

基于二维属性的高维数据聚类算法研究

游芳,姜建国,张坤

(西安电子科技大学,陕西 西安 710071)

摘要:聚类就是按照一定的要求和规律对事物进行区分和分类的过程。在这一过程中没有任何关于类别的先验知识,也没有教师的指导,仅靠事物间的相似性作为类属划分的准则,因此属于无监督分类的范畴。聚类分析则是指用数学的方法研究和处理给定对象的分类。针对目前很多聚类算法只能对低维数据进行聚类的问题,提出了一种改进的相异度量方法对二维属性的高维数据采用层次分裂算法进行聚类,而且根据用户指定的参数聚类,并对传统相异度量度和改进的相异度量度方法的聚类结果进行比较,发现改进的相异度量度方法更适用于二维属性的高维数据的聚类。

关键词:聚类;层次分裂算法;高维数据;相异度

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2009)05-0111-03

Cluster - Algorithm Studies Based on Two - Dimensional Attribute Higher - Dimension Data

YOU Fang, JIANG Jian-guo, ZHANG Kun

(Xidian University, Xi'an 710071, China)

Abstract: Clustering is a process that in accordance with certain requirements and rules of conduct to distinguish between things and classification. In this process there is no priori knowledge on the type nor the guidance of teachers, alone among things of a type similar to a breakdown criteria, are owned to classified areas with no supervision. Cluster analysis refers to research and mathematical methods to deal with the classification of objects. Problem specifically for many cluster - algorithm can only carry out clustering on low thinking data, brought forward one kind adopt arrangement of ideas to divisive analysis cluster - algorithm carrying out clustering to two - dimensional attribute higher - dimension data, and be based on the parameter clustering that the consumer allocates, make comparison on the traditional degree of amounts and the improved degree of amounts method clustering result then discover the improved degree of amounts method improvement applies to two - dimensional attribute higher - dimension data clustering more.

Key words: clustering; divisive analysis cluster - algorithm; higher - dimension data; degree of amounts

0 引言

聚类(Clustering)问题起源于包括机器学习、统计学、数据挖掘、空间数据库技术、生物学及市场营销等多门学科,并在许多领域得到了广泛的应用,例如数据压缩、信息检索、模式识别,尤其是数据挖掘^[1],目前,聚类已经成为该领域非常活跃的研究课题^[2]。从已有的聚类算法应用来看,一个好的聚类算法应具有如下几个特点:能完成对任意形状数据集的聚类;能够处理噪声和孤立点;聚类结果不受输入数据顺序的影响;不需预先给定聚类数目。

已有的聚类算法大致可以分成以下几类:基于划

分的方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法等^[3]。

上述各有所长的算法也分别存在不同的问题,但是有一个普遍的不足之处是:这些算法只适用于低维数据空间。高维数据空间中由于存在维度效应,使得以这些算法无法充分发挥其性能。其原因可以概括为:这些算法所采用的相异度量——聚类的基石在高维数据空间失去了意义,而不是算法的聚类策略存在问题^[4]。

以人体体征数据为聚类依据,而人体体征数据采用二元属性,体征数据的差别以两个个体间不同指标值的数目进行度量,即有几种体征项的指标值不同。针对二元属性的高维数据的相异度提出了新的思想,并使用层次分裂聚类算法对数据进行聚类,并对结果进行了比较,发现新的相异度更加适用于层次分裂聚类算法。

收稿日期:2008-08-04

作者简介:游芳(1985-),女,江西南城人,硕士研究生,研究方向为计算机图形图像处理、网络安全等;姜建国,硕士生导师,教授,研究方向为图形图像处理、网络安全等。

1 相异度度量

1.1 基本数据结构

本节讨论在聚类分析中经常出现的数据类型,以及如何对其进行预处理。大多数聚类算法选择如下两种有代表性的数据结构:

1) 数据矩阵^[5]。

数据矩阵是一种对象与变量结构。它用 P 个变量(也称为度量或属性,即 p 维)来表现 n 个对象。这种数据结构是关系表的形式,或者可以看成是 $n \times p$ 的矩阵,其中每一行为一个向量,代表一个数据对象。

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

2) 相异度矩阵^[6]。

相异度矩阵是一种对象—对象结构。存储 n 个对象两两之间的近似性,表现为一个 $n \times n$ 的对称矩阵。

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

其中 $d(i, j)$ 是对象 i 和 j 之间相似(异)性的量化表示,通常为一个非负的数值。当对象 i 和 j 越相似或越“接近”,其值就越大(越接近 0);反之其值越接近 0(大)。因为 $d(i, j) = d(j, i)$,而且 $d(i, i) = 1$ (0),因此有形如上式的对称矩阵。由于相似度与相异度对于聚类而言是同等度量,因此以下将统一使用相似度这个概念来同时表示相异度或距离。

许多算法以相异度矩阵为基础,如果原始数据是以数据矩阵的形式表现的,使用算法之前需要将数据矩阵经过一定的计算转换成为相异度矩阵^[7]。

1.2 基本相异度度量

传统的相异度度量有两种方法:一种是距离度量;另一种是相似系数。使用距离度量时,往往将数据对象看成是多维空间中的一个点(向量),并在空间中定义点与点之间的距离^[8]。

对象之间的相异度计算涉及到描述对象的变量类型。常见的变量类型有区间标度变量、二元变量、标称、序数和比例变量,以及这些类型的组合型。不同类型变量表示的数据对象之间的相异度计算方法不尽相同^[7]。

1.3 高维数据的相异度度量

根据相异度的意义,两个对象间的相异度就是它们之间的差别的数值表示。以人体体征数据聚类为依

据,而人体体征数据的差别应该以两个个体间不同指标值的数目进行度量,即有几种体征项的指标值不同。因此,文中以二进制按位与的方法进行相异度计算,具体公式如式(1)。

$$d(x_i, x_j) = 1 - \text{Count}(x_i \wedge x_j) / \text{validBitLen} \quad (1)$$

其中,Count(num)是数据 num 为二进制数值所有位中“1”的数目,validBitLen 是 num 的有效二进制位数。从上式看出相异度的值是一个 0 到 1 的数,值越大,两组间的距离越大。对数据表选用不同的相异度计算方法所得的相异度矩阵分别保存为 Diff1 和 Diff2。Diff1 是使用文中设计的相异度矩阵,Diff2 是采用欧几里德距离度量的相异度矩阵。

下节中将两种相异度度量方法所得的矩阵与分裂层次聚类算法相结合,分别得出两种聚类结果,通过对聚类质量的分析,使用本文设计的新的度量方法提高了聚类质量。

2 分裂层次聚类算法描述与实现

为使聚类结果尽量客观,应选取控制参数少、参数设置简单的算法。文中选取基于分裂的层次聚类算法^[9]。

它采用一种自顶向下的策略,首先将所有对象置于一个簇中,然后逐渐细分为越来越小的簇,直到每个对象自成一簇,或者达到了某个终结条件,例如达到了某个希望的簇数目,或者两个最近簇之间的距离超过了某个阈值^[10]。

该算法的输入为用户认为可能的簇数目,同时,它使用下面两种测度方法:

1) 簇的直径:在一个簇中的任意两个数据点都有一个欧氏距离(式 2),

$$d(x_i, x_j) = \|x_i - x_j\| = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^2 \right]^{1/2} \quad (2)$$

这些距离中的最大值是簇的直径。

2) 平均相异度(平均距离)(式 3):

$$d_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} |x - y| \quad (3)$$

算法的描述如下^[11]:

输入:包含 n 个对象的数据库,终止条件簇的数目 k

输出: k 个簇,达到终止条件规定簇数目

1) 将所有对象整个当成一个初始簇

2) FOR($i = 1; i \neq k; i++$) DO BEGIN

3) 在所有簇中挑出具有最大直径的簇

4) 找出所挑中簇里与其他点平均相异度最大的一个点放入 group,剩余的放入 old part 中

5) REPEAT

6)在 old part 里找出到 group 中点的最近距离不大于到 old part 中点的最近距离的点,并将该点加入 group

7)UNTIL 没有新的 old part 的点被分配给 group

8)group 和 old part 为被选中的簇分为的两个簇,与其他簇一起组成新的簇集合

9)END

在实现过程中,需要定义一个称为 Cluster 的数据结构存储分裂所得的中间结果和最终结果,其定义如下:

```
Cluster {
  整数 ID; //簇的标号
  整数 num; //簇中元素的总数目
  实数 diameter; //簇的直径
  Element * head; //存储该簇元素的链表的头指针
}
```

其中 Element 是自定义的表示数据元素的结构类型,其定义如下:

```
Element {
  整数 ID; //该元素对应的登录号
  Element * next; //链表指向下一元素的指针
}
```

3 两种相异度聚类结果的比较

对于层次分裂算法,文中分别针对 Diff1 和 Diff2 进行聚类,从得出的结果可以看出,文中设计的针对二维数据的相异度量方法(Diff1)更适用于对体征数据的处理。

对于聚类结果的评价,从数据挖掘角度来看,有大量不同的方法来衡量聚类的质量,大多都着重于两个方面:每个聚类应该是紧凑的;各个聚类间的距离应该尽可能的远。实现这种直观概念的一种直接方法就是观察聚类 C 的聚类内差异和聚类间差异。聚类内差异衡量了聚类的紧凑性或密集度,而聚类间差异衡量了不同聚类间的距离。

通过簇的直径表示聚类内差异,如式(4)。表示聚类间差异首先要找到该簇的“中心点”,即平均相异度最小的数据点,两个簇的中心点之间的相异度即为两簇的聚类间差异,如式(5)。

$$D_{\text{inner}}(C) = \max(d(x_i, x_j)), \forall x_i, x_j \in C \text{ 且 } i \neq j \quad (4)$$

$$D_{\text{inter}}(C_1, C_2) = \min(d(p_1, p_2)), \forall p_1 \in \text{Center}(C_1), \forall p_2 \in \text{Center}(C_2), \text{Center}(C) = \{x \mid d_{\text{avg}}(x, C) = \min(d_{\text{avg}}(x_i, C)), \forall x_i \in C\} \quad (5)$$

根据上面的定义分别计算所得的不同聚类结果聚类内差异和聚类间差异如表1和表2所示。簇直径最大(小)值是 k 个结果簇的直径中的最大(小)值,簇直径

平均值是 k 个簇直径的算术平均值。簇间差异最大(小)值是 $k!(k \text{ 的阶乘})$ 个差异值中的最大(小)值,平均值是这些差异值的算术平均值。

表1 Diff1 作为输入时不同聚类结果聚类内差异和聚类间差异

k	5	6	7	9
簇直径最大值	0.852	0.843	0.835	0.820
簇直径最小值	0.543	0.541	0.536	0.532
簇直径平均值	0.675	0.653	0.634	0.625
聚类间差异最大值	0.672	0.687	0.715	0.763
聚类间差异最小值	0.356	0.362	0.387	0.402
聚类间差异平均值	0.468	0.485	0.552	0.684

表2 Diff2 作为输入时不同聚类结果聚类内差异和聚类间差异

k	5	6	7	9
簇直径最大值	4.52	4.43	4.35	4.21
簇直径最小值	2.32	2.41	2.30	2.85
簇直径平均值	3.65	3.52	3.41	3.36
聚类间差异最大值	5.58	5.69	6.13	6.56
聚类间差异最小值	4.02	4.34	4.32	4.46
聚类间差异平均值	4.87	4.98	5.21	5.43

根据分别对表1和表2中各个簇的质量指标的比较,可以看出,随着 k 值的增大,聚类间差异在不同程度上变大。根据上表的数据比较,文中认为当 k 值为9时聚类质量最高。同时,对于 Diff1 矩阵,由于其计算公式中已对数值进行了归一化,所有数值都在0和1之间,因此其质量指标的值也是归一化的值,这在比较过程中有利于比较。因此,Diff1 比 Diff2 更适合作为体征聚类的输入,即文中提出的相异度量方法提高了体征数据聚类的质量。

4 结束语

提出了一种改进的相异度量方法对二维属性的高维数据采用层次分裂算法进行聚类,而且根据用户指定的参数聚类,并对传统相异度度量和改进的相异度度量方法的聚类结果进行比较,发现改进的相异度度量方法更适用于二维属性的高维数据的聚类。

参考文献:

- [1] 王 预. 数据仓库与数据挖掘的关系及其安全性问题[J]. 计算机技术与发展, 2008, 18(5): 144-146.
- [2] Hearst M A, Pedersen J O. Reexamining the cluster hypothesis: Scatter/gathe on retrieval results[C]//Frei H P, Harman D, Schauble P, et al. Proc. of the 19th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Re-

(下转第118页)

现高效、高速的构件功能检索过程。同时,研究过程中也发现,基于本体的构件检索,后续工作还有很多,例如本体表示汉化能力,以及对复杂数据结构如 Functor 的推理等。

构件功能检索演示系统 Powered by yw		
检索		
检索结果		
构件信息		
检索项: 构件功能	检索值: register	
构件名称	构件版本	主要功能
RegisterInstance42	v2.0	http://www.owl-ontologies.com/ComponentOntology.owl#register
RegisterInstance23	v1.0	http://www.owl-ontologies.com/ComponentOntology.owl#register
RegisterInstance32	v2.0	http://www.owl-ontologies.com/ComponentOntology.owl#register
RegisterInstance31	v2.0	http://www.owl-ontologies.com/ComponentOntology.owl#register
RegisterInstance21	v1.0	http://www.owl-ontologies.com/ComponentOntology.owl#register
RegisterAndLoginInstance4	v3.0	http://www.owl-ontologies.com/ComponentOntology.owl#register http://www.owl-ontologies.com/ComponentOntology.owl#login
RegisterInstance22	v1.0	http://www.owl-ontologies.com/ComponentOntology.owl#register

图 4 具有 register 功能的构件列表

参考文献:

- [1] Gruber T R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing[J]. Information Journal of Human Computer Studies, 1995, 43: 907-928.
- [2] Gibb F, McCartan C, O'Donnell R, et al. The Integration of Information Retrieval Techniques within a Software Reuse Environment[J]. Journal of Information Science, 2000, 26(4): 520-539.
- [3] 徐正权. 软件复用方法与技术[M]. 武汉: 华中理工大学出版社, 1998.
- [4] Musen M A. Dimensions of Knowledge Sharing and Reuse[J]. Computers and Biomedical Research, 1992, 25: 435-467.
- [5] Stuckenschmidt H. Ontology-Based Information Sharing in Weakly Structured Environments [D]. Amsterdam: AI Department, Vrije Universiteit Amsterdam, 2002.
- [6] Doan A, Madhavan J, Domingos P, et al. Learning to map between ontologies on the Semantic Web [C]//In 11th International World Wide Web Conference. Honolulu, USA: [s. n.], 2002.
- [7] Baader F, Sattler U. Description logics with aggregates and concrete domains [J]. Information Systems, 2003, 28(8): 979-1004.
- [8] Zoe L. Web data retrieval and extraction[J]. Data and Knowledge Engineering, 2003, 44(3): 347-367.
- [9] 李选如, 何洁月. 语义集成: 本体映射方法研究[J]. 计算机技术与发展, 2007, 17(2): 127-130.
- [10] 黄烟波, 张红宇, 李建华, 等. 本体映射方法研究[J]. 计算机工程与应用, 2005(18): 33-35.

(上接第 110 页)

- [1] Proceedings of the SIGCOMM Workshop on Delay Tolerant Networking. [s. l.]: [s. n.], 2005.
- [2] Leguay J, Friedman T, Conan V. DTN Routing in a Mobility Pattern Space[C]//In proc. ACM SIGCOMM 05 Workshop on Delay Tolerant Networking and Related Topics (WDTN-05). [s. l.]: [s. n.], 2005.
- [3] Mukarram M, Tariq B, Ammar M H, et al. Message ferry route design for sparse ad hoc networks with mobile nodes[C]//Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing. [s. l.]: [s. n.], 2006: 37-48.
- [4] 王行甫. 一种 DTN 节点自定位方案及其性能分析[J]. 中国科学院研究生院学报, 2008, 25(3): 367-371.
- [5] Cook W, Applegate D, Bixby R, et al. Concorde: A code for solution of Travelling Salesman problem [J/OL]. 2005. http://www.tsp.gatech.edu/.

(上接第 113 页)

- [1] trieval. [s. l.]: ACM Press, 1996: 76-80.
- [2] 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究, 2007(1): 10-13.
- [3] 谢立宏. 面向高维数据的聚类算法研究[D]. 武汉: 武汉大学, 2002: 10-12, 29-30.
- [4] Makhoul J, Roucos S, Gish H. Vector quantization in speech coding[J]. Proc. of the IEEE, 1985, 7: 1551-1556.
- [5] Karyapis G, Han E H, Kumar V. Chameleon: A hierarchical clustering algorithm using dynamic modeling[J]. IEEE Computer, 1999, 32(8): 68-71.
- [6] 李雄飞, 李军. 数据挖掘与知识发现[M]. 北京: 高等教育出版社, 2003: 11-19.
- [7] 毛国君. 数据挖掘原理与算法[M]. 北京: 清华大学出版社, 2005: 171-175.
- [8] 朱明. 数据挖掘[M]. 合肥: 中国科学技术大学出版社, 2002: 22-32.
- [9] 李秀芳, 李志成. 基于数据挖掘的聚类算法研究[J]. 计算机技术与自动化, 2006(3): 1-2.
- [10] 段明秀, 杨路明. 对层次聚类算法的改进[J]. 湖南理工大学学报, 2008(2): 1-2.