Classifier Chains for Multi-label Classification

Jesse Read, Bernhard Pfahringer, Geoff Holmes, Eibe Frank

University of Waikato
New Zealand

ECML PKDD 2009, September 9, 2009. Bled, Slovenia

Introduction

Multi-label Classification

- Each instance may be associated with multiple labels
- set of instances $X = \{x_1, \cdots, x_m\}$; set of *predefined* labels $L = \{l_1, \cdots, l_n\}$; dataset $(x_1, S_1), (x_2, S_2), \cdots$ where each $S_i \subseteq L$.
- For example, a film can be labeled $\{\texttt{romance,comedy}\}$

Introduction

Multi-label Classification

- Each instance may be associated with multiple labels
- set of instances $X = \{x_1, \cdots, x_m\}$; set of *predefined* labels $L = \{l_1, \cdots, l_n\}$; dataset $(x_1, S_1), (x_2, S_2), \cdots$ where each $S_i \subseteq L$.
- For example, a film can be labeled $\{\texttt{romance,comedy}\}$

Applications

- Scene, Video classification
- Text classification
- Medical classification
- Biology, Genomics

Introduction

Multi-label Classification

- Each instance may be associated with multiple labels
- set of instances $X = \{x_1, \cdots, x_m\}$; set of *predefined* labels $L = \{l_1, \cdots, l_n\}$; dataset $(x_1, S_1), (x_2, S_2), \cdots$ where each $S_i \subseteq L$.
- For example, a film can be labeled $\{\texttt{romance,comedy}\}$

Applications

- Scene, Video classification
- Text classification
- Medical classification
- Biology, Genomics

Multi-label Issues

- label correlations: consider $\{\texttt{romance,comedy}\}$ vs $\{\texttt{romance,horror}\}$
- computational complexity

Prior Work

- Binary relevance method (BR): binary problem for each label
  - simple, efficient
  - does not take into account label correlations

Prior Work

- Binary relevance method (`BR`): binary problem for each label
  - simple, efficient
  - does not take into account label correlations
- Nearest neighbor approaches based on `BR`, e.g. `MLkNN`
- Stacking approaches, e.g. meta level stacking (`MS`)
- Pairwise approaches, e.g. calibrated label ranking

Prior Work

- Binary relevance method (BR): binary problem for each label
  - simple, efficient
  - does not take into account label correlations
- Nearest neighbor approaches based on BR, e.g. MLkNN
- Stacking approaches, e.g. meta level stacking (MS)
- Pairwise approaches, e.g. calibrated label ranking
- Label powerset method: label sets are treated as single labels
  - takes into account label correlations
  - computationally complex

Prior Work

- Binary relevance method (BR): binary problem for each label
  - simple, efficient
  - does not take into account label correlations
- Nearest neighbor approaches based on BR, e.g. MLkNN
- Stacking approaches, e.g. meta level stacking (MS)
- Pairwise approaches, e.g. calibrated label ranking
- Label powerset method: label sets are treated as single labels
  - takes into account label correlations
  - computationally complex
- RAKEL: ensembles of subsets
- EPS: ensembles of pruned sets

Prior Work

- Binary relevance method (BR): binary problem for each label
  - simple, efficient
  - does not take into account label correlations
- Nearest neighbor approaches based on BR, e.g. MLkNN
- Stacking approaches, e.g. meta level stacking (MS)
- Pairwise approaches, e.g. calibrated label ranking
- Label powerset method: label sets are treated as single labels
  - takes into account label correlations
  - computationally complex
- RAKEL: ensembles of subsets
- EPS: ensembles of pruned sets
- Many other methods
  - take into account label correlations
  - complex, prone to overfitting

Binary Relevance (BR)

$L = \{\texttt{romance,horror,comedy,drama,action,western}\}$ $(|L| = 6)$

Classifiers	Classifications
$C_1 : x \rightarrow \{\texttt{romance},\texttt{!romance}\}$	romance
$C_2 : x \rightarrow \{\texttt{horror},\texttt{!horror}\}$	!horror
$C_3 : x \rightarrow \{\texttt{comedy},\texttt{!comedy}\}$	comedy
$C_4 : x \rightarrow \{\texttt{drama},\texttt{!drama}\}$	!drama
$C_5 : x \rightarrow \{\texttt{action},\texttt{!action}\}$	!action
$C_6 : x \rightarrow \{\texttt{western},\texttt{!western}\}$	!western
$Y \subseteq L$	$\{\texttt{romance,comedy}\}$

Binary Relevance (BR)

$L = \{\texttt{romance,horror,comedy,drama,action,western}\}$ $(|L| = 6)$

Classifiers	Classifications
$C_1 : x \rightarrow \{\texttt{romance}, \texttt{!romance}\}$	romance
$C_2 : x \rightarrow \{\texttt{horror}, \texttt{!horror}\}$	!horror
$C_3 : x \rightarrow \{\texttt{comedy}, \texttt{!comedy}\}$	comedy
$C_4 : x \rightarrow \{\texttt{drama}, \texttt{!drama}\}$	!drama
$C_5 : x \rightarrow \{\texttt{action}, \texttt{!action}\}$	!action
$C_6 : x \rightarrow \{\texttt{western}, \texttt{!western}\}$	!western
$Y \subseteq L$	$\{\texttt{romance,comedy}\}$

- simple, intuitive
- efficient
- useful for incremental contexts
- doesn't account for label correlations

Classifier Chains (CC)

$L = \{\texttt{romance},\texttt{horror},\texttt{comedy},\texttt{drama},\texttt{action},\texttt{western}\} \quad (|L| = 6)$

Classifiers	Classifications
$C_1 : x \longrightarrow \{\texttt{romance}, \texttt{!romance}\}$	romance
$C_2 : x \cup \texttt{romance} \longrightarrow \{\texttt{horror}, \texttt{!horror}\}$	!horror
$C_3 : x \cup \texttt{romance} \cup \texttt{!horror} \longrightarrow \{\texttt{comedy}, \texttt{!comedy}\}$	comedy
$C_4 : x \cup \texttt{romance} \cup \texttt{!horror} \cup \texttt{comedy} \longrightarrow \{\texttt{drama}, \texttt{!drama}\}$	!drama
$C_5 : x \cup \texttt{romance} \cup \texttt{!horror} \cup \texttt{comedy} \cup \texttt{!drama} \longrightarrow \cdots$	!action
$C_6 : x \cup \texttt{romance} \cup \texttt{!horror} \cup \texttt{comedy} \cup \texttt{!drama} \cup \cdots$	!western
	$Y \subseteq L = \{\texttt{romance},\texttt{comedy}\}$

Classifier Chains (CC)

$L = \{\texttt{romance},\texttt{horror},\texttt{comedy},\texttt{drama},\texttt{action},\texttt{western}\} \quad (|L| = 6)$

Classifiers	Classifications
$C_1 : x \rightarrow \{\texttt{romance}, \texttt{!romance}\}$	`romance`
$C_2 : x \cup \texttt{romance} \rightarrow \{\texttt{horror}, \texttt{!horror}\}$	`!horror`
$C_3 : x \cup \texttt{romance} \cup \texttt{!horror} \rightarrow \{\texttt{comedy}, \texttt{!comedy}\}$	`comedy`
$C_4 : x \cup \texttt{romance} \cup \texttt{!horror} \cup \texttt{comedy} \rightarrow \{\texttt{drama}, \texttt{!drama}\}$	`!drama`
$C_5 : x \cup \texttt{romance} \cup \texttt{!horror} \cup \texttt{comedy} \cup \texttt{!drama} \rightarrow \cdots$	`!action`
$C_6 : x \cup \texttt{romance} \cup \texttt{!horror} \cup \texttt{comedy} \cup \texttt{!drama} \cup \cdots$	`!western`
	$Y \subseteq L = \{\texttt{romance},\texttt{comedy}\}$

- similar advantages to binary relevance method
- time complexity similar in practice
- takes into account label correlations
- how to order the chain?

Ensembles of Classifier Chains (ECC)

- Ensembles known for augmenting accuracy
- more label correlations learnt, without overfitting
- solves ‘chain order’ issue: each chain has a random order

Ensembles of Classifier Chains (ECC)

- Ensembles known for augmenting accuracy
- more label correlations learnt, without overfitting
- solves 'chain order' issue: each chain has a random order

- For $i \in 1 \cdots m$ iterations:
  - $L' \leftarrow$ shuffle label set L
  - $D' \leftarrow$ subset of training set D
  - train a model $\texttt{CC}_i$ given L' and D'

- Generic vote/score/threshold method for classification:
  - collect votes from models
  - assign a score to each label
  - apply a threshold to determine relevant labels

Ensembles of Classifier Chains (ECC)

- Ensembles known for augmenting accuracy
- more label correlations learnt, without overfitting
- solves 'chain order' issue: each chain has a random order

- For $i \in 1 \cdots m$ iterations:
  - $L' \leftarrow$ shuffle label set L
  - $D' \leftarrow$ subset of training set D
  - train a model $\texttt{CC}_i$ given L' and D'

- Generic vote/score/threshold method for classification:
  - collect votes from models
  - assign a score to each label
  - apply a threshold to determine relevant labels

- Can also be *applied to binary relevance method*, i.e. EBR

Experiments

- WEKA-based framework
- Support Vector Machines as base classifiers
- Multi-label datasets:

	Labels $\|L\|$	Instances $\|D\|$
6 Standard	$6 \cdots 103$	$2407 \cdots 6000$
6 Large	$22 \cdots 983$	$7395 \cdots 95424$

- Multi-label evaluation metrics:
  - accuracy, macro F-measure (label set evaluation)
  - log loss, $AU(\overline{PRC})$ (per-label evaluation)
  - build times, test times
- Method parameters preset to optimise *predictive performance* (ECC requires no additional parameters)
- Experiments:
  1. Compare Classifier Chains (CC) to the Binary Relevance method (BR) and related BR-based methods.
  2. Compare ECC to EBR and modern methods of proven success: RAKEL, EPS, and MLkNN

Results 1

Comparing CC to BR and related methods SM[1] and MS[2].

Table: Standard Datasets: Wins for each evaluation measure.

	CC	BR	SM	MS
Accuracy	**5**	0	1	0
Macro F1	**5**	0	1	0
Micro F1	**3**	1	0	2
Exact Match	**6**	0	0	0
Total wins	**19**	1	2	2

- CC's chaining technique justified over default BR
- CC outperforms other similar methods

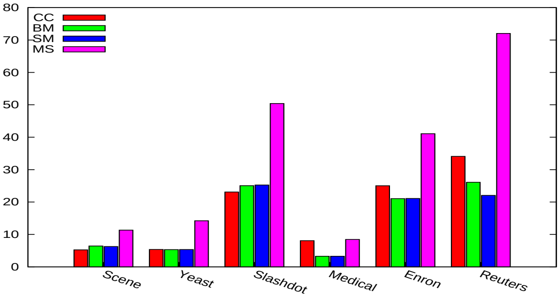
[1] Subset Mapping: maps output of BR to nearest (Hamming distance) known subset

[2] Meta Stacking: stacking the output of BR with meta classifiers

Results 1

Comparing CC to BR and related methods SM[1] and MS[2].

Figure: Standard Datasets: Build times (seconds).



- CC's complexity comparable to BR
  - except for special cases like *Medical* (relatively large label set)

[1]Subset Mapping: maps output of BR to nearest (Hamming distance) known subset

[2]Meta Stacking: stacking the output of BR with meta classifiers

Results 2

Comparing ECC to EBR and methods: RAKEL[3], EPS[4], and MLkNN[5].

Table: Standard Datasets: Wins for each evaluation measure.

	ECC	EBR	RAKEL	EPS	MLkNN
Accuracy	2	0	1	**3**	0
Macro F1	1	0	1	**4**	0
Log Loss	**3**	0	1	1	1
$AU(\overline{PRC})$	**3**	0	0	0	**3**
Total wins	**9**	0	3	8	4

- ECC best at per-label prediction (as a binary method)
- Other methods can sometimes predict better label sets
- ECC rewarded by conservative prediction (log loss)

[3] Tsoumakas and Vlahavas, 2007

[4] Read, Pfahringer, Holmes, 2008

[5] Zhang and Zhou, 2005

Results 2

Comparing ECC to EBR and methods: RAKEL[3], EPS[4], and MLkNN[5].

Table: Large Datasets: Wins for each evaluation measure.

	ECC	EBR	RAKEL†	EPS†	MLkNN
Accuracy	**4**	0	0	1	1
Macro F1	**3**	0	1	1	1
Log Loss	1	1	0	0	**4**
$AU(\overline{PRC})$	**4**	0	0	0	2
Total wins	**12**	1	1	2	8

†Note: 2 DNF for RAKEL and 1 DNF for EPS.

- Binary methods are the best choice for large datasets
- ECC best overall

[3] Tsoumakas and Vlahavas, 2007
[4] Read, Pfahringer, Holmes, 2008
[5] Zhang and Zhou, 2005

Results 2

Comparing build and test times between ECC, RAKEL, and EPS.

Table: All Datasets: Method with fastest Build,Test time†.

Dataset	Build	Test	Dataset	Build	Test
Scene	EPS	RAK	OHSUMED	ECC	ECC
Yeast	ECC	ECC	TMC2007	EPS	ECC
Slashdot	RAK	RAK	Bibtex	ECC	ECC
Medical	RAK	RAK	MediaMill	ECC	ECC
Enron	EPS	ECC	IMDB	RAK	ECC
Reuters	ECC	ECC	Delicious	EPS	EPS

†EBR and MLkNN not included

- ECC's efficiency most noticeable on the larger datasets
- RAKEL most efficient on smaller datasets
- EPS can make large gains by pruning, but occasionally too much

Conclusion

- Ensembles of Classifier Chains
  - classifier chains improves on the binary relevance method
  - takes into account label correlations without overfitting
  - flexible, efficient
  - performs well, especially on large data sets

Conclusion

- Ensembles of Classifier Chains
    - classifier chains improves on the binary relevance method
    - takes into account label correlations without overfitting
    - flexible, efficient
    - performs well, especially on large data sets

Thank you. Any questions?