Data Mining of Population Distribution Rules: an Attribute-Oriented Approach

Liu Deqin, Ma Weijun
Chinese Academy of Surveying and Mapping,
Beijing 100039, China
liudq@casm.ac.cn

Data mining is the extraction of interesting patterns concealed in large databases. Attribute-Oriented induction (AOI) is a set oriented generalization technique used to find various types of rules. The method integrates learning-from-examples techniques with database procedures. In the census database, the vast number of population data is stored. The qualitative method is mainly used for the analysis of the population data in the past, which lead to lack of quantitive answers. The paper presents an attribute-oriented induction approach for the discovery of population distribution rules.

The theory and technology of data mining based on attribute-oriented induction is introduced. The AOI method basically involves three primitives that specify thelearning task. These are collection of initial task-relevant data (Data Collection), use of background knowledge (Domain knowledge) during the mining process and representation of the learning result (Rule formation). The fundamental principle in AOI is to generalise the initial relation to a prime relation and then to a final relation using background knowledge and user-defined threshold. Besides of the population database, the geomorphologic data is also used in the study, which includes 17 classes and 55 sub-classes.

The analysis steps are as follows:

1.   Preparation of the data, which includes the digitization of the geomorphology map, retrieving the county boundary data from the existing database and establishing the population database.
2.   Overlay analysis of the boundary data with the population data.
3.   Calculation of the population data in each geomorphologic sub-classes and obtaining the spatially distributing population data according the geomorphologic classes.
4.   Data mining of rules of population spatially distribution by attribute-oriented approach

Finally, a set of rules of population spatially distribution is obtained, such as, 1) In the north part of china, the population density is in the middle-low level, and its density of population with college and above education attainment is also in the middle-low class; 2) In the high mountain and plateau area of south-west part of china, the population density is in the low class, and its density of population with college and above education attainment is also in the low class, but in the low mountain, hill and mountainous region of south-west part of china, the population density is in the middle class, and its density of population with college and above education attainment is in the low class, etc..

The government has a huge amount of population data from the population census, the important knowledge of population spatial distribution can be obtained through the data mining technology. By these methods, the application of population data will be improved, for the better decision-making assistance in the economic and social development planning.