计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):173−179 [doi: 10.15888/j.cnki.csa.005891]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于互信息的组合特征选择算法①

李叶紫, 周怡璐, 王振友

(广东工业大学 应用数学学院, 广州 510006)
通讯作者: 王振友, E-mail: zhenyouw@gdut.edu.cn

摘　要: 对候选特征进行降维在机器学习领域, 如分类、聚类问题中占有很重要的地位. 现有的方法大多数是基于单一特征对目标 T 的依赖性或特征与特征之间对 Y 影响的关联性, 互补性和冗余性进行特征选择. 然而这些方法几乎都没有考虑到组合特征, 如属性 A, B 仅包含 Y 中的极少量信息, 甚至与 Y 完全独立, 但 A&B 能提供关于 Y 的大量信息, 甚至完全决定 Y. 基于此, 提出了一种能够从特征集合中挖掘到组合特征与单一特征的特征选择算法, 首先对不显著特征进行组合并按照条件概率分布表生成新的候选特征; 然后, 对单一特征和组合特征利用基于最大相关性和最小冗余度的准则进行选择. 最后分别在虚拟和真实数据集上进行实验, 实验结果表明该特征选择算法能够较好的挖掘数据集的组合特征信息, 一定程度上提高了相应的机器学习算法的准确率.

关键词: 组合特征; 特征选择; 最大相关性; 最小冗余度

引用格式: 李叶紫,周怡璐,王振友.基于互信息的组合特征选择算法.计算机系统应用,2017,26(8):173–179. http://www.c-s-a.org.cn/1003-3254/5891.html

Combined Feature Selection Algorithm Based on Mutual Information

LI Ye-Zi, ZHOU Yi-Lu, WANG Zhen-You

(Department of Applied mathematics, Guangdong University of Technology, Guangzhou 510520, China)

Abstract: It is very important to reduce the candidate features in the machine learning such as classification and clustering. Most of the existing methods are based on a single feature on the target T or the association between the feature and the feature on the Y. However, these methods do not take into the combined features, such as attributes A, B contains a little amount of information in Y, and even completely independent of Y, but A & B can provide information on Y lot of information, or even completely determine the Y. Based on this, we can extract an algorithm to find single and combined features from the feature set, firstly combination of non-significant features in accordance with the conditional probability distribution table to generate new candidate features Then, the single feature and the combined features are chosen based on the criterion of the maximum correlation and the minimum redundancy. Finally, the experiment is carried out on the virtual and real data sets respectively, and the experimental results show that the feature selection algorithm can mine the dataset better, Which improves the accuracy of the corresponding machine learning algorithm to a certain extent.

Key words: combined feature; feature selection; max-relevance; min-redundancy

引言

对候选的特征进行维数降维在机器学习领域, 如分类、聚类, 预测中占有很重要的地位, 其学习能力很大程度地依赖特征集的选择. 尽管实验表明[1], 如分类算法在先进行特征选择后往往能比不进行特征选择的预测效果要好, 而且可以很大程度上提高训练速度, 但

① 基金项目: 国家自然科学基金(11401115)
收稿时间: 2016-12-05; 采用时间: 2016-12-26

近十年来, 虽然很多特征选择算法被提出[2-4], 但几乎没有算法能够考虑到组合特征, 如属性 A, B 仅包含 Y 中的极少量信息, 甚至与 T 完全独立, 但 A&B 却能提供关于 Y 的大量信息, 甚至完全决定 Y. 由于 A 与 Y 的关联程度, 依赖程度都很低, 在特征选择的初期往往就会被当做无用的特征被移除, 导致组合特征无法被识别出来.

现时比较常用的适用于 SVR 的特征选择算法几乎都是基于最大依赖性准则(Max-Dependence)[5]. 在特征选择中, 最大依赖性准则目的去寻找一个包含 m 个特征的集合 S, 使得该集合与待预测变量 y 之间存在最大的依赖关系(依赖关系一般使用互信息来评估), 如公式(1)所示:

$$\max_{S\subset X} D(y,S), D = I(y;S) \quad (1)$$

而在实际操作上, 由于候选的特征往往是高维的, 要在高维上对公式(1)进行估算一般是做不到的, 在当维数增加时, 互信息的估算值会与真实值会随之递增. 鉴于这种情况, 一些学者提出了解决方法. 例如 mRMR 算法, 利用最大相关性准则(Max-Relevance)和最小冗余度准则(Min-Redundance)来逼近公式(1); MRMS 算法则利用最小冗余性准则(Min-Redundance)和最大显著性准则(Max-significant)对公式(1)进行概率性估算; MIGS 算法[4]同样利用(条件)互信息对公式(1)的值进行估算. 尽管使用这些特征选择方法后, 如分类算法能够一定程度地提升学习精度和速度, 但都仅仅是对公式(1)的一种逼近, 并没有考虑到组合特征对目标的意义. 这些方法有着一个明显的缺点, 例如 y 为需要预测的目标变量, $X=\{x_1,x_2,x_3,x_4,x_5\}$为 y 的初始特征集, 其中 x1, x2 是 Y 的组合特征, 也就是单独 x1 或者 x2 携带着的关于 Y 的信息量非常少. 在这种情况下, 因为 x3, x4, x5 对 Y 的依赖性要比 x1, x2 大得多, 于是 x1, x2 被完全移除掉. 事实上, x1&x2 可能包含着关于 Y 的一些很重要的信息, 缺少这些信息, 将严重影响算法对 Y 进行较为准确的预测.

本质上, 大多数情况下特征与目标之间存在着一种确定和被确定的因果关系[6,7], 如图 1(a), 一个人的收入被他工作的两个特征, 底薪和提成决定, 可以看到每一个特征跟收入都是不独立的, 一定程度上影响了收入的高低; 但有时候会出现组合特征的情况, 如图 1(b), 日期和地区为气温的两个特征, 但日期跟气温关系并不大, 如南北半球差异, 沙漠, 海洋等不同地域差异非常大, 同样的, 仅知道地区不知道日期也无法对气温进行预测, 而日期与地区是一个组合特征. 现时的大多特征选择算法主要考虑的是每一个特征对目标关联性, 因而组合特征往往会被忽略.

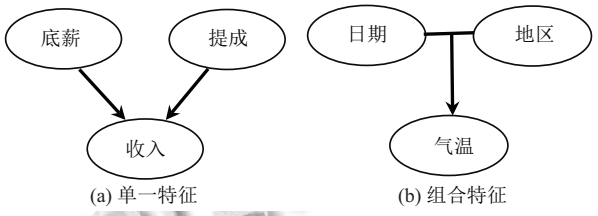

图 1 单一特征与组合特征

组合特征广泛存在于现实生活当中, 比如生物医学领域的合成致死问题[8,9], 两个或多个突变基因(或蛋白质)的相互作用如合并等, 导致细胞的死亡, 然而其中单独一个突变基因不会对细胞活性产生影响. 合成致死作用在抗肿瘤药物筛选中具有重要意义, 有望成为新一代抗肿瘤药物的筛选作用, 其核心就在于如何判断哪些突变基因可能是组合特征, 对细胞活性会产生影响.

然而现存的特征选择算法, 大多是基于单个特征对目标的依赖性或特征与特征之间对目标影响的关联性, 互补性和冗余性进行特征选择. 基于此, 提出一种能够从特征集合中挖掘到组合特征与单一特征的特征选择算法, 该算法首先对不显著特征进行组合并按照条件概率分布表生成新的候选特征; 然后, 利用基于最大相关性, 最小冗余度的准则进行对单一特征和组合特征进行选择.仿真数值实验和在应用在真实数据集的实验结果表明, 该算法应用在分类算法上, 预测的精确度要高于其他特征选择算法.

1 组合特征的特征选择

1.1 组合特征

一般而言, 特征携带着目标的一定量信息, 以此对目标进行分析, 预测, 推断等. 传统的特征方法里一般关注的是单一特征, 对组合特征往往不重视, 组合特征可以表述为以下:

给定一个目标 Y, 与其特征集 X1~Xn, 如果 W 是 X1~X2 的一个组合, 则:

1) 任意 X1 携带关于 Y 的信息非常少, 或者独立于 Y.

2) 最少存在一个 X1 携带关于 Y 的信息非常少, 或者独立于 Y, 一旦加上 X1, 其余的能提供更多关于 Y 的信息.

3) 任意 X1 都携带着关于 Y 的些信息, 也就是任意 X1 不独立于 Y.

显然, 第三种组合特征是很容易被发现的, 所以一般来说挖掘组合特征的目的是要发现第一和第二种组合特征. 而第二种实际上是广泛存在的, 例如图 2, 月份(1~12 月)和日期(1~31 日)是某地气温的两个的特征, 其中显然月份与气温是有着很大关系的, 而日期与气温关系则不大, 然后结合月份与日期的信息, 可以得到关于气温的更多信息, 这类组合特征对目标是有着重要作用的, 然而由于在特征值数目较多的情况下, 尤其是生物信息领域一个目标可能有着上万个基因, 假设有一个如“日期”这样的基因, 那可能得组合特征就是 2^n, 显然是不可计算的, 所以往往这种组合特征在特征集规模较大的情况下, 无法被识别出来.

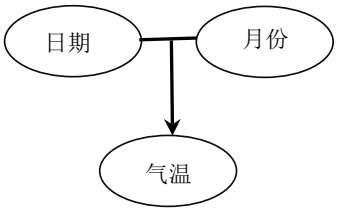

图 2 组合特征

文中, 主要针对第一类组合特征进行分析, 因为这类组合特征一方面包含着目标大量的相关信息, 另一方面, 从计算时间复杂度上来看, 这类特征也是能够用现有的设备进行计算的.

定义 1. 最大依赖性准则(Max-Dependence)

在传统的特征选择方法里, 可以根据最大依赖性准则去寻找与目标变量有着最大依赖性的变量集(一般使用的是基于互信息计算的方法).最大依赖性准则具体形式为(2):

$$\max_{S\subset X} D(y,S), D = I(y;S) \tag{2}$$

其中 D(y, S) 表示变量 y 与变量集 S 的依赖性, I(y, S)表示 y 与 S 之间的互信息大小. 若 m=1, 式(2)等价于最大化其之间的互信息$I(y;x_i)(1\leqslant i\leqslant n)$. 若 m>1, 则采取每次向目标特征候选集添加一个变量的递增搜索策略: 给出 k-1 个变量的变量集S_{k-1}, 则第 k 个变量 x_k 必须能够使得互信息$I(y;S_{k-1}\cup x_k)$达到最大化, 因而其有以下形式(3):

$$\begin{aligned} I(y;S_k) &= \iint p(y,S_k)\log\frac{p(y,S_k)}{p(S_k)p(y)}dydS_k \\ &= \iint p(y,x_k,S_{k-1})\log\frac{p(y,x_kS_{k-1})}{p(x_kS_{k-1})p(y)}dydx_kdS_{k-1} \\ &= \int\cdots\int p(y,x_1,\cdots,x_k)\log\frac{p(y,x_1,\cdots,x_k)}{p(x_1,\cdots,x_k)p(y)}dydx_1\cdots dx_k \end{aligned} \tag{3}$$

定义 2. 最大相关性和最小冗余度准则(Max-Relevance and Min-Redundancy criterion mRMR)

在特征选择方法里, 最大相关性和最小冗余度准则目的是寻找与目标变量同时满足最大相关性和最小冗余度的变量集合, 其采取式(4)的计算规则:

$$S = \arg\max_S[\sum_{x_i\in S} I(y;x_i) - \frac{1}{|S|}\sum_{x_i,x_j\in S} I(x_i;x_j)] \tag{4}$$

并且在实际上, 如果采取的是每次添加一个变量到变量集的选择策略, 则最大相关性和最小冗余度准则等价于最大依赖性准则.

1.2 算法的基本流程

如图 3 所示, 假设 Y 的一个有 5 个候选特征{X_1, X_2, X_3, X_4, X_5}为 Y 的马尔科夫链(一般地, 可以认为目标与其特征之间存在一定的因果关系). 其中 X_1 与 X_2 是 Y 的组合特征, X_5 是 Y 的配偶节点并不携带这关于 Y 的任何信息. 若按照传统的特征排序方法, 仅考虑单个特征之间与 Y 的依赖性关系, 则 X_1 与 X_2 往往会被剔除, 因为单独考虑 X_1 或 X_2 的时候, 其几乎不包含关于 Y 的任何信息, 一定程度上, 其对 Y 的作用往往比不上 X_5, 因为给定 X_4 的时候, X_5 可能携带关于 Y 的大量信息, 也就是条件互信息 $I(Y;X_5|X_4)>0$.

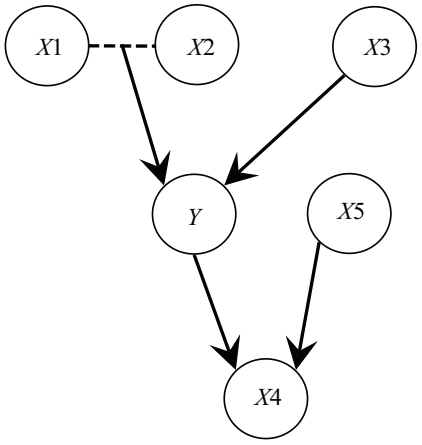

图 3 算法的基本框架

统计学中, 尤其是机器学习领域, 通常对离散数据进行独立性判断是采用卡方检验[10,11]等方法; 算法开始时候, 先令 Y 的特征节点集 D={}, C={}, T={}, S={}.

步骤 1. 应用独立性测试: 基于卡方检验, 测试 X 中 Y 的每一个候选特征$\{x_1, x_2, ..., x_n\}$和 Y 之间的独立性, 若独立性$Ind(y; x_i)$成立, 表明 x_i 没有携带任何关于 Y 的信息, 即 x_i 不可能 Y 的单一特征, 将 x_i 从 Y 中移除, 加入到 I.

步骤 2. 对 C 中所有的元素按照条件概率表进行组合, 构造新的节点, 并加入到 C. 如 I={X1, X2, X3}则 C={X1X2, X1X3, X2X3, X1X2X3}. 其中, X1X2 表示 X1 与 X2 的组合, 如 X1={0, 1}, X2={2, 3}则 X1X2={(0, 2), (0, 3), (1, 2), (1, 3)}.

步骤 3. 合并单一特征 X 和候选组合特征 D, X=CUD. 对任 x_i 属于 I, 测试 I(x_i;Y), 将最大值那个加入 S. 然后对于剩下的每一个特征, 按照以下规律进行选取.

$$1) x_k = \arg \max_{x_i \in X \setminus S_{k-1}} [I(y; x_i) - \frac{1}{k-1} \sum_{x_j \in S_{k-1}} I(x_i; x_j)]$$

$$2) S_k = S_{k-1} \cup x_k$$

步骤 4. 经过以上步骤, 得到一个包含组合特征的特征序列, 因为其是基于最大相关性与最小冗余度进行特征挑选的, 因而该序列任意前 k 个特征与 Y 之间的总体依赖性, 要大于或等于其余的任意 k 个特征与 Y 之间的总体依赖性.

为了方便表述, 上述提出的算法被记为 FOA, 其具体实现方式如下.

Input : n – dimensional vector $X = \{x_1, x_2, ..., x_n\}$,
target variable y,
threshold m.
Output: S_m.

$1 : S_1 = x_i, x_i = \arg\max_{x_i \in X} I(y; x_i)$.

2 : For each variable $x_i \in X$ do.

$$x_k = \arg \max_{x_i \in X \setminus S_{k-1}} [I(y; x_i) - \frac{1}{k-1} \sum_{x_j \in S_{k-1}} I(x_i; x_j)]$$

$3 : S_k = S_{k-1} \cup x_k$.

4 : While $k = m$, end.

针对算法基本流程, 举个简单的例子, 比如: UCI 数据集 Acute Inflammations Data Set, 有 6 个特征, 两个目标特征, 挑选其中的 acute nephritides 特征作为学习的目标特征, 首先基于卡方检验挑选出独立于目标特征的 X5, 将 X5 加入到组合特征候选集中, 然后基于 mRMR 准则进行选择, 得到最终的包含有组合特征候选集.

1.3 算法的完整性和时间复杂度分析

FOA 算法的完整性与正确性分析: FOA 算法是基于最大相关性与最小冗余度的一种特征排序算法, 本质上是对最大依赖性进行逼近, 由于 FOA 是增量式的挑选特征, 从理论上是等价于最大依赖性准则的, 使得任意前 k 个特征与 Y 之间的总体依赖性, 要大于或等于其余的任意 k 个特征与 Y 之间的总体依赖性. 显然, FOA 是满足完整性的.

FOA 算法的时间复杂度分析: 从以上的 FOA 算法步骤可以看出, 该算法运行时间主要消耗在特征排序上面, 也就是与 C 的大小密切相关. 挑选每一个特征加入 S 的时候, 算法计算 N 次互信息, 总共有 N 个特征, 因此算法复杂度为 O(N^2*T)其中 T 为互信息的计算复杂度. 可以见 FOA 是二阶多项式时间复杂度的算法, 即使在特征数量较多的情况下, 在一般得计算机上也能较快得出结果.

2 数值实验

数值实验在 Matlab2010b 中完成, 本文分别用虚拟网络数据和真实数据集对 FOA 进行评价. 而在真实数据集方面, UCI 机器学习数据集数据库[12]Acute Inflammations Data Set 进行测试. 为了进一步体现组合特征在特征选择的重要性, 对 FOA 与两个种经典的算法 mRMR 与 MIGS 进行对比.

2.1 虚拟网络实验

2.1.1 数据来源

在虚拟数据生成阶段, 首先生成图 3 所示的网络结构, 其中 Y 是目标变量, X_1, …, X_5 是 Y 的属性, 其中 X_1X_2 是 Y 的组合特征, 然后按照经典的因果网络数据生成方式生成数据(特别地, 为了更能体现算法的效率, 这里将数据中噪声的权值降低到 0), 样本量为 1000, 并使得 X_1, X_2 独立于 Y, X_1 的 X_2 的条件概率组合不独立于 Y.

2.1.2 数据分析

首先, 对虚拟数据集进行特征排序, 由于 FOA, mRMR, MIGS 都是基于(条件)互信息的选择(排序)方法, 在特征数目较少的情况下, 除开 FOA 的组合特征

外其余特征的排列顺序基本一样. 可以看到尽管 X_1, X_2 单个特征与 Y 之间的互信息要远小于 X_3 与 X_4, 但是组合特征 X_1X_2 却携带着关于 Y 的仅次于 X_4 的信息量(第一个特征选取的是与 Y 之间互信息最大的特征 X_4). X_5 是 Y 的关于 X_4 的配偶节点, 与 Y 是完全独立.由于每次生产数据是随机的, 这里的结果是最大可能出现的排序顺序.

表 1 关于三种特征选择算法的特征排序

算法	特征排序
FOA	$X_4, X_1X_2, X_3, X_1, X_2, X_5$
MIGS	X_4, X_3, X_2, X_5, X_1
mRMR	X_4, X_3, X_2, X_5, X_1

为了进一步体现 FOA 在支持组合特征方面的优异性, 分别提取 3 种算法的排序结果中前 3, 4, 5 个特征做分类, 其中以 MATLAB 工具包中的 CART 决策树算法[13,14](即“classregtree”), 200 样本为训练集, 余下的 800 样本为测试集.

2.1.3 结果讨论

结果如表 2. 可以看到采取了基于 FOA 找出的组合特征能够很大程度上提高分类的准确率, 从图 3 可以知道, 目标变量是由 $X_1\&X_2\&X_3$ 决定的(同时 X_4 也携带者关于 Y 的一定量的信息), 因此, 当挑选 FOA 前 3 个特征的时候, 分类准确率已经比 mRMR, MIGS 要高 6%. 当挑选 FOA 前 4 个的时候, 由于目标变量是由 $X_1\&X_2\&X_3$完全决定的, 添加多一个 X_1 对算法分类作用不大, 因此此时分类准确率并几乎没有变化; 另一方面, 可以看到后面两种算法挑选前 4 个特征的时候, 分类准确率要比之前提高 4% 左右, 但依然低于 FOA 的情况. 最后当 MIGS 与 mRMR 考虑到所有特征的时候, 分类准确率达到一个峰值, 基本与 FOA 持平. 这说明, 在特征数目较多的情况下, 组合特征的采用, 一定程度的降低了算法的时间复杂度, 并且提高了算法的准确率, 因为现有的大部分特征排序, 特征选择算法, 并有考虑到组合特征对目标变量的直接影响.

表 2 关于三种特征选择算法的分类准确率(%)

算法	分类准确率 (前3特征)	分类准确率 (前4特征)	分类准确率 (前5特征)
FOA	98.78	98.64	98.70
MIGS	92.55	96.43	98.15
mRMR	92.55	96.43	98.15

由于 FOA 算法考虑了组合特征, 因此可能要比一般得算法的时间复杂度要高, 下面根据上述图 3 的虚拟网络分别生成{100, 200, 500, 1000, 2000, 5000}样本量的数据, 对三种方法 FOA, MIGS, mRMR 进行对比. 从实验结果图 4 可以看到, MIGS 运行速度最快, 其次 mRMR, 最后是 FOA. 从他们的三条曲线可以看出来, 尽管 FOA 运行时间要比另外两者要久, 但 FOA 与他们之间的差也是一个常数数量级之差, 同时也可以看出来, 在这种差别一般的计算机可处理. 相对于一个常数数量级的运行时间增加, 一般来说更需要注重的是算法的准确率, 也就是是否算法能够找出一组有利于对数据进行分类或聚类等的特征. 因此, 在本组虚拟实验中, 可以看出 FOA 是有着不输于其他算法的特征选择能力. 下一组实验, 将采用真实数据对提出的算法进行评估.

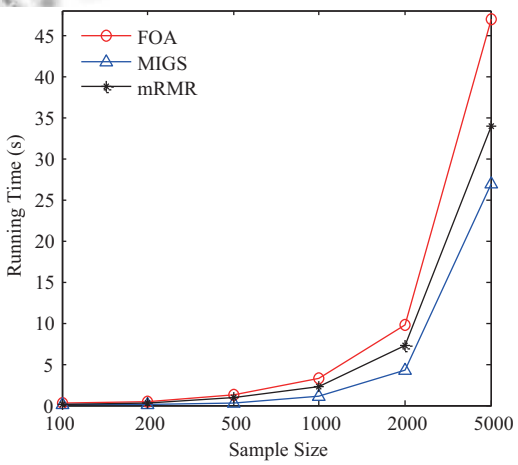

图 4 三种算法在不同样本量下的运行时间

2.2 真实数据实验

2.2.1 数据来源

在真实数据实验方面, 采用 UCI 机器学习数据库数据集 Acute Inflammations Data Set 进行实验. 其中包含着 120 个病人的数据, 特别的有两个目标节点, 指示病人是否患上:

Y_1: acute inflammations of urinary bladder 和/或

Y_2: acute nephritides,

也就是“是”或“非”两个值. 在特征方面, 一共有 6 个特征, 分别是:

X_1: Temperature of patient,

X_2: occurrence of nausea,

X_3: Lumbar pain,

X_4: Urine pushing,

X_5: Micturition pains,

X_6: Burning of urethra, itch, swelling of urethra

outlet.

同样的, 都取值于“是”或“非”.

2.2.2 数据分析

其中, 特征 X_6 与目标症状 Y_1 统计独立, 特征 X_5 与目标症状 Y_2 统计独立, 因此仅有一个特征与目标独立无法应用于第一类组合特征. 特别地, 因为在这组真实数据里, 仅有 6 个特征, 所以可以应用 FOA 寻找第二类组合特征. 由于有两个目标变量 Y_1 与 Y_2, 因此实验分为两组, 第一组实验对所有特征, 以及第二类组合特征针对 Y_1 进行排序, 第二组实验对所有特征, 以及第二类组合特征进行排序, 其中在生成第二类组合特征时候, 选取 X_5(针对 Y_1 时)或 X_6(针对 Y_2 时)与其他 5 个特征的任意 1 或 2 个进行组合, 关于特征 Y_1 的实验结果如下

从表 3 中可以看到找出的前 6 个与目标 Y_1 之间满足最大相关最小冗余度都包含着特征 X_6, 而后两者 mRMR 与 MIGS 都几乎把 X_6 放到了最后, 从本质上, 这与 FOA 的排序性能是有区别, 这可以从他们与 Y_1 之间的确定性关系可以看出.

表 3 在四种特征集下 SVR 算法的预测结果

算法	特征排序(关于特征Y_1)
FOA	$X_4X_5X_6$, $X_2X_5X_6$, $X_3X_4X_6$, $X_1X_4X_6$, $X_2X_4X_6$, $X_2X_3X_6$
MIGS	X_4, X_2, X_5, X_1, X_3, X_6
mRMR	X_4, X_5, X_3, X_1, X_6, X_2

为了进一步说明 FOA 算法在真实数据上是可靠的, 同样地, 使用 matlab 工具包中的 CART 决策树算法为基础算法(即“classregtree”)对 Y_1 进行分类. 由于数据集记录顺序对分类结果可能造成较大影响, 在实验准备阶段, 将数据集的记录排列顺序进行随机打乱, 取前 40 条记录为训练样本训练决策树模型, 后 80 条记录为测试样本, 特征分别选取 FOA 的第一个特征 $X_4X_5X_6$, MIGS 与 mRMR 的前三个特征进行比较(不失公平性), 实验重复 100 次, 准确率为期平均值.

2.2.3 结果讨论

结果如表 4. 可以看到, 选取第一个 FOA 组合特征算法已经要比后两者的准确率要高, 注意到尽管 FOA 的第一个特征 $X_4X_5X_6$ 是三个特征的组合, 但是这个组合在分类算法里面只需求一个特征的计算时间, 从而时间复杂度上要低于后两者的算法.

可以看到, 如表 5(a)所示, 组合特征 $X_4X_5X_6$ 与 Y_1 之间是确定性关系, 也就是说目标 Y_1 被这个组合特征完全确定, 也就是应用决策树等分类方法, 在这个较为简单的数据集中, 容易得到 100% 准确的分类方法; 而从表 5(b)可以看到, MIGS 找出的前三个特征 $X_4\&X_5\&X_3$ 与 Y_1 并非确定性关系, $X_4\&X_5\&X_3$ 数值为{1, 0, 0}的时候无法确定 Y_1 的取值, 因而在分类的时候, 很可能分类效果比用 FOA 得到的前三个特征做分类的时候效果要差.

表 4 在四种特征集下 SVR 算法的预测结果(%)

算法	分类准确率
FOA	95.21
MIGS	89.91
mRMR	91.72

表 5 前三个特征与 Y1 的关系(a)FOA 与(b)MIGS

X_4	X_5	X_6	Y_1
0	0	0	0
0	1	0	0
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

(a) FOA

X_4	X_5	X_3	Y_1
0	0	0	0
0	1	1	0
1	0	0	0
1	0	0	1
1	0	1	1
1	1	1	1

(b)MIGS

类似地, 关于特征 Y_2 的实验结果如表 6, 可以看到 FOA 找出的前 6 个特征均为组合特征, 且都包含 X_5. 可见尽管单一的 X_5 包含着关于 Y_2 的信息量很少, 但对预测 Y_2 起着很大作用. 其余的一些相关分析结果类似于上一实验, 此处不做太多讨论. 事实上, 由于 FOA 考虑到组合特征, 因此要比一般得算法运行速度要慢, 但由于考虑到组合特征, 因此分类准确率往往要比一般得没有考虑组合特征的算法要高. 本组实验, 验证了 FOA 算法在真实数据上的应用也是有效的, 并在组合特征存在的情况下, 一定程度上要优于现有的没有考虑到组合特征的算法.

4 结束语

尽管对候选特征进行维数约减在机器学习领域如分类, 聚类问题中占有很重要的地位, 但现有的方法大

多数是基于单一特征对目标 T 的依赖性而几乎没有考虑到组合特征对目标 T 的影响, 如属性 A 和 B 都仅包含 Y 中的极少量信息, 甚至与 Y 完全独立, 但 A 和 B 组合起来却能提供关于 Y 的大量信息, 甚至完全决定 Y. 基于此, 提出一种能够从特征集合中挖掘到组合特征与单一特征的特征选择算法, 该算法首先对不显著特征进行组合并按照条件概率分布表生成新的候选特征; 然后, 利用基于最大相关性, 最小冗余度的准则进行对单一特征和组合特征进行选择. 虚拟和真实数据集上的实验表明, 尽管考虑到组合特征的特征选择算法要比现存的很多算法在运行时间上要慢, 但该特征选择算法能够挖掘数据集的组合特征, 一定程度上提高相应的机器学习算法的准确率.

表 6 在四种特征集下 SVR 算法的预测结果

算法	特征排序(关于特征Y_2)
FOA	$X_1X_2X_5$, $X_2X_3X_5$, $X_1X_6X_5$, $X_1X_3X_5$, $X_3X_6X_5$, X_1
MIGS	X_1, X_3, X_2, X_4, X_5, X_6
mRMR	X_1, X_3, X_2, X_6, X_4, X_5

参考文献

1 Cao LJ, Chua KS, Chong WK, *et al*. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing, 2003, 55(1-2): 321–336. [doi: 10.1016/S0925-2312(03)00433-8]

2 Peng HC, Long FH, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226–1238. [doi: 10.1109/TPAMI.2005.159]

3 Maji P, Garai P. On fuzzy-rough attribute selection: Criteria of max-dependency, max-relevance, min-redundancy, and max-significance. Applied Soft Computing, 2013, 13(9): 3968–3980. [doi: 10.1016/j.asoc.2012.09.006]

4 Cai R C, Hao ZF, Yang XW, *et al*. An efficient gene selection algorithm based on mutual information. Neurocomputing, 2009, 72(4-6): 991–999. [doi: 10.1016/j.neucom.2008.04.005]

5 Maji P, Paul S. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. International Journal of Approximate Reasoning, 2011, 52(3): 408–426. [doi: 10.1016/j.ijar.2010.09.006]

6 Aliferis CF, Statnikov A, Tsamardinos I, *et al*. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. The Journal of Machine Learning Research, 2010, (11): 171–234.

7 陈一明. 一种基于因果网络的支持向量回归特征选择算法. 湖南师范大学自然科学学报, 2015, 38(4): 90–94.

8 Li XJ, Mishra SK, Wu M, *et al*. Syn-lethality: An Integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies. BioMed Research International, 2014, (2014): 196034.

9 Wu M, Li XJ, Zhang F, *et al*. In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. Cancer Informatics, 2014, 13(S3): 71–80.

10 Peters J, Janzing D, SCHOLKOPF B. Causal inference on discrete data using additive noise models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2436–2450. [doi: 10.1109/TPAMI.2011.71]

11 Chen WQ, Hao ZF, Cai RC, *et al*. Multiple-cause discovery combined with structure learning for high-dimensional discrete data and application to stock prediction. Soft Computing, 2016, 20(11): 4575–4588. [doi: 10.1007/s00500-015-1764-8]

12 UCI 机器学习数据集数据库. http://archive.ics.uci.edu/ml/datasets.html.

13 张亮, 宁芊. CART 决策树的两种改进及应用. 计算机工程与设计, 2015, 36(5): 1209–1213.

14 罗来平, 宫辉力, 刘先林. 基于决策树算法的遥感图像分类研究与实现. 计算机应用研究, 2007, 24(1): 207–209.