

Multi Instance Multi Label

Lorenzo Niccolai

`lorenzo.niccolai3@stud.unifi.it`

Fabio Vittorini

`fabio.vittorini@stud.unifi.it`

Machine Learning

University of Florence, Department of Information Engineering

July 17, 2017



- 1 SVM
 - Classification
 - SVM
- 2 Multi Instance Learning
 - SIL
 - mi-SVM
 - MI-SVM
- 3 Multi Label Learning
- 4 Multi Instance Multi Label Learning
- 5 Our work
 - Results

Classification problem

- Goal:** To produce a classifier able to decide whether an object belongs to one or more classes.
- Idea:** Supervised Learning: Given a dataset of already classified examples, the classifier *learns* a function that solves classification problem.

Dataset is set of classified examples:

$$D = \{(X_i, Y_i) \mid i \in [1, n]\}$$

$$X_i \in \mathbb{R}^f$$

$$Y_i \in \{-1, +1\}$$

- A vector $x \in \mathbb{R}^f$ represents an object using f *relevant* features
- A number $y \in \{-1, +1\}$ indicates whether the example belongs to target class

While learning the target function, the dataset is divided in *training set* and *test set*.

- The idea is to compute the *maximum-margin hyperplane* $w^T x + b$ which best separates positive examples from negative examples
- Samples on the margin are called the *support vectors*

Optimization problem is:

$$\operatorname{argmin}_w \frac{1}{2} \|w\|^2$$

$$Y_i(w^T X_i + b) \geq 1 \quad \forall i \in [1, n]$$

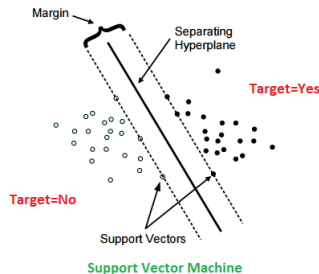


Figure 1: Solution of maximum-margin hyperplane

SVM with slacks

- The examples may not be linearly separable and so the problem would not have any solutions because constraints are not satisfied. Then we introduce slack variables ξ

Optimization problem becomes:

$$\operatorname{argmin}_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i \in [1, n]$$

$$\xi_i \geq 0 \quad \forall i \in [1, n]$$

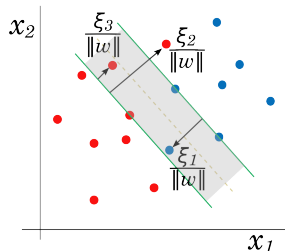


Figure 2: Solution with slacks

Multi instance classification

Motivation:

- Sometimes a complex item can be well represented by a set of *instances*
- A single instance may belong or not to a class or *label* (positive or negative)
- An example, or *bag*, is positive if at least one of its instances is positive, where as a negative bag consists of only negative instances
- A label is provided for the entire bag, not to instances
- We have a *semi-supervised learning* problem

[1]

Notation

Dataset is now a set of bags, where each bag is a set of instances:

$$D = \{(X_i, Y_i) \mid i \in [1, n]\}$$

$$X_i = \{x_{i,k} \mid k \in [1, k_i], x_{i,k} \in \mathbb{R}^f\}$$

Notice that each bag can be made of any number of instances, but every instance has a fixed number of features f .

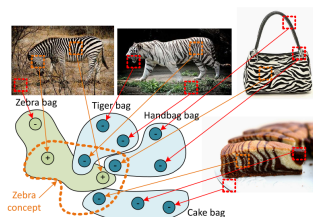


Figure 3: *Example of images instances refer to "zebra concept"*

The first naive approach makes the following label assignment:

- If an instance belongs to a negative bag, sets its label to -1
- If an instance belongs to a positive bag, sets its label to $+1$

The resulting problem can be solved using a regular SVM, treating each instance as a whole document.

Using this approach makes almost useless multi-instance formulation.

Instances label assignment:

- If an instance belongs to a negative bag we can say that its label is -1
- If an instance belongs to a positive bag we don't know for sure its label

This leads to 2 new constraints in SVM problem:

$$y_{i,k} = -1 \text{ if } Y_i = -1$$

$$\sum_{k=1}^{k_i} \frac{y_{i,k} + 1}{2} \geq 1 \text{ if } Y_i = +1$$

Our SVM problem becomes the following:

$$\min_Y \min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

$$y_{i,k}(w^T x_{i,k} + b) \geq 1 - \xi_i \quad \forall i \in [1, n], k \in [1, k_i]$$

$$\xi_i \geq 0 \quad \forall i \in [1, n]$$

$$y_{i,k} = -1 \text{ if } Y_i = -1$$

$$\sum_{k=1}^{k_i} \frac{y_{i,k} + 1}{2} \geq 1 \text{ if } Y_i = +1$$

That is an intractable mixed optimization problem

A feasible algorithm that finds a non optimal solution is the following:

MI-SVM(X, Y)

```
1   $y_{i,k} = -1$  if  $Y_i = -1$ 
2   $y_{i,k} = +1$  if  $Y_i = +1$ 
3  do
4      Solve regular SVM finding  $w, b$ 
5       $y_{i,k} = \text{sign}(w^T x_{i,k} + b)$  if  $Y_i = +1$ 
6      Adjust each positive bag to satisfy constraints
7  while ( $y_{i,k}$  change)
```

This approach uses directly the dataset in its bag form:

$$\operatorname{argmin}_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

$$y_i(\max_k w^T x_{i,k} + b) \geq 1 - \xi_i \quad \forall i \text{ s.t. } Y_i = +1$$

$$y_i(w^T x_{i,k} + b) \geq 1 - \xi_i \quad \forall i \text{ s.t. } Y_i = -1$$

$$\xi_i \geq 0 \quad \forall i \in [1, n]$$

This is possible by selecting a *witness* from each positive bag.

A feasible algorithm that finds a solution is the following:

MI-SVM(X, Y)

- 1 $\bar{x}_i = \text{avg}(x_{i,k}) \forall x_{i,k} \in X_i$ positive bag
- 2 **do**
- 3 Assign $\bar{\alpha}_i \in [0, C]$ to each \bar{x}_i
- 4 Assign $\alpha_{i,j}$ with $\sum_{j=1}^{k_i} \alpha_{i,j} \in [0, C] \forall x_{i,k} \in X_i$ negative bag
- 5 Solve regular SVM finding w, b
- 6 Find new \bar{x}_i by selecting the best one for each positive bag
- 7 **while** (witnesses change)

Multi label classification

Motivation:

- Sometimes a complex item can be well represented by a set of *labels*
- Helps single label classification when the concept is more complicated or general

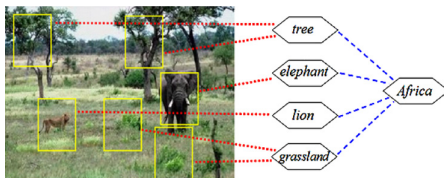


Figure 4: *Multi label example*

A set of labels $L = \{y_1, y_2, \dots, y_l\}$ is given.

Each object contained in the dataset is associated with a set of labels:

$$D = \{(X_i, Y_i) | i \in [1, n]\}$$

$$X_i \in \mathbb{R}^f$$

$$Y_i = \{y_{i,h} | h \in [1, h_i], y_{i,h} \in L, h_i \leq l\}$$

Solutions [4]:

- Problem transformation
- Algorithm adaptation

Problem transformation

Attempt to convert the multilabel problem in a regular binary task.

Two lossy methods:

- Randomly discard each label information except one from each instance
- Remove instances that have actually more than one label

Other solutions:

- Train a binary classifier for each existing combination of labels
- Train a binary classifier for each label (used in this work)

An algorithm adaptation approach

The idea is to focus on ranking rather than binary classification [2]

- Train a classifier $f_l : X \rightarrow \mathbb{R}$ for each label
- For each test example sort label list according to predicted rank
- Set size prediction feeding dataset and thresholds $t(X_i)$ to a classifier

$$t(X_i) = \operatorname{argmin}_t \{k \in Y_i \text{ s.t. } f_k(X_i) \leq t\} + \{k \in \bar{Y}_i \text{ s.t. } f_k(X_i) \geq t\}$$

- Take the best labels according to set size prediction

Introduction to MIML

MIML problems combine motivations of multi instance and multi label ones.

Given a set of labels $L = \{y_1, y_2, \dots, y_l\}$

$$D = \{(X_i, Y_i) | i \in [1, n]\}$$

$$X_i = \{x_{i,k} | k \in [1, k_i], x_{i,k} \in \mathbb{R}^f\}$$

$$Y_i = \{y_{i,h} | h \in [1, h_i], y_{i,h} \in L, h_i \leq l\}$$

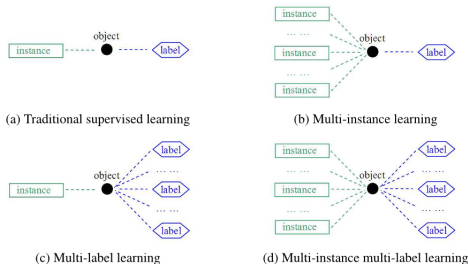


Figure 5: *Different learning frameworks*

SVM Solution

To allow regular SVMs to solve this problem, we use *problem transformation*.

There are 2 possibilities:

- MIML \rightarrow MISL \rightarrow SISL (used in this work)
- MIML \rightarrow SIML \rightarrow SISL

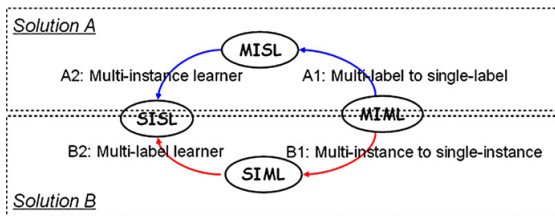


Figure 6: Two possible solutions to implement MIML

Multi label to single label

Excluding lossy approaches, the idea is to train a multi-instance (single label) classifier for each label.

Given a MIML dataset $D = \{(X_i, Y_i) | i \in [1, n]\}$ we produce l datasets as follows:

$$D_{y_j} = \{(X_i, Y_{y_j}) | i \in [1, n]\} \quad \forall j \in [1, l]$$

Where

$$Y_{y_j} = \begin{cases} +1 & \text{if } y_j \in Y_i \\ -1 & \text{otherwise} \end{cases}$$

Then we train L regular multi-instance SVMs and collect their results.

Multi instance to single instance

Given one of MISL datasets produced at previous step, we compared the 3 methods previously exposed:

- SIL
- MI-SVM
- mi-SVM

They all use a standard SISL SVM as subroutine.

The aim of our work is to replicate a part of the results of [5] using the **MIML framework** and compare the different metrics.

- We focused on *text categorization* using text documents (*bags*) belonging to categories (*labels*)
- We choose to use multi-instance learning as the bridge between MIML and SISL
- Bag of words approach to REUTERS-21578 dataset
- Multi instance tasks solved with SIL, MI-SVM and mi-SVM approaches

Documents selection:

- 1 Removed every document with 0 labels
- 2 Removed short documents (less than 30 words)
- 3 Removed randomly documents with 1 label to obtain 2000 examples

Dictionary creation:

- 1 Performed stemming
- 2 Removed stopwords
- 3 Removed rare words keeping 2% of them (about 210)

Multi instance data

- 1 Splitted documents in passages of 50 words max
- 2 Removed empty instances (according to dictionary)

Evaluation criteria

Four criteria are used for performance evaluation:

- **hamming loss:**

$$hloss_S(h) = \frac{1}{p} \sum_{i=1}^p \frac{1}{|\mathcal{Y}|} |h(X_i) \Delta Y_i|$$

- **one-error:**

$$one - error_S(h) = \frac{1}{p} \sum_{i=1}^p [[\arg \max_{y \in \mathcal{Y}} h(X_i, y)] \notin Y_i]$$

- **coverage:**

$$coverages_S(h) = \frac{1}{p} \sum_{i=1}^p \max_{y \in Y_i} rank^h(X_i, y) - 1$$

- **ranking loss:**

$$rloss_S(h) = \frac{1}{p} \sum_{i=1}^p \frac{1}{|Y_i| \cdot |\bar{Y}_i|} |\{(y_1, y_2) \in Y_i \times \bar{Y}_i \text{ s.t. } h(X_i, y_1) \leq h(X_i, y_2)\}|$$

Other metrics used by reference article:

- **average precision:**

$$avgprec_S(h) = \frac{1}{p} \sum_{i=1}^p \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{|\{y' \mid rank^h(X_i, y') \leq rank^h(X_i, y), y' \in Y_i\}|}{rank^h(X_i, y)}$$

- **average recall:**

$$avgrec_S(h) = \frac{1}{p} \sum_{i=1}^p \frac{|\{y \mid rank^h(X_i, y) \leq |h(X_i)|, y \in Y_i\}|}{|Y_i|}$$

- **average F1:**

$$avgF1_S(h) = \frac{2 \times avgprec_S(h) \times avgrec_S(h)}{avgprec_S(h) + avgrec_S(h)}$$

Results

Algorithms	Metrics						
	hloss	one-error	coverage	rloss	aveprec	averecl	aveF1
<i>MimlBoost</i>	.053±.004	.094±.014	.387±.037	.035±.005	.937±.008	.792±.010	.858±.008
<i>MimlSvm</i>	.033±.003	.066±.011	.313±.035	.023±.004	.956±.006	.925±.010	.940±.008
<i>MimlSvm_{mi}</i>	.041±.004	.055±.009	.284±.030	.020±.003	.965±.005	.921±.012	.942±.007
<i>MimlNn</i>	.038±.002	.080±.010	.320±.030	.025±.003	.950±.006	.834±.011	.888±.008
<i>AdtBoost.MH</i>	.055±.005	.120±.017	.409±.047	N/A	.926±.011	N/A	N/A
<i>RankSvm</i>	.120±.013	.196±.126	.695±.466	.085±.077	.868±.092	.411±.059	.556±.068
<i>MISvm</i>	.050±.003	.081±.011	.329±.029	.026±.003	.949±.006	.777±.016	.854±.011
<i>MI — knn</i>	.049±.003	.126±.012	.440±.035	.045±.004	.920±.007	.821±.021	.867±.013
<i>SIL</i>	.072±.002	.129±.017	.104±.036	.025±.004	.865±.012	.797±.020	.829±.016
<i>MISVM</i>	.134±.004	.015±.008	.666±.054	.214±.011	.636±.019	.425±.008	.509±.011
<i>mi — SVM</i>	.111±.003	.041±.010	.295±.073	.121±.013	.746±.025	.534±.014	.626±.018

Considerations

- Scores are quite low, but looking at *one-error* metric we see that our classifier has potential
- Selected labels' frequencies are 520, 434, 283, 222, 223, 220, 187 over 2000 documents.
- Test repeated for best 2 labels with following results

Algorithms	Metrics						
	hloss	one-error	coverage	rloss	aveprec	averecl	aveF1
<i>SIL</i>	.099±.007	.141±.005	.005±.001	.005±.001	.998±.001	.948±.006	.972±.003
<i>MI</i> – <i>SVM</i>	.182±.007	.012±.001	.022±.011	.022±.011	.989±.006	.650±.013	.785±.010
<i>mi</i> – <i>SVM</i>	.126±.013	.036±.012	.015±.010	.015±.010	.993±.005	.785±.018	.877±.013

- [1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003.
- [2] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2002.
- [3] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18, 2010.
- [4] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 2006.
- [5] Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.