

Article

Reduced Data Sets and Entropy-Based Discretization

Jerzy W. Grzymala-Busse ^{1,2,*}, Zdzislaw S. Hippe ² and Teresa Mroczek ² 

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

² Department of Artificial Intelligence, University of Information Technology and Management, 35–225 Rzeszow, Poland; zhippe@wsiz.rzeszow.pl (Z.S.H.); tmroczek@wsiz.rzeszow.pl (T.M.)

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

Received: 13 September 2019; Accepted: 25 October 2019; Published: 28 October 2019



Abstract: Results of experiments on numerical data sets discretized using two methods—global versions of Equal Frequency per Interval and Equal Interval Width—are presented. Globalization of both methods is based on entropy. For discretized data sets left and right reducts were computed. For each discretized data set and two data sets, based, respectively, on left and right reducts, we applied ten-fold cross validation using the C4.5 decision tree generation system. Our main objective was to compare the quality of all three types of data sets in terms of an error rate. Additionally, we compared complexity of generated decision trees. We show that reduction of data sets may only increase the error rate and that the decision trees generated from reduced decision sets are not simpler than the decision trees generated from non-reduced data sets.

Keywords: data mining; numerical attributes; discretization; entropy

1. Introduction

The problem of reducing (or selecting) the set of attributes (or features) is known for many decades [1–6]. It is a central topic in multivariate statistics and data analysis [7]. It was recognized as a variant of the Set Covering problem in [3]. The problem of finding a minimal set of features is crucial for the study of medical and bioinformatics data, with tens of thousands of features representing genes [7]. An analysis of feature selection methods, presented in [7], included filter, wrapper and embedded methods, based on mathematical optimization, e.g., on linear programming. An algorithm for discretization that removes redundant attributes was presented in [8–10]. An example of a MicroArray Logic Analyzer (MALA), where feature selection was based on cluster analysis, was presented in [11]. An approach to feature selection, based on logic, was presented in [12]. A lot of attention has been paid to benefits of feature reduction in data mining, machine learning and pattern recognition, see for example, References [4–6,13–17]. Recently, feature reduction or reducing of the attribute set, combined with discretization of numerical attributes, was discussed in Reference [13,15,16,18,19]. In Reference [18], results of experiments conducted on ten numerical data sets with four different types of reducts were presented. The authors used two classifiers, Support Vector Machine (SVM) [20] and C4.5 [21]. In both cases, results of the Friedman test are inconclusive, so benefits of data reduction are not clear. In experiments presented in Reference [19], genetic algorithms and artificial neural networks were used for a few tasks: discretization, feature reduction and prediction of stock price index. It is difficult to evaluate how feature reduction contributed to final results. In Reference [13] experimental results on ten numerical data sets, discretized using a system called C-GAME, are reported. C-GAME uses reducts during discretization. The authors claim that C-GAME outperforms five other discretization schemes. Some papers, for example, Reference [14–16], discuss reducts combined with discretization. However, no convincing experimental results are included. A related

problem, namely how reduction of the attribute set as a side-effect of discretization of numerical attributes changes an error rate, was discussed in Reference [17].

For symbolic attributes, it was shown [22] that the quality of rule sets induced from reduced data sets, measured by an error rate evaluated by ten-fold cross validation, is worse than the quality of rule sets induced from the original data sets, with no reduction of the attribute set.

2. Reducts

The set of all cases of the data set is denoted by U . An example of the data set with numerical attributes is presented in Table 1. For simplicity, all attributes have repetitive values, though in real-life numerical attribute values are seldom repetitive. In our example $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The set of all attributes is denoted by A . In our example $A = \{Length, Height, Width, Weight\}$. One of the variables is called a *decision*, in Table 1 it is *Quality*.

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

Case	Attributes				Decision Quality
	Length	Height	Width	Weight	
1	4.8	1.2	1.6	0.8	high
2	4.8	1.4	1.8	0.8	high
3	4.8	1.4	1.8	0.8	high
4	4.4	1.4	1.6	1.0	medium
5	4.4	1.2	1.6	1.4	medium
6	4.2	1.2	1.8	1.4	low
7	4.2	1.8	1.8	1.4	low
8	4.2	1.8	1.8	1.4	low

Let B be a subset of the set A of all attributes. The *indiscernibility relation* $IND(B)$ [23,24] 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 $x, y \in U$ and $a(x)$ denotes the value of an attribute $a \in A$ for a case $x \in U$. The relation $IND(B)$ is an equivalence relation. An equivalence class of $IND(B)$, containing $x \in U$, is called a *B-elementary class* and is denoted by $[x]_B$. A family of all sets $[x]_B$, where $x \in U$, is a partition on U denoted by B^* . A union of *B*-elementary classes is called *B-definable*. For a decision d we may define an indiscernibility relation $IND\{d\}$ by analogy. Additionally, $\{d\}$ -elementary classes are called *concepts*.

A decision d depends on the subset B of the set A of all attributes if and only if $B^* \leq \{d\}^*$. For partitions π and τ on U , $\pi \leq \tau$ if and only if for every $Y \in \tau$ there exists $X \in \pi$ such that $X \subseteq Y$. For example, for $B = \{Width, Weight\}$, $B^* = \{\{1\}, \{2, 3\}, \{4\}, \{5\}, \{6, 7, 8\}\}$, $\{d\}^* = \{\{1, 2, 3\}, \{4\}, \{5, 6, 7, 8\}\}$ and $B^* \leq \{d\}^*$. Thus d depends on B .

For Table 1 and for $B = \{Weight\}$, $B^* = \{\{1, 2, 3\}, \{4\}, \{5, 6, 7, 8\}\}$. The concepts $\{4, 5\}$, $\{6, 7, 8\}$ are not *B*-definable. For an undefinable set X we define two definable sets, called *lower* and *upper* approximations of X [23,24]. 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 and $B = \{Weight\}$, $\underline{B}\{6, 7, 8\} = \emptyset$ and $\overline{B}\{6, 7, 8\} = \{5, 6, 7, 8\}$.

A set B is called a *reduct* if and only if B is the smallest set with $B^* \leq \{d\}^*$. The set $\{Width, Weight\}$ is the reduct since $\{Width\}^* = \{\{1, 4, 5\}, \{2, 3, 6, 7, 8\}\} \not\leq \{d\}^*$ and $\{Weight\}^* = \{\{1, 2, 3\}, \{4\}, \{5, 6, 7, 8\}\} \not\leq \{d\}^*$, so B is the smallest set with $B^* \leq \{d\}^*$.

An idea of the reduct is important since we may restrict our attention to a subset B and construct a decision tree with the same ability to distinguish all concepts that are distinguishable in the data set with the entire set A of attributes. Note that any algorithm for finding all reducts is of exponential time complexity. In practical applications, we have to use some heuristic approach. In this paper, we suggest two such heuristic approaches, left and right reducts.

A *left reduct* is defined by a process of a sequence of attempts to remove one attribute at a time, from right to left and by checking after every attempt whether $B^* \leq \{d\}^*$, where B is the current set of attributes. If this condition is true, we remove an attribute. If not, we put it back. For the example presented in Table 1, we start from an attempt to remove the rightmost attribute, that is, *Weight*. The current set B is $\{Length, Height, Width\}$, $B^* = \{\{1\}, \{2, 3\}, \{4\}, \{5\}, \{6\}, \{7, 8\}\} \leq \{d\}^*$, so we remove *Weight* for good. The next candidate for removal is *Width*, the set $B = \{Length, Height\}$, $B^* = \{\{1\}, \{2, 3\}, \{4\}, \{5\}, \{6\}, \{7, 8\}\}$ and $B^* \leq \{d\}^*$, so we remove *Width* as well. The next candidate is *Height*, if we remove it, $B = \{Length\} \leq \{d\}^*$, so $\{Length\}$ is the left reduct since it cannot be further reduced.

Similarly, a *right reduct* is defined by a similar process of a sequence of attempts to remove one attribute at a time, this time from left to right. Again, after every attempt we check whether $B^* \leq \{d\}^*$. It is not difficult to see that the right reduct is the set $\{Width, Weight\}$.

For a discretized data set we may compute left and right reducts, create three data sets: with the discretized (non-reduced) data set and with attribute sets restricted to the left and right reducts and then for all three data sets compute an error rate evaluated by C4.5 decision tree generation system using ten-fold cross validation. Our results show again that reduction of data sets causes increase of an error rate.

3. Discretization

For a numerical attribute a , let a_i be the smallest value of a and let a_j be the largest value of a . In discretizing of a we are looking for the numbers $a_{i_0}, a_{i_1}, \dots, a_{i_k}$, called *cutpoints*, where $a_{i_0} = a_i, a_{i_k} = a_j, a_{i_l} < a_{i_{l+1}}$ for $l = 0, 1, \dots, k - 1$ and k is a positive integer. As a result of discretization, the domain $[a_i, a_j]$ of the attribute a is divided into k intervals

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

In this paper we denote such intervals as follows

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

Discretization is usually conducted not on a single numerical attribute but on many numerical attributes. Discretization methods may be categorized as supervised or decision-driven (concepts are taken into account) or unsupervised. Discretization methods processing all attributes are called global or dynamic, discretization methods processing a single attribute are called local or static.

Let v be a variable and let v_1, v_2, \dots, 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 $i = 1, 2, \dots, n$. An *entropy* $H_S(v)$ is defined as follows

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

All logarithms in this paper are binary.

Let a be an attribute, let a_1, a_2, \dots, a_m be all values of a restricted to S , let d be a decision and let d_1, d_2, \dots, d_n be all values of d restricted to S , where m and n are positive integers. A conditional entropy $H_S(d|a)$ of the decision d given an 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 $a_j; j \in \{1, 2, \dots, m\}$ and $i \in \{1, 2, \dots, n\}$.

As is well-known [25–36], 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 , let a be an attribute and let 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 the set X . Usually, the cutpoint q for which $H_S(d|a)$ is the smallest is considered to be the best cutpoint.

We need how to halt discretization. Commonly, we halt discretization when we may distinguish the same cases in the discretized data set as in the original data set with numerical attributes. In this paper discretization is halted when the *level of consistency* [26], 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 $B = \{Weight\}$,

$$L(B) = \frac{|\underline{B}\{1, 2, 3\}| + |\underline{B}\{4, 5\}| + |\underline{B}\{6, 7, 8\}|}{|U|} = \frac{|\{1, 2, 3\}| + |\emptyset| + |\emptyset|}{8} = 0.375.$$

4. Equal Frequency per Interval and Equal Interval Width

Both discretization methods, Equal Frequency per Interval and Equal Interval Width, are frequently used in discretization and both are known to be efficient [25]. In local versions of these methods, only a single numerical attribute is discretized at a time [31]. The user provides a parameter denoted by k . This parameter is equal to a requested number of intervals. In the Equal Frequency per Interval method, the domain of a numerical attribute is divided into k intervals with approximately equal number of cases. In the Equal Interval Width method, the domain of a numerical attribute is divided into k intervals with approximately equal width.

In this paper we present a supervised and global version of both methods, based on entropy [26]. Using this idea, we start from discretizing all numerical attributes assuming $k = 2$. Then the level of consistency is computed for the data set with discretized attributes. If the level of consistency is sufficient, discretization ends. If not, we select the worst attribute for additional discretization. The measure of quality of the discretized attribute, denoted by a^d and called the *average block entropy*, is defined as follows

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

A discretized attribute with the largest value of $M(a^d)$ is the worst attribute. This attribute is further discretized into $k + 1$ intervals. The process is continued by recursion. The time computational complexity, in the worst case, is $O(m \cdot \log m \cdot n^2)$, where m is the number of cases and n is the number of attributes. This method is illustrated by applying the Equal Frequency per Interval method for the data set from Table 1. Table 2 presents the discretized data set for the data set from Table 1. It is not difficult to see that the level of consistency for Table 2 is 1.

For the data set presented in Table 2, both left and right reducts are equal to each other and equal to {Height^d, Width^d, Weight^d}.

Table 3 presents the data set from Table 1 discretized by the Global Equal Interval Width discretization method. Again, the level of consistency for Table 3 is equal to 1. Additionally, for the data set from Table 3, both reducts, left and right, are also equal to each other and equal to {Length^d, Width^d}.

Table 2. A data set discretized by Equal Frequency per Interval.

Case	Attributes			Weight ^d	Decision Quality
	Length ^d	Height ^d	Width ^d		
1	4.3..4.8	1.2..1.3	1.6..1.7	0.8..1.2	high
2	4.3..4.8	1.3..1.8	1.7..1.8	0.8..1.2	high
3	4.3..4.8	1.3..1.8	1.7..1.8	0.8..1.2	high
4	4.3..4.8	1.3..1.8	1.6..1.7	0.8..1.2	medium
5	4.3..4.8	1.2..1.3	1.6..1.7	1.2..1.4	medium
6	4.2..4.3	1.2..1.3	1.7..1.8	1.2..1.4	low
7	4.2..4.3	1.3..1.8	1.7..1.8	1.2..1.4	low
8	4.2..4.3	1.3..1.8	1.7..1.8	1.2..1.4	low

Table 3. A data set discretized by Equal Interval Width.

Case	Attributes			Weight ^d	Decision Quality
	Length ^d	Height ^d	Width ^d		
1	4.5..4.8	1.2..1.5	1.6..1.7	0.8..1.2	high
2	4.5..4.8	1.2..1.5	1.7..1.8	0.8..1.2	high
3	4.5..4.8	1.2..1.5	1.7..1.8	0.8..1.2	high
4	4.2..4.5	1.2..1.5	1.7..1.8	0.8..1.2	medium
5	4.2..4.5	1.2..1.5	1.7..1.8	1.2..1.4	medium
6	4.2..4.5	1.5..1.8	1.7..1.8	1.2..1.4	low
7	4.2..4.5	1.5..1.8	1.7..1.8	1.2..1.4	low
8	4.2..4.5	1.5..1.8	1.7..1.8	1.2..1.4	low

5. Experiments

We conducted experiments on 13 numerical data sets, presented in Table 4. All of these data sets may be accessed in *Machine Learning Repository*, University of California, Irvine, except for *bankruptcy*. The *bankruptcy* data set was described in Reference [37].

The main objective of our research is to compare the quality of decision trees generated by C4.5 directly from discretized data sets and from data sets based on reducts, in terms of an error rate and tree complexity. Data sets were discretized by the Global Equal Frequency per Interval and Global Equal Interval Width methods with the level of complexity equal to 1. For each numerical data set three data sets were considered:

- an original (non-reduced) discretized data set,
- a data set based on the left reduct of the original discretized data set and
- a data set based on right reduct of the original discretized data set.

The discretized data sets were inputted to the C4.5 decision tree generating system [21]. In our experiments, the error rate was computed using an internal mechanism of the ten-fold cross validation of C4.5.

Additionally, an internal discretization mechanism of C4.5 was excluded in experiments for left and right reduces since in this case data sets were discretized by the global discretization methods, so C4.5 considered all attributes as symbolic.

We illustrate our results with Figures 1 and 2. Figure 1 presents discretization intervals for *yeast* data set, where discretization was conducted by the internal discretization mechanism of C4.5. Figure 2 presents discretization intervals for the same data set with discretization conducted by the global version of the Equal Frequency per Interval method (right reduces and left reduces were identical).

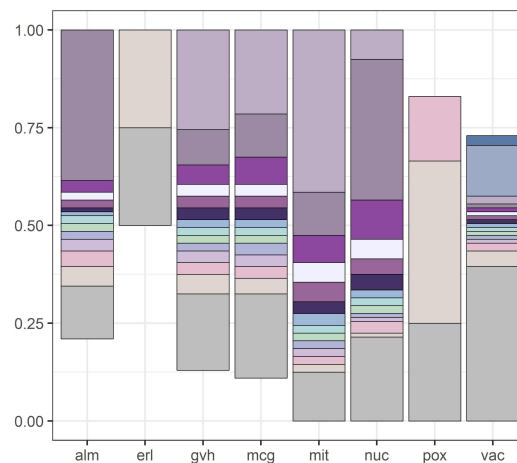


Figure 1. Attribute intervals for the *yeast* data set discretized by the internal discretization mechanism of C4.5.

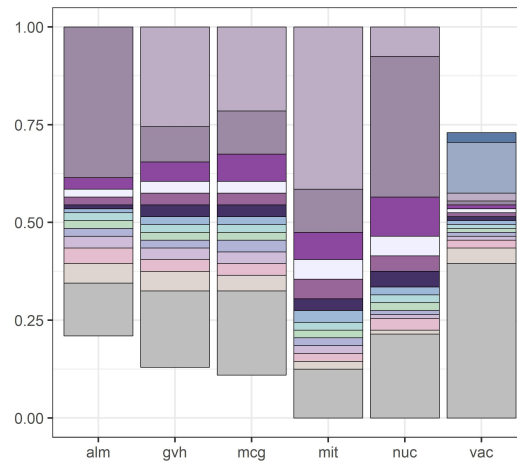


Figure 2. Attribute intervals for the *yeast* data set discretized by the global version of Equal Frequency per Interval method and then computing right reduces.

Results of our experiments are presented in Tables 5–8. These results were analyzed by the Friedman rank sum test with multiple comparisons, with 5% level of significance. For data sets discretized by the Global Equal Frequency per Interval method, the Friedman test shows that there are significant differences between the three types of data sets: the non-reduced discretized data sets and data sets based on left and right reduces. In most cases, the original, non-reduced data sets are associated with the smallest error rates than both left and right reduces. However, the test of multiple comparisons shows that the differences are not statistically significant.

For data sets discretized by the Global Equal Interval Width method results are more conclusive. There are statistically significant differences between non-reduced discretized data sets and data sets based on left and right reduces. Moreover, an error rate for the non-reduced discretized data sets is

significantly smaller than for both types of data sets, based on left and right reducts. As expected, the difference between left and right reducts is not significant.

For both discretization methods and all types of data sets (non-reduced, based on left and right reducts) the difference in complexity of generated decision trees, measured by tree size and depth, is not significant.

Table 8 shows the size of left and right reducts created from data sets discretized by the Global versions of Equal Frequency per Interval and Equal Interval Width methods. For some data sets, for example, for *bupa*, both left and right reducts are identical with the original attribute set.

Table 4. Data sets.

Data Set	Number of		
	Cases	Attributes	Concepts
Abalone	4177	8	29
Australian	690	14	2
Bankruptcy	66	5	2
Bupa	345	6	2
Echocardiogram	74	7	2
Ecoli	336	8	8
Glass	214	9	6
Ionosphere	351	34	2
Iris	150	4	3
Leukemia	415	175	2
Wave	512	21	3
Wine Recognition	178	13	3
Yeast	1484	8	9

Table 5. Error rates for discretized data sets.

Data Set	Equal Frequency per Interval			Equal Interval Width		
	No Reduction	Left Reducts	Right Reducts	No Reduction	Left Reducts	Right Reducts
Abalone	76.87	76.99	77.47	76.90	77.42	77.42
Australian	12.46	14.49	22.46	13.33	30.14	14.64
Bankruptcy	3.03	3.03	3.03	10.61	10.61	10.61
Bupa	35.94	35.94	35.94	34.49	34.49	34.49
Echocardiogram	27.03	27.03	27.03	31.08	39.19	39.19
Ecoli	30.65	28.87	28.87	28.57	28.57	28.57
Glass	41.12	39.25	41.59	33.18	33.64	33.64
Ionosphere	13.11	19.15	20.85	10.83	11.97	15.82
Iris	12.67	12.67	12.67	4.00	4.00	4.00
Leukemia	1.45	2.93	1.60	1.32	1.99	1.59
Wave	25.59	27.15	28.91	27.54	31.46	28.13
Wine Recognition	10.11	10.67	12.36	9.55	8.99	10.67
Yeast	57.82	57.82	57.82	56.54	56.87	56.87

Table 6. Tree size for discretized data sets.

Data Set	Tree Size					
	Equal Frequency per Interval			Equal Interval Width		
	No Reduction	Left Reducts	Right Reducts	No Reduction	Left Reducts	Right Reducts
Abalone	28,236	27,202	24,905	18,711	16,491	16,491
Australian	41	3	12	39	95	3
Bankruptcy	3	3	3	6	6	6
Bupa	17	17	17	27	27	27
Echocardiogram	8	13	13	16	16	16
Ecoli	61	40	40	109	107	107
Glass	70	63	56	186	191	191
Ionosphere	33	53	71	34	24	70
Iris	11	11	11	4	4	4
Leukemia	139	125	132	229	139	174
Wave	62	42	44	107	73	96
Wine Recognition	19	30	30	18	15	18
Yeast	662	678	678	913	914	914

Table 7. Tree depth for discretized data sets.

Data Set	Tree Depth					
	Equal Frequency per Interval			Equal Interval Width		
	No Reduction	Left Reducts	Right Reducts	No Reduction	Left Reducts	Right Reducts
Abalone	2	2	2	3	3	3
Australian	5	1	2	5	4	1
Bankruptcy	1	1	1	1	1	1
Bupa	2	2	2	2	2	2
Echocardiogram	2	2	2	3	3	3
Ecoli	3	2	2	3	3	3
Glass	5	4	5	5	5	5
Ionosphere	6	6	5	6	4	6
Iris	2	2	2	1	1	1
Leukemia	6	4	4	6	4	5
Wave	9	7	9	11	9	6
Wine Recognition	6	6	6	5	4	5
Yeast	4	4	4	3	3	3

Table 8. Number of attributes in left and right reducts for discretized data sets.

Data Set	Discretized Non-Reduced Data Set	Equal Frequency per Interval		Equal Interval Width	
		Left Reducts	Right Reducts	Left Reducts	Right Reducts
Abalone	8	6	6	6	6
Australian	14	8	10	9	9
Bankruptcy	5	4	4	4	4
Bupa	6	6	6	6	6
Echocardiogram	7	5	5	6	6
Ecoli	7	5	5	5	5
Glass	9	8	8	7	7
Ionosphere	34	11	13	11	10
Iris	4	4	4	4	4
Leukemia	15	8	8	11	10
Wave	21	15	14	14	14
Wine Recognition	13	8	8	10	10
Yeast	8	6	6	6	6

6. Conclusions

Our preliminary results [22] show that data reduction combined with rule induction causes an increase of the error rate. The current results, presented in this paper, confirm these results: the reduction of data sets, associated with C4.5 tree generation system, causes the same effect. Decision trees generated from reduced data sets increase the error rate as evaluated by ten-fold cross validation. Additionally, decision trees generated from reduced data sets, in terms of a tree size or tree depth, are not simpler than decision trees generated from non-reduced data sets. Therefore, it is obvious that reduction of data sets (or feature selection) should be used with caution since it may degrade results of data mining.

In the future we are planning to extend our experiments to large data sets and to include other classifiers than systems for rule induction and decision tree generation.

Author Contributions: T.M. designed and conducted experiments, Z.S.H. validated results and J.W.G.-B. 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. Almuallim, H.; Dietterich, T.G. Learning Boolean concepts in the presence of many irrelevant features. *Artif. Intell.* **1994**, *69*, 279–305. [[CrossRef](#)]
2. Kira, K.; Rendell, L.A. The feature selection problem: Traditional methods and a new algorithm. In Proceedings of the 10-th National Conference on AI, San Jose, CA, USA, 12–16 July 1992; pp. 129–134.
3. Garey, M.; Johnson, D. *Computers and Intractability: A Guide to the Theory of NP-Completeness*; W. H. Freeman: New York, NY, USA, 1979.
4. Fuernkranz, J.; Gamberger, D.; Lavrac, N. *Foundations of Rule Learning*; Springer: Berlin/Heidelberg, Germany, 2012.

5. *Feature Selection for Data and Pattern Recognition*; Stanczyk, U., Jain, L.C., Eds.; Springer: Berlin/Heidelberg, Germany, 2015.
6. *Advances in Feature Selection for Data and Pattern Recognition*; Stanczyk, U., Zielosko, B., Jain, L.C., Eds.; Springer: Cham, Switzerland, 2018.
7. Bertolazzi, P.; Felici, G.; Festa, P.; Ficon, G.; Weitschek, E. Integer programming models for feature selection: New extensions and a randomized solution algorithm. *Inf. Sci.* **2016**, *250*, 389–399.
8. Santoni, D.; Weitschek, E.; Felici, G. Optimal discretization and selection of features by association rates of joint distributions. *RAIRO-Oper. Res.* **2016**, *50*, 437–449. [[CrossRef](#)]
9. Liu, H.; Rudy, S. Feature selection via discretization. *IEEE Trans. Knowl. Data Eng.* **1997**, *9*, 642–646.
10. Sharmin, S.; Ali, A.A.; Khan, M.A.H.; Shoyaib, M. Feature selection and discretization based on mutual information. In Proceedings of the IEEE International Conference on Imaging, Vision & Pattern Recognition, Dhaka, Bangladesh, 13–14 February 2017; pp. 1–6.
11. Weitschek, E.; Felici, G.; Bertolazzi, P. MALA: A microarray clustering and classification software. In Proceedings of the International Workshop on Database and Expert Systems Applications, Vienna, Austria, 3–6 September 2012; pp. 201–205.
12. Felici, G.; Weitschek, E. Mining logic models in the presence of noisy data. In Proceedings of the International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, USA, 9–11 January 2012.
13. Tian, D.; Zeng, X.J.; Keane, J. Core-generating approximated minimum entropy discretization for rough set feature selection in pattern classification. *Int. J. Approx. Reason.* **2011**, *52*, 863–880. [[CrossRef](#)]
14. Jensen, R.; Shen, Q. Fuzzy-rough sets for descriptive dimensionality reduction. In Proceedings of the International Conference on Fuzzy Systems FUZZ-IEEE 2002, Honolulu, HI, USA, 12–17 May 2002; pp. 29–34.
15. Nguyen, H.S. Discretization problem for rough sets methods. In Proceedings of the 1-st International Conference RSCTC 1998 on Rough Sets and Current Trends in Computing, Warsaw, Poland, 22–26 June 1998; Springer-Verlag: Berlin/Heidelberg, Germany, 1998; pp. 545–552.
16. Swiniarski, R.W. Rough set methods in feature reduction and classification. *Int. J. Appl. Math. Comput. Sci.* **2001**, *11*, 656–582.
17. Grzymala-Busse, J.W.; Mroczek, T. Attribute selection based on reduction of numerical attribute during discretization. In *Advances in Feature Selection for Data and Pattern Recognition*; Stanczyk, B., Zielosko, B., Jain, L.C., Eds.; Springer International Publishing AG: Cham, Switzerland, 2017; pp. 13–24.
18. Hu, Q.; Yu, D.; Xie, Z. Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognit. Lett.* **2006**, *27*, 414–423. [[CrossRef](#)]
19. Kim, K.j.; Han, I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Syst. Appl.* **2000**, *19*, 125–132. [[CrossRef](#)]
20. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
21. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.
22. Grzymala-Busse, J.W. An empirical comparison of rule induction using feature selection with the LEM2 algorithm. In *Communications in Computer and Information Science*; Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R.R., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 297, pp. 270–279.
23. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [[CrossRef](#)]
24. Pawlak, Z. *Rough Sets. Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publishers: Dordrecht, The Netherlands; Boston, MA, USA; London, UK, 1991.
25. 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.
26. 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](#)]
27. 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](#)]
28. Elomaa, T.; Rousu, J. General and efficient multisplitting of numerical attributes. *Mach. Learn.* **1999**, *36*, 201–244. [[CrossRef](#)]
29. Fayyad, U.M.; Irani, K.B. On the handling of continuous-valued attributes in decision tree generation. *Mach. Learn.* **1992**, *8*, 87–102. [[CrossRef](#)]

30. Fayyad, U.M.; Irani, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Chambéry, France, 28 August–3 September 1993; pp. 1022–1027.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. Stefanowski, J. *Algorithms of Decision Rule Induction in Data Mining*; Poznan University of Technology Press: Poznan, Poland, 2001.
37. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).