View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

基于密度的DBSCAN聚类算法的研究及应用

冯少荣[1,2],肖文俊[1]
FENG Shao-Rong[1,2],XIAO Wen-Jun[1]

1.华南理工大学 计算机科学与工程学院,广州 510640
2.厦门大学 信息科学与技术学院,福建 厦门 361005
1.School of Computer Science and Engineering,South China University of Technology,Guangzhou 510641,China
2.College of Information Science and Technology,Xiamen University,Xiamen,Fujian 361005,China
E-mail:shaorong@xmu.edu.cn

FENG Shao-Rong,XIAO Wen-Jun.Research and application of DBSCAN clustering algorithm based on density.Computer Engineering and Applications,2007,43(20):216-221.

Abstract:This paper first researches DBSCAN clustering algorithm,and analyzes characteristics and existing problems of the DBSCAN algorithm and improved idea.Evaluation method of the traffic accident black spots and an improved thought based on DBSCAN are proposed.In order to illuminate course of processing and feasibility,an example is presented.The experimental result demonstrates that this paper method can greatly enhance the working efficiency of evaluation of the traffic accident black spots.
Key words:clustering analysis;Density Based Spatial Clustering of Applications with Noise(DBSCAN);prone location of traffic;data mining

摘 要:首先对DBSCAN(Density Based Spatial Clustering of Applications with Noise)聚类算法进行了深入研究,分析了它的特点、存在的问题及改进思想,提出了基于DBSCAN方法的交通事故多发点段的排查方法及其改进思路,并且给出了实例以说明处理过程及可行性。实验结果表明本文提出的方法可以大大提高交通事故黑点排查效率。
关键词:聚类分析;DBSCAN;交通事故多发点(段);数据挖掘
文章编号:1002-8331(2007)20-0216-06 文献标识码:A 中图分类号:TP311.13

1 引言

交通事故多发点段可以理解为一条交通道路上发生交通事故密度大的地方[10-17,44-52],DBSCAN算法是基于密度的聚类分析[2-4,6,8-9,18-39]算法。应用在交通事故多发点段的排查中就是基于交通事故密度的交通黑点查找的聚类分析算法,而基于密度的DBSCAN算法[1-9,40-43]基本思想就是通过不断地搜索临近点来使核对象周围的密度逐渐增加,寻找到一个区域内所查找点或对象密度大的地方。算法中所要研究的点可以描述为交通事故发生的地点,对于算法中的ε-近邻区域可以理解为道路的公里数,因此DBSCAN算法在道路交通事故多发点段的智能排查上就可以理解为排查在半径为ε公里内发生MinPts以上交通事故的地点或者路段。这也和我国对于交通事故多发点段的规定不谋而合。所以可以采用DBSCAN技术的方法对交通事故多发点段进行排查。

2 基于DBSCAN算法的交通事故多发点段排查方法

2.1 核心思想[10-17,44-52]

基于DBSCAN算法的交通事故多发点段排查方法的思想是:对于构成交通事故多发点段的每个交通事故,其发生的地点半径ε(邻域)公里范围以内的其它交通事故的个数,必须不小于一个给定的阈值MinPts),也就是说其邻域的密度必须不小于某个阈值。

下面是DBSCAN算法的交通事故多发点段排查方法在交通事故黑点排查中的一些定义:

(1)定义1(核心交通事故点)给定ε,MinPts,若交通事故点p的ε邻域包含的交通事故对象个数|Nε(p)|≥MinPts,则称p是核心交通事故点。

(2)定义2(直接密度可达)给定ε,MinPts,事故p是从事故q出发直接密度可达的,当:

①p∈Nε(q);

②Nε(q)≥MinPts。

(3)定义3(密度可达)给定一个交通事故集合D,当存在一个事故对象链$p_1,p_2,\cdots,p_n,p_1=q,p_n=p$,对$p_i\in D$,$p_{i+1}$是$p_i$关于ε和MinPts直接密度可达的,则称事故对象p从事故对象q关于ε和MinPts密度可达(非对称)。

(4)定义4(密度相连)如果事故对象集合D中存在一个

基金项目:福建省自然科学基金(the Natural Science Foundation of Fujian Province of China under Grant No.A0310008);福建省高新技术研究开放计划重点项目(No.2003H043)。
作者简介:冯少荣(1964-),男,副教授,在职博士研究生,主要研究方向:并行分布数据库、数据仓库、数据挖掘;肖文俊(1950-),男,教授,博士生导师,主要研究方向:网络理论和并行分布式算法,网络和复杂系统及其应用,并行分布式计算及其应用,大规模数据处理。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

事故对象 o,使得事故对象 p 和 q 是从 o 关于 ε和 MinPts 密度可达的,那么事故对象 p 和 q 关于 ε和 MinPts 密度相连(对称)。

(5)簇和噪声):基于密度可达性的最大的密度相连事故对象的集合称为交通事故黑点,不属于任何黑点中的事故对象被认为是'噪声'。

DBSCAN 检查数据库中每个点的 ε-近邻。若一个事故对象 P 的 ε-近邻包含多于 MinPts 个事故点,就要创建包含 P 的新聚类。然后 DBSCAN 根据这些核对象,循环收集'直接密度可达'的对象,其中可能涉及进行若干'密度可达'聚类的合并。当各聚类再无新点(对象)加入时聚类进程结束,最后得到的类就是交通黑点。

2.2 方法的可行性论证

2.2.1 DBSCAN 算法的特点[40-43]

DBSCAN 算法是一种基于密度的空间聚类算法。该算法利用基于密度的聚类(或者类 cluster)概念,即要求聚类空间中的一定区域内所包含对象(点或其它空间对象)的数目不小于某一给定阈值。DBSCAN 算法的显著优点是聚类速度快,且能够有效处理噪声点(outliers)和发现任意形状的空间聚类。但是,由于它直接对整个数据库进行操作,且进行聚类时使用了一个全局性的表征密度的参数,因此也具有两个比较明显的弱点[40]:

(1)当数据量增大时,要求较大的内存支持,I/O 消耗也很大;

(2)当空间聚类的密度不均匀,聚类间距离相差很大时,聚类质量较差。

为了找到一个类,DBSCAN 从 D 中找到任意对象 p,并查找 D 中关于 ε和 MinPts 的从 p 密度可达的所有对象。如果 p 是核心对象,也就是说,p 的半径为 ε的邻域所包含的对象数不小于 MinPts,则根据该算法可以找到一个关于参数 ε和 MinPts 的类。如果 p 是一个边界点,即 p 的半径为 ε的邻域中包含的对象数小于 MinPts,则没有对象从 p 密度可达,p 被暂时地标注为噪声点(noise)。然后,DBSCAN 处理数据库 D 中的下一个对象。

获取从一个核心对象密度可达的所有数据对象是通过反复进行区域查询(region query)来实现一次区域查询返回给定查询区域中的所有对象,这种查询由树型数据结构帮助实施。因此,在进行聚类之前,必须建立存储结构。DBSCAN 要求用户指定一个全局 ε值(为减少计算量,MinPts 经常设定为 4)。为了确定 ε,DBSCAN 需要计算所有数据对象与它的第 k 个(k=4)个最邻近的对象之间的距离,并将结果按距离排序,产生 k-dist 图。k-dist 图中的横坐标表示数据对象与它的第 k 最近的对象间的距离,纵坐标则为对应于某一 k-dist 距离值的数据对象的个数。建立树和绘制 k-dist 图都是非常耗费时间的工作,大规模数据库尤其如此。此外,用户须反复试验,选择合适的 k-dist 值以达到较好的聚类效果。

由于没有做任何预处理而直接对整个数据库进行操作,在用 DBSCAN 算法进行大规模数据库聚类时,一方面,需要大量的内存和 I/O 开销;另一方面,由于使用全局 ε值,因此数据空间中所有对象的邻域大小是一致的。当数据密度和类间距离分布不均匀时,若根据较密的那些类选取较小的 ε值,那么对于客观上相对较稀的那些类中的对象,它们邻域中的数据对象的数目将要小于 MinPts,也就是说这些对象将被认为是边界对象(border objects),从而不被用于所在类的进一步扩展。随之而来的结果是,较稀的类被划分成多个性质相似的类;与此相反,若根据较稀的那些类来选取较大的 ε值,那么离得较近而密度较大的那些类将很可能被合并为同一个类,它们之间的差异也就因为选取较大的 ε值而被忽略。很明显,在上述两种情况下,其实很难选取一个合适的 ε值来进行聚类且得到比较准确的聚类结果。

2.2.2 在实际中的解决参数问题的讨论[40-43]

根据上面所描述的 DBSCAN 聚类算法的特点,可以看出算法的缺点普遍存在于实际计算当中。对于第一个缺点,如果数据库比较大的时候要进行大量的 I/O 开销,在一般聚类算法中是不可避免的,因为在开始计算之前不可预知数据库中的哪些或者哪部分的点属于一个类或密度集中的地方。在一般进行聚类时候要调用空间中点的整个数据库来查找一个类,进行大量的 I/O 操作。特别是访问数据库后要把获得的点存储在一个预定数据结构中,如此又需要大量的内存开销。但是在实际中却避免了进行大量的数据访问。在实际中,要进行交通事故多发点段的排查。作为一个事故多发点段是一条道路的一小段或一个地点。地图上有大量的事故数据点,但是都被一条条街道分割。因此,要排查这个事故黑点,只要求实现查找此条道路的事故点数据,而不需要对整个地图上的事故点访问。在实际中,该算法被用于每次查找一条道路的事故多发点。而建立的数据结构也相对简单,占用内存比较少。作为传统的 DBSCAN 算法还有一个最大的缺点,就是对输入参数十分敏感。算法中有两个主要参数,就是上面提到的 ε和 MinPts。因为事先不能确定数据的密度,而且 ε与 MinPts 是全局唯一的,所以很难选取一个合适的 ε值来进行聚类且得到比较准确的聚类结果。上面介绍了 k-dist 做法来缓和这个问题。但是在实际中,由用户确定这两个参数。对于交通事故多发点的确定,国家有确定的排查标准。对于 ε值来说,在实际交通地图的二维点距上一般在 100 m~2 000 m[50]。也就是说黑点排查时的 ε并不是很大,如果要用 k-dist 算法自动生成 ε,就会出现在交通事故比较稀疏的路段或者根本实际就不能形成黑点的路段生成了比较大的 ε,从而造成排查效果差。由交通专业人士确定 ε还是比较实际的。对于第二个敏感值 MinPts,在国家有关规定中也有确认。由于交通黑点的确定往往要考虑严重程度,所以实际中用户可以选择黑点排查的阈值,单位是损失/(km*年)。在本算法中还加入了对于聚类阈值的判断。后面有关于算法第两种基于事故损失运算模式的介绍。

总之,在实际中算法采用了每次对一条道路的事故多发点段排查聚类方法。参数的确定是由交通部门实践确定,能够灵活改变以适应有关规定。

2.2.3 实际需求对于算法的要求[10-17,44-52]

根据需求分析。实际中要排查交通事故黑点的功能方法,是要根据国家有关规定进行设计的。2001 年公安部交通管理局发布了《全面排查交通事故多发点段工作方案》,其中对公路交通事故多发地点的鉴别标准,做了明确规定。

(1)多发点,为 500 m 范围内,一年之中发生 3 次重大以上交通事故的地点。

(2)多发段,为 2 000 m 范围内或道路桥、涵洞的全程,一年之中发生 3 次重大以上交通事故的路段。

此外,还有一个功能约束:可以由软件系统使用者选定数据进行排查。也就是说用户可以选择排查哪条路什么时间段里

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的数据。

2.2.4 算法中实现实际要求的办法[10-17,44-52]

首先对于第一个要求：算法要符合国家规定的相关标准。可以看到国家对于交通事故黑点的规定中，确定了两个DBSCAN算法敏感的参数，一个是ε-近邻，也就是ε邻域，另一个是MinPts。《全面排查交通事故多发点段工作方案》中规定：多发点，为500 m范围内，一年之中发生3次重大以上交通事故的地点。在算法中500 m就可以理解为ε邻域为500 m，3次重大以上交通事故就可以理解为MinPts是3。多发段，为2 000 m范围内或道路桥、涵洞的全程，一年之中发生3次重大以上交通事故的路段。可以理解为ε邻域为2 000 m，MinPts仍然是3。《2002年预防道路交通事故工作方案》中规定的地、市级排查重点事故多发点段是：2001年以来发生一次死亡3人以上事故的普通公路的点段和3次以上带有规律性死亡事故的点段，那么MinPts仍然是3。

总之，为了适应国家的标准以及未来的变化，交通事故黑点排查的算法中，两个敏感的参数由用户输入调整。事故的严重程度也由用户进行选择。事故按严重程度分为四类：特大事故、重大事故、一般事故、轻微事故。可作为选项。

对于第二个要求，关于功能约束。实现起来相对简单，也是由用户输入排查的数据范围：如时间、地区、路段，然后由程序选择合适的数据库查询语句对数据进行提取。

方法能完全实现用户的需求。

3 基于DBSCAN算法的交通事故多发点段排查方法的改进

3.1 改进方法

基于DBSCAN算法的排查模式（第一种排查模式）的基本思想就是：首先人为确定一个ε邻域（km）和最少事故点MinPts，算法寻找这样的核心事故点，该事故点ε邻域范围内有不少于MinPts个其它点，这些点被划为一类，然后在此类中对其他事故点重复此过程。该排查模式可以相当准确地反映国家对于交通事故黑点排查的规定，但是对于交通事故的严重程度考虑的并不充分。因此可以对其进行改进形成第二种排查模式。

第二种排查模式：在第一种排查模式的基础上，附加了检查过程，从而把事故的严重程度考虑进去。该模式先确定交通事故的各属性权值，再求出每起交通事故的损失情况。在第一种排查模式工作完成后对起排查出来的交通事故黑点进行筛选，它求出一个交通事故黑点类的平均道路损失，与人为给出的交通黑点阈值进行比较来检查损失程度是否符合标准，如果大于阈值就说明是合格的。这种排查模式并不是独立的方法，它是由第一种模式和附加的检查过程组成的。此模式的缺点是，由于交通事故损失密集度的不均衡导致了在检查过程中会漏掉一些交通事故黑点。这些黑点的特征是整体的平均道路损失不高，但是局部平均道路损失高。尽管如此，该排查模式仍然可以满足有关排查标准，并且为交通管理者提供指导性建议。它可以在反应严格标准的基础上，由交通管理者添加附加条件。所以该排查模式适用性更广。对第二种模式的缺点进行改进形成了第三种排查模式。

第三种排查模式：该排查模式是上面两种的改进。该模式首先也要确定ε邻域（km）和最少事故点MinPts，但是在这里MinPts已经不重要，而且可以被设为最低值如2或3。排查的重点是能够使所排查的交通事故黑点的平均道路损失大于黑点的阈值。算法中首先确定一个最初的类，然后在类中寻找核心交通事故点，每当把核心交通事故点ε邻域内所包含的新点加入到该类中的时候都要进行平均道路损失的检查。这样排查的效果是，能够在给定ε邻域（km）和MinPts的情况下找到事故损失密度大于黑点阈值的路点段。该模式充分考虑到了事故的严重程度，为交通事故多发点段排查过程提供支持和扩展。

3.2 算法中关键参数的确定

DBSCAN算法本身就是一种对参数十分敏感的聚类方法，要使算法进行下去首先就需要两个全局统一的参数：ε邻域和MinPts，为了适应相关规定，在第一种排查模式中还设置了事故严重程度的参数以选择排查范围，用户可以选择要排查的事故类型（轻微、一般、重大、特大）。根据国家制定的相关标准，交通管理人员可以把参数设置为（ε邻域=1，MinPts=3（特大））。在模式1中算法将认为，要排查的交通事故多发点段是在半径为1 km的范围内发生3起特大交通事故的点段。

考虑到事故的损失情况，在第2、3种排查模式中设立了交通事故属性和交通黑点阈值。用户可以自己填写交通事故属性的权值，这里有五个与交通事故损失息息相关的属性权值：死亡的权值、重伤的权值、轻伤的权值、失踪的权值、经济损失的权值。

根据1991年12月2日《公安部关于修订道路交通事故等级划分标准的通知》中发布了修订的道路交通事故等级划分标准：

道路交通事故分为以下四类：轻微事故，是指一次造成轻伤1至2人，或者财产损失机动车事故不足1 000元，非机动车事故不足200元的事故；一般事故，是指一次造成重伤1至2人，或者轻伤3人以上，或者财产损失不足3万元的事故；重大事故，是指一次造成死亡1至2人，或者重伤3人以上10人以下，或者财产损失3万元以上不足6万元的事故；特大事故，是指一次造成死亡3人以上，或者重伤11人以上，或者死亡1人，同时重伤8人以上，或者死亡2人，同时重伤5人以上，或者财产损失6万元以上的事故。

当要从重大交通事故中排查交通事故多发点段的时候，根据以上定义的重大交通事故，可以知道一次死亡一人或者至少重伤三人或者至少损失3万元的交通事故才算是重大事故。那么在项目中选择权值的时候就可以提取出权值的比例。

如果想要表示一起重大交通事故，可以：若把经济损失权值设定为1（每万元），则死亡的权值就应该设定为3（每人），重伤的权值就应该设定为1（每人），轻伤的权重并没有明确规定可以设定为0.5（每人），失踪的权值应该和死亡相同为3（每人）。

这样表示一起重大交通事故的权重比例就可以大致确定：

死亡∶重伤∶轻伤∶失踪∶经济损失=6∶2∶1∶6∶2

同样还可以制定出其他等级交通事故的属性权值比例：

轻微事故是（经济损失以千为单位）：死亡∶重伤∶轻伤∶失踪∶经济损失=0∶0∶1∶0∶1

一般事故是（经济损失以万为单位）：死亡∶重伤∶轻伤∶失踪∶经济损失=0∶3∶1∶0∶1

特大事故是（经济损失以万为单位）：死亡∶重伤∶轻伤∶失

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

踪:经济损失=10:3:1:10:5

根据 2001 年公安部交通管理局发布的《全面排查交通事故多发点段工作方案》对公路交通事故多发地点的鉴别标准。如果用算法的模式 2 或 2 计算,在算法排查中的黑点阈值可以确定为每年每公里的损失应该为 3 次重大交通事故的损失。其中权值的确定比例关系是:

死亡:重伤:轻伤:失踪:经济损失=6:2:1:6:2。这说明死亡或者一个人相当于重伤三个人,重伤一个人或者赔偿一万元相当于轻伤两个人。因此,如果权值就按这个确定:死亡为 6,重伤为 2,轻伤为 1,失踪为 6,经济损失为 2。那么黑点阈值就可以由死亡权值确定为 6×3=18(损失/年/km),同样可以由重伤权值确定是 2×9=18,可以由轻伤权值确定为 1×18=18,可以由经济损失权值确定为 2×9=18;总之,无论从哪个交通事故属性来确定黑点阈值都应该是 18。

综上所述,根据国家《公安部关于修订道路交通事故等级划分标准的通知》和公安部交通管理局发布的《全面排查交通事故多发点段工作方案》,实际中选定黑点阈值为 18,死亡、重伤、轻伤、失踪、经济损失的权值分别是 6、2、1、6、2。

参数的选定对于交通管理者是具有指导意义,但是并不是绝对固定的。这也是本方法把参数设置为可选参数的目的,随着社会经济的发展,人们的生活观点也在改变,所以实际中参数的确定也是可变的。例如,根据以前的规定,死亡一个人相当于赔偿 3 万元人们币,但是现代人们社会生活水平的提高,以及人权问题的重视,使得人的生命远远不能用 3 万元来衡量。当然本文只是依靠国家的明文规定,如果以后有所变动,可以根据新规定或者交通管理者的需要来确定。

3.3 结果分析

为了验证算法的正确性,一共选择了 7 个交通事故点,它们的属性如表 1,各属性权重如表 2。

表 1 交通事故属性值表

事故	DEAD	BAD	HURT	LOSS	LOST	POSX	POSY
1	0	0	0	4 000	0	6 057	5 094.84
2	0	0	0	4 400	0	6 057	5 095.37
3	1	0	0	1 000	0	6 057	5 095.42
4	0	0	1	1 000	0	6 057	5 095.56
5	0	0	1	200	0	6 057	5 095.99
6	0	0	0	12 125	0	6 057	5 096.14
7	0	0	0	6 000	0	6 057	5 097.02

实际中的各数据参数(权值):

表 2 各属性权重表

READ	BAD	HURT	LOSS	LOST
6	2	1	2	6

从而求出各个事故点的损失如表 3。

表 3 各交通事故点损失表

事故 1	事故 2	事故 3	事故 4	事故 5	事故 6	事故 7
0.800	0.880	6.200	1.020	1.040	2.425	1.200

损失=DEAD*6+BAD*2+HURT*1+LOSS/10000*2+LOST*6

交通事故点相隔距离如图 1,交通事故点损失与相隔距离如图 2。

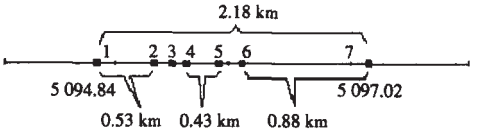

图 1 交通事故点相隔距离示意图

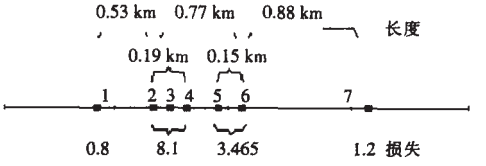

图 2 交通事故点损失与相隔距离示意图

1)用排查模式 1 对以上事故点进行排查

设置:半径为 1 km(ε邻域),发生 3 次(MinPts)以上交通事故。排查结果:该 7 个事故点共同构成一个交通事故黑点。分析:7 个事故点占用了长为 2.18 km 的路段,在这范围内,每个事故点的 1 km 半径范围内都有 3 起以上的交通事故。

设置:半径 600 m 范围内发生 3 次以上交通事故。排查结果:点 1、2、3、4、5、6 这 6 个点构成一个交通事故黑点。分析:因为点 6 与点 7 之间的距离为 880 m,所以点 7 的 600 m 半径范围内无交通事故,同样点 6 的 600 m 范围内也不包含点 7,所以点 7 成为孤立点,可以理解为噪声。

设置:半径 500 m 范围内发生 3 次以上交通事故。排查结果:点 2、3、4、5、6 这 5 个点构成一个交通事故黑点。分析:因为点 1 与 2 之间距离,点 6 与 7 之间距离都大于 500 m,所以,点 1 与点 7 就与其它点相分离,成为孤立点。

设置:半径 400 m 范围内发生 3 次以上交通事故。排查结果:点 2、3、4 这 3 个点构成一个交通事故黑点。分析:点 1 与 2,点 4 与 5,点 6 与 7 之间距离都大于 400 m,所以,点 1、7 成为孤立点。虽然点 5、6 距离近,但是它们的 400 m 半径内只有 2 个事故点,小于规定的 3 次交通事故,所以被孤立。

2)用排查模式 2 对数据进行排查

设置:半径为 1 km(ε邻域),发生 3 次(MinPts)以上交通事故。交通黑点阈值为 18。排查结果:没有构成一个交通事故黑点。分析:按照排查模式 1 的结果 7 个点可以聚为一个类,它们的总损失为 13.565,占用了 2.18 km 的路段,时间间隔是 0.5 年。所以平均道路损失是 13.565/2.18/0.5=12.445<18(阈值),不能构成交通黑点。

设置:半径 600 m 范围内发生 3 次以上交通事故。排查结果:点 1、2、3、4、5、6 这 6 个点构成一个交通事故黑点。分析:根据模式 1 就排查出了由这 6 个点构成的类,因为它们总损失为 12.365,占用了 1.3 km 的路段,时间间隔是 0.5 年。所以平均道路损失是 12.365/1.3/0.5=19.023>18,构成交通黑点。

设置:半径 500 m 范围内发生 3 次以上交通事故。排查结果:点 2、3、4、5、6 这 5 个点构成一个交通事故黑点。分析:因为这 5 个点的总损失为 11.565,占用了 0.77 km 的路段,时间间隔是 0.5 年。所以,平均道路损失为 11.565/0.77/0.5=30.039>18,构成交通黑点。

设置:半径 400 m 范围内发生 3 次以上交通事故。排查结果:点 2、3、4 这 3 个点构成一个交通事故黑点。分析:这 3 个点的总损失为 8.1,占用了 0.19 km 的路段,时间间隔是 0.5 年。所以平均道路损失是 8.1/0.19/0.5=85.263>18,构成交通事故黑点。

3)用排查模式 3 对数据进行排查

设置:半径为1 km(ε邻域),发生 3 次(MinPts)以上交通事故。交通黑点阈值为 18。排查结果:2、3、4、5、6 形成一个交通事

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

故黑点。分析 :在排查模式 2 中同样的设置并没有排查出交通事故黑点,因为第一步计算会把 7 个点全都归为一个类,在检查时候就会产生平均道路损失小于交通黑点阈值的情况。而在排查模式 3 中,算法每次加入新的点都要进行平均道路损失的检查,所以能够发现路段损失密度大的区域。

4 结束语

本文详细探讨了 DBSCAN 算法,介绍了在交通事故多发点段排查问题上算法应用的逻辑思想。设计了三种工作模式,可以适应交通部门的不同需求,以及在排查标准上的变化。三种排查模式都是基于事故数方法,第一种模式完全是数据挖掘中基于密度的聚类方法 DBSCAN 方法的应用。它可以比较准确地排查出国家相关规定中的交通事故黑点。但是它只是在数据库中选择原始数据的时候对交通事故的严重程度进行了筛选,并没有在算法中考虑到。因此,提出了第二种排查模式,就是在第一种排查模式运算以后对结果进行检查,查看是否符合交通黑点阈值。这种排查模式配合了 DBSCAN 算法,是一个进步。但是当管理者主要考虑道路损失,而不是事故数量的时候,第二种方法就会出现漏掉黑点的可能。所以,进一步给出了第三种排查模式,该模式算法中,在进行事故数搜索的同时就检查了黑点阈值,总是保持了当前不断增长的类中事故的平均道路损失大于交通黑点阈值。

这三种模式相互配合,既可以满足国家相关规定的要求又可以考虑到交通事故损失的严重程度。交通管理者利用本文给出的交通事故多发点段智能排查功能可以灵活调整排查标准,找到需要排查的交通事故黑点,大大节省了劳动时间。

由于交通事故的发生会涉及到各种因素和复杂情况,所以对于交通事故多发点段的排查算法还有很多改进的余地,如方法只考虑到了交通事故数据,并没有考虑到交通流量等其他方面的因素。 (收稿日期 :2006 年 10 月)

参考文献:

[1] Ester M,Kriegel H P,Sander J,et al.A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceeding the 2nd International Conference on Knowledge Discovery and Data Mining(KDD),Portland,1996 :226-231.

[2] Ester M,Kriegel H P,Sander J,et al.Clustering for mining in large spatial databases[J].KI,1998,12 (1) :18-24.

[3] Sander J,Ester M,Kriegel H P,et al.Density-based clustering in spatial databases:the algorithm GDBSCAN and its applications[J].Data Mining and Knowledge Discovery,Kluwer Academic Publishers,1998,2 (2).

[4] Hinneburg A,Keim D A.Clustering techniques for large data sets:from the past to the future [C]//Tutorial,Proc Int Conf on Knowledge Discovery in Databases(KDD '99),San Diego,CA,1999.

[5] Introduction to data mining and knowledge discovery[M].3rd ed.Two Crows Corporation,ISBN :1-892095-02-5,1999 :1-36.

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

[7] Braunmüller B,Ester M,Kriegel H P.Similarity queries:a basic DBMS operation for mining in metric databases [J].IEEE Transactions on Knowledge and Data Engineering,2000.

[8] Ertoz L,Steinbach M,Kumar V.Finding clusters of different sizes,shapes,and densities in noisy,high dimensional data,Technical Report[R],2002

[9] Berkhin P.Survey of clustering data mining techniques[J].Accrue Software,2002.

[10] Khisty C J.Transportation engineering:an introduction[M].NJ,USA:Prentice Hall,Englewood Cliffs,1990.

[11] Oppe S.Development of traffic safety global trends and incidental fluctuations[J].Accident Analysis and Prevention,1991,23 (1).

[12] Ogden K W.Saferroads:a guide to road safety engineering[M].England:a Shgate Publishing Company,1996.

[13] Tarko A P,Weiss J V,Sinha K C.Potential crash reduction for identification of hazardous locations[C]//Traffic Congestion and Traffic Safety in the 21st century,Proceedings of the Conference Sponsored by Urban Transportation Division,ASCE,Highway Division ASCE,New York,1997.

[14] Kurauch H.Road accidents in Japan[J].The International Association of Traffic and Safety Science Research,1997,21 (2) :160-161.

[15] Dinesh M.Road accidents in India[J].The International Association of Traffic and Safety Science Research,1999,23 (1) :130-131.

[16] Mannan M S,Sc M,Karim M.Road accidents in metropolitan Dhaka,Bangladesh[J].IATSS Research,1999,23 (2) :91-94.

[17] Doherty S T,Hall A,Swaynos J.Commuter cyclist accident patterns in Toronto and Ottawa[J].Journal of Transportation Engineering,2000,21 (1) :21-26.

[18] Yang Y,Guan X,You J.CLOPE:a fast and effective clustering algorithm for transactional data[C]//Proc ACM Int'l Conf Knowledge Discovery and Data Mining(KDD '02),2002 :682-687.

[19] Wang H,Yang J,Wang W.Clustering by pattern similarity in large data sets[C]//Proc ACM SIGMOD 2002 Conf,2002 :394-405.

[20] Olman V,Xu D,Xu Y.Cubic:identification of regulatory binding sites through data clustering [J].Bioinformatics and Computational Biology,2003,1 (1) :21-40.

[21] Dhillon I S,Mallela S,Modha D S.Information-theoretic co-clustering[C]//Proc SIGKDD '03,2003 :89-98.

[22] Wang J,Zeng H,Chen Z,et al.ReCoM:reinforcement clustering of multi-type interrelated data objects[C]//Proc SIGIR '03,2003.

[23] Law M H,Jain A K,Figueiredo M A T.Feature selection in mixture-based clustering[C]//Proc Advances in Neural Information Processing,2003.

[24] Hermes L,Buhmann J M.Semi-supervised image segmentation by parametric distributional clustering[J].Energy Minimization Methods in Computer Vision and Pattern Recognition,2003 :229-245.

[25] Jiang D,Pei J,Zhang A.DHC:a density-based hierarchical clustering method for time series gene expression data[C]//Proc Third IEEE Symp on Bio-Informatics and Bio-Engineering(BIBE '03),2003.

[26] Dai B-R,Lin C-R,Chen M-S.On the techniques for data clustering with numerical constraints [C]//Proc SIAM Int'l Conf Data Mining(SDM '03),2003.

[27] Merugu S,Ghosh J.Privacy-preserving distributed clustering using generative models [C]//Proc Third IEEE Int'l Conf Data Mining (CDM '03),2003.

[28] Wu W,Yu C T,Doan A,et al.An interactive clustering-based approach to integrating source query interfaces on the deep Web [C]//Proc SIGMOD,2004.

[29] Caverlee J,Liu L,Buttler D.Probe,cluster and discover:focused extraction of QA-pagelets from the deep Web[C]//Proc Int'l Conf

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Data Eng,2004.
[30] Jiang D,Pei J,Zhang A.Mining coherent gene clusters from gene-sample-time microarray data [C]//Proc ACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining(KDD '04),2004.
[31] Lian W,Cheung D,Mamoulis N,et al.An efficient and scalable algorithm for clustering XML documents by structure[J].IEEE Trans Knowledge and Data Eng,2004,16(1).
[32] Özdal,Aykanat C.Hypergraph models and algorithms for data-pattern based clustering [J].Data Mining and Knowledge Discovery,2004,9(1):29-57.
[33] Parsons L,Haque E,Liu H.Subspace clustering for high dimensional data:a review[J].SIGKDD Explorations,2004,6(1):90-105.
[34] Andritsos P,saparas P T,Miller R J,et al.LIMBO:scalable clustering of categorical data [C]//Proc Ninth Int'l Conf Extending DataBase Technology(EDBT),2004:123-146.
[35] Caverlee J,Liu L,Buttler D.Probe,cluster,discover:focused extraction of QA-pagelets from the deep Web [C]//Proc Int'l Conf Data Eng,2004.
[36] Cheung Yiu-ming.Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection[J].IEEE Trans Knowledge and Data Eng,2005,17(6):750-761.
[37] Lin Cheng-ru,Liu Ken-hao,Chen Ming-syan,et al.Dual clustering:integrating data clustering over optimization and constraint domains[J].IEEE Trans Knowledge and Data Eng,2005,17(5):628-637.
[38] Shen Huang,Zheng Chen,Yong Yu,et al.Multitype features coselection for Web document clustering[J].IEEE Trans Knowledge and Data Eng,2006,18(4):448-459.
[39] Hsu Chung-Chian,Wang Sheng-hsuan.An integrated framework for visualized and exploratory pattern discovery in mixed data[J].IEEE Trans Knowledge and Data Eng,2006,18(2):161-173.
[40] 周水庚,周傲英,曹晶.基于数据分区的 DBSCAN 算法[J].计算机研究与发展,2000,37(10):1153-1159.
[41] 李雄飞,李军.数据挖掘与知识发现[M].北京:高等教育出版社,2003:108-110.
[42] 毛国君,段立娟,王实.数据挖掘原理与算法[M].北京:清华大学出版社,2005:175-178.
[43] Han Jia-wei,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,2006:242-243.
[44] 郑柯,冯桂炎.道路交通事故多发点道路状态的技术分析[J].长沙交通学院学报,2000,16(1):63-66.
[45] 任江涛,张毅,李志恒,等.智能交通系统信息特征及亟待解决的相关问题[J].信息与控制,2001,30(6):550-554.
[46] 姜华平,许洪国.基于数理统计原理的交通事故多发点识别[J].济南交通高等专科学校学报,2001,9(3):15-18.
[47] 丁美玲,陈抗生.交通事故处理信息系统的设计与实现[J].计算机工程与应用,2001,37(12):169-172.
[48] 杜心全.交通事故多发地点鉴别标准与方法研究[J].公安大学学报:自然科学版,2002(5):68-71.
[49] 路峰,姜文龙,马社强.交通事故多发点段排查方法[J].长安大学学报,2003(1):87-90.
[50] 秦利燕,邵春福,贾洪飞.高速公路交通事故分析及预防对策研究[J].中国安全科学学报,2003,13(6):64-67
[51] 刘志强,宫镇,蔡东.道路交通事故多发点鉴别[J].交通运输工程学报,2003,3(2):120-123.
[52] 裴玉龙,马骥.道路交通事故道路条件成因分析及预防对策研究[J].中国公路学报,2003,16(4):77-82.

(上接 199 页)

并对其识别结果进行了比较与分析。其中 DCT-PCA 和 Gabor-PCA 方法能够达到最高识别率 77.5% 和 77.9%,与传统的人工选择变换系数的降维方法相比识别率提高了约 10%,实验表明分块 DCT 和全局 DCT 的识别结果相差不大,而 DCT 人工选择方式的识别结果略高于直接使用 PCA 的方法。目前的研究还主要是针对于特定人的识别,接下来的工作是将方法扩展到非特定人的研究领域。本文对唇读中应用 Gabor 小波变换进行特征提取进行了初步尝试,取得了一定成果,以后还需要进行深入的研究。 (收稿日期:2007 年 2 月)

参考文献:

[1] 姚鸿勋,高文,王瑞,等.视觉语言——唇读综述[J].电子学报,2001(1).
[2] Potamianos G.A cascade image transform for speaker independent automatic speechreading [C]//IEEE International Conference on Multimedia and Expo,2:1097-1100.
[3] Potamianos G,Graf H P,Cosatto E.An image transform approach for HMM based automatic lipreading[C]//Proc Int Conf Image Process,Chicago,1998:173-177.
[4] Scanlon P,Reilly R.Visual feature analysis for automatic speechreading [C]//Proc Works Multimedia Signal Processing,Cannes,France,Oct 3-5,2001,2001:625-630.
[5] Matthews et al.Extraction of visual features for lipreading[J].IEEE Transaction on Pattern Analysis and Machine Intelligence,2002,24(2).
[6] Bregler C,Konig Y.Eigenlips for robust speech recognition[C]//Proc Int Conf Acoust Speech Signal Process,Adelaide,1994:669-672.
[7] Duchnowski P.Toward movement-invariant automatic lipreading and speech recognition [C]//Proc Int Conf Acoust Speech Signal Process,Detroit,1995:109-112.
[8] Neti C.Audio-visual speech recognition,Final Summer 2000 Work Shop Report[R].Center for Language and Speech Processing,Baltimore,2000.
[9] Heckmann M.DCT-based video features for audio-visual speech recognition[C]//Proc Int Conf Spoken Lang Process Denver,USA,September 2002,2002:1925-1928.
[10] 山世光.人脸识别中若干关键技术的研究[D].中科院计算所,2004.
[11] Liu C,Wechsler H.Gabor feature based classification using enhanced fisher linear discriminant model for face recognition[J].IEEE Trans Image Processing,2002,11(4):467-476.
[12] Lee T S.Image representation using 2D Gabor wavelets[J].IEEE Trans Pattern Analysis and Machine Intelligence,1996,18(10):959-971.
[13] Yuille A.Deformable templates for face recognition[J].Journal of Cognitive Neuroscience,1991,3(1).
[14] Wiskott L.Face recognition by elastic bunch graph matching[J].IEEE Trans on PAMI,1997,19(7):775-779.
[15] Daugman J G.Uncertainty relation for resolution in space,spatial frequency,and orientation optimized by two-dimensional visual cortical filters[J].J Optical Soc Amer,1985,2(7):1160-1169.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net