# 基于佳点集遗传算法的特征选择方法

贾瑞玉,宁再早,耿锦威,查 丰
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

**摘 要**:针对特征选择中降维效果与分类精度间的矛盾,通过分析传统的特征选择方法中的优点和不足,结合佳点集遗传算法的思想和K最近邻简单有效的分类特性,提出了基于佳点集遗传算法的特征选择方法。该算法对特征子集采用佳点集遗传算法进行随机搜索,并采用K近邻的分类错误率作为评价指标、淘汰不好的特征子集,保存较优的特征子集。通过实验比较看出,该算法可以有效地找出具有较高分类精度的特征子集,降维效果良好,具有较好的特征子集选择能力。

**关键词**:K最近邻算法;特征选择;佳点集遗传算法
**中图分类号**:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2011)01-0050-03

# Feature Selection Method Based on Good Point-Set Genetic Algorithm

JIA Rui-yu, NING Zai-zao, GENG Jin-wei, ZHA Feng
(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

**Abstract**: To address the contradiction between the dimension reduction for feature selection and the precision of classification, by analyzing the strengths and weaknesses of the traditional feature selection method, combines the idea of good point-set genetic algorithm and the simple and effective features of K nearest neighbor classification, presents a new feature selection method based on good point set genetic algorithms. Through a random search of the feature subset with the good point-set genetic algorithm, and using K nearest neighbor classification error rate as the evaluation index, eliminate the bad feature subset, save the optimum feature subset. It can be seen through the comparison experiments that the algorithm can effectively find out those feature subset which has high classification accuracy, and the effect of dimension reduction is good, these show that the algorithm has the better ability to select feature subset.

**Key words**: K-nearest neighbor algorithm; feature selection; good point-set genetic algorithm

## 0 引 言

特征选择是在数据挖掘和模式识别中数据预处理的重要方法之一。原始数据中通常存在着不相关或冗余的特征,特征选择的目的就是在保证处理后所得的数据的数据类的概率分布尽可能和原分布接近的情况下,删除一部分特征,从而减少分类系统的代价和运行时间。特征选择方法根据其是否依赖于机器学习分为filter型和wrapper型两类,filter型的特征选择方法具有计算代价小,效率高但降维效果一般等特点[1],其代表模型有Focus和Relief;wrapper型特征选择算法将归纳算法封装于特征选择算法中,降维效果好,但存在计算代价大,效率低的不足,如文献[2]中采用类间模糊距离和类内的模糊距离的差作为适应度来度量所选择的特征子集的分类能力,不同模式的欧式距离计算量大,并且训练时间较长。文献[3]中提出以不一致标准作为特征子集的评价函数,采用拉斯维加斯(Las-Vegas)算法找出满足可接受的不一致性比例的特征集合。文献[4]中作者把以上两种基于距离度量作为评价函数的方法及基于一致性度量的度量的评价方法和基于分类精度的评价函数进行对比,指出前者评价函数无法反映精确程度,而这一点在特征选择方法中很重要。文献[5~7]对多种特征选择的方法进行比较,如开方拟和检验(CHI)、文档频率(DF)、信息增益(IF)、互信息(MI)、术语强度(TS)等,并通过实验得出CHI、IF和DF的性能较优,文献[8]指出特征选择的任务是求出一组对分类最有效的特征,如何衡量特征对分类的有效性,文献[9]分析指出特征与类别之间服从符合有一阶自由度的$\chi^2$分布,文献[10,11]采用CHI统计方法度量两者之间的相关程度,选出最优的特征,以覆盖算法的分类准确率作为评价函数,这种方法存在特征选择后的样本形成覆盖的难易程度问题。从优化的角度来说,特征选择是一个组合优化及

**收稿日期**:2010-05-15;**修回日期**:2010-08-13
**基金项目**:安徽省高等学校省级自然科学基金(KJ2008B092)
**作者简介**:贾瑞玉(1965-),女,副教授,研究方向为计算机图形学、数据挖掘、人工智能。

多目标优化的问题,解决这类问题的常规方法有遍历搜索、随机搜索,以及启发式搜索,而遗传算法属于随机搜索方法。文献[12]指出遗传算法的本质是一个具有定向制导的随机搜索,其制导的原则是导向以高适度为模式为祖先的"家族"方向,并提出一种在"高适度模式为祖先"的"家族"方向上搜索更好样本的改进遗传算法——佳点集遗传算法,提高了传统遗传算法的效率,文中结合这种佳点集遗传算法的思想,以 K 近邻算法计算适应度,提出一个改进的特征选择算法。

## 1　K 近邻算法

K 近邻算法是一种基于实例的学习法,也称为惰性学习法。这种学习方法当给定训练元组时,只是简单存储它,不构造分类模型,只有当给定一个检验元组时,它才根据该检验元组和存储的元组的相似性进行分类。

K 近邻算法通过找出与检验元组最"邻近"的 $K$ 个元组,并根据这 $K$ 个元组类别信息对 $K$ 个元组进行分组,对检验元组的类别,指派到这 $K$ 个最近邻中的多数类,这种"邻近"性用距离度量,文中采用欧式距离,如式(1)所示,$X_1, X_2$ 分别代表 $L$ 维空间的两个元组:$X_1 = (x_{11}, x_{12}, \cdots, x_{1L}), X_2 = (x_{21}, x_{22}, \cdots, x_{2L})$

$$\mathrm{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{L} (x_{1i} - x_{2i})^2} \tag{1}$$

对检验元组 $X$,找出 $K$ 个与之最邻近的训练元组,然后基于一定的投票机制决定该检验元组的类别。文中以 K 最邻近算法作为分类器,计算特征子集的分类准确率,以此作为特征子集的适应度。其过程可以描述为:首先从数据集 $D$ 的属性集中,生成特征子集 $F$,根据特征子集 $F$ 从数据集 $D$ 取出对应的数据集 $D_f$,最后采用 EvaluateGene 计算 $D_f$ 的准确率,作为特征子集 $F$ 的适应度。

算法 1:EvaluateGene($F$,$K$)

输入:特征子集 $F$,最近邻数 $K$

输出:特征子集的评估值

步骤:

1) 从根据特征子集 $F$ 从存储的数据集 $D$ 中取出数据集 $D_f$;

2) 采用 K 最近邻算法对数据集 $D_f$ 进行分类,统计分类的正确率 Pre;

3) 保存分类的正确率 Pre 作为特征子集 $F$ 的评价指标,即适应度。

## 2　佳点集遗传特征选择

### 2.1　佳点集交叉算法

标准的遗传算法以定向制导的原则(即导向以高适应度模式为祖先的"家族"方向)随机搜索适应值高的后代,而其交叉算法(如单点交叉,多点交叉)都只能保证交叉后的后代是落在"家族"中,却无法保证交叉后的后代具有较高的适应值,而佳点集交叉算法能做到这一点。用佳点集来进行近似积分,误差的阶只与样本数有关,而与维数无关,这对佳点集遗传算法用于高维的近似计算是个很好的优势,不止如此,用佳点集方法取一定数的点,比随机取的点偏差少很多[5],这样佳点集遗传算法的收敛速度更快。令数据集 $D$ 由 $N$ 个 $L$ 维的特征子集组成,即 $D = \{X^1, X^2, \cdots, X^N\}$,$X^i = \{x_1^i, x_2^i, \cdots, x_L^i\}$,$1 \le i \le N$,赌轮法从某代数据集 $D$ 选择中选择两个特征子集:第 $a$ 个染色体 $X^a$ 和第 $b$ 个染色体 $X^b$ 进行佳点操作。

令 $X^a = \{x_1^a, x_2^a, \cdots, x_L^a\}$,$X^b = \{x_1^b, x_2^b, \cdots, x_L^b\}$。

令 $J = \{i \mid X_i^a \ne X_i^b, 1 \le i \le L\}$,$J$ 的大小 $|J| = t$。

$X^a$ 和 $X^b$ 交叉后的后代中第 $i$ 个染色体的 $A^i = \{a_1^i, a_2^i, \cdots, a_L^i\}$,其中,当 $m \notin J$ 时,$a_m^i = x_m^1$;而当 $m \in J$ 时,$a_m^i = \langle r_m \times i \rangle$,$1 \le m \le L$,$r_m = \left\{2\cos\frac{2\pi m}{p}\right\}$,$\{a\}$ 表示 $a$ 的小数部分,$p$ 是满足 $p \ge 2t + 3$ 的最小素数,$\langle b \rangle$ 表示,如果 $b$ 的小数部分小于 0.5,则 $\langle b \rangle = 0$,否则 $\langle b \rangle = 1$。

### 2.2　佳点集遗传特征选择算法

算法 2:基于佳点集的遗传特征选择算法,简记为 GGaKNN 算法

step1:读取样本数据,对样本数据进行归一化处理,采用 10-交叉试验,把样本分为训练集和检验集;

step2:个体采用二进制编码方式,原始特数为 $L$,则编码的长度为 $L$,个体每一个二进制基因位对应于相应次序的特征,当个体的某一基因为 1 时,该基因对应的特征项选中,初始化种群,种群数 $N = 50$,交叉概率 $p_c = 0.8$,变异概率 $p_v = 0.005$,迭代次数 $T = 50$;

step3:K 最近邻算法计算检验集中每个个体 $A_i$ 的适应度 $f_i$,$1 \le i \le N$;

step3.1 取 $A_i$ 中检验集中每个个体 $X_i$,计算训练集中每个个体与 $A_i$ 训练集中样本的欧式距离集 DistCol;

step3.2 从小到大冒泡排序 DistCol,取前 $K$ 个个体;

step3.3 对 $K$ 个个体按类别分类,并计算每个类别的样本数,样本数最大的类别为个体 $A_i$ 的类别;$K = 1$,取第一个个体的类别,判断是否和 $A_i$ 类别一到,若是计数器 Count 加 1;

step3.4 重复以上步骤,对 $A_i$ 中检验集中每个个体 $X_i$ 的类别作出判断,最后 Count 占检验集的比例为 $f_i$ 的值;

step4：以概率 $\mathrm{rel}_i = f_i / \sum_{i=1}^{N} f_i$ 复制个体 $A_i$，复制个体的数目为 $N_i = \mathrm{round}(\mathrm{rel}_i \times N)$ ；round($a$) 表示与 $a$ 距离最小的整数，$N_i = 0$ 的个体被淘汰；

step5：赌轮法选择两个个体 $X_a, X_b$，以概率 $p_c$ 进行佳点集交叉操作；

step5.1 记 $p_i = f_i / \sum_{i=1}^{N} f_i$，随机生成一个[0,1] 内的随机数 $r$；

step5.2 若 $p_1 + p_2 \cdots + p_{i-1} < r \leq p_1 + p_2 + p_3 + \cdots + p_i$，则选择个体 $i$；

step5.3 使用 step5.1，step5.2 两步选择第 $a$ 个染色体 $X_a$ 和对第 $b$ 个染色体 $X_b$；

step5.4 取 $J = \{i \mid X_i^a \neq X_i^b, 1 \leq i \leq L\}$ ，$J$ 的大小 $|J| = t$。$X^a$ 和 $X^b$ 交叉后的后代中第 $i(i = \{a,b\})$ 个染色体的 $A^i = \{a_1^i, a_2^i \cdots, a_L^i\}$ ，其中，当 $m \notin J$ 时，$a_m^i = X_m^1$；而当 $m \in J$ 时，$a_m^i = \langle r_m \times i \rangle, 1 \leq m \leq L, r_m = \left\{2\cos\frac{2\pi m}{p}\right\}$，$\{a\}$ 表示 $a$ 的小数部分，$p$ 是满足 $p \geq 2t + 3$ 的最小素数，$\langle b \rangle$ 表示，如果 $b$ 的小数部分小于 0.5，则 $\langle b \rangle = 0$，否则 $\langle b \rangle = 1$。

step5.5 计算交叉后个体的适应度；

step6：以概率 $p_v$ 进行变异遗传操作；

step7：对经过遗传操作后的染色体放到染色体池中，对新得到的染色体计算适应值，将适应度小的染色体从繁殖池中删除；

step8：重复 step4 ~ step7 直至达到迭代次数 $T$，取得迭代 $T$ 次的适应度最大的染色体，即为所求的个体；

## 3 实 验

为了评估文中算法的有效性，采用 6 组 uci 公用数据集进行验证，数据集的相关信息如表 1 所示。

表 1 实验中用到的 uci 数据集

| 数据集 | 特征数 | 样本数 | 类别数 |
|---|---|---|---|
| diabetes | 8 | 768 | 2 |
| wine | 13 | 178 | 3 |
| vehitle | 19 | 846 | 4 |
| sonar | 60 | 208 | 2 |
| ionosphere | 34 | 351 | 2 |
| iris | 4 | 150 | 3 |

### 3.1 实验结果

实验采用 10-交叉试验，文中提出的佳点集遗传特征方法，记为 GGaKNN，其中 $K = 5$，执行 10 次，实验结果取平均值，与文献[13]提出的 PsoKnn 算法、和 Weka 下实现的决策树的特征选择方法（简记为 C4.5）进行对比。如表 2 所示，表中第一列括号中的数值代表数据集中的特征数目，第 2、3、4 列括号中的数值代表经过特征选择后的特征数目。

表 2 实验数据比较

| 数据集 | PsoKnn | C4.5 | GGaKNN |
|---|---|---|---|
| Diabetes(8) | 76.45%(6) | 75%(4) | 77.53%(4) |
| Wine(13) | 94.12%(9) | 93.82%(11) | 98.95%(8) |
| Vehitle(18) | 74.52%(11) | 68.44%(11) | 76.82%(9) |
| Sonar(60) | 64%(28) | 73.56%(19) | 94.29%(30) |
| Ionosphere(34) | 88%(22) | 89.4%(15) | 90.28%(16) |
| Iris(4) | 96.67%(3) | 96%(2) | 98.67%(2) |

### 3.2 实验结果分析

GGaKNN 算法以分类的精度为适应度，在高适应度模式为祖先的“家族”中，通过以佳点集交叉操作保证导向的后代有较高的适应度。从表 2 实验的结果来看，在相同的数据集下，GGaKNN 最后搜索出的最优特征子集在分类精度高于 C4.5，PsoKnn 方法，在删除冗余和不相关特征上效果较好，体现该算法有较强的特征选择能力。

## 4 结束语

文中在对基于遗传算法的特征方法的基础上，引入佳点集的思想的，同时采用 KNN 特征子集评价方法，通过与 PsoKnn，C4.5 的实验比较分析表明，该方法能搜索出较优特征子集，获得较好的分类精度。GGaKNN 算法是佳点集遗传与 KNN 结合的一次尝试，随着进化计算在数据挖掘领域的进一步广泛运用，今后可进一步研究佳点集遗传在特征权重学习上的应用，以及佳点集遗传与其它算法的结合。

**参考文献：**

[1] 任江涛，孙吴昊，黄焕宇，等. 一种基于信息增益及遗传算法的特征选择算法[J]. 计算机科学，2006，33(10)：193-25.

[2] 刘素华，侯惠芳，李小霞. 基于遗传算法和模拟退火算法的特征选择方法[J]. 计算机工程，2005，31(6)：157-159.

[3] Liu H，Setiono R. A Probabilistic Approach to Feature Selection - A Filter Solution[C]//In：Proceddings of International Conference On Machine Learning. [s. l. ]：[s. n. ]，1996：319-327.

[4] 孙 雷，王 新. 一种基于遗传操作和类内类间距离判据理论的特征选择方法[J]. 计算机工程与应用，2004(21)：178-181.

[5] Yang Y M，Pederson J O. A comparative Study on Feature Selection in Text Categorization [C]//In：Proceeding of the Fourteenth International Conference of Machine Learning (ICML'97). [s. l. ]：[s. n. ]，1997：412-420.

（下转第 57 页）

OWL-S 描述的服务本体分别转化为规划器可以执行的问题描述文件和领域描述文件,以使 JSHOP2 可处理。而且在使用规划算法的基础上同时还提供将组合的服务经过 OWL-S 语义标注后进行发布,以增强本系统的组合能力。

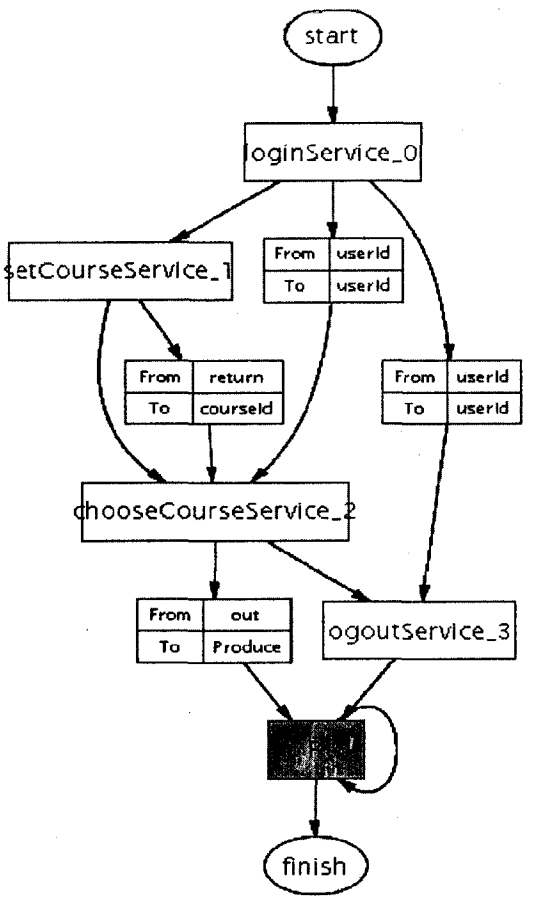


图 2　语义 Web 服务组合器运行结果

本系统在服务匹配过程中未考虑 Web 服务的非功能 Qos 因素,比如服务费用、服务的响应时间、服务的可用性等。它可以与领域本体匹配算法相结合,在首次匹配结果为 Exact 或 Subsume 的情况下,通过 Qos 约束,可获得更准确的、符合用户需求的 Web 服务。这是本课题的下一步研究工作。

参考文献:

[1] W3C. Semantic Web Activity [EB/OL]. 2006. http://www.w3.org/2001/sw/.

[2] Liu Xingwei, Zhao Hui. An AI Planning Based Approach for Automated Web Services Composition[C]//The 2007 International Conference Proceedings on Intelligent Systems and Knowledge Engineering (ISKE2007). Chengdu: [s. n.], 2007.

[3] Koehler J, Srivastava B. Web service composition: current solutions and open problems[C]//The 2003 International Conference on Automated Planning and Scheduling(ICAPS'03). Trento, Italy:[s. n.],2003.

[4] Skogan D, Gronmo R, Solheim I. Web service composition in UMLm[C]//The 8th International IEEE Enterprise Distributed Object Computing Conference (EDOC). Monterey, California:[s. n.],2004.

[5] 李　鹏,战德臣,刘国忠,等. 一种面向用户的 Web 服务组装方法[J]. 计算机应用,2009,29(11):3120-3123.

[6] 吴善明,沈建京,韩　强. 基于领域本体和 OWL-S 的 Web 服务组合方法[J]. 计算机工程,2009,35(21):256-257.

[7] 郑娅峰,鱼　滨. 基于语义 Web 的动态组合服务关键技术研究[J]. 计算机工程与应用,2005(3):45-48.

[8] Martin D, Burstein M, Hobbs J, et al. OWL-S: Semantic Markup for Web Services[EB/OL]. 2004. http://www.w3.org/Submission/OWL-S.

[9] Sirin E, Parsia B, Wu Dan, et al. HTN Planning for Web Service Composition Using SHOP2[J]. Journal of Web Semantics, 2004(4):377-396.

[10] Narayanan, S, McIlraith S. Simulation, verication and automated composition of web services[C]//in: Proceedings of the Eleventh International World Wide Web Conference. Honolulu, Hawaii:[s. n.], 2002.

[11] 方其庆,彭晓明,刘庆华,等. 结合 AI 规划和工作流的动态服务组合框架研究[J]. 计算机科学. 2009, 36(9):110-114.

[12] Johnson R, Hoeller J, Arendsen A, et al. Professional Java Development with the Spring Framework [M]. Beijing: China Machine Press, 2006:233-238.

(上接第 52 页)

[6] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报,2003,18(1):26-32.

[7] 单松巍,冯是聪,李晓明. 几种典型特征选取方法在中文网页分类上的效果比较[J]. 计算机工程与应用,2003(22):146-148.

[8] 万　忠,张燕平,张　玲,等. 基于覆盖算法决策界的特征选择算法[J]. 计算机技术与发展,2006,16(4):84-87.

[9] Dunning T E. Accurate methods or the statistics of surprise and coincidence[J]. Computational Linguistics,1993,19(1):61-71.

[10] 闫　屹,张燕平,耿筱媛. 基于 CHI 值特征选取和覆盖的文本分类方法[J]. 计算机技术与发展,2008,18(5):79-85.

[11] 段　震,王倩倩,张燕平,等. 覆盖算法下文本分类特征选择的研究[J]. 计算机技术与发展,2008,18(11): 29-31.

[12] 张　铃,张　钹. 佳点集遗传算法[J]. 计算机学报,2001,24(9):1-9.

[13] 任江涛,卓晓岚,许盛灿,等. 基于 PSO 面向 K 近邻分类的特征权重学习算法[J]. 计算机科学,2007,34(5):187-189.