A Method for Regularization of Evolutionary Polynomial Regressions

Francisco Coelho[a,c,*], João Pedro Neto[b,c]

[a] *Dept. Informática, Universidade de Évora, Rua Romão Ramalho 58, 7000-671 Évora*
[b] *Dept. Informática, Faculdade de Ciências da Universidade de Lisboa, Campo Grande 1749-016 Lisboa*
[c] *Laboratory of Agent Modelling (LabMAg)*

Abstract

While many applications require models that have no acceptable linear approximation, the simpler nonlinear models are defined by polynomials. The use of genetic algorithms to find polynomial models from data is known as Evolutionary Polynomial Regression (EPR). This paper introduces Evolutionary Polynomial Regression with Regularization (EPR^2), an algorithm that extends the EPR method and describes a set of experiences on common datasets that compare both flavors of EPR and other methods including Linear Regression, Regression Trees and Support Vector Regression.

The empiric conclusion of those experiments is that EPR^2 is able to achieve better fitting than other non-ensemble methods and it has shorter computation time than plain EPR.

Keywords: evolutionary polynomial regression, regularization, feature extraction, dimensionality reduction

1. Introduction

With notable exceptions (*e.g.* neural networks) machine learning regression techniques produce linear models. The linearity assumption has many advantages including reduced computational complexity and strong theoretical framework. However nonlinearity is unavoidable in many application

*Corresponding author
Email addresses: fc@di.uevora.pt (Francisco Coelho), jpn@di.fc.ul.pt (João Pedro Neto)

scenarios, specially those with phase transitions or feedback loops, so common in engineering, ecology, cybernetics and other areas. The kernel trick in Support Vector Machines (SVM) (Schölkopf et al. (1997); Liang and Lee (2012); Bao et al. (2013)) alleviates this problem by allowing special nonlinear transformations of the feature-space. The condition such transformations must meet is known as the *kernel trick*, $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$, where φ is the feature-space transformation and $\langle \cdot, \cdot \rangle$ denotes inner product. The "trick" consists on computing the kernel $k(x, x')$ while avoiding the computation of the inner product and the transformations $\varphi(x), \varphi(x')$. A special case of polynomial transformation, the *polynomial kernel*, $k(x, x') = \langle x, x' \rangle^d$ is commonly used in regression and classification tasks with SVMs. However general polynomial transformations do not verify the kernel trick.

Polynomials, one of the most studied subjects in mathematics, generalize linear functions and define, perhaps, the simplest and most used nonlinear models. Applications include colorimetric calibration (Mendes and Carvalho, 2005), explicit formulæ for turbulent pipe flows (Davidson et al., 1999), computational linguistics (Sánchez et al., 2009) and more recently analytical techniques for cultural heritage materials (Cséfalvayová et al., 2010), liquid epoxy moulding process (Chan et al., 2011), B-spline surface reconstruction (Gálvez et al., 2012), product design (Chan et al., 2012) or forecasting cyanotoxins presence in water reservoirs (García Nieto et al., 2013). These examples not only illustrate the wide spectrum of applications but, additionally, in each one uses, at some point, a Genetic algorithms (GA).

GAs were, arguably, one of the hottest topics of research in the recent decades and with good reason since they outline an optimization scheme easy to conceptualize and with very broad application. If a nonlinear (or otherwise) model requires parameterization, GAs provide a simple and often effective approach to search for locally optimal parameters. Research related to genetic algorithms abound and spans from the 1950s seminal work of Nils Aall Barricelli (Barricelli, 1962) in the Institute for Advanced Study of Princeton to today's principal area of study for thousands of researchers, covered in hundreds of conferences, workshops and other meetings. Perhaps the key impulse to GAs come from John Holland's work and his book "Adaptation in Natural and Artificial Systems" (Holland, 1975).

One interesting variation of genetic algorithms, named *genetic programming* by John Koza (Koza, 1992), proposes the use of GAs to search the syntactic structure of complex functions. Syntactic structure search is also keen to the central ideas of deep learning (Bengio, 2009; Bengio et al., 2013),

a subarea of machine learning actually producing quite promising results (*e.g.* in Tarlow et al. (2013)). It is also related to the work presented in this paper in the sense that, unlike linear models that have a simple structure, $y = \sum_i \beta_i x_i$, nonlinear (in particular polynomial) models pose an additional structure search problem.

The idea of using GAs to find a polynomial regression is not new (Maertens et al., 2006; Yu and Lin, 2008; Wu et al., 2009) but still generates original research (Hofwing et al., 2011; Cetisli and Kalkan, 2011). The modern formulation of the use of GA to find polynomial models is known as Evolutionary Polynomial Regression (EPR) and systematization can be traced back to the work of Davidson, Savic and Walters (Davidson et al. (2003)). Further developments include multi-objective optimizations (Giustolisi and Savic (2009)).

This paper describes an extension of the general EPR method to find a regularized polynomial regression of a given dataset. The optimal regression results from a cost function that accounts for both the root-mean-square (error) and a regularization factor to avoid overfitting. [**Note**[1]]

The next section describes the method's details and is followed by a presentation of some performance results. The last section draws some conclusions and points future research tasks.

2. Genetic Algorithms for Polynomials

This section starts with a brief introduction and outline of the evolutionary polynomial regression algorithm, EPR, and proceeds into core details as the encoding used to represent individual polynomial instances in the GA populations and the regularization of the cost function.

An usual representation of polynomials is through expressions of the form

$$p\left(x_1, \ldots, x_m\right) = \sum_i \theta_i q_i$$

where each $q_i = \prod_j x_j^{\alpha_{ij}}$ is a monomial, the exponents $\alpha_{ij} \in \mathbb{N}_0$ are non-negative integers and the coefficients $\theta_i \in \mathbb{R}$ are real valued. For example $p\left(x_1, x_2, x_3\right) = 2x_1 + x_2x_3 + \frac{1}{2}x_1^2x_3$ has monomials $q_1 = x_1, q_2 = x_2x_3$ and $q_3 = x_1^2x_3$, exponents $\alpha_{1,1} = 1, \alpha_{2,2} = 1, \alpha_{2,3} = 1, \alpha_{3,1} = 2, \alpha_{3,3} = 1$ and all other $\alpha_{ij} = 0$ and coefficients $\theta_1 = 2, \theta_2 = 1$ and $\theta_3 = 1/2$.

[1] **Note:** Explain a bit more the overfitting avoiding method.

The exponents alone can be organized into a matrix $[\alpha_{ij}]$ that defines the monomial structure of the polynomial. For the example above the matrix representation of the monomials is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} x_1 \\ x_2 x_3 \\ x_1^2 x_3 \end{bmatrix}$$

where each row defines a monomial and each column represents a variable. Changing the order of the rows doesn't change the polynomial whereas changing the order of the columns corresponds to changing the respective variables.

This partial representation of polynomials makes the problem of structure search very clear: except for the trivial cases, the number of possible monomials given n variables and a maximum joint degree d grows exponentially with either n or d. But more importantly, by separating the set of monomials from the coefficients, the polynomial regression problem can be naturally split into two subproblems:

1. For a given set of monomials $\mathcal{Q} = \{q_1, \ldots, q_k\}$ find the regression coefficients $\Theta = \{\theta_1, \ldots, \theta_k\}$ that minimize the error on a given dataset;
2. Find the fittest set of monomials, *i.e.* the polynomial that minimizes the error on the same dataset;

More precisely, concerning the first problem, let $\mathcal{D}$ be a dataset with n observations of variables $Y, X_1, \ldots, X_m$ and $\mathcal{Q} = \{q_1, \ldots, q_k\}$ a set of k monomial expressions over $X_1, \ldots, X_m$. Define the hypothesis[2]

$$h_{\Theta,\mathcal{Q}}(x_1, \ldots, x_m) = \sum_{j=1}^{k} \theta_j q_j|_{X_i = x_i, \forall 1 \leq i \leq m}$$

and let the error (as "cost") be

$$J_{\text{fit}}(\Theta; \mathcal{Q}, \mathcal{D}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - h_{\Theta,\mathcal{Q}}\left(x_1^{(i)}, \ldots, x_m^{(i)}\right)\right)^2} \quad (1)$$

[2]The expression "$q|_{X=x}$" reads "*q with all instances of X replaced by x.*"

Algorithm 1 This EPR algorithm uses linear regression for the calculation of the error J and the space of polynomials is searched in the GAs iteration step. At exit the error of the fittest instance is bounded by ϵ or the maximum number of allowed iterations.

```
function EPR(D, pop_0, ε, maxiter)
    pop ← pop_0; err ← 1.0 + ε
    while err > ε ∧ iterations < maxiter do
        pop ← IterateGA(pop)
        pop ← Sort(pop, key = J)    ▷ Sort population by regression error
        err ← J(First(pop))
    end while
    return First(pop)
end function
```

the usual root-mean-square (error) function. Now the first problem can be stated as: *Given a dataset $\mathcal{D}$ and a set of monomials $\mathcal{Q}$ find parameters Θ that minimize the cost $J_{fit}(\Theta; \mathcal{Q}, \mathcal{D})$.*

This is a simple linear regression problem obtained by expanding $\mathcal{D}$ with columns that replicate the monomials in $\mathcal{Q}$. The resulting dataset, $\mathcal{D} \cup \mathcal{Q}(\mathcal{D})$, adds the monomial transformations in $\mathcal{Q}$ to the original dataset $\mathcal{D}$. An alternative formulation would just replace $\mathcal{D}$ by $\mathcal{Q}(\mathcal{D})$. It turns out that the first formulation is a special case of the second (by including the variables in the monomial set) and has better error performance — what is not surprising because it uses more features.

The second problem is treated in the GA setting: Let $\mathcal{D}$ be a dataset as above and $\mathcal{P}$ a set of polynomials. For each polynomial $p \in \mathcal{P}$ let $\mathcal{Q}_p$ be the set of monomials in p (without the coefficients) and define the (anti) fitness

$$\phi_p \quad = \quad \min_{\Theta} J_{\text{fit}}(\Theta; \mathcal{Q}_p, \mathcal{D})$$

by solving the first problem. With a fitness of every instance, the GA genetic operators (usually mutation and crossover) evolve the population $\mathcal{P}$ until a reasonable approximation of a local minimum is found. Notice that the properties of GAs and linear regression entail that Algorithm 1 converges to a polynomial that is a local minimum of the fitness function, encapsulated in the error function J_{fit}.

Subsection 2.1 describes the encoding of individual polynomial instances as chromosomes and other parameters used in the GA implementation. The

regularization of the cost function is discussed in subsection 2.2.

2.1. Polynomial Encoding

The specific encoding (representation) of a set of monomials is an important aspect in the implementation of EPR. The choice described below permits active and inactive monomials for regression purposes. The active (or inactive) state of a monomial might change through mutation or crossover. This simple mechanism enhances variation in the complexity of polynomial expressions by evolutionary operations.

Let $\{q_1, \ldots, q_k\}$ be a set of monomials over the variables $X_1, \ldots, X_m$. The encoding of that set using d bits per exponent is a binary list such that

1. the initial segment of k bits defines the active state of each monomial;
2. the remaining bits are split into k segments of size $m \times d$, each representing a monomial;
3. the bits in each monomial segment are split into m sub-segments of size d. The j^{th} sub-segment is the binary representation of the degree of the variable X_j in the enclosing monomial segment;

This encoding can also be viewed as the flattening of the binary exponents in the matrix representation prefixed by the activation segment. The set $\{x_1^3 x_3, x_3^7, x_1 x_2\}$ (with $m = 3$) has matrix representation

$$\begin{bmatrix} 3 & 0 & 1 \\ 0 & 0 & 7 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 011 & 000 & 001 \\ 000 & 000 & 111 \\ 001 & 001 & 000 \end{bmatrix}_{(2)} \tag{2}$$

where the right matrix is in binary form using $d = 3$ bits. An encoding of this set of monomials with the extra monomial $x_1^6 x_2^2 x_3^5$ inactive, setting $k = 4$, would be

1110; 011, 000, 001; 000,000,111; 001,001,000; 110,010,101

where, for reading purposes, semicolons separate segments and commas separate variables. The first $k = 4$ bits inform that the first, second and third monomials are active while the fourth is not.

While each valid encoding represents a set of monomials the map is not bijective: each set of monomials has multiple encodings, for example by changing d or the order of monomial segments. However, considering the

EPR task, this is a minor issue and a bijective map would add computational complexity and negative impact to the algorithm's performance.

There is one final remark concerning this encoding method. As it is, the activation segment can become all zeros, representing the empty set of monomials. This situation can be avoided with a simple hack: Given an encoding, the first monomial is always considered active, thus restricting the syntactic form of encodings to binary strings starting with 1. In practice, this means that the implementation of the encoding can omit the first bit.

2.2. Cost Function

The polynomial regression error considered so far accounts for the ability to predict the transformed test-set. A known problem of using a cost function based only in the dataset error (and of polynomial regressions in general) is the tendency to overfit training data.

Excessive variance of the estimation method can be reduced by regularizing the error function with a penalty factor. Thus, to reduce polynomial complexity and variance by regularizing the size of the monomial set the error function from equation 1 is multiplied by a factor λ^k

$$J_{\text{reg}}\left(\Theta, \lambda; \mathcal{Q}, \mathcal{D}\right) \;=\; \lambda^k J_{\text{fit}}\left(\Theta; \mathcal{Q}, \mathcal{D}\right) \tag{2}$$

where k is the number of monomials in the polynomial. When $\lambda > 1$ polynomials with more monomials are penalized. The regularized extension of EPR is denoted by Evolutionary Polynomial Regression with Regularization (EPR^2).

A simple exploration on the effect of the regularization parameter is depicted in Figure 1 where it is possible to observe that the typical inflection point lies around $\lambda = 0.8$. This value, favoring "larger" polynomials is justified by the balance of the data's non-linearity and polynomial complexity: below $\lambda = 0.8$, even penalized, larger monomial sets achieve better error performance than smaller ones while above that value the size of the monomial set is excessive. Within this tension the overall error results reduced when compared to the non-regularized EPR version.

2.3. Genetic Algorithm Parameterization

In general GAs offer many possibilities with respect to the choice of genetic operators and respective application rates, population evolution, *etc.* The results found here where obtained using the package `genalg` (Willighagen, 2012) with default parameters, standard operators (crossover and mutation) and population evolution with 20% elitism between generations.

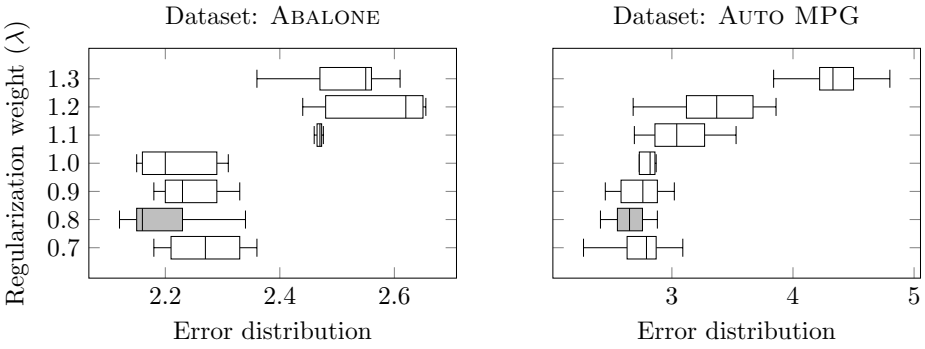


Figure 1: Error distribution by regularization exponent for two common datasets. The box plots summarize error values of ten simulations for each value of λ. The smallest overall error, in grey, is achieved in both datasets when $\lambda = 0.8$. Performance of the non-regularized EPR is plotted in the line $\lambda = 1$.

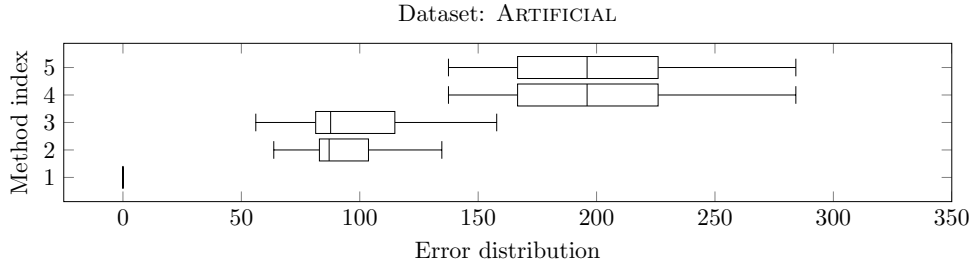
3. Experimental Results

The results were found using `R` programming language (R Core Team, 2013)[3]. To compare this paper's proposed algorithm we applied the exact same train and test samples using several well-known learning algorithms for regression, namely: the classic EPR, Linear Regression, Support Vector Machines (Meyer et al., 2012), Regression Trees (Therneau et al., 2013) and Conditional Inference Trees (Hothorn et al., 2006; Strobl et al., 2007, 2008). To achieve better error results, the SVM and Regression Tree algorithms had their parameters tuned.

In order to train and test the performance of EPR^2 several mainstream datasets were used. For each dataset, we selected 70% for training purposes and the remaining observations to make a test set in order to compute the estimated error. To achieve more robust results, each dataset were processed 25 times, each one with different samples for the train and test sets. For the datasets with attribute values of different magnitudes, a preliminary scaling was executed. The results below are box plots for the test set error predictions over these different runs.

Artificial this is an artificial dataset with four numeric features, $x_1, \ldots x_4$,

[3]The datasets and the `R` code used to produce the results and plots in this paper are available online at `https://github.com/jpneto/GenAlgPoly`.



[Note[4]]

Figure 2: Results for Artificial dataset. As shown, polynomial regression finds the exact polynomial structure that generates the dataset: in line 1, the error box is centered in 0 and has width 0. The regression methods depicted are: 1. EPR^2, 2. Linear Regression, 3. SVM, 4. Regression Trees and 5. Conditional Inference Trees

where x_1, x_3 are outcomes from Poisson random variables, and x_2, x_4 from Normal random variables. The dependent variable y is given by expression $x_2 x_4^2 + x_1^2 x_3 + 5$. The dataset includes $n = 50$ observations.

This dataset was used in order to verify if EPR was able to find the polynomial relation, which the algorithm did (cf. figure 2). The genetic algorithm run with a population of $n = 100$ solutions with 50 iterations for each run.

In the next datasets, the population of the genetic algorithm had size $n = 300$ with 100 iterations for each run. EPR^2 was run with regularization parameter $lambda = 0.8$.

HOUSING: This data set concerns the task of predicting housing values in areas of Boston. There are $m = 13$ continuous attributes and the dependent variable is the median value of owner-occupied homes in \$1000's. There are $n = 506$ observations.

ABALONE This dataset can be used to predict the age of a abalone shell using the given $m = 8$ numeric attributes concerning several physical measurements. There are $n = 4177$ observations.

AUTO MPG This dataset is used to predict fuel consumption in miles per gallon, based on two discrete and five continuous attributes ($m = 7$). There are $n = 398$ observations.

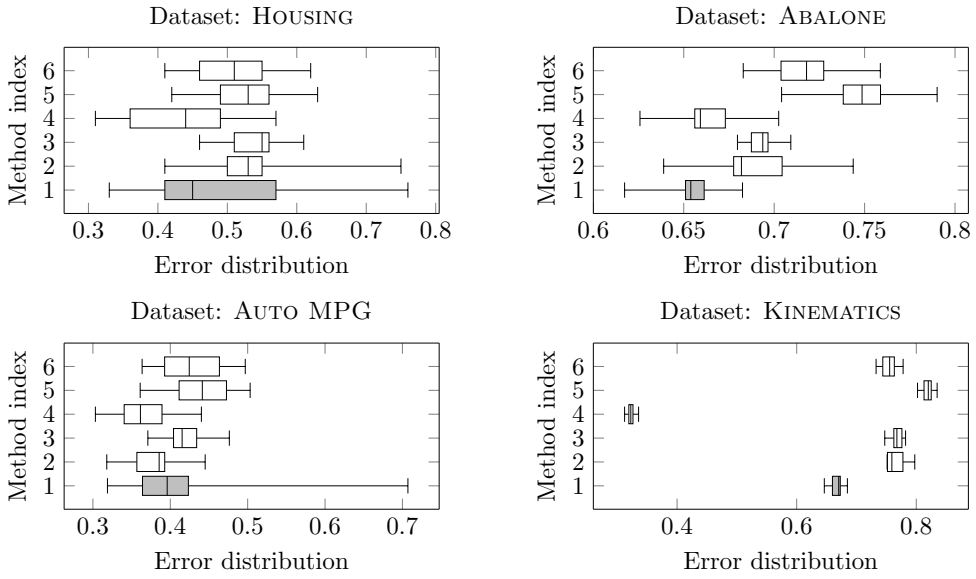


Figure 3: Summary results for different regression methods on the Housing, Abalone, Auto MPG and Kinematics. Although EPR^2 not always achieves the smallest error, performance is on-par with more sophisticated methods. The regression methods depicted in these figures are: 1. EPR^2, 2. EPR, 3. Linear Regression, 4. SVM, 5. Regression Trees and 6. Conditional Inference Trees

Kinematics This dataset is concerned with the realistic simulation of the forward kinematics of an 8 link robot arm. The task is to predict the distance of the end-effector from a target using $m = 8$ continuous attributes. There are $n = 8192$ observations.

3.1. Convergence speed

The GA quickly proceeds in the first 50 to 100 generations to reasonable error rates. Then, it proceeds slower achieving best solutions with marginal error reduction. Since the entire learning process takes some time, in the current R implementation, placing a limit between 50 to 100 generations already achieves good results, relative to higher iteration values. Figure 4 shows a typical error evolution for the dataset Abalone given a regularized version ($\lambda = 0.975$) and a non-regularized ($\lambda = 1.0$).

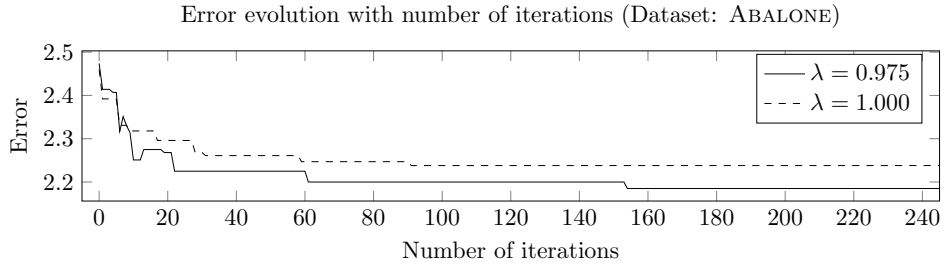


Figure 4: Error progress for Abalone dataset during a single execution of the genetic algorithm. The figure shows the fitness evolution for two different regularization values. The population for both consisted of 200 polynomials. The error values seem to stabilize around iteration 250.

4. Conclusion and Future Work

[**Note**[5]]

Of the regression methods considered, SVM achieved the best results in 3 out of 4 datasets. However, SVM was tuned for each particular dataset while EPR^2 was executed with its parameters preset. Even so, EPR^2 had the best error in the Abalone dataset, and achieved competitive results in the Housing and Auto MPG datasets.

Comparing EPR and EPR^2 – which is the main article's topic – the regularized version achieved much better results at Abalone and especially Kinematics. The Housing improved errors when compared with EPR in a difference in means, resulted in a 95% HDI (Highest Density Interval) equal to [0.001, 0.119] which, while borderline, achieves statistical significance. Only in the Auto MPG dataset EPR achieved better results, even if not that different from EPR^2.

For complexity considerations EPR and EPR^2 demands some processing time. On a quad-core computer, processing the Kinematics dataset (with 8k observations) takes approximately 5 minutes. The processing time can probably be speeded between one to two orders in magnitude if the process is implemented in a low level programming language like `C++`. However, speed optimization was not the focus of this article.

A cross-validation procedure can be implemented to refine the appropri-

[5]**Note:** Discutir a localidade do regressor

ate parameter values to achieve better errors. Namely, the regularization parameter, λ, can be tested with several values, instead of being fixed at 0.8. Other parameters like mutation chance or the amount of elitism could also be tested. However, these type of tests need a low-level, fast implementation of EPR and were not considered.

Acknowledgements

The authors are grateful to the Fundação para a Ciência e Tecnologia (FCT) and the R&D laboratory LabMAg for the financial support given to this work, under the strategic project PEst-OE/EEI/UI0434/2011.

Datasets used herein were selected from Luís Torgo's data repository, `http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html`. Most can also be found in the UCI ML repository at `http://archive.ics.uci.edu/ml/`.

The authors wish to thank professor André Falcão for motivation and useful discussions around the article.

Bibliography

Bao, Y., Hu, Z., Xiong, T., 2013. A pso and pattern search based memetic algorithm for svms parameters optimization. Neurocomputing 117, 98–106.

Barricelli, N.A., 1962. Numerical testing of evolution theories. part i: Theoretical introduction and basic tests. Acta Biotheoretica 16, 69–98.

Bengio, Y., 2009. Learning deep architectures for AI. Foundations and trends in Machine Learning 2, 1–127.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives .

Cetisli, B., Kalkan, H., 2011. Polynomial curve fitting with varying real powers. Electronics and Electrical Engineering 112, 117–122.

Chan, K.Y., Dillon, T.S., Kwong, C.K., 2011. Modeling of a liquid epoxy molding process using a particle swarm optimization-based fuzzy regression approach. Industrial Informatics, IEEE Transactions on 7, 148–158.

Chan, K.Y., Kwong, C., Dillon, T.S., 2012. Development of product design models using fuzzy regression based genetic programming, in: Computational Intelligence Techniques for New Product Design. Springer, pp. 111–128.

Cséfalvayová, L., Pelikan, M., Kralj Cigić, I., Kolar, J., Strlič, M., 2010. Use of genetic algorithms with multivariate regression for determination of gelatine in historic papers based on FT-IR and NIR spectral data. Talanta 82, 1784–1790.

Davidson, J., Savic, D., Walters, G., 1999. Method for the identification of explicit polynomial formulae for the friction in turbulent pipe flow. Journal of Hydroinformatics 1, 115–126.

Davidson, J., Savic, D.A., Walters, G.A., 2003. Symbolic and numerical regression: Experiments and applications. Information Sciences 150, 95–117.

Gálvez, A., Iglesias, A., Puig-Pey, J., 2012. Iterative two-step genetic-algorithm-based method for efficient polynomial b-spline surface reconstruction. Information Sciences 182, 56–76.

García Nieto, P.J., Alonso Fernández, J., de Cos Juez, F., Sánchez Lasheras, F., Díaz Muñiz, C., 2013. Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the trasona reservoir (northern spain). Environmental research .

Giustolisi, O., Savic, D., 2009. Advances in data-driven analyses and modelling using epr-moga. Journal of Hydroinformatics 11, 225–236.

Hofwing, M., Strömberg, N., Tapankov, M., 2011. Optimal polynomial regression models by using a genetic algorithm, in: Proceedings of the Second International Conference on Soft ComputingTechnology in Civil, Structural and Environmental Engineering Conference,(Crete, Greece), 2011009.

Holland, J.H., 1975. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15, 651–674.

Koza, J.R., 1992. Genetic Programming: vol. 1, On the programming of computers by means of natural selection. volume 1. MIT press.

Liang, Z., Lee, Y., 2012. Eigen-analysis of nonlinear pca with polynomial kernels .

Maertens, K., De Baerdemaeker, J., Babuška, R., 2006. Genetic polynomial regression as input selection algorithm for non-linear identification. Soft Computing 10, 785–795.

Mendes, L., Carvalho, P.d., 2005. Adaptive polynomial regression for colorimetric scanner calibration using genetic algorithms, in: Intelligent Signal Processing, 2005 IEEE International Workshop on, IEEE. pp. 22–27.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2012. e1071: Misc functions of the department of statistics (e1071), tu wien URL: `http://CRAN.R-project.org/package=e1071`. r package version 1.6-1.

R Core Team, 2013. R: A language and environment for statistical computing URL: `http://www.R-project.org/`.

Sánchez, L., Otero, J., Couso, I., 2009. Obtaining linguistic fuzzy rule-based regression models from imprecise data with multiobjective genetic algorithms. Soft Computing 13, 467–479.

Schölkopf, B., Smola, A., Müller, K.R., 1997. Kernel principal component analysis, in: Artificial Neural Networks ICANN'97. Springer, pp. 583–588.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9. URL: `http://www.biomedcentral.com/1471-2105/9/307`.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8. URL: `http://www.biomedcentral.com/1471-2105/8/25`.

Tarlow, D., Sutskever, I., Zemel, R.S., 2013. Stochastic k-neighborhood selection for supervised and unsupervised learning. Journal of Machine Learning Research .

Therneau, T., Atkinson, B., Ripley, B., 2013. rpart: Recursive partitioning R package version 4.1-1.

Willighagen, E., 2012. genalg: R based genetic algorithm.

Wu, C.H., Tzeng, G.H., Lin, R.H., 2009. A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. Expert Systems with Applications 36, 4725–4735.

Yu, T.L., Lin, W.K., 2008. Optimal sampling of genetic algorithms on polynomial regression, in: Proceedings of the 10th annual conference on Genetic and evolutionary computation, ACM. pp. 1089–1096.