Methods for On-line Multi-label Classification

Six Months Progress

Jesse Read
(Supervisors: Bernhard Pfahringer, Geoff Holmes)

Start date: March 2007

Outline

1. Introduction
   - Multi-label Classification
   - The Pruned Sets Method (PS)
2. Ensembles of Pruned Sets (EPS)
3. Classification of On-line Multi-label Data
4. A New Method for Multi-label Ranking
5. Future Direction

Introduction

- Multi-label Classification
  - Assigning multiple labels (classes) to instances
  - labels are selected from a *predefined* set
  - instances can represent text, media, biological data, etc ...
- Example Applications
  - a news article can be about `Science` and `Technology`
  - a film can be labeled `Romance` and `Comedy`
  - 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 {`Science`,`Environment`} vs {`Sport`,`Environment`}
  - computational complexity

Problem Setting

Multi-label Classification

- A set of predefined labels: $L = \{l_0, l_1, \cdots, l_n\}$
- A set of instances: $D = \{x_0, x_1, \cdots, x_m\}$

Problem Setting

Multi-label Classification

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

Problem Setting

Multi-label Classification

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

Problem Setting

Multi-label Classification

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

Problem Transformation

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

Problem Setting

Multi-label Classification

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

Problem Transformation

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

Combination Method (CM)

Each label subset $S \subseteq L$ can be treated as single label, thus forming a single-label problem.

Background: The Combination Method (CM)

Example (CM Method)

$L = \{Anim, Family, Comedy, Musical\}$

D	$S \subseteq L$ (Multi-label)
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_3	$\{Anim, Comedy, Family\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_9	$\{Musical, Anim\}$

Background: The Combination Method (CM)

Example (CM Method)

$L' = \{\{Anim, Comedy\}, \{Anim, Family\}, \{Musical\}, \{Anim, Comedy, Family\}, \{Musical, Anim\}\}$

D	$l \subseteq L'$ (Single-label)
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_3	$\{Anim, Comedy, Family\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_9	$\{Musical, Anim\}$

- Each set is a label

Background: The Combination Method (CM)

Example (CM Method)

$L' = \{\{Anim, Comedy\}, \{Anim, Family\}, \{Musical\}, \{Anim, Comedy, Family\}, \{Musical, Anim\}\}$

D	$l \subseteq L'$ (Single-label)
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_3	$\{Anim, Comedy, Family\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_9	$\{Musical, Anim\}$

- Each set is a label
  - creates many possible labels
  - cannot predict new combinations

Previous work: The `Pruned Sets` (PS) Method[1]

Pruned Sets Method (PS)

Infrequently occurring label sets are pruned and decomposed into label subsets which *are* frequent.

[1] Read. *A Pruned Problem Transformation Method*. In Proc. of NZCSRSC'08

Previous work: The Pruned Sets (PS) Method[2]

Example (PS Method)

D	$I \subseteq L'$
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_3	$\{Anim, Comedy, Family\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_9	$\{Musical, Anim\}$

10 examples 5 combinations

1. Prune examples (e.g. where occurrences ≤ 1))

[2] Read. *A Pruned Problem Transformation Method*. In Proc. of NZCSRSC'08

Previous work: The Pruned Sets (PS) Method[2]

Example (PS Method)

D	$I \subseteq L'$
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_3	$\{Anim, Comedy, Family\}$
x_9	$\{Musical, Anim\}$

10 examples 5 combinations

1. Prune examples (e.g. where occurrences ≤ 1))
2. Decompose infrequent label sets into *frequent* sets

[2] Read. *A Pruned Problem Transformation Method*. In Proc. of NZCSRSC'08

Previous work: The Pruned Sets (PS) Method[2]

Example (PS Method)

D	$l \subseteq L'$
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_3	$\{Anim, Comedy\}$
x_3	$\{Anim, Family\}$
x_9	$\{Musical\}$

10 examples 5 combinations

1. Prune examples (e.g. where occurrences ≤ 1))
2. Decompose infrequent label sets into *frequent* sets
3. Reintroduce instances with new label sets

[2] Read. *A Pruned Problem Transformation Method*. In Proc. of NZCSRSC'08

Previous work: The Pruned Sets (PS) Method[2]

Example (PS Method)

D	$l \subseteq L'$
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_3	$\{Anim, Comedy\}$
x_3	$\{Anim, Family\}$
x_9	$\{Musical\}$

11 examples 3 combinations

1. Prune examples (e.g. where occurrences ≤ 1))
2. Decompose infrequent label sets into *frequent* sets
3. Reintroduce instances with new label sets

- More examples, fewer labels

[2] Read. *A Pruned Problem Transformation Method*. In Proc. of NZCSRSC'08

Previous work: The Pruned Sets (PS) Method[2]

Example (PS Method)

D	$l \subseteq L'$
x_0	$\{Anim, Family\}$
x_1	$\{Anim, Comedy\}$
x_2	$\{Anim, Comedy\}$
x_4	$\{Musical\}$
x_5	$\{Musical\}$
x_6	$\{Anim, Comedy\}$
x_7	$\{Anim, Family\}$
x_8	$\{Musical\}$
x_3	$\{Anim, Comedy\}$
x_3	$\{Anim, Family\}$
x_9	$\{Musical\}$

11 examples 3 combinations

1. Prune examples (e.g. where occurrences ≤ 1))
2. Decompose infrequent label sets into *frequent* sets
3. Reintroduce instances with new label sets

- More examples, fewer labels
- Cannot form new combinations

[2]Read. *A Pruned Problem Transformation Method*. In Proc. of NZCSRSC'08

Outline

1. Introduction
   - Multi-label Classification
   - The Pruned Sets Method (PS)
2. Ensembles of Pruned Sets (EPS)
3. Classification of On-line Multi-label Data
4. A New Method for Multi-label Ranking
5. Future Direction

Ensembles of Pruned Sets (EPS)[3]

Ensembles of Pruned Sets (EPS):

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

[3] Read, Pfahringer, Holmes. *Multi-label Classification with Ensembles of Pruned Sets*. To appear in Proc. of ICDM 2008

Ensembles of Pruned Sets (EPS)[3]

Ensembles of Pruned Sets (EPS):

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

Example (predictions for a test instance)

Ensemble: Predictions:	PS_0 $\{M\}$	PS_1 $\{A, F\}$	PS_2 $\{A, C\}$	PS_3 $\{A, F\}$	PS_4 $\{M\}$	PS_5 $\{M\}$
All Pred.:	$\{A_3, M_3, F_2, C_1\}$					
Final Pred.:	$\{A, M, F\}$ (> 1)					

[3] Read, Pfahringer, Holmes. *Multi-label Classification with Ensembles of Pruned Sets*. To appear in Proc. of ICDM 2008

Ensembles of Pruned Sets (EPS)[3]

Ensembles of Pruned Sets (EPS):

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

Example (predictions for a test instance)

Ensemble:	PS_0	PS_1	PS_2	PS_3	PS_4	PS_5
Predictions:	$\{M\}$	$\{A, F\}$	$\{A, C\}$	$\{A, F\}$	$\{M\}$	$\{M\}$
All Pred.:	$\{A_3, M_3, F_2, C_1\}$					
Final Pred.:	$\{A, M, F\}$ (> 1)					

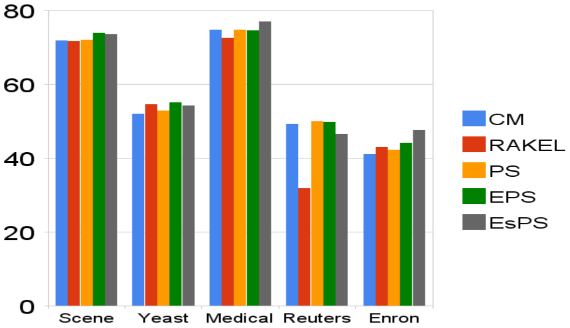
Variations of EPS

- A traditional Bagging scheme
- EsPS: Each PS model trains using a *label* subset

[3]Read, Pfahringer, Holmes. *Multi-label Classification with Ensembles of Pruned Sets*. To appear in Proc. of ICDM 2008

Ensembles of Pruned Sets (EPS)[4]

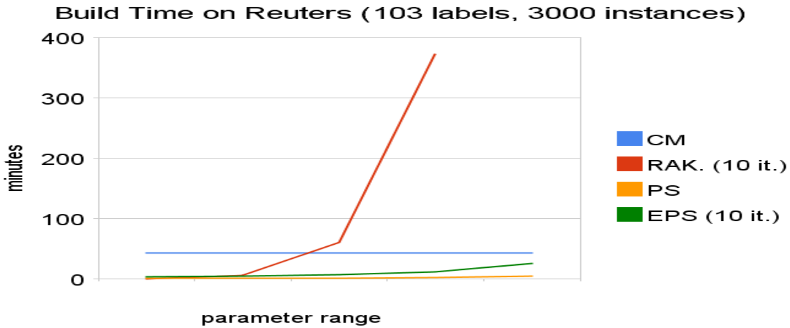
Accuracy on a collection of multi-label datasets



- EPS always statistically similar or better than CM and RAKEL
- Is PS/EPS worth the effort over other methods?

[4] Read, Pfahringer, Holmes. *Multi-label Classification with Ensembles of Pruned Sets*. To appear in Proc. of ICDM 2008

Ensembles of Pruned Sets (EPS)[5]



- More efficient than CM/RAKEL

[5] Read, Pfahringer, Holmes. *Multi-label Classification with Ensembles of Pruned Sets*. To appear in Proc. of ICDM 2008

Outline

On-line Multi-label Classification

On-line data:

- New instances constantly incoming
- Limited processing for each instance
- Concept drift

Goals of on-line algorithms:

Efficiency Learn from new examples quickly and efficiently

Adaptivity Gracefully handle concept drift

Accuracy Strive to maintain a low error rate

- In the multi-label case, it is important to take into account label relationships
- This is relevant to both *Efficiency* and *Adaptivity*

On-line Multi-label Classification

Incremental learning using update-able classifiers

- *Problem Transformation* approaches (e.g. `PS`) can use any single-label classifier
- There already exist update-able single-label classifiers (e.g. `Naive Bayes`)
- But when treating label combinations as single labels (e.g. `CM`, `PS`):
  - incoming instances bring new combinations
  - the label set L' changes over time
  - *PS* must be either rebuilt or reset (e.g. every n instances)

On-line Multi-label Classification

Incremental learning using update-able classifiers

- *Problem Transformation* approaches (e.g. `PS`) can use any single-label classifier
- There already exist update-able single-label classifiers (e.g. `Naive Bayes`)
- But when treating label combinations as single labels (e.g. `CM`, `PS`):
  - incoming instances bring new combinations
  - the label set L' changes over time
  - *PS* must be either rebuilt or reset (e.g. every n instances)
    - *rebuilt*: non-incremental (slow)
    - *reset*: data loss
    - for which n?
  - An ensemble (i.e. `EPS`) can mitigate these issues but limitations remain

o−EPS: On-line multi-label classification

1. Initialise each PS using a random selection of single-labels including an $\emptyset$ label as initial "combinations"

On-line Multi-label Classification

o–EPS: On-line multi-label classification

1. Initialise each PS using a random selection of single-labels including an $\emptyset$ label as initial "combinations"

Example (o–EPS Initialisation (where $L = \{A, C, F, M, X\}$))

Ens.:	PS_0	PS_1	PS_2	PS_3	$\cdots$
init.:	$\{A, C, \emptyset\}$	$\{M, F, \emptyset\}$	$\{X, M, \emptyset\}$	$\{A, X, \emptyset\}$	$\cdots$

On-line Multi-label Classification

o−EPS: On-line multi-label classification

1. Initialise each PS using a random selection of single-labels including an $\emptyset$ label as initial "combinations"
2. Decompose the label set of every incoming instance PS-style and add copies to *relevant* PS models
   - e.g. $(x_0, \{A, X\}) \rightarrow (x_0, A), (x_0, X), (x_0, \emptyset)$

Example (o−EPS Initialisation (where $L = \{A, C, F, M, X\}$))

Ens.:	PS_0	PS_1	PS_2	PS_3	$\cdots$
init.:	$\{A, C, \emptyset\}$	$\{M, F, \emptyset\}$	$\{X, M, \emptyset\}$	$\{A, X, \emptyset\}$	$\cdots$

On-line Multi-label Classification

o−EPS: On-line multi-label classification

1. Initialise each PS using a random selection of single-labels including an $\emptyset$ label as initial "combinations"
2. Decompose the label set of every incoming instance PS-style and add copies to *relevant* PS models
   - e.g. $(x_0, \{A, X\}) \rightarrow (x_0, A), (x_0, X), (x_0, \emptyset)$
3. Whenever significant change in the data is detected, reset one PS model with freq. combinations as labels, and continue...

Example (o−EPS Initialisation (where $L = \{A, C, F, M, X\}$))

Ens.:	PS_0	PS_1	PS_2	PS_3	$\cdots$
init.:	$\{\{A, X\}, F, M, \emptyset\}$	$\{M, F, \emptyset\}$	$\{X, M, \emptyset\}$	$\{A, X, \emptyset\}$	$\cdots$

On-line Multi-label Classification

o−EPS: On-line multi-label classification

1. Initialise each PS using a random selection of single-labels including an $\emptyset$ label as initial "combinations"
2. Decompose the label set of every incoming instance PS-style and add copies to *relevant* PS models
   - e.g. $(x_0, \{A, X\}) \rightarrow (x_0, A), (x_0, X), (x_0, \emptyset)$
3. Whenever significant change in the data is detected, reset one PS model with freq. combinations as labels, and continue...
   - ADWIN (A. Bifet, R. Gavald. 2007): detects change in a sequence of numbers

Example (o−EPS Initialisation (where $L = \{A, C, F, M, X\}$))

Ens.:	PS_0	PS_1	PS_2	PS_3	$\cdots$
init.:	$\{\{A, X\}, F, M, \emptyset\}$	$\{M, F, \emptyset\}$	$\{X, M, \emptyset\}$	$\{A, X, \emptyset\}$	$\cdots$

On-line Multi-label Classification

o−EPS: On-line multi-label classification

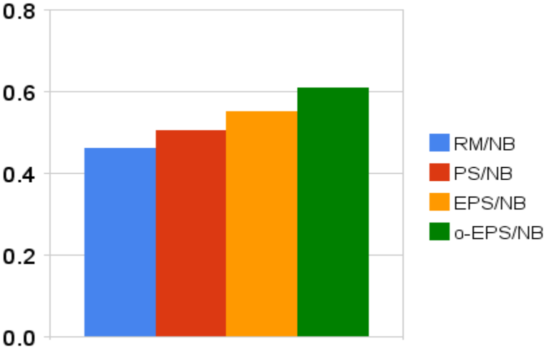
1. Initialise each PS using a random selection of single-labels including an $\emptyset$ label as initial "combinations"
2. Decompose the label set of every incoming instance PS-style and add copies to *relevant* PS models
   - e.g. $(x_0, \{A, X\}) \rightarrow (x_0, A), (x_0, X), (x_0, \emptyset)$
3. Whenever significant change in the data is detected, reset one PS model with freq. combinations as labels, and continue...
   - ADWIN (A. Bifet, R. Gavald. 2007): detects change in a sequence of numbers
   - can use the no. of *freq. combinations* or *closed freq. itemsets*

Example (o−EPS Initialisation (where $L = \{A, C, F, M, X\}$))

Ens.:	PS_0	PS_1	PS_2	PS_3	···
init.:	$\{\{A, X\}, F, M, \emptyset\}$	$\{M, F, \emptyset\}$	$\{X, M, \emptyset\}$	$\{A, X, \emptyset\}$	···

On-line Multi-label Classification

AU(PRC) for *News* dataset



- o−EPS takes into account label combinations; is incremental
- (RM is a method which considers each label independently)
- A work in progress

Outline

A New Algorithm for Multi-label Ranking

```
class Label {

  int index               // index in L
  Instance instance       // a test instance

  int compareTo(other) {
    // get or build binary classifier
    c = classifiers[this.index][other.index]
    if(c == null)
       c.buildBinaryClassifier(this.index,other.index);
    if(c.classifiy(instance) == 0.0)
       return -1
    else if(c.classifiy(instance) == 1.0)
       return 1
  }
}
classifiers = new classifiers[|L|][|L|]
```

A New Algorithm for Multi-label Ranking

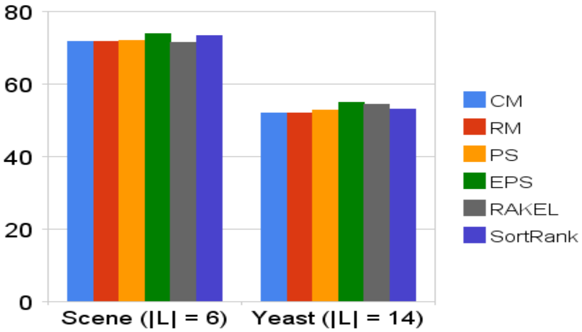
Ranking the label relevance for an instance '`instance`'

```
Label labels[] = new Label[n] // n = |L|
for (i = 0; i < n; i++)
   labels[i] = new Label(i,instance)

Utils.sort(labels)
```

- The sorted array represents the label ranking
- Requires $\frac{|L|(|L|-1)}{2}$ classifiers
- "zero" build time
  - Initially classification is slow
  - Very rapid once all classifiers are build
- Guaranteed complexity (same as the sorting algorithm used)
- Can't threshold (yet) for multi-label classification

A New Algorithm for Multi-label Ranking



- let a = average label set size in the training set
  - Scene = 1.07, Yeast = 4.24
- classifications using the top `math.round`(a) labels
- doesn't work well for large of $|L|$

Outline

Current and Future Direction

The main focus of future work is on the on-line context

- Methods for large and complex multi-label datasets (large D and large L with complex label relationships)
- Reducing the computational complexity of label-combination approaches by using hierarchies, etc ...
- Develop classification methods for `SortRank`
- Further development of `o-EPS`
- Developing multi-label Hoeffding trees

Thanks.