文章编号:1001 - 2486(2006)01 - 0077 - 04

多媒体数据挖掘中数据间的相似性度量研究*

贺 玲,吴玲达 蔡益朝,谢毓湘,雷 震
(国防科技大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:聚类是多媒体数据挖掘的重要任务之一,数据之间的相似性度量是聚类的基础和前提。多媒体数据的特征矢量通常都是数十维甚至数百维的,但传统的相似度量方式一般只适用于低维数据。在分析高维数据特性的基础上,提出了一个新的度量方式。通过使用一个特定的策略对原始数据空间进行网格划分,该方法较好地避免了噪声数据对高维数据相似性度量的影响。实验证明此方法是有效的。

关键词:多媒体数据挖掘;维度灾难;相似度量
中图分类号:TP311 文献标识码:A

Research on Similarity Measurement in Multimedia Data Mining

HE Ling, WU Ling-da, CAI Yi-chao, XIE Yu-xiang, LEI Zhen
(College of Information System and Management, National Univ. of defense Technology, Changsha 410073, China)

Abstract: Clustering is one of the focused problems in multimedia data mining, and similarity measurement among data is fundamental to clustering. In multimedia data clustering, the corresponding vector features are always of high dimensionality. Most traditional measurement methods, however, are only efficient for low dimensional data. This paper, based on an analysis of general characteristics of data presented in high dimensional spaces, proposes a new similarity measurement for multimedia data mining. It used a special strategy to split the original data space before computing the similarity among data points, thus efficiently avoiding the influence of noisy data in high dimensional dimensional spaces. Experiments show that the new method presented is effective.

Key words: multimedia data mining; curse of dimensionality; similarity measurement

在多媒体数据挖掘的研究中,聚类分析的应用最为广泛。所谓聚类,就是把一组个体按照相似性归成若干类别,使得同一类的个体之间的相似度尽可能大、不同类个体之间的相似度可能小。由此可见,数据之间的相似性度量是聚类的重要依据。多媒体数据聚类所处理的目标数据通常是媒体数据对应的特征向量,这些特征向量往往是数十维甚至数百维的。如果仍然用适用于低维数据的相似性度量方式来处理这些高维的特征数据,将得不到理想的结果,这就是所谓的"维度灾难"[1]。

为克服维度灾难的影响,目前在高维数据的索引结构方面有很多研究者提出了很多经典的算法,但是对高维数据之间相似性度量的研究还不是很多。例如文献[2]中提出了一种处理高维数据的方法,它将原始数据空间向各个不同的子空间的组合进行投影。这种方法是目前处理高维数据时最常用的方式之一,它虽然从一定程度上降低了数据的维度,但是这些子空间的组合数目随着维数的增高呈指数级增长,这使得用户几乎不可能列举出所有可能的组合方式,从而很难得到一个最优的聚类结果。

针对上述问题,本文提出了一种基于网格划分的、可用于高维数据的相似性度量方法(相似性度量有两种表现形式,一种是用数据之间的差异度(即距离)来表现,另外一种是用数据之间的相似度来表现,本文的研究针对的是前者)。

1 维度灾难

"维度灾难"这一术语由 Bellman 首次提出,它泛指在数据分析中遇到的由于变量(属性)过多而引起

* 收稿日期:2005 - 10 - 20
基金项目:国家自然科学基金资助项目(60473117)
作者简介:贺玲(1976—),女,博士生。

的一系列问题。

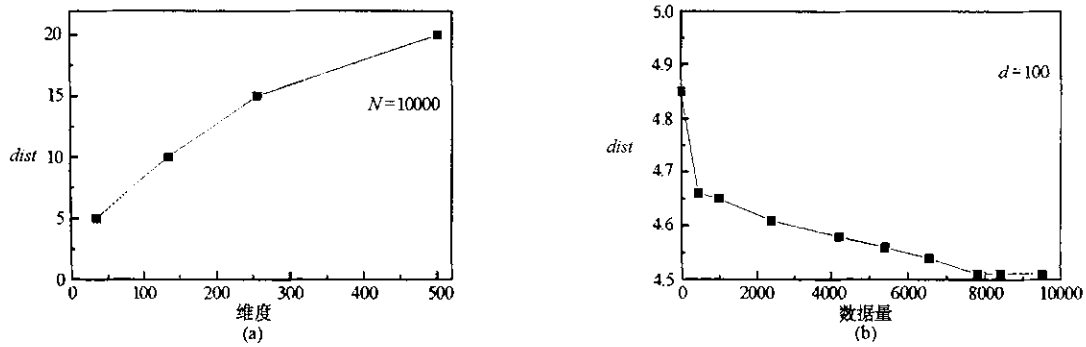


图1 维度灾难示意图

Fig.1 Curse of dimensionality

设一个含有 N 个数据的 d 维单位超球，采用欧氏距离读出的最近邻距离 $dist$ 可以表达为 $dist = 2\left[\frac{d\Gamma(d/2)}{2\pi^{d/2}N}\right]^{1/d}$[3]。图1显示了 $dist$、N 和 d 维数的关系[4]。从图1(a)中可以看出，最近邻距离随着维数的增长呈线性增长趋势。图1(b)则表明，对于一个给定的维数 d，包含最近邻点的最近邻距离随着 N 的增长而仅仅减小 $N^{1/d}$。也就是说，即使是寻求一个近邻点的最近邻查询，相对来说也需要遍历比较大的数据量才能使其最近邻距离较小。

由此可见，虽然欧氏度量是目前效果较好、应用最为广泛的度量方式，但是，由于维度灾难的影响，最近邻概念在这种度量下失去了应有的意义[5]。这是因为在高维空间中，即使是最为相似的两个记录，几乎总存在着一些维，在这些维上，这两个记录的值差别较大，但在欧氏度量中占主导地位的就是这些维。这样，两个记录的相似信息就被淹没在这少数的几个维中，而这些维多是噪声信息。

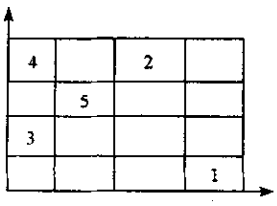
2 一种新的相似度量方式

为更好地克服维度灾难的影响，本文提出了一个新的思路来度量高维数据之间的相似性，即先将高维数据空间按一定规则进行划分，以形成数据空间的网格结构，从而在度量两个数据之间的相似性时，只考虑它们落入相同的网格中的维度信息。

2.1 基于网格划分的相似性度量

划分策略可简单描述如下：对于 d 维的数据空间，首先为其每维指定一个划分位数 b_i，于是该维就被划分为 2^{b_i} 个单元。设 $b = \sum_{i=1}^{d} b_i$，那么整个数据空间被划分为 2^b 个单元。

图2所示是一个二维数据集的划分，数据空间的每维都分配2位，并被均匀地分割成4个部分。这样，在把各维划分成若干区间的同时，就自动创建了一个网格结构，从而原始数据空间中的数据点的每一维都可以对应于一个具体的网格区域。



	原始向量	对应单元
1	(0.9, 0.1)	(11 00)
2	(0.6, 0.8)	(10 11)
3	(0.1, 0.4)	(00 01)
4	(0.2, 0.9)	(00 11)
5	(0.3, 0.7)	(01 10)

图2 二维数据空间的网格划分

Fig.2 Grid Splitting of 2-dimensional data space

设数据集中任意两点为 $Q = (q_1, q_2, \cdots, q_d)$，$X = (x_1, x_2, \cdots, x_d)$，$S[X, Q]$ 为 X 与 Q 的各属性值落入同一区间的那些维的集合，那么 X 与 Q 之间的距离 $D(X, Q) = \left[\sum_{i \in S[X,Q]} \| x_i - q_i \|^2\right]^{1/2}$。

在一些特定的应用领域,当高维数据的各维属性值对数据的相似性度量所产生的影响各不相同时,可以将上式泛化为 $D(X,Q)=\left[\sum_{i\in S[X,Q]} w_i \| x_i - q_i \|^2\right]^{1/2}$,其中 w_i 为经验权值。

该度量方式与欧氏度量的一个重要差别在于,在该函数中占主导地位的是那些 X 与 Q 之间差别较小的维,而且,它们接近的维数越多,其之间的相似性也越高。这显然是符合人们判定数据点之间相似性的习惯的。

2.2　针对随机产生的高维数据的实验及结论

Beyer 在文献[5]中将 $\frac{D_{\max d}-D_{\min d}}{D_{\min d}}$(此处记作 v),即最远距离与最近距离之差与最近距离的比值,作为衡量度量函数是否有意义的准则。若该值随着数据维数的升高以一个常数概率趋向于 0,就说明最远和最近距离的差异不明显,同时也表明与其相对应的度量方式不合理。本文也首先以此表达式作为评判准则,分别随机生成 100 条 50 维、100 维、150 维、200 维、300 维以及 400 维的记录,从而得出与本文度量方式相对应的 v 值随维数 d 变化的曲线图。如图 3 所示。

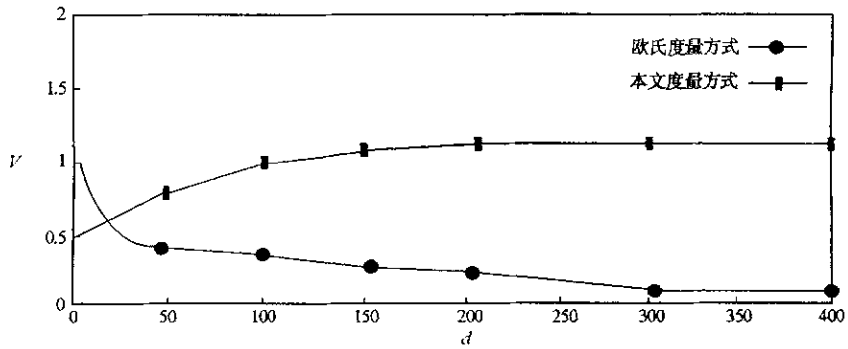


图 3　最远距离和最近距离之差与最近距离的比值随维数变化曲线

Fig.3　Change curve of V with dimensionality

从图 3 可以看出,对于这些随机产生的数据,利用本文提出的度量方式计算所得到的 v 值没有随着数据维度的增高而降低,但直接通过欧氏距离公式计算所得的 v 值却随着数据维数的增高急剧减小。因此,从对最近邻与最远邻的区分能力上来讲,本文提出的高维度量方式要明显优于欧氏度量,它从一定程度上克服了"维度灾难"的影响,从而能够保证高维空间中相似度量的稳定性。

2.3　针对真实图像数据的实验及结论

为进一步验证本文的度量方式在真实数据中的应用效果,文章针对真实的图像特征数据,以常用的 k-means 聚类算法为例,分别采用欧式度量和本文度量方式进行了聚类处理。

实验数据来自 UCI 机器学习数据库(UCI Machine Learning Repository)[6],所选数据集为图像分割数据(Image Segmentation Data)。该数据集包含了 2100 幅室外(outdoor)图像的特征数据,每个特征数据均由 19 维属性组成。图像共分为以下七类:Brickface(B),Sky(S),Foliage(F),Cement(C),Window(W),Path(P)以及 Grass(G),每类图像的数量相同。

实验中本文度量方式的网格划分参数取为 4bit/维,采用本文度量方式和欧式度量所得到的聚类结果分别如表 1 和表 2 所示。

表中,C1 至 C7 依次代表 Brickface 至 Grass 这七个类。以表 1 的第二行为例,其第 2 列表示正确聚类到 Brickface 类中的图像数目,而第 3 至 8 列分别代表了通过聚类处理之后,聚类到 Brickface 类中的所有其它类图像的数目。对于每一类而言,其聚类准确率定义为正确聚类到该类图像的数目与该类中图像的真实数目(即 300)的比值。所有类的聚类准确率的平均值则可视为相应算法的聚类准确率。按照这一计算方式,采用本文度量方式的聚类准确率为 77.91%,而采用欧式度量的聚类准确率仅为 64.19%。从这个意义上说,采用本文度量方式的聚类算法具有较好的聚类性能。

表1 采用本文度量方式的聚类结果

Tab.1 Clustering result by our measure

	B	S	F	C	W	P	G
C1	239	0	6	65	55	0	0
C2	0	295	0	0	0	0	0
C3	8	0	210	2	65	0	0
C4	0	0	10	198	0	60	0
C5	53	0	74	13	180	6	20
C6	0	0	0	22	0	234	0
C7	0	5	0	0	0	0	280

表2 采用欧氏度量的聚类结果

Tab.2 Clustering result by euclidean measure

	B	S	F	C	W	P	G
C1	210	0	15	70	90	0	0
C2	0	280	0	0	0	0	0
C3	24	0	160	0	120	0	2
C4	6	0	45	140	0	60	0
C5	60	0	80	30	86	30	36
C6	0	0	0	60	4	210	0
C7	0	20	0	0	0	0	262

此外,对于本文的度量方式而言,网格划分参数 b_i 是影响度量的有效性,从而进一步影响聚类效果的一个重要参数。实验数据和聚类算法同上,图4显示了 b_i 与平均聚类准确率 Ave-precision 之间的关系。

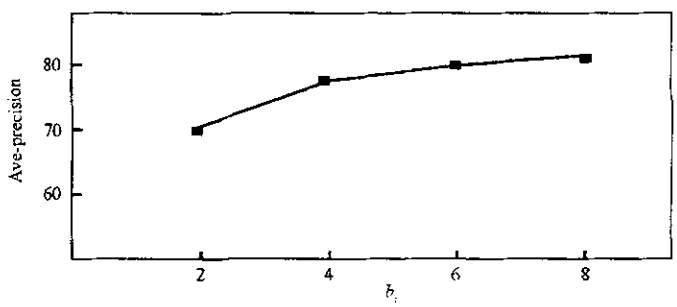


图4 平均聚类准确率随网格划分参数 b_i 的变化曲线

Fig.4 Change of ave-precision with b_i

由图4可见,随着 b_i 的增大,聚类准确率也随之逐渐升高。但是,这也并不意味着 b_i 的值越大越好,因为 b_i 越大,存储和计算的代价也必然会随之增加。以图4中的实验数据为例,在 b_i 取值为4到6的区间上,平均聚类准确率并没有发生很大的变化,但是由于每维空间上划分的网格数均为 2^{b_i},因此其存储和计算复杂度都会产生较大的变化。所以,寻求聚类算法的准确率与算法的时间和空间复杂度之间的折衷,也是一个重要的问题。对于本文的实验数据来说,b_i 取值为4就能获得令人满意的聚类结果。

3 总结与展望

度量函数的意义在于计算、表达空间中特征点之间的相似性。在高维空间中,由于‘维度灾难’的影响,使得那些适用于低维数据的度量方式在应用于高维数据时不能得到稳定的结果。本文提出了一个基于网格划分的距离度量方式,对欧式度量进行了有意义的扩展。该方法将数据空间形成网格结构之后,在度量两个数据之间的相似性时,只考虑它们落入相同区间中的维度信息。通过这种手段,很好地避免了高维空间中噪声数据对度量结果的影响。实验证明,该度量方法是行之有效的。

以本文的研究为基础,今后的工作将主要包括以下几个方面:首先,通过理论分析和反复试验,寻求一个尽可能通用的准则来确定网格划分参数,从而得到聚类算法的准确率与算法的时间和空间复杂度之间的合理折衷;其次,在保证算法性能的前提下,对原度量公式中的可用维度进行有意义的扩展,即寻求一个合适的阈值,把‘落入相同区间’和‘满足相应阈值’的维度信息都作为度量中的有用信息,从而提高度量结果的精确度。最后,还要考虑将融入了经验权值的度量方式应用于实践中。

参考文献:

[1] Zhang R, Ooi B C, Tan K L. Making the Pyramid Technique Robust to Query Types and Workloads[A]. In: Proceedings of the 20th International Conference on Data Engineering[C]. Boston, MA, USA, March 2004: 313 – 324.

[2] Amir A, Kashi R, Netanyahu N S, et al. Analyzing High-dimensional Data by Subspace Validity[A]. In: Proceedings of the Third IEEE International Conference on Data Mining[C]. Melbourne, Florida, USA, November 2003: 473 – 476.

[3] Friedman J H. Flexible Metric Nearest Neighbor Classification [R]. Technical Report, Department of Statistics, Stanford University, 1994.

[4] 汪祖媛,庄镇泉,王煦法. 逐维聚类的相似度索引算法[J]. 计算机研究与发展,2004,41(6): 1003 – 1009.

[5] Beyer K, Goldstein J, Ramarkrishnan R. When is Nearest Neighbor Meaningful? [A]. In: International Conference on Database Theory[C]. Jerusalem, Israel, January 1999: 217 – 225.

[6] http://www.ics.uci.edu/mlearn/MLRepository.