# Chapter 13

# Conclusions and future work

## 13.1 Conclusions

This thesis has explored two common problems in machine learning, the improvement of an existing learning algorithm, in our case the decision tree induction method, and the treatment of a "new" kind of data, in our case temporal data.

Our walk to the first goal was very systematic. First, we studied in depth the behavior of standard decision tree induction in various conditions in order to highlight its main shortcomings. The main drawback of this method turns out to be its variance, be it the prediction variance which affects accuracy or the parameter variance which affects interpretability. From the general considerations about variance reduction techniques in the first part of this thesis, we had at our disposal two main approaches to reduce this variance: regulate the bias/variance tradeoff during induction or combine several perturbed decision trees to decrease variance. The first kind of techniques led us to the evaluation of the well-known pruning algorithm and to the proposal of discretization threshold stabilization methods. As expected, these two approaches only improve slightly accuracy because they inevitably increase the bias. On the other hand, perturb and combine algorithms are very effective in combination with decision trees. As our study of decision tree variance has shown that the parameter variance is very high, we devised an extremely randomized decision tree based algorithm which compares favorably with bagging and boosting algorithms. Trying to propose a compromise between the interpretability of decision trees and the accuracy of perturb and combine algorithm, we reinvented another type of soft tree model as a particular instance of a general variance reduction technique, the dual perturb and combine algorithm.

The comparison of the resulting algorithms given in the closure of the second part of the thesis shows that the proposed methods, from single trees to bagging of soft trees, span a large spectrum of improvements and hence of solutions according to the main criteria used to assess learning algorithms: interpretability, computational efficiency, and accuracy. Nevertheless, we are convinced that there still remains ample room for improvements and further research. In particular, the most promising approaches are certainly the dual perturb and combine algorithm and randomized trees in general. On the other hand, from our experiments with threshold stabilization methods, we do not expect much improvements from variance reduction techniques within the strict hypothesis space of standard decision trees.

Tackling the treatment of time series data was more difficult. First, we defined the general problem of supervised learning with temporal data and then restrained our ambition to the problem of time series classification. An a priori analysis allowed us to highlight the peculiarities of such problems that could prevent standard learning algorithms to succeed. Then, we chose a small number of test problems which were artificially generated to present such peculiarities. As our goal was to propose general approaches, our first experiments logically consisted in applying existing learning algorithms on the low-level attributes with minimum preprocessing. However, even the best variance reduction techniques were not able to tackle all test problems in a satisfactory way. So, we proposed the segment and combine algorithm which allowed to improve the bias/variance tradeoff by artificially providing more information to the learning algorithm while taking into account the temporal ordering of low-level attributes. If the problem of accuracy on our test problems was partially solved by this variant, methods that use low-level attributes were not able to provide really interpretable rules and hence in a second step, we tried to design an interpretable method. To this end, we decided to extend decision tree induction, motivated by the interpretability and the open-minded structure of this method. Our extension focused on the detection of local shift invariant properties. The resulting method gave good results in terms of accuracy and interpretability on problems presenting shift invariant properties but was powerless on other problems.

The main benefits of our study in the last part are, on the one hand, to show that the problem is indeed difficult and hence that it deserves further attention and, on the other hand, to open up two complementary directions for the search of solutions: first adapt perturb and combine algorithms to the properties of temporal data and second try to devise interpretable temporal models. In both directions, we have made several choices (for example, the use of decision trees) but it is clear that many other approaches could have been adopted which might prove superior to the one selected here. We therefore consider that our work is only a first step towards accurate and interpretable automatic learning methods for time series classification.

All in all, our search for solutions to both problems has shown that the design of automatic learning algorithms is a complex multiobjective optimization problem. Among the main conflicting objectives in this context, we have average (and worst-case) accuracy over a large class of problems, interpretability of models, computational efficiency and scalability. Often interpretability means low accuracy and high accuracy means low computational efficiency and no interpretability. While extra-trees show that it is possible to combine efficiency and accuracy in the same algorithm, interpretability and accuracy on the other hand seem to be the most difficult criteria to combine in one algorithm. While dual P&C keeps the interpretability of decision trees, it is not as accurate as other perturb and combine algorithms. Although the extension of decision trees to detect patterns gives accurate results in some types of problems, in average, it is not as good as the segment and combine approach.

Another important compromise which must be taken into account when designing a learning algorithm is the bias/variance tradeoff which regulates the accuracy. Our experiments have shown that thinking problems in terms of bias and variance is useful to understand the behavior of learning algorithms and helps avoiding a priori bad design choices. Although it is not the only way to study learning algorithms, one of its main advantages is that bias and variance can be effectively evaluated (even if this evaluation is cumbersome in some cases). Furthermore, even if the available measures of bias and variance in classification are not fully sound from the theoretical point of view, we have seen that, in combination with the bias/variance decomposition of the regression error on probability estimates, it allows to effectively explain many things.

Eventually, a last conclusion from our experiments is that decision tree induction definitely deserves its place among the most popular learning algorithms. Even if it suffers from several well-known shortcomings (such as its variance), its open-minded structure and the simplicity and efficiency of its induction algorithm makes it the method of choice for many improvements. This thesis has shown for example that the extra-tree extension is a very promising alternative for handling very large datasets and that the power of the tests may be easily extended to handle different kinds of problems. Even if the resulting extensions are sometimes very far from the standard decision tree algorithm, such improvements would not have been easy with more rigid and black-box type algorithms.

## 13.2 Future work directions

We collect here some future research directions. First, we present short term developments which concern improvements or extensions of the presented algorithms. Second, we discuss interesting long term research directions in a more general point of view.

### 13.2.1 Short term investigations

Several possible enhancements of the proposed algorithms have already been mentioned in the course of the thesis. The most important ones are summarized below.

Concerning our method for improving decision trees, the two methods which deserve more attention in the future are certainly the dual perturb and combine algorithm and the extra-tree method.

Concerning the dual P&C algorithm, it would be interesting to investigate how its performances compare with those of bagging and boosting in a broader range of conditions (larger learning samples, larger number of attributes). Although this method can be combined with any learning algorithm, we have studied its effect and implementation only in combination with decision trees. It would be also interesting to investigate the use of this method in combination with other learning algorithms such as neural networks. In this context, a closed-form derivation of the algorithm may further increase its practical interest.

The extra-tree algorithm provides certainly the most promising basis for future research. We would suggest to make a more in depth assessment of its accuracy, in particular, in comparison to techniques such as multilayer perceptrons, and on the basis of broadly accepted datasets, such as for instance those used in the Statlog project [MS94]. The question of computational performance optimization in the context of very large databases also deserves special attention. In particular, among the questions which have not been addressed in the thesis, we mention the storage of extra-trees which sizes might become problematic for very large sample sizes and for some specific applications (such as embedded systems).

Concerning the segment and combine approach for time series classification, it would also be interesting to investigate the behavior of this method on a broader range of problems to better circumscribe its limitations. Depending on the results of such a study, other perturbation schemes could be imagined to tackle problems for which it does not work well enough.

The most attractive extension of pattern extraction techniques is certainly the use of probabilistic models for pattern matching which has been mentioned in the conclusion of Chapter 12. From our experiments in this chapter, we have also concluded that the tree partitioning was maybe responsible for the bad results on the two problems with small learning sample sizes and high number of classes. So, it would be interesting to investigate other ways to take advantage of the pattern extraction algorithm which would be more robust in such conditions (for example, using a set of overlapping rules instead of the recursive partitioning of decision trees).

Finally, although we have focused in this thesis on the treatment of numerical input attributes and symbolic outputs, from a practical point of view it would be necessary to extend our methods to handle other types of data, such as symbolic attributes or regression problems. Notice that, since most of our algorithms rely on decision trees, their extension to regression problems is often trivial (for example, dual P&C and extra-trees are easy to transpose). The extension to symbolic attributes with dual P&C and extra-trees is also direct, although it would need further validation. On the other hand, the treatment of discrete sequences of events with the segment and combine approach or with our extension of decision trees for pattern detection would require more creativity.

### 13.2.2 Long term research

We expose here two long term research directions which we find interesting from a broader point of view.

#### Bias/variance tradeoff and other paradigms

In this thesis, all our improvements have been motivated from a bias/variance point of view. However, the bias/variance decomposition is apparently not able to explain the good behavior of boosting algorithms and furthermore there exist in machine learning other paradigms which have motivated new learning algorithms. For example, recently the machine learning community shows an increasing interest in support vector machines and large margin classifiers which have emerged from Vapnik's work on the statistical learning theory. We have seen that at least qualitatively some connections may be drawn between random tree algorithms and these latter methods (in terms of their ability to overcome the overfitting problem). The Bayesian approach to automatic learning offers another motivation for model averaging techniques. We believe that understanding the relationships between these different paradigms or between the learning algorithms which they have fostered, would be helpful to improve or to design new algorithms and hence certainly deserves further research in the future.

#### Fully automatic method for complex data types

A very challenging and interesting question which is partially addressed in the last part of this thesis concerns the adaptation of automatic learning methods to handle complex data types such as time series, images, texts, or biological sequences. Indeed, more and more databases containing such data become available for automatic learning and there is still a lot of work to be done before these datasets can be handled in an efficient and fully automatic way. According to our preliminary results with time series, we believe that this problem should be tackled from two complementary points of view: first the development of generic approaches, as far as possible independent of the use of prior knowledge about the problem domain, and second, the development of interpretable methods.

With the goal of providing general approaches without consideration about interpretability questions, we believe that variance reduction techniques studied in the first part of this thesis should be methods of choice. In particular, similarly to the segment and combine algorithm, we could devise perturb and combine methods adapted to the considered type of data. Another solution is to extend decision trees with candidate tests adapted to the type of data (just like our pattern detection tests for time series) and then use randomization to avoid the problem of overfitting which will inevitably appear with the increase of flexibility.

On the other hand, the problem of inducing interpretable models seems more challenging. When extending the decision tree method, we have made the hypothesis that the problem may be characterized by local shift invariant properties. This assumption makes the method even competitive with the segment and combine approach on problems which satisfy this hypothesis but however, does not improve the results over the use of low-level attributes on problems which do not satisfy it. Such hypotheses seem to be necessary to give interpretable and accurate models especially in complex domains but prevent the design of very general solutions. One goal of research in this context would be to evaluate to which extent it is possible to free oneself from such a priori hypotheses.