7th INTERNATIONAL SYMPOSIUM ON INDUSTRIAL ENGINEERING

**INDUSTRIAL ENGINEERING DEPARTMENT,
FACULTY OF MECHANICAL ENGINEERING,
UNIVERSITY OF BELGRADE, SERBIA**
&
**STEINBEIS ADVANCED RISK TECHNOLOGIES,
STUTTGART, GERMANY**
&
**INNOVATION CENTER OF THE FACULTY OF
MECHANICAL ENGINEERING,
UNIVERSITY OF BELGRADE**

SIE 2018

PROCEEDINGS

Editors:

Vesna Spasojević-Brkić
Mirjana Misita
Dragan D. Milanović

27th-28th September 2018
Belgrade, Serbia

Editors
Vesna Spasojević-Brkić
Mirjana Misita
Dragan D. Milanović

7th INTERNATIONAL SYMPOSIUM ON INDUSTRIAL ENGINEERING - SIE 2018, PROCEEDINGS

Publisher
Faculty of Mechanical Engineering, Belgrade

Printing firm
"PLANETA PRINT" d.o.o. Beograd

Published 2018
ISBN 978-86-7083-981-6

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

005.22(082)
658.5(082)
006.83:338.45(082)

INTERNATIONAL Symposium of Industrial Engineering (7 ; 2018 ; Beograd)
Proceedings / 7th International Symposium of Industrial Engineering - SIE 2018, 27th-28th September 2018, Belgrade, Serbia ; [organizers] Industrial Engineering Department, Faculty of Mechanical Engineering, University of Belgrade [and] Steinbeis Advanced Risk Technologies, Stuttgart, Germany [and] Innovation Center of The Faculty of Mechanical Engineering, University of Belgrade ; editors Vesna Spasojević-Brkić, Mirjana Misita, Dragan D. Milanović. - Belgrade : Faculty of Mechanical Engineering, 2018 (Beograd : Planeta Print). - [10], 263 str. : ilustr. ; 30 cm

Tekst štampan dvostubačno. - Tiraž 100. - Str. [6]: Preface / editors. - Napomene i bibliografske reference uz radove. - Bibliografija uz svaki rad.

ISBN 978-86-7083-981-6
a) Производња - Организација - Зборници b) Индустријски менаџмент -
Зборници c) Индустрија - Систем квалитета - Зборници
COBISS.SR-ID 267659020

Sponzored by
Government of the Republic of Serbia
Ministry of Education, Science and Technological Development

7 th International Symposium on Industrial Engineering
SIE2018
Supported by: STEINBEIS ADVANCED RISK TECHNOLOGIES & INNOVATION CENTER OF THE FACULTY OF MECHANICAL ENGINEERING

PREDICTION OF SKIING TIME BY STRUCTURED REGRESSION ALGORITHM

Andrija Petrovic[1*], Ugljesa Bugaric[1], Boris Delibasic[2], Igor Ivetic[3]
[1]Faculty of Mechanical Engineering, University of Belgrade, Serbia
[2]Faculty of Organizational Sciences, University of Belgrade, Serbia
[3]Bosch, Serbia
[*]aapetrovic@mas.bg.ac.rs

Abstract. *In this paper, the application of Gaussian conditional random fields (GCRF) in the case of prediction skiing time between ski gates in ski center Kopaonik, is presented. Gaussian conditional random fields is well-known structured regression method that exploits advantages of unstructured predictors and combines them with the information concerning correlation between outputs. Four different unstructured predictors were used: ridge regression, LASSO regression, Random forest regression and support vector machine regression. Even though, only 18 features are used for prediction of skiing time, GCRF achieved better results, concerning R^2 and mean absolute error, compared to unstructured predictors.*

Keywords: *structured regression, Gaussian conditional random fields, GCRF, skiing.*

1. INTRODUCTION

One of the most fundamental problems in machine learning that have numerous application in various application fields is a regression. A wide variety of machine learning algorithms for unstructured regression have been developed [1]. A comparative empirical study on both well established, supervised machine learning technique including regression and classification, was carried out on different tasks and data sets originating from different domains by Singh et al. [2].

Besides unstructured regression predictors that are well known and widely used, an increased amount of information concerning relations between outputs, have made a drastic impact on prediction performances. Due to that, it is necessary to exploit additional information concerning structure between outputs. Gaussian conditional random fields (GCRF) are a widely used structured model for continuous outputs that use multiple unstructured predictors to form its features and in the same time exploits structure among outputs, which is defined by given similarity measure [3]. Furthermore, various adaptations and procedure improvements have been developed in order to extend the algorithm on directed graphs [4], or to reduce the computational cost of the learning procedure [5].

In this paper, GCRF was used to predict average skiing time between two ski gates in ski center Kopaonik. GCRF prediction performances (mean absolute error and coefficient of determination (R^2)) were compared with unstructured predictors performances: ridge regression, LASSO regression, random forest regression and support vector machine regression. The advantages of GCRF are emphasized and experimentally evaluated.

2. RELATED WORK

Machine learning supervised algorithms have been used in a various sports for prediction of: basketball outcomes [6], golf ball trajectories [7], football players selection [8] etc.

Additionally, Akgol et al. [9] used General Regression Neural Networks and Decision Tree forest to predict upper body power, one of the most important determinants of cross-country ski performance. The result shows that gender and oxygen value is the most important parameters for prediction of upper body power. Similarly, Delibasic et al. [10] developed model for ski injury predictive system by analyzing skier transportation data from six consecutive seasons. The predictive system is based on logistic regression and chi-square automatic interaction detection decision tree. The lowest ski injury risk is observed for skiers who spend more time in the ski lift transportation system and ski faster than average skier. In the same manner, a comparison of several models based on data mining, expert modeling and a combination of both have been evaluated in [11]. The analysis showed that expert models are 10-15% less accurate in comparison with data mining models.

In addition, analysis of different ski tasks is up to date research area [12, 13].

In this paper, we showed that structured regression algorithms can significantly improve prediction performances in cases when unstructured predictors scores are poor due to a small number of relevant features.

3. METHODOLOGY

The generalized form of GCRF conditional distribution $P(\mathbf{y}|\mathbf{x},\alpha,\beta)$ is given in form of condtional random field (CRF) and can be expressed as:

$$P(\mathbf{y}|\mathbf{x},\alpha,\beta) = \frac{1}{Z(\mathbf{x},\alpha,\beta)} exp\left(\sum_{i=1}^{N} A(\alpha, y_i, \mathbf{x_i}) + \sum_{i\neq j} I(\beta, y_i, y_j)\right) \quad (1)$$

Two different feature functions are used: *association potential* $A(\alpha, y_i, \mathbf{x})$ to model relations between outputs y_i and corresponding input vector $\mathbf{x_i}$ and *interaction potential* $I(\beta, y_i, y_j)$ to model pairwise relation between nodes. Vectors α and β are parameters of the association potential A and the interaction potential I. The *association potential* is defined as:

$$A(\alpha, y_i, \mathbf{x_i}) = -\sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k \left(y_i - R_k(\mathbf{x})\right)^2 \quad (2)$$

where $R_k(\mathbf{x})$ represents unstructured predictor of y_i for each node in the graph. This unstructured predictor can be any regression model that gives independent prediction of output y_i for given attributes $\mathbf{x}$. K is the total number of unstructured predictors. The *interaction potential* functions is defined as:

$$I(\beta, y_i, y_j) = -\sum_{l=1}^{L} \sum_{k=1}^{K} \beta_l S_{ij}^{l} (y_i - y_j)^2 \quad (3)$$

where S_{ij}^{l} is a value that express the similarity between nodes i and j in graph l. L is the total numbers of graphs (similarity functions). Graphs can express any kind of relations between nodes e.g., spatial and temporal correlations between outputs. Hence, the conditional probability distribution of the presented model is defined as:

$$P(\mathbf{y}|\mathbf{x},\alpha,\beta) = \frac{1}{Z(\mathbf{x},\alpha,\beta)} exp\left(-\sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k (y_i - R_k(\mathbf{x}))^2 - \sum_{l=1}^{L} \sum_{k=1}^{K} \beta_l S_{ij}^{l} (y_i - y_j)^2\right) \quad (4)$$

The quadratic form of interaction and association potential enables conditional distribution $P(\mathbf{y}|\mathbf{x},\alpha,\beta)$ to be expressed as multivariate Gaussian distribution. The canonical form of GCRF is:

$$P(\mathbf{y}|\mathbf{x},\alpha,\beta) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T \Sigma^{-1} (\mathbf{y}-\mu)\right) \quad (5)$$

where precision matrix $\Sigma^{-1} = 2Q$ and distribution mean $\mu = \Sigma\mathbf{b}$ is defined as, respectively:

$$Q = \begin{cases} \sum_k \alpha_k + \sum_h \sum_l \beta_l S_i h^l, & \text{if i = j} \\ -\sum_l \beta_l S_{ij}^{l}, & \text{if i} \neq \text{j} \end{cases} \quad (6)$$

$$b_i = 2\left(\sum_{k=1}^{K} \alpha_k R_k(\mathbf{x})\right) \quad (7)$$

The representation of GCRF is illustrated in Fig. 1.

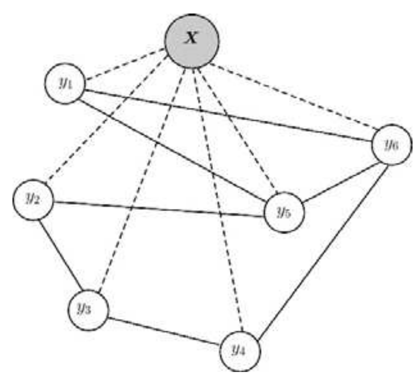

Figure 1. GCRF representation.

Due to the convexity of multivariate Gaussian distribution the inference task $\underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x},\alpha,\beta)$ is straightforward. The maximum posterior estimate of $\mathbf{y}$ is the distribution expectation μ.

The objective of the learning task is to optimize parameters α and β by maximizing conditional log likelihood $\underset{\alpha,\beta}{\operatorname{argmax}} \sum_{\mathbf{y}} logP(\mathbf{y}|\mathbf{x},\alpha,\beta)$. One way to ensure positive definiteness of covariance matrix of GCRF is to impose constraints that all elements of α and β be greater than 0. The derivative od the conditional log likelihood can be expressed in the following form:

$$dlogP = -\frac{1}{2}Tr(d\Sigma^{-1}(\mathbf{y}-\mu)(\mathbf{y}-\mu)^T - 2\Sigma^{-1}(\mathbf{y}-\mu)d\mu^T) - \frac{1}{2}Tr(\Sigma^{-1}d\Sigma) \quad (8)$$

The optimization of the parameters can be obtained by gradient descent method with log transformation of derivatives. Moreover, the optimization can also be performed by a truncated Newton algorithm for nonlinear functions with constraints or sequential quadratic programming. The GCRF code used in this work is publicly available.[1]

4. RESULTS AND DISCUSSION

Kopaonik is one of the largest ski resort in a southern part of Europe with more than 55 km of ski slopes and 25 ski lifts. One of the biggest problem in Kopaonik is crowds on ski lifts, due to that it is necessary to predict average skiing times between two ski gates. With this average skiing times it is possible to predict occurance of rush hours on ski-lifts.

Dataset used in this research includes information on ski lift entrance for a period from 15th to 30th of March for years between 2006-2011. Totally seven ski lifts were considered: Karaman Greben, Mali Karaman, Marine vode, Duboka I, Karaman, Pancicev vrh, Duboka II.

All used features are separated in the three distinct groups:

1. Descriptive features: the total number of skiers, total unique number of skiers, time expressed in hours and minutes
2. Statistical features: mean, tenth percentile, first and second quartile, ninetieth percentile, median, minimum value, maximum value, kurtosis and skewness of average skiers velocity
3. Weather features: wind speed, temperature, dew point, cloud cover and pressure

All features were evaluated by observing shifts in time periods of 5 minutes, whereas prediction was made 15 minutes in advance.

The two graphs for potential interaction between ski lifts (nodes) were used. The first graph is obtained by differences between the history of average skiing time in the period of 30 minutes, whereas the second graph was obtained by differences between the history of average skiing time in slopes for the whole dataset. A total number of instances in the dataset was 4850 for each ski lift, which is totally 33950. The 20% of the dataset was used for testing, whereas the rest was used for training. Half of the training data was used for unstructured predictor learning, whereas the rest was used for optimizing GCRF parameters. All methods are implemented in Python and experiments were run on Windows 16 GB of memory and 2.5 GHz CPU. The results of the learning, concerning score metrics, are presented in Table 1.

Table 1. Prediction performances

	Ridge regression	LASSO regression	Random forest	Support vector machine	GCRF
Mean absolute error	364.87	367.72	343.6	388.05	336.44
R^2	0.578	0.585	0.549	0.484	**0.613**

It can be seen that GCRF outperformed unstructured predictors. Even though best-unstructured predictor Random forest provides satisfactory prediction performances, GCRF has best metrics R^2 and mean absolute error. It is important to emphasize that not only features are important in this particular case, but also the correlation structure between ski lifts.

5. CONCLUSION

In this paper, Gaussian conditional random fields (GCRF), well-known structured regression algorithm is applied on task of predicting skier average time between ski gates. The unstructured predictor that are used as association function in GCRF are: ridge regression, LASSO regression, random forest and support vector machine. It was shown that in this particular case dependencies among outputs are significant, such that GCRF outperformed all other regression algorithms. Further studies should concern comparing GCRF with other structured regression on real-world sports tasks.

[1] https://github.com/andrijaster

6. REFERENCES

[1] F. Stulp, O. Sigaud, Many regression algorithms, one unified model: A review, Neural Networks, 69 (2015) 60-79.

[2] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, IEEE, 2016, pp. 1310-1315.

[3] N. Djuric, V. Radosavljevic, Z. Obradovic, S. Vucetic, Gaussian Conditional Random Fields for Aggregation of Operational Aerosol Retrievals, IEEE Geosci. Remote Sensing Lett., 12 (2015) 761-765.

[4] T. Vujicic, J. Glass, F. Zhou, Z. Obradovic, Gaussian conditional random fields extended for directed graphs, Mach Learn, 106 (2017) 1271-1288.

[5] K. Ristovski, V. Radosavljevic, S. Vucetic, Z. Obradovic, Continuous Conditional Random Fields for Efficient Regression in Large Fully Connected Graphs, AAAI, 2013, pp. 840-846.

[6] F.J. Ruiz, F. Perez-Cruz, A generative model for predicting outcomes in college basketball, Journal of Quantitative Analysis in Sports, 11 (2015) 39-52.

[7] B. Bačić, Predicting golf ball trajectories from swing plane: An artificial neural networks approach, Expert Syst Appl, 65 (2016) 423-438.

[8] S. Robertson, C. Woods, P. Gastin, Predicting higher selection in elite junior Australian Rules football: The influence of physical performance and anthropometric attributes, Journal of Science and Medicine in Sport, 18 (2015) 601-606.

[9] D. Akgöl, M.F. Akay, Predicting upper body power of cross-country skiers using machine learning methods combined with feature selection, Signal Processing and Communications Applications Conference (SIU), 2015 23th, IEEE, 2015, pp. 148-151.

[10] B. Delibašić, S. Radovanović, M. Jovanović, Z. Obradović, M. Suknović, Ski injury predictive analytics from massive ski lift transportation data, Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, (2017) 1754337117728600.

[11] M. Bohanec, B. Delibašić, Data-mining and expert models for predicting injury risk in ski resorts, International conference on decision support system technology, Springer, 2015, pp. 46-60.

[12] M. Jones, C. Walker, Z. Anderson, L. Thatcher, Automatic detection of alpine ski turns in sensor data, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, ACM, 2016, pp. 856-860.

[13] O.M.H. Rindal, T.M. Seeberg, J. Tjønnås, P. Haugnes, Ø. Sandbakk, Automatic classification of sub-techniques in classical cross-country skiing using a machine learning algorithm on micro-sensor data, Sensors-Basel, 18 (2017) 75.