Knowledge Discovery in Databases with Diversity of Data Types

QingXiang Wu
University of Ulster at Magee, UK

Martin McGinnity
University of Ulster at Magee, UK

Girijesh Prasad
University of Ulster at Magee, UK

David Bell
Queen's University, UK

INTRODUCTION

Data mining and knowledge discovery aim at finding useful information from typically massive collections of data, and then extracting useful knowledge from the information. To date a large number of approaches have been proposed to find useful information and discover useful knowledge; for example, decision trees, Bayesian belief networks, evidence theory, rough set theory, fuzzy set theory, kNN (k-nearest-neighborhood) classifier, neural networks, and support vector machines. However, these approaches are based on a specific data type. In the real world, an intelligent system often encounters mixed data types, incomplete information (missing values), and imprecise information (fuzzy conditions). In the UCI (University of California – Irvine) Machine Learning Repository, it can be seen that there are many real world data sets with missing values and mixed data types. It is a challenge to enable machine learning or data mining approaches to deal with mixed data types (Ching, 1995; Coppock, 2003) because there are difficulties in finding a measure of similarity between objects with mixed data type attributes. The problem with mixed data types is a long-standing issue faced in data mining. The emerging techniques targeted at this issue can be classified into three classes as follows: (1) Symbolic data mining approaches plus different discretizers (e.g., Dougherty et al., 1995; Wu, 1996; Kurgan et al., 2004; Diday, 2004; Darmont et al., 2006; Wu et al., 2007) for transformation from continuous data to symbolic data; (2) Numerical data mining approaches plus transformation from symbolic data to numerical data (e.g.,, Kasabov, 2003; Darmont et al., 2006; Hadzic et al., 2007); (3) Hybrid of symbolic data mining approaches and numerical data mining approaches (e.g.,, Tung, 2002; Kasabov, 2003; Leng et al., 2005; Wu et al., 2006). Since hybrid approaches have the potential to exploit the advantages from both symbolic data mining and numerical data mining approaches, this chapter, after discassing the merits and shortcomings of current approaches, focuses on applying Self-Organizing Computing Network Model to construct a hybrid system to solve the problems of knowledge discovery from databases with a diversity of data types. Future trends for data mining on mixed type data are then discussed. Finally a conclusion is presented.

BACKGROUND

Each approach for data mining or knowledge discovery has its own merits and shortcomings. For example, EFNN (Evolving Fuzzy Neural Network based on Tokagi-Sgeno fuzzy rules) (Kasabov, 2003; Takagi and Sugeno, 1985), SOFNN (Leng et al., 2005; Kasabov, 2003; Tung, 2002), dynamic fuzzy neural networks, kNN, neural networks, and support vector machines, are good at dealing with continuous valued data. For example, the EFNN (Kasabov, 2003) was applied to deal with benchmark data sets--the gas furnace times series data and the Mackey-Glass time series data (Jang, 1993). High accuracies were reached in the predictive results. The errors were very small i.e. 0.156 for the

Gas-furnace case and 0.039 for the Mackey-Glass case. However, they cannot be directly applied to symbolic data or to a data set with missing values. Symbolic AI techniques (Quinlan, 1986, Quinlan, 1996, Wu et al., 2005) are good at dealing with symbolic data and data sets with missing values. In order to discover knowledge from a database with mixed-type data, traditional symbolic AI approaches always transform continuous valued data to symbolic data. For example, the temperature is a continuous data, but it can be transformed to symbolic data 'cool', 'warm', 'hot', etc. This is a typical transformation of one dimension of continuous data, which is called *discretization*. The transformation for two or more dimensions of continuous data such as pictures or videos can be regarded as object recognition or content extraction. However, information about distance and neighborhood in continuous valued data is ignored if the discretized values are treated as symbolic values in symbolic AI techniques.

On the other hand, symbolic data can be transformed to numerical data using some encoding scheme. This can be done by statistic, rough sets or fuzzy membership functions. For example, 'high', 'mid', and 'low' can be sorted in a sequence and can be represented by fuzzy member functions. However, it is difficult to encode symbols without an explicit sequence, e.g., symbolic values for furniture: 'bed', 'chair', 'bench', 'desk' and 'table'. If the symbols have to be sorted out in a sequence, some additional information is required. For example, they can be sorted by their size or price if the information of price or size are known. Therefore, correct data transformation plays a very important role in data mining or machine learning.

MAIN FOCUS

The Self-Organizing Computing Network Model (Wu et al., 2006) provides a means to combine the transformations and data mining/knowledge discovery approaches to extract useful knowledge from databases with a diversity of data types, and the knowledge is represented in the form of a computing network. The model is designed using a hybrid of symbolic and numerical approaches. Through an analysis of which data type is suitable to which data mining or machine learning approach, data are reclassified into two new classes -- *order dependent attribute* and *order independent attribute*. Then concepts of fuzzy space, statistical learning, neural networks and traditional AI technologies are integrated to the network model to represent knowledge and self-adapt to an instance information system for decision making.

Proper Data Type Transformation

Usually, data can be categorized in two types, i.e. numerical data and symbolic data. From a data mining or machine learning point of view, attribute values can be separated into two classes. If the values of an attribute can be sorted out in a sequence and a distance between two values is significant to data mining or machine learning, the attribute is called an *order dependent attribute*. Numerical attribute can be separated into two kinds of attributes, i.e. a continuous valued attribute and an encoding numerical attribute. A continuous valued attribute is an *order dependent attribute* because a distance between two values can be used to describe neighbors or similarities in data mining or machine learning algorithms. Some encoding numerical attributes are an *order dependent attribute* such as grade numbers 1 to 5 for student courses. Some encoding numerical data such as product identification numbers are not an *order dependent attribute*. There is a distance between two values, but the distance is not significant to data mining or machine learning algorithms. We cannot say product No.1 and product No.2 certainly have similarity or can be regarded as neighbors. If this distance is used in data mining or machine learning algorithms, the results will be degraded. If the values of an attribute cannot be sorted out in a sequence or a distance between two values is not significant to data mining or machine learning, the attribute is called an *order independent attribute*. For example, some symbolic attributes are an *order independent attribute* in which there is neither definition of a distance between two symbolic values nor definition of value sequences or neighborhoods. However, there are some symbolic data with an explicit sequence; for example, 'high', 'mid', and 'low'. These symbolic values are suitable for transfer to numerical data so that value sequence and neighborhood can be used by data mining or machine learning algorithms. Therefore, two attribute channels are designed in the input layer of the Self-Organizing Computing Network Model to lead an attribute with a given data type to a suitable data mining approach. The first channel is called an *order dependent attribute* channel. The data mining or machine learning approaches, which can take