HCOC: hierarchical classifier with overlapping class groups

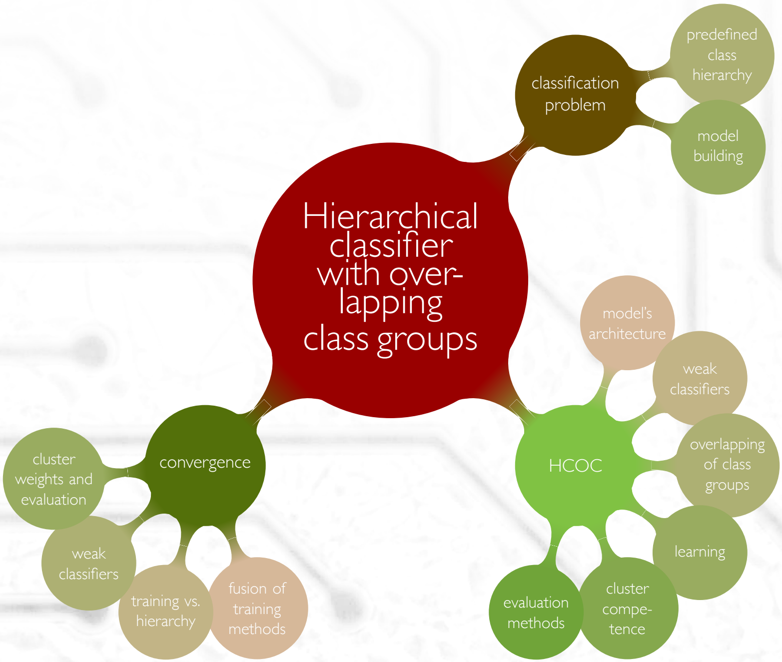
Igor T. Podolak

Group of Machine Learning Methods GMUM
Theoretical Foundations of Machine Learning, Będlewo 2015

20th February 2015

group of machine
gmum
learning research

tfml 2015
theoretical foundations
of machine learning, Będlewo


Hierarchical classifier with over-lapping class groups
classification problem
predefined class hierarchy
model building
HCOC
model's architecture
weak classifiers
overlapping of class groups
learning
cluster compe-tence
evaluation methods
convergence
cluster weights and evaluation
weak classifiers
training vs. hierarchy
fusion of training methods

Problem statement

1. high number of classes
   1.1 right model architecture
   1.2 unbalanced number of class examples
   1.3 divide the problem into simpler ones?

2. what is a hierarchical classification?
   2.1 predefined class hierarchy
   2.2 map natural class groups to the model architecture

3. solve by splitting the output classification space
   3.1 hierarchically group examples from similar classes
   3.2 hipothesis: if examples from classes A and B are frequently mistaken, then they are probably similar
      3.2.1 define the similarity of classes with the frequency of incorrect classifications
   3.3 find the class groups using weak classifiers (hierarchically)

Problem

tasks to solve

4. HCOC: fusion of supervised training in nodes and unsupervised cluster building
   4.1 supervised training returns class probability vectors
      4.1.1 hypothesis: similar classification vectors $\Longrightarrow$ examples hard to differentiate $\Longrightarrow$ classes are similar
   4.2 clustering in classifiers activation space recovers classification errors
   4.3 a classifier trained in supervised mode might be weak

HCOC architecture

classifier tree root

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

HCOC architecture

each node is a separate classifier

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

$P(C = A|x)$

Cl in node returns a class probability vector

Similar activations represent similar classes, thus we may split them into subproblems

HCOC architecture

some classes are classified similarly

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A B E G J X

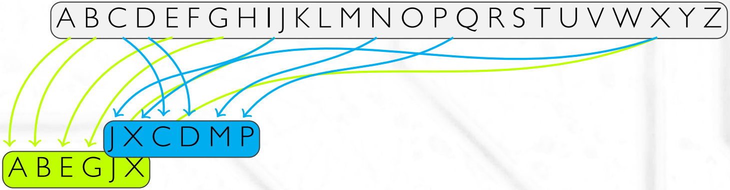
Similarly classified classes are grouped together into clusters

Grouping makes it possible to recover some classification errors later

Clusters may overlap

HCOC architecture

classifiers are **weak**



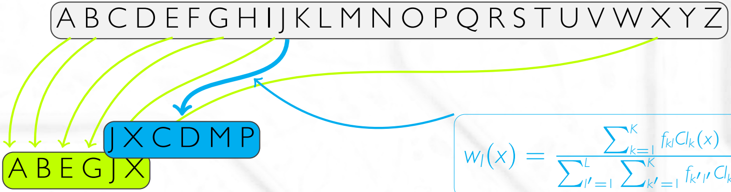
K-class classifier is at least weak if the probability that the activation for the true class is at least $1/K$

K-class Cl is weak iff $\mathbb{E}[Cl_i(x)|true(x) = i]$ for true class is higher than $\alpha(K)$, where $\alpha(K) = \min_{\alpha} \left[\alpha : (-1)^i \binom{K-1}{i} (1 - \frac{i\alpha}{1-\alpha})_+^{K-2} > \frac{1}{K}\right]$

K	2	3	4	5	6	7	10	20
$\alpha(K)$	0.500	0.400	0.333	0.289	0.256	0.230	0.180	0.110

HCOC architecture

cluster weights are computed separately for each given input vector



$$w_l(x) = \frac{\sum_{k=1}^{K} f_{kl} Cl_k(x)}{\sum_{l'=1}^{L} \sum_{k'=1}^{K} f_{k'l'} Cl_{k'}(x)}$$

$$f_{kl} = \begin{cases} 1 & C_k \in Q^l \\ 0 & C_k \notin Q^l \end{cases}$$

$w_l(x)$ corresponds to **softmax**, therefore a model that predicts a cluster is a classifier

different competence measures

HCOC architecture

clustering methods

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

J X C D M P

A B E G J X

$$w_l(x) = \frac{\sum_{k=1}^{K} f_{kl} Cl_k(x)}{\sum_{l'=1}^{L} \sum_{k'=1}^{K} f_{k'l'} Cl_{k'}(x)}$$

SAHN based, Bayesian, GNG based

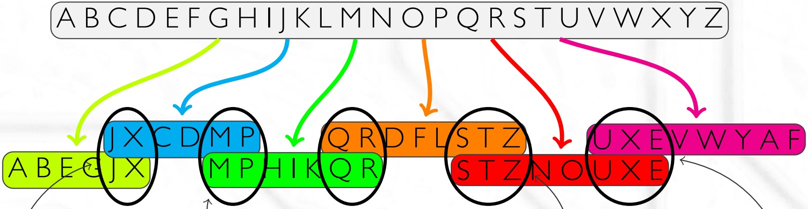
Bayesian: join classes using error matrix

GNG: build clusters online **simultaneously** with classifier training

control **diversity** of clusters and descendant classifiers

HCOC architecture

clusters overlap



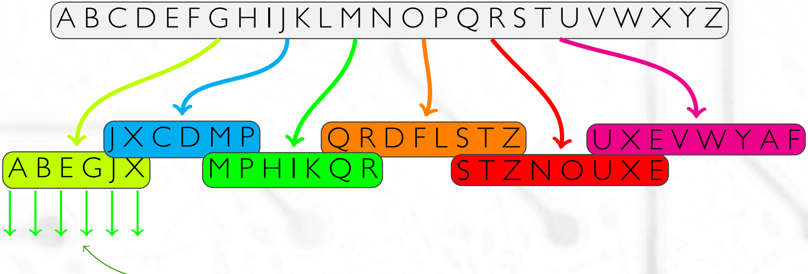
individual classes may belong to several clusters

which clusters do overlap comes from the **inability** of classifier to solve the actual problem: an architecture corresponding to the problem is being built

clusters overlap increases the HCOC accuracy ability

HCOC architecture

independence of HCOC base classifiers

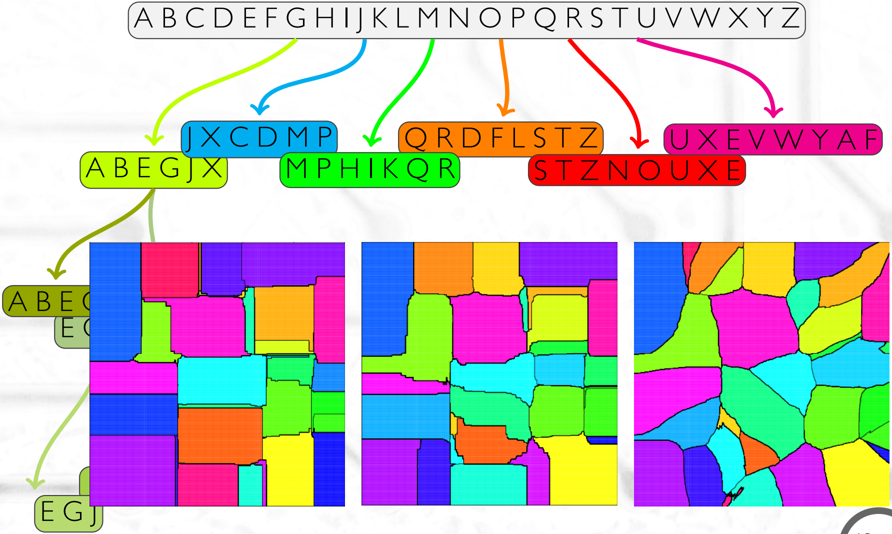


each *CI* solves its own subproblem

subtrees may be built independently in parallel

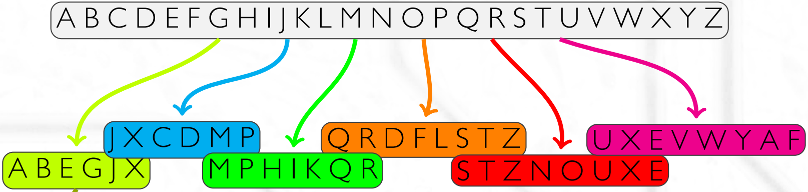
HCOC architecture

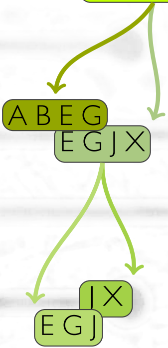
classification on different levels



HCOC architecture

HCOC convergence of training



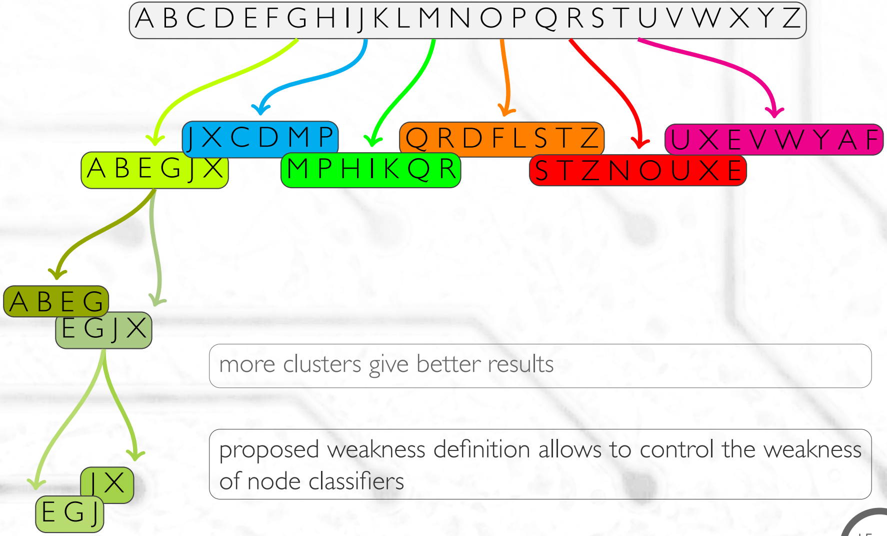


Let HCOC be two-level model with $\ell(x, t, h(x)) = (t - h(x))^2$. The HCOC risk is lower than risk of root Cl^0 provided that classes are spread independently betwen clusters and $\sum_k \sum_l \sum_i p_i f_{kl} m^l_{ii} m^0_{ik}$ is maximised and higher than $\sum_i p_i m_{ii} - \sum_i p_i \sum_k (m_{ik})^2$

HCOC is built recursively: the above statement strengthens with each level added

HCOC architecture

weakness property of base classifiers

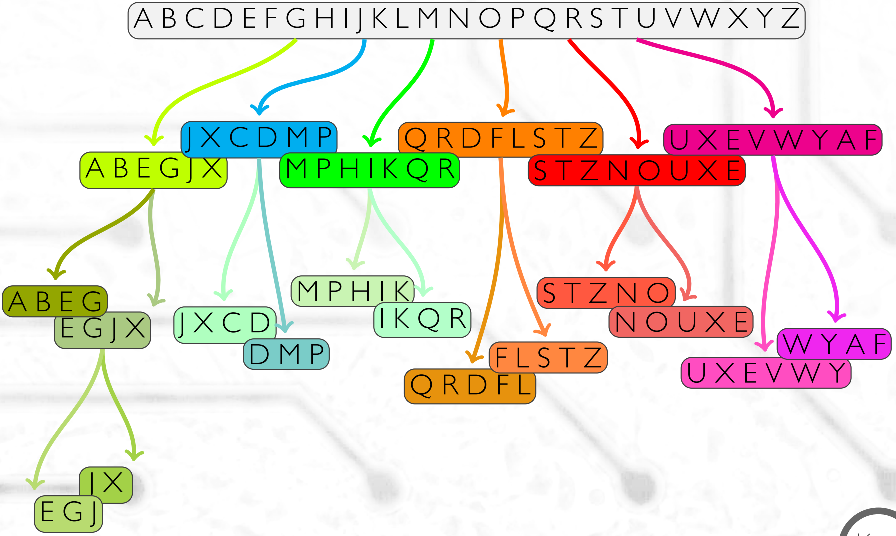


more clusters give better results

proposed weakness definition allows to control the weakness of node classifiers

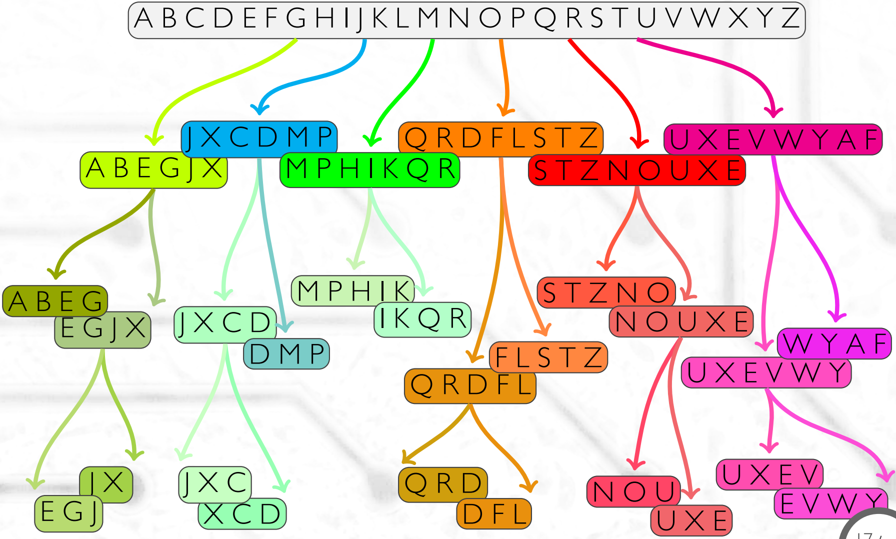
HCOC architecture

it is possible to build several simple classifiers independently



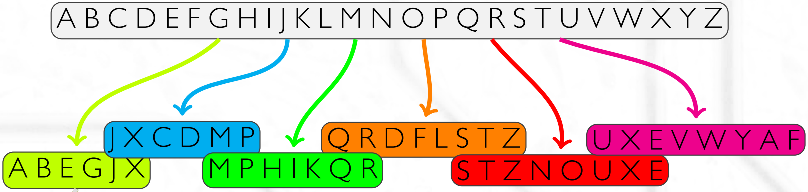
HCOC architecture

complete classifier



HCOC architecture

evaluation of HCOC



$P(C_j|x) = y_j(x) = \sum_{l=1}^{L} w_l(x) y_j^l(x)$ where $y_j^l(x)$ is the return value of descendent classifier with competence $w_l(x)$ for given x

geometric mean in overlaps

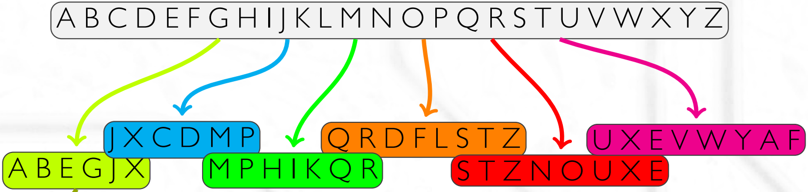
possible methods: **ALL-SUBTREES**, **SINGLE-PATH**, **RESTRICTED** and **α-RESTRICTED**

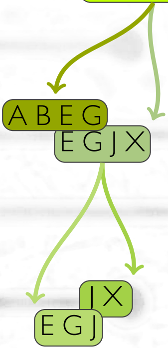
ALL-SUBTREES evaluate all paths

SINGLE-PATH select only the highest competence w_l path

HCOC architecture

Restricted and α-**Restricted** approaches





only some clusters shall have competence higher than *a priori* p_i probability of classes

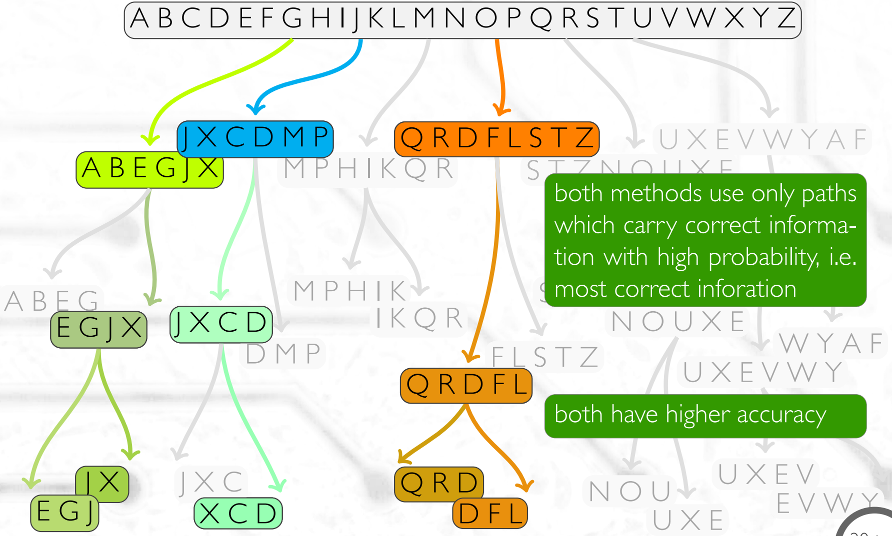
evaluation of others is equal of adding some noise

Restricted use only clusters where at least one class has activation higher than *a priori* p_i

α-**Restricted** use only clusters where $\exists\, C_k\ Cl_k(x) > \alpha(K)$ (weakness condition is being used)

HCOC architecture

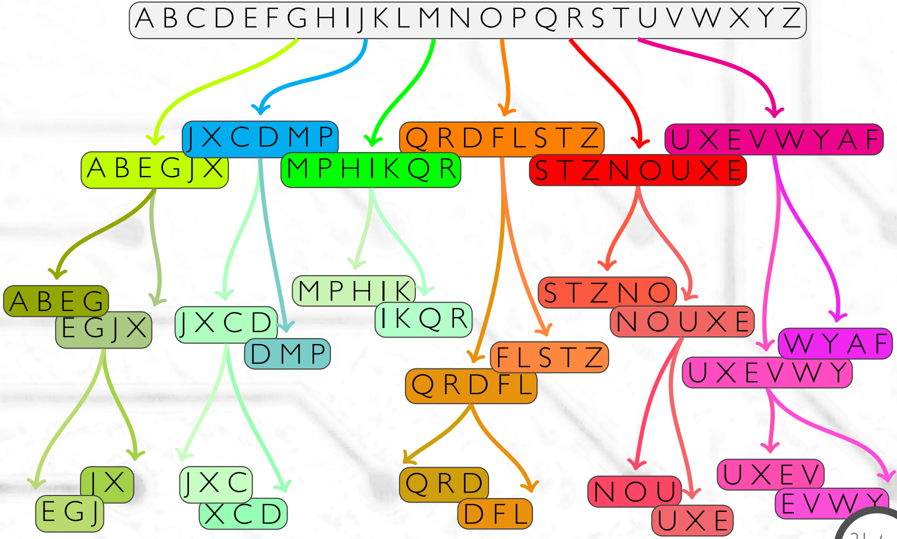
Restricted and α-Restricted



both methods use only paths which carry correct information with high probability, i.e. most correct inforation

both have higher accuracy

HCOC architecture



HCOC

properties

1. fusion of supervised and unsuperised training
   1.1 possible solution for a high number of output classes

2. a split corresponds to complexity of subproblem at a node
   2.1 subproblems overlap, hence improvement of accuracy
   2.2 problem split through unsupervised clustering of class outputs
   2.3 clustering control results in different resulting subproblems
   2.4 different clustering methods
      2.4.1 parallel training

3. weak classifiers in nodes
   3.1 probabilistic measure of classifiers weakness
   3.2 provides for simple weakness control

4. classifiers competence compted separately for each input vector classified

5. different methods of evaluation
   5.1 simple reduction of unimportant information (noise)
   5.2 evaluation is related to classifier weakness

6. HCOC properties
   6.1 classifier risk is minimised with new layers being added
   6.2 control of diversity

tfml 2015

theoretical foundations of machine learning, Będlewo

GMUM

group of machine
gmum
learning research