文章编号: 1004－5422(2013)01－0058－04

多目标监督聚类GA 研究

索 飞, 张洪伟, 邹书蓉

(成都信息工程学院 计算机学院, 四川 成都 610225)

摘 要: 提出了多目标监督聚类GA 算法,即: 根据样本的类标签有监督地将样本聚类,在每个类中根据样本属性的相似性有监督地聚成类簇.如果分属不同类标签的类簇出现相交,则相交类簇再次聚类,直到所有类簇均不相交.适应度矢量函数由类簇数和类内距离2个目标确定,类簇数和类簇中心由目标函数自动确定,从而类簇数和中心就不受主观因素的影响,并且保证了这2个关键要素的优化性质.预测分类时,删去单点类簇,并根据类簇号和离某个类簇中心距离的最近邻法则以及该类簇的类标签进行分类.算法模型采用C#实现,采用3个UCI 数据集进行实例分析,实验结果表明,本算法优于著名的Native Bayes、Boost C4.5 和KNN 算法.

关键词: 多目标GA; 监督聚类; 类标签; 最近邻法则

中图分类号: TP301.6 文献标志码: A

0 引 言

近年来,聚类已经成为数据挖掘领域中一个热门的研究课题[1].一方面,它可以作为一个专门工具来处理数据分布信息;另一方面,也可以作为数据挖掘算法的一个预处理步骤[2].监督聚类分析是聚类分析的一种,它根据样本的先验信息或假设来决定样本的分类,据此建立判别模型,并利用该判别模型对未知对象进行分类.

遗传算法(Genetic Algorithm, GA)是一类借鉴生物界的进化规律演化而来的随机化搜索方法.它由美国J.Holland 教授首先提出.经过多年的研究、改进,遗传算法已经被广泛地应用于组合优化、机器学习及人工生命等领域.现阶段,科研人员仍在不断地研究、改进和拓展遗传算法的应用领域[3－4].

本研究提出了多目标监督聚类GA 算法,该算法主要有以下特点:多目标遗传算法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[5－6],从而解决了难以搜索到非凸解的问题[7].由于适应度矢量函数可自动确定类簇数和类簇中心,而不受主观因素的影响,从而提高预报的可靠性.为减少数据噪声对学习的影响,采取样本归一化处理的方法来提高学习的准确性,并在遗传优化过程中,采用有效的保忧策略以提高收敛速度和泛化能力.

1 多目标监督聚类GA 算法模型

1.1 算法描述

算法的基本思路是:对样本归一化处理,根据样本的类标签有监督地将样本分类,在每个类中按样本属性的相似性有监督地聚成类簇,如果分属不同类标签的类簇出现相交,则相交类簇再次聚类,直到所有类簇均不相交,样本的类簇号构成染色体及种群;根据类簇数最少化及类内距离最小化原则构造适应度矢量函数;利用遗传算法全局寻优的特点对样本进行多目标优化,并找到较好的染色体;进行预测分类时,删去单点类簇,并根据最近邻法则用较优染色体及类标签进行分类.

1.2 算法体系结构

1.2.1 样本归一化.

设第i个样本为, $x_i=(x_{i1}, \cdots, x_{im})$, $x_{ij}\in R$.为了更准确地学习样本,对所有样本进行归一化处理,

$$x'_{ij}=[\max(\{x_{i1}, x_{i2}\cdots x_{im}\})-x_{ij}]/[\max(\{x_{i1}, x_{i2}\cdots x_{im}\})-\min(\{x_{i1}, x_{i2}\cdots x_{im}\})] \quad (1)$$

其中, m为样本属性的个数, 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.经过处理后,所有样本的值都对应到[0, 1]区间.

1.2.2 编码与解码.

1) 编码.输入样本集S,总数为N,可能被分成

收稿日期: 2013－01－21.

作者简介: 索 飞(1988－),男,硕士研究生,从事计算机智能计算技术研究.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的类簇数为 c，设第 i 个样本 x_i 被唯一指定在第 k_i 个类簇中，$x_i \in C_{k_i}$ 则，$S = \bigcup_{1 \leqslant k_i \leqslant c} C_{k_i}$，从而可定义染色体 e 为，

$$(x_1, x_2, \cdots, x_N) \rightarrow e = (k_1, k_2, \cdots, k_N),\ k_i \in \{1, 2, \cdots, c\} \quad (2)$$

其中，基因 k_i 表示第 i 个样本被指定在第 k_i 类簇，染色体 e 即表示这些样本被唯一指定属于某个类簇.

2）解码. 若已知染色体，

$$e = (k_1, k_2, \cdots, k_N) \rightarrow (x_1, x_2, \cdots, x_N) \quad (3)$$

则将样本 x_i 指定属于第 k_i 个类簇 C_{k_i}，并且确定了类簇数 c，即，

$$S(e) = \bigcup_{1 \leqslant k \leqslant c} C_k,\ C_k = \{x_i^k\} \quad (4)$$

1.2.3　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，

$$d(x_i, x_k) = \sum_{j=1}^{m} (x_{ij} - x_{kj})^2 \quad (5)$$

根据类簇数最少和类内距离最小原则定义 2 维适应度矢量函数，

$$Z(e) = \min(c, \sum_{x_i^r \in C_r} d(x_i^r, s_r)),$$
$$s_r = \frac{1}{n_r}(\sum_{i=1}^{n_r} x_{i1}^r, \sum_{i=1}^{n_r} x_{i2}^r, \cdots, \sum_{i=1}^{n_r} x_{im}^r) \quad (6)$$

其中，$x_i^r \in C_r$ 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 是类簇数.

1.2.4　擂台选择法.

本研究采用擂台法（Arena' s Principle，AP）作为选择评价算子[9-10]，其过程为，

1）令 $E = \varnothing$，E 中存放偏序比较后较优的染色体；

2）目标函数集 $Z = \{Z_i = (Z_1^i, Z_2^i)\}$；

3）若 $Z \neq \varnothing$，从 Z 中选出 Z_i 作为擂台主，重复步骤 ① 和 ②，① 从 Z 中选出另一个 Z_j 与 Z_i 做偏序比较，若 Z_j 优于 Z_i，则从 Z 中删除 Z_i，并令 Z_j 作为新擂主，② 若擂台主与 Z 中其他元素比较遍历了 Z 后，将擂台主所对应的染色体增加到 E 中，从 Z 中删除擂台主，返回到 3）；

4）输出 E.

1.2.5　保优策略.

将经过 AP 算法选择后的较优染色体加入到记忆池中，迭代结束后，对记忆池中的染色体再次进行 AP 选择，最终得到的即为经过优化的染色体. 这样做既保留了祖先的优良基因，又经过了全局选择，即不会因为保优而陷入局部最优解[11].

1.2.6　分类方法.

进行预测分类时，删除单点类簇. 由最优染色体计算各类簇中心，根据最近邻法则分类，

$$r_0 = \arg\min_r \{d(x_i, s_r) = \sum_{j=1}^{m} (x_{ij} - s_{rj})^2\} \quad (7)$$

若输入的样本离聚类中心 S_r 的距离最近，则该样本的类簇号和类标签与第 r 类的相同.

1.3　算法步骤

多目标监督聚类 GA 算法具体步骤为：

1）初始化参数，样本归一化处理；

2）根据样本的类标签有监督地将样本分类，在每个类中按样本属性的相似性有监督地聚成类簇，若分属不同类标签的类簇不存在相交，则转到 6）；

3）记录出现相交的类簇的类簇号；

4）将出现相交的类簇再次聚类成多个类簇；

5）将聚好的新类簇再次进行有无相交的判断. 若存在相交，则返回 3），进行再次聚类，直到所有类簇均不存在相交的情况；

6）样本的类簇号构成染色体及种群；

7）计算各染色体的适应度矢量函数值；

8）利用 AP 算法对种群进行选择，被选择出染色体称为父染色体，并将其加入到记忆池中；

9）父染色体进行交叉和变异生成新染色体，其中类别相同的基因进行交叉变异；

10）若新生成的染色体数量小于种群的数量，则返回步骤 2）生成染色体，共同作为下一代种群；

11）若满足结束条件，则对记忆池中的染色体求适应度，并用 AP 算法做选择，输出较优染色体，否则转到 7）.

2　仿真实验分析

2.1　仿真实验数据与结果

为了验证本算法的可行性及有效性，选用实验平台 Windows XP，C ♯语言编程环境，采用国际上专门用来测试机器学习和数据挖掘算法的标准 UCI 数据集中使用频繁的 Iris、Echocardiogram 和 Post-operative-patient 3 组数据集作为测试数据. 3 组数据集的简单描述如表 1 所示.

表 1　数据集的简单描述

数据集	样本个数	连续数值属性	离散数值属性	类数
Iris	150	4	0	3
Echocardiogram	132	12	0	2
Post-operative-patient	90	1	7	3

为了计算分类的准确率，将整个数据集均分为

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

10 份进行交叉验证, 即 9 个子数据集作为训练样本, 一个子数据集作为预测样本, 每个子数据集轮流作为预测样本, 从而保证了对整个数据集的分类.

经过多次试验, 最后确定算法参数如下, 种群规模为 50, 遗传代数为 1000, 交叉概率为 0.8, 变异概率为 0.2.

将算法进行交叉验证, Iris 样本分类结果的平均正确率为 96.66%. 本算法与 Native Bayes、Boost C4.5 和 Bayesian Network (K2)算法的平均分类正确率相比[12], 结果如表 2 所示.

表 2 平均准确率比较表

数据集	Native Bayes	Boost C4.5	Bayesian Netwoik	本文算法
Iris	95.53%	94.33%	93.20%	96.66%

将算法进行交叉验证, Echocardiogram 样本分类结果的平均正确率为 72.73%. 本算法与 C4.5、CN2 和 KNN 算法的平均分类正确率相比[13], 结果如表 3 所示.

表 3 平均准确率比较表

数据集	C4.5	CN2	KNN	本文算法
Echocardiogram	66.64%	63.37%	71.81%	72.73%

将算法进行交叉验证, Post-operative-patient 样本分类结果的平均正确率为 74.17%. 本算法与 Native Bayes、9-NN 和 9-NN^2 算法的平均分类正确率相比[14], 结果如表 4 所示.

表 4 平均准确率比较表

数据集	Native Bayes	9-NN	9-NN^2	本文算法
Post-operative-patient	67.8%	71.1%	68.9%	74.17%

2.2 实验结果分析

通过对 Iris、Echocardiogram 和 Post-operative-patient 3 组数据集进行分类, 由表 2、3、4 可知, 本算法在平均准确率上有了一定的提高. 其中, 在 Iris 数据集中, 本算法的平均准确率比 3 种算法中最高平均准确率提升了 1.13%; 在 Echocardiogram 数据集中, 本算法的平均准确率比 3 种算法中最高平均准确率提升了 0.92%; 在 Post-operative-patient 数据集中, 本算法的平均准确率比 3 种算法中最高平均准确率提升了 3.07%. 同时, 由这 3 组数据集可以说明, 无论样本属性为连续型或者离散型, 本算法都可以有效地进行分类, 此表明了本算法的有效性.

3 结 语

本研究提出的基于多目标监督聚类 GA 算法利用遗传算法全局寻优的特点, 对整个样本空间按属性的相似性和类标签进行分类, 使得具有相似规则知识的样本被划分为同一类簇. 实验结果表明, 算法有较高的准确性和有效性.

参考文献:

[1] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48—61.

[2] 周炜奔, 石跃祥. 基于密度的 K-means 聚类中心选取的优化算法[J]. 计算机应用研究, 2012, 29(5): 1726—1728.

[3] 郭志刚, 邹书蓉. 基于非劣排序的多目标优化免疫遗传算法[J]. 成都: 成都信息工程学院学报, 2012, 27(2): 136—141.

[4] 黄晓滨, 邹书蓉, 张洪伟. 免疫遗传算法及其在 VRP 中的应用[J]. 成都: 成都信息工程学院学报, 2008, 23(6): 637—641.

[5] Gen M, Li Y Z. *Spanning tree-based genetic algorithm for bicriteria fixed charge transportation problem*[C]// *Proceedings of the* 1999 *Congress on Evolutionary Computation*. Washington, DC: IEEE Press, 1999: 2265—2271.

[6] Gen Mitsuo, Cheng Runwei. *Genetic algorithms and engineering optimization*[M]. New York: John Wiley &sons, Inc., 1999.

[7] Das Indraned. *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.

[8] Zhang Hongwei, Yang Zhenyu, Zou Shurong. *Multi-objective supervised clustering GA and megathermal climate forecast*[C]// *IEEE* 2011 *International Conference on Management and Service Science*. Wuhan: IEEE Press, 2011: 20—23.

[9] Zhang Hongwei, Cui Xiaoke, Zou Shurong. *Multi objective transportation optimization based on fmcica*[C]// *The* 2 *nd IEEE International Conference on Information Management and Engineering*. Chengdu: IEEE Press, 2010: 426—430.

[10] 郑金华. 多目标进化算法及其应用[M]. 北京: 科学出版社, 2007.

[11] 雷德明, 严新平. 多目标智能优化算法及其应用[M]. 北京: 科学出版社, 2009.

[12] Kotsiantis S B, Pintelas P E. *Logitboost of simple bayesian classifier*[J]. Informatica, 2005, 29: 53—59.

[13] Todorovski L, Dzeroski S. *Experiments in meta-level learning with ILP*[C]// *Third European Conference, PKDD*' 99. Pragve: Springer Berlin Herdelberg, 1999: 98—106.

[14] Petri K, Jussi L, Petri M, et al. *Unsupervised bayesian visualization of high-dimensional data*[C]// *Proceedings of the* 6*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston: ACM Press, 2000: 325—329.

(下转第 63 页)

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

与 Phase-Toggle 法的特点，通过旋转被测支路的移相器，实现了移相器的全移相状态参与校准，同时其他支路的移相器不参与移相，减小了移相误差对校准的影响，扩大了系统校准时对支路的动态要求，非常适合目前相控阵天线小型化、集成化及共形化的工程应用要求.

参考文献：

[1] Lee J J, Ferren E M, Woollen D P, et al. *Near-field probe used as a diagnostic tool to locate defective elements in an array antenna* [J]. Antennas and Propagation, IEEE Transactions, 1988, 36(6): 884—889.

[2] Davis D. *Analysis of array antenna patterns during test*[J]. Microwave Journal, 1978, 21: 67—72.

[3] Shnitkin H. *Rapid fast fourier transform phase alignment of an electronically scanned antenna* [C]// *20th European Microwave Conference*. Budapest: IEEE Press, 1990: 247—257.

[4] 彭祥龙，石星，刘茁. 基于行波耦合馈线的相控阵幅相校正算法比较[C]//第十届全国雷达学术年会论文集. 北京：国防工业出版社，2008：786—788.

[5] 杨顺平，张云，温剑. 接收机噪声对行波馈电法校准误差影响分析[C]//2009 年全国天线年会. 北京：电子工业出版社，2009：778—780.

Calibration Method of Phased Array Based on Mean of Vector

YANG Shunping

(Southwest China Institute of Electronic Technology, Chengdu 610036, China)

Abstract: A new phased array calibration method-mean of vector is proposed in this paper. The method is verified on a 8-element linear experimental array. A higher accuracy has been obtained than fast Fourier transform (FFT) method under the loss of transmission path with large dynamic range.

Key words: fast Fourier transform; calibration; phased array

(上接第 60 页)

Research of Multi-objective Supervised Clustering GA

SUO Fei, ZHANG Hongwei, ZOU Shurong

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

Abstract: This paper presents a new multi-objective supervised clustering genetic algorithm. Samples are supervisedly clustered into several classes by class labels. In each class, samples are supervisedly clustered into class clusters according to the similarity of the sample properties. If the class clusters which belong to different class labels intersect, these intersecting class clusters are clustered again into class clusters until all the class clusters don't intersect. The fitness vector function is determined by the number of class clusters and within-class distance. The number and center of class clusters can be determined automatically by using the fitness vector function. The two key elements can be unaffected by subjective factors and have optimization natures. During classification forcast, the single-point class cluster is deleted and then classification is done according to the class cluster number, the nearest neighbor rule and the class labels. The algorithm model is implemented with C #, using three UCI data sets as the experiment data. The experimental results indicate that this algorithm is better than Native Bayes, Boost C4.5 and KNN algorithms.

Key words: multi-objective GA; supervised clustering; class label; nearest neighbor rule

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net