第 26 卷第 1 期　　四川理工学院学报(自然科学版)　　Vol. 26　No. 1
2013 年 2 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Feb. 2013
文章编号:1673-1549(2013)01-0033-05　　DOI:10.3969/j.issn.1673-1549.2013.01.008

特征提取与多目标机器学习研究及应用

何　涛, 张洪伟, 邹书蓉

(成都信息工程学院计算机学院, 成都　610225)

摘　要:特征提取与多目标机器学习算法是基于多目标协同 EA 提出的,该算法通过对学习样本多属性进行特征提取找出其核属性,由核属性与其他非核属性组成属性组,从而提高了分类的精度。各属性组再按相似性和类标签进行有监督地聚成类簇,类簇个数和中心根据适应度矢量函数通过机器学习算法自动确定,这样类簇个数和中心就不受主观因素的影响并且保证了这两个关键要素的优化性质。待分类样本的类属是按离某个类簇中心距离最近邻法则和该类簇的类标签来判定。最后,将算法应用到 UCI 数据集中的 Liver Disorders 和 Hepatitis 两个数据集,以及浙江省北部地区夏天异常高温天气预测。通过实验表明,特征提取与机器学习算法优于著名的朴素贝叶斯、C4.5、SVM 算法。

关键词:特征提取;核属性;机器学习;多目标 EA;监督聚类

中图分类号:TP301.6　　**文献标志码**:A

引　言

令 W 是给定世界的有限或无限的所有观测对象的集合,由于我们观察能力的限制,我们只能获得这个世界的一个有限的子集 $Q \subset W$, 称为样本集。机器学习就是根据这个样本集,推算出这个世界的模型,使它对这个世界(尽可能地)为真[1]。

目前,机器学习领域的研究工作主要围绕在以下三个方面进行:

(1)面向任务的研究。研究和分析改进一组预定任务的执行性能的学习系统。

(2)认知模型。研究人类学习过程并进行计算机模拟。

(3)理论分析。从理论上探索各种可能的学习方法和独立于应用领域的算法。

机器学习是继专家系统之后人工智能应用的又一重要研究领域,也是人工智能和神经计算的核心研究课题之一。现有的计算机系统和人工智能系统没有什么学习能力,至多也只有非常有限的学习能力,因而不能满足科技和生产提出的新要求。对机器学习的讨论和机器学习研究的进展,必将促使人工智能和整个科学技术的进一步发展。

本文的特征提取与多目标机器学习算法正是根据机器学习中的聚类算法和进化算法理论提出的新的机器学习算法。算法有以下主要特点:(1)通过特征提取进行降维操作找到核属性,由核属性与其他非核属性组成属性组,从而提高了分类的精度。利用各最小属性组协同进化,能有效的跳出局部最优点而寻找更好的优化解;(2)采用机器学习算法中的多目标进化算法,它是自适应全局优化概率搜索算法,具有简单通用、鲁棒性强、适于并行处理的优点。引入适应度矢量函数和擂台选择法,而不使用聚集函数法[2-3],这就解决了难以搜索到非凸解的问题[4];(3)根据预估类簇数和适应度矢量函数通过机器学习算法自动确定类簇个数和类簇中心,而不受主观因素的影响,从而提高了分类的可靠性;(4)为减少数据噪声对学习的影响,对数据进行预处理,提高学习的准确性。

将算法应用到 UCI 数据集中的 Liver Disorders 和

收稿日期: 2012-12-11

作者简介:何 涛(1988-),男,四川南充人,硕士生,主要从事智能工程方面的研究,(E-mail) vanceho@163.com

Hepatitis 两个数据集,以及浙江省北部地区夏天异常高温天气预报,实验结果表明,该算法具有独特的智能性和较高的准确性和实用性。

1 算法模型

1.1 算法描述

进化算法(Evolutionary Algorithm,简称 EA)作为一类启发式搜索算法,已被成功应用于多目标优化领域,发展成为一个相对较热的研究方向。进化算法通过在代与代之间维持由潜在解组成的种群来实现全局搜索,这种从种群到种群的方法对于搜索多目标优化问题的 Pareto 最优解集是很有用的[5]。

本文提出的算法基本思想是:由特征提取对样本集进行属性分解得到各子属性组。将各子属性组所对应的学习样本集通过学习器进行机器学习。根据学习样本的类簇号构成染色体及种群,且随机生成并修补;根据类内最小化和学习误差最小原则构造适应度矢量函数;利用多目标 EA 全局寻优的特点对学习样本进行多目标优化,找到较好的染色体并将其保存下来以便每次系统运行的时候将其调出进行运算,找到更优的解;用找到的染色体根据最近邻法则用较优染色体及类标签对测试样本进行预测,最后根据学习误差小及类内距离最小化原则,筛选出各子属性组最优染色体,再根据最近邻法则用较优染色体及类标签对测试样本进行协同分类。只要一个子属性组将该样本分为显性类,则该样本为显性。

1.2 特征提取算法

本文采用的特征提取的方法为降维机器学习算法,这是一种自上而下的搜索方法,从全部特征开始每次剔除一个特征。具体规则如下:

(1)初始属性个数 m,去掉的属性个数 $n=0$,当前保留属性集 P,当前去掉属性集 D。

(2) $if(x_{ij}==x(i+1)j)\ j++$,其中 $j\in P$,x_{ij} 表示第 i 行样本的第 j 列属性所对应的值。

(3) $i\in P$ 判断 i,$i+1$ 这两个样本的类标签是否一致,如果不一致则找到学习样本的核属性,属性组 P 则为核属性组,跳到(6);如果一致,$D\neq\phi$ 则转到(5),否则转到(4)。

(4) $n=n+1$,$m=m-1$。

(5) 从 P 中取 n 个属性加入 D 中,转到(2),直至 P 中每 n 个组合都被取出过。

(6) 由核属性组构造各子属性组 $Propertyi$,每个子属性组由核属性组加上一个非核属性构成。

1.3 多目标机器学习算法的体系结构

1.3.1 预估类簇数

本文采用基本 K 均值算法作为多目标机器学习算法预估类簇数的算法,基本 K 均值算法步骤如下:

(1) 选择 K 个点作为初始中心。

(2) 将每个点指派到最近的中心,形成 K 个簇。

(3) 重新计算每个类簇的中心。

(4) 若类簇中心不再发生变化,算法停止;否则,转到步骤(2)。

1.3.2 数据预处理

为了得到更准确无噪音的样本数据,需对所有样本进行去噪的预处理。步骤如下:

(1)如果该样本数据中含有非法值或空值,则直接从样本集中删除该样本。

(2)如果某属性的值在[0,1]区间内,则不做任何处理;否则,对该属性进行归一化处理:

$$x'_{ij}=[\max(\{x_{1j},x_{2j}\ldots x_{nj}\})-x_{ij}]$$
$$/[\max(\{x_{1j},x_{2j}\ldots x_{nj}\})-\min(\{x_{1j},x_{2j}\ldots x_{nj}\})]$$

其中,n 为样本的个数,x_{ij} 表示第 i 行样本的第 j 列属性所对应的值,x'_{ij}表示处理后的值。

1.3.3 编码与解码

编码:输入学习样本总数为 N,可能被分成的类簇数为 c,染色体 e 定义为:

$$e=(k_1,k_2,\cdots,k_N)\quad k_i\in\{1,2,\cdots,c\}$$

其中 k_i 表示每个样本的类簇号,若相同,即表示这些样本属于同一类簇,具有相同的类标签。

解码:$f(x_i)=k_i,i\in\{1,2,\cdots,N\}$,其中 i 表示样本编号,k_i 表示样本 x_i 所对应的类簇号。

1.3.4 类簇类标签判定

(1) 若该类簇所涵盖的学习样本的类标签一致,则该类簇的类标签为学习样本的类标签。

(2) 若该类簇所涵盖的学习样本的类标签不一致,则统计各类标签所涵盖的学习样本的数量,涵盖学习样本最多的类标签即为该类簇的类标签。

1.3.5 修补染色体

一条染色体中,若存在样本 x_1,x_2 的类簇号(基因)相同,但对应的类标签不同,即类标签不同的样本被划分到同一类簇中,则该基因需要修补。修补的思想是将 x_1,x_2 两个样本根据最近邻法则调整到不同的类簇中。具体实现步骤如下:

(1) 选择 x_1 或者 x_2,与其它类簇中和 x_1 或者 x_2 具有相同类标签的样本或聚类中心求距离,并选出距离最小的类簇号 k_0:

$$k_0 = \arg\min_r \{d(x_i^l, x_r^l) = \sum_{j=1}^{m} (x_{ij} - x_{rj})^2\}$$

(2) 根据找到距离最小的类簇号 k_0, 将 x_1 或者 x_2 的类簇号调整为 k_0。

(3) 继续搜索该染色体中其他基因位是否需要修补,若需要则反复执行(1)、(2),否则该染色体修补完成。

1.3.6　适应度矢量函数

m 为样本属性的个数, x_i, x_k 表示第 i, k 号样本。定义距离:

$$d(x_i, x_k) = \sum_{j=1}^{m} (x_{ij} - x_{kj})^2$$

根据类内最小化与学习误差最小原则定义适应度矢量函数:

$$F(e) = \min\{\sum_{r=1}^{c} [nr \sum_{i=1}^{n_r} d(x_i^r, s_r)], f\}$$

$$s_r = \frac{1}{n_r}(\sum_{i=1}^{n_r} x_{i1}^r, \sum_{i=1}^{n_r} x_{i2}^r, \cdots, \sum_{i=1}^{n_r} x_{im}^r)$$

其中:f 是学习误差,即用得到的类簇中心对学习样本分类,分类错误的数量, x_i^r 是属于第 r 类的样本, n_r 为第 r 类的样本个数, s_r 为第 r 类的类簇中心, c 是类簇个数。

1.3.7　类别判定方法

由最优染色体计算各类聚类中心 m_r, 分类是根据最近邻法则:

$$r_0 = \arg\min_r \{d(x_i, m_r) = \sum_{j=1}^{m} (x_{ij} - m_{rj})^2\}$$

若输入的待分类样本离聚类中心 m_r 的距离最近,则该样本的类标签与第 r 类的相同。每个子属性组都进行相同操作,最后根据各子属性组的分类结果协同决策,只要在任一子属性组中该待分类样本的类标签为显性,则将其归类为显性。

1.4　算法步骤

(1)学习样本特征提取,划分为 p 个子属性组。

(2)选择初始化学习样本集 M 和待分类样本集 N,初始各子属性组样本数据及算法运行参数,设置每组子属性的运算次数 *iter*1,样本归一化处理。

(3)设置该子属性组学习迭代次数 *iter*2。

(4)随机生成多个染色体构成种群,并对每个染色体进行修补。

(5)计算各染色体的适应度矢量函数值。

(6)利用 AP 算法[6] 对种群进行选择,被选择出染色体称为父染色体,并将其加入到记忆池中。

(7)父染色体进行交叉和变异生成新染色体,并对该染色体进行修补。

(8)若新生成的染色体的数量小于种群的数量,则随机生成染色体,并对其进行修补,让其作为下一代种群。

(9)若满足 *iter2*,则对记忆池中的染色体求适应度,并用 AP 算法做选择,将较优染色体保存下来,否则转到(5)。

(10)若满足 *iter1*,则取下一组子属性转到(3),直至所有子属性组被取完,否则转到(3)。

(11)将所有子属性组得到的染色体按学习误差小及类内最小化原则,筛选出各子属性组最优染色体,进行协同决策分类。

2　实验与结果分析

2.1　实验数据

为了验证本文提出算法的有效性和实用性,将用 UCI Machine Learning Repository 中的 LDD(Liver Disorders Dataset)和 Hepatitis Dataset 数据集来进行实验,并将算法应用到浙江省北部地区夏天异常高温的预测中。其中,浙江省北部地区夏天异常高温数据来源于文献[7]。对于数据集浙江省北部地区夏天异常高温按照选取前 32 年样本作为学习样本,后 8 年样本作为测试样本。LDD 与 Hepatitis,采用十折交差验证算法,将数据集分为 10 份,轮流让其中的每 1 份作为测试样本,其余 9 份作为学习样本来验证算法。利用机器学习软件 weka 中的 SimpleKMeans 通过多次试验来确定预估类簇数。实验数据描述见表 1。

表 1　实验数据描述

数据集	LDD	Hepatitis	异常高温
条件属性	6	19	9
类标签数	2	2	2
学习样本数	307	72	32
测试样本数	34	8	8
显性类	Patient	Die	高温

2.2　实验结果与分析

2.2.1　Liver Disorders

采用 UCI 数据集的 LDD(判断病人是否为乙肝患者)的数据样本作为算例。该数据样本中总样本数为 345,去掉其中 4 个重复的样本后。患者数为 142 人,非患者数为 199 人,数据样本包括 6 个条件属性和 1 个决策属性。条件属性包括: Mcv (X_1)、Alkphos (X_2)、SGPT (X_3)、SGOT (X_4)、Gammagt (X_5)、Drinks (X_6)。

通过特征提取降维操作,得到核属性为 (X_1, X_4, X_5, X_6)。由其余任一非核属性与核属性构成最小属性组,各属性组见表 2。

表 2 子属性组

子属性组号	属 性
1	$(X_1, X_2, X_4, X_5, X_6)$
2	$(X_1, X_3, X_4, X_5, X_6)$

采用十折交差验证进行实验,算法运行参数如下:预估类簇数:100;每组子属性运行次数:8;种群规模:100;遗传代数:500;交叉概率:0.8;变异概率:0.2。经过 10 折交叉实验验证,测试样本分类结果的平均正确率为 74.12%。通过实例说明该算法在应用到中规模数据样本中,能得到较满意的分类结果。将本文提出的算法与朴素贝叶斯、C4.5 决策树、BP 神经网络、KNN 和 SVM 五种分类算法的平均分类正确率相比[8-9]结果见表 3。

表 3 几种算法的分类正确率对比

算法	NBC	C4.5	BP	KNN	SVM	本文算法
正确率	56.52%	68.69%	71.59%	62.89%	58.26%	74.12%

2.2.2 Hepatitis

采用 UCI 数据集的 Hepatitis 的数据样本作为算例。该数据样本中总样本数为 155,其中,32 个样本为 Die(显性),123 个样本为 Live(隐性)。去掉有缺失数据的样本后,样本总数为 80。数据样本包括 19 个条件属性和 1 个决策属性。条件属性包括:AGE (X_1)、SEX (X_2)、STEROID (X_3)、ANTIVIRALS (X_4)、FATIGUE (X_5)、MALAISE (X_6)、ANOREXIA (X_7)、LIVER BIG (X_8)、LIVER FIRM (X_9)、SPLEEN PALPABLE (X_{10})、SPIDERS (X_{11})、ASCITES (X_{12})、VARICES (X_{13})、BILIRUBIN (X_{14})、ALK PHOSPHATE (X_{15})、SGOT (X_{16})、ALBUMIN (X_{17})、PROTIME (X_{18})、HISTOLOGY (X_{19})。

通过特征提取降维操作,得到核属性为 ($X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{19}$)。由其余任一非核属性与核属性构成最小属性组。

采用十折交差验证进行实验,算法运行参数如下:预估类簇数:30;每组子属性运行次数:5;种群规模:100;遗传代数:500;交叉概率:0.8;变异概率:0.2。经过 10 折交叉实验验证,测试样本分类结果的平均正确率为 90%。通过该实例说明该算法在应用到中规模数据样本中,同样能得到较满意的分类结果。将本文提出的算法与朴素贝叶斯、C4.5 决策树、Bagging、Boost 和 SVM 五种分类算法的平均分类正确率相比[8-11]结果见表 4。

表 4 几种算法的分类正确率对比

算法	NBC	C4.5	Bagging	Boost	SVM	本文算法
正确率	65.7%	78.7%	80.6%	83.8%	81.9%	90%

2.2.3 夏天异常高温天气预报

采用浙江省北部地区夏天异常高温天气数据样本作为算例,根据文献[11],异常高温气候数据样本共 10 个属性,其中 9 个条件属性,1 个决策属性。条件属性包括:上年 7~9 月降水量 (X_1)、上年 7~8 月的降水量 (X_2)、上年年积温 (X_3)、当年 3 月温度 (X_4)、当年 1~6月积温 (X_5)、上年 7~12 月积温增量 (X_6)、上年 4~6月降水量 (X_7)、上年 11 月温度 (X_8)、上年 10~12 月积温 (X_9)。通过特征提取降维操作,得到核属性为 (X_4, X_6)。由其余任一非核属性与核属性构成最小属性组。

文献[11]中共有 1956 年到 1996 年这 40 年的数据样本,选择 1956 年到 1988 年 32 年的数据作为学习样本,1989 年到 1996 年 8 年数据为待分类样本。算法运行参数如下:预估类簇数:7;每组子属性运行次数:3;种群规模:50;遗传代数:500;交叉概率:0.8;变异概率:0.2。各属性组预测结果见表 5。

表 5 异常高温分类结果 (1:高温,0:非高温)

年份	各子属性组的分类结果							实际类别
	1	2	3	4	5	6	7	
1989	0	0	0	0	0	0	0	0
1990	1	1	1	1	1	1	1	1
1991	0	0	0	0	0	0	0	0
1992	1	0	1	0	0	0	0	1
1993	0	0	0	0	0	0	0	0
1994	1	1	1	1	1	1	1	1
1995	0	0	0	0	0	0	0	1
1996	0	0	0	0	0	0	0	0

由各子属性组预报结果,根据各子属性组协同决策分类,则准确率为 87.5%。由此可见,本文所提出的算法在异常高温预报方面具有重要的意义,通过实例的验证,说明该算法的有效性。

3 结束语

本文基于多目标协同 EA 提出特征提取与多目标机器学习算法,并将其应用到 UCI 数据集中的 Liver Disorders 和 Hepatitis 两个数据集,以及浙江省北部地区夏天异常高温天气预报实例中。由实验结果可知,该算法表现出其独特的智能性和准确性以及实用性。如何进一步提高算法的执行效率,将是下一步需要研究的课题。

参 考 文 献:

[1] 王 珏,周志华,周傲英.机器学习及其应用[M].北京:清华大学出版社,2006.

[2] Gen M,Li Y. Spanning tree-based genetic algorithm for bicriteria fixed charge transportation problem[C]. Proc. of the Congress on Evolutionary Computation. Washington: IEEE Press,1999.

[3] Gen Mitsuo, Cheng Runwei. Genetic algorithms and engineering optimization[M]. San Francisco: John Wiley & Sons,Inc,1999.

[4] Das I,Dennis J E.A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems[J]. Structural Optimization,1997,14(1):63-69.

[5] 公茂果,焦李成,杨咚咚.进化多目标优化算法研究[J].软件学报,2009,20(2):271-289.

[6] Zheng Jinhua. Multi-objective evolutionary computations and their applications[M]. Science Press,2007.

[7] 周洪祥.灾害性天气的预测方法[M].北京:气象出版社,2002.

[8] Bendi V R,Prasad Babu M S,Venkateswarlu N B.A critical study of selected classification algorithms for Liver Disease Diagnosis[J]. International Journal of Database Management Systems(IJDMS),2011,3(2):101-114.

[9] Michael L,Raymer,Travis E,et al.Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm[J]. IEEE Transactions on Systems,Man,and Cybernetics,2003(2):33.

[10] Li Jinyan, Wong Limsoon. Using rules to analyse biomedical data:a comparison between C4. 5 and PC[J]. Lecture Notes in Computer Science,2003,2762:254-265.

[11] Elena Smirnova, Ida G, Sprinkhuizen Kuyper, et al. Unanimous voting using support vector machines[EB/OL]. http://www. personeel. unimaas. nl/Smirnov/papers/bnaic04.pdf,2004.

Research and Application of Feature Extraction and Multi-objective Machine Learning

HE Tao, *ZHANG Hong-wei*, *ZOU Shu-rong*

(College of Computer, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: A feature extraction and multi-objective machine learning algorithm is proposed based on multi-objective coevolutionary algorithm. Training samples core attributes are found by feature extraction and attribute groups are composed of core attributes and non-core attributes, so the classified accuracy is improved. All attribute groups are supervised clustering by attribute similarity and class tags. The number and center of class families can automatically determined by using the fitness function in machine learning as the goal; in this way, they can avoid the effect of subjective factors and the two key elements owning optimization nature are guaranteed. The class tag using the nearest neighbor method determines a genus of the unclassified samples. At last, the algorithm is demonstrated by the UCI data sets of Liver Disorders, Hepatitis data sets and summery abnormal megathermal forecast in the north area of Zhejiang province. The experiment results indicate that feature extraction and multi-objective machine learning algorithm is better than the well-known NBC, C4. 5 and SVM.

Key words: feature extraction; core attributes; machine learning; multi-objective EA; supervised clustering