A Theoretical and Empirical Analysis of Support Vector Machine Methods for Multiple-Instance Learning

Gary Doran  
gary.doran@case.edu

Soumya Ray  
sray@case.edu

ECML 2014
Motivation

- Standard supervised SVMs have some nice properties: they are efficiently solvable, and their loss functions appropriately measure risk.

- Most prior work takes for granted that these properties will be satisfied when SVMs are extended to the MI setting.

- However, we show that this cannot be the case for MI SVMs.
Content-based Image Retrieval (CBIR)

Positive Image

✓

Negative Image

✗
Content-based Image Retrieval (CBIR)

Positive Image

Negative Image
Content-based Image Retrieval (CBIR)

Positive Image

Negative Image
Content-based Image Retrieval (CBIR)

Positive Image

Negative Image
Content-based Image Retrieval (CBIR)

Positive Image

Negative Image
Content-based Image Retrieval (CBIR)

Positive Image

Negative Image

✓

✗
### Multiple-Instance Learning

<table>
<thead>
<tr>
<th>Bag–Instance Label Relationship</th>
<th>Bags</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels</td>
<td>Observed</td>
<td>Unobserved</td>
</tr>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>At least one Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>All Negative</td>
</tr>
</tbody>
</table>

4
Support Vector Machines

\[
\begin{align*}
\min_{w, b, \xi} & \quad \frac{1}{2} \|w\|^2 + C \sum_i \xi_i, \\
\text{s.t.} & \quad y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \\
& \quad \xi_i \geq 0
\end{align*}
\]
SVM Properties

Hyperplane: \((w, b)\)

Solution: \((w, b, \xi)\)

Consistent
Zero Loss \((w, b, 0) \in \text{Feasible}\)

Inconsistent
Nonzero Loss \((w, b, 0) \notin \text{Feasible}\)
MI SVM Properties

Positive Bag

Negative Bag

Hyperplane: 
\[(w, b)\]

Solution: 
\[(w, b, \xi)\]

Consistent

Zero Loss 
\((w, b, 0) \in \text{Feasible}\)

Inconsistent

Nonzero Loss 
\((w, b, 0) \not\in \text{Feasible}\)
\textbf{sMIL} \\
(Bunescu & Mooney 2007)

\[
\begin{align*}
\min_{w, b, \xi} & \quad \left\{ \frac{1}{2} \|w\|^2 \right\} + \left\{ \frac{C}{|B^+|} \sum_i \xi^+_i \right\} + \left\{ \frac{C}{|X^-|} \sum_{i,j} \xi^-_{ij} \right\} \\
\text{s.t.} & \quad -\langle w, \phi(x_{ij}) \rangle + b \geq 1 - \xi^-_{ij} \quad \text{if } Y_i = -1 \\
& \quad \frac{1}{|B_i|} \sum_j (\langle w, \phi(x_{ij}) \rangle + b) \geq \frac{2-|B_i|}{|B_i|} - \xi^+_i \quad \text{if } Y_i = 1 \quad \xi_{ij} \geq 0
\end{align*}
\]
sMIL

(Bunescu & Mooney 2007)

\[ \min_{w, b, \xi} \begin{cases} \underbrace{\frac{1}{2} \|w\|^2} & \text{Regularization} \\ + \underbrace{\frac{C}{|B^+|} \sum_i \xi_i^+} & \text{Pos. Bag Loss} \\ + \underbrace{\frac{C}{|X^-|} \sum_{i,j} \xi_{ij}^-} & \text{Neg. Instance Loss} \end{cases} \]

\[ \begin{align*}
\text{s.t.} & \quad - (\langle w, \phi(x_{ij}) \rangle + b) \geq 1 - \xi_{ij}^- \\
& \quad \frac{1}{|B_i|} \sum_j (\langle w, \phi(x_{ij}) \rangle + b) \geq \frac{2 - |B_i|}{|B_i|} - \xi_i^+ \quad \text{if } Y_i = -1 \\
& \quad \text{if } Y_i = 1, \xi_{ij} \geq 0
\end{align*} \]
sMIL

(Bunescu & Mooney 2007)

\[
\min_{w,b,\xi} \left\{ \frac{1}{2} \|w\|^2 \right\} \quad \text{Regularization} \quad + \quad \left\{ \frac{C}{|B^+|} \sum_i \xi_i^+ \right\} \quad \text{Pos. Bag Loss} \quad + \quad \left\{ \frac{C}{|X^-|} \sum_{i,j} \xi_{ij}^- \right\} \quad \text{Neg. Instance Loss}
\]

\[
\text{s.t.} \quad \begin{cases} 
- (\langle w, \phi(x_{ij}) \rangle + b) \geq 1 - \xi_{ij}^- & \text{if } Y_i = -1 \\
\frac{1}{|B_i|} \sum_j \left( \langle w, \phi(x_{ij}) \rangle + b \right) \geq \frac{2-|B_i|}{|B_i|} - \xi_i^+ & \text{if } Y_i = 1
\end{cases}, \quad \xi_{ij} \geq 0
\]

Average instance label in the bag

Standard negative instance constraints
sMIL

(Bunescu & Mooney 2007)

\[
\min_{w,b,\xi} \left\{ \frac{1}{2} \|w\|^2 + \frac{C}{|B^+|} \sum_i \xi_i^+ + \frac{C}{|X^-|} \sum_{i,j} \xi_{ij}^- \right\}
\]

s.t.
\[
\begin{align*}
- (\langle w, \phi(x_{ij}) \rangle + b) & \geq 1 - \xi_{ij}^- \\
\frac{1}{|B_i|} \sum_j \left( \langle w, \phi(x_{ij}) \rangle + b \right) & \geq \frac{2 - |B_i|}{|B_i|} - \xi_i^+ 
\end{align*}
\]

if \( Y_i = -1 \)

if \( Y_i = 1 \)

, \( \xi_{ij} \geq 0 \)

Average instance label in the bag

Expected average label if there is exactly 1 positive instance in the bag

Standard negative instance constraints
sMIL

Inconsistent hyperplane, zero loss!

\[
\frac{1}{|B_i|} \sum_j (\langle w, \phi(x_{ij}) \rangle + b) \geq \frac{2 - |B_i|}{|B_i|} - \xi_i^+
\]

\[
\frac{1}{4} \sum_j (\langle w, \phi(x_{ij}) \rangle + b) \geq -\frac{1}{2} - 0
\]
sMIL

Consistent hyperplane, nonzero loss!

\[
\frac{1}{|B_i|} \sum_j \left( \langle w, \phi(x_{ij}) \rangle + b \right) \geq \frac{2 - |B_i|}{|B_i|} - \xi_i^+ \\
\frac{1}{2} \sum_j \left( \langle w, \phi(x_{ij}) \rangle + b \right) \geq 0 - 0
\]
SVM Properties

• Zero-loss solutions are consistent [Soundness]

• There is a zero-loss solution for each consistent hyperplane [Completeness]

• The objective function and feasible region of the optimization program are convex [Convex]

Some MI SVMs do not have some of these properties!
Other MI SVMs
Other MI SVMs

Andrews et al. 2003

MI-SVM    mi-SVM
Other MI SVMs

Andrews et al. 2003
MI-SVM    mi-SVM

Zhou & Xu 2007
MissSVM
Other MI SVMs

Andrews et al. 2003
ML-SVM  mi-SVM

Zhou & Xu 2007
MissSVM

Bunescu & Mooney 2007
stMIL  sbMIL  sMIL
Other MI SVMs

Andrews et al. 2003
MI-SVM   mi-SVM

Zhou & Xu 2007
MissSVM

Bunescu & Mooney 2007
stMIL    sbMIL    sMIL

Mangasarian & Wild 2008
MICA
Other MI SVMs

Andrews et al. 2003
MI-SVM  mi-SVM

Zhou & Xu 2007
MissSVM

Bunescu & Mooney 2007
stMIL sbMIL sMIL

Mangasarian & Wild 2008
MICA

Li et al. 2009
I-KI-SVM B-KI-SVM
## Other MI SVMs

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrews et al. 2003</td>
<td></td>
<td>MI-SVM, mi-SVM</td>
</tr>
<tr>
<td>Zhou &amp; Xu 2007</td>
<td></td>
<td>MissSVM</td>
</tr>
<tr>
<td>Bunescu &amp; Mooney 2007</td>
<td></td>
<td>stMIL, sbMIL, sMIL</td>
</tr>
<tr>
<td>Li et al. 2009</td>
<td></td>
<td>I-KI-SVM, B-KI-SVM</td>
</tr>
<tr>
<td>Mangasarian &amp; Wild 2008</td>
<td></td>
<td>MICA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Supervised SVM, SIL</td>
</tr>
</tbody>
</table>
MI SVM Properties

Sound
Zero-loss solutions are consistent

Complete
Consistent hyperplanes have zero-loss solutions

Convex

MI-SVM
mi-SVM
MissSVM
MICA

stMIL
sbMIL
SIL

sMIL

B-KI-SVM
I-KI-SVM
MI SVM Properties

Sound
- Zero-loss solutions are consistent

Complete
- Consistent hyperplanes have zero-loss solutions

Convex

MI-SVM, mi-SVM, MissSVM, MICA

stMIL

sbMIL, SIL

sMIL
Main Result

No MI SVM can be sound, complete, and convex.
Main Result

No MI SVM can be sound, complete, and convex.
Main Result

No MI SVM can be sound, complete, and convex.
Main Result

No MI SVM can be sound, complete, and convex.

Zero-Loss Solutions (Convex)
Main Result

No MI SVM can be sound, complete, and convex.

Zero-Loss Solutions (Convex)
Main Result

No MI SVM can be sound, complete, and convex.

Zero-Loss Solutions (Convex)
MI Kernels

Gärtner et al. 2002:

\[ K_{\text{MI}}(B_i, B_j) = \frac{1}{f_{\text{norm}}(B_i) f_{\text{norm}}(B_j)} \sum_{x_i \in B_i} \sum_{x_j \in B_j} k_1^p(x_i, x_j) \]
MI Kernels

Gärtner et al. 2002:

\[ K_{MI}(B_i, B_j) = \frac{1}{f_{\text{norm}}(B_i) f_{\text{norm}}(B_j)} \sum_{x_i \in B_i} \sum_{x_j \in B_j} k^p_I(x_i, x_j) \]

Instance Kernel
MI Kernels

Gärtner et al. 2002:

\[
K_{MI}(B_i, B_j) = \frac{1}{f_{\text{norm}}(B_i)f_{\text{norm}}(B_j)} \sum_{x_i \in B_i} \sum_{x_j \in B_j} k^p_I(x_i, x_j)
\]
MI Kernels

Gärtner et al. 2002:

\[ K_{\text{MI}}(B_i, B_j) = \frac{1}{f_{\text{norm}}(B_i)f_{\text{norm}}(B_j)} \sum_{x_i \in B_i} \sum_{x_j \in B_j} k^p_I(x_i, x_j) \]

MI Kernel

Normalized Set Kernel (NSK)

Normalization

Instance Kernel
MI Kernels

Gärtner et al. 2002:

$$K_{MI}(B_i, B_j) = \frac{1}{f_{\text{norm}}(B_i)f_{\text{norm}}(B_j)} \sum_{x_i \in B_i} \sum_{x_j \in B_j} k^p_I(x_i, x_j)$$

- MI Kernel
- Normalized Set Kernel (NSK)
- Integer Parameter
- Instance Kernel
- Normalization
MI Kernel Properties

• Since the MI kernel is used with a standard SVM, it is convex, sound, and complete with respect to *bag labeling*, so we define properties in terms of the ability of $K_{MI}$ and $K_I$ to separate bags and instances, respectively.

• Completeness for MI Kernels: If $K_I$ separates instances, then $K_{MI}$ separates bags for some $p$ (Gärtner et al. 2002).

• However, the converse is *not* true: there exist bag separators when no instance separators exist (not sound).
MI SVM Properties

Sound
Zero-loss solutions are consistent

Complete
Consistent hyperplanes have zero-loss solutions

Convex

MI-SVM
mi-SVM
MissSVM
MICA

stMIL
sbMIL
SIL

NSK*
KI-SVM

sMIL
Empirical Hypotheses

- Soundness and Completeness ensure that structural risk minimization properly assesses risk through the loss function.
- Soundness is more important as it ensures that inconsistent solutions are penalized.
- We hypothesize that techniques with these properties perform better in general.
Methodology

- Datasets: CBIR, Text Categorization, Drug Activity Prediction (22 Total)
- 12 datasets have instance labels
- Evaluate both instance- and bag-labeling tasks w.r.t. accuracy
Instance-Labeling Results

![Diagram showing rankings and methods: CD, NSK-FS, NSK-AV, MissSVM, SIL, MICA, stMIL, sbMIL, MI-SVM, mi-SVM, B-KI-SVM, I-KI-SVM, sMIL]
Instance-Labeling Results

- Techniques that are only complete perform most poorly (NSK).
- Sacrificing completeness is sometimes beneficial (sbMIL).

Average Ranks

<table>
<thead>
<tr>
<th></th>
<th>Sound</th>
<th>Complete</th>
<th>Convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Techniques</td>
<td>6.1</td>
<td>5.8</td>
<td>7.9</td>
</tr>
<tr>
<td>that are only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>complete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>perform most</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poorly (NSK)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sacrificing</td>
<td>5.5</td>
<td></td>
<td>6.1</td>
</tr>
<tr>
<td>completeness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is sometimes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beneficial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(sbMIL)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bag-Labeling Results

Worst  Rank  Best

CD
sMIL  stMIL  MissSVM  MICA  SIL  I-KI-SVM

NSK-FS  sbMIL  NSK-AV  MI-SVM  mi-SVM  B-KI-SVM
Bag-Labeling Results

• Now, the NSK has the best accuracy (it is sound for bag-labeling).

• Again, sacrificing completeness is sometimes beneficial (sbMIL).

Average Ranks

<table>
<thead>
<tr>
<th></th>
<th>Sound</th>
<th>Complete</th>
<th>Convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Ranks</td>
<td>11.1</td>
<td>6.8</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td></td>
<td>11.2</td>
</tr>
</tbody>
</table>
Conclusions

• The soundness, completeness, and convexity properties enjoyed by supervised SVM can be defined for MI SVMs as well.

• However, no MI SVM can have all three properties.

• The presence or absence of each property has practical implications for performance.

• Our results highlight differences in behavior of the instance- and bag-labeling performance of each approach (Tragante do O et al. 2011).
Future and Ongoing Work

• Can we define similar properties with respect to other metrics (e.g., AUC)?

• Are there bag-level kernels that also perform well on the instance-labeling task?

• What is the relationship between these properties and formal learnability results?