DOI：10.7652/xjtuxb201712008

一种适用于高维非线性特征数据的聚类算法及应用

姜洪权[1]，王岗[2]，高建民[1]，高智勇[1]，高瑞琪[1]，郭旗[2]
（1.西安交通大学机械制造系统工程国家重点实验室，710049，西安；2.西安交通大学第二附属医院，710004，西安）

摘要：针对高维数据聚类分析中数据之间具有多种非线性特征关系，导致数据分布不均、传统相似性度量失效及结果类中心难以精准表征等问题，提出了一种基于核主元分析（KPCA）与密度聚类（DBSCAN）的高维非线性特征数据聚类分析技术。首先，为有效提取高维数据的非线性特征，利用KPCA理论将原始数据映射到更高维数据空间，利用主元分析获得数据变化的方向集合，并进行降维分析；然后，通过重新定义数据样本在主元空间的相似性距离对传统DBSCAN聚类方法进行改进，并利用3δ统计理论对各簇中心的进行表征，从而实现高维数据的精确分类与类中心知识表达。以实际高血压患者群体聚类问题为例对方法进行了有效性验证，实验表明，所提方法可以有效获取原始数据的非线性特征，实现患者个体特征群体的有效划分及簇类中心知识的表达，解决传统DBSCAN聚类方法对高维数据不适用的问题。

关键词：非线性；高维数据；核主元分析；密度聚类
中图分类号：TP391.3 **文献标志码**：A **文章编号**：0253-987X(2017)12-0049-07

A Clustering Algorithm for High-Dimensional Nonlinear Feature Data with Applications

JIANG Hongquan[1], WANG Gang[2], GAO Jianmin[1], GAO Zhiyong[1], GAO Ruiqi[1], GUO Qi[2]
(1. State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710004, China)

Abstract: Aiming at the problems caused by the nonlinear relations between the attributes of high dimensional data in cluster analysis, such as uneven distribution of data, invalidation of traditional similarity measures and difficulty of accurate representation of the result class, a clustering algorithm for high dimensional nonlinear feature data is proposed based on kernel principal component analysis (KPCA) and density clustering (DBSCAN). To extract the nonlinear characteristics of high dimensional data, the KPCA theory is adopted to map the original to a higher dimensional data space, thus a set of directions in principal component space-PCS for extracting the nonlinear characteristics of data and reduced dimensions can be obtained. The similarity distance of data in PCS is defined to improve the traditional DBSCAN clustering algorithm and 3δ statistical theory is used to characterize the clustering results. A case of hypertension group clustering is provided to illustrate the feasibility of the proposed method, and the results show that the proposed method can effectively obtain the nonlinear characteristics of

收稿日期：2017-05-19。 作者简介：姜洪权(1978—)，男，讲师，博士生导师。 基金项目：国家自然科学基金资助项目(51375375)；中央高校基本科研业务费专项资金资助项目(xjj2014108)。
网络出版时间：2017-08-27 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20170827.1943.002.html

the high dimensional data and realize cluster analysis and cluster center knowledge expression to solve the difficulties in the traditional DBSCAN clustering method for cluster analysis of high dimensional data.

Keywords: nonlinear feature; high dimensional data; kernel principal component analysis; density clustering

随着数据采集技术及信息技术的发展,具有高维数据特征的对象聚类分析在如医学数据分析、购物模式分析、图像模式分析等众多领域得到了关注。聚类分析是一种在数据集中寻找相似元素集合的无监督学习过程,其目的是发现数据集中的分布特征[1]。聚类方法通常可以分 5 类[2-3]:①基于密度的聚类方法,如 DBSCAN 和 OPTICS;②基于划分的方法,如 K-means 算法;③基于层次的方法,如 CURE 和 BIRCH 等;④基于格网的方法,如 STING、CLIQU 等;⑤基于模型的方法,如 COBWEB 等。其中,DBSCAN 聚类方法是经典的基于密度的聚类方法,具有可以在数据集中划分出任意形状的簇集合、速度快等特点。然而,由于高维的存在,使得许多适用于低维数据的聚类方法不再有效或有效性大大降低[4]。

传统 DBSCAN 聚类在应用于高维数据时存在以下缺点:当数据维数较高时,由于各维数据间非线性关系导致的分布不均、稀疏性以及空空间现象的存在,使得传统基于欧式距离的对象相似性度量失效,直接影响密度计算的准确性;同时,当完成聚类后,需要获得准确的类知识表达,需要对簇中心进行表征。现有聚类方法缺乏准确的簇中心表征方法,或仅仅以簇的几何中心进行替代,容易受到簇内边界点影响。

针对以上问题,本文提出了一种实现数据非线性特征为核心的高维空间数据聚类方法。通过对高维数据进行核主元分析(KPCA),实现原始数据的非线性特征提取及降维处理,定义了一种适用于高维数据对象相似性的度量方法进行聚类分析,最后利用 3δ 理论对聚类结果进行有效的表征。以医学领域高血压患者群体分析问题为例对本文方法进行了验证,表明该算法解决了高维非线性特征数据下的聚类问题,可以有效辨识患者特征群体。

1 高维数据下 DBSCAN 算法分析

1.1 DBSCAN 算法简介

DBSCAN 是利用基于密度的聚类概念进行分析的,即要求聚类空间的某区域 E 中所包含的对象数目不小于给定值($P_{\min}$),主要概念[5]如下。

定义 1 E 邻域。给定对象半径 E 内的区域称为该对象的 E 邻域。

定义 2 核心对象。给定 E 和 $P_{\min}$,若对象 P 的 E 邻域 $N_E(P)$包含的对象个数 $|N_E(P)|\geqslant P_{\max}$,则称 P 为核心对象。

定义 3 直接密度可达。给定 E 和 $P_{\min}$,当

(1)$P\in N_E(Q)$且

(2)$|N_E(Q)|\geqslant P_{\max}$

则称对象 P 是从对象 Q 出发直接密度可达。

定义 4 密度可达。给定对象集合 D,当存在一个对象链 $P_1,P_2,\cdots,P_m$,且 $P_1=Q,P_n=P$,对 $P_i\in D,P_{i+1}$是从 P_i 关于 E 和 $P_{\min}$直接密度可达,则称对象 P 是从对象 Q 关于 E 和 $P_{\min}$密度可达。

定义 5 簇和噪声。通过密度可达性连接而成的对象集合称为簇,不在任何簇中的对象被认为是噪声。

通过以上定义看出,DBSCAN 算法通过给定邻域 E 和 $P_{\min}$两个参数,按搜索符合密度特征的对象点进行聚类,具有速度快且可发现任意形状聚类的优势。

1.2 高维数据非线性特征对聚类的影响

在高维数据表示的聚类对象中,各维属性之间存在多种关系,如线性关系和非线性关系等。不失一般性,如图 1 所示,在(X,Y)二维数据表示的对象中,若两个维度 X 与 Y 之间仅仅存在线性关系,则对象的分布应该围绕该线性函数进行分布;若在平面上呈现分布不均现象,则说明两个维度的属性指标 X 与 Y 之间存在多种非线性特征关系,即在任意一个对象密度集中区域都是由 X 与 Y 之间的非线性关系形成的。显然,在大于 2 维的高维数据对象中这种特点也是存在的。

由于 DBSCAN 算法是利用对象密度进行聚类分析的,当密度分布不均时,将对分类结果产生重要的影响[6]。如图 1 所示,直观上在(X,Y)二维平面上对象共可分为 5 个类。从对象密度上可以看出,C_1、C_2 和 C_3 密度较大且相邻较近,C_4 和 C_5 密度较小且相邻较近。当采用传统 DBSCAN 算法进行聚类分析时,若按较小的密度值进行聚类,则 C_1、C_2 和 C_3 被分为一类,C_4、C_5 各为一类,即最后聚类结

果是 3 个类；若按较大的密度值进行聚类，则聚类结果将是 C_1、C_2 和 C_3 共 3 个类，C_4 和 C_5 则会被作为噪声点导致簇类丢失。因此，在对高维数据对象进行聚类分析时，如何捕获数据间具有的多种非线性关系特征并指导分类过程，会有效提高传统 DBSCAN 的聚类效果。

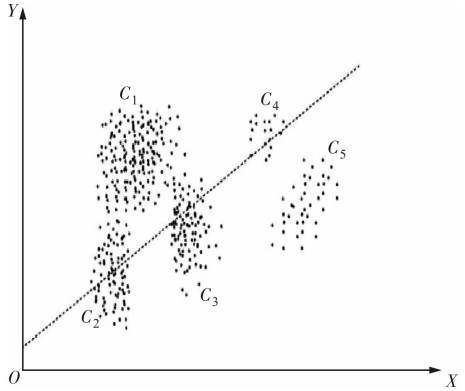

图 1　数据属性间非线性关系对聚类的影响

1.3　高维数据的相似性度量对聚类的影响

聚类方法核心步骤之一是定义数据对象间的相似性度量，即通过定义数据之间的距离对聚类对象进行差异性描述。对与基于密度的 DBSCAN 算法而言，在确定邻域 E 后，需要通过距离建立对象之间的相似性度量，并以此为依据进行 E 邻域内密度计算。由于高维数据稀疏性和空空间现象的存在，导致高维空间产生维度灾难[7]，导致传统的相似度度量对高维数据聚类有效性大大降低。

假设含有 N 个数据的 d 维数据空间，采用欧式距离，确定对象之间的最临近距离 D[8] 为 $2\left[\frac{\mathrm{d}\Gamma(d/2)}{2\pi^{d/2}N}\right]^{1/d}$，其中 N 是数据个数，d 是数据维数。D、N 及 d 之间的关系如图 2 所示，可以看出，最小临近距离 D 随维数 d 的增长而增加（如图 2 中左侧曲线所示）；对于给定维数，最小临近距离 D 随着数据量 N 的增长而减小。可见，对于高维数据，维数及数据量都会对分类对象的最小临近距离产生影响，进而影响对象之间的相似性度量准确性。实际上，在高维数据中，有些维度的相似性较高，而有些维度的相似性较低。DBSCAN 算法以欧式距离为基础进行相似性度量，没有考虑以上高维数据中各维度相似性与否的特点，从而影响相似性及 E 邻域的密度度量，最终导致 DBSCAN 算法分类有效性降低。

1.4　高维数据的类中心表征问题

当完成聚类分析后，需要对得到簇进行表征，即

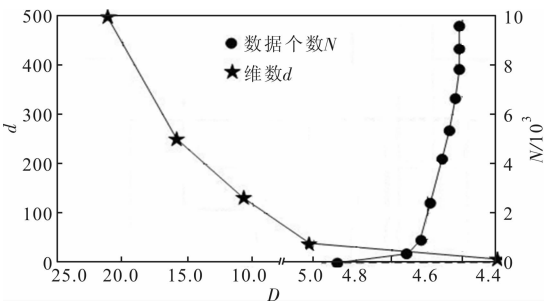

图 2　最小临近距离 D 与维度 d、数据 N 之间的关系

获得簇的特征信息。与其他聚类算法一样，DBSCAN 算法聚类结果仅仅是将有共同特征的对象进行聚类，并没有给出每个簇的表征知识。一般的，较为常用的做法是利用聚类得到的簇均值或者重心特征进行类的表征，但容易受到簇的边界对象点（野点）影响，尤其对高维数据的簇类中心表征时，影响更为重要。

如图 3 所示，A_1，A_2，…，A_n 为原始簇的边界点，C_1 为原始簇的中心表征。C_2 为去除簇的边界点 A_n 后的簇区域的类中心。显然，簇的边界点对类中心 C_1 的确定有着重要的影响，尤其是偏离簇中心较大的边界点 A_1，A_2，…，A_n 的影响。可以看出，去除边界点影响后的 C_2 作为类中心更能准确实现统计意义上的簇中心表征。因此，在确定类中心表征时，需要消除边界点对簇中心的影响。

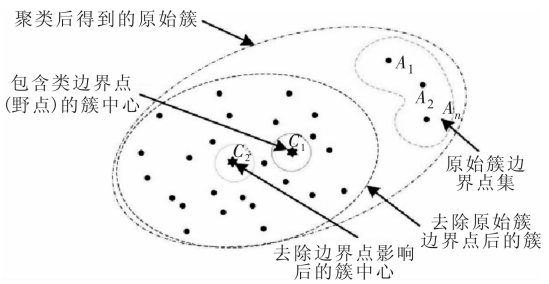

图 3　类中心表征示意图

需要说明的是，图 3 是聚类示意图，由于 A_1，A_2，…，A_n 在所设定密度阈值的约束条件下，所以只能与其他数据形成一个类，不能自身形成一个新类。

2　高维数据的 KDBSCAN 算法

2.1　本文算法的核心思想

依据本文第 1 节所述，针对高维数据具有分布不均、各维度属性之间存在多种非线性关系、以及簇类中心表征不准确等问题，本文提出了一种基于 KPCA 与 DBSCAN 算法的高维非线性特征数据聚类分析技术——KDBSCAN 算法，采用 KPCA 分析

方法对原始数据进行非线性特征提取[9]。在此基础上,定义一种综合考虑对象各维度相似程度的相似性度量方法,代替传统 DBSCAN 算法中的欧氏距离,来度量不同对象之间的相似度,并实现聚类分析。最后,在聚类得到的对象簇基础上,引入 3δ 统计分析理论,对簇内边界点进行去除[10],并建立 3δ 统计意义上的均值,对簇中心进行表征,从而提高簇中心表征的准确性。

2.2 基于 KPCA 的高维数据非线性特征提取及降维

设原始数据 X 的维数为 n,通过非线性函数 Φ(核函数)映射到高维特征空间 $\boldsymbol{H}$,则原始数据 $\boldsymbol{x}_i$ 在映射 $\boldsymbol{H}$ 的像为 $\boldsymbol{\varphi}(\boldsymbol{x}_i)$[11]。假设映射数据是零均值,则映射数据 $\boldsymbol{\varphi}(\boldsymbol{X})$的协方差矩阵可以表示为

$$\bar{\boldsymbol{C}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}(\boldsymbol{x}_i)\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_i) \tag{1}$$

对 $\bar{\boldsymbol{C}}$ 做特征矢量分析,可得特征值为 λ、特征矢量为 $\boldsymbol{V}$,则有 $\lambda\boldsymbol{V}=\bar{\boldsymbol{C}}\boldsymbol{V}$。将每个样本与该式做内积,可得

$$\lambda[\boldsymbol{\varphi}(\boldsymbol{x}_k)\cdot\boldsymbol{V}] = [\boldsymbol{\varphi}(\boldsymbol{x}_k)\cdot\bar{\boldsymbol{C}}\boldsymbol{V}] \tag{2}$$

式中:$k=1,2,\cdots,n$。

此时,存在一组系数 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_3,\cdots,\boldsymbol{\alpha}_n$,使得 $\bar{\boldsymbol{C}}$ 的特征矢量可以用 $\boldsymbol{\varphi}(\boldsymbol{x}_i)$线性表示为

$$\boldsymbol{V} = \sum_{j=1}^{n}\alpha_j\boldsymbol{\varphi}(\boldsymbol{x}_j) \tag{3}$$

联合式(1)~式(3),可得

$$\lambda\sum_{j=1}^{n}\alpha_j[\boldsymbol{\varphi}(\boldsymbol{x}_k)\cdot\boldsymbol{\varphi}(\boldsymbol{x}_j)] = \frac{1}{n}\sum_{j=1}^{n}\alpha_j[\boldsymbol{\varphi}(\boldsymbol{x}_k)\cdot\sum_{i=1}^{n}\boldsymbol{\varphi}(\boldsymbol{x}_i)][\boldsymbol{\varphi}(\boldsymbol{x}_j)\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i)] \tag{4}$$

定义一个 $n\times n$ 对称矩阵 $\boldsymbol{K}$,令

$$\boldsymbol{K}_{ij} = [\boldsymbol{\varphi}(\boldsymbol{x}_j)\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i)] \tag{5}$$

将式(5)代入式(4),则式(4)可以表示为 $n\lambda\boldsymbol{K}\boldsymbol{\alpha}=\boldsymbol{K}\boldsymbol{K}\boldsymbol{\alpha}$,其中 $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n]^{\mathrm{T}}$,即

$$n\lambda\boldsymbol{\alpha} = \boldsymbol{K}\boldsymbol{\alpha} \tag{6}$$

因此,在特征空间中的主元分析就转化为求解式(6),得到特征值 $\lambda_1\geqslant\lambda_2\geqslant\lambda_3\geqslant\cdots\geqslant\lambda_n$ 以及相应的特征向量 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n$,选择前 k 个特征向量对特征空间进行表述,由式(3)可得

$$\boldsymbol{v}_k = \sum_{i=1}^{n}\alpha_i^k\boldsymbol{\varphi}(\boldsymbol{x}_i) \tag{7}$$

$$[\boldsymbol{v}_k\cdot\boldsymbol{v}_k] = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i^k\alpha_j^k k_{ij} = \lambda_k[\boldsymbol{\alpha}_k\cdot\boldsymbol{\alpha}_k] \tag{8}$$

通过 $\boldsymbol{V}$ 的归一化,对 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n$ 进行标准化处理,则式(8)化为

$$[\boldsymbol{v}_k\cdot\boldsymbol{v}_k] = 1 \tag{9}$$

计算映射数据 $\boldsymbol{\varphi}(\boldsymbol{x})$在特征空间 $\boldsymbol{H}$ 中特征向量 $\boldsymbol{v}_k$ 上的投影,即可得到非线性映射数据的主元向量

$$\boldsymbol{T}_k = [\boldsymbol{v}_k\cdot\boldsymbol{\varphi}(\boldsymbol{x})] = \sum_{i=1}^{n}\alpha_i^k[\boldsymbol{\varphi}(\boldsymbol{x}_i)\cdot\boldsymbol{\varphi}(\boldsymbol{x})] = \sum_{i=1}^{n}\alpha_i^k k(\boldsymbol{x}_i\cdot\boldsymbol{x}) \tag{10}$$

式中:$k=1,2,\cdots,n$。

对于主元个数 k 的选择,可通过特征值累积方差贡献率进行确定

$$V_{\mathrm{CPV}} = \sum_{j=1}^{k}\lambda_j\Big/\sum_{j=1}^{n}\lambda_j \tag{11}$$

V_{CPV}的值越大说明该主成分所包含的信息量就越强。一般的,当 $V_{\mathrm{CPV}}\geqslant 0.85$ 时,代表所选择的主元个数已经能够包含原始数据足够的信息量。

通过上述步骤,即可将原始 n 维数据降到 k 维指标数据,且各主元特征向量是正交独立的,实现了原始指标数据的冗余信息消除;同时,由于原始数据是通过核函数进行高维空间映射后,并获取的主元方向的变化度量,因此也实现了原始各维数据之间非线性关系特征提取。

2.3 综合考虑各维相似程度的对象相似性度量

如 1.3 节所述,对象之间的相似性度量是实现聚类分析的关键环节。在高维数据中,若对象的相似的维数越多,则对象相似性程度越高。传统基于欧式距离的对象相似性度量是将对象所有维数等同考虑,无法体现上述高维数据相似性表征这一特点。文献[12]提出了一种基于 $\mathrm{Hsim}(\boldsymbol{X},\boldsymbol{Y})$的对象之间相似性度量方法,体现了以上高维数据间相似性表征的特点。本文以 KPCA 处理后的非线性特征数据进行分析,因此需要重新定义一种适用于高维非线性特征的对象相似性度量方法。

假设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是 d 维空间中的两个点,$\boldsymbol{X}=(x_1,x_2,\cdots,x_d)$,$\boldsymbol{Y}=(y_1,y_2,\cdots,y_d)$,其中 X_i 与 $Y_i(i=1,2,\cdots,d)$是原始数据经 KPCA 处理后在主元方向的投影,则两个对象间的距离为

$$D_{\mathrm{istw}}(X_i,Y_j) = \Big(\sum_{n=1}^{k}\frac{1}{1+|\,t_{in}-t_{jn}\,|^2}\Big)^{1/2}\Big/k \tag{12}$$

式中:X_i、Y_j 为 KPCA 降维处理后的第 i 个和第 j 个对象;k 为主元个数,也是降维后的维度;t_{in} 为第 i 个对象的第 n 个数据。

由式(12)可知:$D_{\mathrm{istw}}(\boldsymbol{X},\boldsymbol{Y})\in[0.1]$,当为最小值 0 时,表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在对象各维上的差都接近于无穷大,即 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的相似性最小;当为 1 时,表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在对象各维上的值相等,即相似性最高。

2.4　聚类后簇中心的计算

高维数据的簇中心易受到簇边界点的影响，导致传统依据各维数据的均值计算类中心不准确。因此，本文采用 3σ 统计意义上的均值 μ' 进行类中心计算，以减小边缘值（过小值或者过大值）的影响，主要步骤如下：

（1）假设 $\boldsymbol{x}_i \in n \times m$ 为一个聚类，求取 $\boldsymbol{X}_i$ 聚类中所有的点坐标的平均值 μ 及标准偏差 σ；

（2）根据 μ 以及 σ，剔除超出 $\mu \pm 3\sigma$ 范围的边缘值，得到 n' 个点，获得新的数据集为 $\boldsymbol{x}'_i \in n' \times m$，按照下式求取平均值 μ' 作为 X_i 聚类的中心

$$\mu' = \frac{1}{n'} \sum_{k=1}^{n'} f_{ki}, \quad i = 1,2,\cdots,m \tag{13}$$

式中：m 为原始数据的维度；f_{ki} 表示矩阵 $\boldsymbol{x}'_i$ 的第 k 行第 i 列，即第 k 个样本的第 i 个变量。

综上所述，本文所提 KDBSCAN 算法的主要步骤如下：

（1）利用 2.2 节对原始数据进行特征提取及降维分析，定义获得结果数据集为 $\boldsymbol{T}$；

（2）给定参数 E 和 $P_{\min}$，对于所有数据对象的距离采用 2.3 节相似性度量进行计算，并依据给定 E 和 $P_{\min}$ 进行聚类分析，主要过程包括检查数据集 T 中未检查过的对象 P，如果 P 未被处理（归入某个簇或标记为噪声），则检查其 E 邻域 $N_E(P)$，若 $N_E(P)$ 包含的对象数不小于 $P_{\min}$，建立新簇 C，将 $N_E(P)$ 中所有点加入 C；

（3）对 C 中所有尚未被处理的对象 Q，检查其 E 邻域 $N_E(P)$，若 $N_E(Q)$ 中包含至少 $P_{\min}$ 个对象，则将 $N_E(Q)$ 中未归入任何一个簇的对象加入 C；

（4）重复步骤 2，继续检查 C 中未处理对象，直到没有新的对象加入当前簇 C；

（5）重复步骤 1～3，直到所有对象都完成聚类分析；

（6）依据 2.4 节所述方法进行簇类中心计算，算法结束。

3　实例分析

本文内容的应用背景来源于工学与医学综合交叉研究，故将以医学中高血压患者个性群体分析问题为例，对本文方法进行说明验证。临床数据来源于某三级甲等医院心内科自 2014 年 1 月至 2016 年 11 月收治入院的 500 例高血压患者的临床指标数据。所有患者对本研究知情并同意参与，且临床信息均已匿名化处理。高血压患者的临床数据包括多项指标，如动态血压、血糖、血脂、胆固醇，属于典型的高维数据。本文选取与患者密切相关的 11 个临床指标作为原始数据，如表 1 所示。由于人作为一个典型生物复杂系统，各指标数据之间必然存在多种非线性关系以及冗余性，因此可以作为本文方法的实例验证数据。

表 1　高血压患者指标数据样例

总胆固醇 /mmol·L^{-1}	甘油三酯 /mmol·L^{-1}	M_{HDL} /mmol·L^{-1}	M_{LDL} /mmol·L^{-1}	M_{VLDL} /mmol·L^{-1}	年龄 /周岁	勺形特征值	收缩压昼夜下降率/%	舒张压昼夜下降率/%	性别特征值	是否吸烟
5.15	2.17	0.95	3.65	0.55	32	3	19.0	22.2	1	2
5.65	2.86	1.03	4.32	0.30	40	1	11.1	10.2	2	1
4.03	5.24	0.84	1.95	1.28	41	2	10.5	5.9	1	1

注：M_{HDL} 表示高密度脂蛋白在血清中的含量；M_{LDL} 表示低密度脂蛋白在血清中的含量；M_{VLDL} 表示极低密度脂蛋白在血清中的含量；勺形特征值中 1 代表反勺形规律，2 代表非勺形规律，3 代表勺形规律；性别特征值中 1 代表男性，2 代表女性；是否吸烟中，1 代表吸烟，2 代表不吸烟。

首先，应用 2.2 节所提方法对原始数据进行特征提取及降维分析，最终得到主元特征数据（部分数据）如下

$$\boldsymbol{T} = \begin{bmatrix} -5.480 & 0.499 & 7.635 & -6.012 & -7.341 & -17.936 \\ 9.689 & -11.213 & 5.013 & -13.787 & -19.941 & -26.750 \\ 13.860 & -1.468 & 4.987 & -15.853 & -17.931 & -21.733 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 5.849 & -12.673 & 16.864 & -38.964 & -26.409 & -39.226 \\ -7.432 & -14.298 & 19.435 & -35.523 & -35.733 & -49.826 \\ 4.679 & -15.003 & 14.231 & -29.472 & -42.362 & -43.396 \end{bmatrix} \tag{14}$$

式中：$\boldsymbol{T}$ 的每个行向量代表每个病例原始数据在主元方向的变化情况。应用式(11)确定主元个数为 6 时即可代表原始数据集 85%的变化情况。

以数据集 $\boldsymbol{T}$ 作为输入，结合 2.3 节所述相似性度量及 2.4 节聚类分析步骤，可将 500 例患者实现聚类分析；同时，将本文 KDBSCAN 算法与传统 DBSCAN 算法进行了对比，表 2 为两种算法的聚类结果。可以看出，KDBSCAN 算法与 DBSCAN 算法都将 500 例患者分为 7 类，即依据数据密度的变化进行分类，两种方法的结果一致，但 KDBSCAN 算法的聚类共含有患者数为 342 例，而 DBSCAN 算法结果只包含 153 例患者。这表明 500 例患者各个指标之间存在强的非线性关系，导致数据在高维空间呈现非均匀分布，而传统 DBSCAN 算法法无法有效捕获高维空间中各维数据间的非线性关系，从而将 347 例患者数据视为噪声信息而没有进行聚类，丢失了大量信息，其信息利用率仅有 30.6%。本文提出的 KDBSCAN 算法的信息利用率为 68.4%，即本文所提算法与传统 DBSCAN 算法具有相同聚类效果，但在保留信息完整性上远远好于传统 DBSCAN 算法。

表 2 两种算法聚类结果对比

聚类算法	聚类结果/例						
	类 1	类 2	类 3	类 4	类 5	类 6	类 7
KDBSCAN	146	92	38	28	16	13	9
DBSCAN	46	37	28	19	11	7	5

表 3 KDBSCAN 算法聚类结果的簇中心值

类别	总胆固醇 /mmol·L^{-1}	甘油三酯 /mmol·L^{-1}	M_{HDL} /mmol·L^{-1}	M_{LDL} /mmol·L^{-1}	M_{VLDL} /mmol·L^{-1}	年龄 /周岁	勺形特征值	收缩压昼夜下降率 /%	舒张压昼夜下降率 /%	性别特征值	是否吸烟
1	4.70	1.59	1.31	2.81	0.60	71.43	1.77	2.41	4.77	1.51	1.80
2	4.66	1.71	1.25	2.75	0.65	60.68	2.37	9.13	11.42	1.42	1.66
3	4.56	1.92	1.19	2.66	0.72	57.61	2.00	5.09	5.62	1.39	1.57
4	5.13	2.41	1.29	2.90	0.91	63.55	2.80	12.09	17.57	1.75	1.90
5	4.61	2.03	1.34	2.80	0.47	57.89	3.00	13.72	16.71	1.44	1.67
6	5.00	1.56	1.34	2.99	0.68	60.75	1.00	−6.48	−6.24	1.50	1.50
7	4.67	1.80	1.30	2.83	0.54	59.00	3.00	14.96	19.86	1.71	2.00

表 4 DBSCAN 算法聚类结果的簇中心值

类别	总胆固醇 /mmol·L^{-1}	甘油三酯 /mmol·L^{-1}	M_{HDL} /mmol·L^{-1}	M_{LDL} /mmol·L^{-1}	M_{VLDL} /mmol·L^{-1}	年龄 /周岁	勺形特征值	收缩压昼夜下降率 /%	舒张压昼夜下降率 /%	性别特征值	是否吸烟
1	4.43	1.71	1.23	2.49	0.70	53.88	2.03	5.90	8.07	1.45	1.57
2	4.37	1.33	1.28	2.56	0.53	74.55	1.97	4.57	6.17	1.60	1.81
3	4.49	1.41	1.26	2.74	0.46	62.18	2.18	8.32	13.01	1.41	1.65
4	4.54	1.48	1.29	2.85	0.40	65.19	1.57	0.03	1.10	1.44	1.69
5	4.84	1.19	1.44	2.95	0.44	66.67	2.00	7.02	5.45	1.50	1.83
6	4.51	1.73	1.39	2.53	0.58	78.33	2.00	2.73	−0.07	1.33	2.00
7	4.69	2.09	1.16	2.97	0.56	78.00	1.50	0.50	10.05	1.75	2.00

表 3 和表 4 分别为采用 KDBSCAN 与 DBSCAN 算法计算后的簇中心表征值。总体上看，KDBSCAN 算法得到的簇中心较 DBSCAN 算法得到的簇中心值更具有分离性。为定量评价这一结果，在此采用聚类通用评价方式对结果进行评价[13]，即聚类结果应该尽可能的使簇间的距离更大、簇内数据

间的距离更小。

表 5 为依据表 3 与表 4 结果计算得到的簇间距离的各项指标。可以看出，本文提出的 KDBSCAN 算法聚类结果虽然在簇间最小距离略小于 DBSCAN 算法的簇间最小距离，但前者簇间最大距离及平均距离都大于后者，说明本文算法的分类结果具有更好的类间分离性。

表 5　聚类结果簇间距离各项指标

聚类算法	最大距离	最小距离	平均距离
KDBSCAN	33.89	3.61	16.18
DBSCAN	25.96	6.68	14.12

表 5 为根据表 3 和表 4 所示的聚类结果直接计算所获得的簇间距离各项指标，不需要对各项指标进行标准化处理。但是，在计算簇内距离时，由于簇内不同对象指标数据量纲的不同对距离计算产生很大的影响，导致对象之间相似性的度量结果可靠性降低。因此，为了消除这种不良影响，首先应该对数据的各项指标进行标准化处理，即归一化处理。其主要过程为：将簇内原始数据减去均值除以标准差，再进行簇内对象之间距离的计算，求取每个对象与簇内其他对象的平均距离，最后求出簇内对象间的平均距离。表 6 所示即为将聚类结果标准化后求取簇内各对象间平均距离的对比结果。从表 6 可以看出，本文提出的 KDBSCAN 算法获得的簇内平均距离为 4.16，采用传统 DBSCAN 的簇内平均距离为 3.99；虽然本文方法得到的簇内聚类大于后者，但两种方法的簇内距离平均差值很小（仅为 0.17）；同时，由于传统 DBSCAN 的簇内对象只包含 153 个患者对象（患者对象总数为 500），丢失大量信息，在总体聚类效果上劣于本文提出的方法。

综上所述，本文所提出的 KDBSCAN 算法与传统 DBSCAN 算法相比，保留了原始数据中更多的数据信息，簇间的距离较大，簇内对象之间的距离大致相等，具有更好的聚类效果和准确度。

表 6　聚类结果簇内平均距离

聚类方法	平均距离							
	类 1	类 2	类 2	类 4	类 5	类 6	类 7	均值
KDBSCAN	4.42	4.39	4.31	4.19	4.02	3.96	3.85	4.16
DBSCAN	4.31	4.26	4.24	4.16	3.72	3.81	3.40	3.99
差值	0.11	0.13	0.07	0.03	0.30	0.14	0.45	0.17

4　结　论

本文对具有高维空间数据特征的对象聚类算法进行研究，具有一定应用普适性。针对传统 DBSCAN 算法对高维非线性特征数据聚类存在的问题，即高维数据具有非线性特征性导致的数据分布不均、传统相似性度量失效、聚类中心表征等问题，提出了一种 KDBSCAN 聚类算法。首先利用核主元分析（KPCA）理论对原始高维空间数据进行非线性特征的提取和降维，然后重新定义相似性函数实现对象间相似性度量，最后利用 3δ 统计理论对各群体中心进行准确表征，从而提高了对高维非线性特征数据分类的精确度和类中心知识表达的准确度。以医学中高血压患者个性群体分析问题为例对本文提出方法进行验证，结果表明，本文提出的 KDBSCAN 算法与传统 DBSCAN 算法相比，更能充分利用原始的数据信息，且具有更好的聚类效果。

由于原始数据的固有特征提取及对象相似性度量是高维数据的聚类分析两个重要环节，后期工作将进一步研究原始数据特征的提取技术，以及如何结合数据固有特征构建相似性度量及分类策略。

参考文献：

[1] 王骏，王士同，邓赵红. 聚类分析研究中的若干问题 [J]. 控制与决策，2012，27(3)：321-327.
WANG Jun，WANG Shitong，DENG Zhaohong. Survey on challenges in clustering analysis research [J]. Control and Decision，2012，27(3)：321-327.

[2] 刘红岩，陈剑，陈国青. 数据挖掘中的数据分类算法综述 [J]. 清华大学学报（自然科学版），2002，42(6)：727-730.
LIU Hongyan，CHEN Jian，CHEN Guoqing. Review of classification algorithms for data mining [J]. Journal of Tsinghua University (Science and Technology)，2002，42(6)：727-730.

[3] 蔡颖琨，谢昆青，马修军. 屏蔽了输入参数敏感性的 DBSCAN 改进算法 [J]. 北京大学学报（自然科学版），2014，40(3)：480-486.

（下转第 90 页）