Minimization of decomposable in U-shaped curves functions defined on poset chains – algorithms and applications

Author: Marcelo S. Reis1
Supervisor: Junior Barrera1

1Institute of Mathematics and Statistics – University of São Paulo
{msreis, jb}@ime.usp.br

June 7, 2013

Abstract. This article is an overview of the doctoral thesis defended by Marcelo S. Reis at the Institute of Mathematics and Statistics of University of São Paulo. His thesis features the U-curve optimization problem, which can be applied to model the classical feature selection problem in Pattern Recognition, and proves that this problem is NP-hard. Besides, the thesis presents three algorithms to tackle the U-curve problem, namely the UCS, UBB and PFS algorithms. Finally, it provides theoretical and experimental evaluation of these three algorithms, from optimal and sub-optimal experiments on simulated and real-world data. This overview summarizes the objectives of the thesis and lists its main contributions.

1. Introduction

The feature selection problem is a central problem in Pattern Recognition that has been studied for many years [Theodoridis and Koutroumbas 2006]. It is a combinatorial optimization problem that consists in finding a subset \( X \) of a finite set of features \( S \) that minimizes a cost function \( c \), defined from \( P(S) \), the power set of \( S \), to \( \mathbb{R}^+ \). If \( X \) is a subset of \( S \) and \( c(X) \) is a minimum (i.e., there does not exist another subset \( Y \) of \( S \) such that \( c(Y) < c(X) \)), then \( X \) is an optimal set of features or a set of features of minimum cost. The model of the feature selection problem is defined by a cost function that has impact in the complexity of the corresponding search algorithms. In realistic models, the cost function depends on the estimation of the joint probability distribution. The feature selection problem is NP-hard [Guyon and Elisseeff 2003], which implies that the search for a subset \( X \) of \( S \) of minimum cost maybe unfeasible (i.e., it may require a computing time that is exponential in the cardinality of \( S \)). However, for some families of cost functions, there is a guarantee that the problem is not NP-hard and, consequently, the search problem is feasible; for instance, when the cost function is submodular [Schrijver 2000]. Moreover, depending on the definition of the cost function, different search algorithms may be employed to solve the problem.

Once it is defined the feature selection model of a given problem, the next step is the choice of an appropriate feature selection algorithm. In the literature, there is a myriad of algorithms and heuristics to tackle the general form of the feature selection problem: some of them explores the branch-and-bound technique [Somol et al. 2004]. One of the most successful heuristics is the Sequential Forward Floating Selection (SFFS) [Pudil et al. 1994], which presents good results within a small computational
time [Ris et al. 2010]. However, none of those techniques explore the fact that there are cost functions that rely on the estimation of a joint probability distribution and describe “U-shaped curves” in the chains of a Boolean lattice of the feature subsets, a well-known phenomenon in Pattern Recognition: for a fixed number of samples, the increase in the number of considered features may have two consequences: if the available sample is enough to a good estimation, then it should occur a reduction of the estimation error, otherwise, the lack of data induces an increase of the estimation error. In practice, the available data are not enough for good estimations. The feature selection problem whose cost function describes U-shaped curves is also known as the U-curve problem. We present in Figure 1(a) an instance of the U-curve problem, consisting in a Boolean lattice of degree three, while in Figure 1(b), we show a graph that highlights a chain of this lattice.

![Hasse diagram](a)

![Cost curve](b)

**Figure 1.** An instance of the U-curve problem. Figure 1(a): the Hasse diagram of a Boolean lattice of degree 3 – the cost of the elements are defined by the numbers beside the nodes; the chain \{∅, c, bc, abc\} is highlighted. The element bc has minimum cost in the chain (and in the Boolean lattice). Figure 1(b): the costs of the elements of the highlighted chain describe a U-shaped curve.

There are important practical applications in Pattern Recognition whose feature selection steps may be understood as a minimization of a decomposable in U-shaped curves cost function: for instance, the identification of predictors during the estimation of Probabilistic Gene Networks (PGNs) [Barrera et al. 2007] and the design of W-operators [Martins-Jr et al. 2006]. Recently, it was proposed a new algorithm to tackle the U-curve problem: the U-CURVE algorithm, homonymous to the problem, takes into account the fact that the cost function is decomposable in U-shaped curves, as well as that the search space may be organized as a Boolean lattice. U-CURVE showed to have a satisfactory performance for several real-world instances [Ris et al. 2010].

After this Introduction, in Section 2, we will list the main objectives of the thesis. Finally, in Section 3, we will present the main theoretical and experimental results, and the possible impact of them in the scientific community.
2. Objectives of the thesis

The thesis had as one of the main objectives to study some aspects of the structure of the U-curve problem, more specifically:

- the computational complexity of the U-curve problem, which has direct impact on the performance of the algorithms that are designed to tackle this problem;

- generalizations of the U-curve problem that allow to model some problems of practical interest, for instance, the classifier design in Pattern Recognition. We studied the case in which not every chain of the Boolean lattice describes an U-shaped curve.

Starting from those studies of the structure of the U-curve problem, we accomplished another central objective of this thesis, namely the development of optimal algorithms to solve the U-curve problem and the generalization described previously. This objective involved:

- improvements on the U-CURVE algorithm presented by Ris et al.;

- development of new optimal algorithms to tackle the U-curve problem;

- theoretical and experimental evaluation of these algorithms, employing both simulated and real-world data.

These works resulted in theoretical advances in the understanding of the U-curve problem and one of its generalizations, as well as in the introduction and evaluation of three new optimal algorithms to tackle this problem. In the next section, we will summarize these results.

3. Theoretical and experimental results

We will present now the main theoretical and experimental results of the thesis. The results were organized by type of contribution: scientific and technological ones. Finally, we will discuss the possible impact of the results in the scientific community.

3.1. Scientific contributions

1. A proof that the U-curve problem is NP-hard. This result was presented in Section 3.2 of the thesis (Theorem 3.2.3).

2. A demonstration that the U-CURVE algorithm developed by Ris et al. has an error that leads to sub-optimal solutions. This result was showed in Section 4.1.

3. The introduction of the U-CURVE-SEARCH (UCS) algorithm, a corrected and improved version of the U-CURVE algorithm. UCS takes into account sufficient conditions to perform pruning in the search space without the risk of losing global minima (Propositions 4.2.1 and 4.2.2). It was also provided time complexity analysis of UCS. These results were presented in Section 4.2.

4. The U-CURVE-BRANCH-AND-BOUND (UBB) algorithm, an optimal branch and bound to solve the U-curve problem. We presented a proof of correctness and time complexity analysis of this algorithm. This contribution was made in Section 6.1.
Figure 2. comparison between UCS, UBB and PFS, in the sub-optimal experiment with simulated data. In Figure 2(a), it is showed a comparison of the average time (in log base 2) that the algorithms spent to solve different instance sizes (i.e., different sizes of the feature set $S$). For each instance size, each algorithm was executed on a hundred different instances. In Figure 2(b), it is showed the number of times that each algorithm found a best solution (i.e., for a given instance, a solution with minimum cost among the solution given by all four algorithms).

5. The introduction of the POSET-FOREST-SEARCH (PFS) algorithm, which gives optimal solutions for the U-curve problem. PFS generalizes the UBB algorithm, in the sense that the search is bidirectional (i.e., the chains of the Boolean lattice are visited in both top-down and bottom-up approach). It was presented time complexity analysis of the PFS algorithm. This result was presented in Section 6.3.

6. The execution of optimal and sub-optimal (i.e., using the algorithms as heuristics, through the adoption of a stop criterion) experiments, using simulated and real-world data. In both types of experiments and in both types of data, there was among the three algorithms a trade off between required the computational time and the number of computed nodes (in optimal experiments) or the quality of the best solution found (in sub-optimal experiments); for instance, in Figure 2, it is showed the results of the execution of the algorithms in a sub-optimal experiment on simulated data; the SFFS heuristic was used as control and had a poor performance in this experiment, since with the increase of the instance size the number of times that it found a best solution had steadily decreased. We provided an analysis of the performance of the algorithms, as well as a verification of the consistency between the theoretical time complexity analysis and the results that were obtained from the experiments. These contributions were presented in Chapter 7.

7. A generalization of the U-curve problem, with the definition of the problem of minimization of partially decomposable in U-shaped curves cost function (Problem 8.1), in which not every chain of the Boolean lattice are decomposable in U-shaped curves. It was proved that this problem models some types of classifier design problem (e.g., when the error measure is the MAE – Mean Absolute Error), as well as that the UCS, UBB and PFS algorithms solve this problem. These results were reported in Chapter 8.
3.2. Technological contributions

- The featsel framework, developed to allow the experimental evaluation of the algorithms presented in this work. The framework, developed under the Object Oriented paradigm, makes available a set of classes and methods to solve search problems that deal with a search space organized as a poset (e.g., a Boolean lattice). It also offers flexibility in programming new algorithms and cost functions. The framework was described in Appendix A.

3.3. Impact of the work in the scientific community

The main contributions of the thesis, which are now available to the scientific community, are three optimal algorithms to tackle the U-curve problem and one of its generalizations. The development of efficient forms to solve the U-curve problem is a relevant question, since the utility of this problem is not limited to the modeling of the feature selection procedure in the context of Pattern Recognition: important problems of other fields may also be solved through a reduction to the U-curve problem; for instance, in Mathematical Morphology, constrained versions of the classic Tailor problem [Reis and Barrera 2013].

References


