Inducing Better Rule Sets by Adding Missing Attribute Values

Jerzy W. Grzymala-Busse[1,2] and Witold J. Grzymala-Busse[3]

[1] Department of Electrical Engineering and Computer Science University of Kansas,
Lawrence, KS 66045, USA
[2] Institute of Computer Science, Polish Academy of Sciences,
01–237 Warsaw, Poland
[3] Touchnet Information Systems, Inc.,
Lenexa, KS 66219, USA

Abstract. Our main objective was to verify the following hypothesis: for some complete (i.e., without missing attribute vales) data sets it is possible to induce better rule sets (in terms of an error rate) by increasing incompleteness (i.e., removing some existing attribute values) of the original data sets. In this paper we present detailed results of experiments on one data set, showing that some rule sets induced from incomplete data sets are significantly better than the rule set induced from the original data set, with the significance level of 5%, two-tailed test. Additionally, we discuss criteria for inducing better rules by increasing incompleteness and present graphs for some well-known data sets.

1 Introduction

In this paper we show that by increasing incompleteness of data sets (i.e., by removing attribute values in a data set) we may improve quality of the rule sets induced from such modified data sets. In our experiments we replaced randomly existing attribute values in the original data sets by symbols that were recognized by the rule induction module as missing attribute values. In other words, the rule sets were induced from data sets in which some values were erased using a Monte Carlo method. The process of such replacements was done incrementally, with an increment equal to 5% of the total number of attribute values of a given data set.

We distinguish three different kinds of missing attribute values: *lost values* (the values that were recorded but currently are unavailable) [1,2,3,4], *attribute-concept values* (these missing attribute values may be replaced by any attribute value limited to the same concept) [5], and *"do not care" conditions* (the original values were irrelevant) [4,6,7,8]. A *concept* (class) is a set of all cases classified (or diagnosed) the same way.

We assumed that for each case at least one attribute value was specified, i.e., they are not missing. Such an assumption limits the percentage of missing attribute values used for experiments; for example, for the *wine* data set, starting

C.-C. Chan et al. (Eds.): RSCTC 2008, LNAI 5306, pp. 160–169, 2008.

Table 1. Data sets used for experiments

Data set	Number of		
	cases	attributes	concepts
Bankruptcy	66	5	2
Breast cancer - Slovenia	277	9	2
Hepatitis	155	19	2
Image segmentation	210	19	7
Iris	150	4	3
Lymphography	148	18	4
Wine	178	12	3

from 70% of randomly assigned missing attribute values, this assumption was violated. Additionally, we assumed that all decision values were specified.

For rule induction from incomplete data we used the MLEM2 data mining algorithm, for details see [9]. We used rough set methodology [10,11], i.e., for a given interpretation of missing attribute vales, *lower* and *upper approximations* were computed for all concepts and then rule sets were induced, *certain* rules from lower approximations and *possible* rules from upper approximations. Note that for incomplete data there is a few possible ways to define approximations, we used *concept* approximations [4,5].

As follows from our experiments, some of the rule sets induced from such incomplete data are better than the rule sets induced form original, complete data sets. More precisely, the error rate, a result of ten-fold cross validation, is significantly lower, with the significance level of 5%, than the error rate for rule sets induced from the original data.

2 Experiments

In our experiments seven typical data sets were used, see Table 1. All of these data sets are available from the UCI ML Repository, with the exception of the *bankruptcy* data set. These data sets were completely specified (all attribute values were completely specified), with the exception of *breast cancer - Slovenia* data set, which originally contained 11 cases (out of 286) with missing attribute values. These 11 cases were removed.

In two data sets: *bankruptcy* and *iris* all attributes were numerical. These data sets were processed as numerical (i.e., discretization was done during rule induction by MLEM2). The *image segmentation* data set was converted into symbolic using a discretization method based on agglomerative cluster analysis (this method was described, e.g., in [12]).

Preliminary results [13] show that, for some data sets by increasing incompleteness we may improve rule sets. Therefore we decided to conduct extensive

Table 2. Wine data set. Certain rule sets.

Percentage of lost values	Average error rate	Standard deviation	Z score
0	7.66	1.32	
5	7.17	1.74	1.22
10	7.13	2.00	1.20
15	8.76	1.85	−2.66
20	7.06	1.38	1.72
25	7.27	1.55	1.06
30	6.20	1.39	**4.17**
35	6.55	1.16	**3.43**
40	6.8	1.28	**2.56**
45	7.73	1.48	−0.21
50	7.21	0.82	1.58
55	8.01	1.29	−1.05
60	7.30	1.00	1.00
65	8.41	0.98	0.98

Table 3. Wine data set. Possible rule sets.

Percentage of lost values	Average error rate	Standard deviation	Z score
0	7.66	1.32	
5	7.21	1.92	1.06
10	7.32	1.34	0.98
15	8.46	1.75	−2.01
20	7.17	1.72	1.23
25	7.64	1.63	0.05
30	6.33	1.15	**4.15**
35	6.57	1.12	**3.44**
40	6.22	1.29	**4.27**
45	7.79	1.30	−0.39
50	7.12	0.68	**2.00**
55	7.68	0.98	−0.06
60	6.89	0.78	**2.74**
65	8.31	1.17	−2.04

Table 4. Wine data set. Size of rule sets.

Percentage of	Certain rule set		Possible rule set	
lost values	Number of		Number of	
lost values	rules	conditions	rules	conditions
0	20	65	20	65
10	25	89	21	73
20	34	108	28	90
30	38	117	46	149
40	47	140	54	166
50	62	246	70	204
60	59	156	61	148

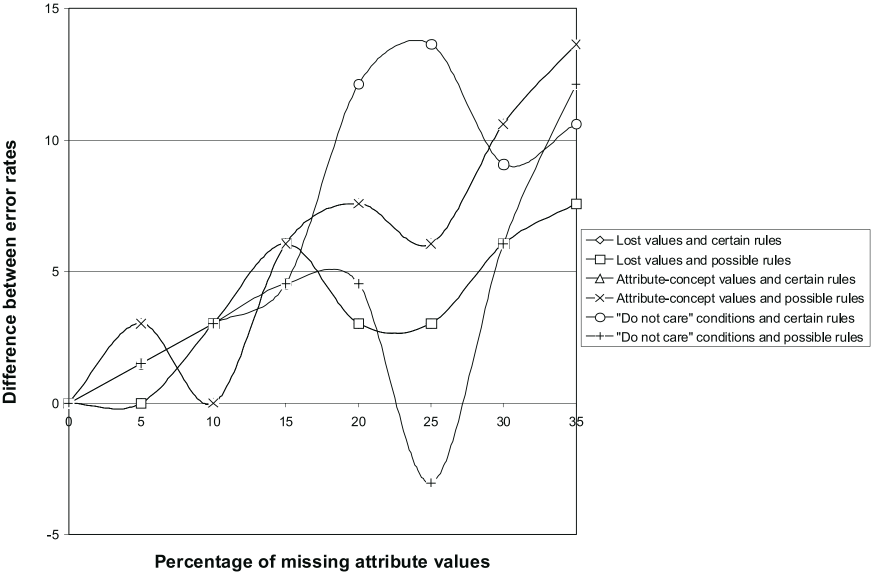

Fig. 1. Bankruptcy data set. Difference between error rates for testing with complete data sets and data sets with missing attribute values.

experiments, repeating 30 times the ten-fold cross validation experiment (changing the random case ordering in data sets) for every percentage of lost values and then computing the Z score using the well-known formula

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2+s_2^2}{30}}},$$

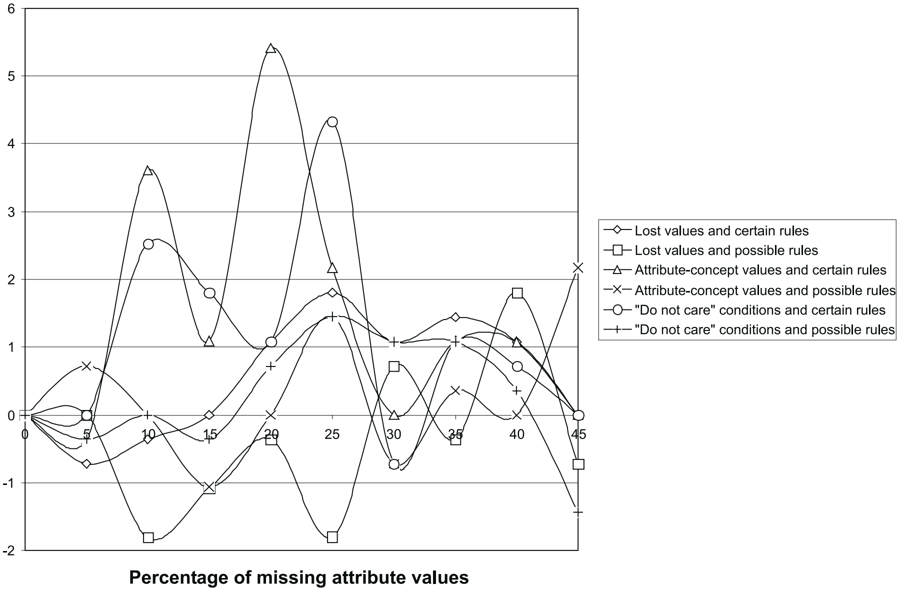

Fig. 2. Breast cancer - Slovenia data set. Difference between error rates for testing with complete data sets and data sets with missing attribute values.

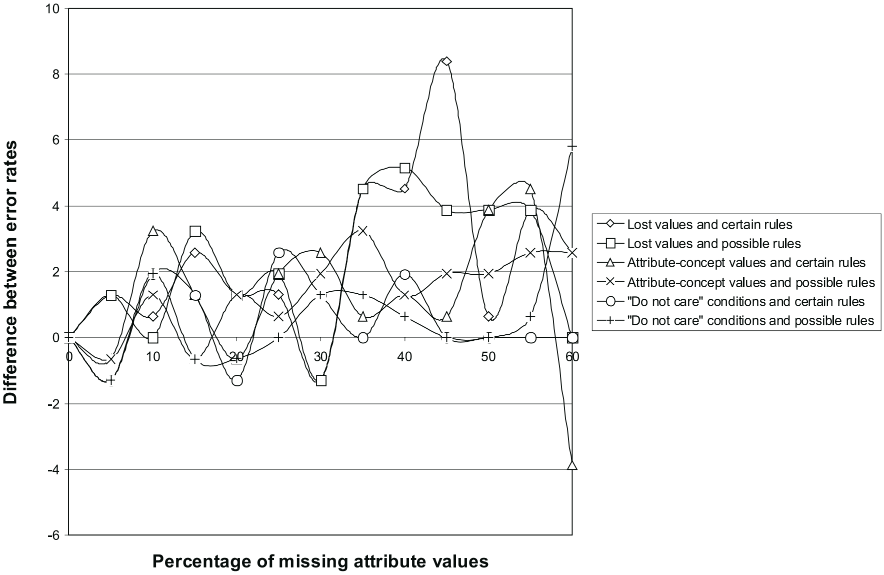

Fig. 3. Hepatitis data set. Difference between error rates for testing with complete data sets and data sets with missing attribute values.

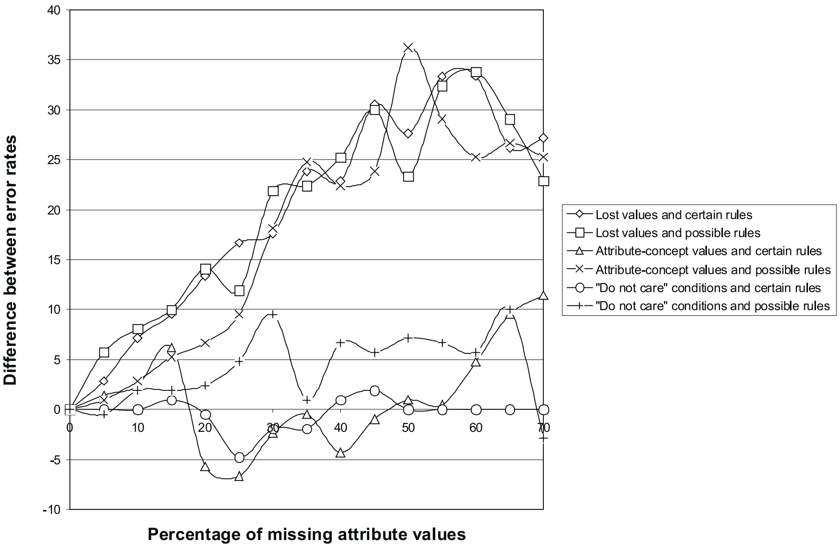

Fig. 4. Image segmentation data set. Difference between error rates for testing with complete data sets and data sets with missing attribute values.

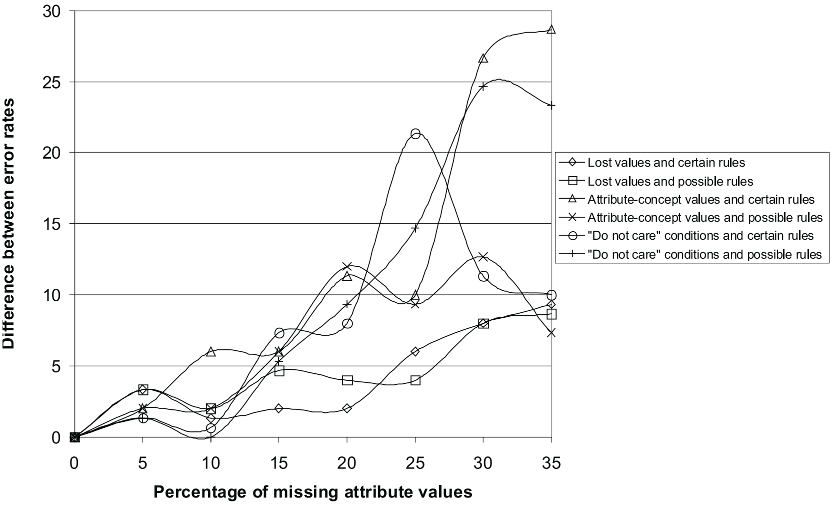

Fig. 5. Iris data set. Difference between error rates for testing with complete data sets and data sets with missing attribute values.

where $\overline{X_1}$ is the mean of 30 ten-fold cross validation experiments for the original data set, $\overline{X_2}$ is the mean of 30 ten-fold cross validation experiments for the data set with given percentage of lost values, s_1 and s_2 are sample standard deviations for original and incomplete data sets, respectively.

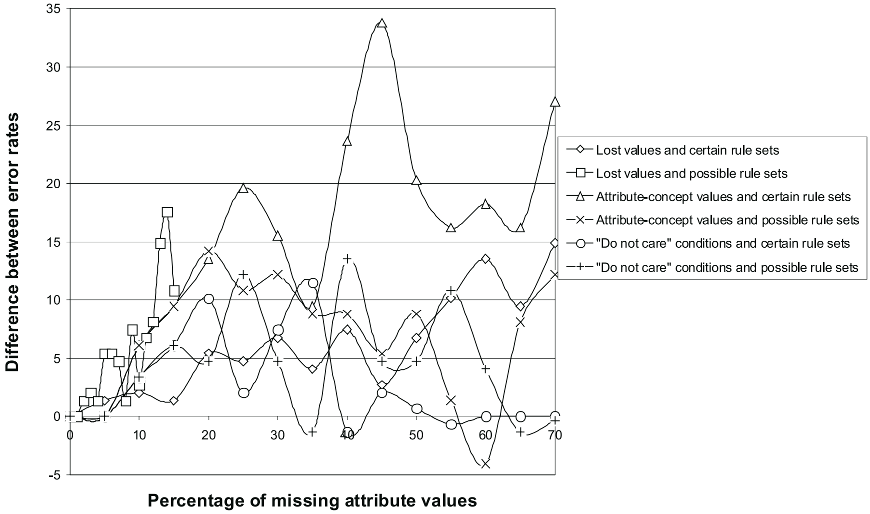

Fig. 6. Lymphography data set. Difference between error rates for testing with complete data sets and data sets with missing attribute values.

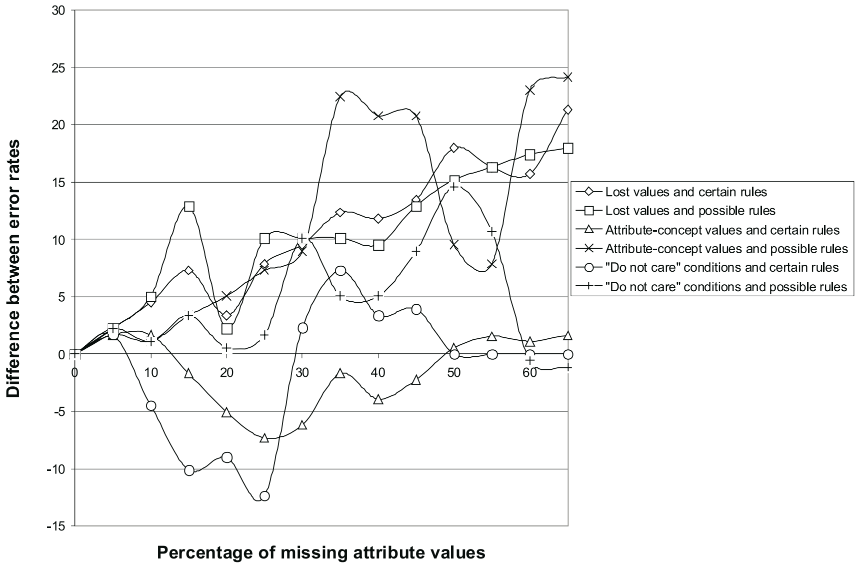

Fig. 7. Wine data set. Difference between error rates for testing with complete data sets and data sets with missing attribute values.

Note that though rule sets were induced from incomplete data, for testing such rule sets the original, complete data were used so that the results for incomplete data are fully comparable with results for the original data sets. Obviously, if

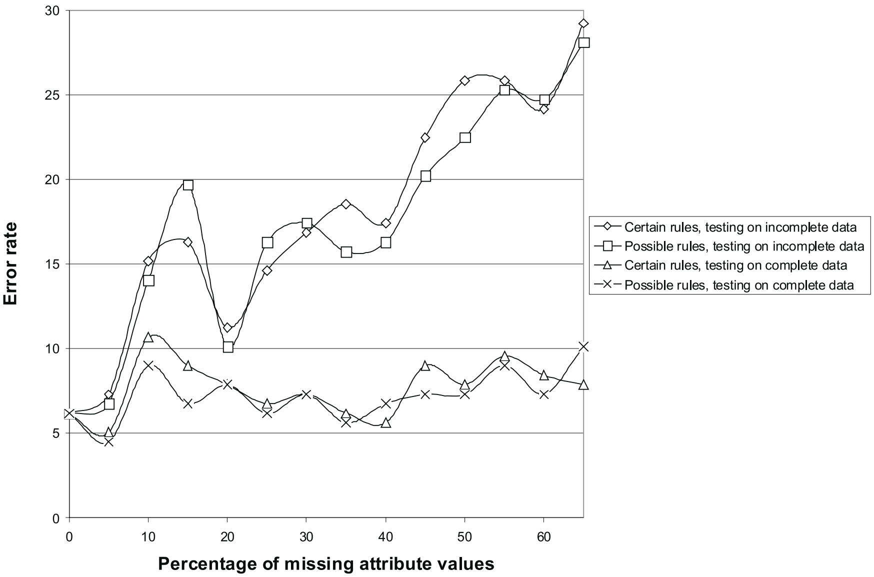

Fig. 8. Wine data set. Testing on complete and incomplete data sets, all missing attribute values are interpreted as *lost*.

the Z score is larger than 1.96, the rule set induced from the data set with given percentage of lost values is significantly better than the corresponding rules set induced from the original data set, with the significance level of 5%, two-tailed test. As follows from Tables 2 and 3, there are three and five rules sets better than the rule sets induced from the original data sets, for certain and possible rule sets, respectively. In Tables 2 and 3, the corresponding Z scores are presented in bold font. Additionally, in only one case for certain rule sets and for two cases for possible rule sets the rule sets induced from incomplete data are worse than the rule sets induced from the original data.

The problem is how to recognize a data set that is a good candidate for improving rule sets by increasing incompleteness. One possible criterion is a large difference between two error rates: one induced from incomplete data and tested on incomplete data and the other induced from incomplete data and tested on the original data set. The corresponding differences of these error rates are presented on Figures 1–7.

Another criterion of potential usefulness of inducing rules from incomplete data is the graph of an error rate for rule sets induced from incomplete data and tested on original, complete data. Such graphs were presented in [13]. In this paper we present these graphs, restricted to the *wine* data set and to *lost values* on Figure 8. A good candidate is characterized by the flat graph, roughly speaking, parallel to the *percentage of missing attribute values* axis. It is clear that the *wine* data set satisfies both criteria. Note that all graphs, presented in Figures 1–8, were plotted for single experiments of ten-fold cross validation.

Because of the space limitation, we cannot present more experimental results in this paper, but it is clear that the main objective of this paper is proven: for some data sets it is possible to improve the quality of rule sets by increasing incompleteness of data sets (or replacing existing attribute values by symbols of missing attribute values).

The question is why sometimes we may improve the quality of rule sets by increasing incompleteness of the original data set. As follows from Table 4, the size of the induced rule sets form incomplete data, both in terms of the number of rules and the total number of conditions, is larger for incomplete data. This fact follows from the MLEM2 algorithm: MLEM2 is less likely to induce simpler rules if the search space is smaller. A possible explanation for occasional improvement of the quality of rule sets is redundancy of information in some data sets, such as *wine* data set, so that it is still possible to induce not only good but sometimes even better rule sets than the rule set induced from the original data set.

3 Conclusions

As follows form our experiments, there are some cases of the rule sets, induced from incomplete data sets, with an error rate (result of ten-fold cross validation) significantly smaller (with a significance level of 5%, two-tailed test) than the error rate for the rule set induced from the original data set. Thus, we proved that there exists an additional technique for improving rule sets, based on increasing incompleteness of the original data set (by replacing some existing attribute values by symbols of missing attribute values). Note that this technique is not always successful. A possible criterion for success are based on large difference between the error rate for rule sets induced from incomplete data and tested on original data and on incomplete data. As follows from Figures 1–7, *image segmentation*, *iris* and *lymphography* data sets are also, potentially, good candidates for improving rule sets based on increasing incompleteness of the original data sets. Another criterion is a flat graph for an error rate versus percentage of missing attribute vales for rule sets induced from incomplete data and tested on original, complete data.

References

1. Grzymala-Busse, J.W., Wang, A.Y.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. In: Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC 1997) at the Third Joint Conference on Information Sciences (JCIS 1997), pp. 69–72 (1997)
2. Stefanowski, J., Tsoukias, A.: On the extension of rough sets under incomplete information. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 73–82. Springer, Heidelberg (1999)
3. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Computational Intelligence 17, 545–566 (2001)
4. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Workshop Notes, Foundations and New Directions of Data Mining, in conjunction with the 3rd International Conference on Data Mining, pp. 56–63 (2003)

5. Grzymala-Busse, J.W.: Three approaches to missing attribute values—a rough set perspective. In: Proceedings of the Workshop on Foundation of Data Mining, in conunction with the Fourth IEEE International Conference on Data Mining, pp. 55–62 (2004)
6. Grzymala-Busse, J.W.: On the unknown attribute values in learning from examples. In: Raś, Z.W., Zemankova, M. (eds.) ISMIS 1991. LNCS, vol. 542, pp. 368–377. Springer, Heidelberg (1991)
7. Kryszkiewicz, M.: Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, pp. 194–197 (1995)
8. Kryszkiewicz, M.: Rules in incomplete information systems. Information Sciences 113, 271–292 (1999)
9. 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, pp. 243–250 (2002)
10. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
11. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
12. Chmielewski, M.R., Grzymala-Busse, J.W.: Global discretization of continuous attributes as preprocessing for machine learning. International Journal of Approximate Reasoning 15, 319–331 (1996)
13. Grzymala-Busse, J.W., Grzymala-Busse, W.J.: Improving quality of rule sets by increasing incompleteness of data sets. In: Proceedings of the Third International Conference on Software and Data Technologies, pp. 241–248 (2008)