Symbolic Data Clustering

Edwin Diday
University of Dauphine, France

M. Narasimha Murty
Indian Institute of Science, India

INTRODUCTION

In data mining, we generate class/cluster models from large datasets. Symbolic Data Analysis (SDA) is a powerful tool that permits dealing with complex data (Diday, 1988) where a combination of variables and logical and hierarchical relationships among them are used. Such a view permits us to deal with data at a conceptual level, and as a consequence, SDA is ideally suited for data mining. Symbolic data have their own internal structure that necessitates the need for new techniques that generally differ from the ones used on conventional data (Billard & Diday, 2003). Clustering generates abstractions that can be used in a variety of decision-making applications (Jain, Murty, & Flynn, 1999). In this article, we deal with the application of clustering to SDA.

BACKGROUND

In SDA, we consider multivalued variables, products of interval variables, and products of multivalued variables with associated weights (Diday, 1995). Clustering of symbolic data (Gowda & Diday, 1991; De Souza & De Carvalho, 2004) generates a partition of the data and also descriptions of clusters in the partition using *symbolic objects*. It can have applications in several important areas coming under data mining:

- **Pattern Classification:** The abstractions generated can be used for efficient classification (Duda, Hart, & Stork, 2001).
- **Database Management:** SDA permits generation of symbolic objects from relational databases (Stéphan, Hébrail, & Lechevallier, 2000). Usage of data in aggregate form where variables assuming interval values can be handy. This not only permits a brief description of a large dataset but also in dealing with privacy issues associated with information of an individual (Goupil, Touati, Diday, & Moult, 2000). An important source of symbolic data is provided by relational databases if we have an application that needs several relations merged (Bock & Diday, 2000).
- **Knowledge Management:** It is possible to extract meaningful conceptual knowledge from clustering symbolic data. It is also possible to use expert knowledge in symbolic clustering (Rossi & Vautrain, 2000).
- **Biometrics:** Clustering is used in a variety of biometric applications, including face recognition, fingerprint identification, and speech recognition. It is also used in protein sequence grouping (Zhong & Ghosh, 2003).

The SDA Community enjoys a right mix of theory and practice. The Symbolic Official Data Analysis System (SODAS) software package developed over the past few years is available for free distribution (Morineau, 2000).

MAIN THRUST

We deal with various components of a symbolic data-clustering system in this section.

Symbolic Data Analysis (SDA)

In SDA the input comes in the form of a table; columns of the table correspond to *symbolic variables,* which are used to describe a set of individual patterns. Rows of the table are *symbolic descriptions* of these individuals. They are different from the conventional descriptions that employ a vector of quantitative or categorical values to represent an individual (Jain & Dubes, 1988). The cells of this symbolic data table may contain data of the following types:

1. **A single quantitative value:** for example, height (John) = 6.2.
2. **A single categorical value:** for example, color_of_eyes (John) = blue.

3. **A set of values or categories:** for example, room_number (John) = {203, 213, 301}, which means that the number of John's room is either 203, 213, or 301.
   An Interval: for example, height (John) = [6.0,6.3], which means that John's height is in the interval [6.0,6.3]; note that the minimum and maximum values of an interval are instances of an interval variable. So the interval is an instance of an ordered pair of interval variables.
5. **An ordered set of values with associated weights:** here we have either a *histogram* or *a membership function.* In the case of the histogram, the weight is the normalized frequency of occurrence, and in the case of membership function, the weight corresponds to the membership of the value in the concept. Note that this definition permits us to deal with variables that have probability distributions as their values, or functions as their values, also.

In addition, it is possible to have logical and structural relationships among these variables. For example, the statement "If the *age* of John is between one and two months, then the *height* of John is between 30 and 40 centimeters" is a logical implication. Two or more variables could be hierarchically related (Bock & Diday, 2000). For example, the variable color is considered to be *light* if it is *yellow, white,* or *metallic*. Similarly, we can describe the make and model of a car if one owns a car, which depicts dependency between variables (Bock & Diday, 2000).

Symbolic objects are formal operational models of concepts. A concept in the real world is mathematically described by a symbolic object; it may use a formula in a classical logic or a multivalued logic to describe the concept. In addition, the symbolic object provides a way to calculate the extent of a concept, which is a set of individuals in the real world associated with the concept (Diday, 2002). The important step in symbolic clustering is to output symbolic objects corresponding to the clustering. These output symbolic descriptions are used in a variety of decision-making situations and can be used again as new units for a higher level analysis or clustering (Bock & Diday, 2000).

Dissimilarity Measures for Symbolic Objects

In conventional clustering, we put similar objects in the same group and dissimilar objects in different groups (Jain et al., 1999). So the notion of similarity/dissimilarity plays an important role in arriving at the partition of the dataset. A good collection of dissimilarity measures is used in dealing with the conventional data consisting of only numerical or categorical variables (Duda et al., 2001). The need for computing dissimilarities between symbolic objects is obvious because we would like to group, for reducing both time and space requirements, symbolic objects that are summaries of groups of objects. An excellent collection of dissimilarity measures between symbolic objects is given in Esposito, Malerba, and Lisi (2000).

It is possible to use a distance function to capture the dissimilarity. A simple view is to accept that similarity can be obtained from dissimilarity between objects. However, it may be inspiring to view similarity and dissimilarity as complimenting each other. A variety of dissimilarity functions are defined and used in symbolic clustering. A most popular dissimilarity measure is the one proposed by De Carvalho (1998). Dissimilarity measures for histograms and probability distributions are reported in Bock and Diday (2000).

Grouping Algorithms

Traditionally, clustering algorithms are grouped into hierarchical and partitional categories (Jainet al., 1999). The hierarchical algorithms are computationally expensive, as they need to either compute and store a proximity matrix of size quadratic in the number of patterns or compute the proximity based on need using time that is cubic in the number of patterns. Even the incremental hierarchical algorithms need time that is quadratic in the number of objects. So even though hierarchical algorithms are versatile, they may not scale up well to handle large datasets.

The partitional algorithms, such as the dynamic clustering algorithm (Diday & Simon, 1976), are better as they take linear time in the number of inputs. So they have been successfully applied to moderately large datasets. The dynamic clustering algorithm may be viewed as a *K*-kernels algorithm, where a kernel could be the mean, a line, multiple points, probability law, and other more general functions of the data. Such a general framework was proposed for the first time in the form of the dynamic clustering algorithm. The well-known *k*-means algorithm (Duda et al., 2001) is a special case of the dynamic clustering algorithm where the kernel of a cluster is the centroid. However, most of these partitional algorithms are iterative in nature and may require a data scan several times.

It will be useful to explore schemes that can help in scaling up the existing symbolic clustering algorithms. The possible solutions are to:

- Use an incremental clustering algorithm (Jain et al., 1999). One of the simplest incremental algo-