基于相似性度量的高维数据聚类算法研究

王晓阳[1],张洪渊[2],沈良忠[2],池万乐[2]
(1. 温州大学 物理与电子信息工程学院,浙江 温州 325035;
2. 温州大学城市学院,浙江 温州 325035)

摘 要:高维数据空间中的高维数据相似性度量问题是一个具有挑战性的课题。针对传统数据相似性度量算法在高维数据空间的不适应性,通过对传统的距离度量方法进行分析,结合高维数据特性,提出了高维数据相似性度量函数 Esim(X, Y)。将其与已有的相似性度量函数 Hsim(X,Y)进行比较,得出改进的算法在高维相似性度量方面的优越性,特别是在高值数据之间与低值数据之间的相对差异方面更具优势。利用数值型数据集进行实验分析,验证了该函数在高维数据空间聚类的有效性和合理性。

关键词:高维数据;相似性度量;数据聚类
中图分类号:TP311.1 **文献标识码**:A **文章编号**:1673-629X(2013)05-0030-04
doi:10.3969/j.issn.1673-629X.2013.05.008

Research on High Dimensional Clustering Algorithm Based on Similarity Measurement

WANG Xiao-yang[1], ZHANG Hong-yuan[2], SHEN Liang-zhong[2], CHI Wan-le[2]
(1. College of Physics & Electronic Information Engineering, Wenzhou University, Wenzhou 325035, China;
2. City College of Wenzhou University, Wenzhou 325035, China)

Abstract: The problem of similarity measurement for high dimensional data between high dimensional spaces is a challenging issue. Aiming at the problems of the inapplicability of the traditional measurement in high dimensional space, the improved function Esim(X,Y) is proposed to measure the similarity between the data in high dimensional space through analyzing and summarizing the traditional measurement with the properties of high dimensional data. Advantages of the improved function are obvious between high dimensional space similarity measurement comparing with Hsim(X,Y), especially in high values and low values. The experiments by numerical dataset demonstrate that the function Esim(X,Y) is effective and reasonable in high dimensional data clustering.

Key words: high dimensional data; similarity measurement; data clustering

0 引 言

随着信息技术的快速发展,社会逐渐步入信息化,各行各业的业务数据随之不断积累。运用数据挖掘技术,可以从这些海量数据中获得潜在的、未知的、有价值的知识。聚类分析[1]是数据挖掘通常采用的关键技术之一,它根据数据集中不同数据对象间的关系,将数据集划分为若干类或者簇,从而使得同一个类中的数据对象间相似度高,不同类间的数据对象相似度低。相似性度量问题在聚类过程中起到关键作用,使用聚类算法在处理高维数据时,在低维数据空间应用的传统相似性度量方法在高维数据空间表现的性能很差,因而聚类效果很不理想。因此,研究适用于高维数据空间的相似性度量方法成为亟待解决的一个关键问题。

收稿日期:2012-08-15;**修回日期**:2012-11-22
基金项目:温州市科技局项目(KZ1102239)
作者简介:王晓阳(1987-),男,硕士研究生,研究方向为数据挖掘;张洪渊,副教授,研究方向为数据挖掘、智能网及信号与信息处理。

1 高维数据空间特性分析

高维数据空间存在着稀疏性和空空间现象,从而导致高维数据空间存在“curse of dimensionality”,也就是维度灾难[2]。这一概念由 Bellman 首次提出,它是指在数据分析中遇到的由于变量(属性)过多而引起的一系列问题。聚类方法用在高维数据空间时,结果不尽人意。一方面,随着数据维度增加,通常与目标簇不相关的维度数目也随之增加,因而与目标簇相关的维度会淹没在大量不相关维度这样的噪声之中;另一方面,高维数据会随着维度增高而变得稀疏,继续采用传统的距离度量方法计算数据对象之间的距离,得到

的结果是数据对象间的距离基本相等。Aggarwal[3] 等人对 L_K –范数和数据维度的关系作了深入的研究,结果证明:随着维度的增加,最近邻和最远邻之间的对比越来越不明显,即点与点之间的距离对比性不复存在。

在低维数据空间中,通常采用传统的基于空间距离的相似度来描述数据之间的差异,但由于高维空间存在着稀疏性、空空间现象等特性,使得这一传统方法在高维空间中使用时其效果大大下降,导致结果变得不稳定[4]。聚类算法多种多样,但其区别主要在于数据对象之间的相互定义关系不同[5]。经典的聚类算法多针对低维数据所设计,当数据的维度增高时,原有聚类方法将不再适用,当然获得的结果也变得毫无意义。

2　高维相似性度量函数的改进

2.1　传统相似性度量函数

相似度方法运用得恰当与否直接影响着数据聚类的效果。针对不同的应用和数据类型,需要不同的相似性度量方法。传统的距离函数有明科夫斯基距离(Minkowsky)、欧氏距离(Euclidean)、曼哈顿距离(Manhattan)、切比雪夫距离(Chebychev)等[6]。具体如下:

设 $\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n$ 为 n 个数据样本,其中 $\boldsymbol{T}_i=(t_{i1},t_{i2},\cdots,t_{id})$,$i=1,2,\cdots,n$,而 d 为数据样本 $\boldsymbol{T}_i$ 的特征项个数也就是维度数,则数据集 $\boldsymbol{T}$ 可以表述为 $\boldsymbol{T}=(\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n)^{\mathrm{T}}$,即:

$$\boldsymbol{T}=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1d} \\ t_{21} & t_{22} & \cdots & t_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nd} \end{bmatrix}$$

对任意两个数据样本,$\boldsymbol{T}_i=(t_{i1},t_{i2},\cdots,t_{id})$,$\boldsymbol{T}_j=(t_{j1},t_{j2},\cdots,t_{jd})$,二者的明科夫斯基距离计算公式为:

$$D(\boldsymbol{T}_i,\boldsymbol{T}_j)=\left[\sum_{m=1}^{d}|t_{im}-t_{jm}|^p\right]^{\frac{1}{p}}$$

上述公式中,当 $p=2$ 时,即为欧氏距离;当 $p=1$ 时,即为曼哈顿距离;当 $p=\infty$ 时,即为切比雪夫距离。

明科夫斯基距离实际上是欧式距离、曼哈顿距离、切比雪夫距离的一种泛化。以上距离函数在处理低维数据空间时能获得较好的效果,但是在处理数据高维空间时,方法很难达到预期效果。

2.2　目前相似性度量研究方法

对于传统的距离度量方法在高维数据空间中失效的问题,很多学者进行了创新研究,提出了众多不同的相似性度量重构函数。

如相似性度量函数 Hsim($\boldsymbol{X},\boldsymbol{Y}$)[7],其具体定义如下:

$$\mathrm{Hsim}(\boldsymbol{X},\boldsymbol{Y})=\frac{\sum_{i=1}^{d}\frac{1}{1+|x_i-y_i|}}{d}$$

其中 $\boldsymbol{X},\boldsymbol{Y}$ 为数据对象,d 为数据的维度数。该重构函数对比传统的相似性度量函数取得了较好的效果,但仍存在两方面的问题:一方面,该函数仅仅考虑了对应维度的属性值之差的绝对值,没有纵向考虑对应属性的相对差值;另一方面没有将各个维度的分布特征考虑在内。

又有学者提出 Close($\boldsymbol{X},\boldsymbol{Y}$)[6] 相似度量函数,其具体定义如下:

$$\mathrm{Close}(\boldsymbol{X},\boldsymbol{Y})=\frac{\sum_{i=1}^{d}\mathrm{e}^{-|x_i-y_i|}}{d}$$

该函数较传统的相似性度量函数有一定的优势,利用了 e 负指数幂函数这一单调递减特性,但同样也存在着对应属性间相对差值考虑不足的问题。

2.3　改进的高维数据相似性度量函数

根据对 Hsim($\boldsymbol{X},\boldsymbol{Y}$) 以及 Close($\boldsymbol{X},\boldsymbol{Y}$) 函数的分析,针对其不足之处提出改进的高维数据相似性度量函数 Esim($\boldsymbol{X},\boldsymbol{Y}$),具体定义如下:

$$\mathrm{Esim}(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{d}\sum_{i=1}^{d}\boldsymbol{\omega}_i\mathrm{e}^{-\frac{|x_i-y_i|}{|x_i-y_i|+|x_i+y_i|/2}}$$

式中 $\boldsymbol{X}=(x_1,x_2,\cdots,x_d)$ 与 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_d)$ 是维度数为 d 的数据空间中两个特征向量,x_i 与 y_i 分别为 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 在第 i 维上的属性值;$\boldsymbol{\omega}_i$ 是每个维度上的属性权重,取值为 0 到 1 之间。Esim($\boldsymbol{X},\boldsymbol{Y}$) 函数具有如下性质:

1) 函数描述了两数据对象之间的相似性程度,函数值介于 0 到 1 之间,越大表示二者越相似。

2) 函数的最小值为 0,表示各个数据对象在对应维度上差别接近无穷大或相似度最低。

3) 函数的最大值为 1,表示各个数据对象在对应维度上差别为零或相似度最高,即相互重合。

4) 若两数据对象同时在某同一维度上都取值为零,则视为在该维度相似度最低,取值为零。

Esim($\boldsymbol{X},\boldsymbol{Y}$) 重构函数利用了 e 负指数幂函数这一单调递减特性。令 $d=|x_i-y_i|$,函数 e^{-d} 比函数 $1/(1+d)$ 在 0 ~ 1 区间上更为陡峭,下降速度更快,这就意味着,在对应维度差值较大的情况下更具有区分度。同时 Esim($\boldsymbol{X},\boldsymbol{Y}$) 在其指数的分母中加入 $|x_i+y_i|/2$,能有效地考虑到高值数据间的相似性与低值数据间的相似性度量上的相对差异,并且随着维度的升高,数据的整体相似性程度也会增高,使其在处理相对差值问题上更贴合实际。

3 有效性及实验分析

3.1 有效性分析

文中提出的 Esim($\boldsymbol{X},\boldsymbol{Y}$) 优越性在于,一方面考虑了不同数据对象对应维度上的相对差值,另一方面是加入维度权重 ω_i,可以根据实际需要来定夺不同维度对相似度的贡献程度,这种具有伸缩性的权重取值来自于领域知识以及实际经验。

衡量距离函数的有效性通常采用文献[8]提出的方法作为判断标准:

$$v = \frac{D_{\max_d} - D_{\min_d}}{D_{\min_d}}$$

式中 $D_{\max_d}$ 和 $D_{\min_d}$ 分别为数据空间中数据对象间的最大距离与最小距离。传统的相似性度量函数在处理高维数据时,这一比值 v 会随着维度的增加而逐渐趋于零,即最远距离与最近距离趋于相等,比如欧式距离、L_k - 范数[9,10]。为证明文中提出的相似性度量函数 Esim($\boldsymbol{X},\boldsymbol{Y}$) 的有效性,实验利用 Matlab 的 normrnd(μ,σ,m,n) 高斯随机函数,按维度呈高斯分布随机生成 10 维,30 维,50 维,100 维,150 维,200 维,300 维,500 维,800 维不同维度的数值型数据集,每种维度数据集含有 100 条数据,得到的比值 v 随维度变化图像如图 1 所示。

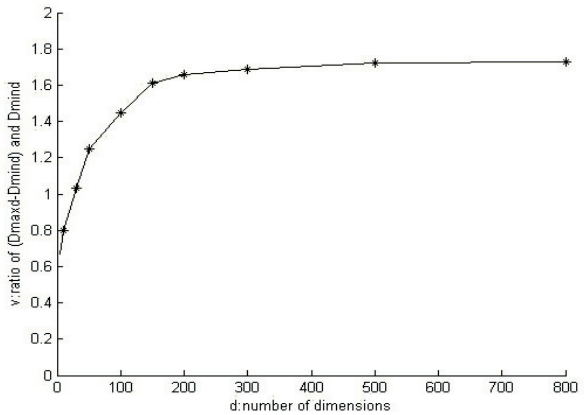

图 1 比值 v 随维数 d 变化的曲线

从图 1 中可以得知,由文中提出的度量方法所得出比值 v 没有随维数 d 的增加而降低,反而随着维度的增加而增加最后趋于稳定。从这一点说明,最远距离没有无限接近最近距离,说明了该度量方法在区分最远距离与最近距离上有显著成效,在某种程度上减小了维度灾难带来的影响,从而验证了 Esim($\boldsymbol{X},\boldsymbol{Y}$) 函数在高维空间适用的有效性。

Esim($\boldsymbol{X},\boldsymbol{Y}$) 函数在计算时保留了数据对象对应维度上的绝对差值,同时也加入了对应维度上的相对差值[11],这个改进取得了良好效果。以下选取一对数据做一分析说明,对比其与 Hsim($\boldsymbol{X},\boldsymbol{Y}$) 以及 Close 函数的效果如表 1 所示。

表 1 Esim($\boldsymbol{X},\boldsymbol{Y}$) 函数与 Hsim($\boldsymbol{X},\boldsymbol{Y}$) 以及 Close($\boldsymbol{X},\boldsymbol{Y}$) 函数的比较

		维 1	维 2	Hsim( )	Close( )	Esim( )
数据对象	A	2	202	0.2	0.018	0.794
	B	6	206			
		A	B	Hsim( )	Close( )	Esim( )
单维相似度	维 1	2	6	0.2	0.018	0.606
	维 2	202	206	0.2	0.018	0.981

有两个数值型二维数据对象,分别为数据 A(2,202),数据 B(6,206)。由表 1 中数据可以看出,数据 A 与数据 B 在对应的维度上的差值同为 4,尽管在维 2 上二者的数值较大,但通过直观能判断出实际上 A 与 B 的相似度是比较高的。然而,由 Hsim($\boldsymbol{X},\boldsymbol{Y}$) 计算得出二者相似度为 0.2,由 Close($\boldsymbol{X},\boldsymbol{Y}$) 函数计算得出二者相似度为 0.018,显然结果偏低,未能符合实际。

究其原因容易得出,Hsim($\boldsymbol{X},\boldsymbol{Y}$) 以及 Close($\boldsymbol{X},\boldsymbol{Y}$) 函数在计算时只考虑了对应维度的绝对差值而忽略了相对差值,因此当对应维度差值相同或相近时,其属性数值的不同数量级高低对相似性度量影响极大。Esim($\boldsymbol{X},\boldsymbol{Y}$) 函数恰恰能避免这一点,结果得出维 1 与维 2 的单维相似度不同,分别为 0.606 与 0.981,其二者平均值即为 A、B 的整体相似度,符合实际结果。通过引入对应属性相对差值,从而对不同维度的属性数值区别对待,因而能更好地表达数据间的相似性。

3.2 实验数据仿真

为进一步验证 Esim($\boldsymbol{X},\boldsymbol{Y}$) 函数的性能在高维数据聚类计算中的表现,实验采用 UCI[12] 提供的 Wine 数据集。Wine 数据集包含 178 条数据对象,数据为数值型共 13 个维度,由三个簇组成。选取 K - means 作为此次聚类算法,采用 Hsim($\boldsymbol{X},\boldsymbol{Y}$) 距离函数、Esim($\boldsymbol{X},\boldsymbol{Y}$) 函数分别进行聚类,聚类结果如表 2、3 所示。

表 2 基于 Hsim($\boldsymbol{X},\boldsymbol{Y}$) 相似性度量函数的聚类结果

类别	总数	实际划分	正确划分	聚类精度
类一	59	49	42	85.7%
类二	71	65	50	76.9%
类三	48	64	33	51.6%

表 3 基于 Esim($\boldsymbol{X},\boldsymbol{Y}$) 相似性度量函数的聚类结果

类别	总数	实际划分	正确划分	聚类精度
类一	59	57	50	87.7%
类二	71	69	59	85.5%
类三	48	52	46	88.5%

由表 2 和表 3 的实验结果得出,基于文中提出的 Esim($\boldsymbol{X},\boldsymbol{Y}$) 相似性度量函数的高维数据聚类方法相对于基于 Hsim($\boldsymbol{X},\boldsymbol{Y}$) 距离函数的聚类方法在三个类别中正确划分的数目明显增加,更接近各类数据的真实分布,同时提高了聚类精度。实验效果表明文中提出的 Esim($\boldsymbol{X},\boldsymbol{Y}$) 相似性度量函数能够有效、合理地用于描述数据之间的相关程度,同时能提高高维数据聚类效果。

4　结束语

文中对高维数据空间的特性进行分析,从相似性度量角度着手来缓解维数灾难所带来的严重问题。通过分析传统的相似性度量函数在处理高维数据时的不适应性,并对当前所应用的相似性度量函数进行总结评价,提出了改进的相似性度量函数。通过有效性及实验分析,该函数优于各类已有的相似性度量方法,从而验证了该函数在处理高维数据空间时是有效的、合理的。

参考文献:

[1] 王丽珍,周丽华,陈红梅,等. 数据仓库与数据挖掘原理及应用[M]. 北京:科学出版社,2005.

[2] Pan Guotao, Huang Decai. Research on Similarity Measurement of High Dimensional[C]//The 2011 International Conference of Youth Communication. Macao, China: [s. n.], 2011.

[3] 谢明霞,郭建忠. 高维数据相似性度量方法研究[J]. 计算机工程与科学,2010,32(5):780-783.

[4] 贺　玲,吴玲达,蔡益朝. 高维空间中数据的相似性度量[J]. 数学的实践与认识,2006,36(9):189-194.

[5] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[6] 邵昌昇,楼　巍,严利民. 高维数据中的相似性度量算法的改进[J]. 计算机技术与发展,2011,21(2):1-4.

[7] 杨风召. 高维数据挖掘技术研究[M]. 南京:东南大学出版社,2007.

[8] Hinneburg A, Aggarwal C C, Keim D. What is the Nearest Neighbor in High Dimensional Spaces? [C]//Proceedings of the 26th International Conference on Very Large Data Bases. Cario, Egypt: UKPMC Press, 2000: 506-515.

[9] Aggarwal C C. Re-designing Distance Functions and Distance Based Applications for High Dimensional Data[J]. ACM SIGMOD Record, 2001, 30(1): 13-18.

[10] Aggarwal C C. On the Effects of Dimensionality Reduction on High Dimensional Similarity Search[C]//Proc of the 20th ACM SIGMOD-SIGACT-SIGART Symp on Principles of Database Systems. New York: ACM Press, 2001: 256-266.

[11] 黄斯达,陈启买. 一种基于相似性度量的高维数据聚类算法的研究[J]. 计算机应用与软件,2009,26(9):102-105.

[12] Forina M. UCI Machine Learning Repository[EB/OL]. 2012. http://archive.ics.uci.edu/ml/datasets/Wine.

(上接第 29 页)

问题还有很多。由于空间数据挖掘的主要对象是 GIS 空间数据库[11,12],所以在以后的研究中,将重点集中在基于 GIS 的空间数据挖掘的研究领域。

参考文献:

[1] Shekhar S, Chawla S. Spatial Databases[M]. 北京:机械工业出版社,2004.

[2] 李德仁,王树良,史文中,等. 论空间数据挖掘和知识发现[J]. 武汉大学学报:信息科学版,2001,26(6):491-499.

[3] 王劲峰. 地图的定性和定量分析[J]. 地球信息科学学报,2009,11(2):169-175.

[4] Chen Y L, Chen J M, Tung C W. A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales[J]. Decision Support Systems, 2006, 42(3): 1503-1520.

[5] 黄杏元,劲　松,汤　勤. 地理信息系统概论[M]. 北京:高等教育出版社,2001.

[6] Ester M, Hans-Peter K, Sander J. Knowledge Discovery in Spatial Database[C]//Proc of Conf. on Artificial Intelligence. Bonn, Germany: [s. n.], 1999.

[7] Ester M, Frommelt A, Kriegel H, et al. Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support[J]. Data Mining and Knowledge Discovery, 2000, 4(2/3): 193-216.

[8] Han Jiawei, Kamber M. 数据挖掘:概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2007.

[9] Bembenik R, Rybifiski H. Mining Spatial Association Rules with No Distance Parameter[C]//Proc. of Intelligent Information Processing and Web Mining. Berlin: Springer, 2006: 499-508.

[10] 张　宏,温永宁,刘爱利,等. 地理信息系统算法基础[M]. 北京:科学出版社,2006.

[11] 李东平,姚　远. GIS 的发展趋势与数字地震应急救灾的实现技术[J]. 计算机技术与发展,2011,21(1):214-217.

[12] 吕曹芳. 基于 GIS 的空间数据挖掘研究进展[J]. 皖西学院学报,2010,26(2):43-46.

基于相似性度量的高维数据聚类算法研究

作者： 王晓阳， 张洪渊， 沈良忠， 池万乐
作者单位： 王晓阳(温州大学 物理与电子信息工程学院,浙江 温州 325035)， 张洪渊,沈良忠,池万乐(温州大学城市学院,浙江 温州 325035)

刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(5)