Adaptive Learning Algorithms for Bayesian Network Classifiers

Gladys Castillo
Departamento de Matematica, Universidade de Aveiro, Aveiro 3810-193, Portugal
E-mail: gladys@mat.ua.pt

This thesis is concerned with adaptive learning algorithms for Bayesian network classifiers (BNCs) in a prequential (on-line) learning scenario. Online learning is particular relevant since in many applications learning algorithms act in environments where the data flows continuously. An efficient supervised learning algorithm in dynamic environments must be able to improve its predictive accuracy while optimizing the cost of updating. However, if the process is not strictly stationary, the target concept could change over time. Hence, the predictive model should be adapted quickly to these changes. The main contribution of this work is a proposal of a unified, adaptive prequential framework for supervised learning called AdPreqFr4SL, which attempts to handle the cost-performance trade-off and deal with concept drift. We experimentally prove the advantages of using the AdPreqFr4SL in comparison against its non-adaptive versions for the particular class of k-Dependence Bayesian Classifiers (k-DBCs).

1. Cost-Performance Management

We chose the class of k-DBCs to illustrate our approach. A k-DBC is a BNC, which contains the structure of the Naïve Bayes (NB) and allows each attribute to have a maximum of k attribute nodes as parents. This class is very suitable for our proposal. By increasing k we can obtain classifiers that move smoothly along the spectrum of attribute dependencies.

The adaptation strategy for incorporating new data is based upon two main policies: bias management and gradual adaptation. Instead of selecting a particular class-model of BNCs (e.g. TAN, BAN, etc.) and using it during all the learning process, we use the class of k-DBCs and start with its more restrictive class-model: the NB (k = 0). We then attempt to reduce the bias of the NB by gradually adding dependencies between the attributes over time. To this end, we use simple control strategies based on the observation of some performance indicators to decide when to increase the k value and to start searching for new attribute dependences. This bias control leads to the selection of the optimal class-model for the current training data (i.e. the optimal k value), thus avoiding the problems caused by either too much bias (underfitting) or too much variance (overfitting). Since updating the structure is a costly task, we reduce the cost of updating by first adapting parameters. We adapt the structure only at sparse time points, when there is some accumulated data and there is evidence that the use of the current structure no longer guarantees the desirable improvement in the performance. Finally, we stop the adaptation process when there is evidence that the use of more training data will not result in significantly improved performance. However, we will continue monitoring the performance. If any significant change in the behaviour is observed, then we will activate the adaptation procedures.

2. Concept Drift Management

The method for handling concept drift is based on Statistical Quality Control. We use a Shewhart P-Chart - an attribute control chart for monitoring the proportion of misclassified examples in each incoming batch of examples. The chart has a center line (CL), an upper control limit (UCL) and an upper warning limit (UWL). We set the center line to the minimum value of the current model
error. At each time point it is observed where the current batch error \( p(t) \) falls. If \( p(t) \) falls above the UCL, an abrupt, concept shift is signaled; if \( p(t) \) falls between the UCL and the UWL for two or more consecutive times then a gradual, concept drift is detected. Moreover, we manipulate a SHORT–MEMORY to store those examples that we suspect belongs to a new concept. If a concept shift is detected then all the examples from the SHORT–MEMORY are used to build a new NB classifier. However, after signaling a concept drift, the new examples are not used to update the model in order to force a great degradation of the performance. This way the P-Chart will more quickly be able to recognize a concept shift and re-build the model.

3. General Results for Learning BNCs.

We first conducted an empirical study to compare the performance of several \( k \)-DBCks for different \( k \) values and scores in a prequential non-adaptive framework. At each time point was invoked a hill-climbing algorithm that rebuilt the current hypothesis using all the data seen so far. By providing a bias-variance decomposition of the test error we see that varying \( k \), the score and the training set size produce different bias-variance behaviors. At each time point, there is an optimal \( k \) value that gives the best performance on test data, which is very dependent on how the chosen score makes the bias-variance trade-off for choosing the appropriate complexity of the model for the available data. These facts raise the question about the difficulty of the selection of the optimal \( k \) value for a particular learning task unless we have some beliefs about the actual degree of attribute dependence. This problem is still more challenging in a prequential learning framework since the training data increases over time. Results show that, in general, increasing the \( k \) value above 0 leads to significant improvements over NB only as training data increases (\( \geq 1000 \) examples). The proposed adaptive strategy based upon bias management and adaptation control aimed at automatically solving this problem, independently of the score used.

We evaluated the adaptive algorithms for \( k \)-DBCks in the AdPreqFr45L using both, artificial and benchmark problems. The use of artificial domains allowed us to test our adaptive strategies knowing the true degree of attribute dependencies and when concept drift actually occurs. Results show that the adaptive algorithms are able to scale up the model’s complexity and to perform an artful bias management during the whole learning process. Moreover, in most cases, the resulting \( k \) values approach the real degree of attribute dependence. Results in simulated concept drift scenarios show that the P-Chart is able to consistently recognize concept changes, both abrupt and gradual, and to adapt quickly to these changes, thus verifying a good recoverability capability.

Results also show that adaptive algorithms work as expected, independently of the score used. By gradually increasing the \( k \) value adaptive algorithms are capable of improving the predictive accuracy of the NB significantly over time. This evidence that modeling attribute dependencies can improve the classification results, much better when there is enough training data to discover them. Results also show that, in most cases, the resulting classifier approaches the best \( k \)-DBC induced with the same underlying learning algorithm and score but using a temporal batch learning approach. On the other hand, a considerable reduction of the cost of updating is achieved by using adaptive algorithms, as shown by the small number of adaptations performed on the structure during the whole learning process. Results evidence that it is more appropriate to perform adaptations on the structure only when there is some accumulated data and the search procedure is actually able to find new dependencies.

4. General Results for Supervised Learning

Our contributions to the field of supervised learning are also general because the controlling methods to handle the cost-performance trade-off and concept drift are mainly based on the monitoring of some performance indicators that are classifier-independent. Moreover, we believe that almost all of the adaptive policies for bias management could be applied to essentially any supervised learning algorithm based on parametric models and discrete search with a hierarchical and increasing control over the complexity of its induced hypotheses. On the other hand, the benefit of our method for handling concept drift is that this is a simple, well-argued, statistically-driven method and independent of the learning algorithm, which makes it broadly applicable.