A Comparison of Two Approaches to Discretization: Multiple Scanning and C4.5

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

Abstract. In a Multiple Scanning discretization technique the entire attribute set is scanned many times. During every scan, the best cut-point is selected for all attributes. The main objective of this paper is to compare the quality of two setups: the Multiple Scanning discretization technique combined with the C4.5 classification system and the internal discretization technique of C4.5. Our results show that the Multiple Scanning discretization technique is significantly better than the internal discretization used in C4.5 in terms of an error rate computed by ten-fold cross validation (two-tailed test, 5 % level of significance). Additionally, the Multiple Scanning discretization technique is significantly better than a variant of discretization based on conditional entropy introduced by Fayyad and Irani called Dominant Attribute. At the same time, decision trees generated from data discretized by Multiple Scanning are significantly simpler from decision trees generated directly by C4.5 from the same data sets.

1 Introduction

Mining numerical data sets requires an additional step called *discretization*. Discretization is a process of transforming numerical values into intervals.

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

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

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

A new discretization technique, called *Multiple Scanning*, introduced in [11,12], was very successful when combined with rule induction and a classification system of LERS (Learning from Examples based on Rough Sets) [9]. The novelty of this paper is a comparison of the C4.5 classification system applied to data discretized using Multiple Scanning with C4.5 applied directly to the original data sets with numeric attributes. Additionally, we compare the Multiple

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 44–53, 2015.
DOI: 10.1007/978-3-319-19941-2_5

Scanning discretization technique with a variant of the well-known discretization based on conditional entropy introduced by Fayyad and Irani [7,8] and called *Dominant Attribute* [11,12].

In Multiple Scanning, during every scan, the entire attribute set is analyzed. For all attributes the best cutpoint is selected. At the end of a scan, some subtables that still need discretization are created. The entire attribute set of any subtable is scanned again, and the best corresponding cutpoints are selected. The process continues until the stopping condition is satisfied or the required number of scans is reached. If the required number of scans is reached and the stopping condition is not satisfied, discretization is completed by Dominant Attribute, in which first the best attribute is selected, then for this attribute, the best cutpoint, again using conditional entropy, is selected. This process continues recursively until the same stopping criterion is satisfied. Multiple Scanning ends up with an attempt to reduce the number of intervals called merging. Since Multiple Scanning uses Dominant Attribute as the last resort, if we skip scanning, or equivalently set the required number of scans to zero, discretization is reduced to Dominant Attribute. Thus we may include a comparison of Multiple Scanning with Dominant Attribute. Typically, in Multiple Scanning the required number of scans should be set to some small number. In our experiments, for all data sets, after six scans the error rate computed using ten-fold cross validation was constant, because new intervals created in consecutive scans were merged together during the last step of discretization. The stopping criterion used in this paper is based on rough set theory.

The main objective of this paper is to compare the quality of two setups: the Multiple Scanning discretization technique combined with the C4.5 classification system and the internal discretization technique of C4.5. For 12 numerical data sets two sets of experiments were conducted: first the C4.5 system of tree induction was used to compute an error rate using ten-fold cross validation, then the same data sets were discretized using Multiple Scanning and for such discretized data sets the same C4.5 system was used to establish an error rate. Thus we may compare two discretization techniques: Multiple Scanning with the internal discretization of C4.5.

Our results show that the Multiple Scanning discretization technique is significantly better than the internal discretization used in C4.5 or the Dominant Attribute discretization 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 from the same data sets.

2 Entropy Based Discretization

Discretization based on conditional entropy of the concept given the attribute is considered to be one of the most successful discretization techniques [2–8,10, 11,13–15,19,20].

An example of a data set 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 *Max_Speed* and *Number_of_Seats* while the decision is *Price*. 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 data set with numerical attributes

Case	Attributes		Decision Price
	Max_Speed	Number_of_Seats	
1	280	2	very-high
2	220	4	small
3	180	5	small
4	220	5	medium
5	220	2	high
6	280	4	medium
7	180	4	small

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

$$I_S(a) = H_S(d) - H_S(d|a)$$

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

$$G_S(a) = \frac{I_S(a)}{H_S(a)}.$$

Both criteria were introduced by J.R. Quinlan, see, e.g., [18] and used for decision tree generation.

Let a be an attribute and q be a cutpoint 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. 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* [16,17]. 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 $B = \{Max\_Speed\}$, B^* = {{1, 6}, {2, 4, 5}, {3, 7}} and $\{d\}^*$ = {{1}, {2, 3, 7}, {4, 6}, {5}}. In general, arbitrary $X \in \{d\}^*$ is not B-definable. For example, the concept {2, 3, 7} 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, $\underline{B}\{2, 3, 7\} = \{3, 7\}$ and $\overline{B}\{2, 3, 7\} = \{2, 3, 4, 5, 7\}$.

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* [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 1.0, since {A}* = {{1}, {2}, {3}, {4}, {5}, {6}, {7}} and, for any X from $\{Price\}^*$ = {{1}, {2, 3, 7}, {4, 6}, {5}}}, we have $\underline{A}X = X$. Additionally, $L(B) \approx 0.286$.

Table 2. Partially discretized Table 1 using Multiple Scanning

Case	Attributes		Decision Price
	Max_Speed	Number_of_Seats	
1	[280, 280]	[2, 4)	very-high
2	[180, 280)	[4, 5]	small
3	[180, 280)	[4, 5]	small
4	[180, 280)	[4, 5]	medium
5	[180, 280)	[2, 4)	high
6	[280, 280]	[4, 5]	medium
7	[180, 280)	[4, 5]	small

2.2 Multiple Scanning Strategy

This 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 cutpoint is computed for each attribute $a \in A$, based on minimum of conditional entropy $H_U(d|a)$, a new discretized attribute set is A^D, and the set U is partitioned into a partition $(A^D)^*$,
- if the number t of scans 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 cutpoint, the best cutpoint for each block $X \in (A^D)^*$ is computed, the best cutpoint, among all such blocks is selected,
- if the requested number t of scans is reached and the data set needs more discretization, the Dominant Attribute technique is used for remaining subtables,
- the algorithm stops when $L(A^D) = 1$, where A^D is the discretized set of attributes.

We illustrate this technique by scanning Table 1 once, i.e., $t = 1$. First we are searching for the best cut-point for both attributes, *Max_Speed* and *Number_of_Seats*. For the attribute *Max_Speed* there exist two potential cutpoints: 220 and 280 with three potential intervals: [180, 220), [220, 280) and [280, 280]. The corresponding conditional entropies are

$$H_{Max\_Speed}(220, U) = \frac{5}{7}((-\frac{1}{5} \cdot \log \frac{1}{5})(3) + (-\frac{2}{5} \cdot \log \frac{2}{5})) + \frac{2}{7}(0) \approx 1.373,$$

$$H_{Max\_Speed}(280, U) = \frac{5}{7}((-\frac{1}{5} \cdot \log \frac{1}{5})(2) + (-\frac{3}{5} \cdot \log \frac{3}{5})) + \frac{2}{7}(1) \approx 1.251.$$

The better cutpoint is 280. Similarly, there are three potential cutpoints for the attribute *Number_of_Seats*: 4 and 5, with three potential intervals: [2, 4), [4, 5) and [5, 5]. The corresponding conditional entropies are

$$H_{Number\_of\_Seats}(4, U) = \frac{5}{7}((-\frac{3}{5} \cdot \log \frac{3}{5}) + (-\frac{2}{5} \cdot \log \frac{2}{5})) + \frac{2}{7}(1) \approx 0.979,$$

$$H_{Number\_of\_Seats}(5, U) = \frac{5}{7}((-\frac{1}{5} \cdot \log \frac{1}{5})(3) + (-\frac{2}{5} \cdot \log \frac{2}{5})) + \frac{2}{7}(1) \approx 1.229.$$

The better cut-point is 4. Table 1, partially discretized this way, is presented as Table 2.

The level of consistency for Table 2 is 0.429 since $A^* = \{\{1\}, \{2, 3, 4, 7\}, \{5\}, \{6\}\}$, we need to distinguish cases 2, 3, and 7 from the case 4. Therefore we need to use the *Dominant Attribute* technique for a subtable, with four cases, 2, 3, 4 and 7. This data set is presented in Table 3.

Table 3. The remaining data set that still needs discretization

Case	Attributes		Decision Price
	Max_Speed	Number_of_Seats	
2	220	4	small
3	180	5	small
4	220	5	medium
7	180	4	small

3 Experiments

Our experiments were conducted on 12 data sets available on the University of California at Irvine *Machine Learning Repository*, with the exception of *bankruptcy*. The *bankruptcy data* set is a well-known data set used by E.I. Altman to predict a bankruptcy of companies [1].

Both discretization methods, Multiple Scanning and C4.5, were applied to all data sets, with the level of consistency equal to 100 %. For a choice of the best attribute, we used gain ratio.

Table 4 presents results of ten-fold cross validation, using increasing number of scans. Obviously, for any data set, after some fixed number of scans, an error rate is stable (constant). For example, for *Australian* data set, the error rate is 14.93 % for the scan number 4, 5, etc. Thus, any data set from Table 4 is characterized by two error rates: minimal and stable [12]. For a given data set, the smallest error rate from Table 4 is called *minimal* and the last entry in the row that corresponds to the data set is called *stable*. For example, for the *Australian* data set, the minimal error rate is 13.48 % and the stable error rate is 14.93 %. For some data sets (e.g., for *bankruptcy*), minimal and stable error rates are identical.

Table 5 presents the size of decision trees generated from all 12 data sets discretized by Multiple Scanning. In Table 6 error rates are shown for decision trees generated directly by C4.5 and for the decision trees generated by C4.5 from data sets discretized by Multiple Scanning, only the minimal error rates are presented with the corresponding scan numbers. Finally, Table 7 presents tree

Table 4. Error rates for Multiple Scanning

Data set	Error rate for scan number						
	0	1	2	3	4	5	6
Australian	14.49	13.48	13.77	14.93			
Bankruptcy	10.61	3.03	3.03				
Bupa	41.74	29.86	30.43	29.28	29.86		
Connectionist bench	27.89	16.83					
Echocardiogram	27.03	14.86	14.86	24.32	22.97		
Glass	34.11	30.84	28.50	24.77	25.23	27.10	26.64
Image segmentation	13.81	18.10	11.90	12.38			
Iris	5.33	5.33	4.67				
Pima	27.73	25.78	24.09	24.61	25.00	26.17	
Wave	32.81	23.05	26.17	24.80			
Wine recognition	7.87	3.93					
Yeast	56.40	53.84	54.65	51.75	51.75	51.75	

Table 5. Tree size for Multiple Scanning

Data set	Tree size for scan number						
	0	1	2	3	4	5	6
Australian	3	13	26	27			
Bankruptcy	14	3	4				
Bupa	13	10	9	11	20		
Connectionist bench	6	31					
Echocardiogram	13	5	10	7	7		
Glass	126	72	67	58	40	40	40
Image segmentation	16	33	24	24			
Iris	6	4	4				
Pima	73	34	27	44	49	48	
Wave	7	55	94	105			
Wine recognition	8	11					
Yeast	414	276	491	362	458	442	

size for decision trees generated directly by C4.5 and for decision trees generated by C4.5 from data sets discretized by Multiple Scanning.

It is clear from Tables 4–7 that the minimal error rate is never associated with 0 scans, i.e., with a special case of the Multiple Scanning discretization technique: Dominant Attribute. Using the Wilcoxon matched-pairs signed-ranks

Table 6. Error rates for C4.5 and the best results of Multiple Scanning

Data set	C4.5	Multiple Scanning	
	Error rate	Error rate	Scan number
Australian	16.09	13.48	1
Bankruptcy	6.06	3.03	1
Bupa	35.36	29.28	3
Connectionist bench	25.96	16.83	1
Echocardiogram	28.38	14.86	1
Glass	33.18	24.77	3
Image segmentation	12.38	11.90	2
Iris	5.33	4.67	2
Pima	25.13	24.09	2
Wave	26.37	23.05	1
Wine recognition	8.99	3.93	1
Yeast	44.41	51.75	3

Table 7. Tree size for C4.5 and the best results of Multiple Scanning

Data set	C4.5	Multiple Scanning
Australian	63	13
Bankruptcy	3	3
Bupa	51	11
Connectionist bench	35	31
Echocardiogram	9	5
Glass	45	58
Image segmentation	25	24
Iris	9	4
Pima	43	27
Wave	85	55
Wine recognition	9	11
Yeast	371	362

test, we conclude that the following three statements are statistically significant (with the significance level equal to 5 % 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 error rate associated with C4.5,
– the size of decision trees generated from data discretized by Multiple Scanning is smaller than size of decision trees generated directly by C4.5.

4 Conclusions

This paper presents results of experiments in which three different techniques were used for discretization: Multiple Scanning, the internal discretization of C4.5, and Dominant Attribute. All techniques were validated by conducting experiments on 12 data sets with numerical attributes. Our discretization techniques were combined with decision tree generation using the C4.5 system. Results of our experiments show that the Multiple Scanning technique is significantly better than discretization included in C4.5 and that decision trees generated from data discretized by Multiple Scanning are significantly simpler than decision trees generated directly by C4.5 from the same data sets (two-tailed test and 0.05 level of significance). Additionally, the Multiple Scanning discretization technique is significantly better than the Dominant Attribute technique. Thus, we show that there exists a new successful technique for discretization.

References

1. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Financ. **23**(4), 189–209 (1968)
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: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 31–38. Springer, Heidelberg (2008)
3. Chmielewski, M.R., Grzymala-Busse, J.W.: Global discretization of continuous attributes as preprocessing for machine learning. Int. J. Approximate Reasoning **15**(4), 319–331 (1996)
4. Clarke, E.J., Barton, B.A.: Entropy and MDL discretization of continuous variables for bayesian belief networks. Int. J. Intell. Syst. **15**, 61–92 (2000)
5. Elomaa, T., Rousu, J.: General and efficient multisplitting of numerical attributes. Mach. Learn. **36**, 201–244 (1999)
6. Elomaa, T., Rousu, J.: Efficient multisplitting revisited: optima-preserving elimination of partition candidates. Data Min. Knowl. Disc. **8**, 97–126 (2004)
7. Fayyad, U.M., Irani, K.B.: On the handling of continuous-valued attributes in decision tree generation. Mach. Learn. **8**, 87–102 (1992)
8. 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, pp. 1022–1027 (1993)
9. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundamenta Informaticae **31**, 27–39 (1997)
10. Grzymala-Busse, J.W.: Discretization of numerical attributes. In: Kloesgen, W., Zytkow, J. (eds.) Handbook of Data Mining and Knowledge Discovery, pp. 218–225. Oxford University Press, New York (2002)
11. Grzymala-Busse, J.W.: A multiple scanning strategy for entropy based discretization. In: Proceedings of the 18th International Symposium on Methodologies for Intelligent Systems, pp. 25–34 (2009)
12. Grzymala-Busse, J.W.: Discretization based on entropy and multiple scanning. Entropy **15**, 1486–1502 (2013)

13. Kerber, R.: Chimerge: discretization of numeric attributes. In: Proceedings of the 10-th National Conference on AI, pp. 123–128 (1992)
14. 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, pp. 114–119 (1996)
15. Nguyen, H.S., Nguyen, S.H.: Discretization methods in data mining. In: Polkowski, L., Skowron, A. (eds.) Rough Sets in Knowledge Discovery 1: Methodology and Applications, pp. 451–482. Physica-Verlag, Heidelberg (1998)
16. Pawlak, Z.: Rough sets. Int. J. Comput. Inform. Sci. **11**, 341–356 (1982)
17. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
18. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
19. Stefanowski, J.: Handling continuous attributes in discovery of strong decision rules. In: Polkowski, L., Skowron, A. (eds.) RSCTC 1998. LNCS (LNAI), vol. 1424, pp. 394–401. Springer, Heidelberg (1998)
20. Stefanowski, J.: Algorithms of Decision Rule Induction in Data Mining. Poznan University of Technology Press, Poznan (2001)