Comparison of Instance Selection Algorithms II. Results and Comments

Marek Grochowski and Norbert Jankowski

Department of Informatics, Nicholaus Copernicus University
ul. Grudziądzka 5, 87-100 Toruń, Poland, http://www.phys.uni.torun.pl/kis
{grochu|norbert}@phys.uni.torun.pl

Abstract. This paper is an continuation of the accompanying paper with the same main title. The first paper reviewed instance selection algorithms, here results of empirical comparison and comments are presented. Several test were performed mostly on benchmark data sets from the machine learning repository at UCI. Instance selection algorithms were tested with neural networks and machine learning algorithms.

1 Introduction

The survey of different algorithms for instance selection was presented in the accompanying article with the same main title.

The performance of instance selection methods is tested here using k-nearest neighbors model, support vectors machine, SSV decision tree, NRBF (a normalized version of RBF network), FSM model and IncNet (see the accompanying paper for references).

Instance selection algorithms were grouped into noise filters (ENN [1], ENRBF [2,3]), condensation algorithms (CNN [4], CA [5], RNN [6], IB3 [7], GE, RNGE [8], ICF [9], ENRBF2, DROP1-5[10]) and prototype selection algorithms (LVQ [11], MC1 & RMHC [12], ELH, ELGrow and Explore [13], DEL [10]).

2 Results

To test the reliability of instance selection algorithms the performance on several datasets was checked. Nearly all tests were prepared on databases from the Machine Learning Repository at UCI Irvine [14]. One database (skin cancer – 250 train and 26 test instances, 14 attributes, 4 classes) was obtained from Z. Hippe [15]. Each benchmark was tested with 10-fold cross-validation, except for the skin cancer dataset which has separate test file. Cross-validation results were repeated and averaged over 10 runs. Standardization of all data was performed before learning. The following UCI repository datasets were used in tests: Wisconsin breast cancer (699 instances, 9 attributes and two classes), Cleveland heart disease (303 instances, 13 attributes, 2 classes), appendicitis (106

L. Rutkowski et al. (Eds.): ICAISC 2004, LNAI 3070, pp. 580–585, 2004.

instances, 8 attributes, two classes), Iris (150 instances, 4 attributes, two classes), wine (178 instances, 13 attributes, 3 classes), Pima indians diabetes (768 instances, 8 attributes, two classes).

Figures 1–6 present information about accuracy on the unseen data and on the compression provided by the selection algorithms. Each figure corresponds to a single classification algorithm (kNN, NRBF, FSM, IncNet, SSV, SVM) tested with several instance selection algorithms (single point); results were averaged over all benchmarks. The horizontal axis shows the compression of the training set in percents (100% = the whole training set). The vertical axis corresponds to accuracy changes on **the test set** for a given instance selection algorithm. The *zero level* is defined by the accuracy obtained by a given classification algorithm trained on the whole training set (kNN, NRBF, etc.). For example "+2" on vertical axis means that the average test accuracy is 2% better than the base algorithm trained using the whole dataset.

In the *dataset reduction* category at the top are MC1, RMHC, LVQ with 1.3% of training set instances left, next are ELGrow (1.35%), Explore (1.43%), DEL (4.91%). It is important that the first three algorithms had compression factor fixed (one instance per class), while algorithms ELGrow, Explore and DEL estimate optimal reduction level automatically depending only on the complexity of the training dataset. As can be seen especially in figures 1 and 2 the performance of the ELH or the ELGrow is worse than that of Explore, and the performance of DEL algorithm falls between them. Next to algorithms Explore and DEL are algorithms ICF, DROP3 and DROP5. Algorithms like LVQ, RMHC, Explore, MC1, DEL and DROP2-4 have got the best performance taking into account accuracy and dataset reduction performance – the upper left parts of figures 1 and 2 – for kNN and NRBF classifiers. It is very important that prototypes extracted from algorithms such as the Explore or RMHC algorithm, that leave only a few instances, can be considered as really simple knowledge representation through prototypes. In most cases Explore extracts extremely small set (a few) of instances and they are very effective (accuracy on unseen data is very high); in cases where Explore is not at the top of accuracy on the unseen data it may be substituted by RMHC, MC1 or DROP2-4 algorithms. For example on the appendicitis database the accuracy of kNN with selection algorithm Explore was 82.7%, and with MC1 was 86.7% (the base performance of kNN was 86.4%)

For kNN and NRBF interesting results were obtained for some prototype selection algorithms, condensation algorithms, as well as for some noise reduction filters, but for FSM, IncNet, SSV or SVM models it is clear that only some noise reduction algorithms (like ENRBF or ENN) can be used, and without any significant gain in accuracy. The selection algorithm ENRBF2 may be considered for models FSM, IncNet or SVM. Noise reduction for these models may stabilize the learning process, however it is not necessary for most of the benchmark. Note that algorithms like Explore, RMHC, ELGrow lead to a complete collapse when used with SSV or SVM.

For tests presented in figures 1–6 LVQ, MC1 and RMHC were configured with one instance per one class.

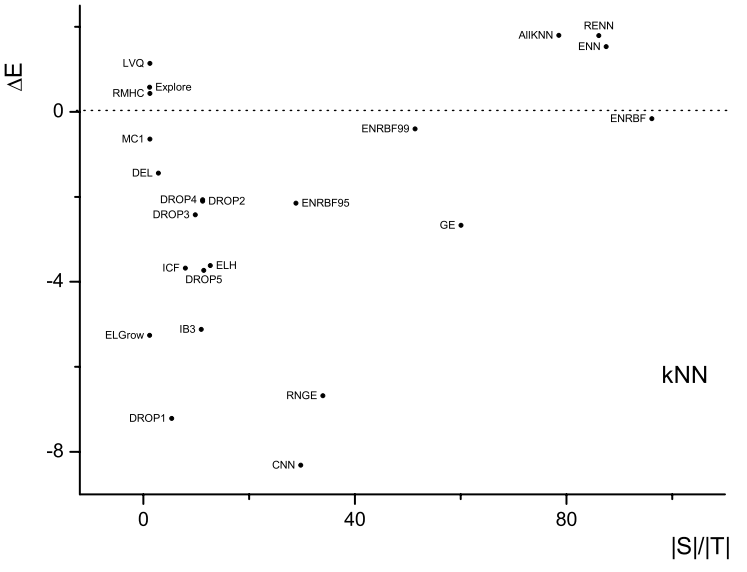

Fig. 1. Classifier: kNN ($\Delta E = 0$ corresponds to accuracy 85.77%)

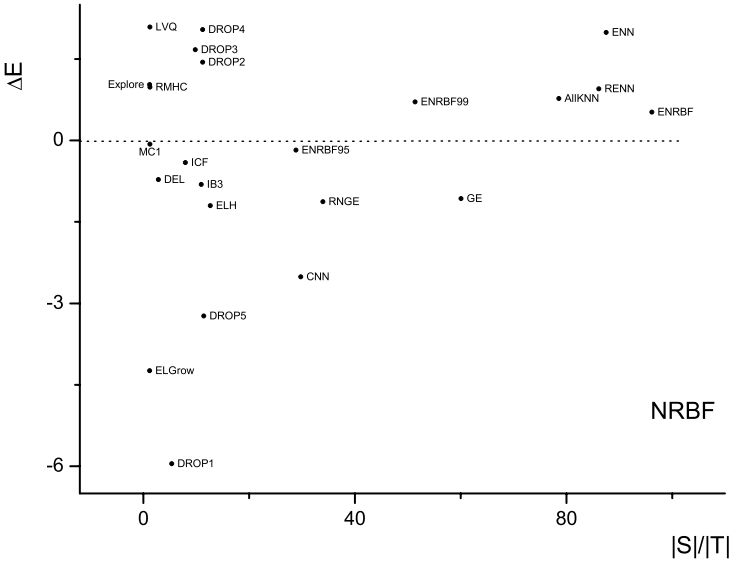

Fig. 2. Classifier: NRBF ($\Delta E = 0$ corresponds to accuracy 84.87%)

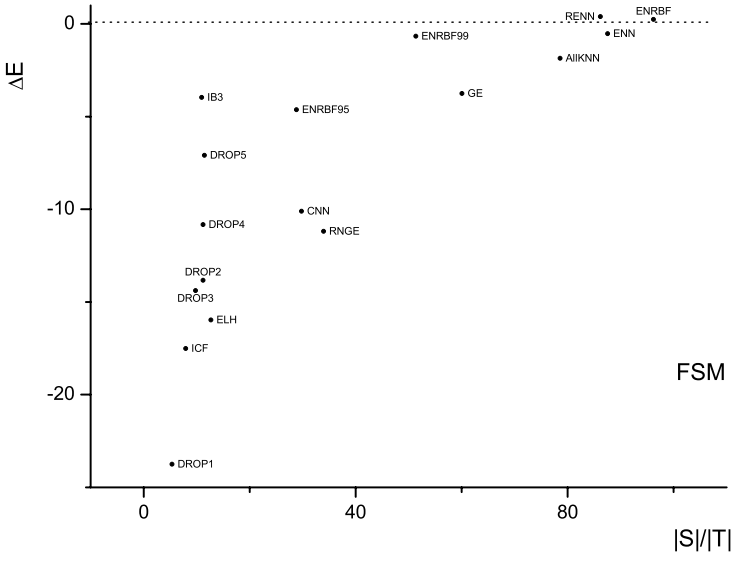

Fig. 3. Classifier: FSM ($\Delta E = 0$ corresponds to accuracy 89.18%)

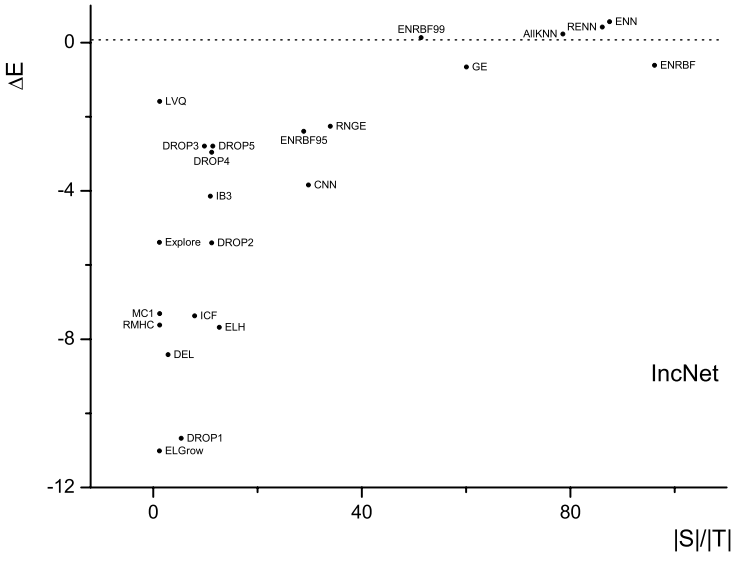

Fig. 4. Classifier: IncNet ($\Delta E = 0$ corresponds to accuracy 88.01%)

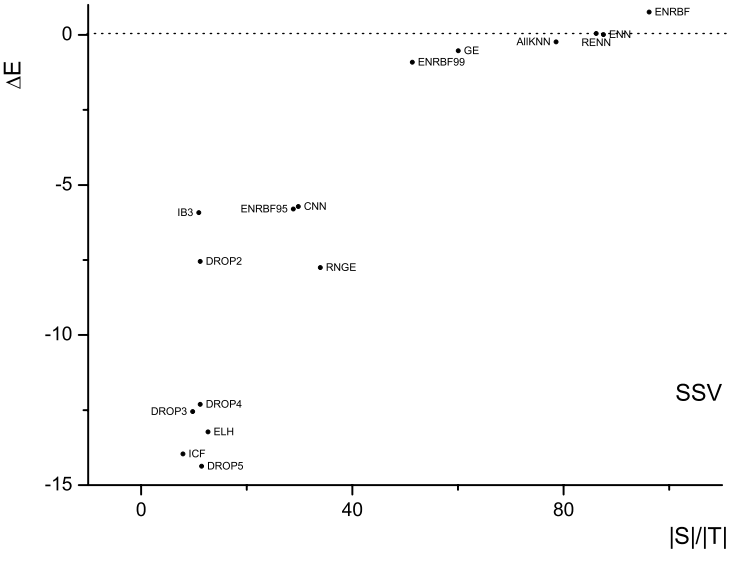

Fig. 5. Classifier: SSV ($\Delta E = 0$ corresponds to accuracy 88.18%)

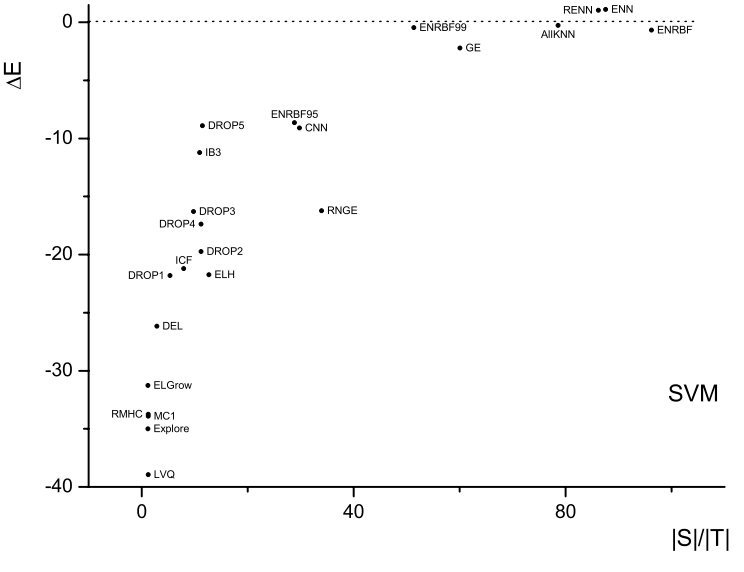

Fig. 6. Classifier: SVM ($\Delta E = 0$ corresponds to accuracy 86.35%)

3 Conclusions

Some of the instance selection algorithms tested in this paper are very interesting. Between the prototype selection algorithms Explore, RMHC, MC1, LVQ, DROP2-4 and DEL, are the most effective. They automatically estimate the number of instances for optimal compression of the training set and reach high accuracy on the unseen data. The RMHC and LVQ algorithms also finished tests with high level of accuracy on the unseen data.

In the group of noise filters ENN algorithm came at the top, especially for kNN & NRBF and ENRBF. These algorithm may stabilize the learning process, except the SSV or SVM models.

References

1. Wilson, D.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics **2** (1972) 408—421
2. Grochowski, M.: Wybór wektorów referencyjnych dla wybranych method klasyfikacji. Master's thesis, Department of Informatics, Nicholas Copernicus University, Poland (2003)
3. Jankowski, N.: Data regularization. In Rutkowski, L., Tadeusiewicz, R., eds.: Neural Networks and Soft Computing, Zakopane, Poland (2000) 209–214
4. Hart, P.E.: The condensed nearest neighbor rule. IEEE Transactions on Information Theory **14** (1968) 515—516
5. Chang, C.L.: Finding prototypes for nearest neighbor classifiers. IEEE Transactions on Computers **23** (1974) 1179—1184
6. Gates, G.: The reduced nearest neighbor rule. IEEE Transactions on Information Theory **18** (1972) 431—433
7. Aha, D.W., Kibler, D., Albert, M.K.: Aha,. Machine Learning **6** (1991) 37—66
8. Bhattacharya, B.K., Poulsen, R.S., Toussaint, G.T.: Application of proximity graphs to editing nearest neighbor decision rule. In: International Symposium on Information Theory, Santa Monica (1981)
9. Brighton, H., Mellish, C.: Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery **6** (2002) 153–172
10. Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-based learning algorithms. Machine Learning **38** (2000) 257—286
11. Kohonen, T.: Learning vector quantization for pattern recognition. Technical Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland (1986)
12. Skalak, D.B.: Prototype and feature selection by sampling and random mutation hill climbing algorithms. In: International Conference on Machine Learning. (1994) 293—301
13. Cameron-Jones, R.M.: Instance selection by encoding length heuristic with random mutation hill climbing. In: Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence. (1995) 99—106
14. Merz, C.J., Murphy, P.M.: UCI repository of machine learning databases (1998) http://www.ics.uci.edu/∼mlearn/MLRepository.html.
15. Hippe, Z.S., Iwaszek, G.: From research on a new method of development of quasi-optimal decision trees. In Kopotek, M., Michalewicz, M., Wierzcho, S.T., eds.: Intelligent Information Systems IX, Warszawa, Institute of computer science (2000) 31–35