Datasets Complexity using Network Science

Dieudonne Ouedraogo

10/04/2018

Abstract

In machine learning, the performance of a classifier is intrinsically related to the task. The structure of the dataset plays an important role. Even though new developments and improvements are done algorithmically, very little is done to understand the dataset structure. In many situations, it can take days even weeks to train and evaluate the models. If we have to try several models, this can be resource and time-consuming, and the current trial and error process for selecting the right algorithm is not efficient. We consider the case of a binary classification problem, but it is possible to generalize the findings into more than two classes problems. In this paper we use network metrics to describe the complexity of a data set relative to a classification task. A dataset is transformed into a graph representation based on the ϵNN algorithm. A data point is a node and an edge exist between two points 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, and their metrics are computed and used as a training dataset to build a predictive model where the outputs are the algorithms competing to be used as the best classifier on a dataset. The algorithms used in this study are Decision Trees, Naive Bayes, SVM, Logistic regression, Artificial Neural Networks,K-nearest neighbors

keywords:

Meta-Learning, Machine Learning, Network Science, Dataset complexity, Classification, Algorithms Selection.

Introduction

In Machine learning a binary classification task difficulty can arise from the following reason: class ambiguity, the shape of the boundary, sample sparsity, feature space dimensionality and the network characteristics that define the dataset Papers published on the subject don't tackle the issue from a network science point of view. This paper is intended to show how vital the characteristics of a network formed by a dataset could be used to explain the performance of a classification task and to determine the competent algorithm to be assigned to the job.

We use binary classification here, but for multi-classes, the work could be easily extended by using one against all approach which will lead to binary classification. We use classification datasets available on UCI machine learning repository and dataset available on Weka.

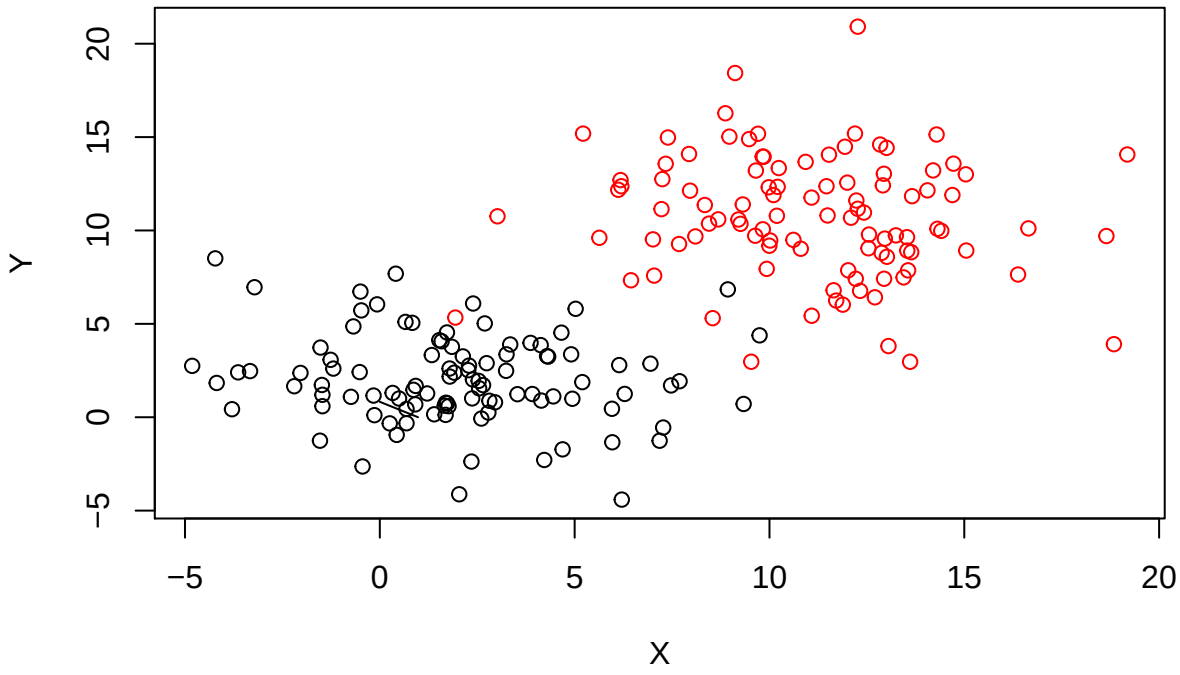
We explore the performance of five of the mainly used algorithms, Decision Trees, Naive Bayes, SVM, Logistic regression, Artificial Neural Networks,K-nearest neighbors.

The two plots below show two binary classifications one with two features (PredictorA and PredictorB) and another with features X and Y.

A task to determine the class for a data point could be relatively easy in one case and more complex in another

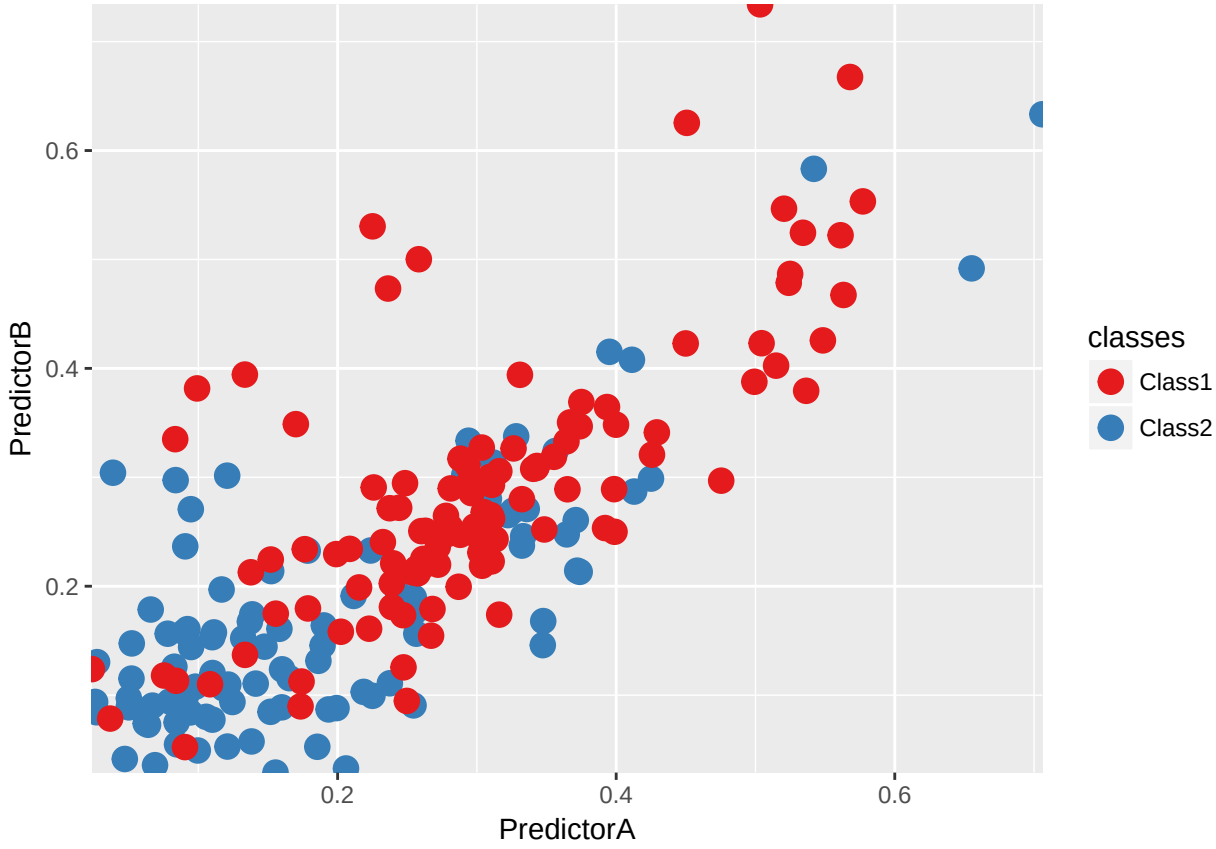
Low Complexity dataset for classification

Here the majority of the points in each class is easily separable by a straight line



High Complexity dataset for classification

Here, it is difficult to separate the two classes by a straight line



Literature review

A classification difficulty dramatically depends on the dataset. Understanding the characteristics associated with the dataset can help define and recommend the right algorithm to improve the performance of the task and to save time.

(Domingos, 2000) explained that the error of a predictor arises from three sources: a bias, from the difficulty of an algorithm to accurately model the relationship present in data; a variance, from the estimation of the correct parameters from the model due to imperfections from the sample used; a fundamental error referred to as noise.

For Ho & Basu (2002), classification difficulty comes from three components: the complexity of the decision boundary, the sample size and dimensionality induced sparsity and the ambiguity of the classes.

Ho (2008), believed that data complexity analysis is essential when comparing algorithms performance in machine learning. Usage of dataset complexity can also be found in combinatorial optimization, Smith-Miles & Lopes, (2012).

Choosing a sufficiently diverse set of problems to explain both strengths and weaknesses of the analyzed algorithms is essential in determining the domain of competence of the algorithm.

Macià et al. (2013), who described how algorithm comparison might be biased by benchmark dataset selection, and showed how complexity measures might guide the choice. Characterizing problem space with some metrics makes it possible to estimate regions in which specific algorithms perform well as detailed by Luengo & Herrera, (2013). This leads to possibilities of meta-learning as described by Smith-Miles et al.,(2014).

Complexity measures could then be used not only as predictors of classifier performance but also as diversity measures capturing various properties of the datasets.

Ho & Basu (2002) introduced complexity measures which were also extended by Ho, Basu & Law (2006) and Orriols-Puig, Macià & Ho (2010). Those measures are often used for algorithm's evaluation as described by Macià et al., (2013) and Luengo & Herrera, (2013), they are also used in meta-learning (Diez-Pastor et al., 2015; Mantovani et al., 2015). Part of these measures focuses on the overlap of values of specific attributes: Fisher's discriminant ratio, the volume of the overlap region, the attribute efficiency, etc.

Another part is toward the class separability; in this section, we have measures such as the fraction of points on the decision boundary, the linear separability, the ratio of intra/interclass distance. Those measures focus on specific properties of the classification problem, measuring the shape of the decision boundary and the amount class overlap. We also have topological measures concerned with data sparsity, such as the ratio of attributes to observations.

Li & Abu-Mostafa (2006) defined dataset complexity using the general concept of Kolmogorov complexity. The measures proposed to use the number of support vectors in the support vector machine (SVM) classifier. They analyzed the problems of data decomposition and data pruning using the above methodology. They defined the representation of the dataset complexity called the complexity-error plot.

Smith, Martinez & Giraud-CARRIER(2013) tackled the problem of complexity in the dataset from a single instance point of view where they analyzed misclassified instances by various algorithms approach to data complexity is to explain. They devised local complexity measures calculated concerning the individual case and explored the correlations of those measures with the global data complexity measures of Ho & Basu (2002). They concluded that they are mainly related to class overlap.

Yin et al. (2013) used a feature selection based on Hellinger distance to described complexity by measuring the similarity between probability distributions. They chose features, which conditional distributions (depending on the class) have a minimal affinity. The authors demonstrated experimentally that, for the high-dimensional imbalanced data sets, their method is superior to popular feature selection methods using the Fisher criterion, or mutual information.

Dataset complexity measures

A-Measure of overlapping

The feature overlapping measures characterize how informative the available fea- tures are to separate the classes

F1 - Maximum Fisher discriminant ratio.

The measure gives the effectiveness of a single feature in separating the classes This measure computes the maximum discriminative power (Fisher ratio) of the attributes. The ratio is defined as

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

where μ_1,μ_2,σ_1^2, σ_2^2 are the means and variances of each class, respectively, in that feature. f is computed for each feature and maximum is taken as F1. A high value of F1 indicates that at least one of the attributes enables the learner to separate the examples of different classes with partitions that are parallel to an axis of the feature space. A low value of this measure does not imply that the classes are not linearly separable, but that they cannot be discriminated by hyperplanes parallel to one of the axis of the feature space.

F1v

Directional-vector maximum Fisher's discriminant ratio (F1v) complements F1 by searching for a vector able to separate two classes after the training examples have been projected into it.

F2

Volume of overlap region,this measure computes the overlap of the tails of distributions defined by the instances of each class. Let $min(f_i, c_j)$ and $max(f_i, c_j)$ be, respectively, the minimum and maximum values of the feature f_i for class c_j. Then, the overlap measure is defined as

$$F2 = \prod \frac{MINMAX_i - MAXMIN_i}{MAXMAX_i - MINMIN_i}$$

where $MINMAX_i = min(max(f_i, c_1), max(f_i, c_2))$ $MAXMIN_i = max(min(f_i, c_1), min(f_i, c_2))$ $MAXMAX_i = max(max(f_i, c_1), max(f_i, c_2))$ $MINMin_i = min(min(f_i, c_1), min(f_i, c_2))$

A low value states that the attributes can discriminate the examples of different classes.

F4

Collective feature efficiency(F4) get an overview on how various features may work together in data separation. First the most discriminative feature according to F3 is selected and all examples that can be separated by this feature are removed from the dataset. The previous step is repeated on the remaining dataset until all the features have been considered or no example remains. F4 returns the ratio of examples that have been discriminated.

B- measures of neighborhood

Neighborhood measures characterize the presence and density of same or different classes in local neighborhoods. The Neighborhood measures analyze the neighborhoods of the data items and try to capture class overlapping and the shape of the decision boundary. They work over a distance matrix storing the distances between all pairs of data points in the dataset. To deal with both symbolic and numerical features, we adopt a heterogeneous distance measure named Gower distance.

N1

Fraction of borderline points (N1) computes the percentage of vertexes incident to edges connecting examples of opposite classes in a Minimum Spanning Tree (MST).

N2

Ratio of intra/extra class nearest neighbor distance (N2) computes the ratio of two sums: intra-class and inter-class. The former corresponds to the sum of the distances between each example and its closest neighbor from the same class. The later is the sum of the distances between each example and its closest neighbor from another class (nearest enemy).

N3

Error rate of the nearest neighbor(N3)classifier corresponds to the error rate of a one Nearest Neighbor (1NN) classifier, estimated using a leave-one-out procedure in dataset.

N4

Non-linearity of the nearest neighbor classifier (N4) creates a new dataset randomly interpolating pairs of training examples of the same class and then induce a the 1NN classifier on the original data and measure the error rate in the new data points.

T1

Fraction of hyper-spheres covering data (T1) builds hyper-spheres centered at each one of the training examples, which have their radios growth until the hyper-sphere reaches an example of another class. Afterwards, smaller hyper-spheres contained in larger hyper-spheres are eliminated. T1 is finally defined as the ratio between the number of the remaining hyper-spheres and the total number of examples in the dataset.

LSCAvg

Local Set Average Cardinality (LSCAvg) is based on Local Set (LS) and defined as the set of points from the dataset whose distance of each example is smaller than the distance from the examples of the different class. LSCAvg is the average of the LS.

C-Measures of linearity

The linearity measures try to quantify if it is possible to separate the classes by a hyper-plane. The underlying assumption is that a linearly separable problem can be considered simpler than a problem requiring a non-linear decision boundary.

L1

Sum of the error distance by linear programming (L1) computes the sum of the distances of incorrectly classified examples to a linear boundary used in their classification.

L2

Error rate of linear classifier(L2)computes the error rate of the linear SVM classifier induced from dataset.

L3

Non-linearity of a linear classifier (L3) creates a new dataset randomly interpolating pairs of training examples of the same class and then induce a linear SVM on the original data and measure the error rate in the new data points.

D- Measures of dimensionality

These measures give an indicative of data sparsity. They capture how sparse a datasets tend to have regions of low density. These regions are know to be more difficult to extract good classification models.

T2

Average number of points per dimension (T2) is given by the ratio between the number of examples and dimensionality of the dataset.

T3

Average number of points per PCA (T3) is similar to T2, but uses the number of PCA components needed to represent 95 variability as the base of data sparsity assessment.

T4

Ratio of the PCA Dimension to the Original (T4) it estimates the proportion of relevant and the original dimensions for a dataset.

E-Measures of class balance

These measures capture the differences in the number of examples per class in the dataset. When these differences are severe, problems related to generalization of the ML classification techniques could happen because of the imbalance ratio.

C1

The entropy of class proportions (C1) measure the imbalance in a dataset based on the pro- portions of examples per class.

C2

The imbalance ratio (C2) is an index computed for measuring class balance. This is a version of the measure that is also suited for multiclass classification problems.

Measures of Network

The network measures represent the dataset as a graph and extract structural information from it. The transformation between raw data and the graph representation is based on the epsilon-NN algorithm. Next, a post-processing step is applied to the graph, pruning edges between examples of opposite classes.

Density

Average Density of the network (Density) represents the number of edges in the graph, divided by the maximum number of edges between pairs of data points.

ClsCoef

Clustering coefficient (ClsCoef) averages the clustering tendency of the vertexes by the ratio of existent edges between its neighbors and the total number of edges that could possibly exist between them.

Hubs

Hubs score (Hubs) is given by the number of connections it has to other nodes, weighted by the number of connections these neighbors have.

3-Methodology

Part1

Creation of the graph from a dataset and extraction of metrics. each data point from the dataset is a node. We transform the dataset into a graph and we extract structural information from it. The transformation between raw data and the graph representation is based on the ϵNN algorithm.nodes i and j are connected by an edge, if the distance $d(i, j) < \epsilon$. The hyper-parameter ϵ controls the neighborhood radius. Next, a post-processing step is applied to the graph, pruning edges between examples of opposite classes.

Dataset1

```
library(pander)
library(RDocumentation)
library(ECoL)
dataset1=twoClass
kable(head(dataset1))
```

PredictorA	PredictorB	classes
0.1582	0.1609	Class2
0.6552	0.4918	Class2
0.7060	0.6333	Class2
0.1992	0.0881	Class2
0.3952	0.4152	Class2
0.4250	0.2988	Class2

```
net1_Metrics=network(classes ~.,dataset1)
pander(net1_Metrics)
```

Density	ClsCoef	Hubs
0.1227	0.6626	0.1902

Dataset2 IRIS dataset

```
dataset2=iris
kable(head(iris))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

```
net2_Metrics=network(Species ~ ., dataset2)
pander(net2_Metrics)
```

Density	ClsCoef	Hubs
0.1669	0.7326	0.2173

Dataset3

```
library(readr)
BreastCancer<- read_csv("http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsi

colnames(BreastCancer) <- c("SampleCodeNumber",
                            "ClumpThickness",
                            "UniformityOfCellSize",
```

```
                              "UniformityOfCellShape",
                              "MarginalAdhesion",
                              "SingleEpithelialCellSize",
                              "BareNuclei",
                              "BlandChromatin",
                              "NormalNucleoli",
                              "Mitosis",
                              "Classes")

BreastCancer$Classes <- ifelse(BreastCancer$Classes == "2", "Benign",
                               ifelse(BreastCancer$Classes == "4", "Malignant", NA))

#Data cleaning
BreastCancer[BreastCancer == "?"] <- NA
#length(which(is.na(BreastCancer)))
#Comparing the effect of the removal of NA
#str(BreastCancer)
#BreastCancer
nrow(BreastCancer[is.na(BreastCancer), ])
```

```
## [1] 16
```

```
#nrow(BreastCancer)
df=BreastCancer[,-c(1,7,11)]
#df
dataset3=BreastCancer[,-c(1,7)]
#str(df2)
#kable(head(dataset3))
net3_Metrics=network (Classes ~ ., dataset3)
pander(net3_Metrics)
```

Density	ClsCoef	Hubs
0.1877	0.615	0.3178

Part2

Creation of the meta Dataset

We pick datasets from UCI and Weka and Kaggle which cover a broad spectrum of characteristics for datasets. We create a meta feature target which value represents the best classifier reported so far by previous studies on that particular dataset. We use the characteristics of our networks as features for the meta-dataset (the dataset formed by the collection of datasets). The set of values of the characteristics of each dataset is an instance(observation) in our meta-dataset.

Meta-Dataset of 3 datasets metrics

```
df1=as.data.frame(rbind(net1_Metrics,net2_Metrics,net3_Metrics))
pander(df1)
```

	Density	ClsCoef	Hubs
net1_Metrics	0.1227	0.6626	0.1902
net2_Metrics	0.1669	0.7326	0.2173
net3_Metrics	0.1877	0.615	0.3178

If we are interested on runing a study on the entire complexity measures we can use this dataset below

```
net1_Total=complexity(classes ~ ., dataset1)
net2_Total=complexity(Species ~ ., iris)
net3_Total=complexity(Classes ~ ., dataset3)
df2=as.data.frame(rbind(net1_Total,net2_Total,net3_Total))
pander(df2)
```

Table 7: Table continues below

	overlapping.F1	overlapping.F1v	overlapping.F2
net1_Total	0.4797	0.9633	0.6654
net2_Total	190.3	104.2	0.006382
net3_Total	5.586	16.55	0.2478

Table 8: Table continues below

	overlapping.F3	overlapping.F4	neighborhood.N1
net1_Total	0.03365	0.04808	0.4808
net2_Total	0.8767	0.9567	0.1067
net3_Total	0.1189	0.2321	0.09169

Table 9: Table continues below

	neighborhood.N2	neighborhood.N3	neighborhood.N4
net1_Total	0.6521	0.3365	0.2692
net2_Total	0.1875	0.06	0.02
net3_Total	0.3588	0.05874	0.04298

Table 10: Table continues below

	neighborhood.T1	neighborhood.LSCAvg	linearity.L1
net1_Total	0.4808	0.02152	0.2252
net2_Total	0.12	0.1836	0.004393
net3_Total	0.1476	0.2417	0.03339

Table 11: Table continues below

	linearity.L2	linearity.L3	dimensionality.T2
net1_Total	0.25	0.2596	104
net2_Total	0.01333	0.006667	37.5
net3_Total	0.03725	0.01576	87.25

Table 12: Table continues below

	dimensionality.T3	dimensionality.T4	balance.C1
net1_Total	104	1	0.9967
net2_Total	75	0.5	1
net3_Total	116.3	0.75	0.9298

	balance.C2	network.Density	network.ClsCoef	network.Hubs
net1_Total	1.009	0.1227	0.6626	0.1902
net2_Total	1	0.1669	0.7326	0.2173
net3_Total	1.212	0.1877	0.615	0.3178

Part3

We run multiclass classification algorithms of our meta-dataset. From this classification, we extract the most predictive features from the network characteristics. We then build a predictive model based on the network characteristics related to datasets

Part4

Algorithm domain of competence With the previous model build, a new dataset could be handled by computing its characteristics and using the model to predict the suitable algorithm which is the class of the meta-data set.

The more datasets we use to train our model the more accurate the heuristic method will work. So finding more diverse classification datasets and feeding our system is essential.

REFERENCES

Ana C Lorena, Ivan G Costa, Newton Spolaor and Marcilio C P Souto. (2012). Analysis of complexity indices for classification problems: Cancer gene expression data. Neurocomputing 75, 1, 33–42.

Gleison Morais and Ronaldo C Prati. (2013). Complex Network Measures for Data Set Characterization. In 2nd Brazilian Conference on Intelligent Systems (BRACIS). 12–18.

Luis P F Garcia, Andre C P L F de Carvalho and Ana C Lorena. (2015). Effect of label noise in the complexity of classification problems. Neurocomputing 160, 108–119.

Ajay K Tanwani and Muddassar Farooq. (2010). Classification potential vs. classification accuracy: a comprehensive study of evolutionary algorithms with biomedical datasets. Learning Classifier Systems 6471, 127–144.

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.

Albert Orriols-Puig, Nuria Macia and Tin K Ho. (2010). Documentation for the data complexity library in C++. Technical Report. La Salle - Universitat Ramon Llull.

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.

R. Leite, P. Brazdil, and J. Vanschoren, “Selecting classification algorithms with active testing,” in MLDM, ser. Lecture Notes in Computer Science, P. Perner, Ed., vol. 7376. Springer, 2012, pp. 117–131.

P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta, Metalearning: applications to data mining. Springer, 2009.

T. K. Ho, M. Basu, and M. H. C. Law, Data Complexity in Pattern Recognition. Springer, 2005, ch. Measures of Geometrical Complexity in Classification Problems.

J. M. Sotoca, R. A. Mollineda, and J. S. Sanchez "A meta-learning framework for pattern classification by means of data complexity measures," Inteligencia Artificial, vol. 10, no. 29, pp. 31–38, 2006.

T. K. Ho and H. S. Baird, "Pattern classification with compact distribution maps," Computer Vision and Image Understanding, vol. 70, no. 1, pp. 101 – 110, 1998.

F. Smith, "Pattern classifier design by linear programming," Computers, IEEE Transactions on, vol. C-17, no. 4, pp. 367–372, 1968.

A. Hoekstra and R. Duin, "On the nonlinearity of pattern classifiers," in Proceedings of the 13th International Conference on Pattern Recognition, vol. 4, 1996, pp. 271–275.

L. Frank and E. Hubert, "Pretopological approach for supervised learn- ing," in Proceedings of the 13th International Conference on Pattern Recognition, vol. 4, 1996, pp. 256–260.

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.

S.-W. Kim and B. J. Oommen, "On using prototype reduction schemes to enhance the computation of volume-based inter-class overlap measures," Pattern Recognition, vol. 42, no. 11, pp. 2695 – 2704, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320309001642

X. Zhu, "Semi-supervised learning with graphs,"Ph.D. thesis, Carnegie Mellon University, 2005, http: //pages.cs.wisc.edu/ jerryzhu/pub/thesis.pdf.

N. Ganguly, A. Deutsch, and A. Mukherjee, Eds., Dynamics On and Of Complex Networks Applications to Biology, Computer Science, and the Social Sciences. Springer, 2009.

E. D. Kolaczyk, Statistical Analysis of Network Data: Methods and Model. Springer, 2009.

A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," SIGKDD Explorations, vol. 11, no. 1, pp. 10–18, 2009.

R. C. Prati, "Combining feature ranking algorithms through rank aggre- gation," in IJCNN. IEEE, 2012, pp. 1–8.

D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," J. Artif. Intell. Res. (JAIR), vol. 6, pp. 1–34, 1997.

G. E. A. P. A. Batista and D. F. Silva, "How k-nearest neighbor parameters affect its performance," in Argentine Symposium on Artificial Intelligence, 2009, pp. 1–12.

C. Soares, "Uci++: Improved support for algorithm selection using datasetoids," in PAKDD, ser. Lecture Notes in Computer Science, T. Theeramunkong, B. Kijsirikul, N. Cercone, and T. B. Ho, Eds., vol. 5476. Springer, 2009, pp. 499–506.

A. Ben-David, "Comparison of classification accuracy using Cohen's Weighted Kappa," Expert Syst. Appl., vol. 34, no. 2, pp. 825–832, 2008.