Stability Selection using a Genetic Algorithm and Logistic Linear Regression on Healthcare Records

Zamuda, Aleš; Zarges, Christine; Stiglic, Gregor; Hrovat, Goran

Published in:
GECCO '17

DOI:
10.1145/3067695.3076077

Publication date:
2017

Citation for published version (APA):
Zamuda, A., Zarges, C., Stiglic, G., & Hrovat, G. (2017). Stability Selection using a Genetic Algorithm and Logistic Linear Regression on Healthcare Records. In *GECCO '17: Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 143-144). Association for Computing Machinery. https://doi.org/10.1145/3067695.3076077

Stability Selection using a Genetic Algorithm and Logistic Linear Regression on Healthcare Records

Aleš Zamuda
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova ul. 17, SI-2000 Maribor, Slovenia
ales.zamuda@um.si

Christine Zarges
Department of Computer Science
Aberystwyth University
Aberystwyth, SY23 3DB, United Kingdom

Gregor Stiglic
University of Maribor,
Faculty of Health Sciences, Zitna ul. 15, SI-2000 Maribor,
Faculty of Electrical Engineering and Computer Science,
Smetanova ul. 17, 2000 Maribor, Slovenia

Goran Hrovat
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova ul. 17, SI-2000 Maribor, Slovenia
goran.hrovat@um.si

ABSTRACT

This paper presents a Genetic Algorithm (GA) application to measuring feature importance in machine learning (ML) from a large-scale database. Too many input features may cause over-fitting, therefore a feature selection is desirable. Some ML algorithms have feature selection embedded, e.g., lasso penalized linear regression or random forests. Others do not include such functionality and are sensitive to over-fitting, e.g., unregularized linear regression. The latter algorithms require that proper features are chosen before learning.

Therefore, we propose a novel stability selection (SS) approach using GA-based feature selection. The proposed SS approach iteratively applies GA on a subsample of records and features. Each GA individual represents a binary vector of selected features in the subsample. An unregularized logistic linear regression model is then trained and tested using GA-selected features through cross-validation of the subsamples. GA fitness is evaluated by area under the curve (AUC) and optimized during a GA run.

AUC is assessed with an unregularized logistic regression model on multiple-subsampled healthcare records, collected under the Healthcare Cost, and Utilization Project (HCUP), utilizing the National (Nationwide) Inpatient Sample (NIS) database.

Reported results show that averaging feature importance from top-4 SS and the SS using GA (GASS), improves these AUC results.

CCS CONCEPTS

•Mathematics of computing → Evolutionary algorithms; *Robust regression; Dimensionality reduction;* **•Information systems → Rank aggregation; Top-k retrieval in databases;** Data mining; **•Computing methodologies → Ranking; Supervised learning by classification; Supervised learning by regression; Genetic algorithms; Feature selection; Cross-validation;** *Learning linear models; Regularization;* **•Applied computing → Health care information systems;**

KEYWORDS

Stability Selection; Genetic Algorithm; Feature Selection; Feature Importance; Cross-validation; Logistic Generalized Linear Regression; Healthcare Cost Utility Project; Disease Risk Prediction; Healthcare Records.

ACM Reference format:
Aleš Zamuda, Christine Zarges, Gregor Stiglic, and Goran Hrovat. 2017. Stability Selection using a Genetic Algorithm and
Logistic Linear Regression on Healthcare Records. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3076077

GECCO '17 Companion, Berlin, Germany
 978-1-4503-4939-0/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3067695.3076077

1 INTRODUCTION

This paper proposes a novel stability selection approach to estimate feature importance, which uses a Genetic Algorithm [3] and unregularized logistic linear regression in combination with top-k stability selection [9].

Namely, as Machine Learning (ML) advances rapidly at all fields, e.g., healthcare, it is used to help people make decisions at more and more tasks. An example of such a task is diagnose prediction, where a diagnose is predicted from available data, e.g., age or tumor size [6]. Data is therefore the main requirement from which ML models are created. The main question arising is how much and which features we need to predict the outcome as accurately as possible. It is desirable to have as much data as we can, however, we need to carefully decide which features to include [4].

The big problem in ML is over-fitting [5], where the learned model performs well on the data used for learning and poorly on new data that is actually used in practice. Selecting only proper data or features, e.g., age or gender, is also desirable to shorten the training time and in some cases to make learned models easier to

interpret. Many algorithms were developed to assist selecting the best features for a given problem [2, 8].

When acquiring the data it is therefore useful to know, which features are more and which are less important. To estimate feature importance, many algorithms are available, e.g., Random Forests [1] and Stability Selection (SS) [7]. The latter feature selection algorithm, where feature importance can be assessed, works as ensemble of many feature selection runs and uses some other feature selection algorithm internally. Thereby, the proposed approach in this paper (see Algorithm 1) provides an important contribution to advances in ML by improving SS and its applicative performance (see Table 1), ordered by combined scores (see Table 2).

Algorithm 1 GASS combined with top-k SS.

Require: MAX_FES (maximum number of function evaluations allocated to each GA run), NP (GA population size), s (number of SS subsamples), k (top-k SS argument), data (data records).
Ensure: $S_{\text{GASS+SS(k)}}$ stability scores for all features

1: **for** $s = 1$ to N_s **do**
2: $\mathbf{F_s}$ = subsample of features;
3: $\mathbf{I_s}$ = subsample of data with only $\mathbf{F_s}$ features;
4: generate uniformly at random and evaluate initial GA binary coded population $\mathbf{x}_{i,0}, \forall i \in \{1, 2, ..., NP\}$;
5: **for** GA generation loop g (while FES < MAX_FES) **do**
6: **for** GA iteration loop i (for each individual $\mathbf{x}_{i,g}$ of the g-th population) **do**
7: GA individual offspring $\mathbf{x}_{i,g+1}$ evolution (mutation, crossover, selection), where $j_1, j_2, \in [1, NP]$ are two parent individuals:
8: $\mathbf{u}_{i,g+1}$ = binary crossover($\mathbf{x}_{j_1,g}, \mathbf{x}_{j_2,g}$);
9: $\mathbf{v}_{i,g+1}$ = binary mutation($\mathbf{u}_{i,g+1}$);
10: GA fitness evaluation of $\mathbf{v}_{i,g+1}$ using cross-validation;
11: **end for**
12: GA selection: propagate fittest individuals stochastically proportional;
13: **end for**
14: $\mathbf{x}_{\text{best}}$ = the best individual obtained in GA;
15: $\hat{S}(\mathbf{I_s})$ = features from $\mathbf{x}_{\text{best}}$;
16: **end for**
17: Calculate top-k SS scores for all p features on the data;
18: **for** i = 1 to p **do**
19: $S_{\text{GASS}}(x_i) = \frac{\sum_{s=1}^{N_s} 1\{x_i \in \hat{S}(\mathbf{I_s})\}}{\sum_{s=1}^{N_s} 1\{x_i \in \mathbf{F_s}\}}$;
20: $S_{\text{SS(k)}}(x_i)$ = score of feature x_i, calculated with top-k SS;
21: $S_{\text{GASS + SS(k)}}(x_i) = \frac{S_{\text{SS(k)}}(x_i) + S_{\text{GASS}}(x_i)}{2}$;
22: **end for**
23: **return** descending sorted $S_{\text{GASS+SS(k)}}(x_i), \forall i \in \{1, 2, ..., p\}$;

ACKNOWLEDGMENTS

This article is based upon work from COST Action CA15140 'Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO)fi and COST Action IC1406 'High-Performance Modelling and Simulation for Big Data Applications (cHiPSet)' supported by COST (European Cooperation in Science and Technology). The work is also supported in part by the Slovenian Research Agency, Programme Unit P2-0041.

Table 1: Logistic linear regression classification AUC statistics for different SS methods, over 100 random subsamples.

Selected features	Average	Median	Min	Max	Std. dev.
GASS	0.8920	0.8923	0.8852	0.8996	0.0027
top-4 SS	0.8895	0.8896	0.8803	0.8965	0.0028
top-3 SS	0.8895	0.8896	0.8803	0.8965	0.0028
top-2 SS	0.8895	0.8896	0.8803	0.8965	0.0028
GASS + top-4 SS	**0.8929**	**0.8930**	**0.8875**	**0.8998**	**0.0026**
All features	0.8685	0.8683	0.8575	0.8771	0.0036
top-1 SS	0.8317	0.8321	0.8192	0.8422	0.0045

Table 2: SS results: top 20 feature importances from the top-4 SS, utilized GASS, and the proposed GASS + top-4 SS.

#	Feature	Description	top-4 SS	GASS	GASS + top-4 SS
1	AGE	Age in years	1.00	1.00	1.00
2	272.4	Other and unspecified hyperlipidemia	1.00	1.00	1.00
3	250.00	Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled	1.00	1.00	1.00
4	V27.0	Outcome of delivery, single liveborn	1.00	1.00	1.00
5	403.90	Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified	1.00	1.00	1.00
6	272.0	Pure hypercholesterolemia	1.00	1.00	1.00
7	585.9	Chronic kidney disease, unspecified	1.00	1.00	1.00
8	585.6	End stage renal disease	1.00	1.00	1.00
9	403.91	Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage V or end stage renal disease	0.76	1.00	0.88
10	530.81	Esophageal reflux	0.75	1.00	0.88
11	414.01	Coronary atherosclerosis of native coronary artery	0.50	1.00	0.75
12	278.00	Obesity, unspecified	0.50	1.00	0.75
13	278.01	Morbid obesity	0.50	1.00	0.75
14	786.59	Other chest pain	0.50	1.00	0.75
15	585.3	Chronic kidney disease, Stage III (moderate)	0.50	1.00	0.75
16	401.0	Malignant essential hypertension	0.50	1.00	0.75
17	402.90	Unspecified hypertensive heart disease without heart failure	0.50	1.00	0.75
18	401.1	Benign essential hypertension	0.50	1.00	0.75
19	402.91	Unspecified hypertensive heart disease with heart failure	0.50	1.00	0.75
20	404.91	Hypertensive heart and chronic kidney disease, unspecified, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified	0.50	0.97	0.74

REFERENCES

[1] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[2] Manoranjan Dash and Huan Liu. 1997. Feature selection for classification. *Intelligent data analysis* 1, 1-4 (1997), 131–156.
[3] David E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company, Reading, Massachusetts.
[4] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.
[5] Douglas M. Hawkins. 2004. The problem of overfitting. *Journal of chemical information and computer sciences* 44, 1 (2004), 1–12.
[6] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13 (2015), 8–17.
[7] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 4 (2010), 417–473.
[8] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *bioinformatics* 23, 19 (2007), 2507–2517.
[9] Jiayu Zhou, Jimeng Sun, Yashu Liu, Jianying Hu, and Jieping Ye. 2013. Patient risk prediction model via top-k stability selection. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 55–63.