2014 IEEE Congress on Evolutionary Computation
(World Congress on Computational Intelligence)

GP-Based Kernel Evolution for L_2-Regularization Networks

Scardapane S., Comminiello D., Scarpiniti M. and Uncini A.

intelligent signal processing and multimedia lab

WCCI 2014

Aim of the presentation

State of the art:

- In *kernel-based* learning, choosing an optimal kernel is an intractable problem.
- Several *Genetic Programming* (GP) algorithms have been proposed to solve it, mostly using *Support Vector Machine* classifiers.

This presentation:

- We present a GP-based algorithm that we apply to a faster kernel algorithm, namely L_2-Regularization Networks (RN).
- We study the problem of *diversity* in the context of kernel evolution.

Table of Contents

Supervised learning

Given a *training set* $S = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, find $f \in \mathcal{H}$ such that:

$$f(\mathbf{x}) \approx y$$

for any unseen pair $(\mathbf{x}, y)$. In the *regularization framework* we have:

$$f^* = \min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2 \tag{1}$$

where $L(\cdot, \cdot)$ is a *loss function* and $\|f\|_{\mathcal{H}}$ the norm in $\mathcal{H}$.

Kernels

If $\mathcal{H}$ is a *Reproducing Kernel Hilbert Space* (RKHS), the *Representer's Theorem* asserts that solutions to Eq. (1) have the form:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}, \mathbf{x}_i) \tag{2}$$

where $k(\cdot, \cdot)$ is the kernel associated to the RKHS, and $\alpha_i, i = 1, \ldots, N$ a set of real scalars.
Conversely, Eq. (2) holds if and only if, for any $\mathbf{z} \in \mathcal{X}$ and *kernel matrix* $\mathbf{K}$, we have:

$$\mathbf{z}^T \mathbf{K} \mathbf{z} \geq 0$$

i.e., $\mathbf{K}$ is positive semi-definite (PSD).

L_2-Regularization Networks

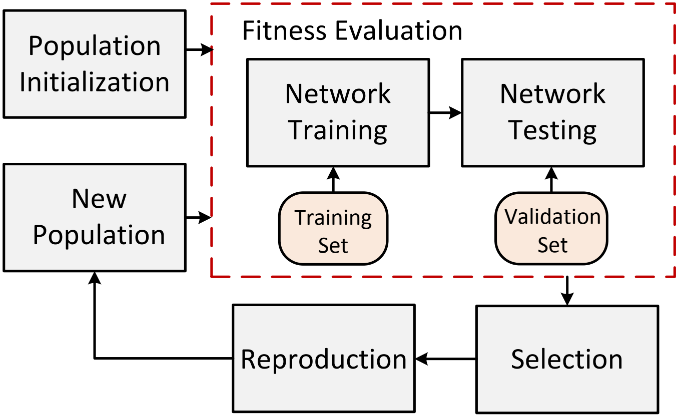
For the *quadratic loss* $L(y, d) = (y - d)^2$ coefficients are computed as:

$$(\mathbf{K} + \lambda \mathbf{I}_N)\boldsymbol{\alpha} = \mathbf{y} \tag{3}$$

Choosing the optimal kernel function is an optimization problem over the space $\mathcal{K}$ of PSD kernels. Possible approaches are:

- Choosing a "standard" kernel (e.g. Gaussian).
- Linearly combining standard kernels (*Multiple Kernel Learning* (MKL)) [XWD13].
- Optimizing the system using a *Genetic Programming* algorithm [DRP10].

Block Diagram of the Algorithm



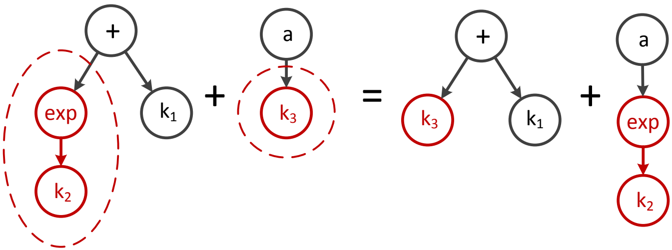
Kernel algebra

A possible kernel is represented as a tree, where:

- A leaf is a basic kernel, i.e. Gaussian, Linear, Polynomial, Hyperbolic.
- An internal node is an operation over one or two kernels:
  - Sum of two kernels: $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$,
  - Scaling of a kernel: $k(\mathbf{x}, \mathbf{y}) = a \times k_1(\mathbf{x}, \mathbf{y})$,
  - Shifting of a kernel: $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) + a$,
  - Kernel product: $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) \times k_2(\mathbf{x}, \mathbf{y})$,
  - Exponentiation of a kernel: $k(\mathbf{x}, \mathbf{y}) = \exp\{k_1(\mathbf{x}, \mathbf{y})\}$,

Kernel algebra ensures that, if k_1 and k_2 are PSD, the resulting kernels are also PSD.

Individuals and Crossing-Over



Peculiarities of our algorithm

1. The use of RNs as basic learning method. In this way, longer evolutions can be considered.
2. More importance is given to basic kernels during initialization.
3. A post-processing phase for fine-tuning the λ parameter.
4. Kernels giving rise to ill-posed problems are easily excluded by checking the 2-*norm condition number* r of the $\mathbf{K}$ matrix.

UCI Datasets

Table : Experimental results. Misclassification rate is shown for datasets I, W and YE, while NRMSE for datasets G and YA.

Dataset	RN-G	RN-P	RN-L	RN-GP
G	0.51	0.52	0.53	**0.49**
I	0.08	0.10	0.19	**0.05**
W	**0.02**	0.06	0.06	**0.02**
YA	0.10	0.05	0.61	**0.04**
YE	0.47	0.49	0.49	**0.45**

Kernels for the W dataset

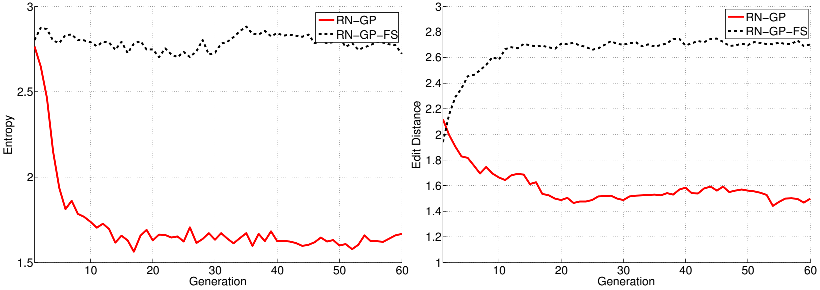
On the W dataset, the GP procedure is consistently converging to a basic Gaussian kernel:

Fold	Kernel
1	$k_1(x,y) = \exp\left\{-0.9\|x-y\|^2\right\}$
2	$k_2(x,y) = \exp\left\{-0.2\|x-y\|^2\right\}$
3	$k_3(x,y) = \exp\left\{-0.2\|x-y\|^2\right\} \cdot \exp\left\{-0.6\|x-y\|^2\right\} \cdot \exp\left\{-0.4\|x-y\|^2\right\}$
4	$k_4(x,y) = 0.85 * \tanh\{0.0002 * \{x,y\} + 1\} \cdot \exp\left\{-0.2\|x-y\|^2\right\} \cdot \exp\left\{-0.6\|x-y\|^2\right\}$
5	$k_5(x,y) = \exp\left\{-0.9\|x-y\|^2\right\} \exp\left\{-0.7\|x-y\|^2\right\}$

Diversity in GP

1. Diversity in a GP algorithm relates to the variety of individuals in a population. It is known that a fast decrease in diversity in the early iterations of the GP procedure can be correlated with poor performance [EN02].
2. We computed diversity using the *entropy*, the *edit-2 distance* (E2D) between each individual and the best solution found so far, and the *phenotypic diversity* (PD).
3. Moreover, we implemented a *diversity-preserving* mechanism in our procedure [EN02].

Typical Evolution for the I dataset



(a) Entropy for the I dataset

(b) Average E2D with best individual for the I dataset

However, adaptively maintaining diversity does not seems to increase performance of the method.

References

L. Dioşan, A. Rogozan, and J.-P. Pecuchet.
Improving classification performance of Support Vector Machine by genetically optimising kernel shape and hyper-parameters.
Applied Intelligence, 36(2):280–294, October 2010.

A. Ekárt and S. Z. Németh.
Maintaining the diversity of genetic programs.
In *Genetic Programming*, volume 2278 of *Lecture Notes in Computer Science*, pages 162–171. Springer Berlin Heidelberg, 2002.

T. Evgeniou, M. Pontil, and T. Poggio.
Regularization networks and support vector machines.
Advances in Computational Mathematics, 13:1–50, 2000.

T. Hofmann, B. Schölkopf, and A. J. Smola.
Kernel methods in machine learning.
The Annals of Statistics, 36(3):1171–1220, June 2008.

X. Xu, Tsang I. W., and Xu. D.
Soft margin multiple kernel learning.
IEEE Transactions on Neural Networks and Learning Systems, 24(5):749–761, May 2013.