数       | FSCC  |       |       |       | FSSU  |       |       |       | FSMI  |       |       |       |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|            | 10    | 20    | 30    | 40    | 10    | 20    | 30    | 40    | 10    | 20    | 30    | 40    |
| SMK-CAN    | 65.15 | 66.7  | 67.15 | 68.46 | 64.8  | 65.78 | 65.89 | 68.99 | 58.17 | 60.85 | 67.77 | 70.15 |
| TOX-471    | 59.7  | 68.7  | 73.16 | 68.87 | 66.14 | 68.46 | 65.7  | 69.5  | 65.15 | 66.44 | 70.3  | 67.15 |
| Arrhythmia | 60.17 | 57.1  | 63.1  | 57.76 | 63.15 | 68.91 | 66.4  | 65.11 | 60.8  | 61.39 | 62.11 | 63.14 |
| Leukemia   | 96.78 | 97.13 | 96.77 | 96.4  | 95.33 | 94.58 | 94.44 | 94.44 | 95.7  | 95.4  | 94.61 | 94.7  |
| PIE10P     | 72.56 | 81.46 | 83.7  | 84.76 | 90.56 | 94.36 | 97.13 | 92.11 | 90.7  | 92.88 | 93.12 | 94.21 |
| ORL10P     | 62.41 | 78.16 | 68.11 | 62.14 | 75.16 | 77.99 | 75.13 | 76.7  | 76.14 | 77.32 | 76.8  | 80.45 |
| PIX10P     | 79.1  | 77.43 | 79.41 | 82.16 | 92.78 | 95.13 | 96.57 | 96.66 | 93.77 | 95.4  | 96.35 | 92.33 |
| AR10P      | 40.88 | 50.49 | 56.7  | 55.86 | 64    | 65.28 | 67.79 | 68.02 | 68.7  | 65.4  | 66.32 | 65.55 |

由上述实验结果可得,线性相关系数适合基因类型数据的特征选择工作,而在图像类型数据上选取特征的分类准确率较差。互信息和对称不确定性更适合处理图像类型的数据,对称不确定性在两种类型的数据上选取特征的分类效果较为稳定。

## 4 结 语

本文选取基因和图像两种特征选择常用类型数据集,对特征选择常用的3种度量——线性相关系数、对称不确定性和互信息在不同数据集上的效果进行研究。为了加快特征选择的效率,同时保证选取特征的分类效果,将3种度量应用到基于相关性的快速特征选择方法中,并提出FSCC、FSSU和FS-MI 3种不同的特征选择方法。使用朴素贝叶斯和SVM两种分类器评价3种不同特征选择方法选取的特征。在选取的8个数据集上的实验表明,线性相关系数更适合于处理基因类型数据,选择的特征能够取得较好的分类效果,而在图像数据集上的效果较差;互信息在图像类型数据上的效果较在基因类型数据上更为突出。对称不确定性在两种类型的数据上的效果较为稳定,且效果较好。

## 参 考 文 献:

- [1] CHANDRASHEKAR G, SAHIN F. A Survey on Feature Selection Methods [J]. Computers & Electrical Engineering, 2014, 40(1): 16-28.
- [2] DESSI N, PES B. Similarity of Feature Selection Methods: An Empirical Study Across Data Intensive Classification Tasks [J]. Expert Systems with Applications, 2015, 42(10): 4632-4642.
- [3] ZHAO Z, LIU H. Searching for Interacting Features [C] // Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 1156-1161.
- [4] DASH M, LIU H, MOTODA H. Consistency Based Feature Selection [C] // Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. Springer-Verlag, 2000: 98-109.
- [5] ZHANG J G, DENG H W. Gene Selection for Classification of Microarray Data Based on the Bayes Error [J]. BMC bioinformatics, 2007, 8(1): 370.
- [6] SOTOCA J M, PLA F. Supervised Feature Selection by Clustering Using Conditional Mutual Information-based Distances [J]. Pattern Recognition, 2010, 43(6): 2068-2081.
- [7] SONG Q, NI J, WANG G. A Fast Clustering-based Feature Subset Selection Algorithm for High-dimensional Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 1-14.
- [8] GUYON I, ELISSEEFF A. An Introduction to Variable and Feature Selection [J]. The Journal of Machine Learning Research, 2003, 3: 1157-1182.
- [9] 谢娟英, 高红超. 基于统计相关性与 K-means 的区分因子集选择算法 [J]. 软件学报, 2014, 25(9): 2050-2075.
- [10] MITRA P, MURTHY C A, PAL S K. Unsupervised Feature Selection Using Feature Similarity [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312.
- [11] YU L, LIU H. Efficient Feature Selection Via Analysis of Relevance and Redundancy [J]. The Journal of Machine Learning Research, 2004(5): 1205-1224.
- [12] PEREIRA R B, PLASTINO A, ZADROZNY B, et al. Information Gain Feature Selection for Multi-Label Classification [J]. Journal of Information and Data Management, 2015, 6(1): 48.
- [13] HOQUE N, BHATTACHARYYA D K, KALITA J K. MIFS-ND: A Mutual Information-based Feature Selection Method [J]. Expert Systems with Applications, 2014, 41(14): 6371-6385.
- [14] LEE S, PARK Y T, d'Auriol B J. A Novel Feature Selection Method Based on Normalized Mutual Information [J]. Applied Intelligence, 2012, 37(1): 100-120.
- [15] FLEURET F. Fast Binary Feature Selection with Conditional Mutual Information [J]. The Journal of Machine Learning Research, 2004, 5: 1531-1555.
- [16] KIRA K, RENDELL L A. The Feature Selection Problem: Traditional Methods and a New Algorithm [C] // Tenth National Conference on Artificial Intelligence. AAAI Press, 1992: 129-134.
- [17] KONONENKO I. Estimating Attributes: Analysis and Extensions of RELIEF [C] // European Conference on Machine Learning on Machine Learning. Springer-Verlag New York, Inc., 1994: 356-361.
- [18] HALL M A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning [C] // Seventeenth International Conference on Machine Learning. Morgan Kaufmann, 2000: 359-366.
- [19] DING C, PENG H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data [J]. Journal of bioinformatics and computational biology, 2005, 3(2): 185-205.
- [20] KOLLER D. Toward Optimal Feature Selection [C] // Proceedings of 13th International Conference on Machine Learning. Morgan Kaufmann, 2000: 284-292.
- [21] 崔自峰, 徐宝文, 张卫丰, 等. 一种近似 Markov Blanket 最优特征选择算法 [J]. 计算机学报, 2007, 30(12): 2074-2081.
- [22] W Xindong, Y Kui, D Wei, et al. Online Feature Selection with Streaming Features [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(5): 1178-1192.

(编辑: 关 毅)