Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

A fast Branch-and-Bound algorithm for U-curve feature selection

Esmaeil Atashpaz-Gargari[a,1], Marcelo S. Reis[b], Ulisses M. Braga-Neto[a,c,*], Junior Barrera[d], Edward R. Dougherty[a,c]

[a] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA
[b] Center of Toxins, Immune-response and Cell Signaling (CeTICS), LECC, Instituto Butantan, São Paulo, Brazil
[c] Center for Bioinformatics and Genomics Systems Engineering, TEES, College Station, TX, USA
[d] Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 21 June 2015
Revised 9 August 2017
Accepted 11 August 2017
Available online 18 August 2017

Keywords:
Branch-and-bound algorithm
Feature selection
U-curve assumption
W-operator design

ABSTRACT

We introduce a fast Branch-and-Bound algorithm for optimal feature selection based on a U-curve assumption for the cost function. The U-curve assumption, which is based on the peaking phenomenon of the classification error, postulates that the cost over the chains of the Boolean lattice that represents the search space describes a U-shaped curve. The proposed algorithm is an improvement over the original algorithm for U-curve feature selection introduced recently. Extensive simulation experiments are carried out to assess the performance of the proposed algorithm (IUBB), comparing it to the original algorithm (UBB), as well as exhaustive search and Generalized Sequential Forward Search. The results show that the IUBB algorithm makes fewer evaluations and achieves better solutions under a fixed computational budget. We also show that the IUBB algorithm is robust with respect to violations of the U-curve assumption. We investigate the application of the IUBB algorithm in the design of imaging *W*-operators and in classification feature selection, using the average mean conditional entropy (MCE) as the cost function for the search.

1. Introduction

Feature selection is the problem of finding an optimal subset of a finite set of features that minimizes a cost function that is correlated to the classification error (e.g., the estimated classification error) [8]. Determining the optimal set of features can be a complicated task, since for a problem with n features, an exhaustive search requires considering all 2^n possible feature sets. The Cover–Campenhout theorem [3] stipulates that to be guaranteed to find the optimal feature set, no algorithm can avoid the exponential complexity of exhaustive search, in a worst-case sense, unless there is extra information about the problem.

Algorithms have been proposed that use heuristics to attempt to find the optimal feature set in fewer evaluations than exhaustive search; among them are feature selection algorithms based on the well-known Branch-and-Bound (BB) paradigm for discrete and combinatorial optimization [5,12]. A BB algorithm uses some property of the cost function, such as monotonicity, to accomplish a systematic enumeration of the features sets in the form of a *tree*. At each step of the algorithm, the tree is traversed (*Branch*) and the cost of the best feature set found until that step is recorded (*Bound*). If the cost of a node is smaller than the bound, its successor nodes are explored further and the bound is updated. Otherwise, the successors of that node can be safely discarded or *pruned*, by exploring the monotonicity of the cost. If the tree is organized in such a way that large sections of it can be pruned en masse, then the BB algorithm is successful. Different improvements have been proposed to enhance the performance of the basic BB algorithm [11]. Yu and Yuan [17] suggest avoiding the evaluation of intermediate single-branching nodes by obtaining a “minimum search tree.” Also ordering the nodes in the tree based on the significance of the features is used in some of the variants of the BB algorithm [11]. In addition, to minimize the number of cost evaluations, some algorithms use analytical properties of the search space [16].

It is well-known that the optimal classification error is monotonically nonincreasing with an increasing number of features [4], making it a perfect candidate for a cost function for a BB algorithm. However, the optimal classifier and optimal classification error are rarely known in practice, and the criterion used is typically the classification error for a classifier designed using sample data, which does not generally decrease monotonically. Rather, increas-

[*] Corresponding author at: Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.

E-mail address: ulisses@ece.tamu.edu (U.M. Braga-Neto).

[1] Present address: School of Engineering and Computing, National University, San Diego, CA, USA.

http://dx.doi.org/10.1016/j.patcog.2017.08.013
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

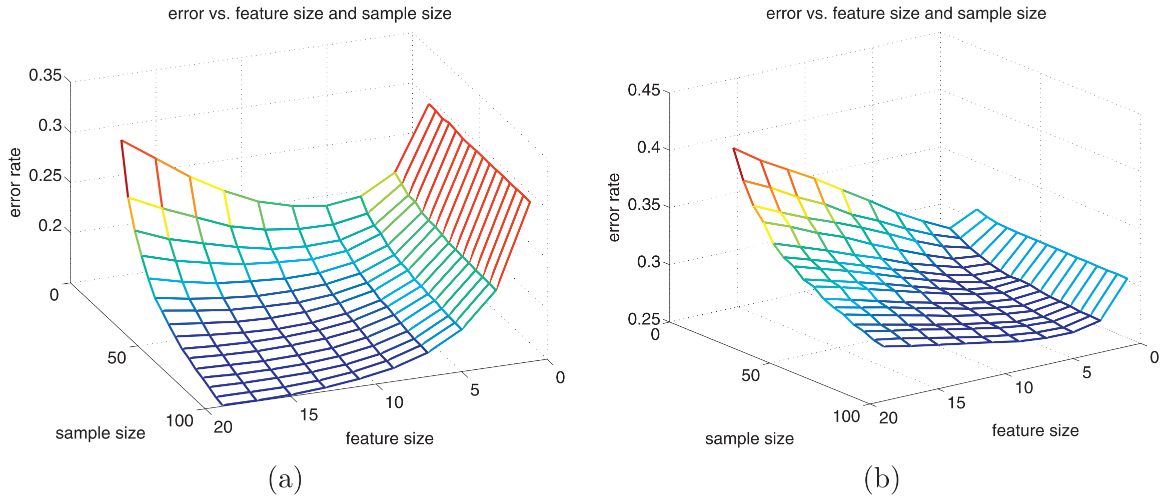

Fig. 1. Peaking phenomenon. (a) Slightly correlated features. $\rho = 0.125$. (b) Highly correlated features. $\rho = 0.5$. Reproduced from [15].

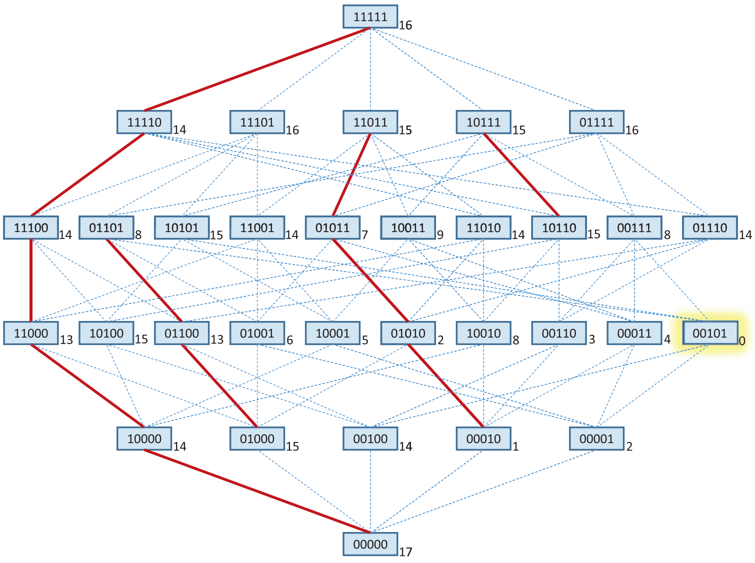

Fig. 2. Lattice for 5 features, with 4 chains highlighted in red. The cost function for this example is decomposable in U-shaped curves. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ing the number of features used to design the classifier, with a fixed sample size, generally makes the expected error of the designed classifier decrease and then increase. This is known as the *peaking phenomenon*, which was first studied in [7].

Fig. 1(a) and (b) shows the peaking phenomenon for the Linear Discriminant Analysis (LDA) classification rule. In Fig. 1(a) the features are slightly correlated. In this case, peaking occurs earlier (i.e., for a smaller number of features) or later depending on the sample size. For example, for sample size 30, peaking occurs with about 6 features, but when sample size increases to 100, peaking occurs at a larger feature size. In Fig. 1(b) the features are highly correlated. As we see in this case, even for a large sample size, peaking occurs early.

Due to the peaking phenomenon, the error of the designed classifier (as opposed to the optimal classification error) is likely to display a U-shaped behavior along a chain of increasing nested feature sets. Thus, it is reasonable to make the assumption that all the chains of the Boolean lattice that represent the search space have U-shaped behavior (U-curve assumption). The U-curve assumption was used by Ris and colleagues to formulate the *U-curve* optimization problem, which in turn can be employed to model the feature selection step of classifier design [14]. To solve this problem, the original BB algorithm, or its variants mentioned previously, are not suitable, as all of these algorithms assume that the cost function is monotone. Hence, the solution found by these algorithms will not necessarily be the globally best possible feature set. A feature selection algorithm based on a U-shaped cost function was proposed in [14]. They also presented some principles for a Branch-and-Bound procedure to tackle the U-curve problem, which were developed by Reis into the U-curve Branch-and-Bound (UBB) algorithm [13].

In this paper, we propose and evaluate a Branch-and-Bound algorithm for the U-curve optimization problem, which outperforms the original UBB algorithm, and investigate its application in the design of imaging *W*-operators and in feature selection for classifier design. Section 2 presents a formal description of the U-curve optimization problem and also reviews the original UBB algorithm. In Section 3, we introduce the Improved UBB (IUBB) algorithm, a