文章编号:1001 - 9081(2006)06 - 1403 - 03

基于相关性分析及遗传算法的高维数据特征选择

任江涛,黄焕宇,孙婧昊,印 鉴
(中山大学 计算机科学系,广东 广州 510275)
(issrjt@ mail. sysu. edu. cn)

摘 要:特征选择是模式识别及数据挖掘等领域的重要问题之一。针对高维数据对象,特征选择一方面可以提高分类精度和效率,另一方面可以找出富含信息的特征子集。针对此问题,提出了一种综合了 filter 模型及 wrapper 模型的特征选择方法,首先基于特征与类别标签的相关性分析进行特征筛选,只保留与类别标签具有较强相关性的特征,然后针对经过筛选而精简的特征子集采用遗传算法进行随机搜索,并采用感知器模型的分类错误率作为评价指标。实验结果表明,该算法可有效地找出具有较好的线性可分离性的特征子集,从而实现降维并提高分类精度。

关键词:特征选择;相关性;遗传算法
中图分类号: TP311.13;TP18 **文献标识码**:A

High-dimensional data feature selection based on relevance analysis and GA

REN Jiang-tao, HUANG Huan-yu, SUN Jing-hao, YIN Jian
(*Department of Computer Science, Zhongshan University, Guangzhou Guangdong* 510275, *China*)

Abstract: Feature selection is one of the important problems in the pattern recognition and data mining areas. For high-dimensional data feature selection not only can improve the accuracy and efficiency of classification, but also can discover informative feature subset. The new feature selection method combining filter and wrapper models was proposed, which first filters featured by feature relevance analysis, and realized the near optimal feature subset search on the compact feature subset by genetic algorithm; and the feature subset was evaluated by the classification inaccuracy of the perceptron model. The experiments show that the proposed algorithm can find the feature subsets with good linear separability, which results in the low-dimensional data and the good classification accuracy.

Key words: feature selection; relevance; Genetic Algorithm(GA)

0 引言

特征选择是模式识别与数据挖掘领域的重要数据处理方法之一。随着模式识别与数据挖掘研究的深入,研究对象越来越复杂,对象的特征维数越来越高。大量高维数据对象的特征空间中含有许多冗余特征甚至噪声特征,这些特征一方面可能降低分类或聚类的精度,另一方面会大大增加学习及训练的时间及空间复杂度。因此,在面对高维数据进行分类或聚类时,通常需要运用特征选择算法找到具有较好可分性的特征子空间,从而实现降维,降低机器学习的时间及空间复杂度[1,2,8]。

根据是否依赖机器学习算法,特征选择算法可以分为两大类,一类为 wrapper 型算法,另一类为 filter 型算法。Filter 型特征选择算法独立于机器学习算法,具有计算代价小,效率高但降维效果一般等特点;而 wrapper 型特征选择算法则需要依赖某种或多种机器学习算法,具有计算代价大,效率低但降维效果好等特点[1,2]。

从优化的观点来看,特征选择问题实际上是一个组合优化问题。通常解决该问题有遍历搜索、随机搜索及启发式搜索等方法。遗传算法在组合优化问题中也有着广泛的应用,属于一种随机搜索方法。近年来,随着对特征选择方法研究的深入,基于遗传算法的特征选择问题也得到了许多研究及应用[4]。目前基于遗传算法的特征选择方法通常基于分类器进行特征子集的评估,依据分类精度给出个体的评价指标及适应度。

但是,原始特征集合中含有许多与分类不相关或弱相关的特征,若直接针对原始特征集合采用遗传算法进行特征选择,可能会收敛到分类性能较差的局部最小点(即分类性能较差的特征子集),另外也会降低搜索的效率。因此,本研究融合了特征选择算法的 filter 模型及 wrapper 模型,提出了一种基于相关性分析及遗传算法的两阶段特征选择方法。首先基于信息增益进行特征相关性评价及筛选,然后针对经过筛选而精简的特征子集采用遗传算法进行随机搜索,并采用感知器模型的分类错误率作为评价指标。另外,在遗传算法编码方面没有采用传统的二进制直接编码方案,而是采用基于区间的二进制编码方案,一方面减小了编码长度、提高了时空效率,另一方面可对选择的特征个数进行灵活控制。

收稿日期:2005 - 12 - 09;**修订日期**:2006 - 02 - 15
基金项目:国家自然科学基金资助项目(60573097);广东省自然科学基金资助项目(04300462,05200302)
作者简介:任江涛(1975 -),男,广西柳州人,讲师,博士,主要研究方向:数据挖掘与知识发现、生物信息学; 黄焕宇(1980 -),男,广东湛江人,硕士研究生,主要研究方向:信息处理和数据挖掘; 孙婧昊(1983 -),女,河南许昌人,硕士研究生,主要研究方向:高维数据挖掘及特征选择、信息处理和电子商务; 印鉴(1968 -),男,湖北武汉人,教授,博士,主要研究方向:人工智能、数据挖掘与数据仓库.

1 基于相关性分析的特征过滤

基于相关性分析的特征过滤是进行特征选择及降维的有效方法之一,其主要思想是基于特定的相关性定义,逐个度量单个特征与类别标签的相关性,即单个特征各自的分类能力,然后根据各特征的分类能力对特征进行降序排序,选出分类能力高的特征子集,从而在一定程度上消除与分类弱相关甚至无关的特征,实现降维。在本研究中,采用在某些决策树算法中广泛采用的信息增益[6]作为特征与类别标签的相关性度量(即分类能力度量),然后根据该度量的降序排序选出分类能力强的一组特征,实现特征子集的精简,下面首先给出信息增益的定义[5,6]。

令 X 为随机变量,则 X 的信息熵定义为:

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \tag{1}$$

通过观测随机变量 Y 随机变量 X 的信息熵变为:

$$H(X \mid Y) = -\sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j)) \tag{2}$$

其中 $P(x_i)$ 代表随机变量 X 的先验概率,$P(x_i \mid y_j)$ 代表观测到随机变量 Y 后随机变量 X 的后验概率。引入随机变量 Y 的信息后,随机变量 X 的信息熵 $H(X \mid Y) \leqslant H(X)$,即引入 Y 后,X 的不确定程度会变小或保持不变。若 Y 与 X 不相关,则 $H(X \mid Y) = H(X)$;若 Y 与 X 相关,则 $H(X \mid Y) < H(X)$,而差值 $H(X) - H(X \mid Y)$ 越大,Y 与 X 的相关性越强。因此如公式(3)定义信息增益 $IG(X \mid Y)$ 为 $H(X)$ 与 $H(X \mid Y)$ 的差值,反映了 Y 与 X 的相关程度,$IG(X \mid Y)$ 越大,则变量 Y 与 X 的相关性越强。

$$IG(X \mid Y) = H(X) - H(X \mid Y) \tag{3}$$

另外,为了对信息增益进行归一化,可采用公式(4):

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

在上述相关性度量定义的基础上,就可以基于特征 f_i 与类别标签 L 之间的相关性进行特征排序及筛选,选出相关性最强的若干个特征形成精简后的特征子集,具体算法流程如算法 1 所示。

算法 1 *FeaturesFiltering*(D,F,m)

输入:带类别标签的数据集 D,原始特征集合 F,保留的特征数 m

输出:精简后的特征子集

步骤:

1)根据公式(1)~(4)计算每个特征 f_i 与类别标签的信息增益 SU_i;

2)根据 SU_i 对特征进行降序排序;

3)选出 SU_i 值最高的前 m 个特征形成精简特征集合。

2 基于遗传算法的特征选择

上文算法 1 通过消除不相关特征实现了原始特征集合的精简后,可以采用基于遗传算法的 Wrapper 型特征选择方法。下面从编码方案、适应度函数及算法流程等方面对该算法进行描述。

2.1 编码方案

编码问题的关键在于能代表所给特征集合的所有可能子集的解空间。常用的方法是采用直接二进制编码,即每一个二进制位对应特征集合中的一个特征,该位为 1 则表示对应的特征入选特征子集,而该位为 0 则表示对应的特征不在选出的特征子集中。在特征维数 d 相对较低时,该表示方法可得到较小的二进制串,提高计算效率。但在特征维数 d 特别高的情况下,该表示方法反而可能导致较长的串,从而降低了计算效率。例如,基因表达数据集 Colon Tumor 的维数为 2 000,采用直接二进制的编码方法就需要长度为 2 000 的二进制串。另外,直接的二进制表示方法不利于对选择出的特征个数进行限制。因此本研究采用基于区间的二进制编码方案。即用一个长度为 l 的二进制数表示所选择的特征在原特征集合中的序号。这样,如果指定要选择的特征个数 j,则这个二进制串长度为 $j*l$。当 $j \ll d$ 时,可得到较小的二进制串。例如,针对 2 000 个特征,每个特征编号需要一个 11 位的二进制数串来表示,即 $l = 11$,假设每次搜索 6 个特征的组合($j = 6$),那么整个编码二进制串的长度为 $j*l = 6*11 = 66$,远小于直接二进制编码的长度 2 000,提高了空间及时间效率。同时,该编码方案可保证每次选择的特征个数可指定,从而实现了对特征子集大小的灵活控制。

2.2 适应度定义

在大多数基于遗传算法的 Wrapper 型特征选择方法中,采用某些分类器模型对所选择的特征集合进行评价,并利用得到的分类精度或分类错误率作为适应度函数。在本研究中,为搜索出线性可分性较好的特征子空间,采用感知器模型作为分类器模型,并采用分类错误率作为适应度,评价算法 Evaluation 的流程由算法 2 给出。

算法 2 *Evaluation*(D,F_s)

输入:带类别标签的数据集 D,特征子集 F_s

输出:特征子集评价值

步骤:

1)根据特征子集 F_s 从数据集 D 中选出一个降维后的数据集 D_F;

2)采用感知器算法对数据集 D_F 进行分类,统计分类错误率 err;

3)输出分类错误率 err 作为特征子集的评价指标,即适应度,算法结束。

3 算法流程

基于标准遗传算法框架,得到一种新的基于相关性分析及遗传算法的特征选择方法(Feature Selection based on Relevance Analysis and GA,FSRAGA),算法具体描述如下。

算法 3 *FSRAGA*(D,F,m,fn,$MaxI$)

输入:带类别标签的数据集 D,原始特征集合 F,保留的特征数 m,选择的特征数 fn,最大迭代次数 $MaxI$

输出:优化的特征子集

步骤:

1)调用算法 *FeaturesFiltering*(D,F,m),进行特征的过滤,形成精简特征子集 F_c;

2)根据上文给出的编码方案,以及选择的特征数 fn,随机产生一组初始个体构成初始种群;

3)根据编码方案,将个体的二进制表达转化为精简特征子集 F_c 中的特征编号,根据这些特征编号进行特征选择,形成特征子集 F_s;

4)根据适应度评价方法,调用函数 *Evaluation*(D,F_s),计算个体适应度;

5)判断是否达到最大迭代次数 $MaxI$,若达到则输出当

前的最优特征子集,否则执行以下步骤;

6) 根据适应度执行选择操作;

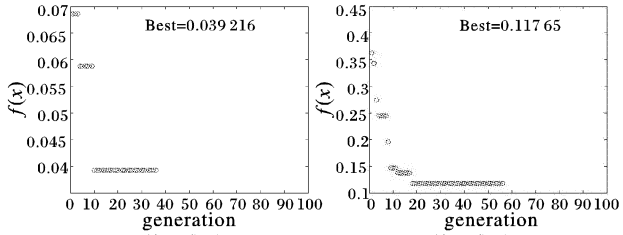
7) 执行交叉操作;

8) 执行变异操作;

9) 返回步骤 3)。

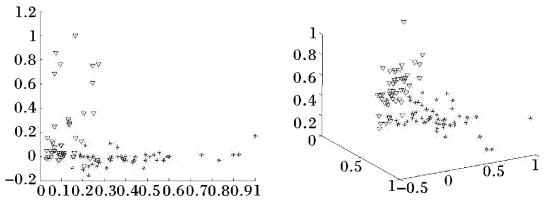
4 实验研究

为了评估上述 算法的有效性,采用基因表达数据集采用了基因表达数据集 Prostate Cancer(前列腺癌)进行测试。Prostate Cancer 数据集有 102 个样本,每个样本均含有 12 600 个基因的表达数据,其中 52 个样本被确诊患有前列腺癌,另外 50 个样本为未患前列腺癌的正常组织样本。



(a)特征集为512　　(b)特征集为12600

图 1　FSRAGA 算法针对 Prostate Cancer 数据集的迭代运行结果



(a)二维特征子空间散点图　　(b)三维特征子空间散点图

图 2　Prostate Cancer 数据集在 FSRAGA 算法选出的散点图

实验首先运用 算法对特征集合进行过滤,从原始的 12 600 个特征中选出类相关性最强的前 512 个特征,然后针对这 512 个特征应用遗传算法进行选择。图 1(a)给出采用 FSRAGA 算法对 Prostate Cancer 数据集进行特征选择的遗传算法迭代运行结果,图中横坐标代表遗传算法的迭代次数,纵坐标代表每一代种群得到的最优结果(即最低的感知器分类错误率)。为了验证进行基于相关分析的特征过滤的效果,图1(b)给出了针对未经过滤的原始12600维的特征集合运用遗传算法进行特征选择的实验结果,这两个实验均选出 2 维特征。从图 1(a)中可以看出,在遗传算法的迭代过程中,2 维 Prostate Cancer 数据集的感知器分类错误率在持续下降,迭代到第 10 次时就收敛到一个较低的错误率 3.92%。而图 1(b)中可以看出,在针对没有过滤的原始特征集合进行特征选择时,遗传算法的效率及效果变差,一方面在迭代近 20 次时才收敛,另一方面收敛时得到的分类错误率高达 11.77%。图 2 给出了 Prostate Cancer 数据集在 FSRAGA 算法所选出的优化的 2 维(图 2(a))及 3 维(图 2(b))特征子空间中的分布散点图,分别用“*”及“▽”代表两类样本,从图中可看出样本集在对应的特征子空间中具有较好的线性可分性,其中 3 维特征子空间中的线性可分性要优于 2 维特征子空间。

5 结语

本文主要针对高维数据的特征选择问题,融合 filter 及 wrapper 特征选择模型,提出了一种基于相关性分析及遗传算法的特征选择算法。实验证明,该算法能较为有效地找出具有较好的可分离性的特征子集,从而实现降维并提高分类精度。

参考文献:

[1] JOHN GH, KOHAVI R, PFLEGER K. Irrelevant Features and the Subset Selection Problem[A]. Proceedings of the Eleventh International Conference on Machine Learning[C]. New Brunswick, NJ, USA, Morgan Kaufmann, 1994. 121 - 129.

[2] KOHAVI R, JOHN GH. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997, 97(1 - 2): 273 - 324.

[3] LIU H, YU L. Toward Integrating Feature Selection Algorithms for Classification and Clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(5): 491 - 502.

[4] YANG J, HONAVAR V. Feature subset selection using a genetic algorithm[J]. IEEE Intelligent Systems, 1998, 13(2): 44 - 49.

[5] YU L, LIU H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Learning Research, 2004, (5): 1205 - 1224.

[6] HAN J, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰译. 北京, 机械工业出版社, 2001.

(上接第 1402 页)

表 1　SGA 与 SRAGA 评估性能对照表

函数	SGA		SRAGA		改进率(%)	
	平均收敛代数	不收敛次数	平均收敛代数	不收敛次数	收敛速度*	收敛率**
F1	62.036	172	33.412	35	185.67	589.29
F2	61.097	45	47.859	9	127.66	123.23

*收敛速度=SGA 平均收敛进化代数/SRAGA 平均收敛进化代数

**收敛率=SRAGA 收敛次数/SGA 收敛次数

3 结语

标准遗传算法(SGA)在遗传算子的作用下,使得一些优秀的基因片段过早丢失,从而整个种群丧失多样性,致使算法停滞,长时间徘徊在局部极值附近而难于跳离,因此常常不能得到满意的全局最优解。本文提出的带基因丢失检测及修复策略的自适应遗传算法(SRAGA)在稳定性、运行效率和全局寻优能力上与 SGA 相比均有较大的优势。

参考文献:

[1] HOLLAND JH. Adaption in Natural and artificial system (second edition)[M]. Cambridge, MA: MIT press, 1992.

[2] DE JONG KA. An analysis of the behavior of a class of genetic adaptive systems[D]. University of Michigan, 1975. 76 - 9481.

[3] GOLDBERG DE. Genetic algorithms in search, optimization & machine Learning[M]. Addison-Wesley Publishing Company, 1989.

[4] HATTA K, WAKABAYASHI S, KOIDE T. Adaptation of genetic operators and parameters of a genetic algorithm based on the elite degree of an individual[J]. Systems and Computers in Japan, 2001, 32(1): 29 - 37.

[5] 段玉倩, 贺家李. 遗传算法及其改进[J]. 电力系统及其自动化学报, 1998, 10(1): 39 - 52.

[6] KIVIJARYI J, FRANTI P, NEVALAINEN O. Self-adaptive genetic algorithm for clustering[J]. Journal of Heuristics, 2003, 9(2): 113 - 129.