

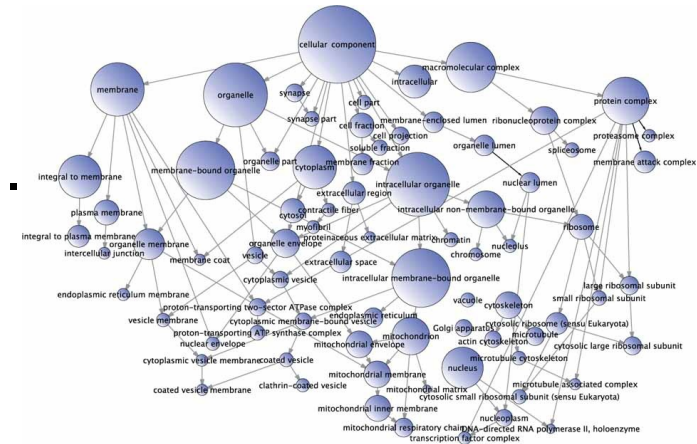
Ensembles for predicting structured outputs

Dragi Kocev

Motivation

■ Increasing amounts of structured data

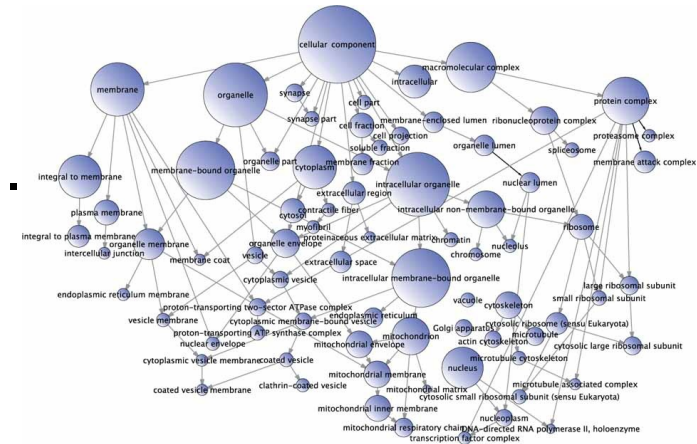
- Vectors
- Hierarchies – trees, DAGs,...
- Sequences – time series



Motivation

■ Increasing amounts of structured data

- Vectors
- Hierarchies – trees, DAGs,...
- Sequences – time series



■ Success of the ensemble methods in simple classification and regression

Outline

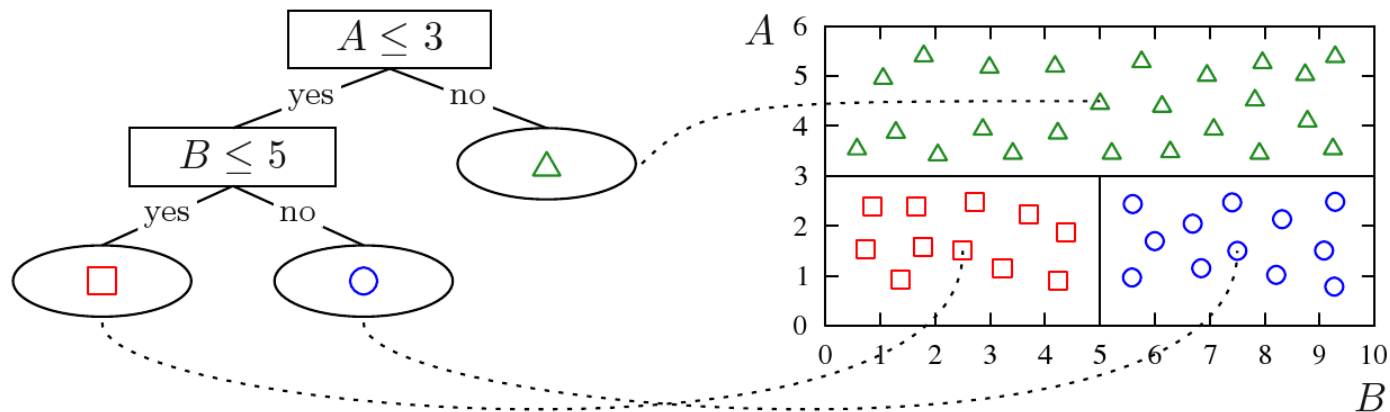
- Background
 - Structured outputs
 - Predictive Clustering Trees (PCTs)
 - PCTs for HMLC
- Ensembles of PCTs
- Experimental evaluation
- Application in functional genomics
- Conclusions

Structured outputs

- Target in supervised learning
 - Single discrete or continuous variable
- Target in structured prediction
 - Vector of discrete or continuous variables
 - Hierarchy – tree or DAG
 - Sequences – time series
- Solutions
 - De-composition to simpler problems
 - Exploitation of the structure

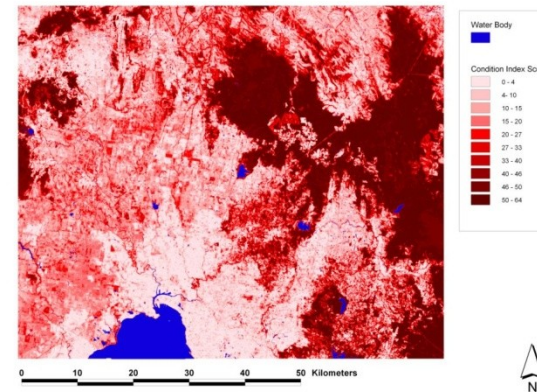
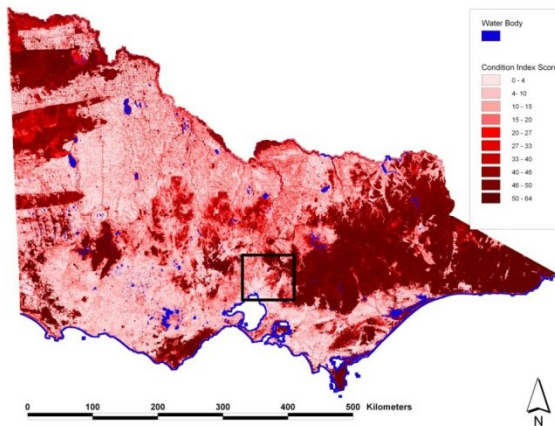
Predictive Clustering Trees

- Standard Top-Down Induction of DTs
- Hierarchy of clusters
- Distance measure: minimization of intra-cluster variance
- Instantiation of the variance for different tasks



PCTs – Multiple numeric targets

- Condition of vegetation in Victoria, Australia
- Habitat Hectares Index
 - Large Trees, Tree Canopy Cover, Understorey, Litter, Logs, Weeds and Recruitment



- Euclidean distance

$$Var(E) = \sum Var(E, y_t)$$

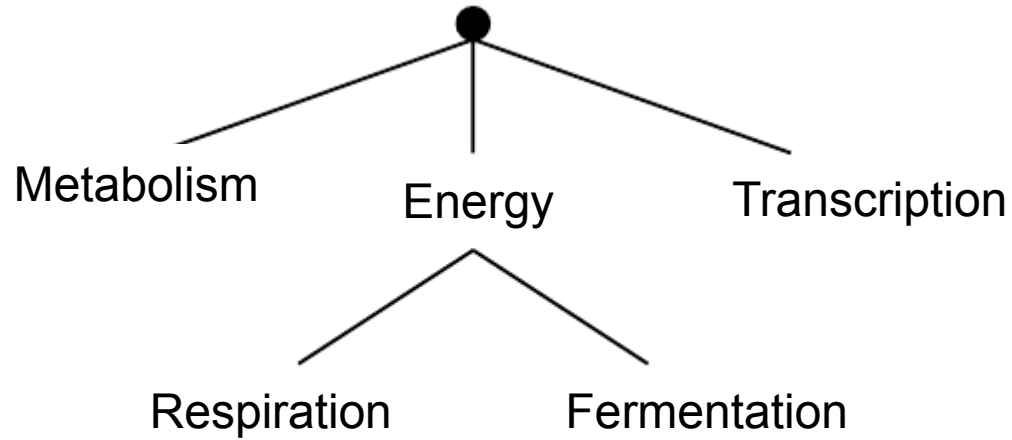
PCTs – Multiple discrete targets

- Mediana – Slovenian media space
 - 8000 Questionnaires about reading and TV habits and life style of Slovenians
- What type of people read:
 - Delo, Dnevnik, Ekipa, Slovenske Novice, Večer

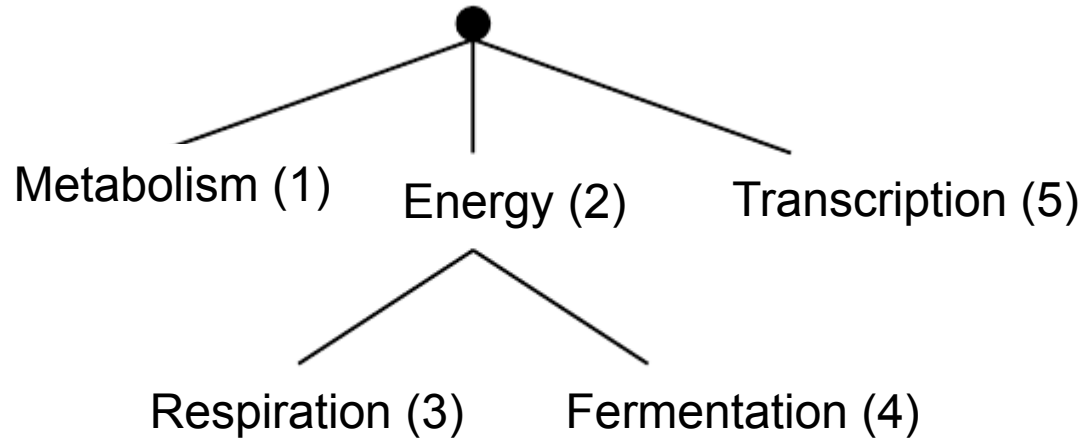
$$Var(E) = \sum Entropy(E, y_t)$$

$$Var(E) = \sum Gini(E, y_t)$$

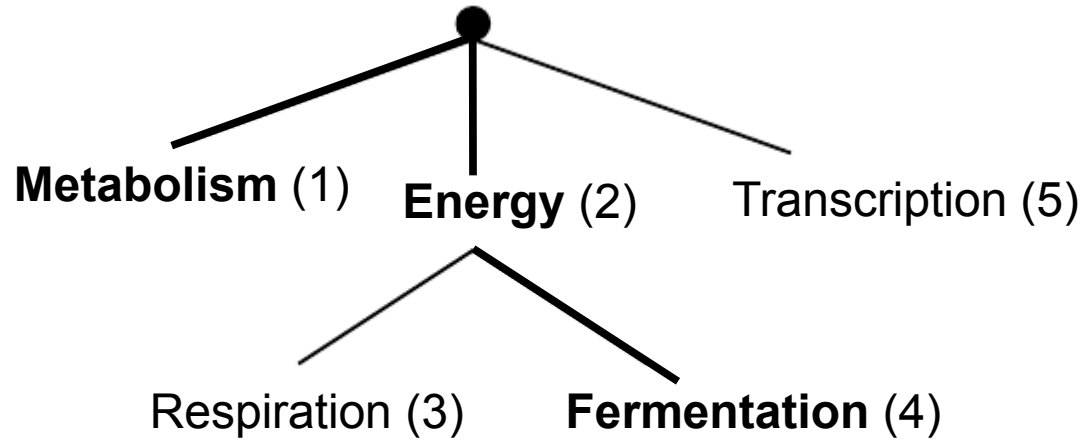
PCTs - HMLC



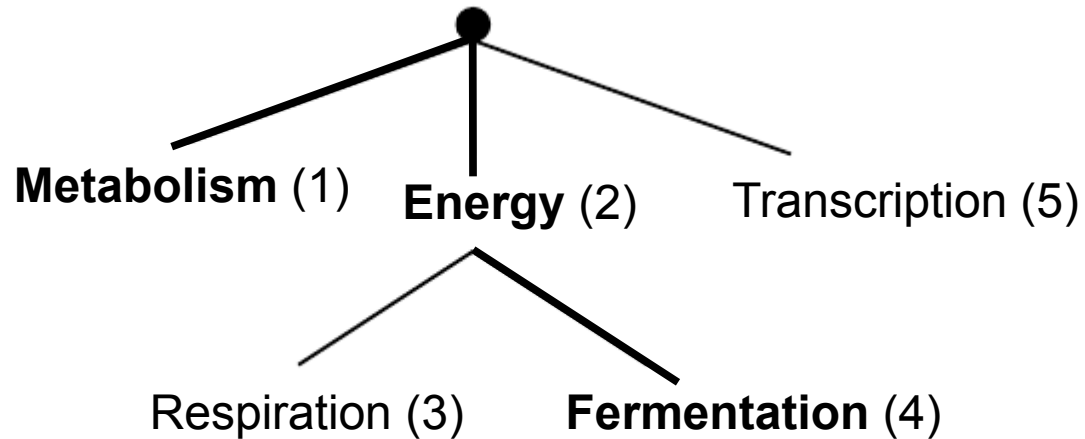
PCTs - HMLC



PCTs - HMLC

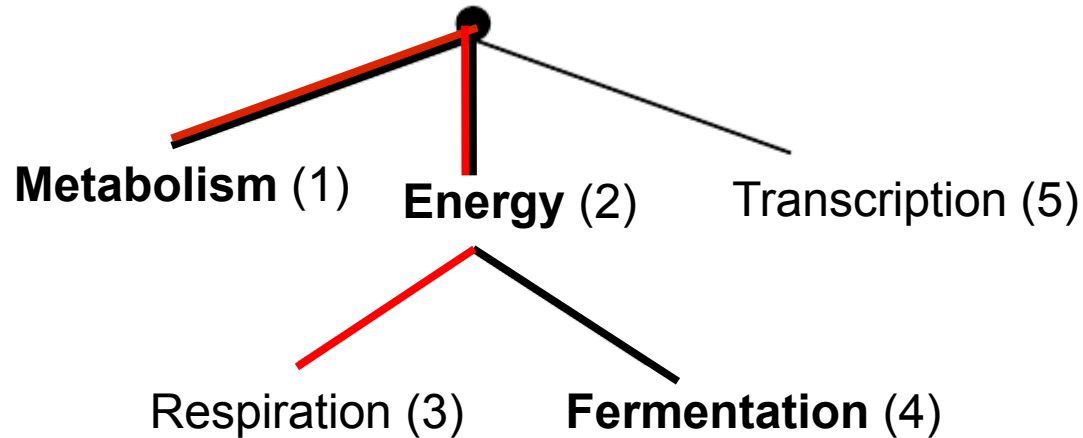


PCTs - HMLC



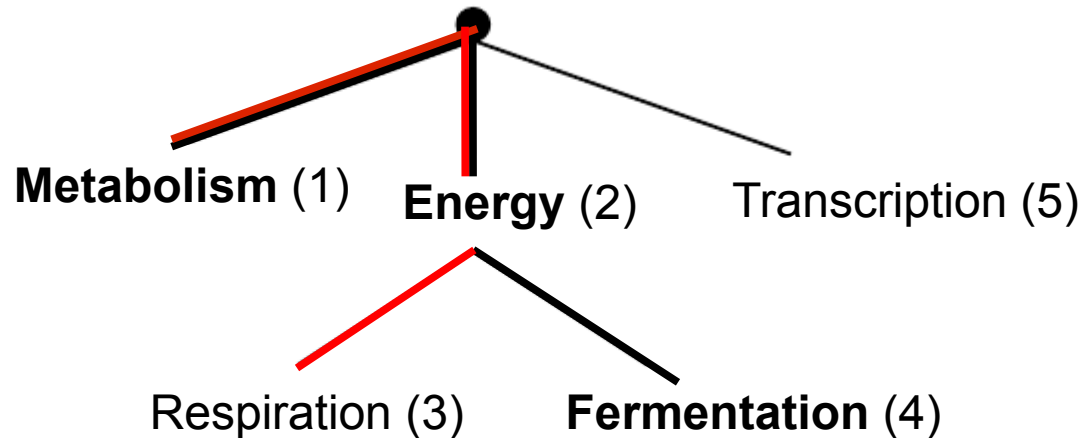
$$v_i = \begin{matrix} (1)(2)(3)(4)(5) \\ [1, 1, 0, 1, 0] \end{matrix}$$

PCTs - HMLC



$$\begin{array}{c} (1)(2)(3)(4)(5) \\ v_i = [1, 1, 0, 1, 0] \\ v_i = [1, 1, 1, 0, 0] \end{array}$$

PCTs - HMLC



$$\begin{matrix} & (1)(2)(3)(4)(5) \\ v_i = & [1, 1, 0, 1, 0] \\ v_i = & [1, 1, 1, 0, 0] \end{matrix}$$

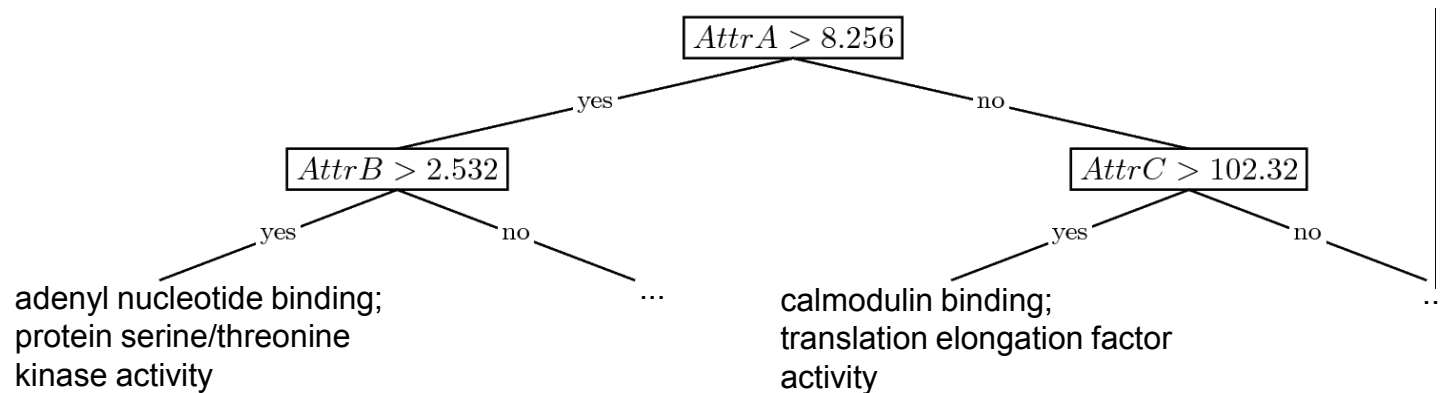
$$Var(E) = \frac{\sum d(v_i, \hat{v})^2}{|E|}$$

$$d(v_i, \hat{v}) = \sqrt{\sum \omega(c_i) \cdot (v_{1,i} - v_{2,i})^2}$$

$$\omega(c_i) = \omega_0 \cdot \omega(par(c_i))$$

PCTs - HMLC

- A leaf stores the mean label: proportion of the examples belonging to each class
- Prediction is made by using user-defined threshold



Outline

- Background
 - Structured outputs
 - Predictive Clustering Trees (PCTs)
 - PCTs for HMLC
- **Ensembles of PCTs**
- Experimental evaluation
- Application in functional genomics
- Conclusions

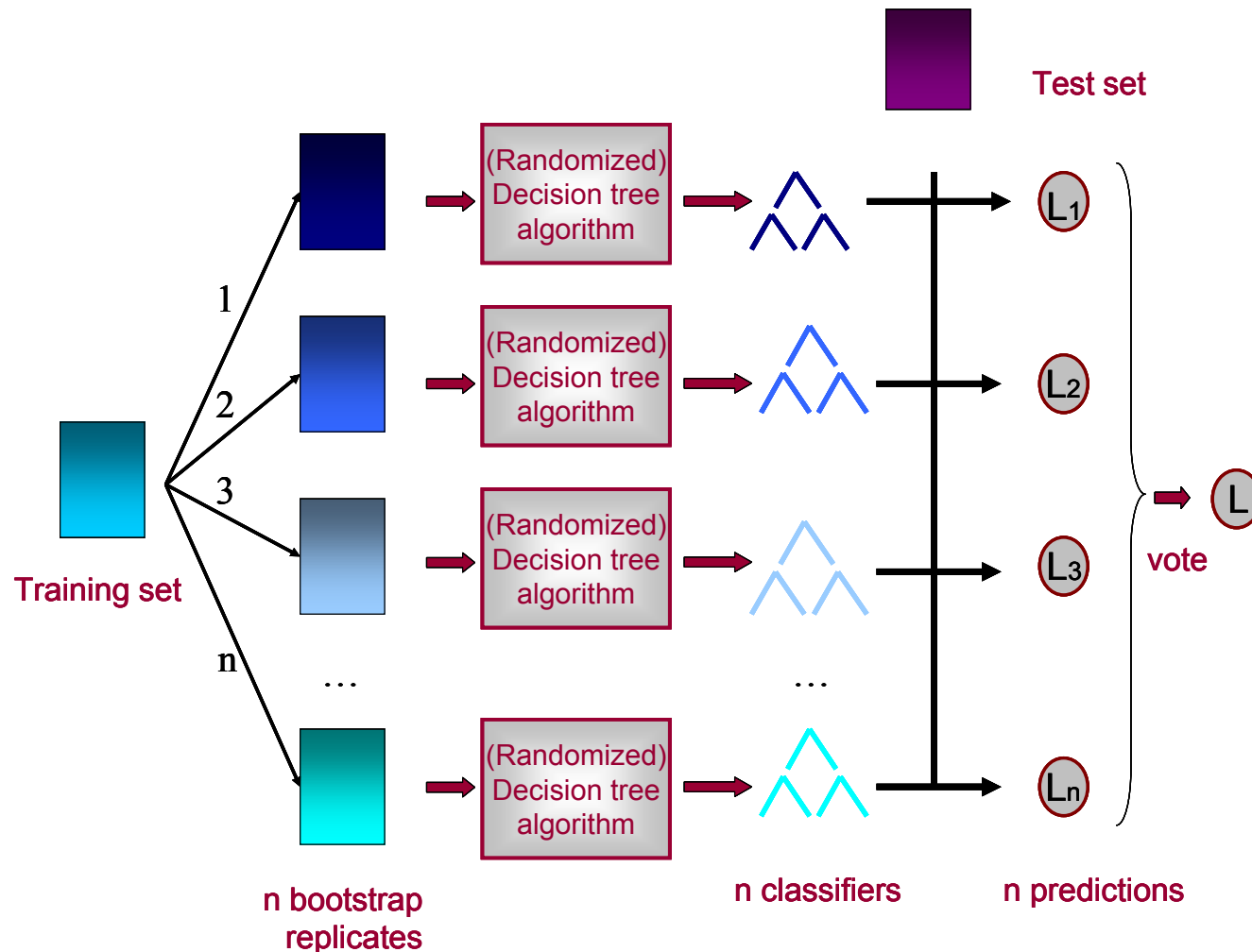
Ensemble Methods

- Ensembles are set of predictive models
 - Unstable base classifiers
- Voting schemes to combine the predictions into a single prediction
- Ensemble learning approaches
 - Modification of the data
 - Modification of the algorithm

Ensemble Methods

- Ensembles are set of predictive models
 - Unstable base classifiers
 - Voting schemes to combine the predictions into a single prediction
 - Ensemble learning approaches
 - Modification of the data ← **Bagging**
 - Modification of the algorithm
- } **Random Forest**

Ensemble Methods: Algorithm



Ensembles for structured outputs

- PCTs as base classifiers
- Voting schemes for the structured outputs
 - MT Classification: majority and probability distribution vote
 - MT Regression and HMLC: average
 - For an arbitrary structure: prototype calculation function

Outline

- Background
 - Structured outputs
 - Predictive Clustering Trees (PCTs)
 - PCTs for HMLC
- Ensembles of PCTs
- **Experimental evaluation**
- Application in functional genomics
- Conclusions

Experimental hypotheses

- How many base classifiers are enough?
- Does the ensembles of PCTs lift the predictive performance of a single PCT?
- Does the ensembles of PCTs are better than ensembles for each sub-component separately?
- Which approach is more efficient in terms of time and size of the models?

Datasets

| | Datasets | Examples | Descriptive attributes | Size of Output |
|-------------------|----------|------------|------------------------|----------------|
| MT Regression | 14 | 154..60607 | 4..160 | 2..14 |
| MT Classification | 11 | 154..10368 | 4..294 | 2..14 |
| HMLC | 10 | 988..10000 | 80..74435 | 36..571 |

- Datasets with multiple targets
 - Mainly environmental data acquired in EU and Slovenian projects
- HMLC
 - Image classification
 - Text classification
 - Functional genomics

Experimental design

- Types of models
 - PCTs and ST trees (with F-test pruning)
 - Ensembles of PCTs and ensembles of ST trees
- PCTs for HMLC weight for the distance - 0.75
- Number of base classifiers (unpruned)
 - Classification: 10, 25, 50, 75, 100, 250, 500, 1000
 - Regression: 10, 25, 50, 75, 100, 150, 250
 - HMLC: 10, 25, 50, 75, 100

Experimental design (ctd.)

- Random forest – feature subset size
 - Multiple Targets: log
 - HMLC: 10%
- 10-fold cross-validation
- MT Classification - Accuracy
- MT Regression - correlation coefficient, RMSE, RRMSE
- HMLC - Precision-Recall (PR) curves, Area under PRCs
- Friedman and Nemenyi statistical tests

Precision-Recall Curves

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

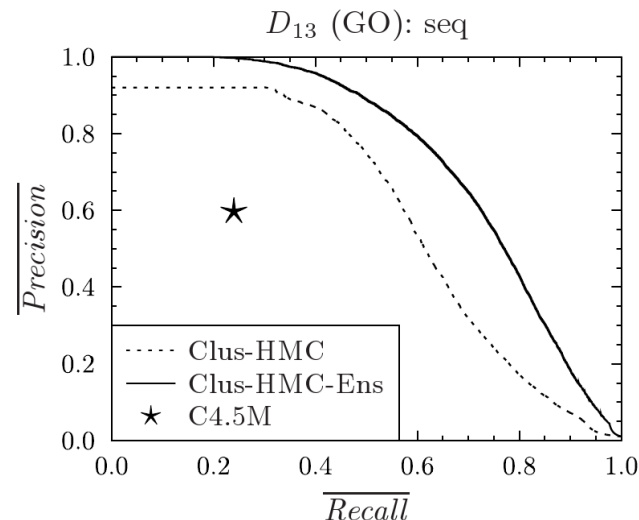
- PR curve plots Precision as a function of the Recall
- Combination of the curves per class
 - Micro-averaging: Area under the average PRC
 - Macro-averaging: Average area under the PRCs

Precision-Recall Curves

$$Precision = \frac{TP}{TP + FP}$$

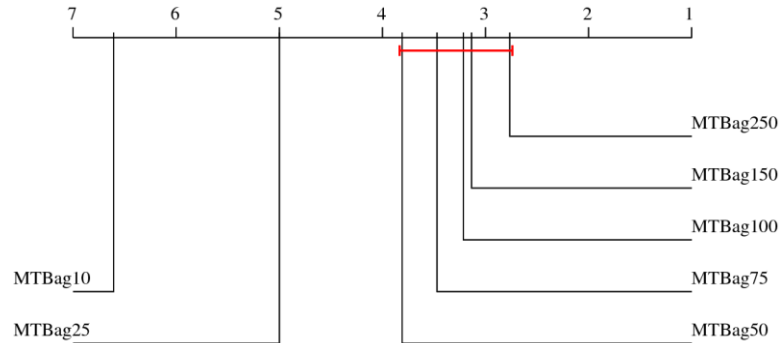
$$Recall = \frac{TP}{TP + FN}$$

- PR curve plots Precision as a function of the Recall
- Combination of the curves per class
 - Micro-averaging: Area under the average PRC
 - Macro-averaging: Average area under the PRCs

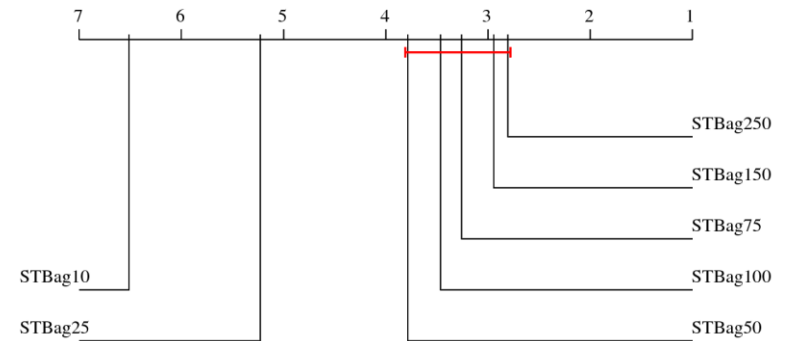


Results – Regression (RRMSE)

MT Bagging

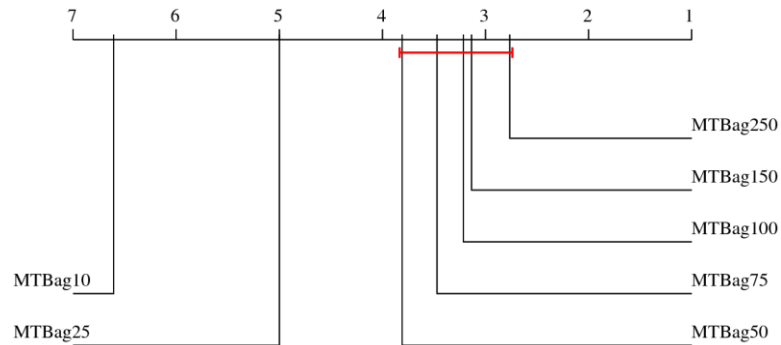


ST Bagging

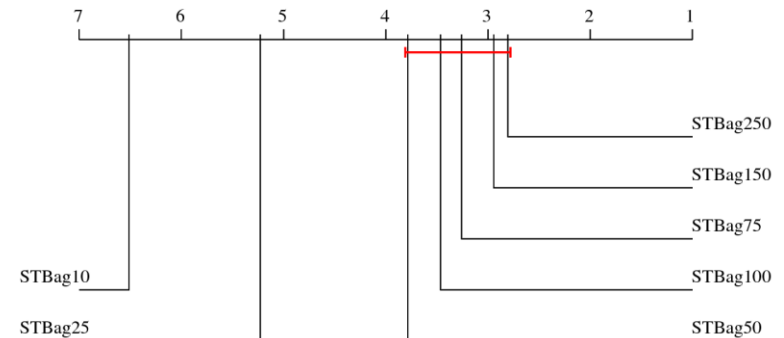


Results – Regression (RRMSE)

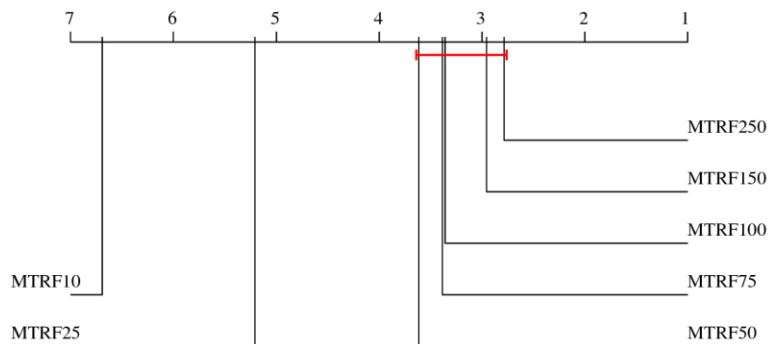
MT Bagging



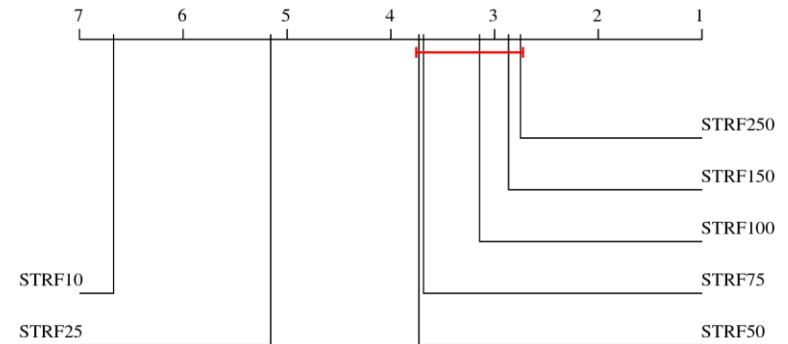
ST Bagging



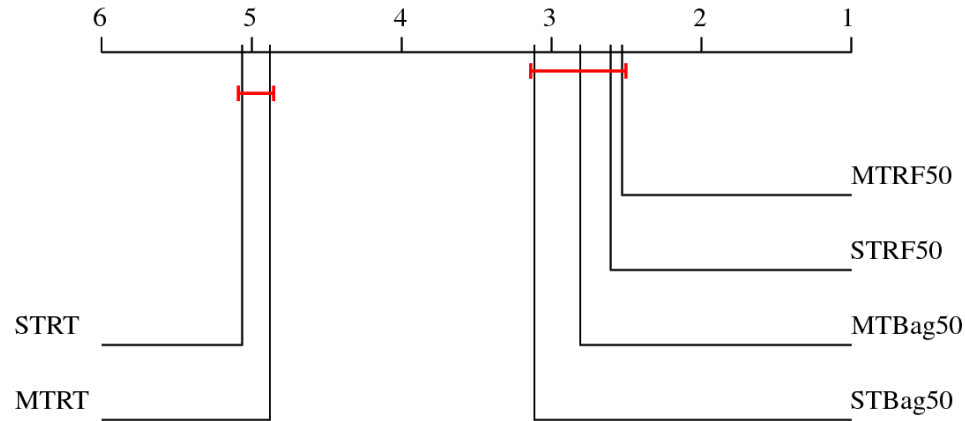
MT Random forest



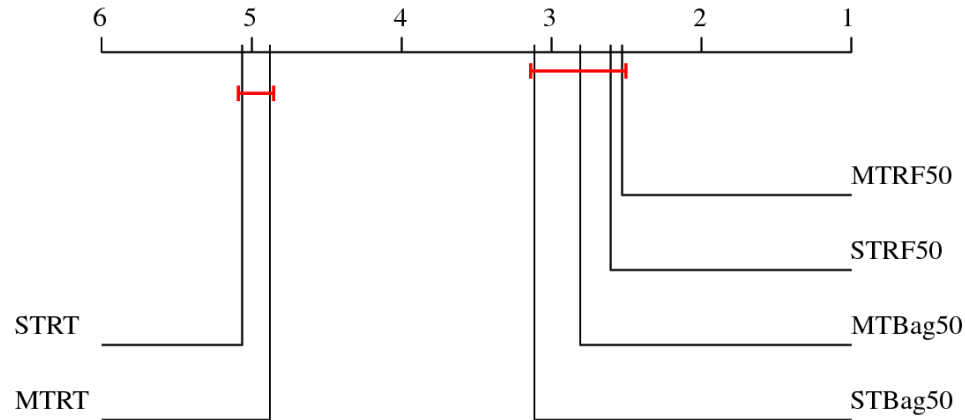
ST Random forest



Results – Regression (RRMSE)



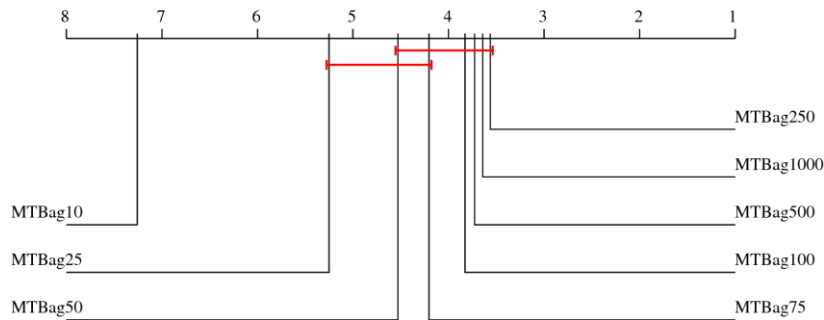
Results – Regression (RRMSE)



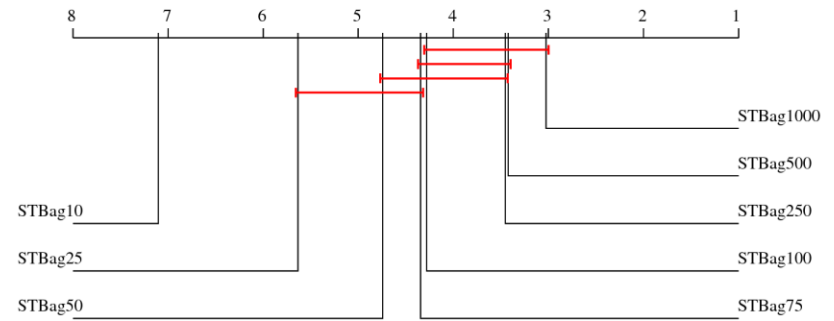
- MT Ensembles perform better than
 - single MT tree (stat. sign.)
 - ST Ensembles
 - faster to learn than ST ensembles (stat. sign.)
 - smaller models than ST ensembles (stat. sign.)

Results – Classification

MT Bagging

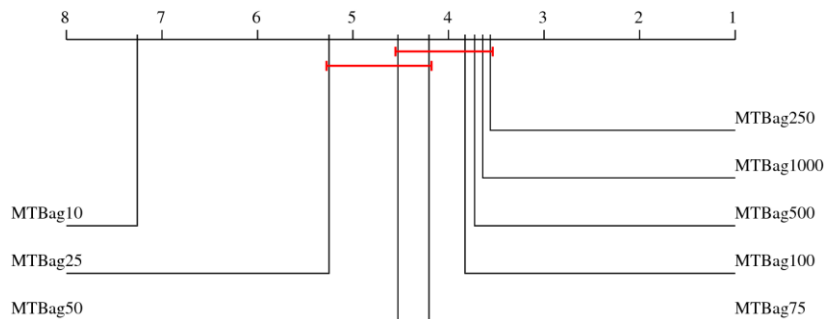


ST Bagging

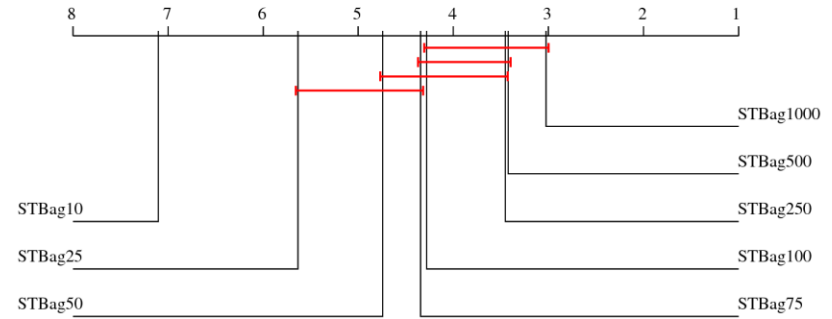


Results – Classification

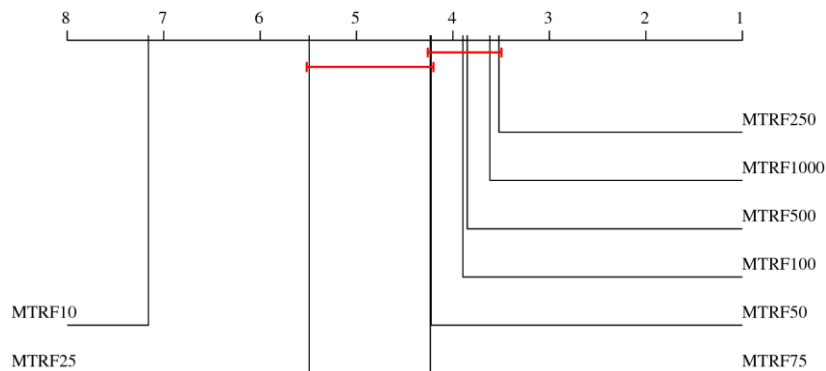
MT Bagging



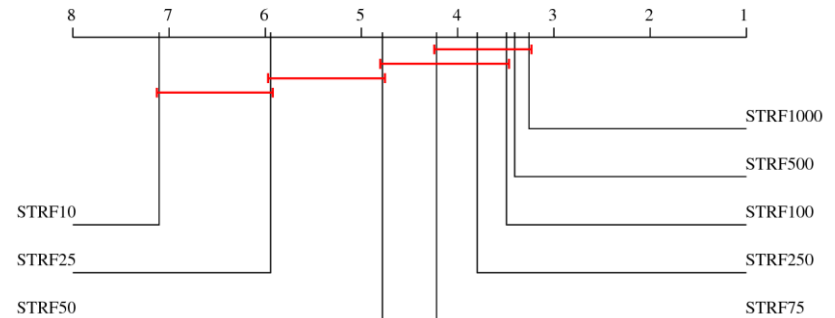
ST Bagging



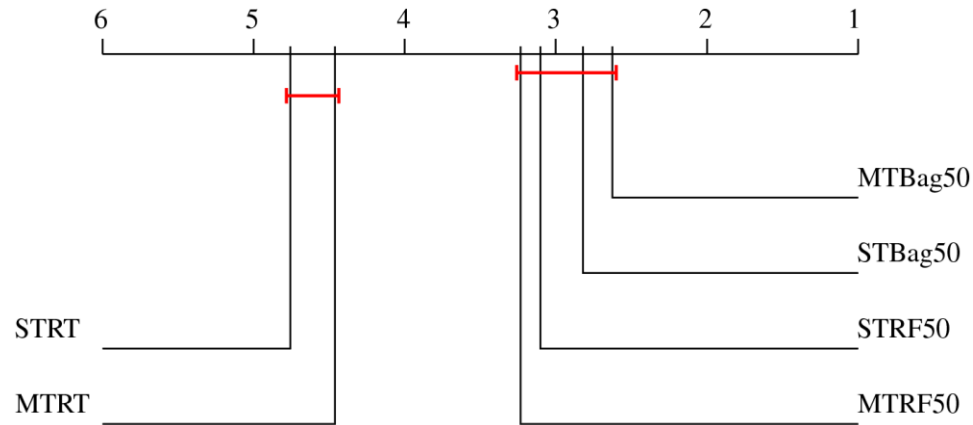
MT Random forest



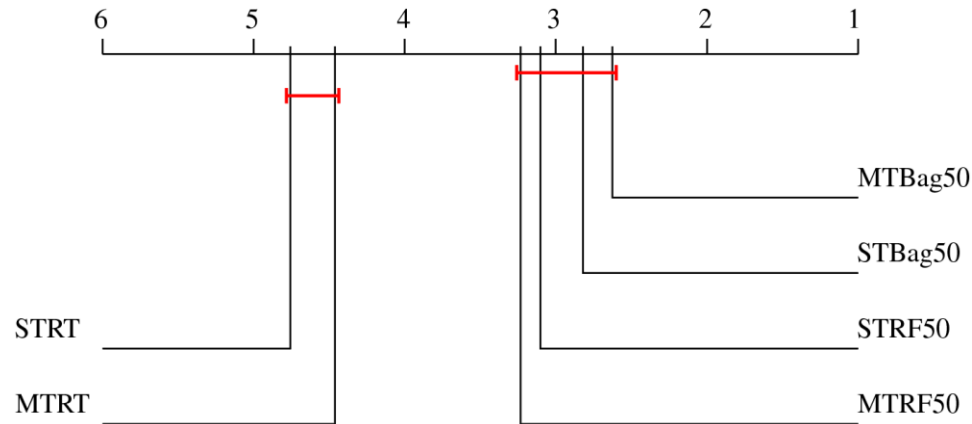
ST Random forest



Results – Classification



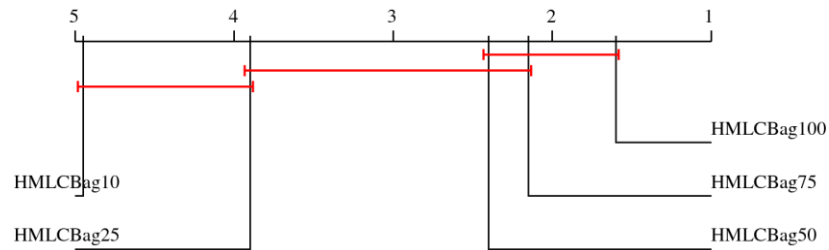
Results – Classification



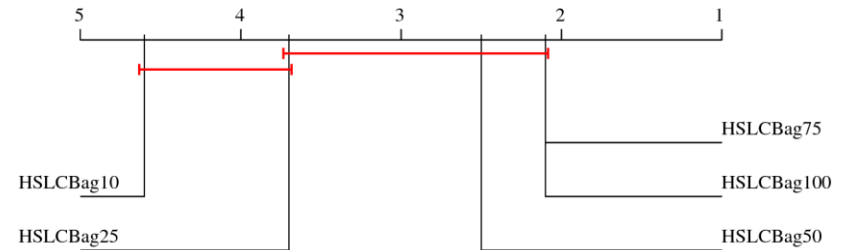
- MT Ensembles perform better than
 - single MT tree (stat. sign.)
 - faster to learn than ST ensembles (stat. sign.)
 - smaller models than ST ensembles (stat. sign.)
- MT Bagging is better than ST Bagging, while MT random forest is worse than ST random forest

Results – HMLC

HMLC Bagging

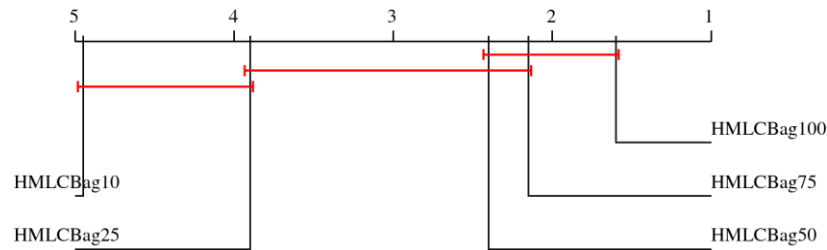


HSLC Bagging

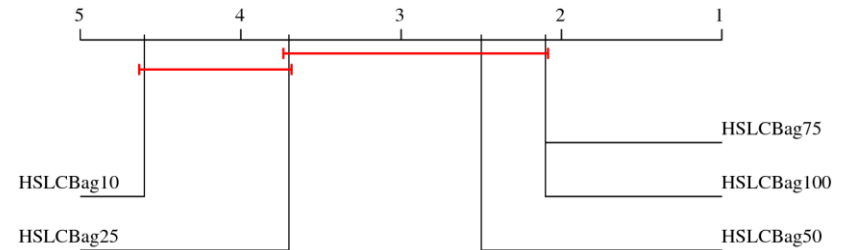


Results – HMLC

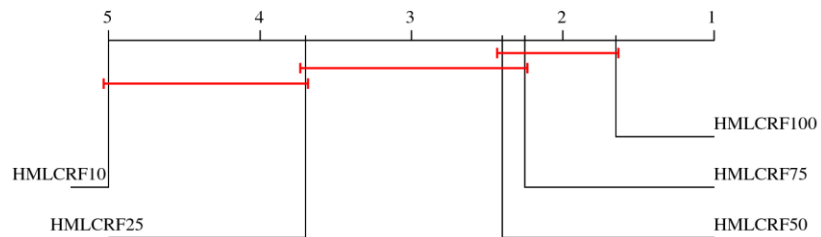
HMLC Bagging



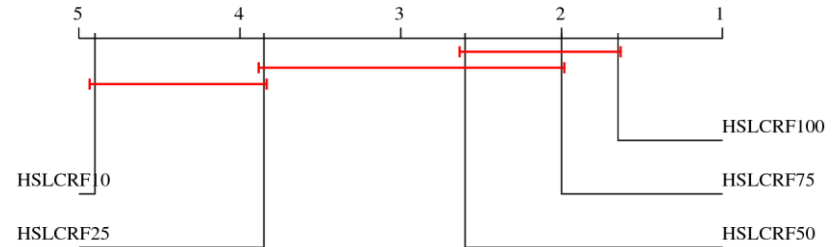
HSLC Bagging



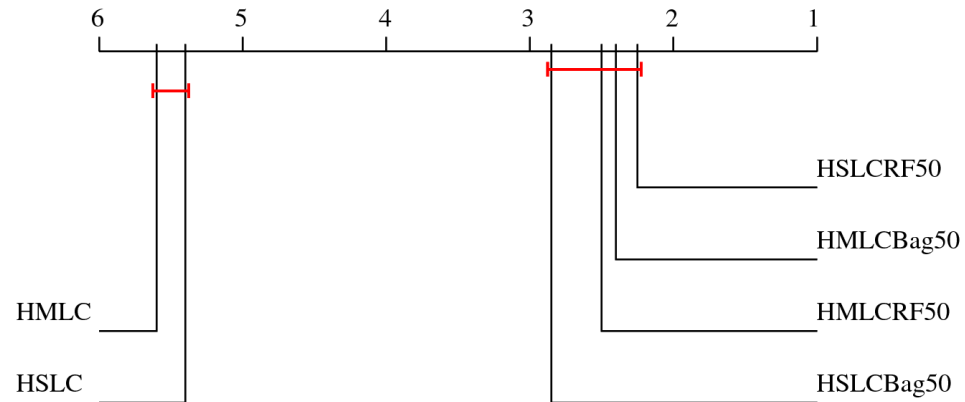
HMLC Random forest



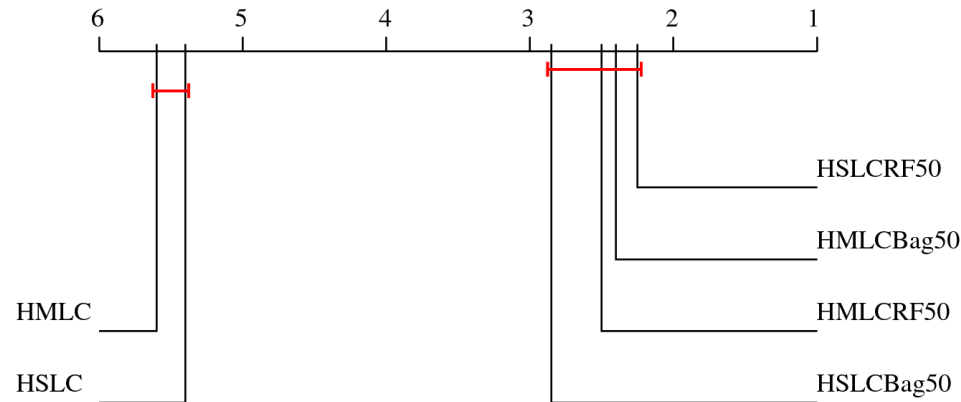
HSLC Random forest



Results – HMLC



Results – HMLC



- HMLC Ensembles perform better than
 - single HMLC tree (stat. sign.)
 - faster to learn than HSLC ensembles (stat. sign.)
 - smaller models than HSLC ensembles (stat. sign.)
- HMLC Bagging is better than HSLC Bagging, while HMLC random forest is worse than HSLC random forest

Results - Summary

Ensembles of PCTs:

- Perform significantly better than single PCT
- Perform better than ensembles for the sub-components
- Smaller and faster to learn than the ensembles for the sub-components
 - For multiple targets the ratio is ~ 3
 - For HMLC the ratio is ~ 4.5

Outline

- Background
 - Structured outputs
 - Predictive Clustering Trees (PCTs)
 - PCTs for HMLC
- Ensembles of PCTs
- Experimental evaluation
- **Application in functional genomics**
- Conclusions

Ensembles of PCTs - functional genomics

- Automatic prediction of the multiple functions of the ORFs in a genome
- Bagging of PCTs for HMLC compared with state-of-the-art approaches
- Datasets for three organisms
 - *S. Cerevisiae*, *A. Thaliana* and *M. musculus*
- Two annotation schemes
 - FunCAT and GO (Gene Ontology)

Background – network based approaches

- Usage of known functions of genes nearby in a functional association network
- **GeneFAS** (Chen and Xu, 2004)
- **GeneMANIA** (Mostafavi et al., 2008) – combines multiple networks from genomic and proteomic data
- **KIM** (Kim et al., 2008) – combines predictions of a network with predictions from Naïve Bayes
- **Funckenstein** (Tian et al., 2008) – logistic regression to combine predictions from network and random forest

Background – kernel based approaches

- **KLR** (Lee et al., 2006) – combination of Markov random fields and SVMs with diffusion kernels and using them in kernel logistic regression
- **CSVM** (Obozinski et al., 2008) – SVM for each function separately, and then reconciliated to enforce the hierarchy constraint
- **BSVM** (Barutcuoglu et al., 2006) – SVM for each function separately, then predictions combined using a Bayesian network
- **BSVM+** (Guan et al., 2008) – extension of BSVM, uses Naïve Bayes to combine results over the data sources

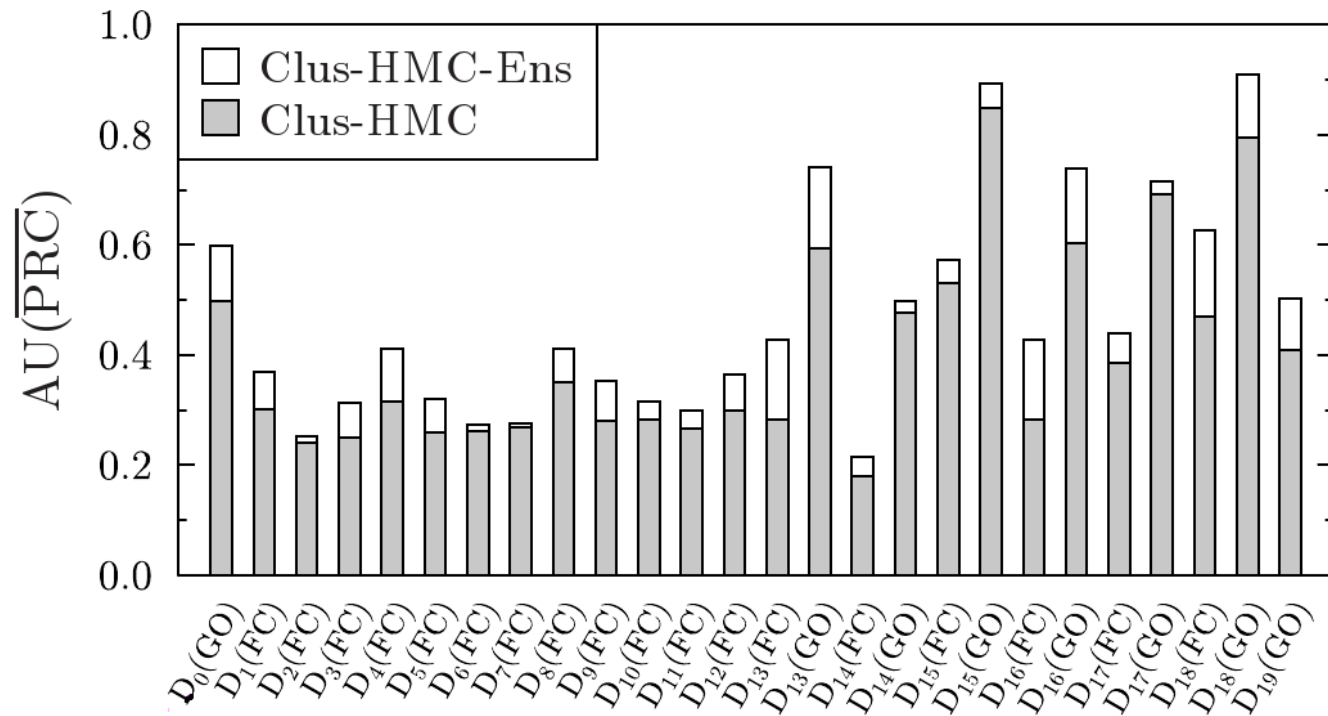
Datasets

- *Saccharomyces Cerevisiae* (D_0 - D_{12})
 - sequence statistics, phenotype, secondary structure, homology and expression
- *Arabidopsis Thaliana* (D_{13} - D_{18})
 - sequence statistics, expression, predicted SCOP class, predicted secondary structure, InterPro and homology
 - Each dataset annotated with FunCAT and GO
- *Mus Musculus* (D_{19})
 - MouseFunc challenge, various data sources

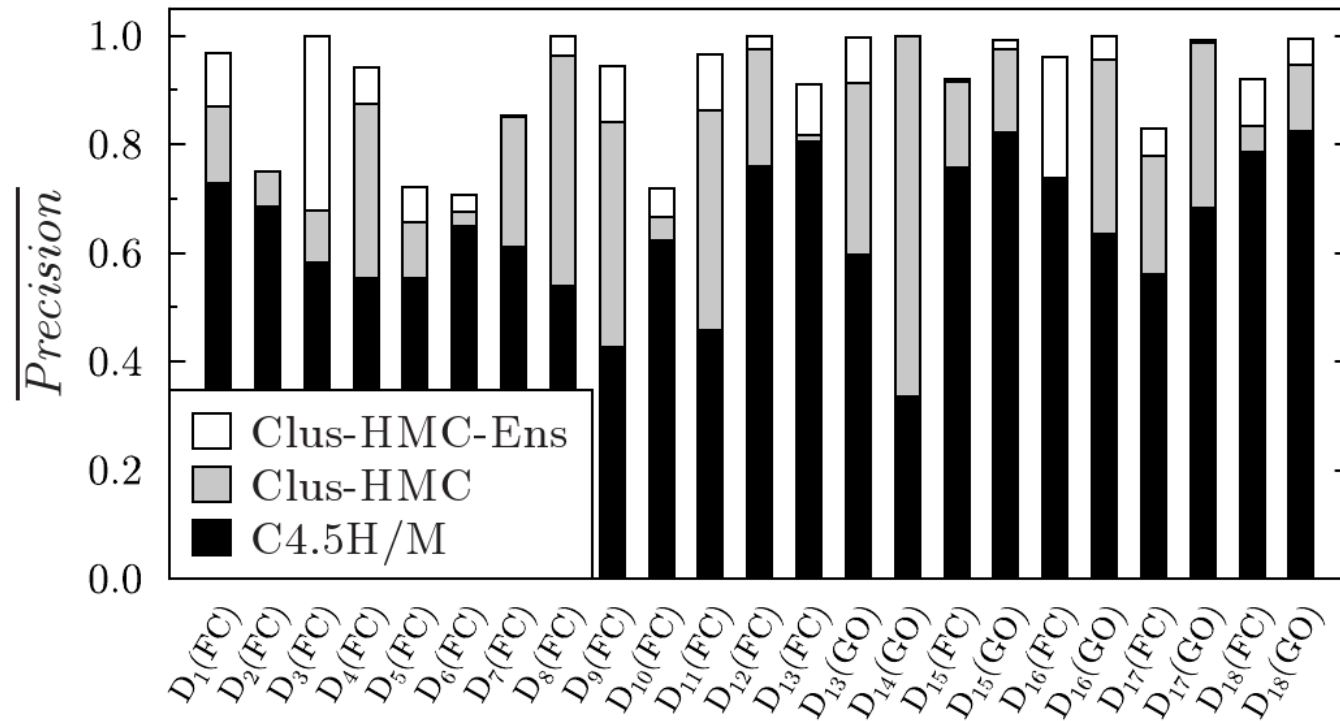
Experimental hypotheses

- Whether Bagging of PCTs are better than single PCT for HMLC and C4.5/H? (D_1 - D_{18})
- Compare Bagging of PCTs with BSVM on D_0
- Compare Bagging of PCTs with the approaches from MouseFunc Challenge (D_{19})

Results – Bagging of PCTs vs. PCTs



Results – Bagging vs. PCTs vs. C4.5H



Results – Bagging of PCTs vs. BSVM

- Comparison by AUROC
- Average over the AUROC for all functions
 - Bagging of PCTs: 0.871
 - BSVM: 0.854
- Bagging of PCTs scores better on 73 of the 105 GO functions than BSVM ($p = 4.37 \cdot 10^{-5}$)

Results – MouseFunc

- Comparison by AUC , \overline{APRC} , \overline{AUROC}
- Different teams use different features
- Division of the dataset
 - Three branches of GO (BP, MF and CC)
 - Four ranges of specificity: number of genes by which each term is annotated with (3-10, 11-30, 31-100 and 101-300)

Results – MouseFunc (2)

$AUROC$

| Subset | CLUS-HMC-Ens | BSVM ⁺ | KLR | CSVM | GENEFAS | GENEMANIA | KIM | Funckenstein |
|------------|--------------|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| BP_3-10 | 0.045 | 0.040 \ominus | 0.028 \ominus | 0.029 \ominus | 0.028 \ominus | 0.071 \oplus | 0.029 \ominus | 0.085 \oplus |
| BP_11-30 | 0.055 | 0.042 \ominus | 0.053 \ominus | 0.017 \ominus | 0.012 \ominus | 0.038 \ominus | 0.031 \ominus | 0.083 \oplus |
| BP_31-100 | 0.109 | 0.100 \ominus | 0.135 \oplus | 0.077 \ominus | 0.033 \ominus | 0.035 \ominus | 0.044 \ominus | 0.190 \oplus |
| BP_101-300 | 0.173 | 0.161 \ominus | 0.174 \oplus | 0.146 \ominus | 0.078 \ominus | 0.055 \ominus | 0.051 \ominus | 0.225 \oplus |
| CC_3-10 | 0.182 | 0.076 \ominus | 0.060 \ominus | 0.046 \ominus | 0.050 \ominus | 0.131 \ominus | 0.128 \ominus | 0.202 \oplus |
| CC_11-30 | 0.207 | 0.085 \ominus | 0.128 \ominus | 0.094 \ominus | 0.038 \ominus | 0.068 \ominus | 0.112 \ominus | 0.167 \ominus |
| CC_31-100 | 0.233 | 0.163 \ominus | 0.161 \ominus | 0.074 \ominus | 0.107 \ominus | 0.046 \ominus | 0.127 \ominus | 0.226 \ominus |
| CC_101-300 | 0.220 | 0.166 \ominus | 0.225 \oplus | 0.157 \ominus | 0.110 \ominus | 0.101 \ominus | 0.094 \ominus | 0.248 \oplus |
| MF_3-10 | 0.266 | 0.243 \ominus | 0.191 \ominus | 0.205 \ominus | 0.174 \ominus | 0.359 \oplus | 0.189 \ominus | 0.368 \oplus |
| MF_11-30 | 0.356 | 0.258 \ominus | 0.285 \ominus | 0.275 \ominus | 0.136 \ominus | 0.270 \ominus | 0.215 \ominus | 0.384 \oplus |
| MF_31-100 | 0.360 | 0.245 \ominus | 0.294 \ominus | 0.231 \ominus | 0.120 \ominus | 0.284 \ominus | 0.191 \ominus | 0.482 \oplus |
| MF_101-300 | 0.368 | 0.283 \ominus | 0.331 \ominus | 0.386 \oplus | 0.184 \ominus | 0.202 \ominus | 0.140 \ominus | 0.485 \oplus |

■ Bagging of PCTs is

- significantly better ($p < 0.01$) than BSVM+, CSVM, GeneFAS and KIM
- better than KLR and GeneMANIA
- Significantly worse ($p < 0.01$) than Funckenstein

Results – MouseFunc (3)

\overline{AUPRC}

| Subset | CLUS-HMC-ENS | BSVM ⁺ | KLR | CSVM | GENEFAS | GENEMANIA | KIM | Funckenstein |
|------------|--------------|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| BP_3-10 | 0.120 | 0.156 \oplus | 0.075 \ominus | 0.075 \ominus | 0.108 \ominus | 0.170 \oplus | 0.108 \ominus | 0.198 \oplus |
| BP_11-30 | 0.110 | 0.141 \oplus | 0.087 \ominus | 0.085 \ominus | 0.074 \ominus | 0.151 \oplus | 0.107 \ominus | 0.162 \oplus |
| BP_31-100 | 0.139 | 0.172 \oplus | 0.158 \oplus | 0.140 \oplus | 0.094 \ominus | 0.177 \oplus | 0.116 \ominus | 0.244 \oplus |
| BP_101-300 | 0.171 | 0.172 \oplus | 0.169 \ominus | 0.173 \oplus | 0.104 \ominus | 0.160 \ominus | 0.056 \ominus | 0.214 \oplus |
| CC_3-10 | 0.319 | 0.249 \ominus | 0.119 \ominus | 0.083 \ominus | 0.233 \ominus | 0.324 \oplus | 0.271 \ominus | 0.316 \ominus |
| CC_11-30 | 0.260 | 0.194 \ominus | 0.212 \ominus | 0.151 \ominus | 0.131 \ominus | 0.235 \ominus | 0.178 \ominus | 0.267 \oplus |
| CC_31-100 | 0.217 | 0.232 \oplus | 0.197 \ominus | 0.161 \ominus | 0.191 \ominus | 0.261 \oplus | 0.144 \ominus | 0.287 \oplus |
| CC_101-300 | 0.244 | 0.217 \ominus | 0.259 \oplus | 0.221 \ominus | 0.177 \ominus | 0.258 \oplus | 0.118 \ominus | 0.279 \oplus |
| MF_3-10 | 0.320 | 0.441 \oplus | 0.258 \ominus | 0.228 \ominus | 0.427 \oplus | 0.465 \oplus | 0.304 \ominus | 0.472 \oplus |
| MF_11-30 | 0.356 | 0.373 \oplus | 0.347 \ominus | 0.393 \oplus | 0.350 \ominus | 0.401 \oplus | 0.302 \ominus | 0.455 \oplus |
| MF_31-100 | 0.269 | 0.289 \oplus | 0.230 \ominus | 0.278 \oplus | 0.242 \ominus | 0.291 \oplus | 0.255 \ominus | 0.416 \oplus |
| MF_101-300 | 0.322 | 0.317 \ominus | 0.321 \ominus | 0.374 \oplus | 0.295 \ominus | 0.391 \oplus | 0.172 \ominus | 0.441 \oplus |

- Bagging of PCTs is
 - significantly better ($p < 0.01$) than KIM
 - not different from BSVM⁺, KLR, CSVM, GeneFAS
 - Significantly worse ($p < 0.01$) than Funckenstein and GeneMANIA

Results – MouseFunc (4)

\overline{AUROC}

| Subset | CLUS-HMC-Ens | BSVM ⁺ | KLR | CSVM | GENEFAS | GENEMANIA | KIM | Funckenstein |
|------------|--------------|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| BP_3-10 | 0.695 | 0.808 \oplus | 0.581 \ominus | 0.588 \ominus | 0.715 \oplus | 0.873 \oplus | 0.813 \oplus | 0.790 \oplus |
| BP_11-30 | 0.748 | 0.808 \oplus | 0.741 \ominus | 0.659 \ominus | 0.767 \oplus | 0.849 \oplus | 0.822 \oplus | 0.796 \oplus |
| BP_31-100 | 0.831 | 0.874 \oplus | 0.846 \oplus | 0.778 \ominus | 0.780 \ominus | 0.872 \oplus | 0.851 \oplus | 0.880 \oplus |
| BP_101-300 | 0.823 | 0.853 \oplus | 0.845 \oplus | 0.813 \ominus | 0.733 \ominus | 0.840 \oplus | 0.795 \ominus | 0.838 \oplus |
| CC_3-10 | 0.748 | 0.845 \oplus | 0.571 \ominus | 0.618 \ominus | 0.782 \oplus | 0.899 \oplus | 0.865 \oplus | 0.837 \oplus |
| CC_11-30 | 0.791 | 0.873 \oplus | 0.790 \ominus | 0.785 \ominus | 0.834 \oplus | 0.907 \oplus | 0.846 \oplus | 0.850 \oplus |
| CC_31-100 | 0.863 | 0.896 \oplus | 0.850 \ominus | 0.851 \ominus | 0.783 \ominus | 0.887 \oplus | 0.863 | 0.849 \ominus |
| CC_101-300 | 0.845 | 0.873 \oplus | 0.851 \oplus | 0.821 \ominus | 0.750 \ominus | 0.842 \ominus | 0.808 \ominus | 0.867 \oplus |
| MF_3-10 | 0.818 | 0.887 \oplus | 0.630 \ominus | 0.681 \ominus | 0.850 \oplus | 0.951 \oplus | 0.880 \oplus | 0.879 \oplus |
| MF_11-30 | 0.842 | 0.903 \oplus | 0.861 \oplus | 0.836 \ominus | 0.865 \oplus | 0.936 \oplus | 0.884 \oplus | 0.909 \oplus |
| MF_31-100 | 0.838 | 0.888 \oplus | 0.892 \oplus | 0.881 \oplus | 0.843 \oplus | 0.887 \oplus | 0.884 \oplus | 0.903 \oplus |
| MF_101-300 | 0.874 | 0.904 \oplus | 0.894 \oplus | 0.884 \oplus | 0.843 \ominus | 0.909 \oplus | 0.844 \ominus | 0.918 \oplus |

- Bagging of PCTs is
 - not different from KLR, CSVM, GeneFAS and KIM
 - Significantly worse ($p < 0.01$) than Funckenstein, BSVM+ and GeneMANIA

Summary – Functional genomics

- Bagging of PCTs outperforms single PCT and C4.5H/M (Clare, 2003)
- Bagging of PCTs outperforms a statistical learner based on SVMs (BSVM) for *S. Cerevisiae*
- Bagging of PCTs is competitive to statistical and network based methods for the *M. Musculus* data

Outline

- Background
 - Structured outputs
 - Predictive Clustering Trees (PCTs)
 - PCTs for HMLC
- Ensembles of PCTs
- Experimental evaluation
- Application in functional genomics
- **Conclusions**

Conclusions

- Lift the predictive performance of single PCT
- Better than ensembles for each of the sub-component of the output
- Competitive with state-of-the-art approaches in functional genomics
- Applicability to wide range of problems
 - Different type and sizes of outputs
 - Small and large datasets



Questions?