INTRODUCTION

A great deal of interesting real-world data is encountered through the analysis of continuous variables, however many of the robust tools for rule discovery and data characterization depend upon the underlying data existing in an ordinal, enumerable or discrete data domain. Tools that fall into this category include much of the current work in fuzzy logic and rough sets, as well as all forms of event-based pattern discovery tools based on probabilistic inference.

Through the application of discretization techniques, continuous data is made accessible to the analysis provided by the strong tools of discrete-valued data mining. The most common approach for discretization is quantization, in which the range of observed continuous valued data are assigned to a fixed number of quanta, each of which covers a particular portion of the range within the bounds provided by the most extreme points observed within the continuous domain. This chapter explores the effects such quantization may have, and the techniques that are available to ameliorate the negative effects of these efforts, notably fuzzy systems and rough sets.

BACKGROUND

Real-world data sets are only infrequently composed of discrete data, and any reasonable knowledge discovery approach must take into account the fact that the underlying data will be based on continuous-valued or mixed mode data. If one examines the data at the UCI Machine-Learning Repository (Newman, Hettich, Blake & Merz, 1998) one will see that many of the data sets within this group are continuous-valued; the majority of the remainder are based on measurements of continuous valued random variables that have been pre-quantized before being placed in the database.

The tools of the data mining community may be considered to fall into the following three groups:

- minimum-error-fit and other gradient descent models, such as: support vector machines (Cristianini & Shawe-Taylor, 2000; Duda, Hart & Stork, 2001; Camps-Valls, Martínez-Ramón, Rojo-Álvarez & Soria-Olivas, 2004); neural networks (Rumelhart, Hinton & Williams, 1986); and other kernel or radial-basis networks (Duda, Hart & Stork, 2001; Pham, 2006)
- Bayesian-based learning tools (Duda, Hart & Stork, 2001), including related random-variable methods such as Parzen window estimation
- statistically based pattern and knowledge discovery algorithms based on an event-based model. Into this category falls much of the work in rough sets (Grzymala-Busse, & Ziarko, 1999; Pawlak, 1982,1992; Singh & Minz, 2007; Slezak & Wroblewski, 2006), fuzzy knowledge representation (Boyen & Wehenkel, 1999; Gabrys 2004; Hathaway & Bezdek 2002; Höppner, Klawonn, Kruse & Runkler, 1999), as well as true statistical methods such as “pattern discovery” (Wang & Wong, 2003; Wong & Wang, 2003; Hamilton-Wright & Stashuk, 2005, 2006).

The methods in the last category are most affected by quantization and as such will be specifically discussed in this chapter. These algorithms function by constructing rules based on the observed association of data values among different quanta. The occurrence of a feature value within particular quanta may be considered an “event” and thereby all of the tools of information theory may be brought to bear. Without
the aggregation of data into quanta, it is not possible
to generate an accurate count of event occurrence or
estimate of inter-event relationships.

**MAIN FOCUS**

The discretization of continuous-valued data can be
seen as a clustering technique in which the ranges
of observed values are assigned to a limited set of $Q$
cluster labels (sometimes referred to as a Borel set).
The success or failure of a quantization structure may
therefore be evaluated in terms of how well each of
the $Q$ clusters represents a homogenous and useful
grouping of the underlying data.

The action of quantization is performed as a first
step towards the discovery of the data topology, and
therefore must frequently be undertaken without a great
deal of knowledge of the underlying structure of the
data. For this reason, such discretization is usually done
using the information available in a single feature. Two
major strategies for this are feature value partitioning
and quantization.

**Feature Value Partitioning**

Partitioning schemes, such as those described in ID3
and C4.5 (Quinlan, 1986; 1993) as well as those used
in the WEKA project (Witten & Frank, 2000) rely
upon an analysis of decisions to be made within a
single feature to provide a classification specific means
of dividing the observed data values between labels.
Quinlan (1993) provides an excellent discussion of an
information-theoretic based approach to the construc-
tion of per-feature partitioning schemes in the discus-
sion of the C4.5 classifier. In any such partitioning
scheme, the placement of the partition is chosen as a
means to optimize a classification decision made based
on a single feature.

Quinlan’s (1993) discussion is particularly salient, as
it is in such tree-based classifiers that this treatment is
the most advantageous, because the primary feature of
a partitioning mechanism is that each feature is treated
independently. This supports classification algorithms
that are based on a decision tree, but does not support
the discovery of multi-feature, high-order events.
Furthermore, note that a classifier label value must be
known in advance in order to use this technique; all
data is therefore viewed in terms of its ability to provide
support for some particular label value.

**Feature Value Quantization**

Quantization, on the other hand, refers to the construc-
tion of a set of range-based divisions of the input feature
space, where each distinct quantization “bin” represents
a projection of an input feature through an aggregation
scheme, independent of label value. By constructing
such a quantization independently of class label values,
it is therefore possible to support the discovery of data
patterns independent of any potential classification
structure. Feature value quantization underlies most
fuzzy systems (Pedrycz, 1995; Pal & Mitra, 1999) as
well as discrete information theoretic based approaches
such as the “pattern discovery” algorithm (Wang &
Wong, 2003; Wong & Wang, 2003; Hamilton-Wright
& Stashuk, 2006).

It is through the introduction of uncertainty-manage-
tment techniques that the “cost of quantization” may be
ameliorated. This cost is inherent in the structure of the
quantization resulting from a particular technique.

**Properties of Quantization**

The main strength of quantization is that by reduc-
ing a continuous domain problem to a discrete one,
the powerful tools of event-based reasoning may be
brought to bear. By taking a problem from the con-
tinuous domain to a discrete one, the random variables
underlying the model of the continuous domain may
be represented by discrete state variables, thereby al-
lowing the representation of the data domain as a series
of events with possible correlation (for an example,
see Ross, 1985).

As an added benefit, such quantization provides
protection against measurement noise (as analysis will
be insensitive to perturbation in measurement value)
provided such perturbation does not change the bin as-
signment. This simplifies the overall problem, reducing
computational complexity, which in turn allows the
application of algorithms that locate and analyze high-
order patterns (i.e.; significant correlations between a
large number of input features) such as the “pattern
discovery” algorithm just mentioned.

All problems arising from the use of quantization
stem from the fact that no quantization scheme can