TomekLinks - Undersampling

def tomeklinks(data, y, option = “majority”, drop_na_col = True, drop_na_row = True, rel_thres = 0.5, rel_method = "auto", rel_xtrm_type = "both", rel_coef = 1.5, rel_ctrl_pts_rg = None):

TomekLinks under-samples the majority/minority/both class(es) by removing TomekLinks.

Function designed to help solve the problem of imbalanced data for regression.

Parameters:

main arguments / inputs:

data: pandas dataframe, the training set.

y: string, response variable y by name. It should be a header name found in the dataframe data.

option: {‘majority’, ‘minority’, ‘both’}, default = ‘majority’. Sampling information to sample the data set.

‘majority’: resample only the majority class;

‘minority’: resample only the minority class;

‘both’: resample both majority and minority class.

drop_na_col: bool, default = ‘True’, if ‘True’, auto drop columns with NaN's.

drop_na_row: bool, default = ‘True’, if ‘True’, auto drop rows with NaN's.

phi relevance function arguments / inputs:

rel_thres: float, positive real number, default = 0.5, define the relevance threshold considered rare in phi relevance function.

rel_method: {‘auto’, ‘manual’}, default = ‘auto’, the relevance method in phi relevance function.

rel_xtrm_type: {‘low’, ‘high’, ‘both’}, default = ‘both’, distribution focus on high, low or both.

rel_coef: float, positive real number, default = 1.5, coefficient for box plot in phi relevance function to consider rare.

rel_ctrl_pts_rg: 2d array, default = None, when rel_method = ‘manual’, it inputs for "manual" rel method.

References:

Branco, P., Torgo, L., Ribeiro, R. (2017). SMOGN: A Pre-Processing Approach for Imbalanced Regression. Proceedings of Machine Learning Research, 74:36-50. http://proceedings.mlr.press/v74/branco17a/branco17a.pdf.

Elhassan, T., & Aljurf, M. (2016). Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. Global J Technol Optim S, 1. https://www.researchgate.net/profile/Mohamed-Shoukri-2/publication/326590590_Classification_of_Imbalance_Data_using_Tomek_Link_TLink_Combined_with_Random_Undersampling_RUS_as_a_Data_Reduction_Method/links/5b96a6a0a6fdccfd543cbc40/Classification-of-Imbalance-Data-using-Tomek-Link-T-Link-Combined-with-Random-Under-sampling-RUS-as-a-Data-Reduction-Method.pdf

Kunz, N. (2019). SMOGN: Synthetic Minority Over-Sampling for Regression with Gaussian Noise (Version 0.1.0). Python Package Index. https://pypi.org/project/smogn.