基于带权多维尺度变换的奇异值挖掘*)

魏 莱 王守觉 徐菲菲
(同济大学计算机科学与技术系 上海 201804)

摘 要 大量的高维数据在分布上表现为一低维流形,试图从这样的数据集中探测出奇异点,传统的奇异点挖掘算法可能失效。本文提出了一种带权重的多维尺度变化,算法通过局部的高维数据集和其低维重构的误差来设定数据点的局部权重,再利用权重之和得到的数据点置信度,以此来进行奇异值的判定。通过实验验证了算法的有效性。

关键词 奇异值,多维尺度变换,带权多维尺度变换,流形学习

Outliers Mining via Weighted Multidimensionality Scaling

WEI Lai WANG Shou-Jue XU Fei-Fei
(Department of Computer Science and Technology, Tongji University, Shanghai 201804)

Abstract Mining outliers from the data set which is distributed on a low dimensional manifold is a hard task. The existing algorithm may not be effective for the situation. So a novel approach called weighted multidimensionality scaling is proposed for outliers mining. It is based on multidimensionality scaling, MDS. Every data point will get a reliability score by the algorithm, then it can be determined whether it is a outlier through the value of its reliability score. The experiments show the efficiency of the algorithm.

Keywords Outliers, MDS, Weighted MDS, Manifold learning

1 引言

在大量的数据中可能包含着一些数据,它们与一般的数据的行为和模型不一致。这些被称为奇异点(outliers)的数据在大部分情况下会被去除,但在一些应用中非正常的事件可能比正常的事件更有趣,因此作为数据挖掘的一个分支,奇异点挖掘也引起了很多学者的关注[1~3]。

通过对奇异值表现的分析,大部分挖掘算法可以归为如下四类:基于统计的检测[3]、基于距离的检测[4,5]、基于偏差的检测[6]以及基于密度的检测[7]。但是随着信息技术的发展,数据的形式和内容与以往有了较大的区别,表现在数据的高维数和数据属性的高度相关。这些高维数据集的分布在高维观测空间形成了一种低维的几何结构,在数学上称为流形,这种几何结构是未知的,且非线性的。因此,以上传统的奇异点检测算法在这种情况下可能失效。为此需要一种新的方法来处理流形上奇异值检测的问题。

本文提出了一种带权重的多维尺度变换奇异值挖掘算法(Outliers Mining via Weighted Multidimensional Scaling, OMWMDS)。该算法利用多维尺度变换(Multidimensional Scaling)的思想,通过计算数据重构的误差来赋予数据点相应的权值,对于分布成为一低维流形的数据集通过局部的Weighted-MDS得到局部权重,遍历数据集后通过以局部权重的和作为每一数据点的置信度,通过置信度的大小来判定数据点是否为奇异点。实验表明算法能够有效挖掘出当正常数据分布成一低维流形时整体数据中的非正常数据,同时每一数据点相应的置信度也可用于改进流行学习算法的鲁棒性[8]。

2 多维尺度变换

2.1 多维尺度变换(Multidimensional Scaling, MDS)[9]

多维尺度变换同主成分分析(Principal Component)一样是一种应用广泛的线性降维算法,在图像处理、计算机视觉等方面有着广泛的应用。MDS不需要知道数据点的具体坐标,它通过对数据之间的距离矩阵的奇异值分解来获得数据的低维重构坐标,从而有效地对数据集进行降维。经典的MDS算法具体步骤如下:

假设数据集为 $X=\{x_1,x_2,\cdots,x_n\}$,其中 $x_i\in R^D$,并且我们有任意两数据点之间距离矩阵 $D=\{d_{ij}\}$,数据点相应的重构坐标为 $Y=\{y_1,y_2,\cdots,y_n\}$,$y_i\in R^d$。

STEP 1. 计算 $S=\{d_{ij}^2\}$;

STEP 2. 取矩阵 $H=\{h_{ij}\}$,满足 $h_{ij}=\delta_{ij}-\frac{1}{n}$,其中

$$\delta_{ij}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases};$$

STEP 3. (双中心化)计算 $\tau(D)=-\frac{HSH}{2}$;

STEP 4. 求矩阵 $\tau(D)$ 的特征值和其对应的特征向量,矩阵 Λ 是对角阵,对角元素是从大到小排列的特征值,矩阵 U 的列为相应的特征向量。

STEP 5. 计算 Y,对 Λ 对角线元素依次取算术平方根,得 $\sqrt{\Lambda}$,那么 $Y=U\sqrt{\Lambda}$。

如果数据分布是由内在低维变量控制,变量个数为 d,那么矩阵 Λ 的第 $d+1$ 个特征值就接近于零,这时就有 $Y=U_d\sqrt{\Lambda_d}$,其中 Λ_d 为 Λ 的前 d 个特征值组成的矩阵,U_d 的列为

*)国家自然科学基金项目(60495019),教育部博士点专项基金(20060247039)。魏 莱 博士研究生,研究方向:流形学习、仿生模式识别、信息几何、粗糙集理论。

相应的特征向量。这里 $Ye^t=0, e(1,1,\cdots,1)$。

事实上 MDS 算法的原理是基于这样的数学推导。我们不知道数据点的具体坐标，假设将数据点均值在坐标原点，定义内积矩阵 $Bx=XX^T$。那么如果能得到 Bx 的值那么相应的数据集坐标也就可以得到了。而从距离矩阵 $D=\{d_{ij}\}$ 出发，令 $S=\{d_{ij}^2\}$，对其进行所谓的双中心化 $\tau(D)=-\frac{HSH}{2}$（H 如上述算法定义相同），这时我们发现有 $Bx=\tau(D)$，再对 $\tau(D)$ 进行奇异值分解，即 $Bx=\tau(D)=U\Lambda U^T$（U^T 表示 U 的转置），很明显 X 的重构坐标就可以表述为 $X'=U\sqrt{\Lambda}$。

由于 MDS 算法的数学直观性和操作简易性，它成为一种经典的数据降维算法。然而正如前文指出，在当数据集结构高度非线性，分布成一低维流形时，MDS 是不能进行有效降维的。近年提出的流形学习算法就是解决数据集具有上述特征时的数据维数约简问题。

数据分布高度非线性结构对奇异值探测提出了挑战。为此作者提出了利用多维尺度变换来进行流形奇异值的探测。

2.2 带权重的多维尺度变换

设数据集为 $X=\{x_1,x_2,\cdots,x_n\}$，其中 $x_i\in R^D$，非奇异点分布在一个低维流形上，如图 1、2 左侧图片所示。我们取任意一点 x_i 的 K 个近邻组成局部邻域 $X_i=\{x_{i0}=x_i,x_{i1},x_{i2},\cdots,x_{iK}\}$，由于流形在局部上与欧式空间同胚，因此通过 MDS 可以得到低维坐标 $Y_i=\{y_{i0},y_{i1},\cdots,y_{iK}\}$，其坐标均值位于坐标原点。我们将原数据集进行平移使其均值点也位于坐标原点，即

$$X_i'=\{x_{i0}-\bar{x},x_{i1}-\bar{x},\cdots,x_{iK}-\bar{x}\}=\{x_{i0}',x_{i1}',\cdots,x_{iK}'\}$$

$$\bar{x}=\frac{1}{K}\sum_{j=0}^{K}x_{ij}$$

定义原数据坐标和相应的重构数据坐标之间的误差为 $\varepsilon_{ij}=|x_{ij}'-y_{ij}|$，邻域中所有点的误差表示成向量形式为：$E_i=\{\varepsilon_{i0},\varepsilon_{i1},\cdots,\varepsilon_{iK}\}$，根据相应的误差我们定义数据点的权重为：$w_j=w(\varepsilon_{ij})$。

由文[10]的思想，可以假设总误差可以表示为 $E_i'=\sum_{j=0}^{K}\rho(\varepsilon_{ij})$，$\rho(\cdot)$ 为一凸函数。这里采用 Huber[10] 函数：

$$\rho(\varepsilon)=\begin{cases}\frac{1}{2}\varepsilon^2, & |\varepsilon|\leqslant c\\ c(|\varepsilon|-\frac{1}{2}c), & |\varepsilon|>c\end{cases}$$

那么权重函数可以定义为：

$$w(\varepsilon)=\frac{\rho'(\varepsilon)}{\varepsilon}=\begin{cases}1, & |\varepsilon|\leqslant c\\ \frac{c}{|\varepsilon|} & |\varepsilon|>c\end{cases}$$

其中 c 是一个参数，设 $c=\frac{1}{2K}\sum_{j=0}^{K}|\varepsilon_{ij}|$。

再利用相应权重更新局部距离矩阵 D，使得 $D=\{w_rw_sd_{rs}\}$。有了这个更新的局部距离矩阵，可以重复上述过程直到当 D 不再变化，此时会得到每一个数据点相应的局部权重。通过遍历所有数据点，将每一数据点的局部权重相加，得到每一数据点的一个置信度(reliability score)，我们发现奇异点的置信度都比较小，而非奇异点的置信度则较大，因此通过设定置信度的阈值，我们就可以判定一个数据点是否为一奇异点。

具体算法步骤如下(Weighted—MDS, WMDS)：

STEP 1：初始化所有数据点的可信度为 0，选取数据点 x_i 的 K 个近邻，组成局部邻域集 X_i；

STEP 2：通过对局部邻域 X_i 做 MDS，得到低维数据集 Y_i，计算误差向量 E_i；

SETP 3：根据误差向量计算局部邻域中每个数据点相应的权重，利用权重更新局部距离矩阵 D_i；

STEP 4：如果 D_i 变化不超过阈值 t，则算法停止，每一数据点的可信度加上输出的权重；否则转 STEP2。

3 实验比较

我们来看一下实验结果，这里我们采用流形学习算法中常用的几个人造实验数据集。令局部距离矩阵 D_i 变化阈值为 0.1。

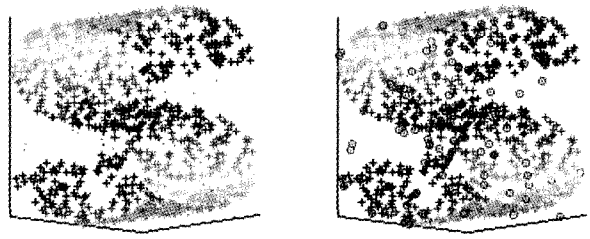

图 1 Scurve；数据点个数 1100，其中奇异点个数 100，置信度阈值 0.5，$K=8$；

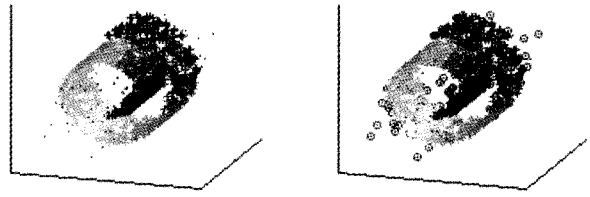

图 2 SwissRoll；数据点个数 1100，其中奇异点个数 100，置信度阈值 0.5，$K=8$；

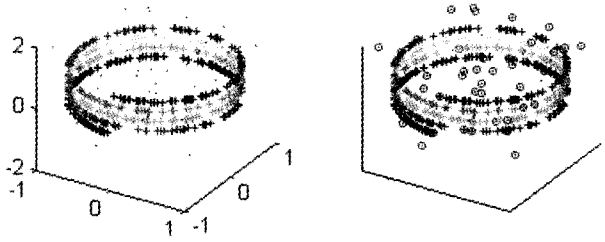

图 3 Helix；数据点个数 500，其中奇异点个数 50，置信度阈值 0.6，$K=12$；

可见直观上，上述算法确实能够探测出伴随着低维流形的奇异值。我们用下面的表格来显示 WMDS 和基于距离探测奇异值算法在探测奇异值方面性能的差异，这里 TPO 是指奇异点被正确识别的概率，FPN 是指非奇异点被错误识别的概率。

表 1

	Scurve		SwissRoll		Helix	
	TPO	FPN	TPO	FPN	TPO	FPN
WMDS	0.54	0.035	0.65	0.037	0.78	0.1
TestByDistance	0.48	0.021	0.63	0.027	0.7	0

结论 奇异值探测是数据挖掘邻域近年兴起的一项研究热点，它试图发现隐藏在常规数据集中的非常规数据点，在各种决策分析和风险规避方面起到了很大的作用。但是随着数据性质以及数据结构的变化，大量的真实数据表现为一种流形结构，因此传统的一些奇异值挖掘算法可能失效，为此本文

提出了一种基于 MDS 的奇异值探测算法，算法赋予每一数据点相应的置信度，以其来进行奇异值的判定，算法稳定，比以往的奇异值挖掘算法效率也有提高。

参 考 文 献

1 Han Jiawei, Kamber M. Data Mining: Concepts and techniques [M]. Morgan Kaufmann Publishers, 2001
2 Grubbs F E. Procedures for detecting outlying observations in samples [J]. Technometrics, 1969, 11: 1～21
3 Barnett V, Lewis T. Outliers in Statistical Data [M]. John Wiley & Sons, 1994
4 Knorr E, Ng R. A unified notion of outliers: Properties and computation. In: Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD'97), 1997. 219～222
5 Knorr E, Ng R. Algorithms for mining distancebased outliers in large datasets. In: Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98), 1998. 392～403
6 Arning A, Agrawal R, Raghavan P. A linear method for deviation detection in large databases. In: Proc. 1996 Int. Conf. Data Mining and KnowledgeDiscovery (KDD'96), 1996. 164～169
7 Breunig M, Kriegel H P, Ng R, Sander J. LOF: Identifying Density-Based Local Outliers. In: Proc. ACM SiGMOD 2000 Int. Conf. on Management of Data, 2000. 93～104
8 Chang Hong, Yeung D Y. Robust locally linear embedding [J]. Pattern Recognition, 2006, 39: 1053～1065
9 Cox T, Cox M. Multidimensional Scaling [M]. London Chapman&Hall, 1994
10 Huber P J. Robust regression: asymptotics, conjectures, and Monte Carlo, Ann. Statist, 1973, 1(5): 799～821

(上接第 186 页)

H(心情/体温)=

-(2/8)((1/2) * log2(1/2)+ (1/2) * log2(1/2))-

-(4/8)((1/4) * log2(3/4)+ (1/4) * log2(3/4))-

-(2/8)((1/2) * log2(1/2)+ (1/2) * log2(1/2))=0.91

H(心情/气温)=

-(2/8)((1/2) * log2(1/2)+ (1/2) * log2(1/2))-

-(4/8)((1/4) * log2(3/4)+ (1/4) * log2(3/4))-

-(2/8)((1/2) * log2(1/2)+ (1/2) * log2(1/2))=0.91

H(心情/天气)=

-(4/8)((2/4) * log2(2/4)+ (2/4) * log2(2/4))-

-(4/8)((1/4) * log2(1/4)+ (3/4) * log2(3/4))=0.905

(3)互信息计算

$I(U)=H(U)-H(U|V)$

I(血压)=0.954-0.94 =0.014;

I(心跳)=0.954-0.8 =0.154;

I(体温)=0.954-0.91 =0.044;

I(气温)=0.954-0.91 =0.044;

I(天气)=0.954-0.905=0.049;

(4)建立决策树的树根和分支

ID3 选择互信息最大的心跳作为树根，在 8 个规则中，对心跳进行分支，3 个分支所对应的子集为

F 心跳慢={3,6,8};F 心跳正常={4};F 心跳快={3,6,8}

F 心跳正常={4}中的例子全部属于"心情好"的类，其余两个集合继续采用该建树方法。

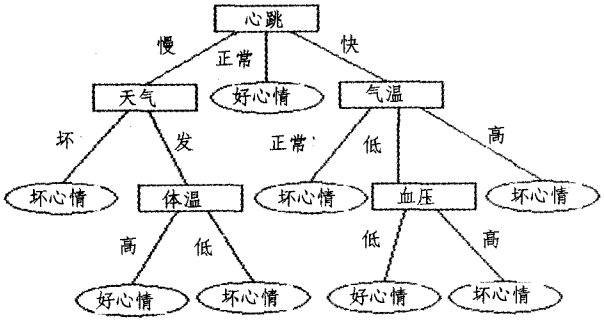

图 3 Agent 的决策树

(5)递归建树

F 心跳慢={3,6,8}，所有的值都是心跳慢，所以 $H(U)=H(U/V)$，故 $I(U)=0$。在剩余的 4 个属性中，I(天气)=0.049 互信息最大，所以作为该分支的根节点。再向下分支，F 天气坏={3}，其全部属于"心情"好的类；F 天气好={6,8}；然后在建立子树。I(体温)=0.044，作为其子树的根节点，F 体温高={6}，属于"好心情"类；F 体温低={8}，属于"坏心情"类。

F 心跳快={1,2,5,7}，所有的值都是心跳慢，所以 $H(U)=H(U|V)$，故 $I(U)=0$。在剩余的 2 个属性中，I(气温)=0.044 的互信息最大，故选择其为子树的根节点。F 气温低={1,7}；F 气温正常={2}，属于坏心情类；F 气温高={5}，属于坏心情类；对于 F 气温低={1,7}，选择 I(血压)=0.014 作为子树的根节点。F 血压低={1}，属于"好心情"；F 血压高={7}，属于"坏心情"。根据计算结果画出的 Agent 决策树，如图 3 所示。

结论 本文根据信息系统的定义，提出了采用集合方法描述的多 Agent 信息系统，研究了在多 Agent 交互中规则的形成过程。用实例验证了采用 ID3 算法挖掘 Agent 规则，发现知识过程的可行性。在后续的研究中，将把知识发现功能融入到 Agent 系统中，使其具备更强的认知能力。

参 考 文 献

1 Wooldridge M, Jennings N R. Intelligent agents: theory and practice. The Knowledge Engineering Review, 1995, 10(2): 115～152
2 Cohen P R, Levesque H J. Intention is choice with commitment. Artificial Intelligence, 1990, 42(2): 213～261
3 Konolige K, Pollack M E. A representation a list theory of intention. In: Bajcsy, R., ed. Proceedings of the 13th International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1993. 390～395
4 Halpern J Y, Moses Y. A guide to completeness and complexity for modal logics of knowledge and belief. Artificial Intelligence, 1992, 54: 319～379
5 Kraus S, Sucara K, Evenchik A. Reaching agreements through argumentation: a logical model and implementation. Artificial Intelligence, 1998, 104～169
6 Allen J F. Towards a general theory of action and time. Artificial Intelligence, 1984, 23(2): 123～154
7 Zhang Wenran, Chen Sushinn. Pool2: A generic system for cognitive map development and decision analysis. IEEE Transactions on Systems, Man and Cybernetics, 1989, 19(1): 31～39
8 Tomohiro T, Michio S. Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man and Cybernetics, 1985, 15(1): 116～132
9 Quinlan J R. Induction of decision trees [J]. Machine Learning, 1986, 1: 81～106
10 Sandholm T W, Lesser V R. Coalition among computationally bounded agents. Artificial Intelligence, 1997, 94: 99～137
11 刘东升. 基于 MobileAgent 的分布式 ID3 挖掘模型. 计算机应用与软件[J], 2005, 22(10): 49～51
12 王熙照, 谢竞博. 基于属性间交互信息的模糊 ID3 算法的扩展. 复以学报(自然科学版) [J], 2004, 43(5): 777～780
13 王澜, 何华灿. 基于广义相关系数的 Agent 行为决策模型. 计算机科学[J], 2005, 32(2): 175～177