面向高维数据的PCA-Hubness聚类方法

葛亮，郎江涛，唐黄，唐允恒

（重庆大学计算机学院，重庆，400044）

**摘要：**hub聚类算法可以解决传统聚类算法无法处理高维数据的问题。然而，由于它未考虑数据中的冗余和噪声特征，从而降低了聚类性能。因此，提出了PCA-Hubness 聚类方法用于提高高维数据的聚类性能。PCA-Hubness聚类方法利用逆近邻数的偏度和本征维度的相互关系，以偏度的变化率为降维依据，保证了在对高维数据降维时不会损失过多的有价值信息，有利于提高聚类效果。此算法在UCI数据集上进行实验，相比hub聚类算法，轮廓系数平均提高了15% 。

**关键字：**Hub聚类；高维数据； 偏度；本征维度；PCA ；

Clustering High-Dimensional Data using PCA-Hubness

Ge Liang，Lang Jiangtao，Tang Huang，Tang Yunheng

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

**Abstract:** The hub-based clustering algorithm can solve high dimensional data problem that traditional clustering algorithm cannot handle. However, since it does not handle redundancy and noise features in high-dimensional data, the clustering performance is reduced. Therefore, PCA-Hubness clustering method is proposed to solve the clustering problem of high-dimensional data. The PCA-Hubness clustering method utilizes the relationship between skewness of anti-nearest-neighborhood’s number and intrinsic dimension. According to the rate of change of the skewness, it is guaranteed that the high dimensional data will not lose too much Information. And it is conducive to improving the clustering effect. This algorithm performs experiments on the UCI data set, and the Silhouette Index are increased by an average of 15% compared to hub-based clustering algorithm.

**Keywords:** Skewness;Intrinsic Dimension; PCA; Hub Clustering; High-dimensional data;

# 1 问题的提出

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，聚类算法的目的是要增加类内的相似性并减小类间的相似性。多年来，已提出多种聚类算法，可以大致分为以下五类：划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法[1]。以上这五类传统聚类算法并不适用于高维数据聚类。虽然hub聚类算法可以对高维数据聚类，然而当存在冗余和噪声数据时，聚类效果表现不佳。传统聚类算法不适用于高维数据聚类主要是由以下两个因素引起的：数据的稀疏性和距离的集中。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[2]。后者是说高维数据空间表示出现了某种程度上的反直觉特性。随着维度增加，数据间的距离趋于相同，这可能会导致基于距离的算法性能变差。这便是机器学习中令人头疼的维数灾难问题。然而，由于本征维数的存在，许多高维空间中的数据集可削减至低维空间数据，而不必丢失重要信息。在高维数据中，某些点易频繁地出现在其它点的*k*近邻列表中，这种现象称为 hubness 现象，那些受“欢迎”的点称之为 hubs[6]。高维数据中存在着的冗余和噪声特征维度对聚类造成了严重的影响，然而无目标的降维又会损失重要的有价值信息。本文利用逆近邻数的偏度和本征维度的相互关系，以偏度的变化率为降维依据，保证了在对高维数据降维时不会损失过多的有价值信息，有利于提高聚类效果，实验结果表明此方法是可行的。

# 2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（Aucouturier and 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（Tony Jebara et al 2009）[3]，Amina M 等人通过将 hub 引入到 k-means 算法中从而形成了hub 聚类分析算法（Amina M et al 2015）[4]。尽管在数据聚类中 hubness 这一现象并没有给予过多的关注，然而 *k*近邻列表却广泛使用在诸多聚类中。*k*近邻列表通过观察 *k* 个最近邻所确定的空间体积来计算密度估计。基于密度的聚类算法的主要目标是寻找被低密度区域分离的高密度区域[5]。在高维空间中，这常常难以估计，因为数据非常稀疏。Hub 聚类算法可以处理高维数据，然而并未对高维数据中的冗余和噪声数据给予关注，从而导致聚类性能不佳。

## 2.1 Hubness 现象

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

在此基础之上，定义 ， 表示为在 空间中， 出现在其它 *k* 近邻列表中的次数，也记为 *k-occurrence*。 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*近邻次数，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[6]。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。偏度是用于衡量实数域中随机变量分布的不对称性。偏度的值有正负之分，偏度为负则表明绝大多数的值（包括中值在内）位于平均值的右侧；偏度为正则表明绝大多数的值（不一定包括中值）位于平均值的左侧；偏度为零则表明数值近似地均匀分布在均值的两侧，却不一定为对称分布。值越大表明数据集的 hubness 现象就越严重，hubs 出现的概率就越高。

## 2.2 Hub聚类算法

具有高 hubness分数的点更易接近簇中心 [6]。将 hubness 视为一种局部中心度量方式，则可以将它应用到聚类中。Hub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具有高 hubness 分数的点作为簇中心，例如，K-hub 聚类算法[9]。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，例如，HPC聚类算法[9]和GHPC聚类算法[9]。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，例如，HPKM聚类算法[9]和GHPKM聚类算法[9]。Kernel 方法在前三者基础上可以对非超球面簇集进行处理，例如，Ker-KM聚类算法[4]和Ker-GHPKM聚类算法[4]。Hub 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。虽然 hub 聚类算法可以处理高维数据，然而高维数据中存在的冗余和噪声特征却并未得到解决，这不利于聚类分析。

# 3 PCA-Hubness聚类算法

## 3.1 算法框架

PCA-Hubness聚类算法的整体流程图如下所示：

图1 PCA-Hubness算法流程图

首先，对数据集进行预处理，将数据的每一维进行归一化；其次，构建KNN 邻域矩阵，计算每个点的逆近邻数。然后，用 PCA 进行降维，在降维的过程中通过偏度的变化率来控制降维的程度，以防损失过多重要的有价值信息。最后，在获取降维数据后利用hub 聚类算法进行聚类分析。

## 3.2 基于偏度的降维方法

关于数据降维的方法有多种，本文采用的是主成分分析法。主成分分析（Principal Components 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[7]。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）。

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。然而，香农熵却无法作为数据有效降维时的衡量标准，因此本文采用了 的偏度这一指标。下文中将会探讨在使用降维技术 PCA 的情况下 的偏度和本征维数的相互作用。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。“因为观察到的 的偏度与与本征维数强烈相关，本征维数对 到数据集的均值或到最接近簇的均值有着积极影响，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”[6]。

实验过程中采用的距离度量方法是闵可夫斯基距离（Minkowski distance），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* 值是 2 和 1, 前者是欧几里得距离（Euclidean distance），后者是曼哈顿距离（Manhattan distance）。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 的偏度和本征维数的相互作用，本文使用了来自加州大学尔湾分校(UCI) 机器学习库[10]的数据集进行观测 的分布。在表1中包含了以下信息：数据集的样本数（，第2列）；数据样本的特征维数（，第3列）；数据样本的类别数（，第4列）。

表1 真实数据集

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | n | d | cls | 数据集 | n | d | cls |
| arrhythmia | 452 | 279 | 10 | mfeat-fou | 2000 | 76 | 10 |
| Ionosphere | 351 | 34 | 2 | spectrometer | 208 | 60 | 10 |
| musk | 476 | 166 | 2 | sonar | 208 | 60 | 2 |
| mfeat\_factors | 2000 | 216 | 10 |  |  |  |  |

图 2 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，若降维后本征维数未发生明显变化，那么降维并不会对 hubness 这一现象有显著影响。

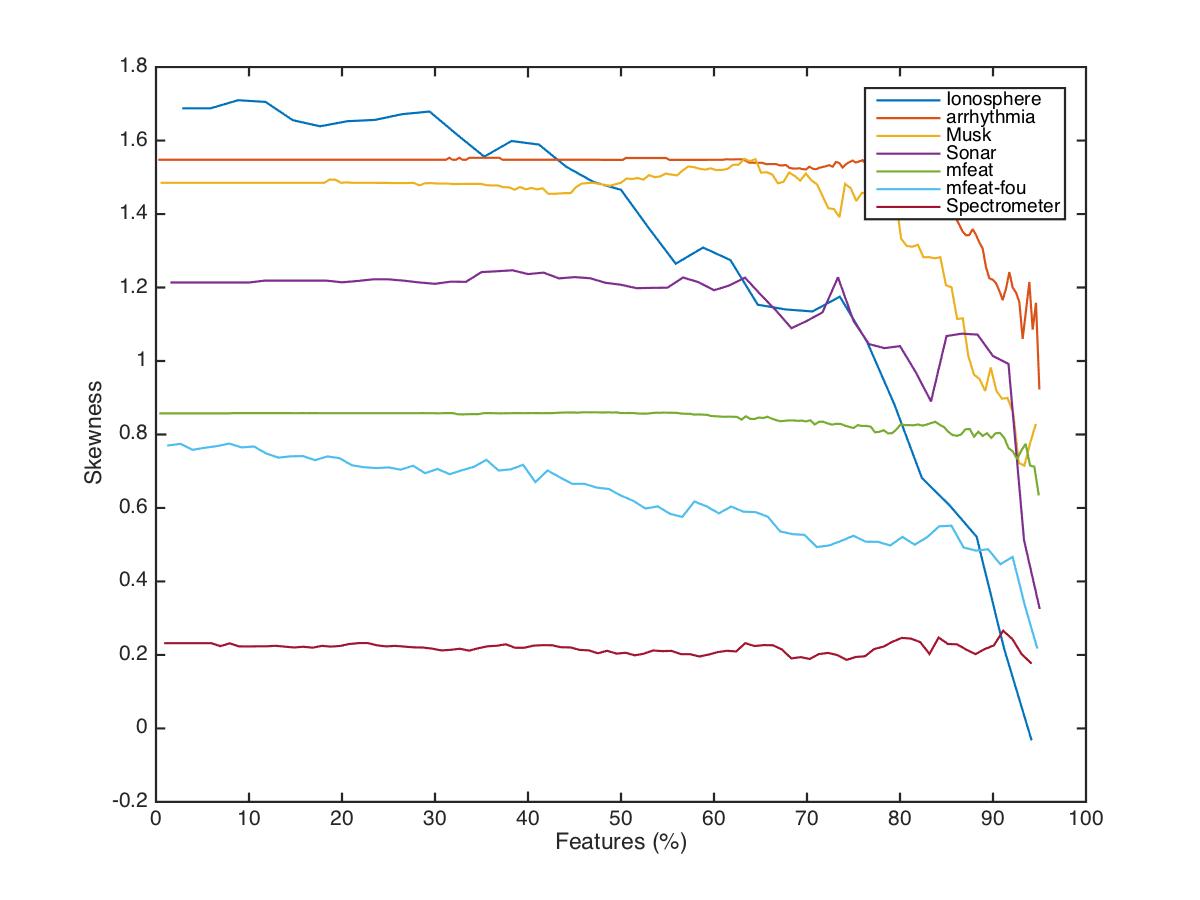


图 2 特征维度与偏度的关系

# 4 实验结果

实验数据来源于加州大学尔湾分校(UCI) 机器学习库。表2中第5列为真实数据集的偏度值 ，其中10代表*k*近邻数。从表中数据可以看出，对于大多数数据集的 的分布发生了倾斜。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。采用轮廓系数（Silhouette Index）作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。本文方法与KMEANS[9]、GHPKM[9]、Ker-KM[4]和 Ker-KM[4] 方法进行了比较，其中PH-KM为本文的聚类方法。实验结果如表 2 所示，下表中加粗的数据表示当前数据集的最优值。

表2 轮廓系数

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| data set | size | d | cls |  | dist | KMEANS  [9] | GHPKM  [9] | Ker-KM  [4] | Ker-GHPKM  [4] | PH-KM |
| Ionosphere | 351 | 34 | 2 | 1.72 |  | 0.28 | 0.28 | 0.28 | 0.25 | **0.41** |
| mfeat\_factors | 2000 | 216 | 10 | 0.83 |  | 0.18 | 0.20 | 0.17 | 0.18 | **0.24** |
| musk | 476 | 166 | 2 | 1.33 |  | 0.28 | 0.28 | 0.29 | 0.29 | **0.31** |
| parkinsons | 195 | 22 | 2 | 0.73 |  | 0.42 | 0.44 | 0.64 | 0.21 | **0.88** |
| sonar | 208 | 60 | 2 | 1.35 |  | 0.20 | 0.21 | **0.26** | 0.17 | 0.22 |
| wpbc | 198 | 33 | 2 | 0.86 |  | 0.16 | 0.16 | **0.32** | 0.22 | 0.31 |
| AVG-UCI | | | | | | 0.25 | 0.26 | 0.33 | 0.22 | **0.39** |

对于每一个数据集而言，取KMEANS、GHPKM、Ker-KM以及Ker-GHPKM聚类算法中轮廓系数的最大值作为经典聚类算法的最优值，然后同本文的PH-KM聚类算法进行比较。实验结果表明，相比之前的聚类算法，本文提出的PH-KM聚类算法在轮廓系数上平均提高了15%。从实验结果可以看出， 在数据集缺乏 hubness 特性的情况下，GHPKM、Ker-GHPKM等hub 聚类算法表现不佳，其性能接近于 KMEANS算法；然而当数据集呈现出较高的hubness特性时，GHPKM、Ker-GHPKM等hub 聚类算法的表现要优于KMEANS算法。同时，本文提出的PH-KM聚类算法无论数据集是否呈现出较高的hubness特性，均可以取得不错的聚类效果，相比之前的聚类算法适用范围更广，聚类性能更佳。

# 5 结论

在高维数据空间中，传统的聚类算法已变得不再适用。虽然hub聚类算法可以处理上述问题，但是它却忽略了高维数据中的冗余和噪声数据，从而导致聚类效果不佳。本文以 的偏度与本征维数强烈正相关为理论基础，通过构建数据集的KNN邻域矩阵，以偏度的变化率作为降维依据，最后再对降维后的数据集进行聚类分析。实验结果表明，无论数据集是否含有较高的hubness特性，本文提出的PH-KM聚类算法均可以取得不错的聚类效果，相比之前的聚类算法，轮廓系数平均提高了15%。

参考文献：

[1] Jiawei Han．数据挖掘概念与技术[C] 机械工业出版社. 2012.

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

[3] Tony Jebara，Jun Wang，Shih-Fu Chang. Graph construction and b-matching for semi-supervised learning[J]. In Proceedings of the 26th International Conference on Machine Learning(ICML)， pages 441–448. 2009.

[4] Amina M，Syed Farook K. A Novel Approach for Clustering High-Dimensional Data using Kernel Hubness[J]. International Confenrence on Advances in Computing and Communication. 2015.

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

[6] Milosˇ Radovanovic ́，Alexandros Nanopoulos，Mirjana Ivanovic ́. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data[J]， Journal of Machine Learning Research 11 (2010) 2487-2531. 2010

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

[8] Peter J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis[J]. Computational and Applied Mathematics. 20: 53–65. 1987.

[9] Nenad Toma sev，Milo s Radovanovi c，Dunja Mladeni c，and Mirjana Ivanovi c. The Role of Hubness in Clustering High-Dimensional Data[J]，IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 3. 2014

[10] Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013

**作者简介**：郎江涛（1990-），男，山西人，硕士生，本科学士学位，研究方向计算机应用技术

**联系地址**：重庆市重庆大学计算机学院

**邮编**：400044

**电子邮箱：shuiyibu.ja@gmail.com**

**联系电话：**18883701550

**QQ：**296017297