Multi-label Classification using Ensembles of Pruned Sets

Jesse Read, Bernhard Pfahringer, Geoff Holmes

University of Waikato
New Zealand

ICDM 2008, December 15, 2008. Pisa, Italy

Introduction

- A set of instances: $D = \{x_0, x_1, \cdots, x_m\}$
- A set of *predefined* labels: $L = \{l_0, l_1, \cdots, l_n\}$
- Single-label Classification: Each instance is assigned a label: $(x, l \in L)$
- Multi-label Classification: Each instance is assigned a subset of labels: $(x, S \subseteq L)$

Introduction

- A set of instances: $D = \{x_0, x_1, \cdots, x_m\}$
- A set of *predefined* labels: $L = \{l_0, l_1, \cdots, l_n\}$
- Single-label Classification: Each instance is assigned a label: $(x, l \in L)$
- Multi-label Classification: Each instance is assigned a subset of labels: $(x, S \subseteq L)$

- Example Applications
  - a film can be labeled `Romance` and `Comedy`
  - a news article can be about `Science` and `Technology`
  - an image can contain `Beach`, `Sunset` and `Mountains`
  - a patient's symptoms may correspond to *various ailments*
  - a collection of genes can have *multiple functions*

Introduction

- A set of instances: $D = \{x_0, x_1, \cdots, x_m\}$
- A set of *predefined* labels: $L = \{l_0, l_1, \cdots, l_n\}$
- Single-label Classification: Each instance is assigned a label: $(x, l \in L)$
- Multi-label Classification: Each instance is assigned a subset of labels: $(x, S \subseteq L)$

- Example Applications
  - a film can be labeled `Romance` and `Comedy`
  - a news article can be about `Science` and `Technology`
  - an image can contain `Beach`, `Sunset` and `Mountains`
  - a patient's symptoms may correspond to *various ailments*
  - a collection of genes can have *multiple functions*

- Some Multi-label-centric Issues
  - label correlations
    - consider {`Romance`,`Comedy`} vs {`Romance`,`Horror`}
  - computational complexity

Problem Transformation

Problem Transformation

Any multi-label problem can be transformed into one or several single-label problems. Any single-label classifier can be used.

- Problem transformation is core to most multi-label classification, even "algorithm adaption" methods
- There are several "base" methods common to many works
  - e.g. Combination Method (`CM`)

Problem Transformation

Problem Transformation

Any multi-label problem can be transformed into one or several single-label problems. Any single-label classifier can be used.

- Problem transformation is core to most multi-label classification, even "algorithm adaption" methods
- There are several "base" methods common to many works
  - e.g. Combination Method (`CM`)

Combination Method (`CM`)

Each label subset $S \subseteq L$ is treated as a single label, thus forming a single-label problem. The distinct label sets are the possible single labels.

- takes into account label correlations
- many single labels to choose from
- cannot predict new combinations

The Pruned Sets Method (PS)

- Multi-label data:
  - Some label correlations are very frequent
  - Most label correlations are very *infrequent*

The Pruned Sets Method (PS)

- Multi-label data:
  - Some label correlations are very frequent
  - Most label correlations are very *infrequent*

The Pruned Sets Method (PS)

- Treat each label set as a single-label (as per CM)
  - preserves label correlation information
- Prune away infrequent sets and;
- decompose these sets into frequent sets
  - e.g. $(movie_i, \{\texttt{Romance,Comedy,Horror}\})$ (infrequent) $\rightarrow$(movie$_i$, $\{\texttt{Romance,Comedy}\}$), $(movie_i, \{\texttt{Comedy,Horror}\})$ ...
  - represents only the core label sets as single-labels
  - fewer single labels to learn/choose from (efficient/less error prone)

The Pruned Sets Method (PS)

- Multi-label data:
  - Some label correlations are very frequent
  - Most label correlations are very *infrequent*

The Pruned Sets Method (PS)

- Treat each label set as a single-label (as per `CM`)
  - preserves label correlation information
- Prune away infrequent sets and;
- decompose these sets into frequent sets
  - e.g. $(movie_i, \{\texttt{Romance}, \texttt{Comedy}, \texttt{Horror}\})$ (infrequent) $\rightarrow$(movie$_i$, $\{\texttt{Romance}, \texttt{Comedy}\}$), $(movie_i, \{\texttt{Comedy}, \texttt{Horror}\}) \ldots$
  - represents only the core label sets as single-labels
  - fewer single labels to learn/choose from (efficient/less error prone)
  - cannot predict new combinations
  - prone to over-fitting the data

Ensembles of Pruned Sets (EPS)

Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on *subsets* of the training data
  - introduces variation
- The predictions are combined to form new combinations
  - reduces over-fitting
  - more robust

Ensembles of Pruned Sets (EPS)

Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on *subsets* of the training data
  - introduces variation
- The predictions are combined to form new combinations
  - reduces over-fitting
  - more robust

Example (EPS - Classification Phase)

Ensemble	PS_0	PS_1	PS_2	PS_3	PS_4	PS_5
SL Predictions	(M)	(A,F)	(A,C)	(A,F)	(M)	(M)

Ensembles of Pruned Sets (EPS)

Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on *subsets* of the training data
  - introduces variation
- The predictions are combined to form new combinations
  - reduces over-fitting
  - more robust

Example (EPS - Classification Phase)

Ensemble	PS_0	PS_1	PS_2	PS_3	PS_4	PS_5
SL Predictions	(M)	(A,F)	(A,C)	(A,F)	(M)	(M)

Counts	
A	3
M	3
F	2
C	1

Ensembles of Pruned Sets (EPS)

Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on *subsets* of the training data
  - introduces variation
- The predictions are combined to form new combinations
  - reduces over-fitting
  - more robust

Example (EPS - Classification Phase)

Ensemble	PS_0	PS_1	PS_2	PS_3	PS_4	PS_5
SL Predictions	(M)	(A,F)	(A,C)	(A,F)	(M)	(M)
Classif.($\subseteq L$)			$\{A, M, F\}$			

Counts	
A	0.375
M	0.375
F	0.250
$t = 0.2$	
C	0.125

- *Reuters* dataset ($|D| = 6000, |L| = 103$) 50/50 train/test split
- BM: Binary Method (one binary classifier per label)
- CM: Combination Method (each set is a single-label)
- EPS,RAKEL: 10 models, auto-tuned threshold, varying p,k
  - e.g. $p = 3$: only label sets occurring > 3 times are *frequent*
- *All using Support Vector Machines as single-label classifiers*

BM	
Time	Acc.
123	32.48

CM	
Time	Acc.
1,379	48.75

EPS		
p	Time	Acc.
5	194	48.01
4	277	48.51
3	408	48.40
2	719	48.71
1	1,553	49.97

RAKEL		
k	Time	Acc.
2	10	10.05
25	350	36.66
50	3,627	44.70
61*	22,337	47.35
102	DNF	DNF

Conclusions

- Ensembles of Pruned Sets: A new problem transformation method
  - classifier independent
  - improved performance over BM, CM, and RAKEL
  - efficient in practice
- Main contribution: focus on core label correlations
  - pruning infrequent sets
  - set decomposition into frequent sets
  - flexible pruning parameter p
  - can be combined easily with other methods

End

End

	$\|D\|$	$\|L\|$	$LC(D)$	$PD(D)$	Description.
Scene	2407	6	1.07	0.006	still scenes
Yeast	2417	14	4.24	0.082	protein function
Medical	978	45	1.25	0.096	medical text
Enron	1702	53	3.38	0.442	e-mail corpus
Reuters	6000	103	1.46	0.147	newswire stories

- D = full dataset
- L = label set
- LC = Label Cardinality. Average number of labels per instance in D
- PD = Percent Dinstinct. The percentage of instances with a distinct label set

End

- Framework
  - WEKA[1] framework
  - using Support Vector Machines (SVM) as single-label classifiers (default parameters)
  - 5×2 Cross Validation (CV)
- Problem Transformation parameters
  - trialled in order according to theoretical complexity
  - under $5 \times CV$ on training set
  - cut off: 1 hour per parameter combination
- Evaluation Methods
  - $Accuracy(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|S_i \cap Y_i|}{|S_i \cup Y_i|}$
  - Micro $F_1(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{2 \times prec_i \times recall_i}{prec_i + recall_i}$
  - $Hamming\ loss(D) = 1 - \frac{1}{|D| \times |L|} \sum_{i=1}^{|D|} |S_i \oplus Y_i|$

[1] http://www.cs.waikato.ac.nz/ml/weka/

End

- CM: Combination Method
- BM: Binary Method
- RM: Ranking Method
  - tune threshold $t = \{0.1, \cdots, 0.9\}$
- PS: Pruned Sets method
  - tune parameter $p = \{5, 4, 3, 2, 1\}$
  - tune parameter $s = \{-, A_1, A_2, A_3, B_1, B_2, B_3\}$
- EPS: Ensembles of Pruned Sets
  - tune parameters using a single PS method
  - tune threshold $t = \{0.1, \cdots, 0.9\}$
- RAKEL: RAndom K labEL subsets
  - parameter range as per paper
  - tune threshold $t = \{0.1, \cdots, 0.9\}$

End

	BM	[CM]	RAKEL	PS	EPS
Scene	58.28$\searrow$	71.81	71.58	71.93	73.80
Yeast	49.64$\searrow$	51.98	54.49	52.82	55.03
Medical	73.00	74.71	72.55	74.63	74.45
Enron	31.91	41.02	42.98	42.15	44.09
Reuters	38.64$\searrow$	49.17	31.80	49.83	49.80

- Accuracy Measure
- Paired t Test (against `CM`)
  - $\nearrow$, $\searrow$ statistically significant improvement,degradation

End

	BM	[CM]	RAKEL	PS	EPS
Scene	0.671 $\searrow$	0.729	0.735	0.730	0.752 $\nearrow$
Yeast	0.630	0.633	0.664 $\nearrow$	0.643	0.655 $\nearrow$
Medical	0.791 $\nearrow$	0.767	0.784	0.766	0.764
Enron	0.504	0.502	0.543 $\nearrow$	0.520	0.543 $\nearrow$
Reuters	0.421 $\searrow$	0.482	0.418 $\searrow$	0.496	0.499 $\nearrow$

- F_1 Measure
- Paired t Test (against `CM`)
  - $\nearrow$, $\searrow$ statistically significant improvement,degradation