Computer Science and Application 计算机科学与应用, 2018, 8(6), 961-969
Published Online June 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.86107

Hans汉斯

The Research on an Attribute-Weighted Multi-Kernel Fuzzy Clustering Algorithm

Yun Kan, Zhenqiang Bao, Zhaoyue Zhang

College of Imformation Engineering, Yangzhou University, Yangzhou Jiangsu
Email: 1346600780@qq.com, yzbzq@163.com, 947756158@qq.com

Received: Jun. 7th, 2018; accepted: Jun. 22nd, 2018; published: Jun. 29th, 2018

Abstract

Considering the problem of unsatisfactory clustering effect of single kernel function for multiple data sources or heterogeneous data sets, and taking into account the difference in importance of different attributes for different categories, this paper proposes an attribute-weighted multi-kernel fuzzy clustering algorithm (AWMKFCM)). The algorithm combines the multi-kernel fuzzy clustering algorithm with the attribute-weighted single-kernal fuzzy clustering algorithm. It not only can handle the problem that the single kernel function can not meet the clustering accuracy requirements of the clustering data set, but also can adjust the importance of each attribute dynamically to different categories according to the specific characteristics of different types in the clustering process. Clustering experiments show that the attribute-weighted multi-kernal fuzzy clustering algorithm has higher clustering accuracy than the attribute-weighted single-kernel fuzzy clustering algorithm and multi-kernal fuzzy clustering algorithm under the premise of a certain amount of running time and iterations.

Keywords

Fuzzy Clustering, Multi-Kernal, Attribute-Weighted

属性加权多核模糊聚类算法研究

阚　云，包振强，张照岳

扬州大学信息工程学院，江苏　扬州
Email: 1346600780@qq.com, yzbzq@163.com, 947756158@qq.com

收稿日期：2018年6月7日；录用日期：2018年6月22日；发布日期：2018年6月29日

摘　要

针对多数据源或异构数据集，采用单个核函数的聚类效果不理想的问题，以及考虑到不同属性对不同类

文章引用: 阚云, 包振强, 张照岳. 属性加权多核模糊聚类算法研究[J]. 计算机科学与应用, 2018, 8(6): 961-969.
DOI: 10.12677/csa.2018.86107

别重要性的差异，本文提出了一种属性加权多核模糊聚类算法(WMKFCM)。该算法将多核模糊聚类算法与属性加权核模糊聚类算法相结合，不仅能够处理单个核函数不能满足待聚类数据集聚类准确度要求的问题，而且能在聚类过程中根据不同类的具体特性动态调整各个属性对于不同类别的重要性。聚类实验表明，在牺牲一定的运行时间和迭代次数的前提下，相比于属性加权核模糊聚类算法和多核模糊聚类算法，属性加权多核模糊聚类算法具有更高的聚类准确度。

关键词

模糊聚类，混合核函数，属性加权

1. 引言

聚类算法作为一种无监督学习方法，为识别数据内在结构提供了一种有用的工具，是数据挖掘的重要研究内容之一。聚类分析的一般过程是，根据数据之间的相似性将数据划分到不同的簇集，结果使得同一簇中的数据具有较高的相似度，而不同簇间的数据具有较低的相似度[1] [2] [3] [4]。1974 年，Bezdek 建立了模糊聚类理论[5]。模糊聚类分析采用隶属度函数表示样本间的亲疏程度，能够更加全面的体现数据集的结构[6]。2002 年，Girolami 将聚类与核方法相结合，首次提出了核聚类方法，为之后的核聚类研究奠定了基础[7]。2004 年，伍忠东，高新波等将核方法的思想推广到模糊 c-均值算法，构造了基于核方法的模糊核 c-均值算法[8]。2009 年，赵犁丰，李新等人针对多类样本数据，提出一种多核模糊聚类算法。通过选取子核函数及其参数构造多核函数，使得输入空间的样本经多核映射后增大不同类别样本间的差别，提高核函数的学习能力和泛化能力[9]。2011 年，王栋将加权多宽度高斯核学习引入到聚类分析中，提出了一种加权多宽度高斯核聚类算法[10]。

上述模糊聚类算法主要通过改进核函数来提高聚类效果。然而这些算法将所有特征或样本视为平等，都未考虑待聚类数据集样本属性间的不平衡性的缺陷问题，从而降低了聚类性能。2012 年，蔡威提出了一种改进的属性加权核模糊聚类算法(WKFCM)，该方法充分体现了各特征属性对聚类结果重要性的差异性，改进了当时模糊聚类算法的不足[11]。2017 年，曹喆提出了一种基于样本和特征加权的模糊聚类，同时考虑样本和特征的重要性对聚类的影响，建立了基于样本和特征的模糊聚类模型，并将核函数引入该聚类算法，建立了基于样本和特征加权的核模糊聚类模型[12]。然而上述加权模糊聚类算法均是采用单个核函数进行聚类，对于多数据源或异构数据集，聚类算法的聚类效果并不理想。

因此，本文对属性加权核模糊聚类算法进行改进，将该聚类算法与多核模糊聚类算法相结合，使其在聚类过程中，不仅能动态调整各属性对不同类别的重要性，而且通过将全局核函数和局部核函数线性组合，增强了核函数的泛化能力和学习能力，能更加准确的反映数据集的结构特征。

2. 属性加权多核模糊聚类算法

2.1. 混合核函数

不同的核函数会将输入空间的数据集映射到不同的特征空间，相似度度量矩阵就会不同，对于基于核的模糊聚类方法，核函数起着关键的作用。常见的核函数大致被分成两类，即局部核函数和全局核函

数。局部核函数就是在数据点远离测试点时函数取值很小，比如高斯核函数就是最常用的局部核函数，具有较好的学习能力。而多项式核函数是一个典型的全局核函数，对离测试点相距较远的数据点也会从产生作用，但局部性较差，保证了良好的泛化能力。构造一个合适的核函数对于核模糊聚类的聚类结果具有重要影响。只使用一个核函数有时并不能满足待聚类数据集的需求，而不同的核函数结合起来使用，会有更好的特性，比如将具有局部核函数和全局核函数相结合，这就是混合核函数的基本思想[13]。

根据定理 2-1 如果 $k(\cdot,\cdot)$ 是一个有效的核函数(也称 Mercer 核函数)，当且仅当输入数据集 $D=\{x_1,x_2,\cdots,x_n\}$ 对应的核矩阵是半正定和对称的。

设 k_1,k_2 是 $R^d\times R^d\to R$ 上的核函数， $a\in R^+$，根据定理 2-1 容易得出以下是核函数：

$$k(x,z)=k_1(x,z)+k_2(x,z) \tag{1}$$

$$k(x,z)=ak_1(x,z) \tag{2}$$

容易证明函数 $k(x,z)=\lambda k_1(x,z)+(1-\lambda)k_2(x,z),\lambda\in[0,1]$ 也是核函数。

本文将学习能力强的高斯核函数和泛化能力强的多项式核函数进行线性组合构造了混合核函数，如下式所示：

$$k(x,x_i)=\lambda\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)+(1-\lambda)\left(\gamma\|x-x_i\|+r\right)^d \tag{3}$$

并将该混合核函数用于以下的聚类算法中。

2.2. 算法理论

以往的聚类算法中，默认所有的特征属性在聚类过程中起着同等的作用，但实际上不同属性对各个类别的重要性并不相同。属性加权多核模糊聚类算法充分考虑了各特征属性对聚类结果贡献度的差异，在聚类过程中，根据不同类的具体特性动态调整各个属性对于不同类别的重要性。并引入混合核函数，将学习能力强的局部核函数和泛化能力强的全局核函数进行线性组合以此取代原来的单个核函数，以此处理单个核函数不能满足待聚类数据集的需求的问题。将该方法充分考虑了属性间的不平衡性，能够更真实的反映待聚类问题的真实情况。

设聚类样本集为 $X=\{x_1,x_2,\cdots,x_j,\cdots,x_n\}$， $x_j=\{x_{j1},x_{j2},\cdots,x_{jk},\cdots,x_{jd}\},j=1,2,\cdots,n$，样本集经过非线性变化 φ 映射到高维核特征空间后

$$\begin{aligned}X^{\varphi}&=\{\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_j),\cdots,\varphi(x_n)\}\\ \varphi(x_j)&=\{\varphi(x_{j1}),\varphi(x_{j2}),\cdots,\varphi(x_{jk}),\cdots,\varphi(x_{jd})\}\end{aligned},$$

则 WMKFCM 模糊聚类算法准则为使以下目标函数最小：

$$J^{\varphi}_{\text{WMKFCM}}=\sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{d}u_{ij}^{b}w_{ik}^{a}\left\|\varphi(x_{jk})-m_{ik}^{\varphi}\right\|^2 \tag{4}$$

S.t.

$$u_{ij}\in[0,1]\ \text{and}\ \sum_{i=1}^{c}u_{jk}=1,\ j=1,2,\cdots,n \tag{5}$$

$$w_{ik}\in[0,1]\ \text{and}\ \sum_{k=1}^{d}w_{ik}=1,\ i=1,2,\cdots,c \tag{6}$$

其中， w_{ik} 表示第 k 个属性对于第i类的重要程度，称为属性权值； $a>1$ 表示权重因子。

根据约束优化问题的 Lagrange 求解，用混合核函数推导结果如下：

$$u_{ij}=\frac{1}{\left(\sum_{i=1}^{c}\left(\frac{1}{\sum_{k=1}^{d}w_{ik}^{\frac{a}{b-1}}\left\|\varphi\left(x_{jk}\right)-m_{ik}^{\varphi}\right\|^{\frac{2}{b-1}}}\right)\right)\left(\sum_{k=1}^{d}w_{ik}^{\frac{a}{b-1}}\left\|\varphi\left(x_{jk}\right)-m_{ik}^{\varphi}\right\|^{\frac{2}{b-1}}\right)} \tag{7}$$

$$w_{ik}=\frac{1}{\left(\sum_{k=1}^{d}\left(\frac{1}{\sum_{j=1}^{n}u_{ij}^{\frac{b}{a-1}}\left\|\varphi\left(x_{jk}\right)-m_{ik}^{\varphi}\right\|^{\frac{2}{a-1}}}\right)\right)\left(\sum_{j=1}^{n}u_{ij}^{\frac{b}{a-1}}\left\|\varphi\left(x_{jk}\right)-m_{ik}^{\varphi}\right\|^{\frac{2}{a-1}}\right)} \tag{8}$$

$$m_{ik}=\frac{\sum_{j=1}^{n}u_{ij}\varphi\left(x_{jk}\right)}{\sum_{j=1}^{n}u_{ij}} \tag{9}$$

将混合核函数代入，得以下公式：

$$u_{i,j}=\frac{\left[\sum_{k=1}^{d}w_{ik}^{a}\left(k\left(x_{jk},m_{ik}\right)-2k\left(x_{jk},m_{ik}\right)+k\left(m_{ik},m_{ik}\right)\right)\right]^{\frac{1}{1-b}}}{\sum_{i=1}^{c}\left(\sum_{k=1}^{d}w_{ik}^{a}\left(k\left(x_{jk},x_{jk}\right)-2k\left(x_{jk},m_{jk}\right)+k\left(m_{jk},m_{jk}\right)\right)\right)^{\frac{1}{1-b}}} \tag{10}$$

$$w_{ik}=\frac{\left[\sum_{j=1}^{n}u_{ik}^{b}\left(k\left(x_{jk},x_{jk}\right)-2k\left(x_{jk},m_{ik}\right)+k\left(m_{ik},m_{ik}\right)\right)\right]^{\frac{1}{1-b}}}{\sum_{i=1}^{c}\left(\sum_{j=1}^{n}u_{ij}^{b}\left(k\left(x_{jk},x_{jk}\right)-2k\left(x_{jk},m_{ik}\right)+k\left(m_{ik},m_{ik}\right)\right)\right)^{\frac{1}{1-b}}} \tag{11}$$

$$m_{ik}=\frac{\sum_{j=1}^{n}u_{ij}k\left(x_{jk},m_{ik}\right)x_{jk}}{\sum_{j=1}^{n}u_{ij}k\left(x_{jk},m_{ik}\right)} \tag{12}$$

其中，k 表示混合核函数。

2.3. 算法步骤

属性加权多核模糊聚类算法(WMKFCM)模糊聚类算法具体实现步骤如下：

Step 1 设定初试迭代聚类参数：聚类类别数 c、模糊加权指数 b、属性权重因子 a、混合核函数比例 λ，高斯核函数参数 σ，多项式核函数参数 γ,r,d，迭代终止阈值 ε 和迭代次数 I；

Step 2 随机初试化 m_{ik} 和 w_{ik}；

Step 3 根据上述推导公式，依次更新 u_{ij},w_{ik},m_{ik} 和 $J_{\mathrm{WMKFCM}}^{\varphi}$；

Step 4 重复 Step 3 的计算，直到 $\left|J_{\mathrm{WMKFCM}}^{\varphi}\left(I+1\right)-J_{\mathrm{WMKFCM}}^{\varphi}\left(I\right)\right|\le\varepsilon$；

Step 5 根据最大隶属度原则，由隶属度矩阵U确定最终聚类结果。

3. 实验及结果分析

为了测试属性加权多核模糊聚类算法(WMKFCM)的性能，本次研究将该算法与另外两种典型的求解大规模数据的聚类算法属性加权核模糊聚类和多核模糊聚类在同一数据集 Haberman’s Survival 数据集上聚类，并对比聚类性能。

Haberman’s Survival 数据集，选自 UCI 机器学习数据库，该数据集包含 306 个样本，具备 3 种属性：手术时病人的年龄(Age of patient at time of operation)，病人手术的年份(Patient’s year of operation)，检测

到正腋窝节点的数量(Number of positive axillary nodes detected)。一共分为 2 类：1——这个病人活了 5 年或更长(1—the patient survived 5 years or longer)，如下图蓝色点，2——病人在 5 年内死亡(2—the patient died within 5 year)，如下图 1 红色点。

将属性加权多核模糊聚类中的混合核函数比例系数的取值区间设为[0, 0.2 ,0.5,0.6,0.7,0.8,0.9,1]。其他参数设置相同。本次实验分别从聚类准确率，CPU 运行时间，以及迭代次数三方面进行比对。其中聚类准确率的计算方式如下：

$$P=\frac{\mathrm{SS}+\mathrm{DD}}{\mathrm{SS}+\mathrm{SD}+\mathrm{DS}+\mathrm{DD}} \tag{13}$$

上式中，SS 对记录属于相同的真实分组和相同的聚类；DD 对记录属于不同的真实分组和不同的聚类；SD 对记录属于相同的真实分组，但不同的聚类；DS 对记录属于不同的真实分组，但相同的聚类。如下图 2 所示。

属性加权多核模糊聚类算法在 Haberman's Survival 数据集中的运行结果如下表 1 所示。

如表 1 所示。当 $\lambda=0$ 的时候，此时聚类算法为属性加权核模糊聚类算法，核函数为多项式核函数；

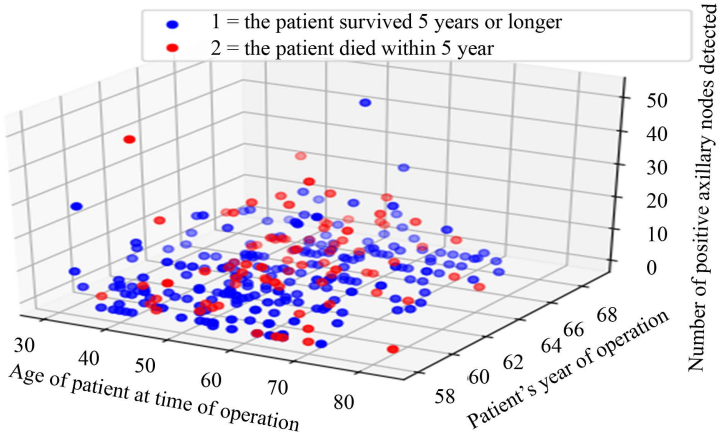


Figure 1. The dataset of habarman's survival
图 1. Habarman's Survival 数据集

		聚类	
		Same Cluster	Different Clusters
真实分组	Same Cluster	*SS*	*SD*
	Different Clusters	*DS*	*DD*

Figure 2. The chart of clustering accruacy
图 2. 聚类准确图解释图

Table 1. The result table of attribute-weighted multi-kernel fuzzy clustering algorithm
表 1. 属性加权多核模糊聚类算法运行结果表

混合核函数比例	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
迭代次数	9	9	13	10	11	14	14	17	42	107	45
cpu 运行时间	3.1356	3.276	4.6956	3.7128	3.962	5.616	4.992	6.0372	15.5065	32.4207	16.0901
聚类准确度	0.5001	0.5005	0.5067	0.5069	0.504	0.5003	0.577	0.5578	0.5534	0.5795	0.5031

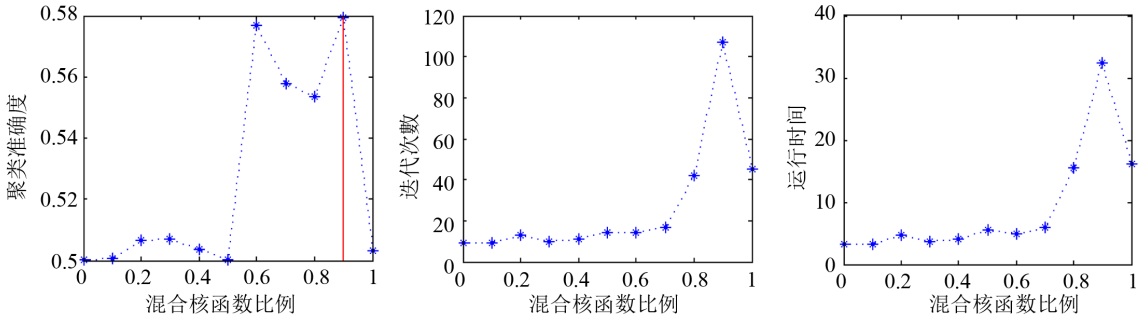
当 $\lambda=1$ 的时候，此时聚类算法为属性加权核模糊聚类算法，核函数为高斯核函数。当 $\lambda\in(0,1)$ 时，此时聚类算法为属性加权多核模糊聚类算法，核函数为混合核函数。

将上述数据用折线图的形式表示更为直观。如下图 3 图示。

如图所示，3(a)为聚类准确度与混合核函数关系图，3(b)为迭代次数与混合核函数比例关系图，3(c)为运行时间与混合核函数比例关系图。与属性加权核模糊聚类算法相比，属性加权多核模糊聚类的迭代次数和运行时间较高。但当混合核函数比例系数 $\lambda\in(0,1)$ 的时候，其聚类准确度明显高于 $\lambda\in(0,1)$ 时。其中，当混合核函数比例系数 $\lambda=0.9$ 的时候，属性加权多核模糊聚类算法的聚类准确度最高。由此可见，在牺牲一定的运行时间和迭代次数上，属性加权多核模糊聚类算法的聚类准确度要明显优于属性加权核模糊聚类算法。

如图 4~6 所示，属性加权多核模糊聚类算法相比于多核模糊聚类算法，其运行时间和迭代次数有一定的增长，但是其聚类准确度普遍优于多核模糊聚类算法。因此，在牺牲一定运行时间和迭代次数的前提下，属性加权多核模糊聚类算法的准确度优于多核模糊聚类算法。

将属性加权多核模糊聚类孙发与多核模糊聚类算法相对比，其他参数设置相同，实验结果如表 2 所示。将表 2 数据以折线图的形式表示更为直观，如图 4~6 所示。



(a)聚类准确度~混合核函数比例　(b)迭代次数~混合核函数比例　(c)运行时间~混合核函数比例

Figure 3. The result graph of attribute-weighted multi-kernel fuzzy clustering algorithm
图 3. 属性加权多核聚类算法运行结果图

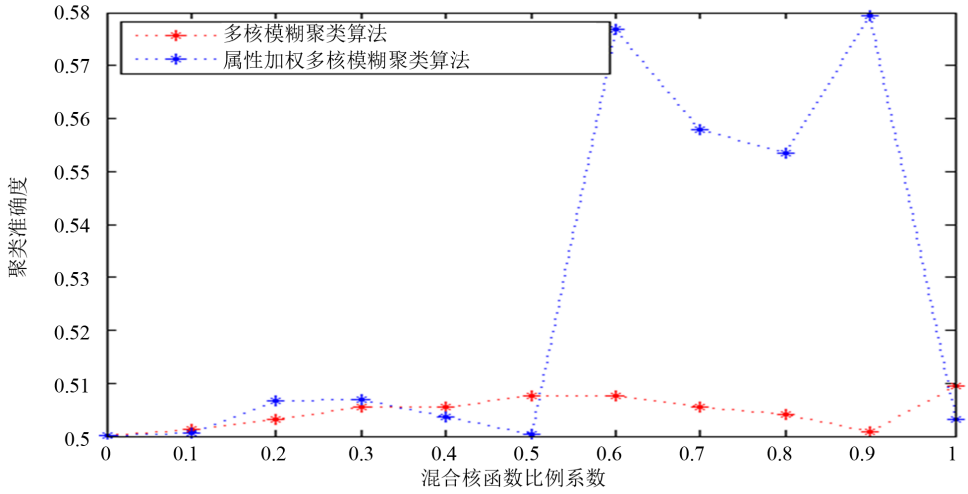


Figure 4. Contrast graph of clustering accruacy of multi-kernel fuzzy clustering algorithm and attribute-weighted multi-kernel fuzzy clustering algorithm
图 4. 多核模糊聚类与属性加权多核模糊聚类准确度对比图

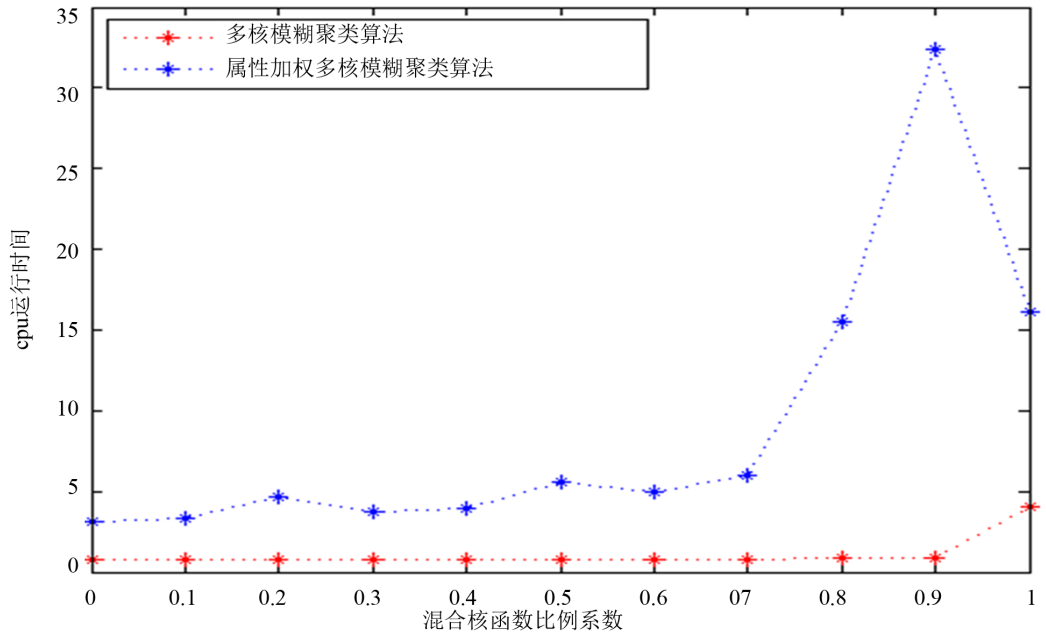


Figure 5. Contrast graph of CPU time of multi-kernel fuzzy clustering algorithm and attribute-weighted multi-kernel fuzzy clustering algorithm
图 5. 多核模糊聚类与属性加权多核模糊聚类 Cpu 运行时间对比图

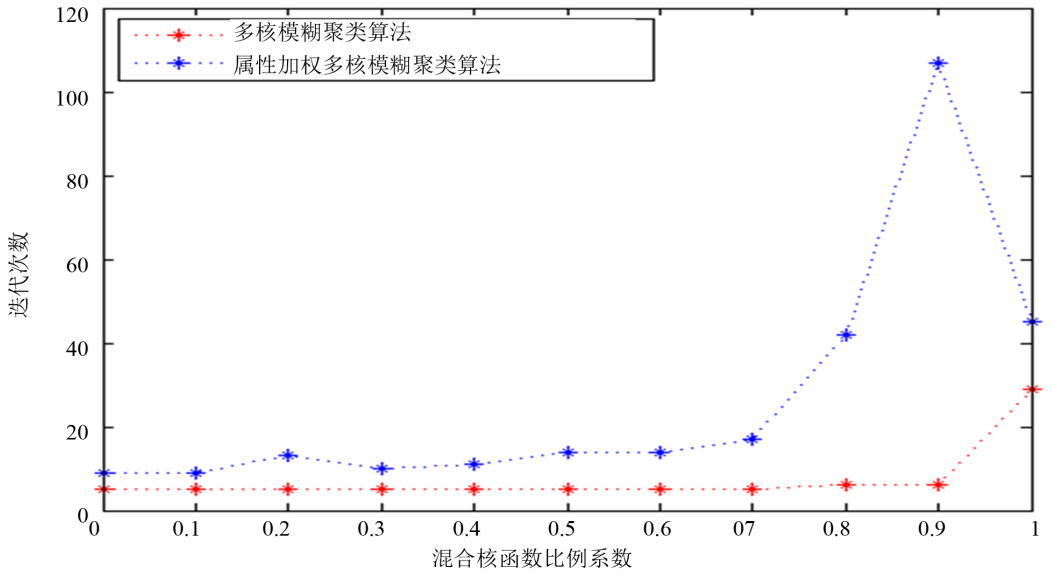


Figure 6. Contrast graph of the number of iterations of multi-kernel fuzzy clustering algorithm and attribute-weighted multi-kernel fuzzy clustering algorithm
图 6. 多核模糊聚类与属性加权多核模糊聚类迭代次数对比图

4. 结束语

本次研究通过将多核模糊聚类算法与属性加权核模糊聚类算法相结合，提出了属性加权多核模糊聚类算法。通过多个核函数的线性组合以解决单个核函数不能满足待聚类数据集的需求，并且，充分考虑了各特征属性对聚类结果贡献度的差异，在聚类过程中，根据不同类的具体特性动态调整各个属性对于

DOI: 10.12677/csa.2018.86107

Table 2. Contrast table of attribute-weighted multi-kernel fuzzy clustering algorithm and multi-kernel fuzzy clustering (multi-kernel fuzzy clustering/attribute-weighted multi-kernel fuzzy clustering algorithm)

表 2. 属性加权多核模糊聚类算法与多核模糊聚类算法对比表(多核模糊聚类算法/属性加权多核模糊聚类算法)

混合核函数比例	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
迭代次数	5/9	5/9	5/13	/5/10	5/11	5/14	5/14	/5/17	6/42	6/107	29/45
cpu 运行时间	1.78/3.1356	1.749/3.276	1.7644/4.6956	1.7644/3.7128	1.7644/3.962	1.7488/5.616	1.7176/4.992	1.7332/6.0372	1.8580/15.5065	1.8580/32.4207	5.0716/16.0901
聚类准确度	0.5002/0.5001	0.5014/0.5005	0.5031/0.5067	0.5055/0.5069	0.5055/0.504	0.5077/0.5003	0.5077/0.577	0.5055/0.5578	0.5042/0.5534	0.5008/0.5795	0.5094/0.5031

不同类别的重要性。最后通过在数据集上进行测试，有效地证明了属性加权多核模糊聚类算法的可行性。

参考文献

[1] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.

[2] Ding, S., Zhang, J., Jia, H., *et al*. (2016) An Adaptive Density Data Stream Clustering Algorithm. *Cognitive Computation*, **8**, 30-38. https://doi.org/10.1007/s12559-015-9342-z

[3] Jain, A. (2010) Data Clustering: 50 Years beyond k-Means. *Pattern Recognition Letters*, **31**, 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

[4] Du, M., Ding, S. and Jia, H. (2016) Study on Density Peaks Clustering Based on k-Nearest Neighbors and Principal Component Analysis. *Knowledge-Based Systems*, **99**, 135-145. https://doi.org/10.1016/j.knosys.2016.02.001

[5] 严峻. 模糊聚类算法应用研究[D]: [硕士学位论文]. 浙江大学, 2006.

[6] Chen, M.S., Wang, W., *et al*. (1999) Fuzzy Clustering Analysis for Optimizing Fuzzy Membership Function. *Fuzzy Sets and Systems*, **103**, 239-254. https://doi.org/10.1016/S0165-0114(98)00224-3

[7] Girolami, M. (2002) Mercer Kernel-Based Clustering in Feature Space. *IEEE Transactions on Neural Networks*, **13**, 780-784. https://doi.org/10.1109/TNN.2002.1000150

[8] 伍忠东, 高新波, 谢维信. 基于核方法的模糊聚类算法[J]. 西安电子科技大学学报(自然科学版), 2004, 31(4): 533-537.

[9] 赵犁丰, 李新, 王栋. 多核模糊聚类算法的研究[J]. 中国海洋大学学报(自然科学版), 2009, 39(5): 1047-1050.

[10] 王栋. 基于加权多宽度高斯核函数的支持向量机聚类算法研究[D]. 中国海洋大学, 2011.

[11] 蔡威. 模糊聚类在数据挖掘中的应用研究[D]. 兰州交通大学, 2012.

[12] 曹喆. 样本和特征加权的模糊聚类算法研究[D]. 河北大学, 2017.

[13] 樊淑炎. 核 k-means 优化方法研究[D]. 中国矿业大学, 2017.