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 15, 2017

Overview

1. SVM
   - Binary 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

Binary classification

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.

Notation

- A vector $x \in \mathbb{R}^f$ represents an object using f *relevant* features.
- A vector $y \in \{-1, +1\}^l$ indicates wether the example belong to each of the l label classes.

The input of a classification problem is a dataset $D = \{X, Y\}$ where $X \in \mathbb{R}^{n \times f}$ is a set of examples and $Y \in \mathbb{R}^{n \times l}$ is a set of labels.
While learning the target function, the dataset is divided in *training set* and *test set*.

SVM

- For 1-class problems we have to compute the *maximum-margin hyperplane* $w^T x + b$ which best separates positive examples from negative examples.

Optimization problem is:

$$argmin_w \frac{1}{2}||w||^2$$

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

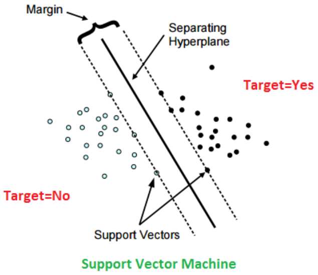

Figure 1: *Solution of maximum-margin hyperplane*

SVM with slacks

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

Optimization problem becames:

$$argmin_{w,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1} n\xi^{(i)}$$

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

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

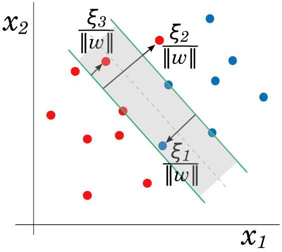

Figure 2: *Solution with slacks*

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
- An example is positive if at least one of its instances is positive, it's negative otherwise
- Dataset labels are assigned to examples, 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:

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

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

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

IMMAGINE MI?

SIL

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.

mi-SVM

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:

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

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

mi-SVM

Our SVM problem becames the following:

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

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

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

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

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

That is an intractable mixed optimization problem

mi-SVM algorithm

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

MI-SVM (X, Y)

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

$y_{i,k} = +1$ *if* $Y_i = +1$

do

  Solve regular SVM finding w, b

  $y_{i,k} = sign(w^T x_{i,k} + b)$ *if* $Y_i = +1$

  Adjust each positive bag to satisfy constraints

while ($y_{i,k}$ change)

MI-SVM

This approach uses directly the dataset in its bag form:

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

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

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

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

MI-SVM algorithm

A feasible algorithm that finds a solution is the following:

```
MI-SVM(X, Y)
x̄_i = avg(x_{i,k}) ∀x_{i,k} ∈ X_i positive bag
do
    Assign ᾱ_i ∈ [0, C] to each x̄_i
    Assign α_{i,j} with Σ_{j=1}^{k_i} α_{i,j} ∈ [0, C] ∀x_{i,k} ∈ X_i negative bag
    Solve regular SVM finding w, b
    Find new x̄_i by selecting the best one for each positive bag
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

Solutions:

- Problem transformation
- Algorithm adaptation

Notation

A set of labels $L = \{y_1, y_2, ...y_l\}$ is given.
Each object contained in the dataset is associated with a set of labels:

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

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

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

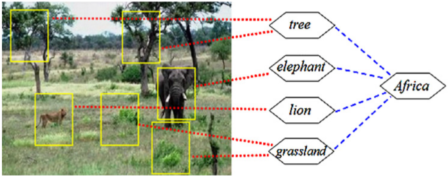

Figure 3: *Multi label example*

IMMAGINE ML?

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)

Algorithm adaptation

Regular algorithms are modified to support multi-label tasks.
Sometimes they use problem transformation at the core.
An example using SVM-related approach based on ranking and label set size prediction.
???

Another multi label approch

[2]

Introduction to MIML

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

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

$$X_i = \{x_{i,k} | k \in [1, k_i], x_{i,k} \in \mathbb{R}^f\} Y_i = \{y_{i,h} | h \in [1, h_i], y_{i,h} \in L, h_i \leq l\}$$

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

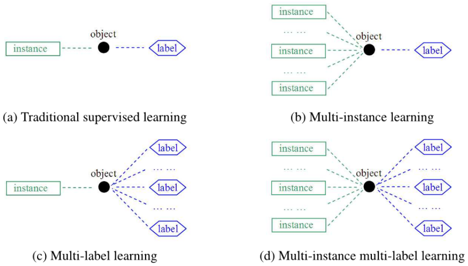

Figure 4: *Different learning frameworks* [4]

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

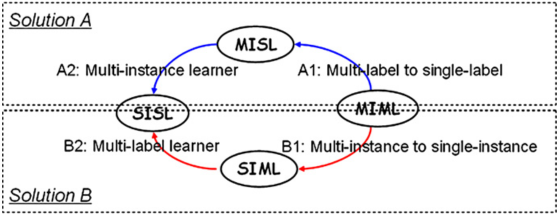

Figure 5: *Two possible solutions to implement MIML*

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:

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

Where

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

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.

Our work

The aim of our work is to replicate a part of the results of [4] using the **MIML framework** and compare the different metrics.

- We focused on the *text categorization* using text documents (*bags*) belonging to categories (*labels*)
- We have choose to use the MimlBoost solution using multi-instance learning as the bridge
- ... ALTRO?

Issues

CI SI METTE?

Evaluation criteria

Four criteria are used for evaluating the performances:

- **hamming loss**: $hloss_S(h) = \dfrac{1}{p} \sum_{i=1}^{p} \frac{1}{|\mathcal{Y}|} |h(X_i) \Delta Y_i|$
  where Δ stands for the symmetric difference between two sets
- **one-error**: $one - error_S(h) = \dfrac{1}{p} \sum_{i=1}^{p} [[\arg\max_{y \in \mathcal{Y}} h(X_i, y)] \notin Y_i]$
- **coverage**: $coverage_S(h) = \dfrac{1}{p} \sum_{i=1}^{p} \max_{y \in Y_i} rank^h(X_i, y) - 1$
- **ranking loss**: $rloss_S(h) = \dfrac{1}{p} \sum_{i=1}^{p} \frac{1}{|Y_i||\bar{Y}_i|} |(y_1, y_2) \; | h(X_i, y_1) \leq h(X_i, y_2), (y_1, y_2) \in Y_i \times \bar{Y}_i|$
  where $\bar{Y}_i$ denotes the complementary set of Y_i in $\mathcal{Y}$

More metrics

We have also used other metrics... NEWS

- **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**:
$$avgrecl_S(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|\{y \mid rank^h(X_i,y) \leq |h(X_i)|,\ y \in Y_i)\}|}{|Y_i|}$$
- **average F1**: $avgF1_S(h) = \dfrac{2 \times avgprec_S(h) \times avgrecl_S(h)}{avgprec_S(h) + avgrecl_S(h)}$

[3]

Test

We have implement a *text categorization* using the dataset REUTERS-21578 selecting **7** most frequent categories on **2000** best documents removing texts that do not have labels or that have a few words.

COSA AGGIUNGERE?

Results

Future works

CI SI METTE? IO DIREI DI NO

References

[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] 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.