Efficient Multi-label Classification

Jesse Read
(Supervisors: Bernhard Pfahringer, Geoff Holmes)

November 2009

Outline

1. Introduction
2. Pruned Sets (PS)
3. Classifier Chains (CC)
4. Related Work
5. Experiments
6. Scaling up to Large Datasets
7. Summary

Multi-label Classification

- Single-label (Multi-class) Classification
  - Examples: $D = \{x_1, \cdots, x_n\}$ Labels: $L = \{l_1, \cdots, l_m\}$
  - Each example is associated with one label: $(x, l \in L)$
- Multi-label Classification
  - Examples: $D = \{x_1, \cdots, x_n\}$ Labels: $L = \{l_1, \cdots, l_m\}$
  - Each example is associated with a *subset* of labels: $(x, S \subseteq L)$

Multi-label Classification

- Single-label (Multi-class) Classification
  - Examples: $D = \{x_1, \cdots, x_n\}$ Labels: $L = \{l_1, \cdots, l_m\}$
  - Each example is associated with one label: $(x, l \in L)$
- Multi-label Classification
  - Examples: $D = \{x_1, \cdots, x_n\}$ Labels: $L = \{l_1, \cdots, l_m\}$
  - Each example is associated with a *subset* of labels: $(x, S \subseteq L)$

Multi-label Data - Dataset Statistics

	$\|D\|$	$\|L\|$	$avg.\|S\|$	$uniq.S$	Type
Music	593	6	1.87	0.046	media
Scene	2407	6	1.07	0.006	media
Yeast	2417	14	4.24	0.082	biology
Genbase	661	27	1.25	0.048	biology
Medical	978	45	1.25	0.096	medical text
Slashdot	3782	22	1.18	0.041	news
Lang.Log	1460	75	1.18	0.208	forum
Enron	1702	53	3.38	0.442	e-mail
Reuters(avg)	6000	103	1.46	0.147	news
OHSUMED	13929	23	1.66	0.082	medical text
tmc2007	28596	22	2.16	0.047	text
Media Mill	43907	101	4.38	0.149	media
Bibtex	7395	159	2.40	0.386	text
IMDB	95424	28	1.92	0.036	text
del.icio.us	16105	983	19.02	0.981	text

Multi-label Data - Label Correlations

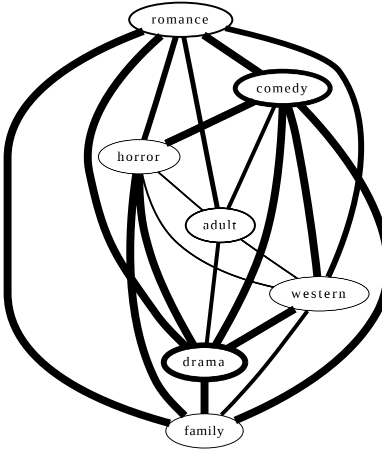

Figure: Label correlations: *IMDB* subset.

Freq. /403756	Combination ($S \subseteq L$) $\|L\| = 28$
56620	drama
43968	short
42024	documentary
36794	adult
35849	comedy
27713	documentary,short
25268	comedy,short
19634	drama,short
10031	animation,short
6550	action
...	...
4360	crime,drama
4042	horror
...	...
337	documentary,war
342	comedy,western
361	action,sci-fi
...	...
29	horror,romance
21	adult,western
...	...

Multi-label Data - Label Correlations

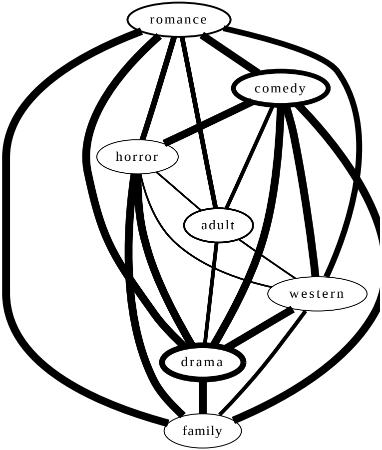

Figure: Label correlations: *IMDB* subset.

Freq. /403756	Combination ($S \subseteq L$) $\vert L\vert = 28$
56620	drama
43968	short
42024	documentary
36794	adult
35849	comedy
27713	documentary,short
25268	comedy,short
19634	drama,short
10031	animation,short
6550	action
...	...
4360	crime,drama
4042	horror
...	...
337	documentary,war
342	comedy,western
361	action,sci-fi
...	...
29	horror,romance
21	adult,western
...	...

where {**horror,romance**} $\subseteq S$
Thriller,{horror, romance, short, thriller, music}
Love at First Bite,{horror, comedy, romance}
Kondom des Grauens,{comedy, horror, romance}
...

Multi-label Data - Label Correlations

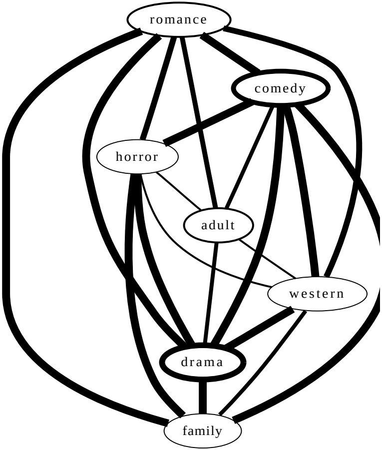

Figure: Label correlations: *IMDB* subset.

Freq. /403756	Combination ($S \subseteq L$) $\|L\| = 28$
56620	drama
43968	short
42024	documentary
36794	adult
35849	comedy
27713	documentary,short
25268	comedy,short
19634	drama,short
10031	animation,short
6550	action
...	...
4360	crime,drama
4042	horror
...	...
337	documentary,war
342	comedy,western
361	action,sci-fi
...	...
29	horror,romance
21	adult,western
...	...

where {**horror,romance**} $\subseteq S$
Thriller, {horror, romance, short, thriller, music}
Love at First Bite, {horror, comedy, romance}
Kondom des Grauens, {comedy, horror, romance}
...

where {**adult,western**} $\subseteq S$
Hard on the Trail , {action, adult, western}
Good the Bad the Nasty, The, {adult, comedy, western}
Ride a Wild Stud , {adult, drama, western}
...

Multi-label Evaluation

For each test example $(x_i, S_i \subseteq L)$, classifier $H : x \rightarrow Y_i \subseteq L$.

e.g.: $L = \{$horror,romance,comedy,western,drama,family,$\ldots\}$
$S_i = \{$romance,comedy,drama$\}$
$Y_i = \{$romance,comedy$\}$

- Evaluation by example? $S_i = Y_i$? *too harsh*
- Evaluation by label? $\sum_j^{|L|} l_j = k_j ? | l_j \in S_i, k_j \in Y_i$ *too lenient*

Multi-label Evaluation

For each test example $(x_i, S_i \subseteq L)$, classifier $H : x \to Y_i \subseteq L$.

e.g.: $L = \{$horror,romance,comedy,western,drama,family,$\ldots\}$
$S_i = \{$romance,comedy,drama$\}$
$Y_i = \{$romance,comedy$\}$

- Evaluation by example? $S_i = Y_i$? *too harsh*
- Evaluation by label? $\sum_j^{|L|} l_j = k_j ? | l_j \in S_i, k_j \in Y_i$ *too lenient*

Combine with other Multi-label Evaluation Metrics:

- F-measure macro/micro averaged by $|D|/|L|$
- Subset accuracy
- One error, Coverage, Rank loss
- Log loss, Average Area Under the Precision Recall Curve

Algorithm Adaption or Problem Transformation

Algorithm Adaption

- Adapt a single-label algorithm for multi-label classification
- e.g. Multi-label Naive Bayes:
  - Posterior probabilities: $H : x \to P(l_1), P(l_2), \cdots, P(l_{|L|})$
  - Classify $l_j \in Y$ where $P(l_j) > 0.5$

Problem Transformation

- Transform a multi-label problem into single-label problems
- Use any single-label classifier for classification
- Flexible, involved in algorithm adaption anyway
- e.g. Binary Relevance method, Label Combination method

Problem Transformation: Binary Relevance Method (BR)

- One binary classifier predicts the association of each label
- $|L|$ Binary classifiers $H_1, \cdots, H_{|L|}$,
  each $H_j : x \rightarrow l_j / \neg l_j$ where all $l_j \in Y$

Example

- Multi-label example $(x, \{comedy, romance\})$
- Single-label example (1): $(x, \neg$ horror$)$
- Single-label example (2): $(x,$comedy$)$
- Single-label example (3): $(x,$romance$)$
- Single-label example (4): $(x, \neg$ western$)$
- Single-label example (.): $(x, \ldots)$
- Single-label example ($|L|$)): $(x, \ldots)$

- simple, intuitive, fast
- ignorant of label correlations

Problem Transformation: Label Combination Method (LC)

- Each unique label combination in the training set is treated as a single class-label
- One single-label classifier: $H : x \to Y | \exists (x, Y) \in D$

Example

- Multi-label example: $(x, \{\texttt{comedy}, \texttt{romance}, \texttt{western}\})$
- Single-label example: $(x, \texttt{comedy+romance+western})$

- Takes into account label correlations
- $2^{|L|}$ possible combinations

Pruned Sets (PS)[1]

Problem (Combination Method (LC)):

- $2^{|L|}$ combinations
  - slow
  - skewed
- e.g. (x,comedy+romance+western)

[1] Jesse Read, Bernhard Pfahringer, Geoff Holmes. *Multi-label Classification using Ensembles of Pruned Sets*. Proc. of IEEE International Conference on Data Mining (ICDM 2008), Pisa, Italy, 2008.

Pruned Sets (PS)[1]

Problem (Combination Method (LC)):

- $2^{|L|}$ combinations
    - slow
    - skewed
- e.g. (x,comedy+romance+western)

Solution (Pruned Sets (PS)):

- prune and reformat outlying combinations
- e.g. (x,comedy+romance),(x, western)

[1] Jesse Read, Bernhard Pfahringer, Geoff Holmes. *Multi-label Classification using Ensembles of Pruned Sets*. Proc. of IEEE International Conference on Data Mining (ICDM 2008), Pisa, Italy, 2008.

Pruned Sets (PS)[1]

Problem (Combination Method (LC)):

- $2^{|L|}$ combinations
  - slow
  - skewed
- e.g. (x,comedy+romance+western)

Solution (Pruned Sets (PS)):

- prune and reformat outlying combinations
- e.g. (x,comedy+romance),(x, western)

Expansion (Ensembles of PS (EPS)):

- form new combinations at classification time
- i.e. $H_1, \cdots, H_m : x \to Y_1, \cdots, Y_m \to Y \subseteq L$

[1] Jesse Read, Bernhard Pfahringer, Geoff Holmes. *Multi-label Classification using Ensembles of Pruned Sets*. Proc. of IEEE International Conference on Data Mining (ICDM 2008), Pisa, Italy, 2008.

Classifier Chains (CC)[2]

Problem (Binary Method (BM)):

- ignorance of label combinations
- e.g. (x,adult), (x, family), (x, $\neg$drama), ...

[2] Jesse Read, Bernhard Pfahringer, Geoff Holmes, Eibe Frank. *Classifier Chains for Multi-label Classification*. In Proc. of 20th European Conference on Machine Learning (ECML 2009). Bled, Slovenia, September 2009.

Classifier Chains ($\mathtt{CC}$)[2]

Problem (Binary Method ($\mathtt{BM}$)):

- ignorance of label combinations
- e.g. (x,$\mathtt{adult}$), (x, $\mathtt{family}$), (x, $\neg\mathtt{drama}$), ...

Solution (Classifier Chains ($\mathtt{CC}$)):

- chaining mechanism (efficient form of stacking)
- e.g. (x,$\mathtt{adult}$) $\rightarrow$ (x,?) $\rightarrow$ (x,?)

[2] Jesse Read, Bernhard Pfahringer, Geoff Holmes, Eibe Frank. *Classifier Chains for Multi-label Classification*. In Proc. of 20th European Conference on Machine Learning (ECML 2009). Bled, Slovenia, September 2009.

Classifier Chains (CC)[2]

Problem (Binary Method (BM)):

- ignorance of label combinations
- e.g. (x,adult), (x, family), (x, $\neg$drama), ...

Solution (Classifier Chains (CC)):

- chaining mechanism (efficient form of stacking)
- e.g. (x,adult) $\rightarrow$ (x,?) $\rightarrow$ (x,?)

Expansion (Ensembles of CC (ECC)):

- different chain orderings for each model

[2] Jesse Read, Bernhard Pfahringer, Geoff Holmes, Eibe Frank. *Classifier Chains for Multi-label Classification*. In Proc. of 20th European Conference on Machine Learning (ECML 2009). Bled, Slovenia, September 2009.

Related Work

- Problem Transformataion:
  - Binary pairwise classification e.g. (Fürnkranz et al., 2008)
  - Ensemble `LC` e.g. (Tsoumakas and Vlahavas, 2007)
  - Meta `BR` stacking e.g. (Godbole, Sarawagi, 2004)
- Algorithm Adaption:
  - `lazy/kNN`-based e.g. (Zhang and Zhao, 2007),
  - `NN`-based, `SVM`-based e.g. (Elisseeff and Weston, 2002)
  - `C4.5`-based e.g. (Clare and King, 2001)
  - `Boosting` e.g. (Schapire and Singer, 2000)
  - `Bayesian` / probabilistic e.g. (McCallum, 1999)

Experiments - Predictive Performance

Table: Number of wins over 9 datasets. SVMs used as base classifier for problem transformation methods.

Measure	ECC	EPS	CLR[3]	RAkEL[4]		MLkNN[5]	IBLR[6]
Subs.acc.	1.0	4.0	0.0	2.5	1.0	0.0	0.0
AU(PRC)	3.5	1.0	0.0	0.0	0.0	1.0	3.0
LogLoss	1.5	0.0	0.0	0.0	0.0	4.0	3.0
F1-micro	1.0	1.5	0.0	4.0	1.5	0.0	0.0
E-match	3.3	2.0	0.0	0.8	1.3	0.0	1.0
Total	10.3	8.5	0.0	7.3	3.8	5.0	7.0

[3] CLR: Calibrated Label Ranking (binary pairwise) (Fürnkranz et al., 2008)

[4] RAkEL: RAndom *k*-labEL subsets (Ens. LC) (Tsoumakas and Vlahavas, 2007) - two parameter configurations

[5] MLkNN: kNN + Bayes inference (Zhang and Zhou, 2007)

[6] IBLR: kNN + logistic regression (Cheng and Hüllermeier, 2009)

Experiments - Computational Complexity - $|L|$

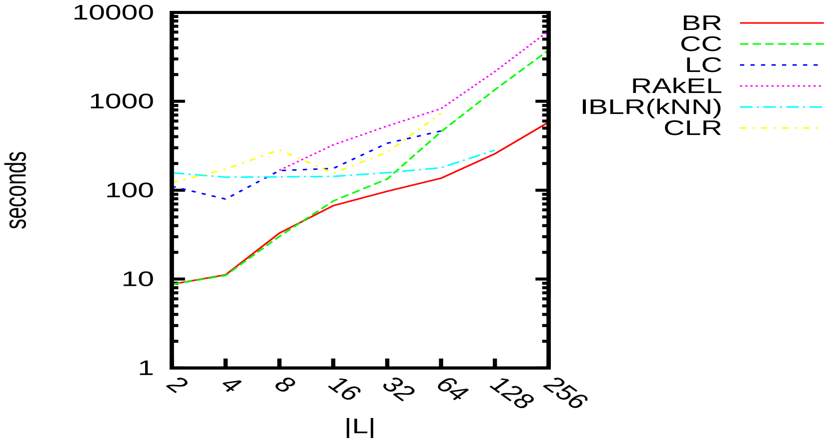

Figure: Synthetic data: $|L| = 2, 4, \cdots, 256$, $|D| = 3000$, $|X| = 500$. `j48` used as the base-classifier for problem transformation methods.

Experiments - Computational Complexity - $|D|$

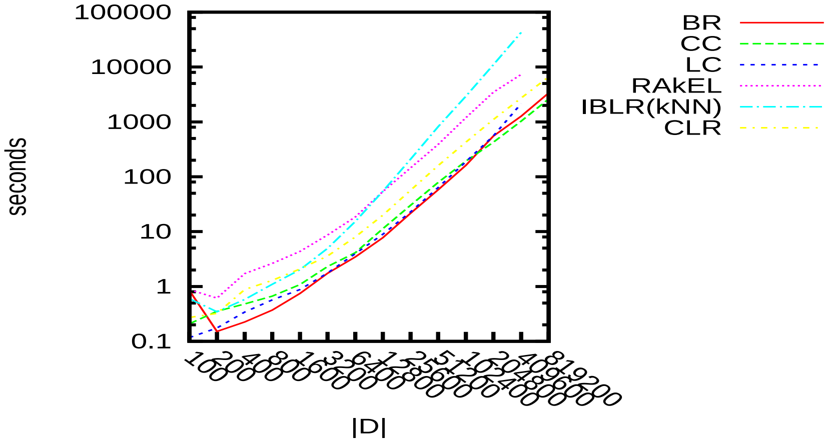

Figure: Synthetic data: $|D| = 100, 200, \cdots, 819200$, $|L| = 10$, $|X| = 20$. j48 used as the base-classifier for problem transformation methods.

Efficient Multi-label Classification: Scaling Up

large `EPS` efficient version of `LC`

larger `ECC` scales up to large datasets

largest Can adapt `EPS` and `ECC` to *incremental* settings

- using e.g. `NaiveBayes`, `HoeffdingTrees` as base classifier
- for very large datasets, data streams

Experiments - Very Large Datasets

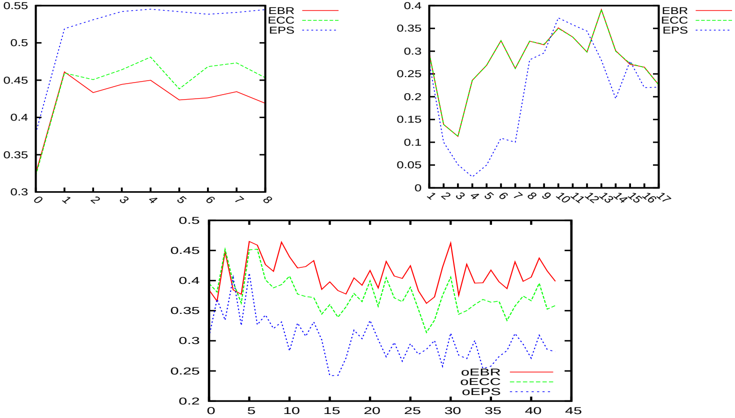

Figure: o−EPS, o−ECC, o−EBR on different datasets, x axis in *thousands*. Left: 'regular' labelling: binary methods become skewed. Right: concept drift: combination-based method slow to adapt. Bottom: 'irregular' labelling: label combinations become distracting.

How Far has Multi-label Classification Come?

- Beginning 2007:
    - about 5 good relevant papers
    - 1 - 3 datasets used in evaluation
    - 1 evaluation measure
    - parameters optimised on test data
    - comparison only to the binary relevance method
    - scalability largely ignored
- End of 2009:
    - about 5 relevant papers per conference
    - 10+ datasets used in evaluation
    - 4+ evaluation measures
    - various statistical significance tests
    - comparison to a wide range of algorithms
    - scalability to large datasets a hot topic

Thank you

- Thanks for your attention
- Datasets, Links, `MOA`-based and `WEKA`-based Software: `http://www.cs.waikato.ac.nz/~jmr30/`