Combining Robustness and Flexibility in Learning Drifting Concepts

Gerhard WIDMER
Department of Medical Cybernetics and Artificial Intelligence,
University of Vienna, and
Austrian Research Institute for Artificial Intelligence,
Schottengasse 3, A-1010 Vienna, Austria
e-mail: gerhard@ai.univie.ac.at

Abstract

The paper deals with incremental concept learning from classified examples. In many real-world applications, the target concepts of interest may change over time, and incremental learners should be able to track such changes and adapt to them. The problem is known in the literature as *concept drift*. The paper presents a new method for learning in such changing environments. In particular, it addresses the problem of learning drifting concepts from noisy data. We present an algorithm that is both robust against noise and quick at recognizing and adapting to changes in the target concepts. The method has been implemented in a system named *FLORA4*, the latest member of a whole family of learning algorithms. Experiments demonstrate significant improvement over previous results, both in noise-free and noisy situations.

1 Introduction

In many real-world domains, the context on which some concepts of interest depend may change, resulting in more or less abrupt and radical changes in the definition of the target concept. A typical example are weather prediction rules, which may vary radically with the change of seasons. As another example, consider measuring devices or sensors which may alter their characteristics over longer periods of time, resulting in a perceived change of the world and the necessity to modify prediction rules that rely on these measurements. Incremental learning algorithms operating in such environments should be capable of adapting to and tracking such changes. The problem has been termed *concept drift* and has been recognized in the machine learning literature for quite some time (see, e.g., Schlimmer and Granger, 1986). Recently, the notions of context-dependence and concept drift have received renewed interest by a number of researchers (e.g., Kilander and Jansson, 1993; Salganicoff, 1993a; Turney, 1993; Widmer and Kubat, 1992, 1993).

A difficult problem in incremental learning is distinguishing between 'real' concept drift and slight irregularities that are due to *noise* in the training data. Methods designed to react quickly to the first signs of concept drift may be misled into over-reacting to noise, erroneously interpreting it as concept drift. This leads to unstable behaviour and low predictive accuracy in noisy environments. On the other hand, an incremental learner that

is designed primarily to be highly robust against noise runs the risk of not recognizing real changes in the target concepts and may thus adjust to changing conditions very slowly, or only when the concepts change radically. An ideal learner should combine stability and robustness against noise with flexible and effective context tracking capabilities. However, on the face of it, these two requirements seem diametrically opposed.

This paper reports on a new learning method designed to achieve exactly this combination of seemingly incompatible capabilities, at least to a higher degree than previously possible. The central idea of the approach is to combine a generalization and selection strategy based on statistical confidence measures with a time-based forgetting operator. The generalization strategy will provide noise resistance, and the forgetting operator together with a heuristic that controls the amount of forgetting will enable the algorithm to adjust very rapidly to changes in the target concept.

This work is based on previous research on the *FLORA* family of incremental learning algorithms (Widmer and Kubat, 1992; 1993), which were designed to track concept drift effectively. The new method has been implemented in a system by the name of *FLORA4*. To set the stage, we will first briefly review the main components of the *FLORA* strategy, as previously realized in the system *FLORA3*. The main part of the paper will then descibe the new learning algorithm *FLORA4*, and two sets of experiments will be presented that demonstrate the effectiveness of the method.

2 A quick review of *FLORA3*

Let us start by briefly describing the essential components of the *FLORA* approach to learning as they were realized in our last system *FLORA3*. An understanding of these is necessary, as *FLORA4* will be based on the same architecture. We can only give a rather cursory account of the basic method here; the interested reader is referred to (Widmer and Kubat, 1993) for the details of the *FLORA3* algorithm.

FLORA3 assumes an incremental concept learning scenario, where a stream of training examples (positive and negative instances of a target concept) is coming in. Examples are processed one by one, and the system updates its concept hypothesis after each instance. The representation language is propositional: examples are described by attribute–value pairs, and generalizations/hypotheses are sets or disjunctions of conjunctive expressions. The system is specifically targeted at learning problems exhibiting concept drift. The main components of the *FLORA3* method are:

- concept representation in the form of three description sets;
- a forgetting operator and a time window over the incoming examples to control forgetting;
- a heuristic algorithm that automatically and dynamically adjusts the size of this window during learning; and
- a method to store concepts and re-use them in new contexts.

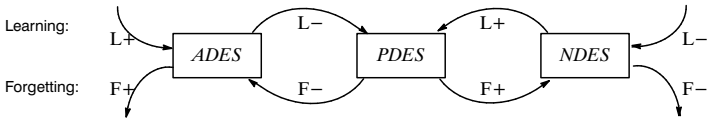

Figure 1: Transitions among the description sets in *FLORA3*.

The concept hypothesis is represented by *three description sets* called *ADES*, *PDES*, and *NDES*. *ADES* contains generalizations that are consistent with the examples; i.e., *ADES* corresponds to the current positive hypothesis and is used to classify new examples. In effect, it represents a propositional hypothesis in disjunctive normal form. Similarly, *NDES* contains generalizations that consistently describe the negative instances. It is used to summarize the negative information seen so far and to prevent over-generalization of hypotheses in *ADES*. Finally, *PDES* is a set of generalizations that describe positive examples, but also some negative ones, i.e., hypotheses that are more or less incorrect or overly general but were once useful and might become relevant again. *PDES* acts as a reservoir of possible alternative hypotheses. Each generalization in these sets is accompanied by explicit *match counts* that count how many positive and negative examples are covered by an expression. Learning in this framework consists in generalizing hypotheses (in *ADES* and *NDES*) in response to incoming examples with a simple incremental *generalization operator*, but also in moving hypotheses from one of the three sets to another under certain circumstances, or dropping some hypothesis altogether. The match counts are used to decide when to move a generalization from one set to another. For instance, an expression in *ADES* is moved to *PDES* as soon as it covers a negative example. Figure 1 indicates the possible migrations of items between the description sets in *FLORA3* (where *L+* and *L-* denote possible transitions after learning from a new positive or negative instance, and *F+/F-* denote possible changes after an example is dropped from the window).

The approach to adapting to concept drift is based on the idea that in dynamic environments, recent information is more trustworthy than older instances, and hence that old examples that maybe pertain to an outdated context should be *forgotten*, i.e., erased from memory, along with generalizations based on them. This is realized in the form of a *time window* that moves over the stream of examples. Incoming examples are added to the window, and old examples are dropped. The description sets are updated so that they only describe instances currently in the window. Hypotheses covering no example in the window are dropped. This mechanism provides the basic capability of adjusting to changes in the concept definition.

Clearly, the effectiveness of this learning method depends crucially on the size of the window. If the window is too narrow, relevant examples and generalizations are forgotten too early, and the result is unstable predictive performance. If the window is too wide, the system will hang on to irrelevant or outdated information too long and will be slow in reacting to concept drift. For this reason, *FLORA3* includes a method for automatically adjusting (growing and shrinking) the window during learning, embodied in a *Window Adjustment Heuristic* (*WAH*). The basic goal of the heuristic is to let the window grow until

the system's hypotheses seem stable, to keep it fixed when learning proceeds smoothly, and most importantly, to successively *shrink* the window when concept drift is suspected. The idea is that when the hidden target concept changes, more of the old examples in the window that still represent the old concept and may now contradict the new instances should be discarded so that the learner can concentrate on the new information and converge more quickly to a good hypothesis for the new concept. The *WAH* takes several indicators into account when guessing whether concept drift is occurring. The two most important ones are the *predictive accuracy* of the current hypothesis, monitored by first trying to classify incoming examples before learning from them, and the *complexity* of the current hypothesis (the set *ADES*). The heuristic has been shown to be quite powerful. (Again, see Widmer and Kubat, 1992 and 1993, for a detailed presentation.)

Finally, another component realized in *FLORA3* is a strategy to *store* stable concept hypotheses and *re-use* them in new contexts. This is especially useful in tasks where contexts and corresponding concepts may reappear. However, as the topic of this paper is concept drift and noise and the experiments described in section 4 do not involve recurring contexts, we will not describe this component here.

3 *FLORA4*: Overcoming brittleness

FLORA3 has been demonstrated to be very effective in adjusting to changes in the target concept, thanks to its highly reactive window adjustment strategy. However, the system turned out to have problems when the input data are *noisy*, which is frequently the case in realistic applications. The factor mainly responsible for this brittleness is the strict *consistency condition* used to decide which generalizations to keep in *ADES*. As hypotheses in *ADES* (and *NDES*) must be strictly consistent with the examples (e.g., an expression in *ADES* must not cover any negative instances), one negative example is sufficient to invalidate a hypothesis and cause it to be moved from *ADES* to *PDES*, even if this hypothesis covers a large number of positive examples. This can lead to somewhat unstable behaviour even in noise-free domains, especially when a concept drift is taking place, but it is particularly problematic when the input data are noisy, i.e., when some of the training examples may be mislabelled.

To counter this problem, our new system *FLORA4* drops this strict consistency condition and replaces it with a 'softer' notion of reliability or predictive power of generalizations. The idea is to continuously monitor the *predictive accuracy* of each generalization in the description sets and to statistically evaluate the confidence of these accuracy estimates: *FLORA4* uses its current generalizations to classify each incoming example before learning from it, and a classification record is kept for each generalization. Statistical *confidence intervals* with a given confidence level are then constructed around these measures. Decisions concerning when to move a hypothesis from one set to another or when to drop it altogether are now based on the relation between these confidence intervals and the observed class frequencies: a hypothesis is kept in *ADES* as long as its predictive accuracy is higher (with high confidence) than the observed frequency of the class it predicts.

More precisely, let μ = required confidence level (parameter); assume that each generalization is associated with two numbers α_l and α_u that represent the lower and upper endpoints, respectively, of the statistical *confidence interval* (with confidence μ) around the generalization's classification accuracy, computed over the instances in the current window; and let γ_l and γ_u be the lower and upper endpoints, respectively, of the confidence interval (with confidence μ) around the relative frequency of the positive class observed so far. *FLORA4* then uses the following criteria to maintain its description sets:

- a generalization G is kept in *ADES* if the lower endpoint of its accuracy confidence interval is greater than the class frequency interval's upper endpoint ($\alpha_l > \gamma_u$); similarly, any G in *PDES* that satisfies this condition is moved to *ADES* — we say that G is (temporarily) *accepted* as a predictor;

- a generalization G in *ADES* whose accuracy interval overlaps with the class frequency interval ($\alpha_u > \gamma_l$) is moved to *PDES* — G is a *mediocre* predictor; expressions in *PDES* are not used for classification;

- a generalization G is dropped completely if the upper endpoint of its accuracy interval is lower than the class frequency interval's lower endpoint ($\alpha_u < \gamma_l$) — G is *rejected*;

- generalizations in *NDES* are kept as long as they are acceptable predictors of negative instances ($\alpha_l > \gamma_u$, computed over the negative examples in the window). In contrast to *FLORA3*, there is no migration of generalizations between *NDES* and *PDES*. Unacceptable hypotheses in *NDES* are simply dropped.

This general approach to deciding which hypotheses to trust has been adopted from the instance-based learning method *IB3* (Aha et al., 1991), which also uses statistical confidence measures to distinguish between reliable and unreliable predictors (*exemplars* in *IB3*). The terms *accepted*, *mediocre*, and *rejected* are used here to highlight this similarity. In all our experiments with *FLORA4*, we used a confidence level $\mu = 80\%$.

The main effect of this new strategy is that generalizations in *ADES* and *NDES* may be permitted to cover some negative or positive instances, respectively, and still to remain in *ADES* or *NDES* if their overall predictive accuracy warrants it. *PDES* is a reservoir of alternative generalizations that are recognized as unreliable at the moment, either because they cover too many negative examples, or because the absolute number of instances they cover is still too small (and thus the confidence intervals are large). The rest of the *FLORA3* strategy, including the generalization operator, has been adopted unchanged in *FLORA4*. After every learning step the window adjustment heuristic is invoked and may decide to grow the window or shrink it, thus dropping a number of old instances. Predictive accuracy of hypotheses is always computed with respect to the current window. In this way, *FLORA4* combines the advantages of the windowing approach—effective adjustment to new contexts by quickly getting rid of old, outdated information—with a less brittle strategy for maintaining relevant generalizations, which should make the system more robust against noise.

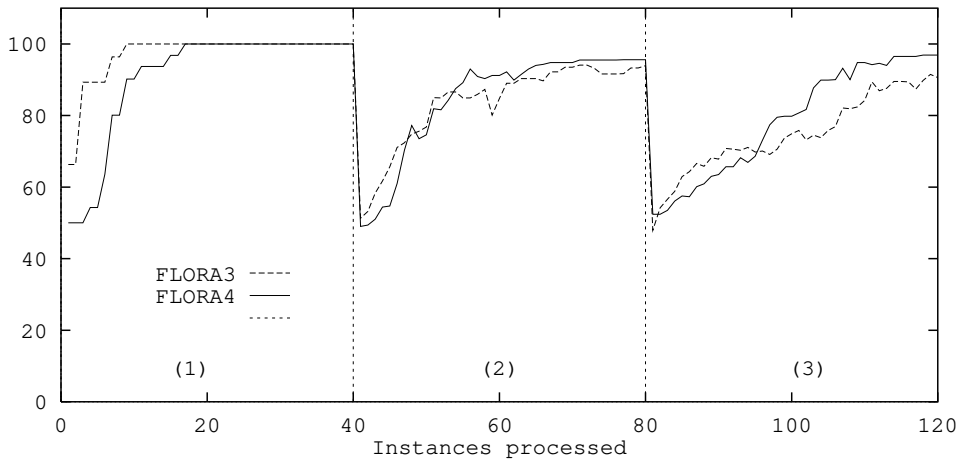

Figure 2: *FLORA3* vs. *FLORA4* on sequence of three concepts.

4 Experiments

This section describes two sets of experiments that were designed to test the effectiveness of *FLORA4* and its improvement over *FLORA3* both in noise-free and noisy concept drift scenarios. For both experiments, we used the same artificial domain as in the articles on *FLORA2* and *FLORA3* (Widmer and Kubat, 1992, 1993). The concepts were initially introduced by Schlimmer and Granger (1986) to demonstrate *STAGGER*'s concept drift tracking abilities. In a simple blocks world, we define a sequence of three target concepts (1) $size = small \wedge color = red$, (2) $color = green \vee shape = circular$ and (3) $size = (medium \vee large)$. The (hidden) target concept will switch from one of these definitions to the next at certain points, creating situations of extreme concept drift.

4.1 Tracking concept drift: *FLORA4* vs. *FLORA3*

The first experiment compares *FLORA4* to *FLORA3* on the basic noise-free drift tracking task. A sequence of training instances was generated randomly according to the hidden concept, and after processing each instance, the predictive accuracy was tested on an independent test set of 100 instances, also generated randomly. The underlying concept was made to change after every 40 training examples. The results in this and all other experiments are averaged over 10 runs. Figure 2 plots the predictive accuracy of *FLORA3* and *FLORA4* in this task.

The characteristic difference between the two systems that is immediately obvious from this result, and that appeared very clearly in all experiments, is that *FLORA4* is initially a bit slower in reacting to the change in the target concept, but then soon picks up and eventually regains high accuracy faster than *FLORA3*, and usually with a smoother curve. The explanation is to be found in *FLORA4*'s statistical confidence measure. *FLORA4*

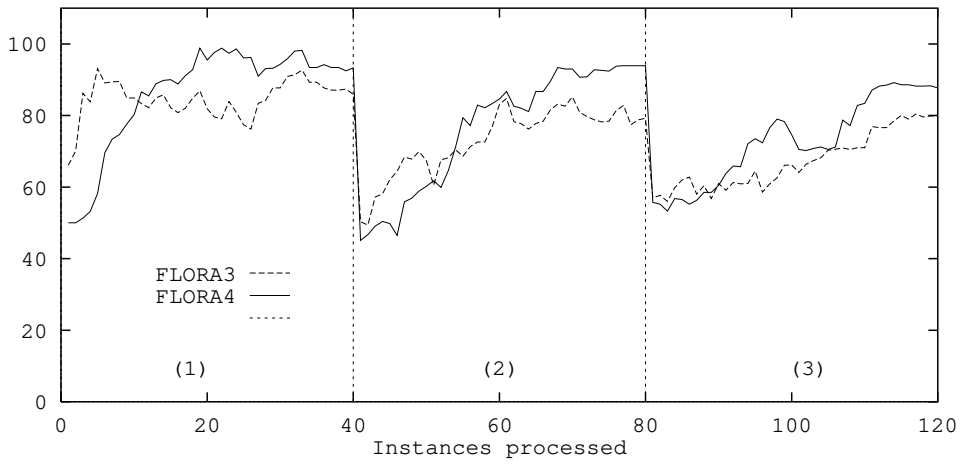

Figure 3: Performance of *FLORA3* and *FLORA4* at 20% noise level.

reacts more reluctantly initially because several contradicting examples are necessary to invalidate a hitherto stable hypothesis in *ADES*, while *FLORA3* will drop a description item as soon as the first contradicting instance appears. (This, of course, is also the source of *FLORA3*'s problem with noisy data, as the next section will show.) The same observation also explains why *FLORA4* later reaches high accuracy faster than its predecessor: a consequence of *FLORA3*'s strict consistency conditions is that one old negative instance (pertaining to the outdated context) erroneously still in the window may prevent a good generalization from being included in *ADES*. *FLORA4*, with its 'softer' consistency condition, is less disturbed by remnants of the old context still in the window and thus readjusts faster to the new context.

4.2 Distinguishing between noise and concept drift

FLORA4's strengths should come out even more clearly when the training data are noisy. Distinguishing between noise and concept drift is inherently difficult, as both problems make themselves known to the learner in the form of prediction errors. Here we expect that the combination of the statistical confidence measures and the window adjustment heuristic will come to bear. The statistical measures provide a certain robustness against noise, especially in relatively stable situations, and the window adjustment heuristic should recognize persistent misclassifications that indicate a concept change, and should lead to effective adjustment by shrinking the window in such situations.

In the following experiment, the same target concepts were used, but the training data were corrupted with various levels of classification noise. *FLORA4* was compared with *FLORA3* throughout and turned out to be consistently and significantly superior. For instance, figure 3 compares the performance of the two systems at a noise level of 20%

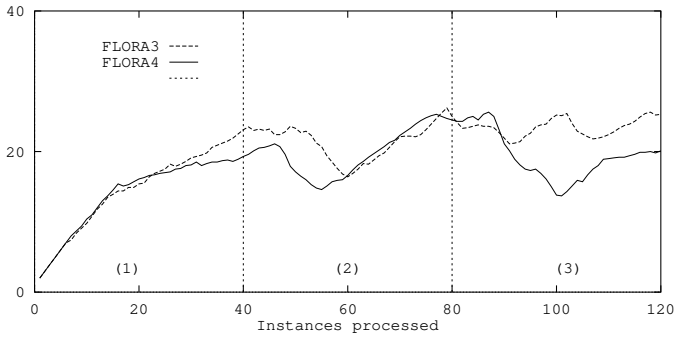

Figure 4: Average window size at 20% noise level.

in the training data.[1] Here again, we see the characteristic difference: *FLORA4* is a bit slower in its initial reaction to the concept change, but then soon outperforms *FLORA3*. However, the difference between the two curves is markedly greater than in the noise-free case. *FLORA3* has obvious problems with noise, while *FLORA4*'s accuracy quickly rises to a mark that corresponds roughly to the given level of noise (remember that 20% noise means 10% misclassified instances on average in a two-class learning task).

A comparison of the average window sizes in this experiment (figure 4) illustrates the workings of the window adjustment heuristic and also the effect of the statistical strategy of the generalizer. The expected characteristic shape of the curve (growing the window to a reasonable size in relatively stable situations, shrinking it in response to a perceived concept drift) comes out clearer in the *FLORA4* case. As the *WAH* reacts to factors (e.g., the number and complexity of accepted expressions in *ADES*) that are also affected by the generalizer's strategy, there is a *synergy* between the two components: the generalizer's robustness against noise prevents the *WAH* from erroneously growing or shrinking the window. This effect is clearly visible in the third phase in figure 4, where *FLORA3* grows and shrinks the window in the middle of a phase of concept stability, which is obviously due to irritation by noisy examples.

The stability of *FLORA4*'s behaviour under different noise conditions is illustrated in figure 5, which shows *FLORA4*'s performance at various noise levels (10, 20, and 40%). The qualitative shape of the performance curves remains unchanged. The rapid drop in accuracy after a concept change is followed by relatively fast re-convergence toward a quasi-optimal prediction accuracy. In no case does the performance really collapse. (The closest it comes to collapsing is with the third (disjunctive) concept in the 40% noise situation, where *FLORA4*'s convergence is rather slow, but still recognizable. This part of the concept seems to be inherently more difficult to adjust to than the other two, as evidenced by *FLORA3*'s and *FLORA4*'s slower convergence even in the noise-free case (figure 2)).

[1] In this article, η% class noise means that with probability $\eta/100$, the class label of an instance will be assigned randomly. Thus, completely random data will be generated when $\eta = 100$.

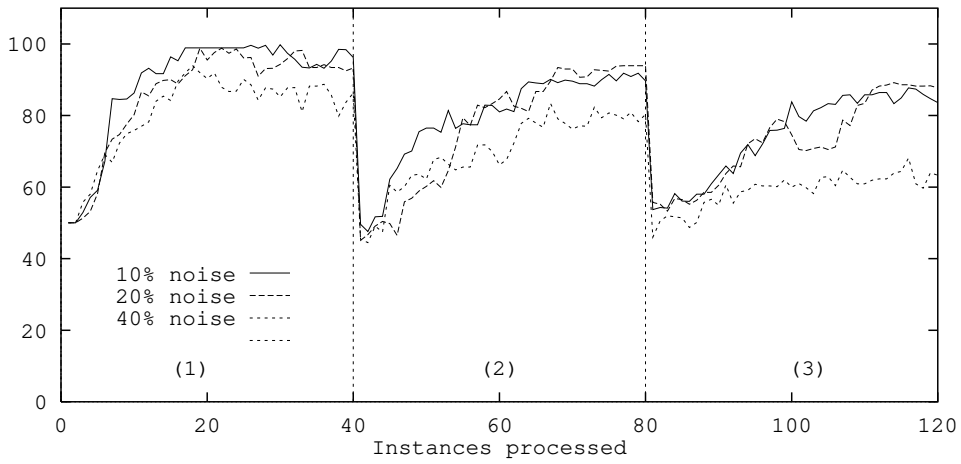

Figure 5: Performance of *FLORA4* at various noise levels.

In summary, it seems justified to say that the combination of statistical performance evaluation with window-based forgetting realized in *FLORA4* produces a system that is at the same time robust against noise and flexible in the recognition of and reaction to concept drift. Additional experiments that were performed but cannot be reported here due to space limitations have confirmed these characteristics also for various *rates of concept drift* (abrupt changes vs. gradual drift).

5 Conclusion

To summarize, the power of *FLORA4* in dealing with both noise and concept drift derives from the fact that it integrates two different learning strategies: the statistical criteria used to distinguish between reliable and unreliable generalizations make it robust against noise, and the 'forgetting' of outdated information, controlled by reactive automatic window adjustment, enables it to quickly adapt to new contexts and concept drift. In terms of the framework of Salganicoff (1993b), *FLORA4* can be characterized as integrating "performance-error weighted forgetting" and "time-weighted forgetting".

The relation between the *FLORA* method and other approaches to learning in the presence of concept drift (especially *STAGGER*) has been discussed at length in (Widmer and Kubat, 1993), especially with respect to the aspect of *forgetting*. Here we will reflect briefly on the relation between *FLORA4* and the instance-based learning algorithm *IB3* (Aha et al., 1991), because the statistical decision criteria of *FLORA4* were adapted from *IB3*'s learning strategy. *IB3* has a certain capability of adjusting to concept drift, even though it was not designed explicitly for this purpose.

Our experience from comparative experiments with the publicly available implementation of *IB3* is that *IB3* requires significantly more examples to converge to a high pre-

dictive accuracy, and especially that it is slower in recovering from changes in the target concept. The first effect is due to the general instance-based learning method. The latter phenomenon—faster re-adjustment of *FLORA4*—is clearly attributable to the combination in *FLORA4* of *IB3*'s statistical confidence measures with a highly reactive window-based forgetting strategy, which permits the system to get rid of irritating information much faster. As a side note, one could also point out that a symbolic generalizer like *FLORA4* has certain advantages over an instance-based learner in terms of the comprehensibility of the results of learning.

Generally, it is interesting to see how two conceptualizations of learning—the three description sets of the *FLORA* family of learners (first introduced in Kubat, 1989) and Aha's categories of accepted, mediocre, and rejected predictors—that were developed out of quite different motivations (the basic motivation and interpretation framework for the description sets in the original *FLORA* approach was *Rough Set Theory* (Pawlak, 1982)) now converge on a common interpretation. We may take this as another indication that one fruitful strategy for achieving powerful machine learning algorithms is to actively search for promising methods in machine learning research and try to combine or integrate them in a more general framework.

Acknowledgments

I would like to thank Miroslav Kubat for his continuing cooperation on the *FLORA* family and fruitful discussions. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry for Science and Research.

References

Aha, D., Kibler D., and Albert, M.K (1991). Instance-Based Learning. *Machine Learning 6(1)*, pp.37–66.

Kilander, F. and Jansson, C.G. (1993). COBBIT - A Control Procedure for COBWEB in the Presence of Concept Drift. In *Proceedings of the European Conference on Machine Learning (ECML-93)*, Vienna, Austria.

Kubat, M. (1989). Floating Approximation in Time-Varying Knowledge Bases. *Pattern Recognition Letters 10*, pp.223–227.

Pawlak, Z. (1982). Rough Sets. *International Journal of Computer and Information Sciences 11*, pp.341–356.

Salganicoff, M. (1993a). Density-Adaptive Learning and Forgetting. In *Proceedings of the 10th International Conference on Machine Learning*, Amherst, MA.

Salganicoff, M. (1993b). Explicit Forgetting Algorithms for Memory-Based Learning. Report MS-CIS-93-80, Dept. of Computer and Information Science, University of Pennsylvania.

Schlimmer, J.C. and Granger, R.H. (1986). Incremental Learning from Noisy Data. *Machine Learning 1*, pp.317–354.

Turney, P.D. (1993). Robust Classification with Context-Sensitive Features. In *Proceedings of the Sixth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-93)*, Edinburgh, Scotland.

Widmer, G. and Kubat, M. (1992). Learning Flexible Concepts from Streams of Examples: *FLORA2*. In *Proceedings of the European Conference on Artifical Intelligence (ECAI-92)*, Vienna, Austria.

Widmer, G. and Kubat, M. (1993). Effective Learning in Dynamic Environments by Explicit Context Tracking. In *Proceedings of the 6th European Conference on Machine Learning (ECML-93)*, Vienna, Austria.