TECHNISCHE
UNIVERSITÄT
DARMSTADT

Towards Multilabel Rule Learning

Eneldo Loza Mencía, Frederik Janssen
Knowledge Engineering Group, TU Darmstadt
{eneldo,janssen}@ke.tu-darmstadt.de

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Multilabel and Rule Learning

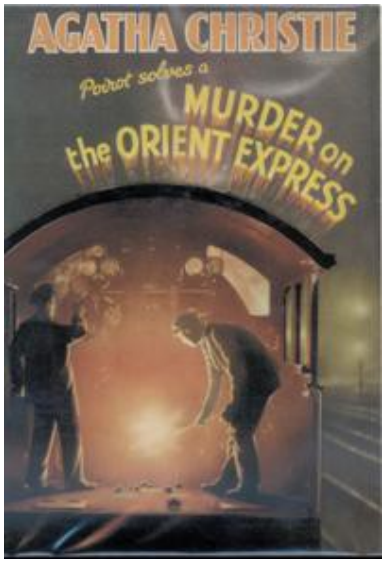
Multilabel Rules

Stacked Bootstrapping

Experiments

Conclusions and Outlook

Multilabel setting



Summary: Returning from an important case in Syria, Hercule Poirot boards the Orient Express in Istanbul. The train is unusually crowded for the time of year. Poirot secures a berth only with ...

Author: Agatha Christie

Genres:

Crime, Mystery, Thriller

- assignment of an object x to a **subset** of a set of label Y
- in contrast to
  - multiclass classification: mapping to exactly one class
  - binary classification: mapping to one of only two classes

Typical application areas

- text: tagging/indexing of news, web pages, blogs, … with keywords, topics, genres, authors, languages, writing styles, …
- multimedia: detection of scenes/object (images), instruments, emotions, music styles (audio)
- biology: classification of functions of genomes and protein

Rule Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

A rule consists of

- body: tests on attribute-value pairs (conditions)
- head: the consequence, sets a class attribute to a certain value

Common approach: Separate and Conquer

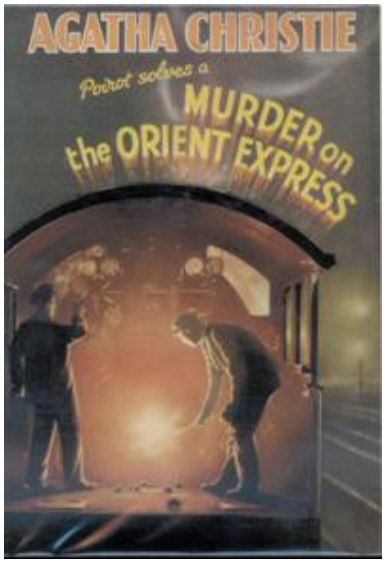
1. find a good rule on training set (conquer)
2. remove training instances covered by rule (separate)
3. repeat 1-2 until (positive) instances covered

Advantages:

- interpretability and comprehensibility
- competitive prediction performance
- natural support of *missing/unknown* states of attributes

Motivation and Goals: Multilabel Rule Learning

TECHNISCHE UNIVERSITÄT DARMSTADT



Summary: Returning from an important case in Syria, Hercule Poirot boards the Orient Express in Istanbul. The train is unusually crowded for the time of year. Poirot secures a berth only with ...

Author: Agatha Christie

Genres:

Crime, Mystery, Thriller

Multilabel: **dependencies between labels!**

- conditional and unconditional
- better comprehension of implication between labels
  - direct representation by rules
  - in contrast to e.g. statistical approaches

1. Traditional Rule: Single label in the head

author="Christie" → **Crime**

2. Multilabel Rule: multiple labels in the head

author="Christie" & text."Poirot"=1 → **Crime**, **Mystery**, **Thriller**

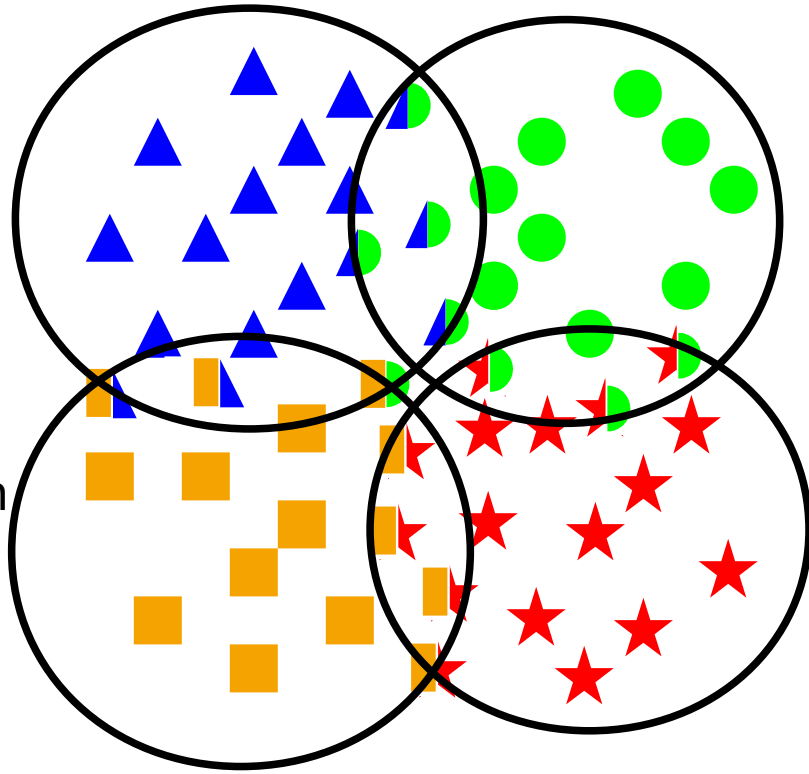
3. "Recursive" Multilabel Rule: Labels also appear in the body

Thriller & text."murder"=1 → **Crime**

Case 1: Attributes → Label

TECHNISCHE UNIVERSITÄT DARMSTADT

- learn one rule set per label
  - referred to as *binary relevance* decomposition (BR)
    - aka one-against-all in multiclass
  - corresponds to concept learning
    - learn each label as separate concept learning problem
  - learns a description of a label based on features describing the instances
- not considering label dependencies
  - each rule set is learned separately
  - corresponds to learning marginal class probabilities $P(y_i \mid x)$



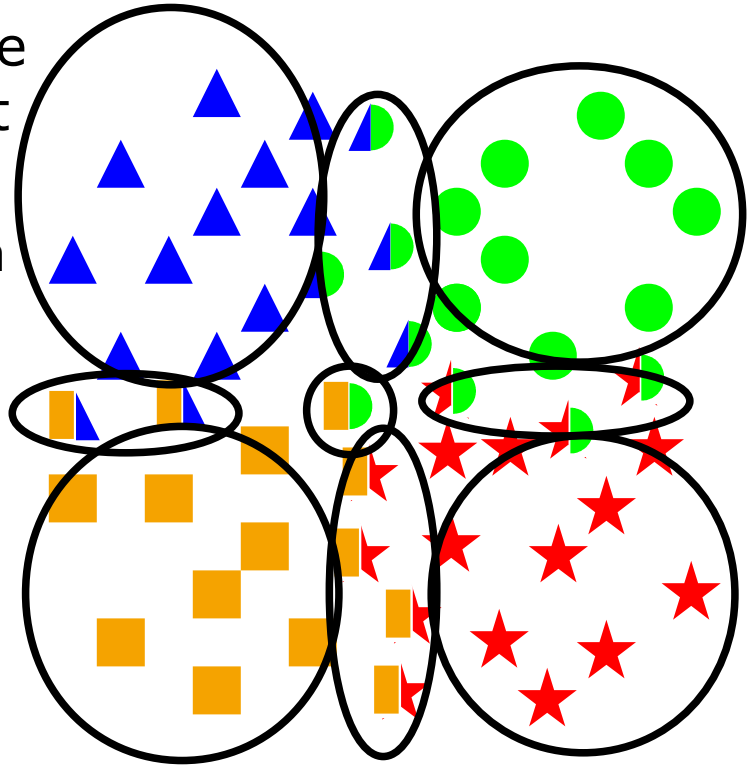
box → orange
triangle → blue
...

Case 2: Attributes → Labelset

TECHNISCHE UNIVERSITÄT DARMSTADT

Straight-forward approach: create one meta-class for each occurring labelset

- Label Powerset (LP) approach
- train a multiclass learner, i.e. learn each label*set* independently
- corresponds to learning the joint class probabilities $P(y_1,\ldots,y_n \mid x)$



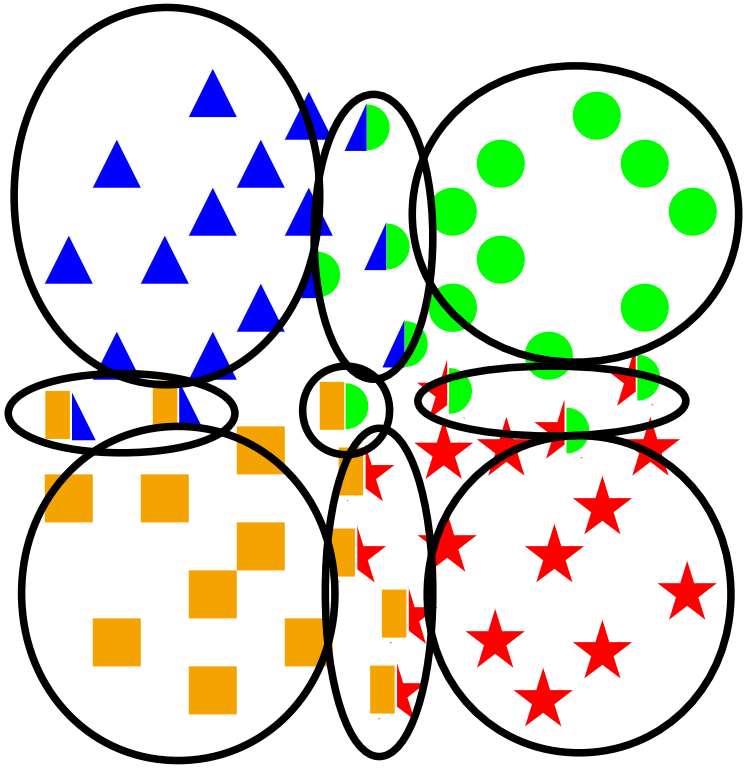
box,~~triangle~~ → orange
box,triangle → orange,blue
triangle,~~box~~ → blue
...

Case 2: Attributes → Labelset

TECHNISCHE UNIVERSITÄT DARMSTADT

Disadvantages

- learn co-occurrences, but no explicit interdependencies ("implications")
- computationally expensive: possible labelsets grow exponentially
- prediction of unseen label combinations in training data impossible
  - no discovery of new relationships
  - we may miss correct labelsets in unseen data



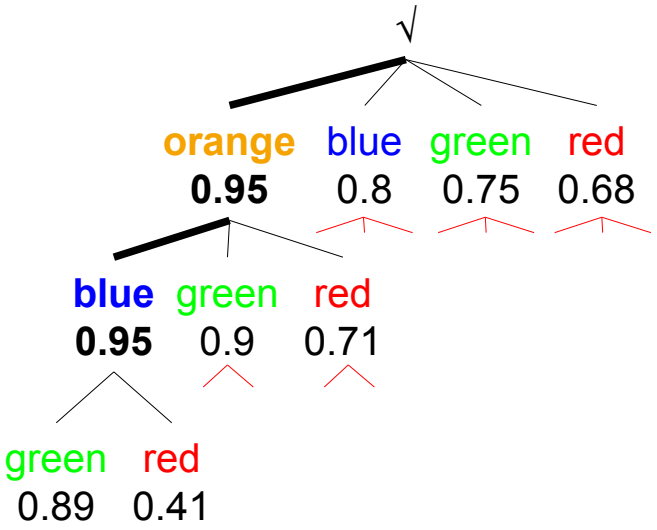
box,~~triangle~~ → orange
box,triangle → orange,blue
triangle,~~box~~ → blue
...

Case 2: Attributes → Labelset

TECHNISCHE UNIVERSITÄT DARMSTADT

Our proposal: adapt Separate and Conquer Algorithm

- find for each rule candidate the best possible head
  - instead for each head the best rule candidate
- much more efficient:
  - consider label combinations as a tree
  - branches (labelsets) can be safely pruned away if used heuristic measure decreases
    - anti-monotonicity of positive coverage
- prediction of unseen label combinations possible

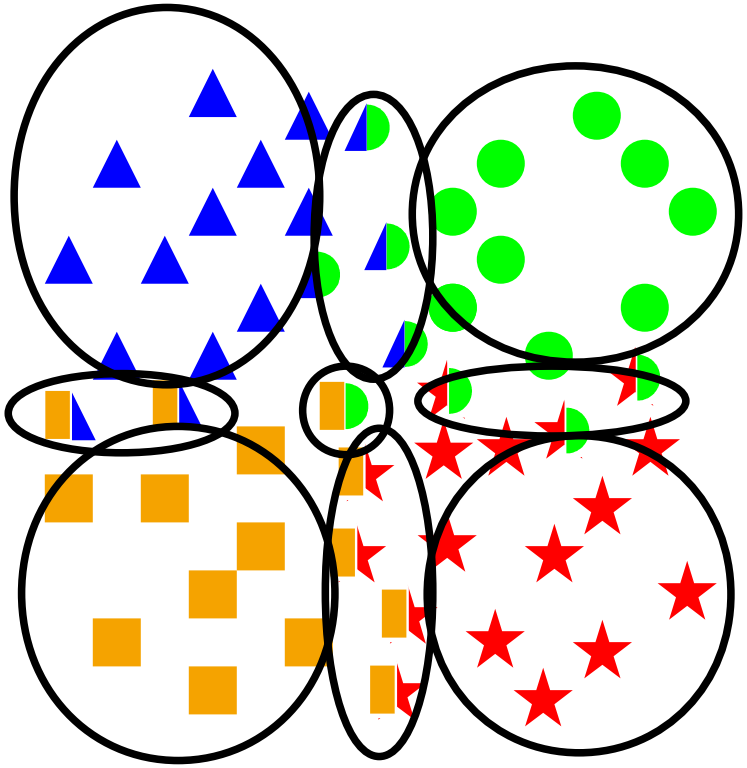


Case 2: Attributes → Labelset

TECHNISCHE UNIVERSITÄT DARMSTADT

Open Questions to investigate

- produces mostly single label heads?
- allow negative label heads (i.e. ~~blue~~)?
  - contrary to notion of concept learning
- different heuristics necessary?
  - current heuristics are developed to work well for finding single-label rules
- effective in predicting unseen labelsets?

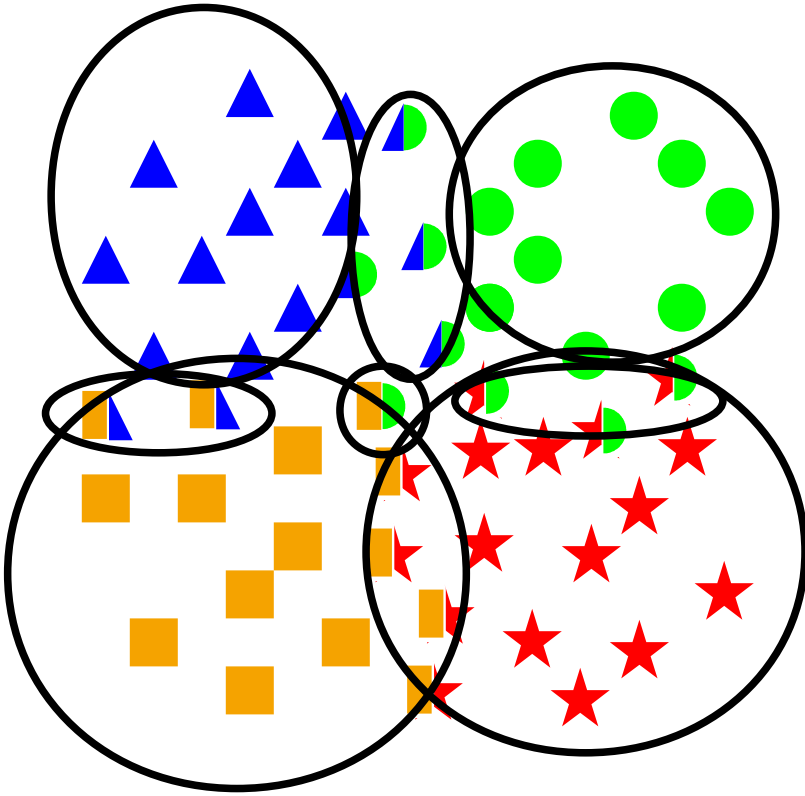


Case 3: Attributes, Labels → Labels

TECHNISCHE UNIVERSITÄT DARMSTADT

Learn dependencies between labels

- create explicit knowledge about label relationships
  - direct, symbolic representation of dependencies ($x, y_1 \rightarrow y_2$)
  - in contrast to LP, not only model co-occurences ($x \rightarrow y_1, y_2$)
    - note: can obtain $x \rightarrow y_1, y_2$ from concatenation of recursive rules
  - can also model unconditional dependencies (e.g. $y_1 \rightarrow y_2$)



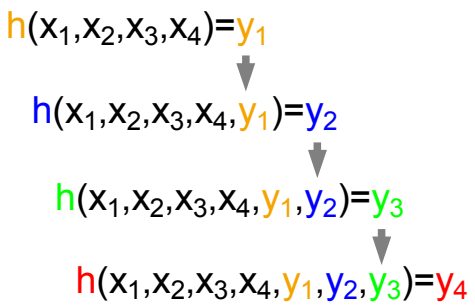
box → orange
triangle,orange → +blue
...

Case 3: Attributes, Labels → Labels

TECHNISCHE UNIVERSITÄT DARMSTADT

Starting point: Classifier Chains

- very popular recent multilabel approach from Read et al. (2009)
- stacking predictions of previous binary single-label (BR) classifiers
- explicitly model label dependencies
- but: fixed ordering, **learns dependencies only in one direction**
- corresponds to learning conditional label probabilities $P(y_i \mid y_1 \ldots y_{i-1}, x)$
  - only dependencies $y_1 \ldots y_{i-1} \rightarrow y_i$



Our First Naive Approach: Stacked Bootstrapping

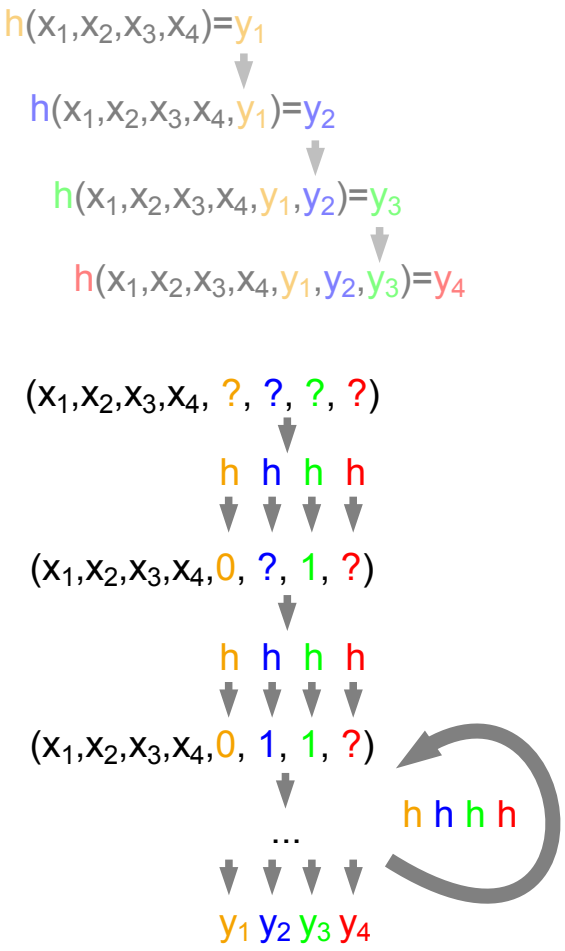
TECHNISCHE UNIVERSITÄT DARMSTADT

Training:

- learn binary single-label classifiers, but include labels as additional features

Prediction:

- fill up label features of test instances with:
  - "?" ("unknown", "missing"…), or
  - prediction of conventional BR classifiers, or ...
- use single-label classifiers to predict labels
- repeat prediction until some stopping criterion
- corresponds to learning conditional class probabilities $P(y_i \mid y_1 \ldots y_{i-1}, y_{i+1} \ldots y_n, x)$



Stacked Bootstrapping

TECHNISCHE UNIVERSITÄT DARMSTADT

Advantages from using rule learners

- direct, symbolic representation of dependencies (Case 3 rules)
- rule learners naturally support to abstain from predicting
  - predict "?" if no rule fires (default rule)
- in contrast to e.g. SVMs, which cannot handle this as input/output

Open questions

- if init with "?", only rules with no label attributes in body can fire
  - deadlock if not enough local evidence in data for such rules

Future perspectives

- directly include stacking in SeCo process
- basic idea: don't remove covered instances, but re-include them with set label attributes from the learned head of the covering rule
- no bootstrapping, no iterations, no chaining needed, no deadlocks

Stacked Bootstrapping: Preliminary Experiments

TECHNISCHE UNIVERSITÄT DARMSTADT

Large experimental setting:

- 7 small to mid-sized multilabel datasets
- several symbolic learners and settings
  - C4.5, Ripper, SimpleSeCo-learner
- three basic configurations
  - binary relevance, bootstrapping with ?-init & BR-init

dataset name	domain	#instances m	#attributes a	#labels n	labelset size d	density $\frac{d}{n}$	distinct $\lvert\{P_{\mathbf{x}}\}\rvert$
scene	image	2407	294	6	1.074	17.9 %	15
emotions	music	593	72	6	1.869	31.1 %	27
yeast	biology	2417	103	14	4.237	30.3 %	198
genbase	biology	662	1186	27	1.252	4.6 %	32
medical	text	978	1449	45	1.245	2.8 %	94
enron	text	1702	1001	53	3.378	6.4 %	753
CAL500	music	502	68	174	26.044	0.150 %	502

Stacked Bootstrapping: Preliminary Experiments

TECHNISCHE UNIVERSITÄT DARMSTADT

Preliminary Results

- bootstrapping with BR-init works better than BR
- ?-init clearly worse
  - too many deadlocks

- results (on the same datasets) comparable to

  Montañesa, Senge, Barranquero, Quevedo, del Coz, Hüllermeier:
  Dependent binary relevance models for multi-label classification
  Pattern Recognition, to appear
  (developed in parallel without knowledge of each other)
  - corresponds to BR-init variant with only one bootstrapping iteration
  - but: they use logistic regression as base learner

Preliminary Experiments: Example of learned models

TECHNISCHE UNIVERSITÄT DARMSTADT

approach	yeast	enron
binary relevance (Ripper)	x23 > 0.08, x49 < -0.09 → Class4 x68 < 0.05, x33 > 0.00, x24 > 0.00, x66 > 0.00, x88 > -0.06 → Class4 x3 < -0.03, x71 > 0.03, x91 > -0.01 → Class4 x68 < 0.03, x83 > -0.00, x44> 0.029, x93 < 0.01 → Class4 x96 < -0.03, x10 > 0.01, x78< -0.07 → Class4	"mail", "fw", "didn" → **Joke**
stacked bootstrapping	Class3, ~~Class2~~ → Class4 Class5, ~~Class6~~ → Class4 Class3, Class1, x22 > -0.02 → Class4	**Personal**, "day", "mail" → **Joke**
logistic regression (BR)	x1 -1.4589 x2 -2.2719 x3 1.1241 x4 -1.5108 … 95 more ... x100 13.5979 x101 11.4313 x102 6.7264 x103 31.8507	"0" 1.1635 "00" 0.7809 "000" -2.1408 "01" 0.3419 … 992 more ... "year" -0.6546 "years" 1.0814 "yesterday" 1.0011 "york" 2.229

Conclusions and Outlook

TECHNISCHE UNIVERSITÄT DARMSTADT

Two new perspectives for multilabel rule learning

- Attributes → Labelset
- Attributes, Labels → Labels

Stacked Bootstrapping

- promising first results: good performance, interpretability
- but actually only "proof-of-concept" for direct integration

Open questions:

- also good impact on length of the rules?
- setting changes when applying rule learning on multilabel data:
  - overfitting, pruning, heuristics, "multiclass" decision lists, notion of positive and negative examples?
- combination of both approaches: attributes, labels → labelset

Thanks!

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Questions? Suggestions? Hints?