Detection of Minimum Biomarker Features via Bi-level Optimization Framework by Nested Hybrid Differential Evolution

Kai-Cheng Hsu
Department of Neurology,
National Taiwan University Hospital,
Taipei 10002, Taiwan
edwardfirst@gmail.com

Feng-Sheng Wang
Department of Chemical Engineering,
National Chung Cheng University,
Chiayi 62102, Taiwan
chmfsw@ccu.edu.tw

ABSTRACT

Support vector machine (SVM) using full features is a common approach for classifying diseases in healthcare systems. However, little literature reported to use it towards determining minimum features of biomarkers. This study introduced a bilevel mixed-integer optimization framework to detect minimum biomarker features for SVM. We proposed the two-population nested hybrid differential evolution (NHDE) to solve the problem. In case studies, two dominant biomarkers were found. The two-population NHDE algorithm was more efficient to achieve minimum biomarkers compared with one-population NHDE and traditional genetic algorithm.

KEYWORDS

Precision medicine, machine learning, biomarker detection, evolutionary optimization

1 INTRODUCTION

Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis and has been used in many pattern recognition problems. The SVM classifier is recently used in computational biology to discover biomarkers [1, 2]. Biomarkers using in biomedicine may increase the accuracy of diagnosis and allow disease classifications effectively targeted for precision medicine [3]. The most biomarkers are generally designed by a SVM using full features. However, such a SVM should use all features of a patient so that the diagnostic cost is expensive. Moreover, we are difficult to understand which feature or metabolite is dominant in the system because some features are correlated and dependent.

2 COMPUTATIONAL METHOD

2.1 Bi-level Optimization Framework

This study proposes a bilevel mixed-integer optimization problem (BLMIOP) to determine minimum biomarkers towards reducing diagnostic costs with similar accuracies and achieving the dominant features. This BLMIOP is referred to as the minimizing biomarker detection problem, and expressed detection problem, and expressed as

$$\max_{z_k}\left[\frac{\sum_{i\in\Omega^{TN}} T_i+\sum_{i\in\Omega^{TS}} T_i}{L}+\frac{n-\sum_{k=1}^{n} z_k}{n-1}\right] \quad (1)$$

$$\begin{cases}\min\limits_{\boldsymbol{\alpha}}\left[\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha}-\mathbf{e}^T\boldsymbol{\alpha}\right]\\ \text{subject to}\\ \sum\limits_{i=1}^{L}\alpha_i y_i=0,\\ 0\le\alpha_i\le C, i=1,...,L\end{cases}$$

where Ω^{TN} and Ω^{TS} are the set of the training and testing data, respectively. The prediction indicator T_i is defined as

$$T_i=\begin{cases}1, \text{if } \operatorname{sgn}\left(d(\mathbf{x}_i)\right)=\operatorname{sgn}\left(y_i\right); i\in\Omega^{TN}\cup\Omega^{TS}\\ 0, \text{if } \operatorname{sgn}\left(d(\mathbf{x}_i)\right)\neq\operatorname{sgn}\left(y_i\right)\end{cases} \quad (2)$$

where sgn is the signum function, and a linear discriminant function is defined as $d(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. The outer maximizing objective function in the BLMIOP is applied as a measure to select the features which can achieve higher average accuracy as the first term in the objective function and less number of features in the second term. The kernel in the matrix $\mathbf{H}$ of BLMIOP is different from the inner product of all features in the problem (1), and is in terms of the selected features, *i.e.*

$$K(\mathbf{x}_i,\mathbf{x}_j)=\sum_{k=1}^{n}(1-z_k)x_{ki}x_{kj} \quad (3)$$

where z_k is a binary variable for the k^{th} feature, that is equal to one if the feature is selected as the input.

A bilevel optimization problem (BLOP) is a special type of multi-objective optimization (MOO) problem, and the objectives between

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/10.1145/3205651.3208205

the upper and lower level are hierarchical relationships so that conventional MOO methods cannot be directly applied to solve BLOPs. Numerous conventional algorithms have been proposed to solve BLOPs; these algorithms can be classified into two categories: Kuhn–Tucker transformation and evolutionary algorithms [4]. Kuhn–Tucker algorithms have been employed to reduce a BLOP to a single-level optimization problem by using optimality conditions. However, the computation time when such an approach is used can increase exponentially when the number of decision variables is increased.

2.2 Nested Hybrid differential Evolution

A few studies have considered solving BLOPs through evolutionary optimization, and most of the methods proposed are nested in nature, as discussed in the current article [4]. Differential evolution (DE) has been applied to solve BLOPs at both inner and outer levels [5]. However, such an algorithm requires lot of computations to determine an optimal solution for large-scale BLOPs. Wang and Wu [6] have introduced a nested hybrid differential evolution (NHDE) to solve a genome-scale growth-coupled production strain design problem to overcome such a drawback. One population of individuals was employed in the NHDE algorithm to determine minimum number of knocked out genes. However, the premature minimum number could be achieved by the one population approach. This study proposes two populations of individuals in the NHDE algorithm to surmount the weakness.

The computational concept of the NHDE algorithm is based on differential evolution (HDE), which was extended from the original DE algorithm. The original version of NHDE use one population to represent the decision variables. In this study, we introduce two populations representation for NHDE to minimize the biomarker detection problem as shown in Fig.1. The first population is to code the selected dimension of features in SVM and the second one is to represent which features are applied for computing SVM classification. The core procedure of the NHDE algorithm is the "selection and evaluation" operation, which differs from DE and HDE algorithms. The selection and evaluation operation for NHDE involves two additional steps. The first evaluation step solves each SVM problem under the selected features that is posed by the inner optimization of each individual. An optimal solution for each candidate individual is achieved when the SVM problem is convergent, the set of which comprises a feasible solution to the BLMIOP. By contrast, the fitness of the outer problem is penalized, if it results in an infeasible solution. Thereafter, one-on-one fitness competition is used to select which trial individuals survive.

3 RESULTS AND DISCUSSION

We used the presynaptic dopamine overactivity and deficiency, and uric acid overexpression as case studies. The dopamine metabolic network consisted of 34 metabolites, 18 independent variables, and 68 target enzymes, and the purine metabolic network consisted of 16 dependent variables, one diet control variable, one constant variable and 28 independent variables for modulating enzyme activities. They are used to generate a set of training data in order to identify the minimum biomarkers using the proposed SVM classifier, respectively. In case studies, the accuracies of classification by SVM using full biomarkers, 34 biomarkers for dopamine metabolism and 16 biomarkers for purine metabolism, were nearly identical to that of 2 biomarkers selected by the minimizing feature approach. Furthermore, the approach could determine that the dopamine packed in vesicle in the presynaptic dopamine overactivity case and S-Adenosyl-L-homocysteine in deficient case were the dominant biomarkers, respectively. The two-population NHDE algorithm was more efficient to achieve minimum biomarkers compared with one-population NHDE and traditional genetic algorithm. In addition, NHDE can also be applied to design a genome-scale growth-coupled production strain [6]. We are recently applying the two-population NHDE to infer concogenes in a genome-scale metabolic network of hepatocytes, and will also present the results on the conference.

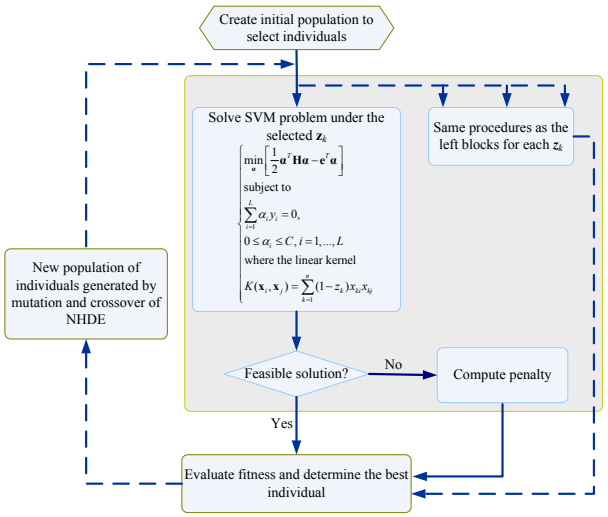

Fig. 1 Flowchart of the one/two-population NHDE algorithm for solving bilevel mixed-integer optimization problems.

ACKNOWLEDGMENTS

The financial support from Ministry of Science and Technology of Taiwan (Grant MOST106-2221-E-194-049-MY3 and MOST106-2627-M-194-001), is highly appreciated.

REFERENCES

[1] W.S. Noble 2006. What s A support vector machine? Nature biotechnology. 24(12), 1565-7.

[2] A.L. Tarca, *et al.* 2007. Machine learning and its applications to biology. PLoS computational biology. 3(6), e116.

[3] F.S. Collins and H. Varmus. 2015. A new initiative on precision medicine. N Engl J Med. 372(9), 793-795.

[4] A. Sinha, P. Malo and K. Deb. 2017. Evolutionary bilevel optimization: An introduction and recent advances. Cham: Springer International Publishing. 71–103.

[5] J.S. Angelo, E. Krempser, H.J. Barbosa, (Ed.). 2013. Differential evolution for bilevel programming. Evolutionary Computation (CEC), 2013 IEEE Congress.

[6] F.S. Wang and W.H. Wu. 2015. Optimal design of growth-coupled production strains using nested hybrid differential evolution. Journal of the Taiwan Institute of Chemical Engineers. 54, 57-63.