高维空间中的离群点发现*

魏 藜, 宫学庆, 钱卫宁, 周傲英

(复旦大学 计算机科学与工程系,上海 200433)
E-mail: {lwei,xqgong,wnqian,ayzhou}@fudan.edu.cn
http://www.fudan.edu.cn

摘要: 在许多 KDD(knowledge discovery in databases)应用中,如电子商务中的欺诈行为监测,例外情况或离群点的发现比常规知识的发现更有意义.现有的离群点发现大多是针对数值属性的,而且这些方法只能发现离群点,不能对其含义进行解释.提出了一种基于超图模型的离群点(outlier)定义,这一定义既体现了“局部”的概念,又能很好地解释离群点的含义.同时给出了 HOT(hypergraph-based outlier test)算法,通过计算每个点的支持度、隶属度和规模偏差来检测离群点.该算法既能够处理数值属性,又能够处理类别属性.分析表明,该算法能有效地发现高维空间数据中的离群点.

关 键 词: 数据挖掘;离群点;超图模型;聚类
中图法分类号: TP311 文献标识码: A

KDD(knowledge discovery in databases)是从大量数据中发现正确的、新颖的、潜在有用并能够被理解的知识的过程[1].现有的KDD研究大多集中于发现适用于大部分数据的常规模式.但在一些应用中,如电子商务和金融服务领域中的欺诈等犯罪行为监测,有关例外情况的信息比常规模式更有价值.目前,这样的研究正得到越来越多的重视.

KDD 中多数聚类算法(CLARANS[2],DBSCAN[3],BIRCH[4],STING[5],WaverCluster[6],DenClue[7],CLIQUE[8])能够发现一些例外情况.但是,因为聚类算法的主要目标是发现簇,而不是发现离群点(outlier),聚类算法或者对这些例外情况不敏感,或者忽视这些例外情况.最近,有一些研究是专门针对离群点发现的,例如文献[9~13].

现有的离群点发现方法大多是针对数值属性的,而且只能发现离群点,不能对其含义进行解释.本文提出了一种基于超图模型的离群点检测方法 HOT(hypergraph-based outlier test),它具有如下特点:

· 既能够处理数值属性,又能够处理类别(categorical)属性;

· 能有效并且高效地处理高维数据;

· 离群点是在“窗口”中定义的,而窗口中的其他点与该点有许多相似之处,既体现了数据的局部性,又体现了属性的局部性,同时也能很好地解释离群点的物理含义——正是窗口规定的这些属性造成了它的离群.

本文第 1 节简单介绍了超图聚类,传统的离群点发现方法和针对高维数据的离群点发现方法.第 2 节详细描述了发现离群点的问题,并给出了支持度、隶属度和规模偏差的定义.寻找离群点的具体算法步骤及算法复杂度分析在第 3 节中给出.第 4 节讨论 HOT 算法的特点.第 5 节总结全文,并给出了本文的后续工作.

* 收稿日期: 2001-04-20; 修改日期: 2001-09-20
基金项目: 国家自然科学基金资助项目(60003016;60003008);国家重点基础研究发展规划 973 资助项目(G1998030404)
作者简介: 魏藜(1978-),女,江西南昌人,硕士生,主要研究领域为数据挖掘技术;宫学庆(1974-),男,黑龙江饶河人,讲师,主要研究领域为 WEB 数据管理,数据挖掘;钱卫宁(1976-),男,浙江上虞人,博士生,主要研究领域为数据挖掘,Web 数据管理;周傲英(1965-),男,安徽人,博士,教授,博士生导师,主要研究领域为 Web 数据管理,数据挖掘.

1 相关工作

1.1 超图模型聚类

文献[14]提出了一种基于超图(hypergraph)模型的,对高维空间数据进行聚类的方法.该方法将数据集中的每一条记录看作超图中的一个点,把具有公共频繁项集的点归结到一条超边中,并用基于关联规则的概念来衡量超边的权重.因此,该方法能够将数据之间的关系映射到超图上,其中超边表示数据点之间的关系,超边的权重反映这种关系的强弱.建立了超图模型以后,使用超边分割方法,每次打断权重最小的超边,直到每个分割中的数据都密切相关为止,最终得到的分割就是聚类的结果.在进行超边分割的同时,用点与簇之间的连通度来评价得到的簇,因此可以有效地排除噪声数据对聚类结果的影响.

1.2 传统的寻找离群点的方法

到目前为止,离群点还没有一个正式的、为人们普遍接受的定义.Hawkin 的定义[15]揭示了离群点的本质:“离群点的表现与其它点如此不同,不禁让人怀疑它们是由另外一种完全不同的机制产生的.”(“An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.”)现有的发现离群点的方法大多建立在统计学的基础上,大致可以分为 4 类:基于分布的、基于深度的、基于距离的和基于密度的,每种方法都给出了自己的离群点的定义.

在基于分布的方法中,数据集满足的分布是已知的,根据该分布对数据集中的每个点进行“不一致性测试”(discordancy test),如果与分布不符合,就认为它是一个离群点.现有的“不一致性测试”方法有 100 多种[16],大部分只能对单变量分布(univariate)进行测试,这是基于分布的方法的一个缺点.尽管有些方法可以对多变量(multivariate)分布进行测试,也并没有使情况好转,因为对大多数 KDD 应用来说,数据集的多变量分布是无法预知的.基于分布的方法的另一个缺点是,对每个点进行测试,代价太大,而且有时并不能获得令人满意的结果.

在基于深度的方法中,每个数据对象被映射为 k 维空间中的一个点,并赋予了一个深度值(深度的定义有多种,参见文献[17,18]).深度小的数据对象是离群点的可能性比较大.基于深度的方法对二维和三维空间上的数据比较有效,但对四维及四维以上的数据,处理效率比较低.

基于距离的离群点的概念是由 Knorr 和 Ng 提出的[11,12].他们认为,如果一个点与数据集中大多数点之间的距离都大于某个阈值,就是一个离群点.这一定义包含并扩展了基于分布的思想,当数据集不满足任何标准分布时,基于距离的方法仍能有效地发现离群点.而且,这种方法能够处理任意维的数据.当空间维数比较高时,算法的效率比基于密度的方法要高.[19]对基于距离的离群点的概念进行了扩展,根据 k-最近邻距离对离群点进行排序,并给出了一种有效的方法,计算排在最前面的 n 个离群点.

基于密度的离群点的定义是在距离的基础上建立起来的.这种方法将点之间的距离和某一给定范围内点的个数这两个参数结合起来,得到“密度”的概念,根据密度来判断一个点是否是离群点.其他方法对离群点的定义都是二值的,即一个点或者是离群点、或者不是离群点,不存在中间状态.而文献[20,21]提出了离群点因子的概念,该因子定义了点的离群程度.同时,一个点的离群程度与它周围的点有关,这体现了“局部”的概念.

上述方法的重点都放在离群点的“识别”方面.实际上,识别出离群点之后,还需要进一步揭示离群点的含义——“为什么这个点是离群点?”“它与其他点到底有什么不同?”这才是用户关心的问题,也是我们寻找离群点的最终目的.文献[12]进行了这方面的研究.他们认为,一个点的某些属性与其它点有很大差异,就足以使它成为离群点了,关键在于发现是哪些属性导致了它的离群.根据内涵的不同,文献[12]将离群点分为强离群点和弱离群点两类,并给出了计算它们内涵信息的算法.

1.3 高维空间中的离群点发现

研究发现,高维数据的特性完全不同于低维数据,因此高维空间中的离群点发现方法势必不同于传统的离群点发现方法.

首先,与低维空间不同,高维空间中的数据分布得比较稀疏,这使得高维空间中数据之间的距离尺度及区域密度不再具有直观的意义.从一个数据点来看,其他点到它的距离落在一个范围很小的区间内,很难给出一个合

适的近似度阈值,来确定哪些点是与它相似的,哪些点不是[22].

为了解决这一问题,一些新的数据挖掘方法开始尝试将高维空间的数据投影到子空间中,在子空间中计算数据的相似度,这样能够获得比较有意义的相似度.如,文献[23]不再平等地看待各个维,而是通过某些标准选择出一些维,在这些维组成的子空间中寻找最近邻.文献[8]提出了一种在子空间中进行聚类的算法,能够对聚类的结果给出比较直观的解释.文献[24]提出了一种更广泛的子空间聚类方法,可以通过维度的任意组合来产生子空间,再将数据投影到子空间中进行聚类.

现有的研究表明,将数据投影到子空间再进行数据挖掘是可行的.但这带来了另一个问题——随着数据维数的增加,对维度进行组合得到的子空间个数呈指数级增长.因此我们不可能采用穷举法,对每一个可能的子空间进行投影,再从中选择效果最好的子空间,因为这样的计算代价太大.这时,如何有效地选择出最优的子空间就成了问题的关键.

文献[13]提出,为了使算法在高维空间上也有效,离群点发现算法必须满足以下要求:

(1) 能够有效地解决高维空间中数据的稀疏问题;

(2) 能够解释是什么原因造成了数据的异常;

(3) 能够找到合适的衡量办法,给出 k 维子空间中离群点的物理意义;

(4) 对高维空间中的数据仍然是计算高效的;

(5) 判断一个点是否为离群点时,要考虑到数据点的局部行为.

同时,依据这些标准,文献[13]提出了一种新的方法,通过观察数据投影后的密度分布来发现离群点.该方法用演化计算来寻找最优的子空间,并根据数据的特点对选择、交换、变异算子进行了调整,能够比较有效地找到高维空间中的离群点.

2 问题定义

2.1 高维数据集的超图模型

假定数据集 D 是一张包含 l 个属性 $A_1,A_2,\ldots,A_l$ 的关系表,共有 n 条记录.属性可以是类别属性,也可以是数值属性.(数值属性需预先作离散化.有很多算法可以对连续属性进行离散化,可参考文献[25])每个(属性名,属性值)称为项,(属性名,属性值)的集合称为项集.

将这 n 条记录映射为高维空间中的 n 个点(在不引起混淆的情况下,下文将不区分“记录”和“点”),并对其建立超图模型 $H=(V,E)$.其中,$V=\{v_1,v_2,\ldots,v_n\}$是超图的顶点集合,$E=\{e_1,e_2,\ldots,e_m\}$是超图模型的超边集合(设共有 m 条超边).

超图是对图的一种扩展,其中的每条超边可以连接两个或两个以上的顶点.超边是超图中一组顶点的集合,即超边 $e_i=\{v_{i_1},v_{i_2},\ldots,v_{i_{n_i}}\}$,其中 $v_{i_1},v_{i_2},\ldots,v_{i_{n_i}}\in V$.在我们的超图模型中,顶点集 V 对应于要测试的数据集,其中每个点 $v\in V$ 对应于数据集中 D 的一条记录,记录 v 包含的所有项的集合记为 $I(v)$.每条超边 $e\in E$ 对应于一组相关的项集,记为 $I(e)$.超边 e 包含的顶点集合记为 $P(e)$.超边 e 的支持数(support count)定义为

$$sc(e)=\left|\{v\mid v\in V,I(e)\subseteq I(v)\}\right|.$$

如果超边 e 的支持数 $sc(e)$大于给定的阈值 minsc,就称超边 e 为一个窗口,窗口实际上是一条频繁的超边.同样,$P(w)$表示窗口 w 包含的顶点的集合.

项集 I 上的关联规则表示了 I 的子集之间的关系,关联规则的形式是 $X\Rightarrow Y$,其中 $X\subseteq I,Y\subseteq I$.

规则的支持度(support)定义为 $sup(X\Rightarrow Y)=\dfrac{sc(X\cup Y)}{\left|\{v\mid I\subseteq I(v)\}\right|}$.

规则的置信度(confidence)定义为 $conf(X\Rightarrow Y)=\dfrac{sc(X\cup Y)}{sc(X)}$.

给定超图 H 中的一条超边 e,可以找出该超边 e 包含的项集 $I(e)$上的关联规则,这些规则的支持度和置信度分别大于给定的阈值 minsup 和 min$Conference$ 超边 e 上所有满足这些条件的关联规则记为 $R(e)$.

得到了超边 e 上所有满足这些条件的关联规则 $R(e)$后,即可定义超边 e 的权重:

超边 e 的权重(weight)定义为 $wei(e)=\frac{\sum_{r} conf(r)}{|R(e)|}$,其中 $r\in R(e)$.

在超图模型中,超边表示数据点之间的关系,超边的权重反映了这种关系的强弱.

得到了超图模型及超边的权重后,可以对数据集 D 中的 n 条记录进行聚类(聚类方法将在第 3.1 节中详细介绍),假定共得到 q 个簇 $C_1,C_2,...,C_q$.$P(C)$表示簇 C 包含的顶点的集合,$W(C)$表示簇 C 包含的顶点能够构成的窗口的集合.

2.2 定 义

定义 **1**. 点(簇)对窗口的支持度(support).

设 w 为一个窗口,v 为 w 中的一个点,即 $v\in P(w)$;v 所属的簇记为 C_v,即 $v\in P(C_v)$,则点 v(簇 C_v)相对于窗口 w 的支持度定义为

$$S_w(v)=S_w(C_v)=\frac{|P(w)\cap P(C_v)|}{|P(w)|}.$$

定义 **2**. 点对簇的隶属度(belongingness).

设有顶点 v,v 所属的簇记为 C_v,即 $v\in P(C_v)$,则点 v 相对于簇 C_v 的隶属度定义为

$$B_{C_v}(v)=\frac{|\{e\mid e\in W(C_v),v\in P(e)\}|}{|W(C_v)|}.$$

定义 **3**. 点(簇)对窗口的规模偏差(deviation of size).

设窗口 w 中的点共属于 t 个簇 $C_1,C_2,...,C_t$,窗口 w 中的点 v 所属的簇记为 C_v,即 $v\in P(C_v)$,则点 v(簇 C_v)相对于窗口 w 的规模偏差定义为

$$D_w(v)=D_w(C_v)=\frac{|P(w)\cap P(C_v)|-\frac{|P(w)|}{t}}{\sqrt{\sum_{i=1}^{t}(|P(w)\cap P(C_i)|-\frac{|P(w)|}{t})^2}}.$$

定义 **4**. 离群点(outlier).

给定窗口 w 及其中的顶点 $v\in P(w)$,设 v 属于簇 C_v,给定阈值 S_t,B_t,D_t,如果 $S_w(v)<S_t,B_{Cv}(v)>B_t$, $D_w(v)<D_t$,则称 v 是一个相对于窗口 w 的离群点.

定义 **5**. 噪声(noise).

给定窗口 w 及其中的顶点 $v\in P(w)$,设 v 属于簇 C_v,给定阈值 S_t,B_t,D_t,如果 $S_w(v)<S_t,B_{Cv}(v)>B_t$, $D_w(v)>D_t$,则称 v 是一个相对于窗口 w 的噪声.

2.3 解 释

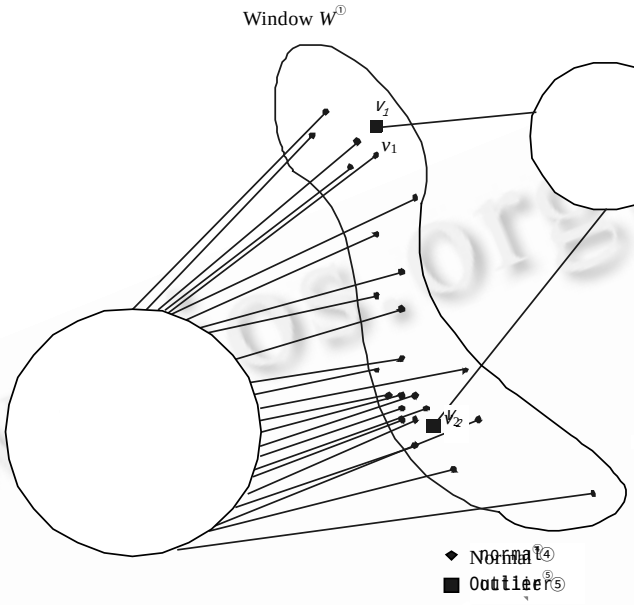
首先,我们解释点(簇)对窗口的支持度、点对簇的隶属度和点(簇)对窗口的规模偏差的实际意义.

设 w 为一个窗口,v 为 w 中的点,即 $v\in P(w)$,v 所属的簇记为 C_v,即 $v\in P(C_v)$,则点 v(簇 C_v)相对于窗口 w 的支持度表示窗口 w 中属于簇 C_v 的点的个数与窗口 w 中所有点的个数之比.支持度越大,说明属于簇 C_v 的点在窗口 w 中占的比例越大.点 v 对簇 C_v 的隶属度 $B_{C_v}(v)$ 表示簇 C_v 中包含点 v 的频繁超边的条数与簇 C_v 中所有频繁超边的条数之比.隶属度越大,说明点 v 与簇 C_v 之间的频繁超边越多,点 v 与簇 C_v 之间的联系越紧密.在点 v(簇 C_v)相对于窗口 w 的规模偏差的定义中,$\frac{|P(w)|}{t}$ 表示的是:如果窗口 w 中的点平均分布在 t 个簇中,每个簇中的顶点个数.点 v(簇 C_v)相对于窗口 w 的规模偏差表示窗口 w 中属于簇 C_v 的点的个数与这一平均值之间的差异.规模偏差越小,说明窗口 w 中属于簇 C_v 的点异常的少,与其它点的相异性越大.

了解了这 3 个参数的含义后,我们可以分析在此基础上给出的离群点和噪声的定义.

给定窗口 w 及其中的顶点,设 v 属于簇 C_v,如果点 v(簇 C_v)相对于窗口 w 的支持度 $S_w(v)<S_t$,说明属于簇

C_v的点在窗口 w 中占的比例比较小;点 v 对簇 C_v 的隶属度 $B_{C_v}(v) > B_t$,说明点 v 与簇 C_v 之间的联系很紧密,将点 v 归入簇 C_v 是合理的;当规模偏差的阈值 D_t 取值为一个大小适当的负数时,若点 v(簇 C_v)相对于窗口 w 的规模偏差 $D_w(v) < D_t$,说明窗口 w 中属于簇 C_v 的点比平均值少很多,即,窗口 w 中的点 v 属于簇 C_v,但窗口 w 中的大多数点都不属于簇 C_v,而是集中地属于一个或几个簇,并且窗口 w 中簇 C_v 的规模比其他簇的规模小很多,如图 1 所示(图 1 给出了一种极端情况,除了点 v_1, v_2 外,窗口中的其他点都属于另一个簇),因此我们认为这样的点 v 相对于窗口 w 中的其他点来说,是一个离群点.



①窗口 w,②簇 C_v,③簇 C',④正常点,⑤离群点.

Fig.1 Outliers in Window w

图 1 窗口 w 中的离群点

同理,如果点 v(簇 C_v)相对于窗口 w 的支持度 $S_w(v) < S_t$,点 v 对簇 C_v 的隶属度 $B_{C_v}(v) > B_t$,而点 v(簇 C_v)相对于窗口 w 的规模偏差 $D_w(v) > D_t$,说明尽管属于簇 C_v 的点在窗口 w 中占的比例比较小,点 v 也确实应该归入簇 C_v 中,但窗口 w 中簇 C_v 的规模与其他簇的规模差不多(因为 $S_w(v) < S_t$,所以可以排除属于簇 C_v 的点的个数远大于平均值的情况),也就是说,窗口 w 中的点分属于许多不同的簇,如图 2 所示,我们说这样的点 v 相对于窗口 w 来说是噪声.

根据以上定义得到的离群点是局部的.以上讨论的离群点和噪声都是在窗口范围中进行的.前面说过,窗口实际上是一条频繁的超边,而超边对应于一组相关属性值集合,窗口中的点都是具有这些属性值的,这使得它们具有可比性,因此可以把它们作为一个讨论的整体,找到的离群点是相对于窗口这个局部点集合的.同时,由于高维空间中距离是没有意义的,所以不能基于距离来寻找离群点.而根据以上定义找到的离群点仅考虑窗口中的属性,所以能够较好地应用到高维空间上.

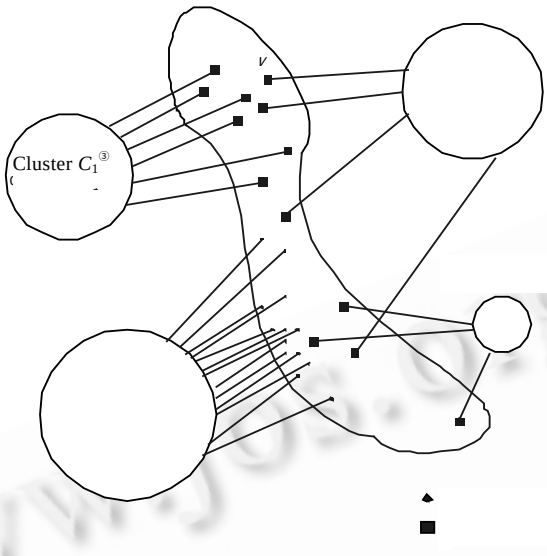
同时要说明的是,我们只在窗口中寻找离群点.因为一条超边,如果它不是窗口,即不是一条频繁超边,说明它包含的点不够多,我们对这样的点的集合不感兴趣.

简而言之,我们要解决的问题是对数据集中的每个窗口,找出其中的离群点.

3 HOT 算法

我们的寻找离群点的 HOT(hypergraph-based outlier test)算法对高维空间中的每个点进行测试,以确定它是不是离群点.HOT 算法主要分为 3 个步骤:第 1,为数据集中的记录建立超图;第 2,根据超图对数据点聚类;第 3,测试数据集中的每一点,判断它是否是相对于某个窗口的离群点.下面我们详细解释每个步骤的具体执行过程

并分析算法复杂度.



①窗口 w,②簇 C_v,③簇 C_1,④簇 C_2,⑤簇 C_3.

Fig. 2 Noise in window w

图 2 窗口 w 中的噪声

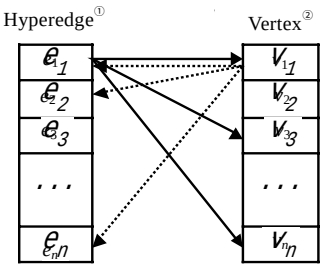
3.1 算法步骤

步骤 1. 建立超图模型

假定数据集是一张包含 l 个属性 $A_1,A_2,...,A_l$ 的关系表,共有 n 条记录.我们的目标是找出所有频繁的超边(窗口),即找出所有支持数大于 minsc 的超边.得到频繁超边之后,超边对应的项集也确定了.对每条超边 e,找到其项集 $I(e)$中包含的支持度和置信度分别满足 minsup 和 min$conf$ 的关联规则,并计算每条超边的权重.

寻找频繁超边的问题实际上是数据挖掘中经典的计算大项集的问题.目前有很多算法可以用来计算大项集,其中 Apriori 算法[26]是比较有效的一种.假定我们使用 Apriori 算法,可以得到数据集中所有的大项集,每个大项集就是一条频繁超边,所有具有大项集中的属性值的点都属于这条超边.已知超边,找关联规则的问题也能用文献[26]中的方法解决,这里不再赘述.

在计算出超边及超边上的关联规则后,即可得超边的权重.用如图 3 所示的结构记录下每条超边包含哪些点,每个点属于哪些超边以及每条超边的权重等信息.



①超边,②点.

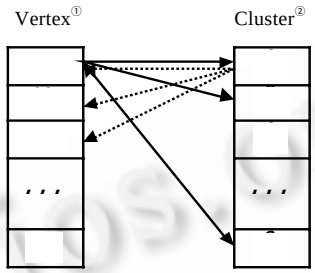
Fig. 3 The relationship between hyperedges and vertices

图 3 超边与点的对应关系

步骤 2. 聚类

得到超图以后,就得到了各个数据点之间的联系(用超边表示),我们要在此基础上对数据点进行聚类.聚类方法使用超图分割算法 HMETIS[27],每次将超图分成两部分,并保证被截断的超边的权重最小.超边的权重越小,说明超边表示的关系越不重要.反复使用超图分割算法,直到每个分割内部都紧密联系为止,得到的分割就是簇.具体的基于超图的聚类算法参见文献[14].

聚类完毕后,超图中的每个点都被分到且仅被分到一个簇中去,这时,可以用如图 4 所示的结构记录每个点属于哪个簇,每个簇包含哪些点以及每个簇对应的频繁超边等信息.



①点,②簇.

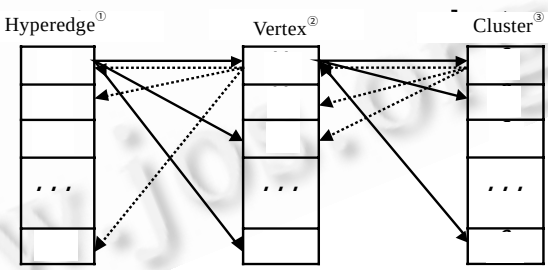
Fig. 4 The relationship between vertices and clusters

图 4 点与簇的对应关系

步骤 3. 测试离群点

对超图中的每一个窗口 w 及窗口中的每一个点 v,设 v 属于簇 C_v,计算点 v(簇 C_v)相对于窗口 w 的支持度 $S_w(v)$,点 v 对簇 C_v 的隶属度 $B_{C_v}(v)$ 以及点 v(簇 C_v)相对于窗口 w 的规模偏差 $D_w(v)$,如果 $S_w(v)<S_t$,$B_{Cv}(v)>B_t$,$D_w(v)<D_t$ 就认为 v 是一个相对于窗口 w 的离群点.

前两个步骤结束以后,可以得到如下所示的图 5,所以支持度、隶属度和规模偏差都很容易计算.下面我们分别进行说明.



①超边,②点,③簇.

Fig. 5 The relationship among hyperedges, vertices and clusters

图 5 超边-点-簇的对应关系

点 v 相对于窗口 w 的支持度 $S_w(v)=S_w(C_v)=\dfrac{|P(w)\cap P(C_v)|}{|P(w)|}$.其中$|P(w)|$是窗口 w 中点的个数,这在第 1 步计算超边时已经得到了,并存在了相应的超边节点上.$|P(w)\cap P(C_v)|$ 是既属于窗口 w 又属于簇 C_v 的点的个数,容易通过扫描窗口 w 包含的点得到.

点 v 相对于簇 C_v 的隶属度 $B_{C_v}(v)=\dfrac{|\{e\,|\,e\in W(C_v),v\in P(e)\}|}{|W(C_v)|}$.其中$|W(C_v)|$是簇 C_v 中频繁超边的条数,这在第 2 步聚类时已经得到了,并存在了相应的簇节点上.$|\{e\,|\,e\in W(C_v),v\in P(e)\}|$ 是点 v 与簇 C_v 之间频繁超边的条数,从

图 5 来看,只要通过点 v 得到包含点 v 的超边,再判断有多少条超边属于簇 C_v 对应的频繁超边集即可.

点 v 相对于窗口 w 的规模偏差 $D_w(v)=D_w(C_v)=\dfrac{\left|P(w)\cap P(C_v)\right|-\dfrac{|P(w)|}{t}}{\sqrt{\sum_{i=1}^{t}(\left|P(w)\cap P(C_i)\right|-\dfrac{|P(w)|}{t})^2}}$.其中$|P(w)|$是窗口 w 中点的个数,如前所述,可以从相应的超边节点上得到.t 是窗口 w 的点属于的簇的个数,通过扫描窗口 w 中的点可以统计出来.$\left|P(w)\cap P(C_v)\right|$是窗口 w 中属于簇 C_v 的点的个数,扫描窗口 w 中的每个点,判断它是否属于簇 C_v 即可.

通过以上分析,可知每个点的 3 个参数都能够通过扫描图 5 的结构方便地得到.得到了这 3 个参数以后,根据它们与阈值之间的关系,就可以判断这个点是不是离群点了.

3.2 算法复杂度分析

HOT 算法包含 3 个步骤,下面我们分别对这 3 步的复杂度进行分析.

研究表明,已经有大量算法能够高效地从大规模数据集中得到频繁项集以及项集上满足给定支持度和置信度阈值的关联规则.当给定的支持度阈值足够大时,Apriori 算法能够快速地找到超大规模数据库中的关联规则.当然,如果支持度阈值降低,计算量将大幅度上升.

文献[27]对超图分割算法 HMETIS 进行了研究.进行 k-way 分割时,HMETIS 算法的复杂度是 $O((V+E)\log K)$,其中 V 是顶点个数,E 是超边条数.这里,顶点个数等于数据集中记录的条数,超边条数等于满足给定的 minsc 的大项集的个数.注意到,大项集的个数(即超边条数)并不与记录条数(即顶点个数)成正比.

接下来是计算每个点的 3 个参数值.详细的计算过程在第 3.1 节中已经介绍过,这里主要分析计算复杂性.点 v(簇 C_v)相对于窗口 w 的支持度 $S_w(v)=S_w(C_v)=\dfrac{\left|P(w)\cap P(C_v)\right|}{|P(w)|}$.其中$|P(w)|$是窗口 w 中点的个数,在第 1 步计算超边时已经得到了.$\left|P(w)\cap P(C_v)\right|$ 是既属于窗口 w 又属于簇 C_v 的点的个数,要通过扫描图 5 得到,代价是 $O(max\{|P(w)|\})$.点 v 相对于簇 C_v 的隶属度 $B_{C_v}(v)=\dfrac{\left|\{e\mid e\in W(C_v),v\in P(e)\}\right|}{\left|W(C_v)\right|}$,该值在 HMETIS 算法聚类的过程中已经计算过.从 n 个顶点中查找一个给定顶点对其所属簇的隶属度的复杂度为 $O(n)$.点 v(簇 C_v)相对于窗口 w 的规模偏差 $D_w(v)=D_w(C_v)=\dfrac{\left|P(w)\cap P(C_v)\right|-\dfrac{|P(w)|}{t}}{\sqrt{\sum_{i=1}^{t}(\left|P(w)\cap P(C_i)\right|-\dfrac{|P(w)|}{t})^2}}$,其中$|P(w)|$是窗口 w 中点的个数,如前所述,可以从相应的超边节点上得到.t 是窗口 w 的点属于的簇的个数,可以通过扫描窗口 w 中的点统计,因此计算 $\dfrac{|P(w)|}{t}$ 的复杂度为 $O(\max\{|P(w)|\})$.而 $\left|P(w)\cap P(C_v)\right|$ 是窗口 w 中属于簇 C_v 的点的个数,聚类后已知,并且对窗口中的每个簇 C_i,$\left|P(w)\cap P(C_i)\right|$ 只需计算一次.因此计算规模偏差的复杂度为 $O(\max\{|P(w)|\})$.

显然,以上的计算对每个窗口和每个簇都要且只要进行一次.因此,HOT 算法的总的计算代价是 $O(|w|\cdot|C_v|\cdot\max\{|P(w)|\}+n)$,其中$|w|$是模型中窗口的个数,$|C_v|$是簇的个数,$n$ 是模型中顶点的个数.注意到,$|w|$,$|C_v|$和 $\max\{|P(w)|\}$都不会太大,因此算法的效率较高.

4 讨 论

根据我们提出的新的离群点定义以及 HOT 算法,对照文献[13]提出的 5 个标准,我们认为该方法能够有效地解决高维空间中的离群点发现问题.

首先,算法并没有在整个空间上寻找离群点,而是在窗口中寻找离群点.而窗口是频繁的超边,这保证了在窗口决定的子空间中,数据是相对稠密的.因此算法能够有效地解决高维空间中数据分布比较稀疏的问题.

其次,离群点是在窗口中定义的,窗口中的点具有某些共同的属性,这些共同属性提供了一个观察离群点的

视角.聚类的结果造成了点的离群.因此算法能够有效地解释数据异常的原因.

第三,对窗口中的每个点,算法都计算它对窗口的支持度、对簇的隶属度和对窗口的规模偏差.根据第 2.3 节的解释可知,这 3 个参数有着明显的物理意义,能对离群点的含义做出合理的解释.

第四,通过第 3.2 节的算法分析可以看到,我们的算法对高维空间中的数据仍然是计算高效的.

最后,但同样重要的是,算法在判断一个点是否为离群点时,考虑了数据点的局部行为.离群点是在窗口中定义的,而窗口中的其他点与该点有许多相似之处,这既体现了数据的局部性,又体现了属性的局部性,同时也能很好地解释离群点的实际意义.

通过以上讨论可以看出,我们的算法满足文献[13]提出的 5 个标准,是一种寻找高维空间中的离群点的有效方法.

5 结 论

本文提出了一种新的基于超图模型的离群点定义.离群点是在超图模型的“窗口”(即频繁超边)中定义的,窗口中的点有着紧密的联系.在窗口中寻找离群点,一方面体现了“局部”的概念(该点相对于与它相似的其他点来说是离群点),一方面也能很好地解释离群点(正是窗口规定的这些属性造成了它的离群).同时,本文还给出了基于超图模型的寻找高维数据库中离群点的 HOT 算法,该算法既能够处理数值属性,也能够处理类别属性.分析表明,HOT 算法能够有效地寻找高维空间中的离群点.进一步的工作包括验证算法在真实空间上的有效性,研究我们的离群点定义与文献[13]介绍的离群点定义之间的关系,以及探索针对大规模数据的改进算法.

References:

[1] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. Knowledge discovery and data mining: towards a unifying framework. In: Simoudis, E., Han, J., Fayyad, U.M., eds. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press, 1996. 82~88.

[2] Ng, R. T., Han, J. Efficient and effective clustering methods for spatial data mining. In: Bocca, J.B., Jarke, M., Zaniolo, C., eds. Proceedings of the 20th International Conference on Very Large Data Bases. Santiago: Morgan Kaufmann, 1994. 144~155.

[3] Ester, M., Kriegel, H.-p., Sander, J., *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad, U.M., eds. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press, 1996. 226~231.

[4] Zhang, T., Ramakrishnan, R., Linvy, M. BIRCH: an efficient eata clustering method for very large databases. In: Jagadish, H.V., Mumick, I.S., eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Montreal: ACM Press, 1996. 103~114.

[5] Wang, W., Yang, J., Muntz, R. STING: a statistical information grid approach to spatial data mining. In: Jarke, M., Carey, M.J., Dittrich, K.R., *et al*., eds. Proceedings of the 23rd International Conference on Very Large Data Bases. Athens, Greece: Morgan Kaufmann, 1997. 186~195.

[6] Sheikholeslami, G., Chatterjee, S., Zhang, A. WaveCluster: a multi-resolution clustering approach for very large spatial databases. In: Gupta, A., Shmueli, O., Widom, J., eds. Proceedings of the 24th International Conference on Very Large Data Bases. New York : Morgan Kaufmann, 1998. 428~439.

[7] Hinneburg, A., Keim, D.A. An efficient approach to clustering in large multimedia databases with noise. In: Agrawal, R., Stolorz, P.E., Piatetsky-Shapiro, G. eds. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998. 58~65.

[8] Agrawal, R., Gehrke, J., Gunopulos, D., *et al*. Automatic subspace clustering of high dimensional data for data mining applications. In: Haas, L.M., Tiwary, A., eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Seattle, Washington, D C: ACM Press, 1998. 94~105.

[9] Ruts, I., Rousseeuw, P. Computing depth contours of bivariate point clouds. Journal of Computational Statistics and Data Analysis, 1996,23:153~168.

[10] Arning, A., Agrawal, R., Raghavan, P. A linear method for deviation detection in large databases. In: Simoudis, E., Han, J., Fayyad, U.M., eds. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press, 1996. 164~169.

[11] Knorr, E.M., Ng, R.T. Algorithms for mining distance-based outliers in large datasets. In: Gupta, A., Shmueli, O., Widom, J., eds. Proceedings of the 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann, 1998. 392~403.

[12] Knorr, E.M., Ng, R.T. Finding intensional knowledge of distance-based outliers. In: Atkinson, M.P., Orlowska, M.E., Valduriez, P., eds. Proceedings of the 25th International Conference on Very Large Data Bases. Edinburgh, Scotland: Morgan Kaufmann, 1999. 211~222.

[13] Aggarwal, C.C., Yu, P. Outlier detection for high dimensional data. In: Aref, W.G., eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Santa Barbara, CA: ACM Press, 2001. 37~47.

[14] Han, E. H., Karypis, G., Kumar, V., *et al*. Clustering in a high-dimensional space using hypergraph models. Technical Report, TR-97-063, Minneapolis: Department of Computer Science, University of Minnesota, 1997.

[15] Hawkins, D. Identification of Outliers . London: Chapman and Hall, 1980.

[16] Barnett V., Lewis T. Outliers in Statistical Data. New York: John Wiley and Sons, Inc., 1994.

[17] Tukey, J.W. Exploratory Data Analysis. MA: Addison-Wesley, 1977.

[18] Preparata, F., Shamos, M. Computational geometry: an introduction. New York: Springer-Verlag, 1988.

[19] Ramaswamy, S., Rastogi, R., Kyuseok, S. Efficient algorithms for mining outliers from large data sets. In: Chen, W., Naughton, J.F., Bernstein, P.A., eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas: ACM Press, 2000. 427~438.

[20] Breunig, M.M., Kriegel, H.P., Ng, R.T., *et al*. OPTICS-OF: identifying local outliers. In: Zytkow, J.M., Rauch, J., eds. Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science 1704, Prague: Springer, 1999. 262~270.

[21] Breunig, M. M., Kriegel, H. P., Ng, R. T., *et al*. LOF: identifying density-based local outliers. In: Chen, W., Naughton, J.F., Bernstein, P.A., eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas: ACM Press, 2000. 93~104.

[22] Beyer, K., Goldstein, J., Ramakrishnan, R., *et al*. When is nearest neighbors meaningful? In: Beeri, C., Buneman, P., eds. Proceedings of the 7th International Conference on Data Theory. Lecture Notes in Computer Science 1540, Jerusalem: Springer, 1999. 217~235.

[23] Hinneburg, A., Aggarwal, C.C., Keim, D.A.. What is the nearest neighbor in high dimensional spaces? In: Abbadi, A.E., Brodie, M.L., Chakravarthy, S., *et al*. eds. Proceedings of the 26th International Conference on Very Large Data Bases. Cairo: Morgan Kaufmann, 2000: 506~515.

[24] Aggarwal, C. C., Yu, P. Finding generalized projected clusters in high dimensional spaces. In: Chen, W., Naughton, J.F., Bernstein, P.A., eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas: ACM Press, 2000. 70~81.

[25] Dougherty, J., Kohavi, R., Sahami, M. Supervised and unsupervised discretization of continuous features. In: Prieditis, A., Russell, S.J., eds. Proceedings of the 12th International Conference on Machine Learning. Tahoe, CA: Morgan Kaufmann, 1995. 194~202.

[26] Agrawal, R., Srikant, R. Fast algorithms for mining association rules in large databases. In: Bocca, J.B., Jarke, M., Zaniolo, C., eds. Proceedings of the 20th International Conference on Very Large Data Bases. Santiago: Morgan Kaufmann, 1994. 487~499.

[27] Karypis, G., Aggarwal, R., Kumar, V., *et al*. Multilevel hypergraph partitioning: application in VLSI design. In: Proceedings of the ACM/IEEE Design Automation Conference. Anaheim, CA: ACM Press, 1997. 526~529.

Finding Outliers in High-Dimensional Space*

WEI Li, GONG Xue-qing, QIAN Wei-ning, ZHOU Ao-ying

(*Computer Science and Engineering Department, Fudan University, Shanghai* 200433, *China*)

E-mail: {lwei,xqgong,wnqian,ayzhou}@fudan.edu.cn

http://www.fudan.edu.cn

Abstract: For many KDD (knowledge discovery in databases) applications, such as fraud detection in E-commerce, it is more interesting to find the exceptional instances or the outliers than to find the common knowledge. Most existing work in outlier detection deals with data with numerical attributes. And these methods give no explanation to the outliers after finding them. In this paper, a hypergraph-based outlier definition is presented, which considers the locality of the data and can give good explanation to the outliers, and it also gives an algorithm called HOT (hypergraph-based outlier test) to find outliers by counting three measurements, the support, belongingness and deviation of size, for each vertex in the hypergraph. This algorithm can manage both numerical attributes and categorical attributes. Analysis shows that this approach can find the outliers in high-dimensional space effectively.

Key words: data mining; outlier; hypergraph model; clustering

* Received April 20, 2001; accepted September 20, 2001

Supported by the National Natural Science Foundation of China under Grant Nos.60003016, 60003008; the National Grand Fundamental Research 973 Program of China under Grant No.G1998030404

第 1 届计算机图形学与空间信息系统应用国际学术会议

通　　知

由中国国家自然科学基金委员会、中国科学院地理科学与资源研究所、北京大学、浙江大学、国家遥感中心农业应用部、中国农学会计算机农业应用分会、中国自动化学会计算机图形学及辅助设计专业委员会联合组织的首届计算机图形学与空间信息系统应用国际学术会议将于 2002 年 8 月 6 日~9 日在北京召开。本次会议是首次技术学科与应用学科同台交流,是一个多学科综合交流的国际学术会议。其目的是让应用学科走技术学科研究成果的捷径,跨越前沿,高水平地发展;让技术学科从应用研究需求探索新的技术与方法研究。会议将评选优秀论文在相关国家一级刊物上发表。本次会议同时设有会议成果展览。

详情请查看会议网页:http://www.cgconference2002.com

一、联系人

(1) 陈宝雯

电话: 86-10-64889810; 86-10-64877339;　　传真: 86-10-64854230

E-mail: bwchen@cgconference2002.com; bwchen@cern.ac.cn

(2) 丛升日

地址: 北京大学计算机科学技术系　　邮政编码: 100871

电话: 86-10-62754939; 传真: 86-10-62754939　　E-mail: srcong@263.net

二、会议筹备处

中国科学院地理科学与资源研究所;北京大学