基于属性值相关距离的KNN算法的改进研究

肖辉辉 段艳明

(河池学院 宜州 546300)

摘 要 样本距离机制的定义直接影响到KNN算法的准确性和效率。针对传统KNN算法在距离的定义及类别决定上的不足,提出了利用属性值对类别的重要性进行改进的KNN算法(FCD-KNN)。首先定义两个样本间的距离为属性值的相关距离,此距离有效度量了样本间的相似度。再根据此距离选取与待测试样本距离最小的K个近邻,最后根据各类近邻样本点的平均距离及个数判断待测试样本的类别。理论分析及仿真实验结果表明,FCD-KNN算法较传统KNN及距离加权-KNN的分类准确性要高。

关键词 KNN算法,相关距离,属性值,样本距离机制

中图法分类号 TP301.6 **文献标识码** A

Improved the KNN Algorithm Based on Related to the Distance of Attribute Value

XIAO Hui-hui DUAN Yan-ming

(Department of Computer and Information Sciences, Hechi University, Yizhou 546300, China)

Abstract Definition of the samples will directly impact on the accuracy and the efficiency of KNN. In view of disadvantages to the traditional KNN algorithm on the distance the definition and categories of decision, proposed the use of attribute importance to category to improve KNN algorithm (FCD-KNN). At first, a distance of the two samples is defined as the correlation distance of the same attribute values. The distance can effectively measure the similarity degree of the two sample. Secondly, According to this distance selects the k nearest neighbors. Finally, the category of the test sample is decided by the average distance and the numbers on the respective category. The theoretical analysis and the simulation experiment show that compared with KNN and-KNN, raised the rate of accuracy enormously in classification.

Keywords KNN algorithm, Correlation distances, Attribute, Sample distance mechanism

1 引言

K最近邻(k-Nearest Neighbor, KNN)分类算法是最简单的机器学习算法之一,理论上比较成熟。KNN算法首先将待分类样本表达成和训练样本一致的特征向量;然后根据距离函数计算待测试样本和每个训练样本的距离,选择距离最小的K个样本作为近邻样本;最后根据K个近邻样本判断待分类样本的类别[1]。其中,样本距离机制的定义直接影响到KNN算法的准确性和效率。传统KNN的距离机制是以样本的属性值为参数的,计算距离时只简单计算待测样本与训练集样本的相同属性值个数[2],一旦数据分布不均,大类别样本就会占有密度优势,从而导致其包含的属性参数的频率也随之提高,这将导致不同类别的近邻样本点个数相等、因K取值的不同而错误判断类别等问题。然而在实际应用中不同的属性参数对分类的影响力是不相同的,即属性参数频率的高低与分类相关度强弱关联不大。KNN算法中近邻样本的正确选取是分类准确的关键因素之一,而近邻样本是通过计算测试样本与每个训练集样本的距离来选定的。不同的距离度量选取的近邻样本也不相同,不合适的距离函数下选取的近邻样本,不仅有可能干扰分类,甚至可能产生错误的分类结果。因此,定义合适的距离是KNN正确分类的前提[3]。目前,国内外很多学者对KNN算法中的距离进行了大量的研究,但一般都没有考虑特征属性值对类别判断的重要性。两样本间属性的相关距离可用来度量属性值对类别的重要性,相关距离熵越小,两样本的相似程度越大,类可信度越大。

本文在上述研究的基础上,将样本距离重新定义为任意两个样本间属性间的相关距离,提出一种基于属性值相关距离的KNN改进算法(Feature Correlation Difference KNN, FCD-KNN)。利用UC I数据库中的Iris数据集进行实验,在Matlab中进行仿真,结果表明FCD-KNN算法在分类效率上和准确性上优于传统-KNN算法和距离加权-KNN算法。

本文受广西教育厅科研基金项目(201106LX577, 201106LX604),国家自然科学基金项目(40971234),河池学院青年科研项目(2012B-N005, 2012B-N007)资助。

肖辉辉(1977—),男,硕士生,主要研究领域为人工智能、数据库,E-mail:whoamireal@126.com;段艳明(1978—),女,硕士生,主要研究领域为人工智能、数据库,E-mail:yanhui0920@126.com(通信作者)。

2 KNN 算法与距离加权-KNN 算法

KNN 分类算法的主要思想是：先计算待分类样本与已知类别的训练样本之间的距离，从中找出距离最近的 K 个邻居；再根据这 K 个邻居样本所属的类别来判断待分类样本数据的类别：K 个样本中的大多数属于某一个类别，则该样本也属于这个类别[4]。KNN 算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。KNN 方法虽然从原理上也依赖于极限定理，但在类别决策时，只与极少量的相邻样本有关[5]。当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的 K 个邻居中大容量类的样本占多数。

传统-KNN 算法的距离只是简单计算待测样本与训练集样本的属性值相同的个数，极易产生不同类别的近邻样本点个数相等的情况。这种情况下，利用 KNN 算法无法判断其类别。鉴于此问题的存在，将其改进为在确定类别时对距离进行加权，即距离加权-KNN 算法。该算法采用与传统 KNN 分类方法相同的距离选取近邻样本，但投票时考虑各类近邻样本与待分类样本的平均距离，从一定程度上克服了 KNN 算法的上述不足，但仍未解决各类近邻样本的个数及平均距离均相同时的测试问题。以上两种算法的弊端在于样本之间距离的定义过于简单，没有考虑不同属性值对类别判断时的作用大小[6]。即两种算法的距离机制仅局限于样本间同一属性下的相同特征，而未考虑不同的特征，从而导致省略掉的特征针对分类的一些极端情况，在大多数情况下会导致样本间距离比实际要小。

3 基于属性值相关距离的 KNN 算法改进

综上所述，KNN 算法和距离加权-KNN 算法精度不高的主要原因是基本上都没有考虑到属性值对类别判断的重要性，忽略了不同属性值间对分类的相关性。例如：电脑和计算机，是两个不同的属性值，在 KNN 算法和距离加权-KNN 算法的距离机制中，这两个属性值的距离很大，但我们知道电脑和计算机其实是同一个概念，所在的样本属于同一类的概率较大，利用属性值之间的相关距离可以避免这种错误分类。

3.1 相关系数与相关距离

相关系数是衡量随机变量 X 与 Y 相关程度的一种方法，相关系数的取值范围是[−1,1]。相关系数的绝对值越大，则表明 X 与 Y 相关度越高。当 X 与 Y 线性相关时，相关系数取值为 1(正线性相关)或−1(负线性相关)。

相关系数的定义：

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{1}$$

相关距离的定义：

$$D_{xy}=1-\rho_{XY} \tag{2}$$

样本间的相关系数和相关距离的计算可以利用 Matlab 的 corrcoef()函数和 pdist()函数。例如：Matlab 计算(3,6,2,4)与(3,8,4,6)之间的相关系数与相关距离如下：

```
X=[3 6 2 4 ;3 8 4 6]
C=corrcoef(X')    %将返回相关系数矩阵
D=pdist(X,'correlation')
```

结果：

```
C=1.0000   0.9022
  0.9022   1.0000
D=0.0978
```

其中 0.9022 就是相关系数，0.0978 是相关距离。显然 pdist 值都会在 0～1 之间，值越小，越能体现两个样本的强相关性，0 则表示绝对相关。

3.2 基于属性值相关距离的 KNN 算法改进

设数据集为 S，$X_i(i=1,2,\cdots,m)$是训练样本集，它包含了 n 个不同类别 $C_r(r=1,2,\cdots,n)$，每个训练样本有 m 个属性 V，每个属性 V 具有 i 个不同值$\{v_1,v_2,\cdots,v_i\}$。

定义 1(距离)　测试样本 X(属性为：$V_{x1},V_{x2},\cdots,V_{xm}$)与训练样本 Y(属性为：$V_{y1},V_{y2},\cdots,V_{ym}$)的相关距离定义如下：

$$d(Y,X)=p_{dist}([V_{x1},V_{x2},\cdots,V_{xm};V_{y1},V_{y2},\cdots,V_{ym}],\ \text{'correlation'}) \tag{3}$$

由定义可知，X 与 Y 的距离为任意两个样本间属性值的相关距离，此距离有效度量了两个样本间的相似程度。当 X 与 Y 相同属性值的相近值越多时，相关距离越小，则两个样本的相似程度越高。

定义 2(类可信度)　采用文献[7]判断类别的方法，设 C_r 代表类别，Y 为待测样本，X_r 为近邻中属于 C_r 类的样本，N 为近邻样本总数，N_r 为近邻样本属于 C_r 类的样本个数。称 $T(C_r,Y)$为 Y 对 C_r 的类可信度。计算方法如下：

$$T(C_r,Y)=\frac{N-N_r}{N}\times\frac{1}{N_r}\sum_{r=1}^{N_r}d(Y,X_r) \tag{4}$$

其中，$\frac{1}{N_r}\sum_{r=1}^{N_r}d(Y,X_r)$为 Y 于 C_r类的平均距离。

式(4)由两部分组成：待测样本与各类的平均距离和各类的近邻样本点数。平均距离越小，Y 判别为类的可能性越大；当 C_r 类的入选点样本点数 N_r 越多时，$\frac{N-N_r}{N}$值越小，则 $T(C_r,Y)$越小，Y 判别为 C_r 类的可信度越高。

式(4)对 Y 的判断不仅基于近邻样本中属于 C_r 类的个数，更重要的是基于 Y 与类 C_r 的平均距离。

FCD-KNN 算法步骤如下：

(1)在 Matlab 中利用 corrcoef()函数和 pdist()函数计算出测试样本与训练样本间的相关系数和相关距离。

(2)根据 K 值大小选取距离较小的前 K 个的训练样本，计算各类近邻样本个数 N_r。

(3)利用式(4)计算待测样本与各类的类可信度 $T(C_r,Y)$。

(4)比较类可信度 $T(C_r,Y)$的大小，选取 $T(C_r,Y)$较小的近邻样本的类别为待测样本的类别。

下面用一个例子说明 FCD-KNN 算法的实现过程。

例 1　表 1 是 Iris 数据集的一部分，前 15 条数据(X_1 −

X_{15})为训练集，后一条(X_{16})为测试数据，A 至 D 为条件属性，E 为 class。以下测试均取 $K=6$，即 6 个近邻。

下面用 FCD-KNN 算法来判别 X_{16} 的类别。

(1)计算测试样本 X_{16} 与训练样本间的相关距离，确定 X_{16} 的近邻样本。由式(3)得到 X_{16} 与训练样本 X_1-X_{15} 的相关距离分别为：{0.2572，0.2779，0.2686，0.1980，0.0063，0.0003，0.0003，0.0017，0.0012，0.0064，0.0023，0.0118，0.0266，0.0249，0.0192}。

(2)$K=6$，取相关距离小的 6 个训练样本 X_5、X_6、X_7、X_8、X_9 和 X_{11}。其中，X_5 为 Iris Setosa，X_6、X_7、X_8 和 X_9 为 Iris Versicolour，X_{11} 为 Iris Virginica。

表 1　训练集与测试样本

	A (Sepal length)	B (Sepal wrowNoth)	C (Petal length)	D (Petal wrowNoth)	E (class)
X_1	5.0	3.3	1.4	0.2	Iris Setosa
X_2	5.0	3.5	1.6	0.6	Iris Setosa
X_3	5.0	3.2	1.2	0.2	Iris Setosa
X_4	5.0	3.0	1.6	0.2	Iris Setosa
X_5	4.5	2.1	1.8	0.3	Iris Setosa
X_6	5.5	2.3	4.0	1.3	Iris Versicolour
X_7	6.5	2.8	4.6	1.5	Iris Versicolour
X_8	6.6	2.9	4.6	1.3	Iris Versicolour
X_9	6.9	3.1	4.9	1.5	Iris Versicolour
X_{10}	6.4	3.2	4.5	1.5	Iris Versicolour
X_{11}	6.5	3.2	5.1	2.0	Iris Virginica
X_{12}	7.2	3.6	6.1	2.5	Iris Virginica
X_{13}	5.0	2.5	4.5	1.7	Iris Virginica
X_{14}	5.8	2.2	5.0	1.4	Iris Virginica
X_{15}	5.0	2.0	4.1	1.5	Iris Virginica
X_{16}	5.0	2.0	3.5	1.0	Iris Versicolour

(3)利用式(4)计算测试样本 X_{16} 的类可信度 $T(C_r,Y)$。$N=6$，$N_1=1$，$N_2=4$，$N_3=1$，则：

$$T(\text{Iris Setosa},X_{16})=\frac{6-1}{6}\times 0.0063=0.0053$$

$$T(\text{Iris Versicolour},X_{16})=\frac{6-4}{6}\times\frac{1}{4}\times(0.0003+0.0003+0.0017+0.0012)=0.0003$$

$$T(\text{Iris Virginica},X_{16})=\frac{6-1}{6}\times 0.0023=0.0019$$

(4)比较上一步求得的类可信度，可判断 X_{16} 为 Versicolour，得到正确的类别判断。

从表 1 中的属性值可以得到：测试样本 X_{16} 与训练样本 X_1、X_2、X_3、X_4 和 X_{12} 各有 1 个属性值相同，与 X_{15} 有 2 个属性值相同，除此外与其他训练样本均没有相同的属性值。若按传统的 *KNN* 算法测试样本 X_{16}，则 $K=6$ 个近邻样本，应为 X_1、X_2、X_3、X_4、X_{12} 和 X_{15}，其中 4 个为 Iris Setosa 类别，2 个为 Iris Virginica 类别，可判断样本 X_{16} 为 Iris Setosa 类别，判断错误。若按距离加权-KNN 算法测试 X_{16}，同样判断 X_{16} 为 Iris Setosa 类，判断错误。

由上述例子的演算结果可知，传统-KNN 算法和距离加权-KNN 算法在分类中准确率不高，其主要原因是样本间的距离机制过于简单，未考虑样本属性值之间的相关性[8]。而本文提出的 FCD-KNN 算法把样本间的距离改进为样本属性值的相关距离，顾全了样本属性值对分类的重要性，通过重要属性值相同能够更有效地找到近邻样本，提高了分类的准确效。同时，在进行分类时，通过类可信度综合考虑各类别的近邻样本点个数及测试样本与训练样本间的类别平均距离，使测试样本尽可能接近各类中的类可信度，提高正确的分类准确率。

4　实验结果与分析

选用 UCI 数据集上的 Iris 数据在 Matlab 中结合 VC++ 进行 FCD-KNN 算法的仿真。Iris 共有 150 组数据，为了操作方便，对各组数据添加 rowNo 属性，且第一组 rowNo=1。考虑到训练数据集的随机性和多样性，选择 rowNo 模 3 不等于 0 的 100 组作为训练数据集，剩下的 50 组做测试数据集。取 $K=6$ 个最邻近数据。为了验证 FCD-KNN 算法的效率，分两种情况进行仿真实验：(1)训练样本数据不同的情况；(2)不同 K 值的情况。

实验 1　测试训练样本大小对分类结果的影响。设定 K 值为 6，该实验的训练样本数从 10 开始，依次取到 100，其得到的准确率如图 1 所示。根据图 1 可知，当训练样本数小于 40 时，FCD-KNN 明显优于传统的马氏-KNN 和距离加权-KNN。该仿真实验说明在训练样本数较少的情况下，FCD-KNN 的分类准确率较高，能获取较高的信息。当训练样本数上升时，三者的分类准确率都提高，但 FCD-KNN 的准确率都高于马氏-KNN 和距离加权-KNN，并且随着训练样本数的增加，FCD-KNN 算法准确率表现更为平稳。

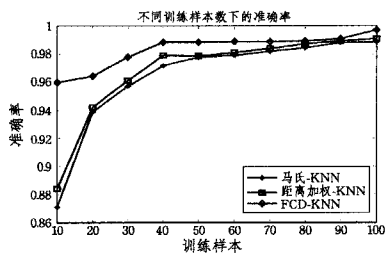

图 1　不同训练样本数下的准确率

实验 2　测试不同 K 值对分类结果的影响。该实验训练样本取 150 个，K 的取值从 5 开始，依次取到 32。在 Matlab 中仿真得到的准确率如图 2 所示。从图 2 的实验结果可知，随着 K 值的变化，FCD-KNN 明显优于马氏-KNN 和距离加权-KNN，而且 FCD-KNN 对分类效果有明显的改进，准确率也较为平稳。

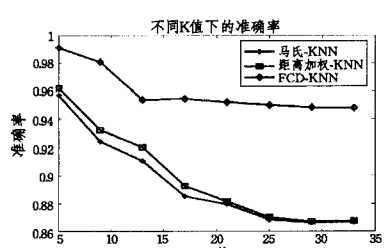

图 2　不同 K 值时的准确率

(下转第 187 页)

byte2hex(unicode()newHash). equals(byte2hex(signedData. getContent())))(6)

(1)计算待上报数据的消息摘要,选择与客户端相同的SHA1算法;

(2)消息摘要转换为大写;

(3)消息摘要转换成字节;

(4)获得数字签名者信息;

(5)验证数字签名证书的有效性;

(6)把SignedData中提取出的content(此处为消息摘要)和重新计算的消息摘要均转换成16进制字符串,再判断是否相等;

结束语 在网络上实现信息的安全传输,成为越来越多人的普遍诉求。本文提出了一种实现信息安全传输的方案。在客户端(浏览器端),通过引入CAPICOM技术对待传输的数据实现数字签名和数字信封。使用CAPICOM技术,使得浏览器对数据的加密操作变得十分简单和高效。在服务端,使用第三方库IAIK来解析PKCS#7格式的数字信封,并验证数字签名。同时,作为CAPICOM的基础证书管理部分,信息安全传输系统搭建了EJBCA系统,完成对证书的各种管理。此信息安全传输系统,通过结合数字证书、加密、数字签名、数字信封、PKI等各种网络信息安全传输的技术,实现了信息传输的机密性、完整性、真实性、不可否认性。

参 考 文 献

[1] 宋玲,李陶深,陈拓. 用CAPICOM组件实现应用系统安全性的方法[J]. 计算机工程,2004,30(16):128-129

[2] 李志民. 数字签名和数字信封的比较与应用[J]. 经济师,2006,4:137-138

[3] 张敏,卢巍. 基于CAPICOM的文档管理系统[J]. 计算机应用,2012,32(S1):56-57

[4] 陈勤,凌青山,丁宏. 安全CA实例——EJBCA的研究[J]. 计算机工程与设计,2005,26(12):3222-3224

[5] 谭文学,张健钦,王细萍. 密码中间件CAPICOM的应用研究[J]. 微计算机信息,2006,22(11):112-114

[6] 唐辉天. 将微软CAPICOM组件引入J2EE平台进行数字签名的研究与实现[D]. 西南石油大学,2006

[7] 张文奇,肖衡,段斌,等. Linux平台上基于PKCS#11的PKCS#7 Signed-data的实现[J]. 计算机应用,2003,23(11):103-105

[8] 张青凤,张凤琴. CryptoAPI在基于数字证书身份认证系统中的应用[J]. 现代计算机,2011(24)

[9] 周志刚,徐芳,肖晓华,等. 在Java中进行数字签名的一种实现方法[J]. 科学技术与工程,2006,6(17):2752-2754

[10] 李刚. 轻量级Java EE企业应用实战[M]. 北京:电子工业出版社,2010

[11] 梁栋. Java加密与解密的艺术[M]. 北京:机械工业出版社,2010

[12] 马臣云,王彦. 精通PKI网络安全认证技术与编程实现[M]. 北京:人民邮电出版社,2008

[13] 黄智诚,谢静贤,黄恺昕. 中文WORD 2000使用指南[M]. 北京:中国石化出版社,2000

[14] 胡凯,李腊元. 一个电子商务模型的认证和加密的设计与实现[J]. 计算机工程与设计,2003,24(2):41-43

[15] 王细萍,谭文学,张健钦. CAPICOM在安全电子商务中的应用研究[Z]. 计算机与信息技术,14-17

[16] 王金伟,孙德兵. 基于OpenSSL和CAPICOM身份认证的研究与实现[J]. 福建电脑,2010(08):5-7

[17] 张小江,高翔. CAPICOM组件技术实现数字签名的研究[J]. 商业现代化,2007,19

[18] 徐歆恺,梁军. 巧用CAPICOM进行安全通信[Z]. 计算机与网络,2009(18):53-55

[19] 吴艳. 使用数字证书进行PKCS#7数字签名[J]. 电脑编程技巧与维护,2011(16):130-134

[20] IAIK-JCE 3.13 API:Documentation[OL]. http://javadoc.iaik.tugraz.at/iaik_jce/3.13/overview-summary.html

[21] CAPICOM Reference(Windows) [OL]. http://msdn.microsoft.com/en-us/library/windows/desktop/aa375732(v=vs.85).aspx

[22] EJBCA[OL]. http://www.ejbca.org/

(上接第159页)

结束语 针对传统-KNN算法和距离加权-KNN算法在距离的定义及类别决定上的不足而导致准确率的问题,对这两类算法的特点和不足进行分析,提出了一种基于属性值相关距离的FCD-KNN算法。该算法考虑属性值对分类的重要性,定义样本间的距离为属性值的相关距离,此距离能有效度量两个样本间的相似程度,最大程度地提取与待测样本相似的近邻样本。仿真实验结果表明,该算法能够较大幅度地提高分类的准确率,且在相同测试条件下其并分类的准确率都高于马氏-KNN和距离加权-KNN,证实了FCD-KNN算法的有效性。但由于在实际应用中样本数据集都存在着不同程度的模糊性,这会对算法中样本属性值间的相关度的计算造成较大的干扰,如何解决这一问题并提高FCD-KNN的健壮性是下一步的研究重点。

参 考 文 献

[1] 王增民,王开珏. 基于熵权的K最临近算法改进[J]. 计算机工程与应用,2009,45(30):129-131

[2] 周靖,刘晋胜. 特征联合熵的一种改进k近邻分类算法[J]. 计算机应用,2011,37(7):1787-1792

[3] 陆微微,刘晶. 一种提高k-近邻算法效率的新算法[J]. 计算机工程与应用,2008,44(4):163-165

[4] 周靖,刘晋胜. 一种采用类相关度优化距离的KNN算法[J]. 微计算机应用,2010,31(11):7-12

[5] 杨立,左春,王裕国. 基于语义距离的K-最近邻分类方法[J]. 软件学报,2005,16(12):2054-2062

[6] Wu Xin-dong,Kumar V,Quinlan J R,et al. Top 10 Algorithms in Data Mining[J]. Knowledge and Information Systems,2008,14(1):1-37

[7] 童先群,周忠眉. 基于属性值信息熵的KNN改进算法[J]. 计算机工程与应用,2010,46(3):114-117

[8] 周靖,刘晋胜. 基于特征熵相关度差异的KNN算法[J]. 计算机工程,2011,37(17):146-148