On Facts Versus Misconceptions about Rough Sets

Igor Kononenko
University of Ljubljana
Faculty of Computer and Information Science
Tržaška 25, SI-1000 Ljubljana, Slovenia
E-mail: igor.kononenko@fri.uni-lj.si

Keywords: rough set theory, critical analysis, machine learning

Edited by: Matjaž Gams
Received: December 2, 1996 **Revised:** December 4, 1996 **Accepted:** December 5, 1996

This note is a response to the paper Rough Sets: Facts Versus Misconceptions *by J.Grzymala-Busse, J.Stefanowski and W. Ziarko, Informatica, this volume, which is in turn the response to the paper (Kononenko and Zorc, 1994). I clarify some points from our original paper that were mistakenly interpreted by Grzymala-Busse et al. and stress points from our original paper that were ignored by Grzymala-Busse et al. I conclude that with additions to the Rough Sets theory one can achieve good performance, which is however not due to Rough Sets but due to the additions, and that the use of Rough Sets is an unnecessary burden for machine learning algorithms.*

1 Introduction

In the paper (Kononenko and Zorc, 1994) we critically analyzed the Rough Sets Theory (RST) approach to machine learning (ML). We analyzed the following drawbacks of the RST approach to machine learning:

- complicated formalization of rather trivial notions and sometimes strange terminology that confuses the point,
- inflexibility of the knowledge representation,
- ad-hoc solutions, and
- comparison of the RST approach to machine learning with other approaches.

We concluded:

"... *It seems that many authors have no overview of the work that is going on in machine learning and that may be the reason for many reinventings and also plenty of ad-hoc solutions. ... Complicated formalization in RST adds confusion with numerous new notions and unusual terminology that prevents global overview of the RST and prevents systematic analysis. ... The problems with noise and incomplete data disables RST from providing efficient solutions for complex real-world problems.*"

Grzymala-Busse et al. (this volume) tried to "correct the inaccuracies and respond to unfounded claims" in our paper. In this note I clarify some points from our original paper that were mistakenly interpreted by Grzymala-Busse et al. and stress points from our original paper that were ignored by Grzymala-Busse et al.

2 Experimental comparison in (Kononenko and Zorc, 1994)

In the paper (Kononenko and Zorc, 1994) we compared the performance of one RST-based algorithm for inducing decision rules with one classical algorithm for generating decision trees called Assistant (Cestnik et al., 1987) and the naive Bayesian classifier. We reproduce the results in Tables 1 (classification accuracy) and 2 (information score, (Kononenko and Bratko, 1991)).

In (Kononenko and Zorc, 1994) we have interpreted the results as follows:

"*All differences in the classification accuracy (Table 1) that are less than 4 % are*

Table 1: Comparison of the classification accuracy (%) of different classifiers on various data sets.

domain	Assistant	Bayes	RST
primary tumor	44	50	35
breast cancer	77	79	80
thyroid diseases	73	72	61
rheumatology	65	69	66
hepatitis	82	87	81
lymphography	79	84	77
criminology	61	61	63
fresh concrete	61	63	61

Table 2: Comparison of the average information score (bit) of different classifiers on various data sets.

domain	Assistant	Bayes	RST
primary tumor	1.38	1.57	0.96
breast cancer	0.07	0.18	-0.04
thyroid diseases	0.87	0.85	0.46
rheumatology	0.46	0.58	0.16
hepatitis	0.15	0.42	0.12
lymphography	0.67	0.83	0.51
criminology	0.06	0.27	0.03
fresh concrete	0.70	0.89	0.59

statistically insignificant (confidence level is 0.99 using two-tailed t-test). Other differences are significant. However, note that for breast cancer, rheumatology, and criminology, where the differences are the lowest, the classification accuracy is practically equal to the proportion of the majority class. For those data sets the information score is a better measure. The majority of differences in information score (Table 2) are statistically significant (the exceptions are the differences between Assistant and RST in hepatitis and criminology)."

However, Grzymala-Busse et al. comment:

"*Their ad hoc methodology of comparison is based on removal of "differences in the classification accuracy that are less than 4 percent" and then grading the performance by looking at the remaining differences. Obviously, populations are not normal, so t-distribution will not produce valid results. Moreover, samples are not selected in a random way, etc. What should be used to compare RST with Assistant is a standard non-parametric test, e.g., the Wilcoxon matched-pairs signed rank test. Using Wilcoxon test, the critical value of T for the case of seven pairs of accuracies (one of the pairs should be removed because the results are identical) and one-tailed test at the significance level of 5% is equal to 3. The calculated value of T is 10, so Kononenko and Zorc's own evidence does not permit to reject the null hypothesis. Thus, the correct conclusion is that the performance of both classifiers, RST and Assistant, does not differ significantly.*

The second criterion, "average importance score", used by Kononenko and Zorc, is not convincing either. The criterion is invented by one of these authors and is not widely used by others, and as such, still requires some independent validation."

The above discussion is meaningless. Four points need clarification:

1. We didn't compare the average performance of Assistant with the average performance of the RST algorithms in all domains as such comparison doesn't make sense at all. Obviously one percent in one domain may be a significant improvement while 3 percents in other domain may not be very significant, especially if the deviation of the performance is higher than 3 percents in that domain. Comparing the average performance in various domains is missleading and can easily lead to wrong conclusions. What we have compared is the perfomance of two algorithms on each domain separately.

2. When comparing the performance of two algorithms in one domain, we conducted 10 runs with different training/testing splits. Although we presented only the average results, for evaluating the significance of the difference we, of course, used 10 pairs of results to test the significance level. It turned out, as we stated in our discussion, that all differences in the classification accuracy in Table 1 that are less than 4 % are statistically insignificant and the other differences are significant. The use of 4 % in this statement is merely for brevity reasons. So we didn't use

ad-hoc threshold of 4 %, as Grzymala-Busse et al. wrongly concluded.

3. The average information score has nice properties (see (Kononenko and Bratko, 1991)) that allows it to appropriately deal with probabilistic answers and to take into account the prior probabilities of classes. Although in the majority of experiments the classification accuracy and the information score are highly correlated (which is the main reason why so few other authors use information score, personal communication with many authors from ML community), they contain different messages for the user. The former states merely the percentage of correct answers while the latter provides the estimate of the average information contained in the classifiers answers and in domains with high variances of prior probabilities of classes this may be of significant value (Kononenko and Bratko, 1991). Still, however, there are some authors that cite and/or use this measure (e.g. Bailey and Elkan, 1993; Brazdil et al., 1994; Bruha and Kockova, 1993; Eisenstein and Alemi, 1994; Fürnkranz, 1994; Kodratoff et al., 1994; Michie et al. 1994; Moustakis, 1995; Reich, 1995; Tirri et al., 1996; Zheng, 1993), including even, surprisingly, Grzymala-Busse (1992).

4. In our paper (Kononenko and Zorc, 1994) we give the best results for the RST-based algorithm where we tried different values of the parameter α for the majority class limit. For other algorithms the default values of all parameters were used. Therefore, the results in Tables 1 and 2 are an *overestimation* of the performance of the RST-based algorithm. This fact was ignored by Grzymala-Busse et al.

3 Drawbacks of the RST approach to ML

In this section we stress the drawbacks of the RST as described in (Kononenko and Zorc, 1994):

The lack of experimental comparison: In the previous section we clarified our experimental comparison of one RST-based machine learning algorithm. In our previous paper we claimed that too few experimental evaluation of the RST approach to machine learning exists. Grzymala-Busse et al. disclaim this fact by citing 19 references that appeared later or the same year as our paper. One certainly should look at all those references, however, this cannot serve as argumentation against claims about state-of-the-art of the paper that appeared in 1994. With modifications/extentions of the RST approach to machine learning it is obvious that one can achieve good performance. However, this fact is obvious also for any approach to machine learning.

Complicated formalization: of rather trivial notions and sometimes strange terminology that confuses the point clearly indicates that the RST is an unnecessary burden for the machine learning algorithms. Skipping the RST staff from the RST-based machine learning algorithms would make the algorithms more simple and more readable and would make algorithms more easily extensible to deal with noise and incomplete data.

Grzymala-Busse et al. fail to comment this issue in their paper as well as they fail to comment the definition of complete independence of two attributes X and Y (Pawlak et al., 1988):

$$H(Y|X) = \log m$$

where m is the number of values of attribute Y.

Inflexibility of the knowledge representation: of the RST approach to machine learning was recently overcome in a bunch of different ways by extending the basic RST apporach by more or less ad-hoc solutions. Therefore, the same argument applies for this issues as for the experimental comparison: With modifications/extentions of the RST approach to machine learning it is obvious that one can obtain more flexible knowledge representation. However, this fact is obvious also for any approach to machine learning. Again, the RST approach is an unnecessary burden that can be easily skipped.

Ad-hoc solutions: The conclusions from our original paper are still valid: Instead of using well known results from the probability theory and the information theory, authors from the RST community often use ad-hoc definitions and solutions. There is plenty of parameters and thresholds with poor theoretical background. The same methodology is used also in the recent extentions of the RST approach to ML.

4 Conclusion

From the discussion above I conclude that with additions to the Rough Sets theory one can certainly achieve good performance. However, good performance of such systems is not due to the Rough Sets Theory but due to the additions, and that the use of Rough Sets is an unnecessary burden for machine learning algorithms. Therefore, of conclusions from our original paper only the latter ("The problems with noise and incomplete data disables RST from providing efficient solutions for complex real-world problems.") was invalidated by Grzymala-Busse et al.

References

[1] Bailey T.M., Elkan C. (1993) Estimating accuracy of learned concepts, *Proc. Int. Joint Conf. on Artificial Intelligence IJCAI-93*, pp.895-900.

[2] Brazdil P., Gama J., Henery B. (1994) Characterizing the applicability of classification algorithms using meta-level learning. *Proc. Eurp. Conf. on Machine Learning ECML-94*, Catania, April 1994, pp. 83-102.

[3] Bruha I., Kockova S. (1993) Quality of decision rules: empirical and statistical approaches. *Informatica*, 17:233-243.

[4] Eisenstein E.L., Alemi F.(1994) An Evaluation of Factors Influencing Bayesian Learning- Systems, *J. of the American Medical Informatics Association*, 1:72-284.

[5] Fürnkranz J. (1994) Top down pruning in relational learning, *Proc. Europ. Conf. on Artificial Intelligence ECAI-94*, Amsterdam. pp.453-457.

[6] Grzymala-Busse J.W. (1992) Lers- a system for learning from examples based on rough sets. In. R.Slowinski (ed.) *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publ.

[7] Grzymala-Busse J.W., Stefanowski J. and Ziarko W. (1996) Rough Sets: Facts Versus Misconceptions, *Informatica*, this volume.

[8] Kodratoff Y., Moustakis V., Graner N. (1994) Can machine learning solve my problem, *Applied Artificial Intelligence*, 8:1-31.

[9] Kononenko I. and Bratko I. (1991) Information based evaluation criterion for classifier's performance, *Machine Learning Journal*, 6:67-80.

[10] Kononenko I., Zorc S. (1994) A critical analysis of rough sets approach to machine learning, *Informatica*, 18:305-313

[11] Michie D., Spiegelhalter D.J., Taylor C.C. (eds.) (1994) *Machine learning, neural and statistical classification*, Ellis Horwood.

[12] Moustakis V.S. (1995) CEG - A case-based decision modeling architecture, *European J. of Operational Research*, 84:170-191.

[13] Pawlak Z., Wong S.K.M., Ziarko W. (1988) Rough sets: probabilistic versus deterministic approach. *Int. J. Man-Machine Studies*, 29: 81-95.

[14] Reich Y. (1995) Measuring the value of knowledge, *Int. J. of Human-Computer Studies*, 42:3-30.

[15] Tirri H., Kontkanen P., Myllymaki P. (1996) Probabilistic instance-based learning, *Proc. Int. Conf. on Machine Learning ICML-96*, Bari, Italy, July 1996, pp.507-515.

[16] Zheng Z. (1993) A benchmark for classifier learning, *Proc. Australian Joint Conf. on AI*, 16-19. Nov. 1993.