Chapter I
Discretization of Rational Data

Jonathan Mugan
University of Texas at Austin, USA

Klaus Truemper
University of Texas at Dallas, USA

ABSTRACT

Frequently, one wants to extend the use of a classification method that, in principle, requires records with True/False values, so that records with rational numbers can be processed. In such cases, the rational numbers must first be replaced by True/False values before the method may be applied. In other cases, a classification method in principle can process records with rational numbers directly, but replacement by True/False values improves the performance of the method. The replacement process is usually called discretization or binarization. This chapter describes a recursive discretization process called Cutpoint. The key step of Cutpoint detects points where classification patterns change abruptly. The chapter includes computational results, where Cutpoint is compared with entropy-based methods that, to date, have been found to be the best discretization schemes. The results indicate that Cutpoint is preferred by certain classification schemes, while entropy-based methods are better for other classification methods. Thus, one may view Cutpoint to be an additional discretization tool that one may want to consider.

INTRODUCTION

One often desires to apply classification methods that, in principle, require records with True/False values to records that, besides True/False values, contain rational numbers. For ease of reference, we call rational number entries rational data and refer to True/False entries as logic data. In such situations, a discretization process must first convert the rational data to logic data. Discretization is also desirable in another setting. Here, a classification method in principle can process records with rational numbers directly, but its performance is improved when the rational data are first converted to logic data.
This chapter describes a method called Cutpoint for the discretization task, and compares its effectiveness with that of entropy-based methods, which presently are considered to be the best discretization schemes.

Define nominal data to be elements or subsets of a given finite set. In Bartnikowski et al. (Bartnikowski, Granberry, Mugan, & Truemper, 2004), an earlier version of Cutpoint is described and used for the transformation of some cases of nominal data to logic data. Specifically, the nominal data are first converted to rational data, which are then transformed to logic data by Cutpoint.

We focus here on the following case. We are given records of two training classes, $A$ and $B$, that have been randomly selected from two populations $A$ and $B$, respectively. We want to derive a classification scheme from the records of $A$ and $B$. Later, that scheme is to be applied to records of $A - A$ and $B - B$.

For the purpose of a simplified discussion in this section, we assume for the moment that the records have no missing entries. That restriction is removed in the next section.

**Abrupt Pattern Changes and Cutpoint**

Generally, the discretization may be accomplished by the following, well-known approach. One defines, for a given attribute, $k \geq 1$ breakpoints and encodes each rational number of the attribute by $k$ True/False values, where the $j$th value is True if the rational number is greater than the $j$th breakpoint, and is False otherwise. The selection of the $k$ breakpoints requires care if the records of $A - A$ and $B - B$ are to be classified with good accuracy.

A number of techniques for the selection of the breakpoints have been proposed, and later in this section, we provide a review of those methods. Suffice it to say here that the most effective methods to date are based on the notion of entropy. In these methods, the breakpoints are so selected that the rational numbers of a given attribute can be most compactly classified by a decision tree as coming from $A$ or $B$. In contrast, Cutpoint is based on a different goal. Recall that the records of the sets $A$ and $B$ are presumed to be random samples of the populations $A$ and $B$. Taking a different viewpoint, we may view each record of $A - A$ and $B - B$ to be a random variation of some record of $A$ or $B$, respectively. The goal is then to select the breakpoints so that these random variations largely leave the True/False values induced by the breakpoints unchanged.

Cutpoint aims for the stated goal by selecting breakpoints, called markers, that correspond to certain abrupt changes in classification patterns, as follows. First, for a given attribute, the rational numbers are sorted. Second, each value is labeled as $A$ or $B$, depending on whether the value comes from a record of $A$ or $B$, respectively. For the sake of a simplified discussion, we ignore, for the moment, the case where a rational number occurs in both a record of $A$ and a record of $B$. Third, each entry with label $A$ (resp. $B$) is assigned a class value of 1 (resp. 0). Fourth, Gaussian convolution is applied to the sequence of class values, and the midpoint between two adjacent entries, where the smoothed class values change by the largest amount, is declared to be a marker.

For example, if the original sorted sequence, with class membership in parentheses, is ..., 10.5($A$), 11.7($A$), 15.0($A$), 16.7($A$), 19.5($B$), 15.2($B$), 24.1($B$), 30.8($B$),..., then the sequence of class values is ..., 1, 1, 1, 1, 0, 0, 0, 0,..., Note the abrupt transition of the subsequence of 1s to the subsequence of 0s. When a Gaussian convolution with small standard deviation $\sigma$ is performed on the sequence of class values, a sequence of smoothed values results, which exhibits a relatively large change at the point where the original sequence changes from 1s to 0s. If this is the largest change for the entire sequence of smoothed class values, then the original entries 16.7($A$) and 19.5($B$), which correspond to that change, produce a marker with value $(16.7 + 19.5)/2 = 18.1$. 

2