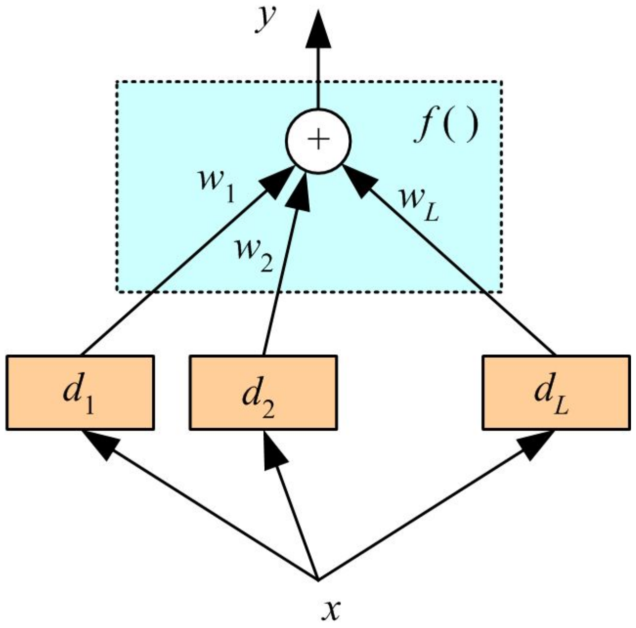


Ensemble Approaches

for classification and Regression

Nikesh Bajaj

DITEN, University of Genova

http://nikeshbajaj.in

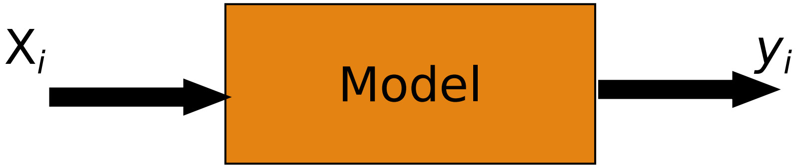
Overview

1. Introduction to Ensemble approaches
   - key ideas
2. Approaches
   1. Voting
   2. Mixture of Experts, Stacking
   3. Bagging, Boosting, Cascade
3. Case Study
   1. NDSB - Classification
   2. Liberty Mutual - Regression

Before we go….

Termonology

1. Training set
2. Validation set
3. Testing set
4. Features
5. Underfit and Overfit



class->	A	B	C
y_i	0.7	0.1	0.2

What is Ensemble？

Combining multiple models, learners or Estimators

- What models to combine?
- How to combine?

Key Ideas

1. Combine different Learners/Estimators/Models

   --- Average, weighted Average etc **---Voting**

   --- Mixture of experts

   --Stacking

2. **NOT** to combine strong or very accurate (should be complement)

   ---- different base learners/estimators

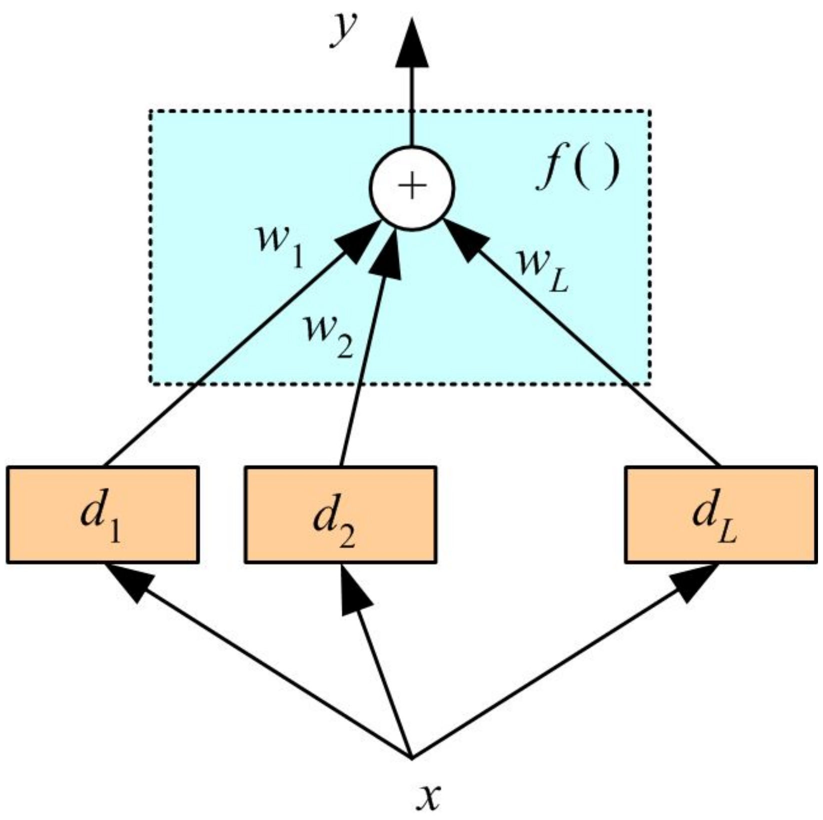
   ---- different hyperparameters of same learners/estimatos

3. Different learners/estimators/modals with <u>different set of features</u>

4. Different learners/estimators/modals with <u>different training sets</u>

   ---- **Bagging, Boosting, Cascading**

Voting



$$y_i = \sum_j w_j d_{ji} \text{ where } w_j \geq 0, \sum_j w_j = 1$$

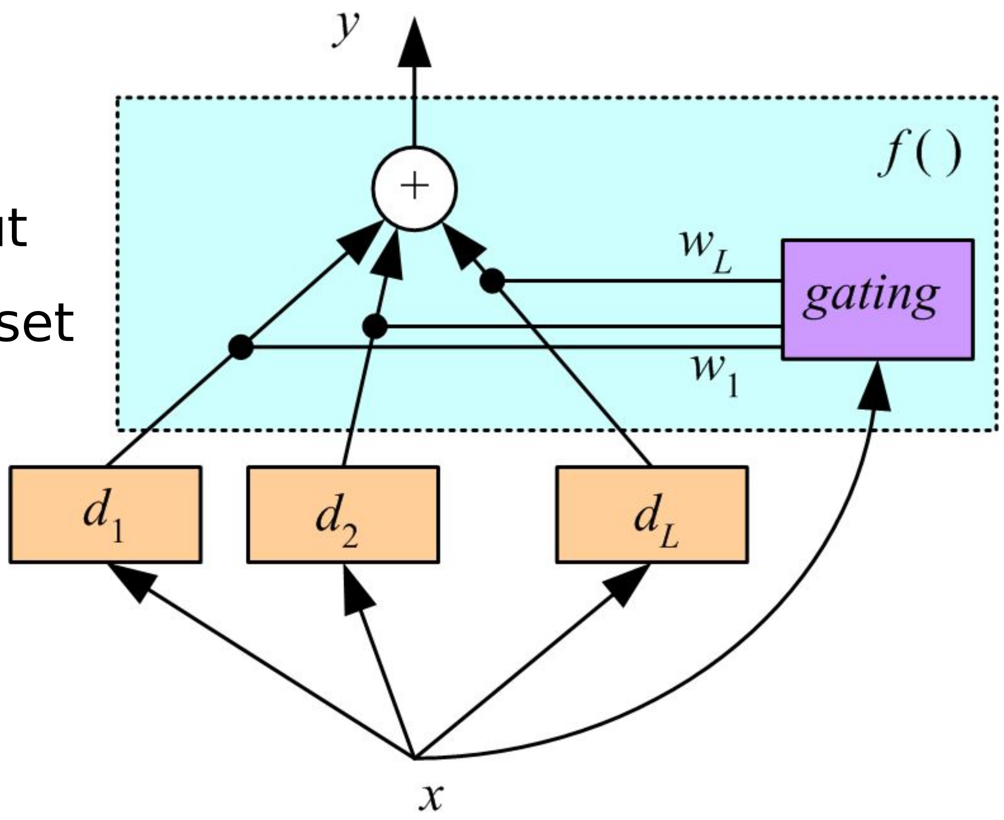
Rule	Fusion function $f(\cdot)$
Sum	$y_i = \frac{1}{L}\sum_{j=1}^{L} d_{ji}$
Weighted sum	$y_i = \sum_j w_j d_{ji}, w_j \geq 0, \sum_j w_j = 1$
Median	$y_i = \text{median}_j d_{ji}$
Minimum	$y_i = \min_j d_{ji}$
Maximum	$y_i = \max_j d_{ji}$
Product	$y_i = \prod_j d_{ji}$

Mixture of Experts

(Jacobs et al., 1991)

- Experts or gating can be nonlinear
- Weights may be different for different input
- Each learner become experts of different set of inputs

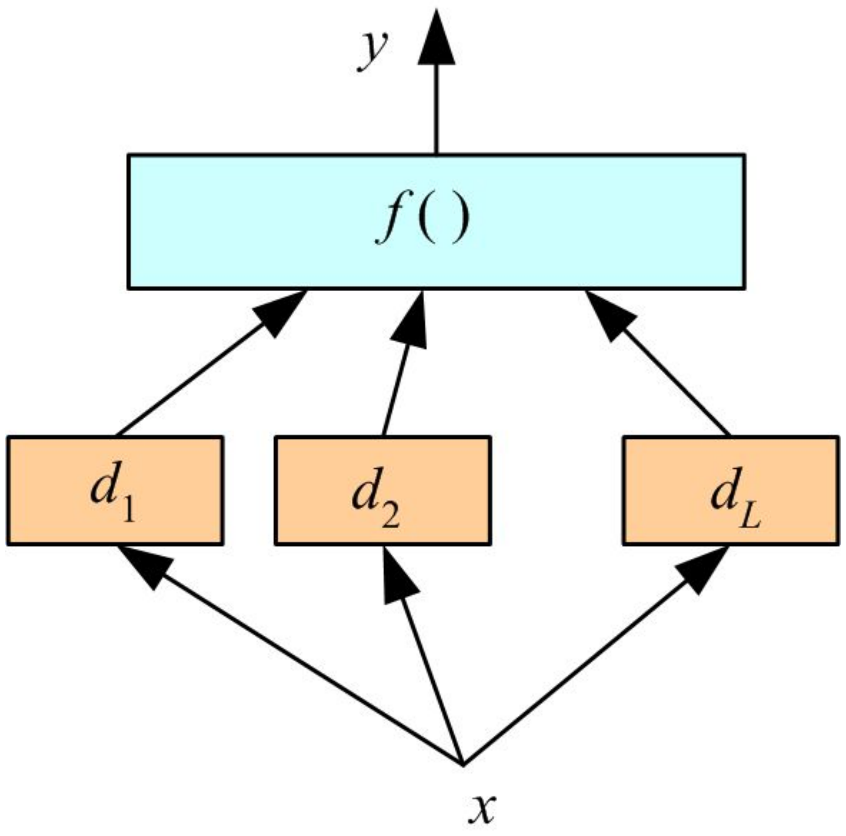
$$y = \sum_{j=1}^{L} w_j(x) d_j$$



Stacking

(Wolpert, 1992)

- Combiner f( ) is also learner
- f( ) may not be even linear
- Need to be trained on non-training data



Bagging

(Breiman 1996)

- L learner are trained with slightly different L training sets
- X^1, X^2 ... .. X^L $\subseteq X$
- Done by boostrap
- Some sample might be many times and some might not be at all
- All sets of samples X^i almost similar, but slightly different
- For large training set, simple approach is to divide with overlapping sets
- Works better when learning algo is unstable (sensitive to small change)

Boosting – Adaboost

Original Idea - (Schapire 1990), Adaboost - (Freund and Schapire 1996)

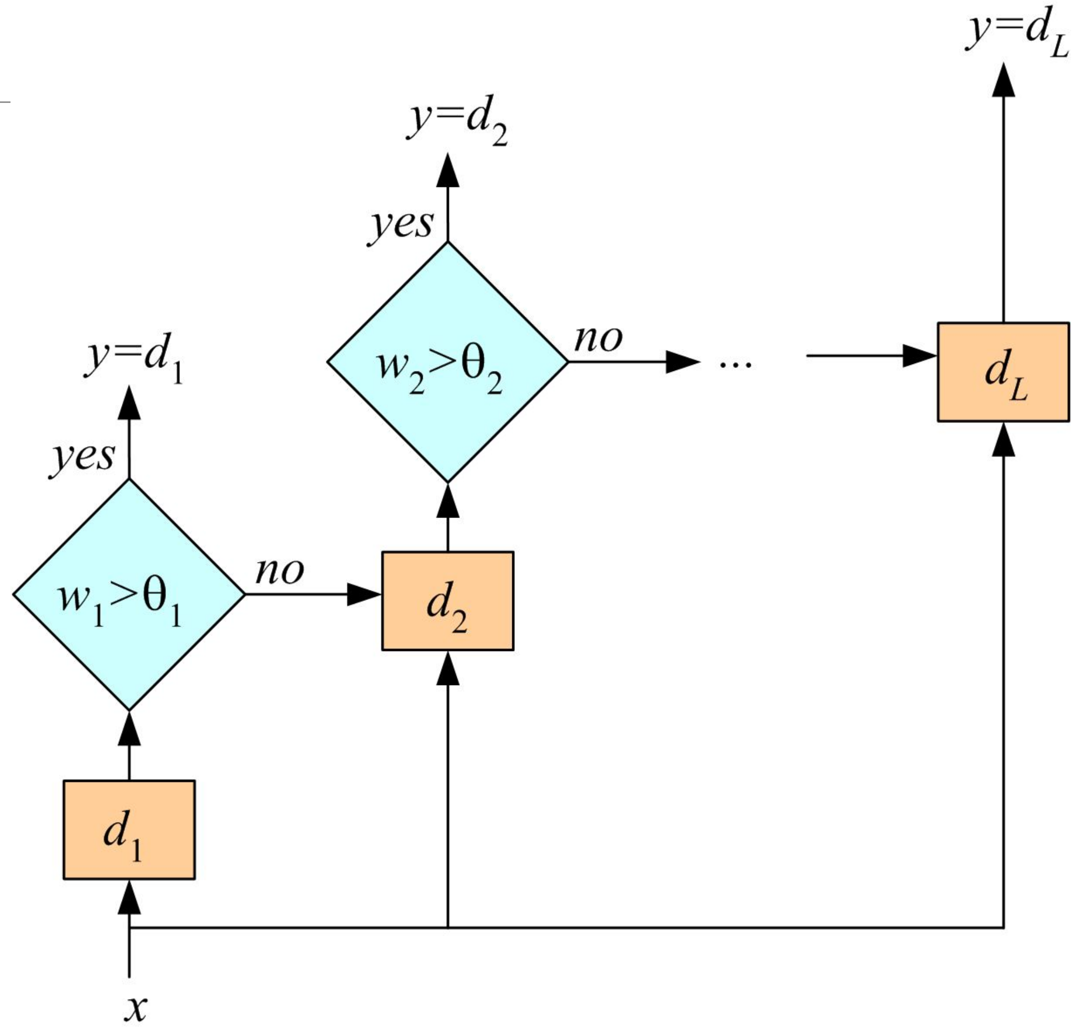
- In Bagging - learners being complemantory depends on chances
- In Boosting, complemantory leaners are actively generated
- L Leaners: d_1 d_2 .. .. . d_L
- d_{j+1} focus on instance more which was misclassified by d_j

Cascading

(Kaynak and Alpaydın 2000)

- Almost same idea as boosting

- Only difference is next model is trained when previous model was not confident enough (unlike misclassifies)



Case Study

1. NDSB – Classification

(http://www.datasciencebowl.com/)

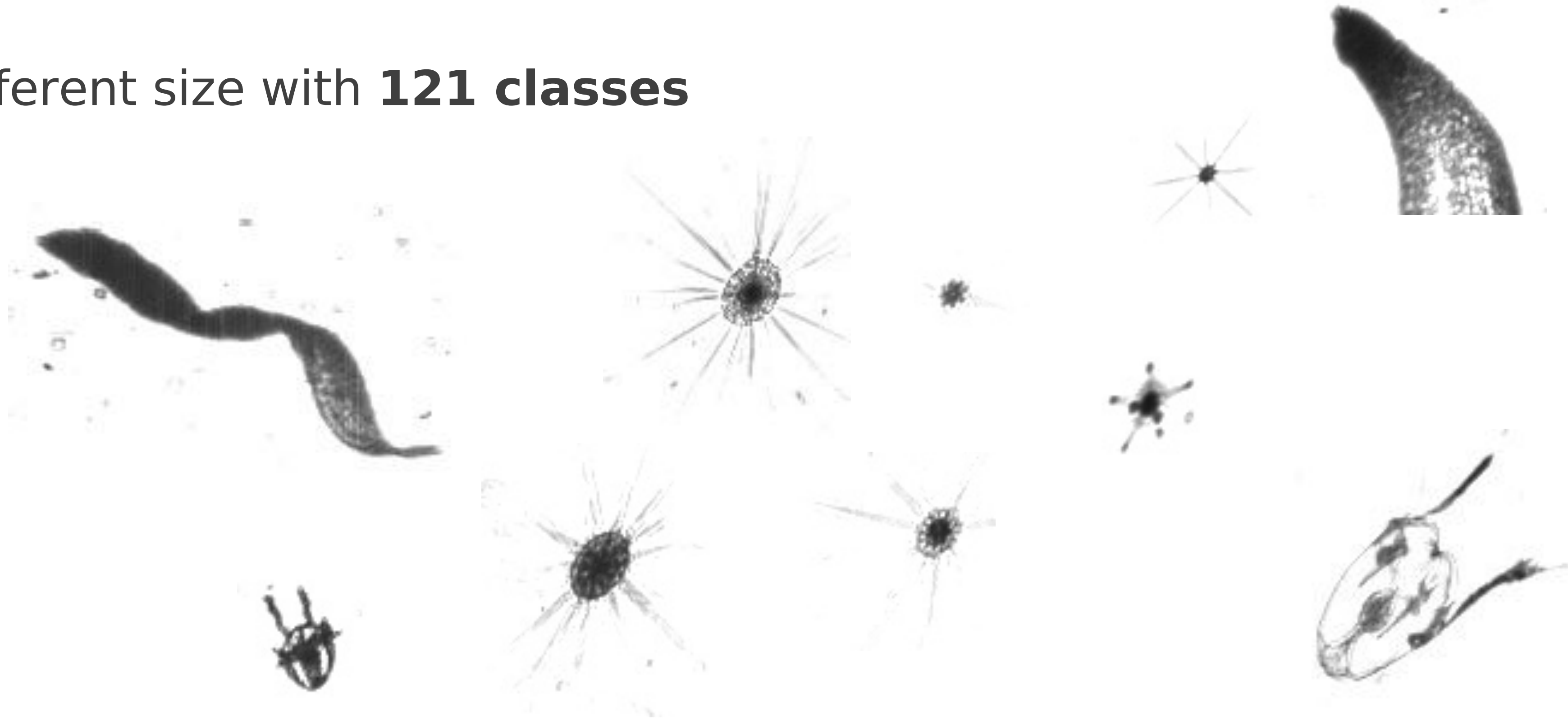
Goal : Classify images of plankton (kaggle)

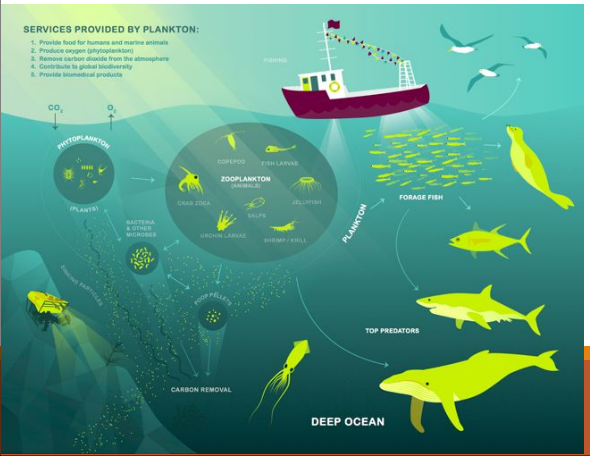
Training data :

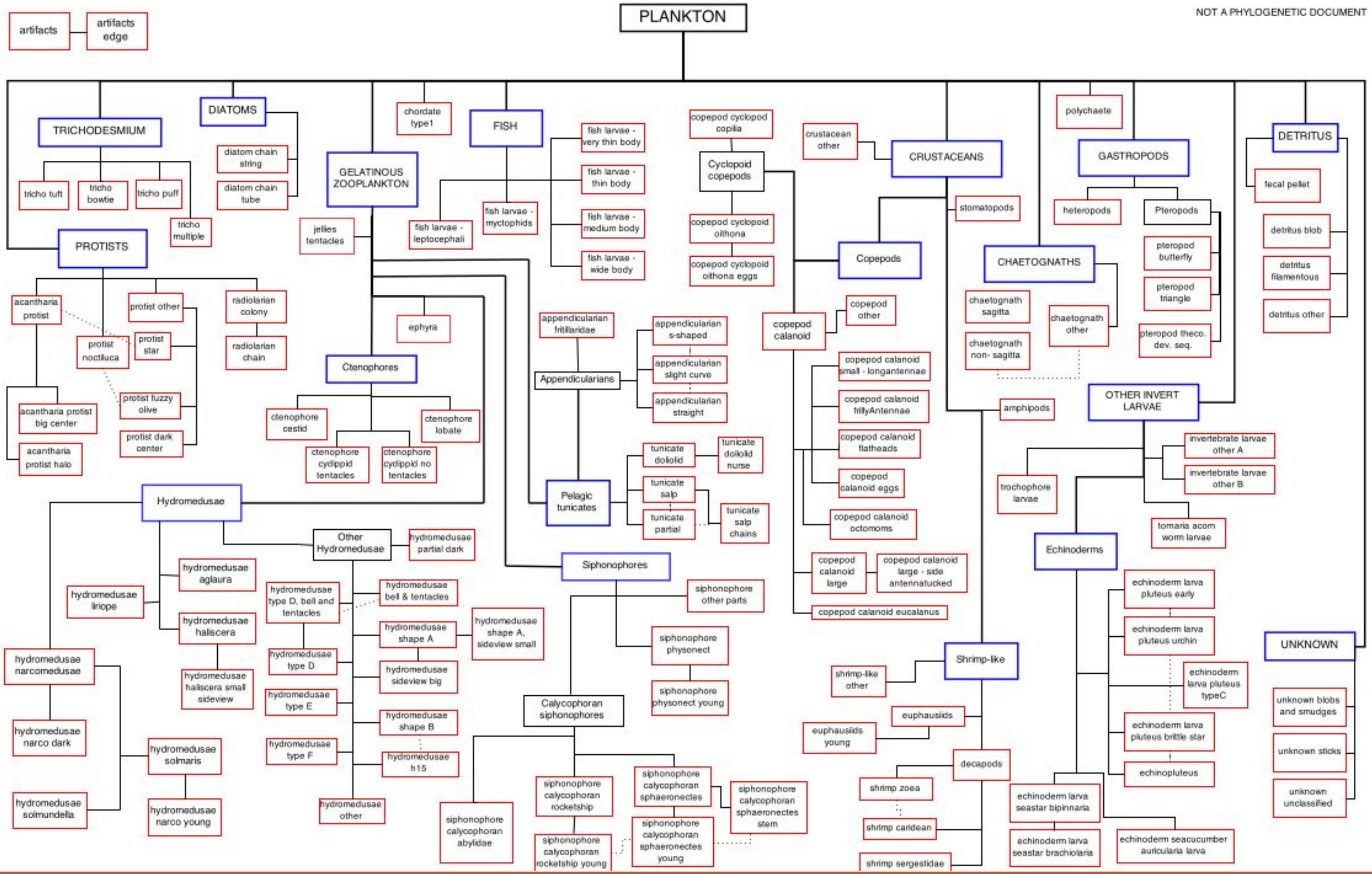
30K images of different size with **121 classes**

Test data :

120K images






PLANKTON
NOT A PHYLOGENETIC DOCUMENT
artifacts
artifacts edge
TRICHODESMIUM
tricho tuft
tricho bowtie
tricho puff
tricho multiple
DIATOMS
diatom chain string
diatom chain tube
PROTISTS
acantharia protist
protist noctiluca
protist star
protist other
protist fuzzy olive
protist dark center
radiolarian colony
radiolarian chain
acantharia protist big center
acantharia protist halo
GELATINOUS ZOOPLANKTON
jellies tentacles
ephyra
chordate type1
Ctenophores
ctenophore cestid
ctenophore lobate
ctenophore cydippid tentacles
ctenophore cydippid no tentacles
Hydromedusae
Other Hydromedusae
hydromedusae partial dark
hydromedusae aglaura
hydromedusae liriope
hydromedusae haliscera
hydromedusae haliscera small sideview
hydromedusae narcomedusae
hydromedusae narco dark
hydromedusae solmaris
hydromedusae solmundella
hydromedusae narco young
hydromedusae type D, bell and tentacles
hydromedusae bell & tentacles
hydromedusae shape A
hydromedusae shape A, sideview small
hydromedusae type D
hydromedusae sideview big
hydromedusae type E
hydromedusae shape B
hydromedusae type F
hydromedusae h15
hydromedusae other
FISH
fish larvae - very thin body
fish larvae - thin body
fish larvae - medium body
fish larvae - wide body
fish larvae - myctophids
fish larvae - leptocephali
appendicularian fritillaridae
Appendicularians
appendicularian s-shaped
appendicularian slight curve
appendicularian straight
Pelagic tunicates
tunicate doliolid
tunicate doliolid nurse
tunicate salp
tunicate partial
tunicate salp chains
Siphonophores
siphonophore other parts
siphonophore physonect
siphonophore physonect young
Calycophoran siphonophores
siphonophore calycophoran abylidae
siphonophore calycophoran rocketship
siphonophore calycophoran rocketship young
siphonophore calycophoran sphaeronectes
siphonophore calycophoran sphaeronectes young
siphonophore calycophoran sphaeronectes stem
copepod cyclopoid copilia
Cyclopoid copepods
copepod cyclopoid oithona
copepod cyclopoid oithona eggs
Copepods
copepod calanoid
copepod other
copepod calanoid small - longantennae
copepod calanoid frillyAntennae
copepod calanoid flatheads
copepod calanoid eggs
copepod calanoid octomoms
copepod calanoid large
copepod calanoid large - side antennatucked
copepod calanoid eucalanus
crustacean other
CRUSTACEANS
stomatopods
amphipods
Shrimp-like
shrimp-like other
euphausiids
euphausiids young
decapods
shrimp zoea
shrimp caridean
shrimp sergestidae
polychaete
GASTROPODS
heteropods
Pteropods
pteropod butterfly
pteropod triangle
pteropod theco. dev. seq.
CHAETOGNATHS
chaetognath sagitta
chaetognath non- sagitta
chaetognath other
OTHER INVERT LARVAE
invertebrate larvae other A
invertebrate larvae other B
trochophore larvae
tornaria acorn worm larvae
Echinoderms
echinoderm larva pluteus early
echinoderm larva pluteus urchin
echinoderm larva pluteus typeC
echinoderm larva pluteus brittle star
echinopluteus
echinoderm larva seastar bipinnaria
echinoderm larva seastar brachiolaria
echinoderm seacucumber auricularia larva
DETRITUS
fecal pellet
detritus blob
detritus filamentous
detritus other
UNKNOWN
unknown blobs and smudges
unknown sticks
unknown unclassified

acantharia_protist_big_center	**diatom_chain_tube**	**protist_noctiluca**	**crustacean_other**	**trichodesmium_tuft**
acantharia_protist_halo	echinoderm_larva_pluteus_brittlestar	protist_other	ctenophore_cestid	trochophore_larvae
acantharia_protist	echinoderm_larva_pluteus_early	protist_star	ctenophore_cydippid_no_tentacles	tunicate_doliolid_nurse
amphipods	echinoderm_larva_pluteus_typeC	pteropod_butterfly	ctenophore_cydippid_tentacles	tunicate_doliolid
appendicularian_fritillaridae	echinoderm_larva_pluteus_urchin	pteropod_theco_dev_seq	ctenophore_lobate	tunicate_partial
appendicularian_s_shape	echinoderm_larva_seastar_bipinnaria	pteropod_triangle	decapods	tunicate_salp_chains
appendicularian_slight_curve	echinoderm_larva_seastar_brachiolaria	radiolarian_chain	detritus_blob	tunicate_salp
appendicularian_straight	echinoderm_seacucumber_auricularia_larva	radiolarian_colony	detritus_filamentous	unknown_blobs_and_smudges
artifacts_edge	echinopluteus	shrimp_caridean	detritus_other	unknown_sticks
artifacts	ephyra	shrimp_sergestidae	diatom_chain_string	unknown_unclassified'
chaetognath_non_sagitta	euphausiids_young	shrimp_zoea	jellies_tentacles	hydromedusae_typeF
chaetognath_other	euphausiids	shrimp-like_other	polychaete	invertebrate_larvae_other_A
chaetognath_sagitta	fecal_pellet	siphonophore_calycophoran_abylidae	protist_dark_center	invertebrate_larvae_other_B
chordate_type1	fish_larvae_deep_body	siphonophore_calycophoran_rocketship_adult	protist_fuzzy_olive	
copepod_calanoid_eggs	fish_larvae_leptocephali	siphonophore_calycophoran_rocketship_young	hydromedusae_narco_young	
copepod_calanoid_eucalanus	fish_larvae_medium_body	siphonophore_calycophoran_sphaeronectes_stem	hydromedusae_narcomedusae	
copepod_calanoid_flatheads	fish_larvae_myctophids	siphonophore_calycophoran_sphaeronectes_young	hydromedusae_other	
copepod_calanoid_frillyAntennae	fish_larvae_thin_body	siphonophore_calycophoran_sphaeronectes	hydromedusae_partial_dark	
copepod_calanoid_large_side_antennatucked	fish_larvae_very_thin_body	siphonophore_other_parts	hydromedusae_shapeA_sideview_small	
copepod_calanoid_large	heteropod	siphonophore_partial	hydromedusae_shapeA	
copepod_calanoid_octomoms	hydromedusae_aglaura	siphonophore_physonect_young	hydromedusae_shapeB	
copepod_calanoid_small_longantennae	hydromedusae_bell_and_tentacles	siphonophore_physonect	hydromedusae_sideview_big	
copepod_calanoid	hydromedusae_h15	stomatopod	hydromedusae_solmaris	
copepod_cyclopoid_copilia	hydromedusae_haliscera_small_sideview	tornaria_acorn_worm_larvae	hydromedusae_solmundella	
copepod_cyclopoid_oithona_eggs	hydromedusae_haliscera	trichodesmium_bowtie	hydromedusae_typeD_bell_and_tentacles	

Approaches

Prepocessing

- Resizing
- Augmentation
- Edge detaction - Canny filter and BW

Feature Extraction

- PCA
- GLCM - Haralick Features
- DCT

Models **Ensembling**

- SVM - RandomForestClassifier
- Neural Network - xgboost, ExtraTree,
- Dicision Tree - Adaboost
- KNeighborsClassifier
-

% GLCM Features (Soh, 1999; Haralick, 1973; Clausi 2002)
% f1. Uniformity / Energy / Angular Second Moment (done)
% f2. Entropy (done)
% f3. Dissimilarity (done)
% f4. Contrast / Inertia (done)
% f5. Inverse difference
% f6. correlation
% f7. Homogeneity / Inverse difference moment
% f8. Autocorrelation
% f9. Cluster Shade
% f10. Cluster Prominence
% f11. Maximum probability
% f12. Sum of Squares
% f13. Sum Average
% f14. Sum Variance
% f15. Sum Entropy
% f16. Difference variance
% f17. Difference entropy
% f18. Information measures of correlation (1)
% f19. Information measures of correlation (2)
% f20. Maximal correlation coefficient
% f21. Inverse difference normalized (INN)
% f22. Inverse difference moment normalized (IDN)

Results

Primilary Results (Error - Logloss)

C =10

SVC(degree=5, kernel='linear')

('Cross Val Error: ', 0.4770399735711926)

('Training Error: ', 0.31020812685827553)

('Score', 0.52296002642880735)

SVC(degree=3, kernel='rbf')

('Cross Val Error: ', 0.47092831185992734)

('Training Error: ', 0.08462724369562824)

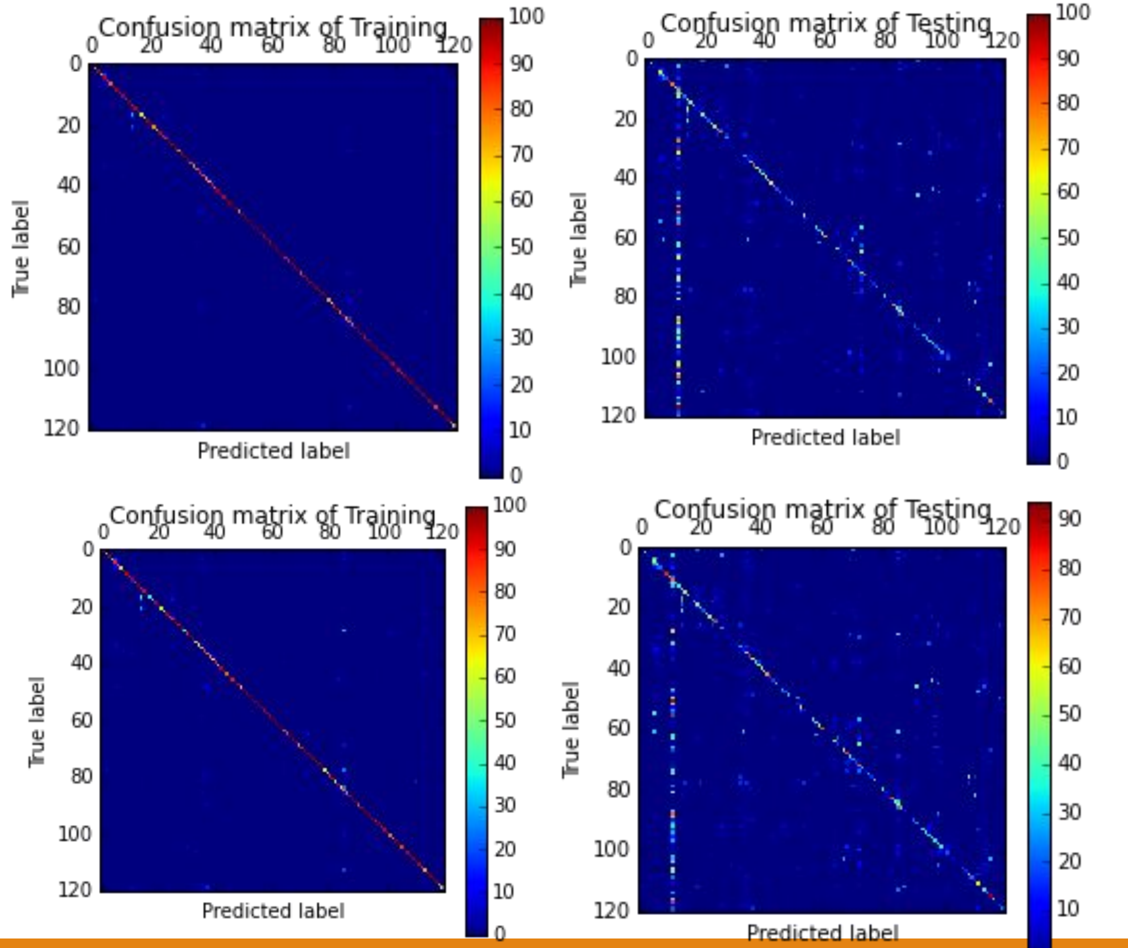
('Score', 0.52907168814007266)

SVC(degree=5, kernel='rbf')

('Cross Val Error: ', 0.45201519656425504)

('Training Error: ', 0.12884043607532211)

('Score', 0.54798480343574496)



Results

Log loss

-- 7.977

-- 6.57

-- 6.2

-- 6.19

-- 4.7

-- 4.4

-- 2.18

-- 1.82

Log loss

-- 1.77

-- 1.73

-- 1.533

-- 1.47

-- 1.39

-- 1.355

-- **1.3856** --Last Achieved

2. Mutual Liberty – Regression

Lesson Learned

- Not to combine very accurate models
- Try different features for different models
- Boost models with different training sets
- Combine with cascading

Any Question????

Thank You..