entropy

MDPI

Article

# Merging of Numerical Intervals in Entropy-Based Discretization

**Jerzy W. Grzymala-Busse [1,2,*] and Teresa Mroczek [2]**

[1] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA

[2] Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, Rzeszow 35-225, Poland; tmroczek@wsiz.rzeszow.pl

* Correspondence: jerzy@ku.edu; Tel.: +1-785-864-4488; Fax: +1-785-864-3226

Received: 25 September 2018; Accepted: 13 November 2018; Published: 16 November 2018


**Abstract:** As previous research indicates, a multiple-scanning methodology for discretization of numerical datasets, based on entropy, is very competitive. Discretization is a process of converting numerical values of the data records into discrete values associated with numerical intervals defined over the domains of the data records. In multiple-scanning discretization, the last step is the merging of neighboring intervals in discretized datasets as a kind of postprocessing. Our objective is to check how the error rate, measured by tenfold cross validation within the C4.5 system, is affected by such merging. We conducted experiments on 17 numerical datasets, using the same setup of multiple scanning, with three different options for merging: no merging at all, merging based on the smallest entropy, and merging based on the biggest entropy. As a result of the Friedman rank sum test (5% significance level) we concluded that the differences between all three approaches are statistically insignificant. There is no universally best approach. Then, we repeated all experiments 30 times, recording averages and standard deviations. The test of the difference between averages shows that, for a comparison of no merging with merging based on the smallest entropy, there are statistically highly significant differences (with a 1% significance level). In some cases, the smaller error rate is associated with no merging, in some cases the smaller error rate is associated with merging based on the smallest entropy. A comparison of no merging with merging based on the biggest entropy showed similar results. So, our final conclusion was that there are highly significant differences between no merging and merging, depending on the dataset. The best approach should be chosen by trying all three approaches.

**Keywords:** data mining; discretization; numerical attributes; entropy

## 1. Introduction

Discretization of numerical attributes is an important technique used in data mining. Discretization is the process of converting numerical values of data records into discrete values associated with numerical intervals defined over the domains of the data records. As is well known, discretization based on entropy is very successive [1–26]. Additionally, many new techniques have been proposed, e.g., discretization using statistical and logical analysis of data [27], discretization using low-frequency values and attribute interdependency [28], discretization based on rough-set theory [29], a hybrid scheme of frequency and expected number of so-called segments of examples [30], and an oversampling technique combined with randomized filters [31]. Entropy-based discretization was also used for special purposes, e.g., for ranking [32] and for stock-price forecasting [33].

As follows from recent research [13,34,35], one of the discretization methods, called multiple scanning and based on entropy, is especially successful. An important step of such discretization is

merging intervals, conducted as the last step of discretization. As a result, some pairs of intervals are replaced by new, larger intervals. In this paper, we compare two methods of merging numerical intervals, based on the smallest and biggest entropy by skipping merging, i.e., no merging at all. Our results show that such interval merging is crucial for quality of discretization.

The multiple-scanning discretization method, as the name indicates, is based on scanning the entire set of attributes many times. During every scan, for every attribute, the best cutpoint is identified. The quality of a cutpoint is estimated by the conditional entropy of the decision given an attribute. The best cutpoint is associated with the smallest conditional entropy. For a specific scan, when all best cutpoints are selected, a set of subtables is created; each such subtable needs additional discretization. Every subtable is scanned again, and the best cutpoints are computed. There are two ways to end this process: either the stopping condition is satisfied, or the requested number of scans is achieved. If the stopping condition is not satisfied, discretization is completed by another discretization method called Dominant Attribute [34,35].

Dominant-attribute discretization uses a different strategy than multiple scanning, but it is also using many step approach to discretization. In every step, first the best attribute is selected by using the minimum of the conditional entropy of decision given attribute condition. Then, the best cutpoint is identified using the same principle. Discretization is complete when the stopping condition is satisfied.

The multiple-scanning methodology is better than two well-known discretization methods: Equal Interval Width and Equal Frequency per Interval enhanced to globalized methods [34]. In Reference [34], rule induction was used for data mining. Additionally, four other discretization methods, namely, the original C4.5 approach to discretization, and the same globalized versions of Equal Interval Width and Equal Frequency per Interval methods, and Multiple Scanning were compared in Reference [35]; this time, data mining was based on the C4.5 generation of decision trees. Again, it was shown that the best discretization method is Multiple Scanning.

## 2. Discretization

Let $a$ be a numerical attribute, $a_i$ be the smallest value of $a$, and $a_j$ be the largest value of $a$. Discretization of $a$ is based on finding the numbers $a_{i_0}$, $a_{i_1}$, ..., $a_{i_k}$, called cutpoints, where $a_{i_0} = a_i$, $a_{i_k} = a_j$, $a_{i_l} < a_{i_{l+1}}$ for $l = 0, 1, \ldots, k-1$, and $k$ is a positive integer. Thus, domain $[a_i, a_j]$ of $a$ is partitioned into $k$ intervals

$$\{[a_{i_0}, a_{i_1}), [a_{i_1}, a_{i_2}), \ldots, [a_{i_{k-2}}, a_{i_{k-1}}), [a_{i_{k-1}}, a_{i_k}]\}.$$

In the remainder of this paper, such intervals are denoted as follows:

$$a_{i_0} \ldots a_{i_1}, a_{i_1} \ldots a_{i_2}, \ldots, a_{i_{k-2}} \ldots a_{i_{k-1}}, a_{i_{k-1}} \ldots a_{i_k}.$$

In practical applications, discretization is conducted on many numerical attributes. Table 1 presents an example of a dataset with four numerical attributes: Length, Height, Width, and Weight, and eight cases. An additional symbolic variable, Quality, is the decision. Attributes are independent variables, while the decision is a dependent variable. The set of all cases is denoted by $U$. In Table 1, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

Let $v$ be a variable and let $v_1, v_2, \ldots, v_n$ be values of $v$, where $n$ is a positive integer. Let $S$ be a subset of $U$. Let $p(v_i)$ be a probability of $v_i$ in $S$, where $i = 1, 2, \ldots, n$. An *entropy* $H_S(v)$ is defined as follows:

$$H_S(v) = -\sum_{i=1}^{n} p(v_i) \cdot \log p(v_i).$$

In this paper, we assume that all logarithms are binary.

**Table 1.** An example of a dataset with numerical attributes.

| Case | Attributes | | | | Decision |
|---|---|---|---|---|---|
| | Length | Height | Width | Weight | Quality |
| 1 | 4.7 | 1.8 | 1.7 | 1.7 | high |
| 2 | 4.5 | 1.4 | 1.8 | 0.9 | high |
| 3 | 4.7 | 1.8 | 1.9 | 1.3 | high |
| 4 | 4.5 | 1.8 | 1.7 | 1.3 | medium |
| 5 | 4.3 | 1.6 | 1.9 | 1.7 | medium |
| 6 | 4.3 | 1.4 | 1.7 | 0.9 | low |
| 7 | 4.5 | 1.6 | 1.9 | 0.9 | very-low |
| 8 | 4.5 | 1.4 | 1.8 | 1.3 | very-low |

Let $a$ be an attribute, let $a_1, a_2, \ldots, a_m$ be all values of $a$ restricted to $S$, let $d$ be a decision and let $d_1, d_2, \ldots, d_n$ be all values of $d$ restricted to $S$. Conditional entropy $H_S(d|a)$ of the decision $d$ given attribute $a$ is defined as follows:

$$-\sum_{j=1}^{m} p(a_j) \cdot \sum_{i=1}^{n} p(d_i|a_j) \cdot log\ p(d_i|a_j),$$

where $p(d_i|a_j)$ is the conditional probability of the value $d_j$ of the decision $d$ given $a_j$; $j \in \{1, 2, \ldots, m\}$ and $i \in \{1, 2, \ldots, n\}$.

As is well known [1,4,5,7,9,10,12,13,16,21,23,24], discretization that uses conditional entropy of the decision-given attribute is believed to be one of the most successful discretization techniques.

Let $S$ be a subset of $U$, $a$ be an attribute, and $q$ be a cutpoint splitting the set $S$ into two subsets, $S_1$ and $S_2$. The corresponding conditional entropy, denoted by $H_S(d|a)$ is defined as follows:

$$\frac{|S_1|}{|U|} H_{S_1}(a) + \frac{|S_2|}{|U|} H_{S_2}(a),$$

where $|X|$ denotes the cardinality of set $X$. Usually, cutpoint $q$ for which $H_S(d|a)$ is the smallest is considered to be the best cutpoint.

We need a condition to stop discretization. Roughly speaking, the most obvious idea is to stop discretization when we may distinguish the same cases in the discretized dataset that were distinguishable in the original dataset with numerical attributes. The idea of distinguishability (indiscernibility) of cases is one of the basic ideas of rough-set theory [36,37]. Let $B$ be a subset of set $A$ of all attributes, and $x, y \in U$. Indiscernibility relation $IND(B)$ is defined as follows:

$$(x, y) \in IND(B) \text{ if and only if } a(x) = a(y) \text{ for any } a \in B,$$

where $a(x)$ denotes the value of the attribute $a \in A$ for the case $x \in U$. Obviously, $IND(B)$ is an equivalence relation. For $x \in U$, the equivalence class of $IND(B)$ is denoted by $[x]_B$, and is called a B-elementary set.

A family of all sets $[x]_B$, where $x \in U$, is a partition on $U$, denoted by $B^*$. Additionally, for a decision $d$, a $\{d\}^*$-elementary set is called a *concept*. For Table 1, and for $B$ = {*Length*}, $B^*$ ={{1, 3}, {2, 4, 7, 8}, {5, 6}} and $\{d\}^*$= {{1, 2, 3}, {4, 5}, {6}, {7, 8}}. None of the concepts {1, 2, 3}, {4, 5}, {6}, {7, 8} is $B$-definable. It is a usual practice in rough-set theory to use for any $X \in \{d\}^*$ two sets, called lower and upper approximations of $X$. The lower approximation of $X$ is defined as follows:

$$\{x \mid x \in U, [x]_B \subseteq X\}$$

and is denoted by $\underline{B}X$. The upper approximation of $X$ is defined as follows:

$$\{x \mid x \in U, [x]_B \cap X \neq \emptyset\}$$

and is denoted by $\overline{B}X$. For Table 1, $\underline{B}\{1, 2, 3\} = \{1, 3\}$ and $\overline{B}\{1, 2, 3\} = \{1, 2, 3, 4, 7, 8\}$.

Usually, discretization is stopped when so-called level of consistency [4], defined as follows:

$$L(A) = \frac{\sum_{X \in \{d\}^*} |\underline{A}X|}{|U|}$$

and denoted by $L(A)$, is equal to 1. For Table 1, $A^* = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}\}$, so $\underline{A}X = X$ for any concept $X$ from $\{d\}^*$. On the other hand, for $B = \{Length\}$,

$$L(B) = \frac{|\underline{B}\{1,2,3\}| + |\underline{B}\{4,5\}| + |\underline{B}\{6\}| + |\underline{B}\{7,8\}|}{|U|} = \frac{|\{1,3\}| + |\emptyset| + |\emptyset| + |\emptyset| + |\emptyset|}{8} = 0.25.$$

### *2.1. Multiple Scanning*

Special parameter $t$, selected by the user and called the total number of scans, is used in multiple-scanning discretization. During the first scan, for any attribute $a$ from the set $A$, the best cutpoint is selected using the criterion of smallest entropy $H_U(d|q)$ for all potential cutpoints splitting $U$, where $d$ is the decision. Such cutpoints are created as the averages of two consecutive values of sorted attribute $a$. Once the best cutpoint is found, a new binary attribute $a^d$ is created, with two intervals as vales of $a^d$, the first interval is defined as containing all original numerical values of $a$ smaller than the selected cutpoint $q$, and the second interval contains the remaining original values of $a$. Partition $\{A^d\}^*$ is created, where $A^d$ is the set of all partially discretized attributes. For the next scans, starting from $t = 2$, set $A$ is scanned again: for each block $X$ of $\{A^d\}^*$, for each attribute $a$, and for each remaining cutpoint of $a$, the best cutpoint is computed, and the best cutpoint among all blocks $X$ of $\{A^d\}^*$ is selected as the next cutpoint of $a$. If parameter $t$ is reached and $L(A^d) \neq 1$, another discretization method, Dominant Attribute, is used. In the dominant-attribute strategy, the best attribute is first selected among partially discretized attributes, using the criterion of smallest conditional entropy $H(d|a^d)$, where $a^d$ is a partially discretized attribute. For the best attribute, best cutpoint $q$ is selected, using the criterion of smallest entropy $H_S(d|a^d)$, where $q$ splits $S$ into $S_1$ and $S_2$. For both $S_1$ and $S_2$, we select the best attribute and then the best cutpoint, until $L(A^d) = 1$, where $A^d$ is the set of discretized attributes.

We illustrate the multiple-scanning discretization method using the dataset from Table 1. Since our dataset was small, we used just one scan. Initially, for any attribute $a \in A$, all conditional entropies $H_a(q, U)$ should be computed for all possible cutpoints $q$ of $a$. The set of all possible cutpoints for Length is {4.4, 4.6}. Similarly, the sets of all possible cutpoints for Height, Width, and Weight were {1.5, 1.7}, {1.75, 1.85} and {1.1, 1.5}, respectively. Furthermore,

$$H_{Length}(4.4, U) = \frac{2}{8}(-\frac{1}{2} \cdot log\frac{1}{2})2 + \frac{6}{8}(-\frac{3}{6} \cdot log\frac{3}{6} - \frac{2}{6} \cdot log\frac{2}{6} - \frac{1}{6} \cdot log\frac{1}{6}) = 1.344,$$

$$H_{Length}(4.6, U) = \frac{6}{8}((-\frac{2}{6} \cdot log\frac{2}{6})2 + (-\frac{1}{6} \cdot log\frac{1}{6})2) + \frac{2}{8}(0) = 1.439.$$

The best cutpoint is 4.4. In a similar way, we selected the best cutpoints for the remaining attributes, Height, Width, and Weight. These cutpoints are 1.5, 1.75, and 1.1, respectively. Thus, the partially discretized dataset, after the first scan, is presented in Table 2.

The dataset from Table 2 needs an additional discretization since $(A^d)^* = \{\{1, 4\}, \{2\}, \{3\}, \{5\}, \{6\}, \{7\}, \{8\}\}$, $\{d\}^* = \{\{1, 2, 3\}, \{4, 5\}, \{6\}, \{7, 8\}\}$ and

$$L(\{Length^d,\ Height^d,\ Width^d,\ Weigth^d\}) = \frac{2+1+1+2}{8} = 0.75 < 1.$$

As follows from Table 2, Cases 1 and 4 need to be distinguished. A dataset from Table 1, restricted to Cases 1 and 4, is presented in Table 3.

**Table 2.** Partially discretized dataset after the first scan. ($^d$ is the decision)

| Case | Attributes | | | | Decision |
|---|---|---|---|---|---|
| | Length $^d$ | Height $^d$ | Width $^d$ | Weight $^d$ | Quality |
| 1 | 4.4..4.7 | 1.5..1.8 | 1.7..1.75 | 1.1..1.7 | high |
| 2 | 4.4..4.7 | 1.4..1.5 | 1.75..1.9 | 0.9..1.1 | high |
| 3 | 4.4..4.7 | 1.5..1.8 | 1.75..1.9 | 1.1..1.7 | high |
| 4 | 4.4..4.7 | 1.5..1.8 | 1.7..1.75 | 1.1..1.7 | medium |
| 5 | 4.3..4.4 | 1.5..1.8 | 1.75..1.9 | 1.1..1.7 | medium |
| 6 | 4.3..4.4 | 1.4..1.5 | 1.7..1.75 | 0.9..1.1 | low |
| 7 | 4.4..4.7 | 1.5..1.8 | 1.75..1.9 | 0.9..1.1 | very-low |
| 8 | 4.4..4.7 | 1.4..1.5 | 1.75..1.9 | 1.1..1.7 | very-low |

**Table 3.** A subset of the dataset presented in Table 1.

| Case | Attributes | | | | Decision |
|---|---|---|---|---|---|
| | Length | Height | Width | Weight | Quality |
| 1 | 4.7 | 1.8 | 1.7 | 1.7 | high |
| 4 | 4.5 | 1.8 | 1.7 | 1.3 | medium |

Cases 1 and 4 from Table 3 may be distinguished by any of the two following attributes: Length and Weight. Both attributes are of the same quality, as a result of a heuristic step we selected Length. A new cutpoint for Length was equal to 4.6. Thus, attribute Length has two cutpoints, 4.4 and 4.6. Table 4 presents the next partially discretized dataset.

**Table 4.** Discretized dataset.

| Case | Attributes | | | | Decision |
|---|---|---|---|---|---|
| | Length $^d$ | Height $^d$ | Width $^d$ | Weight $^d$ | Quality |
| 1 | 4.6..4.7 | 1.5..1.8 | 1.7..1.75 | 1.1..1.7 | high |
| 2 | 4.4..4.6 | 1.4..1.5 | 1.75..1.9 | 0.9..1.1 | high |
| 3 | 4.6..4.7 | 1.5..1.8 | 1.75..1.9 | 1.1..1.7 | high |
| 4 | 4.4..4.6 | 1.5..1.8 | 1.7..1.75 | 1.1..1.7 | medium |
| 5 | 4.3..4.4 | 1.5..1.8 | 1.75..1.9 | 1.1..1.7 | medium |
| 6 | 4.3..4.4 | 1.4..1.5 | 1.7..1.75 | 0.9..1.1 | low |
| 7 | 4.4..4.6 | 1.5..1.8 | 1.75..1.9 | 0.9..1.1 | very-low |
| 8 | 4.4..4.6 | 1.4..1.5 | 1.75..1.9 | 1.1..1.7 | very-low |

For the dataset from Table 4, $(A^d)^* = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\}\}$ and $L(A) = 1$.

### 2.2. Interval Merging

In general, it is possible to simplify the result of discretization by interval merging. The idea is to replace two neighboring intervals, *i*...*j* and *j*...*k*, of the same attribute by one interval, *i*...*k*. It can be conducted using two different techniques: safe merging and proper merging. In safe merging, for a given attribute, any two neighboring intervals *i*...*j* and *j*...*k* are replaced by interval *i*...*k*, if for both intervals the decision value is the same.

In proper merging, two neighboring intervals *i*...*j* and *j*...*k* of the same attribute are replaced by interval *i*...*k*, if the levels of consistency before merging and after merging are the same. A question is how to guide the search for such two neighboring intervals. In experiments described in this paper, two search criteria were implemented based on the smallest and the largest conditional entropy $H_S(d|a)$. Another possibility, also taken into account, is ignoring any merging at all.

It is clear that, for Table 4, for the Length attribute, we may eliminate Cutpoint 4.4. As a result, a new data set, presented in Table 5 is created. For the dataset from Table 4, $(A^d)^* = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\}\}$ and $L(A) = 1$.

**Table 5.** Discretized dataset after interval merging.

| Case | Attributes | | | | Decision |
|---|---|---|---|---|---|
| | Length $^d$ | Height $^d$ | Width $^d$ | Weight $^d$ | Quality |
| 1 | 4.6..4.7 | 1.5..1.8 | 1.7..1.75 | 1.1..1.7 | high |
| 2 | 4.3..4.6 | 1.4..1.5 | 1.75..1.9 | 0.9..1.1 | high |
| 3 | 4.6..4.7 | 1.5..1.8 | 1.75..1.9 | 1.1..1.7 | high |
| 4 | 4.3..4.6 | 1.5..1.8 | 1.7..1.75 | 1.1..1.7 | medium |
| 5 | 4.3..4.4 | 1.5..1.8 | 1.75..1.9 | 1.1..1.7 | medium |
| 6 | 4.3..4.6 | 1.4..1.5 | 1.7..1.75 | 0.9..1.1 | low |
| 7 | 4.3..4.6 | 1.5..1.8 | 1.75..1.9 | 0.9..1.1 | very-low |
| 8 | 4.3..4.6 | 1.4..1.5 | 1.75..1.9 | 1.1..1.7 | very-low |

## 3. Experiments

Experiments described in this paper were conducted on 17 datasets with numerical attributes. These datasets presented in Table 6 and are accessible in the Machine-Learning Repository, University of California, Irvine, except for bankruptcy. The bankruptcy dataset was given in Reference [38].

**Table 6.** Datasets.

| Dataset | Cases | Number of Attributes | Concepts |
|---|---|---|---|
| Abalone | 4177 | 8 | 29 |
| Australian | 690 | 14 | 2 |
| Bankruptcy | 66 | 5 | 2 |
| Bupa | 345 | 6 | 2 |
| Connectionist Bench | 208 | 60 | 2 |
| Echocardiogram | 74 | 7 | 2 |
| Ecoli | 336 | 8 | 8 |
| Glass | 214 | 9 | 6 |
| Image Segmentation | 210 | 19 | 7 |
| Ionoshere | 351 | 34 | 2 |
| Iris | 150 | 4 | 3 |
| Leukemia | 415 | 175 | 2 |
| Pima | 768 | 8 | 2 |
| Spectrometry | 25,931 | 15 | 2 |
| Wave | 512 | 21 | 3 |
| Wine Recognition | 178 | 13 | 3 |
| Yeast | 1484 | 8 | 9 |

For discretization, we applied the multiple-scanning method. The level of consistency was set to 1. We used three approaches to merging intervals in the last stage of discretization:

- no merging at all,
- proper merging based on the minimum of conditional entropy, and
- proper merging based on the maximum of conditional entropy.

The discretized datasets were processed by the C4.5 decision-tree generating system [39]. Note that the C4.5 system builds a decision tree using conditional entropy as well. The main mechanism of selecting the most important attribute $a$ in C4.5 is based on the maximum of mutual information, which in C4.5 is called an information gain. The mutual information is the difference between marginal entropy $H_S(d)$ and conditional entropy $H_S(d|a)$, where $d$ is the decision. Since $H_S(d)$ is fixed, the maximum of mutual information is equivalent to the minimum of conditional entropy $H_S(d|a)$. In our experiments, an error rate was computed using internal tenfold cross validation of C4.5.

Our methodology is illustrated by Figures 1–8, all restricted to the yeast dataset, one of 17 datasets used for experiments. Figure 1 presents an error rate for three consecutive scans conducted on the yeast dataset. Figure 2 shows the number of discretization intervals for three scans on the same dataset.

Figure 3 shows domains of all attributes for the yeast dataset, and Figures 4–8 show intervals of all attributes during interval scanning and merging.

Table 7 shows error rates for the three approaches to merging. Note that, for any dataset, we included only the smallest error rate with a corresponding scan number. The error rates were compared using the Friedman rank sum test combined with multiple comparison, with 5% level of significance. As follows from the Friedman test, the differences between all three approaches are statistically insignificant.

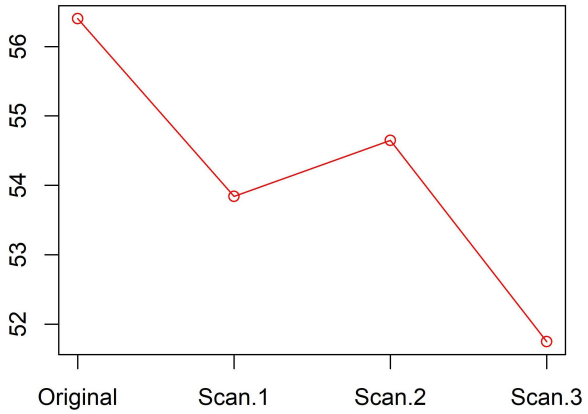


**Figure 1.** Error rate for consecutive scans for the yeast dataset.

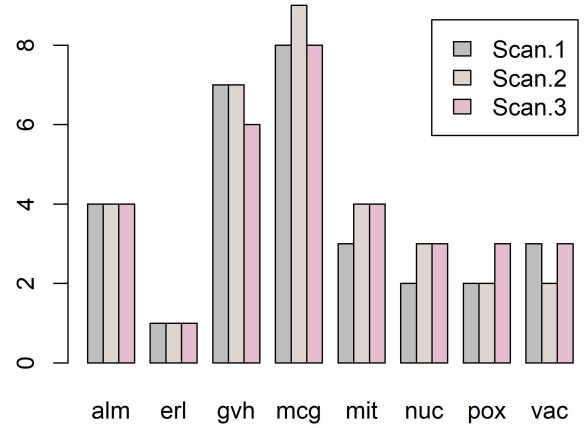


**Figure 2.** Number of discretization intervals for consecutive scans for the yeast dataset.

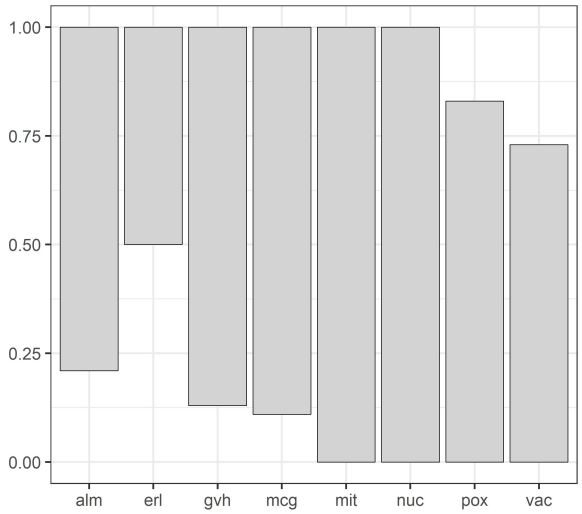


**Figure 3.** Domains of all attributes for the yeast dataset.

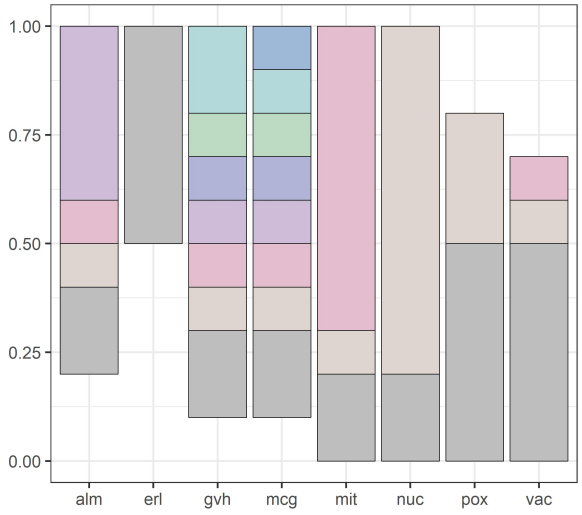


**Figure 4.** Intervals for all attributes after the first scan for the yeast dataset.

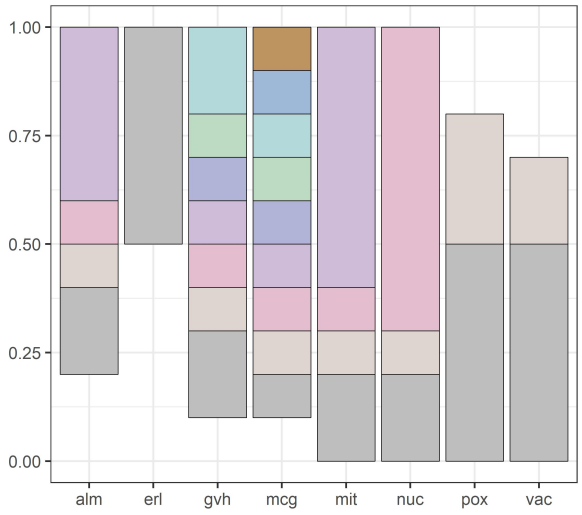


**Figure 5.** Intervals for all attributes after the second scan for the yeast dataset.

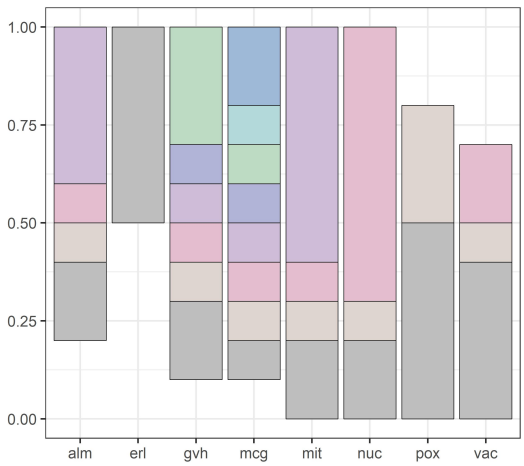


**Figure 6.** Intervals for all attributes after the third scan for the yeast dataset.

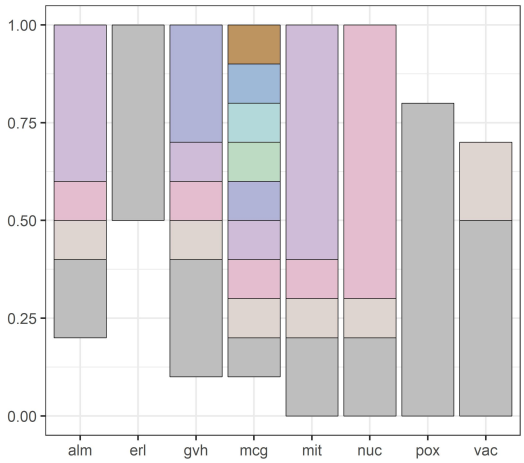


**Figure 7.** Intervals for all attributes after merging based on minimal entropy for the second scan for the yeast dataset.

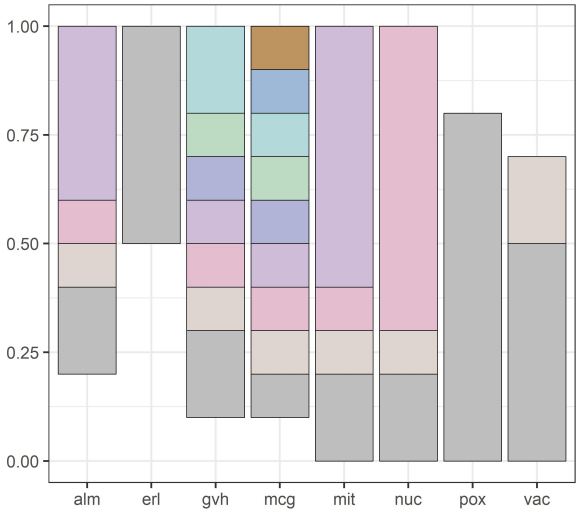


**Figure 8.** Intervals for all attributes after merging based on maximal entropy for the second scan for the yeast dataset.

**Table 7.** Error rates for three approaches to merging.

| Dataset | No Merging | Scan Number | MIN Entropy | Scan Number | MAX Entropy | Scan Number |
|---|---|---|---|---|---|---|
| Abalone | 75.58 | 5 | - | - | - | - |
| Australian | 13.48 | 1 | 12.61 | 3 | 13.04 | 1 |
| Bankruptcy | 3.03 | 1 | - | - | - | - |
| Bupa | 29.28 | 3 | 30.43 | 2 | 30.43 | 2 |
| Connectionist Bench | 16.83 | 1 | 24.04 | 1 | 24.04 | 1 |
| Echocardiogram | 14.86 | 1 | 14.86 | 2 | 14.86 | 1 |
| Ecoli | 22.02 | 0 | 17.86 | 0 | 20.54 | 2 |
| Glass | 24.77 | 3 | 23.36 | 2 | 23.36 | 2 |
| Image Segmentation | 11.90 | 2 | - | - | 13.81 | 0 |
| Ionoshere | 5.98 | 2 | 5.98 | 1 | 5.98 | 4 |
| Iris | 4.67 | 2 | - | - | - | - |
| Leukemia | 21.20 | 2 | 26.27 | 2 | 21.20 | 2 |
| Pima | 24.09 | 2 | 24.48 | 0 | 24.61 | 0 |
| Spectrometry | 1.13 | 2 | 1.15 | 5 | 1.19 | 3 |
| Wave | 23.04 | 1 | 24.02 | 1 | 23.44 | 3 |
| Wine Recognition | 3.93 | 1 | 3.37 | 1 | 3.37 | 1 |
| Yeast | 51.75 | 3 | 49.12 | 5 | 49.93 | 2 |

Thus, there is no universally best approach among no merging, merging based on minimum of conditional entropy, and merging based on maximum of conditional entropy.

Our next objective was to test the difference between all three approaches for a specific dataset. We conducted extensive experiments, with the repetition of 30 tenfold cross validations for every dataset and recorded averages and standard deviations in order to use the standard test for difference between averages. The corresponding $Z$ scores are presented in Table 8. It is quite obvious that the choice of the correct approach to merging is highly significant in most cases, with the level of significance at 0.01, since the absolute value of the corresponding $Z$-score is larger than 2.58. For example, for the ecoli dataset, merging of intervals based on minimum of conditional entropy is better than no merging, while for the leukemia dataset, it is the other way around. Similarly, for the ecoli dataset, no merging is better than merging based on the maximum of conditional entropy, while for the pima dataset it is the opposite.

**Table 8.** $Z$ scores for the test of differences between averages of error rates associated with three approaches to merging.

| Dataset | No Merging − Merging with MIN Entropy | Scan Number | No Merging − Merging with MAX Entropy | Scan Number |
|---|---|---|---|---|
| Abalone | - | - | - | - |
| Australian | 6.90 | 3 | 5.20 | 3 |
| Bankruptcy | - | - | - | - |
| Bupa | 22.90 | 0 | 12.75 | 0 |
| Connectionist Bench | −9.09 | 1 | −8.73 | 1 |
| Echocardiogram | −6.71 | 0 | −7.84 | 0 |
| Ecoli | 31.75 | 0 | −140.05 | 0 |
| Glass | −7.28 | 1 | 11.92 | 0 |
| Image Segmentation | −0.33 | 0 | −14.71 | 2 |
| Ionoshere | −41.36 | 2 | −8.85 | 3 |
| Iris | - | - | - | - |
| Leukemia | −51.18 | 0 | −45.40 | 0 |
| Pima | 16.34 | 0 | 27.16 | 2 |
| Spectrometry | 20.94 | 5 | −14.55 | 3 |
| Wave | 6.92 | 2 | 10.20 | 3 |
| Wine Recognition | −0.73 | 0 | 8.94 | 1 |
| Yeast | 25.68 | 2 | 23.14 | 2 |

Our future research plans include a comparison of our main methodology, multiple-scanning discretization, with discretization based on binning using histograms and chi-square analysis.

## 4. Conclusions

The main contribution of our paper is showing that postprocessing discretization based on merging intervals is extremely important for the discretization quality. Results of our experiments indicate that there is no universally best approach to merging intervals. However, there are statistically highly significant differences (with 1% significance level) between these three approaches, depending on the dataset. Therefore, it is very important to use the best choice among the three approaches during multiple-scanning discretization of datasets with numerical attributes.

**Author Contributions:** T.M. designed and conducted experiments, J.W.G.-B. validated results and wrote the paper.

**Funding:** This research received no external funding.

**Acknowledgments:** The authors would like to thank the editor and referees for their helpful suggestions and comments on the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Blajdo, P.; Grzymala-Busse, J.W.; Hippe, Z.S.; Knap, M.; Mroczek, T.; Piatek, L. A comparison of six approaches to discretization—A rough set perspective. In Proceedings of the Rough Sets and Knowledge Technology Conference,Chengdu, China, 17–19 May 2008; pp. 31–38.
2. Chan, C.C.; Batur, C.; Srinivasan, A. Determination of quantization intervals in rule based model for dynamic. In Proceedings of the IEEE Conference on Systems, Man, and Cybernetics, Charlottesville, VA, USA, 13–16 October 1991; pp. 1719–1723.
3. Chen, M.Y.; Chen, B.T. Online fuzzy time series analysis based on entropy discretization and a fast Fourier transform. *Appl. Soft Comput.* **2014**, *14*, 156–166. [CrossRef]
4. Chmielewski, M.R.; Grzymala-Busse, J.W. Global discretization of continuous attributes as preprocessing for machine learning. *Int. J. Approx. Reason.* **1996**, *15*, 319–331. [CrossRef]
5. Clarke, E.J.; Barton, B.A. Entropy and MDL discretization of continuous variables for Bayesian belief networks. *Int. J. Intell. Syst.* **2000**, *15*, 61–92. [CrossRef]
6. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and unsupervised discretization of continuous features. In Proceedings of the 12-th International Conference on Machine Learning, Tahoe, CA, USA, 9–12 July 1995; pp. 194–202.
7. Elomaa, T.; Rousu, J. General and efficient multisplitting of numerical attributes. *Mach. Learn.* **1999**, *36*, 201–244. [CrossRef]
8. Elomaa, T.; Rousu, J. Efficient multisplitting revisited: Optima-preserving elimination of partition candidates. *Data Min. Knowl. Discov.* **2004**, *8*, 97–126. [CrossRef]
9. Fayyad, U.M.; Irani, K.B. On the handling of continuous-valued attributes in decision tree generation. *Mach. Learn.* **1992**, *8*, 87–102. [CrossRef]
10. Fayyad, U.M.; Irani, K.B. Multiinterval discretization of continuous-valued attributes for classification learning. In Proceedings of the Thirteenth International Conference on Artificial Intelligence, Chambéry, France, 28 August–3 September 1993; pp. 1022–1027.
11. Garcia, S.; Luengo, J.; Saez, J.A.; Lopez, V.; Herrera, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 734–750. [CrossRef]
12. Grzymala-Busse, J.W. Discretization of numerical attributes. In *Handbook of Data Mining and Knowledge Discovery*; Kloesgen, W., Zytkow, J., Eds.; Oxford University Press: New York, NY, USA, 2002; pp. 218–225.
13. Grzymala-Busse, J.W. A multiple scanning strategy for entropy based discretization. In Proceedings of the 18th International Symposium on Methodologies for Intelligent Systems, Prague, Czech Republic, 14–17 September 2009; pp. 25–34.
14. Grzymala-Busse, J.W. Mining numerical data—A rough set approach. *Trans. Rough Sets* **2010**, *11*, 1–13.
15. Grzymala-Busse, J.W.; Stefanowski, J. Three discretization methods for rule induction. *Int. J. Intell. Syst.* **2001**, *16*, 29–38. [CrossRef]
16. Kohavi, R.; Sahami, M. Error-based and entropy-based discretization of continuous features. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 114–119.
17. Kerber, R. ChiMerge: Discretization of numeric attributes. In Proceedings of the 10-th National Conference on AI, Menlo Park, CA, USA, 12–16 July 1992; pp. 123–128.
18. Kotsiantis, S.; Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *32*, 47–58.
19. Kurgan, L.A.; Cios, K.J. CAIM discretization algorithm. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 145–153. [CrossRef]
20. Liu, H.; Hussain, F.; Tan, C.L.; Dash, M. Discretization: An enabling technique. *Data Min. Knowl. Discov.* **2002**, *6*, 393–423. [CrossRef]
21. Nguyen, H.S.; Nguyen, S.H. Discretization methods in data mining. In *Rough Sets in Knowledge Discovery 1: Methodology and Applications*; Polkowski, L., Skowron, A., Eds.; Physica-Verlag: Heidelberg, Germany, 1998; pp. 451–482.
22. Sang, Y.; Qi, H.; Li, K.; Jin, Y.; Yan, D.; Gao, S. An effective discretization method for disposing high-dimensional data. *Inf. Sci.* **2014**, *270*, 73–91. [CrossRef]

23. Stefanowski, J. Handling continuous attributes in discovery of strong decision rules. In Proceedings of the First Conference on Rough Sets and Current Trends in Computing, Warsaw, Poland, 22–26 June 1998; pp. 394–401.
24. Stefanowski, J. *Algorithms of Decision Rule Induction in Data Mining*; Poznan University of Technology Press: Poznan, Poland, 2001.
25. Wong, A.K.C.; Chiu, D.K.Y. Synthesizing statistical knowledge from incomplete mixed-mode data. *IEEE Trans. Pattern Anal. Mach. Intell.* **1987**, *9*, 796–805. [CrossRef] [PubMed]
26. Yang, Y.; Webb, G. Discretization for naive-Bayes learning: managing discretization bias and variance. *Mach. Learn.* **2009**, *74*, 39–74. [CrossRef]
27. Bruni, R.; Bianchi, G. Effective classification using a small training set based on discretization and statistical analysis. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 2349–2361. [CrossRef]
28. Rahman, M.D.; Islam, M.Z. Discretization of continuous attributes through low frequency numerical values and attribute interdependency. *Expert Syst. Appl.* **2016**, *45*, 410–423. [CrossRef]
29. Jiang, F.; Sui, Y. A novel approach for discretization of continuous attributes in rough set theory. *Knowl.-Based Syst.* **2015**, *73*, 324–334. [CrossRef]
30. Wang, R.; Kwong, S.; Wang, X.Z.; Jiang, Q. Segment based decision tree induction with continuous valued attributes. *IEEE Trans. Cybern.* **2015**, *45*, 1262–1275. [CrossRef] [PubMed]
31. Dimic, G.; Rancic, D.; Spalevic, P. Improvement of the accuracy of prediction using unsupervised discretization method: Educational data set case study. *Tech. Gaz.* **2018**, *25*, 407–414.
32. De Sa, C.R.; Soares, C.; Knobbe, A. Entropy-based discretization methods for ranking data. *Inf. Sci.* **2016**, *329*, 921–936. [CrossRef]
33. Chen, M.Y.; Chen, B.T. A hybrid fuzzy time series model based on granular computing for stock price forecasting. *Inf. Sci.* **2015**, *294*, 227–241. [CrossRef]
34. Grzymala-Busse, J.W. Discretization based on entropy and multiple scanning. *Entropy* **2013**, *15*, 1486–1502. [CrossRef]
35. Grzymala-Busse, J.W.; Mroczek, T. A comparison of four approaches to discretization based on entropy. *Entropy* **2016**, *18*, 69. [CrossRef]
36. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
37. Pawlak, Z. *Rough Sets. Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publishers: Dordrecht, The Netherlands; Boston, MA, USA; London, UK, 1991.
38. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
39. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.