Chapter 3
Data Discovery Approaches for Vague Spatial Data

Frederick E. Petry
Naval Research Laboratory, USA

ABSTRACT

This chapter focuses on the application of the discovery of association rules in approaches vague spatial databases. The background of data mining and uncertainty representations using rough set and fuzzy set techniques is provided. The extensions of association rule extraction for uncertain data as represented by rough and fuzzy sets is described. Finally, an example of rule extraction for both types of uncertainty representations is given.

INTRODUCTION

Data mining or knowledge discovery generally refers to a variety of techniques that have developed in the fields of databases, machine learning (Alpaydin, 2004) and pattern recognition (Han & Kamber, 2006). The intent is to uncover useful patterns and associations from large databases. For complex data such as that found in spatial databases (Shekar & Chawla, 2003) the problem of data discovery is more involved (Lu et al., 1993; Miller & Han, 2009).

Spatial data has traditionally been the domain of geography with various forms of maps as the standard representation. With the advent of computerization of maps, geographic information systems (GIS) have come to fore with spatial databases storing the underlying point, line and area structures needed to support GIS (Longley et al., 2001). A major difference between data mining in ordinary relational databases (Elmasri & Navathe, 2010) and in spatial databases is that attributes of the neighbors of some object of interest may have an influence on the object and therefore have to be considered as well. The explicit location and extension of spatial objects define implicit relations of spatial neighborhood (such as topological, distance and direction relations), which are used by spatial data mining algorithms (Ester et al 2000).

Additionally when wish to consider vagueness or uncertainty in the spatial data mining process

DOI: 10.4018/978-1-4666-2455-9.ch003

(Burrough & Frank 1996; Zhang & Goodchild, 2002), an additional level of difficulty is added. In this chapter we describe one of the most common data mining approaches, discovery of association rules, for spatial data for which we consider uncertainty in the extraction rules as represented by both fuzzy set and rough set techniques.

BACKGROUND

Data Mining

Although we are primarily interested here in specific algorithms of knowledge discovery, we will first review the overall process of data mining (Tan, Steinbach & Kumar, 2005). The initial steps are concerned with preparation of data, including data cleaning intended to resolve errors and missing data and integration of data from multiple heterogeneous sources. Next are the steps needed to prepare for actual data mining. These include the selection of the specific data relevant to the task and the transformation of this data into a format required by the data mining approach. These steps are sometimes considered to be those in the development of a data warehouse (Golfarelli & Rizzi, 2009), i.e., an organized format of data available for various data mining tools. There are a wide variety of specific knowledge discovery algorithms that have been developed (Han & Kamber, 2006). These discover patterns that can then be evaluated based on some interestingness measure used to prune the huge number of available patterns. Finally as true for any decision aid system, an effective user interface with visualization/alternative representations must be developed for the presentation of the discovered knowledge.

Specific data mining algorithms can be considered as belonging to two categories - descriptive and predictive data mining. In the descriptive category are class description, association rules and classification. Class description can either provide a characterization or generalization of the data or comparisons between data classes to provide class discriminations. Association rules are the main focus of this chapter and correspond to correlations among the data items (Hipp et al., 2000). They are often expressed in rule form showing attribute-value conditions that commonly occur at the same time in some set of data. An association rule of the form $X \rightarrow Y$ can be interpreted as meaning that the tuples in the database that satisfy the condition X also are "likely" to satisfy Y, so that the "likely" implies this is not a functional dependency in the formal database sense. Finally, a classification approach analyzes the training data (data whose class membership is known) and constructs a model for each class based on the features in the data. Commonly, the outputs generated are decision trees or sets of classification rules. These can be used both for the characterization of the classes of existing data and to allow the classification of data in the future, and so can also be considered predictive.

Predictive analysis is also a very developed area of data mining. One very common approach is clustering (Mishra et al., 2004). Clustering analysis identifies the collections of data objects that are similar to each other. The similarity metric is often a distance function given by experts or appropriate users. A good clustering method produces high quality clusters to yield low inter-cluster similarity and high intra-cluster similarity. Prediction techniques are used to predict possible missing data values or distributions of values of some attributes in a set of objects. First, one must find the set of attributes relevant to the attribute of interest and then predict a distribution of values based on the set of data similar to the selected objects. There are a large variety of techniques used, including regression analysis, correlation analysis, genetic algorithms and neural networks to mention a few.

Finally, a particular case of predictive analysis is time-series analysis. This technique considers a large set of time-based data to discover regularities and interesting characteristics (Shasha &