Multiple-Instance Learning

for weakly supervised visual classification

Yannis Kalantidis

National Technical University of Athens

February 2011

Outline

Multiple-Instance Learning

an example (from Dietterich et al. 1996)

- there exists a keyed lock on the door to the supply room in an office
- each staff member has a key chain containing several keys
- *one* key on *every* key chain can open the supply room door
- for some staff members their supply room key may open one or more other doors

Multiple-Instance Learning

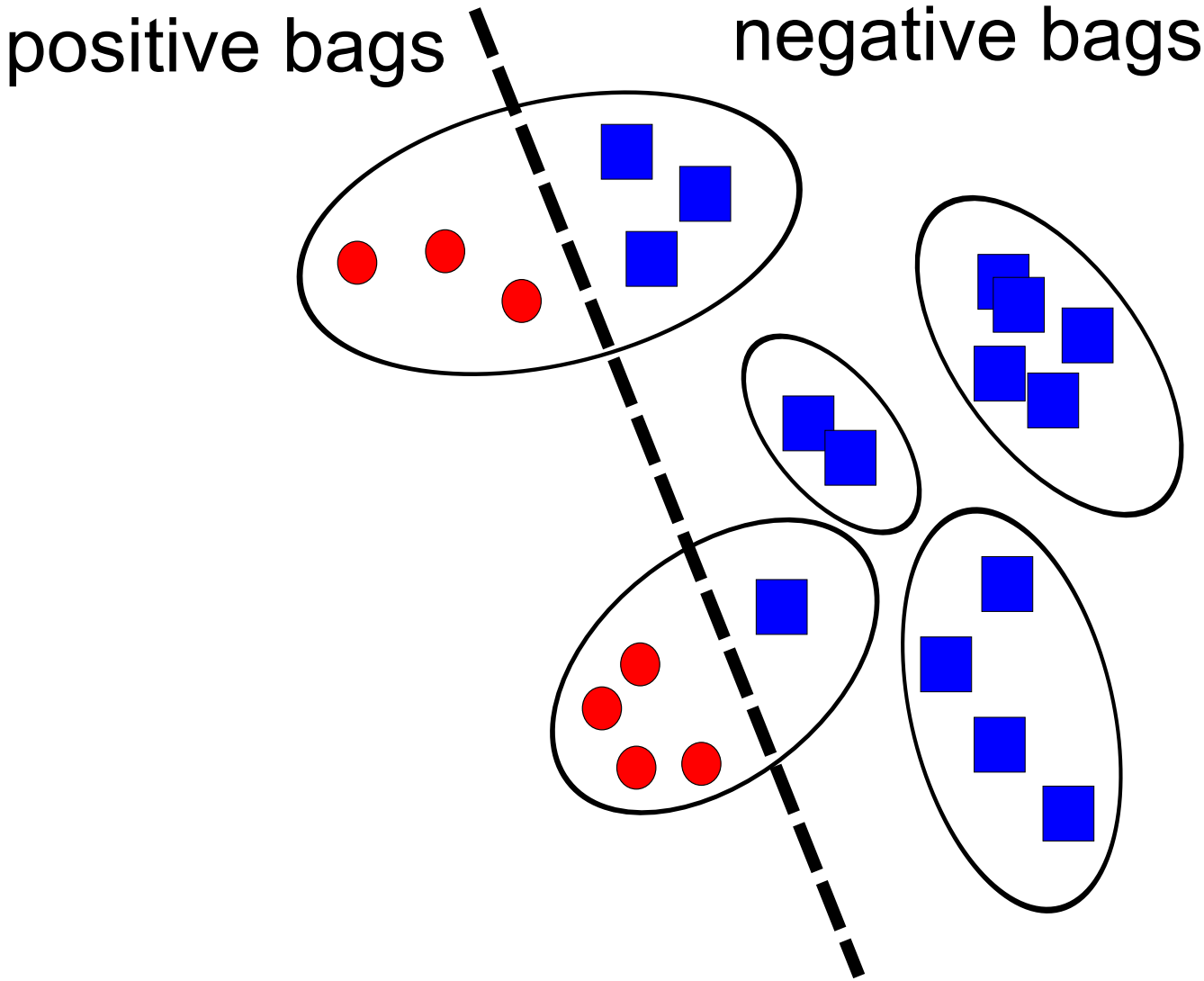
an example (from Dietterich et al. 1996)

- you are a locksmith, attempting to infer the most general required shape that a key must have in order to open the supply room door
- if you knew this required shape, you could predict, by examining any key, whether that key could unlock the door
- the staff members are uncooperative: they just hand you their entire key chain
- you are not given access to the supply room door, so you can't try out the individual keys

Multiple-instance learning (MIL)

- a generalization of supervised classification in which training class labels are associated with sets of patterns, or *bags*, instead of individual patterns
- pattern labels are only indirectly accessible through labels attached to bags
- a set receives a particular label, if at least one of the patterns in the set possesses the label

Multiple-instance learning – Example



Multiple-instance learning (MIL)

the key challenge in MIL

- to cope with the ambiguity of not knowing which of the patterns in a positive bag are the actual positive examples and which ones are not

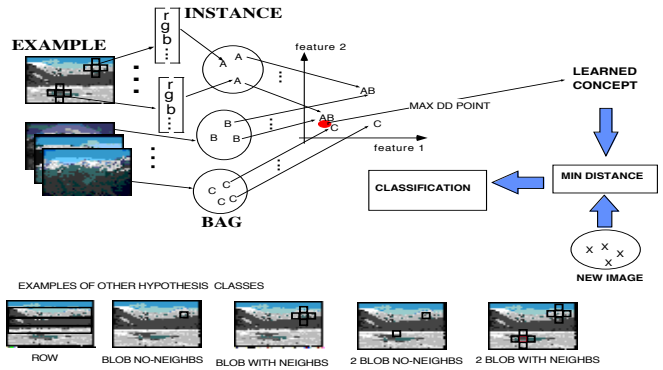
Multiple-instance learning – Applications

- the classification of molecules in the context of drug design
  - each molecule is represented by a bag of possible conformations—alternative shapes that the molecule can adopt by rotating its bonds
  - the eficacy of a molecule can be tested experimentally, but there is no way to control for individual conformations
- image indexing for content-based image retrieval
  - an image can be viewed as a bag of local image patches or image regions
  - annotating whole images is far less time consuming then marking relevant image regions
- text categorization
  - labels are rarely available on the passage level and are most commonly associated with the document as a whole

Open source code for MIL

- MATLAB
  - MILL: A Multiple Instance Learning Library
    - http://www.cs.cmu.edu/ juny/MILL/index.html
- Java
  - MILK: A Multi-Instance Learning Kit in Java (now part of WEKA)
    - http://www.cs.waikato.ac.nz/ml/milk/
    - http://www.cs.waikato.ac.nz/ml/weka/

Multiple-Instance Learning for Natural Scene Classification



Maron, Ratan – ICML, 1998

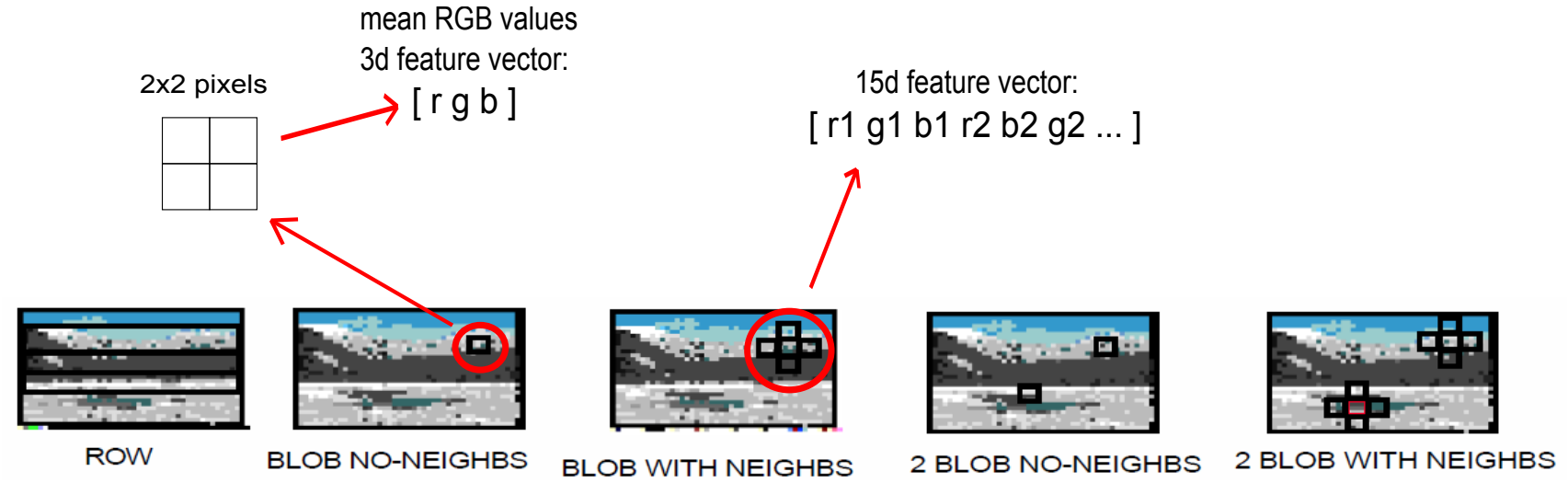
Multiple-Instance Learning for Natural Scene Classification

MIL for Natural Scene Classification

apply the MIL framework to the problem of learning how to classify natural images

- images are inherently ambiguous
- each image is a bag
- the instances are various sub-regions in the image

Features: blobs

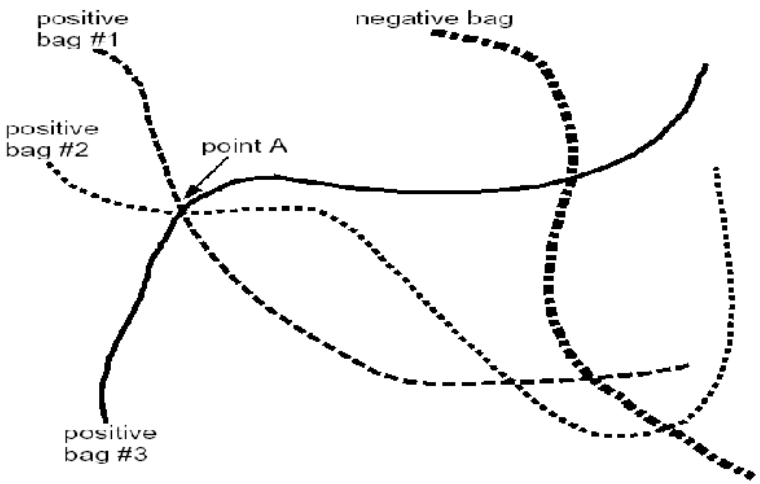


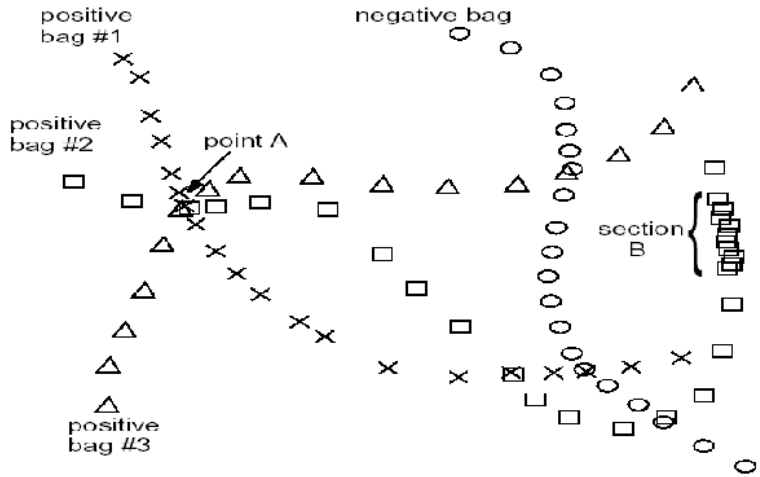
MIL for Natural Scene Classification

use the Diverse Density algorithm

- it regards each bag as a manifold, which is composed of many instances, *i.e.* feature vectors
- If a new bag is positive then it is believed to intersect all positive feature-manifolds without intersecting any negative feature-manifolds
- Intuitively, diverse density at a point in the feature space is defined to be a measure of how many different positive bags have instances near that point, and how far the negative instances are from that point
- the task of multi-instance learning is transformed to the search for points in the feature space with high diverse density

Diverse Density algorithm



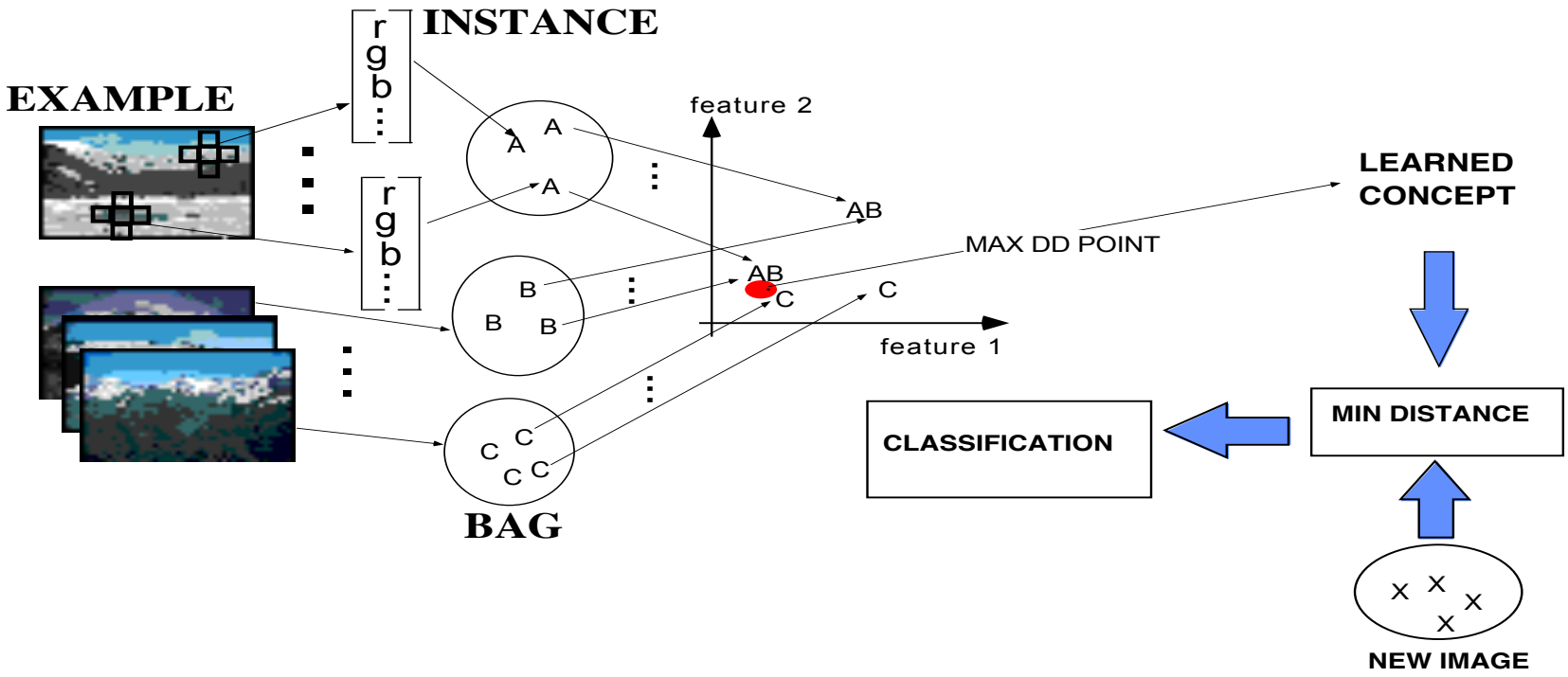


MIL for Natural Scene Classification

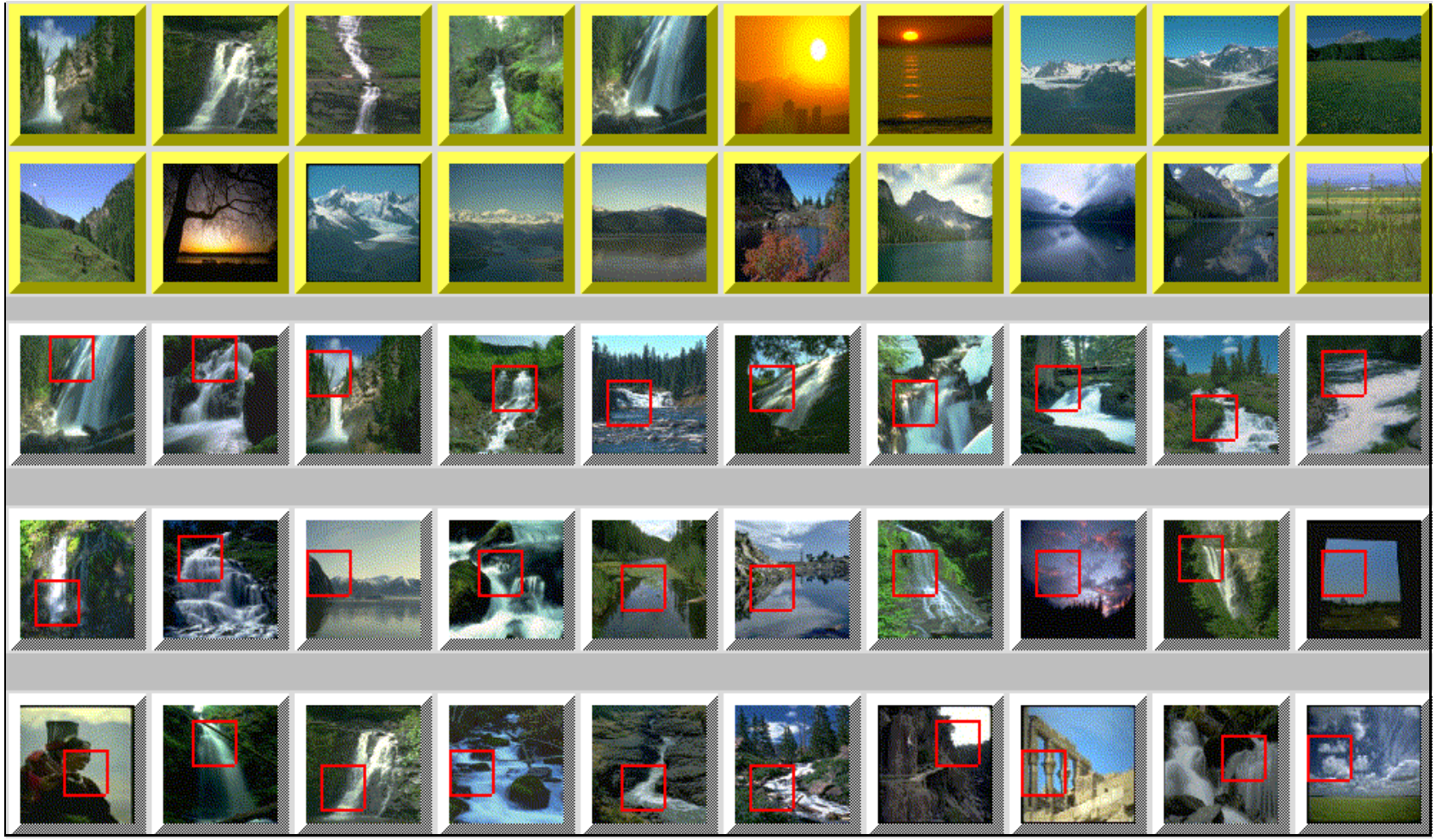
finding the maximum diverse density

- use gradient ascent with multiple starting points (that are instances from positive bags)
- time-consuming
- combine with EM-algorithm (DD-EM) to increase performance (*Zhang et al.*)

MIL for Natural Scene Classification – Overview



MIL for Natural Scene Classification – Results



Sparse Multiple-Instance Learning (sMIL)

minimize:

$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{L_n} \sum_{x \in \tilde{\mathcal{X}}_n} \xi_x + \frac{C}{|\mathcal{X}_p|} \sum_{X \in \mathcal{X}_p} \xi_X$$

subject to:

$$\begin{aligned} w\,\phi(x) + b &\leq -1 + \xi_x, && \forall x \in \tilde{\mathcal{X}}_n \\ w\,\frac{\phi(X)}{|X|} + b &\geq \frac{2 - |X|}{|X|} - \xi_X, && \forall X \in \mathcal{X}_p \quad (*) \\ \xi_x &\geq 0, \xi_X \geq 0 \end{aligned}$$

Bunescu, Mooney – ICML, 2007

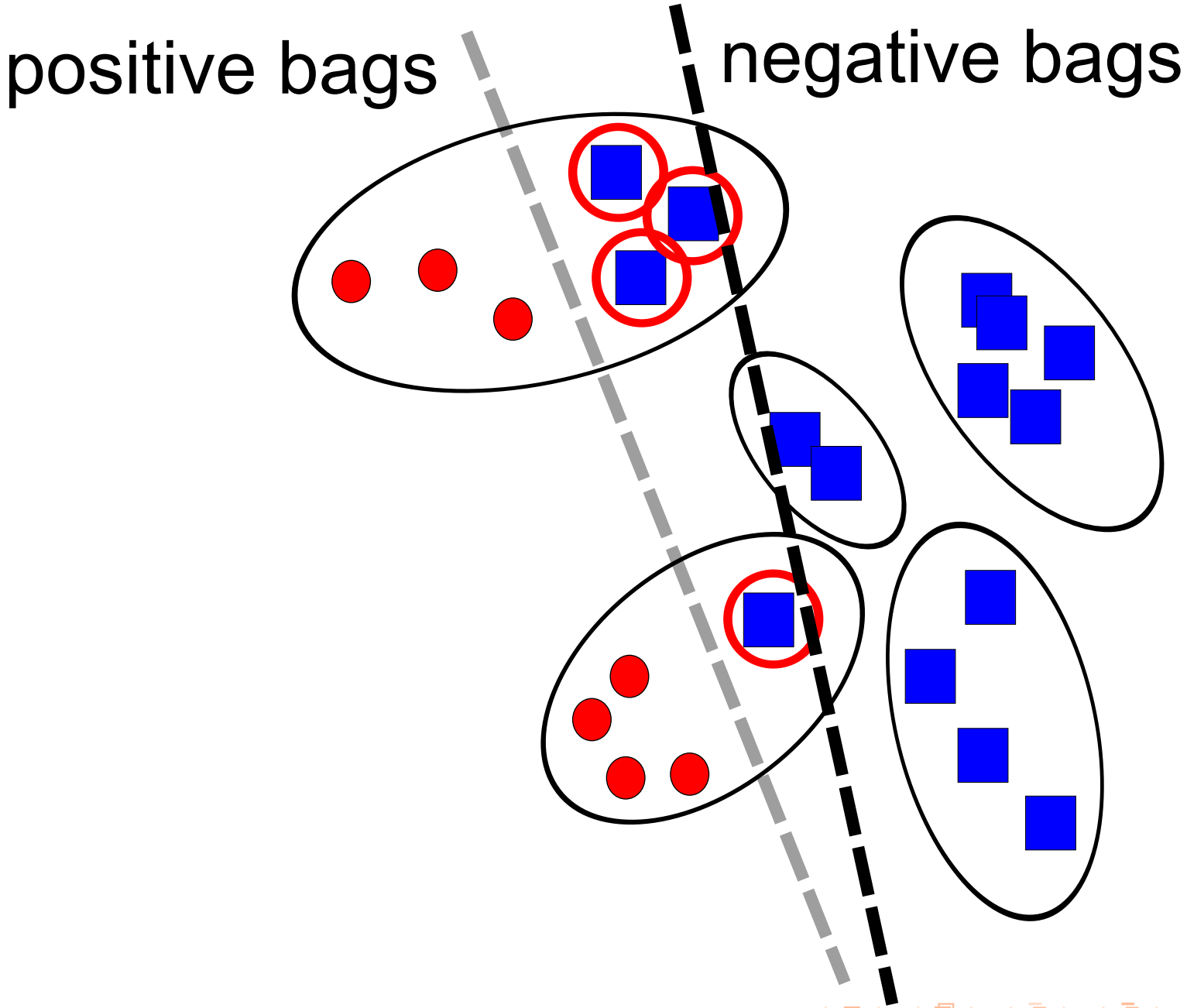
Multiple-Instance Learning for Sparse Positive Bags

Simpler solutions to MIL problems

Single Instance Learning (SIL)

- transform the problem into a standard supervised learning problem by just labeling all instances from positive bags as positive
- often achieves competitive results when compared with other more sophisticated MIL methods (*Ray & Craven, 2005*)
- however, SIL and many other MIL methods, rely on the positive bags being *fairly rich* in positive examples

SIL - example



Single Instance Learning using SVM

SIL-SVM

- apply a normal (soft-margin) SVM on the resulting dataset

Notation

- Let $\mathcal{X}$ be the set of bags used for training, $\mathcal{X}_p \subseteq \mathcal{X}$ the set of positive bags, and $\mathcal{X}_n \subseteq \mathcal{X}$ the set of negative bags.
- Let $\tilde{\mathcal{X}}_p = \{x | x \in X \in \mathcal{X}_p\}$ and $\tilde{\mathcal{X}}_n = \{x | x \in X \in \mathcal{X}_n\}$ be the set of instances from positive bags and negative bags, respectively.
- Let $L = L_p + L_n = |\tilde{\mathcal{X}}_p| + |\tilde{\mathcal{X}}_n|$ be the total number of instances.
- For any instance $x \in X$ from a bag $X \in \mathcal{X}$, let $\phi(x)$ be the feature vector representation of x.
- $\phi(X) = \sum_{x \in X} \phi(x)$ is the feature vector representation of bag X.

SIL-SVM Optimization

minimize:

$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{L}\sum_{X \in \mathcal{X}}\sum_{x \in X}\xi_x$$

subject to:

$$\begin{aligned} & w\,\phi(x) + b \leq -1 + \xi_x, \quad \forall x \in \tilde{\mathcal{X}}_n \\ & w\,\phi(x) + b \geq +1 - \xi_x, \quad \forall x \in \tilde{\mathcal{X}}_p \quad (*) \\ & \xi_x \geq 0 \end{aligned}$$

Normalized Set Kernel (NSK)

Normalized Set Kernel (NSK-SVM), Gartner et al. 2002

- a bag is represented as the sum of all its instances, normalized by its 1 or 2-norm
- the resulting representation is used in training a traditional SVM

NSK-SVM Optimization

minimize:

$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{|\mathcal{X}|}\sum_{X \in \mathcal{X}} \xi_X$$

subject to:

$$w\,\frac{\phi(X)}{|X|} + b \leq -1 + \xi_X, \;\; \forall X \in \mathcal{X}_n$$

$$w\,\frac{\phi(X)}{|X|} + b \geq +1 - \xi_X, \;\; \forall X \in \mathcal{X}_p$$

$$\xi_X \geq 0$$

tight NSK-SVM Optimization

minimize:

$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{L_n} \sum_{x \in \tilde{\mathcal{X}}_n} \xi_x + \frac{C}{|\mathcal{X}_p|} \sum_{X \in \mathcal{X}_p} \xi_X$$

subject to:

$$\begin{aligned}
w\,\phi(x) + b \;&\leq -1 + \xi_x, \quad \forall x \in \tilde{\mathcal{X}}_n \\
w\,\frac{\phi(X)}{|X|} + b &\geq +1 - \xi_X, \quad \forall X \in \mathcal{X}_p \quad (*) \\
\xi_x \geq 0, \xi_X &\geq 0
\end{aligned}$$

Maximum pattern margin formulation (mi-SVM)

refinement of the Statistic Kernel SVM, Andrews et al. 2003

- start by training a SIL-SVM
- relabel the instances in positive bags using the learned decision hyperplane
- if a positive bag contains no instances labeled as positive, then the instance that gives the maximum value of the decision function for that bag is relabeled as positive
- retrain SVM with the new dataset and repeat until no labels are changed

Maximum bag margin (MI-SVM)

refinement of the Normalized Set Kernel SVM, Andrews et al. 2003

- start by training a NSK-SVM
- for every positive bag, use the learned decision function to select the bag instance that gives the maximum value
- replace the bag representation (that was initially an average of all bag instances) with this instance
- retrain SVM with the new dataset and repeat until no bag representation is changed

Transductive SVMs

multi-instance kernels ignore individual instances from positive bags

- instances from positive bags can be treated as unlabeled data, with the potential of further improving the generalization accuracy when used in the framework of transductive support vector machines
- desired constraint: that at least one of the instances in a positive bag is positive, so we further constrain all bag instances to be classified far away from the decision hyperplane,
- use the framework of transductive SVMs
- introduce a balancing constraint, in order to ensure that unlabeled examples are assigned to both classes

Sparse Multiple-Instance Learning (sMIL)

[t] Let $y(x) = \pm 1$ be the hidden (*i.e.* unknown) label of an instance x from a positive bag X

$$w \frac{\phi(X)}{|X|} + b \geq +1 - \xi_X, \;\; \forall X \in \mathcal{X}_p$$

can be written as

$$\sum_{x \in X} \frac{w\,\phi(x) + b}{|X|} \geq \sum_{x \in X} \frac{y(x)}{|X|} - \xi_X$$

$$y(x) = 1, \; \forall x \in X$$

Sparse Multiple-Instance Learning (sMIL)

[t] Let $y(x) = \pm 1$ be the hidden (*i.e.* unknown) label of an instance x from a positive bag X

$$w \frac{\phi(X)}{|X|} + b \geq +1 - \xi_X, \ \ \forall X \in \mathcal{X}_p$$

can be written as

$$\sum_{x \in X} \frac{w\,\phi(x) + b}{|X|} \geq \sum_{x \in X} \frac{y(x)}{|X|} - \xi_X$$

$$y(x) = -1, \ \forall x \in X \backslash \{\hat{x}\}$$

$$y(\hat{x}) = +1$$

Sparse MIL Optimization

minimize:

$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{L_n} \sum_{x \in \tilde{\mathcal{X}}_n} \xi_x + \frac{C}{|\mathcal{X}_p|} \sum_{X \in \mathcal{X}_p} \xi_X$$

subject to:

$$\begin{aligned}
w\,\phi(x) + b &\leq -1 + \xi_x, && \forall x \in \tilde{\mathcal{X}}_n \\
w\,\frac{\phi(X)}{|X|} + b &\geq \frac{2 - |X|}{|X|} - \xi_X, && \forall X \in \mathcal{X}_p \quad (*) \\
\xi_x &\geq 0, \xi_X \geq 0
\end{aligned}$$

Sparse transductive SVM

- enforce an upper bound on the scores of negative instances inside a positive bag
- add the transductive constraint
- non-convex optimization problem, in which the objective function is rewritten as a difference of convex functions, and then solved using the Concave Convex Procedure

Sparse transductive MIL

minimize:

$$\mathbf{J}(\cdot) = \frac{\|w\|^2}{2} + \frac{C}{L_n} \sum_{x \in \tilde{\mathcal{X}}_n} \xi_x + \frac{C^*}{L_p} \sum_{x \in \tilde{\mathcal{X}}_p} \xi_x + \frac{C}{|\mathcal{X}_p|} \sum_{X \in \mathcal{X}_p} \xi_X$$

subject to:

$$\begin{aligned}
w\,\phi(x) + b &\leq -1 + \xi_x, && \forall x \in \tilde{\mathcal{X}}_n \\
|w\,\phi(x) + b| &\geq +1 - \xi_x, && \forall x \in \tilde{\mathcal{X}}_p \\
w\,\frac{\phi(X)}{|X|} + b &\geq \frac{2 - |X|}{|X|} - \xi_X, && \forall X \in \mathcal{X}_p \quad (*) \\
\xi_x &\geq 0, \xi_X \geq 0
\end{aligned}$$

Sparse balanced MIL

Input:

- training bags $\mathcal{X}_n$ and $\mathcal{X}_p$
- feature representation $\phi(x)$
- capacity parameter C from sMIL
- balance parameter $\eta \in (0, 1]$

Output:

- decision function $f(x) = w\,\phi(x) + b$

Procedure:

- $(w, b) = \text{solve\_sMIL}(\mathcal{X}_n, \mathcal{X}_p, \phi, C)$
- order instances $x \in \tilde{\mathcal{X}}_p$ using $f(x)$
- label instances in $\tilde{\mathcal{X}}_p$:
  - the top $\eta|\tilde{\mathcal{X}}_p|$ as positive
  - the rest $(1 - \eta)|\tilde{\mathcal{X}}_p|$ as negative
- $(w, b) = \text{solve\_SIL}(\tilde{\mathcal{X}}_n, \tilde{\mathcal{X}}_p, \phi, C)$
- return (w, b)

Experimental Results

Table 1. Average area under ROC curve for each SVM method on each dataset.

Dataset	**SIL-SVM**	**NSK**	**STK**	**sMIL**	**sbMIL**	**stMIL**
AIMED	57.44	87.11	N/A	87.19	87.99	**92.11**
AIMED$\frac{1}{2}$	45.86	54.06	N/A	54.08	67.66	**72.94**
TIGER	76.65	79.07	80.80	81.12	**82.95**	74.48
ELEPHANT	85.08	82.94	85.22	87.98	**88.58**	81.64
FOX	52.72	64.01	62.14	66.13	**69.78**	60.67
MUSK1	87.82	85.61	69.44	86.91	**91.78**	79.46
MUSK2	87.33	**90.78**	61.01	81.19	87.74	68.41
TST1	96.25	97.16	96.19	97.29	**97.41**	96.81
TST2	85.37	**90.60**	86.87	87.97	90.57	88.55