Instance Pruning with the SBL-PM-M-EkP System

Karol Grudziński
E-mail: karol.grudzinski@wp.pl

Department of Physics, Kazimierz Wielki University
Plac Weyssenhoffa 11, 85-072 Bydgoszcz, Poland
and
Institute of Applied Informatics
University of Economy
Garbary 2, 85-229 Bydgoszcz, Poland

Abstract. A prototype (case-based) way of data explanation is a powerful method for data analysis and understanding. Interesting instance vectors (prototypes) are usually generated by a training set pruning with various partial memory learners. This approach is the alternative to the rule induction techniques for knowledge discovery and understanding. In this paper a completely new system, SBL-PM-M-EkP is introduced. The study of suitability of SBL-PM-M-EkP for training data compression has been studied on ? datasets. As an underlying classifier we have chosen the well known IB1 system (1-Nearest Neighbor classifier). We compare the generalization ability of our system to the performance of IB1 trained on the entire training data. The results indicate that even with only few prototypes per class which has been generated by SBL-PM-M-EkP system, on ? datasets we have obtained statistically indistinguishable results from those attained with IB1. In several cases the generalization ability has been improved by our system over the IB1 method.

1 Introduction.

Data mining is commonly used in many domains. A case-based way of data explanation is very popular among researchers. Such an approach to knowledge discovery and understanding is particularly often employed in medicine, where a medical doctor makes a diagnosis by referring to other similar cases in a database of patients.

Interesting instance vectors (prototypes) are usually generated by a training set pruning with various partial memory learners. The term 'Partial Memory Learning' (PML) is most often reserved for on-line learning systems that select and store a portion of the past learning examples. In this paper we stick to this naming convention (i.e. PML), but this methodology is called also 'instance selection', 'training data compression, reduction or pruning'. The idea behind this machine learning paradigm is that only a small fraction of a usually much larger, original training set is used for a final classification of unseen samples. [1–10].

The acronym SBL-PM-M-EkP is short for **S**imilarity-**B**ased-**L**earner-**P**artial-**M**emory-**M**inimization-**E**xacly-k-**P**rototypes. We however want to stress here that our new system is completely different from our earlier model, SBL-PM-M.

References

1. Maloof, M. A., Michalski, R. S.: AQ-PM: A System for Partial Memory Learning. Intelligent Information Systems, Ustroń, Poland (1999) 70-79
2. Maloof, M., Michalski, R. S.: Selecting Examples for Partial Memory Learning. Machine Learning, **41**, (2000) 27-52
3. Maloof, M., Michalski, R. S.: Incremental Learning with Partial Instance Memory. Proceedings of the Thirteenth International Symposium on Methodologies for Intelligent Systems, Lyon, France, (2002), In Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence, (2366), Berlin:Springer-Verlag, 16-27
4. Wilson, D. R., Martinez, T. R.: Instance Pruning Techniques. In Fisher, D.: Machine Learning: Proceedings of the Fourteenth International Conference. Morgan Kaufmann Publishers, San Francisco, CA, (1997), 404-417.
5. Wilson, D. R., Martinez, T. R.: Reduction Techniques for Instance-Based Learning Algorithms. Machine Learning, **38**, (2000), 257-286
6. Grochowski, M.: Selecting Reference Vectors in Selected Methods for Classification. MSc. Thesis, Nicholaus Copernicus University, Department of Applied Informatics, Toruń, Poland, (2003) (In Polish)
7. Jankowski N., Grochowski, M.: Comparison of Instances Selection Algorithms I: Algorithms Survey. Artificial Intelligence and Soft Computing ICAISC 2004, in Lecture Notes in Artificial Intelligence (LNAI 3070), (Springer), 598-603.
8. Grochowski, M., Jankowski N.: Comparison of Instance Selection Algorithms II: Results and Comments. Artificial Intelligence and Soft Computing ICAISC 2004, in Lecture Notes in Artificial Intelligence (LNAI 3070), (Springer), 580-585.
9. Grudziński, K., Duch, W.: SBL-PM: A Simple Algorithm for Selection of Reference Instances for Similarity-Based Methods. Intelligent Information Systems, Bystra, Poland (2000), in Advances in Soft Computing, Physica-Verlag (Springer), 99-108
10. Grudziński, K.: SBL-PM-M: A System for Partial Memory Learning. Artificial Intelligence and Soft Computing ICAISC 2004, in Lecture Notes in Artificial Intelligence (LNAI 3070), (Springer), 586-591.
11. Nelder, J. A., Mead, R.: A simplex method for function minimization. Computer Journal **7** (1965), 308-313
12. Ortega J., Koppel M., Argamon S.: Arbitrating Among Competing Classifiers Using Learned Referees. Knowledge and Information Systems 3 (2001), 470-490.
13. Bauer E., Kohavi R.: An empirical comparison of voting classification algorithms: bagging, boosting and variants. Machine Learning 36 (1999), 105-142
14. Duch, W.: Similarity Based Methods: a general framework for classification, approximation and association. Control and Cybernetics 29 **4**, (2000), 1-30
15. SBL, Similarity Based Learner, Software Developed by Karol Grudziński. Nicholaus Copernicus University: 1997-2002, Kazimierz Wielki Univeristy: 2002-2008, University of Economy: 2005-2008.
16. Mertz, C. J., Murphy, P. M.: UCI repository of machine learning databases. http://www.ics.uci.edu/pub/machine-learning-data-bases.
17. Dudani S. A.: The distance-weighted k-nearest-neighbor-rule. IEEE Transactions on Systems, Man and Cybernetics 6 (4) 1975, 325-327.

Tables

Table 1. 10-Fold Cross-Validation Results Obtained on the Selected Datasets with the IB1 System Trained on the Whole Training Partitions vs. the Results Attained by the SBL-PM-M-EkP System with 1, 2 and 3 Prototypes Per Class.

#	**Dataset**	**IB1(%)**	**s.d.(%)**	**EkP-1(%)**	**s.d.(%)**	**EkP-2(%)**	**s.d.(%)**	**EkP-3(%)**	**s.d.(%)**
1.	Appendicitis	80.3	10.8	86.3	10.7	86.4	9.2	**87.1**	9.0
2.	Balance-Scale	78.2	5.0	79.3	4.8	79.7	5.0	**80.5**	4.6
3.	Breast-C.-Liubliana	68.6	7.5	**73.6**	6.4	73.0	5.8	73.1	5.2
4.	Breast-C.-Wisc	95.7	2.4	95.5	2.5	96.1	2.3	**96.2**	2.3
5.	Horse-Colic	78.2	5.0	**79.3**	4.8	78.3	6.1	78.7	6.6
6.	Credit-German	**71.9**	3.7	69.7	1.5	69.8	3.1	69.4	3.2
7.	Credit-Rating	81.6	4.6	79.6	6.1	81.4	5.0	**82.5**	4.6
8.	Heart-Cleveland	76.1	6.8	80.3	7.2	**80.5**	6.7	79.6	7.1
9.	Heart-Hungarian	78.3	7.5	**82.5**	6.5	82.0	7.1	82.3	6.7
10.	Heart-Statlog	76.1	6.8	**80.3**	7.2	80.0	7.3	80.1	7.5
11.	Hepatitis	81.4	8.6	80.7	9.2	83.2	9.4	**83.3**	8.9
12.	Iris	**95.4**	4.8	92.1	7.1	94.5	5.7	94.1	5.9
13.	Leukemia, 8 attrbs.	78.6	13.7	75.8	14.6	78.0	15.1	**78.9**	14.3
14.	Leukemia, 15 attrbs.	**91.4**	9.8	89.4	11.7	90.4	11.7	90.0	10.1
15.	Pima-Diabietes	70.6	4.7	70.4	5.9	**71.1**	5.7	**71.1**	4.9