A boosting approach to multiview classification with cooperation

Sokol Koço Cécile Capponi

CNRS – LIF UMR 6166 – Aix-Marseille University

September 6, 2011

Outline

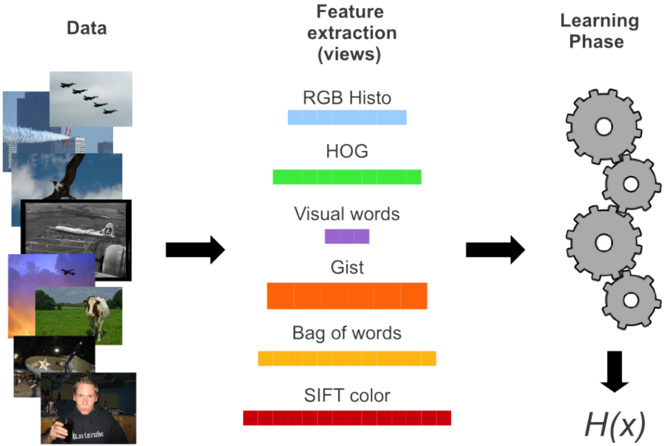
Introduction
- Multi view learning
- Boosting

Mumbo

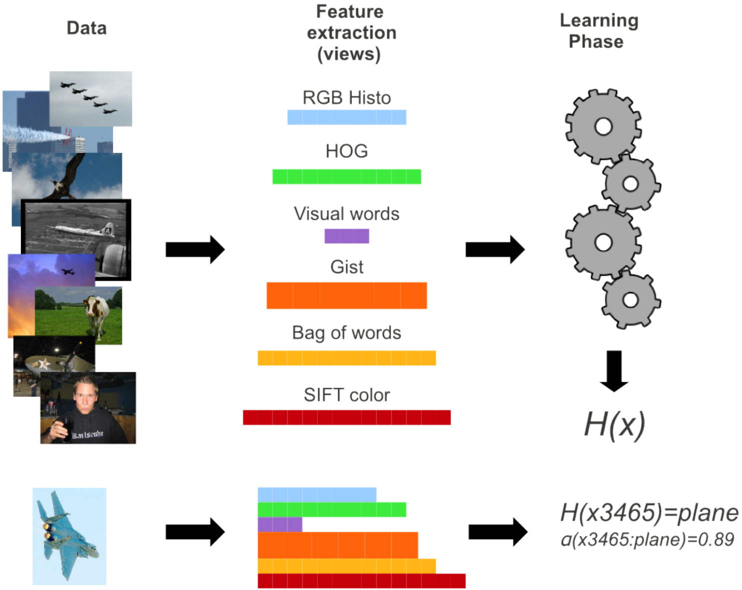
Results

Conclusion et Perspectives

Multi view learning for image classification

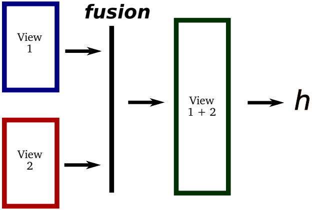


Multi view learning for image classification



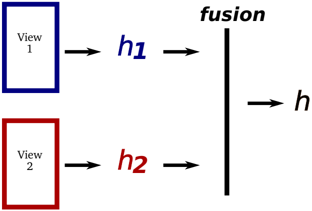
Multi view learning : Fusion based methods

- Early fusion :



- $X = \{x_1, \ldots, x_n\}$ the feature space
- $X_1 \subseteq X,\ X_2 \subseteq X$, s.t. $X_1 \cap X_2 = \emptyset$
- $h : X_1 \odot X_2 \rightarrow Y$

- Late fusion :



- $X = \{x_1, \ldots, x_n\}$ the feature space
- $X_1 \subseteq X,\ X_2 \subseteq X$, s.t. $X_1 \cap X_2 = \emptyset$
- $h = h_1 \circledast h_2$ s.t. $h_1 : X_1 \rightarrow Y$, $h_2 : X_2 \rightarrow Y$

Motivation

Drawbacks of the fusion based methods

- The learnt classifiers do not communicate their failures to each other
- The views must be independent in order for combined classifiers to be most accurate
- The fusion based methods are not effective with weak views

Possible improvement : cooperation among the views

- A classifier learnt on a view gives up on the most difficult examples and entrusts them to the other views
- This should affect only a limited number of examples
- Each example is processed by the most appropriate views

Weak views

Definition(view)

A view is a representation of an example consisting of a set of features

Definition (weak view)

Let V be a view and σ_V the lower bound of the error of $h*$, then V if called a strong view if σ_V is near 0 and V is called a weak view if $\gamma_V = \rho - \sigma_V$ is near 0

Multiclass boosting

Two different approaches :

- divide the multiclass problem in several 1 vs 1 problems
- divide the multiclass problem in several 1 vs all problems

Multiclass boosting

Two different approaches :

- divide the multiclass problem in several 1 vs 1 problems
- divide the multiclass problem in several 1 vs all problems
- Inconvenient : there is no formal definition of weak learning condition

Multiclass boosting

Two different approaches :

- divide the multiclass problem in several 1 vs 1 problems
- divide the multiclass problem in several 1 vs all problems
- Inconvenient : there is no formal definition of weak learning condition

Multi class boosting (Mukherjee et al [NIPS 2010])

- Replace the weight of an example with the cost of classification of the example
- Defines the weak learning condition :

$$\mathbf{C} \cdot \mathbf{1}_h \leq \mathbf{C} \cdot \mathbf{B}$$

Outline

Introduction

Mumbo

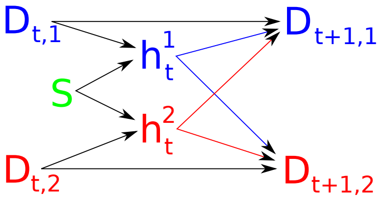
- Main ideas
- The Algorithm
- Empirical error bound
- Generalization error bound

Results

Conclusion et Perspectives

The core of MuMBo

- Several views of different strengths defined on the data
- At each iteration, $m + 1$ cost matrices are maintained, one per view and a global one
- Each classifier communicates its failures to the others
- For each view, the cost matrix is updated using the results of the learnt classifier and the failures of the others



Schema of the algorithm

for $i = 1$ **to** T **do**

Train WL using $\mathbf{C}_{t,1}, \ldots, \mathbf{C}_{t,m}$

for $j = 1$ **to** m **do**

Get $h_{t,j}$with edge $\delta_{t,j}$ on $\mathbf{C}_{t,j}$, and $\alpha_{t,j} = \frac{1}{2} \ln \frac{1+\delta_{t,j}}{1-\delta_{t,j}}$

end for

Update cost matrices (for each view)

Choose

$$\begin{cases} h_t = \underset{h_{t,j}}{\text{argmax}}(\text{edge } h_{t,j} \text{ on } \mathbf{C}_{t,G}) \\ \delta_t = \{edge\ of\ h_t\ on\ \mathbf{C}_{t,G}\} \end{cases}$$

Compute $\alpha_t = \frac{1}{2} \ln \frac{1+\delta_t}{1-\delta_t}$

Update $\mathbf{C}_{t,G}$, the global cost matrix

end for

Output final hypothesis :

$$H(x) = \underset{l \in 1,\ldots,k}{\text{argmax}} f_T(x,l), \quad \text{where} \quad f_T(i,l) = \sum_{t=1}^{T} \mathbb{1}[h_t(i) = l]\alpha_{t,m}$$

Schema of the algorithm : matrix update formula

The matrices are updated using the following formula :

$$\mathbf{C}_{t,j}(i,l) = \begin{cases} \exp(f_{t,j}(i,l) - f_{t,j}(i,y_i)) & \text{if } l \neq y_i \\ -\sum_{p=1;p\neq y_i}^{k} \exp(f_{t,j}(i,p) - f_{t,j}(i,y_i)) & \text{if } l = y_i \end{cases}$$

For $j \in \{1,\ldots,m\}$, $f_{t,j}$ is defined as follows :

- $f_{t,j}(i,l) = \sum_{z=1}^{t} \mathbb{1}[h_{z,j}(i) = l]\alpha_{z,j} d_{z,j}(i)$ and
- $d_{z,j}(i) = \begin{cases} 1 & \text{if } h_{z,j}(i) = y_i \textit{ or } \nexists q \in \{1,\ldots,m\}, h_{z,q}(i) = y_i \\ 0 & \text{else} \end{cases}$

For the global cost matrix (j=G), $f_{t,j}$ is defined as follows :

- $f_{t,G}(i,l) = \sum_{z=1}^{t} \mathbb{1}[h_{z,m}(i) = l]\alpha_{z,m}$

A bound per view

Bounding the empirical error in view m

- Let $\mathbf{C}_{t,m}$ be the cost matrix of the view m and $h_{t,m}$ be the returned classifier for the view m and time t
- Assuming $h_{t,m}$ satisfies the edge condition, then choosing a weight $\alpha_{t,m} > 0$ for $h_{t,m}$ makes the error $\epsilon_{t,m}$ at most a factor

$$\begin{aligned}\tau_{t,m} &= 1 - \frac{1}{2}\left(\exp(\alpha_{t,m}) - \exp(-\alpha_{t,m})\right)\delta_{t,m} \\ &\quad + \frac{1}{2}\left(\exp(\alpha_{t,m}) + \exp(-\alpha_{t,m}) - 2\right)\end{aligned}$$

of the loss before choosing $\alpha_{t,m}$.

Proof of the bound

- Using the edge condition $\mathbf{C_{t,m}} \cdot \mathbf{1_{h_{t,m}}}_h \leq \mathbf{C_{t,m}} \cdot \mathbf{U}_{\delta_{\mathbf{t},\mathbf{m}}}$ and the update formulas, we obtain :
$$\sum_{i \in S_+} L_{t-1,m}(i) - \sum_{i \in S_- \cup S_{-+}} \exp(\zeta_{t-1,m}(i, h_{t,m}(x_i))) \geq \delta_{t,m} \sum_{i \in S} L_{t-1,m}(i)$$
where, $L_{t,m}(i) = \sum_{l \neq y_i} \exp(f_{t,m}(i,l) - f_{t,m}(i,y_i))$
- In order to obtain the loss, we compute :
$$\Delta_+ = (1 - \exp(-\alpha_{t,m})) \sum_{i \in S_+} L_{t-1,m}(i)$$
$$\Delta_- = (\exp(\alpha_{t,m}) - 1) \sum_{i \in S_-} \exp(\zeta_{t-1,m}(i, h_{t,m}(x_i))) \text{ and } \Delta_{-+} = 0$$
- The drop in loss $\Delta = \Delta_+ - \Delta_- - \Delta_{-+}$ at round t is :
$$\begin{aligned}
&\geq \left(\frac{\exp(\alpha_{t,m}) - \exp(-\alpha_{t,m})}{2}\right)\left(\sum_{i \in S_+} L_{t-1,m}(i) - \sum_{i \in S_- \cup S_{-+}} \exp(\zeta_{t-1,m}(i, h_{t,m}(i)))\right) \\
&\quad - \left(\frac{\exp(\alpha_{t,m}) + \exp(-\alpha_{t,m}) - 2}{2}\right)\left(\sum_{i \in S_+} L_{t-1,m}(i) + \sum_{i \in S_- \cup S_{-+}} \exp(\zeta_{t-1,m}(i, h_{t,m}(i)))\right) \\
&\geq \left(\frac{\exp(\alpha_{t,m}) - \exp(-\alpha_{t,m})}{2}\right)\delta_{t,m} \sum_i L_{t-1,m}(i) - \left(\frac{\exp(\alpha_{t,m}) + \exp(-\alpha_{t,m}) - 2}{2}\right)\sum_i L_{t-1,m}(i) \\
&\geq \left(\frac{\exp(\alpha_{t,m}) - \exp(-\alpha_{t,m})}{2}\delta_{t,m} - \frac{\exp(\alpha_{t,m}) + \exp(-\alpha_{t,m}) - 2}{2}\right)\sum_i L_{t-1,m}(i)
\end{aligned}$$

Generalization error bound

Generalization error bound for multi-class problems [Schapire et al.]

Let $\mathcal{D}$ be a distribution, S a sample of n examples, d the VC-dimension of $\mathcal{H}$ and $\delta > 0$. Then with probability $1 - \delta$, for all $\theta > 0$:

$$\mathbf{P}_{\mathcal{D}}[margin(f, x, y) \leq 0] \leq \mathbf{P}_S[margin(f, x, y) \leq \theta] + O\left(\frac{1}{\sqrt{n}} + \left(\frac{d\ \log^2(nk/d)}{\theta^2} + log(1/\delta)\right)^{1/2}\right)$$

To prove that the generalization error of Mumbo decreases with the number of iterations, it suffices to show that $\mathbf{P}_S[margin(f, x, y) \leq \theta]$ decreases.

Lemma

$$\mathbf{P}_S[margin(f, x, y) \leq \theta] \leq \frac{(k-1)}{n}\left(\prod_{t=1}^{T}(1+\delta_t)^{\frac{1+\theta}{2}}(1-\delta_t)^{\frac{1-\theta}{2}}\right)$$

Proof of the bound

Let $l = \underset{y' \neq y}{argmax} f(x, y')$. The margin of an example is defined as : $margin(f, x, y) = f(x, y) - f(x, l) = \frac{\sum_t h_t(x,y)\alpha_t}{\sum_t \alpha_t} - \frac{\sum_t h_t(x,l)\alpha_t}{\sum_t \alpha_t}$

Hence,

- $margin(f, x, y) < \theta \Leftrightarrow \theta \sum_t \alpha_t - \left(\sum_t h_t(x, y)\alpha_t - \sum_t h_t(x, l)\alpha_t \right) \geq 0$
- $\mathbf{P}[margin(f, x_i, y) \leq \theta] = 1 \Leftrightarrow$
  $- \left(\sum_t \alpha_t h_t(x_i, y) - \sum_t \alpha_t h_t(x_i, l) \right) + \theta \sum_t \alpha_t \geq 0$
- $\mathbf{P}_S[margin(f, x, y) \leq \theta] \leq$
  $\frac{1}{n} \sum_{i=1}^{n} \exp(- \left(\sum_t \alpha_t h_t(x_i, y) - \sum_t \alpha_t h_t(x_i, l) \right)) \exp(\theta \sum_t \alpha_t)$
  $\leq \frac{1}{n} \sum_{i=1}^{n} \sum_{y' \neq y} \exp\left(f_T(x_i, y') - f_T(x_i, y)\right) \exp(\theta \sum_t \alpha_t)$
  $= \frac{1}{n} \exp(\theta \sum_t \alpha_t) \epsilon_t$

Using the bound on the empirical error, we obtain :

- $\mathbf{P}_S[margin(f, x, y)] \leq \frac{(k-1)}{n} \left(\prod_t (1 + \delta_t)^{\frac{1+\theta}{2}} (1 - \delta_t)^{\frac{1-\theta}{2}} \right)$

Outline

Introduction

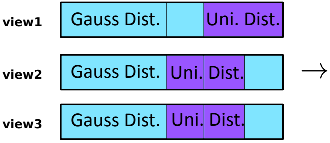
Mumbo

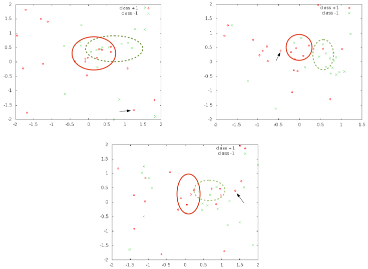
Results

Synthetic data

Conclusion et Perspectives

Results on synthetic data





$\lvert S \rvert = 80,\ \eta_M$	0.5	0.38	0.25	0.12	0
Early+SVM	0.390	0.410	0.437	0.396	0.389
SVM+Late	0.246	0.229	0.263	0.254	0.232
Early+Adaboost	0.415	0.420	0.403	0.364	0.358
Mumbo	0.148	0.152	0.168	0.174	0.164

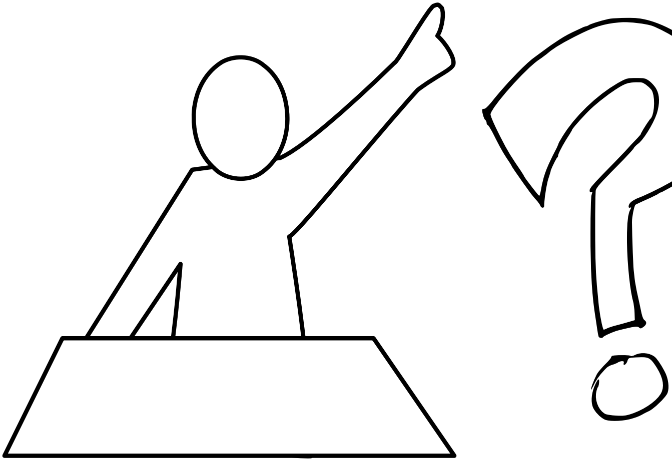
Outline

Conclusion

- We present a boosting-like algorithm for the multi view setting, which promotes the collaboration between views
- The views are of different strengths
- The cost update formulas allow the views to focus on examples that are hard to classify for the others
- The bounds on the generalization error and the empirical error are proved
- The results on synthetic data confirm the theoretical properties

Perspectives

- Include the performances of the classifiers in the update formulas
- Find tighter bounds for the empirical and the generalization errors
- Test this algorithm on speech recognition data



This work is supported by the French agency ANR, Project DECODA, contract no 2009-CORD-005-01, and the French business clusters Cap Digital and SCS. For more information about the DECODA project, please visit the project home-page, `http://decoda.univ-avignon.fr`.