DOI：10.7652/xjtuxb201803019

混合测量子空间聚类算法的研究

金利英，赵升吨
（西安交通大学机械工程学院，710049，西安）

摘要：针对闵可夫斯基子空间聚类算法对特征权重分配的问题，提出了一种混合测量子空间聚类算法(iMWK-HD)，以实现调节特征权重因子和提高算法性能的目的。利用闵可夫斯基距离与余弦相结合的混合测量来分配特征权重，构造新的目标函数；在聚类迭代过程中，采用智能 K-means 进行初始化来解决选择正确类数的问题；根据新的目标函数，使用拉格朗日乘子法求解新的隶属度和特征权重更新公式，使类中心更加稳定，从而促进特征空间转换，获取数据集最优聚类结果。采用 UCI 数据集设计了对比实验，实验结果表明，iMWK-HD 算法优于 iK-means、iWK-means、iMWK-means 这 3 个现有的聚类算法，所提算法能有效提升聚类精确度和聚类结果的稳定性。

关键词：闵氏距离；子空间聚类算法；特征权重；混合测量
中图分类号：TP181 **文献标志码**：A **文章编号**：0253-987X(2018)03-0139-06

A Hybrid Clustering Algorithm for the Measurement of Subspace

JIN Liying, ZHAO Shengdun
(School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: This study presents a hybrid clustering algorithm for the measurement of subspace. This approach is adopted to distribute the feature weight in Minkowski algorithm to adjust the feature weighting factor and improve the performance of the algorithm. The feature weight is assigned by using hybrid dissimilarity measurement of Minkowski distance and Cosine dissimilarity and a new objective function is designed. In the process of clustering iteration, the problem of selecting correct class number is solved by using intelligent K-means initialization. According to the new objective function, the Lagrange multiplier method is used to solve the new membership degree and the feature weight iteration updating formula, so that the class center is more stable and the optimal clustering result of dataset is obtained by promoting the transformation of feature space. Adopting UCI dataset to design the experiments, the results showed that compared with the other three algorithms, i.e., iK-means, iWK-means and iMWK-means algorithms, the proposed algorithm can effectively improve the clustering accuracy and the stability of clustering results.

Keywords: Minkowski distance; subspace clustering algorithm; feature weight; hybrid dissimilarity measurement

子空间聚类算法是模式识别和数据挖掘等应用领域的重要工具，随着新技术的高速发展，应用环境

收稿日期：2017-07-09。 作者简介：金利英(1985—)，女，博士生；赵升吨(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(51335009)；国家重大科技专项及国家重点研发计划资助项目(HZ201602)。
网络出版时间：2018-01-13 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20180113.0917.006.html

不断变化，大规模数据与复杂结构数据对子空间聚类算法的需求越来越紧迫。迄今为止，已开发出许多聚类算法[1-2]，以实现数据聚类的任务，这些算法在如何定义集群以及如何识别集群方面有很大的不同。K-means 算法是最经典且应用最广的聚类算法之一[3]，但对聚类中心的初始化非常敏感，影响聚类结果和收敛速度。Huang 等在 K-means 算法的基础上提出了权重 K-means 算法(WK-means)[4]，为每个特征分配一个权重，特征权重与特征尺度因子相关[5]。使用闵可夫斯基距离自动计算每个集群的特征权重，以确保这些权重可以看作是特征调节因子。

现有的大多数子空间聚类算法仅使用一个距离函数来评估各个样本之间的距离，该距离函数难以处理具有复杂内部结构的数据集。通过特征之间权重方向旋转特征空间可进一步提高聚类性能，在向量空间模型中[6]，余弦相异性比欧式距离更重要。在特征空间(ESCHD)中[7]，余弦相异性可提高软子空间聚类的性能。

受 iMWK-means 和 ESCHD 算法的启发，本文提出了混合测量子空间聚类算法(iMWK-HD)。该算法将余弦相异性引入到距离函数中，使闵可夫斯基指数与分配特征权重指数相符，构造新的目标函数。使用该目标函数推导迭代过程中更新规则，促进特征空间的扩展、收缩和旋转，iMWK-HD 算法可提高聚类精度和聚类结果的稳定性。

1 iMWK-HD 算法

1.1 闵可夫斯基距离和智能 K-means 算法初始化类中心

设数据集 $\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 由 N 个 M 维样本 $\{\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{iM})\}_{i=1}^{N}$ 构成，样本 $\boldsymbol{X}_i=[x_{i1},x_{i2},\cdots,x_{iM}]^{\mathrm{T}}$ 与 $\boldsymbol{X}_j=[x_{j1},x_{j2},\cdots,x_{jM}]^{\mathrm{T}}$ 之间的闵可夫斯基距离定义为

$$d(x_i,x_j)=\Big[\sum_{k=1}^{M}\mid x_{ik}-x_{jk}\mid^{\beta}\Big]^{1/\beta} \tag{1}$$

式中：β 为用户定义的参数，表示权重对距离贡献的影响率。$d(x_i,x_j)$ 因 β 值的不同而变化，指数 $\beta\neq 2$ 不再是特征比例因子。Huang 等扩展了相应的闵可夫斯基测量[4]，用于测量距离，以确保闵可夫斯基指数与特征权重指数一致。

智能 K-means 初始化[8](iK-means)是在运行 K-means 算法之前找出异常集群，即对数据集进行全局加权距离的计算并排序，取最大距离样本点为异常集群，将其设置为初始类中心，在迭代过程中找出距离它最近的样本点，然后从数据集中依次取出；在剩余数据集里取距离最远的样本为初始化类中心，在迭代过程中找出距离它最近的样本点，然后从这个数据集里移出；依次类推，直到剩余数据集为空，之后选取样本点个数最多的异常簇的异常集群为类中心。

1.2 混合测量

基于距离测量准则，在处理高维样本点时会出现维数灾难的情况[9]，影响聚类结果的稳定性。将余弦相异性代入优化目标函数中，结合距离计算样本点之间的距离，可促进特征空间转换，样本点 x_i 和第 j 维 k 类中心 v_k 之间的距离定义为

$$d_\beta(x_{ij},v_{kj})=\Big(\sum_{j=1}^{M}\boldsymbol{w}_{kj}^{\beta}\big[(\boldsymbol{x}_{ij}-\boldsymbol{v}_{kj})^{\beta}-\eta\Big(\frac{\boldsymbol{v}_{kj}\times\boldsymbol{x}_{ij}}{\|\boldsymbol{v}_k\|\times\|\boldsymbol{x}_i\|}-\frac{1}{M}\Big)\big]\Big)^{1/\beta} \tag{2}$$

式中：$\boldsymbol{w}_{kj}$ 为对角特征值距阵。数据分析的有效性随着 β 值的不同而改变，当参数 $\beta=1$ 时，闵可夫斯基距离退化为曼哈顿距离或出租车距离；当参数 $\beta=2$ 时，闵可夫斯基距离退化为欧氏距离；当参数 $\beta=\infty$ 时闵可夫斯基距离退化为切比雪夫距离或最大值距离。式(2)引入余弦相异性，使特征权重在特征空间里旋转，$\|\boldsymbol{v}_k\|$ 是类中心向量 $\boldsymbol{v}_k$ 的范数，$\|\boldsymbol{x}_i\|$ 是样本点向量 $\boldsymbol{x}_i$ 的范数，η 是正则项系数，为了保持闵可夫斯基距离和余弦之间的平衡，$\sum_{j=1}^{M}\frac{1}{M}=1$ 是为了确保余弦不等式中的第 2 项大于 0。

混合测量参数较少、易于实现，可控制数据集在特征空间中的形状、大小和方向，能减小冗余和无关特征对数据集的影响，进而区分特征在不同子空间中的权重大小，提高全局搜索能力。

1.3 目标函数设计

目标函数是评估聚类性能的重要指标，直接影响算法的搜索方向和收敛速度。假设有 N 个 M 维数据集 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_N\}$，第 i 个样本点表示为 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{iM})$；聚类中心矩阵 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_C\}$，$C$ 是 N 个数据集划分类数。利用混合测量求解目标函数最小值，定义为

$$\begin{aligned}&J_{\text{iMWK-HD}}(\boldsymbol{U},\boldsymbol{W},\boldsymbol{V})=\sum_{k=1}^{C}\sum_{i=1}^{N}u_{ik}d_\beta(x_{ij},v_{kj})\\&\text{s.t.}\quad \boldsymbol{u}_{ik}\in\{0,1\},\quad\sum_{k=1}^{C}\boldsymbol{u}_{ik}=1,\quad 1\leqslant i\leqslant N\\&\qquad\boldsymbol{w}_{kj}\in[0,1],\quad\sum_{j=1}^{M}w_{kj}=1,\quad 1\leqslant k\leqslant C\end{aligned} \tag{3}$$

式中：矩阵 $\boldsymbol{V}=[v_{kj}](1\leqslant k\leqslant C,1\leqslant j\leqslant M)$ 为第 k 类在第 j 维特征上的类中心值；集群 $\boldsymbol{U}=[u_{ik}](1\leqslant i\leqslant N,1\leqslant k\leqslant C)$ 为第 i 个样本对第 k 类的隶属度；矩阵 $\boldsymbol{W}=[w_{jk}](1\leqslant j\leqslant M,1\leqslant k\leqslant C)$ 为第 j 维特征对第 k 类的重要度。

iMWK-HD 算法优化目标函数，矩阵 $\boldsymbol{U}$、$\boldsymbol{W}$ 的最优解可由下式给出，即

$$\left.\begin{aligned} u_{ik} &= 1, \quad \sum_{j=1}^{M} w_{kj}^{\beta} D_{kj} \leqslant \sum_{j=1}^{M} w_{lj}^{\beta} D_{lj} \\ u_{il} &= 0, \quad 1\leqslant l\leqslant C, l\neq k \end{aligned}\right\} \tag{4}$$

$$w_{kj} = \frac{1}{\sum_{m=1}^{M} (D_{kj}/D_{km})^{1/(\beta-1)}} \tag{5}$$

式中：，$D_{kj}=[(\boldsymbol{x}_{ij}-\boldsymbol{v}_{kj})^{\beta}-\eta(\frac{\boldsymbol{v}_{kj}\times\boldsymbol{x}_{ij}}{\|\boldsymbol{v}_k\|\times\|\boldsymbol{x}_i\|}-\frac{1}{M})]$，$u_{ik}=1$ 表示第 i 个样本 x_i 被分配到第 k 类 v_k；$D_{kj}=\sum_{i=1}^{N}\sum_{k=1}^{C}u_{ki}[(\boldsymbol{x}_{ij}-\boldsymbol{v}_{kj})^{\beta}-\eta(\frac{\boldsymbol{v}_{kj}\times\boldsymbol{x}_{ij}}{\|\boldsymbol{v}_k\|\times\|\boldsymbol{x}_i\|}-\frac{1}{M})]$。

矩阵 $\boldsymbol{V}$ 的求解可由文献[5]中 iMWK-means 算法进行迭代更新。式(3)的无约束最小化为 $L(\boldsymbol{U},\boldsymbol{W},\boldsymbol{V},\lambda_k)=\sum_{j=1}^{M}w_{kj}^{\beta}D_{kj}-\sum_{k=1}^{C}\lambda_k\left(\sum_{j=1}^{M}w_{kj}-1\right)$，其中 λ_k 是与等式约束有关的 Lagrange 乘子。对 $L(\boldsymbol{U},\boldsymbol{W},\boldsymbol{V},\lambda_k)$ 中的变量 w_{kj} 求偏导得 $\partial L(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\lambda_k)/\partial w_{kj}=\beta w_{kj}^{\beta-1}D_{kj}-\lambda_k=0$，$w_{kj}=(\lambda_k/\beta D_{kj})^{1/(\beta-1)}$，由约束条件得 $\sum_{j=1}^{M}(\lambda_k/\beta D_{kj})^{1/(\beta-1)}=1$，从而得出 $(\lambda_k/\beta)^{1/(\beta-1)}=1/\sum_{j=1}^{M}(1/D_{kj})^{1/(\beta-1)}$，iMWK-HD 算法流程如下：

$\beta\in\{1.2,2.0,2.3,2.7,3.0,4.9\}$，$\eta\in[0,10]$，$\varepsilon=10^{-5}$，$G=500$

iK-means initialize $\boldsymbol{V}$ and $\boldsymbol{W}$ with $w_{jk}=1/M$

for $t=1$:G

Updata $\boldsymbol{U}^{t+1}$ according to (4) with $\boldsymbol{W}^{t}$ and $\boldsymbol{V}^{t}$;

Updata $\boldsymbol{V}^{t+1}$ according to $\boldsymbol{W}^{t}$ and $\boldsymbol{U}^{t+1}$;

Updata $\boldsymbol{W}^{t+1}$ according to (5) with $\boldsymbol{U}^{t+1}$ and $\boldsymbol{V}^{t+1}$;

Calculate the objective function $J(\boldsymbol{U}^{t+1},\boldsymbol{V}^{t+1},\boldsymbol{W}^{t+1})$ with (3);

if $|J(\boldsymbol{U}^{t+1},\boldsymbol{V}^{t+1},\boldsymbol{W}^{t+1})-J(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})|<\varepsilon$

break;

End for

Output $\boldsymbol{W}^{t+1}$，$\boldsymbol{V}^{t+1}$ and $\boldsymbol{U}^{t+1}$。

2　实验设计和结果分析

实验环境为 Intel(R) Core(TM) i3－4170 CPU，Windows7，软件开发平台是 Matlab 13.0。为验证 iMWK-HD 算法有效性，与现有的 iK-means，iWK-means 和 iMWK-means 算法进行对比。

2.1　实验数据和评估指标

选取 UCI 数据库[10]中 Wine、Hepatitis、Australian、Liver、Wave、Sonar、Musk 这 7 个数据集来进行对比实验，其基本信息如表 1 所示。在实验中，使用精确度 A_{CC}[11]和 R_I[12]指标来评估聚类结果的性能，R_I 定义为

$$R_I=\frac{T+E}{T+F+E+P}\times 100\%$$

式中：T 为在同一类的一对样本点被分到同一类内；E 为不在同一类的一对样本点被分到不同的类里；F 为不在同一类的一对样本点被分到同一类内；P 为在同一类的一对样本点被分到不同类里。A_{CC}、R_I 取值范围为[0,1]，数值越接近 1，说明聚类性能越好。

表 1　数据描述

数据集	样本数	维数	类数
Hepatitis	155	19	2
Australian	690	15	2
Liver	345	6	2
Wave	5 000	40	3
Wine	178	13	3
Sonar	208	60	2
Musk	476	167	2

2.2　实验结果和分析

算法对 7 个数据集聚类的最优精度、最优解如表 2、表 3 所示。iMWK-HD 算法和 iMWK-means 算法之间是闵可夫斯基指数 β 取相同值时对数据集进行实验比较，表中平衡系数 η 为 iMWK-HD 算法获得最优实验结果时的取值。

由表 2、表 3 可知：iMWK-HD 算法在 7 个数据集上均获得最优的 A_{CC}、R_I，高出其他 3 种算法数个百分点；在聚类结果的稳定性、聚类精确度上，iMWK-HD 算法更具优势。这是因为本文将闵可夫斯基距离、余弦相结合构成新的目标函数，使其更好评价特征子集的质量，能更好找到子空间的内部结构，促进样本在特征空间转换，提高全局搜索能力，从而

获得更好的聚类结果。

表 2 算法对 7 个数据集聚类的最优精度

数据集	A_{CC}/%			
	iK-means	iWK-means	iMWK-means	iMWK-HD
Hepatitis	65.161 3	64.516 1	66.451 6	**70.967 7**
Australian	51.304 3	50.144 9	53.913 0	**55.797 1**
Liver	54.492 8	55.362 3	54.492 8	**57.101 4**
Wave	51.200 0	51.140 0	51.260 0	**75.400 0**
Wine	94.943 8	94.382 0	91.573 0	**96.067 4**
Sonar	53.365 4	53.846 2	59.100 0	**64.423 1**
Musk	49.900 0	51.491 2	50.581 4	**57.600 0**

表 3 算法对 7 个数据集聚类的最优解

数据集	R_I/%			
	iK-means	iWK-means	iMWK-means	iMWK-HD
Hepatitis	54.302 5	53.917 1	55.123 6	**58.525 3**
Australian	49.961 5	49.927 9	50.234 1	**50.600 5**
Liver	50.259 5	50.431 4	50.259 5	**50.866 2**
Wave	66.684 3	66.682 8	66.665 7	**74.105 0**
Wine	93.182 3	92.572 8	89.081 4	**94.667 7**
Sonar	49.986 1	50.005 7	51.435 3	**53.939 1**
Musk	50.361 2	49.938 4	49.900 0	**51.040 0**

β、η 是 iMWK-HD 算法的重要参数，对于大规模数据 Wave、高维数据 Musk，在 β 取不同值时实验获得的 A_{CC} 随 η 的变化趋势如图 1、图 2 所示。

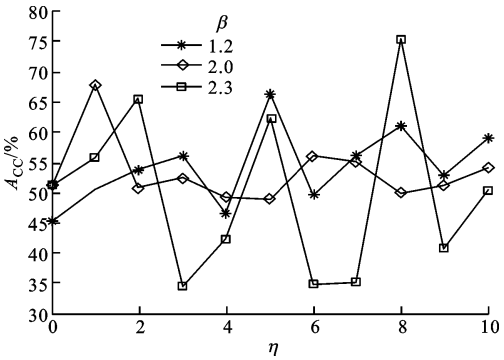


图 1 iMWK-HD 算法在 β=1.2,2.0,2.3 时的聚类结果

由图 1 中可知，当 β=1.2,2.0 时，A_{CC} 随着平衡系数 η 的不同取值而变化，且绝大多数 A_{CC} 优于未引入余弦时的聚类结果，说明引入余弦可增强算法的鲁棒性；当 β=2.3 且 η=3,4,6,7,9,10 时，A_{CC} 低于 η=0 时的值，说明对大规模数据而言，β 在 2.3 以内引入余弦能很好地调节特征权重因子，促进特征空间转换，从而提高聚类精确度；当 η=0 且 β 分别取 1.2、2.0、2.3 时，实验获得的 A_{CC} 分别为 45.1%、51.1%、51.3%；当 η=1 时，A_{CC} 分别为 50.5%、68%、55.8%；当 η=2 时，A_{CC} 分别为 53.7%、50.7%、65.4%；当 η 取定值时，在 β 取值范围内，实验获得的 A_{CC} 与 β 取值无关，这说明余弦的引入与 β 不冲突。

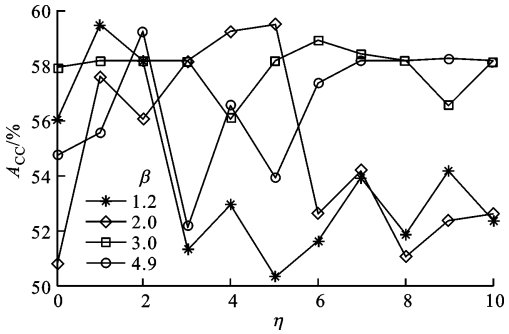


图 2 iMWK-HD 算法在 β=1.2,2.0,3.0,4.9 时的聚类结果

由图 2 中可知，当 β=1.2 时，A_{CC} 随着 η 的不同取值而变化，且当 η=3 时，A_{CC} 均小于 η=0 时的值，说明余弦的引入在 β=1.2 时的聚类结果不稳定。当 β=2.0,3.0,4.9 时，大部分 A_{CC} 优于未引入余弦时的聚类结果，对于高维数据，β 取值大于 2.0 时余弦的引入能很好调节特征权重因子，促进特征空间转换，从而提高聚类 A_{CC} 和算法的鲁棒性。因此，iMWK-HD 算法在原子空间聚类算法参数下，有且仅增加一个控制参数 η，能最大程度保证算法在高维数据中的聚类精度和聚类结果的稳定性。

iMWK-HD 算法与 iWK-means、iMWK-means 算法进行实验对比，结果如表 4～7 所示。

表 4 Wave 数据集精确度 A_{CC}

算法		A_{CC}/%		
		β=1.2	β=2.0	β=2.3
iWK-means		50.600 0	51.140 0	51.120 0
iMWK-means		47.220 0	51.060 0	51.260 0
iMWK-HD	η=0	47.220 0	51.060 0	51.260 0
	η=1	50.460 0	**68.000 0**	**55.840 0**
	η=2	**53.700 0**	50.720 0	**65.420 0**
	η=3	**56.040 0**	**52.640 0**	34.500 0
	η=4	46.780 0	49.060 0	42.420 0
	η=5	**66.460 0**	49.020 0	**62.260 0**
	η=6	49.640 0	**55.900 0**	34.880 0
	η=7	**56.260 0**	**55.160 0**	35.120 0
	η=8	**60.940 0**	49.780 0	**75.400 0**
	η=9	**52.860 0**	**51.240 0**	40.600 0
	η=10	**59.160 0**	**54.240 0**	**50.360 0**

表 5　Wave 数据集 R_I

算法		R_I/%		
		$\beta=1.2$	$\beta=2.0$	$\beta=2.3$
iWK-means		66.626 5	66.682 8	66.681 3
iMWK-means		64.467 1	66.689 9	66.665 7
iMWK-HD	$\eta=0$	64.467 1	66.689 9	66.665 7
	$\eta=1$	66.294 5	**68.225 1**	67.356 3
	$\eta=2$	64.946 8	65.825 0	68.479 5
	$\eta=3$	64.272 0	66.097 1	34.201 0
	$\eta=4$	63.284 7	62.984 6	49.612 6
	$\eta=5$	66.570 0	63.306 0	67.200 0
	$\eta=6$	60.192 4	64.838 9	34.663 8
	$\eta=7$	64.954 5	64.743 4	35.595 9
	$\eta=8$	**68.205 1**	64.226 8	**74.105 0**
	$\eta=9$	61.357 0	65.990 4	51.962 1
	$\eta=10$	66.143 1	65.532 5	60.701 8

表 7　Msuk 数据集 R_I

算法		R_I/%		
		$\beta=2.0$	$\beta=3.0$	$\beta=4.9$
iWK-means		49.943 2	50.241 2	49.941 5
iMWK-means		49.943 2	50.698 0	50.240 7
iMWK-HD	$\eta=0$	49.943 2	**50.698 0**	50.240 7
	$\eta=1$	**50.612 4**	**50.698 0**	**50.314 2**
	$\eta=2$	**50.412 5**	**50.698 0**	**51.000 0**
	$\eta=3$	**50.712 5**	**50.698 0**	50.000 0
	$\eta=4$	**51.000 0**	50.412 5	**50.412 5**
	$\eta=5$	**51.000 0**	**50.698 0**	50.100 0
	$\eta=6$	**50.000 0**	**50.914 5**	**50.581 2**
	$\eta=7$	**50.100 0**	**50.782 1**	**50.698 0**
	$\eta=8$	49.943 2	**50.698 0**	**50.698 0**
	$\eta=9$	**50.000 0**	50.412 5	**50.698 0**
	$\eta=10$	**50.000 0**	**50.698 0**	**50.698 0**

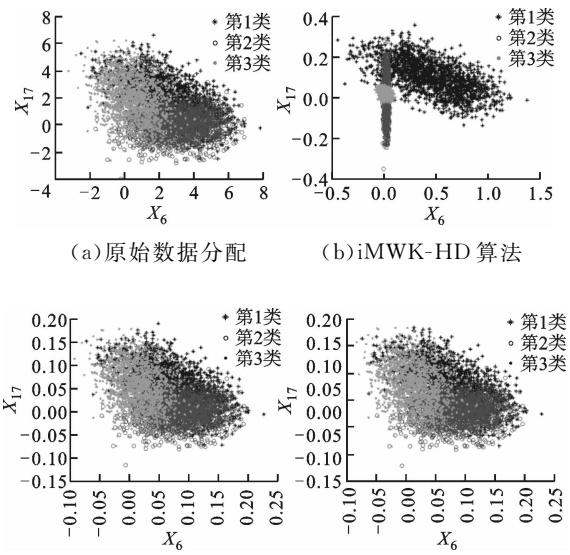
由表 4、5 可知，当 $\beta=1.2, 2.0$ 且 $\eta=0$，即未引入余弦时，iWK-means 算法取得 A_{CC} 最优值，但 iMWK-HD 算法的 A_{CC} 与最优结果相差不大，而引入余弦时，iMWK-HD 算法的 A_{CC} 大部分优于 iWK-means 和 iMWK-means 算法。当 $\beta=2.3$ 且 $\eta=0$ 时，iMWK-HD 和 iMWK-means 算法的 A_{CC} 相同；而引入余弦时，iMWK-HD 算法的最优值高出 iWK-means 和 iMWK-means 算法分别为 24.24%、24.1%，说明 iMWK-HD 算法具有较强鲁棒性和较高聚类精度。A_{CC} 所示的最佳聚类性能并不总是与 R_I 表示一致，因此有必要对不同指标聚类算法的性能进行评估。

表 6　Msuk 数据集精确度 A_{CC}

算法		A_{CC}/%		
		$\beta=2.0$	$\beta=3.0$	$\beta=4.9$
iWK-means		51.491 2	53.812 4	51.324 1
iMWK-means		50.614 5	56.314 2	53.825 6
iMWK-HD	$\eta=0$	50.614 5	56.314 2	53.825 6
	$\eta=1$	**56.091 4**	**56.498 1**	**54.398 1**
	$\eta=2$	**54.791 2**	**56.498 1**	**57.435 1**
	$\eta=3$	**56.524 8**	**56.498 1**	51.746 0
	$\eta=4$	**57.381 2**	54.812 4	**55.320 0**
	$\eta=5$	**57.635 8**	**56.498 1**	53.214 2
	$\eta=6$	**52.110 0**	**57.097 1**	**55.897 1**
	$\eta=7$	**53.441 2**	**56.710 0**	**56.498 1**
	$\eta=8$	50.798 6	**56.498 1**	**56.498 1**
	$\eta=9$	**51.935 8**	55.321 4	**56.642 1**
	$\eta=10$	**52.146 2**	**56.498 1**	**56.498 1**

由表 6 可知，当 $\eta=0$，即未引入余弦时，iMWK-HD 和 iMWK-means 算法在 β 取不同值时 A_{CC}、R_I 均相同；当引入余弦时，iMWK-HD 算法在大多数情况下的实验结果均优于 iMWK-means 和 iWK-means 算法，只有少数实验结果低于它们，且实验结果相差不大，说明本文所提 iMWK-HD 算法具有较强鲁棒性。Musk 数据虽拥有较高维数，iMWK-HD 算法也能有效地提取对应维数信息，再一次体现出 iMWK-HD 算法的聚类结果稳定性好、聚类精度高的优点。

3 种算法根据特征权重进行转换的数据分布情况如图 3、图 4 所示。由图 3、图 4 可知，集群之间的

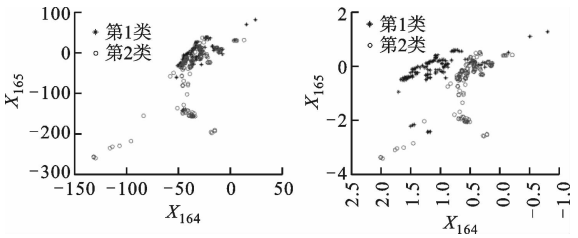


(a)原始数据分配　(b)iMWK-HD 算法

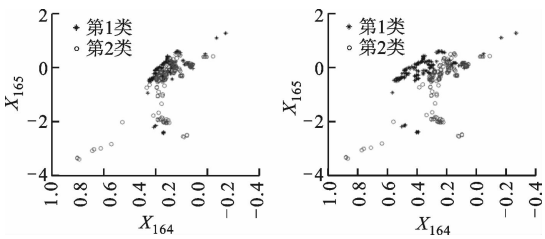
(c)iMWK-means 算法　(d)iWK-means 算法

图 3　Wave 数据集中 3 种算法根据特征权重进行转换的数据分布情况

重叠较少，在总体性能上优于另外两种算法，说明 iMWK-HD 算法聚类结果的稳定性好。这是因为 iMWK-HD 算法的目标函数中引入了余弦相异性，使特征权重在特征空间不仅能扩展、收缩，还能旋转，iMWK-HD 算法能有效提高解的质量。



(a)原始数据分配 (b)iMWK-HD 算法



(c)iMWK-means 算法 (d)iWK-means 算法

图 4 Musk 数据集中数据分布情况

对比实验结果说明了本文所提 iMWK-HD 算法的鲁棒性及聚类结果的稳定性。子空间聚类算法能更好发现子空间的内部结构，展现良好的聚类结果。对于 Wine 数据，各个子空间聚类算法的内部结构如表 8 所示，将其各维标上对应编号。由表 8 中前 6 位特征重要性排序可知，所得各个数据簇的特征排列顺序在每个子空间聚类算法中有所不同，也可在其他数据集中得到类似发现，这种发现子空间的能力可提高聚类有效性。

表 8 Wine 数据集上的特征重要性排序

算法	类数	每类中排名(取前 6 位特征序号)
iMWK-HD	1	10,11,6,2,7,8
	2	13,10,11,1,12,3
	3	7,6,12,1,9,8
iMWK-means	1	11,7,10,6,9,12
	2	7,13,12,3,5,11
	3	10,13,7,1,6,4
iWK-means	1	7,11,10,6,9,12
	2	7,13,12,11,3,5
	3	10,13,7,1,4,6

综上对比实验结果，分析 iMWK-HD 算法最优的原因：①初始化阶段使用智能 K-means 算法求解聚类中心点，解决了随机选取类中心导致若干聚类合并及正确选择类数的问题；②混合测量导出样本与中心点的距离，解决了特征空间旋转问题；③利用拉格朗日乘子求解新的隶属度和特征权重迭代更新公式，保证了类中心的稳定性，从而促进特征空间转换，获取有效的聚类结果。实验结果验证了 iMWK-HD 算法的优越性。

3 结 论

本文提出了一种混合测量子空间聚类算法，该算法是在 iMWK-means 框架下，将余弦相异性引入到子空间聚类的目标函数中，与闵氏距离相结合，构造了新的目标函数，该目标函数控制参数少，易于实现；利用拉格朗日乘子求解样本点新的隶属度和特征权重迭代更新距阵，实现特征权重因子自动调节，促进特征空间转换，从而有利于隶属度矩阵 $\boldsymbol{U}$ 的划分和类中心矩阵 $\boldsymbol{V}$ 的精准度，最终获取样本点的最优聚类结果，提高聚类性能。实验结果表明，与 iK-means、iWK-means 和 iMWK-means 三种算法相比，iMWK-HD 算法聚类结果稳定性好，聚类精确度高。

参考文献：

[1] PAN Y, WANG J, DENG Z. Alternative soft subspace clustering algorithm [J]. Journal of Information & Computational Science, 2013, 10: 3615-3624.

[2] CHEN X, YE Y, XU X, et al. A feature group weighting method for subspace clustering of high-dimensional data [J]. Pattern Recognition, 2012, 45: 434-446.

[3] MACQUEEN J. Some methods for classification and analysis of multivariate observations [J]. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1: 281-297.

[4] HUANG Z, NG M K, RONG H, et al. Automated variable weighting in k-means type clustering [J]. IEEE Transactions on Pattern Analysis and Machine Learning, 2005, 27(5): 657-668.

[5] RENATO C D, MIRKIN B. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering [J]. Pattern Recognition, 2012, 45: 1061-1075.

[6] KIM D. Group-theoretical vector space model [J]. International Journal of Computer Mathematics, 2015, 92(8): 1536-1550.

(下转第 167 页)