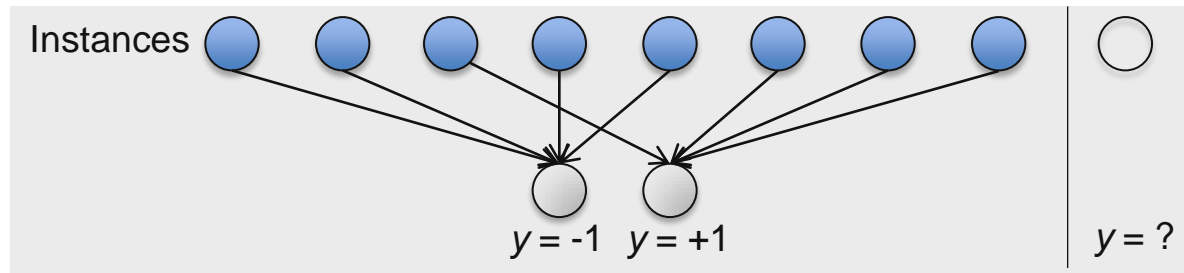


Multiple-Instance Learning with Instance Selection via Dominant Sets

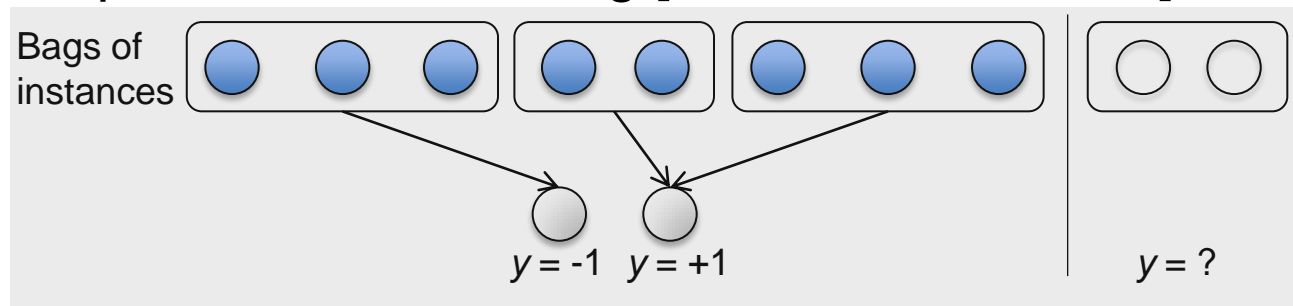
Aykut Erdem and Erkut Erdem
Hacettepe University,
Ankara, Turkey

Multiple-Instance Learning (MIL)

- Traditional Supervised (single instance) learning



- Multiple-instance learning [Dietterich et al. '97]

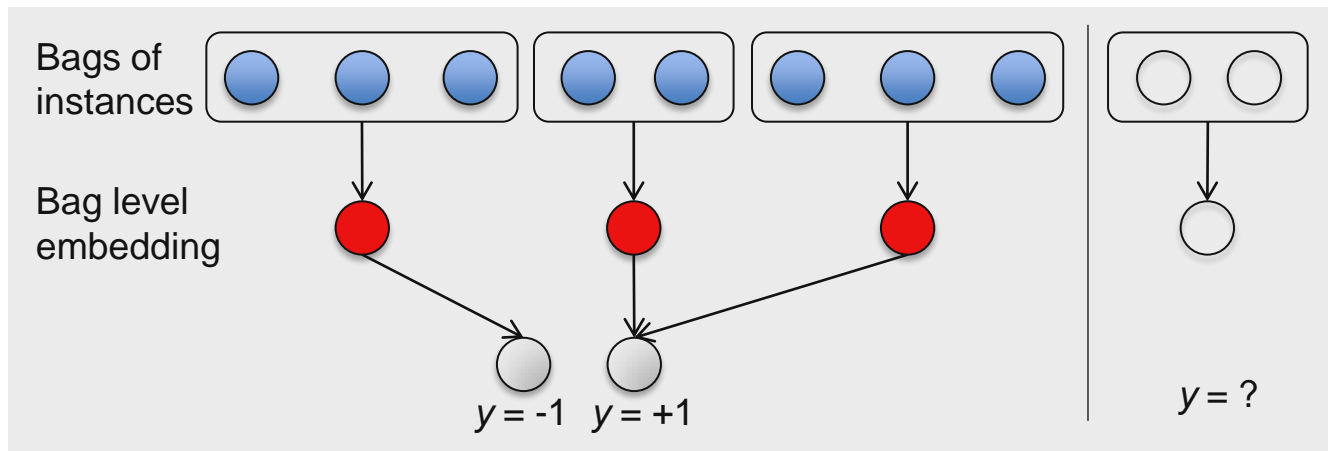


- Two MIL assumptions:
 - A bag is negative if all of its members are negative
 - A bag is positive if it contains at least one positive instance

learning with ambiguously labeled data

Instance-Selection based MIL

- Transform a MIL problem into a standard single-instance learning problem



- How to form the embedding space?
 - Select a set of representative instances (prototypes)
 - A similarity-based representation

Which instances best model the data? How many prototypes are needed? Robustness to outliers and labeling noise?

A comparison of Instance-Selection based MIL methods

Method	Prototypes	Classifier	Drawback
DD-SVM [Chen and Wang, 2004]	one from each training bag <i>DD function</i>	SVM + RBF	sensitive to labeling noise
MILES [Chen et al., 2006]	all the instances in the training bags	1-norm SVM <i>implicit instance selection</i>	exponentially expensive as the volume of the training data increases
MILD [Li and Yeung, 2010]	one from each pos. training bag <i>A conditional prob.model</i>	SVM + RBF	no neg. prototype
MILIS [Fu et al., 2011]	one from each training bag <i>A pdf for neg. instances based on KDE</i>	linear SVM	alternating instance selection and training (expensive)

Our suggestion: A clustering based approach
(based on *dominant sets* [Pavan and Pelillo, 2003, 2007])

Dominant Sets

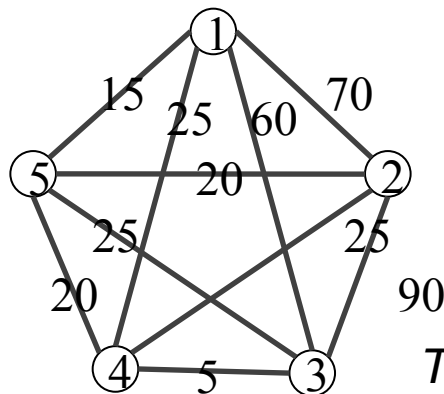
- a pairwise clustering approach
- makes no assumption on the underlying data representation
- detects the proper number of clusters and is very robust to outliers
- imposes no constraint on the structure of the similarity matrix, being able to naturally deal with asymmetric and negative similarities alike.
- can handle unseen data in a principled way

Dominant Sets

- A generalization of a maximal clique to edge-weighted graphs

Definition 1. A nonempty subset of vertices $S \subseteq V$ such that $\sum_{i \in T} w_T(i) > 0$ for any nonempty $T \subseteq S$, is said to be dominant if:

1. $w_S(i) > 0$, for all $i \in S$, (*internal homogeneity*)
2. $w_{S \cup \{i\}}(i) < 0$, for all $i \notin S$. (*external inhomogeneity*)



The set {1,2,3} is dominant

- *Dominant Sets \equiv Clusters*

Dominant Sets

Theorem 1. *Dominant sets can be computed as the support $\sigma(\mathbf{x})$ of the local maximizers of*

$$(1) \quad \begin{aligned} & \text{maximize} && f(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} \\ & \text{subject to} && \mathbf{x} \in \Delta \end{aligned}$$


where A is the weighted adjacency matrix of edge-weighted graph $G=(V, E, w)$, $\Delta=\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0} \text{ and } \mathbf{e}^T \mathbf{x}=1\}$ is the standard simplex in \mathbb{R}^n with \mathbf{e} being a vector of ones of appropriate dimension, and $\sigma(\mathbf{x})$ is defined as the set of indices corresponding to its positive component, i.e. $\sigma(\mathbf{x}) = \{i \in V \mid x_i > 0\}$.

- The objective function $f(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$ gives a measure of the cohesiveness of a cluster.
- The components of \mathbf{x} provides a measure of the participation of the corresponding data points in the cluster.
- The similarity of an element j to a cluster can be directly computed by the weighted similarity $(A\mathbf{x})_j$.

MIL with Instance Selection via Dominant Sets (MILDS)

- **Observation**: No ambiguity in the negative bags
- **Assumption**: Negative instances form clusters
 - may not be always valid (*outliers, labeling noise, etc.*)

Method	Prototypes	Classifier
DD-SVM [Chen and Wang, 2004]	one from each training bag <i>DD function</i>	SVM + RBF
MILES [Chen et al., 2006]	all the instances in the training bags	1-norm SVM <i>implicit instance selection</i>
MILD [Li and Yeung, 2010]	one from each pos. training bag <i>A conditional prob.model</i>	SVM + RBF
MILIS [Fu et al., 2011]	one from each training bag <i>A pdf for neg. instances based on KDE</i>	linear SVM
Our Approach (MILDS)	+ one from each cluster extracted from the negative data + one from each pos. training bag <i>Dominant Set clusters</i>	linear SVM



MILDS – Basic Notations

$B_i = \{B_{i1}, \dots, B_{ij}, \dots, B_{in_i}\}$ *i^{th} bag of instances*

$y_i \in \{+1, -1\}$ *label of i^{th} bag*

$B = \left\{ \underbrace{B_1^+, \dots, B_{m^+}^+}_{\text{positive bags}}, \underbrace{B_1^-, \dots, B_{m^-}^-}_{\text{negative bags}} \right\}$ *the set training bags*

$N = \{I_i \mid i = 1, \dots, M\}$ *the collection of neg. instances*
 $= \{B_{ij}^- \in B_i^- \mid i = 1, \dots, m^-\}$ *from all of the neg. training bags*

MILDS – Instance Selection (1)

- Pairwise similarities

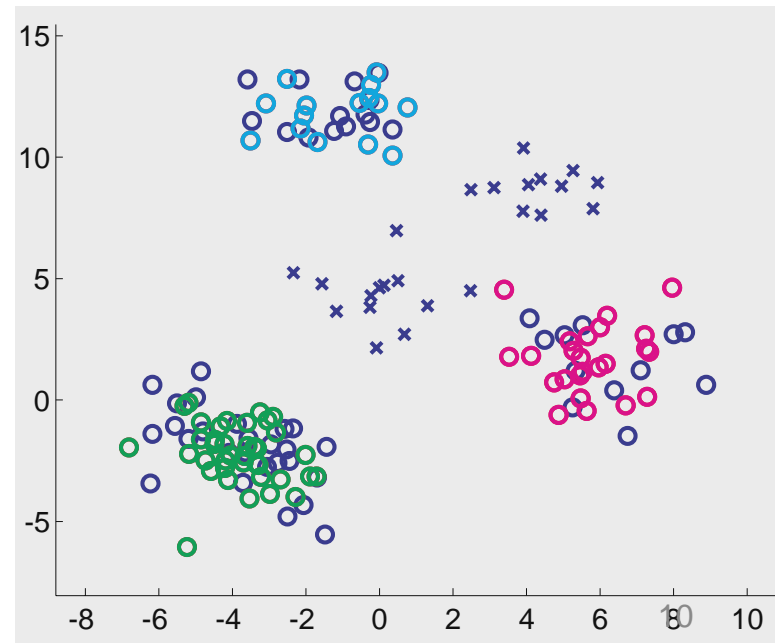
$$A = [a_{ij}] \quad a_{ij} = \begin{cases} \exp\left(-\frac{d(I_i, I_j)^2}{2\sigma^2}\right) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

- Extract the dominant sets (clusters)

from $N = \{B_{ij}^- \in B_i^- \mid i = 1, \dots, m^-\}$

- *A peeling-off strategy is used*
- *max m^- clusters*
(with the highest internal coherencies)

$$C = \{C_1, \dots, C_k\} \quad k \leq m^-$$

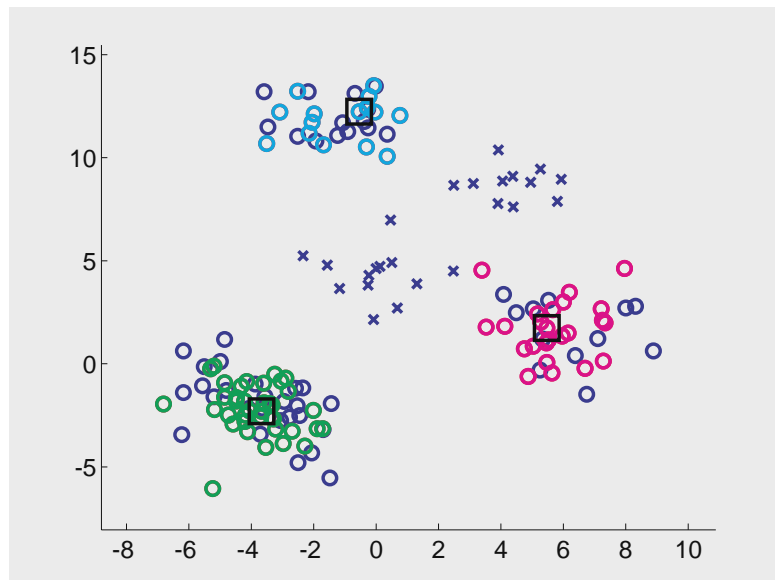


MILDS – Instance Selection (2)

- Select one prototype from each cluster $C_i \in \mathcal{C}$

$$z_i^- = l_{j^*} \quad \text{with } j^* = \arg \max_{j \in \sigma(x^{C_i})} x_j^{C_i}$$

The components of x^{C_i} gives us a measure of the participation of the corresponding instances in the cluster

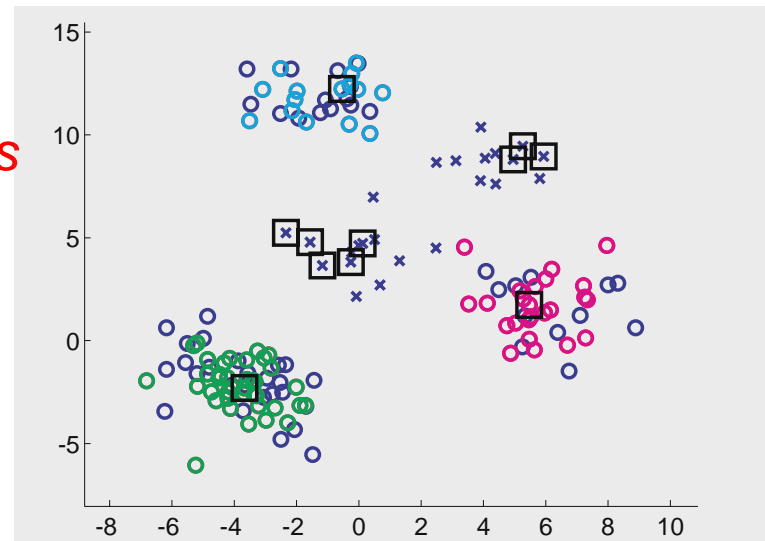


MILDS – Instance Selection (3)

- Ambiguity in positive bags
→ *clustering based selection strategy does not make sense!*
- Select the most positive (least negative) instance from each positive bag $B_i^+ = \{B_{i1}^+, \dots, B_{in_i^+}^+\}$

$$z_i^+ = B_{ij^*}^+ \quad \text{with } j^* = \arg \min_{j=1, \dots, n_i^+} \frac{\sum_{\ell=1, \dots, k} (A^T x^{C_\ell})_j \times |C_\ell|}{\sum_{\ell=1, \dots, k} |C_\ell|}$$

*the most distant instance from
the extracted (negative) clusters*



Classification

- Set of prototypes $Z = \{z_1^-, \dots, z_k^-, z_1^+, \dots, z_{m^+}^+\}$
- A similarity measure of a bag to an instance prototype:

$$s(z, B_i) = \max_{B_{ij} \in B_i} \exp\left(-\frac{d(z, B_{ij})^2}{2\sigma^2}\right)$$

based on the distance between z and its nearest neighbor in B_i

- An embedding function:

$$\phi(B) = [s(z_1^-, B), \dots, s(z_k^-, B), s(z_1^+, B), \dots, s(z_{m^+}^+, B)]^T$$

- The classifier: linear SVM $f(B; w) = w^T \phi(B) + b$
 $y(B) = \text{sign}(f(B; w))$

One-vs-rest Multi-Class MILDS

- Train C binary classifiers
 - one for each class against all other classes.
- Classification:

$$y(B) = \arg \max_{i=1, \dots, c} f_i(B; \mathbf{w}_i)$$

*A different instance-based embedding
for each binary subproblem*

mILDS (1)

- A second multi-class extension of MILDS
- Construct an embedding space common for all classes

→ *the same selected set of instances for all classes*

$$\varphi(B) = [s(z_1^1, B), s(z_2^1, B), \dots, s(z_{m_1}^1, B), \\ s(z_1^2, B), s(z_2^2, B), \dots, s(z_{m_2}^2, B), \\ \vdots \\ s(z_1^c, B), s(z_2^c, B), \dots, s(z_{m_c}^c, B)]$$

*training data is kept the same for all binary sub-problems
(only the labels differ)*

→ *makes the training phase much more efficient!*

mIDS (2)

- For each class k ,
 - Consider $I^k = \{I_i^k \mid i = 1, \dots, M_k\}$
 $= \{B_{ij} \in B_i \mid \text{for all } B_i \in B \text{ with } y(B_i) = k\}$
 - Extract the clusters in I^k
 $C^k = \{C_1^k, \dots, C_{m_k}^k\}$
 - Select one prototype from each extracted cluster C_i^k

$$z_i^k = I_{j^*}^k \quad \text{with } j^* = \arg \max_{j \in \sigma(x^{C_i^k})} \frac{x_j^{C_i^k}}{\beta_{ik}(j)}$$

the degree of participation
the similarity to all the remaining classes

The most dissimilar instance to the other training data from other classes

$$\beta_{ik}(j) = \max_{\substack{m=1, \dots, c \\ m \neq k}} \frac{\sum_{C_\ell^m \in C^m} (A_{km} x^{C_\ell^m})_j \times |C_\ell^m|}{\sum_{C_\ell^m \in C^m} |C_\ell^m|}$$

consider only the similarity to the most closest class

Experiments

- Two kinds of tasks
 - Benchmark data sets (*2-class*)
 - Image classification (*multi-class*)

Benchmark Data Sets

- 10 times 10-fold cross validation

[*except MIForest (over 5 runs) and MILIS and MIO (over 15 runs)]*

Table 2. Classification accuracies of various MIL algorithms on standard benchmark data sets. The best performances are indicated in bold typeface.

Algorithm	<i>Musk1</i>	<i>Musk2</i>	<i>Elephant</i>	<i>Fox</i>	<i>Tiger</i>
MILDS	90.9	86.1	84.8	64.3	81.5
MILD_B [13]	88.3	86.8	82.9	55.0	75.8
MILIS [8]	85.6	86.5	78.7	61.6	83.1
MILES [4]	83.3	91.6	84.1	63.0	80.7
DD-SVM [5]	85.8	91.3	83.5	56.6	77.2
MILD_I [13]	89.9	88.7	83.2	49.1	73.4
MIForest [10]	85.0	82.0	84.0	64.0	82.0
MIO [12]	88.3	87.7	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>
Ins-KI-SVM [14]	84.0	84.4	83.5	63.4	82.9
Bag-KI-SVM [14]	88.0	82.0	84.5	60.5	85.0
mi-SVM [1]	87.4	83.6	82.2	58.2	78.9
MI-SVM [1]	77.9	84.3	81.4	59.4	84.0
EM-DD [24]	84.8	84.9	78.3	56.1	72.1

- The performance of MILDS is competitive with all the state-of-the-art MIL methods.

Benchmark Data Sets – The Dimensions of the Embedding Spaces

- MILES has the highest embedding space dimension.
- MILD_B has the lowest dimension but its performance is poor.
- As compared to MILIS and DD-SVM, MILDS has dimensions ~6—23% smaller (except for Musk2 and Fox)

Table 1. the MIL benchmark data sets

data set	bags		dim
	pos./neg.	avg. inst./bag	
<i>Musk1</i>	47/45	5.17	166
<i>Musk2</i>	39/63	64.69	166
<i>Elephant</i>	100/100	6.96	230
<i>Fox</i>	100/100	6.60	230
<i>Tiger</i>	100/100	6.10	230

Table 3. The dimensions of the embedding spaces averaged over 10 runs of 10-fold CV

Algorithm	<i>Musk1</i>	<i>Musk2</i>	<i>Elephant</i>	<i>Fox</i>	<i>Tiger</i>
MILDS	75.0	92.0	169.4	180.0	139.2
MILD_B	42.4	35.2	90.0	90.0	90.0
MILIS	83.0	92.0	180.0	180.0	180.0
MILES	429.4	5943.8	1251.9	1188.0	1098.0
DD-SVM	83.0	92.0	180.0	180.0	180.0

- The dimensions can be further reduced by employing 1-norm linear SVM in the training step, or eliminating the dimensions with very small weights.

Image Categorization

- 2000 images from 20 categories, each having 100 examples
- Each image (bag) is segmented and then represented with regions of interest (instances in the bag)
- Two groups of experiments:
 - 1000-Image → Only the first 10 categories
 - 2000-Image → All the 20 categories
- Two possible extensions: MILDS and mildS

Fig. 2. Example images randomly drawn from the COREL data set. For each category, the average number of regions per image is given inside the parentheses.

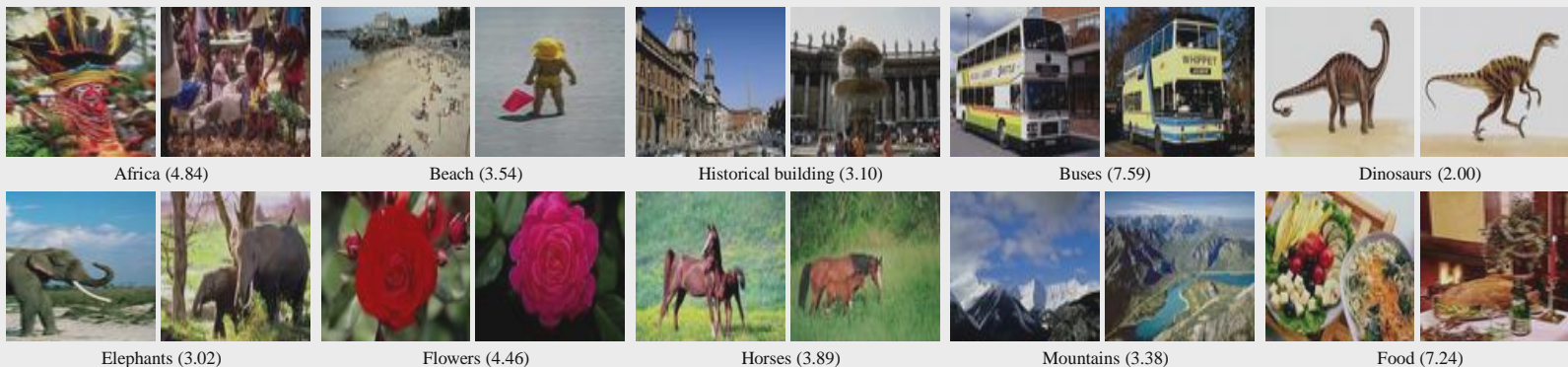


Image Categorization

- 5 times 2-fold cross validation

Table 4. Classification accuracies of various MIL algorithms on COREL *1000-Image* and *2000-Image* data sets. The best performances are indicated in bold typeface.

Algorithm	<i>1000-Image</i>	<i>2000-Image</i>
miIDS	82.2	70.6
MILDS	83.0	69.4
MILD_B [13]	79.6	67.7
MILIS [8]	83.8	70.1
MILES [4]	82.6	68.7
DD-SVM [5]	81.5	67.5
MIForest [10]	59.0	66.0
MissSVM [26]	78.0	65.2
mi-SVM [1]	76.4	53.7
MI-SVM [1]	74.7	54.6

- The performance of MILDS and miIDS are competitive.
- For *2000-Image*, miIDS gives the best result.

Image Categorization – Selected Instance Prototypes (MILDS)

- In MILDS, each classifier is trained for distinguishing a specific category from the rest.
 - A different embedding space is built for each subproblem
 - The set of selected prototypes varies in every subproblem
- Positive prototypes are mostly selected from the discriminative regions for that class.

Fig 3. Sample instance prototypes selected by the *MILDS* algorithm. For each image category, the first row shows a sample training image from that category, and the bottom row illustrates the selected prototype region (shown in white) on the corresponding segmentation map.

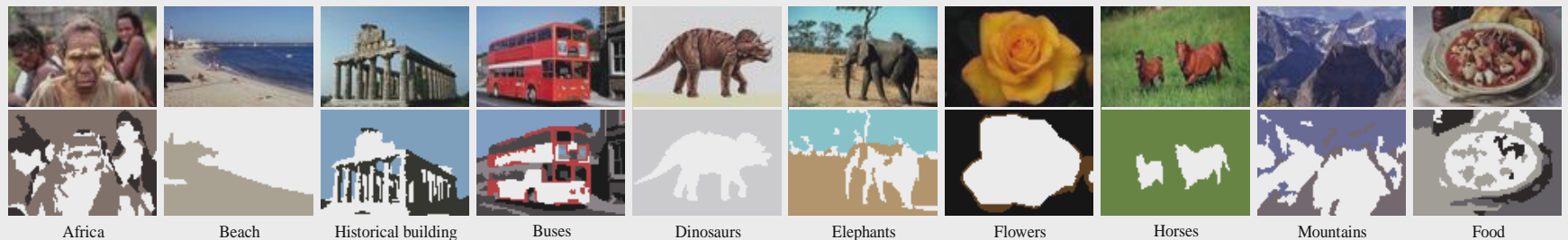
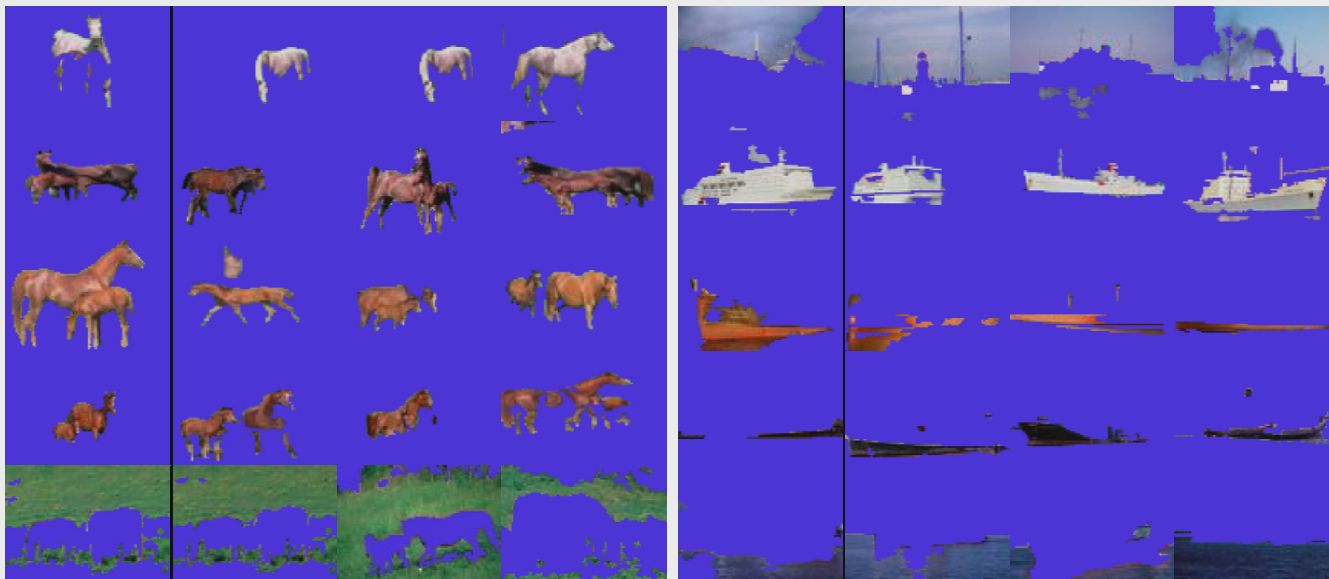


Image Categorization – Selected Instance Prototypes (miDS)

- In miDS, the set of selected instance prototypes is the same for all the subproblems
 - provides a rich way to include context
 - resembles the vocabulary generation step of bag-of-words*

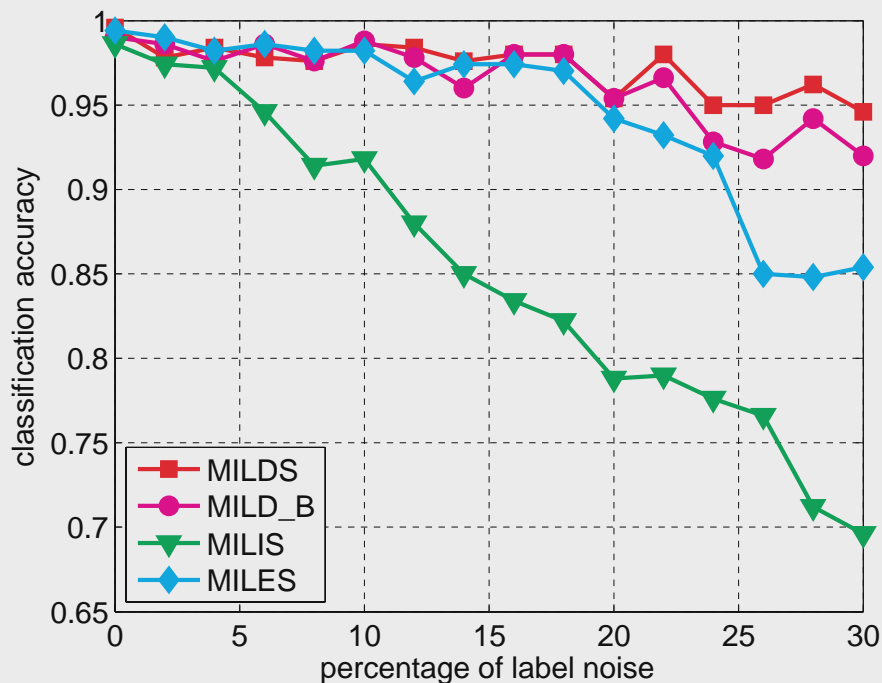
Fig. 4. Sample instance prototypes selected by the *miDS* algorithm for the *Horse* and the *Battle ships* categories. The leftmost columns are the prototypes. The rightmost three columns show other sample regions from the corresponding extracted clusters. The regions in each cluster share similar visual characteristics.



Sensitivity to Labeling Noise

- *Historical buildings vs. Horses (2 class)*
- Noisy labels
 - For each noise level, $d\%$ of pos. and $d\%$ of neg. images are randomly selected from the training set and then their labels are exchanged
- 5 times 2-fold cross validation
- At low levels ($d \leq 5\%$), there is no considerable difference in the performances
- At high levels ($d \geq 25\%$), MILDS is the most robust one

Fig. 5. Sensitivity of various MIL algorithms to labeling noise.



→ Dominant sets is quite robust to outliers

Summary and Future Directions

- A new instance selection strategy based on dominant sets
- Identifies the most representative examples in the positive and negative training bags
- Competitive with state-of-the-art MIL methods
- Quite robust to labeling noise
- Future directions
 - Multi-instance multi-label learning
[Zhou and Zhang, 2006, Zha et al., 2008]
 - Non-i.i.d. samples
[Zhou et al., 2009, Warrell and Torr, 2011]

- Any questions?