基于PSO面向K近邻分类的特征权重学习算法*)

任江涛 卓晓岚 许盛灿 印 鉴
(中山大学计算机科学系 广州 510275)

摘 要 特征权重学习是基于特征赋权的K近邻算法需要解决的重要问题之一,传统上提出了许多启发式的学习方法。近年来,随着进化计算技术在模式识别及数据挖掘领域的广泛应用,基于进化计算的权重学习和距离学习方法也得到越来越多的重视。本研究针对基于特征赋权的K近邻算法的权重学习问题,提出了一种基于PSO进行权重学习的算法PSOKNN,通过与传统KNN、GAKNN及ReliefKNN的实验比较分析表明,该方法可有效地搜索出合适的特征权重,获得较好的分类精度并淘汰冗余或无关的特征。

关键词 特征赋权,K近邻分类,粒子群算法

PSO Based Feature Weighting Algorithm for KNN

REN Jiang-Tao ZHUO Xiao-Lan XU Shen-Chan YIN Jian
(Department of Computer Science, Zhongshan University, Guangzhou 510275)

Abstract Feature weighting is one of the important problems for feature weighting based KNN algorithm, and many heuristic methods have been employed to solve the problem traditionally. Recently, with the wide applications of evolutionary computation in pattern recognition and data mining areas, the evolutionary computation based feature weighting and distance learning method have got more and more attention. According to the feature-weighting problem of KNN, the paper proposes a PSO based feature weighting algorithm named PSOKNN. In comparison with other methods such as traditional KNN, GAKNN and ReliefKNN, the experiments show that the PSOKNN can get suitable feature weights, result in good classification accuracy and get rid of more redundant or irrelevant features.

Keywords Feature weighting, KNN, PSO

1 引言

K近邻分类是一种简单有效的分类方法,广泛应用于模式识别及数据挖掘的各个领域。给定一个待分类样本x,该算法首先找出与x最接近或最相似的k个已知类别标签的训练集样本,然后根据这k个训练样本的类别标签确定样本的x类别。K近邻算法的关键技术主要有两个方面,一个是距离度量技术,另一个则是根据近邻样本的类别标签投票确定待分类样本类别标签的技术,本文主要讨论距离度量技术[1]。

欧氏距离是应用得最广泛的距离度量方法之一,该度量简单易行,具有较为广泛的适用性。但该度量将所有的特征赋予相同的权重,而事实上,在当前许多高维数据中,存在许多冗余的、不相关的甚至错误的特征,将这些特征与重要的相关特征在距离计算中赋予相同的权重将导致采用该度量的K近邻算法分类精度的降低。而且,即使是与分类相关的重要特征之间,重要性也不尽相同。为此,人们提出了基于特征赋权的欧式距离度量,即给数据集的每个特征赋以一定的权重,权重大小表示特征在分类时的重要程度。当前,人们提出了许多特征赋权算法,如著名的Relief算法等,这些特征赋权算法都在一定程度上提高了K近邻算法的分类精度。

PSO算法是一种新兴的优化技术,其思想来源于人工生命及演化计算理论。PSO通过粒子追随自身找到的最优解及整个粒子群的最优解来进行优化[2,3]。近年来,粒子群算法(PSO)在模式识别与数据挖掘领域得到了许多应用,如二进制PSO算法用于特征选择[4,5],PSO算法还被应用于解决分类问题[6]。

针对K近邻算法的特征权重学习问题,本文提出了一种基于PSO的特征权重学习算法PSOKNN。论文的第2部分首先介绍了基于特征赋权的K近邻算法,第3部分介绍了PSO算法的基本原理,第4部分讨论了所提出的PSO-KNN算法,第5部分给出了实验研究结果,最后是对本文的总结。

2 基于特征赋权的K近邻算法

K近邻分类是一种基于案例的分类方法,所有的训练样本存放在n_f维(每个样本有n_f个特征属性)模式空间中。给定一个未知类别标签样本,k近邻分类器搜索其模式空间,找出最接近该未知类别标签样本的k个训练样本,作为未知样本的K个"近邻",然后根据这些近邻训练样本的类别信息基于一定的投票机制判定该未知类别标签样本的类别。

一般K近邻算法采用欧氏距离度量样本之间的距离或相似程度,其定义如公式(1)所示,其中X_1、X_2分别代表两个样本向量。

$$d(X_1,X_2)=\sqrt{\sum_{i=1}^{n_f}(x_{1i}-x_{2i})^2} \quad (1)$$

在这种度量中,每一种特征的地位是平等的,对距离度量的贡献同等重要。而在实际问题中,衡量两个样本之间的相似性时,不同的特征属性其重要程度往往是不相同的。解决方法是给每一个特征属性赋予不同的权重来体现其重要程度

*)本文研究得到国家自然科学基金资助(60573097)、广东省自然科学基金资助(04300462、05200302)。任江涛 博士,讲师,主要研究方向:数据挖掘与知识发现、生物信息学、商务智能。

的不同，即将(1)改进为：

$$d(X_1,X_2)=\sqrt{\sum_{i=1}^{n_f}w_i(x_{1i}-x_{2i})^2} \quad (2)$$

其中 w_i 代表特征权重，基于该距离度量的 K 近邻算法称为基于特征赋权的 K 近邻算法。该算法最重要的问题，就是特征权重学习，即距离学习问题，有许多启发式算法，如 Relief 算法。但进化计算方法也很适合于此问题，本研究中发现用 PSO 算法来解决此问题有一定的优势。

3 PSO 算法

粒子群优化算法(PSO)是一种进化计算技术，源于对鸟群捕食的行为模拟，由 Eberhart 和 Kennedy 提出。系统初始化为一组随机解，通过迭代搜寻最优值。但是并不同于遗传算法使用的交叉以及变异，而是粒子在解空间追随当前最优的粒子进行搜索。同遗传算法相比，PSO 的优势在于简单易实现并且没有许多参数需要调整。目前已经广泛应用于函数优化，神经网络训练，模糊系统控制以及其他遗传算法的应用领域。

PSO 中，将每个粒子也就是优化问题的解看作是搜索空间中的一点。所有的粒子都有一个表示当前在解空间中位置的属性 $X_i=(x_{i1},x_{i2},\cdots,x_{iD})$，并由评价函数计算其适应度，每个粒子还有一个速度 $V_i=(v_{i1},v_{i2},\cdots,v_{iD})$ 决定它们运动的方向和距离。粒子之间通过共享当前最优粒子的信息，在解空间中搜索。

首先粒子群初始化为一群随机粒子(随机解)，然后通过迭代来寻找最优解。在每一轮的迭代中，粒子通过速度更新当前位置，并通过适应值函数计算出其适应值，然后粒子根据以下的速度更新公式进行计算，更新其当前速度。

$$v_{id}=v_{id}+c_1\times rand\times(p_{id}-x_{id})+c_2\times rand\times(p_{gd}-x_{id}) \quad (3)$$

$$x_{id}=x_{id}+v_{id} \quad (4)$$

公式(3)是粒子的速度更新公式，$P_i=(p_{i1},p_{i2},\cdots,p_{iD})$ 表示粒子 i 的局部最优值，即粒子 i 目前为止在搜索空间中到过的最佳点，$P_g=(p_{g1},p_{g2},\cdots,p_{gD})$ 表示整个粒子群的全局最优值，即整个粒子群到目前为止在搜索空间中到过的最佳点。c_1 和 c_2 是两个正的常数，称为学习因子。$rand$ 表示 0 到 1 之间的随机数，公式(4)是粒子的位置更新公式。

根据 PSO 的粒子定义及粒子更新公式(3)、(4)，可知 PSO 的粒子(问题解)采用连续值，因此较适合于求解连续值域上的数值优化问题。而另一种进化计算方法 GA，则需进行二进制编码，一方面使得解的变化不连续，另一方面限制了解的精度。而 PSO 不但无需进行二进制编码与十进制连续值之间的转换，节省了计算量，又避免了编码过程中造成的精度损失。而本文需要解决的是基于特征赋权的 K 近邻算法中各特征的权重学习问题，而权重一般是分布在 0～1 之间的连续值，因此本研究提出了一种基于 PSO 进行权重学习的算法，如第 4 节所述。

4 PSOKNN 特征赋权及分类算法

下面从粒子设计、适应度函数及算法流程等方面分别介绍 PSOKNN 算法。

4.1 粒子设计

粒子用一个 n_f 维特征向量表示，n_f 代表数据集的特征数目，每个特征对应一个权重。权重取值范围设定为[0,1]，若粒子在移动过程中，其某些维超出[0,1]这个取值范围时，如超过上界时取 1，超过下界时取 0。

4.2 适应度函数

由于该算法目标为优化特征权重以提高 K 近邻算法的分类正确率，因此粒子的适应度函数定义为根据粒子所代表的权重向量所构造的 K 近邻算法的分类正确率。

4.3 算法流程

算法 1 $PSOKNN(D,c_1,c_2,PN,n_f,N)$

输入：数据集 D，学习因子 c_1、c_2，粒子数目 PN，特征维数 n_f、最大迭代次数 N

输出：优化的特征权重向量及相应的分类精度

步骤：

1)初始化粒子群；

2)将每个粒子所代表的特征权重向量归一化，根据粒子编码方案构造基于特征赋权的 KNN 分类器，测试并记录分类精度，作为粒子适应度；

3)根据粒子适应度，更新 P_i 及 P_g；

4)根据公式(3)，更新粒子的速度；

5)根据公式(4)，更新粒子的位置；

6)判断是否达到最大迭代次数 N，若达到则输出当前最优的特征向量及分类精度，否则返回步骤 2)继续迭代。

5 实验研究

5.1 实验数据及实验环境

为了评估算法的有效性，采用多个 UCI 数据集进行实验研究，数据集的名称及相关信息如表 1 所示。本研究的所有实验在采用 Intel 奔腾 4 3.0G CPU、内存为 512M 的计算机上进行，算法用 Matlab 编程实现。

表 1 实验中用到的 UCI 数据集

数据集	类别数	样本数	特征维数
Pima-Indian diabetes	2	768	8
Heart disease	2	270	13
Ionosphere	2	351	34
Sonar	2	208	60
Iris	3	150	4
Wine	3	178	13
Vehicle	4	846	19

5.2 实验评价方法及指标

实验采用交叉验证法(Cross Validation)进行分类精度评价。数据集被随机分成 k 个子集，PSOKNN 算法在每个数据集上运行 k 次，每次取一个子集作为测试集，其余 $k-1$ 个子集合并为训练集，然后取 k 次实验结果的均值作为该数据集的分类结果，本研究中取 $k=10$。采用分类正确率作为分类精度评价指标。

另外，为了与所提出的 PSOKNN 算法进行比较分析，还采用传统的 K 近邻算法、Relief 算法及基于 GA 的特征赋权算法对上述数据集进行实验，采用与 PSOKNN 相同的评价方法与指标，上述三种方法分别简记为 KNN、ReliefKNN 和 GAKNN。

5.3 实验参数设置

实验中算法的参数设置如下：$c_1=c_2=2$，粒子速度的初始值范围是[0,1]。粒子数 PN 随着数据集特征维数 n_f 的高低而不同，具体见表 2。这是因为当数据集的特征维数高时，

搜索空间大，所以需要较多的粒子数目；反之，则可以采用较少的粒子数目。迭代次数 N 设置为 30。

表 2 各个数据集采用的粒子数目

数据集	特征维数 n_f	粒子数 PN
Pima-Indian diabetes	8	100
Heart disease	13	100
Ionosphere	34	200
Sonar	60	300
Iris	4	50
Wine	13	100
Vehicle	18	150

5.4 实验结果

表 3、4、5 分别给出了 $K=1$、3、5 时传统 KNN、PSOKNN、GAKNN 及 ReliefKNN 针对不同数据集的分类精度，各表中第 1 列括号中的数值代表数据集的特征数目，第 3、4、5 列括号中的数值代表经过特征赋权后权重非 0 的特征数目。由于权重为 0 意味着该特征未被采用，因此 PSOKNN、GAKNN 及 ReliefKNN 算法具有一定的特征选择能力，权重非 0 的特征数即为所选择的特征数。其中所取得的最佳分类精度采用加粗字体标识。

5.5 实验结果分析

从实验结果来看，$K=1$，即采用最近邻算法时，PSOKNN 的分类精度普遍优于其它算法；$K=3$ 时，GAKNN 略优于 PSOKNN；$K=5$ 时，PSOKNN 略优于 GAKNN。在大部分数据集上，PSOKNN 及 GAKNN 优于传统 KNN 及 ReliefKNN。从权重不为 0 的特征数，即从特征选择的角度来看，PSOKNN 在分类精度高于 GAKNN 及 ReliefKNN 或相似的情况下，其选择的特征数量最少，即淘汰的冗余或无关特征较多，体现出较强的特征选择能力。综合上述分析，从整体看，PSOKNN 的性能要优于本研究中所对比的其它方法。

表 3 $K=1$ 时 KNN、PSOKNN、GAKNN 及 ReliefKNN 实验结果

数据集	KNN	PSOKNN	GAKNN	ReliefKNN
Pima-Indian diabetes(8)	0.59518	**0.728947(6)**	0.718421(6)	0.6776(7)
Heart-stalog disease (13)	0.52926	**0.814815(6)**	0.729630(12)	0.6000(11)
Ionosphere(34)	0.74775	**0.937143(14)**	0.877143(32)	0.8657(33)
Sonar(60)	0.66935	**0.675000(33)**	0.610000(57)	0.5200(49)
Iris(4)	0.89333	**0.966667(4)**	**0.966667(4)**	0.9600(4)
Wine(13)	0.65239	**0.947059(6)**	0.864706(12)	0.6824(6)
Vehicle(18)	0.40769	**0.722619(12)**	0.694048(17)	0.5595(18)

表 4 $K=3$ 时 KNN、PSOKNN、GAKNN 及 ReliefKNN 实验结果

数据集	KNN	PSOKNN	GAKNN	ReliefKNN
Pima-Indian diabetes(8)	0.69087	0.746053(5)	**0.752632(8)**	0.7171(7)
Heart-stalog disease (13)	0.61222	**0.825926(8)**	0.766667(9)	0.6111(11)
Ionosphere(34)	**0.89179**	0.862857(20)	0.862857(33)	0.8343(33)
Sonar(60)	**0.81009**	0.675000(28)	0.575000(59)	0.5100(49)
Iris(4)	0.95867	**0.960000(2)**	**0.960000(4)**	0.9467(4)
Wine(13)	0.7403	**0.941176(10)**	**0.941176(10)**	0.6294(6)
Vehicle(18)	0.69237	0.709524(13)	**0.710714(18)**	0.5631(18)

表 5 $K=5$ 时 KNN、PSOKNN、GAKNN 及 ReliefKNN 实验结果

数据集	KNN	PSOKNN	GAKNN	ReliefKNN
Pima-Indian diabetes(8)	0.70962	0.764474(6)	**0.785185(5)**	0.7132(7)
Heart-stalog disease (13)	0.62852	**0.870370 (6)**	0.755556(10)	0.6519(11)
Ionosphere(34)	**0.89564**	0.880000(22)	0.837143(32)	0.8171(33)
Sonar(60)	**0.83957**	0.640000(28)	0.610000(56)	0.4900(49)
Iris(4)	0.95667	**0.966667(3)**	0.960000(3)	0.9467(4)
Wine(13)	0.72713	**0.941176 (9)**	0.911765(11)	0.6353(6)
Vehicle(18)	0.70213	**0.745238(11)**	0.735714(16)	0.5536(18)

结论 本研究针对基于特征赋权的 K 近邻算法的权重学习问题，提出了一种基于 PSO 进行权重学习的算法 PSOKNN，通过与传统 KNN、ReliefKNN 及 GAKNN 的实验比较分析表明，该方法可有效地搜索出合适的特征权重，获得较好的分类精度并淘汰冗余或无关的特征。

参 考 文 献

1 Han Jiawei, Kamber M 著. 范明，孟小峰译. 数据挖掘概念与技术. 机械工业出版社，2001

2 Kennedy J, Eberhart R C. Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ, 1995. 1942～1948

3 Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm. In: Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics 1997, Piscataway, NJ. 1997. 4104～4109

4 Liu Yu, Qin Zheng, Xu Zenglin, He Xingshi. Feature Selection with Particle Swarms, Computational and Information Science. In: First International Symposium, CIS 2004, Shanghai, China, Proceedings, LNCS3314, December 2004. 425～430

5 Wang Ling, Yu Jinshou. Fault Feature Selection Based on Modified Binary PSO with Mutation and Its Application in Chemical Process Fault Diagnosis, ICNC LNCS3612, 2005. 832～840

6 De Falco I, Della Cioppa A, Tarantino E. Facing classification problems with Particle Swarm Optimization. Applied Soft Computing. In Press, Available online 3, February 2006