基于基因表达式编程的核k近邻分类算法

柳秋云，王翰虎
（贵州大学 计算机科学与技术学院，贵州 贵阳 550025）

摘 要：核k近邻分类算法在生物信息学和蛋白质结构预测等领域中的应用受到人们极大的关注。核函数在核k近邻分类算法的分类性能中起着重要的作用，如果核函数及其参数选择得当，则将获得较高的分类准确率。为了自动产生合适的核函数，提高分类的准确率，提出了一种基于基因表达式编程的核k近邻分类算法GEPKNN。该算法的基本思想是用基因表达式编程搜索与训练数据相关的核函数及其参数，在进化过程中用k折交叉验证评估个体的适应度。该算法克服了核k近邻算法的主观性和不确定性，能自动产生合适的核函数并提高分类的准确率。

关键词：数据挖掘；进化计算；基因表达式编程；核k近邻分类器
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1673-629X(2009)08-0019-04

A Kernel KNN Classifier Based on Gene Expression Programming

LIU Qiu-yun, WANG Han-hu
(School of Computer Science and Technology, Guizhou University, Guiyang 550025, China)

Abstract: The kernel KNN classifier becomes an attractive and interest topic in application of bioinformatics and protein structure prediction. Performence of the kernel KNN is strongly dependent on the kernel function. A better classification performance could be achieved by choosing the kernel function and its parameters carefully. Describes a kernel KNN classifier based on gene expression programming (GEPKNN), which adopts gene expression programming to search for any kernel function that is related to the training data. K cross-validation is used to assess the fitness values of the individuals in the current population. The method can automatically construct a proper kernel function and overcome the subjectivity and uncertainty of kernel KNN classifier, and the accuracy can be also raised.

Key words: data mining; evolution computation; gene expression programming; kernel KNN classifier

0 引 言

分类算法一直是数据挖掘和模式识别领域关注的焦点，在过去的研究工作中已经陆续提出许多精巧的算法，但是找出一种适用于所有数据的分类方法是不现实的。基于实例的方法虽是分类算法中较为古旧的一员，但由于这类算法的简单高效，所以直到现在它们仍在文本分类等领域中发挥着重要作用，这类方法的典型代表是k近邻算法(KNN)。

核k近邻分类器是对k近邻分类器的一种改进，它把核方法引入到KNN中，部分解决了KNN在高维空间中性能显著降低的问题，同时也继承了KNN的高效率。核KNN在生物信息学和蛋白质结构预测等领域中都发挥了重要的作用并取得了一些好的结果，近来受到人们极大的关注[1]。

核KNN的主要不足在于其分类性能对核函数的敏感性，使用恰当的核函数能显著提升核KNN的分类性能。选择核函数通常依赖于经验和直觉，带有强烈的主观随意性，且获得的核函数通常与问题空间无关，所以其分类性能一般不会是最优的。

为了降低选择核函数时的不确定性，提出了一种基于基因表达式编程的核KNN算法，简记为GEPKNN。算法的基本思路是利用GEP的函数空间搜索能力为核KNN自动构造与训练数据相关的核函数。

文中提出的方法克服了核KNN算法的主要不足并具有如下特点：

(1)自动化核函数及其参数的选择，消除了对专家经验和直觉的依赖，使核KNN的分类性能优势趋于稳定；

(2)实验表明GEPKNN的分类性能优于C4.5、KNN等传统算法，而且它继承了核KNN的优点，如结构简单、分类速度快，在高维空间上仍然保持较好的分类性能等；

收稿日期：2008-12-29；**修回日期**：2009-03-20
基金项目：贵阳市科技攻关项目(2006,16-6号)
作者简介：柳秋云(1981-)，男，硕士研究生，研究方向为数据挖掘技术；王翰虎，教授，CCF高级会员，主要研究方向为数据库技术和分布式系统。

(3)相对于 GEP 分类器和 SVM 而言,它能够直接处理多分类问题,不必引入“一对一”和“一对多”等复杂机制。

1 相关工作

1.1 核方法

核方法的基本思想是把数据通过非线性变换(通常称其为特征映射)投影到更高维甚至维度无穷的特征空间中,这个变换的结果是原空间中非线性可分的数据在特征空间中可能成为线性可分,或不同类之间的分类间隔被拉大,籍此提高分类算法的分类准确率。核技巧则使我们不需要知道具体的特征映射就能通过核函数求出特征空间的内积,从而避免了在高维空间中求内积引起的计算开销[2]。

为了便于后面的叙述,下面介绍核方法中几个重要的定理和定义。

定义 1(核函数) 对于空间 X 上的任何一个二元对称函数 k,如果它在 X 的任意 m 个点上的 Gram 矩阵是半正定的,则根据 Mercer 定理,k 是一个核函数,且存在特征映射 Φ,使得 $k(x,x') = <\Phi(x),\Phi(x')>$。

以下是几个在分类器设计中经常使用的核函数:

(1) 齐次多项式核函数

$k_1 = (<x_1,x_2>+1)^n$

(2) 高斯核函数

$k_2 = e^{\frac{(-1)*d(x_1,x_2)}{\sigma^2}}$

(3) $k_3 = (<x_1,x_2>+\beta)^{-0.5}$

根据定义 1 和欧氏距离的定义,容易证明定理 1。

定理 1(特征空间上的欧氏距离) 设 Φ 是从原空间 X 到特征空间 C 的特征映射,k 是其对应的核函数,则 X 中任意两点 x_1 和 x_2 的核距离为:

$$d(\phi(x_1),\phi(x_2)) = \sqrt{k(x_1,x_1)+k(x_2,x_2)-2*k(x_1,x_2)}$$

定理 2(核函数的封闭性) 令 k_1,k_2 是 $X\times X$ 上的核函数,其中 $X\subseteq R^n$,则

① $k_1(x,y)+k_2(x,y)$

② $k_1(x,y)*k_2(x,y)$

③ $e^{k_1(x,y)}$

④ $e^{k_2(x,y)}$

等函数都是核函数。

1.2 基因表达式编程

基因表达式编程(GEP)继承了遗传编程[3,4]和遗传算法[5]两者的优点。GEP 明确地将个体的基因型和表现型分离,个体的基因型是在遗传算法中常见的线性串(染色体),各种遗传算子直接作用在基因型上;个体的表现型则可以是任意复杂的树结构,如数学表达式、决策树甚至神经网络,它的作用主要体现在计算个体的适应度上。GEP 通过一个简单的映射规则(广度优先遍历)把基因型和表现型联系起来。相比 GP 而言,GEP 特殊的编码方式使其容许更多更复杂的遗传算子作用于其染色体,而不必拘泥于传统的交叉和变异算子[6~8]。

2 GEPKNN 分类器

本节描述所提出的 GEPKNN 算法,算法的基本思路是利用核函数的封闭性质(定理 2),依靠 GEP 在搜索复杂表达式空间方面的优势,为核 KNN 自动构造与数据相关的核函数,以期减小人工选择核函数的主观性,达到提升核 KNN 的分类性能的目的。

2.1 核 KNN 算法

核 KNN 与 KNN 的主要区别在于使用了不同的距离度量[1],如果将 KNN 所用的 X 上的欧氏距离替换为定理 1 中的核距离,就得到了一个核 KNN 分类器。

2.2 GEP 中的符号

在 GEPKNN 算法中,令终结符集合为 $\{k_1,k_2,k_3\}$,分别代表 1.1 节中的 3 个常用核函数,这里称其“基本核函数”。令非终结符集合为{+,*,exp},其中 exp 代表指数函数,它是一元函数。

这里需要对上面的终结符集合作一个说明。大多数文献中 GEP 的终结符集通常是输入值或常量,少见以函数作为终结符的报道。但是如果我们根据核函数的封闭性质,通过几个基本核函数构造出复杂的核函数,进而把构造出的复杂核函数转换为表达式树,可见基本核函数始终在树的叶节点上,因此 k_1,k_2,k_3 就是所需的最小的组成单元,把它们作为终结符处理是合理的。

实际上,如果把原空间 X 中的输入向量作为终结符,把 k_1,k_2,k_3 归入非终结符集合,就必须引入语法约束,强制最末一层内部节点为 k_1,k_2 或 k_3;这样本质上并无不同,还引入一些不灵活的限制。

2.3 编 码

由基本核函数构成的表达式和其中每个基本核函数的参数都必须编码在基因型中,以便同步进化。设计了一种特殊的基因型编码以达到上述要求:GEPKNN 的染色体分为两部分,第一部分是由头和尾组成的表达式域(EDOM),它编码核函数,紧接着的串存储出现在表达式域中的基本核函数的参数,这个串称为

参数域(PDOM),它的长度是 $p = t * m$,t 是尾部长度,m 是参数最多的基本核函数的参数数目。从基因型构造表现型时,按照基本核函数在染色体中出现的顺序把参数域中的参数逐个赋给基本核函数。为了使参数域与表达式域同时进化,借鉴 GEP 中表达常量的方法 —— 为每个染色体关联一个常数池,参数域中的各参数值随机地在常数池中选取。

综上可知,每条染色体的长度为:

$$h + (h * (n - 1) + 1) * (m + 1)$$

2.4 遗传算子

GEP 的大部分遗传算子可以不加改变地应用到 GEPKNN 中,需要注意的是应用遗传算子于染色体时必须保持 EDOM 和 PDOM 的边界,防止产生无效的后代;还要引入两个特殊的算子用于进化 PDOM 域:其中 PDInversion 算子转置 PDOM 中的随机子串,而 CPMutation 改变常数池中随机位置上的常数。

2.5 适应度函数

为了计算个体 I 的适应度,在给定的数据集上对核 KNN 进行 k 折交叉验证,采用 k 折交叉验证的优点是能有效防止核函数过拟合训练数据。其中折数 k 是可调整的参数。为了缩短训练时间,在算法中统一取 $k = 3$。

求个体适应度的具体过程是:个体 I 的两个域(表达式域和参数域)解码后组装成核函数 k,把 k 载入核 KNN 后就可以在其上作 k 折交叉验证了。假设经过 k 折交叉验证得到的平均错误率是 e,令个体的适应度为

$$\text{fitness} = 1000 * (1 - e)$$

求适应度函数时需要频繁计算核距离(定理 2),因此缩短计算核距离的时间是提高算法效率的一个有效手段。基于此提出了下面的定理 3,该定理很容易用数学归纳法证明。

定理 3(求核距离的快速方法) 令 $K = \{k_1, k_2, k_3\}$,其中 k_1、k_2、k_3 是上面提到的 3 个常用核函数,S 是 K 上的加法、乘法、指数运算等运算符的集合。k 是由 K 和 S 中的元素构成的任一符合语法规范的算术表达式,由定理 2 可知 k 是核函数。如果原空间 X 中的内积进行了规范化,则存在实数 SPID,对于 X 中任意点 x 都有 $k(x,x) = \text{SPID}$。

此时核距离的公式可以修改为

$$d(\phi(x_1),\phi(x_2)) = \sqrt{2 * (\text{SPID} - k(x_1,x_2))}$$

用上式计算核距离可以使求适应度的时间缩短约 2/3,同时通过检查 SPID 是否有效实数能快速检测出那些求值时会发生溢出的核函数/染色体,处理这些无效染色体的方法是将其适应度设置为较小的值(如 45%)。

2.6 GEPKNN 算法

综合上面的分析,GEPKNN 的伪码如下:

```
算法(GEPKNN)
输入 T //训练集
输出 核 KNN 分类器
Init(p(0)) //初始化种群
t=0
while(t<maxGeneration){//未到达最大进化代数
    p(t+1)=GEP(p(t)) //产生下代种群
    for(individual I in p(t+1)){
        k=decode(I) //k 是核函数
        e=crossvalidation(T, 核 KNN(k))
        //用核函数 k 构造核 KNN,在训练集上作交叉验证
        I.fitness=1000*(1-e)
    }
    t++
    if( bestFitness>threshold)
        Break
}
k=decode(p(t)中的最好个体 I)
Return 核 KNN(k)
```

3 实　验

3.1 实验结果

为验证 GEPKNN 算法的有效性,在 UCI 的 wisconsin - breast - canser、iris、diabetes 和 glass 四个标准数据集上比较了 GEPKNN,KNN 和 C4.5 等算法的分类性能。对每个数据集,随机抽取其中 65% 作为训练集,剩下的 35% 作为测试集。表 1 显示了数据集的明细情况。GEPKNN 的参数设置汇总在表 2 中。在每个数据集上做 5 次实验,取分类准确率的平均值作为输出结果。

表 1　实验数据明细

名称	属性数	训练实例	测试实例
wisconsin - breast - canser	9	454	245
iris	5	98	52
diabetes	9	499	269
glass	9	139	75

表 2　参数设置

参数名称	参数值
非终结符号集合	{+,*,exp}
终结符号集合	$\{k_1,k_2,k_3\}$
基因头部长度	6
参数域长度	7
最大进化代数	100
变异率(表达式域和参数域)	0.044
基因转置(表达式域和参数域)	0.1
交叉	0.3
ISTransposition 和 RISTransposition	0.1

实验程序用 Java 和 Weka 实现,实验平台为 jdk1.

6,pentium4 1.8GHz 处理器,512M 内存,Windows XP 操作系统。

实验结果归纳在表 3 中,GEPKNN 算法在四个数据集上的分类正确率都优于 KNN 和 C4.5 决策树。表中最末一列是使用高斯核函数的核 KNN 在四个数据集上的分类准确率,其中的参数 δ 设为 15,这个分类器在 diabetes 上的表现比较差,在另外几个数据集上的分类性能与 KNN 相当,这表明核 KNN 的分类性能与核函数紧密关联,核函数及参数要经过仔细调校才能凸显核 KNN 的分类性能优势。

表 3 实验结果

数据集名称	GEPKNN	C4.5	KNN	高斯核 KNN
bw	0.987	0.945	0.951	0.947
iris	0.993	0.963	0.950	0.956
diabetes	0.793	0.735	0.701	0.661
glass	0.775	0.668	0.719	0.681

3.2 讨 论

GEPKNN 算法目前还有几个问题有待进一步研究,较显著的问题是其训练时间比传统分类算法长(在 diabets 数据集上算法的训练时间长达 5.17min),这个问题的成因是在计算个体的适应度时采用了比较费时的 k 折交叉验证,在种群规模较大或数据较多时 GEPKNN 需要花费较多时间更新个体的适应度。

4 结束语

文中提出的 GEPKNN 算法较好地解决了为核 KNN 选择核函数及其参数的问题,它依靠 GEP 强大的函数发现能力为核 KNN 自动构造最优的核函数,实验结果表明 GEPKNN 算法是有效的。

GEPKNN 算法的主要不足是训练时间长,下一步研究的主要内容包括开发更高效的适应度函数和进化策略,在避免过拟合训练数据的同时降低算法的时间复杂度。

参考文献:

[1] Xiong Huilin, Zhang Ya, Chen Xue - Wen. Data Dependent Kernel Machines for Microarray Data Classification[J]. Transactions on Computational Biology and Bioinformatics, 2007, 4(4):583 - 595.

[2] 饶 鲜,杨绍全,魏 青,等.核的最近邻算法及其仿真[J].系统工程与电子技术,2007,29(3):470 - 471.

[3] 王 东,吴湘滨.遗传编程运行期个体多样性分析方法及应用[J].计算机技术与发展,2006,16(9):18 - 20.

[4] 李 钧,王忠群,刘 涛.基于遗传编程的网格资源调度算法[J].计算机技术与发展,2008,18(2):129 - 132.

[5] 余新宁,王文鹏,张 俊.遗传算法程序的模块化设计[J].微机发展(现更名:计算机技术与发展),2003,13(3):4 - 7.

[6] 李 曲,蔡之华,蒋思伟,等.基因表达式程序设计在预测中的应用研究[C]//第五届全球智能控制与自动化大会.杭州:[出版者不详],2004.

[7] Ferreira C. Gene Expression Programming in Problem Solving [C]//invited tutorial of the 6th Online World Conference on Soft Computing in Industrial Applications. Berlin: [s. n.], 2001.

[8] Peng Jing, Tang Chang - jie, Zhang Jing, et al. Evolutionary algorithm based on overlapped gene expression[C]//ICNC - LNCS3612. Berlin: Springer, 2005:194 - 204.

(上接第 18 页)

参考文献:

[1] 郑少仁,王海涛,赵志峰.Ad Hoc 网络技术[M].北京:人民邮电出版社,2005.

[2] 陈 勇.Ad Hoc 网络的节能路由算法的研究[D].杭州:浙江大学,2008.

[3] 何昆鹏,李腊元.Ad Hoc 网络中按需路由协议的仿真与性能分析[J].计算机技术与发展,2008,18(3):81 - 84.

[4] Johnson D B, Maltz D A, Broch J. DSR: The Dynamic Source Routing Protocol for Multi - Hop Wireless Ad Hoc Networks. in Ad Hoc Networking[M]. Boston, MA, USA: Addison - Wesley Longman Publishing Co., Inc, 2001:139 - 172.

[5] Haquel T, Assi C. OLEAR: Optimal Localized Energy Aware Routing in Mobile Ad Hoc Networks[C]//IEEE International Conference on Communications (ICC'06). Istanbul, Turkey: IEEE Communications Society, 2006:3548 - 3553.

[6] 王敏强,郑宝玉.一种新的应用于 Ad Hoc 网络的能量感知路由协议[J].南京邮电学院学报,2005,25(1):13 - 17.

[7] Senouci S M, Naimi M. New Routing for Balanced Energy Consumption in Mobile Ad Hoc Networks[C]//Proc of the 2nd ACM international workshop on Performance evaluation of wireless ad hoc, sensor, and ubiquitous networks. NY, USA: [s. n.], 2005:238 - 241.

[8] Tarique M, Tepe K E, Naserian M. Energy Saving Dynamic Source Routing for Ad Hoc Wireless Networks[C]//Third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks. Montreal, Quebec, Canada: WIOPT, 2005:305 - 310.

[9] 李 鹏.基于能量的移动自组网路由协议研究[D].武汉:华中科技大学,2006.