MULTI-OBJECTIVE APPROACH FOR SUPPORT VECTOR MACHINE PARAMETER OPTIMIZATION AND VARIABLE SELECTION IN CARDIOVASCULAR PREDICTIVE MODELING

Christina Brester[1,3], Ivan Ryzhikov[1,3], Tomi-Pekka Tuomainen[2], Ari Voutilainen[2], Eugene Semenkin[3], Mikko Kolehmainen[1]

[1]*Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, Finland*
[2]*Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland*
[3]*Institute of Computer Sciences and Telecommunication, Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia*
christina.brester@gmail.com, ryzhikov-88@yandex.ru, tomi-pekka.tuomainen@uef.fi. ari.voutilainen@uef.fi, eugenesemenkin@yandex.ru, mikko.kolehmainen@uef.fi

Keywords: Support vector machine, cardiovascular predictive modeling, multi-objective evolutionary algorithm, parameter optimization, variable selection

Abstract: We present a heuristic-based approach for Support Vector Machine (SVM) parameter optimization and variable selection using a real-valued cooperative Multi-Objective Evolutionary Algorithm (MOEA). Due to the possibility to optimize several criteria simultaneously, we aim to maximize the SVM performance as well as minimize the number of input variables. The second criterion is important especially if obtaining new observations for the training data is expensive. In the field of epidemiology, additional model inputs mean more clinical tests and higher costs. Moreover, variable selection should lead to performance improvement of the model used. Therefore, to train an accurate model predicting cardiovascular diseases, we decided to take a SVM model, optimize its meta and kernel function parameters on a true population cohort variable set. The proposed approach was tested on the Kuopio Ischemic Heart Disease database, which is one of the most extensively characterized epidemiological databases. In our experiment, we made predictions on incidents of cardiovascular diseases with the prediction horizon of 7–9 years and found that use of MOEA improved model performance from 66.8% to 70.5% and reduced the number of inputs from 81 to about 58, as compared to the SVM model with default parameter values on the full set of variables.

1 INTRODUCTION

Cardiovascular diseases (CVDs) are one of the most frequent causes of people's deaths around the world for today. Stress, unhealthy diet, physical inactivity, harmful use of tobacco and alcohol increase the risk of CVDs significantly. Nowadays, even young people suffer from CVDs. According to the report of the World Health Organization, in 2015 about 17.7 million people died from CVDs (it was 31% of all global deaths) (World Health Organization, 2017). Early detection of a high CVD risk for a patient is considered to be the main issue for doctors to undertake appropriate measures in time and prevent non-fatal or fatal incidents of CVDs.

In this paper, we develop a predictive system based on a Support Vector Machine (SVM) and a Multi-Objective Evolutionary Algorithm (MOEA), which is applied to tune SVM meta and kernel function parameters and select a proper set of input variables. There are many comparative studies in which SVMs outperform other predictive models on different problems, including medical diagnostics (Bellazzi and Zupan, 2008; Yu *et al*., 2005). Moreover, this model copes successfully with a high-dimensional set of input variables (Ghaddar and Naoum-Sawaya, 2018), which is important for our study because the data used contains 81 variables.

A number of successful studies are devoted to optimizing SVM parameters by various heuristic approaches. In most cases, however, only one-criterion optimization algorithms are used (Ren and Bai, 2010; Liao *et al*., 2015). In our study, we

employ a real-valued cooperative MOEA to optimize two criteria at once: the model predictive ability and the number of input variables (Chao and Hoang, 2017; Zhao *et al*., 2011). In epidemiology, reducing the number of input variables is quite important because it leads to fewer clinical tests and lower costs for patients.

In medical data mining studies, researches often compare their proposals with conventional models on several test problems from repositories (Brameier and Banzhaf, 2001; Cheng *et al.*, 2006; Tu *et al.*, 2009). However, our goal is not to compare different models but to take the first step in building a predictive model based on real "raw" high-dimensional data. The presented model has been trained and tested on the Kuopio Ischemic Heart Disease (KIHD) dataset, which is one of the most properly characterized epidemiological study populations with a huge variety of variables: biomedical, psychosocial, behavioral, clinical and other. Previously, some of these variables have been pre-selected and used in risk factor analysis. In our approach, we do not involve experts to pre-select explanatory variables but perform variable selection algorithmically.

The next sections describe the approach proposed in detail, experimental results, conclusions, and future plans.

2 PREDICTIVE MODELING

SVM models are widely used in machine learning to solve classification and regression problems from various practical areas (Boser *et al.*, 1992). Generally, training SVM models is performed by minimizing the error function (1):

$$\frac{1}{2}\mathbf{w}^T \cdot \mathbf{w} + C \cdot \sum_{i=1}^{N} \xi_i \rightarrow \min, \qquad (1)$$

which is subject to the constraints: $y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0, \quad i = 1,..., N$, where C is an adjustable parameter of regularization, ξ_i expresses an error $\max(0, 1 - y_i \cdot (\mathbf{w}^T \cdot \mathbf{x}_i + b))$ on training examples $(\mathbf{x}_i, y_i), y_i \in \pm 1, i = 1,..., N$, N is the number of training examples.

These models are more complex in comparison with a linear regression and allow reflecting non-linear dependencies due to the 'kernel trick', i.e. kernel functions map points into a higher dimensional space, where a linear separation is applicable. In this work, we use a Radial Basis Function as a kernel (2):

$$K(\mathbf{x}_{i1}, \mathbf{x}_{i2}) = \exp(-\sigma \cdot \left\| \mathbf{x}_{i1} - \mathbf{x}_{i2} \right\|^2). \qquad (2)$$

SVMs have a strong theoretical background and, in general, training these models is reduced to solving a dual quadratic programming problem with one global optimum.

By this time, a number of effective approaches to solve this quadratic programming problem have been proposed, for example, Sequential Minimal Optimization (SMO) (Platt, 1998). However, there are still some parameters which require proper tuning. They are meta (C) and kernel function parameters (σ) and, as can be found in other studies (Liao *et al*., 2015; Syarif *et al*., 2016), an arbitrary choice of their values may lead to a significant deterioration in the solution quality. The easiest way to tune these parameters is to use a grid search, but it requires quite a lot of computational time to check a good amount of values. As an alternative, heuristic optimization methods might be applied to find a pseudo-optimal combination of parameter values.

Evolutionary Algorithms (EA) operate with a set of candidate-solutions, which allows investigating a search space in a parallel way. One of important benefits of using EAs in optimizing SVM parameters is a possibility to incorporate variable selection into parameter tuning. This leads not only to finding proper values of SVM meta and kernel function parameters but also to determination of an effective input variable set corresponding to these particular SVM parameters. In our study, we propose to apply a MOEA to optimize two criteria simultaneously (3). The first criterion reflects the model predictive ability and the second one expresses the number of selected variables $N_{selected}$:

$$\begin{aligned} f_1 &= 1 - F\_score \rightarrow \min; \\ f_2 &= N_{selected} \rightarrow \min. \end{aligned} \qquad (3)$$

In the first criterion, we estimate the F-score metric (Goutte and Gaussier, 2005), specifically the F_1-measure with equally weighted precision and recall.

To solve this two-criterion optimization problem (3), we have developed a real-valued cooperative modification of the Strength Pareto Evolutionary Algorithm 2 (SPEA2) (Zitzler *et al*., 2002). In this algorithm, a chromosome consists of real-valued genes which code SVM parameters C, σ, and input