Importance of Network Metrics in Classification

Dieudonne Ouedraogo

System Science and Industrial Engineering Department
State University of New York at Binghamton, USA

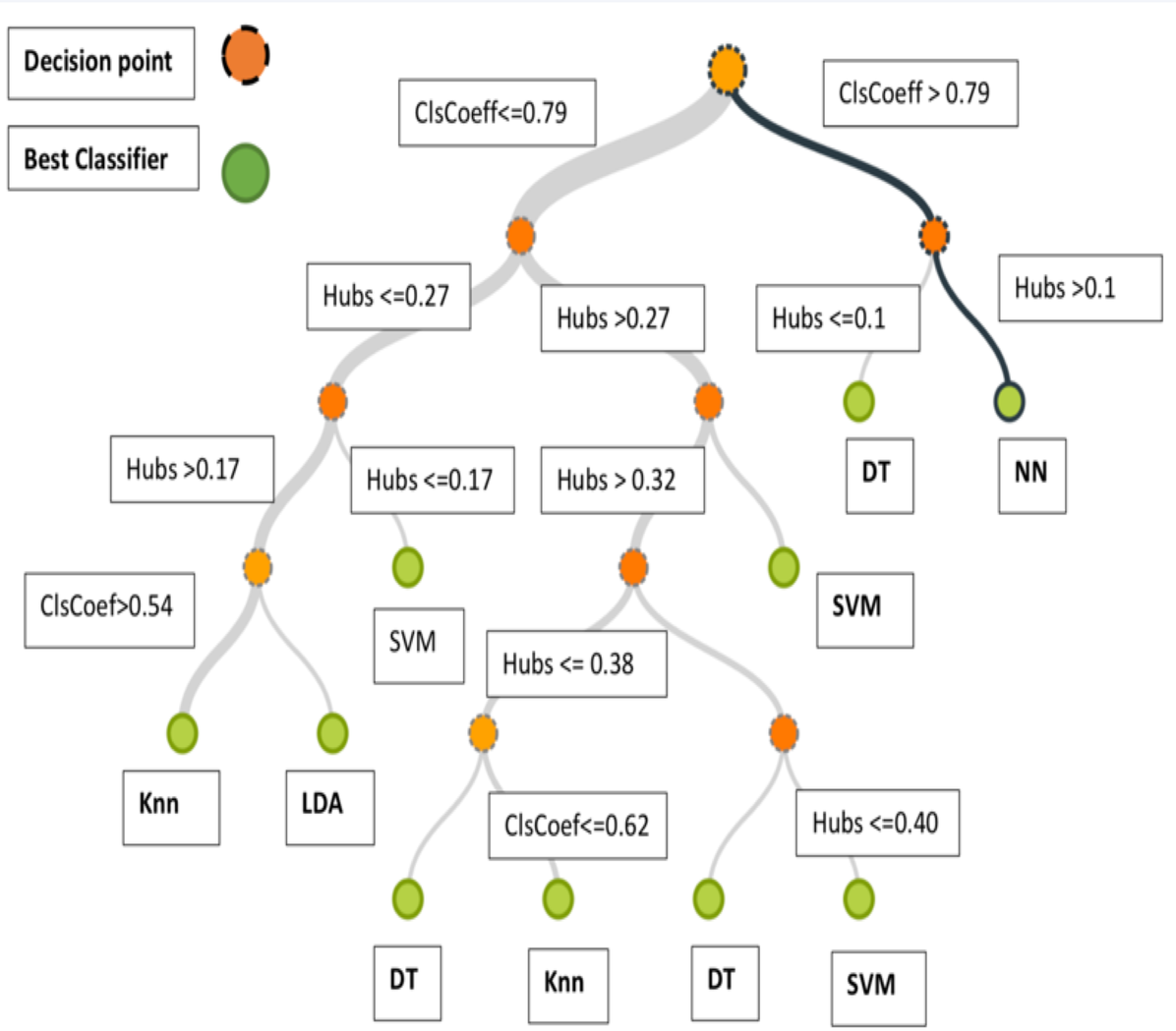
ABSTRACT

In machine learning, the performance of a classifier is intrinsically related to the task. The structure between data points within the dataset plays an important role. In this paper, we explore the usage of network metrics to describe the selection of a machine learning algorithm for a classification task concerning a specific dataset. A dataset is transformed into a graph representation based on the ϵNN algorithm. A data point is a node, and an edge exists between two points $\mathbf{i, j}$ **if** $d(i,j) < \epsilon$. A post-processing step is applied to the graph, pruning edges between examples of different classes. The structural information such as density, clustering coefficient, and hubs are extracted. Various data sets are collected, their network metrics are computed. A predictive model is built to investigate the possible relationship between networks characteristics and the classifier used to be used on the machine learning task. Results show that network metrics such as clustering coefficient, hubs, density are very informative in predicting the classifier to be used on the task.

RESULTS

The results show that network metrics such as clustering coefficient, hubs, density are very informative in predicting the classifier to be used on the task. For clustering coefficient greater than 0.79 or hubs greater 0.1, neural network are the best performers; for hubs less than 0.1, using decision trees gives the best performance.

Decision Tree with network metrics as variables and algorithms as outputs



Computed Network Metrics and Best Classifier

Data Set	Density	ClsCoef	Hubs	Classifier
Data Set 1	0.12267744	0.66259662	0.19018967	Knn
Data Set 2	0.16689038	0.73260435	0.21731327	Knn
Data Set 3	0.18770169	0.61501807	0.31778641	SVM
Data Set 4	0.08261087	0.75370635	0.1407504	SVM
Data Set 5	0.12462275	0.46805337	0.42546745	DT
Data Set 6	0.15489668	0.55026316	0.37760867	SVM
Data Set 7	0.16296607	0.46244857	0.32698233	Knn
Data Set 8	0.10032129	0.54130901	0.27980588	SVM
Data Set 9	0.13096887	0.52377361	0.25825507	LDA
Data Set 10	0.18939032	0.6414742	0.3495547	DT
Data Set 11	0.11474631	0.8304949	0.16130215	NN
Data Set 12	0.19112739	0.60846942	0.37585083	Knn
Data Set 13	0.13113155	0.55996607	0.26080425	Knn
Data Set 14	0.02252658	0.84263901	0.0376098	DT
Data Set 15	0.16892021	0.64951819	0.21242524	Knn
Data Set 16	0.12711158	0.70428715	0.21882891	Knn
Data Set 17	0.04958247	0.88500332	0.21287715	NN

DISCUSSION & FUTURE STUDY

A multiclass decision tree algorithm is run on our meta-dataset formed by the set of network metrics, based on previous work the best classifier is known. The results show how the network metrics in the structure of the dataset could infer the best classifier.

Machine learning pipelines involved long testing and comparison of performances of several algorithms with different parameters and steps. Our method is a meta-learning tool, to learn which machine learning algorithm to use without spending much time in testing.

In this study we tackle classification, future research will address regression problems where the output is continuous; we will also explore unsupervised algorithms such as clustering.

REFERENCES

[1] E. Mansilla and T. K. Ho, "On classifier domains of competence," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 1, 2004, pp. 136–139 Vol.1.

[2] Tin K Ho and Mitra Basu. (2002). Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 3, 289–300.

[3] Enrique Leyva, Antonio Gonzalez and Raul Perez. (2014). A Set of Complexity Measures Designed for Applying Meta-Learning to Instance Selection. IEEE Transactions on Knowledge and Data Engineering 27, 2, 354–367.