Edited Nearest Neighbor (ENN)

def enn(data, y, samp_method = "balance", 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, k=3, n_jobs=1):

Function designed to help solve the problem of imbalanced data for regression, ENN under-samples the majority class.

Parameters:
main arguments / inputs:

data: pandas dataframe, the training set.

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

samp_method: {'balance', 'extreme'}, default = 'balance', specified method to determine over / under sampling percentage.

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: positive real, 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: positive real, 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.

KNeighborsClassifier attribute:

k: positive integer, default = 3, number of neighbors for k-NN.

n_jobs: positive integer, default = 1, the number of parallel jobs to run for neighbors' search.

k_neighbors_classifier: KNeighborsClassifier, default = None, allow users to customize the attributes of the classifier.

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.

Kunz, N., (2019). SMOGN. https://github.com/nickkunz/smogn

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, (3), 408-421. https://ieeexplore.ieee.org/abstract/document/4309137.