Chapter 21
On the MDBSCAN Algorithm in a Spatial Data Mining Context

Gabriella Schoier
Università di Trieste, Italy

ABSTRACT

The rapid developments in the availability and access to spatially referenced information in a variety of areas, has induced the need for better analysis techniques to understand the various phenomena. In particular, spatial clustering algorithms, which group similar spatial objects into classes, can be used for the identification of areas sharing common characteristics. The aim of this chapter is to present a density based algorithm for the discovery of clusters of units in large spatial data sets (MDBSCAN). This algorithm is a modification of the DBSCAN algorithm (see Ester (1996)). The modifications regard the consideration of spatial and non spatial variables and the use of a Lagrange-Chebychev metrics instead of the usual Euclidean one. The applications concern a synthetic data set and a data set of satellite images

INTRODUCTION

The development of new techniques and tools that support the humans in transforming data into useful knowledge has been the focus of a relatively new and interdisciplinary research area: knowledge discovery in databases (KDD), term coined to describe the process for finding relations among observed data.

Its heart is Data Mining which consists in the process of selection, modeling and application of algorithms to discover relations among large quantities of data. In particular Spatial Data Mining can be used for browsing spatial databases (see e.g. Bailey (1996), Koperski (1998)), understanding spatial data, discovering spatial relationships, optimizing spatial queries. Recently, clustering techniques have been recognized as primary Data Mining methods for knowledge discovery in spatial databases, i.e. databases managing 2D or 3D points, polygons etc. or points in some *d*-dimensional feature space. The well-known clustering algorithms, however, have some drawbacks when applied to large spatial databases (see e.g.

DOI: 10.4018/978-1-4666-2455-9.ch021

Han (2001)). On one side traditional algorithms seems to be inefficient when managing spatial data, on the other side problems arise when considering spatial and non-spatial data together for clustering. Algorithms for spatial data detects clusters in the geographical distribution of data but not always seem to be suited for considering also their attributes, as intensity, frequency or other characteristics of the observed phenomena (see e.g. Cressie (1993)).

In particular the application to large spatial databases involves the following requirements for clustering algorithms:

1. Minimal requirements of domain knowledge to determine the input parameters, because appropriate values are often not known in advance when dealing with large databases.
2. Discovery of clusters with arbitrary shape, because the shape of clusters in spatial databases may be spherical, drawn-out, linear, elongated etc..
3. Good efficiency on large databases, i.e. on databases of significantly more than just a few thousand objects.

In this paper we present some modifications of the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm proposed in Ester (1996) by considering spatial and non spatial variables (see Schoier (2004), Schoier (2007)) and by applying a Lagrange-Chebychev metric instead of the usual Euclidean distance. This allows improving the efficiency as regards time of execution, even augmenting the dimension or the number of elements of the data set.

Our proposal has been applied to synthetic datasets in particular to a problem of digital bitmat images reconstruction and to a problem of satellite images reconstruction which has been used widely in fields like urban studies, medicine, ecology etc...

SPATIAL DATA MINING

Data mining, or knowledge discovery in databases (KDD), is the technique of analyzing data to discover previously unknown information. The goal is to reveal regularities and relationships that are non-trivial.

Data Mining is often presented as a revolution in information processing. Two very well known definitions taken from the literature are:

1. Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad (1996)).
2. Data Mining consists in the discovery of interesting, unexpected, or valuable structures in large data sets (Hand (2000)).

The development of Data Mining is related with the availability of very large databases and the need of exploiting these bases in a new way. In the last fifteen years the Data Base Management (DBMS) community has become interested in using DBMS for decision support. They realized that data which have been collected for management organisation such as recording transactions might contain useful information for e.g. improving the knowledge of, and the service to customers.

Spatial data mining differs from regular data mining in parallel with the differences between non-spatial data and spatial data. The attributes of a spatial object stored in a database may be affected by the attributes of the spatial neighbors of that object. In addition, spatial location, and implicit information about the location of an object, may be exactly the information that can be extracted through Spatial Data Mining.

In order to successfully explore the massive amounts of spatial data being collected it is necessary to develop database primitives to manipulate the data.