第 32 卷 第 16 期 计 算 机 工 程 2006 年 8 月
Vol.32 № 16 Computer Engineering August 2006
·博士论文· 文章编号：1000—3428(2006)16—0017—03 文献标识码：A 中图分类号：TP18

基于对象组特征向量的聚类与分类的实现

吴 萍[1,2]，张利萍[1]
（1. 北京理工大学计算机科学技术学院，北京 100081；2. 兰州理工大学计算机与通信学院，兰州 730050）

摘 要：高维稀疏数据的聚类分析是目前数据挖掘领域内亟待解决的问题之一。传统的聚类方法中，大部分不适用于高维稀疏数据，不能得到满意的结果。该文借助对象组相似度和对象组的特征向量，提出了一种实现聚类的方法。根据聚类结果后，根据聚类集合的上确界和下确界给出新对象的分类。该方法思想明了，实现起来简单轻松，结果准确可靠。
关键词：高维稀疏二态数据；对象组相似度；对象组特征向量；聚类；分类

Realization of Object Clustering and Classification Based on OSF

WU Ping[1,2], ZHANG Liping[1]
(1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081;
2. School of Computer and Communnication, Lanzhou University of Technology, Lanzhou 730050)

【Abstract】 The clustering for high-dimensional sparse data is one of important problem which need to be solved in the field of data mining. Most of the traditional clustering methods do not adapt to high-dimensional data. A clustering method is proposed based on set similarity (SS) and object set feature (OSF), and according to the supremum and infimum of clustering set, the new object can be distributed to different clusters. The idea of this kind method is very clear and easily implemented with reliable and exact results.
【Key words】 High-dimensional sparse binary data; Object set similarity; Object set feature; Clustering; Classification

在聚类分析的研究中，有许多问题亟待解决，尤其是针对高维稀疏数据的处理能力方面更为突出。目前，高维数据是普遍存在的一种数据形式，特别是在 Web 个性化服务的处理过程中，会经常遇到大量高维数据的处理事件。

聚类是数据挖掘领域最为常见的技术之一，已有的聚类方法有很多，如K-means、BIRCH、CURE、DBSCAN、OPTICS等。但是经过分析后，可以发现[1, 2]：

(1)低维数据对象之间相似度的计算方法不适用于高维聚类问题。有些情况下，得出的结果不能真实地反映对象间的相似程度，有时甚至得到相反的结果。

(2)频繁的数据输入/输出操作或是对数据的多次扫描，造成算法效率降低。

(3)一些算法的计算复杂度对数据属性维数的增长非常敏感。

(4)为了提高算法的效率，有些算法进行了数据压缩。有时，由于个别比较重要的属性特征被裁减，直接影响到最终的聚类结果，而且在属性裁剪过程中，不仅需要专家的介入，而且这类算法通常不具备通用性。

本文提出了一种基于高维稀疏二态数据的聚类方法，定义了对象组相似度和对象组特征向量进行对象集的数据描述和相似程度计算，进而完成对象聚类。根据聚类结果，利用聚类集合的上/下确界，可以有效而快速准确地实现新对象的分类。

1 高维稀疏二态数据

1.1 高维稀疏二态数据的定义

假设有 n 个对象，描述每个对象的属性有 m 个，m 比较大时，就成为高维数据。如果其中大部分的属性值为 0 或为空值，则称之为高维稀疏数据。进一步，如果每个对象的属性值为二态变量，即取值为{0，1}，或者即便每个对象的属性值为多态变量或是区间变量，但处理过程中，对具体的数据值不感兴趣，这时可以处理为二态变量，即进行如下的变换：

$$y_{ij}=\begin{cases}1, & \text{如果} x_{ij}\geq\delta\\ 0, & \text{如果} x_{ij}<\delta\end{cases} \tag{1}$$

而对于这类典型数据的聚类处理，称作是高维稀疏二态数据的聚类。

在个性化服务中，假设用户访问某一个功能完善的网站，我们希望能够根据用户的访问，对用户进行聚类，以便对具有相似浏览行为的用户提供兴趣提示。每个用户作为一个对象，而每个网页节点就是对象的属性。众所周知，任何用户都不可能在一次会话中把站点的所有页面都访问到。所以实际处理中，设定：如果用户访问了某页面则其属性值为 1，否则为 0。显见，这个过程就属于一个典型的高维稀疏二态数据问题。

1.2 传统二态变量处理方法

传统二态变量的处理方法主要是通过差异矩阵来实现。首先，进行属性数据的统计，也被称作是条件表的获取，如表 1。

表 1 二态变量条件

		对象 j		
		1	0	合计
对象 i	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	合计	*a+c*	*b+d*	*a+b+c+d*

作者简介：吴 萍(1972—)，女，讲师、博士，主研方向：Web 挖掘，CSCW 等；张利萍，副教授
收稿日期：2005-11-27 **E-mail：**wuping@bit.edu.cn

其次，根据条件表进行差异度的计算。基于对称二态变量计算得出的差异性：

$$d(i,j)=\frac{b+c}{a+b+c+d} \tag{2}$$

基于非对称二态变量计算得差异性，即 Jaccard 相关系数：

$$d(i,j)=\frac{b+c}{a+b+c} \tag{3}$$

2 基于对象组特征向量的聚类

2.1 对象组相似度

设有对象 n 个，每个对象具有 m 个属性，O 为其中一个对象子集，$|O|$表示对象个数，a 表示该子集中所有对象属性值均为 1 的属性个数，b 表示对象子集中的对象间对应属性值不相同的属性个数，则对应于对象子集 O 而言，对象组相似度 SS 为

$$SS(O)=\frac{a}{|O|\times(a+b)} \tag{4}$$

显见，对象组相似度表明了组内各对象之间的相似程度。

2.2 对象组的特征向量

假设有n个对象，每个对象有m个属性，O为由若干对象组成的一个对象子集，$|O|$表示该子集内的对象个数，在子集O中所有对象属性值都为 1 的属性个数记为k，$T=\{t_1, t_2, \ldots, t_k\}$，$t_i$为对应第$i$个属性值为 1 的属性在原来属性集中的标识，而在子集O中所有对象属性值都为 0 的属性个数记为l，$F=\{f_1, f_2, \ldots, f_l\}$，$f_j$为对应第$j$个属性值为 0 的属性在原来属性集中的标识。由此，定义对象组O的特征向量为四元组：

$$\overline{OSF(O)}=(T(O),F(O),|O|,SS(O)) \tag{5}$$

其中，$T(O)$为对象子集 O 中所有对象对应属性值为 1 的属性标识组成的集合；$F(O)$为对象子集 O 中所有对象对应属性值为 0 的属性标识组成的集合；$SS(O)$为对象子集 O 的对象组相似度，相应的 $a=|T(O)|$，$a+b=|\overline{F(O)}|$，则

$$SS(O)=\frac{a}{|O|\times(a+b)}=\frac{|T(O)|}{|O|\times|\overline{F(O)}|}$$

由此对于一个对象组，无须存储组中所有对象的每一个属性值，只须保存对象组的特征向量即可。另外，可以证明这种特征向量具有一个重要的特性——可加性。由此可以在对象合并时精确地计算新对象组的特征向量，并且很容易得到合并后的新对象组的组相似度。

2.3 对象组的特征向量的可加性

定义 1 对象组的特征向量的加法。

对象子集 X 和 Y 满足 $X\cap Y=\Phi$，有

$$OSF(X)+OSF(Y)=(T,\overline{F},N,SS)$$

其中：

$$T=T(X)\cap T(Y),\quad \overline{F}=\overline{F(X)\cap F(Y)}$$

$N=|X|+|Y|$，$SD=|F|/(N\times|T|)$。

定理 1 对象组的特征向量的可加性。

假设有 n 个对象，每个对象有 m 个属性，X 和 Y 为两个不相交的对象子集，即满足 $X\cap Y=\Phi$，$X\cup Y$ 为集合 X 和 Y 的合集，有

$$OSF(X\cup Y)=OSF(X)+OSF(Y) \tag{6}$$

证明 (1)第 1 个分量：$T(X\cup Y)$ 与 T 要证明集合 $T(X\cup Y)$ 与集合 $T(X)\cap T(Y)$ 的等价，可以依次证明二者互为对方的子集。

因为 $T(X\cup Y)$ 中的元素是集合 $X\cup Y$ 中所有对象的对应的属性值为 1，$\forall t\in T(X\cup Y)$，则说明 $X\cup Y$ 中所有对象的第 t 个属性的属性值为 1。显而易见，$t\in T(X)$，且 $t\in T(Y)$，即 $t\in T(X)\cap T(Y)$，从而说明 $T(X\cup Y)\subseteq(T(X)\cap T(Y))$。

又因为 $T(X)$ 和 $T(Y)$ 中的元素分别是集合 X 和 Y 中所有对象的对应的属性值为 1，则 $\forall t\in T(X)\cap T(Y)$，同时也表明集合 $X\cup Y$ 中所有对象的第 t 个属性的属性值为 1，即 $t\in T(X\cup Y)$，所以 $(T(X)\cap T(Y))\subseteq T(X\cup Y)$。

综上，可以得出 $T(X\cup Y)=T$。

(2)第 2 个分量：$\overline{F(X\cup Y)}$ 与 $\overline{F}$。

类似于第 1 个分量的证明，可以得出

$$\overline{F(X\cup Y)}=\overline{F(X)\cap F(Y)}$$

由此可得 $\overline{F(X\cup Y)}=\overline{F}$。

(3)第 3 个分量：$|X\cup Y|$ 与 N。

已知集合 X 和 Y 为不相交的两个子集，则有

$$|X\cup Y|=|X|+|Y|=N$$

(4)第 4 个分量：$SS(X\cup Y)$ 与 $SS=\frac{|T|}{N\times|F|}$

根据前面 3 个分量的等价性证明，可以得出

$$SS(X\cup Y)=SS$$

综上，因为向量 $OSF(X)+OSF(Y)$ 和向量 $OSF(X\cup Y)$ 中对应分量均被证明是等价的，则

$$OSF(X\cup Y)=OSF(X)+OSF(Y)$$

由此，可以在不相交子集合并后，通过这个特点精确地计算出大子集的组特征向量，从而降低了计算量和数据存储量。

3 算法描述

假设 n 个对象，共有 m 种属性，每个对象的属性值为二态变量或是经过二态化处理，给出对象组相似度的最小阈值为 θ。下面，详细描述了聚类算法的处理过程：

(1)进行数据二态化预处理。如果本身就是二态数据，则直接从下步开始。

(2)以每个对象为元素建立一个只包含自身的组，记作 X_i，$i\in\{1,2,\ldots,n\}$。

(3)根据特征向量的可加性，计算

$$OSF(X_1\cup X_2)=OSF(X_1)+OSF(X_2)$$

1) $SS(X_1\cup X_2)\geq\theta$，则二者合并为新组 X'_1；

2) $SS(X_1\cup X_2)\prec\theta$，则不合并，重新标识为 X'_1 和 X'_2。

记录聚类个数记为 num。

(4)依次处理 X_i，$i\in\{3,4,\ldots,n\}$，计算

$$OSF(X_i\cup X'_k)=OSF(X_i)+OSF(X'_k),\ k\in\{1,2,\ldots,num\}$$

1) $SS(X_i\cup X'_k)\geq\theta$，取 $\max(SS(X_i\cup X'_k))$ 为新组的相似度，合并新组 X'_k；

2) $SS(X_i\cup X'_k)\prec\theta$，则 X_i 重新标识为 X'_{num+1}。

对聚类个数进行累加 $num=num+1$。

(5)新集合 $X'=\{X'_k\mid k\in\{1,2,\ldots,num\}\}$。

(6)如果 $|X'_k|\leq\varepsilon$，则把 X'_k 从 X' 中删减掉，并对相应的 num 进行更新。

4 新对象的分类问题

4.1 相关概念的数据定义

任意实数 a，b 属于实数集 R，存在以下次序关系，即有如下性质成立：自反性：$a\leq a$，$b\leq b$；反对称性：

$a \le b, b \le a \Rightarrow a = b$ ；传递性： $a \le b, b \le c \Rightarrow a \le c$ 。

定义 2 半序关系，半序集。

如果一个集合 X 中的元素之间满足上述 3 个性质，称为半序关系，现记为 $\le$ 。而集合 X 称为半序集，可以表示为 $(X, \le)$ 。显然，子集关系“ $\subseteq$ ”满足前述 3 个性质，是一种半序关系。

定义 3 上界和上确界。

设集合 $X \subset R$ ，常数 $m \in R$ ，集合 X 中的任意元素 a 都满足半序关系 $a \le m$ ，则 m 是集合 X 的一个上界。若 m 是所有上界中最小的一个，则为上确界，记作 $m = \sup(X)$ 。

定义 4 下界和下确界。

设集合 $X \subset R$ ，常数 $n \in R$ ，集合 X 中的任意元素 a 都满足半序关系 $a \le n$ ，则称 n 是集合 X 的一个下界。若 n 是所有下界中最大的一个，则为下确界，记作 $n = \inf(X)$ 。

定义 5 幂集。

给定集合 X，由 X 的所有子集为元素构成的集合，称作是 X 的幂集，记为 $\Omega(X)$ 。例如，假设集合 $X=\{a, b, c\}$，则 $\Omega(X)=\{\Phi, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$ 。而（$\Omega(X)$，$\subseteq$）为半序集。

4.2 聚类集合的确界表示

给定高维稀疏数据中有 n 个对象，对象集合

$Obj = \{Obj_i | i \in \{1,2,...,n\}\}$

属性标识集合

$Lab = \{1, 2, \ldots, m\}$

将属性值为 1 的属性标识组合成集合 $T(Obj_i), i \in \{1,2,...,n\}$ ；聚类后，假设得到 k 个聚类 $C = \{C_j | j \in \{1,2,...,k\}\}$ ；每个类中所包含所有对象的集合

$O_j = \{T(Obj_l) | Obj_l \in C_j, l \in \{1,2,...,n\}, j \in \{1,2,...,k\}\}$

进而，可以得出以下结论：

(1)($\Omega(X)$，$\subseteq$)为半序集；

(2) $T(Obj_i) \subseteq \Omega(Lab)$ ；

(3)由上/下确界的定义可以得到

$\sup(O_j) = \bigcup_{T(Obj_l) \in O_j} T(Obj_l)$ 和

$\inf(O_j) = \bigcap_{T(Obj_l) \in O_j} T(Obj_l)$ ，且

$\sup(O_j) \subseteq \Omega(Lab)$ ， $\inf(O_j) \subseteq \Omega(Lab)$

(4) $O_j \subseteq \Omega(Lab)$ 。

用 $OSF(O_j) = \{T(O_j), F(O_j), |O_j|, SD(O_j)\}$ 描述每个聚类 O_j 的信息特征。结合各项定义，可以得出

$$T(O_j) = \bigcap_{T(Obj_l) \in O_j} T(Obj_l) = \inf(O_j) \tag{7}$$

$$T(O_j) \cup F(O_j) = \bigcup_{T(Obj_l) \in O_j} T(Obj_l) = \sup(O_j) \tag{8}$$

由此将聚类分析使用集合的确界表示，得出各个集合的最大和最小子集，利用聚类的上/下确界进行新对象的分类。

4.3 新对象的分类算法

设聚类集合 $X' = \{X'_k | k \in \{1,2,...,num\}\}$ ，对于新对象 Y，若要将 Y 分配到以上得出的聚类中，具体的分类过程如下：

(1)按照 $|T(X'_k)|$ 降序， $|\overline{F(X'_k)}|$ 升序， $|X'_k|$ 升序对 X'_k 进行排序，将排序结果对应记作 $C_i, i \in \{1,2,...,num\}$ 。

(2)判断 $\inf(C_i) \subseteq T(Y) \subseteq \sup(C_i)$ ， $i \in \{1,2,...,num\}$ 。

1)如果此包含关系成立，则 Y 可以归并到该 C_i 组中；

2)反之，Y 不能归并到任何组，成为孤立对象。

5 具体实例

5.1 实例描述

表 2 假设了某一网站服务器日志中用户的访问情况，而且这是一组经过用户识别后得出的结果。其中属性值为 1 表示用户访问了该页面，反之表示用户没有访问该页面。

表 2 用户对站点页面的访问情况

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
U1	1	1	0	0	0	1	1	1	0	0
U2	1	1	0	1	0	1	1	1	0	0
U3	1	0	1	0	1	1	1	0	0	1
U4	1	0	1	1	1	1	0	0	1	0
U5	1	1	1	1	1	0	0	0	1	1
U6	1	0	0	1	0	1	1	1	1	0
U7	1	0	1	0	0	1	0	1	1	1
U8	1	0	1	0	0	0	0	0	0	0

取阈值 θ=0.25，并有新对象 U9 和 U10，其中 U9 和 U10 访问的页面标识集合分别为{1，4，6，8}和{1，2，6，8}。整个处理过程如下：

(1)只显示属性值为 1 的属性标识(见表 3)。

表 3 对象在网站上的访问信息

对象	浏览页面的标识
U1	1，2，6，7，8
U2	1，2，4，6，7，8
U3	1，3，5，6，7，10
U4	1，3，4，5，6，9
U5	1，2，3，4，5，9，10
U6	1，4，6，7，8，9
U7	1，3，6，8，9，10
U8	1，3

(2)聚类得出如结果表 4。

表 4 聚类结果

组	组成员	T	$\overline{F}$	SS
X'_1	U1, U2, U6	{1, 4, 6, 7, 8}	{1, 2, 4, 6, 7, 8, 9}	0.412
X'_2	U3, U4, U5, U7	{1, 3, 6, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	0.25
X'_3	U8	{1, 3}		0

因为 X'_3 只包含 U3 一个对象，将其删除。由此得出两个聚类 $X' = \{X'_1, X'_2\}$ 。

(3)按照 $|T(X'_k)|$ 降序， $|\overline{F(X'_k)}|$ 升序， $|X'_k|$ 升序对 X'_k 进行排序，将排序结果对应记作 $C_i, i \in \{1,2\}$ ，如表 5 所示。

表 5 排序结果

组	组成员	T	$\overline{F}$	N
C_1	U1，U2，U6	{1，4，6，7，8}	{1，2，4，6，7，8，9}	3
C_2	U3，U4，U5，U7	{1，3，6，9，10}	{1，2，3，4，5，6，7，8，9，10}	4

(4)新对象 U9 和 U10，判断 $\inf(C_i) \subseteq T(X_k) \subseteq \sup(C_i)$ 是否成立，其中 $i \in \{1,2\}$ ， $k \in \{9,10\}$ 。可以得出，U9 归并到 C_1 中，而 U10 不可以归并到 C_1 或 C_2 中，成为孤立对象。

5.2 结果分析

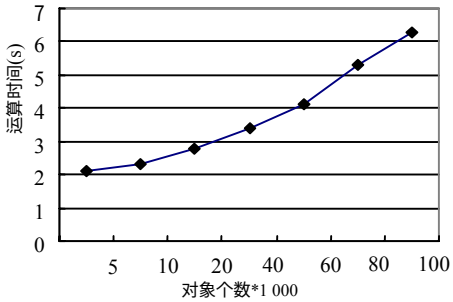

图 1 各种聚类算法中对象个数与运算时间的关系

(下转第 57 页)