对象间矢量感应聚类算法

李雄飞[1], 孙 涛[2], 武佳薇[1]
(1. 吉林大学计算机科学与技术学院符号计算与知识工程教育部重点实验室, 吉林长春 130012;
2. 中国科学院长春光学精密机械与物理研究所, 吉林长春 130033)

摘 要: 从万有引力角度考虑, 质点之间相互影响包括距离和方向两个方面. 本文讨论数据之间的矢量感应, 并将其应用于聚类算法VICA. 引入对象的标量感应函数和方向感应函数, 提出矢量感应函数概念. 并给出确定方向感应函数的两个方法: 方向相似度法和累加法. 将核心对象邻域中的对象投影, 进行向量单位化, 考察核心对象的邻域均匀感应程度, 将与均匀感应核心对象均匀感应密度可达的对象聚成一个簇. 理论分析和实验结果表明, 算法可以处理任意形状的簇, 有效地排除了稀疏感应对象这类噪声, 并且可以解决高维数据聚类边界区分不明显、密度分布不均、类边界噪声对象多的问题, 提高了聚类精度. 由于感应函数是一个泛化定义, 算法具有通用性和可扩展性. 将半结构化数据变换到欧式空间时, 容易出现边界稀疏对象, 算法可以有效处理噪声. 因此, 算法适用于大规模的高维数据集合, 也可用于半结构化数据聚类.

关键词: 矢量感应; 聚类; 投影点; 稀疏感应对象; 均匀感应邻域
中图分类号: TP313 文献标识码: A 文章编号: 0372-2112 (2011) 06-1347-06

Clustering Algorithm Concerning Vector Influence Between Objects

LI Xiong-fei[1], SUN Tao[2], WU Jia-wei[1]
(1. *Key Laboratory of Symbolic Computation and Knowledge Engineer of Ministry of Education, Department of College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China;*
2. *Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, Changchun, Jilin 130033, China*)

Abstract: Considering from the law of gravity, the influence between particles includes distance and direction. After discussing the vector influence between data objects, it is applied in clustering algorithm. Vector influence function is presented from the scalar influence function and direction influence function. Two methods—similarity and sum are introduced to compute the direction influence. The algorithm deals with the core point by getting the projection of the points in its neighborhood to judge whether it is uniformity influence. Only uniformity influence points can be expanded to form clusters. The theoretical analysis and experimental results indicate that the algorithm can discover clusters with arbitrary shape and can effectively eliminate noise such as boundary sparse points. It solves the difficulties of clustering high dimensional spatial data such as the spatial distribution of the data, not obvious boundary between clusters, too many noise data points and the phenomenon that the distance between the nearest and farthest neighbors of a data point goes to zero etc. The algorithm improves the accuracy of clustering and offers better results of clustering on various data sets. It executes effectively and efficiently. The algorithm is scalable and general. While transforming the semi-structure data into Euclid space, it will always appear boundary sparse objects, VICA can deal with the noise effectively. Therefore, the algorithm is proper with the high dimension data set, and also can be applied in the semi-structure data clustering.

Key words: vector influence; clustering; projection points; sparse influence points; uniformity influence neighborhood

1 引言

聚类就是将数据对象分组成为多个簇(cluster), 同一个簇中的对象相似度较高, 不同簇中的对象则差别较大. 只考虑方向和相对位置的数据是方向数据[1]; 具有几十、几百甚至成千上万个属性的数据称为高维数据, 数据特征是稀疏性和维数灾难性[2]等; XML等先有数据后有结构的数据是半结构化数据. 面向上述数据的聚类技术是目前的研究热点与难点.

DBSCAN[3]算法(Density Based Spatial Clustering of Applications with Noise)是一种基于密度的空间聚类算法, 利用簇的高密度连通性可以快速、有效地发现任意

收稿日期: 2009-09-07; 修回日期: 2010-09-23
基金项目: 国家科技支撑计划项目(No. 2006BAK01A33); 吉林省科技发展计划项目(No. 20090704)

形状的簇.但存在以下弱点:当数据分布不均匀时聚类的效果差;不能精确地判断簇边界点;缺乏对边界噪声点的监测;高维数据聚类困难.原因在于算法不适合数据密度分布差异较大以及簇边缘频繁出现噪声的情况.

本文在研究半结构化数据聚类问题中总结出一个密度聚类算法 VICA(Vector Influence Clustering Algorithm).算法以数据对象之间相互作用为指导,从标量和方向两个角度讨论密度聚类问题,提出累加向量法和方向相似度法等计算矢量感应函数的方法.实验结果表明,VICA 算法能较好地处理密度分布不均且含有边界噪声的数据.

2 相关工作

目前,基于密度的聚类算法[4]主要有 GDBSCAN[5]、OPTICS[6]、DBCLASD[7]、DENCLUE[8] 等.其中,DENCLUE 算法提出数据库中对象之间的感应函数概念,并给出高斯函数等感应函数.文献[9]提出 k-PCLDHD 和 k-LDCHD 算法,旨在解决高维数据空间的聚类问题.文献[10]提出的基于数据分区的 DBSCAN 算法、文献[11]提出的 CURD 算法等从数据密度分布上改进了 DBSCAN 算法.此外,近年来还有一些针对高维密度聚类的研究成果[12~15].

数据对象之间不仅存在方向上的相似性,也存在方向上的相互影响.已有的聚类算法没有充分利用方向上的相互影响信息.

3 相关定义

3.1 处理半结构化数据聚类遇到的问题

半结构化数据是高维、维空间不统一的数据.对属性空间不同的高维数据集进行聚类时,传统做法有两种:一种是统一维度空间,构建关于所有属性的二维数据表,对缺少属性值的数据对象进行缺失值填充(如取均值);另一种方法是用主成分提取、粗集属性约简等方法对数据进行选维或降维.这些方法通常会改变数据的分布密度,影响聚类质量.高维属性空间中的稀疏特征和无关属性干扰了数据的聚类趋势.影响了现有聚类算法的质量.

3.2 基本思想

从万有引力定律看,自然界中任何两个物体都是相互吸引的.数据库中的每个对象对最终的聚类结果都有贡献.这种贡献可用对象间的感应(influence)来衡量,具体表现为距离和方向上的感应影响.将数据集中某对象受到其它对象的作用泛化为矢量感应函数,用于聚类过程中.

3.3 相关定义

定义 **1** 矢量感应函数 $\varphi(x)$.某对象 x 的矢量感应函数 $\varphi(x)$ 可用两部分表达:标量感应函数和方向感应函数.

其中,标量感应函数衡量对象之间的距离或相似性.图 1 表示 DBSCAN 算法应用到某数据集上的聚类结果,左图中算法虽然将下方的数据分为两个簇,而上方的稀疏数据均被划为了噪声对象;通过调整参数可将上方的数据归为一簇,但同时会将下方本应聚为两簇的高密度区域归并为一个簇,并且将右下方的噪声对象 p' 错误地纳入簇中.可见,仅从距离上衡量对象之间的感应影响具有局限性,难以在将高密度区域聚集的同时,又能准确地排除噪声.

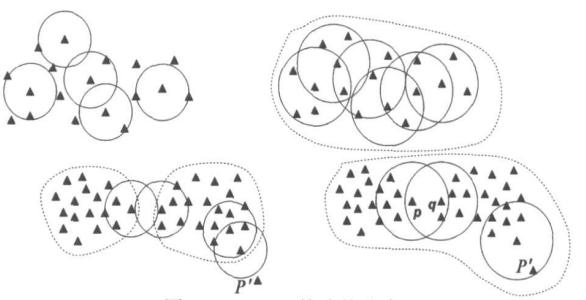

图1 DBSCAN算法的聚类效果

定义 **2** 方向影响函数.d 维属性空间 F^d 中一个数据对象 y 的方向影响函数 $g_B^y(x)$ 由一个基本影响函数 g_B 定义:

$$g_B^y(x)=g_B(x,y)$$

定义 **3** 方向感应函数.设 N 个数据对象所组成的数据集 $D=\{x_1,\cdots,x_N\}\subset F^d$,某个对象 x 的方向感应函数为所有数据对象对该对象的方向影响函数之和,即:

$$g_B^D(x)=\sum_{i=1}^{n} g_B^{x_i}(x)=\sum_{i=1}^{n} g_B(x,x_i)$$

定义 **4** 邻域集合 $near(x,\sigma)$.任意对象 $x\in D$,则有 $near(x,\sigma)=\{y\in D\mid dist(x,y)\leqslant\sigma\}$ 为 x 的 σ 邻域集合.

定义 **5** 邻域方向感应函数.对象 x 的邻域 $near(x)$ 方向感应函数为

$$\hat{g}^D(x)=\sum_{y\in near(x)} g_B^y(x)$$

3.4 相关理论及证明

设数据集 D,对象 p、q、$p'\in D$.下面给出两种确定方向感应函数 g_B 的方法:相似度法和累加法.

定义 **6** 投影点.设对象 $p\in near(q,\sigma)$,点 p' 在以 q 为起始点且经过点 p 的射线上,且 $|\overrightarrow{qp'}|=1$,则称点 p' 为对象 p 在 $near(q,\sigma)$ 上的投影点.邻域 $near(q,\sigma)$ 内对象的投影点集合用 P_q 表示.

图 2 示意三维空间邻域内对象的投影.

定义 **7** 方向相似度.对于对象 $p\in near(q,\sigma)$,$p'\in P_q$,有集合 $S=\{q'\mid q'\in P_q,\text{且}|\overrightarrow{p'q'}|<L_c\}$,$|S|$ 为点 p 在 $near(q,\sigma)$ 上的方向相似度.

对象的方向相似度为邻域内的某方向上对象个数, 用于衡量该方向的感应程度.

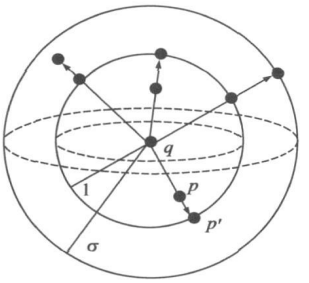

图2　三维空间投影示意图

定义 **8**　均匀感应邻域. 对所有对象 $p \in near(q, \sigma)$, 若满足 $|S|/|near(q, \sigma)| < \eta$, 则称 σ_{-q} 为均匀感应邻域否则, 称 σ_{-q} 为非均匀感应邻域.

均匀感应邻域考察对象在邻域内的分布状况. 若分布较均匀, 就为均匀感应邻域; 否则, 就为非均匀感应邻域. 方向相似度可以判断对象的邻域均匀感应程度. 其中, 参数的取值范围为 $L_c \in (0, 2)$, $\eta \in (0, 1)$. 邻域均匀感应程度的本质是从方向感应的思想出发衡量一个对象邻域内的各对象对它在方向上的影响. 图 3 中两个对象 p、q, 在给定的参数 L_c、η 下, 可知 σ_{-p} 邻域感应均匀, σ_{-q} 邻域感应不均匀.

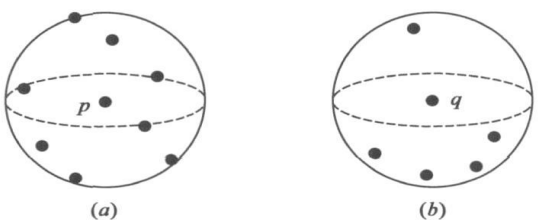

图3　对象邻域感应均匀和感应不均匀图示

相似度法求方向感应函数的过程是: 先将 $near(p, \sigma)$ 内的所有对象投影到单位曲面上, 然后依次判断各个投影对象的方向相似度, 从而得到该对象的方向感应函数, 最后确定对象 p 的邻域均匀感应程度.

定义 **9**　累加向量. 设向量 $\boldsymbol{V} = \sum_{1}^{n} p_i (p_i \in P_q)$, 则称 $\boldsymbol{V}$ 为累加向量.

累加法是将投影向量求和, 当 $|\boldsymbol{V}| > \lambda$ 时, 对象 p 邻域感应不均匀. 用户给定 λ

通过对象间的相互作用得出矢量感应函数为标量感应函数 $f^D(p)$ 和方向感应函数 $\hat{g}^D(p)$. 其中标量感应函数的阈值为 ξ

定义 **10**　均匀感应核心对象. 对于对象 p, 若有 $f^D(p) > \xi$ 且 $\hat{g}^D(p) < \eta$, 即 σ_{-p} 邻域感应均匀, 则称对象 p 为均匀感应核心对象.

定义 **11**　均匀感应直接密度可达. 对 $\forall p, q \in D$, 如果 $p \in near(q, \sigma)$, 并且 q 是一个均匀感应核心对象, 则对象 p 从对象 q 出发是均匀感应直接密度可达的.

定义 **12**　均匀感应密度可达. 如果存在数据对象链 $p_1, p_2, \cdots, p_n (p_1 = q, p_n = p)$, 对 $p_i \in D (1 \leqslant i \leqslant n)$, p_{i+1} 是从 p_i 出发均匀感应直接密度可达的, 则对象 p 是从对象 q 均匀感应密度可达的.

定义 **13**　簇. 数据集 D 关于 σ、ξ、L_c 和 η 的簇 C (其中 $|C| > 1$) 为满足下列条件的非空集合:

(1) 对 $\forall p, q \in D$, 若 $p \in C$ 且 q 是 p 关于 σ、ξ、L_c 和 η 均匀感应密度可达的对象, 则 $q \in C$ (极大性);

(2) 对 $\forall p, q \in C$, p、q 是关于 σ、ξ、L_c 和 η 均匀感应密度可达的 (连通性).

定义 **14**　稀疏感应对象. 若对象 p 满足 $|near(p, \sigma)| < \xi$ 且其邻域内对象均不是均匀感应核心对象, 则该对象为稀疏感应对象.

定义 **15**　噪声对象. 已知数据集 D 按均匀感应密度可达划分为 m 个簇 $C_i (1 \leqslant i \leqslant m)$, 若对象 $p \in D$ 且 $\forall (1 \leqslant i \leqslant m) \supset p \notin C_i$, 则称 p 是噪声对象.

定理 **1**　稀疏感应对象为聚类中的噪声.

证明　设对象 p 为稀疏感应对象, 则根据定义 14 其邻域内其它对象都不是均匀感应核心对象, 故对象 p 无法通过均匀感应密度可达扩展形成聚类. 设 $p_1, p_2, \cdots, p_i \in near(p, \sigma)$, 且 $p_1, p_2, \cdots, p_i$ 都不是均匀感应核心对象, 则有:

(1) 若 p_i 为非核心对象, 则 p_i 无法通过密度可达概念扩展形成聚类, 因此 p 也无法被聚类所包含;

(2) 若 p_i 是核心对象, 但是其邻域感应不均匀, 则 p 不是 p_i 的均匀感应密度可达对象, 由定义 13 中簇的极大性可知 p 不属于该聚类.

因此, 稀疏感应对象 p 不被任何聚类所包含, 是聚类中的噪声. 证毕.

定理 **2**　设 C 和 C' 分别为定义 13 和 DBSCAN 算法中的簇, 则有:

(1) $C \subseteq C'$;

(2) 若稀疏感应对象 $q \in C'$, 则 $q \notin C$.

证明　(1) 在 σ、ξ 参数取值相同时, $\forall m, n \in C$, 依据定义 13 中的连通性有 m 和 n 均匀感应密度可达. 而 DBSCAN 算法形成簇 C' 的过程不需要进行均匀感应的判断, 即在相同的参数条件下, m 和 n 也满足密度可达, 故 $m, n \in C'$. 因此, 有 $C \subseteq C'$.

(2) 设对象 $q \in C'$; 满足 $|near(q, \sigma)| < \xi$, 且 $near(q, \sigma)$ 均为邻域感应不均匀的核心对象. 设 $p \in near(q, \sigma)$, 由 DBSCAN 算法可知 p 是一个核心对象, 且 $|near(p, \sigma)| < \xi$, p 可以扩展形成簇, q 被归入簇中. 但由定义 14 可知, q 为稀疏感应对象. 由定理 1 知 q 为噪声. 对定义 13 得到的簇 C, 有 $p \in C$, $q \notin C$, 故定义 13 中的簇剔除了 q 这类噪声. 证毕.

4　VICA 算法

4.1　算法思想和描述

VICA 算法扫描数据库, 对于对象 p, 如果 $f^D(p) >$

ξ 且 $\hat{g}^D(p) < \eta$ 标记为簇, 并且关于参数(σ, ξ, L_c, η)扩展其均匀感应密度可达对象; 否则暂时标注为噪声, 处理数据集合中的下一个对象. 相似度法求取对象方向感应函数的算法如下:

Step 1 判断 对于对象 p, if $f^D(p) > \xi$ Yes: 转到 step2; No: 暂时标为噪声, 转到 step4;

Step 2 投影 以对象 p 为中心建立坐标系, 将 A_p 中的所有对象向量单位化形成 P_p;

Step 3 遍历 遍历 P_p, if $\hat{g}^D(p) < \eta$, Yes: 标为簇, 继续扩展簇; No: 暂时标为噪声, 到 step4;

Step 4 继续处理数据集中下一个对象;

累加法求取对象方向感应函数的过程只需将Step 3 改为如下形式:

Step 3 累加: 求 $V = \sum_1^n p_i (p_i \in P_q)$, if $\hat{g}^D(p) < \lambda(p)$, Yes: 标记为簇, 继续扩展簇; No: 暂时标为噪声, 到 step4;

感应函数是一般化的定义, 可以根据实际情况选择标量感应函数和矢量感应函数.

4.2 参数问题

参数 σ 规定对象的邻域 $near(p)$, ξ 确定标量感应函数的阈值, 用于找出"最薄"的簇. 参数 L_c 和 η 用来衡量受邻域内对象方向上的影响. 实验分析表明, 在相同 σ 和 ξ 取值下, VICA 算法能够比 DBSCAN 算法获得更加优化的聚类结果. 无论是二维还是高维数据集, 参数取值 $L_c = \sqrt{2}$、η = 0.75 时是较好的选择. λ 根据数据分布而变化, 本文实验中的取值范围是区间[1.0, 5.0].

4.3 算法的时间复杂度

DBSCAN 采用 R^*-树时有 $O(N\log N)$. VICA(方向相似度法)需要计算核心对象的邻域均匀感应程度. 设 M 为核心对象邻域内对象个数的均值, VICA 算法的时间复杂度为 $O(N\log N) + O(\sum_1^K \frac{M(M-1)}{2})$. K 为核心对象的个数. 最坏情况是 $K = N$, 时间复杂度近似为 $O(N((\log N) + \frac{M(M-1)}{2}))$, 当 $O(\log N) = M^2$ 时, 对象集中分布于数据空间中的狭窄区域, 极为罕见, 时间复杂度相对于 DBSCAN 算法呈指数级增长. 所以, 认为 VICA 的时间复杂度相对于 DBSCAN 的时间复杂度呈线性增长. 邻域空间的规模与数据集空间的规模的比值越小, 算法的额外计算量越少.

VICA(累加法)的时间复杂度为: $O(N\log N) + O(\sum_1^K M)$, 相对于 DBSCAN 的时间复杂度呈线性增长, 当面对较大规模的数据库时, 它们的时间复杂度相当.

5 实验结果

测试数据采用 UCI 的 Iris、2D2K 以及 New_thyroid 数据集. 用 C# 编写 VICA, 硬件配置为 PIV 2.6GHz, 1GB 内存, 80GB.

聚类模型评估采用文献[16]提供的方法. 取 $L_c = \sqrt{2}$、η = 0.75 时测试了 VICA 算法的 S_Dbw 指标.

表 1 Iris 数据集的测试结果(ξ = 5, λ = 1.0)

Input	S_Dbw		
σ	DBSCAN	$VICA_{(Similarity)}$	$VICA_{(Sum)}$
0.9	0.1014	0.0605	0.0344
1	0.1069	0.0825	0.0371
1.1	0.1604	0.0952	0.04207
1.2	0.1604	0.0985	0.04897
1.3	0.1604	0.0944	0.05863

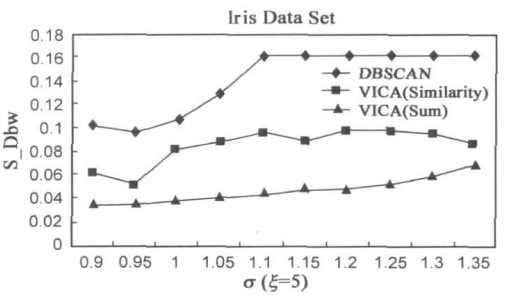

图4 Iris数据集的聚类质量比较

表 2 2D2K 数据集的测试结果(ε = 0.03, λ = 4.5)

Input	S_Dbw		
ξ	DBSCAN	$VICA_{(Similarity)}$	$VICA_{(Sum)}$
4	0.4063	0.1306	0.2722
6	0.564	0.2722	0.1822
8	0.2542	0.1822	0.2178
10	0.2994	0.2883	0.2804
12	0.2812	0.2719	0.2807

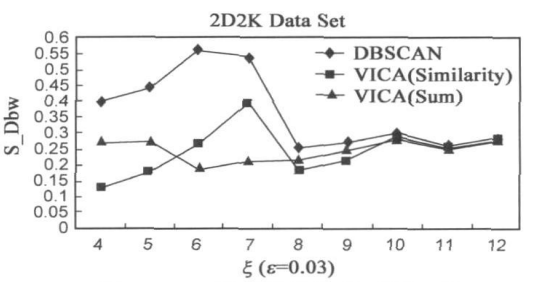

图5 2D2K数据集的聚类质量比较

表 3 New_thyroid 数据集的测试结果(MinPts = 6)

Input	S_Dbw	
ε	DBSCAN	VICA
3.80	1.0000	0.8010
3.90	1.0000	0.6059
4.00	0.4930	0.4230
4.10	0.7828	0.5516
4.20	1.0000	0.5184

可见, VICA 两种方法的 S_Dbw 明显优于 DBSCAN, 聚类准确率高于 DBSCAN, 在解决了边界噪声的问题后, 取得更高的聚类质量. 当面对不同的数据集合时,

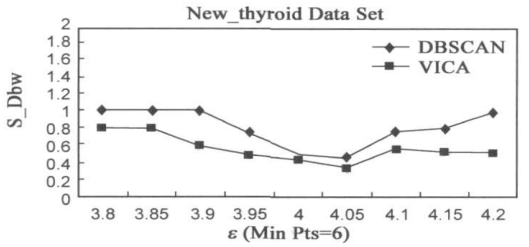

图6　New_thyroid数据集的聚类质量比较

相似度法和累加法的表现不同, 但都比 DBSCAN 的结果好. VICA 算法同样适用于多维数据集.

VICA 算法利用对象之间的标量和矢量感应, 提高了对不同数据密度区域的区分能力, 排除了稀疏感应对象这类噪声, 聚类质量明显提高.

表 4 是算法时间开销的实验结果. 对相同数据集, 所列数据均为DBSCAN 算法和 VICA 算法得到较好聚类结果下的各自所用的时间, 分别记为 T_{DBCSAN}、$T_{VICA(similarity)}$和 $T_{VICA(sum)}$.

$$T_{DS1} = T_{VICA(similarity)} / T_{DBCSAN}$$

$$T_{DS2} = T_{VICA(sum)} / T_{DBCSAN}$$

从表 4 和图 7 可以看出: 在相同数据规模下, VICA (相似度法)比 DBSCAN 的时间开销略高, 但随着数据规模的增加, 二者的时间开销均增加, 并且两种算法运行时间的差距缩小, $T_{DS1} \approx 1$. VICA(累加法)与 DBCSAN 算法的时间开销基本持平, $T_{DS2} \approx 1$, 在不同数据集上, T_{DS2} 有小波动, 算法实验结果与 4.3 节的时间复杂度分析结果相符.

表 4　不同数据规模下的算法时间开销比较

Data Scale	Time Spending(s)		
	T_{DBCSAN}	$T_{VICA(similarity)}$	$T_{VICA(sum)}$
600	0.0313	0.0781	0.0313
1075	0.0625	0.1094	0.07813
2000	0.5	0.625	0.5313
20000	8.9	11.09	9.3
40000	32.64	38.6	33.25
80000	125.4	147	132.8
250000	1266	1473.62	1356

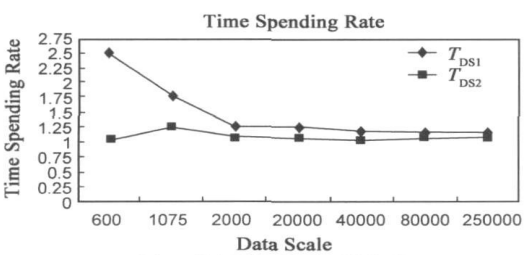

图7　算法的时间开销之比

综合考虑聚类质量和时间开销, VICA 的两种方法均优于 DBSCAN. 由于累加系数需要视具体情况而定, 当数据规模小时, 应首选相似度法, 发挥其开销小, 聚类质量高的优势; 当数据规模较大时, 可以采用累加法, 在短时间内获得相对优化的结果.

6　结束语

本文讨论了数据之间的矢量感应, 并将其应用于聚类算法. 从标量感应和方向感应两个角度设计了矢量感应函数. 提出同时考察对象间距离和方向影响的矢量感应聚类算法 VICA. 算法从对象之间的矢量感应影响出发, 将 $near(q, \sigma)$ 中的对象投影到单位圆上, 实现向量单位化, 进而考察核心对象的邻域均匀感应程度. 讨论了对象之间矢量感应的泛化函数并给出相似度法和累加法两种计算方向感应函数的方法. 理论分析和实验结果表明, VICA 能够得到数据集更好的簇划分, 明显地提高了聚类质量. 可以切实选择合适的标量感应函数和方向感应函数, 因而 VICA 具有很好的通用性和可扩展性.

将半结构化数据变换到欧式空间时, 容易出现边界稀疏对象, VICA 算法可以有效处理边界噪声, 提高聚类质量. 因此, 算法适用于大规模的高维数据集合, 也可用于半结构化数据聚类.

参考文献:

[1] Mardia K V, Jupp P. Directional Statistics(2nd Edition)[M]. Chichester, U K: John Willey and Sons Ltd, 2000.

[2] David L Donoho. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality [D]. USA: Department of Statistics Stanford University, 2000.

[3] Ester M, et al. A density—based algorithm for discovering clusters in large spatial databases with noise [A]. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining [C]. Oregon, Portland, USA: AAAI Press, 1996. 226 – 231.

[4] Xu R, et al. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645– 678.

[5] Sander J, Ester M, Kriegel H-P, Xu X. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 169– 194.

[6] Ankerst M, Breunig M M, Kriegel H-P, et al. Optics: ordering points to identify the clustering structure [A]. Proceedings of the ACM SIGMOD' 99 International Conference on Management of Data [C]. Philadelphia, Pennsylvania, USA: ACM Press, 1999. 49– 60.

[7] Xu X, Ester M, Kriegel H, Sander J. A distribution-based clustering algorithm for mining in large spatial databases [A]. Proceedings of the 14th International Conference on Data Engineering [C]. Orlando, Florida, USA: IEEE Computer Society Press,

1998. 324– 331.

[8] Hinneburg A, Keim D A. An efficient approach to clustering in multimedia databases with noise [A]. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining [C]. New York, USA: AAAI Press, 1998. 58– 65.

[9] 倪巍伟, 孙志挥, 陆介平. k-LDCHD——高维空间 k 邻域局部密度聚类算法[J]. 计算机研究与发展, 2005, 42(5): 784– 79.
NI W W, SUN Z H, LU J P. k-LDCHD—A local density based neighborhood clustering algorithm for high dimensional space [J]. Journal of Computer Research and Development, 2005, 42(5): 784– 791. (in Chinese)

[10] 周水庚, 周傲英, 等. 基于数据分区的 DBSCAN 算法[J]. 计算机研究与发展, 2000, 37(10): 1153– 1159.
ZHOU S G, ZHOU A Y, et al. A data-partitioning based DBSCAN algorithm [J]. Journal of Computer Research and Development, 2000, 37(10): 1153– 1159. (in Chinese)

[11] 马帅, 王腾蛟, 唐世渭, 等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报, 2003, 14(6): 1089– 1095.
MA S, WANG T J, TANG S W, et al. A fast clustering algorithm based on reference and density [J]. Journal of Software, 2003, 14(6): 1089– 1095. (in Chinese)

[12] 王玲, 薄列峰, 焦李成. 密度敏感的谱聚类[J]. 电子学报, 2007, 35(8): 1577– 1581.
WANG L, BO L F, JIAO L C. Density-sensitive spectral clustering[J]. Acta Electronica Sinica, 2007, 35(8): 1577– 1581. (in Chinese)

[13] 倪巍伟, 陈耿, 吴英杰, 孙志挥. 一种基于局部密度的分布式聚类算法[J]. 软件学报, 2008, 19(9), 2339– 2348.
NI W W, CHEN G, WU Y J, SUN Z H. Local density based distributed clustering algorithm[J]. Journal of Software, 2008, 19(9): 2339– 2348. (in Chinese)

[14] 谌德荣, 孙波, 陶鹏, 宫久路. 基于核光谱角余弦的高光谱图像空间邻域聚类方法[J]. 电子学报, 2008, 36(10): 1992– 1995.
CHEN D R, SUN B, TAO P, GONG J L. Clustering technology for high dimensional data based on semantics [J]. Acta Electronica Sinica, 2008, 36(10): 1992– 1995. (in Chinese)

[15] 刘铭, 王晓龙, 刘远超. 基于语义的高维数据聚类技术[J]. 电子学报, 2009, 37(5): 925– 929.
LIU M, WANG X L, LIU Y C. Clustering technology for high dimensional data based on semantics[J]. Acta Electronica Sinica, 2009, 37(5): 925– 929. (in Chinese)

[16] Halkidi M, Vazirgiannis M. Clustering validity assessment: finding the optimal partitioning of a data set [A]. Proceedings of the 2001 IEEE International Conference on Data Mining [C]. California, USA: IEEE Computer Science Press, 2001. 187– 194.

作者简介:

李雄飞 男, 教授、博士生导师. 1963 年出生于吉林省吉林市. 主要从事知识发现、数据库理论等方向的研究.
E-mail: lxf@jlu.edu.cn

孙 涛 男, 博士, 1980 年 6 月出生于吉林省长春市. 2003 年毕业于吉林大学数学学院, 2003 年起就读于吉林大学计算机科学与技术学院. 研究方向主要为: 数据挖掘相关方向.
E-mail: suntao_23@hotmail.com

武佳薇 女, 1984 年出生于山东省潍坊市. 2003 至 2009 年就读于吉林大学计算机科学与技术学院, 获得工学学士和硕士学位. 主要研究方向为: 聚类技术、分布式数据挖掘.