Available online at www.sciencedirect.com
ScienceDirect
Knowledge-Based Systems 20 (2007) 419–425
Knowledge-Based SYSTEMS
www.elsevier.com/locate/knosys

A Hellinger-based discretization method for numeric attributes in classification learning

Chang-Hwan Lee *

Department of Information and Communications, DongGuk University, 3-26 Pil-Dong, Chung-Gu, Seoul 100-715, Republic of Korea

Received 14 February 2004; accepted 3 June 2006
Available online 20 October 2006

Abstract

Many classification algorithms require that training examples contain only discrete values. In order to use these algorithms when some attributes have continuous numeric values, the numeric attributes must be converted into discrete ones. This paper describes a new way of discretizing numeric values using information theory. Our method is context-sensitive in the sense that it takes into account the value of the target attribute. The amount of information each interval gives to the target attribute is measured using Hellinger divergence, and the interval boundaries are decided so that each interval contains as equal amount of information as possible. In order to compare our discretization method with some current discretization methods, several popular classification data sets are selected for discretization. We use naive Bayesian classifier and C4.5 as classification tools to compare the accuracy of our discretization method with that of other methods.

Keywords: Machine learning; Discretization; Data mining; Knowledge discovery

1. Introduction

Discretization is a process which changes continuous numeric values into discrete categorical values. It divides the values of a numeric attribute into a number of intervals, where each interval can be mapped to a discrete categorical or nominal symbol. Most real-world applications of classification algorithm contains continuous numeric attributes. When the feature space of data includes continuous attributes only or mixed type of attributes (continuous type along with discrete type), it makes the problem of classification vitally difficult. For example, classification methods based on similarity-based measures are generally difficult, if not possible, to apply to such data because the similarity measures defined on discrete values are usually not compatible with similarity of continuous values. Alternative methodologies such as probabilistic modeling, when applied to continuous data, require an extremely large amount of data.

In addition, poorly discretized attributes prevent classification systems from finding important inductive rules. For example, if the ages between 15 and 25 mapped into the same interval, it is impossible to generate the rule about the legal age to start military service. Furthermore, poor discretization makes it difficult to distinguish the non-predictive case from poor discretization. In most cases, inaccurate classification caused by poor discretization is likely to be considered as an error originated from the classification method itself. In other words, if the numeric values are poorly discretized, no matter how good our classification systems are, we fail to find some important rules in databases.

In this paper, we describe a new way of discretizing numeric attributes. We discretize the continuous values using a minimum loss of information criterion. Our discretization method is supervised one since it takes into consideration the class values of examples, and adopts

* Tel.: +82 2 2260 3801; fax: +82 2 2285 3343.
E-mail address: chlee@dgu.ac.kr

0950-7051/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2006.06.005

information theory as a tool to measure the amount of information each interval contains. A number of typical machine learning data sets are selected for discretization, and these are discretized by both other current discretization methods and our proposed method. To compare the correctness of the discretization results, we use the naive Bayesian classifier and C4.5 as the classification algorithms to read and classify data.

The structure of this paper is as follows. Section 2 introduces some current discretization methods. In Section 3, we explain the basic ideas and theoretical background of our approach. Section 4 explains the brief algorithm and correctness of our approach, and experimental results of discretization using some typical machine learning data sets are shown in Section 5. Finally, conclusions are given in Section 6.

2. Related work

Although discretization influences significantly the effectiveness of classification algorithms, not many studies have been done because it usually has been considered a peripheral issue. Among them, we describe a few well-known methods in machine learning literature.

A simple method, called equal distance method, is to partition the range between the minimum and maximum values into N intervals of equal width. Thus, if L and H are the low and high values, respectively, then each interval will have width $W = (H - L)/N$. However, when the outcomes are not evenly distributed, a large amount of information may be lost after discretization using this method. Another method, called equal frequency method, chooses the intervals so that each interval contains approximately the same number of training examples; thus, if $N = 10$, each interval would contain approximately 10% of the examples. These algorithms are very simple, easy to implement, and in some cases produce a reasonable discretization of data. However, there are many cases where they cause serious problems. For instance, suppose we are to discretize attribute age, and reason about the retirement age of a certain occupation. If we use the equal distance method, ages between 50 and 70 may belong to one interval, which prevents us from knowing what the legal retirement age is. Similarly, if we use the equal frequency method to discretize attribute weight, the weights greater than 180 pounds may belong to one interval, which prevents us to reason about the health problem of the persons who are overweight.

With both of these discretizations it would be very difficult or almost impossible to learn certain concepts. The main reason for this is that they ignore the class values of the examples, making it very unlikely that the interval boundaries will just happen to occur in the places which best facilities accurate classification.

Some classification algorithms such as C4.5 [14], CART [3], and PVM [19] take into account the class information when constructing intervals. For example, in C4.5, an entropy measure is used to select the best attribute to branch on at each node of the decision tree. And that measure is used to determine the best cut point for splitting a numeric attribute into two intervals. A threshold value, T, for the continuous numeric attribute A is determined, and the test $A \leqslant T$ is assigned to the left branch while $A > T$ is assigned to the right branch. This cut point is decided by exhaustively checking all possible binary splits of the current interval and choosing the splitting value that maximizes the entropy measure. CART, developed by [3], takes into account the class information as well but it just splits the range into two intervals. It selects the interval boundary which makes the information gain gap between the two intervals maximum. This process is carried out as part of selecting the most discriminating attribute.

Fayyad [8] has extended the method of binary discretization in CART [3] and C4.5 [14], and introduced multi-interval discretization using minimal description length (MDL) technique. In this method, the data are discretized into two intervals and the resulting class information entropy is calculated. A binary discretization is determined by selecting the cut point for which the entropy is minimal amongst all candidates. The binary discretization is applied recursively, always selecting the best cut point. A minimum description length criterion is applied to decide when to stop discretization. This method is implemented in this paper, and used in our experimental study.

Fuzzy discretization, proposed by Kononenko [10], initially forms k equal-width intervals using equal width discretization. Then it estimates $p(a_i < X_i \leqslant b_i | C = c)$ from all training instances rather than from instances that have value of X_i in (a_i, b_i). The influence of a training instances with value v of X_i on (a_i, b_i) is assumed to be normally distributed with the mean value equal to v. The idea behind fuzzy discretization is that small variation of the value of a numeric attribute should have small effects on the attribute's probabilities, whereas under non-fuzzy discretization, a slight difference between two values, one above and one below the cut point can have drastic effects on the estimated probabilities. The number of initial intervals k is a predefined parameter and is set as 7 in our experiments. This method is also implemented and used in our experimental study.

BRACE [18] concentrates on finding the natural boundaries between intervals and creates a set of possible classifications using these boundaries. All classifications in the set are evaluated according to a criterion function and the classification that maximizes the criterion function is selected. It creates a histogram of the data, finds all local minima, and ranks them according to size. The largest is then used to divide the data into a two-interval classification. A three-interval classification is then created using the two largest valleys and so on until a v-interval classification has been created (where v is the number of local minima in the histogram). These classifications are then used to predict the output class of the data, and the classification with the best prediction rate is selected.