第44卷第7期 中国科学技术大学学报 Vol. 44, No. 7
2014年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2014

文章编号:0253-2778(2014)07-0570-06

基于 Finsler 几何的 k-means 算法

许 晴,李凡长,邹 鹏

(苏州大学计算机科学与技术学院,江苏苏州 215006)

摘要:针对 k-means 算法存在的相似性度量、准则函数优化效果不理想及多维流形数据分析性能效果不好等问题,引入 Finsler 几何中的 Finsler 度量,提出了一种基于 Finsler 几何的 k-means 算法,并在 UCI 数据集和 ORL 人脸数据库上与传统 k-means 算法及 SBKM 算法进行了比较,实验结果验证了该算法的可行性和有效性.

关键词:Finsler 几何;Finsler 度量;k-means 算法;相似性度量;准则函数

中图分类号:TP311 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2014.07.005

引用格式: Xu Qing, Li Fanzhang, Zou Peng. The k-means algorithm based on Finsler geometry[J]. Journal of University of Science and Technology of China, 2014,44(7):570-575.

许晴,李凡长,邹鹏. 基于 Finsler 几何的 k-means 算法[J]. 中国科学技术大学学报,2014,44(7):570-575.

The k-means algorithm based on Finsler geometry

XU Qing, LI Fanzhang, ZOU Peng

(*School of Computer Science and Technology, Soochow University, Suzhou* 215006, *China*)

Abstract: The problems with the k-means algorithm that the optimization effect of similarity measure and criterion function is not ideal and the analysis performance of multi-dimensional manifold data is ineffective, a modified version based on Finsler geometry was proposed, which introduces Finsler metric. Experimental results in comparison with traditional k-means algorithm and SBKM algorithm on UCI data sets and ORL face image sets show the feasibility and effectiveness of the algorithm.

Key words: Finsler geometry; Finsler metric; k-means algorithm; similarity measure; criterion function

0 引言

k-means 算法是 MacQueen[1] 提出的一种无监督聚类算法,该算法已在机器学习、数据分析等领域得到广泛关注.它的核心思想是把 n 个数据对象划分为 k 个聚类,使得每个聚类中的数据点到该聚类中心的平方和最小.该算法原理简单且便于处理大批量数据,具有收敛速度快、聚类效果好的优势.同时 k-means 算法也存在不足之处,如 k-means 算法多采用欧式距离度量数据对象之间的相似度,通常只能发现数据对象分布较均匀的球状簇,而无法发现任意簇;另外欧式距离在现实的二维及三维空间中有明确的几何意义,但在高维空间情况易会失去几何直观意义.针对 k-means 算法的不足,许多学者

收稿日期:2014-03-21;**修回日期**:2014-06-15

基金项目:国家自然科学基金(61033013,60775045),东吴学者计划,苏州大学敬文书院"3I 工程"重点项目资助.

作者简介:许晴,女,1990 年生,硕士生. 研究方向:机器学习. E-mail: 20124227022@suda.edu.cn

通讯作者:李凡长,教授. E-mail: lfzh@suda.edu.cn

做了许多改进[2-10]. 如文献[5]考虑了类结构中的多种对称性质，提出了一种改进的基于类对称的距离度量，可以有效发现超球面以外的具有较对称形状的簇；文献[9]用马氏距离取代传统 k-means 算法中的欧式距离度量，可检测超椭球集群；文献[10]提出了一种新的相似性度量函数，较好地克服了传统距离函数在高维空间的缺点.

这些成果虽然对单流形结构有较好的效果，但对多流形问题就显得不是很有效了，因此，本文尝试利用 Finsler 几何来研究多流形结构上的 k-means 问题. Finsler 几何是研究具有 Finsler 度量的空间几何性质的学科. 在数学定义中，度量是用来测量空间中两点之间的“距离”，从一点到另一点的距离函数称为度量. 常见的度量有：欧氏度量、黎曼度量等. 欧氏度量是目前最常见的度量，衡量的是多维空间中两点之间的绝对距离，但是不能直接地用在流形上. 黎曼度量正是为了在微分流形上建立曲线的长度等概念而产生的，它的度量形式是一个正定的(至少是非退化的)二次微分式. 相比于这些度量，Finsler 几何中的 Finlser 度量更具有一般性，它不受二次式度量形式的限制，且度量本身受高维空间的影响很小，这正好弥补了上述 k-means 算法中相似性度量、准则函数优化效果不理想及多维流形数据分析性能效果不好的缺陷.

1　k-means 算法回顾

k-means 算法的基本思想是：首先指定需要划分的簇的个数值 k，然后随机地选择 k 个初始数据对象作为初始的聚类中心；计算其余的各数据对象到这 k 个初始聚类中心的距离(这里一般采用欧式距离作为相似性度量)，把数据对象划归到距离它最近的那个中心所处在的簇类中；最后，调整新类并且重新计算出新类的中心，如果两次计算出来的聚类中心未曾发生任何的变化，那么就可以说明数据对象的调整已经结束，也就是说聚类采用的准则函数(这里一般采用的是误差平方和准则函数)是收敛的，表示算法结束.

k-means 算法中假定所有数据对应于 n 维空间 R^n 中的点，数据对象与聚类中心间的距离一般是根据欧式距离定义的. 数据对象 $\boldsymbol{x}$ 表示为特征向量 $\boldsymbol{x}=(x^1, x^2, \cdots, x^p)$，$x^b$ 表示样本 $\boldsymbol{x}$ 的第 p 个特征的值.

那么数据对象 x_i 与聚类中心 q_j 间距离表示为：

$$d(x_i, q_j)=\sqrt{\sum_{l=1}^{p}(x_i^l-q_j^l)^2} \tag{1}$$

误差平方和准则函数则表示为：

$$J=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\| x_i-q_j\|^2 \tag{2}$$

式中，q_j 是 k 个集合的中心，可以用来代表 k 个类型，q_j 的值即类 j 中样本的均值：$q_j=\frac{1}{n_j}\sum_{i=1}^{n_j}x_i$，$j=1,2,\cdots,k$.

2　Finsler 几何的基本概念

Finsler 几何是较黎曼几何学更具一般性的一种度量几何学，下面先对 Finsler 几何的基本概念进行介绍.

定义 2.1　Finsler 函数

我们知道黎曼几何学是基于正定的(至少是非退化的)二次微分式 $\mathrm{d}s^2=g_{ij}(u)\mathrm{d}u^i\otimes\mathrm{d}u^j$ 的度量几何学，其中 u^i 是局部坐标，$g_{ij}=g_{ji}$ 是该流形上的光滑函数. 而 Finsler 几何则考虑更一般的情形，不受度量形式是二次式的限制. 为此假定 $\mathrm{d}s=F(u^1,\cdots,u^m;\mathrm{d}u^1,\cdots,\mathrm{d}u^m)$. 其中，$F(x;y)$ 是有 $2m$ 个自变量的非负光滑函数，并且仅当 $y=0$ 时为零，称为 Finsler 函数. 同时还要求 $F(x;y)$ 关于自变量 y 是一阶齐次函数，即

$$F(x^1,\cdots,x^m;\lambda y^1,\cdots,\lambda y^m)=|\lambda|F(x^1,\cdots,x^m;y^1,\cdots,y^m),\ \forall\lambda\in R \tag{3}$$

定义 2.2　Finsler 度量

设 M 是一个 n 维光滑流形，$F:\mathrm{TM}\to[0,+\infty)$ 是其切丛上的非负函数. 如果 F 满足如下条件：

(Ⅰ) 正齐性：$F(x,\lambda y)=\lambda F(x,y)$，$\forall\lambda>0$；

(Ⅱ) 光滑性：在带孔切丛 $\mathrm{TM}\backslash\{0\}$ 上 $F(x;y)$ 是 C^∞ 函数；

(Ⅲ) 正则性：对于任意非零向量 $y\neq0$，

$$g_{ij}(x,y)=\frac{1}{2}\frac{\partial^2F^2}{\partial y^i\partial y^i}(x,y)=\frac{1}{2}[F^2]_{y^iy^i}$$

构成正定的矩阵，则称 F 为 M 上的一个 Finsler 度量.

3　基于 Finsler 几何的 k-means 算法设计

Finsler 度量是较黎曼度量更一般的一种度量，由于其定义的一般性，我们可以构造出很多 Finsler

度量.国内外很多学者对 Finsler 几何进行了研究,并给出了一些具有良好性质或易于计算的 Finsler 度量[11-17].(α,β)度量便是其中一类丰富的可计算的 Finsler 度量,它们在 Finsler 几何中扮演着非常重要的角色.

令 $F=\alpha\phi(s)$, $s=\dfrac{\beta}{\alpha}$.其中,$\alpha=\sqrt{a_{ij}(x)y^iy^i}$是黎曼度量,$\beta=b_i(x)y^i$ 为非零的 1 形式, $\phi=\phi(s)$定义在开区间$(-b_0,b_0)$且满足 $\phi(0)=1$, $\phi(s)>0$, $\phi(s)-s\phi'(s)+(b^2-s^2)\phi''(s)>0$, $|s|\leqslant b\leqslant b_0$,则对任意 $x\in M$,F 是流形 M 上正定的 Finsler 度量,称作(α,β)度量.可以很容易验证(α,β)度量是满足 Finsler 度量的定义的.证明如下:

令

$$\alpha_\lambda=\alpha(x^1,\cdots,x^m;\lambda y^1,\cdots,\lambda y^m),$$
$$\beta_\lambda=\beta(x^1,\cdots,x^m;\lambda y^1,\cdots,\lambda y^m);$$

$$\because\quad \alpha_\lambda=\alpha(x^1,\cdots,x^m;\lambda y^1,\cdots,\lambda y^m)=\sqrt{a_{ij}(x)\lambda y^i\lambda y^j}=|\lambda|\sqrt{a_{ij}(x)y^iy^j},$$
$$\beta_\lambda=\beta(x^1,\cdots,x^m;\lambda y^1,\cdots,\lambda y^m)=b_i(x)\lambda y^i=\lambda(b_i(x)y^i);$$

$$\therefore\quad F(x^1,\cdots,x^m;\lambda y^1,\cdots,\lambda y^m)=\alpha_\lambda\phi(\beta_\lambda/\alpha_\lambda)=|\lambda|\sqrt{a_{ij}(x)y^iy^j}\phi\left(\frac{\lambda}{|\lambda|}\frac{\beta}{\alpha}\right).$$

所以只需满足函数 ϕ 为偶函数且恒大于 0,即 $\phi(-s)=\phi(s)$且 $\phi(s)\geqslant 0$,就可以保证(α,β)度量为 Finsler 度量.

我们取(α,β)度量的简洁形式,即令 $\phi(s)=1+\dfrac{\beta}{\alpha}$,则有 $F=\alpha+\beta$, α,β 分别取:

$$\alpha=\frac{\sqrt{|y|^2+\mu(|x|^2|y|^2-\langle x,y\rangle^2)}}{1+\mu|x|^2},$$
$$\beta=\frac{\sqrt{-u}\langle x,y\rangle}{1+u|x|^2}.$$

其中,

$$u<0,\ \langle x,y\rangle=\sum_{i=1}^{n}x^iy^i,$$
$$|x|=\sqrt{\sum_{i=1}^{n}(x^i)^2},\ |y|=\sqrt{\sum_{i=1}^{n}(y^i)^2}.$$

至此,Finsler 度量公式可写为:

$$F=\frac{\sqrt{|y|^2+\mu(|x|^2|y|^2-\langle x,y\rangle^2)}}{1+\mu|x|^2}+\frac{\sqrt{-\mu}\langle x,y\rangle}{1+\mu|x|^2}\tag{4}$$

k-means 算法中数据对象 x_i 到聚类中心 q_j 的相似性度量可定义为:

$$d_{\text{Finsler}}(x_i,q_j)=\frac{\sqrt{|q_j|^2+\mu(|x_i|^2|q_j|^2-\langle x_i,q_j\rangle^2)}}{1+\mu|x_i|^2}+\frac{\sqrt{-\mu}\langle x_i,q_j\rangle}{1+\mu|x_i|^2}\tag{5}$$

其中,

$$u<0,\ \langle x_i,q_j\rangle=\sum_{l=1}^{p}x_i^lq_j^l,$$
$$|x_i|=\sqrt{\sum_{l=1}^{p}(x_i^l)^2},\ |q_j|=\sqrt{\sum_{l=1}^{p}(q_j^l)^2}.$$

相应地,k-means 算法中的准则函数可以定义为:

$$J=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(d_{\text{Finsler}}(x_i,q_j))^2\tag{6}$$

基于 Finsler 几何的 k-means 算法的基本流程见算法 3.1.

算法 3.1(FGK-means) 基于 Finsler 几何的 k-means 算法(Finsler geometric k-means, FGK-means)

输入:聚类数 k;包含 n 个数据对象的数据集 X

输出:k 个聚类

Step 1 从数据集 X 中任意选取 k 个对象作为初始聚类中心$(q_1^{(1)},q_2^{(1)},\cdots,q_k^{(1)})$;

Step 2 第 r 次迭代时,根据式(5)分别计算每个对象 x_i, $i=1,2,\cdots,n$ 到各个聚类中心的 Finsler 距离;

Step 3 找出每个数据对象关于聚类中心的最小距离,并将其归入与其距离最小的聚类中心所在的簇中;

Step 4 更新各簇的聚类中心 $q_j^{(r)}=\dfrac{1}{n_j}\sum_{i=1}^{n_j}x_{ji}$;

Step 5 根据式(6),计算准则函数 J_r,并与前一次 J_{r-1}相比,若$|J_{r+1}-J_r|<\delta$,则算法结束;否则转入 Step 2 进行再一次迭代.

4 实验结果与分析

为了验证本文算法的有效性,我们进行了两组实验.其中,第一组实验采用二维仿真数据集,以便对算法的聚类效果进行直观分析;第二组采用实物图像,分别在 UCI 数据库[18]以及 ORL 人脸数据库[19]上进行实验,并与传统 k-means 算法及文献

[5]中提出的 SBKM 算法进行比较. 本文实验均在 matlab 平台上实现.

4.1　实验一的结果比较与分析

实验 1:为了直观分析本文算法的聚类效果以及与传统 k-means 算法和 SBKM 算法进行比较,我们设计了一组人工数据作为输入数据集 X. 该人工数据是随机生成的三类二维连通图集数据,每次生成数据时每类样本是 100 个,我们给出了一次随机生成的数据,将其可视化如图 1(a)所示. 其中符号"$*$","°"和"▽"分别表示不同的类别,图中每个点的坐标表示的是每个二维数据所在的位置.

利用本文提出的 FGK-means 算法对数据集 X 进行聚类,我们输入聚类类别数 k 为 3,并任意选取初始聚类中心点((−0.002 3, −0.491 3),(0.019 8,0.123 5),(−0.041 2,0.152 0)),经过多次迭代,结果如图 1(b)所示. 传统 k-means 算法和 SBKM 算法对数据集 X 聚类后的结果分别如图 1(c)和 1(d)所示.

图 1(a)是原始生成数据的可视化图,其中随机生成的三类数据有较多重叠部分. 图 1(b),(c),(d)分别是用本文提出的 FGK-means 算法,传统 k-means 算法及 SBKM 算法对原始生成数据进行聚类得到的结果. 由图可以发现,本文算法相对另外两种算法可以较好地将三类数据分开.

4.2　实验二的结果比较与分析

实验 2:我们采用 UCI 数据库上的 Iris 及 Statlog(vehicle silhouettes)数据集作为待聚类样本,分别采用传统 k-means 算法、SBKM 算法及本文提出的 FGK-means 算法进行聚类,且为了消除样本选择的随机独立性,独立重复进行 100 次实验,取平均识别率作为最终识别率.

UCI 数据库是一个专门用于测试机器学习、数据挖掘算法的数据库,库中的数据都有确定的分类,可以直观表示聚类结果的质量. 其中,Iris 数据集取自 3 类鸢尾属植物的花朵样本,共含 150 个数据,每个数据有 4 维属性. Statlog 数据集取自 4 类交通工具轮廓样本,共含 946 个数据,每个数据有 18 维属性. UCI 数据集上的聚类识别率分别如图 2(a)和图 2(b)所示.

从图 2(a)和 2(b)可以看出,FGK-means 算法

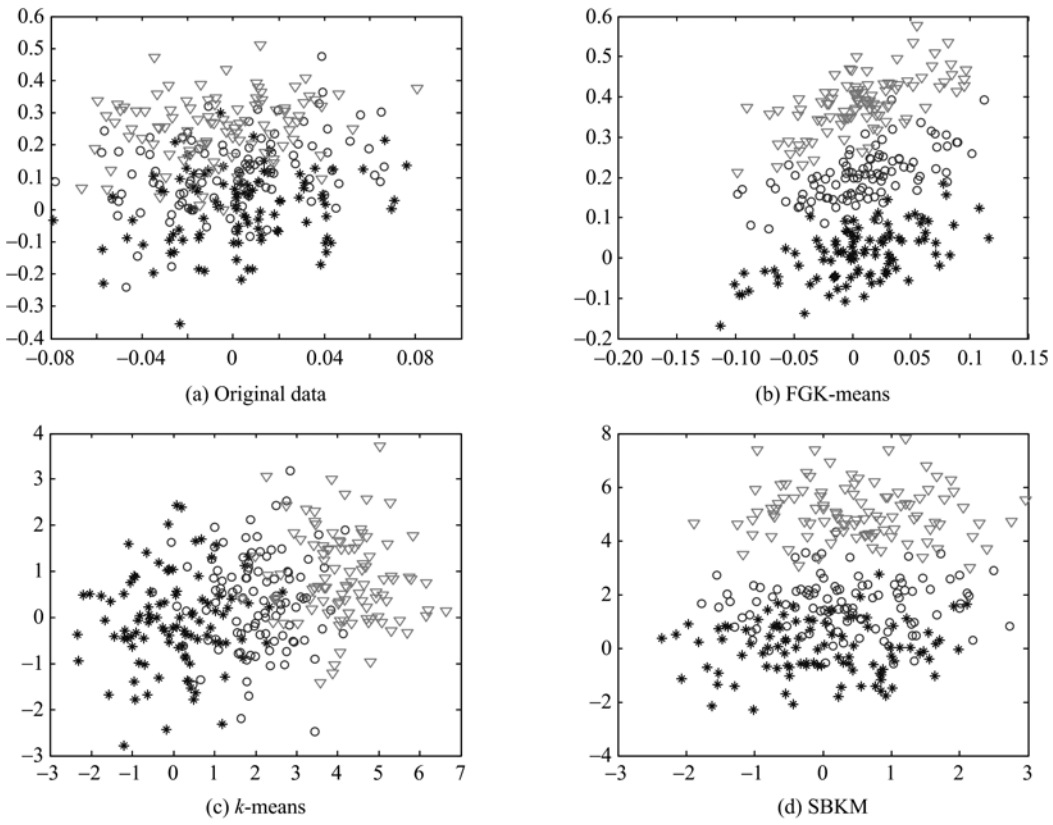

图 1　随机数据和可视化结果

Fig. 1　Random data and the result of visualization

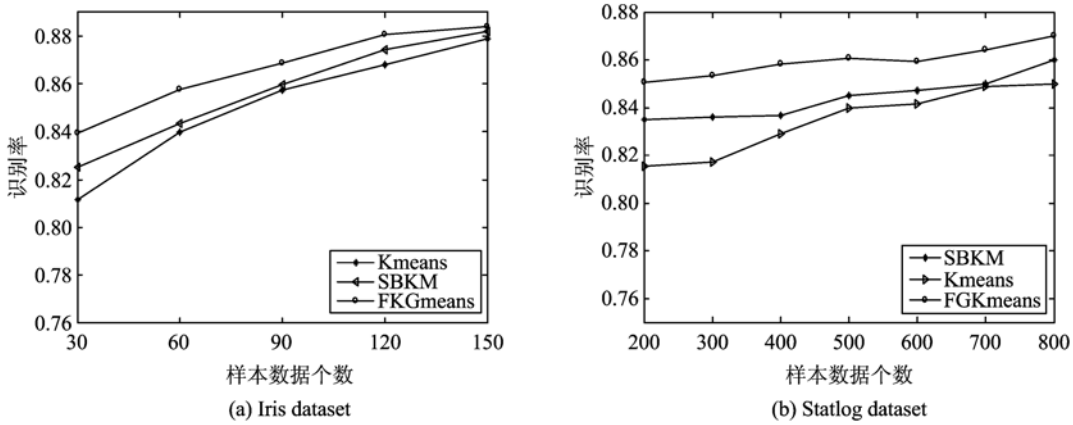

图 2 UCI 数据集上的聚类性能

Fig. 2 Clustering performance on the UCI dataset

的聚类性能明显优于传统 k-means 算法，且随着数据样本的增多，FGK-means 算法也比 SBKM 算法具有更好的识别精度. 以上结果表明，引入 Finsler 度量后，能够提高算法的聚类性能.

4.3 实验三的结果比较与分析

实验 3：我们验证 FGK-means 算法在 ORL 人脸数据库上的实验效果. ORL 人脸数据库由剑桥大学 AT&T 实验室创建，包含 40 个不同的人，每个人有 10 幅人脸面部图像，分别在不同时间，不同光照条件，不同姿势表情的情况下拍摄得到的，每幅图像灰度级为 256，分辨率为 112×92，部分图像如图 3 所示.

图 3 ORL 人脸数据库部分图像

Fig. 3 Part of the images in ORL

实验中，我们从 ORL 人脸库中随机选取一定数量的样本作为输入数据集，大小从 120 变化到 320，且为了消除样本选择的随机独立性，我们独立重复进行 100 次实验，并取平均识别率作为最终识别率. 图 4 给出了传统 k-means 算法，SBKM 算法以及本文提出的 FGK-means 算法在 ORL 人脸数据库上不同训练样本个数情况下的实验结果.

从图 4 可以看出，随着样本数据量的增加，3 个算法的识别率整体呈上升趋势，虽然本文提出的 FGK-means 算法未能保持稳定的聚类能力，但与传统 k-means 算法和 SBKM 算法相比较，本文提出的

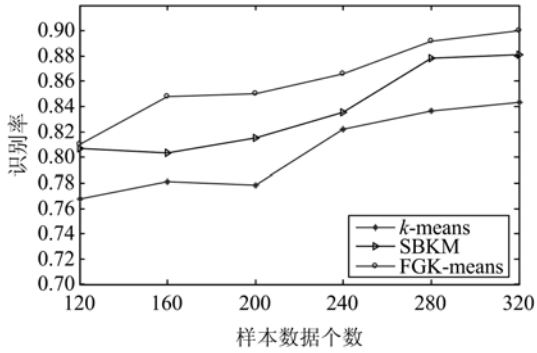

图 4 ORL 人脸库上的聚类性能

Fig. 4 Clustering performance on the ORL face image sets

FGK-means 算法的聚类性能更优.

5 结论

本文针对 k-means 算法相似性度量、准则函数优化效果不理想及多维流形数据分析性能效果不好等不足，引入 Finsler 几何中度量形式不受二次式限制的 Finsler 度量，优化了 k-means 算法的相似性度量、准则函数及多维流形结构上分析数据的性能，提出了基于 Finsler 几何的 k-means 算法（FGK-means），并在 UCI 数据集和 ORL 人脸数据库上与传统 k-means 算法及文献[5]中提出的 SBKM 算法进行了对比实验. 实验结果表明，FGK-means 算法相较传统 k-means 算法及 SBKM 算法在 UCI 数据集和 ORL 人脸数据库上的聚类效果得到了有效的改进，但随着样本数据量的增加，FGK-means 算法未能保持稳定的聚类性能. 因此如何选取更一般的 Finsler 度量函数，且充分运用 Finsler 几何的其他知识，将是我们下一步工作的重点.

参考文献(References)

[1] MacQueen J B. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. London, UK: Cambridge University Press, 1967: 281-297.

[2] Jain A K. Data clustering: 50 years beyond k-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[3] Kashima H, Hu J Y, Ray B, et al. k-means clustering of proportional data using L_1 distance[C]//Proceedings of the 19th International Conference on Pattern Recognition. Tampa, USA: IEEE Press, 2008: 1-4.

[4] Banerjee A, Merugu S, Dhillon I S, et al. Clustering with Bregman divergences [J]. The Journal of Machine Learning Research, 2005, 6: 1 705-1 749.

[5] Su M C, Chou C H. A modified version of the k-means algorithm with a distance based on cluster symmetry [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6):674-680.

[6] Wang L, Bo L F, Jiao L C. A modified k-means clustering with a density-sensitive distance metric[C]// Proceedings of the First International Conference on Rough Sets and Knowledge Technology. Berlin Heidelberg: Springer, 2006: 544-551.

[7] Liu C R, Hu T M, Ge Y, et al. Which distance metric is right: An evolutionary k-Means view [C]// Proceedings of the SIAM International Conference on Data Mining. Anaheim, USA: SIAM Press, 2012: 907-918.

[8] Linde Y, Buzo A, Gray R M. An algorithm for vector quantizer design [J]. IEEE Transactions on Communications, 1980, 28(1): 84-95.

[9] Mao J C, Jain A K. A self-organizing network for hyperellipsoidal clustering (HEC) [J]. IEEE Transactions on Neural Networks, 1996, 7(1): 16-29.

[10] Yang F Z, Zhu Y Y. An efficient method for similarity search on quantitative transactions data [J]. Journal of Computer Research and Development, 2004, 41(2): 361-368.
杨风召,朱扬勇.一种有效的量化交易数据相似性搜索方法[J].计算机研究与发展,2004, 41(2): 361-368.

[11] Chen G Z, Cheng X Y, Yuan M G. On a class of projectively flat Finsler metrics with weakly isotropic flag curvature [J]. Periodica Mathematica Hungarica, 2013, 67(2): 155-166.

[12] Matsumoto M. On Finsler spaces with Randers' metric and special forms of important tensors [J]. Journal of Mathematics of Kyoto University, 1974, 14(3): 477-498.

[13] Cui N W, Shen Y B. Projective change between two classes of (α, β)-metrics [J]. Differential Geometry and its Applications, 2009, 27(4): 566-573.

[14] Chern S S, Shen Z M. Riemann-Finsler Geometry [M]. Singapore: World Scientific, 2005.

[15] 沈一兵, 沈忠民. 现代芬斯勒几何初步[M]. 北京: 高等教育出版社, 2013.

[16] 陈省身, 陈维桓. 微分几何讲义 [M]. 二版, 北京: 北京大学出版社, 2001.

[17] 李凡长, 张莉, 杨季文, 等. 李群机器学习[M]. 合肥: 中国科学技术大学出版社, 2013.

[18] Machine Learning Repository [EB/OL]. http://archive.ics.uci.edu/ml/datasets.html.

[19] http://cs.nyu.edu/~roweis/data/olivettifaces.mat.