Abstract. The need to analyze the vast quantities of weather data collected has led to the development of new data mining tools and techniques. Mining this data can produce new insights into weather, climatological and environmental trends that have both scientific and practical significance. This chapter discusses the challenges posed by weather databases and examines the use of fuzzy clustering for analyzing such data. It proposes the extension of the fuzzy K-Means clustering algorithm to account for the spatio-temporal nature of weather data. It introduces an unsupervised fuzzy clustering algorithm, based on the fuzzy K-Means and defines a cluster validity index which is used to determine an optimal number of clusters. These techniques are validated on weather data in the South Central US, and global climate data (sea level pressure). It is seen that the algorithm is able to identify and preserve interesting phenomena in the weather data.

5.1. Introduction

Satellites, weather stations, and sensors are collecting large volumes of geospatial data on a wide range of parameters. Examples of such geospatial data include earth science data describing spatiotemporal phenomena, multi-dimensional sequences of images in geographic regions, the evolution of natural phenomena, etc. However, despite the importance of such geospatial data sets it is only recently that there have been efforts to develop appropriate data mining and knowledge discovery in databases tools and techniques suitable for data analysis.

The spatio-temporal domain is complex (Gahegan 2001) and characterized by high volumes of data. For example several global landcover/landuse maps have terabytes of information, requiring computationally intense analysis techniques. Interesting signals and data are often masked by stronger signals caused by local effects, such as seasonal variations. The coupling between different regions of the globe also introduces complexities in behavior. These effects make analysis of this data difficult. Non-uniformity in data gathering and sampling sometimes require indirect measurements and interpolation, which lead to the introduction of
model artifacts. Formulating geographic knowledge, and applying it to knowledge discovery and data mining is also difficult. This level of complexity is evident in data mining and knowledge discovery in databases techniques for this domain which have encompassed a wide range of models and techniques (Roddick and Spiliopoulou 1999). These techniques vary from the visual oriented approach (Openshaw 1984), the generation of geographic association rules and sequence rules (Koperski and Han 1995), clustering (Steinbach et al. 2003; Smyth et al. 2000), to combinational approaches (Gahegan et al. 2001).

The application of fuzzy logic techniques to data mining and knowledge discovery in the weather domain has several advantages. Imprecision and uncertainty in this domain is present at several levels. Attribute ambiguity occurs when class membership is partial or unclear. Attribute ambiguity is a severe problem in remotely sensed data (Mohan 2000), such as aerial photography, which is often interpreted inconsistently. Spatial vagueness emerges when the sampling resolution is not fine enough to identify boundary locations exactly, where gradual transitions occur between classes, or when there is location uncertainty. Clustering is a technique which helps in the analysis of such large data sets through the definition of regions that have similar properties. However, conventional hard clustering (Steinbach et al. 2003) is inappropriate when faced with the ambiguities in weather data measurement. The data often is incomplete or has errors in measurement, and the spatial and attribute ambiguities that are characteristic to this data introduce further difficulties into the analysis. Fuzzy clustering is more appropriate for this data with its ability to naturally incorporate these real world issues. The ability to produce soft boundaries permits improved interpretation capabilities.

There are several interesting applications for this research effort. At a regional scale, these clustering techniques may identify micro-climatic regions. The knowledge of these regions can be used to improve operations planning and decision support. The presence of such microclimatic regions is subjective even for domain experts. Fuzzy clustering techniques are better suited to providing a basis for these interpretations in comparison to hard partitioning. The use of clustering techniques is also useful for tuning weather prediction models. These models while good predictors of long term weather fail spectacularly in predicting short term weather patterns (the Santa Ana winds, for instance). Terrain and diurnal heating effects cannot be effectively modeled in these models and the ability to recognize such weather anomalies can be used as corrections to the forecasts produced by the weather models. The development of data mining and knowledge discovery in databases techniques can thus lead to the understanding and prediction of such effects.

Clustering techniques have been widely used in data mining and knowledge discovery in databases and are ideal for understanding weather data. Fuzzy clustering is an extension of the classical clustering technique and has been used to solve numerous problems in the areas of pattern recognition and fuzzy model identification (MacQueen 1967). A variety of fuzzy clustering methods have been proposed and several of them are based upon distance criteria. Fuzzy K-Means clustering has been widely used for understanding patterns especially where the