Chapter XV
An Approach to Mining Crime Patterns

Sikha Bagui
The University of West Florida, USA

ABSTRACT

This paper presents a knowledge discovery effort to retrieve meaningful information about crime from a U.S. state database. The raw data were preprocessed, and data cubes were created using Structured Query Language (SQL). The data cubes then were used in deriving quantitative generalizations and for further analysis of the data. An entropy-based attribute relevance study was undertaken to determine the relevant attributes. A machine learning software called WEKA was used for mining association rules, developing a decision tree, and clustering. SOM was used to view multidimensional clusters on a regular two-dimensional grid.

INTRODUCTION

This article discusses an approach to mining; that is, seeking interesting crime patterns in a type of census data — U.S. state data. With an increasing crime rate and enormous amounts of data being stored in crime databases, it is becoming increasingly important to discover knowledge about crime from databases; that is, mining crime databases. Several other trends in data mining applications also can be found in Chen and Liu (2005) and Hu et al. (2005) (the latter paper discusses usage of data mining in the clinical area). By mining data, we refer to a process of nontrivial extraction of implicit, previously unknown, and

potentially useful information, such as knowledge rules, constraints, regularities, and so forth (Agrawal, Imielinski, & Swami, 1993).

The dataset used in this study was retrieved from http://www.unl.edu/SPPQ/datasets.html, a State Politics and Policy Quarterly Data Resource Web site in the United States. This data set is collected and maintained by the U.S. State Data Center in partnership with the U.S. Census Bureau. The U.S. State Data Centers are official sources of demographic, economic, and social statistics produced by the U.S. Census Bureau. This Web site also contains the data dictionary for this data set.

Our main goal in this article was to discover knowledge on crime in the U.S. from this database by using several data mining techniques. First, data cubes were used to derive and to present quantitative generalizations from the data using *t-weights* and *d-weights*. Then, since our data were unsupervised (where the data had no predefined classes), we used advanced data mining techniques like (1) association rules mining to find the attribute or set of attributes related to crime; (2) decision tree analysis to help us develop classification rules for low/high crime based on a decision tree; and (3) clustering algorithms to see if any natural groupings could be found in the data. Neural networks and other advanced classification techniques would be useful, if the data were supervised (if we had predefined classes); hence, we did not use them for our data analysis.

The rest of the article is organized as follows. The second section deals with preprocessing the data. Preprocessing involves cleaning the data, data generalization in the form of concept hierarchy generation, and developing data cubes. Concept hierarchies and the data cubes allow data to be viewed from multiple angles and present generalized views of the data at multiple levels of abstraction. Data preprocessing mainly was done using SQL (Bagui & Earp, 2004). In the third section, we present derived generalizations, *t_weights* and *d_weights*, calculated using our concept hierarchies and data cubes. In the fourth section, an entropy-based attribute relevance study was undertaken to determine the relevant attributes in the data set. The fifth section explores the idea of discovering association rules. Association rules are generally used with basket, census, or financial data. The sixth section discusses the use of a decision tree to come up with classification rules for crime from this U.S. state data set. In the seventh section, clustering techniques were applied to find the number of clusters, and the centroids of the clusters were analyzed. The results of each of these sections are presented at the end of the respective sections. Finally, in the eighth section, we present the overall conclusions of this study.

A machine learning software, WEKA, was used for mining association rules, developing the decision tree, and clustering. WEKA, which stands for Waikato Environment for Knowledge Analysis, is a collection of machine learning algorithms for solving real-world data mining problems. Written in Java, WEKA runs on almost any platform and is available on the Web at www.cs.waikato.ac.nz/ml/weka (Witten & Frank, 2000). An Excel-based SOM tool (available at www.geocities.com/adotsaha/NN/SOMinExcel.html) was used to generate a SOM graph to view multidimensional clusters on a regular two-dimensional grid.

PREPROCESSING THE DATA

This U.S. state dataset contains various census items like employment, unemployment, area population, crime and law enforcement, and so forth. Our preprocessing and knowledge discovery efforts concentrated on the attributes related to crime — year, abrev (which stands for, state), murder, rape, robbery, assault, burglary, larceny, mvtheft (motor vehicle theft), police (meaning, police population), and pop (meaning, area population) for the years 1980 to 1989. Our preprocessing