UNIVERSITY of HOUSTON
DEPARTMENT OF COMPUTER SCIENCE

CS@UH Data Mining & Machine Learning Group
Computer Science Department at University of Houston
Dr. Eick's Group — Data Mining is the effort of finding needle in a haystack

Supervised Taxonomies –Algorithms and Benefits

Paul K. Amalaman, Christoph F. Eick
UH-DMML Group

Abstract

Given a set of examples belonging to different classes, the goal of supervised taxonomy generation is to identify class-uniform groups of organisms with close phylogenetic relatedness that corresponds to subclasses of the underlying class structure. By extracting such sub-trees from supervised taxonomies, prediction models that infer class membership of unlabeled organisms can ultimately be developed. Additionally, by analyzing the generated tree structure, a biologist can interpret the clustering result and gain insights into how the biological groupings are related.

Problem Statement and Research Goals

- ❑ Traditional phylogenetic-based taxonomy generation approaches rely on gene sequence proximities to merge clusters into larger clusters
- ❑ ***S**upervised **T**axonomies **(S.T.)*** on the other hand, are generated using both sequence proximity and background knowledge in the form of class labels.
- ❑ Given organisms belonging to different classes, supervised taxonomy aims to identify class-uniform groups that correspond to subclasses of the underlying class structure.

(a) Dataset

D. melanogaster	0	0.138	0.199	0.248	0.238	0.211	0.245	0.245	0.234	0.252	0.251	O
D. pseudoobscura	0.138	0	0.199	0.24	0.241	0.197	0.235	0.232	0.227	0.243	0.227	O
S. lebanonensis	0.199	0.199	0	0.256	0.259	0.211	0.244	0.247	0.239	0.245	0.234	R
S. albovittata	0.248	0.24	0.256	0	0.092	0.213	0.18	0.176	0.167	0.178	0.173	P
D. crassifemur	0.238	0.241	0.259	0.092	0	0.214	0.182	0.192	0.185	0.19	0.182	P
D. mulleri	0.211	0.197	0.211	0.213	0.214	0	0.177	0.177	0.177	0.18	0.18	O
D. affinidisjuncta	0.245	0.235	0.244	0.18	0.182	0.177	0	0.037	0.06	0.06	0.091	O
D. heteroneura	0.245	0.232	0.247	0.176	0.192	0.177	0.037	0	0.052	0.059	0.083	P
D. mimica	0.234	0.227	0.239	0.167	0.185	0.177	0.06	0.052	0	0.056	0.075	R
D. adiastola	0.252	0.243	0.245	0.178	0.19	0.18	0.06	0.059	0.056	0	0.091	R
D. nigra	0.251	0.227	0.234	0.173	0.182	0.18	0.091	0.083	0.075	0.091	0	R

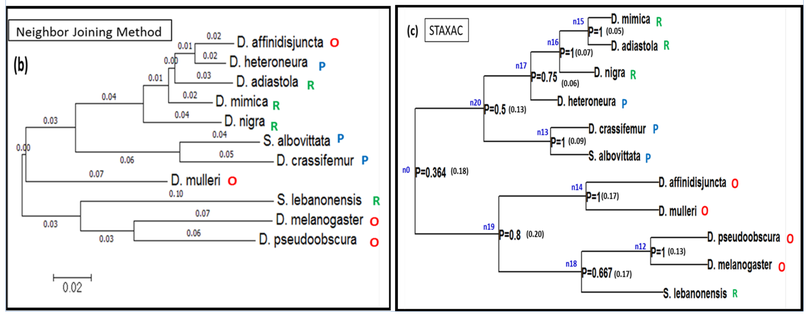

Fig. 1: Example of Tree Generated by S.T. Vs Tree Generated by Traditional Method (N.J)

Algorithms

Two Supervised Taxonomies (S.T.) generation algorithms are proposed:
Input: Dissimilarity matrix *(pairwise distances among the taxa)*
Output: Binary tree

- ❑ Bottom up approach called STAXAC (**S**upervised **T**axonomy **A**gglomerative **C**lustering)
- ❑ Top Down approach called MMCSC (**M**edoid of the **M**ajority **C**lass **S**upervised **C**lustering)

Algorithm 1: *STAXAC*

Input: examples with class labels and their dissimilarity matrix D.
Output: Hierarchical clustering
1. Start with a clustering X of one-object clusters.
2. $\forall C,C' \in X$ merge-candidate(C',C)$\leftrightarrow$ (1-NN$_X$(C') = C or 1-NN$_X$(C)=C')
3. WHILE there are merge-candidates (*C,C**) left
   BEGIN
   a. Merge the pair of merge-candidates (*C,C**) obtaining a new cluster C'=C$\cup$C* and
   a new clustering X' for which Purity(X') has the largest value
   b. Update merge-candidates:
   $\forall C''$ merge-candidate(*C',C''*) $\Leftrightarrow$ (merge-candidate(*C,C''*) *or* merge-candidate(*C*,C*))
   Extend dendrogram by drawing edges from C and C* to C'
   END
4. Return constructed dendrogram

Efficient implementation

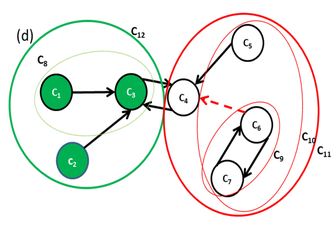

Fig. 2: S.T. Approach Generates High Purity Clusters

Definition 1: Given C $\in X$, merge-candidate set of C
$\mathcal{M}$ (C) = { C' | C' $\in$ 1-NN$_x$(C) $\vee$ C' $\in$ 1-NN$_x^{-1}$(C)} (1)

Definition 2: Given C $\in X$,C' $\in X$ and C* $\in X$ *such that C*= C' $\cup$ C*
$\mathcal{M}$ (C' $\cup$ C*) = $\mathcal{M}$ (C') $\cup$ $\mathcal{M}$ (C*)–{C',C*} (2)

Keep track of 1NN$_X$ relationships as follows
After each merge operation
update the $\mathcal{M}$ (C' $\cup$ C*) using equation (2)
If $\mathcal{M}$ (C' $\cup$ C*) =Ø then
$\mathcal{M}$ (C' $\cup$ C*) $\leftarrow$ compute (1NN$_X$(C))

Algorithm 2: *MMCSC*

Inputs: a set of examples with class labels and their dissimilarity matrix D; purity threshold θ.
Function: MMSCS(D, θ)
IF Purity(D)> θ terminate;
Separate the examples in tree node D into two subsets: one containing only sequences of the majority class D*majority* and the other composed of the remaining sequences, D*not_majority*
Compute the medoids of both sets: medoid(D*majority*), medoid(D*not_majority*)
#Split node D in two new nodes:D1,D2
FOR each d in D
IF d is closer to medoid(D*majority*) than it is to medoid(D*not_majority*)
assign d to D1
ELSE assign d to D2
Call MMCSC(D1, θ)
Call MMCSC(D2, θ)
END Function

Experimental Results

We evaluate STAXAC, and MMCSC against each other and against NJ (Neighbor Joining) a popular distance-based Phylogenetic approach using two criteria:
- ❑ Runtime
- ❑ Tree Topology
  1. Average Purity Per Depth
  2. Number of Sub-family Changes
     (From root node to a leaf-node, count number of changes in majority class label)
  3. Edited Tree Sizes
     (Merge leaf-nodes from bottom up until leaf-node purity <=threshold ϑ)

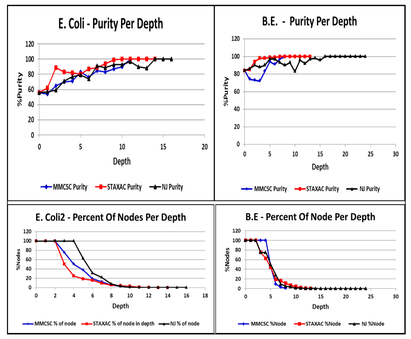

Fig. 3: Average Purity Per Depth

	Time in seconds	
Art#1 Sizes	MMCSC	STAXAC
20	0	0
50	0	0
100	0	0
150	0	0
200	0	1
250	2	2
300	3	3
500	14	22
800	59	90
1200	260	320
1800	780	1140

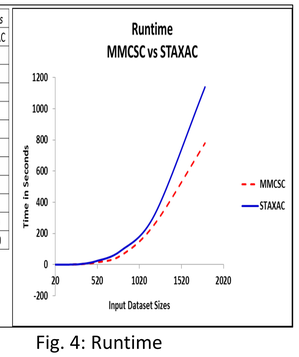

Fig. 4: Runtime

TABLE I: Number Of Sub-family Changes

	MMCSC	STAXAC	NJ
E. coli	34	**18**	35
BE	15	**7**	21
BV	53	**26**	56
BH	8	**4**	8
Art#2	13	**4**	49
Art#3	36	**6**	61

TABLE II: Edited Tree Sizes

Dataset name	MMCSC			STAXAC			NJ		
θ =>	100%	95%	85%	100%	95%	85%	100%	95%	85%
E. coli	103	103	101	**55**	**55**	**48**	123	123	123
BE	57	57	51	**25**	**22**	**11**	73	73	71
BV	197	197	197	**111**	**51**	**51**	273	273	269
BH	57	57	57	**19**	**17**	**3**	75	73	65
Art#2	71	71	71	**15**	**15**	**12**	139	139	139
Art#3	109	109	105	**17**	**17**	**17**	153	153	153

Conclusion & Future Work

Both algorithms can help a biologist interpret the clustering result and gain insights into how the clustering (biological groupings) are related in the context of an underlying class structure.

Future Work:
- Predictive Models using as input trees generated by S. T. approach
- Using trees generated by S. T. to determine if a phenotype (class label) within a closely related species is genetically induced.