基于遗传算法的多任务学习*)

孟浩华　李国正
(上海大学计算机工程与科学学院　上海 200072)

摘　要　机器学习中冗余特征会降低学习器的性能，而特征选择方法可以去掉一些冗余特征。然而，冗余特征也包含有用信息，因此可以利用多任务学习的概念，通过重复利用冗余特征提高预测精度。但是，如何确定哪些特征作为输入和输出仍然是一个待解决的问题。之前的工作是在多任务学习当中，运用遗传算法来确定哪些特征作为输入，哪些作为输出，取得了较好的效果，但是该算法不足之处是没有考虑到不相关特征。现将特征分为三部分：输入的特征、输出的特征和不相关特征，提出了对一个特征进行双位编码的遗传算法搜索策略。在基因芯片数据上的实验结果表明，提出的新算法 e-GA-MTL 比已有基于遗传算法的 GA-MTL 和其它启发式方法效果更好。

关键词　遗传算法，多任务学习，双位编码

Study on Multi-task Learning Based on Genetic Algorithm

MENG Hao-hua　LI Guo-zheng
(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)

Abstract　Redundant features hurt the performance of learning methods. Feature selection methods were developed to remove some redundant features; however, the redundant features contain useful information, therefore, multi-task learning was developed to employ the removed redundant information to improve prediction accuracy. Adding which features to the target and/or the input during multi-task learning is still an open issue. The previous study on multi-task learning uses genetic algorithm to determine the features for the target and/or input, and which has been proved effective. In this paper, we classified the features into three parts: the input features, the output features and irrelevant features. We proposed a new search strategy of the genetic algorithm which encodes double bits for one feature. Experimental results on the microarray data sets show that the novel algorithm e-GA-MTL obtains better performance than the previous algorithm GA-MTL and the other heuristic algorithms.

Keywords　Genetic algorithm, Multi-task learning, Microarray data analysis

1　引言

随着数据收集和计算能力的提高，越来越多的数据可以用来作为机器学习。但是，太多的冗余特征降低了学习器的性能，因此研究者提出了特征选择方法来去除冗余或者不相关特征。在机器学习领域，特征选择变成了一个热点主题[1,2]。目前已经提出了许多特征选择的方法来去除掉冗余的特征，分为过滤式、包装式和嵌入式等方法[1-4]。尽管有些去除掉的特征是冗余的和弱相关的，但是它们也包含了有用的信息，可以用来提高预测精度。多任务学习(Multi-Task Learning, MTL)就是一个利用冗余信息的方法，它的做法是将去除掉的特征添加到学习器的输出端[5]。尽管 MTL 改进的效果很有限，但是在处理像基因和多元校正等真实世界的问题时是有意义的[6]。

之前多任务学习的搜索方法主要是基于启发式搜索[5,6]，这种方法对于选择作为输入/输出的特征个数有一定的武断性。当把搜索方法作为一种组合优化问题时，随机的搜索方法也是一种，其中遗传算法[8]是一种易用且有效的方法，实验结果表明遗传算法在特征选择方面能够获得满意的结果[7]。GA-MTL[8]是一种将遗传算法运用到多任务学习的算法，它是利用遗传算法作为特征选择的方法，特征分为输入特征和输出特征。但是该算法的一个关键问题是将特征用 1 位来表示，每个特征只能被选中为输入或输出，不会被去除。可以看出，它没有考虑到不相关特征，因此，本文提出了一种新的算法：e-GA-MTL，它可以将特征分为三部分：不相关的特征、输入端特征、输出端特征，其中不相关的特征会被剔除掉。

本文第 2 节简要介绍多任务学习，然后详细介绍 e-GA-MTL 及相关方法。第 3 节，在基因芯片数据集上将展示 e-GA-MTL 与其他相关方法的比较，并讨论 e-GA-MTL 的效果。最后是结论。

2　多任务学习

多任务学习(Multi-task learning, MTL)[5]用于冗余特征重用的基本思想就是把特征选择过程中选择的特征作为输入特征，而把一部分去除的特征重新收回作为目标输出。

现有几种多任务学习的启发式搜索方法[5,6]，比如 Caruana 和 de SA[5]提出了一种基于熵的过滤式特征选择模型和一种基于核回归的嵌入式模型，使用前 N 个(个数是事先定义好的)特征作为输入特征集，剩下的作为目标输出。Li 等人[6]用聚类算法来进行选择特征，其做法是先使用 Kohonen 神经网络聚类，然后选择离聚类中心点近的特征作为输入特

*)本文得到国家自然科学基金资助(20503015)。**孟浩华**　硕士研究生；**李国正**　博士，副研究员，研究方向为机器学习和生物信息学。

征子集，其他未被选中的特征按照离输入特征的欧式距离排列，距离较小的前几个特征作为目标输出。

因为上面几种方法对于特征个数的确定是启发式的，很任意，为了自动确定特征个数，Li 等人[8]提出一种基于遗传算法的搜索方法 GA-MTL，作者将 GA-MTL 与上述启发式 MTL 方法和单纯的特征选择方法进行了比较，显示有较好的效果。但是该方法仍然存在一个问题，在 GA-MTL 中对特征采用一位二进制编码，特征要么选中作为输入，要么作为输出，不相关特征无法去除，因此本文在 GA-MTL 基础上进一步提出了采用双位二值制编码的策略，使特征有被去除的机会，算法称为 e-GA-MTL(Enhanced version of Genetic Algorithm Based Multi-task Learning)。下面分别介绍已有的 GA-MTL 算法、启发式算法和本文提出的新算法。

GA-MTL。在 GA-MTL 中，构造一个二进制的染色体，其长度为特征向量的长度，每个基因为对应的一个特征，当相应基因是 1 时表示此特征被选择作为输入，是 0 时，表示此特征作为目标输出。适应度函数定义如下：

$$fitness=\frac{1}{3}E_{dr}+\frac{2}{3}E_{dv}$$

其中 E_{dr} 是基学习器的训练错误率，E_{dv} 是在在验证集上的预测错误率。

e-GA-MTL。在 e-GA-MTL 中，同样是构造一个二进制的染色体，与 GA-MTL 不同的是，每个基因体上的位数是 2 位。当某个特征对应的二进制位是 00 时，这个特征将被删除，当为 01 时，将被作为输出，当为 10 时，将被作为输入，当为 11 时，既作为输入也作为输出。e-GA-MTL 的算法描述如图 1 所示。

输入：训练集 D_t，验证集 D_V，测试集 D_S，和基学习器

步骤：

1. 产生一个 2 位二进制的权重向量
2. 按照式(1)的适应度函数在 Dr 和 DV 对权重向量 w 进行演化
3. w* 为进化过程中最好的权重向量
4. 在 D_S 进行测试，w* 中为 00 的特征丢掉，w* 中为 01 的作为输出，w* 中为 10 的作为输入，w* 中为 11 的既作为输入，又作为输出

输出：在测试集 D_S 上的测试精度

图 1　e-GA-MTL 算法

H-MTL。为了与其它启发式多任务学习做比较，这里在 Caruana 和 de Sa[5] 的想法的基础上，用一种嵌入式特征选择方法，基于预报风险的特征选择方法[8-10]，来对特征进行排序，其做法是通过计算每个特征被其平均值替代前后的训练误差的差值来对各个特征排序：

$$S_i=ERR(\bar{x}^i)-ERR \tag{1}$$

其中，ERR 是训练误差。$ERR(\bar{x}^i)$ 是在如下的训练集上的测试误差：

$$ERR(\bar{x}^i)=\frac{1}{N}\sum_{j=1}^{N}(\tilde{y}(x_j^1,\cdots,\bar{x}^i,\cdots,x_j^M)\neq y_j) \tag{2}$$

其中，M,N 分别是特征和样本个数，$\bar{x}^i$ 是第 i 个特征的平均值，$\tilde{y}()$是对第 j 个样本(这个样本的第 i 个特征被其平均值替代)的预测值，S_i 值为 0 的特征将被作为不相关特征被去掉，因为这个特征在学习过程中是无用的。

算法命名为 H-MTL(Heuristic Multi-Task Learning)。在 H-MTL 算法中，运用风险预报准则将特征按照升序排列。值为 0 的特征将被删除，前面四分之一的特征作为输出，剩下是四分之三的特征作为输入。

GA-FS。为了显示多任务学习方法的有效性，只用特征选择方法选择特征，其它特征去除，算法表示为 GA-FS(Genetic Algorithm based Feature Selection)，在 GA-MTL 中用的 GA 在这里仅用作特征选择。GA-FS 和 GA-MTL 唯一区别在于二进制染色体中为 0 的位的处理方式不同；在 GA-MTL 中，当某一个特征所对应的染色体位为 0 时，这个特征会被选择作为目标输出，而在 GA-FS 中，这个特征将被删除。

基学习器。作为示例，人工神经网络是一种使用频率很高的有效的方法，不失一般性这里使用改进的多层感知器神经网络[11]作为基学习器。

3　实验

3.1　实验数据集

本文使用如表 1 所示的 4 个基因数据集[12]对算法进行比较。

表 1　用于比较的基因数据集

数据集	特征个数	样本个数	类别个数
Colon	2000	62	2
Breast Cancer	24481	97	2
Leukemia	7129	72	2
Ovarian	15154	253	2

1)Colon：Alon 等人使用 Affy 矩阵核苷酸序列对超过 6500 个人的基因进行模拟，其中样本中有 40 人有瘤，22 个有正常的大肠组织。实验数据保留了 2000 个拥有最小密度的在 62 个组织上的基因样本。

2)Breast Cancer：这个数据集是由 Van't Veer 等人公布的。训练集包含了 78 个病人样本；在测试集中有 12 个病人样本旧病复发了，7 个病人样本康复了。基因的个数为 24481。

3)Leukemia：这个急性白血病数据集是由 Golub 等人公布的。训练数据集包括了 38 个骨髓样本(27 个正常以及 11 个急性骨髓白血病)，从 6817 年人类基因中使用了超过 7129 个探针。此外，还有 34 个样品的测试数据，其中 20 个是正常和 14 个是急性骨髓白血病。

4)Ovarian：这个数据集是用来通过鉴定血清中的蛋白质模型来区分是否得了卵巢癌。

对于所有的数据集，首先将符号型的特征转化为数字型，然后将所有的特征变换到区间[−1,1]之间。最后，把每个数据集合并后均分为两部分，一份作为训练样本 D_r，另一份用作测试样本 D_s。这样的操作一共进行 50 次。对于所有的数据集，将从训练集中随机的抽出四分之一作为遗传算法的验证集 D_v。

3.2　实验结果

将本文提出的 e-GA-MTL 在基因芯片数据集上与已有的几种方法 GA-MTL，H-MTL 和 GA-FS 进行比较。其中，遗传算法的相关参数是这样的：种群个数为 50，遗传代数为 20，进化率为 0.6，突变率为 0.001。神经网络的隐层结点数为 10 个。

分类错误率为定义如下：

$$ERR=\frac{1}{N}\sum_{j=1}^{N}(\tilde{y}_j\neq y_j)*100\%$$

(下转第 203 页)

thesizing Web pages // Fifteenth National Conference on Artificial Intelligence. Madison, 1998
[6] 杨怡玲，管旭东，尤晋元. 基于页面内容和站点结构的页面聚类挖掘算法[J]. 软件学报，2002，13(3)：467-469
[7] Fayyad U M, Piatetsky S G, Smyth P. The KDD process for extracting useful knowledge from volumes of data[J]. Communications of the ACM, 1996, 39(11): 27-34
[8] Fu Y, Creado M, Shih M Y. Adaptive Web Sites by Web Usage Mining // International Conference on Internet Computing 2001. Las Vegas, USA, 2001
[9] Doyle J K, Graver J E. Mean distance in a graph[J]. Discrete Math., 1977, 17: 147-154
[10] West D B. Introduction to Graph Theory (2nd Edition) [M]. Published by Prentice Hall 1996, 2001: 67-107
[11] Broder A, Kumar R, et al. Graph structure in the Web[C]. Amsterdam, Netherlands, 2000
[12] Kumar R, Raghavan P, Rajagopalan S, et al. The Web as a graph [C] // Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Dallas, Texas, United States, 2000: 1-10
[13] Borges J, Levene M. A Clustering-Based Approach for Modelling User Navigation With Increased Accuracy // Proceedings of the Second International Workshop on Knowledge Discovery from Data Streams (IWKDDS) in conjunction with PKDD 2005. Porto, Portugal, Outubro, 2005

(上接第 187 页)

其中 N 为测试集样本个数，$\tilde{y}_j$ 为预测值。

表 2　分类错误率统计结果(%)

数据集	All Feature	GA-FS	H-MTL	GA-MTL	e-GA-MTL
Colon	29.0±9.2	27.5±9.1	28.4±9.7	26.8±7.0	25.0±7.8
Breast	23.2±2.6	22.7±4.6	23.1±3.5	22.5±5.7	21.5±2.7
Leukemia	30.0±3.2	24.4±3.3	23.3±5.2	23.0±3.9	22.7±4.2
Ovarian	25.8±3.3	23.3±4.3	21.3±4.8	21.2±3.5	21.1±3.1
平均值	27.0±4.6	24.5±5.3	24.0±5.8	23.4±5.0	22.56±4.5

从表 2 中可以看出，5 个算法的分类错误率在平均值上从小到大依次是 e-GA-MTL＜ GA-MTL ＜ GA-FS＜ H-MTL＜ All Feature。除了 Breast 数据集上 GA-FS 结果比 H-MTL 好之外，其它所有数据集上这个结论也都成立。这个结果表明，1)使用了特征选择的 e-GA-MTL，GA-MTL，H-MTL 和 GA-FS 比没有做特征选择的要好；2)使用了多任务学习的 e-GA-MTL，GA-MTL，H-MTL 比没有使用 GA-FS 的要好；3)在多任务学习的基础上，在用来确定哪些特征作为输入/输的方法中，使用遗传算法的 e-GA-MTL 和 GA-MTL 要比预先定义阈值的 H-MTL 方法要好；4)去掉不相关特征的方法 e-GA-MTL 比没有去掉 GA-MTL 的要好。

3.3　讨论

从实验结果中可以看到，本文提出的算法 e-GA-MTL 要好于以前的算法：All Feature，GA-FS，H-MTL 和 GA-MTL。特别是要好于 GA-MTL。

我们知道，不同的特征对于提高学习器的性能会有不同的贡献[3]。有的特征对于学习器的性能的提高效果很明显，其它特征不能替代，可以归类为强相关特征；有的特征效果一般，有替代的特征，但是对学习器性能也有所提高，可以归类为弱相关特征；还有一些特征可以说对于学习器的性能没有提高，反而会损害学习器性能，这种特征就是不相关特征。

e-GA-MTL 能够通过遗传算法自动地把特征分为上述三类，即二进制染色体对应位是 00 的特征是不相关特征，在训练模型之前会被去掉；二进制染色体上对应基因是 10 的特征是强相关特征，在训练模型时作为 MTL 的输入端特征；二进制染色体上对应基因是 01 的特征是弱相关特征，在训练模型时作为 MTL 的输出端；二进制染色体上对应基因是 11 的特征既是强相关特征，又是弱相关特征，也就是临界于强、弱相关特征之间的，这种特征在训练模型时既作为 MTL 的输入，又作为输出。

GA-MTL 表现得没有 e-GA-MTL 好，是由于 GA-MTL 没有考虑不相关特征的影响，而不相关特征会降低多任务学习的泛化能力和鲁棒性。

结束语　本文提出了一个基于遗传算法的多任务学习算法，e-GA-MTL，较之已有的搜索算法，有效提高了多任务学习的预测精度。实验表明，e-GA-MTL 在预测精度上要高于 GA-MTL 以及之前的其他算法，这是由于 e-GA-MTL 不仅能够自动地决定哪些特征作为输入，哪些特征作为输出，而且它能够自动地去掉那些不相关的特征。不相关的特征是 GA-MTL 没有考虑到的。

本文仅探讨了 e-GA-MTL 在分类问题上的有效性，对于回归问题还有待评测。另外，由于特征是被划分为 3 类：输入、输出、删除，而算法中对特征的指派却有 4 个状态：00，01，10，11。特别是对状态为 11 的特征，我们目前的处理是既作为输入，又作为输出，这个问题还可以做进一步的探讨。

参 考 文 献

[1] Liu Huan , Yu Lei . Toward integrating feature selection algorithms for classification and clustering. IEEE Trans on Knowledge and Data Engineering, 2005, 17(3): 1-12
[2] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research, 2003, 3: 1157-1182
[3] Yu Lei, Liu Huan. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 2004, 5 (10): 1205-1224
[4] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312
[5] Carnana R, de Sa V R. Benefiting from the variables that variable selection discards. Journal of machine learning research, 2003, 3: 1245-1264
[6] Li G Z, Yang J, Lu J, et al. On multivariate Calibration problems // Lecture Notes on Computer Science 3173. Springer, August 2004: 389-394
[7] Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 1998, 13: 44-49
[8] Li Guo-zheng, Liu Tian-yu. Improving Generalization Ability of Neural Networks Ensemble with Multi-Task Learning. Journal of Computational Information Systems, 2006, 2(4): 1235-1239
[9] Verikas A, Bacauskiene M. Feature selection with neural networks. Pattern Recognition Letters, 2002, 23: 1323-1335
[10] Li Guo-zheng, Yang Jie, Liu Guo-ping, et al. Feature selection for multi-class problems using support vector machines // Lecture Notes on Artificial Intelligence 3173. Auckland, New Zealand, Springer, 2004: 292-300
[11] Foresee F D , Hagan M T . Gauss - newton approximation to Bayesian regularization // Proceedings of the 1997 International Joint Conference on Neural Networks. 1997: 1930-1935
[12] Li J, Liu H. Kent Ridge Biomedical Data Set Repository. Available at: http://sdmc-lit. org. sg/GEDatasets/Datasets. html, 2002