基于互信息和遗传算法的两阶段特征选择方法*

裘国永†，王　娜，汪万紫
（陕西师范大学 计算机科学学院，西安 710062）

摘　要：为了在特征选择过程中得到较优的特征子集，结合标准化互信息和遗传算法提出了一种新的两阶段特征选择方法。该方法首先采用标准化的互信息对特征进行排序，然后用排序在前的特征初始化第二阶段遗传算法的部分种群，使得遗传算法的初始种群中含有较好的搜索起点，从而遗传算法只需较少的进化代数就可搜寻到较优的特征子集。实验显示，所提出的特征选择方法在特征约简和分类等方面具有较好的效果。

关键词：标准化互信息；遗传算法；特征选择；特征约简
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2012)08-2903-03
doi:10.3969/j.issn.1001-3695.2012.08.026

Two-stage feature selection algorithm based on mutual information and genetic algorithm

QIU Guo-yong†, WANG Na, WANG Wan-zi
(*School of Computer Science, Shaanxi Normal University, Xi'an* 710062, *China*)

Abstract: To get better feature subset in the feature selection process, this paper proposed a new two-stage feature selection algorithm based on normalized mutual information and genetic algorithm. First it ranked features by normalized mutual information. Then to provide the genetic algorithm with better starting point it used the front ranking features to initialize the population, thus got better feature subset after only a few evolution times. The test results on benchmark datasets show the effectiveness of the algorithm, in terms of dimensionality reduction and classification performance.

Key words: normalized mutual information; genetic algorithm; feature selection; dimensionality reduction

特征选择在模式识别、数据挖掘以及更多的机器学习任务中非常关键，特别是处理高维数据时更是如此。它根据给定准则，在删除一组特征中不相关的、冗余的特征后，从中挑选出对分类最有效的那些特征以降低特征空间维数[1,2]。

特征选择方法大致可分为过滤方法(filter)和封装方法(wrapper)两类。过滤方法主要是根据训练集中数据的特性提取出一个特征子集，它不依赖任何学习算法，运行效率较高，因此适用于大规模数据集[3,4]；而封装方法则要引入某种学习算法(分类器)的效果评价待选特征子集的优劣，因此准确率较高。但由于要训练一个分类器，计算复杂度很高，不适合大规模数据[4]。文献[5,6]将过滤和封装方法组合起来进行特征选择，这种组合方法能够使封装方法充分利用过滤方法得到的结果，加快封装算法的收敛，并得到能产生更高分类性能的特征子集。

互信息方法是一种基于相关性的过滤特征选择方法，它借助互信息度量特征与类别标签的相关性选择特征。其特点是速度快，但得到的特征子集可能不是最优的，分类精度较低。而遗传算法作为一种自适应的全局搜索方法，具有并行性和适合于解决多目标优化问题等特性。为了利用遗传算法的优点，同时改正基于遗传算法的封装方法对于高维数据运行效率低的缺点，本文将过滤方法和封装方法的思想有机地结合起来，提出一种基于标准化互信息(SU)和遗传算法(GA)的两阶段特征选择算法(SU-GA)。

由于互信息每次只选一个与目标变量相关性大的特征，因此它是一种增量算法。已有实验证明，这样选出来的特征不一定能构成最优的特征子集[7,8]。若将过滤方法排序在前的特征作为封装式特征选择方法的出发点，这样有些虽然与目标类的相关性小，但可能是其他特征的补充，与其他特征组合起来对分类有较大贡献的特征，采用过滤模式进行过滤很可能将被滤除，即使后面再采用封装模式进行特征选择也无法被找回。因此，本文的两阶段特征选择算法(SU-GA)包括一个初始化算法。该初始化算法既考虑了单个特征与目标变量的相关性，也考虑了多个特征的组合与目标变量的相关性，克服了不能发现多个特征的组合与目标变量相关性的增量搜索算法存在的局限性。

1　互信息和遗传算法

1.1　熵和互信息

熵和互信息是度量随机变量信息量的重要方法。熵是对随机变量不确定性的一种度量。假设 X 是一个离散随机变量，其取值范围是集合 S，$p(x)$ 是 X 的概率密度分布函数，那么 X 的熵 $H(X)$ 定义为

$$H(X) = -\sum_{x\in S} p(x)\ \log_2 p(x) \tag{1}$$

已知变量 Y 后，X 的不确定性用条件熵来衡量，如式(2)所示，其中 $p(x|y)$ 表示 Y 取值为 y 时，X 取值为 x 的概率，Y 取

收稿日期：2011-12-27；**修回日期**：2012-02-13　　**基金项目**：陕西省自然科学基金资助项目(2010JM8039)

作者简介：裘国永(1964-)，男(通信作者)，浙江嵊州人，副教授，博士，主要研究方向为智能信息处理、数据挖掘(qgyqgy@snnu.edu.cn)；王娜(1983-)，女，新疆阿克苏人，硕士，主要研究方向为数据挖掘；汪万紫(1986-)，男，浙江衢州人，硕士，主要研究方向为数据挖掘.

值范围是集合 T:

$$H(X|Y) = -\sum_{y\in T} p(y) \sum_{x\in S} p(x|y)\ \log_2 p(x|y) \tag{2}$$

互信息概念是从熵的概念引申出来的。设有离散随机变量 X、Y,$p(x,y)$ 表示它们的联合概率密度,则它们的互信息量 $I(X;Y)$ 定义为

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in S} \sum_{y\in T} p(x,y)\ \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

互信息量刻画的是两个随机变量之间共有的信息量,这个值越大,说明两个变量之间的相关程度越高。如果两个变量的互信息量为 0,则说明两个变量是完全不相关的。因为互信息对具有较多值的变量有偏倚,用它们对应的熵对其进行标准化,得到互信息的标准化形式,即 symmetrical uncertainty (SU)[9],计算公式为

$$\mathrm{SU}(X,Y) = 2\left[\frac{I(X;Y)}{H(X)+H(Y)}\right] \tag{4}$$

标准化后的互信息纠正了互信息对有较多值的变量的偏倚,并且把值的范围限制在区间[0,1]。值为 1 表示已知两个变量中的任何一个变量的值可以完全预测另一个变量的值;值为 0 表示 X 和 Y 是不相关的。

本文提出的 SU-GA 算法中,将先使用 SU 计算每个特征与目标变量的相关性,然后根据得到的相关性值的大小对特征进行降序排序,并将这个排序结果用于后续遗传算法的种群初始化。算法如下:

算法 1 基于互信息的特征排序

输入:训练数据集 D,原始特征集合 $F = \{f_1, f_2, \cdots, f_L\}$。

输出:降序排序的特征集 F'。

a)将 F'初始化为空集;

b)for $i = 1$ to L

根据式(1)~(4)计算每个特征 f_i 与类别标签的 SU_i 值;

c)根据特征与类别标签的 SU_i 值对特征进行降序排序,将排好序的特征放到 F'中。

1.2 互信息引导的遗传算法

遗传算法是一种随机搜索策略,通过选择、遗传和变异等进化操作,实现各个个体适应性的提高,进而解决各种复杂问题[10]。由于遗传算法不是对单一特征项进行评价,而是对一个可能的特征子集进行优劣评价,考虑了多个特征的组合与目标变量的相关性,所以保证了所选特征子集的组合最优化,也省去了考虑各特征之间相关性的工作。

1)特征表示

编码问题的关键就是要使编码能够代表所给特征集的所有可能子集的解空间。本文采用经典的二进制编码,该编码简单且非常有效[11]。遗传算法的个体是长度为 L 的二进制串(L 为候选特征数目),第 i 位为 0 表示不选择特征 i,为 1 表示选择特征 i。这样遗传算法的每个个体表示一种特征选择方案。

2)初始种群

为了加快遗传算法的收敛速度,同时为了弥补互信息只评价单个特征与目标变量相关性的不足,初始种群由两部分组成,即标准化互信息排序在前的特征以及为了弥补可能已被过滤掉的有用特征而随机产生的特征。由于遗传算法是一种不确定搜索算法,其性能受初始种群影响很大。互信息评估相当于为遗传算法提供了问题的先验知识,使人们能够得到一个较好的初始种群。

设初始种群的大小为 P,L 为个体的长度;C、I 为用户自定义的参数,$1 \leqslant C \leqslant L$, $1 \leqslant I \leqslant P$。其中,$C$ 表示取出的互信息排序在前的特征数;I 表示初始种群中由 C 确定的个体数,即对初始种群中 I 个个体,把第一阶段互信息排序在前的 C 个特征在这些个体中对应的基因置为 1,其余的基因随机置为 0 或 1。初始种群剩下的 $P-I$ 个个体随机初始化。

算法 2 遗传算法种群初始化

输入:训练数据集 D,用户自定义的参数 C, I。

输出:遗传算法初始种群。

a)调用算法 1;

b)取 F'中排好序的前 C 个特征放入集合 S_0;

c)用集合 S_0 中的特征初始化种群中 I 个个体;

for all $f_i \in S_0$,将初始种群中 I 个个体中特征 f_i 对应的第 i 位设置为 1;

else 随机设置个体第 i 位为 0 或 1;

d)随机初始化剩下初始种群的 $P-I$ 个个体。

3)适应度函数

$$f(X_i) = \mathrm{accuracy}(X_i) - \lambda(\mathrm{nfeatures}(X_i)/L) \tag{5}$$

其中:$\mathrm{accuracy}(X_i)$ 是用个体 X_i 对应的特征子集进行分类的准确率;$\mathrm{nfeatures}(X_i)$ 表示被选择的特征子集的特征个数,即个体 X_i 中基因是 1 的个数;λ 是一个用户自定义的参数,用来平衡适应度函数的前后两项,即分类准确率和特征数目对适应度函数的贡献。

4)遗传算子

选择算法则采用轮盘赌的比例选择法;交叉操作采用单点交叉,交叉点位置随机确定。为了加快遗传算法的收敛速度,变异操作时,首先验证变异后的个体的适应度是否比原个体的适应度高,如果高,则保留变异个体;否则该变异操作不进行,保留原个体。终止条件为达到最大迭代次数或连续五次最优解保持不变[10]。

2 两阶段特征选择方法过程

综上所述,为了提高遗传算法的计算效率并得到分类精度更高的特征子集,本文将结合标准化互信息和遗传算法提出一种两阶段的特征选择方法:先基于标准互信息对特征进行评估,然后将评估结果用于遗传算法的种群初始化。算法的过程如下:

a)使用标准化的互信息计算特征与目标变量的相关性,然后按照其值对特征降序排序。目的是得到与类相关性大的特征,使得遗传算法的初始种群中含有较好的初始点。

b)运行遗传算法。其中遗传算法的初始种群由两部分构成:第一阶段排序在前的特征初始化种群中的部分个体,其余个体随机产生。遗传算法使用上一章介绍的适应度函数和遗传算子,遗传算法迭代结束时,将适应度最高的个体所对应的特征集作为特征选择的结果。

算法 3 两阶段特征选择算法

输入:训练数据集 D, 最大迭代代数 $\max I$,用户定义参数 λ。

输出:优化的特征子集。

a)调用算法 2,得到遗传算法的初始种群;

b)根据编码方案,将种群中的个体编码成二进制串;

c)根据上一章给出的适应度评价方法,即式(5),计算个体适应度;

d)判断是否达到最大迭代次数 $\max I$ 或连续五次保持最优解不变,若是,则输出当前的最优特征子集,否则执行以下步骤;

e)根据个体适应度执行选择操作;

f)执行交叉操作;

g)执行变异操作;

h)返回 d)。

3　实验和结果分析

3.1　实验数据

实验数据选自 UCI 机器学习数据库[12],从中选取七个数据集进行测试。表 1 列出了这七个基准数据集的名称(dataset)及其相关信息:特征的数目(features)、数据集的大小(instances)、类别个数(classes)。

表 1　实验数据集

dataset	features	instances	classes
Dermatology	33	366	6
Lung cancer	56	32	2
Breast cancer	30	569	2
Soybean-large	35	307	19
Ionosphere	34	351	2
Anneal	38	898	6
Sonar	60	208	2

3.2　实验方法

为了验证所提出的两阶段特征选择方法的有效性,本文选取较具代表性的过滤方法标准化互信息算法和封装方法遗传算法,与 SU-GA 算法作比较。实验时,SU 算法的阈值设为 0.15,即标准化互信息小于该值的特征将被去除。除了没有基于标准互信息的预处理阶段,遗传算法与 SU-GA 算法采用相同的参数设置,只不过 GA 直接运行于未经过基于标准互信息的预处理阶段的数据特征空间,选取适应度最高的个体作为特征选择的结果。

实验中采用 Naive Bayes(NB)为分类器。对于表 1 中的每一个数据集,分别运行 SU、GA、SU-GA 算法,记录下各算法得到的特征子集大小(表 2);然后通过十折交叉验证法,将这些特征子集用在 NB 中对七个数据集进行分类,计算出相应的分类准确率(表 3)。在实验中,遗传算法种群大小 $P=20$,最大迭代次数 $\max I=20$,交叉和变异概率分别为 0.6 和 0.033,$\lambda=0.1$,C 取每个数据集特征数的 30%,I 取种群大小的 30%。

表 2　三种特征选择算法得到的特征子集大小

dataset	all features	SU	GA	SU-GA
Dermatology	33	24	10	9
Lung cancer	56	12	25	4
Breast cancer	30	18	8	3
Soybean-large	35	21	17	13
Ionosphere	34	32	9	10
Anneal	38	21	8	5
Sonar	60	14	16	6
Average	40.86	20.29	13.29	7.14

从表 2、3 可以看出,SU、GA 和 SU-GA 三种算法在七个数据集上的平均准确率均高于使用全部特征集的平均准确率,说明特征选择确是提高分类器分类性能的有效方法。相比 SU、GA 算法,由于 SU-GA 算法利用互信息的特征排序结果构建了遗传算法的初始种群的部分个体,使得 GA 能有一个更好的搜索起点,从而得到了更优的特征子集,保证了 NB 分类器的分类效率。此外,SU-GA 算法在七个数据集上的平均准确率高于 SU 和 GA,且仅在 Soybean-large 数据集上比 GA 低,在其余数据集上均不低于 SU 和 GA。因此,SU-GA 特征选择算法在保证分类精度前提下,较大幅度降低了数据集的特征数目,达到了较好的特征约简效果。

表 3　NB 分类器在各特征子集上分类精度对比　/%

dataset	all features	SU	GA	SU-GA
Dermatology	97.26	97.27	98.91	98.91
Lung cancer	84.38	87.5	90.63	90.63
Breast cancer	92.97	92.97	95.96	96.84
Soybean-large	92.18	92.18	93.81	92.84
Ionosphere	82.62	82.91	92.02	92.59
Anneal	98.44	98.66	98.78	98.78
Sonar	71.15	77.40	76.44	82.69
Average	88.43	89.84	92.36	93.33

4　结束语

本文提出了一种基于标准化互信息和遗传算法的两阶段特征选择方法。标准化互信息(过滤类)在对特征进行初步筛选后,去除不相关的特征,从而为遗传算法搜索提供了更好的搜索起点,加快了遗传算法(封装类)的搜索速度,最终提高了遗传算法的运行效率。这样的算法综合了过滤方法和封装方法的优点,实验结果显示了该算法的有效性,既能保证分类精度,又较大幅度地降低了数据集的特征数目。当然算法中涉及的参数具体选择标准还有待进一步研究。

参考文献:

[1] LIU Huan, YU Lei. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(3): 491-502.

[2] 张晓光,孙正,徐桂云,等. 一种类内方差与相关度结合的特征选择算法[J]. 哈尔滨工业大学学报, 2011, 43(2): 133-136.

[3] ZHANG Dao-qiang, CHEN Song-can, ZHOU Zhi-hua. Constraint score: a new filter method for feature selection with pair-wise constraints[J]. Pattern Recognition, 2008, 41(5): 1440-1451.

[4] 蒋盛益, 王连喜. 基于特征相关性的特征选择[J]. 计算机工程与应用, 2010, 46(20): 153-156.

[5] Van DIJCK G, Van HULLE M M. Speeding up the wrapper feature subset selection in regression by mutual information relevance and redundancy analysis [C]//Lecture Notes in Computer Science, vol4131. Berlin: Springer-Verlag, 2006: 31-40.

[6] YANG Cheng-huei, CHUANG Li-yeh, YANG Cheng-hong. IG-GA: a hybrid filter/wrapper method for feature selection of microarray data [J]. Journal of Medical and Biological Engineering, 2009, 30(1): 23-28.

[7] COVER T M. The best two independent measurements are not the two best[J]. IEEE Trans on Systems, Man, and Cybernetics, 1974, 4(1): 116-117.

[8] HSU H H, HSIEH C W, LU Ming-da. Hybrid feature selection by combining filters and wrappers[J]. Expert Systems with Applications, 2011, 38(7): 8144-8150.

[9] YU Lei, LIU Huan. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004, 5(12): 1205-1224.

[10] 王小平, 曹立明. 遗传算法理论应用与软件实现[M]. 西安: 西安交通大学出版社, 2002.

[11] VAFAIE H, De JONG K. Genetic algorithms as a tool for feature selection in machine learning[C]//Proc of the 4th IEEE International Conference on Tools with AI. Washington DC: IEEE Computer Society, 1992: 200-203.

[12] WITTEN I H, FRANK E. Data mining: practical machine learning tools and techniques[M]. 2nd ed. San Francisco: Morgan Kaufmann Publisher, 2005.