Comparison of Two-Criterion Evolutionary Filtering Techniques in Cardiovascular Predictive Modelling

Christina Brester[1,2], Jussi Kauhanen[3], Tomi-Pekka Tuomainen[3], Eugene Semenkin[2]
and Mikko Kolehmainen[1]

[1]Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, Finland
[2]Institute of Computer Sciences and Telecommunication, Siberian State Aerospace University, Krasnoyarsk, Russia
[3]Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland
christina.brester@gmail.com, {jussi.kauhanen, tomi-pekka.tuomainen}@uef.fi, eugenesemenkin@yandex.ru, mikko.kolehmainen@uef.fi

Keywords: Feature Selection, Two-Criterion Filtering, Cooperative Multi-Objective Genetic Algorithm, Cardiovascular Modelling.

Abstract: In this paper we compare a number of two-criterion filtering techniques for feature selection in cardiovascular predictive modelling. We design two-objective schemes based on different combinations of four criteria describing the quality of reduced feature sets. To find attribute subsystems meeting the introduced criteria in an optimal way, we suggest applying a cooperative multi-objective genetic algorithm. It includes various search strategies working in a parallel way, which allows additional experiments to be avoided when choosing the most effective heuristic for the problem considered. The performance of filtering techniques was investigated in combination with the SVM model on a population-based epidemiological database called KIHD (Kuopio Ischemic Heart Disease Risk Factor Study). The dataset consists of a large number of variables on various characteristics of the study participants. These baseline measures were collected at the beginning of the study. In addition, all major cardiovascular events that had occurred among the participants over an average of 27 years of follow-up were collected from the national health registries. As a result, we found that the usage of the filtering technique including intra- and inter-class distances led to a significant reduction of the feature set (up to 11 times, from 433 to 38 features) without detriment to the predictive ability of the SVM model. This implies that there is a possibility to cut down on the clinical tests needed to collect the data, which is relevant to the prediction of cardiovascular diseases.

1 INTRODUCTION

Nowadays, due to the tremendous capacity of data storage, it is becoming more and more popular to collect as much information as possible to design an accurate model. However, in the case of applying the model as a diagnostic tool it means a huge quantity of medical tests are needed to gather the same high-dimensional feature vector for all of the patients who should be checked. Therefore, in this study we observe a number of feature selection techniques which might be effectively used in the predictive modelling of cardiovascular diseases.

We propose several filtering schemes based on various two-objective optimization models. There are four criteria describing the relevance of attribute subsets and, combining them, we produce different feature selection techniques.

Moreover, we engage a cooperative multi-objective genetic algorithm with parallel implementation as an optimizer. It allows us to save computational time and avoid the choice of the most effective heuristic for the current problem.

The performance of the developed filtering techniques is investigated on the high-dimensional KIHD (Kuopio Ischemic Heart Disease) database. This dataset contains state vectors of 433 patients' characteristics, which were measured at the baseline time point, and information about their cardiovascular diseases, including lethal cases, approximately for the next 27 years. Support Vector Machine is applied as a predictive model.

As a result, after a comparison of four filtering schemes, we have found that using the two-criterion

Brester, C., Kauhanen, J., Tuomainen, T-P., Semenkin, E. and Kolehmainen, M.
Comparison of Two-Criterion Evolutionary Filtering Techniques in Cardiovascular Predictive Modelling.
In *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016) - Volume 1*, pages 140-145
ISBN: 978-989-758-198-4

technique based on intra- and inter-class distances, it is possible to reduce the dimensionality of the input vector from 433 down to 38 features without detriment to the predictive ability of the SVM model.

The remainder of the paper is organized as follows: in Section II there is a description of two-criterion filtering techniques for feature selection and a cooperative multi-objective genetic algorithm, which is applied as an optimizer. In Section III we describe the KIHD database. The experiments conducted, the results obtained and the main inferences are included in Section IV. The conclusions and future work are presented in Section V.

2 PROPOSED APPROACH

2.1 Two-Criterion Filtering Techniques

In general, feature selection procedures might be designed based on any of two common schemes – *filter* or *wrapper* (Kohavi *et al.*, 1997).

The filter approach operates with criteria describing a dataset from the perspective of consistency, dependency and distance metrics. It ignores model performance on the reduced feature set and, consequently, requires less computational resources because it does not involve a learning algorithm every time a possible attribute combination should be assessed.

As opposed to the filter method, the wrapper strategy uses a model (for example, a classifier) to estimate the quality of a feature subset, and therefore, it needs many more calculations. In addition to the main criterion indicating model performance (such as the relative classification accuracy), some other metrics might be included.

However, in the case of high-dimensional databases, filtering is most effective in the sense of the time spent on modelling. Therefore, in this paper we observe feature selection techniques based on the filter scheme.

In this study, we implement several two-objective filtering schemes based on diverse combinations of the following criteria (Venkatadri *et al.*, 2010):

1. *The Inter-Class Distance*:

$$IE = \frac{1}{n}\sum_{r=1}^{k} n_r d(p_r, p) \rightarrow max, \quad (1)$$

2. *The Intra-Class Distance*:

$$IA = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r} d(p_j^r, p_r) \rightarrow min, \quad (2)$$

where p_j^r is the *j*-th example from the *r*-th class, p is the central example of the data set, $d(...,...)$ denotes the Euclidian distance, p_r and n_r represent the central example and the number of examples in the *r*-th class.

3. *Attribute Class Correlation* (the dependency measure):

$$AC = \frac{\sum w_i \cdot C(i)}{\sum w_i} \rightarrow max, \quad (3)$$

$$C(i) = \frac{\sum_{j1 \neq j2} \| x_{j1}(i) - x_{j2}(i) \| \cdot \varphi(x_{j1}(i), x_{j2}(i))}{n(n-1)/2},$$

where $x_j(i)$ is the value of the *i*-th feature in the *j*-th case; *n* denotes the number of cases in the database; *m* is the number of features; w_i is equal to 1 if the *i*-th feature is selected, or 0 otherwise; $\varphi(...,...) = 1$ if the *j1*-th and *j2*-th cases are from different classes, or $\varphi(...,...) = 0$ otherwise; $\|\ldots\|$ is the module function; $i = \overline{1,m}$ and $j = \overline{1,n}$.

4. *The Laplacian Score* (the distance-based measure) (He *et al.*, 2005):

$$LS = \sum LS(i) \rightarrow max, \quad (4)$$

$$LS(i) = \frac{\tilde{x}(i)^T \cdot L \cdot \tilde{x}(i)}{\tilde{x}(i)^T \cdot D \cdot \tilde{x}(i)},$$

$$\tilde{x} = x(i) - \frac{x(i)^T \cdot D \cdot l}{l^T \cdot D \cdot l},$$

where $x(i) = [x_1(i), x_2(i), ..., x_n(i)]^T$; $l = [1,1,...,1]^T$; the *D* matrix is defined as $D = diag(S \cdot l)$; *L=D–S*; *S* is a weight matrix of the edges in the nearest neighbour graph *G*: $S_{j1,j2} = e^{-\frac{\|x_{j1}-x_{j2}\|}{t}}$, if nodes *j1* and *j2* are connected, or $S_{j1,j2} = 0$, otherwise. The *G* graph has *n* nodes: the *j*-th node