UNIVERSIDAD
D
CÓRDOBA

KDIS Lab

Multiple Instance Learning: An Introduction

Sebastián Ventura

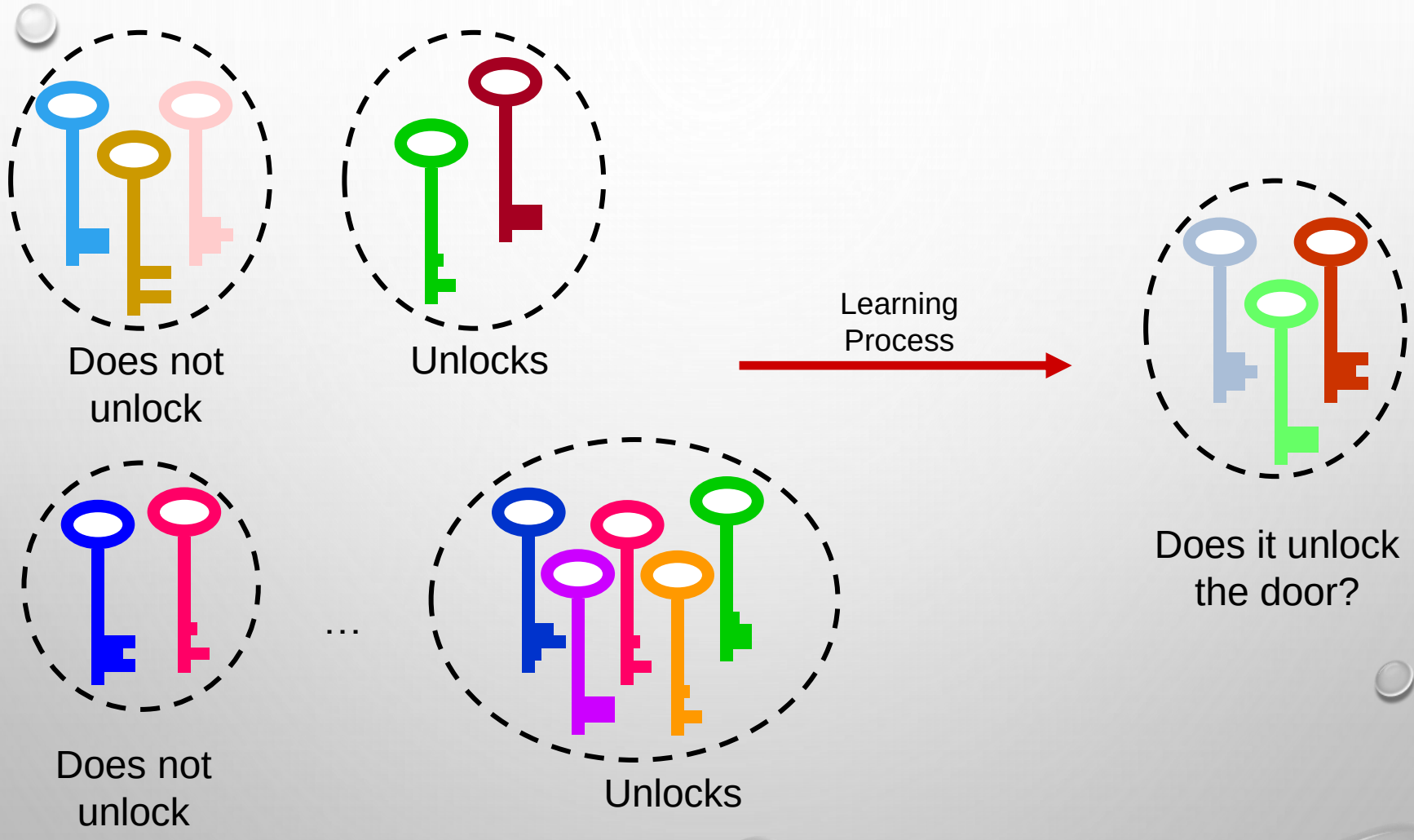
Knowledge Discovery and Information Systems (KDIS) Research Group
University of Córdoba. Córdoba (SPAIN)

Contents

- Introduction
  - Multiple Instance Learning (MIL)
  - Multiple Instance Learning Problems/Paradigms
  - Some Interesting Applications of MIL
- MIC Algorithms
  - Foundations
  - A Recent Taxonomy for MIC algorithms
    - Instance Space Paradigm
    - Bag Space Paradigm
    - Embedded Space Paradigm
  - Conclusions
- Other MIL paradigms
  - Multiple Instance Regression
  - Multiple Instance Clustering
  - Multi-instance Multilabel Classification
- Open Problems in MIL
- Internet Resources

Introduction

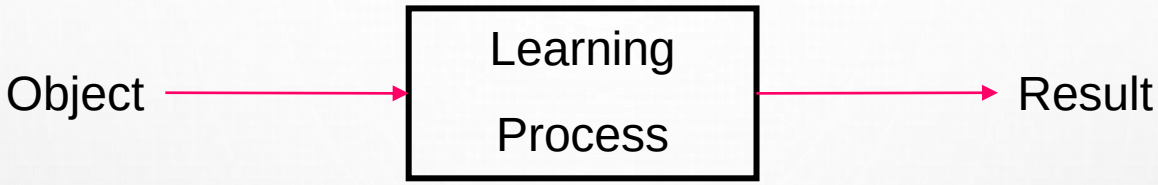
An Easy Example



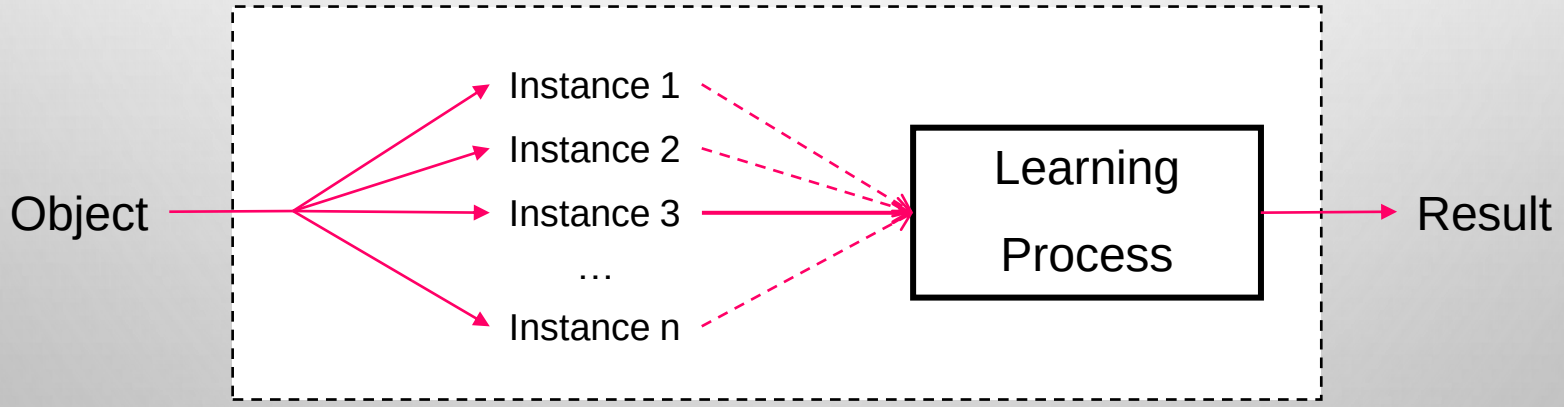
Multiple Instance Learning

- Multiple instance (or multi-instance) learning appears in complex applications of machine learning where the learner has partial or incomplete knowledge about each training example.
- Each training example can be represented by means of a bag or objects composed of one or several feature vectors. Each bag represent a different view (global o partial) of the object.
- The learner only knows that each example can be represented by one of a set of potential feature vectors instead of knowing which particular instance or set of them represent the concept which we want to learn.

Classical Learning vs. Multi-instance Learning



Classical learning



Multiple instance learning

Multi-instance Learning Problems/Paradigms

- *Multi-instance Classification*. The objective is to predict unseen bag labels:
  - *Binary Classification*: Binary label
  - *Multiple Classification*: Nominal (non-binary) label
  - *Multi-Label Classification*: Multiple labels (MI-MLL)
- *Multi-instance Regression*. The objective is to predict the continuous label of unseen bags.
- *Multi-instance Clustering*. Grouping similar objects of bags in clusters (unsupervised learning task).

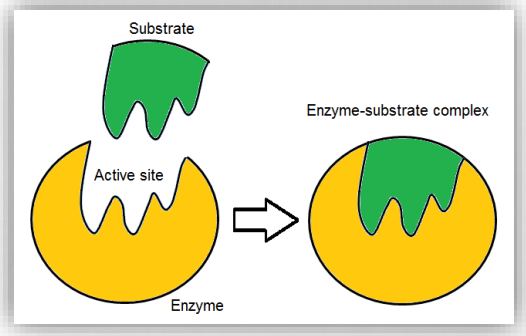
This presentation is focussed on Multi-instance Binary Classification

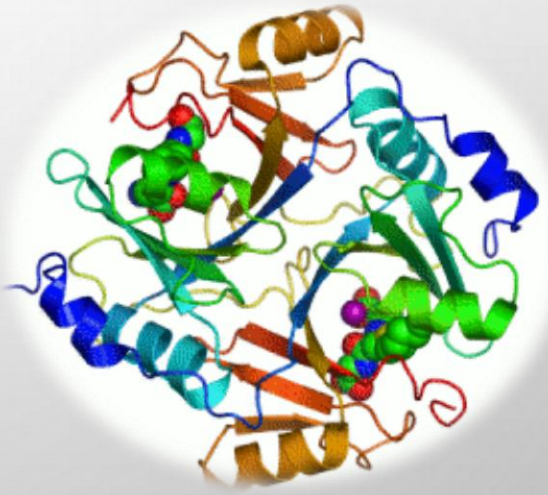
Some Real-World Applications of Multiple Instance Classification

Introduction

Prediction of Pharmacological Activity

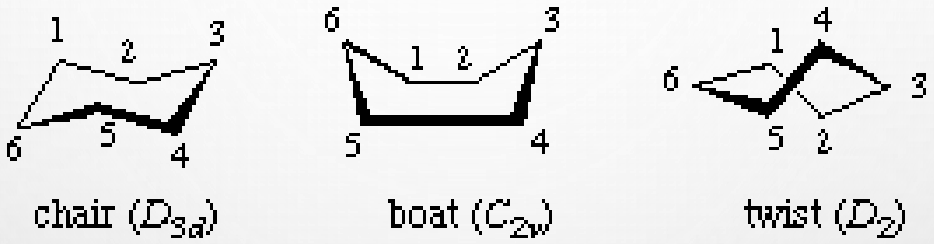
- The first paper on MIL (Dietterich et al., 1997) was motivated by the problem of determining whether a drug molecule exhibits a given activity.
- A molecule presents a given pharmacological activity when it is able to bind with an enzyme or protein. This is only possible if the molecule has certain spatial properties (*key-lock mechanism*).





Prediction of Pharmacological Activity(II)

- A molecule may adopt a wide range of shapes or *conformations*, due to the rotation of its bonds.



- If a conformation can bind/connect to a pharmacological activity center, the whole molecule exhibits the activity under research. Otherwise, the molecule does not exhibit this activity.
- In Dietterich's paper, the property under study was musk. Substances with this property are employed in the manufacture of perfumes and other cosmetic products.

Prediction of Pharmacological Activity(III)

- This problem can be represented by multi-instances in a very natural way:
  - Each molecule is a bag
  - Each conformation is an instance
- Dietterich et al. studied two different datasets:
  - Musk-1: 92 molecules (47 positive y 45 negative), 476 instances and 166 attributes.
  - Musk-2: 102 molecules (39 positive y 63 negative), 6598 instances y 166 attributes.
- There exist other benchmarks related to the pharmacological activity prediction problem. For instance, in mutagenesis dataset, the property under study is the ability to produce mutations.
  - Mutagenesis 1: 188 molecules, 10,468 instances, 7 attributes
  - Mutagenesis 2: 42 molecules, 2,132 instances, 7 attributes

These an other bechmark datasets can be found at http://www.uco.es/grupos/kdis/mil/dataset.html

Content-based Image Classification and Retrieval

- The key to the success of image retrieval and classification is the ability to identify the intended target object(s) in images.
- This problem is complicated when the image contains multiple and possibly heterogeneous objects.
- This problem can fit into the MIL setting well:
  - Each image itself is considered as a bag.
  - A region or segment in an image is considered to be an instance.

Content-based Image Classification and Retrieval (II)

The first paper that describes this kind of application is that of Maron and Lozano-Pérez (1998). These authors face the problem of describing persons from complex images.

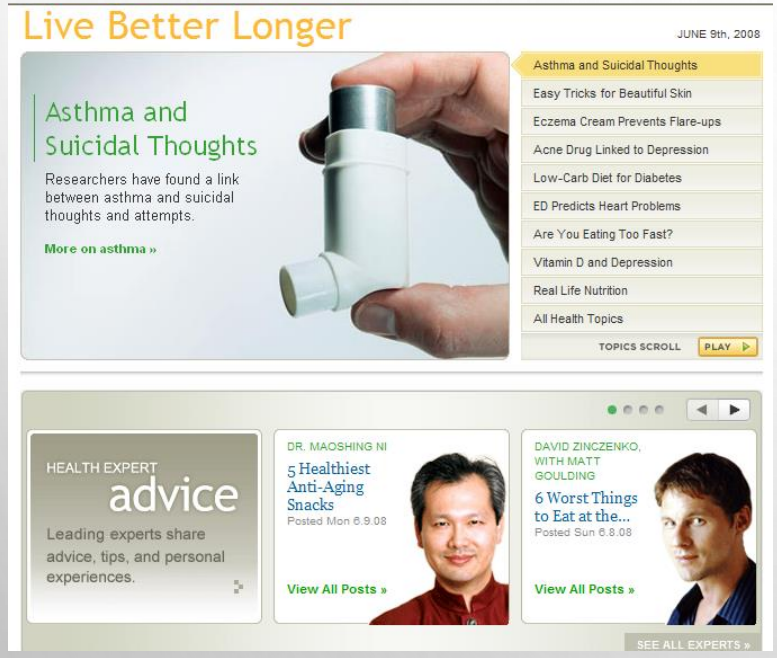
O. Maron & T. Lozano Pérez. A Framework for Multiple-Instance Learning *In Proc. of the 1997 Conference on Advances in Neural Information Processing Systems* (1998) pg. 570-576.

Text categorization

- Andrews et al (2002) use MIL to categorize documents taken from the TREC9 dataset (a benchmark in text categorization problems).
- They divide each document in 50 word length overlapping sets (authors do not specify what this overlapping is like). In this case, each 50 word set represents an instance and the whole document is a training pattern (bag of instances).

S. Andrews, I. Tsochantaridis & T. Hofmann. Support vector machines for Multiple Instance Learning. In *Advances in Neural Information Processing Systems (NIPS 15)*, pp 1-8, 2002

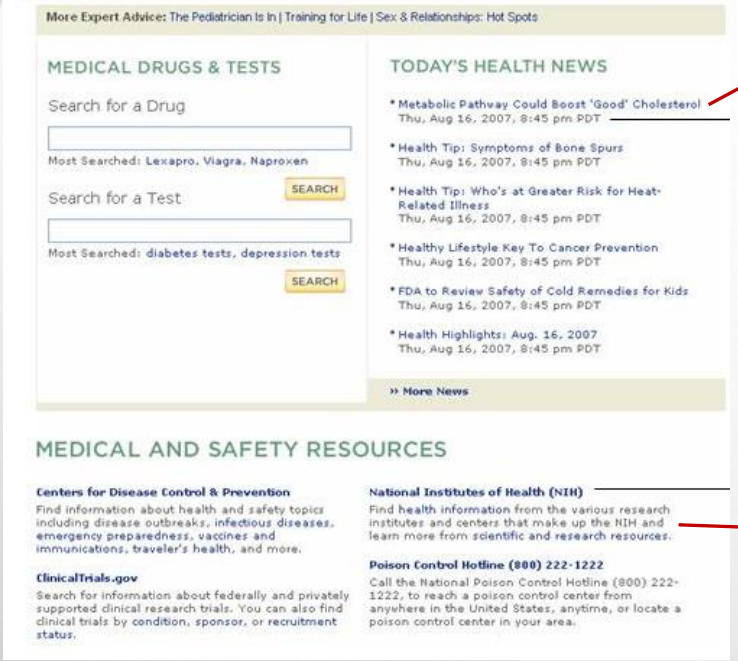
Web index recommendation



- Web index pages are pages that provide titles or brief summaries and leave the detailed presentation to their linked pages.
- The problem of recommending web index pages consists of determining what pages a given user is interested in.
- In general, if a web index page contain links that the user considers interesting, the user will be attracted to it.
- The problem is that we do not have information about links, but about the page as a whole.

Web index recommendation (II)

PATTERN/ BAG



INSTANCES (one per link)

A web index page is represented as a bag, and each link in its web index page is represented by an instance.

Detecting Fraudulent Users in Personal Banking

- The idea is to detect fraudulent use of credit cards from a transactions dataset
- For each user, we have one or more transactions defined by several attributes:
  - amount
  - time
  - transaction interval
  - service ID
  - merchant type
  - …
- This problem can be also represented as a multi-instance where a bag represents all the transactions of a given user and the label will tell if the use has been fraudulent or not

MIC Algorithms

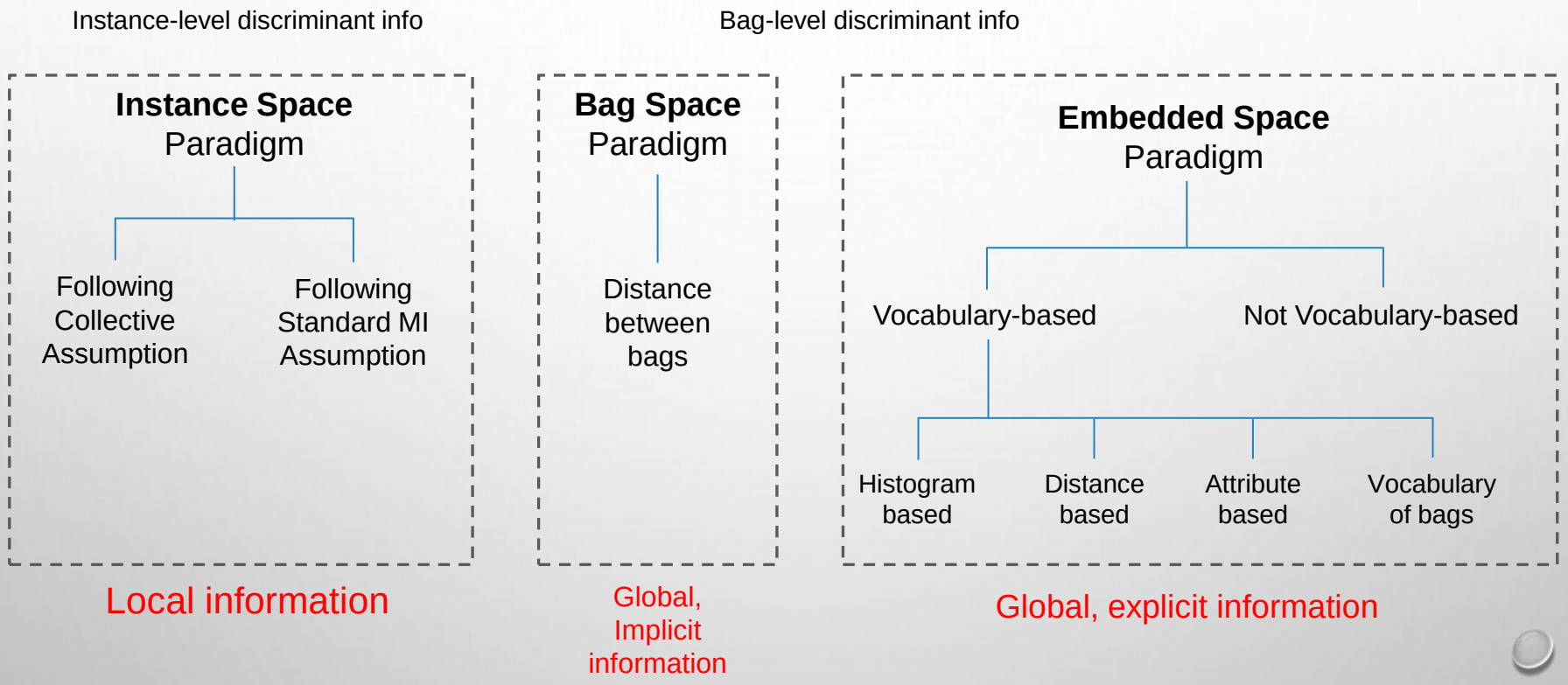
Definitions

- A bag X is a set $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$ where $\vec{x}_i$ are feature vectors, called *instances* and the cardinality *N* can vary across the bags.
- All the instances $\vec{x}_i$ live in a d-dimensional feature space $\vec{x}_1 \in \mathcal{R}^d$, called *instance space*.
- The objective of MIC problems is to learn a model, at training time, that can be used to predict class labels of unseen bags.
- In binary classification, we will learn a function $F(X) \epsilon [0,1]$ that provides the likehood that X is positive.
- In order to obtain $F(X)$, we are given a training set $\mathcal{T}$ with *M* bags and their corresponding labels

$$\mathcal{T} = \{(X_1, y_1), (X_2, y_2), \dots, (X_N, y_N)\}$$

where $y_i \epsilon \{0,1\}$ is the label of X_i.

A New Taxonomy For Multi-instance Algorithms



J. Amores. Multiple Instance Classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201, 81–105 (2013)

Instance-Space Paradigm

MIC Algorithms

Introduction

- The idea is to infer an instance based classifier $f(\vec{x})\epsilon[0,1]$ from the training data.
- Bag level classification $F(X)\epsilon[0,1]$ is constructed as an aggregation of instance-level responses:

$$F(X) = \frac{f(\vec{x}_1) \circ f(\vec{x}_2) \circ \cdots \circ f(\vec{x}_N)}{Z}$$

where $\circ$ represents an *aggregation operator* and Z is a normalization factor

- These methods have to solve the problem of how to infer an instance-level classifier **without** having access to a training set of labelled instances.
- To do this, some *hypothesis* has to be made about the relationship that exists between the label of the bags and the labels of the instances contained in the bags.
- There are two main hypothesis:
  - Standard hypothesis
  - Collective hypothesis

Standard Hypothesis

- Every positive bag contains at least one positive instance, while in every negative bag all the instances are negatives.
- The methods following this hypothesis try to identify the type of instance that make the bag positive.
- There are several classical methods that follow this hypothesis:
  - Learning of Axis-Parallel Rectangles (APR)
  - Diverse Density
  - MI-SVM
  - Sparse MIL and Sparse Balanced MIL
  - Adaptations of Single Instance Learning (SIL) algorithms to MIL. According Z.-H. Zhou (2009), SIL algorithms can be adapted to MIL including the standard hypothesis in their development.
    - Decision Trees
    - Rule Based Learning
    - G3P-MI and MO G3P-MI

Learning of Axis Parallel Rectangles (APRS)

- The first solution to the multiple instance learning problem was proposed by Dietterich et al. 1997

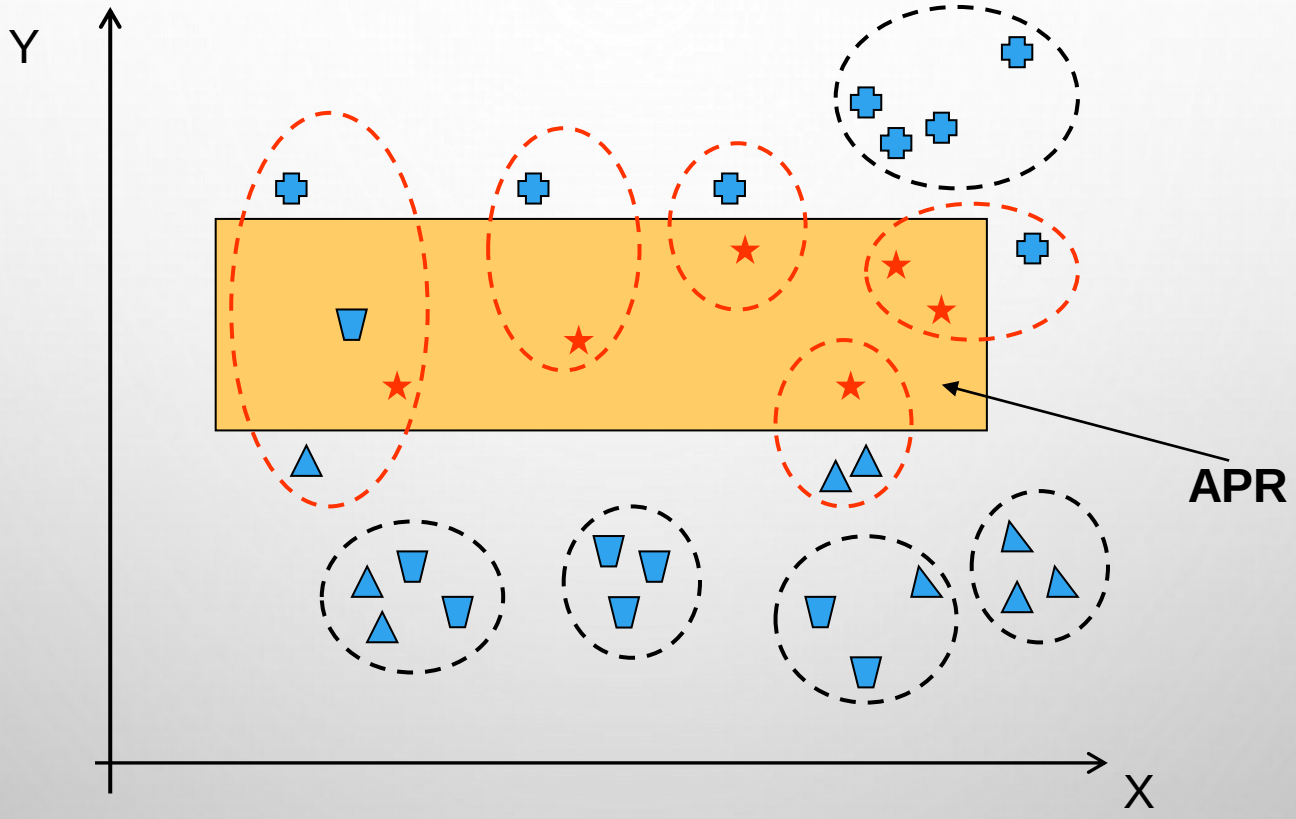
  T. G. Dietterich, R.H Lathrop & T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89:1-2 (1997), pp 31-71

- They propose representing the concept to be learned by axis parallel rectangles (APR) in the feature space. Intuitively, this APR should contain at least one instance from each positive example and meanwhile exclude all the instances from negative examples.

$$f(\vec{x};\mathcal{R}) = \begin{cases} 1 & if\ \vec{x} \in \mathcal{R} \\ 0 & otherwise \end{cases}$$

$$F(X) = \max_{\vec{x} \in X} f(\vec{x})$$

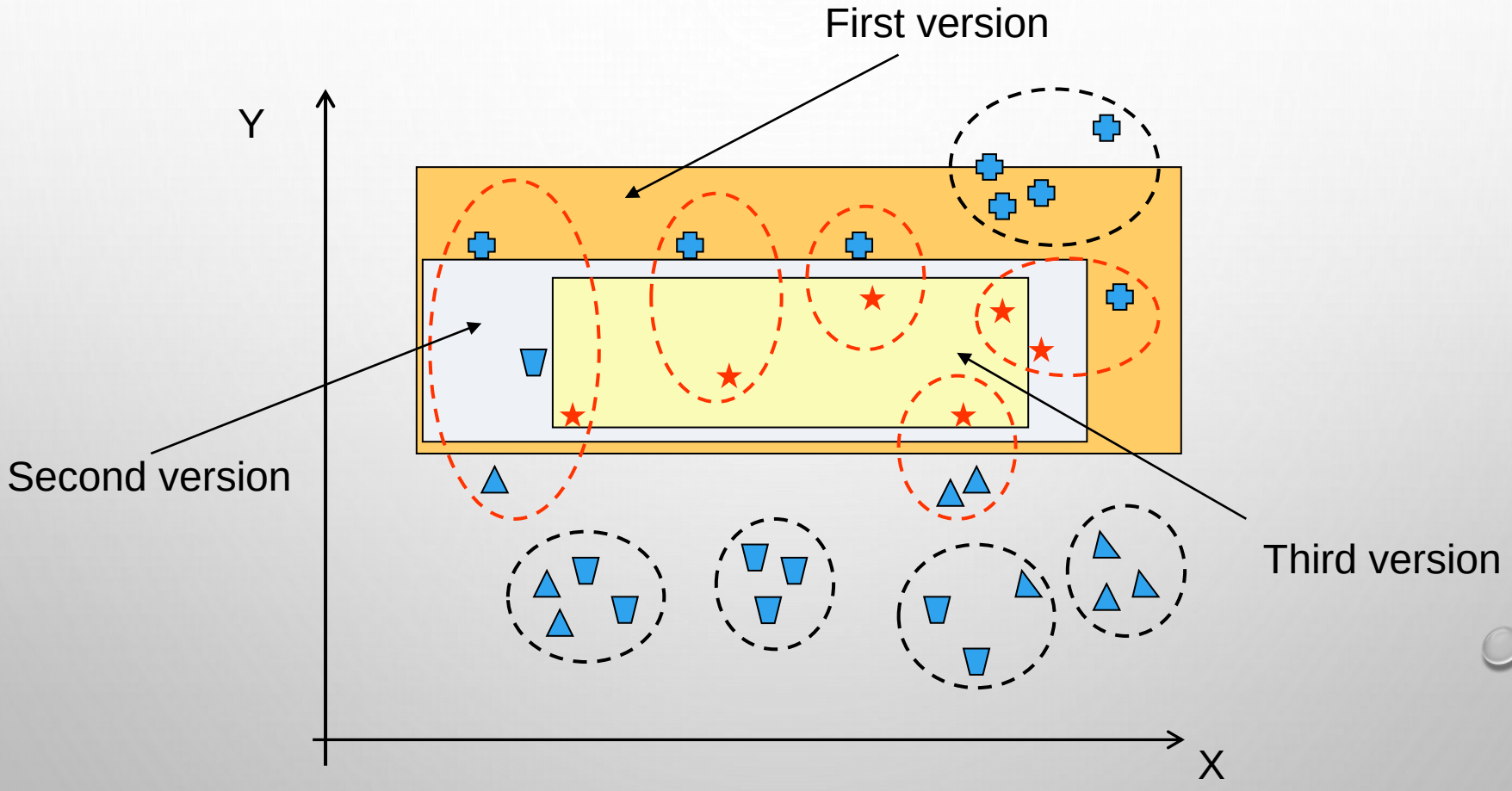
Graphical Description of APR



Variants of the APR Algorithm

- Dietterich's article considers three general designs for APR algorithms:
  - A noise-tolerant "standard" algorithm. The naive APR algorithm just forms the smallest APR that bounds the positive examples.
  - An "outside-in" algorithm. This algorithm is a variation on the "standard" algorithm. It constructs the smallest APR that bounds all of the positive examples and then shrinks this APR to exclude false positives.
  - An "inside-out" algorithm. This algorithm starts with a seed point in the feature space and "grows" a rectangle with the goal of finding the smallest rectangle that covers at least one instance of each positive example and no instances of any negative example.
- Authors apply these algorithms and several supervised learning algorithms (C4.5 and BP-ANN) to three datasets: Musk 1, Musk 2 and a synthetic dataset.
- In general, results show that APR algorithms outperform supervised learning algorithms in this kind of problem.

Variants of APR
(Graphical Description)



APR Versus Non-MIL Algorithms

Dataset Musk-1

Algorithm	TP	FN	FP	TN	Errors	% correct
Iterated discrim APR	42	5	2	43	7	92.4
GFS elim-kde APR	46	1	7	38	8	91.3
GFS elim-count APR	46	1	8	37	9	90.2
GFS all-positive APR	47	0	15	30	15	83.7
All-positive APR	36	11	7	38	18	80.4
Backpropagation	45	2	21	24	23	75.0
C4.5 (pruned)	42	5	24	21	29	68.5

APR Versus Non-MIL Algorithms

Dataset Musk-2

Algorithm	TP	FN	FP	TN	Errors	% correct
Iterated discrim APR	30	9	2	61	11	89.2
GFS elim-kde APR	32	7	13	50	20	80.4
GFS elim-count APR	31	8	17	46	25	75.5
All-positive APR	34	5	23	40	28	72.6
Backpropagation	16	23	10	53	33	67.7
GFS all-positive APR	37	39	32	31	34	66.7
C4.5 (pruned)	32	7	35	28	42	58.8

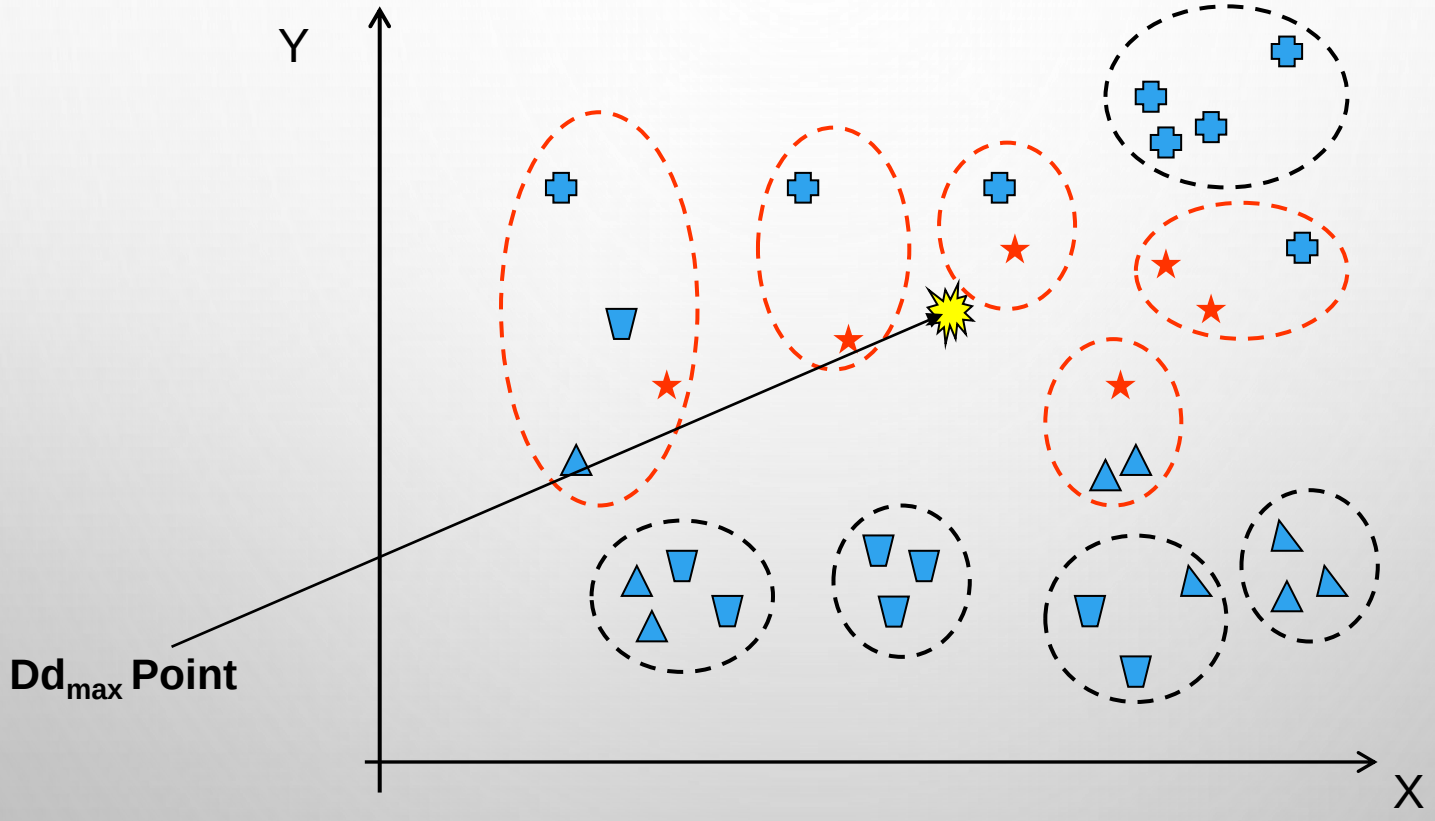
Diverse Density

- One of the most popular learning algorithms in multi-instance learning is diverse density (DD), proposed by Maron & Lozano Pérez in 1998.

  O. Maron & T. Lozano Pérez. A Framework for Multiple-Instance Learning *In Proc. of the 1997 Conference on Advances in Neural Information Processing Systems* (1998) pp 570-576.

- The main idea of the DD approach is to find a concept point in the feature space that is close to at least one instance from every positive example and meanwhile far away from instances in negative examples.
- The optimal concept point is defined as the one with the maximum ***diversity density***, which is a measure of how many different positive bags have instances near the point, and how far the negative instances are away from that point.

Diverse Density
(Graphical Description)



Diverse Density
(Mathematical Formulation)

Let B_i^+ be a positive object and y B_i^- a negative object, then diverse density is defined as

$$DD(x) = P(x \mid B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_m^-)$$

and the point $x=x_{max}$, that represents the concept that has to be learned, has a maximum value for *DD*.

Using Bayes rule and assuming prior uniformity at the concept location, this is equivalent to maximizing the following likelihood

$$P(B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_m^- \mid x=x_{max})$$

Decision Trees and Rule Based Systems

- Y. Chevaleyre and J.D. Zucker adapted the algorithms ID3 (decision trees) and RIPPER (rule induction) to the multiple instance learning paradigm.

  Y. Chevaleyre & J.-D. Zucker. Solving Multiple-Instance and Multiple-Part Learning Problems with Decision Trees and Rule Sets. Application to the Mutagenesis Problem. In E. Stroulia & S. Matwin (Eds): *AI 2001*, LNAI 2056, pp 204-214, 2001.

  Y. Chevaleyre & J.-D. Zucker. A Framework for Learning Rules from Multiple Instance Data. In L. de Raedt & P. Flach (Eds.) *ECML 2001*, LNAI 2167, pp 49-60, 2001.

- These adaptations are based on adapting the concepts of entropy and information gain to the multi-instance context.

Decisión Trees and Rule Based Sytems (II)

$$\text{Entropy}(S) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$\text{Entropy}_{multi}(S) = -\frac{u(S)}{u(S)+v(S)}\log_2\frac{u(S)}{u(S)+v(S)} - \frac{v(S)}{u(S)+v(S)}\log_2\frac{v(S)}{u(S)+v(S)}$$

u = set of positive bags
v = set of negative bags

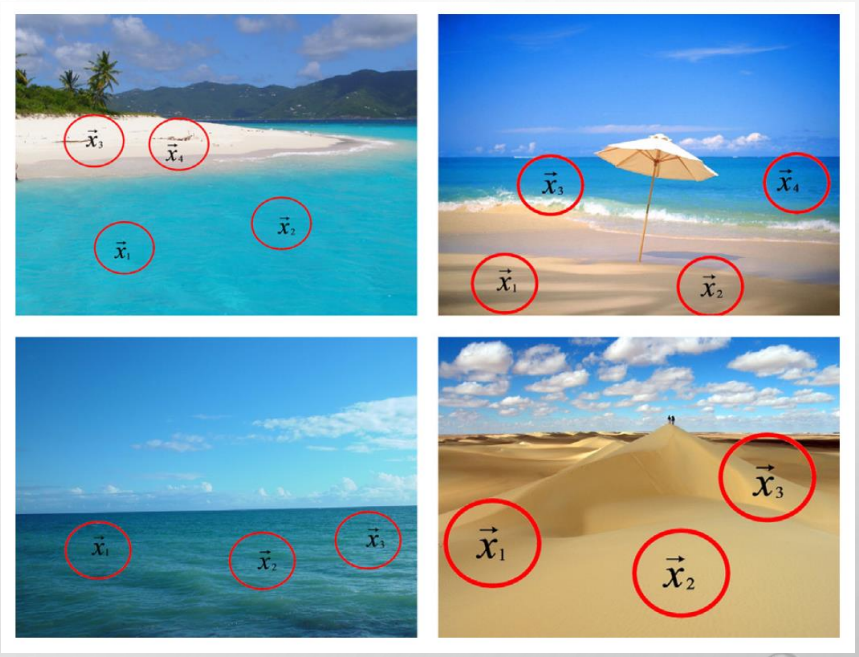
$$\text{InfoGain}(S,F) = \text{Entropy}(S) - \sum_{v \in \text{Values}(F)} \frac{p_v + n_v}{p+n}\text{Entropy}(S_v)$$

$$\text{InfoGain}_{multi}(S,F) = \text{Entropy}_{multi}(S) - \sum_{v \in \text{Values}(F)} \frac{u(S_v)+v(S_v)}{u(S)+v(S)}\text{Entropy}_{multi}(S_v)$$

For adapting the RIPPER, these authors adapt the concept of coverage to the context of multi-instance objects (bags).

Collective Hypothesis

- There are problems where the standard hypothesis does not yield good results.
- The collective hypothesis assumes that **all instances contributes equally to the bag's label**
- Collective hyphotesis can also work well in problems like Musk, because all the instances inside a bag might contribute in certain way to concept associated to the bag.



Concept *beach* is associated with the appearance of <u>both</u> *sand* and *water*, not one of them only

Collective Hypothesis (II)

Algorithms based on the collective hypothesis operate as follows:

1. They use a training set where all the instances inherits the label of the bag where it lies.
2. Then, they train a supervised learning classifier $f(\vec{x})$ using this dataset.
3. Finally, they build the bag classifier $F(X)$ aggregating the instance level predictions.

They different algorithms described in the bibliography only differ in the aggregation method they use to build $F(X)$.

SIL and Wrapper MI Algorithms

SIL

- This is the simplest collective algorithm
- Uses the sum as aggregation rule

$$F(X) = \frac{1}{|X|}\sum_{\vec{x}\in X} f(\vec{x})$$

Wrapper MI

- Uses a weighted sum as aggregation method

$$F(X) = \frac{1}{|X|}\sum_{\vec{x}\in X} w(\vec{x})f(\vec{x}) \qquad w(\vec{x}) = \frac{S}{|X|}$$

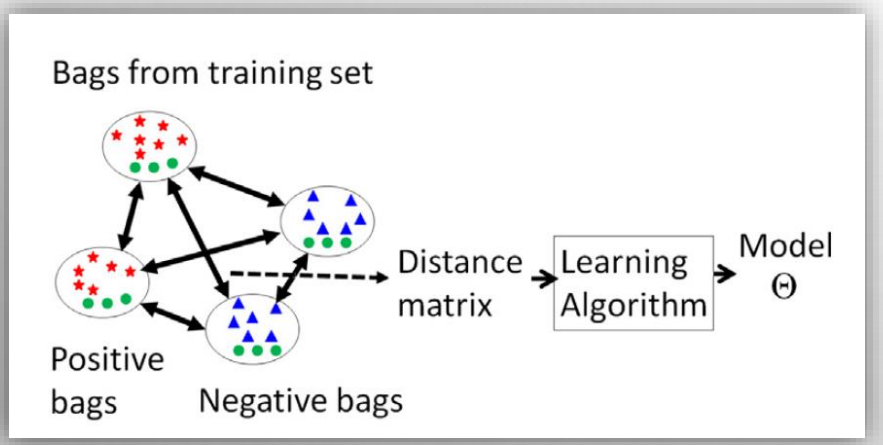
- Authors insist in the importance of these weights, as it makes the different bag of the training set have the same total weight

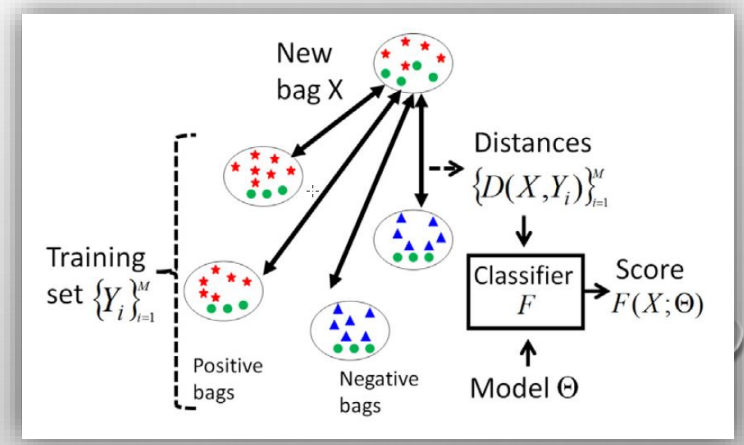
Bag-Space Paradigm

MIC Algorithms

Introduction

- This paradigm treat bags as a whole, and the discriminat learning process is performed in the space of bags
- As bags are not vector entities, we have to define a *distance function* $D(X, Y)$ that compare 2 bags X and Y and plug this distance into a standard distance-based classifier like kNN or SVM.





Introduction

Distances commonly employed

Minimal Hausdorff distance

$$D(X,Y) = \min_{\vec{x}\in X, \vec{y}\in Y} \|\vec{x} - \vec{y}\|$$

Earth Movers distance

$$D(X,Y) = \frac{\sum_i \sum_j w_{ij} \left\|\vec{x}_i - \vec{y}_j\right\|}{\sum_i \sum_j w_{ij}}$$

Chamfer distance

$$D(X,Y) = \frac{1}{|X|}\sum_{\vec{x}\epsilon X} \min_{\vec{y}\epsilon Y} \|\vec{x} - \vec{y}\| + \frac{1}{|Y|}\sum_{\vec{y}\epsilon Y} \min_{\vec{x}\epsilon X} \|\vec{x} - \vec{y}\|$$

Other systems use kernel functions that measure similarity instead of distance

K-NN Algorithms

- Wang y Zucker proposed a k-NN algorithm to MIC problems.

  J. Wang & J.-D. Zucker Solving the multiple-instance problem: a lazy learning approach. *In Proc of 17th International Conference on Machine Learning* (2000), pp 1119-1125

- These authors proposed using the Hausdorff distance as the bag-level distance metric
- The application of k-NN using this metric did not yield good results.

K	K nearest neighbors	# positive	#negative	Total
1	{P}	41	9	50
	{N}	6	36	42
2	{P,P}	41	3	44
	{P,N}	5	15	20
	{N,N}	1	27	28
3	{P,P,P}	40	2	42
	{P,P,N}	**5**	**13**	**18**
	{P,N,N}	2	9	11
	{N,N,N}	0	21	21

- Positive bags also contain negative instances, which attract the negative bags towards themselves.
- There are two ways to solve this problem:
  1. Giving more weight to negative objects.
  2. Defining new ways of combining neighbors to achieve the correct result.

Bayesian K-NN

- Conventional k-NN is based on the votation scheme, that can be represented by

$$\underset{c \in \{\text{positive, negative}\}}{\arg\max} \sum_{i=1}^{k} \delta(c, c_i)$$

- Bayesian k-NN proposes using the probabilities an object has of belonging to the c class, given k nearest neighbors

$$\underset{c \in \{\text{positive, negative}\}}{\arg\max} \; p(c \mid \{c_1, c_2, \ldots, c_k\}) =$$

$$\underset{c \in \{\text{positive, negative}\}}{\arg\max} \; \frac{p(\{c_1, c_2, \ldots, c_k\} \mid c)\; p(c)}{p(\{c_1, c_2, \ldots, c_k\})} =$$

$$\underset{c \in \{\text{positive, negative}\}}{\arg\max} \; p(\{c_1, c_2, \ldots, c_k\} \mid c)\, p(c)$$

These probabilities are calculated from the real distribution of data

Citation K-NN

- This algorithm proposes using, besides the nearest neighbors (called in this case references), the objects that this pattern considers to be a nearest neighbor (called citers).
- Citation k-NN uses R references and C citers and, to decide whether an object is positive or negative, it calculates the following values

$$p = R_p + C_p \qquad n = R_n + C_n$$

where R_p, C_p, R_n y C_n are, respectively, the number of references and citers that has a positive and negative label.

If p>n, the object is labeled as positive. Otherwise, the object is labeled as negative.

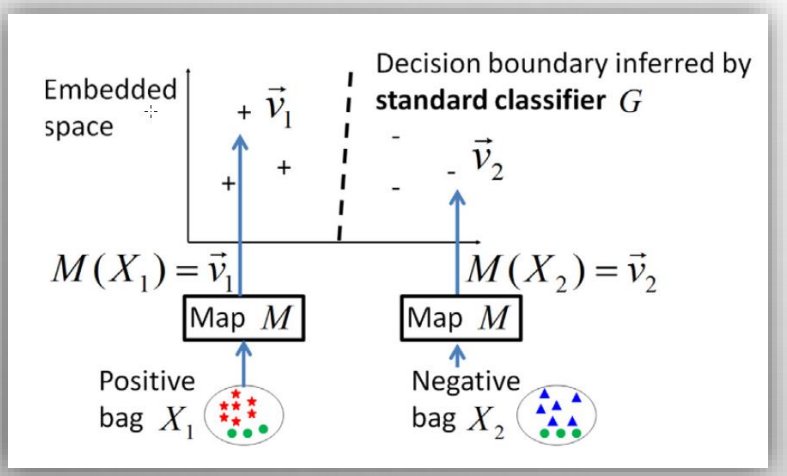
K-NN Algorithms Performance

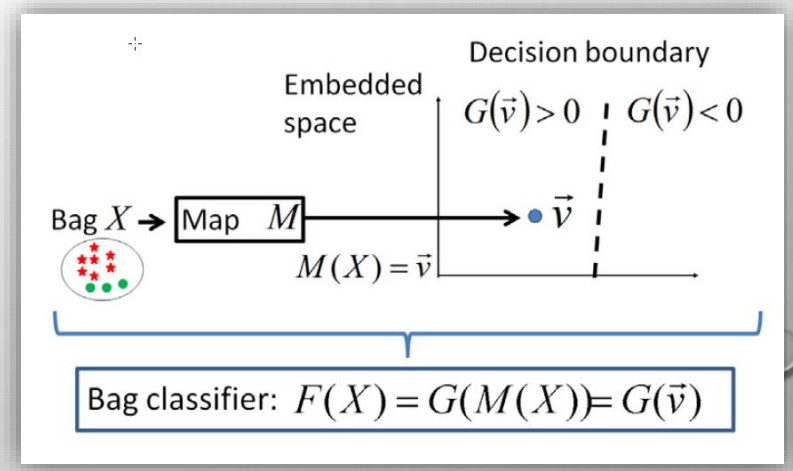
Algorithms	Musk1 %correct	Musk2 %correct
iterated-discrim APR	92.4	89.2
Citation-kNN	92.4	86.3
Bayesian-kNN	90.2	82.4
Diverse Density	88.9	82.5
RELIC	83.7	87.3
MULTINST	76.7	84.0
TILDE	N/A	79.4

Embedded-Space Paradigm

Introduction

- Embedded space methods use the information of bags to perform the discriminative process, like bag space methods.
- Instead of using a distance to compare bags, they define a *mapping function* $\mu: X \rightarrow \vec{v}$ from the bag X to a feature vector $\vec{v}$, wich summarizes the characteristics of the whole bag.





Introduction

We can split embedded methods in two categories:

- Methods that simply *aggregate statistics* of all the instances inside the bag

- Methods that analyze how the instances of the bag match certain prototypes that have been previously discovered in the data (*vocabulary-based methods*).

Methods Without Vocabularies

- This methods aggregate the statistics about the attributes of all the instances without making differenciation among these instances.
- Examples:
- **Simple MI** (Dong et al, 2011). Maps each bag to the average of the instances inside:

$$\mu(X) = \frac{1}{|X|}\sum_{\vec{x}\in X}\vec{x}$$

- Gartner et al (2002) propose to map each bag X to a *min-max vector*,

$$\mu(X) = (a_1, \dots, a_d, \dots, b_1, \dots, b_d)$$

$$a_j = \min_{\vec{x}\in X}\vec{x} \qquad b_j = \max_{\vec{x}\in X}\vec{x} \qquad j=1\dots d \qquad \text{(instance dimensionality)}$$

Vocabulary-Based Methods

- In these methods bags are related with concepts defined previously (**vocabulary**).
- Embedded space contains information about the relationship between bags and concepts in the vocabulary.
- Many times vocabulary concepts are obtained automatically in a unsupervised way (by clustering).
- There is a mapping from bag space to embedded space where vocabulary concepts play a key role.

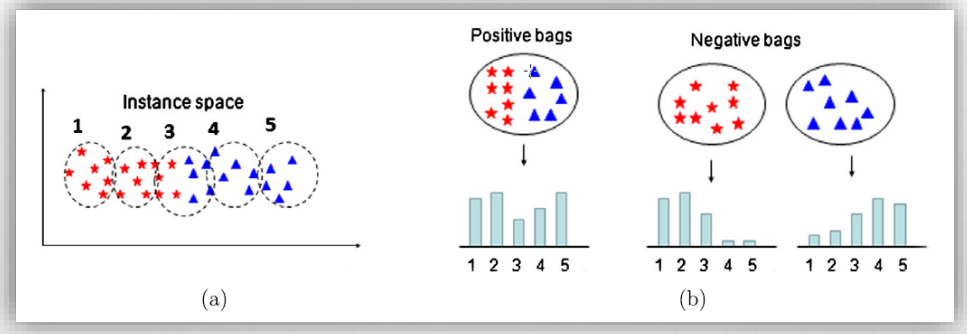
Elements in a Vocabulary-based Method

- **Vocabulary**
  - Stores K concepts.
  - Most of the times the term concept means *class of instances*.
- **Mapping function**
  - Given a bag X and a vocabulary V, this mapping function $\mathcal{M}(X, V) = \vec{v}$ obtains a *D*-dimensional vector that match between the instances $\vec{x}_i \in X$ and the concepts $C_j \in V$
- **Standard supervised classifier.** Classifies the feature vector in the embedded space, using a training set $\mathcal{T}_{\mathcal{M}} = \{(\vec{v}_1, y_1), \ldots, (\vec{v}_N, y_N)\}$

Vocabulary-based methods differs in vocabulary and mapping function

Histogram-based Methods

- These methods use a function $\mathcal{M}$ that maps each bag X into a histogram $\vec{v} = (v_1, \ldots, v_K)$ where the j-th bin v_j counts how many instances of X fall into the j-th vocabulary class C_j.



- Classes (vocabulary) are automatically generated by a clustering algorithm
- Mapping function:

$$\mathcal{M}(X, V) = (v_1, \ldots, v_K) \qquad v_j = \frac{1}{Z} \sum_{\vec{x}_i \in X} f_j(\vec{x}_i)$$

Examples of Histogram-based Methods

Bag-of-words method (Sivic, 2003)

- Clustering method: K-Means
- Mapping function: $f_j(\vec{x}_i) = \begin{cases} 1 & if \quad j = \arg\min_{k=1\ldots,K} \|\vec{x}_i - \vec{p}_k\| \\ 0 & otherwise \end{cases}$

YARDS algorithm (Foulds, 2008)

- There are as many clusters as instances
- Mapping function: $f_j(\vec{x}_i) = \exp\left(-\frac{\|\vec{x}_i - \vec{p}_j\|^2}{\sigma^2}\right)$

Distance-Based methods

Instead of counting the number of instances that fall into class C_j, distance-based methods measure the distance $d_j(\vec{x}_i)$ between a given instance $\vec{x}_i \in X$ and the j-th concept. That is:

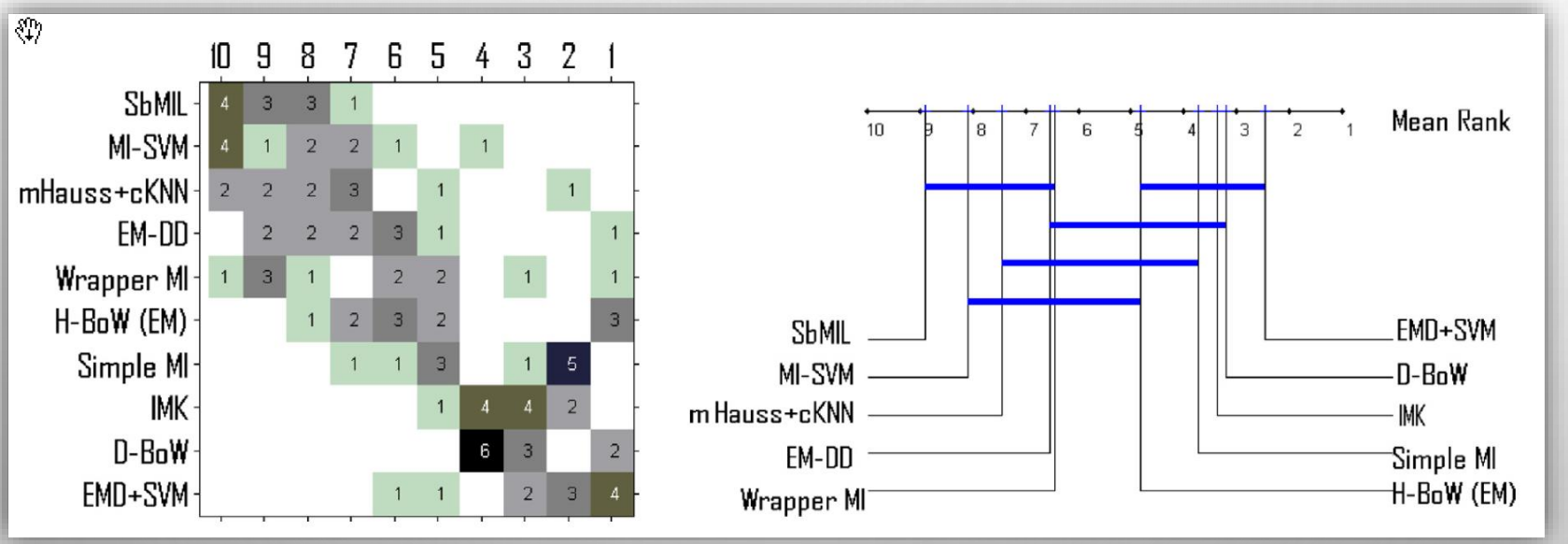
$$\mathcal{M}(X, V) = (v_1, \ldots, v_K) \qquad v_j = \min_{\vec{x}_i \in X} d_j(\vec{x}_i) \qquad j = 1, \ldots, K$$

- Several authors developed similar proposals:
  - Auer et al, 2004
  - DD-SVM (Chen & Wang, 2004)
  - MILES (Chen et al, 2006)
- Concepts definition:
  - Hard assignment Clustering, like K-Means
  - Concepts defined explicitely by authors
- Distance functions:
  - Euclidean distance
  - Mahalanobis distance

Distance Based Bag-of-Words

Comparison of MIL paradigms

- In a recent study, J. Amoros compares 10 different algorithms belonging to the categories previously presented.



- Although the experimental setting does not show a clear winner, there are several interesting conclusions that must be taken into account .

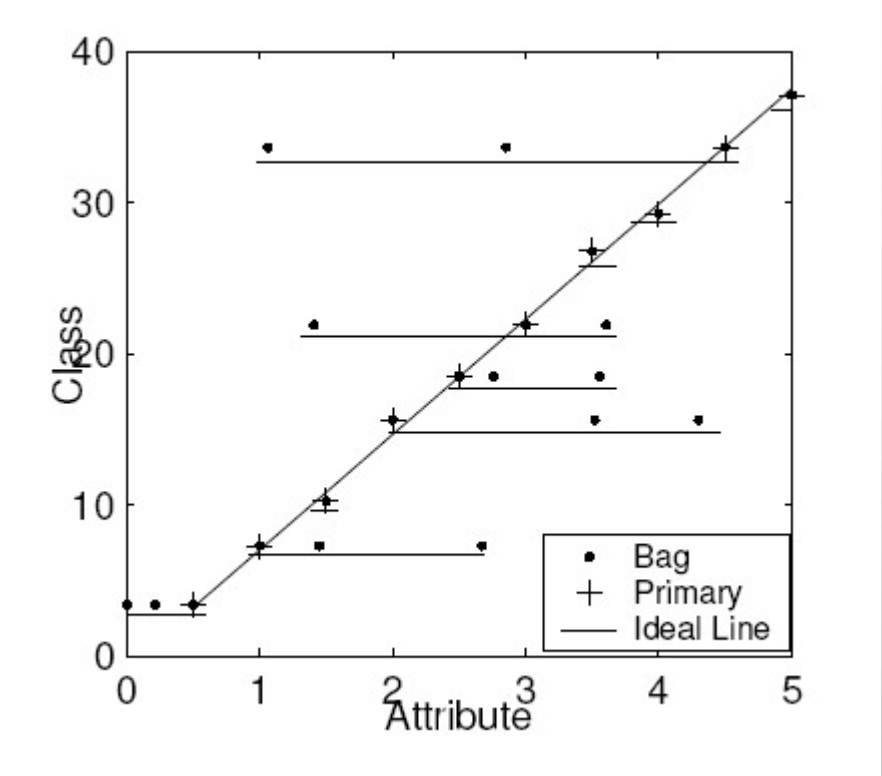
Lesson Learned

- Using global information becomes fundamental in order to obtain a good accuracy
- SMI is not a good framework.
- IS behave similar, regarless of wether they follow the SMI or the collective hypothesis. A similar effect happens with BS methods.
- BS & ES are appropriate methods for exploiting information from the whole bag.
  - Choosing an appropriate distance function is crucial in BS
  - Choosing an appropriate mapping function is crucial in ES
  - Distance function has to evaluate if every instance in one bag has a similar instance in the other (the only metric that does not follo this is Hausdorff)
- BS is infeasible in problems where the number of instances per bag is large. In these cases is recommended the use of ES

Other MIL paradigms

Multiple Instance Regression

- There are few proposals for learning real-valued labels from multi-instance data.
- They assume that one of the instances is responsible of the concept under study (similar to the Dietterich hyphotesis).
- The algorithm search the points that best fit to the hyperplane that represents the concept.



S. Ray and D. Page. Multiple instance regression. In Proceedings of the 18th International Conference on Machine Learning, pages 425–432, 2001

Multiple Instance Clustering

- M.-L. Zhang and Z.-H. Zhou published in 2009 the first paper about clustering of multi-instance data

  M.-L. Zhang & Z.-H. Zhou. Multi-instance clustering with applications to multi-instance prediction. *Appl. Intell.* 31, 47-68, 2009.

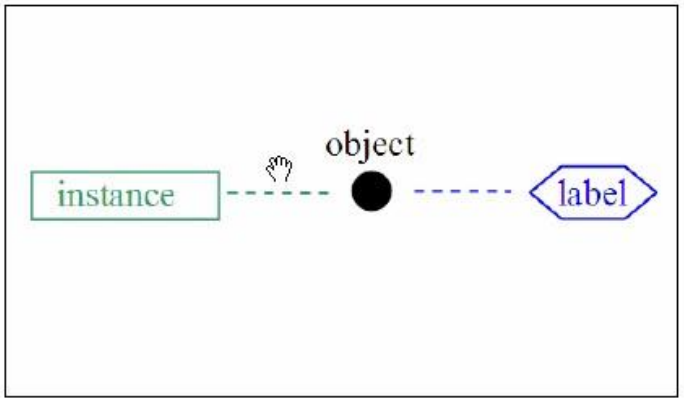
- The algorithm proposed, called BAMIC, is based on the k-medoids algorithm, modifying the metric used to set the distance between examples (bags). They tried three different distances:
  - Minimal Hausdorff distance.
  - Maximal Hausdorff distance
  - **Average Hausdorff distance**

Multi-instance Multi-label Classification

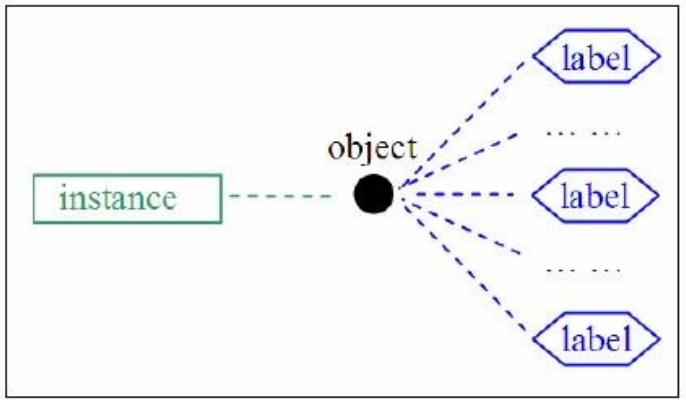
- In this new learning paradigm, learner have to learn **a set of labels** for a given object (represented as multi-instances).
- The objective here is dealing with ambiguity both in the input and output spaces.
  - Input space ambiguity. There are several instances per input object
  - Output space ambiguity. There are several labels for the same object
- This paradigm is a generalization of two previous learning paradigms:
  - Multiple instance learning
  - Multi-label learning (or multi-label classification)
- First reference about this topic

Z.-H. Zhou. Mining ambiguous data with multi-instance multi-label representation. *Lecture Notes in Computer Science* 4632, 2007

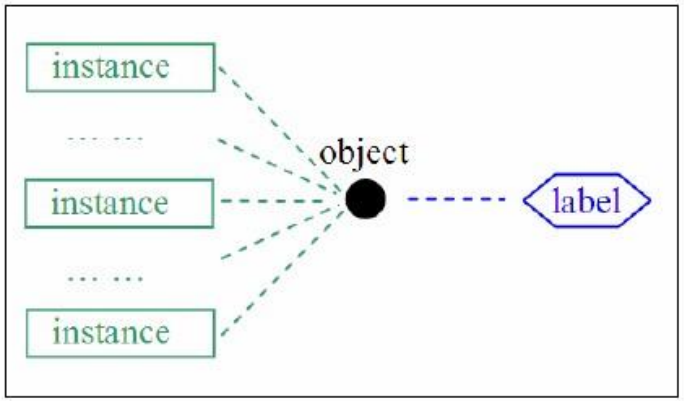
Comparing MIML with other Classification Paradigms



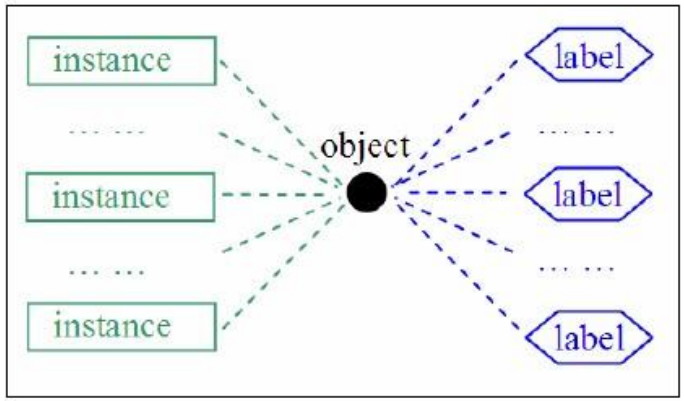
Traditional supervised learning



Multi-label learning



Multi-instance learning



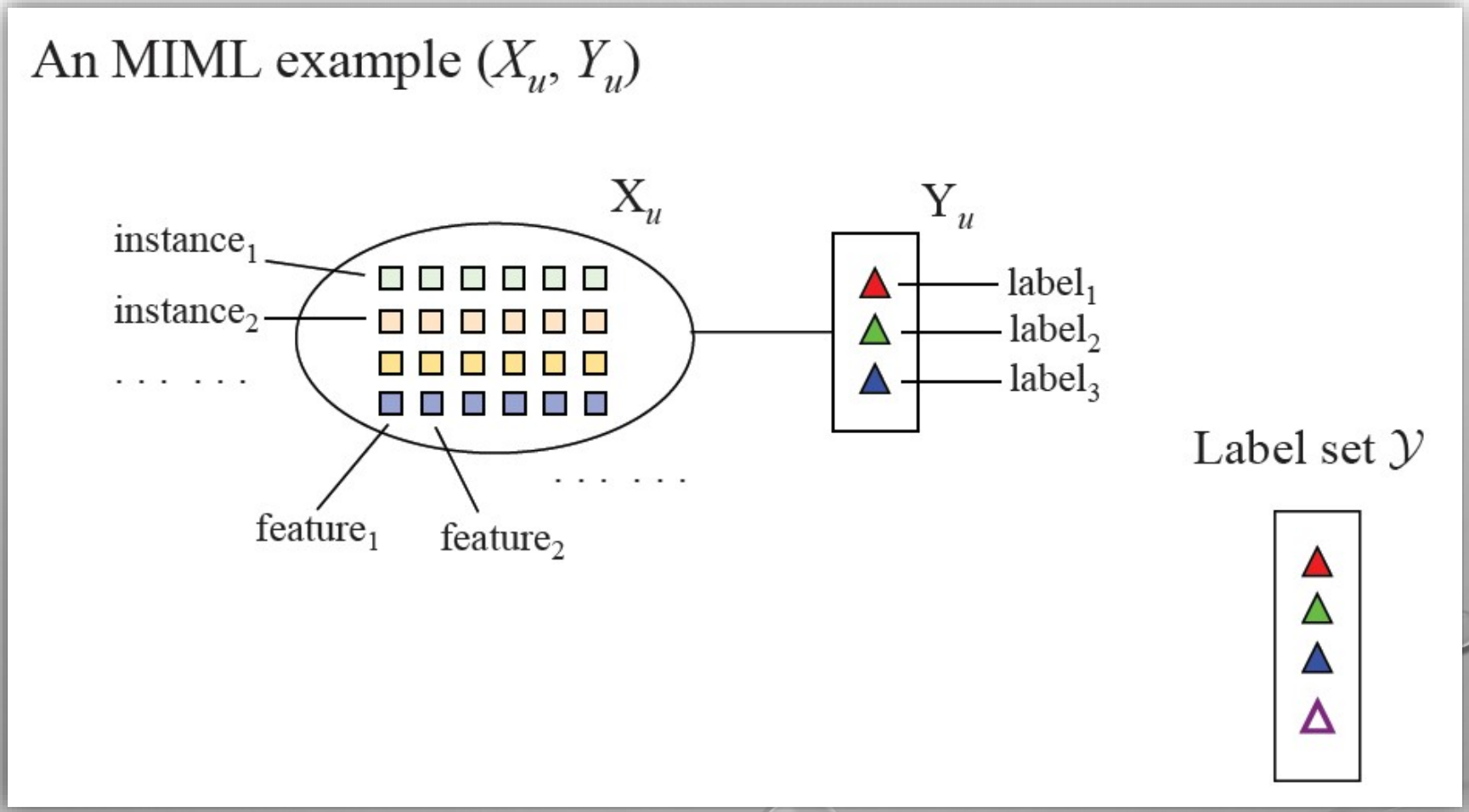
Multi-instance multi-label learning

Methodologies

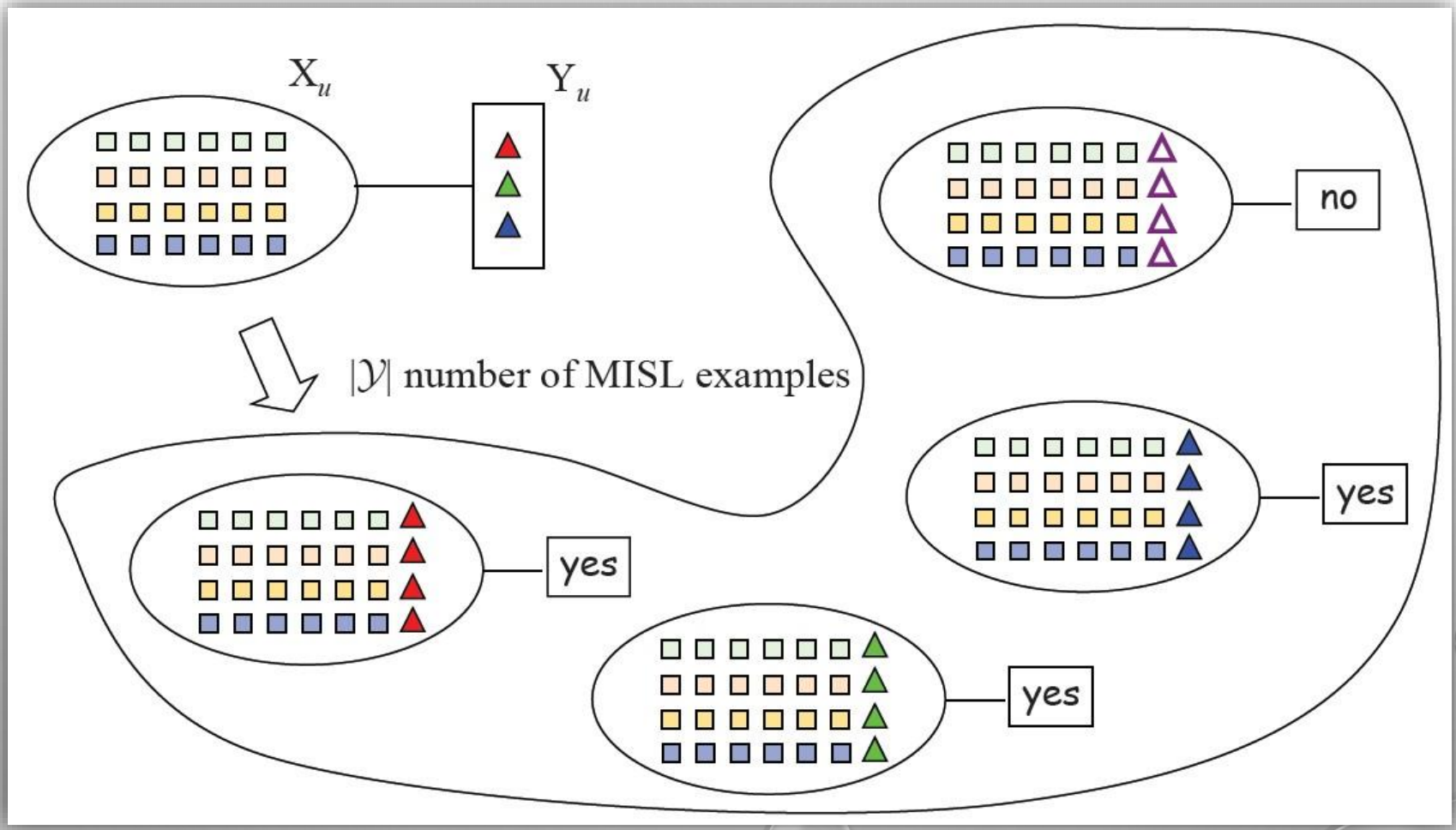
This problem can be solved applying two different methodologies:

- Identifying its equivalence in traditional learning framework, using multi-instance learning as a bridge
  - Step 1: MIMLL → MIL
  - Step 2: MIL → SIL
- Identifying its equivalence in the traditional supervised learning framework, using multi-label learning as a bridge
  - Step 1: MIMLL → MLL
  - Step 2: MLL → SIL

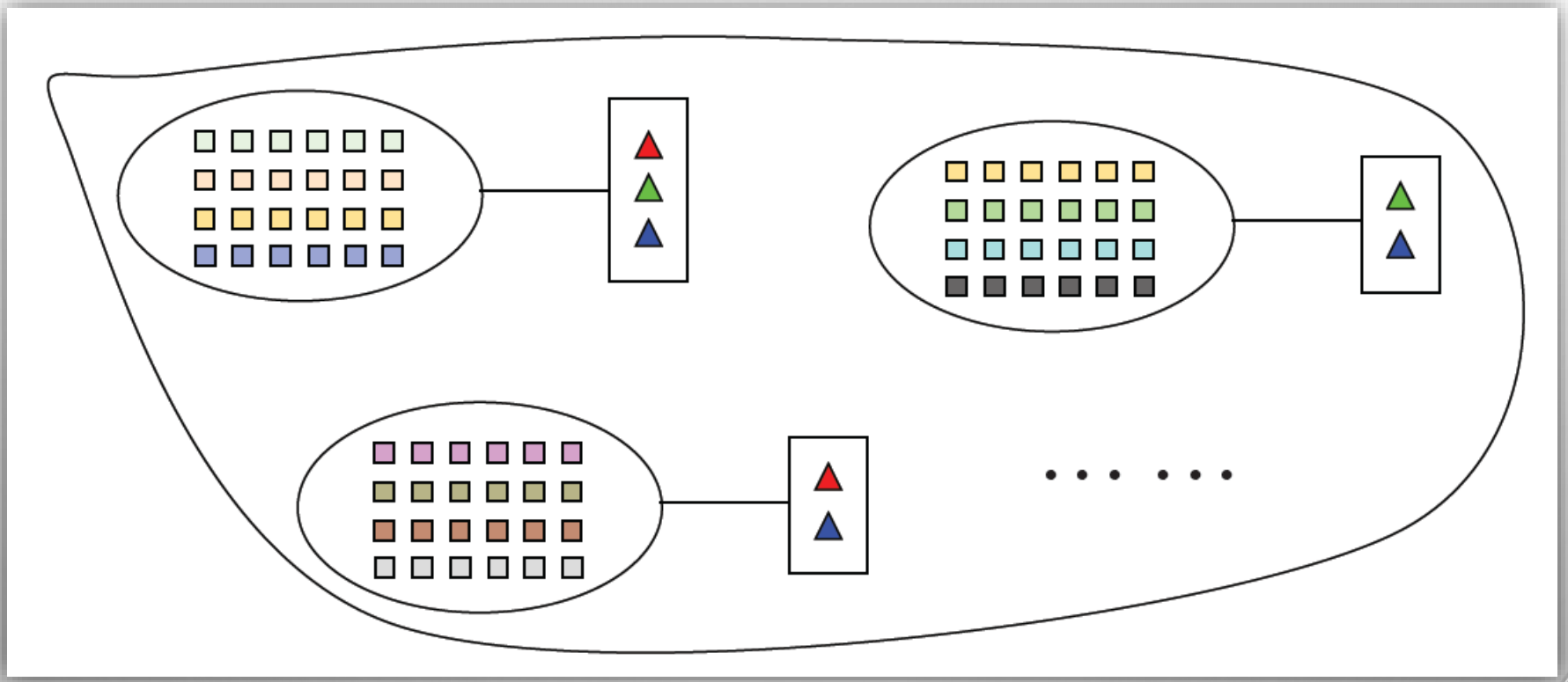
First methodology. Category-wise decomposition



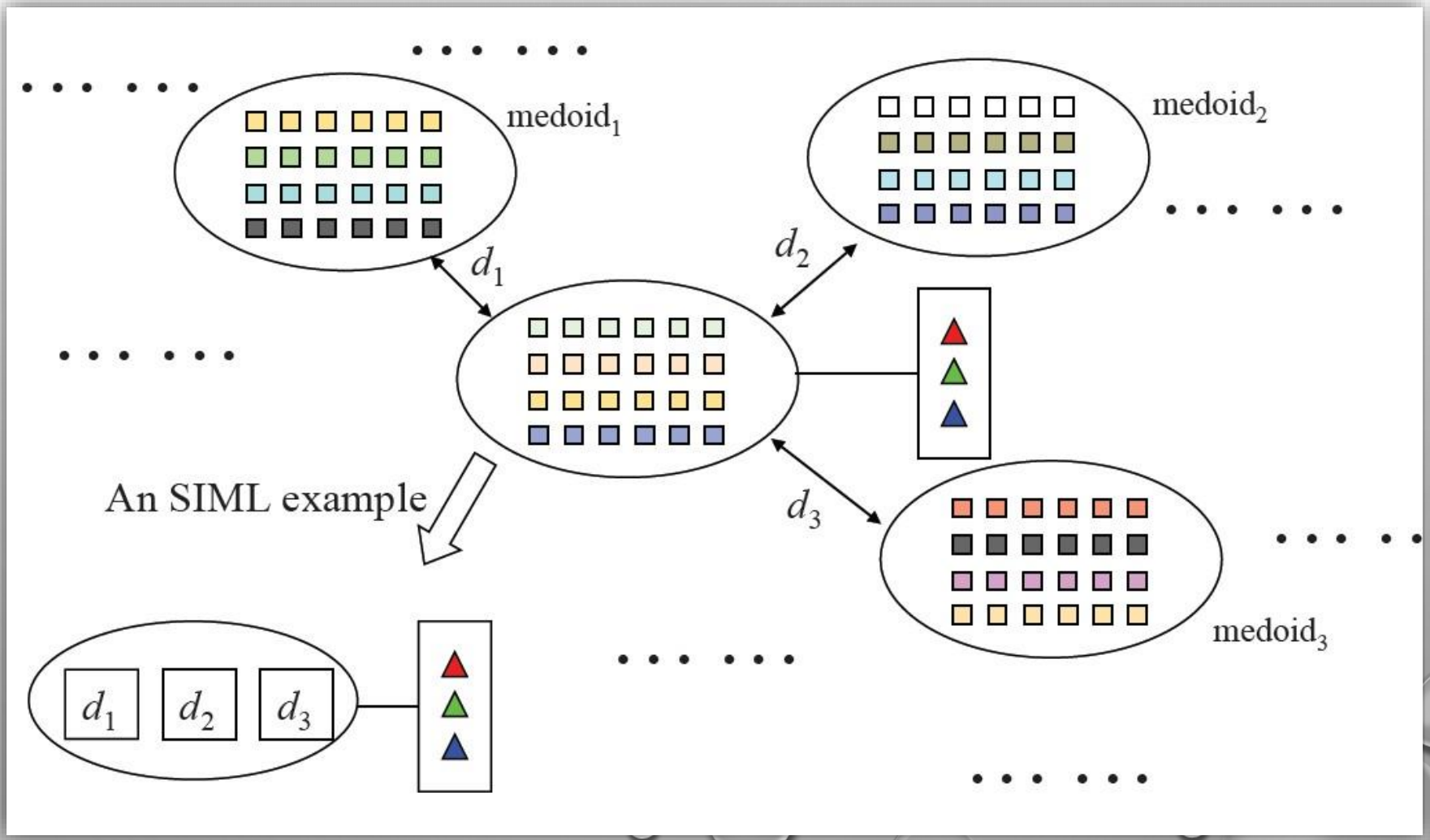
First methodology. Category-wise decomposition (II)



Second methodology. Representation transformation



Second methodology. Representation transformation (II)



Open Problems in MIL

Open Problems in MIL

- What is the best paradigm for MIC?
  - Paradigms based on local information (instances)
  - Paradigms based on global information (bags)
- Feature selection in MIL
  - There are several proposal, but they are based on the standard hypothesis
  - New proposals? Based on different hypothesis?
- Instance selection in MIL
  - Selection of instances (standard hypothesis)
  - Selection of bags

Open Problems in MIL (II)

- MIMLC
  - New approaches and new problems…
- MI Clustering and Regression approaches are based on the standard hypothesis.
  - New approaches based on collective hypothesis
  - New approaches based on global information
- New real-world problems that fit to the MI framework

Internet Resources

Bibliography

1. Z.-H. Zhou. **Multiple instance learning: A survey**. Technical report, Nanjing university, 2004.
2. B. Babenko. **Multiple instance learning: algorithms and applications**. 2009.
3. S. Ray and M. Craven. **Supervised versus multiple instance learning: an empirical comparison.** In *ICML* 2005.
4. J. Foulds and E. Frank. **A review of multi-instance assumptions.** *Knowledge engineering review*, 25(1):1-25, 2010.
5. J. Amores. **Multiple instance classification: review, taxonomy and comparative study.** *Artificial intelligence*, 201:81-105, 2013[.

Software

- **MILK** – A Java MIC kit. Is part of the WEKA suite since version 3.5
- **G3P-MI** and **MO G3P-MI** – both algorithms have been developed using the JCLEC framework for evolutionary computation (http://jclec.sf.net).
- **A KEEL module for mil**. This toolkit includes several algorithms available in MILK, as well as the G3P-MI, MO G3P-MI and others
- **Multiple Instance Learning Toolbox.** Matlab code developed the Pattern Recognition Laboratory at the University of Delf. Available at http://prlab.tudelft.nl/david-tax/mil.html.

Datasets

- http://www.cs.waikato.ac.nz/ml/milk/
  - MILK project address. Contains several MIL datasets, concretely, *MUSK 1, MUSK 2, Mutagenesis 1* and *Mutagenesis 2*. All datasets are in ARFF format.
- http://www.cs.columbia.edu/~andrews/mil/datasets.html
  - Text categorization data (TREC9)
  - Image classification data
- http://www.cs.wustl.edu/~sg/multi-inst-data/
  - Content based image retrieval
  - Drug activity prediction.
- http://www.cs.wisc.edu/~sray/MIPage.html
  - This page contains several datasets and a bibliography about MIL.
- http://www.uco.es/grupos/kdis/mil
  - Several datasets in ARFF format. Data have been partitioned into 10 folds.

Dzięki!

Multiple Instance Learning: An Introduction

Any Questions?

Sebastián Ventura

Knowledge Discovery and Information Systems (KDIS) Research Group
University of Cordoba. Córdoba (SPAIN)