entropy

MDPI

Article

A Comparison of Four Approaches to Discretization Based on Entropy †

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

† This paper is an extended version of our paper published in Pattern Recognition and Machine Intelligence (PReMI). In proceedings of the 6th International Conference on PReMI, Warsaw, Poland, 30 June–3 July 2015.

Academic Editor: Andreas Holzinger
Received: 25 November 2015; Accepted: 14 February 2016; Published: 25 February 2016

Abstract: We compare four discretization methods, all based on entropy: the original C4.5 approach to discretization, two globalized methods, known as equal interval width and equal frequency per interval, and a relatively new method for discretization called multiple scanning using the C4.5 decision tree generation system. The main objective of our research is to compare the quality of these four methods using two criteria: an error rate evaluated by ten-fold cross-validation and the size of the decision tree generated by C4.5. Our results show that multiple scanning is the best discretization method in terms of the error rate and that decision trees generated from datasets discretized by multiple scanning are simpler than decision trees generated directly by C4.5 or generated from datasets discretized by both globalized discretization methods.

Keywords: data mining; discretization; numerical attributes; entropy

1. Introduction

Mining data with numerical attributes requires discretization. Among many discretization techniques, discretization based on entropy is one of the most successful methods [1–31]. Entropy was used for discretization applied to ranking data [7]. A special kind of discretization for data with many attributes was presented in [25]. Discretization combined with semi-supervised learning was presented in [3]. Many papers emphasizing the importance of discretization to data mining were recently published [13,17,18,28,29,31].

In this paper, we present the results of our experiments conducted on 17 numerical datasets using the C4.5 decision tree generation system, combined with four discretization methods: the original C4.5 approach to discretization, two globalized methods, known as equal interval width and equal frequency per interval, and a relatively new method for discretization called *multiple scanning*. The original approach to discretization included in the C4.5 system, as well as discretization based on equal interval width and equal frequency per interval are well known. Multiple scanning, introduced in [15,32], was very successful when combined with rule induction and a classification system of LERS (learning from examples based on rough sets) [33].

In multiple scanning, during every scan, the entire attribute set is analyzed. For all attributes, the best cut point is selected. At the end of a scan, some sub-tables that still need discretization are created. The entire attribute set of any sub-table is scanned again, and the best corresponding cut points are selected. The process continues until the stopping condition is satisfied or the required number of scans is reached. If necessary, discretization is completed by another discretization technique, called *dominant attribute*[15,32]. In the dominant attribute method, initially we select the best attribute. For this attribute, the best cut point is selected using conditional entropy. This process continues until the same stopping criterion is satisfied. The stopping criterion used in this paper is based on rough set theory.

The main objective of our research is to compare the quality of these four discretization methods using two criteria: an error rate evaluated by ten-fold cross-validation and the size of the decision tree generated by C4.5. Experimental results presented in [32] show that multiple scanning is the best discretization method among these four discretization methods. In [32], four discretization techniques were compared using a rule-based methodology. Experiments were conducted using the MLEM2 (modified learning from examples module, Version 2) rule induction algorithm [34] and the LERS classification system. There is a possibility that the results of [32] depend on the choice of experimental setup. Therefore, to remove this bias, we changed the original setup and conducted new experiments using the standard C4.5 decision tree generation methodology. Our new results fully support the results of [32]. For 17 numerical datasets, four sets of experiments were conducted: first, the C4.5 system was used to compute an error rate using ten-fold cross-validation; then, the same datasets were discretized using two globalized methods (equal interval width and equal frequency per interval) and multiple scanning, and for such discretized datasets, the same C4.5 system was used to establish an error rate.

The same methodology, based on computing the C4.5 error rate, was used in [23] to compare nine successful and well-known discretization methods using 11 datasets. Seven of these 11 datasets (*australian*, *bupa*, *glass*, *ionosphere*, *iris*, *pima* and *wine recognition*) were also used in our experiments. For any of these seven datasets, the best result accomplished using our methods is better than the corresponding best result cited in [23]. Thus, our choice for the four discretization methods is well justified: we used very efficient methods. Our results show that the multiple scanning discretization technique is significantly better than the internal discretization used in C4.5 and two globalized discretization methods: equal interval width and equal frequency per interval in terms of the error rate computed by ten-fold cross-validation (two-tailed test, 5% level of significance). Additionally, decision trees generated from data discretized by multiple scanning are significantly simpler than decision trees generated directly by C4.5 and decision trees generated from datasets discretized and both globalized discretization methods.

The main idea of the multiple scanning method, giving the same chance to all attributes, is highly successful. In each consecutive step of this method, every attribute is taken into account. In other discretization methods, some attributes may be eliminated to begin with.

2. Discretization

Let a be a numerical attribute with domain $[a_i, a_j]$. A partition of the domain $[a_i, a_j]$ into k intervals:

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

where $a_{i_0} = a_i$, $a_{i_k} = a_j$, and $a_{i_l} < a_{i_{l+1}}$ for $l = 0, 1, ..., k-1$, determines a discretization of a. The numbers a_{i_1}, a_{i_2}, ..., $a_{i_{k-1}}$ are called *cut points*. In this paper, corresponding intervals are denoted as follows:

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

An example of a dataset with numerical attributes is presented in Table 1. In this table, all cases are described by variables called *attributes* and one variable called a *decision*. The set of all attributes is denoted by A. The decision is denoted by d. The set of all cases is denoted by U. In Table 1, the attributes are *length*, *width* and *height*, while the decision is *quality*. Additionally, U = {1, 2, 3, 4, 5, 6, 7}.

For a subset S of the set U of all cases, an entropy of a variable v (attribute or decision) with values v_1, v_2, ..., v_n is defined by the following formula:

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

where $p(v_i)$ is a probability (relative frequency) of value v_i in the set S, $i = 0, 1, ..., n$. All logarithms in this paper are binary.

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

Case	Attributes			Decision
	Length	**Width**	**Height**	**Quality**
1	3.9	1.7	1.4	very-low
2	3.9	1.8	1.4	low
3	4.3	1.8	1.5	medium
4	4.3	1.9	1.6	medium
5	4.7	1.7	1.5	medium
6	4.7	1.9	1.6	high
7	4.7	1.8	1.6	high

The conditional entropy of the decision d given an attribute a is:

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

where $p(a_j)$ is the probability of value a_j of the attribute a and $p(d_i|a_j)$ is the conditional probability of the value d_j of the decision d given a_j; $a_1, a_2, ..., a_m$ are all values of a, and $d_1, d_2, ..., d_n$ are all values of d. Discretization based on the conditional entropy of the concept given the attribute is considered to be one of the most successful discretization techniques [2,5,6,9,11,12,14,15,19,24,26,27].

Let a be an attribute and q be a cut point that splits the set S into two subsets, S_1 and S_2. The conditional entropy $H_S(d|q)$ 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 the set X. The cut point q for which the conditional entropy $H_S(d|q)$ has the smallest value is selected as the best cut point.

2.1. Stopping Criterion for Discretization

A stopping criterion of the process of discretization, described in this paper, is the *level of consistency* [5], based on *rough set theory* [35,36]. For any subset B of the set A of all attributes, an *indiscernibility* relation $IND(B)$ is defined, for any $x, y \in U$, in the following way:

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

where $a(x)$ denotes the value of the attribute $a \in A$ for the case $x \in U$. The relation $IND(B)$ is an equivalence relation. The equivalence classes of $IND(B)$ are denoted by $[x]_B$ and are called *B-elementary sets*. Any finite union of B-elementary sets is *B-definable*.

A partition on U constructed from all B-elementary sets of $IND(B)$ is denoted by B^*. $\{d\}$-elementary sets are called *concepts*, where d is a decision. For example, for Table 1, if $B = \{length\}$, B^* = {{1, 2}, {3, 4}, {5, 6, 7}} and $\{d\}^*$ = {{1}, {2}, {3, 4, 5}, {6, 7}}. In general, arbitrary $X \in \{d\}^*$ is not B-definable. For example, the concept {1} is not B-definable. However, any $X \in \{d\}^*$ may be approximated by a *B-lower approximation* of X, denoted by $\underline{B}X$ and defined as follows:

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

and by *B-upper approximation* of X, denoted by $\overline{B}X$ and defined as follows:

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

In our example, $\underline{B}\{1\}$ = ∅ and $\overline{B}\{1\}$ = {1, 2}. The B-lower approximation of X is the greatest B-definable set contained in X. The B-upper approximation of X is the least B-definable set containing X. A *level of consistency* [5], denoted by $L(A)$, is defined as follows:

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

Usually, the requested level of consistency for discretization is 1.0, *i.e.*, we want the discretized dataset to be *consistent*. For example, for Table 1, the level of consistency $L(A)$ is equal to 1.0, since {A}* = {{1}, {2}, {3}, {4}, {5}, {6}, {7}} and, for any X from {*quality*}* = {{1}, {2}, {3, 4, 5}, {6, 7}}, we have $\underline{A}X = X$. Additionally, $L(B) \approx 0.286$, where B = {*length*}.

2.2. Equal Interval Width and Equal Frequency per Interval

Some discretization methods are obvious. To this category belong the *equal interval width* and *equal frequency per interval* methods [14]. These methods are applied to a single attribute at a time, so they are called *local* [14]. Note that a discretization method is called *global* if it depends on all attributes [14]. In the local discretization method, the user must specify a positive integer k, a number of intervals required for discretization. In the former method, the domain of a numerical attribute a should be divided into k intervals that are approximately equal. In the latter method, the domain of the numerical attribute a should be divided into k intervals, each containing approximately an equal number of cases.

These two methods were converted to global methods, using the idea of entropy, in [5]. First, all numerical attributes are discretized using k = 2. After that, we need to compute the level of consistency for the set of all discretized attributes. If the level of consistency satisfies the requirement, the discretization is done. If not, the worst attribute must be selected for further discretization.

For an attribute a, let a^d denote the discretized attribute. Let A^d denote the set of all discretized attributes. For any partially-discretized attribute a^d, we define a measure of quality, called the *average block entropy*, in the following way:

$$M(a^d) = \frac{\sum_{B \in \{a^d\}^*} \frac{|B|}{|U|} H(B)}{|\{a^d\}^*|}$$

A partially-discretized attribute a^d with the largest $M(a^d)$ is the worst attribute [5]. The worst attribute is the subject of additional discretization for $k + 1$ intervals. The rest of the algorithm is defined by recursion. These new methods of discretization are called the *globalized version of equal interval width* and the *globalized version of equal frequency per interval*. As follows from [2], both methods are quite successful.

We will illustrate the globalized version of equal frequency per interval method by applying this method to Table 1. We need to compute the values of cut points for the first iteration. These cut points are: 4.5 for *length*, 1.85 for *width* and 1.55 for *height*. Table 2 presents the partially-discretized dataset. The partitions on U corresponding to partially-discretized attributes are:

$\{length^d\}^*$ = {{1, 2, 3, 4}, {5, 6, 7}},
$\{width^d\}^*$ = {{1, 2, 3, 5, 7}, {4, 6}},
$\{height^d\}^*$ = {{1, 2, 3, 5}, {4, 6, 7}}.

The partition $(A^d)^*$ is {{1, 2, 3}, {4}, {5}, {6}, {7}}.

Thus, the level of consistency, after the initial discretization, is $L(A) \approx 0.571$. In order to find the worst attribute, we compute the average block entropy for all initially discretized attributes. For the attribute *length*^{d},

$$M(length^d) = \frac{1}{2}(\frac{4}{7}((-\frac{1}{4} \cdot log\frac{1}{4})(2) + (-\frac{1}{2} \cdot log\frac{1}{2})) + \frac{3}{7}(-\frac{1}{3} \cdot log\frac{1}{3} - \frac{2}{3} \cdot log\frac{2}{3}) \approx 0.625.$$

Table 2. Partially-discretized Table 1 using equal frequency per interval: Part I.

Case	Attributes			Decision
	Length^d	**Width^d**	**Height^d**	**Quality**
1	3.9..4.5	1.7..1.85	1.4..1.55	very-low
2	3.9..4.5	1.7..1.85	1.4..1.55	low
3	3.9..4.5	1.7..1.85	1.4..1.55	medium
4	3.9..4.5	1.85..1.9	1.55..1.6	medium
5	4.5..4.7	1.7..1.85	1.4..1.55	medium
6	4.5..4.7	1.85..1.9	1.55..1.6	high
7	4.5..4.7	1.7..1.85	1.55..1.6	high

Additionally, $M(width^d) \approx 0.829$ and $M(height^d) \approx 0.625$. The worst attribute is *width*^{d}. We need to compute new cut points for *width*^{d} with k = 3. The new cut point is 1.75. This time: $\{width^d\}^*$ = {{1, 5}, {2, 3, 7}, {4, 6}}, so $(A^d)^*$ = {{1}, {2, 3}, {4}, {5}, {6}, {7}}, and the new level of consistency $L(A^d)$ is 0.714. The average block entropy, for *width*^{d} with k = 3, is 0.417. This time, the worst attribute is *length*^{d}. The third cut point for *length*^{d} is 4.1. This time $(A^d)^*$ = {{1}, {2}, {3}, {4}, {5}, {6}, {7}}. Table 3 presents the new discretized table. Furthermore, the level of consistency $(A^d)^*$ for Table 3 is one, so the process of discretization is done.

Table 3. Partially-discretized Table 1 using equal frequency per interval: Part II.

Case	Attributes			Decision
	Length^d	**Width^d**	**Height^d**	**Quality**
1	3.9..4.1	1.7..1.75	1.4..1.55	very-low
2	3.9..4.1	1.75..1.85	1.4..1.55	low
3	4.1..4.5	1.75..1.85	1.4..1.55	medium
4	4.1..4.5	1.85..1.9	1.55..1.6	medium
5	4.5..4.7	1.7..1.75	1.4..1.55	medium
6	4.5..4.7	1.85..1.9	1.55..1.6	high
7	4.5..4.7	1.75..1.85	1.55..1.6	high

2.3. *Multiple Scanning*

In the *multiple scanning* discretization method, the entire attribute set is scanned t times; t is a parameter selected by the user. In our experiments, we applied t = 1, 2, ..., until the error rate, a result of ten-fold cross-validation, computed by C4.5, was the same for two consecutive values of t. Initially, for each attribute, the best cut point is selected, using the minimum of conditional entropy $H_S(d|q)$, for all possible values of q. During the next scans (*i.e.*, for t = 2, 3, ...), the entire attribute set is scanned again; for each attribute, we identify one cut point: for each block X of $(A^d)^*$, the best cut point is selected, the best cut point among all such blocks is accepted as the best cut point for the attribute. If the requested parameter t is reached and the dataset needs more discretization since $L(A^d) \neq 1$, the dominant attribute technique is used for remaining discretization.

Let us discretize Table 1 using the multiple scanning method. First, we need to compute the conditional entropy $H_U(d|q)$ for each attribute q and for all possible cut points for each attribute. The first attribute is *length*, with two possible cut points: 4.1 and 4.5. The corresponding conditional entropies are:

$$H_{length}(4.1, U) = \frac{2}{7}(-\frac{1}{2} \cdot \log \frac{1}{2})(2) + \frac{5}{7}(-\frac{2}{5} \cdot \log \frac{2}{5} - \frac{3}{5} \cdot \log \frac{3}{5}) \approx 0.980,$$

$$H_{length}(4.5, U) = \frac{4}{7}((-\frac{1}{4} \cdot \log \frac{1}{4})(2) - \frac{1}{2} \cdot \log \frac{1}{2}) + \frac{3}{7}(-\frac{1}{3} \cdot \log \frac{1}{3} - \frac{2}{3} \cdot \log \frac{2}{3}) \approx 1.250.$$

For the attribute length, the better cut point is 4.1. For the attribute *width*, there are two possible cut points: 1.75 and 1.85, with $H_{length}(1.75, U)$ = 1.373 and $H_{length}(1.85, U)$ = 1.659; the better cut point is 1.75. For the attribute *height*, there are two possible cut points: 1.45 and 1.55, with $H_{height}(1.45, U)$ = 0.980 and $H_{height}(1.55, U)$ = 1.251; the better cut point is 1.45. A partially-discretized dataset, after the first scan, is presented in Table 4.

For Table 4, $(A^d)^*$ = {{1}, {2}, {3, 4, 6, 7}, {5}}, and $L(A^d)$ = 0.429. The remaining discretization is conducted using the dominant attribute method for the sub-table presented in Table 5.

Table 4. Partially-discretized Table 1 using multiple scanning, after the first scan.

Case	Attributes			Decision
	Length^d	**Width^d**	**Height^d**	**Quality**
1	3.9..4.1	1.7..1.75	1.4..1.45	very-low
2	3.9..4.1	1.75..1.9	1.4..1.45	low
3	4.1..4.7	1.75..1.9	1.45..1.6	medium
4	4.1..4.7	1.75..1.9	1.45..1.6	medium
5	4.1..4.7	1.7..1.75	1.45..1.6	medium
6	4.1..4.7	1.75..1.9	1.45..1.6	high
7	4.1..4.7	1.75..1.9	1.45..1.6	high

Table 5. A sub-table of the dataset presented in Table 1.

Case	Attributes			Decision
	Length	**Width**	**Height**	**Quality**
3	4.3	1.8	1.5	medium
4	4.3	1.9	1.6	medium
6	4.7	1.9	1.6	high
7	4.7	1.8	1.6	high

It is clear that the attribute *length* is the best attribute (with the smallest entropy) and that the remaining cut point is 4.5 for the attribute *length*. As a result, we obtain Table 6, for which $L(A^d)$ = 1.

Table 6. Partially-discretized Table 1 using multiple scanning and the dominant attribute.

Case	Attributes			Decision
	Length^d	Width^d	Height^d	Quality
1	3.9..4.1	1.7..1.75	1.4..1.45	very-low
2	3.9..4.1	1.75..1.9	1.4..1.45	low
3	4.1..4.5	1.75..1.9	1.45..1.6	medium
4	4.1..4.5	1.75..1.9	1.45..1.6	medium
5	4.5..4.7	1.7..1.75	1.45..1.6	medium
6	4.5..4.7	1.75..1.9	1.45..1.6	high
7	4.5..4.7	1.75..1.9	1.45..1.6	high

2.4. Interval Merging

In the discretization techniques presented in this paper, except the internal discretization method of C4.5, the last step is an attempt to merge intervals. During merging intervals, we want to reduce the number of intervals while preserving consistency. The corresponding algorithm has two steps:

- *safe merging*: for any discretized attribute a^d and for any two neighboring intervals $i..j$ and $j..k$, if both intervals belong to the same concept, these intervals are merged (or replaced by the interval $i..k$);
- *proper merging*: for any discretized attribute a^d and for any two neighboring intervals $i..j$, $j..k$, if the new interval $i..k$, the result of merging, does not reduce the level of consistency $L(A^d)$, these intervals are merged (or replaced by the new interval $i..k$).

In Table 3, we may merge intervals 1.75..1.85 and 1.85..1.9 of the attribute *width*^{d}. Additionally, we may merge the only two intervals 1.4..1.55 and 1.55..1.6 of the attribute *height*^{d}, so the attribute *height*^{d} becomes redundant. Table 7 presents the final, discretized table by the equal frequency per interval method.

In Table 6, we may merge both intervals of the attribute *height*^{d}; so finally, the table discretized by the multiple scanning method is identical to Table 7.

Table 7. Table 1 discretized using equal frequency per interval.

Case	Attributes		Decision
	Length^d	Width^d	Quality
1	3.9..4.1	1.7..1.75	very-low
2	3.9..4.1	1.75..1.9	low
3	4.1..4.5	1.75..1.9	medium
4	4.1..4.5	1.75..1.9	medium
5	4.5..4.7	1.7..1.75	medium
6	4.5..4.7	1.75..1.9	high
7	4.5..4.7	1.75..1.9	high

3. Experiments

We conducted experiments on 17 datasets with numerical attributes presented in Table 8. These datasets, except *bankruptcy*, were taken from the *Machine Learning Repository* stored at the University of California, Irvine. The *bankruptcy* dataset is a well-known dataset used by Altman to predict the bankruptcy of companies [37].

Table 8. Datasets.

Dataset	Number of		
	Cases	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
Ionosphere	351	34	2
Iris	150	4	3
Leukemia	415	175	2
Pima	768	8	2
Spectrometry	25,931	152	–
Wave	512	21	3
Wine Recognition	178	13	3
Yeast	1484	8	9

Four discretization methods were used:

- Original C4.5 discretization scheme,
- Globalized version of the equal interval width,
- Globalized version of the equal frequency per interval, and
- Multiple scanning.

For all four discretization methods, an error rate was estimated using the ten-fold cross-validation procedure of C4.5, with the level of consistency equal to 100%. Table 9 presents error rates for all four discretization methods.

Table 10 shows the size of decision trees generated by the C4.5 system.

Table 9. Error rates for C4.5, globalized versions of equal interval width (GVEIW), equal frequency per interval (GVEFI) and multiple scanning (MS).

Data Set	C4.5	GVEIW	GVEFI	MS
Abalone	80.58	76.90	76.87	75.58
Australian	16.09	13.33	12.46	13.48
Bankruptcy	6.06	10.61	3.03	3.03
Bupa	35.36	34.49	35.94	29.28
Connectionist Bench	25.96	25.00	29.33	16.83
Echocardiogram	28.38	31.08	27.03	14.86
Ecoli	22.02	28.57	30.65	22.02
Glass	33.18	33.18	41.12	24.77
Image Segmentation	12.38	18.10	19.52	11.90
Ionosphere	10.54	10.83	13.11	5.98
Iris	5.33	4.00	12.67	4.67
Leukemia	21.45	24.34	21.20	21.20
Pima	25.13	24.87	27.34	24.09
Spectrometry	0.99	1.32	1.45	1.13
Wave	26.37	27.54	25.59	23.05
Wine Recognition	8.99	9.55	10.11	3.93
Yeast	44.41	56.54	57.82	51.75

Table 10. Tree size for C4.5, globalized versions of equal interval width (GVEIW), equal frequency per interval (GVEFI) and multiple scanning (MS).

Data Set	C4.5	GVEIW	GVEFI	MS
Abalone	2367	18,711	28,236	9220
Australian	63	39	41	13
Bankruptcy	3	6	3	3
Bupa	51	27	17	11
Connectionist Bench	35	48	35	31
Echocardiogram	9	16	8	5
Ecoli	43	109	61	57
Glass	45	186	70	58
Image Segmentation	25	37	47	24
Ionosphere	35	34	33	24
Iris	9	4	11	4
Leukemia	61	53	68	33
Pima	43	33	37	27
Spectrometry	151	229	139	62
Wave	85	107	62	55
Wine Recognition	9	18	19	11
Yeast	371	913	662	362

For the analysis of the experimental results, we used the Friedman rank sum test combined with multiple comparisons, with a 5% level of significance. We conclude that the multiple scanning discretization method is associated with a significantly smaller error rate than all three remaining discretization methods: the original C4.5 discretization method and the globalized versions of the equal interval width and equal frequency per interval discretization methods. The differences between the performance of C4.5, the globalized versions of the equal interval width and equal frequency per interval discretization methods and multiple scanning are statistically insignificant. Additionally, decision trees generated by C4.5 from datasets discretized by multiple scanning are simpler than decision trees generated by C4.5 from datasets discretized by both globalized versions of equal interval width and equal frequency per interval.

4. Conclusions

We present results of our experiments using four different discretization techniques based on entropy. These discretization techniques were validated by conducting experiments on 17 datasets with numerical attributes. Our results show that the multiple scanning discretization technique is significantly better than the internal discretization used in C4.5 and two globalized discretization methods: equal interval width and equal frequency per interval in terms of an error rate computed by ten-fold cross-validation (two-tailed test, 5% level of significance). Additionally, decision trees generated from data discretized by multiple scanning are significantly simpler than decision trees generated directly by C4.5 and decision trees generated from discretized datasets and both globalized discretization methods.

Our results show that multiple scanning is the best discretization method in terms of the error rate and that decision trees generated from datasets discretized by multiple scanning are simpler than decision trees generated by both global discretization methods.

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

Author Contributions: Jerzy W. Grzymala-Busse and Teresa Mroczek wrote the paper while Teresa Mroczek conducted the experiments. Both authors have read and approved the final manuscript.

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

References

1. Grzymala-Busse, J.W.; Mroczek, T. A comparison of two approaches to discretization: Multiple scanning and C4.5. In Proceedings of the 6th International Conference on Pattern Recognition and Machine Learning, Warsaw, Poland, 30 June–3 July 2015; pp. 44–53.
2. 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 3rd International Conference on Rough Sets and Knowledge Technology, Chengdu, China, 17–19 May 2008; pp. 31–38.
3. 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.
4. 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.
5. 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.
6. 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.
7. De Sa, C.R.; Soares, C.; Knobbe, A. Entropy-based discretization methods for ranking data. *Inf. Sci.* **2016**, *329*, 921–936.
8. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and unsupervised discretization of continuous features. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; pp. 194–202.
9. Elomaa, T.; Rousu, J. General and efficient multisplitting of numerical attributes. *Mach. Learn.* **1999**, *36*, 201–244.
10. Elomaa, T.; Rousu, J. Efficient multisplitting revisited: Optima-preserving elimination of partition candidates. *Data Min. Knowl. Discov.* **2004**, *8*, 97–126.
11. Fayyad, U.M.; Irani, K.B. On the handling of continuous-valued attributes in decision tree generation. *Mach. Learn.* **1992**, *8*, 87–102.
12. Fayyad, U.M.; Irani, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the 13th International Joint Conference on Artificial Intelligence, San Mateo, CA, USA, 28 August–3 September 1993; pp. 1022–1027.
13. Garcia, S.; Luengo, J.; Sáez, J.A.; Lopez, V.; Herera, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 734–750.
14. Grzymala-Busse, J.W. Data Reduction: Discretization of Numerical Attributes. In *Handbook of Data Mining and Knowledge Discovery*; Klöesgen, W., Zytkow, J.M., Eds.; Oxford University Press: New York, NY, USA, 2002; pp. 218–225.
15. 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.
16. Grzymala-Busse, J.W. Mining Numerical Data—A Rough Set Approach. In *Transactions on Rough Sets XI*, Springer: Berlin/Heidelberg, Germany, 2010; pp. 1–13.
17. Grzymala-Busse, J.W.; Stefanowski, J. Three discretization methods for rule induction. *Int. J. Intell. Syst.* **2001**, *16*, 29–38.
18. Jiang, F.; Sui, Y. A novel approach for discretization of continuous attributes in rough set theory. *Knowl. Based Syst.* **2015**, *73*, 324–334.
19. Kohavi, R.; Sahami, M. Error-based and entropy-based discretization of continuous features. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 2–4 August 1996; pp. 114–119.
20. Kerber, R. ChiMerge: Discretization of numeric attributes. In Proceedings of the 10th National Conference on Artificial intelligence, San Jose, CA, USA, 12–16 July 1992; pp. 123–128.
21. Kotsiantis, S.; Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *32*, 47–58.
22. Kurgan, L.A.; Cios, K.J. CAIM discretization algorithm. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 145–153.

23. Liu, H.; Hussain, F.; Tan, C.L.; Dash, M. Discretization: An enabling technique. *Data Min. Knowl. Discov.* **2002**, *6*, 393–423.
24. Nguyen, H.S.; Nguyen, S.H. Discretization Methods in Data Mining. In *Rough Sets in Knowledge Discovery 1: Methodology and Applications*; Physica-Verlag: Heidelberg, Germany, 1998; pp. 451–482.
25. 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.
26. Stefanowski, J. Handling continuous attributes in discovery of strong decision rules. In Proceedings of the 1st Conference on Rough Sets and Current Trends in Computing, Warsaw, Poland, 22–26 June 1998; pp. 394–401.
27. Stefanowski, J. *Algorithms of Decision Rule Induction in Data Mining*; Poznan University of Technology Press: Poznan, Poland, 2001.
28. Rahman, M.D.; Islam, M.Z. Discretixation of continuous attributes through low frequency numerical values and attribute interdependency. *Expert Syst. Appl.* **2016**, *45*, 410–423.
29. 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.
30. 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.
31. Yang, Y.; Webb, G. Discretization for naive-Bayes learning: Managing discretization bias and variance. *Mach. Learn.* **2009**, *74*, 39–74.
32. Grzymala-Busse, J.W. Discretization based on entropy and multiple scanning. *Entropy* **2013**, *15*, 1486–1502.
33. Grzymala-Busse, J.W. A new version of the rule induction system LERS. *Fundam. Inform.* **1997**, *31*, 27–39.
34. Grzymala-Busse, J.W. MLEM2: A new algorithm for rule induction from imperfect data. In Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Annecy, France, 1–5 July 2002; pp. 243–250.
35. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356.
36. Pawlak, Z. *Rough Sets. Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publishers: Norwell, MA, USA, 1991.
37. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609.