# 大\_面向高维数据的PCA-Hubness聚类方法

郎江涛

（重庆大学计算机学院，重庆，400044）

# 摘要

机器学习（Machine Learning）是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。是通过机器自主学习的方式处理人工智能中的问题。近几十年机器学习在概率论、计算复杂性理论、统计学、逼近论等领域均有发展，已形成一门多领域交叉学科 。机器学习通过设计和分析让机器可以自主“学习”的算法以便从海量数据中自动分析 有价值的模式或规律，从而对未知数据进行预测。机器学习可以大致分为下面四种类别：监督学习（Supervised Learning）、无监督学习（Unsupervised Learning）、半监督学习（Semi-supervised Learning）以及增强学习（Reinforcement Learning）。监督学习是从已知的训练数据集中获得某种函数用于预测未知的数据集。监督学习训练集中的目标是由人为标注的。常见的监督学习算法包括回归分析（Regression Analysis）和统计分类（Statistical Classification ）；无监督学习与监督学习的不用之处在于训练集是没有人为标注的。常见的无监督学习算法有聚类分析（Cluster Analysis）；半监督学习介于监督学习与无监督学习之间；增强学习是基于环境而行动，从而获得最大化的预期利益。机器学习已广泛应用于诸多领域：数据挖掘、计算机视觉、搜索引擎、自然语言处理、语音和手写识别、生物特征识别、DNA序列测序、医学诊断、检测信用卡欺诈和证券市场分析等。

聚类分析（Cluster Analysis，亦称为群集分析）是把相似的对象通过静态分类的方法分成不同的簇或子集，使得在同一个簇中的对象都具有某些相似的属性。传统的聚类分析计算方法主要有如下五种：划分方法（Partitioning Methods）、层次方法(Hierarchical Methods)、基于密度的方法(Density-Based Methods)、基于网格的方法(Grid-Based Methods)和基于模型的方法(Model-Based Methods)。传统的聚类分析适用于低维数据的聚类问题。然而由于现实世界中数据的复杂性，在使用传统聚类算法处理诸多问题和任务时其表现效果不佳，尤其是对于高维数据和大型或海量数据而言。这是因为在高维数据空间中利用传统算法聚类时会碰到下述两个问题：（1）高维数据存在大量冗余、噪声的特征使得不可能在所有维中均存在簇；（2）高维空间中的数据分布十分稀疏，其数据间的距离几乎相等。显然，基于距离的传统聚类方法无法在高维空间中基于距离来构建簇。这便是机器学习中令人头疼的维数灾难（Curse of Dimensionality）问题[d1]。

近年来，“维数灾难”已成为机器学习的一个重要研究方向。与此同时，“维数灾难”通常是用来作为不要处理高维数据的无力借口。随着科技的发展使得数据获取变得愈加容易，导致数据库规模越来越大、复杂性越来越高，如各种类型的贸易交易数据、Web 文档、基因表达数据等，它们的维度（[属性](http://baike.baidu.com/view/77730.htm)）通常可以达到成百上千维，甚至更高。但是，受“维度效应”的影响，许多在低维数据空间表现良好的聚类方法运用在[高维空间](http://baike.baidu.com/view/738106.htm)上往往无法获得好的聚类效果。[高维](http://baike.baidu.com/view/2019275.htm)数据聚类分析是聚类分析中一个非常活跃的领域，同时它也是一个具有挑战性的工作。[高维](http://baike.baidu.com/view/2019275.htm)数据聚类分析在市场分析、信息安全、金融、娱乐、反恐等方面都有很广泛的应用。

1、划分方法(partitioning methods)

给定一个有N个元组或者纪录的数据集，分裂法将构造K个分组，每一个分组就代表一个聚类，K<N。而且这K个分组满足下列条件：（1） 每一个分组至少包含一个数据纪录；（2）每一个数据纪录属于且仅属于一个分组（注意：这个要求在某些模糊[聚类算法](http://baike.baidu.com/view/69222.htm)中可以放宽）；对于给定的K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都较前一次好，而所谓好的标准就是：同一分组中的记录越近越好，而不同分组中的纪录越远越好。使用这个基本思想的算法有：[K-MEANS算法](http://baike.baidu.com/view/31854.htm)、K-MEDOIDS算法、CLARANS算法；

大部分划分方法是基于距离的。给定要构建的分区数k，划分方法首先创建一个初始化划分。然后，它采用一种迭代的重定位技术，通过把对象从一个组移动到另一个组来进行划分。一个好的划分的一般准备是：同一个簇中的对象尽可能相互接近或相关，而不同的簇中的对象尽可能远离或不同。还有许多评判划分质量的其他准则。传统的划分方法可以扩展到子空间聚类，而不是搜索整个数据空间。当存在很多属性并且数据稀疏时，这是有用的。为了达到全局最优，基于划分的聚类可能需要穷举所有可能的划分，计算量极大。实际上，大多数应用都采用了流行的启发式方法，如k-均值和k-中心算法，渐近的提高聚类质量，逼近局部最优解。这些启发式聚类方法很适合发现中小规模的数据库中小规模的数据库中的球状簇。为了发现具有复杂形状的簇和对超大型数据集进行聚类，需要进一步扩展基于划分的方法。2(undefined)

2、层次方法(hierarchical methods)

这种方法对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。例如在“自底向上”方案中，初始时每一个数据纪录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。代表算法有：BIRCH算法、CURE算法、CHAMELEON算法等；

层次聚类方法可以是基于距离的或基于密度或连通性的。层次聚类方法的一些扩展也考虑了子空间聚类。层次方法的缺陷在于，一旦一个步骤（合并或分裂）完成，它就不能被撤销。这个严格规定是有用的，因为不用担心不同选择的组合数目，它将产生较小的计算开销。然而这种技术不能更正错误的决定。已经提出了一些提高层次聚类质量的方法。2(undefined)

3、基于密度的方法(density-based methods)

基于密度的方法与其它方法的一个根本区别是：它不是基于各种各样的距离的，而是基于密度的。这样就能克服基于距离的算法只能发现“类圆形”的聚类的缺点。这个方法的指导思想就是，只要一个区域中的点的密度大过某个[阀值](http://baike.baidu.com/view/648990.htm)，就把它加到与之相近的聚类中去。代表算法有：DBSCAN算法、OPTICS算法、DENCLUE算法等；

4、基于网格的方法(grid-based methods)

这种方法首先将数据空间划分成为有限个单元（cell）的[网格结构](http://baike.baidu.com/view/17502.htm),所有的处理都是以单个的单元为对象的。这么处理的一个突出的优点就是处理速度很快，通常这是与目标数据库中记录的个数无关的，它只与把数据空间分为多少个单元有关。代表算法有：STING算法、CLIQUE算法、WAVE-CLUSTER算法；

很多空间数据挖掘问题，使用网格通常都是一种有效的方法。因此，基于网格的方法可以和其他聚类方法集成。2(undefined)

5、基于模型的方法(model-based methods)

基于模型的方法给每一个聚类假定一个模型，然后去寻找能够很好的满足这个模型的数据集。这样一个模型可能是数据点在空间中的密度[分布函数](http://baike.baidu.com/view/843170.htm)或者其它。它的一个潜在的假定就是：目标数据集是由一系列的[概率分布](http://baike.baidu.com/view/45323.htm)所决定的。通常有两种尝试方向：统计的方案和[神经网络](http://baike.baidu.com/view/5348.htm)的方案。

当然聚类方法还有：[传递闭包](http://baike.baidu.com/view/6178000.htm)法，[布尔矩阵](http://baike.baidu.com/view/3213062.htm)法，直接聚类法，[相关性分析](http://baike.baidu.com/view/5996049.htm)聚类，基于统计的聚类方法等。

维度灾难（Curse of Dimensionality）在各个领域为各种机器学习的方法和任务带来了诸多挑战。在高维空间中，由于数据的稀疏性以及数据点之间的易区分的距离的难度增加，从而导致聚类变得异常困难。因此本文，我们对高维数据的聚类问题采取新的观点。利用高维数据的内在特性，更具体地说，将hubness——— 高维数据倾向于或者易于包含某些频繁出现在其它点的k近邻列表中的点（Hubs）——— 结合PCA降维算法应用到聚类中。在高维数据聚类过程中，hubness可视为一种有效的方法用于检测点的中心性（PointCentrality）。而PCA主要用于减少特征数，减少噪音和冗余，减少过度拟合的可能性。本文通过hubness这一高维数据的内在特性，对PCA的降维程度进行了较为适宜的调控，从而提高了聚类效果。

**关键字：**

聚类；高维数据；Hubness；PCA

# 0 引言

通常在无监督学习过程中，聚类是将元素分成不同的组别或者更多的子集，使得分配到相同簇中的元素彼此之间比其它的数据点更为相似，也就是说，要增加类内的相似性并减小类间的相似性，然而该目标的实现过程中却有诸多障碍。多年来，已提出多种聚类算法，可以大致分为以下四类：partitional，hierarchical，density- based 和 subspace 算法。其中 subspace 算法是在原始数据的低维投影中进行聚类，当原始数据为高维时此方法更为适宜。之所以选择subspace算法是因为在实验过程中观测到：随着数据维度的增大，维度灾难出现的概率也随之增大，而这一性质使得诸多标准的机器学习算法表现不佳。此问题主要由以下两个因素引起的：空的空间现象（EmptySpace Phenomenon）和距离集中（Concentration Of Distances）。前者指的是当维数提高时，空间的体积提高太快，因而可用数据变得很稀疏[1]。后者是说高维数据空间表示出现了某种程度上的反直觉的特性——— 随着维度增加，数据点之间的所有距离趋向于变得更加难以区分，这可能会导致基于距离的算法性能变差。

“维数灾难”通常是用来作为不要处理高维数据的无力借口。由于本征维度（IntrinsicDimensionality）的存在，其概念是指任意低维数据空间可简单地通过增加空余（如复制）或随机维将其转换至更高维空间中，相反地，许多高维空间中的数据集也可削减至低维空间数据，而不必丢失重要信息。这一点也通过众多降维方法的有效性反映出来，如应用广泛的主成分分析（PCA）方法。针对距离函数和最近邻搜索，当前的研究也表明除非其中存在太多不相关的维度，带有维数灾难特色的数据集依然可以处理，因为相关维度实际上可使得许多问题（如聚类分析）变得更加容易。在本文中，我们阐述了hubness———高维数据倾向于或者易于包含某些频繁出现在其它点的k近邻列表中的点（hubs），并将其应用到聚类中。探讨了k-occurrences的偏度与本征维度的相互关系。

# 1 相关工作

近年来在涉及声音和图像数据的若干应用领域中观察到hubness 现象（Aucouturierand Pachet, 2007; Doddington et al., 1998; Hicklin et al., 2005）， 此外，Jebara等人简要地描述了在半监督学习的邻域图构造过程中出现的hubness现象（TonyJebara et al 2009）[2]。Amina M 等人通过将 hub 引入到 k-means 算法中从而形成了基于hubness 的算法（AminaM et al 2015）[3]。尽管在数据聚类中 hubness 这一现象并没有给予过多关注，然而 k-nearest-neighbor 列表却广泛使用在诸多聚类中。k-nearest-neighbor列表通过用于观察由k 个最近邻所确定的空间的体积来计算密度估计。基于密度的聚类方法通常依赖于这种密度估计。基于密度的聚类算法主要的目标是寻找被低密度区域分离的高密度区域[4]。在高维空间中，这常常难以估计，因为数据非常稀疏。此外，选择适当的邻域大小也尤为重要，因为过小和过大的k 值都可能导致基于密度的方法失败。k-nearest-neighbor列表常被用于构造k-NN 图并以此用于图聚类。

## 1.1 Hubness 现象

令 表示一组数据点，其中 为数据集 的元素。令 表示在 空间中的一个距离函数 ，其中 如下定义：

$p\_{i,k}=\begin{cases} 1, & \text{if $x$is among k nearest neighbours of $x\_i$, according to $dist$} \\0  & \text{otherwise} \end{cases}$

在此基础之上，定义， 表示为在 空间中， 出现在其它 k-nearest neighbor 列表中的次数，也记为K-occurrence，仅根据数据点的 K-occurrence 的大小无法确定 hubness 对实验结果有何种影响。数据点的bad k-occurrences 表示为 ，是指数据点 x作为数据集D 中其它的点的k-nearest neighbor次数，并且 x 点的标签和那些点的标签不匹配。数据点的 good k-occurrences表示为， 是指点 x 的标签与那些点的标签相匹配[5]。为了表征 的非对称性，我们使用 k-occurrences 分布的标准第三矩，

其中 和 分别是 的均值和标准差。

一方面，虽然高维数据已表现出hubness这一现象，然而它的偏度（skewness） 和分布却因数据的不同而差异颇大。因此，hubs甚至有时可以在被检索时被视为噪声。另一方面，hubness与距离集中现象密切相关——— 一种与高维数据的反直觉特性（随着维度增加数据点之间的所有距离趋向于变得更加难以区分）[3]。

## 1.2 基于hub的聚类

更接近簇均值的点易倾向于具有比其它点更高的hubness 分数[5]。将hubness 视为一种局部中心度量，则可以以各种方式使用hubness 进行聚类。基于hub的聚类算法主要有以下4种：deterministic, probabilistic, hybrid 和 kernel。这4种方法均为 k-means 算法的扩展。在deterministic 方法中，首先确定簇的数量然后使用 k-means 算法进行聚类，在每次聚类的过程中将当前簇中的具有高的hubness 分数的点作为其中心。Probabilistic方法使用模拟退火算法以一定概率 选择高 hubness 分数的点作为当前簇的中心。Deterministic 和probabilistic 方法只依赖于距离矩阵而不必关心数据的表现形式。为了尽可能地获取数据的中心位置则需要使用 hybrid 方法。在hybrid 方法中，使用数据点的hubness 分数来指导搜索，但最终会形成基于质心的簇结构。kernel方法在前三者基础上可以对非超球面簇集进行处理。基于hub 的聚类算法用于高维数据，由此可见随着维度的增加聚类时间和迭代次数也随之增加。

# 2 PCA-Hubness

主成分分析（Principalcomponents analysis，PCA）经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面[6]。主成分分析主要是通过对协方差矩阵进行特征分解，以得出数据的主成分（即特征向量）与它们的权值（即特征值）。这可以理解为对原数据中的方差做出解释：哪一个方向上的数据值对方差的影响最大？换而言之，PCA提供了一种降低数据维度的有效办法；如果分析者在原数据中除掉最小的特征值所对应的成分，那么所得的低维度数据必定是最优化的（也即，这样降低维度必定是失去讯息最少的方法）。

通过使用降维来保存大部分数据信息的主成分分析的观点是不正确的。确实如此，当没有任何假设信息的信号模型时，主成分分析在降维的同时并不能保证信息的不丢失，其中信息是由香农熵来衡量的。因此，下文中将会探讨在使用降维技术PCA 的情况下 的偏度和本征维度的相互作用。此研究的主要目的在于探讨降维是否能够缓解 的偏度这一问题。“The observed skewness of ,besides being strongly correlated with ,is even more strongly correlated with the intrinsic dimensionality dmle.Moreover, intrinsic dimensionality positively affects the correlations betweenNk and the distance to the data-set mean / closest cluster mean, implying thatin higher (intrinsic) dimensions the positions of hubs become increasinglylocalized to the proximity of centers.”[5]。因为观察到的 的偏度与与本征维数强烈相关此外，本征维数对 到数据集的均值或者与最接近簇的均值的距离有着积极影响，这意味着在较高（本征）维度中，hubs变得越来越接近数据集或最接近簇的中心。

实验过程中采用的距离度量方法是闵可夫斯基距离（Minkowskidistance），它是衡量数值点之间距离的一种非常常见的方法，假设数值点P 和Q 坐标如下：

，

那么，闵可夫斯基距离定义为：

该距离最常用的 p 是 2 和 1, 前者是欧几里得距离（Euclidean distance），后者是曼哈顿距离（Manhattandistance）。可夫斯基距离比较直观，但是它与数据的分布无关，具有一定的局限性，如果x 方向的幅值远远大于y 方向的值，这个距离公式就会过度放大x 维度的作用。所以，在计算距离之前，我们可能还需要对数据进行z-transform 处理，即减去均值，除以标准差：

其中 是该维度上的均值， 是该维度上的标准差。可以看到，上述处理开始体现数据的统计特性了。这种方法在假设数据各个维度不相关的情况下利用数据分布的特性计算出不同的距离。

为了探究在使用降维技术的情况下 的偏度和本征维度的相互作用，我们使用了来自 UCI 多维度的9个数据库进行观测 的分布。在表1中包含了以下信息：数据集的大小（，第2列）；维数（，第3列）；类别数（，第4列）；距离的度量方法（，Euclidean或者Cityblock，第6列）。表1对应的第5列列出了真实数据集的经验值，表明对于大多数数据集的 的分布向右倾斜。虽然 的值是固定的，但是使用其它的 值也可得到类似的结果。

表1 真实数据集。数据来源于 University of California, Irvine (UCI) MachineLearning Repository

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| data set | size | d | cls |  | dist |
| wpbc | 198 | 33 | 2 | 0.86 |  |
| Ionosphere | 351 | 34 | 2 | 1.72 |  |
| musk | 476 | 166 | 2 | 1.33 |  |
| parkinsons | 195 | 22 | 2 | 0.73 |  |
| sonar | 208 | 60 | 2 | 1.35 |  |
| spectrometer | 531 | 100 | 10 | 0.59 |  |
| mfeat-fou | 2000 | 76 | 10 | 1.28 |  |
| mfeat\_factors | 2000 | 216 | 10 | 0.83 |  |
| Arraythmia | 452 | 279 | 16 | 1.98 |  |

图 1 维度下降与偏度的关系

图 1 描述了针对若干个真实数据集（musk, sonar, mfeat-fou等）通过降维方法获得的维数占原有数据集维数的百分比与 之间的相互关系。数据之间距离的度量方法为Minkowski 距离，其中 的取值分别为：2（Euclidean distance）。 从左往右观察，对于大部分数据集而言利用PCA降维算法， 保持相对恒定直到降维后留下的特征的百分比较小时才会陡然下降。因此，当达到数据集的本征维数时若继续减小维数则会导致有价值的信息丢失。针对PCA 方法对数据进行降维时，若降维后的维数在本征维数之上那么降维并不会对hubness 这一现象有显著影响。

算法思想如下：

• 数据预处理；

• 构建 KNN 邻域矩阵，并计算出点的逆近邻；

• 利用 PCA 对数据进行处理，降维后维数的下限是其偏度不小于原始偏度的80%（这是一个比较不错的经验值，可自行调试）；

• 利用点的 找出 hubs， 并将其应用到相关的聚类算法。

# 3 实验结果

在实验之前数，据集中的所有单个特征均作了归一化处理。数据集是一些不甚复杂的、含若干个簇的数据，实验的结果如表2 所示。轮廓系数（SilhouetteIndex）为聚类结果的评测指标[7]，其计算公式如下所示：

其中， 表示 向量到同一簇内其他点不相似程度的平均值， 表示 向量到其他簇的平均不相似程度的最小值。可见轮廓系数的值总是介于 [-1,1]，越趋近于1代表内聚度和分离度都相对较优。将所有点的轮廓系数求平均，就是该聚类结果总的轮廓系数。

表2 轮廓系数（Silhouette Index）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| data set | size | d | cls |  | dist | KM++[8] | GHPKM[8] | Ker-KM[8] | PH-KM | PH-KM |
| Ionosphere | 351 | 34 | 2 | 1.72 |  | 0.28 | 0.27 | 0.34 | **0.39** | **0.41** |
| mfeat\_factors | 2000 | 216 | 10 | 0.83 |  | 0.17 | 0.18 | 0.15 | **0.24** |  |
| mfeat-fou | 2000 | 76 | 10 | 1.28 |  | 0.07 | 0.07 | -0.03 | **0.23** |  |
| musk | 476 | 166 | 2 | 1.33 |  | **0.28** | 0.27 | **0.28** | 0.22 | **0.31** |
| parkinsons | 195 | 22 | 2 | 0.73 |  | 0.37 | 0.37 | 0.45 | **0.88** | **0.88** |
| sonar | 208 | 60 | 2 | 1.35 |  | **0.19** | 0.15 | 0.13 | **0.19** | **0.22** |
| spectrometer | 531 | 100 | 10 | 0.59 |  | 0.23 | **0.25** | 0.15 | 0.15 |  |
| wpbc | 198 | 33 | 2 | 0.86 |  | 0.16 | 0.16 | 0.17 | **0.36** | **0.32** |

在缺乏hubness 特性的情况下，基于hubness 的方法表现不佳，其性能接近于KM++。与此同时观察到一些有趣的现象：一些数据集虽然有较高的hubness，但是在利用PCA 降维的过程中其本征维数损失较大从而导致聚类结果表现不佳；还有一些数据集虽然hubness 本身并不是很高，但在降维的过程中其本征维数几乎保持恒定顾聚类结果较佳。

# 4 结论

此前并没有使用hubness 结合PCA 对数据进行聚类的相关工作。本文通过探讨在使用降维技术PCA 的情况下 的偏度和本征维数的相互作用从而说明：对于大部分数据集而言利用PCA 降维算法， 保持相对恒定直到降维后留下的特征的百分比较小时才会陡然下降。基于 hub 的算法是专为高维数据设计的。这是一个不寻常的属性，因为大多数标准聚类算法的性能随着维数的增加而降低。另一方面，hubness是高维数据的本征属性，通过利用降维技术PCA 而又基本保持本征维数不变的情况下对数据进行实验从而提高聚类结果。10

# 参考文献：

[d1] Richard Ernest Bellman. Dynamic Programmin. Courier Dover Publications. 2003.

[1] Houle, M. E.，Kriegel, H. P.，Kröger, P.，Schubert,E.，Zimek.A. Scientific and Statistical Database Management[J]，Lecture Notes in ComputerScience **6187**: 482. 2010.

[2] Tony Jebara，Jun Wang，Shih-Fu Chang. Graphconstruction and b-matching for semi-supervised learning[J]. In Proceedings ofthe 26th International Conference on Machine Learning(ICML)， pages441–448，2009.

[3] Amina M，Syed Farook K. A Novel Approach forClustering High-Dimensional Data using Kernel Hubness[J]. InternationalConfenrence on Advances in Computing and Communication. 2015.

[4] Ester Martin，Kriegel Hans-Peter，Sander,Jörg，Xu, Xiaowei，Simoudis Evangelos，Han,Jiawei，FayyadUsama M., eds. A density-based algorithm for discovering clusters in largespatial databases with noise[J]. Proceedings of the Second InternationalConference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp.226–231.

[5] Milosˇ Radovanovic ́，Alexandros Nanopoulos，MirjanaIvanovic ́. Hubs in Space: Popular Nearest Neighbors in High-DimensionalData[J]，Journal of Machine Learning Research 11 (2010) 2487-2531. 2010

[6] Abdi. H，Williams L.J. Principal componentanalysis[J]. Wiley Interdisciplinary Reviews: Computational Statistics. 2 (4):433–459. 2010

[7] Peter J. Rousseeuw. Silhouettes: a Graphical Aid tothe Interpretation and Validation of Cluster Analysis[J]. Computational andApplied Mathematics. **20**: 53–65.1987.

[8] Nenad Toma sev，Milo s Radovanovi c，DunjaMladeni c，andMirjana Ivanovi c. The Role of Hubness in Clustering High-Dimensional Data[J]，IEEETRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 3，2014

# Clustering High-Dimensional Data using PCA-Hubness

Lang Jiangtao

(School of Computer Science, Chongqing University,Chongqing, 400044)

**Abstract:**

Curse of dimensionality presents a number of challenges invarious fields for machine learning methods and tasks. In high-dimensionalspace, the difficulty of data sparseness and distinguishing between data pointsincreases, which makes it difficult to cluster. Therefore, we take a new viewon the clustering in high-dimensional data. Using the intrinsic characteristicsof high-dimensional data, and more specifically, hubness, which points tend toappear frequently in list of k-nearest neighbors of other points inhigh-dimensional data. Hubness can be regarded as an effective method fordetecting point centrality in high-dimensional data . The PCA is mainly toreduce the number of features, reduce noise and redundancy, reduce thepossibility of over-fitting. In this paper, using the intrinsic characteristicsof hubness with the proper dimensionality reduction improve performance ofclustering.

**Keywords:**

Clustering; High-dimensional data; Hubness; PCA