A Pruned Problem Transformation Method for Multi-label Classification

Jesse Read

jmr30@cs.waikato.ac.nz

University of Waikato

Outline

- Single-label classification
- Multi-label classification
- Problem Transformation
  - Binary Method
  - Combination Method
- PPT: A Pruned Problem Transformation method
- Experiments I
- PPT-ext: PPT extended
- Experiments II
- Summary

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy..."	
"Antarctic food chain in danger..."	
"Top sports stars fuelling success..."	
"Steeled for ironman..."	
"Greens claim report doctored..."	
"Revealed: Polluting impact of humans on the oceans..."	
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy..."	$Science$
"Antarctic food chain in danger..."	
"Top sports stars fuelling success..."	
"Steeled for ironman..."	
"Greens claim report doctored..."	
"Revealed: Polluting impact of humans on the oceans..."	
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy. . . "	$Science$
"Antarctic food chain in danger. . . "	$Science$
"Top sports stars fuelling success. . . "	
"Steeled for ironman. . . "	
"Greens claim report doctored. . . "	
"Revealed: Polluting impact of humans on the oceans. . . "	
"Union muzzled while awaiting poll watchdog's ruling. . . "	
"Technology pushes sporting boundaries. . . "	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy…"	$Science$
"Antarctic food chain in danger…"	$Science$
"Top sports stars fuelling success…"	$Sport$
"Steeled for ironman…"	
"Greens claim report doctored…"	
"Revealed: Polluting impact of humans on the oceans…"	
"Union muzzled while awaiting poll watchdog's ruling…"	
"Technology pushes sporting boundaries…"	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy…"	$Science$
"Antarctic food chain in danger…"	$Science$
"Top sports stars fuelling success…"	$Sport$
"Steeled for ironman…"	$Sport$
"Greens claim report doctored…"	
"Revealed: Polluting impact of humans on the oceans…"	
"Union muzzled while awaiting poll watchdog's ruling…"	
"Technology pushes sporting boundaries…"	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy..."	$Science$
"Antarctic food chain in danger..."	$Science$
"Top sports stars fuelling success..."	$Sport$
"Steeled for ironman..."	$Sport$
"Greens claim report doctored..."	$Politics$
"Revealed: Polluting impact of humans on the oceans..."	
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy..."	$Science$
"Antarctic food chain in danger..."	$Science$
"Top sports stars fuelling success..."	$Sport$
"Steeled for ironman..."	$Sport$
"Greens claim report doctored..."	$Politics$
"Revealed: Polluting impact of humans on the oceans..."	$Environment$
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
“NZ scientists help discover solar system in our galaxy. . . ”	$Science$
“Antarctic food chain in danger. . . ”	$Science$
“Top sports stars fuelling success. . . ”	$Sport$
“Steeled for ironman. . . ”	$Sport$
“Greens claim report doctored. . . ”	$Politics$
“Revealed: Polluting impact of humans on the oceans. . . ”	$Environment$
“Union muzzled while awaiting poll watchdog’s ruling. . . ”	$Politics$
“Technology pushes sporting boundaries. . . ”	

Single-label (Multi-class) Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label $l \in L$
- Single-label representation: (d, l)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($l \in L$)
"NZ scientists help discover solar system in our galaxy…"	$Science$
"Antarctic food chain in danger…"	$Science$
"Top sports stars fuelling success…"	$Sport$
"Steeled for ironman…"	$Sport$
"Greens claim report doctored…"	$Politics$
"Revealed: Polluting impact of humans on the oceans…"	$Environment$
"Union muzzled while awaiting poll watchdog's ruling…"	$Politics$
"Technology pushes sporting boundaries…"	$Science$

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy..."	
"Antarctic food chain in danger..."	
"Top sports stars fuelling success..."	
"Steeled for ironman..."	
"Greens claim report doctored..."	
"Revealed: Polluting impact of humans on the oceans..."	
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy…"	$\{Science\}$
"Antarctic food chain in danger…"	
"Top sports stars fuelling success…"	
"Steeled for ironman…"	
"Greens claim report doctored…"	
"Revealed: Polluting impact of humans on the oceans…"	
"Union muzzled while awaiting poll watchdog's ruling…"	
"Technology pushes sporting boundaries…"	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy..."	$\{Science\}$
"Antarctic food chain in danger..."	$\{Science, Environment\}$
"Top sports stars fuelling success..."	
"Steeled for ironman..."	
"Greens claim report doctored..."	
"Revealed: Polluting impact of humans on the oceans..."	
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy..."	$\{Science\}$
"Antarctic food chain in danger..."	$\{Science, Environment\}$
"Top sports stars fuelling success..."	$\{Sport\}$
"Steeled for ironman..."	
"Greens claim report doctored..."	
"Revealed: Polluting impact of humans on the oceans..."	
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy..."	$\{Science\}$
"Antarctic food chain in danger..."	$\{Science, Environment\}$
"Top sports stars fuelling success..."	$\{Sport\}$
"Steeled for ironman..."	$\{Sport\}$
"Greens claim report doctored..."	
"Revealed: Polluting impact of humans on the oceans..."	
"Union muzzled while awaiting poll watchdog's ruling..."	
"Technology pushes sporting boundaries..."	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy…"	$\{Science\}$
"Antarctic food chain in danger…"	$\{Science, Environment\}$
"Top sports stars fuelling success…"	$\{Sport\}$
"Steeled for ironman…"	$\{Sport\}$
"Greens claim report doctored…"	$\{Politics, Environment\}$
"Revealed: Polluting impact of humans on the oceans…"	
"Union muzzled while awaiting poll watchdog's ruling…"	
"Technology pushes sporting boundaries…"	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
“NZ scientists help discover solar system in our galaxy...”	$\{Science\}$
“Antarctic food chain in danger...”	$\{Science, Environment\}$
“Top sports stars fuelling success...”	$\{Sport\}$
“Steeled for ironman...”	$\{Sport\}$
“Greens claim report doctored...”	$\{Politics, Environment\}$
“Revealed: Polluting impact of humans on the oceans...”	$\{Environment, Science\}$
“Union muzzled while awaiting poll watchdog’s ruling...”	
“Technology pushes sporting boundaries...”	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy. . . "	$\{Science\}$
"Antarctic food chain in danger. . . "	$\{Science, Environment\}$
"Top sports stars fuelling success. . . "	$\{Sport\}$
"Steeled for ironman. . . "	$\{Sport\}$
"Greens claim report doctored. . . "	$\{Politics, Environment\}$
"Revealed: Polluting impact of humans on the oceans. . . "	$\{Environment, Science\}$
"Union muzzled while awaiting poll watchdog's ruling. . . "	$\{Politics\}$
"Technology pushes sporting boundaries. . . "	

Multi-label Classification

- Set of documents D. Set of labels L.
- For each $d \in D$, select a label subset $S \subseteq L$
- Multi-label representation: (d, S)

e.g. $L = \{Sport, Environment, Science, Politics\}$:

Document ($d \in D$)	**Label** ($S \subseteq L$)
"NZ scientists help discover solar system in our galaxy..."	$\{Science\}$
"Antarctic food chain in danger..."	$\{Science, Environment\}$
"Top sports stars fuelling success..."	$\{Sport\}$
"Steeled for ironman..."	$\{Sport\}$
"Greens claim report doctored..."	$\{Politics, Environment\}$
"Revealed: Polluting impact of humans on the oceans..."	$\{Environment, Science\}$
"Union muzzled while awaiting poll watchdog's ruling..."	$\{Politics\}$
"Technology pushes sporting boundaries..."	$\{Sport, Science\}$

Applications of ML Classification

Using Machine Learning to train from manually multi-labelled documents, and learn to automatically classify new documents with multi-labels (AKA ‘tags’).

- News articles
- Encyclopedia articles
- Academic papers (categories, key words)
- Emails
- Internet forum posts
- Web pages (as bookmarks, web directories)
- RSS feeds
- Biological applications (genes, etc. . .)

Problem Transformation

Single-label classification:

- Analyse a document, make a classification.

Multi-label classification:

- Analyse a document, . . . ?
- Solution 1.: Make several (single-label) decisions
- Solution 2.: Make one (single-label) decision involving multiple labels
- This involves: Transforming a multi-label problem into one or more single-label problems (and back again) **i.e. Problem Transformation**.

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or $\neg$relevant (1/0)).

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or $\neg$relevant (1/0)).

Label set: $L = \{Sport, Environment, Science, Politics\}$

Multi-label D_{train}; $(d, S \subseteq L)$
d_1,{$Sports$,$Politics$}
d_2,{$Science$,$Politics$}
d_3,{$Sports$}
d_4,{$Environment$,$Science$}

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or $\neg$relevant (1/0)).

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \{0,1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_1, 1)$	$(d_1, 0)$	$(d_1, 0)$	$(d_1, 1)$
$(d_2, 0)$	$(d_2, 0)$	$(d_2, 1)$	$(d_2, 1)$
$(d_3, 1)$	$(d_3, 0)$	$(d_3, 0)$	$(d_3, 0)$
$(d_4, 0)$	$(d_4, 1)$	$(d_4, 1)$	$(d_4, 0)$

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or $\neg$relevant (1/0)).

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \{0,1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_1, 1)$	$(d_1, 0)$	$(d_1, 0)$	$(d_1, 1)$
$(d_2, 0)$	$(d_2, 0)$	$(d_2, 1)$	$(d_2, 1)$
$(d_3, 1)$	$(d_3, 0)$	$(d_3, 0)$	$(d_3, 0)$
$(d_4, 0)$	$(d_4, 1)$	$(d_4, 1)$	$(d_4, 0)$

d_x = “Revealed: Polluting Impact of Humans on the Oceans”

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or $\neg$relevant (1/0)).

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \{0, 1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_1, 1)$	$(d_1, 0)$	$(d_1, 0)$	$(d_1, 1)$
$(d_2, 0)$	$(d_2, 0)$	$(d_2, 1)$	$(d_2, 1)$
$(d_3, 1)$	$(d_3, 0)$	$(d_3, 0)$	$(d_3, 0)$
$(d_4, 0)$	$(d_4, 1)$	$(d_4, 1)$	$(d_4, 0)$

d_x = “Revealed: Polluting Impact of Humans on the Oceans”

Single-label Test; $(d, l \in \{0, 1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_x, ?)$	$(d_x, ?)$	$(d_x, ?)$	$(d_x, ?)$

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or $\neg$relevant (1/0)).

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \{0,1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_1, 1)$	$(d_1, 0)$	$(d_1, 0)$	$(d_1, 1)$
$(d_2, 0)$	$(d_2, 0)$	$(d_2, 1)$	$(d_2, 1)$
$(d_3, 1)$	$(d_3, 0)$	$(d_3, 0)$	$(d_3, 0)$
$(d_4, 0)$	$(d_4, 1)$	$(d_4, 1)$	$(d_4, 0)$

d_x = "Revealed: Polluting Impact of Humans on the Oceans"

Single-label Test; $(d, l \in \{0,1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_x, 0)$	$(d_x, 1)$	$(d_x, 1)$	$(d_x, 0)$

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or ¬relevant (1/0)).

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \{0, 1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_1, 1)$	$(d_1, 0)$	$(d_1, 0)$	$(d_1, 1)$
$(d_2, 0)$	$(d_2, 0)$	$(d_2, 1)$	$(d_2, 1)$
$(d_3, 1)$	$(d_3, 0)$	$(d_3, 0)$	$(d_3, 0)$
$(d_4, 0)$	$(d_4, 1)$	$(d_4, 1)$	$(d_4, 0)$

d_x = "Revealed: Polluting Impact of Humans on the Oceans"

Multi-label Test; $(d, S \subseteq L)$
d_x,$\{Environment, Science\}$

Solution 1. Binary Method

Several single-label classifiers make several binary decisions (a label is relevant, or $\neg$relevant (1/0)).

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \{0,1\})$			
C_{Sport}	$C_{Envir.}$	$C_{Science}$	$C_{Politics}$
$(d_1, 1)$	$(d_1, 0)$	$(d_1, 0)$	$(d_1, 1)$
$(d_2, 0)$	$(d_2, 0)$	$(d_2, 1)$	$(d_2, 1)$
$(d_3, 1)$	$(d_3, 0)$	$(d_3, 0)$	$(d_3, 0)$
$(d_4, 0)$	$(d_4, 1)$	$(d_4, 1)$	$(d_4, 0)$

d_x = “Revealed: Polluting Impact of Humans on the Oceans”

Multi-label Test; $(d, S \subseteq L)$
d_x,$\{Environment, Science\}$

- Assumes that all labels are independent

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Label set: $L = \{Sport, Environment, Science, Politics\}$

Multi-label D_{train}; $(d, S \subseteq L)$
d_1,$\{Sports$,$Politics\}$
d_2,$\{Science$,$Politics\}$
d_3,$\{Sports\}$
d_4,$\{Environment$,$Science\}$

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \texttt{distinct}(l \in SLD_{train}))$
d_1,$Sports\_Politics$
d_2,$Science\_Politics$
d_3,$Sports$
d_4,$Environment\_Science$

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in$ `distinct`$(l \in SLD_{train}))$
d_1,$Sports\_Politics$
d_2,$Science\_Politics$
d_3,$Sports$
d_4,$Environment\_Science$

d_x = “Revealed: Polluting Impact of Humans on the Oceans”

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \texttt{distinct}(l \in SLD_{train}))$
d_1,$Sports\_Politics$
d_2,$Science\_Politics$
d_3,$Sports$
d_4,$Environment\_Science$

d_x = “Revealed: Polluting Impact of Humans on the Oceans”

Single-label Test $(d, l \in \texttt{distinct}(l \in SLD_{train}))$
d_x,?

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in$ `distinct`$(l \in SLD_{train}))$
d_1,$Sports\_Politics$
d_2,$Science\_Politics$
d_3,$Sports$
d_4,$Environment\_Science$

d_x = “Revealed: Polluting Impact of Humans on the Oceans”

Single-label Test $(d, l \in$ `distinct`$(l \in SLD_{train}))$
d_x,$Environment\_Science$

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \texttt{distinct}(l \in SLD_{train}))$
d_1,$Sports\_Politics$
d_2,$Science\_Politics$
d_3,$Sports$
d_4,$Environment\_Science$

d_x = "Revealed: Polluting Impact of Humans on the Oceans"

Multi-label Test $(d, S \subseteq L)$
d_x,$\{Environment$,$Science\}$

Solution 2. Combination Method

One decision involves multiple labels. Each label combination becomes an atomic label.

Label set: $L = \{Sport, Environment, Science, Politics\}$

Single-label D_{train}; $(d, l \in \texttt{distinct}(l \in SLD_{train}))$
d_1,$Sports\_Politics$
d_2,$Science\_Politics$
d_3,$Sports$
d_4,$Environment\_Science$

d_x = “Revealed: Polluting Impact of Humans on the Oceans”

Multi-label Test $(d, S \subseteq L)$
d_x,$\{Environment$,$Science\}$

- May generate many labels from a few examples
- Can only predict combinations seen in the training set

Initial Conclusions

- The Combination Method does best, because it incorporates information about the relationships between labels, e.g.:
  - label X may only ever occur by itself
  - labels X and Y may occur together often
  - labels X and Y may never occur together
- But, it. . .
  - often generates too many labels
  - becomes overwhelmed by so many labels
- How can we improve?
  - 90% of label combs. only found in 10% of the data
  - concentrate on the key label combinations!

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

Doc.	**Labels** ($S \subseteq L$)
d_1	{Sports,Science}
d_2	{Environment,Science,Politics}
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

Doc.	Labels ($S \subseteq L$)
d_1	{Sports,Science}
d_2	{Environment,Science,Politics}
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

Doc.	Labels ($S \subseteq L$)
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Doc.	Labels ($S \subseteq L$)
d_1	{Sports,Science}
d_2	{Environment,Science,Politics}

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

Doc.	Labels ($S \subseteq L$)
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Doc.	Labels ($S \subseteq L$)
d_1	{Sports,Science}
d_2	{Environment,Science,Politics}

- Lost 20% of data. Can we save any of that data?

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

Doc.	**Labels** ($S \subseteq L$)
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Doc.	**Labels** ($S \subseteq L$)
d_1	{Sports,Science}
d_2	{Environment,Science,Politics}

- Lost 20% of data. Can we save any of that data?
- Yes. By splitting up S into more frequent sub*sub*sets

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

Doc.	Labels ($S \subseteq L$)
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Doc.	Labels ($S \subseteq L$)
d_1	{Sports,Science}
d_1	{Sports}
d_1	{Science}
d_2	{Environment,Science,Politics}
d_2	{Environment,Science}
d_2	{Politics}

- Lost 20% of data. Can we save any of that data?
- Yes. By splitting up S into more frequent sub*sub*sets

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

Doc.	**Labels** ($S \subseteq L$)
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Doc.	**Labels** ($S \subseteq L$)
d_1	{Sports}
d_1	{Science}
d_2	{Environment,Science}
d_2	{Politics}

- Lost 20% of data. Can we save any of that data?
- Yes. By splitting up S into more frequent sub*sub*sets

Pruned Problem Transformation (PPT)

Prune away all examples with infrequent label subsets. e.g. 10 examples, 6 combinations:

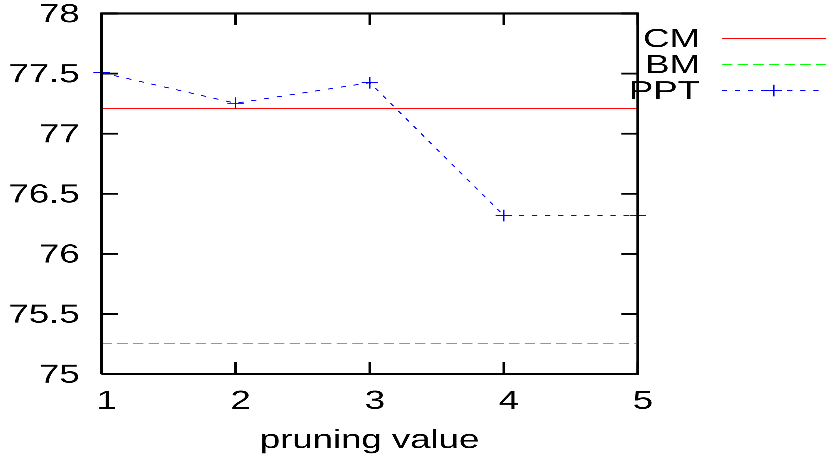
Doc.	**Labels** ($S \subseteq L$)
d_1	{Sports}
d_1	{Science}
d_2	{Environment,Science}
d_2	{Politics}
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

Pruned Problem Transformation (PPT)

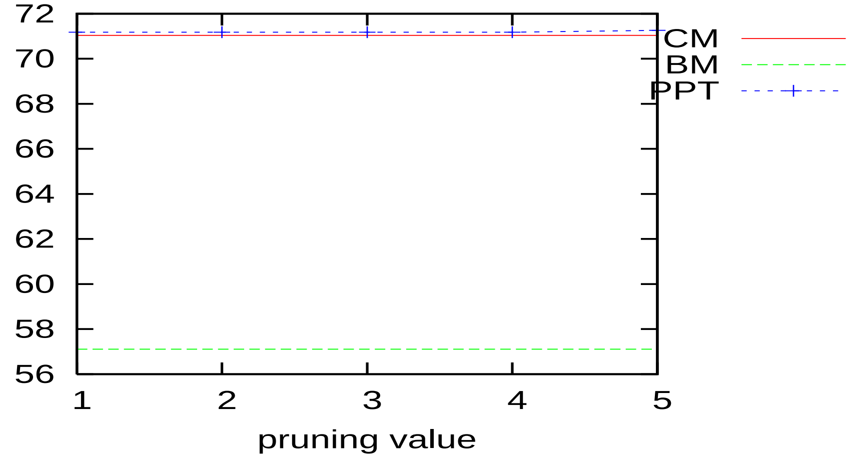
Prune away all examples with infrequent label subsets. e.g. 12 examples, 4 combinations:

Doc.	**Labels** ($S \subseteq L$)
d_1	{Sports}
d_1	{Science}
d_2	{Environment,Science}
d_2	{Politics}
d_3	{Sports}
d_4	{Environment,Science}
d_5	{Science}
d_6	{Sports}
d_7	{Environment,Science}
d_8	{Politics}
d_9	{Politics}
d_{10}	{Science}

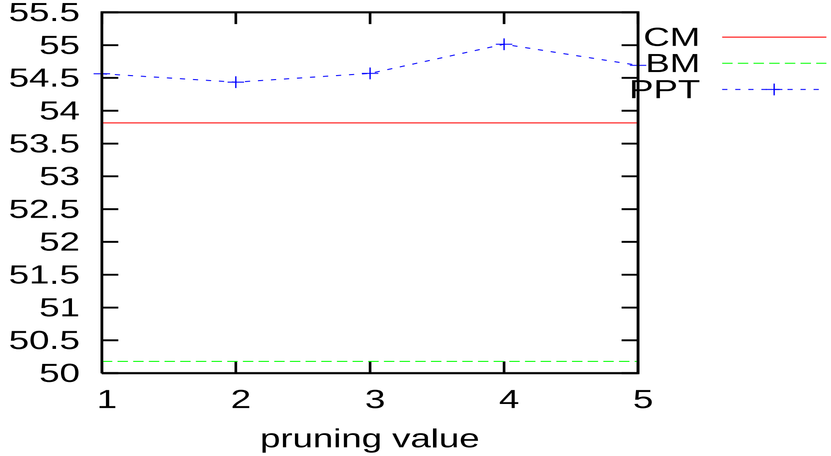
Experiments I. Accuracy.



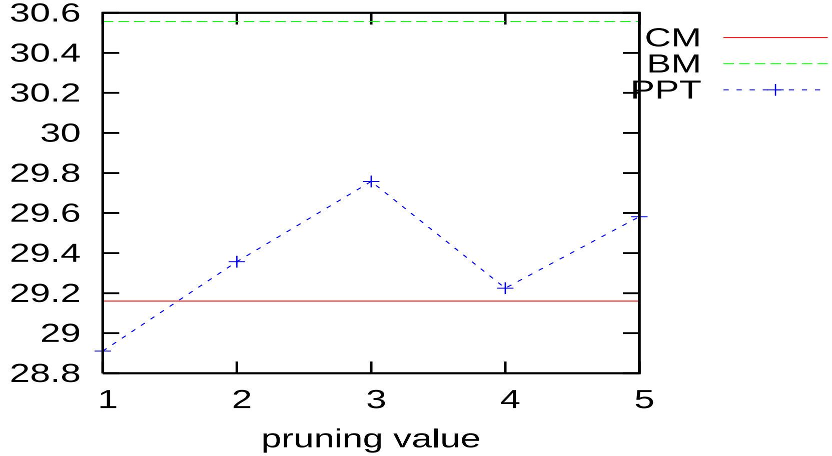
Medical Dataset



Scene Dataset

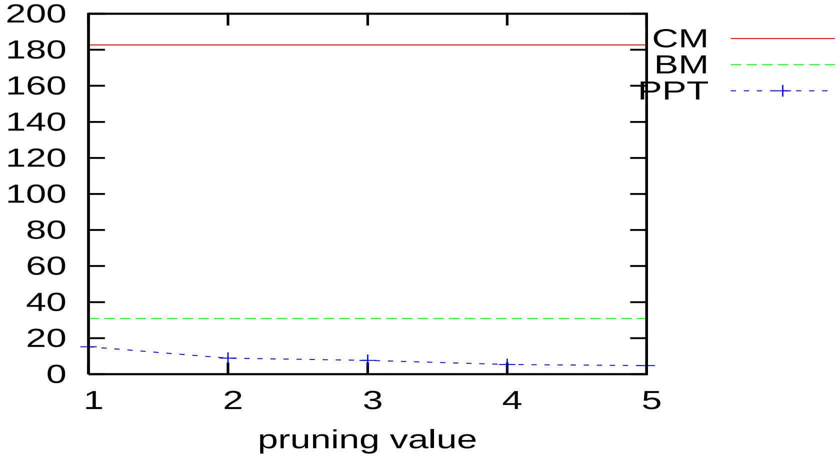


Yeast Dataset

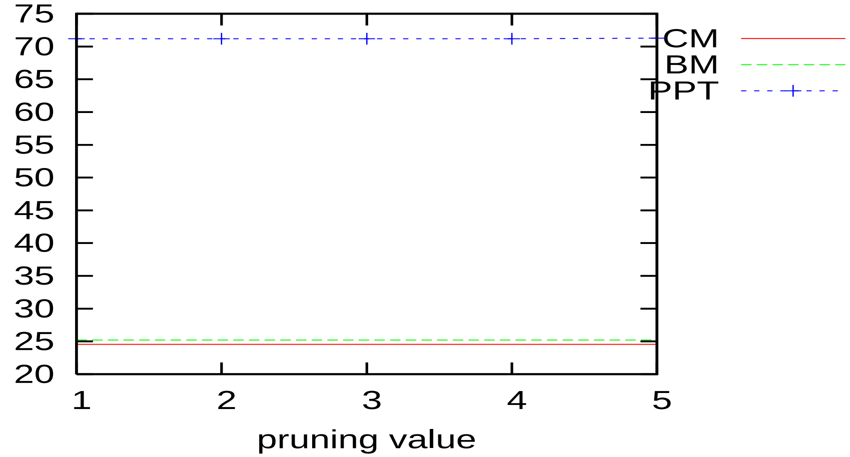


Enron Dataset

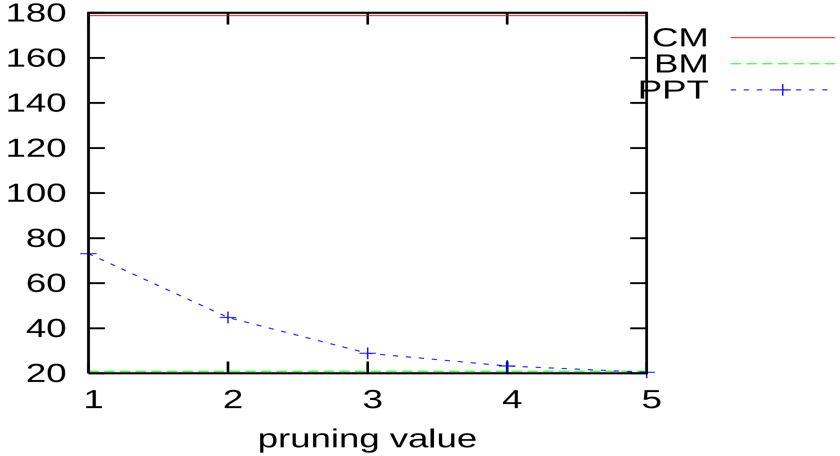
Experiments I. Build Time.



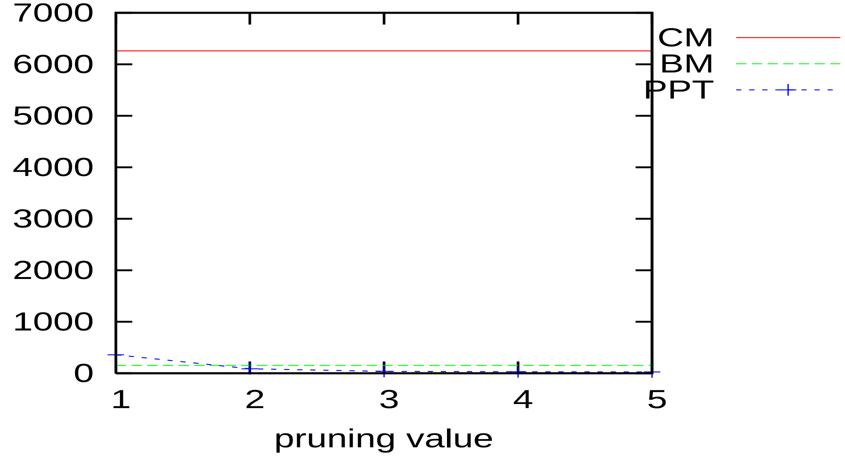
Medical Dataset



Scene Dataset



Yeast Dataset



Enron Dataset

PPT: Initial Conclusions

- Fast
- Superior to BM and CM for some pruning range
- . . . except Enron, where
  - labelling is very irregular (44% as many distinct label combinations as total examples)
  - PPT can't form new combinations
  - the Binary Method can (and *does better* because of this)
- The Binary Method combines several single labels to create a multi-label prediction
- Can we combine multiple labels to create *new multi-label predictions*?

PPT Extended (PPT-ext)

Yes–Given a test example d_x (about $Sports$ and $Science$) ...

- Look at a posterior P*robability* for each possible existing combination:

Combination (S)	$P(S\|d_x)$
$\{Sports, Politics\}$	0.2
$\{Science, Politics\}$	0.2
$\{Sports\}$	0.3
$\{Enviro., Science\}$	0.3

PPT Extended (PPT-ext)

Yes–Given a test example d_x (about $Sports$ and $Science$) ...

- Look at a posterior *Probability* for each possible existing combination:

Combination (S)	$P(S \mid d_x)$
$\{Sports, Politics\}$	0.2
$\{Science, Politics\}$	0.2
$\{Sports\}$	0.3
$\{Enviro., Science\}$	0.3

Label	Score
$Sports$	0.5
$Science$	0.5
$Politics$	0.4
$Enviro.$	0.3

- We can sum these probabilities for each label

PPT Extended (PPT-ext)

Yes–Given a test example d_x (about $Sports$ and $Science$) ...

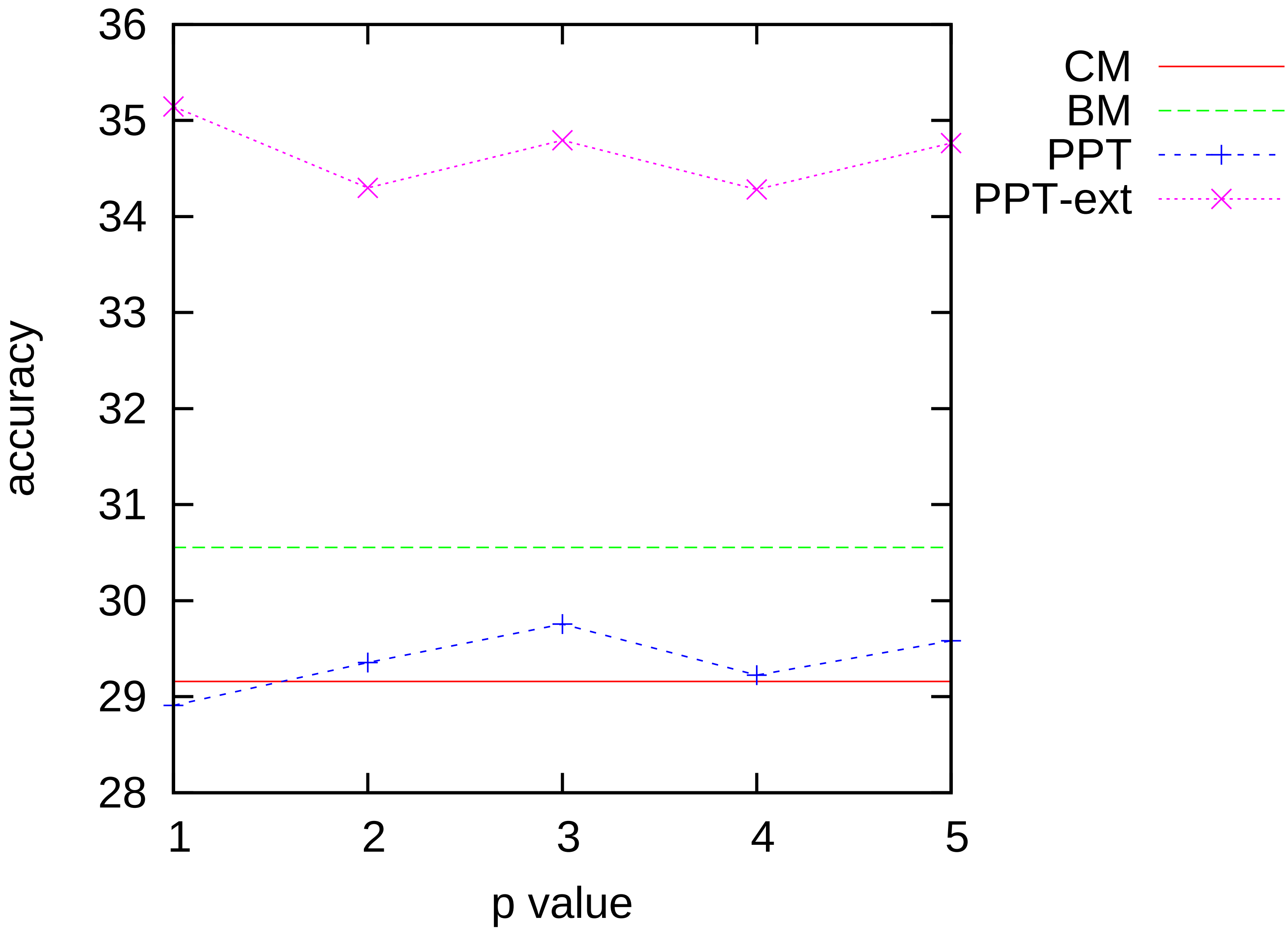
- Look at a posterior *Probability* for each possible existing combination:

Combination (S)	$P(S\|d_x)$
$\{Sports, Politics\}$	0.2
$\{Science, Politics\}$	0.2
$\{Sports\}$	0.3
$\{Enviro., Science\}$	0.3

Label	Score
$Sports$	0.5
$Science$	0.5
$Politics$	0.4
$Enviro.$	0.3

- We can sum these probabilities for each label
- Using a threshold of ≥ 0.5, gives us: $\{Sports, Science\}$

Experiments II. Accuracy



Enron Dataset. Accuracy (no change to build time!)

Summary

- Multi-label Classification via Problem Transformation
- Two standard approaches: CM, and BM
- CM: relationships between labels are important, but too many label combinations causes problems (and can't form new combinations)
- PPT: focus on key relationships
- PPT-ext: able to form new multi-label combinations
- Experiments: PPT and PPT-ext superior to CM and BM

Summary

- Multi-label Classification via Problem Transformation
- Two standard approaches: CM, and BM
- CM: relationships between labels are important, but too many label combinations causes problems (and can't form new combinations)
- PPT: focus on key relationships
- PPT-ext: able to form new multi-label combinations
- Experiments: PPT and PPT-ext superior to CM and BM

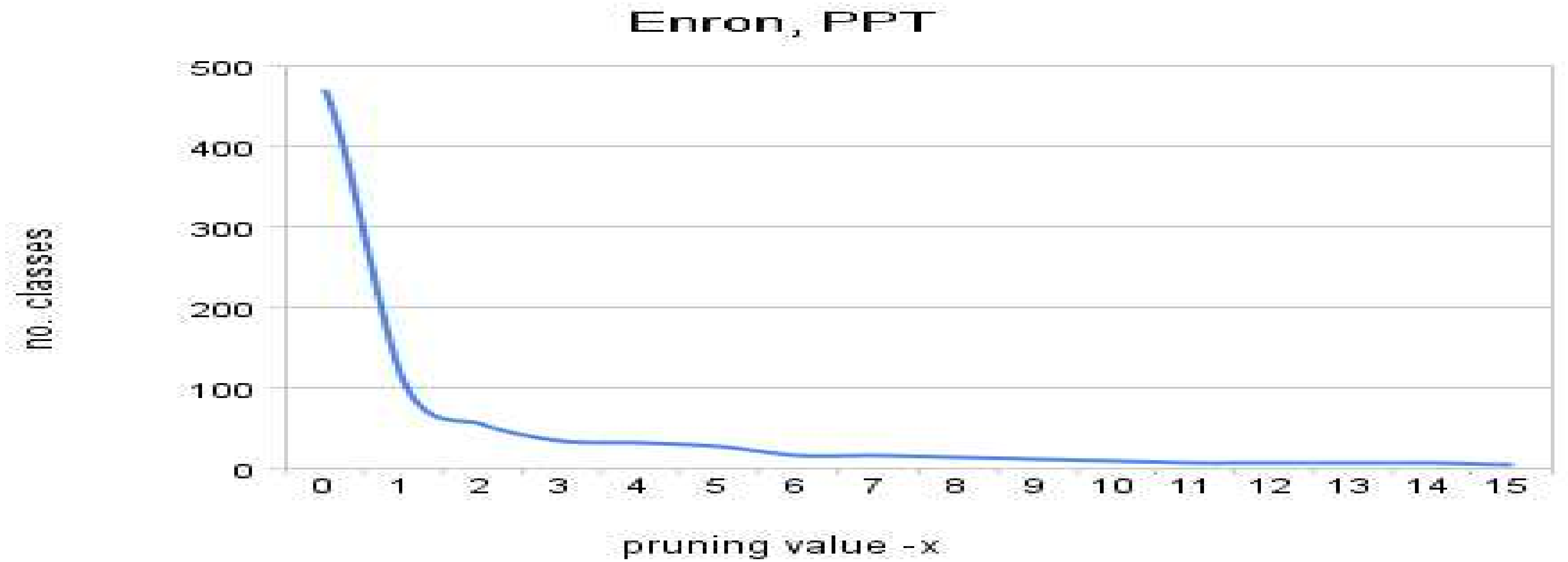
The End. – Questions? / Comments?

Appendix 1. Datasets

	$\|D\|$	$\|L\|$	$LCard(D)$	$PDist(D)$
Medical	978	45	1.25	0.096
Scene	2407	6	1.07	0.006
Yeast	2417	14	4.24	0.082
Enron	1702	53	3.38	0.442

$LCard(D)$ = average size of number of labels per document
$PDist(D)$ = the percentage of documents which are distinct

Appendix 2. Combination Popularity



Appendix 3. Evaluation

- Accuracy: $\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|S_i \cap Y_i|}{|S_i \cup Y_i|}$
- Hamming loss: $\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{Y_i \Delta S_i}{|L|}$ (Δ = symmetrical difference)
- $F1$: $\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{2*p*r}{p+r}$ ($p$recision,$r$ecall of Y_i from S_i)

E.g.:
$Y =$ `0100100010` (predicted)
$S =$ `0100101000` (actual)

Accuracy	$2/4$	0.50	(best = 1.00)
Hamming loss	$2/10$	0.20	(best = 0.00)
F1	$(2 * \frac{2}{3} * \frac{2}{3}/(\frac{2}{3} + \frac{2}{3}))$	0.67	(best = 1.00)

Appendix 4. Experiments III

	Medical		Enron	
	RAKEL	PPT	RAKEL	PPT
$F1$	0.776	0.789	0.457	0.503
Ham. Loss	0.012	0.011	0.067	0.074
Accuracy	0.743	0.776	0.323	0.353
McNemar's	$p = 0.295$		$p = 0.000$	
Build Time	190s	15s	3163s	195s
RAKEL par.	$m = 20, k = 27, t = .5$		$m = 80, k = 21, t = .5$	
PPT par.	$p = 1, -N_A$		$p = 5, -N_A, -J, t = .21$	

Appendix 5. Graph View

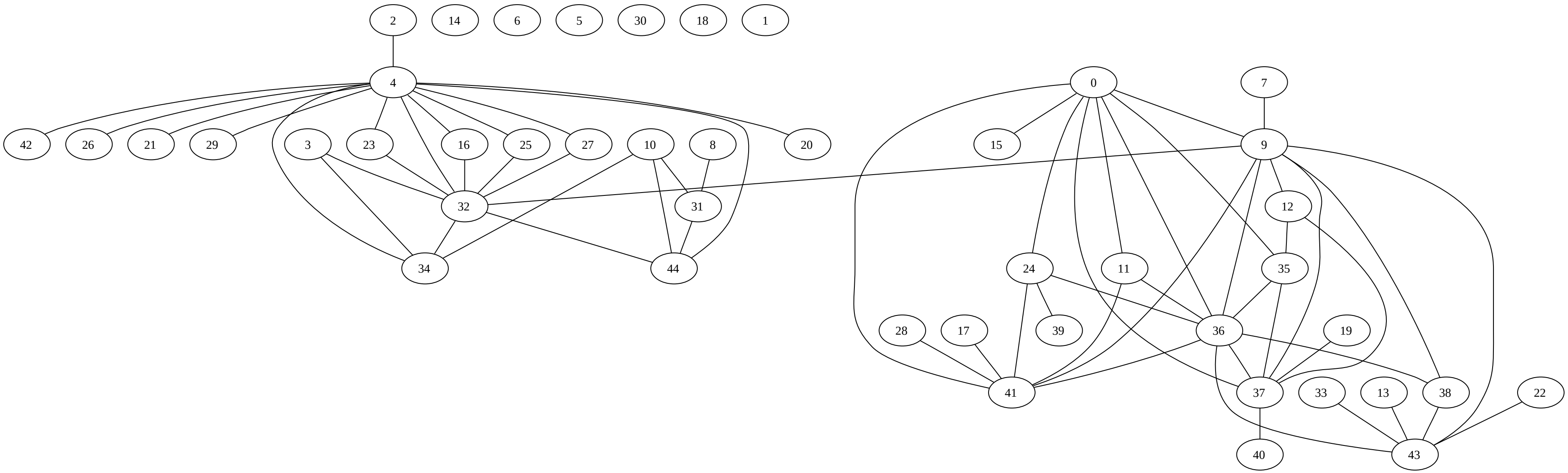


Figure 1: A multi-label dataset. Each node is a label. Each edge represents at least 1 co-occurrence of the two labels it connects

What if we ignore very infrequent co-occurrences between labels?

Appendix 5. Graph View

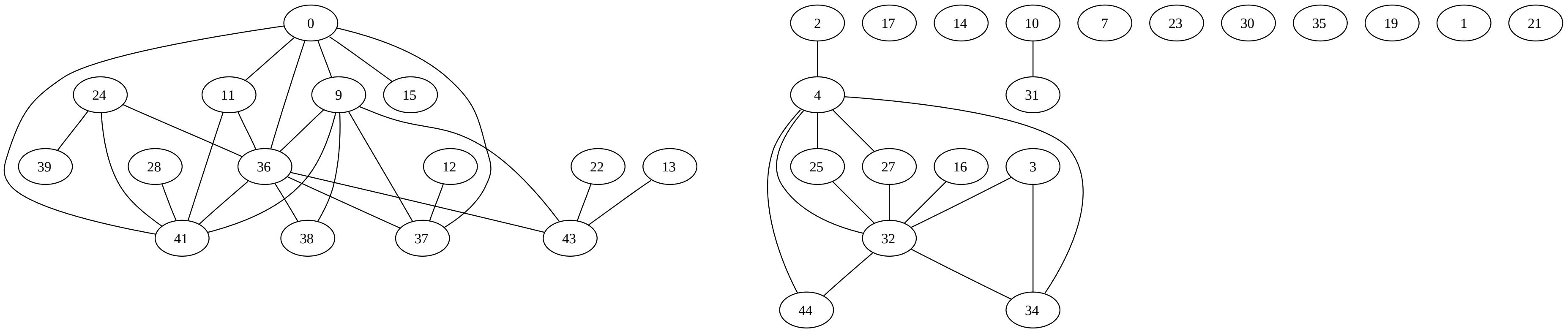


Figure 2: A multi-label dataset. Each node is a label. Each edge represents at least 2 co-occurrences of the two labels it connects (covers 97% of 978 examples)

Appendix 5. Graph View

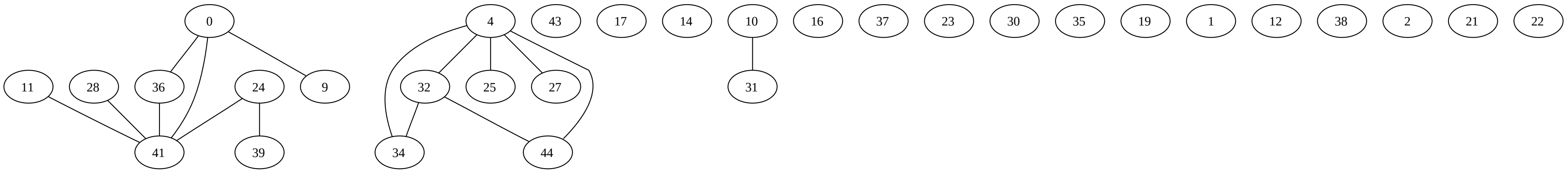


Figure 3: A multi-label dataset. Each node is a label. Each edge represents at least 3 co-occurrences of the two labels it connects (covers 92% of 978 examples)

These are the key label relationships.