第 35 卷 第 5 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.5 Computer Engineering March 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)05—0268—03 文献标识码：A 中图分类号：TP301.6

基于 DBSCAN 算法的营运车辆超速点聚类分析

刘卫宁[1]，曾婵娟[1]，孙棣华[2]
(1. 重庆大学计算机学院，重庆 400044；2. 重庆大学自动化学院，重庆 400044)

摘 要：针对挖掘营运车辆超速点过程中存在的问题，提出一种基于密度的聚类方法。该方法依据车载 GPS 实时监控数据，挖掘超速多发点段，通过区域查询搜索超速点邻域内所有超速事件，寻求超速密度大于阈值的点或地段，并创建密度可达最大值的超速点聚类。同时利用简单直观的邻接表替代 R*-树，简化了数据结构的建立过程，减少内存占用。实验结果表明，该方法有效。

关键词：数据挖掘；聚类；营运车辆；安全管理；超速

Clustering Analysis of Overspeed Spots for Commercial Vehicles Based on DBSCAN

LIU Wei-ning[1], ZENG Chan-juan[1], SUN Di-hua[2]
(1. College of Computer, Chongqing University, Chongqing 400044; 2. College of Automation, Chongqing University, Chongqing 400044)

【Abstract】Aiming at the problems existed in the process of mining overspeed spots for commercial vehicles, a density-based clustering method is proposed, which searches for the spots where the overspeed usually happens in high frequency according to the GPS real-time data. The overspeed spots with high density can be found by searching in all overspeed in the neighborhood of each spot. The maximum overspeed spots can be identified by clustering. To simplify the data structure building process and reduce the memory space it occupied, the method is improved with the adjacency list replaced R*-Tree. Experimental results show this method is effective.

【Key words】data mining; clustering; commercial vehicles; safety management; overspeed

1 概述

在引发道路运输安全事故的众多因素中，营运车辆的违规超速行驶是主要原因之一。深入分析排查营运车辆超速多发点段，全面揭示超速事件的空间分布规律，对于加强道路运输安全管理极具意义。

近年来，发达国家管理部门十分重视对道路安全因素及其规律的分析。例如，澳大利亚通过识别交通系统中的不安全因素提出年度道路安全程序，同时规定整治交通黑点为道路设计者必须完成的工作[1]。

在国内，道路运输安全管理工作通常是对安全事故进行事后考核和分析，并根据管理部门的经验，参照以往的统计数据或其他同级机构所设定的标准，设定一个管理目标值[2]。目前，GPS 技术已应用于营运车辆的监控管理，但对其如何服务于运输安全管理，仍停留在超速数据的简单统计和报表阶段，尚缺乏较深入的研究。

由于GPS营运车辆监控系统已经积累了大量的车辆超速报警数据，为定量分析营运车辆的超速规律创造了条件。因此，本文针对营运车辆超速问题，引入数据挖掘(Data Mining)技术中的聚类方法，着重研究营运车辆超速多发点段的规律。

2 基于 DBSCAN 的营运车辆超速点聚类

2.1 营运车辆超速特征

营运车辆超速现象严重，一方面是由于某些经营业主为了经济利益最大化不惜超速、超车、超载，另一方面也因为营运线路具有“点多、线长、面广、路况复杂”等特点[2]，所以监督管理困难。交通部门统计数据表明，营运车辆超速具有一定规律性。一些涵洞、弯道、坡道等路段，尤其是弯道和坡道结合处，如果线型设计不合理，极易导致超速，使得超速事件发生相对集中，发生率明显高于其他路段。因此，全面掌握营运线路道路状况，查明超速多发点段，对降低营运车辆安全事故发生率具有非常重要的价值。

传统的超速多发点段评定主要通过计算机筛选超速情况统计资料，结合管理经验，得到超速多发点段。但这样得到的路段往往是等长的，缺乏客观真实性，甚至由于路段长度选择不合理，可能漏选一些真正危险的路段，因此不利于安全隐患的排查。

数据挖掘技术可以从大量数据中发掘隐含规律，适于研究车辆超速问题和全面把握营运车辆的超速规律。

2.2 聚类及 DBSCAN 算法

聚类是数据挖掘中用来发现数据分布和隐含模式的一项关键技术。所谓聚类，就是把大量数据点的集合分成若干类，使得每个类中的数据之间最大程度地相似，而不同类中的数据最大程度地不同[3]。DBSCAN 是个基于密度的聚类算法，通过不断地搜索邻近点来使该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。该算法将具有足够高密度的区域划分为簇，并可以在带有“噪声”的数据中发现任意形状的聚类[4]。

2.3 基于 DBSCAN 的营运车辆超速点聚类

超速多发点段可理解为一条高速路上发生超速事件密度

基金项目：重庆市智能交通系统示范工程及其关键技术基金资助项目(7742)

作者简介：刘卫宁(1965－)，女，教授、博士生导师，主研方向：智能交通，电子商务与现代物流；曾婵娟，硕士研究生；孙棣华，教授、博士生导师
收稿日期：2008-09-30 **E-mail：**zengchanjuan163@163.com

大的地点(或路段)。因此，可以引入 DBSCAN 算法分析超速点聚类规律。此处的点描述为超速事件发生点，邻域(neighborhood)为道路的公里数。

基于 DBSCAN 的营运车辆超速点聚类分析方法的核心思想为：对于构成超速多发点段的每个超速报警，其发生的地点半径邻域 *Eps* 公里范围内的其他超速报警的个数必须不小于给定的阈值 *MinPts*，即邻域的密度必须不小于某个阈值。所以，超速多发点段为半径 *Eps* 公里内发生 *MinPts* 以上超速事件的地点或者路段。

下面给出算法中涉及的定义[3]：

(1)核心超速点：给定 *Eps, MinPts*，若超速点 *p* 的 *Eps* 邻域包含的超速对象个数 $|Neps(p)| \geqslant MinPts$，则称 *p* 是核心超速点。

(2)直接密度可达：给定 *Eps, MinPts*，点 *p* 是从点 *q* 出发直接密度可达的，当且仅当 $p \in Neps(q)$，$|Neps(p)| \geqslant MinPts$。

(3)密度可达：给定一个超速集合 *D*，当存在一个对象链 $p_1, p_2, \cdots, p_n$，$p_1 = q$，$p_n = p$，对 $p_i \in D$，p_{i+1} 是 p_i 关于 *Eps* 和 *MinPts* 直接密度可达的，则称对象 *p* 从对象 *q* 关于 *Eps* 和 *MinPts* 密度可达(非对称)。

(4)密度相连：如果对象集合 *D* 中存在一个对象 *o*，使得对象 *p* 和 *q* 是从 *o* 关于 *Eps* 和 *MinPts* 密度可达的，那么对象 *p* 和 *q* 关于 *Eps* 和 *MinPts* 密度相连(对称)。

(5)超速黑点：基于密度可达的最大密度相连对象的集合称为超速黑点。

(6)噪声点：不属于任何黑点的对象被认为是噪声点。

图 1 为算法的应用实例。

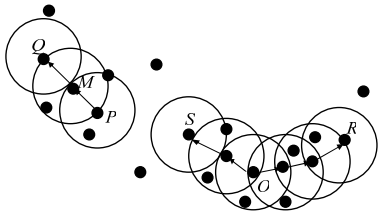

图 1　DBSCAN 算法应用实例

图 1 中所有点均表示超速点，*Eps* 用 1 个相应的半径表示，设 *MinPts*=3。由此可知：(1)由于有标记的各点 *M, P, Q* 和 *R* 的 *Eps* 近邻均包含 3 个以上的点，因此它们都是核心超速点；(2)*M* 从 *P* 直接密度可达，而 *Q* 从 *M* 直接密度可达；(3)基于上述结果，*Q* 从 *P* 密度可达，但 *P* 从 *Q* 无法密度可达(非对称)。类似地，*S* 和 R 从 *O* 密度可达；(4)*O, R* 和 *S* 均是密度相连的。

基于密度聚类是组密度相连的对象，以实现最大化的密度可达。不包含在任何聚类中的对象为噪声数据。因此，图中的 2 个类代表 2 个超速黑点。

DBSCAN 算法检查数据库中每个点的 *Eps* 近邻。若 1 个对象 *P* 的 *Eps* 近邻包含多于 *MinPts* 个超速点，就要创建包含 *P* 的新聚类，然后算法根据这些核对象，循环收集直接密度可达对象，其中可能涉及若干密度可达聚类的合并。当各聚类无新点(对象)加入时，聚类进程结束，最后得到的类就是超速黑点。DBSCAN 算法的计算复杂度为 $O(n\log_x n)$，其中，*n* 为数据库中对象数目。

3　DBSCAN 算法

3.1　DBSCAN 算法的适用性及改进

DBSCAN 算法的显著特点是聚类速度快，且能够有效处理噪声点和发现任意形状的空间聚类。而传统 DBSCAN 算法具有几个较明显的缺点[5]：

(1)算法直接对整个数据库进行操作，当数据量增大时需要较大内存支持和 I/O 消耗；

(2)聚类前需要建立所有数据的 R*-树，耗时且实现繁琐；

(3)由于使用了 1 个全局性的表征密度的参数 *Eps*，因此当空间聚类的密度不均匀或聚类间距离相差很大时，聚类质量较差。

针对第 1 个缺点，在实际应用中，由于地图上大量的超速数据点被一条条道路分割，每次只需考察所排查道路的超速点数据，因此数据量大大减少。

其次，本文采用邻接表代替 R*-树建立数据结构。传统 DBSCAN 算法的基本数据结构 R*-树构建相当费时，且构建 R*-树的时间复杂度和算法总的时间复杂度($O(n\log_x n)$)与数据总量呈非线性关系。由于当通过区域查询排查超速黑点时，算法以每 *Eps* 公里至少要有 *MinPts* 起超速事件的原则进行搜索，因此只需关心每个超速点临近的 *Eps* 范围内的超速情况，超速点集由于邻域的存在而形成一种相邻关系，即可以采用图论中的邻接表来代替 R*-树。邻接表作为数据结构简单直观，且占用内存较少，如图 2 所示。

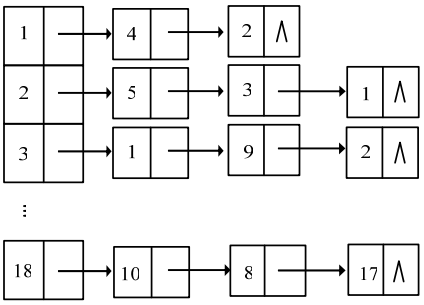

图 2　超速事件邻接表结构

邻接表由基表和其后的链表组成，基表表示所有超速事件数组，链表表示该超速点邻域内的所有其他超速对象。例如，第 3 行第 1 个元素代表第 3 起超速事件，且该超速事件点的 *Eps* 半径范围内有 1, 9, 2 这 3 起超速事件。这样的邻接表既节省了空间，又节省了 DBSCAN 算法的计算步骤。

最后，对于 *Eps* 与 *MinPts* 参数的选取可以根据每条道路车辆超速情况的统计报表来确定其值。不同道路的超速情况有所不同，故选取不同参数值，具有一定科学性与合理性。同时可以根据聚类分析结果好坏对参数值进行修正，提高聚类质量。

3.2　算法描述

Step1 建立原始超速数据集 *Origin*，在属性数据库中增加 1 个新字段 ClusterID(数值型)，用于存储聚类结果，初始化所有记录的 ClusterID 值为零。定义搜索数据集 *Search*，用于临时存储检索结果，初始化参数 *MinPts* 和 *Eps*。

Step2 遍历 *Origin*，依次搜索每个点的 *Eps* 邻域，并为该点建立邻接表存储邻域内所有超速数据点。

Step3 遍历 *Origin*，依次将每个点作为种子点进行考察：1)对于点 p_i，如果 ClusterID 为零，则搜索其邻接表；如果邻接表链表中超速报警数目超过 *MinPts*，则点 p_i 为核心点，将其 ClusterID 设置为 *cluster*，同时将 p_i 的邻接表链表包含的所有点存入 *Search* 中。2)遍历 *Search*，将每个点作为种子点进行考察，对于点 q_i，如果 ClusterID 为零，搜索其邻接表，

如果邻接表链表中超速报警数目超过 *MinPts*，则 q_i 也是个核心点，同时它是点 p_i 的直接密度可达点，与 p_i 属于同一类，将 q_i 的 ClusterID 设置为 *cluster*；否则 q_i 为边界点，但 q_i 的 ClusterID 也设为 *cluster*。如果 q_i 是核心点，点 *o* 存在于 q_i 邻接表链表中并且 *o* 不属于 *Search*，则将点 *o* 存入 *Search* 中。最后将点 q_i 从 *Search* 中删除。3)考察 *Search* 中的下一个点，如果 *Search* 非空，则执行步骤 2)。

Step4 考察数据集 *Origin* 中的下一个点，并将 *cluster* 加 1，执行 Step2，直至遍历完数据集。

Step5 删除搜索数据集 *Search*。

至此聚类结束，对超速数据集仅搜索一次即可得到最终结果。*Origin* 属性数据库中记录了聚类结果，其中，字段 ClusterID 值为零的点为噪声点。

4 实验及结果分析

对重庆市电子地图和重庆市运管局监控中心数据库 3 月 12 日~4 月 12 日成渝高速报警信息进行实验，共计 6 778 条。参数选取 *Eps*=1 km, *MinPts*=20。实验采用的软件平台为 SQL Server 2000，开发工具 VB6.0 和 MapX5.0。

图 4、图 5 分别为成渝高速超速点聚类结果局部图。图中黑点表示聚类算法挖掘出的超速黑点范围内的超速点。

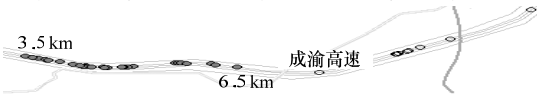

图 4 成渝高速超速点聚类局部图(西往东方向)

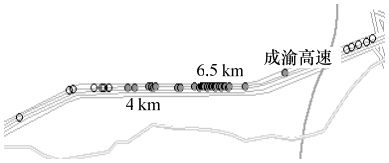

图 5 成渝高速超速点聚类局部图(东往西方向)

从聚类分析的结果来看，图 4 显示了第 9 段石竹镇立交往永州河桥 3.5 km~6.5 km 处(西往东方向)为超速事件多发路段；图 5 显示了成渝高速第 7 段来凤立交往马坊派出所 4 km~6.5 km 处(东往西方向)为超速事件多发路段。

此外，整个聚类图还挖掘出其他 4 处超速事件多发路段，即青杠派出所立交往来凤立交 0.5 km~2 km 处、3.5 km~6.5 km 处(西往东方向)、来凤立交往马坊派出所 3.5 km~5.5 km 处(东往西方向)、大足荣昌交界往新峰派出所 13 km~15 km 处(西往东方向)和新峰派出所往下 1 个立交方向 0.5 km~2 km 处、3 km~5 km 处(西往东方向)为超速事件多发路段。

5 结束语

本文提出的营运车辆超速点聚类分析方法能够快速、有效地对超速点进行聚类，排查出超速多发点段，更全面地揭示超速事件的空间分布规律，为运输管理部门制定超速防治策略和安全管理措施提供科学依据。

参考文献

[1] Paul P. Latest Brifen V'RSF Installations Along Tasmania's Bass Highway[J]. Highway Engineering in Australia, 1999, 30(8): 5-11.

[2] 吴章龙. 坚持人本管理重在关键环节——关于道路运输企业安全管理的探讨[J]. 交通企业管理, 2006, 21(6): 21-22.

[3] 荣秋生, 颜君彪, 郭国强. 基于 DBSCAN 聚类算法的研究与实现[J]. 计算机应用, 2004, 24(4): 45-46, 61.

[4] Kambr 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, 2001.

[5] Martin E. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]//Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Portland, Oregon, USA: AAAI Press, 1996.

编辑　陈　文

（上接第 256 页）

2) $\hat{\boldsymbol{v}}_{j+1}=\boldsymbol{A}\boldsymbol{v}_j-\sum_{i=1}^{j}h_{i,j}\boldsymbol{v}_i$；3) $h_{j+1,j}=\left\|\hat{\boldsymbol{v}}_{j+1}\right\|$；4) $\boldsymbol{v}_{j+1}=\hat{\boldsymbol{v}}_{j+1}/h_{j+1,j}$。

(3)结束：应用 Hessenberg 矩阵[$\boldsymbol{h}_{i,j}$]和基{$\boldsymbol{v}_i$}得到近似解 x_k。这一求解方程组 $\boldsymbol{M}(x)=\boldsymbol{0}$ 的方法称为 Newton-Krylov 方法。

7 结果分析

为了验证拟合的优度，选用没有参加拟合的另外几组数据进行仿真计算，并与实际样本浓度进行比较如表 1 所示。

表 1 实际样本浓度值与计算浓度值

序号	样本浓度/(g·L^{-1})		
	CD-3B	CD-R	CD-丈青
1	0.90	1.8	0.9
2	1.60	0.9	0.8
3	1.20	1.3	2.5
4	0.65	1.3	0.6
5	0.70	0.8	1.5
序号	计算浓度/(g·L^{-1})		
	CD-3B	CD-R	CD-丈青
1	0.900 709 299	0.799 325 738	0.899 462 873
2	1.599 961 802	0.899 865 196	0.800 070 150
3	1.199 324 421	1.301 175 775	2.505 414 913
4	0.649 275 005	1.299 125 104	0.599 269 613
5	0.698 578 391	0.799 564 665	1.499 868 693
序号	绝对误差		
	CD-3B	CD-R	CD-丈青
1	0.000 709 299	0.000 674 262	0.000 537 127
2	0.000 038 198	0.000 134 804	0.000 070 149
3	0.000 675 579	0.001 175 775	0.005 414 913
4	0.000 724 995	0.000 874 896	0.000 730 387
5	0.001 421 609	0.000 435 335	0.000 131 307

由表 1 对样本浓度和计算浓度值进行比较，可以看出误差值在 1%之内，这个误差已经达到了较好的水平。经过对计算结果进行再次打样并与原样进行比较，发现颜色误差在 CIE $L^*a^*b^*$色差要求范围内，达到了预期的效果。因此，证明了所建立的基于三拼色的数学模型的配色方法是可行的。

8 结束语

本文提出了一种基于数值分析的织物染色配色方法。实验表明，此方法能够进行准确高效的配色仿真计算，具有一定的理论研究价值和实际应用价值。

参考文献

[1] 徐海松. 颜色信息工程[M]. 杭州: 浙江大学出版社, 2005.

[2] 王喜昌, 华　臻, 官严军. 基于线性数据库的色差权重因子计算机配色[J]. 光学学报, 2004, 24(9): 224-1228.

[3] Dupont D. Study of the Reconstruction of Reflectance Curves Based on Tristimulus Values: Comparison of Methods of Optimization[J]. Color Research and Application, 2002, 27(2): 88-99.

[4] 程正兴, 李水根. 数值逼近与常微分方程数值解[M]. 西安: 西安交通大学出版社, 2000.

[5] Quarteroni A, Sacco R, Saleri F. Numerical Mathematics[M]. 北京: 科学出版社, 2006.

编辑　顾逸斐