Variable selection in continuous optimization: Some possible directions

Marc Schoenauer

Équipe TAO – INRIA Futurs – Orsay, France

Marc.Schoenauer@inria.fr
Overview

- Optimization vs Classification
- Evolution Strategies
  - Adaptive and self-adaptive Gaussian mutations
  - The Covariance Matrix Adaptation
  - Variable Selection using the Covariance Matrix?
- Feature selection in Machine Learning
  - A survey
  - Feature ranking with ROGER
- Conclusion
We are interested in optimization problems

Machine Learning and Data Mining have designed many methods for Feature Selection . . .

for classification problems

So what?
Evaluate the population
Sort according to fitness
Label as *Good* the best third, and as *Bad* the worst third
Learn a classifier from those examples
Generate next population by sampling the *Good* region

Can be viewed as an *Estimation of Distribution Algorithm*, that evolve a distribution on the search space.
Two possible directions:

- Use Data Mining tools for Feature Selection on the successive classification problems as defined in LEM.

- Use the state-of-the-art Evolutionary Algorithm (CMA) that learns the Covariance Matrix of the objective function.
Overview

- Optimization vs Classification
- Evolution Strategies
  - Adaptive and self-adaptive Gaussian mutations
  - The Covariance Matrix Adaptation
  - Variable Selection using the Covariance Matrix?
- Feature selection in Machine Learning
  - A survey
  - Feature ranking with ROGER
- Conclusion
Learning from examples

- Given a set of examples
  \[(x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}\]

- Find an hypothesis \(H\) s.t.
  \[H(x_i) < 0 \text{ if } y_i = 0 \text{ and } H(x_i) > 0 \text{ if } y_i = 1\]

  or minimizing \[\sum (H(x_i) - y_i)^2\], or . . .

- Such algorithm is called a learner.

- Well-known examples:
  - ID3, AQ15, . . .
  - Neural networks, SVMs, . . .

    Symbolic learners

    Numerical learners
Feature Selection: A hot topic

- Data are growing in size
- Genetic data, medical data, Web data, ...
- Most learners do not scale up well with the number of features

Still (almost) up-to-date: 
M. Dash and H. Liu, Feature selection for classification, 

Shift of paradigm

- Find the subset of features that gives the same empirical accuracy or does not decrease accuracy too much
- Find the optimal subset of size M w.r.t. accuracy
- Find the minimal size for a given accuracy
Whatever the target, it is a combinatorial problem

- Try all subsets
- Forward selection: add features one by one
- Backward selection: remove features one by one
- Stochastic: e.g. using Evolutionary Algorithms

In any case, need for a criterion
Feature Selection: Criteria

- Use a learner to compute accuracy
  - Results depend on the learner
  - Costly
- Use a measure on the feature space
  - Entropy
  - Correlation

Wrapper method
The most discriminant feature w.r.t. class
between features
- Does not try to “fit” the data

- Optimizes the Area under the ROC curve using Evolution Strategies

- Look for $\omega_i$ and $h_i$ s.t. $H(x) = \sum |w_i x_i - h_i|$ optimizes the ranking of the examples

- Look at the weights for each feature across different evolutionary runs
Overview

- Optimization vs Classification
- Evolution Strategies
  - Adaptive and self-adaptive Gaussian mutations
  - The Covariance Matrix Adaptation
  - Variable Selection using the Covariance Matrix?
- Feature selection in Machine Learning
  - A survey
  - Feature ranking with ROGER
- Conclusion


- $\mu$ parents
- generate $\lambda$ offspring
- using normal mutations

\[ X := X + \sigma N(0, C) \]

- deterministically choose who will survive
  - Best $\mu$ among $\lambda$ offspring
  - Best $\mu$ among $\mu$ parents plus $\lambda$ offspring

**Issue:** Tune $\sigma$ (the step-size) and $C$ (the covariance matrix)
Adaptation of Gaussian mutation

History

- \( \sigma \propto t^{-1} \)
  Not adaptive

- \( \sigma \propto t^{-1} \) fitness
  Adaptive, individual

- The 1/5\( ^{t} \)h rule: Modify \( \sigma \) w.r.t. # successful mutations
  Adaptive, population

- Self-adaptive mutations, allele or individual

- Covariance Matrix Adaptation
  Adaptive, population
Self-adaptive mutations

- **Isotropic:** One $\sigma$ per variable, $C = I_d$

\[
\begin{align*}
\sigma &:= \sigma \, e^{\tau N_0(0,1)} \\
X_i &:= X_i + \sigma N_i(0,1) \quad i = 1, \ldots, d
\end{align*}
\]

- **Non-isotropic:** $d$ $\sigma$’s per individual, $C = \text{diag}(\sigma_1, \ldots, \sigma_d)$

\[
\begin{align*}
\kappa &= \tau N_0(0,1) \\
\sigma_i &:= \sigma_i \, e^{\kappa + \tau' N_i(0,1)} \quad i = 1, \ldots, d \\
X_i &:= X_i + \sigma_i N_i'(0,1) \quad i = 1, \ldots, d
\end{align*}
\]

$N_i$ and $N_i1$ are independent
Self-adaptive mutations

- **Correlated**: \( C \) positive definite:

\[
\mathbf{N}(0, C(\bar{\sigma}, \bar{\alpha})) = \prod_{i=1}^{d-1} \prod_{j=i+1}^{d} R(\alpha_{ij}) \mathbf{N}(0, \bar{\sigma})
\]

\[
\begin{align*}
\sigma_i &= \sigma_i e^{\tau' N_0(0,1) + \tau N_i(0,1)} & i &= 1, \ldots, d \\
\alpha_j &= \alpha_j + \beta N_j(0,1) & j &= 1, \ldots, d(d - 1)/2 \\
\mathbf{X} &= \mathbf{X} + \mathbf{N}(0, C(\bar{\sigma}, \bar{\alpha}))
\end{align*}
\]

- From Schwefel: \( \tau \propto \frac{1}{\sqrt{2 \sqrt{d}}} \), \( \tau' \propto \frac{1}{\sqrt{2d}} \), \( \beta = 0.0873 \ (=5^\circ) \)

Isotropic mutation

Non–isotropic mutation

Correlated mutation
Experiments on dynamic landscape

Slightly elliptic function with random moves of minimum every $K$ generations

Fitness and $\sigma_{15}$ for non-isotropic mutation
Self-adaptivity – discussion

- Stability: Correlated mutation and DE require crossover
- Sensitivity to the characteristic basis of the fitness: Correlated mutation (and DE) perform poorly on rotated (elliptic) functions
- Speed: Adaptation can be very slow

But what covariance matrix should be learned?

Good reasons to believe it’s \( \left( \frac{1}{2} H \right)^{-1} \)

\( H \) Hessian matrix of fitness
What does SA-ES learn?

$\frac{1}{2} \sum_{i=1}^{n} (10^6)^{\frac{i-1}{n-1}} x_i^2$

$(\frac{1}{2}H)^{-1} = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 10^{-6}
\end{pmatrix}$

Fitness and square-root of eigenvalues for SA-ES on $f_H$

$n = 10$
The actual path contains local information on the landscape.

It is lost through the self-adaptive mutation process.

→ **Derandomized Evolution Strategies**

Consecutive steps in colinear directions → increase step-size and vice-versa.

Add direction information to the covariance matrix.

Also, use a \((\mu/\mu, \lambda) - ES\) better with small populations Scheel, 85

i.e. generate offspring from

\[ <X>^{n+1} = \sum_{i=1}^{\mu} w_i X^n_{i:\lambda} \]

\[ X^n_{i:\lambda}, i = 1, \ldots, \mu: \text{best } \mu \text{ offspring from the } \lambda \text{ mutations of } <X>^n \]
A $(\mu/\mu, \lambda) - ES$ with covariance matrix $I_n$

- Compute the cumulative path $p_n^n$ using

$$p_{\sigma}^{n+1} = (1 - c_\sigma) p_{\sigma}^n + \sqrt{c_\sigma(2 - c_\sigma)} \frac{<X>^{n+1} - <X>^n}{\sigma^n}$$

- Update the step-size by e.g. isotropic

$$\sigma^{n+1} = \sigma^n \exp \left( \frac{1}{d_\sigma} \left( \frac{||p_{\sigma}^{n+1}||}{E(||N(0, I_d)||)} - 1 \right) \right).$$

- Rationale:
  - if $p_{\sigma}^n \sim N(0, I_d)$ and $\frac{<X>^{n+1} - <X>^n}{\sigma^n} \sim N(0, I_d)$
    and they are independent,
  then $p_{\sigma}^{n+1} \sim N(0, I_d)$ e.g. $\lambda = \mu$
  - if there is no selection
    then $\sigma^{n+1} = \sigma^n$ Nothing should happen
Covariance Matrix Adaptation

A $(\mu/\mu, \lambda) - ES$ with full covariance matrix $C^n$

- Update the (global) step-size
  \[
  p_{\sigma}^{n+1} = (1 - c_\sigma)p_{\sigma}^n + \sqrt{c_\sigma(2 - c_\sigma)}(C^n) - \frac{1}{2} <X>_{\mu}^{n+1} - <X>_{\mu}^n \\
  \sigma^{n+1} = \sigma^n \exp \left( \frac{1}{d_\sigma} \left( \frac{||p_{\sigma}^{n+1}||}{E(||\mathcal{N}(0, I_d)||)} - 1 \right) \right).
  \]

- Rationale: idem CSA with a full covariance matrix $C^n$

- Note: $E[||\mathcal{N}(0, I_d)||] = \sqrt{2\Gamma\left(\frac{n+1}{2}\right)/\Gamma\left(\frac{n}{2}\right)}$ is approximated practically by $\sqrt{d}(1 - \frac{1}{4d} + \frac{1}{21d^2})$
Update the Covariance Matrix:

compute the cumulated path

\[ p_{n+1}^c = (1 - c_c)p_n^c + \sqrt{c_c(2 - c_c)} \frac{<X>_{n+1}^\mu - <X>_{n}^\mu}{\sigma_n^\mu} \]

\[ C_{n+1} = (1 - c_{cov})C_n + c_{cov}p_{n+1}^c p_{n+1}^c T \]

Rationale:

- \( p_{n+1}^c \) is (roughly) the descent direction
- \( C_n \) is updated with the rank 1 matrix \( p_{n+1}^c p_{n+1}^c T \) whose eigenvector is \( p_{n+1}^c \)
Use all $\mu$ best offspring to update $C^n$:

$$U^{n+1} = \sum_{i=1}^{\mu} \frac{(X_i:\lambda - <X>^n_\mu)(X_i:\lambda - <X>^n_\mu)^T}{(\sigma^n)^2}$$

Increase the speed of adaptation in high dimensions

CMA parameters

$$c_c = \frac{4}{d + 4}, \quad c_\sigma = \frac{10}{d + 20}, \quad d_\sigma = \max(1, \frac{3\mu}{d + 10}) + \frac{1}{c_\sigma}$$

$$c_{\text{cov}} = \frac{1}{\mu} \left( \frac{2}{(d + \sqrt{2})^2} + (1 - \frac{1}{\mu}) \min(1, \frac{2\mu - 1}{(d + 2)^2 + \mu}) \right)$$

with initial values $p^0_\sigma = 0, p^0_c = 0$ and $C^0 = I_d$. 
CMA-ES at work

Sphere function, $n = 2$, initial point $(0, 10^9)$. Fitness, (global) step-size, sign($x_2$) and sqrt(eigenvalues)
What does CMA-ES learn?

Fitness and square-root of eigenvalues for CMA-ES on $f_H$

$n = 10$

$$f_H = \frac{1}{2} \sum_{i=1}^{n} (10^6)^{\frac{i-1}{n-1}} x_i^2$$

$$(\frac{1}{2}H)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \cdots & 10^{-6} \end{pmatrix}$$
Toward variable selection

Idea

- Once the covariance matrix has been learned
- select the eigenvectors with the smallest eigenvalues

But

- How good is the approximation? Error criterion Auger, PhD 04
- What threshold? Cross-validate with other measures Entropy, covariance, ...
- A moving target The interesting variables might change along evolution
Nothing yet