Ensembles for predicting structured outputs

Dragi Kocev
outline

- Predictive Clustering Trees (PCTs)
- Bagging and random forests for PCTs
- Beam-search induction of trees
- Applications
- Summary
Motivation

- Increasing amounts of structured data
  - Vectors
  - Hierarchies – trees, DAGs, ...
  - Sequences – time series

- Success of the ensemble methods in simple classification and regression
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 targets

- Multiple target regression
  - Euclidean distance

\[ \text{Var}(E) = \sum \text{Var}(E, y_t) \]

- Multiple target classification

\[ \text{Var}(E) = \sum \text{Gini}(E, y_t) \]
\[ \text{Var}(E) = \sum \text{Entropy}(E, y_t) \]
PCTs - HMLC
PCTs - HMLC

1 (1)  2 (2)  3 (5)

2.1 (3)  2.2 (4)
PCTs - HMLC

\[
\nu_i = [1, 1, 0, 1, 0]
\]
PCTs - HMLC

\[ v_i = [1, 1, 0, 1, 0] \]

\[ v_i = [1, 1, 1, 0, 0] \]
Hierarchical Multi-Label Classification

\[
V_{ar}(E) = \frac{\sum d(l_i, \hat{l})^2}{|E|}
\]
\[
d(l_i, \hat{l}) = \sqrt{\sum \omega(c_i) \cdot (l_{1,i} - l_{2,i})^2}
\]
Ensemble Methods

- Set of predictive models
  - Voting schemes to combine the predictions into a single prediction

- Unstable base classifiers

- Ensemble learning
  - Modification of the data
  - Modification of the algorithm

- Bagging

- Random forests
Ensembles for structured outputs

- PCTs as base classifiers
- Voting schemes for the structured targets
  - MT Classification: majority and probability distribution vote
  - MT Regression and HMLC: average
  - For an arbitrary structure: prototype calculation function
- Predictive performance
  - Classification: accuracy
  - Regression: correlation coefficient, RMSE, RRMSE
  - HMLC: Precision-Recall curve (PRC), Area under PRCs
Experimental design

- **Datasets**

<table>
<thead>
<tr>
<th></th>
<th>Datasets</th>
<th>Examples</th>
<th>Descriptive attributes</th>
<th>Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT Regression</td>
<td>14</td>
<td>154..60607</td>
<td>4..160</td>
<td>2..14</td>
</tr>
<tr>
<td>MT Classification</td>
<td>11</td>
<td>154..10368</td>
<td>4..294</td>
<td>2..14</td>
</tr>
</tbody>
</table>

- F-test pruning for the single trees
  - Internal 3-fold cross validation

- **Number of bags**
  - 10, 25, 50, 75, 100

- **Random Forest**
  - Feature subset size: logarithmic wrt attributes

- 10-fold cross validation
Experimental hypotheses

- Saturation curves for bagging and random forests
  - Number of bags

- Comparison of the ensembles from PCTs to
  - PCTs for each component separately
  - ensembles for each component separately

- Friedman and Nemenyi tests for statistical significance
Results - Regression

- Relative RMSE
- MT Bagging vs. ST Bagging
Results - Regression

- Relative RMSE
- MT Bagging vs. ST Bagging
  - MTBag10
  - MTBag25
  - MTBag75
  - MTBag50

- MT Random forest vs. ST Random forest
  - MTRF10
  - MTRF25
  - MTRF75
  - MTRF50
  - STRF10
  - STRF25
  - STRF75
  - STRF50
Results - Regression

- Relative RMSE @ 75 bags
Results - Classification

- Probability distribution voting
- MT Bagging vs. ST Bagging
Results - Classification

- Probability distribution voting
- MT Bagging vs. ST Bagging
- MT Random forest vs. ST Random forest
Results - Classification

- Probability distribution voting @ 50 bags
Results - Summary

Ensembles for multiple targets:

- Converge faster
- Perform significantly better than single PCT
- Perform better than ensembles for single target
- Smaller and faster to learn
  - Size and time ratio ~ 2.5-3.0
  - More emphasized in bigger datasets
Ensembles for HMLC

- **Datasets**
  - 3 from image classification
  - 3 from text classification
  - 3 from functional genomics

- **Preliminary results show that ensembles for HMLC are:**
  - Better than single PCT for HMLC
  - Better than learning an ensemble for each label separately
  - Significant speed up (~4.5-5.0) wrt learning ensemble for each label separately
Feature Ranking for structured outputs

- Estimating variable importance using random forest
- Uses out-of-bag error estimate and random permutations of the features
- The rationale is: if a feature is important for the target concept(s) then the error rate should increase when its values are randomly permuted
- Obtain feature ranking for
  - Multiple targets: avoid aggregation of ranks
  - Hierarchies (both trees and DAGs)
Outline

- Predictive Clustering Trees
- Bagging and random forests from predictive clustering trees
- Beam-search induction of trees
- Applications
- Summary
Beam Search Algorithm

- Output: k models

- Tree induction perspective
  - Take the tree with best score and the rest as good alternatives (domain knowledge)
  - Addressed the myopia of the standard TDIDT

- Ensemble learning perspective
  - Combine the trees in ensemble and let them vote
Beam-search algorithm

```
  yes
  /\  \
 /   \
no
```
Beam-search algorithm

\[ A < a_0 \]

yes \quad \text{no}
Beam-search algorithm

\[ A < a_0 \]

yes  no

2.1  1.3
Beam-search algorithm

\[ A < a_0 \]

2.1

- yes
- no

- yes
- no
Beam-search algorithm

2.1

\[ A < a_0 \]

\[
\begin{array}{c}
\text{yes} \\
\text{no}
\end{array}
\]

\[ B \text{ in } \{b_0, b_1\} \]

\[
\begin{array}{c}
\text{yes} \\
\text{no}
\end{array}
\]
Beam-search algorithm

\[ A < a_0 \]

yes

2.1

no

no

3.4

B in \( \{b_0, b_1\} \)

yes

0.3

no
Beam-Search

\[ A < a_0 \]

- yes: 2.1
- no: 1.3

\[ A < a_0 \]

- yes: 2.1
- no: \( B \in \{b_0, b_1\} \)
  - yes: 3.4
  - no: 0.3
Stopping criteria: beam no longer changes or user constraints
Beam-Search Heuristic score

\[ h(T, I) = \left( \sum_{\text{leaf } \in T} \frac{|I_{\text{leaf}}|}{|I|} \text{Var}(I_{\text{leaf}}) \right) + \alpha \cdot \text{size}(T) \]
Beam-Search Heuristic score

\[
h(T, I) = \left( \sum_{\text{leaf } \in T} \frac{|I_{\text{leaf}}|}{|I|} \, \text{Var}(I_{\text{leaf}}) \right) + \alpha \cdot \text{size}(T)
\]

Performance

Soft size constraint
Beam-Search Algorithm Summary

- Easy to push user constraints
  - Hard size constraint
- Competitive results as compared to TDIDT
  - Beam-width is set to 10

- Problem: the trees in the beam are quite similar to each other
- Solution: similarity constraints
Similarity constraints: take one

- Enforce diversity in the beam
- First experiments:
  - Change the heuristic score
Similarity constraints: take one

- Enforce diversity in the beam
- First experiments:
  - Change the heuristic score

\[
h(T, I) = \left( \sum_{\text{leaf } \in T} \frac{|I_{\text{leaf}}|}{|I|} \text{Var}(I_{\text{leaf}}) \right) + \alpha \cdot \text{size}(T)
\]

\[
h_s(T, \text{beam}, I) = \left( \sum_{\text{leaf } \in T} \frac{|I_{\text{leaf}}|}{|I|} \text{Var}(I_{\text{leaf}}) \right) + \alpha \cdot \text{size}(T) + \beta \cdot \text{sim}(T, \text{beam}, I)
\]
Similarity constraints: take one

- Enforce diversity in the beam
- First experiments:
  - Change the heuristic score

\[ h(T, I) = \left( \sum_{\text{leaf } \in T} \frac{|I_{\text{leaf}}|}{|I|} \text{Var}(I_{\text{leaf}}) \right) + \alpha \cdot \text{size}(T) \]

\[ h_s(T, \text{beam}, I) = \left( \sum_{\text{leaf } \in T} \frac{|I_{\text{leaf}}|}{|I|} \text{Var}(I_{\text{leaf}}) \right) + \alpha \cdot \text{size}(T) + \beta \cdot \text{sim}(T, \text{beam}, I) \]

\[ \text{sim}(T, \text{beam}, I) = 1 - \frac{d(T, T_{\text{cand}}, I) + \sum_{T_i \in \text{beam}} d(T, T_i, I)}{|\text{beam}|} \]

\[ d(T_1, T_2, I) = \frac{1}{\eta} \cdot \sqrt{\frac{\sum_{t \in I} d_p(p(T_1, t), p(T_2, t))^2}{|I|}}, \]
Similarity constraints: take one

- The trees in the beam are more different to each other
- Better results for regression tasks
  - Problem with the classification tasks is the hit/miss distance that we used
Similarity constraints: take two

- Include the similarity in the test selection procedure
- For classification use distance over the probability distributions
Similarity constraints: take two

- Include the similarity in the test selection procedure
- For classification use distance over the probability distributions

\[ Heuristic(T, beam, I) = \sum_{\text{leaf } \in T} \sum_{(x,y) \in I} d^2(y, \mu_l) - \beta \cdot \frac{1}{k} \sum_{\text{leaf } \in T} \sum_{(x,y) \in I} \sum_{i=1}^{k} d^2(\mu_l, T_i(x)) \]
Similarity constraints: take two

- Include the similarity in the test selection procedure
- For classification use distance over the probability distributions

\[
\text{Heuristic}(T, \text{beam}, I) = \sum_{\text{leaf} \in T} \sum_{(x, y) \in I_t} d^2(y, \mu_l) + \beta \cdot \frac{1}{k} \sum_{\text{leaf} \in T} \sum_{(x, y) \in I_t} \sum_{i=1}^{k} d^2(\mu_l, T_i(x))
\]

Performance

Similarity to the other trees
Beam Search Algorithm - Summary

- Tree induction point of view
  - More than one tree as an answer
  - Competitive with TDIDT
- Ensembles point of view
  - Direct control of the ensemble diversity
  - “Interpretable” ensembles
- Experiments yet to be performed
outline

- Predictive Clustering Trees
- Bagging and random forests from predictive clustering trees
- Beam-search induction of trees
- Applications
- Summary
Case Studies

- Indigenous Vegetation
- Functional Genomics
- Image Classification
Indigenous Vegetation

- 16967 sites in Victoria State, Australia
- Each sample is described with:
  - 40 variables: GIS and remote-sensed data
  - Habitat Hectares Score: Large Trees, Tree Canopy Cover, Understorey, Lack of Weeds, Recruitment, Logs, and Organic Litter
Functional Genomics

- Predicting gene functions of *S. cerevisiae*, *A. thaliana* and *M. Musculus*
- Two annotation schemes: FunCat and Gene Ontology
- Ensembles for HMLC are competitive with other algorithms
Image Classification

- Image CLEF 2008 data
- IRMA coding system with four axes
  - Anatomical, Biological, Directional and Technical
- 12000 annotated X-Ray images
- 1000 not-annotated X-Ray images

![Graphs showing precision-recall curves for CLEF 2008 - Axis 2 and Axis 3](image)

- HMC-DT (AUC = 0.6975)
- HMC-RF (AUC = 0.9030)
- HMC-Bagging (AUC = 0.9064)
- HMC-DT (AUC = 0.5371)
- HMC-RF (AUC = 0.8246)
- HMC-Bagging (AUC = 0.8257)
Summary

- Ensemble methods for predicting structured outputs
  - Exploitation of the structure of the output
  - Bagging and random forest
  - Produce ranking for structured outputs
- Beam-search induction of trees
  - Output multiple possible answers
  - Easy to push user constraints
- Beam-search induction of ensembles
  - Control the diversity in the ensemble
  - “Interpretable” ensembles
- Methods scalable to other types of structured outputs
- Applications in different domains
Publication statistics (2008/09)

Published SCI journal papers:


Conference/workshop papers


Publication statistics (2008/09)
Publication statistics (2008/09)

Drafts of journal papers:


- Andreja Naumoski, Dragi Kocev, Nataša Atanasova, Kosta Mitreski, Svetislav Krstić, Sašo Džeroski, "*Modelling the Relationship between Diatom Abundances and Physico-chemical Parameters in Lake Prespa*", Ecological Informatics …


- Jérôme Cortet, Dragi Kocev, Christophe Schwartz, Caroline Ducobu, Sašo Džeroski, Marko Debeljak, "*Modelling agronomic and environmental soil properties following wastes application in arable crops: results of 10 years management in the field*", Soil journal….
Questions?