基于 PCA 及属性距离和的孤立点检测算法

张忠平,宋少英,宋晓辉
ZHANG Zhong-ping,SONG Shao-ying,SONG Xiao-hui

燕山大学 信息科学与工程学院,河北 秦皇岛 066004
College of Information Science and Engineering,Yanshan University,Qinhuangdao,Hebei 066004,China
E-mail:juegood@yahoo.com.cn

ZHANG Zhong-ping,SONG Shao-ying,SONG Xiao-hui.Algorithm for outlier detection based on principal component analysis and sum of attributes distance.Computer Engineering and Applications,2009,45(17):139-141.

Abstract: An outlier detection algorithm based on principal component analysis and the sum of attributes distance is proposed. The algorithm firstly extracts the principal components from many attributes satisfying accumulative contribution rate.Simultaneously,by the PCA matrix original dataset is transformed to a new feature space composed of principal component.Then outliers are detected using the approach of the sum of attributes distance in the transformed datasets.The results of the experiment show that the outlier detection algorithm based on principal component analysis and the sum of attributes distance is effective.

Key words: outlier;principal component analysis;accumulative contribution rate;the sum of attributes distance

摘 要:提出了一种基于主分量分析和属性距离和的孤立点检测算法。该方法首先通过主分量分析方法从众多属性中提取出满足累计贡献率的主分量,同时利用 PCA 变换矩阵把原始数据集转换到由主分量组成的新的特征空间上,之后对转换后的数据集用属性距离和的方法对孤立点进行检测。实验结果证明了基于主分量分析和属性距离和的孤立点检测算法的有效性。

关键词:孤立点;主分量分析;累计贡献率;属性距离和
DOI:10.3778/j.issn.1002-8331.2009.17.042 **文章编号**:1002-8331(2009)17-0139-03 **文献标识码**:A **中图分类号**:TP311

1 引言

孤立点检测现已成为数据挖掘领域的研究热点之一。其目的是发现那些异常于数据集中绝大部分对象行为的数据对象。通常对异常信息进行挖掘比挖掘常规模式更有价值。因为每一个异常数据对象都代表着一种不同的规则或模式。目前,孤立点检测已经广泛地应用于诸如网络入侵检测、信用卡欺诈、天气预报、医药等领域中。

孤立点迄今为止还没有一个人们广泛接受的定义,Hawkins[1]对孤立点的定义从一定意义上揭示了孤立点的本质:孤立点与其他点如此不同,以至于让人们怀疑这些孤立点是由另外一个不同的机制产生的。近年来科研人员提出了许多孤立点检测算法,根据算法的不同,策略可以分为基于统计的方法、基于距离的方法、基于偏离的方法[2]。其中,基于距离的方法由于算法思想直观,容易实现而得到了广泛的研究和应用。基于距离的孤立点检测算法由 Konrr 与 Ng 首先提出,他们对孤立点的定义是:数据集 D 中,至少有 p 部分对象与对象 O 的距离大于 d,那么对象 O 就是一个带参数 p 和 d 的基于距离的孤立点,记为 $DB(p,d)$;在此基础上设计了基于单元(Cell-Based,CB)的孤立点检测算法[3-4]。之后提出的在线性时间内对任意顺序数据集的基于距离的孤立点挖掘和一个简单的剪枝规则[5]、RBRP(Recursive Binning and Re-Projection)[6]等算法都对基于距离的算法进行了改进,以便使其适用于高维数据集的孤立点挖掘。

本文提出了一种基于主分量分析和属性距离和的孤立点检测方法。该算法首先利用主分量分析方法从正常行为数据集中提取满足累计贡献率的主分量,最大限度地用最少的属性表现样本数据最大的信息;在此基础之上,用属性距离和的方法对转换到新的特征空间上的数据集进行孤立点检测。由用户给出需要检测的孤立点数目,从而突破了基于距离的方法需要由专家设置参数的限制,避免了结果对参数的敏感性。实验结果证明了算法的可行性和有效性。

2 主分量分析

2.1 主分量分析原理

一般的数据集所包含的属性从十几个到上百个不等,但含有上千个属性的数据集也在日益增多。然而,异常行为往往只集中在少部分属性上,如果将算法应用在全部的属性上,不仅

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60773100);教育部科学技术研究重点项目(the Science and Technologies Research Plan of the Ministry of Education under Grant No.205014);河北省教育厅科研计划项目(the Science Research Plan of the Office of Education of Hebei under Grant No.2006143)。

作者简介:张忠平(1972-),男,博士后,副教授,硕士生导师,CCF 会员,主要研究方向:数据挖掘,语义网,XML 数据库,网格技术;宋少英(1981-),女,硕士研究生,主要研究方向:孤立点检测;宋晓辉(1983-),男,硕士研究生,主要研究方向:数据挖掘。

收稿日期:2008-04-02 **修回日期**:2008-06-17

会耗费时间,增加计算的复杂性,还会影响到数据分类的正确性。虽然数据集的每个属性都提供了一定的信息,但其提供信息量的多少及重要性是有差别的,而且在许多情况下,属性间存在着不同程度的相关性,导致这些属性所提供的信息必然有一定的重叠,因此人们希望从这些属性中提取出主要属性,用较少的互不相关的新变量来分析问题。主分量分析[7](Principle Components Analysis,PCA)正好能满足这一要求,它能够很好的处理高维数据,使得低维数据能够在平方和最小的意义下描述高维原始数据[8-9]。

设 $X_1,X_2,\cdots,X_p$ 为实际问题所涉及的 p 维变量,记作 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$,其均值 $E(\boldsymbol{X})=0$,设:

$$\boldsymbol{L}=(L_1,L_2,\cdots,L_p)^{\mathrm{T}}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pp} \end{bmatrix} \tag{1}$$

考虑以下线性组合:

$$\begin{cases} \boldsymbol{Y}_1=\boldsymbol{L}_1^{\mathrm{T}}\boldsymbol{X}=l_{11}X_1+l_{12}X_2+\cdots+l_{1p}X_p \\ \boldsymbol{Y}_2=\boldsymbol{L}_2^{\mathrm{T}}\boldsymbol{X}=l_{21}X_1+l_{22}X_2+\cdots+l_{2p}X_p \\ \quad\vdots \\ \boldsymbol{Y}_p=\boldsymbol{L}_p^{\mathrm{T}}\boldsymbol{X}=l_{p1}X_1+l_{p2}X_2+\cdots+l_{pp}X_p \end{cases} \tag{2}$$

Y_i 也是均值为 0 的随机变量,其方差为:

$$\boldsymbol{E}(\boldsymbol{Y}_i^2)=\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{R}_{XX}\boldsymbol{U}_i \quad i=1,2,\cdots,p \tag{3}$$

式(3)中,$\boldsymbol{R}_{XX}$ 为自相关阵,由于 $E(\boldsymbol{X})=0$,$\boldsymbol{R}_{XX}$ 也是协方差阵。其中,

$$\boldsymbol{R}_{XX}=(\sigma_{ij})_{p\times p}=\boldsymbol{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}] \tag{4}$$
$$\mu=E(\boldsymbol{X}),\sigma_{ij}=\boldsymbol{E}(\boldsymbol{X}_i\boldsymbol{X}_j)$$

设 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$ 及 $\boldsymbol{E}=(e_1,e_2,\cdots,e_p)$ 分别表示协方差矩阵的特征值以及相应的正交化单位特征向量。有

$$Var(\boldsymbol{Y}_i)=\boldsymbol{E}_i^{\mathrm{T}}\sum\boldsymbol{E}_i=\boldsymbol{E}\lambda_i\boldsymbol{E}_i^{\mathrm{T}}\boldsymbol{E}_i=\lambda_i \quad i=1,2,\cdots,p \tag{5}$$

$$Cov(\boldsymbol{Y}_i,\boldsymbol{Y}_j)=Cov\boldsymbol{E}_i^{\mathrm{T}}\sum\boldsymbol{E}_j=\lambda_j\boldsymbol{E}_i^{\mathrm{T}}\boldsymbol{E}_j=0 \quad i,j=1,2,\cdots,p \tag{6}$$

从以上的公式可知,各主分量是由样本集中各属性的线性组合而成的,它们之间互不相关。如果用 $\boldsymbol{Y}_1$ 来代替 $\boldsymbol{X}$ 的 p 个变量,这就要求 $\boldsymbol{Y}_1$ 尽可能地反映原来 p 个变量的信息,即使 $Var(\boldsymbol{Y}_1)=\boldsymbol{L}_1^{\mathrm{T}}\sum\boldsymbol{L}_1$ 达到最大,由此确定的 $\boldsymbol{Y}_1=\boldsymbol{L}_1\boldsymbol{X}$ 称为 $\boldsymbol{X}$ 的第 1 主分量。如果第 1 主分量还不足以表示原变量的信息,则可进一步地求 $Y_2,Y_3,\cdots,Y_m$,直到满足要求为止。此时称 Y_i 为 X 的第 i 个主分量。

在对正常数据进行特征提取的过程中,数据集中的一条记录就是描述某类正常行为特征的一个样本。由于原始数据集里包含着异常数据,若直接从中得到协方差矩阵,最后得到的主分量会受异常数据的影响而不准确,不能充分反映原变量的信息,因此,需要计算样本均值、样本协方差才能得到准确的 PCA 转换矩阵。

2.2 贡献率

不同的主分量对样本数据集的信息量表示程度是不同的,信息量的多少主要由主分量所对应的特征值 λ 来度量。特征值越大,说明特征空间中对应方向上分布的信息就越多,该主分量所包含的信息量就越大,所对应的特征向量的贡献率也就越大。由于 PCA 变换矩阵的特征值是按从大到小的顺序排列的,而主分量对应于特征值,因此主分量是按其重要性来排序的。第 i 个主分量的贡献率为:

$$\alpha_i=\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}\times 100\% \tag{7}$$

前 m 个分量对样本表达信息的充分程度用累计贡献率表示:

$$\phi(m)=\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_j}\times 100\% \tag{8}$$

$\phi(m)$的值越接近于 1,则前 m 个主分量对数据的表达越充分。主分量的个数与累计贡献率的大小有关,在实际应用中应大于 95%。

3 基于主分量分析和属性距离和的孤立点检测算法

数据集中,每个对象的属性值并非是纯粹的数字,它能反应出对象本身的一些状态。若是正常数据,它们对应的属性值都应分布在一定范围内,顺序相邻的两条记录间属性的距离小;而一个异常对象与正常对象在对应属性上的距离会很大。在一个属性上,同样数目的正常行为对象的属性值的距离和一定小于包含孤立点对象的属性值的距离和。基于以上分析,在主分量分析的基础上提出了一种基于属性距离和的孤立点检测算法。该算法省去了基于单元的算法对参数 p 和 d 设置,避免了结果对参数的敏感性,从而降低了对用户的要求。

该算法将经 PCA 矩阵转换到由主分量组成的新的特征空间中的待测数据集 D 和期望得到的孤立点数目 k 作为输入,输出 k 个被识别的孤立点。算法基本思想是:首先,将待测数据集中的数据都看作为非孤立点。其次,算法需要扫描 k 次来确定 k 个孤立点,即每一次扫描识别一个孤立点。在每一次扫描数据集时,对每一个非孤立点都做以下操作:(1)将一个非孤立点(一条记录)暂时移出数据集;(2)重新计算剩余记录的各相应属性的距离和(绝对值距离)。给每个属性距离和值乘一个权重,该权重是各主分量的贡献率。将加权后的属性距离和值再求和,其中与最小的属性距离和值相对应的移出数据集的记录即为孤立点。将该记录标记为孤立点,下次不再参与计算,一次扫描结束。当识别的孤立点数目达到 k 时,算法结束。基于主分量分析和属性距离和的孤立点检测算法见算法 1。

算法 1:Based_PCA_SumAttributeDistance(PCA_ SAD)

输入:数据集 D(已由主分量表示),孤立点个数 k

输出:k 个被识别的孤立点。

Begin

(1)由数据集 D 构造属性差值矩阵 $\boldsymbol{W}$ 及属性距离和数组S;

(2)对数据集进行一次扫描;

(3)count++;

(4) for every non-outlier

(5) 移出一条数据记录并重新计算剩余数据集的属性距离和;

(6) 每个属性距离和乘相应的权重;

(7) 标记与 mim{$\sum m$ 个主分量属性距离和}相对应移出数据集的点为孤立点;

(8)Repeat(2)直到识别出 k 个孤立点;

End

各主分量的贡献率表示了它们对训练数据集信息表达的

充分程度,在主分量上计算的属性距离和也同样只表示了原始数据集的一部分信息,因此用各主分量的贡献率作为每个属性距离和的权重是准确的。将各主分量加权后的属性距离和相加,实际上是每个主分量对该条记录影响的一个线性表达。它综合考虑了各主分量对数据的影响,不会使方差小的主分量受控于方差大的主分量,从全局的角度检测最优孤立点。

4　算法实验及结果分析

4.1　数据集描述

为了评价算法的有效性,本文在主频为 2.5 GHz、256 MB内存和 Windows XP Professional 的 PC 机上用 NHL 数据集测试了算法的性能。所选用的样本数据集 NHL 95-96[10]是美国曲棍球联合会 861 名职业球员的 1995~1996 年度统计数据,一共有 19 个属性,其中包括 3 个字符属性,16 个数值属性,实验中仅取能用距离度量的数据型属性。文献[11]已检测出该数据集中有两条记录是孤立点,因此样本数据集共有 859 条记录。在本实验中所选用的测试数据集为 NHL data'96-97[10]数据集,其中包含 847 条记录。

4.2　数据预处理

样本集中的属性是区间标度变量,他们具有不同的量纲,而取值的分散程度也较大,在计算过程中对相异度会有不同程度的影响,甚至会造成不合理的分类结果。为了消除不同量纲可能带来的负面影响,需要先对各分量进行标准化处理:

(1)计算第 f 个分量的平均绝对偏差 s_f:

$$s_f=\frac{1}{n}(|x_{1f}-m_f|+|x_{2f}-m_f|+\cdots+|x_{nf}-m_f|) \tag{9}$$

其中,$m_f=\frac{1}{n}(x_{1f}+x_{2f}+\cdots+x_{nf})$是第 f 个变量的均值。

(2)计算标准化的度量值(z-$score$)

$$z_{if}=\frac{x_{if}-m_f}{s_f} \tag{10}$$

经上述标准化后的样本数据集,区间标度变量都在同一量纲内,保证了实验结果的正确性。

4.3　实验结果

首先对样本数据集进行主分量提取,表 1 列出了 PCA 分析后的特征值、贡献率及累计贡献率。实验中将累积贡献率设为 95%。可以看出,前 2 个分量的累积贡献率已经达到95.758 7。

表 1　主分量特征

主分量	特征值(λ)	贡献率/%	累计贡献率/%
1	7 212.2	91.337 6	91.337 6
2	349.1	4.421 1	95.758 7
3	158.7	2.009 8	97.768 5
4	86.5	1.095 5	98.864 0
5	57.5	0.728 2	99.592 2
6	22.4	0.283 7	99.592 2
7	3.2	0.040 5	99.916 4
8	2.4	0.030 4	99.946 8
9	2.3	0.029 1	99.975 9
10	0.9	0.011 4	99.987 3
11	0.5	0.006 3	99.993 7
12	0.2	0.002 5	99.996 2
13	0.2	0.002 5	99.998 7
14	0.1	0.001 3	100.000 0
15	0	0.000 0	100.000 0
16	0	0.000 0	100.000 0

用提出的基于主分量分析和属性距离和的方法对NHL96-97 数据集进行孤立点检测。为简便起见,用 CB 代表基于单元的算法,用 PCA_SAD 代表提出的算法。结果如表 2 所示。

表 2　NHL 96-97 的实验结果

	属性距离和		孤立点
PCA_SAD	16 446.8		Sergei Zubov
	16 426.2		Matthew Barnaby
	p	d	孤立点
CB	0.997	6.1 165	Sergei Zubov
			Matthew Barnaby

实验结果表明,提出的孤立点检测算法与基于单元算法结果一致,而且不受数据维数的限制。同时,基于单元的算法对定义中 p 和 d 非常敏感,为了检测孤立点,用户需要通过多次试错来确定参数的合适设置,而本算法并不需要设置参数来返回指定个数的孤立点,从而避免了结果对参数的敏感性。其实用户可以任意指定孤立点的个数,来察看每个对象与其他对象距离的远近程度,检测结果只是提供给用户一个参考,哪些对象是孤立点还需由用户最终确定。

5　结束语

主分量分析已作为一种有效处理高维数据的数学方法被广泛应用。本文将其与基于距离的孤立点检测方法结合,提出了一种基于主分量分析和属性距离和的孤立点检测算法。该算法用较少的主分量表示最大的数据信息,能够达到对原始数据属性上的截断在均方差意义下为最优[7];将待测数据集投影到主分量组成的特征空间,用每条记录的各主分量的加权属性距离和来判断该条记录是否为孤立点。由于引入主分量分析方法,去掉了数据集的冗余属性,分析问题时会更加准确直观。而之后提出的基于属性距离和的方法,从全局的角度综合考虑了各主分量对数据的影响,使孤立点判断更为合理。实验结果表明,该算法省去了用户对参数的设置,提高了维数的扩展性,与经典的基于单元的孤立点检测方法同样有效。

参考文献:

[1] Hawkins D.Identification of outliers[M].London:Chapman and Hall,1980.

[2] Han Jiawei,Kanrber M.数据挖掘概念与技术[M].范明,孟小峰,等译.北京:机械工业出版社,2002:223-259.

[3] Knorr E M,Ng R T.Algorithms for mining distance-based outliers in large datasets[C]//Proc of Int Conf Very Large Databases(VLDB'98),New York,USA,1998:392-403.

[4] Knorr E M,Ng R T,Tucakov V.Distance-based outliers:Algorithms and applications[J].The VLDB Journal:Very Large Databases,2000,8(3-4):237-253.

[5] Bay S D,Schwabacher M.Mining distance-based outliers in near linear time with randomization and a simple pruning rule[C]//The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD2003),Washington,DC,USA,2003:29-38.

[6] Ghoting A,Parthasarathy S,Otey M.Fast mining of distance-based outliers in high dimensional datasets,Technical Report,TR71,CSE[R].The Ohio State University,2005:608-612.

(下转 243 页)