Available online at www.sciencedirect.com
ScienceDirect
Computers, Environment and Urban Systems
31 (2007) 4–18
Computers, Environment and Urban Systems
www.elsevier.com/locate/compenvurbsys

A dual approach to cluster discovery in point event data sets

Allan J. Brimicombe *

Centre for Geo-Information Studies, University of East London, University Way, London E16 2RD, UK

Received 30 April 2004; received in revised form 4 July 2005; accepted 4 July 2005

Abstract

Spatial data mining seeks to discover meaningful patterns in data where a prime dimension of interest is geographical location. Consideration of a spatial dimension becomes important where data either refer to specific locations and/or have significant spatial dependence which needs to be considered if meaningful patterns are to emerge. For point event data there are two main groups of approaches to identifying clusters. One stems from the statistical tradition of classification which assigns point events to a spatial segmentation. A popular method is the *k*-means algorithm. The other broad approach is one which searches for 'hot spots' which can be loosely defined as a localised excess of some incidence rate. Examples of this approach are GAM and kernel density estimation. This paper presents a novel variable resolution approach to 'hot spot' cluster discovery which acts to define spatial concentrations within the point event data. 'Hot spot' centroids are then used to establish additional distance variables and initial cluster centroids for a *k*-means classification that produces a segmentation, both spatially and by attribute. This dual approach is effective in quickly focusing on rational candidate solutions to the values of *k* and choice of initial candidate centroids in the *k*-means clustering. This is demonstrated through the analysis of a business transactions database. The overall dual approach can be used effectively to explore clusters in very large point event data sets.

Keywords: Spatial data mining; Point events; Clustering; Hot spots; Geocomputation; Robust normalisation

* Tel.: +44 20 8223 2352; fax: +44 20 8223 2918.
E-mail address: a.j.brimicombe@uel.ac.uk

0198-9715/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2005.07.004

1. Introduction

Prior to the 1990s the spatial sciences, and the application of geographical information systems (GIS) in particular, suffered from a paucity of digital data sets. The 1990s were a period of transition into data-richness, a trend which accelerates today. Digital spatial data sets have grown rapidly in coverage, volume of records and numbers of attributes per record (Gahegan, 2003; Miller & Han, 2001). This state change has come about as a result of:

- improved technology and wider use of GPS, remote sensing and digital photogrammetry for collecting data on topographic and other physical objects;
- the introduction of new approaches to obtaining lifestyle and preference data such as through loyalty cards;
- dramatic increases in computing power to process raw data coupled with falling unit costs of data storage and data processing;
- the advent of data warehousing technologies;
- more efficient means of accessing and delivering data on-line.

The technical advances in hardware, software and data have been so profound that they have fundamentally affected the range of problems studied and the methodologies used to do so (Macmillan, 1998). An exponential rise in the size of databases, their increasing complexity and the rate at which they can accumulate on a daily basis have therefore lead to an urgent need for techniques that can mine very large databases for the knowledge they contain. Consequently, an active area of research has focused on *spatial data mining* which can be defined as *techniques for the discovery of meaningful patterns from large data sets where a prime dimension of interest is geographical location*. This paper focuses on clustering as a central aspect of spatial data mining and seeks to demonstrate the benefits of using 'hot spot' approaches to clustering in tandem with segmentation approaches to clustering. This is demonstrated using a case study analysis of a business transactions database. The following section discusses the theoretical perspectives and the dichotomy between the two different approaches to clustering. A form of 'hot spot' type clustering is then introduced and is subsequently used in the case study to guide a *k*-means classification of spatial and non-spatial attributes for a customer database. This forms the basis of a dual approach to cluster discovery as alluded to in the title of the paper.

2. Cluster detection in point event data

Transactions databases, be they for business, crime or health, can be regarded as point event data sets if each record has a specific geographical identifier such that geocoding can be achieved at the resolution of an address or postcode. From a location perspective the point event is a binary occurrence – either it happened there or it did not. From a data perspective, the binary occurrence may have added dimensions of attributes that describe the nature or content of the transaction which may relate to the location, the individual or the event that has been recorded. The traditional approach to non-spatial analyses of attributes may reveal apparently meaningful knowledge but may well be lacking in perspicacity or may even be misleading if underlying spatial distributions and dependencies are ignored. The exploratory analysis of point event data seeks to identify patterns using all