



**SIBERIAN STATE AEROSPACE UNIVERSITY**  
Krasnoyarsk, Russia

**UNIVERSITY OF EASTERN FINLAND**  
Kuopio, Finland



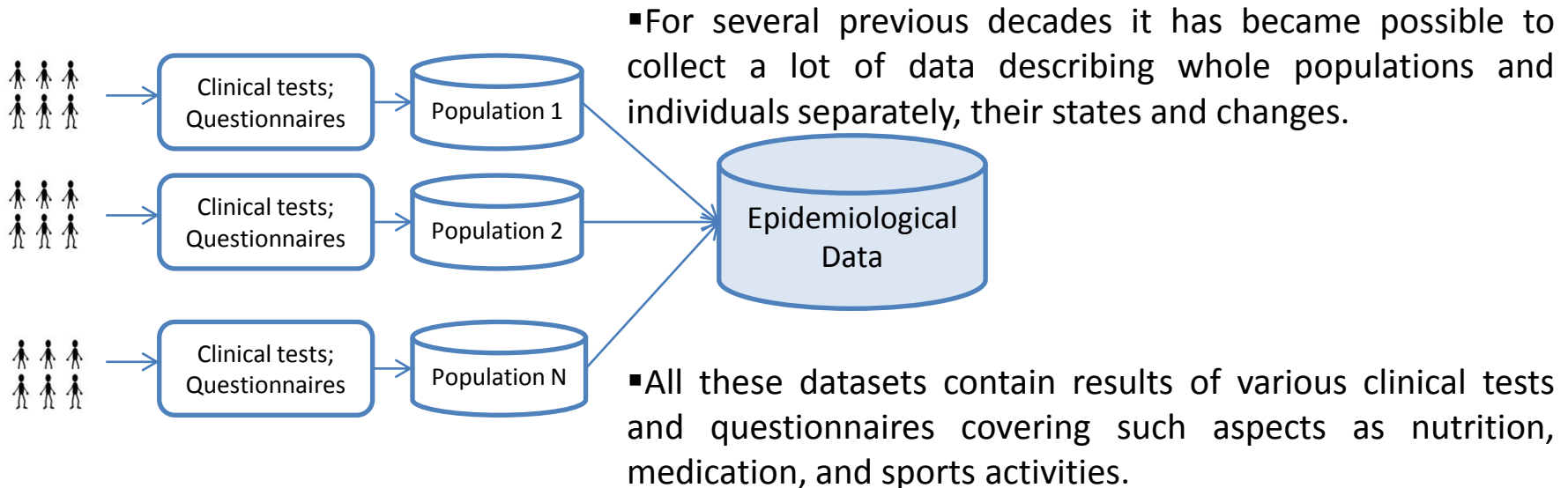
# **Comparison of Two-Criterion Evolutionary Filtering Techniques in Cardiovascular Predictive Modeling**

**Christina Brester, Jussi Kauhanen, Tomi-Pekka Tuomainen,  
Eugene Semenkin, Mikko Kolehmainen**

**Presenter: Vlamimir Stanovov**



# Motivation



## ■ Problems:

- 1) The model performance is not improved by adding more and more variables. **Redundancy of information** is becoming one of the crucial issues for epidemiologists.
- 2) 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.
- 3) **Conventional approaches** such as forward selection, backward elimination, and stepwise selection add or remove one variable at a time, therefore, their effectiveness is limited, especially if the number of features is very large.

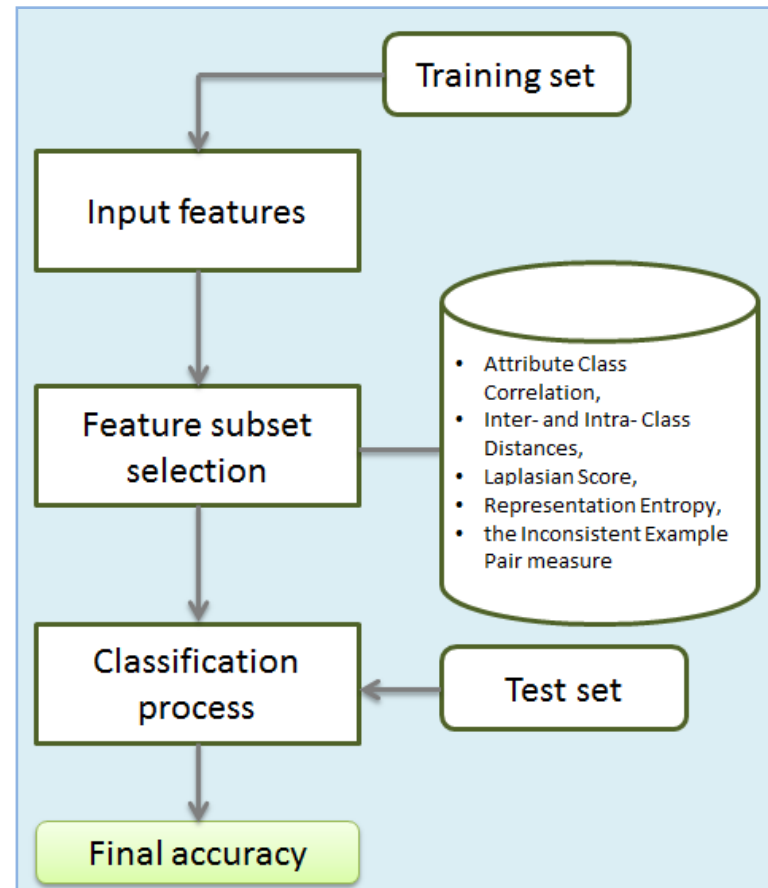
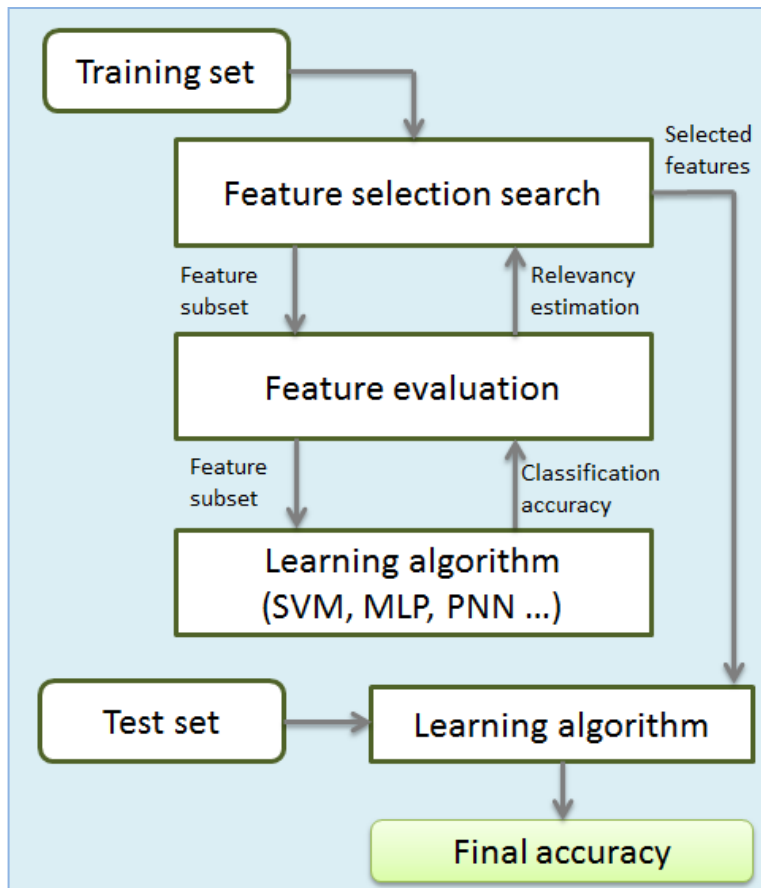
## ■ Therefore ...

High-dimensional datasets require **new effective variable selection methods** to be developed.

# Outline

- **Motivation**
  - Why do you need feature selection in epidemiological modeling?
- **Proposed Approach**
  - Two-criterion Filtering Approach
  - Cooperative Multi-Objective Genetic Algorithm
- **Database Description**
  - 'Kuopio Ischemic Heart Disease' database
- **Results and Discussion**
  - Experiments
  - Results
- **Conclusions**

# Two main feature selection concepts: Wrapper vs Filter



# Two main feature selection concepts

## Wrapper ...

- ✓ involves classification models to evaluate the relevancy of each feature subset: **adjustment to an applied classifier**;
- X requires **high computational resources**.

vs

## Filter ...

- ✓ needs significantly **fewer calculations** therefore it is rather effective in the sense of computational effort;
- ✓ might be effectively used in **combination with an ensemble** of diverse classifiers (MLP, SVM, Logit)\*;
- X **does not cooperate with a learning algorithm** and so ignores its performance entirely.

# Criteria used in our study

1. Inter-Class Distance
2. Intra-Class Distance



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,  $d(\dots, \dots)$  denotes the Euclidian distance,  $p_r$  and  $n_r$  represent the central example and the number of examples in the  $r$ -th class.

# Criteria used in our study

1. Inter-Class Distance
2. Intra-Class Distance
3. Attribute Class Correlation 

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;  $\varphi(\dots, \dots) = 1$  if the  $j1$ -th and  $j2$ -th cases are from different classes, or  $\varphi(\dots, \dots) = 0$  otherwise;  $\|\dots\|$  is the module function;  $i = 1, m$  and  $j = 1, n$ .

# Criteria used in our study

1. Inter-Class Distance
2. Intra-Class Distance
3. Attribute Class Correlation
4. Laplacian Score\*



4. The Laplacian Score (the distance-based measure):

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

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

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

where  $x(i) = [x_1(i), x_2(i), \dots, x_n(i)]^T$ ;  $l = [1, 1, \dots, 1]^T$ ; the  $D$  matrix is defined as  $D = \text{diag}(S \cdot l)$ ;  $L = D - S$ ;  $S$  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.

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



# Feature selection search

## Main concepts:

- An optimization model with **binary representation**:

1	0	0	...	1
---	---	---	-----	---

*unit* corresponds to the relevant attribute;  
*zero* denotes the irrelevant attribute.

- **Evolutionary (genetic) algorithms** as a technique for optimizing both **discrete** and **continuous** criteria.
- **The cooperation of evolutionary algorithms** as a strategy to avoid the choice of an appropriate algorithm for the problem considered.

# Multi-Objective Genetic Algorithms (MOGAs)

- Generate the **initial population**
- Evaluate **criteria** values
- While (stop-criterion!=true), do:
  - {
  - Estimate **fitness-values**;
  - Choose the most appropriate individuals with the mating **selection** operator based on their fitness-values;
  - Produce new candidate solutions with **recombination**;
  - Modify the obtained individuals with **mutation**;
  - Compose the new population (**environmental selection**);
  - }

# Multi-Objective Genetic Algorithms

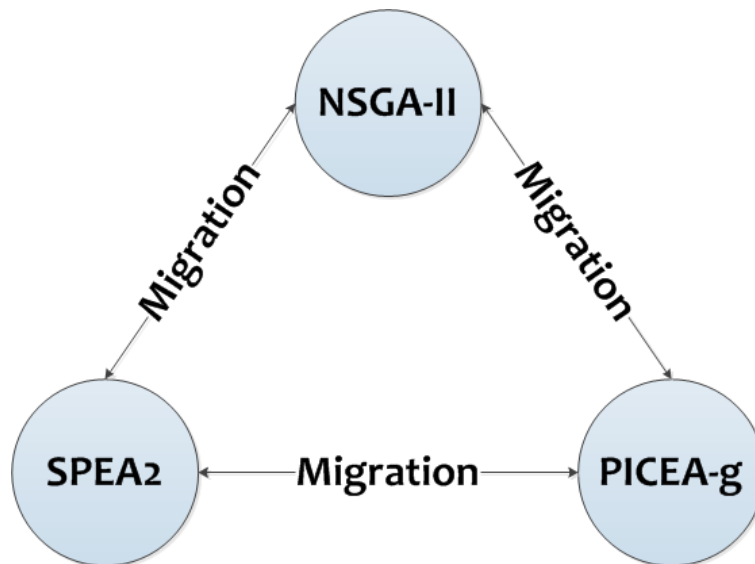
Designing a MOGA, researchers are faced with some issues:

- fitness assignment strategies,
  - diversity preservation techniques,
  - ways of elitism implementation.
- **Possible solution:** Cooperation of genetic algorithms which are based on different concepts

## Basic features of the MOGA 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

# Cooperative Multi-objective Genetic Algorithm



The island model

## Island model ...

- ✓ is based on parallel work of islands;
- ✓ has an ability to preserve genetic diversity;
- ✓ could be applied to separable problems.

At each  $T$ -th generation algorithms exchange the best solutions (**migration**).

There are two parameters:

**migration size**, the number of candidates for migration;

**migration interval**, the number of generations between migrations.

# Database Description

The **KIHD (Kuopio Ischemic Heart Disease)** study is an ongoing prospective population-based cohort study designed to investigate risk factors for CVD (cardiovascular disease), atherosclerosis and related outcomes in middle-aged men from eastern Finland, the population with one of the highest recorded rates of CHD (coronary heart disease).

→ We used this data to predict CVDs in appr. 13 years

The study population is a random sample of men living in the Kuopio city and neighbouring rural communities, stratified and balanced into four strata: 42, 48, 54, or 60 years at the **baseline** examination.

A total of 2682 participants (82.9 % those eligible), were enrolled in the study between 1984 and 1989.

• **Four-year examinations** for the KIHD study were carried out during 1991 to 1993 for 1038 men.

• **Eleven-year examinations** were carried out in 1998 to 1999 for men and women and 20-year examinations in 2006-8.

## KUOPIO ISCHAEMIC HEART DISEASE RISK FACTOR STUDY SEPELVALTIMOTAUDIN VAARATEKIJÄTUTKIMUS (SVVT, KIHD)

BL	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
<b>BASELINE</b> 1984 - 1989				<b>4-YEAR EXAMINATIONS</b> 1991-1993				<b>11-YEAR EXAMINATIONS</b> 1998-2001				<b>20-YEAR EXAMINATIONS</b> 2005-2008							
n = 2682 (men, 42-60 y)				n = 1038 (men)				n = 1774 (854 men + 920 women)				n = 1860 (1241 + 634)							
<ul style="list-style-type: none"> <li>• 4-DAY FOOD RECORDS</li> <li>• BIOCHEMICAL MEASUREMENTS</li> <li>• OTHER RISK FACTORS</li> <li>• IMT MEASUREMENT OF CAROTID ARTERY WALL</li> <li>• SCALP HAIR MeHg</li> <li>• <b>ABOUT 8000 VARIABLES</b></li> </ul>				<ul style="list-style-type: none"> <li>• FOOD FREQUENCY QUESTIONNAIRE</li> <li>• BIOCHEMICAL MEASUREMENTS</li> <li>• OTHER RISK FACTORS</li> <li>• IMT MEASUREMENT</li> <li>• GENETIC DATA</li> <li>• <b>ABOUT 5000 VARIABLES</b></li> </ul>				<ul style="list-style-type: none"> <li>• 4-DAY FOOD RECORDS</li> <li>• BIOCHEMICAL MEASUREMENTS</li> <li>• OTHER RISK FACTORS</li> <li>• IMT MEASUREMENT</li> <li>• PUBIC HAIR MeHg</li> <li>• <b>ABOUT 3000 VARIABLES</b></li> </ul>				<ul style="list-style-type: none"> <li>• <b>BIOCHEMICAL MEASUREMENTS</b></li> <li>• <b>OTHER RISK FACTORS</b></li> <li>• <b>IMT MEASUREMENT</b></li> <li>• <b>ABOUT 750 VARIABLES</b></li> </ul>							

### DURING THE FOLLOW-UP PERIOD:

- 966 ANY DEATHS
- 452 CVD DEATHS
- 597 AMI CASES
- 2335 STROKE CASES
- 588 CANCER CASES
- 199 PROSTATA CANCERS
- 83 LUNG CANCER CASES
- 262 CANCER DEATHS



# Experiments

- **Support Vector Machine (SVM)** was used as a predictive model
- We used the **5-fold cross-validation** procedure
- The result was estimated with the **F-score** measure
- We compared the predictive ability of the SVM model, firstly, **on the full feature set**, and, secondly, **after feature selection**:

# Experiments

- **Support Vector Machine (SVM)** was used as a predictive model
- We used the **5-fold cross-validation** procedure
- The result was estimated with the **F-score** measure
- We compared the predictive ability of the SVM model, firstly, **on the full feature set**, and, secondly, **after feature selection**:
  - *on the full feature set*: F-score = 64.35% (433 features)



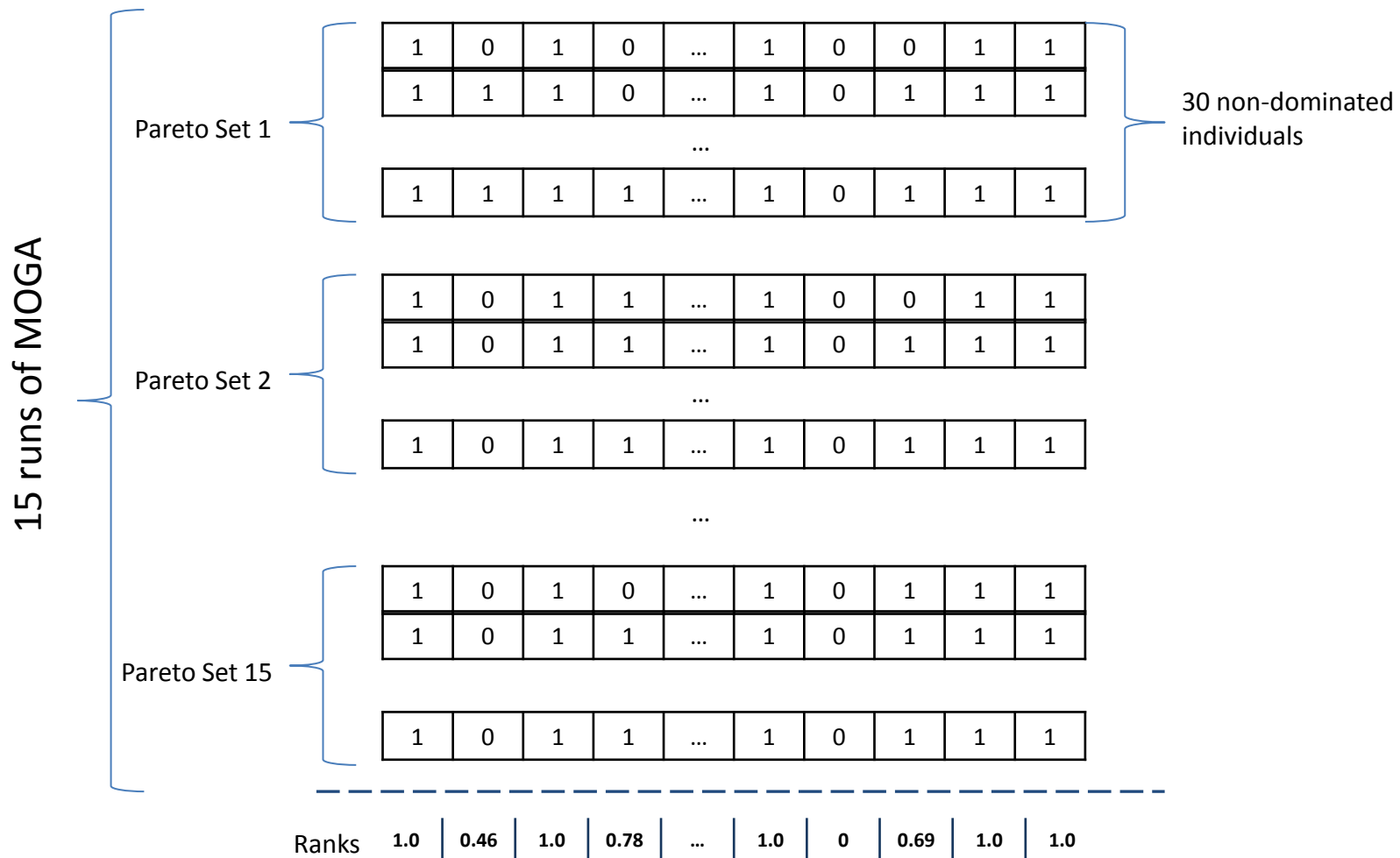
# Feature selection

Possible combinations of the introduced criteria

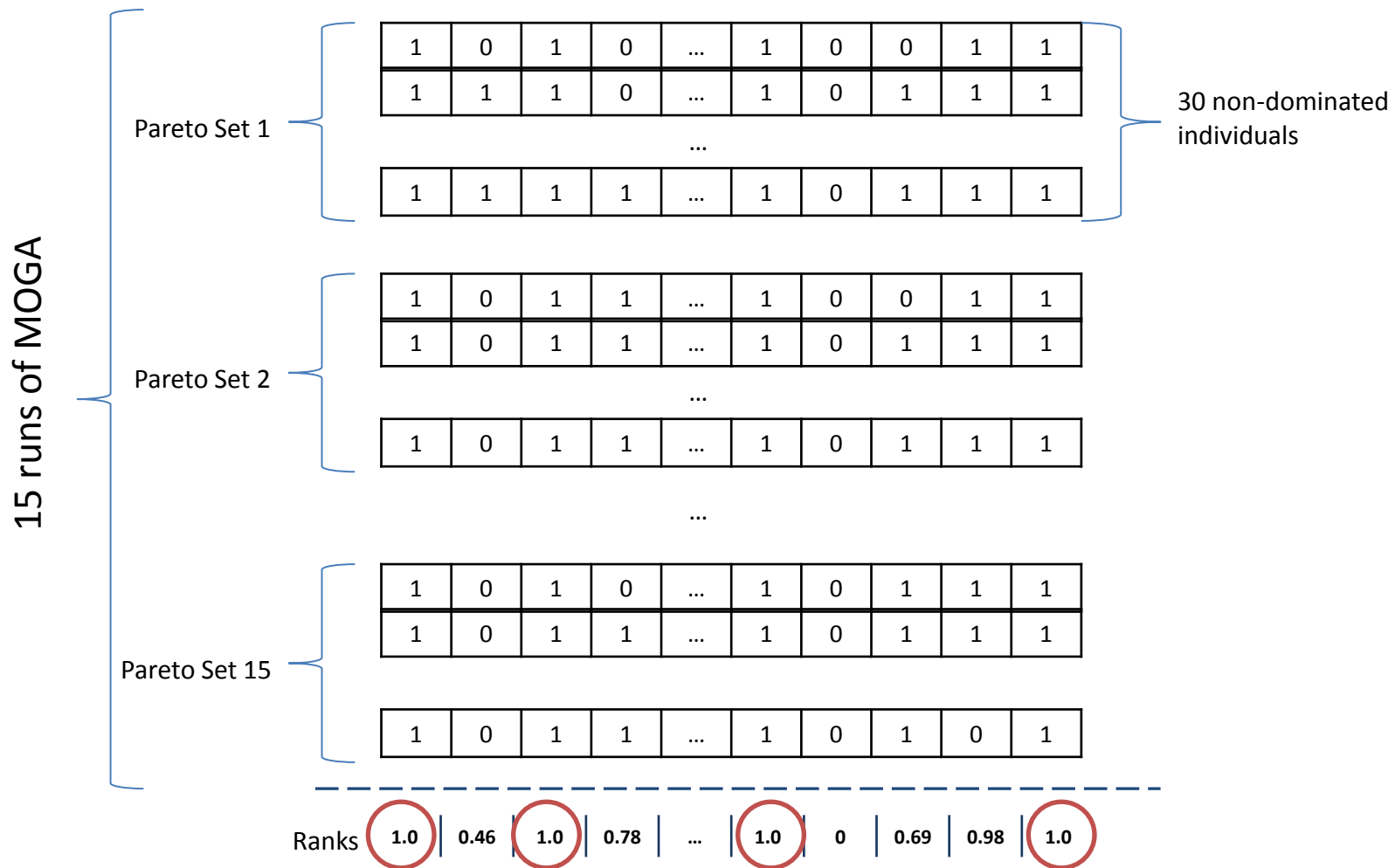
	IE	IA	AC	LS
IE		+	+	
IA				+
AC				+
LS				

- ❖ 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.

# Feature selection. How to form the final feature vector?



# Feature selection. How to form the final feature vector?



Features with absolute ranks form the final vector

# MOGA settings and resources


For each component of the MOGA (NSGA-II. PICEA-g. and SPEA2) the following settings were defined:

- binary tournament selection;
  - uniform recombination;
  - the mutation probability  $p_m=1/n$ , where  $n$  is the length of the chromosome.
- 

All islands had an equal **amount of resources**:

- **90 generations** and **150/3 = 50 individuals** in populations;
  - **the migration size** was equal to 10 (in total each island got 6 points from two others);
  - **the migration interval** was equal to 10 generations.
-

# Results



Method	F-score, %	The number of features
SVM	64.35	433.0
<b>SVM and (IE+IA)</b>	<b>66.37</b>	<b>37.8</b>
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.

# Results

Method	F-score, %	The number of features
SVM	64.35	433.0
<b>SVM and (IE+IA)</b>	<b>66.37</b>	<b>37.8</b>
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

We also compared these two-criterion filtering techniques with **Principal Component Analysis** (the conventional attribute selection method) with the threshold values 0.75 and 0.95.

\* 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.

# Conclusions

- 1) We presented a number of **two-criterion filtering techniques** as a feature selection tool in the predictive modelling of cardiovascular diseases.
- 2) 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.
- 3) 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**.
- 4) 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.

**Thanks a lot!**

