第 28 卷 第 4 期
2002 年 7 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 4
July, 2002

短文

FCM-VKNN 聚类算法的研究[1)]

张洪刚 刘 刚 郭 军

(北京邮电大学信息工程学院 北京 100876)

(E-mail: guojun@bupt.edu.cn)

摘 要 提出了一种新的 K 值可以变化的 FCM-VKNN(Fuzzy C-Means Variable K-Nearest Neighbor)聚类算法. FCM-VKNN 聚类算法充分吸取了 FCM 算法和 KNN 准则的长处,使本算法不受初始值的影响和固定值 K 的束缚. 新的目标准则函数考虑了数据集样本的模糊隶属关系和样本几何分布两个方面的因素,使算法的鲁棒性和分类的正确性大大加强. 最后给出了几组具有代表性数据的聚类结果,实验结果表明了这种算法的有效性.

关键词 FCM-VKNN,准则函数,模糊隶属度

中图分类号 TP391

AN ALGORITHM OF FCM-VKNN CLUSTERING

ZHANG Hong-Gang LIU Gang GUO Jun

(*School of Information Engineering, Beijng University of Post and Telecommunications, Beijing* 100876)

(E-mail: guojun@bupt.edu.cn)

Abstract In this paper, a new FCM-VKNN algorithm of clustering is proposed. It inherits the good virtue of FCM and KNN, which can be immune to initial guesses and get rid of the influence of constant K. Two factors are considered in cluster-validity criterion to enhance the robustness of algorithm and the validity of clustering, one is fuzzy membership and the other is geometric property. The finality of the paper gives clustering result of some representative data sets. Experimental results show the validity of the new clustering algorithm.

Key words FCM-VKNN, cluster-validity function, fuzzy membership

1 引言

聚类算法广泛应用在模式识别、图像处理、自动控制等领域. 目前所用的聚类算法基本

1) 国家"863"高技术研究发展计划(2001AA4080)资助

收稿日期 2000-08-04 收修改稿日期 2002-02-28

上是 KNN(K-Nearest Neighbor)算法和其它基于 KNN 的算法[1],这些传统的聚类算法存在着很多不足和缺陷. 主要是对聚类结果对初始值极为敏感,难以取得全局最优. 为解决聚类结果对初始聚类中心敏感的问题,最近有学者提出了基于模拟退火算法的动态聚类和基于遗传算法的聚类算法. 它们不仅可以使聚类结果不依赖于初始聚类中心,而且具有全局最优性和很高的搜索效率. 但是这些算法复杂,若参数定义不正确,也无法取得正确的分类结果. 现在,国外有很多学者还致力于传统算法的研究[2,3],他们的目标是在于寻找一个可靠有效的聚类准则函数,研究的对象是各类样本个数分布比较均匀,而没有考虑分布极端的情况,算法缺乏一定的鲁棒性.

本文提出的聚类算法采用仍是传统的基于 KNN 的 FCM(Fuzzy C-Means)聚类算法,充分吸取传统的聚类算法的优点,同时对它的不足和缺陷进行了很大的改进. 不同之处是 K 不再是固定值,而是可以随聚类分布而变化,聚类的结果也不再由算法得出,而是在可能的个数区域内 $2\sim\sqrt{n}$ (n 为样本个数)反复迭代,最后根据结果得出最优的聚类个数和样本分类. FCM-VKNN 聚类算法解决了对初始值敏感的问题;反复迭代取最优解,又避免陷入局部最优;K 值可以变化,既保持了样本分布均匀时的有效性,又解决了样本分布不均匀时,将样本错误划分的问题.

本文所用一种新的目标准则函数,它考虑了两方面的因素,一方面是数据集样本的几何分布,主要体现聚类后的类间距离和类内距离;另一方面是样本的模糊隶属关系,并应用一定的模糊操作,反映样本分布的离散度和集中度. 我们对几组具有代表性数据进行了实验,实验结果是令人满意的. 为进一步说明算法的有效性,我们用 FCM-KNN 算法的结果进行了比较,比较的结果也反映了 FCM-VKNN 算法分类的准确性.

2 模糊分类[3,4]

假定给定数据集 $X=\{\boldsymbol{x}_1,\ \boldsymbol{x}_2,\cdots,\boldsymbol{x}_k,\ \cdots,\ \boldsymbol{x}_n\}$,其中 $\boldsymbol{x}_k$是第 k 个数据样本,x_k^j是第 k 个数据样本的第 j 个分量. 若将数据划分到 c 个类中,$\boldsymbol{x}_k$在类别 $i(1\leqslant i\leqslant c)$的隶属度定义为 μ_{ik},μ_{ik}有如下性质

$$\mu_{ik}\in[0,1],\quad \sum_{i=1}^{c}\mu_{ik}=1,\forall\ k,\quad 0<\sum_{k=1}^{n}\mu_{ik}<n,\ \forall\ i \tag{1}$$

利用准则函数 $J(U,V)$: $J(U,V)=\sum_{k=1}^{n}(\mu_{ik})^m(d_{ik})^2$, 得

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}},\qquad \boldsymbol{V}_i=\frac{\sum_{k=1}^{n}(\mu_{ik})^m\boldsymbol{x}_k}{\sum_{k=1}^{n}(\mu_{ik})^m} \tag{2}$$

其中 $m>1$ 为模糊加权指数,$d_{ik}=\|\boldsymbol{x}_k-\boldsymbol{V}_i\|$ 为欧式距离.

3 FCM-VKNN 聚类算法

3.1 目标准则函数的定义

FCM 算法实际上是一个寻找全局最优解的问题,怎样才能得到全局最优解,这就需要

一个准则函数的指导.本文所用的准则函数考虑到两个方面的因素,一方面是样本的几何分布,另一方面是样本的模糊隶属关系.

首先,由第 2 节和模糊集并的定义:$E_{\cup}=\{\boldsymbol{u}_1 \cup \boldsymbol{u}_2 \cup \cdots \cup \boldsymbol{u}_c\}$,根据最大隶属度,可得 c 个不同的模糊集的并

$$E_{\cup}=\{\max_i u_{i1},\ \max_i u_{i2},\cdots \max_i u_{ik}\cdots \max_i u_{in}\},\ 1\leqslant i\leqslant c,$$

其中$\max_i u_{ik}=\max(u_{1k},\ u_{2k}\cdots u_{ck})$.

$\max_i u_{ik}$是样本 k 隶属 c 个类中心的最大隶属度,也就是说,在 c 个聚类中心中,k 最应该隶属于哪一个类.如果样本 k 离第 j 个聚类中心最近,那么 $u_{jk}>u_{ik}(j\neq i)$,$\max_i u_{ik}=u_{jk}$,如果$\max_i u_{ik}$越接近最大值 1,样本 k 就越应该隶属于模糊类 i.反之,若样本隶属每个类的可能性都一样,则$\max_i u_{ik}$就越接近 $1/c$,样本的隶属就最不确定.$E_{\cup}$可以作为衡量聚类效果的标志之一,因此得到以下定义

定义 1.

$$compact(c)=\sum_{k=1}^{n}\max_{1\leqslant i\leqslant c} u_{ik} \tag{3}$$

$compact(c)$的大小反映出聚类效果的好坏.

其次,再考虑两个模糊集的交集,假设 u_i,u_j是两个模糊类,它们之间交集定义为

$$E_{\cap}^{ij}=u_i \cap u_j=\{\min(u_{i1},u_{j1}),\min(u_{i2},u_{j2})\cdots \min(u_{ik},\ u_{jk})\cdots \min(u_{in},u_{jn})\}.$$

交集也可以用来反映聚类结果的情况.我们可以得出另一个用来衡量聚类结果的定义

定义 2.

$$Separate(c)=\sum_{i=1}^{c-1}\sum_{l=1}^{c-i}\sum_{k=1}^{n}\min(u_{ik},u_{jk}),\quad j=l+i \tag{4}$$

$Separate(c)$聚类后样本的隶属关系.

以上是从模糊隶属的角度对数据集的分布作了数学上的表示,没有考虑到数据集的几何特性,衡量数据集的几何特性主要是样本间的距离和类之间的距离,再作以下定义

定义 3.

$$Dis(c)=\frac{D_{\max}}{D_{\min}}\sum_{k=1}^{c}\Big(\sum_{z=1}^{c}\|v_k-v_z\|\Big)^{-1} \tag{5}$$

其中 $D_{\max}=\max\{\|v_i-v_j\|\}$,$\forall\ i,j\in\{2,3,\cdots,c\}$.

综合各方面的考虑,我们定义准则函数为

$$J_{SC}(U,V)=Separate(c)/Compact(c)+\alpha Dis(c) \tag{6}$$

其中 α 为加权系数.

3.2　聚类算法描述

给定数据集 $X=\{\boldsymbol{x}_1,\ \boldsymbol{x}_2,\cdots,\boldsymbol{x}_k,\cdots,\boldsymbol{x}_n \mid \boldsymbol{x}_k\in R^p\}$,在没有给定任何样本分布信息的情况下进行聚类,我们采用迭代法.一般情况下,最佳的聚类个数不会超过 $\sqrt{n}$ 个,算法迭代的范围从 2 到 $\sqrt{n}$,每个类的样本个数在 K 基础上,根据数据集的几何分布自由变动.

算法流程如下

1) 初始化. $c=2$; $K=\dfrac{n}{c}-1$,$J_{SC}^{*}=-\infty$,$c^{*}=1$, 其中 c 为类的个数迭代变量,K 为聚类前每个类的平均样本个数,J_{SC}^{*}为 J_{SC}的最大值,c^{*} 为最佳的分类个数.

2) VKNN 分类

- 按照 KNN 算法进行分类,样本按距离分到 c 个类中,每个类中的样本个数为 k 个.
- 计算模糊中心和各样本的模糊隶属度.
- 根据模糊隶属度,重新调整各样本的类属,各类样本个数在 k 的基础上作上下浮动.
- 重新计算类中心和各样本的隶属度.

3) 利用式(6)计算 J_{SC},如果 $J_{SC}>J_{SC}^{*}$,则 $J_{SC}^{*}\leftarrow J_{SC}$,$c^{*}\leftarrow c$

4) $c=c+1$;如果 $c\leqslant c_{\max}$,转向第二步;否则聚类结束.

5) c^{*} 为所得的最佳的聚类结果.

4 实验结果分析与比较

4.1 实验结果及分析

为验证 FCM-VKNN 算法的有效性,我们对 4 组具有代表性的数据进行了实验. $s1$,$s2$ 是样本分布基本相同的数据集,$s1$ 的各类样本分布不均匀,$s2$ 的样本分布均匀. $s3$ 样本分布比较集中,IRIS 是包含三个类,每类有 50 个四维的样本数据,样本数据有所重叠. 正确的分类个数是 $s1$,$s2$,$s3$,IRIS 均分为三类.

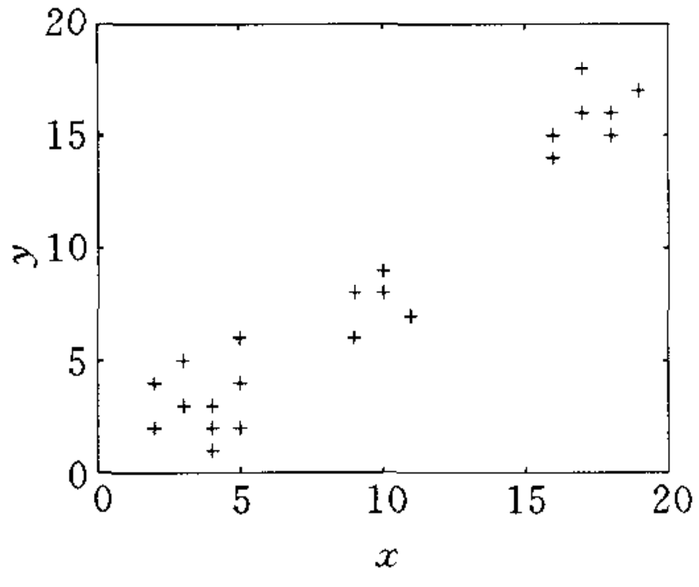

图 1 数据集 $s1$

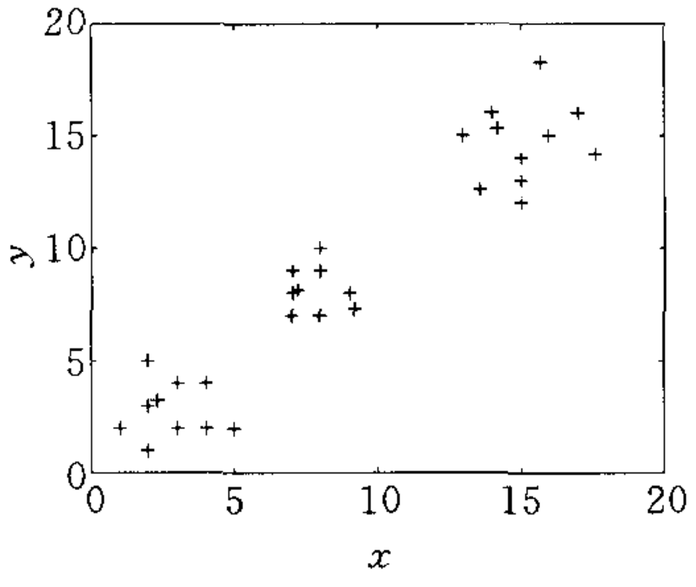

图 2 数据集 $s2$

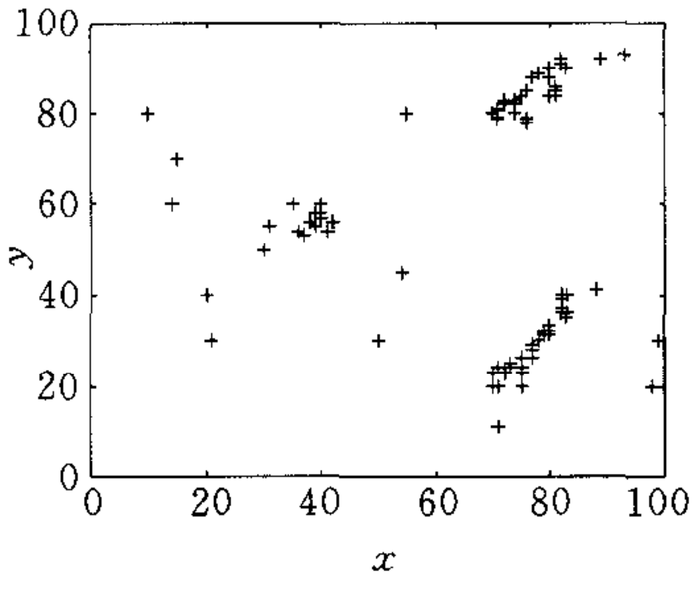

图 3 数据集 $s3$

表 1 中最大的 J_{SC}值对应的 c 值是实验得到的类的个数. 从上表可以看出对不同的数据集,分类的结果是正确的. 实验结果表明:对于样本分布均匀和不均匀的数据集 $s1$,$s2$,FCM-VKNN 算法都能正确的将各类分开,对于各类的样本分布比较收敛的 $s3$,J_{SC}的值递减的快,对于各类样本分布比较分散的 $s4$,J_{SC}的值递减的慢.

表 1 FCM-VKNN 算法实验结果

c	$s1$	$s2$	$s3$	IRIS
2	1.539 36	1.689 54	2.325 40	3.584 27
3	6.047 03	5.875 46	8.362 01	6.235 66
4	3.557 77	4.235 64	4.236 81	4.302 13
5	1.265 48	2.584 67	2.360 23	3.203 05
6	1.023 52	1.258 93	1.203 53	2.032 56
7	0.245 02	0.748 06	0.369 87	1.203 25
8	—	—	0.132 56	0.230 69

4.2　实验结果比较

研究 FCM-VKNN 的目的,不仅是获得正确分类的个数,还要得到各样本的正确划分. 我们用 FCM-KNN 对数据做了同样的实验,对部分样本也可以得到正确的分类个数,但是,在样本划分上却存在错误. 我们以数据集 $s1$ 为例,FCM-KNN 也可以得到 3 个类的结果,但样本划分和 FCM-VKNN 截然不同(见图 4 与图 5). 图 4 是 FCM-KNN 的样本分类结果,图 5 是 FCM-VKNN 的样本分类结果(结果中三类样本分别用 $*$,$+$ 和 x 表示). 从图 4 可以看出,使用 FCM-KNN 算法,由于 K 是固定的,中间一类的样本个数很少,FCM-KNN 错误的将其它类的离自己最近的部分样本划分为中间的一类,从而在样本划分上出现错误. 当使用 FCM-VKNN 算法时,由于 K 的灵活可变性,就不会出现错误的划分,其分类结果见图 5.

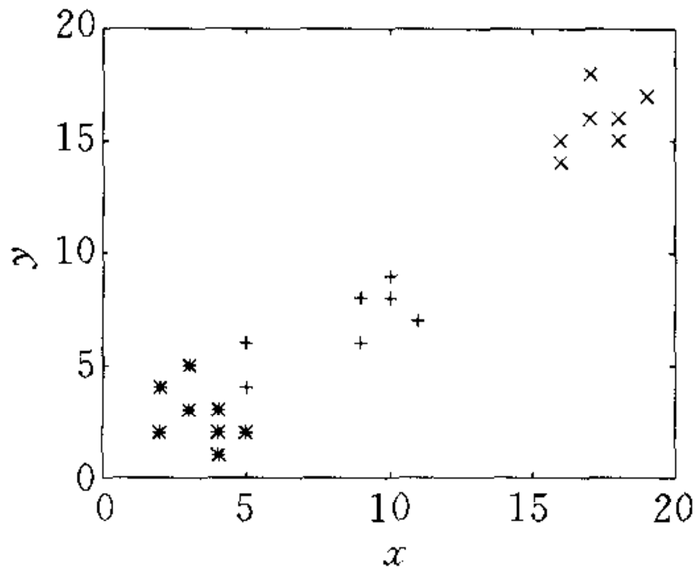

图 4　FCM－KNN 聚类结果

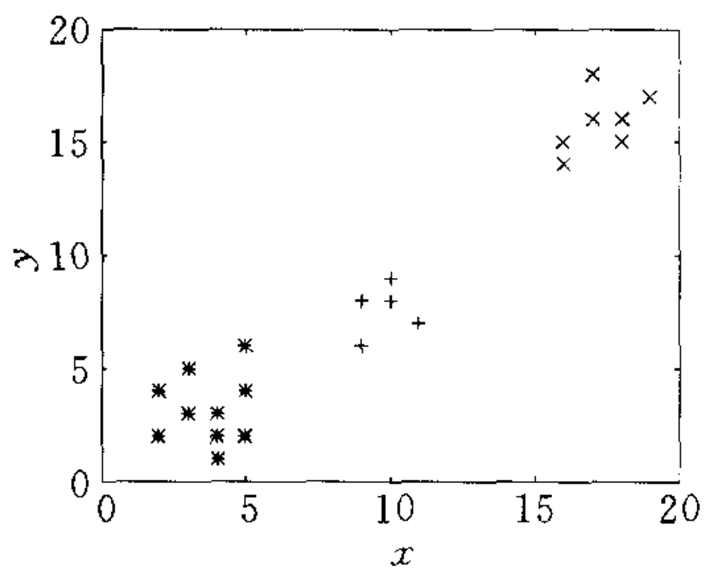

图 5　FCM－VKNN 聚类结果

4.3　算法的应用比较

在手写汉字识别领域,聚类算法广泛应用在样本训练过程中. 由于手写汉字变形大,一个汉字仅有一种特征向量表示是不够的,现在普遍采用多种特征模板,每种特征模板为该汉字的一种变形,反映到特征空间就是一个类中心. 特征训练就是对训练样本进行聚类分析,我们用 FCM-KNN 和 FCM-VKNN 两种不同的聚类算法训练了不同的特征模板,为了便于比较,采用同样的识别样本及识别算法. 识别算法采用到模板识别算法,识别样本及训练样本是北京邮电大学信息工程系建立的 HCL2000 汉字数据库,该库包含 2000 个人书写的一级汉字,也就是说每个汉字有 2000 个不同的样本. 我们取其中的 300 个样本集作为训练样本,5 个样本集作为识别样本. 识别率的比较见表 2.

表 2　FCM-KNN 与 FCM-VKNN 训练特征的识别率比较

识别样本集 \ 识别率	FCM-KNN 特征的识别率(%)	FCM-VKNN 特征的识别率(%)
样本 1	87.3	89.4
样本 2	94.8	96.1
样本 3	75.6	76.9
样本 4	82.1	85.3
样本 5	79.2	82.4

通过上表可以看出:FCM-VKNN 的特征模板的识别率要比 FCM-KNN 的高出 2%以上. 这说明 FCM-VKNN 聚类的结果基本上反映了手写汉字几种常见的变形,在样本划分上没有出现错误分类的情况;而 FCM-KNN 在聚类时,出现了样本错误分类的情况,因而训练的特征不能准确反映手写汉字特征,识别率略低. 在手写汉字识别领域,可以看出 FCM-VKNN 算法的有效性.

5 结论

通过对各种类型的数据集的测试，实验结果表明了这种聚类算法的有效性，另一方面也说明了本文所采用的准则函数的有效性. 准则函数的选取对聚类的结果有直接的影响，本文所采用的准则函数是从样本的模糊隶属关系和样本的几何分布两个方面进行考虑，也可以考虑从其它方面的内容，如：划分系数、划分熵、同一数据函数等. 这样可以对数据集作全面的数学表示，使分类更加准确.

参 考 文 献

1 Keller Gray, Givens M R. A fuzzy K-nearest neighbors algorithm. *IEEE Trans. System Man Cybernetics*, 1985, **15**: 580～585

2 Zahid H, Abouelala O. Unsupervised fuzzy clustering. *Pattern Recognition Letters*, 1999, **20**(2): 123～129

3 Ramze Rezaee M, Lelieveldt B P F. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters*, 1998, **19**(3,4): 237～246

4 Jun Guo, Ning Sun *et al*. Algorithm for recognition of handwritten characters using pattern transformation with cosine function. *IEICE Trans*, 1993, **J76-D-II**(4): 835-842

5 杨伦标，高英仪. 模糊数学. 广州：华南理工大学出版社，1992

张洪刚 北京邮电大学博士生. 主要研究方向模式识别.

刘 刚 北京邮电大学博士生. 主要研究方向语音识别.

郭 军 北京邮电大学教授，博士生导师. 主要研究方向模式识别、网络管理.