大数据中一种基于语义特征阈值的层次聚类方法

罗恩韬*①② 王国军①
①(中南大学信息科学与工程学院 长沙 410083)
②(湖南科技学院电子与信息学院 永州 425006)

摘 要：云计算、健康医疗、街景地图服务、推荐系统等新兴服务促使数据的种类和规模以前所未有的速度增长，数据量的激增会导致很多共性问题。例如数据的可表示，可处理和可靠性问题。如何有效处理和分析数据之间的关系，提高数据的划分效率，建立数据的聚类分析模型，已经成为学术界和企业界共同亟待解决的问题。该文提出一种基于语义特征的层次聚类方法，首先根据数据的语义特征进行训练，然后在每个子集上利用训练结果进行层次聚类，最终产生整体数据的密度中心点，提高了数据聚类效率和准确性。此方法采样复杂度低，数据分析准确，易于实现，具有良好的判定性。

关键词：大数据；数据抽取；层次聚类；聚类分析
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2015)12-2795-07
DOI: 10.11999/JEIT150422

A Hierarchical Clustering Method Based on the Threshold of Semantic Feature in Big Data

Luo En-tao①② Wang Guo-jun①
①(*School of Information Science and Engineering, Central South University, Changsha* 410083, *China*)
②(*School of Electronics and Information Engineering, Hunan University of Science and Engineering, Yongzhou* 425006, *China*)

Abstract: The type and scale of data has been promoted with a hitherto unknown speed by the emerging services including cloud computing, health care, street view services recommendation system and so on. However, the surge in the volume of data may lead to many common problems, such as the representability, reliability and handlability of data. Therefore, how to effectively handle the relationship between the data and the analysis to improve the efficiency of classification of the data and establish the data clustering analysis model has become an academic and business problem, which needs to be solved urgently. A hierarchical clustering method based on semantic feature is proposed. Firstly, the data should be trained according to the semantic features of data, and then is used the training result to process hierarchical clustering in each subset; finally, the density center point is produced. This method can improve the efficiency and accuracy of data clustering. This algorithm is of low complexity about sampling, high accuracy of data analysis and good judgment. Furthermore, the algorithm is easy to realize.

Key words: Big data; Data extraction; Hierarchical clustering; Clustering analysis

1 引言

云计算、健康医疗、推荐系统、社交网络等新兴服务促使人类社会的数据种类和规模正以前所未有的速度增长。而在以上服务中，为了让商品更具有吸引力，让购买客户更快地找到自己所需要的商品、让社交网络个人信息更加个性化等，产生了大规模的、具有特定特征的、离散的词汇来描述数据原型的基本形状特征和结构细节。而这些庞大的数据信息在大数据环境中由于受本身特性影响，往往造成信息过载，这使得数据的后续存储，分类非常不便。因此，如何有效地对这些具有某种特定特征的数据进行分析，挖掘数据之间的关系，建立可靠的数据聚类模型，从而为新型数据计算提供更个性化的决策支持和信息服务就显得尤为必要。

层次聚类最终会产生一棵聚类树，这样就更方便了解聚类过程，因此有益于大数据使用者作出决策。目前，国内外学者对大数据的聚类进行了大量研究，现有的方法较多的是将数据抽样方法、矩阵

收稿日期：2015-04-10；改回日期：2015-09-01；网络出版：2015-11-01
*通信作者：罗恩韬 cs_entaoluo@csu.edu.cn
基金项目：国家自然科学基金(60173037, 6272496, 61272151)，湖南省教育厅科研项目(2015C0589)和湖南科技学院重点学科项目
Foundation Items: The National Natural Science Foundation of China (60173037, 6272496, 61272151); The Hunan Provincial Education Departmant of China (2015C0589); Key Discipline Project of Hunan University of Science and Engineering

分解同经典聚类算法等进行结合来处理大数据的数据样本。例如：文献[1]提出了计算用户权重，从而抽取用户特征的思想。文献[2,3]提出了指数机制，该机制主要是设计一些打分函数来对数据处理，打分越高，数据的特性就越接近。文献[4–7]利用了数据抽样方式对数据进行聚类分析。文献[8,9]利用分布式计算来进行谱聚类。文献[10]利用了矩阵分解的思想来求得特征值。文献[11]利用了TOP-K的方法对大数据进行查询。但是以上方法因为对样本抽取的随意性，没有综合考虑数据本来的特征值，因此很有可能丢失数据的关键特性，而利用矩阵分解思想求得的特征值对聚类分析的帮助较为有限。因此，将上述思想或者算法运用到大数据的聚类分析上并不能取得良好的效果。

基于以上分析，本文结合了大数据环境中的语义文本数据的相似度、凝聚度特征，利用潜在概率语义分析(Probabilistic Latent Semantic Analysis, PLSA)算法的奇异值分解对大数据环境中的语义文本重新建模，结合数据样本抽样，抽取具有相似语义特征的文本来获得词语间的语义关系，再根据样本特征，利用N-game统计语言模型和联合解码算法进行分词，最后通过期望最大化算法(EM)在数据挖掘中的优势来训练目标簇类(潜在类)，从而求解最佳阈值。

利用这个求得的阈值，结合层次聚类能够在不同聚类层次视角下获取不均匀聚类结果的优势，来控制聚类的聚类层次，在压缩训练集规模的同时尽可能地保证分类精度和泛化性能，从而获得相对合适的聚类结果。

本文的创新点如下：(1)利用N-game统计语言模型和联合解码分词算法对大数据文本数据进行解析。(2)提出一种大数据环境下面向语义特征相似性的阈值求解模型。(3)基于语义特征阈值，构建合适的聚类层次，提出了一种新的处理大数据聚类的层次聚类方法。

2 相关定义

2.1 层次聚类定义

聚类分析是数据挖掘领域的一个重要方法，主要用于发现数据中的不同类别及从数据中识别出特定的分布或模式[12]。目前在电子商务、电影、音乐、社交网络(如微信、微博)中得到了广泛的应用，其中的特征信息也广泛用于群体的特征划分，见表1，相对其他聚类算法，传统领域层次聚类方法对所有数据采用相同的度量，所得到的最终聚类结果是一棵聚类树[13]，实现了数据间基于分布的均匀划分，

表 1 数据语义特征聚类信息

	商品	电影	音乐	社交网络关系
聚类样本特征信息	个人电脑	爱情故事电影	流行音乐	同事
	汽车	科幻电影	爵士音乐	亲属
	智能手机	喜剧电影	说唱音乐	朋友
	手表	战争电影	古典音乐	夫妻

数据构成也比较单一。这样，最终形成的聚类层次之间的各个节点就不能显示彼此的特征关系。

2.2 凝聚度的定义

对基于余弦相似度度量的簇的凝聚度可用簇内相似度均值来度量[14]，即

$$T(\mathrm{Ck})=\frac{1}{m}\times\sum_{1\le i\le j\le n}u_{ij} \tag{1}$$

其中Ck为第k个簇，u_{ij}为Ck类第i个对象和第j个对象的余弦相似度，$m=n\times(n-1)/2$，n为簇类对象数。

2.3 相似度定义

相似度的定义是聚类分析的基础[15]，本文相似度的定义采用余弦相似度，如果两个向量的相似度为1，则表示两个向量完全相同，0 则表示两个向量没有相似之处。定义余弦相似度的公式为

$$u_{ij}=\frac{d_i\times d_j}{|d_i|\times|d_j|}=\frac{\sum_{k=1}^{n}(w_{ki}\times w_{kj})}{\sqrt{\sum_{k=1}^{n}w_{ki}^2\times\sum_{k=1}^{n}w_{kj}^2}} \tag{2}$$

其中u_{ij}表示u_i和u_j的相似度，$d_i=(w_{i1},w_{i2},\cdots,w_{in})$表示数据簇，$|d_i|$表示样本点$d_i$在欧式空间的长度。

2.4 N-game 统计语言模型和联合解码分词算法

在数据文本分词时首先利用 N-game 统计语言模型反映词和词之间的转移关系[16]，当两个最近的词(或字)相同时，映射两个词到同一个等价类。根据最大似然估计，语言模型的参数：

$$p(w_i\,|w_1\,w_2\cdots w_{i-1})=\frac{C(w_1w_2\cdots w_i)}{C(w_1w_2\cdots w_{i-1})} \tag{3}$$

其中，$C(w_1w_2\cdots w_i)$表示样本$w_1w_2\cdots w_{i-1}$在训练数据中出现的次数。

其次，将它们的切分词分别作为候选结果加入到一个统一的概率空间，通过求解联合概率空间最大概率下的切分路径，得到最终的切分结果的联合概率分布。

$$P=\prod_i P_{ci}\prod_j P_{wj} \tag{4}$$

P_{ci}为每个字的信息概率模型，P_{wj}为每个词的信息概率模型；而字模型和词模型则利用最大熵模型，进行共同决策。如果输出结果为k，那么结合

式(4)可以表示为

$$k = \arg\max p = \arg\max \prod_i P_{ci} \prod_j P_{wj} \tag{5}$$

对于一个数据文本的分词集合，通常需要计算它对于所有可能输出 k 的条件概率，其中选择最大概率的这个过程通常称为解码，因此，根据式(4)，式(5)对联合概率空间的字词联合解码可以表示为

$$\lg p = \sum_i \lg p_{ci} + \sum_j \lg p_{wj} \tag{6}$$

而字词的联合概率模型可以按照所有概率模型的共同作用选择最终的输出结果。字词联合概率模型可以表示为

$$k = \arg\max \left\{ \sum_i \lg p_{ci} + \sum_j \lg p_{wj} \right\} \tag{7}$$

3　基于潜在语义层次聚类算法

3.1 潜在概率语义分析(PLSA)阈值求解过程

假设大数据测试环境中存在一组具有相似对象的集合 $D=(d_1,d_2,\cdots,d_n)$ 和对象之中关键字的集合 $W=(w_1,w_2,\cdots,w_n)$。且同时存在一组具有某种特征的集合 $Z=(z_1,z_2,\cdots,z_k)$，假设对象中的关键字呈均匀分布，那么文档和词的联合分布概率可以表示为

$$\begin{aligned} p(w_i,d_i) &= p(d_i)p(w_j|d_i) \\ &= p(d_i)\sum_{k=1}^{K} p(w_i|z_k)p(z_k|d_j) \\ &= \sum_{k=1}^{K} p(z_k)p(w_j|z_k)p(d_i|z_k) \end{aligned} \tag{8}$$

式(8)中，$p(d_i)$ 表示第 j 个数据对象概率，$p(w_i|z_k)$ 表示为特征语义发生条件下关键字发生的条件概率，$p(z_k|d_j)$ 为数据对象中存在某种特征的条件概率。模型参数则可以利用极大似然函数来进行EM (Expectation Maximization)算法迭代，从而估计阈值L的最优解。

$$L = \sum_{i=1}^{M}\sum_{j=1}^{N} n(w_i,d_i)\lg p(w_i,d_i) \tag{9}$$

其中 $n(w_i,d_i)$ 是关键字 w 在数据对象中 d 中出现的次数，相对于潜在语义分析中的SVD分解，EM算法具有线性收敛速度，可以使似然函数达到局部最优。

对于抽取出来的小数据集，希望它能和大数据集一样包含了所有的自然簇(设总数为T)。估算样本大小的方法如下：

$$\begin{aligned} s = \frac{n}{n_i}\sqrt{\lg\left(\frac{1}{\theta}\right)\times 2\times k\times n_i + \lg\left(\frac{1}{\theta}\right)^2} \\ + \lg\left(\frac{1}{\theta}\right)\times\frac{n}{n_i} + k\times n \end{aligned} \tag{10}$$

式中 k 为抽取数据簇的比例 $(0<k\le 1)$，n 为数据总量，n_i 为数据空间里的预计簇类个数，式(10)表示按照概率 k 从簇类中抽取 $k\times n_i$ 个数据样本 s 的大小。

3.2 基于潜在语义层次聚类算法

在本文算法中，主要考虑合并条件下的层次聚类，聚类产生的结果来自于上一次的两个聚类的合并，从而来减少聚类中心的数量。当两个类的距离少于设定的阈值时，则将两个类进行合并；否则，进入下一步的探测。

假设大数据环境中存在一个数据集 $\boldsymbol{X}$，其中的元素为 $\boldsymbol{X}=\{\boldsymbol{x}_i, i=1,\cdots,5\}$, $\boldsymbol{x}_1=[1.0,1.0]^{\mathrm{T}}$, $\boldsymbol{x}_2=[2.0,1.0]^{\mathrm{T}}$, $\boldsymbol{x}_3=[5.0,4.0]^{\mathrm{T}}$, $\boldsymbol{x}_4=[6.0,5.0]^{\mathrm{T}}$, $\boldsymbol{x}_5=[6.5,6.0]^{\mathrm{T}}$，那么数据集 $\boldsymbol{X}$ 的样本矩阵可以描述为

$$\boldsymbol{D}(\boldsymbol{X}) = \begin{bmatrix} 1.0 & 1.0 \\ 2.0 & 1.0 \\ 5.0 & 4.0 \\ 6.0 & 5.0 \\ 6.5 & 6.0 \end{bmatrix}$$

经过欧式距离进行计算，可以得到其相应矩阵为

$$\boldsymbol{P}(\boldsymbol{X}) = \begin{bmatrix} 0.0 & 1.0 & 5.0 & 6.4 & 7.4 \\ 1.0 & 0.0 & 4.2 & 5.7 & 6.7 \\ 5.0 & 4.2 & 0.0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0.0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1.1 & 0.0 \end{bmatrix}$$

定义为 $D_{i,j}$ 数据类中 N_i 所有样品和 N_j 类中所有样本间的最小距离，即

$$D_{i,j} = \min(d_{i,j}) \tag{11}$$

利用欧式空间的两点间的距离公式可以得到

$$d_{i,j} = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2} \tag{12}$$

(x_n,y_n) 和 (x_m,y_m) 分别为两个连通区域 m,n 边缘点的坐标。当两坐标之间的距离 $d_{i,j}=\sqrt{(x_m-x_n)^2+(y_m-y_n)^2}$ 比设定的值小时，则将这两个区域进行合并，否则探索下一个坐标。以前已经有算法证明，虽然该方法为数据的特征提取提供了较有效的方法，但是在数据过大，离散数据较多的时候，算法的效率较低，因此对连通域的边缘进行采样以减小算法时间复杂度。将连通域的边缘坐标作为采样数据总体，由于每个边缘坐标作为个体是独立的，同时为满足等概率被抽取，故采用等距抽样。下面讨论在给定层次上如何确定最佳聚类数，它等价于确定用多少个聚类能达到最适合数据聚类效果，以往直观的方法是在近邻树图中搜索生命周期长的聚类。结合以上研究，本文提出的构建语义特征的层次聚类分类算法表述如下：

令 $g(c_i,c_j)$ 为所有可能的 $\boldsymbol{X}$ 聚类对的函数，此函数用于测量两个聚类之间的近邻性，用 t 表示当前聚类的层次级别。通用合并算法的伪码描述如表2所示。

表 2 通用合并算法的伪码

算法 基于语义特征的层次聚类算法
输入：经过PLSA算法训练的样本点
输出：聚类系统树
(1)初始化样本数据；
(2)选择 $R=\{c_i=\{x_i\},i=1,2,\cdots,n\}$ 作为初始聚类；
(3)令 $t=0$；
(4) do:
(5) $t=t+1$；
(6)Select (c_i,c_j) from R_{t-1} according to distance sets (c_r,c_s)，满足

$$g(c_i,c_j)=\begin{cases}\min_{r,s}\, g(c_r,c_s)\\ \max_{r,s}\, g(c_r,c_s)\end{cases}$$

若 g 为不相似函数时取最小值，若 g 为相似函数时取最大值；
(7)定义 $c_q=c_i\cup c_j$，并产生新聚类 $R_t=(R_{t-1}-\{c_i,c_j\})\cup\{c_q\}$；
(8)直到所有向量都被加入一个聚类中；
(9) Update distance sets；
(10)While t 获得最优值，Return；

3.3 新聚类算法的复杂度分析

对于新算法在运行上的空间开销，可以考虑的因素是等待聚类的大数据抽样的特征样本为 k 个，假设聚类采用单链接的聚类方法，每个待聚类的簇按照串行的实现方式进行摆放，那么聚类的总时间开销为 $O(n^3)$，同时又假设抽取样本所花费开销为 m，那么算法总的空间复杂度 $S(n)$ 可以表示为

$$S(n)=m+O(n^3) \tag{13}$$

根据前文所提到的计算方法特征词的提取规则，具有相同或者相似特征的样本规模非常小，因此新算法的空间复杂度为 $O(n)$。同时，数据样本的分词算法的时间复杂度为 $O(n)$。因此算法的时间复杂度主要与前文的提取特征，分词，层次聚类有关。抽样数据如果采用线性排列，假设分词所需时间为 t，提取特征的时间为 l，那么 k 个样本分词的总时间为 $k\times t$，提取特征的总时间为 $k\times l$，聚类的时间复杂度 $k\times n^3$，那么算法的总时间复杂度可以表示为

$$T(n)=k\times t+k\times l+k\times n^3 \tag{14}$$

可见新算法的时间开销是线性的，为 $O(n)$。

4 实验分析

为验证经过提取特征值判定的层次聚类算法在处理不规则形状簇的正确性和有效性，比较本文所给出的语义特征算法模型在给定数据区间获得聚类的能力。本节构建了具有不规则形状的数据集，并在该数据集上检验算法的聚类正确率。实验数据来自芬兰大学(http://cs.joensuu.fi/sipu/datasets/)上的聚类测试集。我们选取一组样本点进行了人工数据合成，共包括65536个特征值元组，使用MATLAB2014搭建实验环境，硬件配置为，CPU：I5-4590 CPU3.3 GHz, RAM 4 GB进行试验仿真。

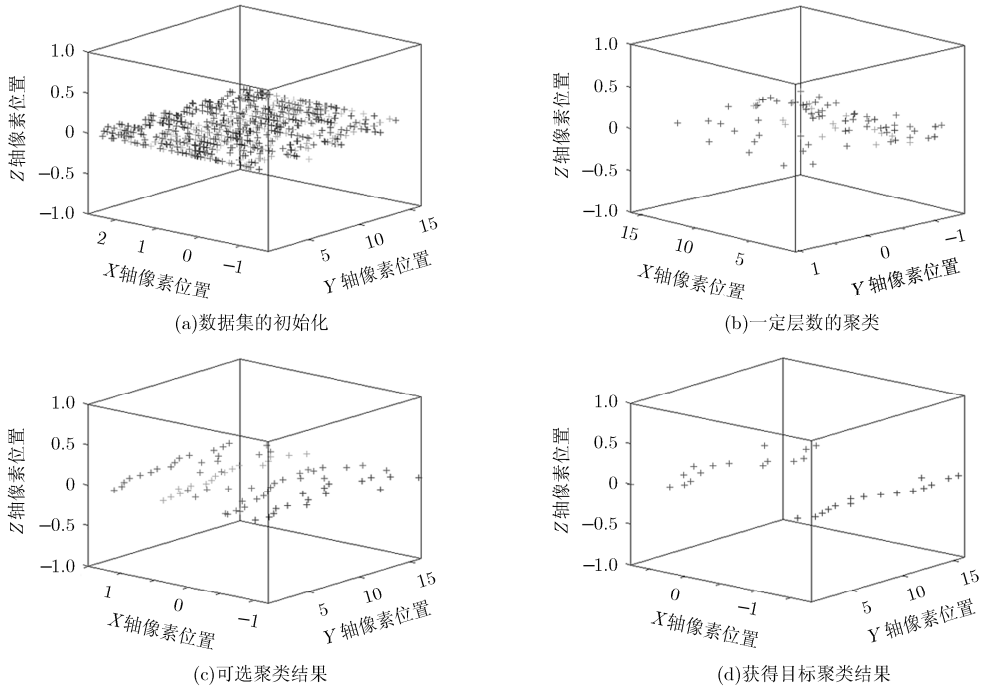
图 1 测试数据集1

4.1 数据集测试

在本实验中，聚类最终输出是根据语义特征求的阈值 L 在该 L 层获得较合适的聚类结果，也就是说在该层上基本可以体现语义特征的区分群体，根据式(9)求得的特征阈值分别比较了不同特征值产生等价类的计算时间及存储空间，实验数据是通过投影在3维空间的样本向量进行表示。参数 L 决定了最终聚类树的大小，由于本实验主要观测聚类树的构建过程，因此为了避免超过阈值 L 而形成完全聚类树的构建，本实验将先通过递增法来表示聚类树的形成过程。3次实验中 L 的取值分别为500, 1000, 2000。

4.2 数据聚类仿真

实验结果见图2，图2(a)表示对数据集的初始化，可以看到数据是无序散列的，图2(b)表示进行了一定的聚类，但是没有得到具体的聚类期望，图2(d)表示根据特征阈值获得了目标聚类，算法终止。可见利用特征值进行干扰下的层次聚类算法，可以利用语义特征的提取减少聚类次数，这样明显减少了读写磁盘的次数，从而节约了运算空间和提高了计算效率。在这里值得一提的是，我们也可以选取图2(c)的聚类结果作为最终聚类，因为该图已经基本体现了了特征分类，而且计算时间要比图2(d)小一倍，当然这主要是根据用户的具体需要进行判定。



(a)数据集的初始化　(b)一定层数的聚类

(c)可选聚类结果　(d)获得目标聚类结果

图2 语义特征聚类仿真结果

4.3 算法效率分析

目前针对于大数据算法分析的研究方法主要有模糊信息粒化，粗糙集近似，多维数据去重法，每一种数据的处理方法针对不同数据的特征有自身特征的应用场合，其算法的时间复杂性比较见表 3。

为验证语义特征阈值算法的正确性和有效性，我们选取了芬兰大学的测试数据集(http://cs.joensuu.fi/sipu/datasets/)进行了对比实验测试，数据集见图 3。

实验结果表明，本文所提出的基于语义特征阈值的层次聚类方法在不同数据集上均可以得到正确的聚类数，并且能够在相对较短的时间完成响应。我们选取了Toy Problem, R15, Flame, Aggregation实验结果进行说明，见图4。

为了验证本文方法在计算时间和算法开销上的有效性，本文方法和多维数据去重聚类方法在Flame框架数据集进行了运算时间的比较测试，经过比较可以看出，本算法在数据的计算效率上优于多维数据去重聚类算法，并且这种时间优势随数据量规模的递增而增加。实验结果见表4和图5。

实验结果证明，基于语义特征阈值的层次聚类方法所产生的等价类查询时间比在一般意义上的层次聚类算法进行查询响应时间要快，这是因为以等价类方式保存的存储空间比传统意义的层次聚类算法要小；同时，它对任意形状的数据集均有效，对输入参数具有弱依赖性，并且具有处理噪声的能力，能够找到满足特定的聚类的约束。经过试验测试，通过对数据集进行特征划分，然后利用概率估计计算出来的阈值，可以较快地获得聚类的最佳层数。

5 结束语

目前大数据的应用非常广泛，特别是在商业智能、政府决策、公共服务、医疗健康等领域已经成为研究热点。以医疗健康领域为例，当前医疗与健

表 3 大数据典型分析方法适应性比较

大数据分析方法	典型时间复杂性	主观性表达方式	数据分析能力	对大数据处理特性(巨量/异构/高速的适应能力)
模糊信息粒化法	$O(nl)$, l 为粒化的属性数	隶属度函数的设定	较弱	对巨量性适应能力较强，但是对复杂关系适应能力较弱
粗糙集近似法	$O(n^2s)$, s 为聚类分类数	相似关系的设定	较强	已有针对大数据的增量并行处理算法，适应能力较强
多维数据去重法	$O(n^2 \lg n)$	数据去重的设定	强	对大数据中的非结构化数据，重复数据有较好的处理能力
语义特征阈值层次聚类法	$O(n)$	语义特征阈值求解	强	对大数据中的语义分类划分具有较好的处理能力

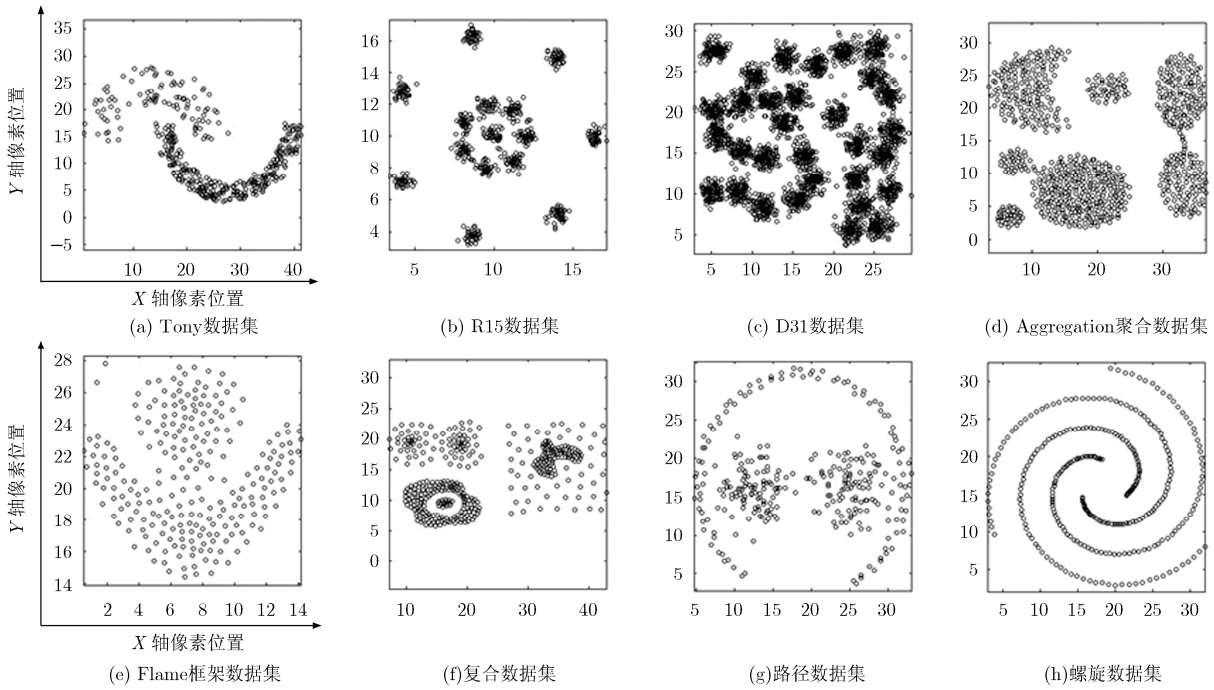


图3 测试数据集2

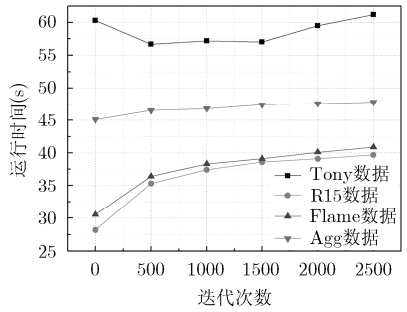


图 4 不同数据集的语义特征聚类时间图

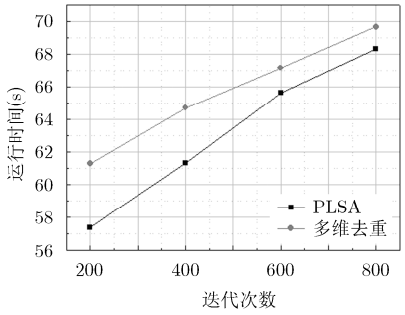


图 5 语义特征阈值方法和多维数据去重方法时间对比

表 4 不同数据量计算时间比较

数据	数据迭代次数	数据特征元组	PLSA 聚类算法时间(s)	多维数据去重算法时间(s)
1	200	26433	57.36	61.28
2	600	52196	61.33	64.71
3	1000	102575	65.61	67.16
4	1400	145017	68.31	69.67

康服务的数据分析面临了前所未有的挑战，具体表现有，医疗患者诊疗方案的数据冗余，因为对于相同的病情，不同的医生描述有相当大的差别，这导致医疗系统中数据冗余非常离散和庞大，而如此大规模的不规则数据为数据分析和处理带来了不便，因此开展研究数据的价值挖掘和数据分析就显得尤为必要。本模型着重关注数据的语义特征，通过算法将数据进行处理和分析，最后对相似语义的数据进行聚类，减少数据在计算和分析过程中的计算开销和存储开销，对于大数据分析具有研究意义，为推动新型医疗健康服务和其他领域的信息化变革，具有重要的参考价值。

同时，本文利用实际数据进行大量的实验分析，试验结果表明本方法可以通过数据出现的特征选择进行降噪和数据简约降低数据的复杂度，根据数据关键字的出现频率来确定聚类特征。在聚类阶段，利用似然估计所计算出来的阈值进行聚类终止。实践证明，该算法比传统的数据层次聚类分析算法具有更好的数据分析性能，能更好地提高数据分析的可视化，为求解数据的多维度之间的映射关系提供了一种新的思路。当然这个方法还有改进的地方，例如数据去重的精确度还不太高，数据的聚类分析的扩展度不够，而这方面的工作是下一步的研究重点。

参 考 文 献

[1] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J].

软件学报, 2014, 25(9): 1889−1909.
Cheng Xue-qi, Jin Xiao-long, Wang Yuan-zhuo, *et al*.. Survey on big data system and analytic technology[J]. *Journal of Software*, 2014, 25(9): 1889−1909.

[2] Du Y, He Y, Tian Y, *et al*.. Microblog bursty topic detection based on user relationship[C]. IEEE 6th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2011, 1: 260−263.

[3] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48−61.
Sun Ji-gui, Liu Jie, and Zhao Lian-yu. Clustering algorithms research[J]. *Journal of Software*, 2008, 19(1): 48−61.

[4] Choromanska A, Jebara T, Kim H, *et al*.. Fast spectral clustering via the nyström method[C]. Proceedings of the 24th International Conference, Algorithmic Learning Theory 2013, Singapore, 2013: 367−381.

[5] Hearn T A and Reichel L. Fast computation of convolution operations via low-rank approximation[J]. *Applied Numerical Mathematics*, 2014, (75): 136−153.

[6] Gajjar M R, Sreenivas T V, and Govindarajan R. Fast computation of Gaussian likelihoods using low-rank matrix approximations[C]. 2011 IEEE Workshop on Signal Processing Systems (SiPS), Beirut, Lebanon, 2011: 322−327.

[7] 崔颖安, 李雪, 王志晓, 等. 社会化媒体大数据多阶段整群抽样方法[J]. 软件学报, 2014, 25(4): 781−796.
Cui Ying-an, Li Xue, Wang Zhi-xiao, *et al*.. Sampling online social media big data based multi stage cluster method[J]. *Journal of Software*, 2014, 25(4): 781−796.

[8] Chen W Y, Song Y, Bai H, *et al*.. Parallel spectral clustering in distributed systems[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(3): 568−586.

[9] 丁世飞, 贾洪杰, 史忠植. 基于自适应 Nyström 采样的大数据谱聚类算法[J]. 软件学报, 2014, 25(9): 2037−2049.
Ding Shi-fei, Jia Hong-jie, and Shi Zhong-zhi. Spectral clustering algorithm based on adaptive Nyström sampling for big data analysis[J]. *Journal of Software*, 2014, 25(9): 2037−2049.

[10] Chen X and Cai D. Large scale spectral clustering with landmark-based representation[C]. Proceedings of the 25th AAAI Conference on Artificaial Inteligence, San Francisco, USA, 2011: 313−318.

[11] 慈祥, 马友忠, 孟小峰. 一种云环境下的大数据 Top-K 查询方法[J]. 软件学报, 2014, 25(4): 813−825.
Ci Xiang, Ma You-zhong, and Meng Xiao-feng. Method for Top-K query on big data in cloud[J]. *Journal of Software*, 2014, 25(4): 813−825.

[12] Horng S J, Su M Y, Chen Y H, *et al*.. A novel intrusion detection system based on hierarchical clustering and support vector machines[J]. *Expert Systems with Applications*, 2011, 38(1): 306−313.

[13] Bahmani B, Moseley B, Vattani A, *et al*.. Scalable k-means++[J]. *Proceedings of the VLDB Endowment*, 2012, 5(7): 622−633.

[14] Zhang X and You Q. Clusterability analysis and incremental sampling for Nyström extension based spectral clustering[C]. IEEE 11th International Conference on Data Mining (ICDM) , Vancouver, Canada, 2011: 942−951.

[15] Zhang K and Kwok J T. Clustered Nyström method for large scale manifold learning and dimension reduction[J]. *IEEE Transactions on Neural Networks*, 2010, 21(10): 1576−1587.

[16] Vlachou A, Doulkeridis C, Kotidis Y, *et al*.. Monochromatic and bichromatic reverse top-k queries[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2011, 23(8): 1215−1229.

罗恩韬：1978 年生，男，副教授，博士生，研究方向为云安全、大数据聚类分析、隐私保护.

王国军：1970 年生，男，教授，博士生导师，研究方向为可信计算、大数据健康医疗、隐私保护.