计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180572
56(12):2600-2611,2019

粒向量与 K 近邻粒分类器

陈玉明　　李　伟
(厦门理工学院计算机与信息工程学院　福建厦门　361024)
(cym0620@163.com)

Granular Vectors and K Nearest Neighbor Granular Classifiers

Chen Yuming and Li Wei
(*College of Computer and Information Engineering*, *Xiamen University of Technology*, *Xiamen*, *Fujian* 361024)

Abstract　K nearest neighbor (KNN) classifier is a classical, simple and effective classifier. It has been widely employed in the fields of artificial intelligence and machine learning. Aiming at the problem that traditional classifiers are difficult to deal with uncertain data, we study a technique of neighborhood granulation of samples on each atom feature, construct some granular vectors, and propose a K nearest neighbor classification method based on these granular vectors in this paper. The method introduces a neighborhood rough set model to granulate samples in a classification system, and the raw data can be converted into some feature neighborhood granules. Then, a granular vector is induced by a set of neighborhood granules, and several operators of granular vectors are defined. We present two metrics of granular vectors which are relative granular distance and absolute granular distance, respectively. The monotonicity of distance of granular vectors is proved. Furthermore, the concept of K nearest neighbor granular vector is defined based on the distance of granular vectors, and K nearest neighbor granular classifier is designed. Finally, the K nearest neighbor granular classifier is compared with the classical K nearest neighbor classifier using several UCI datasets. Theoretical analysis and experimental results show that the K nearest neighbor granular classifier has better classification performance under suitable granulation parameters and k values.

Key words　K nearest neighbor (KNN) classifier; granular computing; granular vector; granular distance; granular classifier

摘　要　K 近邻(K nearest neighbor, KNN)分类器是一种经典的分类器,它简单而又有效,已经在人工智能与机器学习领域得到了广泛的应用.针对传统分类器难以处理不确定性数据的问题,研究样本单特征邻域粒化技术,构造粒的向量形式,提出一种基于粒向量的 K 近邻分类方法.该方法引入邻域粗糙集模型,对分类系统中的样本进行单特征邻域粒化,形成特征邻域粒子.并由多个特征邻域粒子构成一个粒向量,定义了多种粒向量运算算子,提出了 2 种粒向量距离:相对粒距离与绝对粒距离,证明了粒向量距离的单调性原理.进一步,基于粒向量距离定义了 K 近邻粒向量概念,提出了 K 近邻粒分类器.最后,结合 UCI 数据集,采用 K 近邻粒分类器与经典 K 近邻分类器进行比较测试.理论分析和实验表明:针对合适的粒化参数与 k 值,K 近邻粒分类器具有较好的分类性能.

收稿日期:2018-08-14;**修回日期**:2019-05-22
基金项目:国家自然科学基金项目(61573297,61976183);福建省出国留学奖学基金项目;福建省自然科学基金项目(2016J01198,2019J01850);福建省教育厅 A 类项目(JA15363);厦门市科技计划指导项目(3502Z20179038)
This work was supported by the National Natural Science Foundation of China (61573297, 61976183), the Scholarship for Studying Abroad Program of Fujian Province, the Natural Science Foundation of Fujian Province of China (2016J01198, 2019J01850), the Class A Project Fujian Provincial Education Department (JA15363), and the Science and Technology Planning Guidance Project of Xiamen (3502Z20179038).

关键词 KNN 分类器;粒计算;粒向量;粒距离;粒分类器

中图法分类号 TP181

KNN 是最简单的分类算法,由 Hart[1] 于 1968 年提出,主要思想是计算待分类样本与训练样本之间的差异性,并按照差异由小到大排序,选出前面 k 个差异最小的类别,并统计在 k 个类别中出现次数最多的类别为最相似的类,最终将待分类样本分到与训练样本最相似的类中.KNN 算法在众多领域得到了广泛的应用,例如人脸识别[2]、文字识别[3]、聚类[4]、大数据[5-6]、多标签学习[7]等.经典 KNN 算法存在时间和空间复杂度高、k 个近邻样本的同权重影响分类精度、噪声敏感、不均衡样本分类精度低、k 值难以确定等不足.众多学者从多个方面提出了许多改进算法[8-23],提高了 KNN 算法的效率.

KNN 算法存储训练集的所有样本数据,造成了极大的存储开销和计算代价.已有很多文献提出了减少计算的方法,这些方法大致可分为 2 类:1)减少训练集的大小,删去部分冗余的样本,或通过聚类的方式选择部分样本[8];2)采用快速算法[9]搜索到 k 个近邻样本,以及引入高效的索引方法.比较常用的方法有 K-D 树[10]、局部敏感 Hash[11]等.

经典 KNN 算法采用欧氏距离计算相似度,而且赋予每个特征同等的权重,这种方法造成 KNN 算法对噪声特征非常敏感.为此,许多改进算法在度量相似度的距离公式中给特征赋予不同权重,可根据特征在整个训练样本库中的分类作用得到特征权重[12],也可根据训练样本库中的局部样本靠近待测样本的距离得到样本权重[13].当各类样本分布不均衡时,存在 KNN 算法分类性能下降的问题.目前改进的方法有均匀化样本分布密度[14]、优化判决策略.文献[15]赋予稀少样本更高的权重,使得样本相对更均匀,从而改善了近邻判决规则.

KNN 的分类效果很大程度上依赖于 k 值的选择.k 值选择过小,得到的近邻数过少,会降低分类的精度,同时放大了噪声数据的干扰;而 k 值选择过大,把实际上并不相似的数据也包含进来,造成噪声增加而导致分类效果降低.如何选择恰当的 k 值也成为 KNN 研究的热点[16-18].

除上述改进算法之外,也有研究者将 KNN 和其他分类算法进行集成.例如 KNN 与 SVM 进行集成[19-20]、KNN 与 PSO 集成[21]、深度学习与 KNN 集成[22]以及模糊 KNN[23]等方法,有效提高了 KNN 分类算法的分类性能.

KNN 算法是一个性能优秀的分类算法,许多学者从不同角度提出了多种 KNN 改进算法,我们从全新的角度出发,在集合论与单特征邻域信息粒化的基础上定义了粒向量,并提出一种新的分类器模型:K 近邻粒分类器.信息粒的概念最初由 Zadeh[24] 定义;粒计算首次由 Lin 等人[25-26]提出;苗夺谦等人[27]从集合论角度讨论了粒计算的结构;王国胤等人[28-29]分析了粒计算中的不确定性度量及在大数据中的应用;Yao 等人[30-31]提出了邻域系统及邻域粒计算;Hu 等人[32-34]分析了邻域约简和分类;Pedrycz 等人[35-37]设计了多种超盒模糊粒分类器.我们从分类系统的单特征邻域粒化出发,定义了粒向量距离度量,提出了 K 近邻粒向量的概念,从而将分类问题转化为 K 近邻粒向量的搜索问题,构建了 K 近邻粒分类器模型.进一步设计了 K 近邻粒分类器,并进行了实验验证.理论分析与实验结果表明,K 近邻粒分类器可以在合适的粒化参数及 k 值情况下,取得较好的分类性能.

1 邻域粒化与粒向量

波兰数学家 Pawlak[38] 提出的粗糙集理论是分类系统采用最为广泛的模型之一.在粗糙集理论中,等价类视为一个基本粒子.对于现实世界广泛存在的实数型数据,需要进行离散化过程,而离散化过程容易造成分类信息的丢失.为此,Yao[30] 提出了邻域粗糙集模型,Hu 等人[32-34]应用于数据分类领域,其邻域粒化是从整个特征空间进行,下面以单个特征为标准进行邻域粒化并构造粒向量.

定义 1. 设 $C=(S,F,L)$ 为一分类系统,其中 $S=\{x_1,x_2,\cdots,x_n\}$ 为样本集合,$F=\{f_1,f_2,\cdots,f_m\}$ 是特征集合,L 表示样本的标签或类别.样本在特征集 F 上的值是实数型的数据,在标签 L 上的值为符号型或离散型的数据.

定义 2. 设分类系统为 $C=(S,F,L)$,对于样本 $\forall x,y\in S$ 和单原子特征 $\forall a\in F$,定义样本 x,y 在单原子特征 a 上的距离函数为

$$D_a(x,y)=|v(x,a)-v(y,a)|,$$

其中,$v(x,a)$ 表示样本 x 在特征 a 上的值.

定义 3. 设分类系统为 $C=(S,F,L)$ 和邻域粒化参数为 δ,对于样本 $\forall x\in S$,单原子特征 $\forall a\in F$,则 x 在 a 上的 δ 邻域粒子定义为

$$g_a(x)_\delta=\{y\mid x,y\in S,D_a(x,y)\leqslant\delta\}.$$

性质 1. 根据邻域粒子的定义,x 在 a 上的 δ 邻域粒子 $g_a(x)_\delta$ 满足 4 个性质:

1) $g_a(x)_\delta\neq\varnothing$;

2) $x\in g_a(x)_\delta$;

3) $x\in g_a(y)_\delta\Leftrightarrow y\in g_a(x)_\delta$;

4) $\bigcup\limits_{x\in S}g_a(x)_\delta=S$.

定义 4. 设分类系统为 $C=(S,F,L)$ 和邻域粒化参数为 δ,对于样本 $\forall x\in S$,单原子特征 $\forall a\in F$,邻域粒子 $g_a(x)_\delta$ 的大小定义为

$$M(g_a(x)_\delta)=|g_a(x)_\delta|,$$

$|\cdot|$ 表示集合的基数.易知邻域粒子的大小满足:$1\leqslant|g_a(x)_\delta|\leqslant|S|$.

定义 5. 设分类系统为 $C=(S,F,L)$ 和邻域粒化参数为 δ,对于样本 $\forall x\in S$,特征子集 $\forall P\subseteq F$,设 $P=\{a_1,a_2,\cdots,a_m\}$,则 x 在特征子集 P 上的 δ 邻域粒向量定义为

$$\boldsymbol{V}_P(x)_\delta=(g_{a_1}(x)_\delta,g_{a_2}(x)_\delta,\cdots,g_{a_m}(x)_\delta).$$

$g_a(x)_\delta$ 是样本 x 在特征 a 上的 δ 邻域粒子,是一个集合的形式,它称为粒向量的一个元素,简称为粒元素.$\boldsymbol{V}_P(x)_\delta$ 则为粒向量,由粒元素组成.因此,粒向量的元素是集合,与传统向量不一样,传统向量的元素是一个实数.当粒向量的元素都为空集时,称为空粒向量,记为 $\boldsymbol{V}_{\text{null}}$;当粒向量的元素都是样本集,称为满粒向量,记为 $\boldsymbol{V}_{\text{full}}$.

定义 6. 设分类系统为 $C=(S,F,L)$ 和邻域粒化参数为 δ,对于样本 $\forall x\in S$,特征子集 $\forall P\subseteq F$,设 $P=\{a_1,a_2,\cdots,a_m\}$,则 x 在特征子集 P 上的 δ 邻域粒向量 $\boldsymbol{V}_P(x)_\delta$ 的大小定义为

$$|\boldsymbol{V}_P(x)_\delta|=\sqrt{\sum_{i=1}^{m}|g_{a_i}(x)_\delta|^2}.$$

粒向量 $\boldsymbol{V}_P(x)_\delta$ 的大小也称为粒向量的模.

定理 1. 设分类系统为 $C=(S,F,L)$,对于样本 $\forall x\in S$,单特征 $\forall a\in F$,$g_a(x)_\gamma$,$g_a(x)_\delta$ 为 x 在 a 上的邻域粒子,若 $0\leqslant\gamma\leqslant\delta\leqslant1$,则 $g_a(x)_\gamma\subseteq g_a(x)_\delta$.

证明.对于 $\forall x\in S$,根据邻域粒子的定义,则有 $g_a(x)_\gamma=\{y\mid x,y\in S,D_a(x,y)\leqslant\gamma\}$ 和 $g_a(x)_\delta=\{y\mid x,y\in S,D_a(x,y)\leqslant\delta\}$.因 $0\leqslant\gamma\leqslant\delta\leqslant1$,易知 $g_a(x)_\gamma\subseteq g_a(x)_\delta$. 证毕.

定理 2. 设分类系统为 $C=(S,F,L)$,对于样本 $\forall x\in S$,特征子集 $\forall P\subseteq F$,$g_P(x)_\gamma$,$g_P(x)_\delta$ 为 x 在 P 上的邻域粒元素,若 $0\leqslant\gamma\leqslant\delta\leqslant1$,则 $|g_P(x)_\gamma|\leqslant|g_P(x)_\delta|$.

证明.对于 $\forall a\in F$ 和 $0\leqslant\gamma\leqslant\delta\leqslant1$,根据定理 1,则 $g_a(x)_\gamma\subseteq g_a(x)_\delta$.因此,$|g_a(x)_\gamma|\leqslant|g_a(x)_\delta|$ 成立.对于 $\forall P\subseteq F$,根据定义 6,可知:

$$|\boldsymbol{V}_P(x)_\delta|=\sqrt{\sum_{i=1}^{m}|g_{a_i}(x)_\delta|^2},$$

$$|\boldsymbol{V}_P(x)_\gamma|=\sqrt{\sum_{i=1}^{m}|g_{a_i}(x)_\gamma|^2}.$$

由 $|g_a(x)_\gamma|\leqslant|g_a(x)_\delta|$,则 $|g_P(x)_\gamma|\leqslant|g_P(x)_\delta|$ 成立. 证毕.

例 1. 分类系统 $C=(S,F,L)$ 如表 1 所示,$S=\{x_1,x_2,x_3,x_4\}$ 为样本集合,$F=\{a,b,c\}$ 为特征集合,$L=\{l\}$ 为类别标签.设邻域粒化参数 $\delta=0.1$.

Table 1 A Neighborhood Classification System

表 1 邻域分类系统

S	a	b	c	l
x_1	0.1	0.2	0.1	1
x_2	0.2	0.5	0.2	1
x_3	0.3	0.3	0.3	0
x_4	0.7	0.1	0.3	0

样本集 $S=\{x_1,x_2,x_3,x_4\}$,若按照特征 a 进行邻域粒化,则邻域粒子分别为 $g_1=g_a(x_1)_{0.1}=\{x_1,x_2\}$,$g_2=g_a(x_2)_{0.1}=\{x_1,x_2,x_3\}$,$g_3=g_a(x_3)_{0.1}=\{x_2,x_3\}$,$g_4=g_a(x_4)_{0.1}=\{x_4\}$.

若按照特征 b 进行邻域粒化,则邻域粒子分别为 $g_5=g_b(x_1)_{0.1}=\{x_1,x_3,x_4\}$,$g_6=g_b(x_2)_{0.1}=\{x_2\}$,$g_7=g_b(x_3)_{0.1}=\{x_1,x_3\}$,$g_8=g_b(x_4)_{0.1}=\{x_1,x_4\}$.

若按照特征 c 进行邻域粒化,则邻域粒子分别为 $g_9=g_c(x_1)_{0.1}=\{x_1,x_2\}$,$g_{10}=g_c(x_2)_{0.1}=\{x_1,x_2,x_3,x_4\}$,$g_{11}=g_c(x_3)_{0.1}=\{x_2,x_3,x_4\}$,$g_{12}=g_c(x_4)_{0.1}=\{x_2,x_3,x_4\}$.

若 $P=\{a,b,c\}$,则 x_1 在 P 上的粒向量为

$$\boldsymbol{V}_P(x_1)_\delta=(g_1,g_5,g_9)=$$
$$(g_a(x_1)_{0.1},g_b(x_1)_{0.1},g_c(x_1)_{0.1})=$$
$$(\{x_1,x_2\},\{x_1,x_3,x_4\},\{x_1,x_2\}).$$

该粒向量的大小为

$$|\boldsymbol{V}_P(x_1)_\delta|=\sqrt{2\times2+3\times3+2\times2}=4.123.$$

x_2 在 P 上的粒向量为

$$\boldsymbol{V}_P(x_2)_\delta=(g_a(x_2)_{0.1},g_b(x_2)_{0.1},g_c(x_2)_{0.1})=$$
$$(\{x_1,x_2,x_3\},\{x_2\},\{x_1,x_2,x_3,x_4\}),$$

该粒向量的大小为

$$|\boldsymbol{V}_P(x_2)_\delta|=\sqrt{3\times3+1\times1+4\times4}=5.099.$$

x_3 在 P 上的粒向量为

$$\boldsymbol{V}_P(x_3)_\delta=(g_a(x_3)_{0.1},g_b(x_3)_{0.1},g_c(x_3)_{0.1})=(\{x_2,x_3\},\{x_1,x_3\},\{x_2,x_3,x_4\}).$$

该粒向量的大小为

$$|\boldsymbol{V}_P(x_3)_\delta|=\sqrt{2\times2+2\times2+3\times3}=4.123.$$

x_4 在 P 上的粒向量为

$$\boldsymbol{V}_P(x_4)_\delta=(g_a(x_4)_{0.1},g_b(x_4)_{0.1},g_c(x_4)_{0.1})=(\{x_4\},\{x_1,x_4\},\{x_2,x_3,x_4\}).$$

该粒向量的大小为

$$|\boldsymbol{V}_P(x_4)_\delta|=\sqrt{1\times1+2\times2+3\times3}=3.742.$$

2 基于粒向量的 K 近邻粒分类器

采用邻域粒化,将 1 个样本粒化为 1 个邻域粒向量,邻域粒向量和类别标签组成 1 条粒规则,所有样本的粒规则构成粒规则库.通过定义粒距离度量,提出 K 近邻粒子的概念,从而将分类问题转化为 K 近邻粒向量的搜索问题.1 个测试样本粒化为 1 个邻域粒向量,在粒规则库中搜索该粒向量的 K 近邻粒向量,K 近邻粒向量中数量最多的类别标签即为测试样本的预测标签.

2.1 粒向量的运算

定义 7. 设分类系统为 $C=(S,F,L)$. $\forall x\in S$,存在 F 上的 δ 邻域粒向量 $\boldsymbol{V}_F(x)_\delta$,则在 F 上所有粒向量的集合称为粒向量组,定义为

$$Z_F\delta=\{\boldsymbol{V}_F(x)\delta\mid\forall x\in S\}.$$

定义 8. 设分类系统为 $C=(S,F,L)$,其中特征集为 $F=(a_1,a_2,\cdots,a_m)$.对于 $\forall x,y\in S$,存在 F 上的 2 个 δ 邻域粒向量为

$$\boldsymbol{V}_F(x)_\delta=(g_{a_1}(x)_\delta,g_{a_2}(x)_\delta,\cdots,g_{a_m}(x)_\delta),$$

$$\boldsymbol{V}_F(y)_\delta=(g_{a_1}(y)_\delta,g_{a_2}(y)_\delta,\cdots,g_{a_m}(y)_\delta),$$

则 2 个粒向量的交、并、减与异或运算定义为

$$\boldsymbol{V}_F(x)_\delta\wedge\boldsymbol{V}_F(y)_\delta=(g_{a_1}(x)_\delta\wedge g_{a_1}(y)_\delta,g_{a_2}(x)_\delta\wedge g_{a_2}(y)_\delta,\cdots,g_{a_m}(x)_\delta\wedge g_{a_m}(y)_\delta);$$

$$\boldsymbol{V}_F(x)_\delta\vee\boldsymbol{V}_F(y)_\delta=(g_{a_1}(x)_\delta\vee g_{a_1}(y)_\delta,g_{a_2}(x)_\delta\vee g_{a_2}(y)_\delta,\cdots,g_{a_m}(x)_\delta\vee g_{a_m}(y)_\delta);$$

$$\boldsymbol{V}_F(x)_\delta-\boldsymbol{V}_F(y)_\delta=(g_{a_1}(x)_\delta-g_{a_1}(y)_\delta,g_{a_2}(x)_\delta-g_{a_2}(y)_\delta,\cdots,g_{a_m}(x)_\delta-g_{a_m}(y)_\delta);$$

$$\boldsymbol{V}_F(x)_\delta\oplus\boldsymbol{V}_F(y)_\delta=(g_{a_1}(x)_\delta\oplus g_{a_1}(y)_\delta,g_{a_2}(x)_\delta\oplus g_{a_2}(y)_\delta,\cdots,g_{a_m}(x)_\delta\oplus g_{a_m}(y)_\delta).$$

2.2 粒向量的距离度量及粒规则

定义 9. 设分类系统为 $C=(S,F,L)$,其中特征集为 $F=(a_1,a_2,\cdots,a_m)$.对于 $\forall x,y\in S$,存在 F 上的 2 个 δ 邻域粒向量为

$$\boldsymbol{V}_F(x)_\delta=(g_{a_1}(x)_\delta,g_{a_2}(x)_\delta,\cdots,g_{a_m}(x)_\delta),$$

$$\boldsymbol{V}_F(y)_\delta=(g_{a_1}(y)_\delta,g_{a_2}(y)_\delta,\cdots,g_{a_m}(y)_\delta),$$

则 2 个粒向量的相对距离定义为

$$d(\boldsymbol{V}_F(x)_\delta,\boldsymbol{V}_F(y)_\delta)=\frac{1}{|F|}\sum_{i=1}^{m}\frac{|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|}{|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta|}=\frac{1}{|F|}\left(\frac{|g_{a_1}(x)_\delta\oplus g_{a_1}(y)_\delta|}{|g_{a_1}(x)_\delta\vee g_{a_1}(y)_\delta|}+\cdots+\frac{|g_{a_m}(x)_\delta\oplus g_{a_m}(y)_\delta|}{|g_{a_m}(x)_\delta\vee g_{a_m}(y)_\delta|}\right).$$

定理 3. 任意 2 个邻域粒向量 $\boldsymbol{P}=\boldsymbol{V}_F(x)_\delta$,$\boldsymbol{Q}=\boldsymbol{V}_F(y)_\delta$ 的相对距离满足:$0\leqslant d(\boldsymbol{P},\boldsymbol{Q})\leqslant1$.

证明.设 $s=g_{a_i}(x)_\delta$,$t=g_{a_i}(y)_\delta$ 由 $|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|=|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta-g_{a_i}(x)_\delta\wedge g_{a_i}(y)_\delta|$,可知:

$$\frac{|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|}{|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta|}=\frac{|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta-g_{a_i}(x)_\delta\wedge g_{a_i}(y)_\delta|}{|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta|},$$

则 $0\leqslant\frac{|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|}{|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta|}\leqslant1$.由 $F=(a_1,a_2,\cdots,a_m)$,可知 $|F|=m$.因此,$0\leqslant\sum_{i=1}^{m}\frac{|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|}{|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta|}\leqslant|F|$,则 $0\leqslant\frac{1}{|F|}\sum_{i=1}^{m}\frac{|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|}{|g_{a_i}(x)_\delta\vee g_{a_i}(y)_\delta|}\leqslant1$.由定义 9,因 $\boldsymbol{P}=\boldsymbol{V}_F(x)_\delta$,$\boldsymbol{Q}=\boldsymbol{V}_F(y)_\delta$,则 $0\leqslant d(\boldsymbol{P},\boldsymbol{Q})\leqslant1$ 成立. 证毕.

定义 10. 设分类系统为 $C=(S,F,L)$,其中特征集 $F=\{a_1,a_2,\cdots,a_m\}$.对于 $\forall x,y\in S$,存在 F 上的 2 个 δ 邻域粒向量为 $\boldsymbol{V}_F(x)_\delta=(g_{a_1}(x)_\delta,g_{a_2}(x)_\delta,\cdots,g_{a_m}(x)_\delta)$,$\boldsymbol{V}_F(y)_\delta=(g_{a_1}(y)_\delta,g_{a_2}(y)_\delta,\cdots,g_{a_m}(y)_\delta)$,则 2 个粒向量的绝对距离定义为

$$h(\boldsymbol{V}_F(x)_\delta,\boldsymbol{V}_F(y)_\delta)=\frac{1}{|F|\times|S|}\sum_{i=1}^{m}|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|=\frac{1}{|F|\times|S|}(|g_{a_i}(x)_\delta\oplus g_{a_i}(y)_\delta|+\cdots+|g_{a_m}(x)_\delta\oplus g_{a_m}(y)_\delta|).$$

定理 4. 任意 2 个邻域粒向量 $\boldsymbol{P}=\boldsymbol{V}_F(x)_\delta$,$\boldsymbol{Q}=\boldsymbol{V}_F(y)_\delta$ 的绝对距离满足:$0\leqslant h(\boldsymbol{P},\boldsymbol{Q})\leqslant1$.

证明.设 $s=g_{a_i}(x)_\delta$,$t=g_{a_i}(y)_\delta$,由 $|g_{a_i}(x)_\delta\oplus$

$g_{a_i}(y)_\delta|=|g_{a_i}(x)_\delta \vee g_{a_i}(y)_\delta - g_{a_i}(x)_\delta \wedge g_{a_i}(y)_\delta|$,可知,$0\leqslant|g_{a_i}(x)_\delta \oplus g_{a_i}(y)_\delta|\leqslant|S|$.由 $F=(a_1,a_2,\cdots,a_m)$,知 $|F|=m$,则 $0\leqslant\sum_{i=1}^{m}|g_{a_i}(x)_\delta \oplus g_{a_i}(y)_\delta|\leqslant|F|\times|S|$.由:

$$h(\boldsymbol{V}_F(x)_\delta,\boldsymbol{V}_F(y)_\delta)=\frac{1}{|F|\times|S|}\sum_{i=1}^{m}|g_{a_i}(x)_\delta \oplus g_{a_i}(y)_\delta|,$$

则 $0\leqslant h(\boldsymbol{V}_F(x)_\delta,\boldsymbol{V}_F(y)_\delta)\leqslant 1$ 成立.因 $\boldsymbol{P}=\boldsymbol{V}_F(x)_\delta$,$\boldsymbol{Q}=\boldsymbol{V}_F(y)_\delta$,则 $0\leqslant h(\boldsymbol{P},\boldsymbol{Q})\leqslant 1$ 成立. 证毕.

定义 11. 设分类系统为 $C=(S,F,L)$,$\forall x\in S$,存在 F 上的 δ 邻域粒向量 $\boldsymbol{V}_F(x)_\delta$.设 $l_x\in L$ 为样本 x 的类别标签,则一条邻域粒向量规则定义为序对:$r(x)=\langle\boldsymbol{V}_F(x)_\delta,l_x\rangle$,邻域粒向量规则库定义为:$B=\{r(x)\mid\forall x\in S\}$.

分类系统通过邻域参数粒化为单特征邻域粒子,多个单特征邻域粒子构成邻域粒向量.邻域粒向量与其类别标签形成一条粒向量规则,粒向量规则的集合构成了粒向量规则库.因此,分类过程则可转化为粒向量规则库中的推理、搜索与匹配过程.

邻域粒向量的距离可以作为粒向量的相似性度量,表示邻域粒向量的相似程度,粒向量距离越小,2 个粒向量的相似度越大;反之,粒向量距离越大,2 个粒向量的相似度越小.

2.3 K 近邻粒向量组

定义 12. 设分类系统为 $C=(S,F,L)$,Z 为 F 上的邻域粒向量组,$k>0$ 的正整数,对于任一 δ 邻域粒向量 $\boldsymbol{P}\in Z$,则该粒向量 $\boldsymbol{P}$ 的 K 近邻粒向量组定义为

$$V_{KNN}(\boldsymbol{P},Z)=\{I\subseteq Z\mid\forall\boldsymbol{T}_i\in I,\forall\boldsymbol{T}_j\in Z-I,(|I|=k)\wedge(d(\boldsymbol{P},\boldsymbol{T}_i)\leqslant d(\boldsymbol{P},\boldsymbol{T}_j))\}.$$

邻域粒向量组是所有邻域粒向量的集合,邻域粒向量 $\boldsymbol{P}$ 的 K 近邻粒向量组是邻域粒向量组的子集,只是部分邻域粒向量的集合,即为邻域粒向量组中离邻域粒向量 $\boldsymbol{P}$ 最近的 k 个邻域粒向量.根据邻域粒向量的距离可以定义了 2 种 K 近邻粒向量组,第 1 种是基于绝对距离的 K 近邻粒向量组,第 2 种是基于相对距离的 K 近邻粒向量组.

定义 13. 设分类系统为 $C=(Tr\cup Te,F,L)$,其中 Tr 为训练集,Te 为测试集.设 Z 为训练集 Tr 上的粒向量组.对于测试样本 $\forall x\in Te$,单特征 $\forall a_i\in F$,测试样本 x 在训练集 Tr 中特征 a_i 上的 δ 邻域测试粒子为 $g_{a_i}(x)_\delta=\{y\mid x\in Te,y\in Tr,D_{a_i}(x,y)\leqslant\delta\}$,则测试样本 x 在训练集 Tr 中的测试粒向量为 $\boldsymbol{R}=\boldsymbol{V}_F(x)_\delta=(g_{a_1}(x)_\delta,\cdots,g_{a_i}(x)_\delta,\cdots,g_{a_m}(x)_\delta)$,则测试粒向量 $\boldsymbol{R}$ 的 K 近邻粒向量组定义为

$$V_{KNN}(\boldsymbol{R},Z)=\{I\subseteq Z\mid\forall\boldsymbol{T}_i\in I,\forall\boldsymbol{T}_j\in Z-I,(|I|=k)\wedge(d(\boldsymbol{R},\boldsymbol{T}_i)\leqslant d(\boldsymbol{R},\boldsymbol{T}_j))\}.$$

测试粒向量 $\boldsymbol{R}$ 是测试样本 x 在训练集粒化而形成的邻域粒向量.测试粒向量 $\boldsymbol{R}$ 的 K 近邻粒向量组是训练集粒向量组中离测试粒向量 $\boldsymbol{R}$ 最近的 k 个邻域粒向量.

2.4 K 近邻粒分类器

通过对测试粒向量 $\boldsymbol{R}$ 的 K 近邻粒向量组定义,我们可知,测试粒向量的分类可转为在粒向量组中寻找与匹配 k 个最近邻粒向量的过程.根据粒向量组和粒向量规则库的定义可知,带有类别标签的粒向量的集合即为粒向量规则库.因此,测试粒向量的分类可转化为在粒向量规则库中的搜索与匹配 k 个最近邻粒向量的过程.

定义 14. 设分类系统为 $C=(Tr\cup Te,F,L)$,其中 Tr 为训练集,Te 为测试集.设 B 为训练集 Tr 在特征集 F 上粒向量规则库.对于测试样本 $\forall x\in Te$,测试样本 x 在训练集 Tr 中 δ 邻域测试粒向量为 $\boldsymbol{R}=\boldsymbol{V}_F(x)_\delta=(g_{a_1}(x)_\delta,g_{a_2}(x)_\delta,\cdots,g_{a_i}(x)_\delta,\cdots,g_{a_m}(x)_\delta)$,其中 $g_{a_i}(x)_\delta=\{y\mid x\in Te,y\in Tr,D_{a_i}(x,y)\leqslant\delta\}$,则测试粒向量 $\boldsymbol{R}$ 的 K 近邻粒向量规则组定义为

$$V_{KNL}(\boldsymbol{R},B)=\{I\subseteq B\mid\forall\boldsymbol{T}_i\in I,\forall\boldsymbol{T}_j\in B-I,(|I|=k)\wedge(d(\boldsymbol{R},\boldsymbol{T}_i)\leqslant d(\boldsymbol{R},\boldsymbol{T}_j))\}.$$

定义 15. 设分类系统为 $C=(Tr\cup Te,F,L)$,其中 Tr 为训练集,Te 为测试集.对于测试样本 $x\in Te$ 的邻域测试粒向量 $\boldsymbol{R}$,$L_d(\boldsymbol{R})$ 为测试粒向量 $\boldsymbol{R}$ 的 K 近邻粒向量规则组 $V_{KNL}(\boldsymbol{R},B)$ 中的类别标签集合,则 $L_d(\boldsymbol{R})$ 中最多的类别标签为测试样本 x 的类别标签.

3 基于粒向量的 K 近邻粒分类器设计

K 近邻粒分类器是一种基于集合运算的分类器,分为粒化、匹配与分类过程.下面论述 K 近邻粒分类器的原理,并给出具体的 K 近邻粒分类算法.

3.1 K 近邻粒分类器的原理

K 近邻粒分类器没有训练过程,只有粒化过程、匹配过程和分类过程.粒化过程包括:数据预处理、划分训练集和测试集、训练集粒化为粒向量规则库、测试集粒化为测试粒向量集合.粒匹配过程包括:测试粒向量与粒向量规则的距离计算,按照粒距

离进行排序,选出 k 个最近邻的粒向量规则.分类过程则包括:判定测试粒向量的类别.具体的粒化、匹配与分类过程为:

1) 数据集预处理.删去存在缺失值的数据,对数据集归一化为 0~1 之间的数值.

2) 划分训练集与测试集.按照训练集 80%与测试集 20%的规则进行划定.

3) 邻域粒化训练集.设定粒化参数,根据单特征粒化训练集,形成粒向量规则库.

4) 邻域粒化测试集.取出一个测试样本,在每个单特征上计算该测试样本与每个训练样本的距离,形成一个测试粒向量,所有的测试样本粒化为测试粒向量集合.

5) 粒向量的搜索与匹配.取出一个测试粒向量,计算该测试粒向量与规则库中每条粒向量规则的距离,按照距离升序排序,选出 k 个最近邻的粒向量规则.

6) 测试粒向量的类别判断.k 个最近邻的粒向量规则中最多的类别标签则为测试粒向量的类别标签.

7) 所有测试粒向量的分类.转步骤 5,进行下一个测试粒向量的分类,直到所有测试粒向量分类完毕.

从分析可知,K 近邻粒分类器类似于传统 KNN 分类器,不需要使用训练集进行训练,训练时间复杂度为 0.

3.2 K 近邻粒分类算法

根据前述 K 近邻粒分类器的原理与步骤,从而可设计出 K 近邻粒分类器,具体的 K 近邻粒分类算法描述如算法 1 所示:

算法 1. K 近邻粒分类算法(VKNG).

输入:训练集 $C=(Tr,F,L)$、测试样本 t,k 值及邻域参数 δ;

输出:测试样本 t 的类别标签 l_t.

1) 训练集和测试样本归一化:$Tr,t\in[0,1]$;

2) 对每一个训练样本 $x\in Tr$ 循环执行步骤 3)~6):

3) 在每个单原子特征 $a_i\in F$ 上分别进行 δ 邻域粒化为 $g_{a_i}(x)_\delta$;

4) 形成 x 的 δ 邻域粒向量

$\boldsymbol{V}_F(x)_\delta=(g_{a_1}(x)_\delta,g_{a_2}(x)_\delta,\cdots,g_{a_m}(x)_\delta)$;

5) 获取 x 的类别标签 l_x;

6) 构造粒向量规则 $R(x)=\langle\boldsymbol{V}_F(x)_\delta,l_x\rangle$,插入到粒向量规则库 B 中;

7) 在训练集中对测试样本 t 进行粒化,形成测试粒向量 $\boldsymbol{V}_F(t)_\delta$;

8) 对每个粒向量规则 $R(x)$循环执行步骤 9),10);

9) 根据定义 9 或定义 10 计算粒向量距离 $d(\boldsymbol{V}_F(x)_\delta,\boldsymbol{V}_F(t)_\delta)$或 $h(\boldsymbol{V}_F(x)_\delta,\boldsymbol{V}_F(t)_\delta)$;

10) 将粒向量规则$\langle\boldsymbol{V}_F(x)_\delta,l_x\rangle$和粒向量距离 d 或 h 插入临时变量 T 中;

11) 对 T 中的粒向量按粒向量距离进行升序排序;

12) 从 T 中选出排在前面的 k 个粒向量及其类别标签插入到目标变量 W 中;

13) 目标变量 W 中类别标签最多的那一类别判定为测试样本 t 的类别,设为 l_t;

14) 返回测试样本的类别标签 l_t.

在算法 VKNG 中,主要涉及邻域粒化过程.步骤 3)中训练集的邻域粒化采用 Hash 排序算法,时间复杂度为 $O(m\times n)$,其中 m 为特征的个数,n 为训练样本的个数;步骤 2)~6)时间复杂度则为 $O(m\times n^2)$;步骤 7)为测试集的邻域粒化,时间复杂度为 $O(m\times t^2)$,其中 t 为测试样本的个数,$t<n$;步骤 8)~14)都是线性复杂度,为 $O(m)$.因而,最坏情况下,VKNG 算法的时间复杂度为 $O(m\times n^2)$.根据采用相对粒距离与绝对粒距离的不同,VKNG 算法分为 VKNGR 算法与 VKNGA 算法.

4 实验分析

本文采用 UCI 数据集中 8 个数据集作为实验测试的数据源,如表 2 所示:

Table 2 Descriptions of Datasets

表 2 数据集描述

Datasets	Samples	Features	Categories
Ecoli	336	7	8
Glass	214	9	6
Iris	150	4	3
Pima	768	8	2
Seeds	210	7	3
Segmentation	210	19	7
WDBC	569	30	2
Wine	178	13	3

由于表 2 中数据集的值域不同,需要对数据集进行归一化预处理.我们采用最大最小值法,以确保

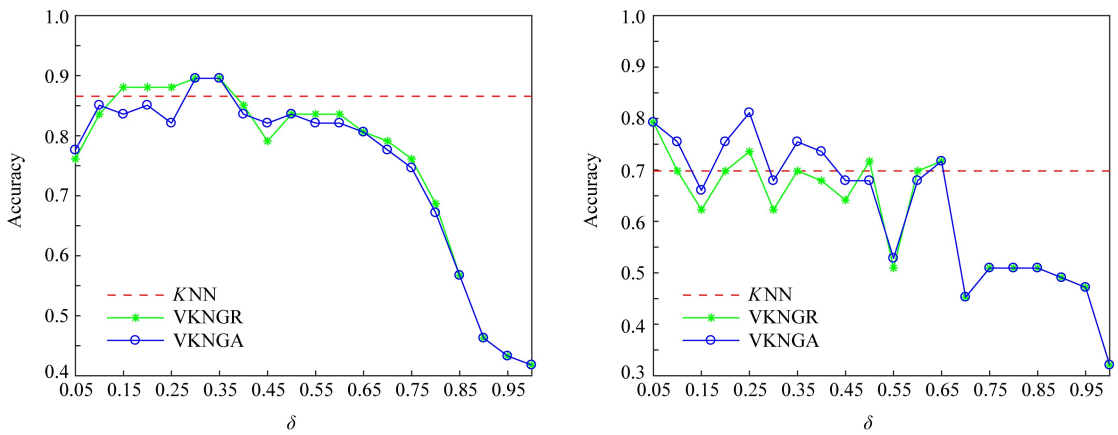
所有数据都能转化为[0,1]之间的数据.最大最小值归一化公式为

$$f(x_i)=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}}.$$

数据在每个单原子特征上进行邻域粒化,形成粒向量.分类算法分别采用传统的 KNN 分类器,基于相对粒向量距离的 K 近邻粒分类器 VKNGR 和基于绝对向量距离的 K 近邻粒分类器 VKNGA.为测试 K 近邻粒分类器的分类精度,每个数据集随机分成 5 份,其中一份为测试集,剩下的为训练集.然后再选另一份为测试集,剩下的为训练集,共测试 5 次,分类精度为 5 次的平均值.

4.1 邻域参数的影响

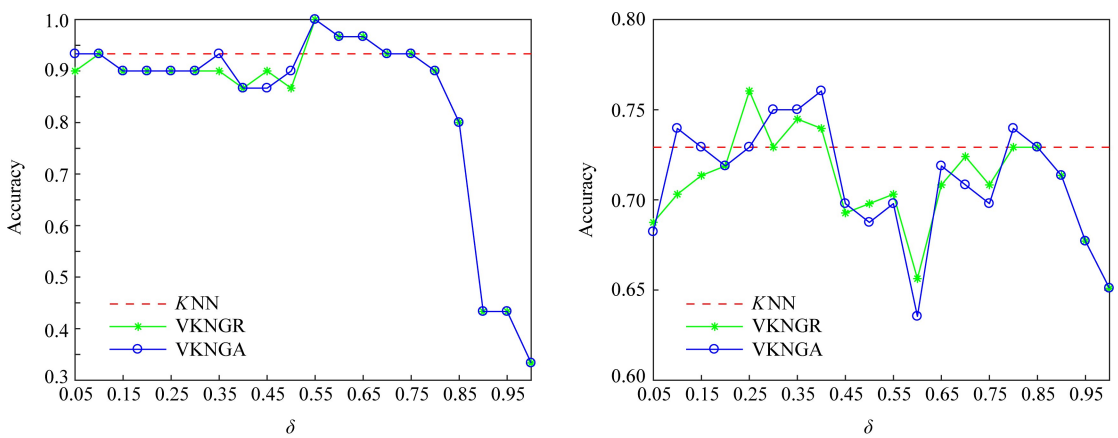
邻域粒化过程需要设置邻域粒化的参数,实验中邻域粒化参数以 0.05 为起点,0.05 为间隔,直到 1 为止.本节实验主要测试邻域参数的影响,k 值则固定,具体数值由实验确定.8 个 UCI 数据集的分类结果实验如图 1～4 所示:



(a) Ecoli(k=7) (b) Glass(k=1)

Fig. 1 Classification accuracy of different δ on datasets Ecoli and Glass

图 1 在数据集 Ecoli 和 Glass 上不同邻域参数的分类精度

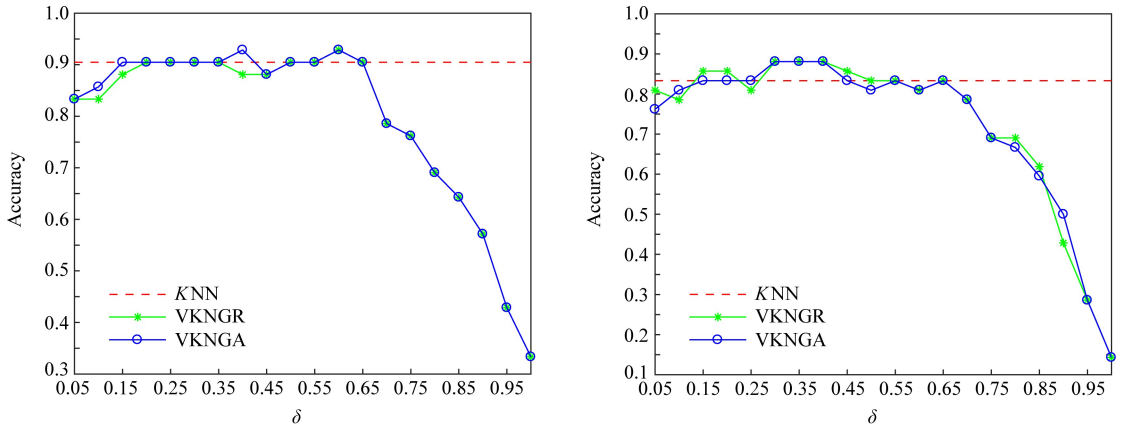


(a) Iris(k=1) (b) Pima(k=9)

Fig. 2 Classification accuracy of different δ on datasets Iris and Pima

图 2 在数据集 Iris 和 Pima 上不同邻域参数的分类精度

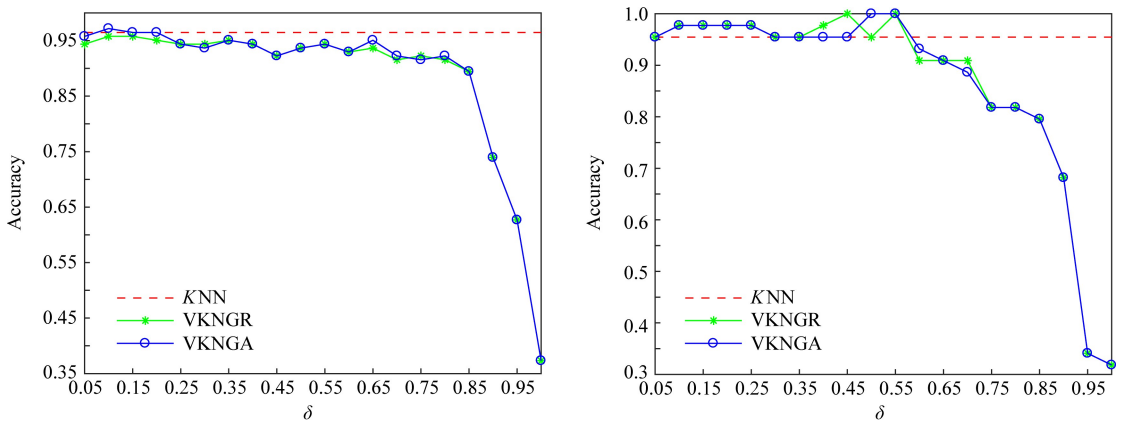
从图 1(a)可知,对于 Ecoli 数据集,传统 KNN 分类器在 k=7 时,分类精度为 0.8657;而对于 K 近邻粒分类器,邻域参数为 0.3 和 0.35,且 k=7 时,VKNGR 分类精度达到最大值为 0.8955,VKNGA 分类精度达到最大值为 0.8955.VKNGR 算法的分类精度略好于 VKNGA 算法.在邻域参数为 0.3～0.35 时,K 近邻粒分类器的分类效果较好,在邻域参数较小或较大的情况下分类效果不佳.从图 1(b)可知,对于 Glass 数据集,k=1 时,KNN 算法的分类精度为 0.6981,VKNGR 算法的最大分类精度为 0.7925,VKNGA 算法的最大分类精度为 0.8113,VKNGA 算法的分类精度略好于 VKNGR 算法.当



(a) Seeds(k=3) (b) Segmentation(k=4)

Fig. 3 Classification accuracy of different δ on datasets Seeds and Segmentation

图 3 在数据集 Seeds 和 Segmentation 上不同邻域参数的分类精度



(a) WDBC(k=4) (b) Wine(k=13)

Fig. 4 Classification accuracy of different δ on datasets WDBC and Wine

图 4 在数据集 WDBC 和 Wine 上不同邻域参数的分类精度

$k=1$ 时,在合适邻域参数下,K 近邻粒分类器好于 KNN.对于 K 近邻粒分类器,在邻域参数较小的情况下分类精度较好.

从图 2(a)可知,对于 Iris 数据集,当 $k=1$ 时,KNN 算法分类精度为 0.933 3,VKNGR 算法和 VKNGA 算法在邻域参数为 0.55 时分类精度为 1.VKNGA 算法略好于 VKNGR 算法.从图 2(b)可知,对于 Pima 数据集,当 $k=9$ 时,KNN 算法分类精度为 0.729 2;VKNGR 算法在邻域参数为 0.25 时,分类精度达到最大值,为 0.760 4;VKNGA 算法在邻域参数为 0.4 时分类精度达到最大值,为 0.760 4.可知在合适邻域参数下,VKNGA 和 VKNGR 粒分类器好于传统的 KNN 算法.

从图 3(a)可知,对于 Seeds 数据集,当 $k=3$ 时,KNN 算法分类精度为 0.904 8;VKNGR 算法在邻域参数为 0.6 时分类精度达到最大值,为 0.928 6;VKNGA 算法在邻域参数为 0.5 和 0.6 时分类精度达到最大值,为 0.928 6.从图 3(b)可知,对于 Segmentation 数据集,当 $k=4$ 时,KNN 算法分类精度为 0.833 3;VKNGR 和 VKNGA 算法在邻域参数为 0.3~0.4 时分类精度都达到最大值,为 0.881,且 VKNGR 算法略好于 VKNGA 算法.在合适邻域参数下,VKNGR 和 VKNGA 算法好于传统的 KNN 算法.

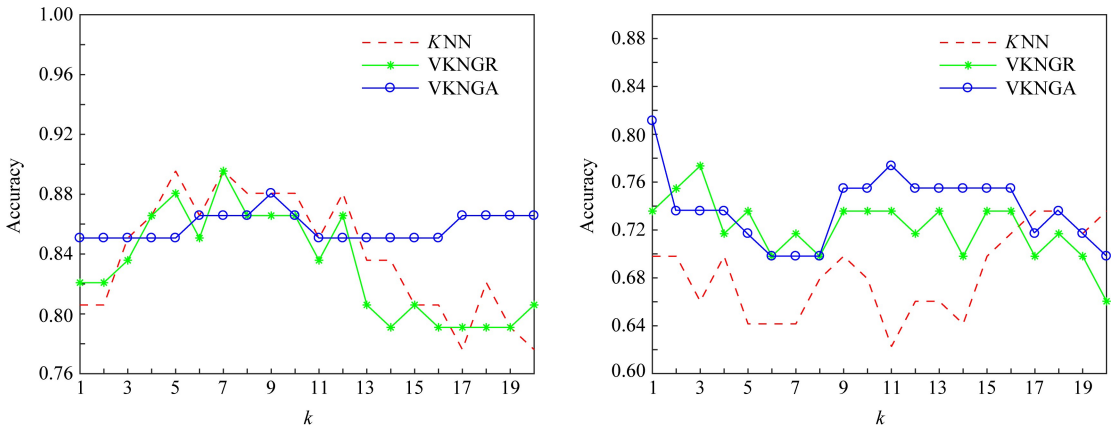
从图 4(a)可知,对于 WDBC 数据集,当 $k=4$ 时,KNN 算法分类精度为 0.964 8;VKNGA 算法在邻域参数为 0.1 时分类精度达到最大值,为 0.9718;VKNGA 算法略好于 VKNGR 算法.从图 4(b)可知,对于 Wine 数据集,当 $k=13$ 时,KNN 算法分类精度为 0.954 5;VKNGR 算法在邻域参数为 0.45 和

0.55 时分类精度达到最大值,为 1;VKNGA 算法在邻域参数为 0.5～0.55 时分类精度达到最大值,也为 1.可知在合适邻域参数下,VKNGR 算法和 VKNGA 算法好于传统的 KNN 算法.

从图 1～4 的实验分析可知,当数据集的类别数或特征数较多时,比如 Ecoli 和 WDBC 数据集,大部分情况下,粒分类器的分类效果不佳,只有某个邻域参数下,分类效果才好于 KNN 算法;当数据集的类别数较小时,比如 Iris,Pima,Seeds,Wine,粒分类器的分类效果较好,多数邻域参数情况下好于 KNN 算法.当邻域参数较大时,数据集的分类精度都较低,分类效果不理想.因此,邻域参数是粒分类器分类精度高低的关键因素之一.大部分情况下,VKNGA 算法的分类精度略好于 VKNGR 算法,说明粒向量的绝对距离度量效果略好于粒向量的相对距离度量.

4.2 k 值的影响

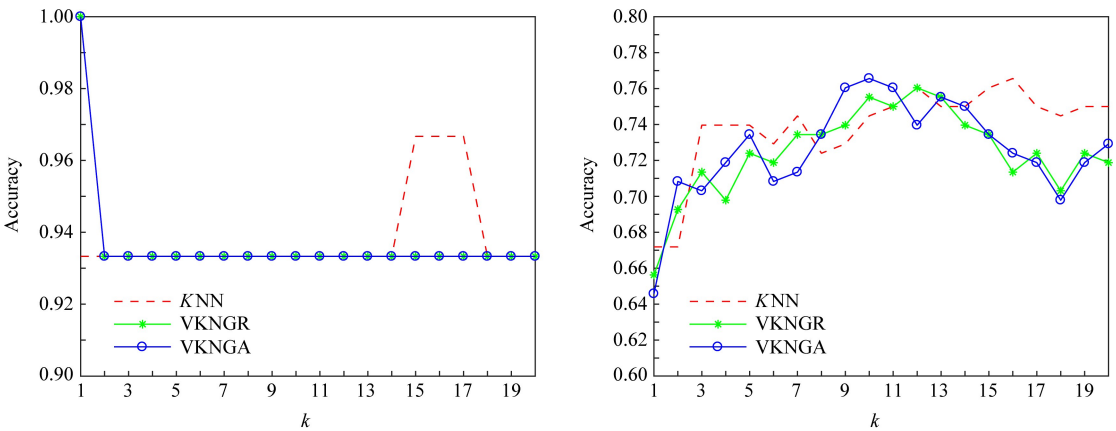
KNN 分类器中,k 值的选择影响着分类的精度.本节实验主要测试 k 值的影响,邻域参数则固定,具体数值由 4.1 节实验中分类精度最好情况下的邻域参数值.实验中 k 值从 1 变化到 20 为止,8 个 UCI 数据集的分类结果实验如图 5～8 所示.



(a) Ecoli(δ=0.3) (b) Glass(δ=0.25)

Fig. 5 Classification accuracy of different k on datasets Ecoli and Glass

图 5 在数据集 Ecoli 和 Glass 上不同 k 值的分类精度

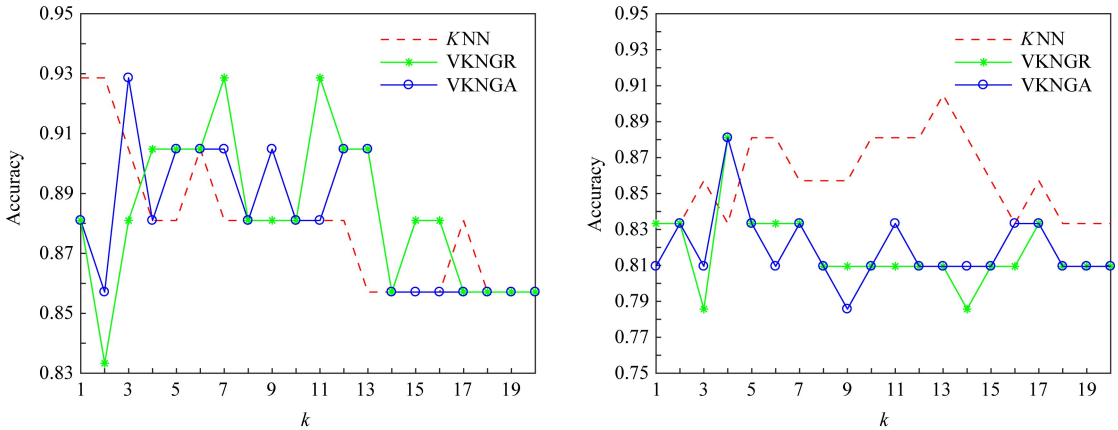


(a) Iris(δ=0.55) (b) Pima(δ=0.4)

Fig. 6 Classification accuracy of different k on datasets Iris and Pima

图 6 在数据集 Iris 和 Pima 上不同 k 值的分类精度

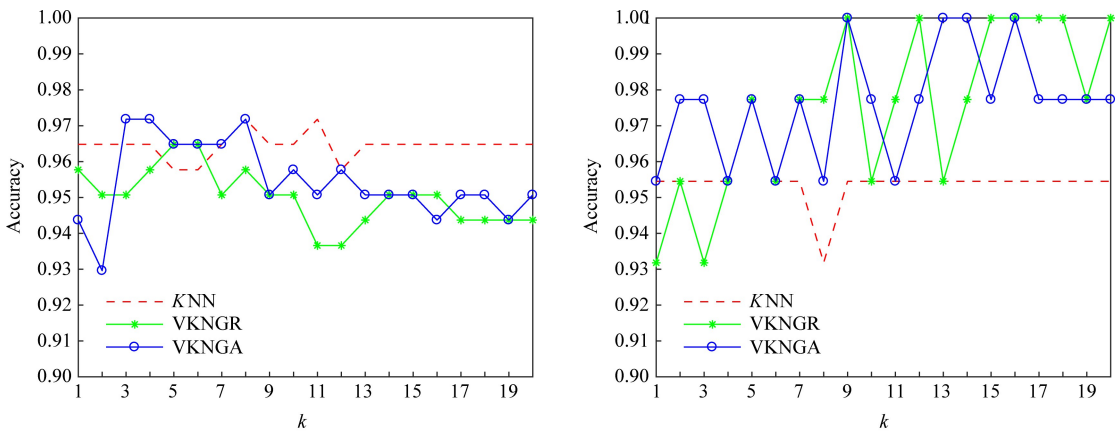
从图 5(a)可知,对于 Ecoli 数据集,k 值偏小时,VKNGR 算法和 VKNGA 算法的分类精度好于 KNN 算法;k 值处于 5～12 时,KNN 算法的分类精度好于 VKNGR 算法和 VKNGA 算法;k 值处于 13～20 时,VKNGA 算法好于 KNN 算法,而 KNN 算法好于 VKNGR 算法.大部分情况下,VKNGA 算法的分类精度好于 VKNGR 和 KNN 算法.从图 5(b)可知,对于 Glass 数据集,VKNGA 和 VKNGR



(a) Seeds(δ=0.4) (b) Segmentation(δ=0.35)

Fig. 7 Classification accuracy of different k on datasets Seeds and Segmentation

图 7 在数据集 Seeds 和 Segmentation 上不同 k 值的分类精度



(a) WDBC(δ=0.1) (b) Wine(δ=0.5)

Fig. 8 Classification accuracy of different k on datasets WDBC and Wine

图 8 在数据集 WDBC 和 Wine 上不同 k 值的分类精度

算法的分类精度都有 16 次高于 KNN 算法,而 VKNGA 算法的分类精度有 14 次高于 VKNGR 算法.因此,在不同的 k 值情况下,VKNGA 算法好于 KNGR 算法,KNGR 算法好于 KNN 算法.

从图 6(a)可知,对于 Iris 数据集,在 $k=1$ 时,VKNGA 算法和 VKNGR 算法的分类精度高于 KNN 算法.在 k 为 15,16 或 17 时,KNN 算法的分类精度高于 VKNGA 算法和 VKNGR 算法.从图 6(b)可知,对于 Pima 数据集,在 k 值为 6～13 时,VKNGA 算法和 VKNGR 算法的分类精度高于 KNN 算法;其他 k 值时,KNN 算法的分类精度高于 VKNGA 算法和 VKNGR 算法.因此,在合适的 k 值情况下,VKNGA 算法和 VKNGR 算法优于 KNN 算法.

从图 7(a)可知,对于 Seeds 数据集,VKNGR 算法的分类精度有 8 次高于 KNN 算法,VKNGA 算法的分类精度有 7 次高于 KNN 算法,而 KNN 算法的分类精度只有 2 次高于 VKGR 和 VKGA 算法.因此,大部分 k 值情况下,VKNGR 和 VKNGA 算法好于 KNN 算法.从图 7(b)可知,对于 Segmentation 数据集,大部分 k 值情况下,KNN 算法的分类精度好于 VKNGR 算法和 VKNGA 算法,VKNGA 算法的分类精度略好于 VKNGR 算法.

从图 8(a)可知,对于 WDBC 数据集,在 k 值为 3～8 时,VKNGA 算法的分类精度好于 VKNGR 算法和 KNN 算法;在 k 值偏大的情况下,KNN 算法的分类精度好于 VKNGR 和 VKNGA 算法;大部分情况下,VKNGA 算法的分类精度好于 VKNGR 算法.从图 8(b)可知,对于 Wine 数据集,VKNGA 算法的分类精度有 16 次高于 KNN 算法;VKNGR 算

法的分类精度有 12 次高于 KNN 算法.因此,VKNGA 算法和 VKNGR 算法都好于 KNN 算法.

从图 5～8 实验分析可知,不同 k 值的情况下,当数据集的类别数较少时,例如 Iris,Pima,Seeds,Wine 数据集,粒分类器的分类效果较好,大部分情况好于 KNN 算法;当数据集的类别数或特征数较多时,例如 Segmentation 和 WDBC,粒分类器的分类效果差于 KNN 算法.大部分情况下,VKNGA 算法的分类精度略好于 VKNGR 算法,说明粒向量的绝对距离度量效果好于粒向量的相对距离度量.k 值的选择也是分类的关键,设置过小会降低分类精度,设置过大则会增加噪声,降低分类效果.一般 k 值取 $1\sim n^{0.5}$(n 为训练集样本个数)的范围,然后采用交叉验证法来选取最优的 k 值.

5 总结与展望

传统的分类器是数值的计算,未涉及集合的运算,本文从研究样本的邻域粒化出发,提出了一种新型的集合形式的分类器:K 近邻粒向量分类器.首先,引入邻域粗糙集粒化方法,在分类系统中构建粒向量和粒规则,并定义了粒向量的大小度量与运算规则.

进一步提出了粒向量的 2 种距离度量:粒向量绝对距离与粒向量相对距离,并定义了 K 近邻粒向量的概念,设计了 K 近邻粒分类器.实验结果表明:新提出的 K 近邻粒分类器能够成功对样本进行分类,并在合适粒化参数下能够取得较好的分类性能.在未来的工作中,引入神经网络的方法进行参数调参,获取优化的邻域粒化参数,用于 K 近邻粒分类器的构建.还可以研究局部数据的粒化,构建局部邻域粒向量,将本文提出的分类方法应用于大数据系统的分类领域.

参 考 文 献

[1] Hart P E. The condensed nearest neighbor rule[J]. IEEE Transactions on Information Theory, 1968, 14(3): 515-516

[2] Kamencay P, Zachariasova M, Hudec R, et al. A novel approach to face recognition using image segmentation based on SPCA-KNN method[J]. Radioengineering, 2013, 22(1): 92-99

[3] Tan Songbo. An effective refinement strategy for KNN text classifier[J]. Expert Systems with Applications, 2006, 30(2): 290-298

[4] Liu Yaohui, Ma Zhengming, Yu Fang. Adaptive density peak clustering based on K nearest neighbors with aggregating strategy[J]. Knowledge-Based Systems, 2017, 133: 208-220

[5] Gallego A J, Calvo-Zaragoza J, Valero-Mas J J, et al. Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation[J]. Pattern Recognition, 2018, 74: 531-543

[6] Maillo J, Ramírez S, Triguero I, et al. kNN-IS: An iterative spark-based design of the k-nearest neighbors classifier for big data[J]. Knowledge-Based Systems, 2017, 117: 3-15

[7] Zhang Minling, Zhou Zhihua. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048

[8] Du Mingjing, Ding Shifei, Jia Hongjie. Study on density peaks clustering based on k-nearest neighbors and principal component analysis[J]. Knowledge-Based Systems, 2016, 99: 135-145

[9] Liu Rui, Wang Hong, Yu Xiaomei. Shared-nearest-neighbor-based clustering by fast search and find of density peaks[J]. Information Sciences, 2018, 450: 200-226

[10] Yuan Jiankui, Chen Weimin. A γ dose distribution evaluation technique using the k-d tree for nearest neighbor searching [J]. Medical Physics, 2010, 37(9): 4868-4873

[11] Li Can, Qian Jiangbo, Dong Yihong, et al. M2LSH: An LSH based technique for approximate nearest neighbor searching on high dimensional data[J]. Acta Electronica Sinica, 2017, 45(6): 1431-1442 (in Chinese)
(李灿, 钱江波, 董一鸿, 等. M2LSH:基于 LSH 的高维数据近似最近邻查找算法[J]. 电子学报, 2017, 45(6): 1431-1442)

[12] Bhattacharya G, Ghosh K, Chowdhury A S. Granger causality driven AHP for feature weighted kNN[J]. Pattern Recognition, 2017, 66: 425-436

[13] Hwang Wenjyi, Wen Kuowei. Fast kNN classification algorithm based on partial distance search[J]. Electronics Letters, 1998, 34(21): 2062-2063

[14] Li Ronglu, Hu Yunfa. A density-based method for reducing the amount of training data in kNN text classification[J]. Journal of Computer Research and Development, 2004, 41(4): 539-545 (in Chinese)
(李荣陆, 胡运发. 基于密度的 kNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展, 2004, 41(4): 539-545)

[15] Tan Songbo. Neighbor-weighted K nearest neighbor for unbalanced text corpus[J]. Expert Systems with Applications, 2005, 28(4): 667-671

[16] Ouyang Desheng, Li Dong, Li Qi. Cross-validation and non-parametric k nearest-neighbour estimation[J]. Econometrics Journal, 2010, 9(3): 448-471

[17] Yang Liu, Yu Jian, Jing Liping. An adaptive large margin nearest neighbor classification algorithm[J]. Journal of Computer Research and Development, 2013, 50(11): 2269-2277 (in Chinese)

(杨柳, 于剑, 景丽萍. 一种自适应的大间隔近邻分类算法[J]. 计算机研究与发展, 2013, 50(11): 2269-2277)

[18] Ghosh A K, Azen S P. On optimum choice of k in nearest neighbor classification[J]. Computational Statistics & Data Analysis, 2006, 50(11): 3113-3123

[19] Li Rong, Ye Shiwei, Shi Zhongzhi. SVM-KNN Classifier: A new method of improving the accuracy of SVM classifier[J]. Acta Electronica Sinica, 2002, 30(5): 745-748 (in Chinese)
(李蓉, 叶世伟, 史忠植. SVM-KNN 分类器: 一种提高 SVM 分类精度的新方法[J]. 电子学报, 2002, 30(5): 745-748)

[20] Aburomman A A, Reaz M B I. A novel SVM-kNN-PSO ensemble method for intrusion detection system[J]. Applied Soft Computing, 2016, 38: 360-372

[21] Kar S, Sharma K D, Maitra M. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive k-nearest neighborhood technique[J]. Expert Systems with Applications, 2015, 42(1): 612-627

[22] Sun Xiao, Pan Ting, Ren Fuji. Facial expression recognition using ROI-KNN deep convolutional neural networks[J]. Acta Automatica Sinica, 2016, 42(6): 883-891 (in Chinese)
(孙晓, 潘汀, 任福继. 基于 ROI-KNN 卷积神经网络的面部表情识别[J]. 自动化学报, 2016, 42(6): 883-891)

[23] Keller J M, Gray M R, Givens J A. A fuzzy k-nearest neighbor algorithm[J]. IEEE Transactions on Systems Man & Cybernetics, 1985, 15(4): 580-585

[24] Zadeh L A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets and Systems, 1997, 90(2): 111-127

[25] Lin Tsauyoung. Data mining: Granular computing approach [G] //LNCS 1574: Proc of the Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 1999: 24-33

[26] Lin Tsauyoung, Zadeh L A. Special issue on granular computing and data mining[J]. International Journal of Intelligent Systems, 2004, 19(7): 565-566

[27] Miao Duoqian, Xu Feifei, Yao Yiyu, et al. Set-theoretic formulation of granular computing[J]. Chinese Journal of Computers, 2012, 35(2): 351-363 (in Chinese)
(苗夺谦, 徐菲菲, 姚一豫, 等. 粒计算的集合论描述[J]. 计算机学报, 2012, 35(2): 351-363)

[28] Wang Guoyin, Zhang Qinghua, Ma Xiao, et al. Granular computer models for knowledge uncertainty[J]. Journal of Software, 2011, 22(4): 676-694 (in Chinese)
(王国胤, 张清华, 马希骜, 等. 知识不确定性问题的粒计算模型[J]. 软件学报, 2011, 22(4): 676-694)

[29] Xu Ji, Wang Guoyin, Yu Hong. Review of big data processing based on granular computing[J]. Chinese Journal of Computers, 2015, 38(8): 1497-1517 (in Chinese)
(徐计, 王国胤, 于洪. 基于粒计算的大数据处理[J]. 计算机学报, 2015, 38(8): 1497-1517)

[30] Yao Yiyu. Relational interpretations of neighborhood operators and rough set approximation operators[J]. Information Sciences, 1998, 111(1): 239-259

[31] Yao Yiyu, Zhang Nan, Miao Duoqian, et al. Set-theoretic approaches to granular computing[J]. Fundamenta Informaticae, 2012, 115: 247-264

[32] Hu Qinghua, Yu Daren, Xie Zongxia. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866-876

[33] Hu Qinghua, Yu Daren, Liu Jinfu, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594

[34] Zhu Pengfei, Hu Qinghua. Adaptive neighborhood granularity selection and combination based on margin distribution optimization[J]. Information Sciences, 2013, 249: 1-12

[35] Pedrycz W, Park B J, Oh S K. The design of granular classifiers: A study in the synergy of interval calculus and fuzzy sets in pattern recognition[J]. Pattern Recognition, 2008, 41(12): 3720-3735

[36] Roh S B, Pedrycz W, Ahn T C. A design of granular fuzzy classifier[J]. Expert Systems with Applications, 2014, 41(15): 6786-6795

[37] Zhong Chunfu, Pedrycz W, Wang Dan, et al. Granular data imputation: A framework of granular computing[J]. Applied Soft Computing, 2016, 46: 307-316

[38] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Sciences, 1982, 11(1): 341-356

Chen Yuming, born in 1977. PhD, professor. His main research interests include machine learning, rough sets, and granular computing.

Li Wei, born in 1979. PhD, associate professor. His main research interests include artificial intelligence, computer graphics, and granular computing.