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*

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.
DOI: 10.5220/0005971101400145
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*:

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

2. *The Intra-Class Distance*:

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

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):

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

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

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; $\|...\|$ 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):

$$LS = \sum LS(i) \to max, \qquad (4)$$

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

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

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 corresponds to x_j. x_{j1} and x_{j2} are connected if x_{j1} is among *k* nearest neighbours of x_{j2} or x_{j2} is among *k* nearest neighbours of x_{j1}; *t* and *k* are adjusted parameters.

To find feature subsets meeting the introduced criteria in an optimal way, we offer to apply a multi-objective genetic algorithm with the binary representation of candidate solutions (Figure 1).

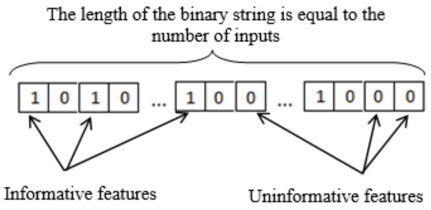

Figure 1: The binary representation of a reduced feature set.

The next subsection contains a brief description of the optimization procedure used.

2.2 Multi-Objective Genetic Algorithm

Genetic algorithms (GAs), inspired by evolutionary theory, are widely used to solve optimization problems. They imitate the alternation of generations based on principles of natural selection: a fitness function reflects an optimized criterion and genetic operators are applied to produce offspring (new candidate-solutions) and direct a search towards promising regions. Due to their unique abilities, GAs remain the only applicable tool for many complex problems: GAs can be effectively used for high-dimensional domains with different types of variables in the dynamic environment.

In this study, we consider a *Multi-Objective Genetic Algorithm* (MOGA) as an optimization procedure in filtering techniques. MOGAs allow us to take into account several criteria at once and fulfil a search on the set of reduced feature subsystems. However, designing a MOGA, researchers are faced with some issues which are related to fitness assignment strategies, diversity preservation techniques and ways of elitism implementation. It is almost impossible to know in advance which algorithm is the most suitable for the current problem. Therefore, to avoid the choice of the most effective heuristic, in our research we employ a modified MOGA, which is based on an 'island' model and includes various search strategies (Whitley *et al.*, 1997). This cooperative method comprises a number of 'islands' working in a parallel way and exchanging the best solutions from time to time.

After the analysis of different MOGAs, three algorithms were chosen as cooperation components:

- Non-Sorting Genetic Algorithm II (NSGA-II) (Deb *et al*., 2002);
- Preference-Inspired Co-Evolutionary Algorithm with goal vectors (PICEA-g) (Wang, 2013);
- Strength Pareto Evolutionary Algorithm 2 (SPEA2) (Zitzler *et al*., 2002).

These algorithms have demonstrated high effectiveness while solving many real problems, and, moreover, they are based on diverse heuristic mechanisms, which is advantageous in the sense of the abilities of the 'island' model (Table 1).

Table 1: Basic features of the MOGAs used.

MOGA	Fitness Assignment	Diversity Preservation	Elitism
NSGA-II	Pareto-dominance (*niching mechanism*) and diversity estimation (*crowding distance*)	Crowding distance	Combination of the previous population and the offspring
PICEA-g	Pareto-dominance (*with generating goal vectors*)	Nearest neighbour technique	The archive set and combination of the previous population and the offspring
SPEA2	Pareto-dominance (*niching mechanism*) and density estimation (*the distance to the k-th nearest neighbour in the objective space*)	Nearest neighbour technique	The archive set

The cooperative MOGA (Brester *et al*., 2015) includes *L* 'islands' working in a parallel way. The number of individuals *M* is shared among subpopulations equally: $M_i=M/L$, $i=1,...,L$. At each *T*-th generation 'islands' exchange the best solutions (a process called *migration*). There are two parameters: the migration size, the number of individuals for migration, and the migration interval, the number of generations between migrations. In our implementation we use a fully connected topology which implies that each algorithm sends its best solutions to all other algorithms in the model. This cooperation allows a higher level of genetic diversity to be preserved. Various heuristics might be useful in different rounds of optimization.

Then, owing to the parallel structure of the MOGA, the time spent on the algorithm execution is decreased significantly.

Additionally, its performance was thoroughly investigated on the set of test functions CEC2009 (Zhang *et al.*, 2008). The experimental results proved the high effectiveness of the cooperative algorithm, which was one more reason to apply it as an optimizer in two-criterion filtering techniques.

3 DATABASE DESCRIPTION

The epidemiologic follow-up study, KIHD, was launched in 1984 and is still continuing. It comprises of a population sample of 2,682 middle aged men recruited in 1984-1989, and 920 ageing women recruited in 1998-2001 from the city of Kuopio and its surrounding communities in Eastern Finland. The KIHD study was primarily set to study the determinants and outcomes of common non-communicable diseases in the general population. The sample is one of the most thoroughly characterized epidemiologic study populations in the world, with thousands of biomedical, psychosocial, behavioural, clinical and other variables in its dataset. Over the past 30 years, the KIHD study has proven to be a valuable source for epidemiologic research, and it has yielded over 500 original peer reviewed articles in international scientific journals. The focus in the KIHD study originally was on cardiovascular diseases, and especially on ischemic heart disease, but it has also examined a wide range of other health outcomes (Virtanen *et al.*, 2016; Kurl *et al.*, 2015; Tolmunen *et al.*, 2014).

Outcome variables and a representative subset of the features were preselected by an experienced epidemiologist as a starting point for the modelling work. This resulted in 433 variables out of several thousand possible ones that were in the database.

In this research we consider only cardiovascular diseases. The dataset was preprocessed in the following way:

- patients who had any cardiovascular problems in their anamnesis (before or at the baseline time point) were excluded so that we got feature vectors for people who were healthy at the beginning of the observation period;
- patients who had been healthy at the baseline time point but then died due to any non-cardiovascular reasons, were also excluded from our consideration;
- the final step of preprocessing resulted in two main groups of patients: people who had any cardiovascular problems for the observation period were labelled as ‘unhealthy’, and patients who did not face cardiovascular diseases for the same period were labelled as ‘healthy’.

The next section contains experimental results of predictive modelling aimed at distinguishing between these two groups of people (‘healthy’ and ‘unhealthy’).

4 EXPERIMENTS AND RESULTS

We involved Support Vector Machine (*WEKA* implementation) (Hall *et al.*, 2009) as a predictive model in our experiments: generally, based on the feature vector measured at the baseline time point (1984-1989), the model should distinguish patients, who will have cardiovascular problems (including lethal cases) for the next several decades (till 2013), from people, who will be healthy in the sense of cardiovascular diseases.

To evaluate the predictive ability of the trained model we estimate the results of the 5-fold cross-validation procedure with the F-score metric (Goutte *et al.*, 2005): the more effective the model that we use, the higher F-score value we obtain.

Firstly, we used the full feature set to adjust the SVM model (the number of features was 433). The F-score value was equal to 64.35%.

Then, according to the goal of the study, we decided to apply feature selection techniques to decrease the number of inputs as much as possible. In this experiment we investigated filtering schemes including various criteria, which were introduced in Section 2.1. Table 2 contains possible two-objective optimization models: IE+IA, IE+AC, IA+LS, and AC+LS. Also there are several combinations such as IA+AC and IE+LS which cannot be used due to the specificity of these criteria:

- While optimizing both IA and AC, it is much easier for the MOGA to find candidate-solutions which allow the Intra-Class Distance (IA) to be minimized, which implies that the evolutionary search tends to reduce the number of features as much as possible to the detriment of the Attribute Class Correlation measure (AC).

- The 'IE+LS' combination (Inter-Class Distance and the Laplacian Score) tends to keep all the features in the dataset because this variant allows both these criteria to be maximized.

Table 2: Possible combinations of the introduced criteria.

	IE	IA	AC	LS
IE	▨	+	+	☒
IA	▨	▨	☒	+
AC	▨	▨	▨	+
LS	▨	▨	▨	▨

As a result, four two-objective filtering techniques were tested in combination with the SVM model. The cooperative MOGA described in Section 2.2 was used as an optimizer.

The outcome of MOGAs is a set of non-dominated points which form the Pareto set approximation. Non-dominated candidate-solutions cannot be preferred to each other and, taking into account this fact, we propose a way to derive the final solution based on all the points from the Pareto set. In our experiments the final Pareto set approximation contained 30 points. However, GAs are heuristic and may lead to different solutions in each run. Therefore, to get statistically significant results, we ran the cooperative MOGA 15 times on the training set of every fold. So, we collected 30*15=450 binary strings which code reduced feature sets. Then, for each feature, we estimated the percentage of cases, when it was chosen, and included attributes with absolute ranks (i.e. 100%) in the final attribute set.

We repeated the same experiment for all of the two-criterion filtering schemes: the MOGA was provided with an equal amount of resources (150 generations and 150/3 = 50 individuals in each of three populations), the migration size was equal to 10 (in total each 'island' got 20 points from two others), and the migration interval was equal to 10 generations. The following types of genetic operators were defined for each component of the cooperative MOGA: binary tournament selection, uniform recombination and the mutation probability $p_m=1/n$, where n is the length of the chromosome.

Finally, we assessed the number of selected features averaged over 5 folds and F-score values achieved with the SVM model on the reduced attribute sets. The results obtained are presented in Table 3.

We found that in all of the cases the application of the SVM model after feature selection led to the same predictive ability. However, the dimensionality of input vectors varied significantly: from 37.8 to 194.4. The 'IE+IA' filtering scheme allowed us to gain feature sets with the lowest number of attributes.

Table 3: Experimental results.

Method	F-score, %	The number of features
SVM	64.35	433.0
SVM and (IE+IA)	**66.37**	**37.8**
SVM and (IE+AC)	65.28	134.2
SVM and (AC+LS)	64.80	194.4
SVM and (IA+LS)	▨	0
SVM and PCA (0.75)	65.09	100.2
SVM and PCA (0.95)	64.74	213.0

For the 'IA+LS' combination it was impossible to get the final feature set using the same strategy, because there were no attributes with absolute ranks.

We also decided to compare these two-criterion filtering techniques with Principal Component Analysis (the conventional attribute selection method) with the threshold values 0.75 and 0.95. In both cases the number of principle components was rather high (100.2 and 213.0 correspondingly). Moreover, it should be taken into account that principle components are estimated based on all of the features. Therefore, a lower number of principle components does not mean a decrease in medical tests and expenses.

5 CONCLUSIONS

In this paper we introduced a number of two-criterion filtering techniques as a feature selection tool in the predictive modelling of cardiovascular diseases. The *filter* scheme (compared to the *wrapper* one) is more beneficial in terms of the computational resources needed for its work. Also filtering relates to the preprocessing stage and so after its application various predictive models might be used. In our research we combined filtering techniques with SVM.

To optimize two criteria at once we applied the cooperative MOGA with the binary representation. This evolutionary method was based on an island model which included a number of different heuristics and, therefore, we managed to avoid the choice of the most appropriate MOGA for the current problem. Moreover, due to the parallel work of 'islands' it became possible to save computational time.

We compared four different two-criterion schemes and revealed that the usage of the 'IE+IA' combination led to a significant reduction of the feature set: from 433 to 38 attributes on average. Thus, the same predictive ability of the SVM model might be achieved with far fewer inputs and, definitely, this implies diminishing costs of clinical tests.

In addition, we are planning to use feature selection procedures to define informative attributes of various cardiovascular problems separately. This analysis may lead towards a deeper understanding of diverse diseases, determine their common risk factors and expose specific ones.

REFERENCES

Brester, Ch., Semenkin, E., 2015. Cooperative multi-objective genetic algorithm with parallel implementation. *ICSI-CCI 2015, Part I, LNCS* 9140, pp. 471–478.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6 (2), pp. 182-197.

Goutte, C., Gaussier, E., 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *ECIR'05 Proceedings of the 27th European conference on Advances in Information Retrieval Research*, pp. 345–359.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, Volume 11, Issue 1.

He, X., Cai, D., Niyogi, P., 2005. Laplacian score for feature selection. Adv. in Neural Inf. Proc. Syst., pp. 507 – 514.

Kohavi, R., John G.H., 1997. Wrappers for feature subset selection. *Artificial Intelligence*, 97, pp. 273-324.

Kurl, S, Jae, SY, Kauhanen, J, Ronkainen, K, Laukkanen, JA, 2015. Impaired pulmonary function is a risk predictor for sudden cardiac death in men. Ann Med, 47(5), pp. 381–385.

Tolmunen, T, Lehto, SM, Julkunen, J, Hintikka, J, Kauhanen, J, 2014. Trait anxiety and somatic concerns associate with increased mortality risk: a 23-year follow-up in aging men. Ann Epidemiol, 24(6), pp. 463-468.

Venkatadri, M., Srinivasa Rao, K., 2010. A multiobjective genetic algorithm for feature selection in data mining. *International Journal of Computer Science and Information Technologies*, vol. 1, no. 5, pp. 443–448.

Virtanen, JK, Mursu, J, Virtanen, HE, Fogelholm, M, Salonen, JT, Koskinen, TT, Voutilainen, S, Tuomainen, TP, 2016. Associations of egg and cholesterol intakes with carotid intima-media thickness and risk of incident coronary artery disease according to apolipoprotein E phenotype in men: the Kuopio Ischemic Heart Disease Risk Factor Study. Am J Clin Nutr, 103(3), pp. 895-901.

Wang, R., 2013. Preference-Inspired Co-evolutionary Algorithms. A thesis submitted in partial fulfillment for the degree of the Doctor of Philosophy, University of Sheffield.

Whitley, D., Rana, S., and Heckendorn, R., 1997. Island model genetic algorithms and linearly separable problems. *Proceedings of AISB Workshop on Evolutionary Computation, vol.1305 of LNCS*, pp. 109–125.

Zhang, Q., Zhou, A., Zhao, S., Suganthan, P. N., Liu, W., Tiwari, S., 2008. Multi-objective optimization test instances for the CEC 2009 special session and competition. *University of Essex and Nanyang Technological University, Tech. Rep*. CES-487, 2008.

Zitzler, E., Laumanns, M., Thiele, L., 2002. SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization. *Evolutionary Methods for Design Optimisation and Control with Application to Industrial Problems EUROGEN* 2001 3242 (103), pp. 95–100.