

Article

Discretization Based on Entropy and Multiple Scanning

Jerzy W. Grzymala-Busse ^{1,2}

¹ Department of Electrical Engineering and Computer Science, University of Kansas, 3014 Eaton Hall, Lawrence, KS 66045, USA; E-Mail: jerzy@ku.edu; Tel.: +1-785-864-4488; Fax: +1-785-864-3226

² Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, Rzeszow 35-225, Poland

Received: 28 February 2013; in revised form: 16 April 2013 / Accepted: 18 April 2013 /

Published: 25 April 2013

Abstract: In this paper we present entropy driven methodology for discretization. Recently, the original entropy based discretization was enhanced by including two options of selecting the best numerical attribute. In one option, Dominant Attribute, an attribute with the smallest conditional entropy of the concept given the attribute is selected for discretization and then the best cut point is determined. In the second option, Multiple Scanning, all attributes are scanned a number of times, and at the same time the best cut points are selected for all attributes. The results of experiments on 17 benchmark data sets, including large data sets, with 175 attributes or 25,931 cases, are presented. For comparison, the results of experiments on the same data sets using the global versions of well-known discretization methods of Equal Interval Width and Equal Frequency per Interval are also included. The entropy driven technique enhanced both of these methods by converting them into globalized methods. Results of our experiments show that the Multiple Scanning methodology is significantly better than both: Dominant Attribute and the better results of Globalized Equal Interval Width and Equal Frequency per Interval methods (using two-tailed test and 0.01 level of significance).

Keywords: numerical attributes; entropy; discretization; data mining

1. Introduction

Discretization of numerical attributes is one of the basic techniques of data mining. During this process, numerical values are transformed into intervals. Among the many discretization techniques,

discretization based on the conditional entropy of the concept given an attribute is considered to be one of the most successful methods [1–21].

For a numerical attribute a with an interval $[i, j]$ as a range, a partition of the range into k intervals

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

where $i_0 = i$, $i_k = j$, and $i_l < i_{l+1}$ for $l = 0, 1, \dots, k - 1$, defines a discretization of a . The numbers i_1, i_2, \dots, i_{k-1} are called *cut points*. Our discretization system denotes such intervals as $i_0..i_1, i_1..i_2, \dots, i_{k-1}..i_k$.

Discretization methods that may be applied only to one variable at a time are called *local*. Methods with all attributes processed during discretization are called *global*.

We present an enhanced version of the original discretization based on entropy presented in [22]. In this version there is a choice between two options. The first option is called *Dominant Attribute*. First the current best attribute is selected, then, for this attribute, the best cut point, using conditional entropy again, is selected. This process continues until a stopping criterion is satisfied. In the second option, called *Multiple Scanning*, the entire attribute set is scanned. For any attribute, the best cut point is selected, then 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 the Dominant Attribute technique. The same stopping criterion, based on rough set theory, was used in all experiments. The quality of each discretization technique was evaluated by an error rate computed as a result of ten-fold cross validation (with the exception of the *spectrometry* data set where hold-out was used). Our experiments were conducted on 17 benchmark data sets.

For comparison, results of experiments on the same data sets using well-known discretization methods of Equal Interval Width and Equal Frequency per Interval are also included. These two methods are local, but we converted both methods to global using entropy as well. Results of our experiments show that the Multiple Scanning methodology is significantly better than both: Dominant Attribute and the better results of Globalized Equal Interval Width and Equal Frequency per Interval methods (using two-tailed test and 0.01 level of significance).

A preliminary version of this paper was presented at the ISMIS 2009, the 18th International Symposium on Methodologies for Intelligent Systems [11].

2. Entropy Based Discretization

We are assuming that the input data set is given in a form of the table exemplified by Table 1. In such a table all cases are described by variables called *attributes* and one variable is called a *decision* (or class) and is denoted by d . The set of all attributes will be denoted by A . The set of all cases will be denoted by U . In Table 1 the attributes are *Weight*, *Length* and *Height* while the decision is *Price*. Additionally, $U = \{1, 2, 3, 4, 5, 6, 7\}$. An entropy of a variable v (attribute or decision) with values v_1, v_2, \dots, v_n is defined by the following formula

$$H_v(U) = - \sum_{i=1}^n p(v_i) \cdot \log p(v_i)$$

where U is the set of all cases in a data set and $p(v_i)$ is a probability (relative frequency) of value v_i in the set U , $i = 0, 1, \dots, n$. All logarithms in this paper are binary.

Table 1. An example of a data set with numerical attributes.

Case	Attributes			Decision Price
	Weight	Length	Height	
1	0.8	0.3	7.2	very small
2	0.8	1.1	7.2	small
3	0.8	1.1	10.2	medium
4	1.2	0.3	10.2	medium
5	1.2	2.3	10.2	medium
6	2.0	2.3	10.2	high
7	2.0	2.3	15.2	very high

A conditional entropy of the decision d given an attribute a is

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

where a_1, a_2, \dots, a_m are all values of a and d_1, d_2, \dots, d_n are all values of d . There are two fundamental criteria of quality based on entropy. The first is an *information gain* associated with an attribute a and defined by

$$I(a) = H_d(U) - H(d|a)$$

the second is *information gain ratio*, for simplicity called *gain ratio*, defined by

$$G(a) = \frac{I(a)}{H_a(U)}$$

Both criteria were introduced by J. R. Quinlan, see, e.g., [22] and used for decision tree generation.

For a cut point q for an attribute a , the conditional entropy, defined by a cut point q that splits the set U of all cases into two sets S_1 and S_2 , is defined as follows

$$H_a(q, U) = \frac{|S_1|}{|U|} H_a(S_1) + \frac{|S_2|}{|U|} H_a(S_2)$$

where $|X|$ denotes the cardinality of the set X . The cut point q for which the conditional entropy $H_a(q, U)$ has the smallest value is selected as the best cut point. The corresponding information gain is the largest.

2.1. Stopping Criterion for Discretization

A stopping criterion of the process of discretization, described in this paper, is the *level of consistency* [3], based on *rough set theory* [23,24]. 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 $x \in U$. For example, in Table 1, $Weight(1) = 0.8$.

For Table 1, let $B = \{Weight, Length\}$. Cases 2 and 3 are B -indiscernible, *i.e.*, $(2, 3) \in IND(B)$, since $Weight(2) = Weight(3)$ and $Length(2) = Length(3)$. On the other hand, $(1, 2) \notin IND(B)$ since $Length(1) \neq Length(2)$.

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. For Table 1, B -elementary classes are $\{1\}$, $\{2, 3\}$, $\{4\}$, $\{5\}$ and $\{6, 7\}$.

Any finite union of B -elementary sets is B -definable. A partition on U constructed from all B -elementary sets of $IND(B)$ will be denoted by B^* . $\{d\}$ -elementary sets are called *concepts*, where d is a decision. For example, for Table 1, $\{d\}^* = \{\{1\}, \{2\}, \{3, 4, 5\}, \{6\}, \{7\}\}$. In general, arbitrary $X \in \{d\}^*$ is not B -definable. For example, the concept $\{3, 4, 5\}$ 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, $\underline{B}\{3, 4, 5\} = \{4, 5\}$ and $\overline{B}\{3, 4, 5\} = \{2, 3, 4, 5\}$.

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 B -rough set of X is the family of all subsets of U having the same B -lower and B -upper approximations of X . For the set $X = \{3, 4, 5\}$, the corresponding B -rough set is $\{\{2, 4, 5\}, \{3, 4, 5\}\}$. Finally, a *level of consistency* [3], denoted by $L(A)$, is defined as follows

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

Practically, the requested level of consistency for discretization is 1.0, *i.e.*, we want the discretized data set to be *consistent*. For example, for Table 1, the level of consistency $L(A)$ is equal to one, since $\{A\}^* = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}\}$ and, for any X from $\{Price\}^* = \{\{1\}, \{2\}, \{3, 4, 5\}, \{6\}, \{7\}\}$, we have $\underline{A}X = X$. On the other hand, $L(B) = 0.857$.

2.2. Dominant Attribute Strategy

We will discuss two basic discretization techniques based on entropy. The first discretization technique is called *Dominant Attribute* (or *Starting from One Attribute* [11]). A similar idea for a decision tree generation is used in the C4.5 system [22]. In addition, for numerical attributes, C4.5 uses discretization similar to the Dominant Attribute approach, yet C4.5 not only does not use the measure $L(A)$ as stopping condition but also does not use merging of intervals created during such discretization either.

The Dominant Attribute algorithm is recursive:

- for a given set of cases (initially it is U), we identify the best attribute (*i.e.*, the attribute with the largest information gain or the attribute with the largest gain ratio),
- for the best attribute, we are looking for the best cut point, *i.e.*, the cut point with the smallest conditional entropy,
- the best cut point divides the data set into two smaller data sets, S_1 and S_2 .
- we apply the same strategy for both smaller data sets separately,
- the algorithm stops when $L(A^D) = 1$, where A^D is the discretized set of attributes,

We need to take into account that the order in which we process smaller data sets may affect the discretization. We will illustrate this method by discretizing the data set from Table 1. We will use the information gain as the criterion to select the best attribute.

The conditional entropy $H(Price|Weight)$ is

$$\frac{3}{7}(-\frac{1}{3} \cdot \log \frac{1}{3})(3) + \frac{2}{7}(0) + \frac{2}{7}(-\frac{1}{2} \cdot \log \frac{1}{2})(2) = 0.965$$

Similarly, the conditional entropies $H(Price|Length) = 1.250$ and $H(Price|Height) = 0.749$. The minimal conditional entropy is associated with attribute *Height*. What is the best cut point for attribute *Height* is the next question. This attribute has two potential cut points (averages between sorted values of the attribute *Height*): 8.7 and 12.7. The conditional entropy $H_{Height}(8.7, U)$ is

$$\frac{2}{7}(-\frac{1}{2} \cdot \log \frac{1}{2})(2) + \frac{5}{7}(-\frac{3}{5} \cdot \log \frac{3}{5} + (-\frac{1}{5} \cdot \log \frac{1}{5})(2)) = 1.265$$

similarly, the conditional entropy $H_{Height}(12.7, U) = 1.536$. Thus we will select the cut point 8.7. Obviously, the current discretization of attribute *Height* into two intervals 7.2..8.7 and 8.7..15.2 is not sufficient, since if we will use only discretized attribute $Height^D$ and $A = \{Height^D\}$, $\{A\}^* = \{\{1, 2\}, \{3, 4, 5, 6, 7\}\}$, we have $\underline{A}X = \emptyset$ for any member X of $\{Price\}^*$, so $L(A) = 0$. The current discretization splits Table 1 into two sub-tables, Tables 2 and 3.

It is also obvious that for Table 2 the only attribute that may be discretized is *Length*, with the cut point is equal to 0.7. Table 4 presents the current situation: discretized are attributes *Length* and *Height*, with cut points 0.7 and 8.7, respectively.

For Table 4 and $A^D = \{Length^D, Height^D\}$, where $Length^D$ and $Height^D$ denote currently discretized attributes, $(A^D)^* = \{\{1\}, \{2\}, \{3, 5, 6, 7\}, \{4\}\}$, and $L(A^D) = 0.429$, so further discretization is necessary. However, by analysis of Table 4 we may easily discover that all what we need to do is to distinguish cases 3 and 5 from cases 6 and 7 and that cases 3 and 4 do not need to be distinguished.

Thus, our next table to be discretized is presented as Table 5 (note that Table 5 is simpler than Table 3). We will continue discretization by recursion. Our final choice of cut points is 1.6 for *Weight*, 0.7 for *Length*, and 8.7 and 12.7 for *Height*. The final discretized table is presented as Table 6.

Table 2. The first sub-table of Table 1.

Case	Attributes			Decision Price
	Weight	Length	Height	
1	0.8	0.3	7.2	very small
2	0.8	1.1	7.2	small

Table 3. The second sub-table of Table 1.

Case	Attributes			Decision Price
	Weight	Length	Height	
3	0.8	1.1	10.2	medium
4	1.2	0.3	10.2	medium
5	1.2	2.3	10.2	medium
6	2.0	2.3	10.2	high
7	2.0	2.3	15.2	very high

Table 4. Table 1, partially discretized, with attributes Length and Height.

Case	Attributes		Decision Price
	Length	Height	
1	0.3..0.7	7.2..8.7	very small
2	0.7..2.3	7.2..8.7	small
3	0.7..2.3	8.7..15.2	medium
4	0.3..0.7	8.7..15.2	medium
5	0.7..2.3	8.7..15.2	medium
6	0.7..2.3	8.7..15.2	high
7	0.7..2.3	8.7..15.2	very high

Table 5. A new sub-table of Table 1.

Case	Attributes			Decision Price
	Weight	Length	Height	
3	0.8	1.1	10.2	medium
5	1.2	2.3	10.2	medium
6	2.0	2.3	10.2	high
7	2.0	2.3	15.2	very high

Table 6. Table 1 discretized by the Dominant Attribute discretization technique.

Case	Attributes			Decision Price
	Weight	Length	Height	
1	0.8..1.6	0.3..0.7	7.2..8.7	very small
2	0.8..1.6	0.7..2.3	7.2..8.7	small
3	0.8..1.6	0.7..2.3	8.7..12.7	medium
4	0.8..1.6	0.3..0.7	8.7..12.7	medium
5	0.8..1.6	0.7..2.3	8.7..12.7	medium
6	1.6..2.0	0.7..2.3	8.7..12.7	high
7	1.6..2.0	0.7..2.3	12.7..15.2	very high

2.3. Multiple Scanning Strategy

The second discretization technique needs some parameter denoted by t and called the total number of scans. In Multiple Scanning Algorithm,

- for the entire set A of attributes the best cut point is computed for each attribute $a \in A$ based on minimum of conditional entropy $H(d|a)$, a new discretized attribute set is A^D , and the original data set is partitioned into a partition $(A^D)^*$,
- if the number of scans t is not reached, the next scan is conducted: we need to scan the entire set of partially discretized attributes again; for each attribute we need only one cut point, the best cut point for each block $X \in (A^D)^*$ is computed, the best cut point among all such blocks is selected,
- if the requested number of scans t is reached and the data set needs more discretization, the Dominant Attribute technique is used for the remaining sub-tables,
- the algorithm stops when $L(A^D) = 1$, where A^D is the discretized set of attributes.

We will illustrate this technique by scanning all attributes, *Weight*, *Length*, and *Height* once. First we are searching for the best cut point for attributes *Weight*, *Length*, and *Height*. The best cut points are 1.6, 1.7, and 8.7, respectively. Table 1, partially discretized this way, is presented as Table 7.

Table 7. Partially discretized Table 1 using Multiple Scanning.

Case	Attributes			Decision Price
	Weight	Length	Height	
1	0.8..1.6	0.3..1.7	7.2..8.7	very small
2	0.8..1.6	0.3..1.7	7.2..8.7	small
3	0.8..1.6	0.3..1.7	8.7..15.2	medium
4	0.8..1.6	0.3..1.7	8.7..15.2	medium
5	0.8..1.6	1.7..2.3	8.7..15.2	medium
6	1.6..2.0	1.7..2.3	8.7..15.2	high
7	1.6..2.0	1.7..2.3	8.7..15.2	very high

The level of consistency for Table 7 is 0.429 since $A^* = \{\{1, 2\}, \{3, 4\}, \{5\}, \{6, 7\}\}$, we need to distinguish cases 1 and 2, and, separately, cases 6 and 7. Therefore we need to use the *Dominant Attribute* technique for two sub-tables, first with two cases, 1 and 2, and second with also two cases, 6 and 7. As a result, we will select cut points 0.7 and 12.7 for attributes *Length* and *Height*, respectively.

2.4. Globalized Versions of Equal Interval Width and Equal Frequency per Interval

Two discretization methods, called *Equal Interval Width* and *Equal Frequency per Interval* are frequently used in data mining. Both methods are local. In the Equal Interval Width method, the domain of a numerical attribute a is divided into k equal intervals, where k is a real number set up by the user. In the Equal Frequency per Interval method, attribute values are distributed in such a way that in all k intervals the number of attribute values is approximately equal to each other. This method is sometimes called a Maximum Entropy Discretization [21].

For our experiments, both methods were converted to global by using entropy. In this approach, the first step is to discretize all attributes, by computing a cut point for all attributes, assuming $k = 2$. If the level of consistency satisfies requirements, the process is completed. If not, we need to select an attribute a whose initial distribution is the worst. A measure of quality for such distribution is the *average block entropy* of an attribute defined by the following formula

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

An attribute a for which $M(a)$ is maximum is considered to be the worst and, as such, is selected for re-discretization with k incremented by one. As follows from [12], the Globalized Versions of Equal Interval Width and Equal Frequency per Interval methods are successful and competitive.

Let us discretize Table 1 using the globalized version of the Equal Interval Width method. First we need to compute cut points for all attributes using the Equal Interval Width principle. Such cut points are: 1.4 for *Weight*, 1.3 for *Length* and 11.2 for *Height*. The corresponding partially discretized table is presented in Table 8. Partitions on U defined by partially discretized attributes are:

$$\{Weight\}^* = \{1, 2, 3, 4, 5\}, \{6, 7\},$$

$$\{Length\}^* = \{1, 2, 3, 4\}, \{5, 6, 7\},$$

$$\{Height\}^* = \{1, 2, 3, 4, 5, 6\}, \{7\},$$

the level of consistency is $L(A) = 0.429$ since $A^* = \{\{1, 2, 3, 4\}, \{5\}, \{6\}, \{7\}\}$. We need to compute the average block entropy for all attributes. For example,

$$M(Weight^D) = \frac{1}{2} \left(\frac{5}{7} \left(-\frac{1}{5} \cdot \log \frac{1}{5} \right) (2) + \left(-\frac{3}{5} \cdot \log \frac{3}{5} \right) \right) + \frac{2}{7} \left(-\frac{1}{2} \cdot \log \frac{1}{2} \right) (2) = 0.632$$

Table 8. Partially discretized Table 1 using Equal Interval Width.

Case	Attributes			Decision Price
	Weight	Length	Height	
1	0.8..1.4	0.3..1.3	7.2..11.2	very small
2	0.8..1.4	0.3..1.3	7.2..11.2	small
3	0.8..1.4	0.3..1.3	7.2..11.2	medium
4	0.8..1.4	0.3..1.3	7.2..11.2	medium
5	0.8..1.4	1.3..2.3	7.2..11.2	medium
6	1.4..2.0	1.3..2.3	7.2..11.2	high
7	1.4..2.0	1.3..2.3	11.2..15.2	very high

Table 9. Table 1 discretized by Equal Interval Width.

Case	Attributes			Decision Price
	Weight	Length	Height	
1	0.8..1.4	0.3..0.967	7.2..9.867	very small
2	0.8..1.4	0.967..1.633	7.2..9.867	small
3	0.8..1.4	0.967..1.633	9.867..12.533	medium
4	0.8..1.4	0.3..0.967	9.867..12.533	medium
5	0.8..1.4	1.633..2.3	9.867..12.533	medium
6	1.4..2.0	1.633..2.3	9.867..12.533	high
7	1.4..2.0	1.633..2.3	12.533..15.2	very high

Similarly, $M(Length^D) = 0.768$ and $M(Height^D) = 0.768$. The worst attributes are *Length* and *Height*. We need to select, heuristically, one attribute, say the first one, *i.e.*, *Length*. Obviously, any heuristic step like this one changes the outcome, but we cannot explore all possibilities due to computational complexity. We compute new cut points for *Length* and $k = 3$. These cut points are 1.2 and 1.6. This time $\{Length\}^* = \{1, 4\}, \{2, 3\}, \{5, 6, 7\}$, so $A^* = \{\{1, 4\}, \{2, 3\}, \{5\}, \{6\}, \{7\}\}$ and the new

level of consistency $L(A)$ is still 0.429 and the average block entropy, for the new attribute ($Length^D$), with three intervals, is $M(Length^D) = 0.417$. The worst attribute is *Height*. The new cut points for *Height* are 9.867 and 12.533. The level of consistency for the new discretized table is $L(A) = 1.0$ since $A^* = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}\}$, so the discretization is completed. Table 9 presents the final, discretized table.

2.5. Interval Merging

In all our discretization techniques, the last step of discretization was the merging of intervals, to reduce their number and, at the same time, preserve consistency. The algorithm for merging intervals consists of two steps:

- *safe merging*: for any attribute and for any two neighboring intervals $i..j$ and $j..k$ of the same discretized attribute, if both intervals are labeled by the same decision value, both intervals are merged, *i.e.*, replaced by a new interval $i..k$,
- *proper merging*: for any attribute and for any two neighboring intervals $i..j$, $j..k$ of the same discretized attribute, if a result $i..k$ of merging does not reduce the level of consistency $L(A^D)$, where A^D is the current set of discretized attributes, both intervals are merged (replaced by a new interval $i..k$).

If neighboring intervals $i..j$ and $j..k$ are merged, we say that the cut point j was eliminated. Obviously, the order in which pairs of neighboring intervals are selected for proper merging affects the final outcome. In our experiments, we selected two neighboring intervals with the smallest conditional entropy, taking all attributes and all intervals into account. Using interval merging, we may eliminate the cut points 1.7 and 1.633 for attribute *Length*, in Tables 7 and 9, respectively.

2.6. LEM2 Algorithm for Rule Induction

The discretized data were fed into the data system LERS (Learning from Examples based on Rough Sets) [25] for rule induction. LERS uses rough set theory to compute lower and upper approximations for concepts involved in conflicts with other concepts [23].

Rules induced from the lower approximation of the concept *certainly* describe the concept, hence such rules are called *certain*. On the other hand, rules induced from the upper approximation of the concept describe the concept *possibly*, so these rules are called *possible*.

The LEM2 algorithm (Learning from Examples Module, version 2) of LERS is most frequently used for rule induction. LEM2 explores the search space of attribute–value pairs. Its input data set is a lower or upper approximation of a concept, so its input data set is always consistent. In general, LEM2 computes a local covering and then converts it into a rule set [25]. Recently, a new, improved version of LEM2, called MLEM2, was developed [26].

3. Experiments

Our experiments were conducted on 17 data sets, summarized in Table 10. All of these data sets, with the exception of *bankruptcy*, *brain* and *spectrometry*, are available on the University of California

at Irvine *Machine Learning Repository*. The *bankruptcy data* set is a well-known data set used by E. I. Altman to predict a bankruptcy of companies [27]. The *leukemia* data set describes penetration across the blood–brain barrier [28]. The *spectrometry* data set describes human proteins used in the mass spectrometry [29].

Table 10. Data sets.

Data set	Number of		
	cases	attributes	concepts
Abalone	4,177	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	15	2
Wave	512	21	3
Wine	178	13	3
Yeast	1484	8	9

Every discretization method was applied to all data sets, with the level of consistency equal to 100%. For a choice of the best attribute, we used gain ratio. Rule sets were induced using the LEM2 algorithm of the LERS data mining system.

Table 11 presents results of ten-fold cross validation, for all data sets except *spectrometry*, using increasing number of scans. For the *spectrometry* data set hold-out, with split of the original data set into 70% for training and 30% for testing was applied due to its size. Obviously, for any data set, after some fixed number of scans, an error rate is stable (constant). For example, for the *Australian* data set, the error rate will be 15.65% for the scan number 4, 5, etc. Thus, any data set from Table 11 is characterized by two error rates: minimal and stable. For a given data set, the smallest error rate from Table 11 will be called *minimal* and the last entry in the row that corresponds to the data set will be called *stable*. For

example, for the *Australian* data set, the minimal error rate is 14.93% and the stable error rate is 15.65%. For some data sets (e.g., for *bankruptcy*), minimal and stable error rates are identical.

Table 11. Error rates for Multiple Scanning.

Data set	Error rate for scan number						
	0	1	2	3	4	5	6
Abalone	76.92	78.91	78.48	77.95	77.90	77.83	78.12
Australian	34.49	15.22	14.93	15.65			
Bankruptcy	3.03	9.09	1.52				
Bupa	31.30	29.28	30.14	26.67			
Connectionist Bench	29.33	27.88					
Echocardiogram	24.32	16.22					
Ecoli	19.64	20.54	18.75	20.83	21.43	20.54	20.83
Glass	24.77	34.58	20.56	25.70	24.77	25.70	26.64
Image Segmentation	29.52	19.52	16.19	17.14			
Ionosphere	10.83	6.27	9.69	7.12			
Iris	5.33	2.67	4.67				
Leukemia	22.41	19.52	20.48				
Pima	27.21	26.04	25.65	26.30	26.82	26.69	26.43
Spectrometry	6.04	1.92	1.74	1.83	1.84	1.95	
Wave	27.10	19.53	20.70	19.53	24.77	19.53	
Wine Recognition	11.24	2.81					
Yeast	56.74	50.47	48.99	48.92	51.28	52.83	

It is clear from Table 11 that the minimal error rate is associated with 0 scans (*i.e.*, with the method *Dominant Attribute*) only for the *abalone* data set. Using the Wilcoxon matched-pairs signed-ranks test, we conclude that the following two statements are statistically highly significant (*i.e.*, the significance level is equal to 1% for a two-tail test):

- the minimal error rate associated with *Multiple Scanning* is smaller than the error rate associated with *Dominant Attribute*,
- the minimal error rate associated with *Multiple Scanning* is smaller than the smaller error rate associated with *Globalized Equal Interval Width* and *Globalized Equal Frequency per Interval* (Table 12).

Table 12. Error rates for Globalized Equal Interval Width and Globalized Equal Frequency per Interval.

Data set	Error rate	
	Equal Width	Equal Frequency
Abalone	78.33	77.50
Australian	15.94	14.93
Bankruptcy	7.58	3.03
Bupa	33.33	39.71
Connectionist Bench	23.08	22.60
Echocardiogram	34.23	35.14
Ecoli	25.60	27.68
Glass	32.71	35.05
Image Segmentation	20.48	20.48
Ionosphere	14.81	13.96
Iris	5.33	10.67
Leukemia	22.54	22.65
Pima	30.60	30.21
Spectrometry	1.71	2.21
Wave	27.93	24.80
Wine Recognition	11.24	6.18
Yeast	57.68	55.39

For completeness we present run time for all four approaches: Dominant Attribute, Multiple Scanning, and Globalized Equal Interval Width and Equal Frequency per Interval (Table 13). Our experiments were conducted on a machine with 34 GB of RAM with Inter(R) Xeon Processor X5650 (12 MB cache, 2.66 GHz, 6 Cores) under Fedora 17 Linux operating system. For large data sets, such as *abalone* and *spectrometry*, both Dominant Attribute and Multiple Scanning algorithms, based on entropy, were faster than Globalized Equal Interval Width and Equal Frequency per Interval methods.

Additionally, the effects of scanning during discretization are presented in Tables 14 and 15. We selected the *echocardiogram* and *iris* data sets not only because their all attributes are numerical with real numbers as values but also because they have small number of attributes. For example, for the *echocardiogram* data set, for 0 scans, *i.e.*, for *Dominant Attribute*, it is clear that attribute *Age* was selected as the best and that during discretization eight cut points were selected. After a single scan, the same attribute was selected as the best attribute. The *Wall-score* attribute was redundant for 0 scans, but it became essential after the first scan.

Table 13. Run time for all four approaches: Dominant Attribute, Multiple Scanning, Globalized Equal Interval Width and Globalized Equal Frequency per Interval.

Data set	Run time			
	Dominant Attribute	Multiple Scanning	Globalized Equal Width	Globalized Equal Frequency
Abalone	0 m 11. 620 s	0 m 26.277 s	8 m 13.661 s	4 m 8.623 s
Australian	0 m 0.915 s	0 m 0.185 s	0 m 0.613 s	0 m 0.226 s
Bankruptcy	0 m 0.004 s	0 m 0.004 s	0 m 0.007 s	0 m 0.002 s
Bupa	0 m 0.023 s	0 m 0.036 s	0 m 0.007 s	0 m 0.040 s
Connectionist Bench	0 m 0.075 s	0 m 0.811 s	0 m 0.439 s	0 m 0.415 s
Echocardiogram	0 m 0.005 s	0 m 0.005 s	0 m 0.003 s	0 m 0.004 s
Ecoli	0 m 0.024 s	0 m 0.040 s	0m 0.065 s	0 m 0.047 s
Glass	0 m 0.024 s	0 m 0.042 s	0 m 0.026 s	0 m 0.024 s
Image Segmentation	0 m 0.035 s	0 m 0.174 s	0 m 0.087 s	0 m 0.054 s
Ionosphere	0 m 0.087 s	0 m 0.480 s	0 m 0.538 s	0 m 0.498 s
Iris	0 m 0.006 s	0 m 0.003 s	0 m 0.008 s	0 m 0.009 s
Leukemia	0 m 0.520 s	0 m 15.304 s	0 m 8.860 s	0 m 8.950 s
Pima	0 m 0.131 s	0 m 0.134 s	0 m 0.494 s	0 m 0.203 s
Spectrometry	3 m 36.673 s	3 m 58.913 s	8 m 46.789 s	7 m 58.871 s
Wave	0 m 0.155 s	0 m 0.479 s	0 m 0.128 s	0 m 0.086 s
Wine Recognition	0 m 0.011 s	0 m 0.034 s	0 m 0.020 s	0 m 0.017 s
Yeast	0 m 0.915 s	0 m 0.970 s	0 m 3.336 s	0 m 1.662 s

Table 14. Number of intervals for scanning data set *echocardiogram*.

Attribute	Number of scans			
	0		1	
	before merging	after merging	before merging	after merging
Age	8	6	6	6
Pericardial	1	1	2	2
Fractional	3	3	3	2
EPSS	2	2	2	1
LVDD	4	3	2	2
Wall-score	1	1	2	2
Wall-index	2	2	2	2

Table 15. Number of intervals for scanning data set *iris*.

Attribute	Number of scans					
	0		1		2	
	before merging	after merging	before merging	after merging	before merging	after merging
Sepal length	4	3	4	3	6	5
Sepal width	5	4	4	3	4	3
Petal length	4	4	4	3	4	4
Petal width	4	2	3	3	3	2

4. Conclusions

This paper presents results of experiments in which four different techniques were used for discretization. All four techniques were validated by conducting experiments on 17 data sets with numerical attributes. Our discretization techniques were combined with rule induction using the LEM2 rule induction algorithm. The results of our experiments show that the Multiple Scanning technique is significantly better than: Dominant Attribute and the better results of Globalized Equal Interval Width and Equal Frequency per Interval methods (using two-tailed test and 0.01 level of significance). Thus, we show that there exists a new successful technique for discretization.

Acknowledgments

The author would like to thank the anonymous reviewers for their valuable suggestions.

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. 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.
4. 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.
5. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and Unsupervised Discretization of Continuous Features. In Proceedings of the 12th International Conference on Machine Learning, Lake Tahoe, CA, USA, 9–12 July 1995; pp. 194–202.
6. Elomaa, T.; Rousu, J. General and efficient multisplitting of numerical attributes. *Mach. Learn.* **1999**, *36*, 201–244.
7. Elomaa, T.; Rousu, J. Efficient multisplitting revisited: Optima-preserving elimination of partition candidates. *Data Min. Knowl. Discov.* **2004**, *8*, 97–126.

8. Fayyad, U.M.; Irani, K.B. On the handling of continuous-valued attributes in decision tree generation. *Mach. Learn.* **1992**, *8*, 87–102.
9. Fayyad, U.M.; Irani, K.B. Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. In Proceedings of the Thirteenth International Conference on Artificial Intelligence, Chambéry, France, 28 August–September 3 1993; pp. 1022–1027.
10. 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.
11. Grzymala-Busse, J.W. A Multiple Scanning Strategy for Entropy Based Discretization. In Proceedings of the ISMIS-09, 18th International Symposium on Methodologies for Intelligent Systems, Prague, Czech Republic, 14–17 September 2009; pp. 25–34.
12. Grzymala-Busse, J.W. Mining numerical data—A rough set approach. *Trans. Rough Sets* **2010**, *11*, 1–13.
13. 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, Oregon, USA 2–4 August 1996; pp. 114–119.
14. Kerber, R. ChiMerge: Discretization of Numeric Attributes. In Proceedings of the 10th National Conference on AI, IEEE International Conference on Systems, Man and Cybernetics, San Jose, CA, USA, 12–16 July 1992; pp. 123–128.
15. Kotsiantis, S.; Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *32*, 47–58.
16. Kurgan, L.A.; Cios, K.J. CAIM discretization algorithm. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 145–153.
17. Liu, H.; Hussain, F.; Tan, C.L.; Dash, M. Discretization: An enabling technique. *Data Min. Knowl. Discov.* **2002**, *6*, 393–423.
18. 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.
19. 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, Poznan-Kiekrz, Poland, 2–4 September 1998; pp. 394–401.
20. Stefanowski, J. *Algorithms of Decision Rule Induction in Data Mining*; Poznan University of Technology Press: Poznan, Poland, 2001.
21. 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.
22. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.
23. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356.
24. Pawlak, Z. *Rough Sets. Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publishers: Dordrecht, The Netherlands; Boston, MA, USA; London, UK, 1991.

25. Grzymala-Busse, J.W. A new version of the rule induction system LERS. *Fundam. Inform.* **1997**, *31*, 27–39.
26. 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.
27. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 189–209.
28. Li, H.; Yap, C.W.; Ung, C.Y.; Xue, Y.; Cao, Z.W.; Chen, Y. Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 1376–1384.
29. Keller, A.; Nesvizhskii, A.I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.

© 2013 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).