Reducing Dimensionality in Multiple Instance Learning with a Filter Method

Amelia Zafra, Mykola Pechenizkiy, *Sebastián Ventura*

Department of Computer Sciences and Numerical Analysis. University of Cordoba

Overview

- Introduction
- Motivation
- ReliefF Algorithm
- ReliefF-MI Algorithm
- Experimental Results
- Conclusions

Introduction

Introduction

- Why select a subset of “relevant” input variables?
  - Naive theoretical view:
    - More features.
      - More information.
      - More discrimination power.
  - In practice:
    - In domains with many features the underlying probability distribution can be very complex and very hard to estimate (e.g. dependencies between variables)
    - Irrelevant and redundant features **can confuse learners**!
    - Limited training data
    - Limited computational resources

Approaches

- There are a large number of research reports in this area for traditional supervised learning considering
  - Wrapper Methods
    - It is used as a part of evaluation function and also to induce the final learning model
  - Filter Methods
    - It is a pre-processing step, which is independent of the learning algorithm
  - Embedded Methods
    - Feature selection is a part of the training procedure of a classifier
- Studies show that feature selection can **significantly improve** a learning algorithm's performance!

Motivation

- In Multiple Instance Learning is very difficult to find studies about this topic.
  - There are not more of three proposals classified all of them as wrapper methods.
- The aim is to investigate the issues raised by introducing a filter method of feature selection to deal with the multiple-instance problem.

$$\mathbf{R}^{p} \longrightarrow \boxed{\text{Feature selection}} \longrightarrow \underset{s \ll p}{\mathbf{R}^{s}} \longrightarrow \boxed{\text{Classifier design}}$$

Motivation

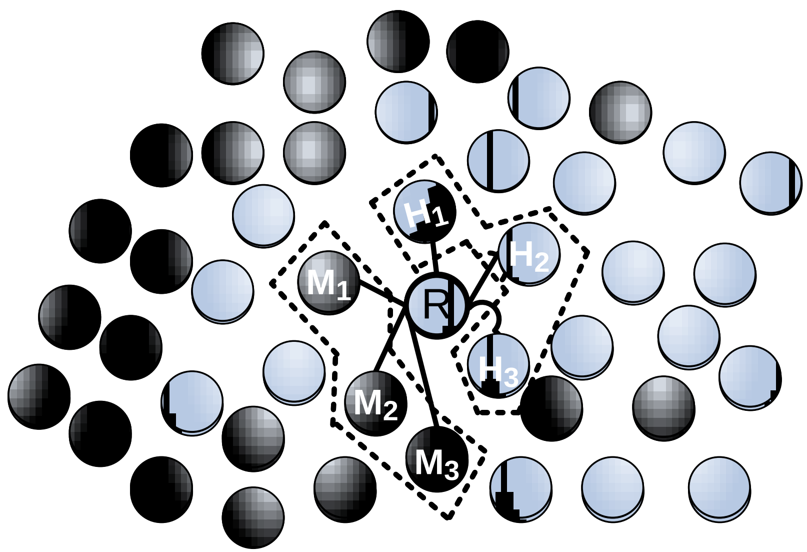
- We propose an effective feature selection approach to MIL called ReliefF-MI.
- This method extends ReliefF algorithm to MIL.
- The main features of ReliefF family are:
    - It may be applied in all situations,
    - It has low bias,
    - It includes interaction among features
    - It may capture local dependencies that other methods miss.

Adapted from J. Fridlyand

ReliefF Algorithm

ReliefF Feature Selection Method

k = 3



Wa = W

$$\mathrm{diff}(A, I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{max(A) - min(A)}$$

m = numb

Positive pattern

Negative pattern

ReliefF-MI Algorithm

Supervised learning vs. multi-instance learning

Supervised Learning

Training Set

Object ≅ Instance

$V(v_1, v_2, \ldots v_n)$ -

I know the class of the object

The instance is negative

Multi-Instance Learning

Training Set

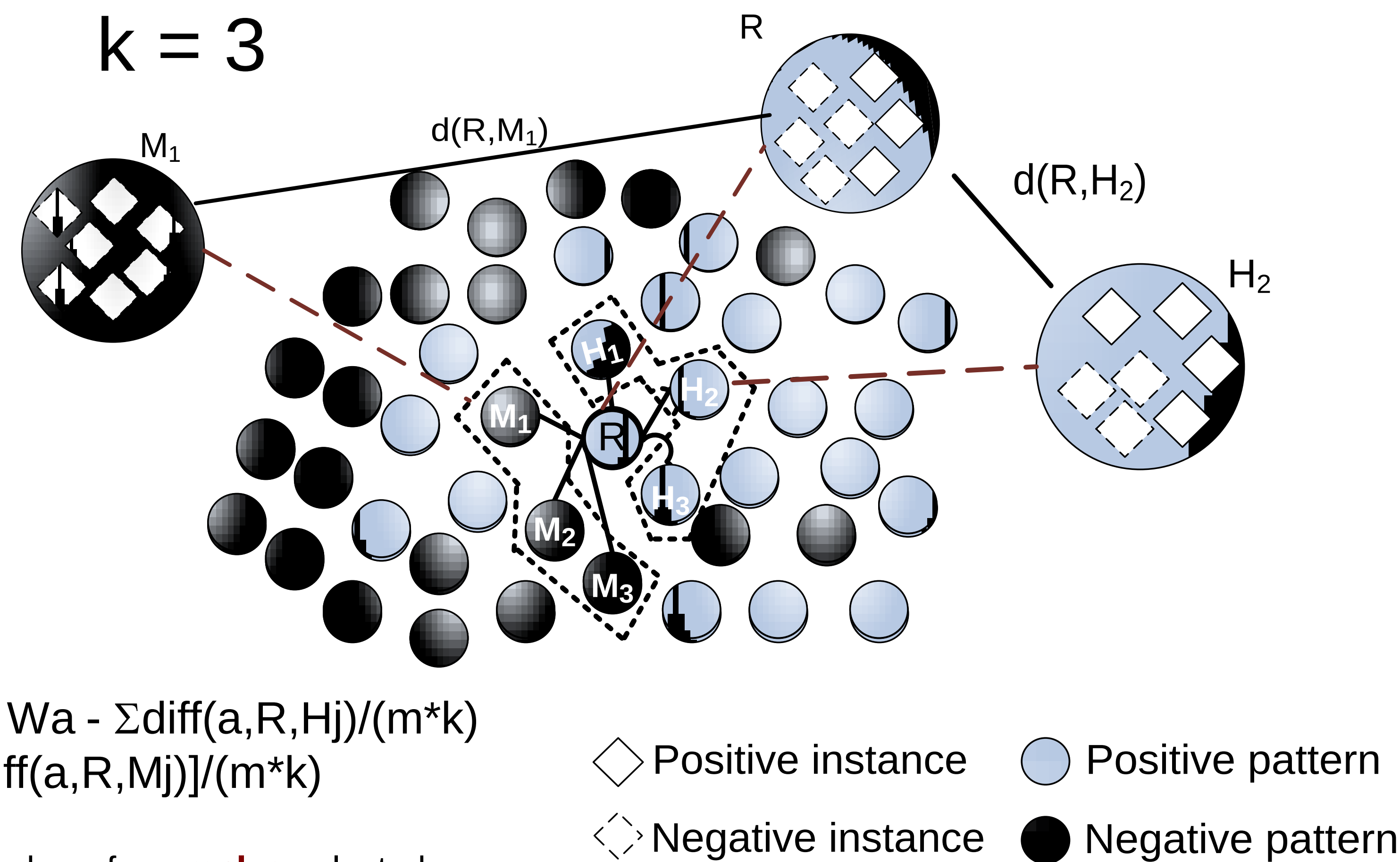
Object ≅ Several Instances

$V_1(v_{11}, v_{12}, \ldots v_{1n})$ -
$V_2(v_{21}, v_{22}, \ldots v_{2n})$ -
…
$V_p(v_{p1}, v_{p2}, \ldots v_{pn})$ -

I know the class of the object, but I do not know the class of each instance

The bag is negative, I know that all instances are negatives

ReliefF Feature Selection Method



Wa = Wa - Σdiff(a,R,Hj)/(m*k)
+ Σ diff(a,R,Mj)]/(m*k)

m = number of **examples** selected

Maximal Hausdorff Distance

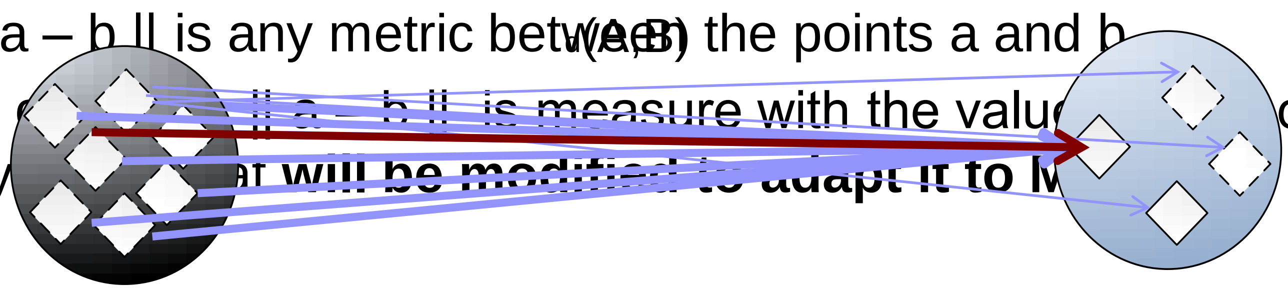
- Given two sets of points A = {a1,...,am} and B = {b1,...,bn} the Maximal Hausdorff Distance is defined as

$$H_{max}(A,B) = \max(h(A,B), h(B,A))$$

where

$$h(A,B) = \max_{a \in A} \min_{b \in B} || a - b ||$$

- || a – b || is any metric between the points a and b
- In o[illegible] || a – b || is measure with the valu[illegible] defined by [illegible] at **will be modified to adapt it to M[illegible]**

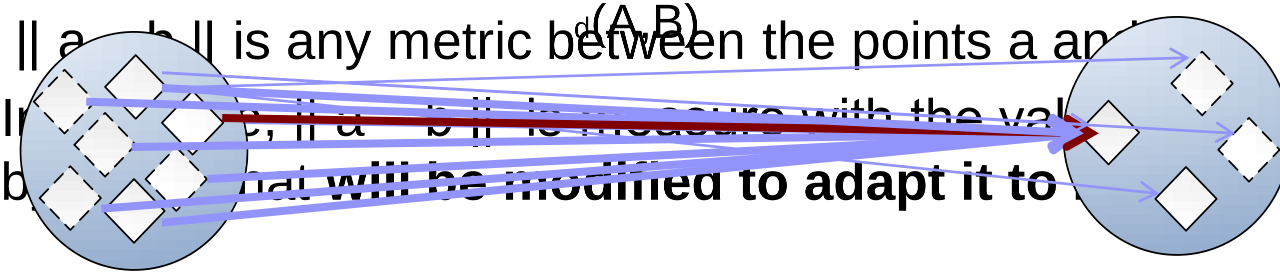


Minimal Hausdorff Distance

- Given two sets of points A = {a1,...,am} and B = {b1,...,bn} the Minimal Hausdorff Distance is defined as

$$H_{min}(A,B) = \min_{a \in A} \min_{b \in B} || a - b ||$$

- || a [illegible] b || is any metric between the points a an[illegible]
- In [illegible]e, || a [illegible] b || is [illegible] with the val[illegible]" defined b[illegible]at **will be modified to adapt it to** [illegible]



Average Hausdorff Distance

- Given two sets of points A = {a1,...,am} and B = {b1,...,bn} the Average Hausdorff Distance is defined as

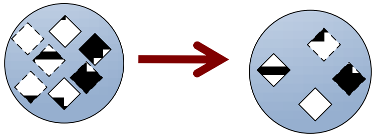
$$\text{Havg}(A,B) = \frac{\sum_{a \in A} \min_{b \in B} \| a - b \| + \sum_{b \in B} \min_{a \in A} \| b - a \|}{| A | + | B |}$$

- || a – b || is any metric between the points a and b.
- In our case, || a – b || is measure with the value of “diff” defined by Relief that **will be modified to adapt it to MIL**

Adapted Hausdorff Distance

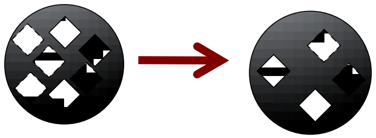
- Given two sets of points A = {a1,...,am} and B = {b1,...,bn} the Adapted Hausdorff Distance is defined as
  - If A & B are positive:

    $H_{adapted}(A,B) = H_{min}(A,B)$

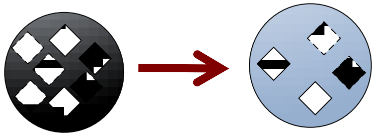
  - If A & B are negative:

    $H_{adapted}(A,B) = H_{avg}(A,B)$

  - If A is positive and B negative or viceverse:

    $H_{adapted}(A,B) = H_{max}(A,B)$

Experimental Study

Introduction

- To show the effectiveness of the method proposed we consider
  - The most relevant algorithms in MIL (seventeen algorithms)
  - Application to different real problems

- Experiment results try to show
  - A comparison between the different metrics used
  - A comparison between algorithm performance that does or does not use ReliefF-MI as a pre-processing step.

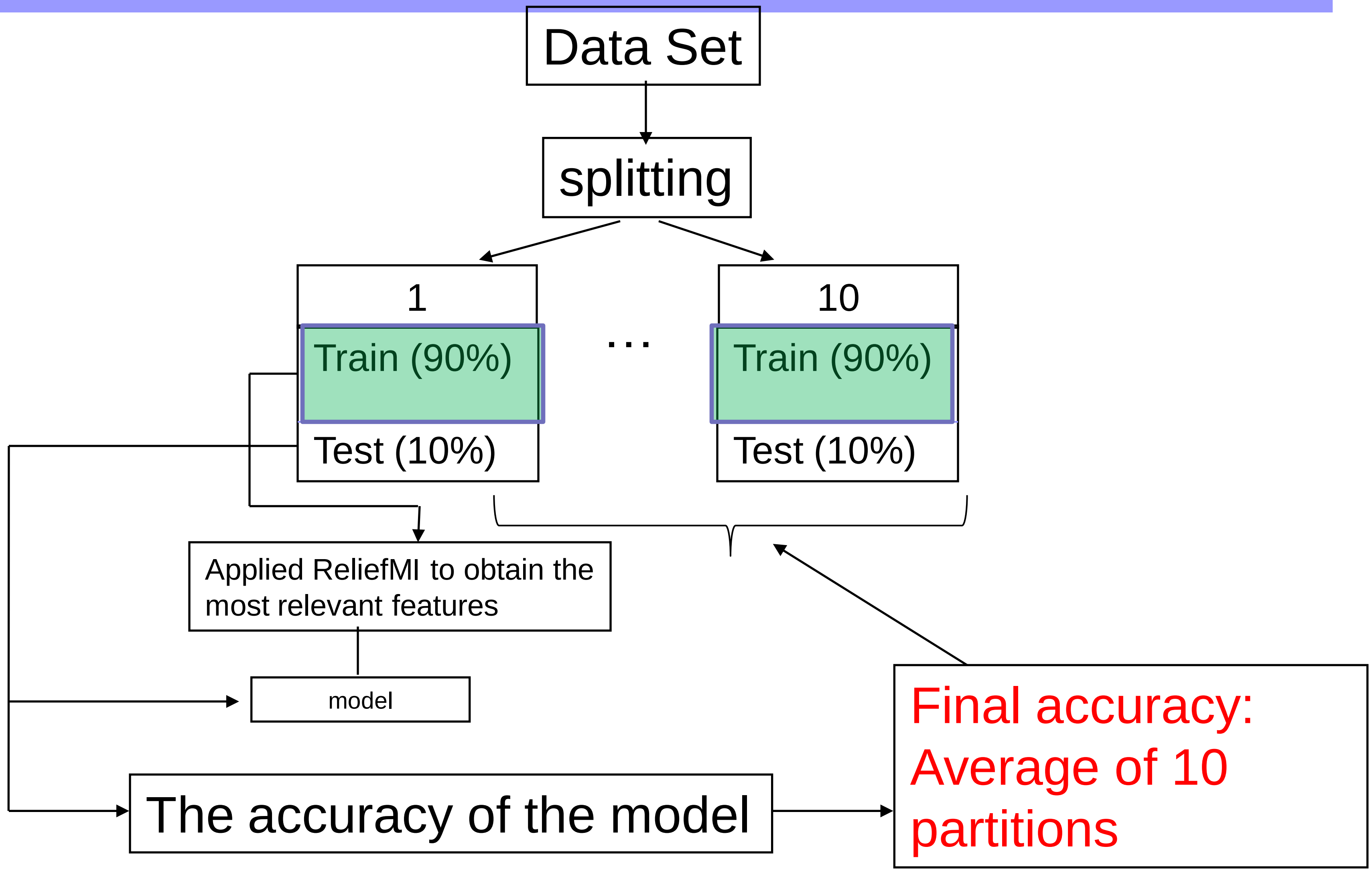
Test Datasets

- Three data set are considered

Dataset	BAG Positive	Negative	Total	Attribute	Instances	Average bag size
Elephant	100	100	200	230	1391	6.96
Tiger	100	100	200	230	1220	6.10
Fox	100	100	200	230	1320	6.60

- 17 algorithms are considered including:
  - Rule based system, regresion logistic system, decision tree, support Vector Machine, naive bayes, …
- Four version of ReliefF-MI are developed considering the four different metrics of Hausdorff distance.

10 fold cross validation



Comparative between distance metrics

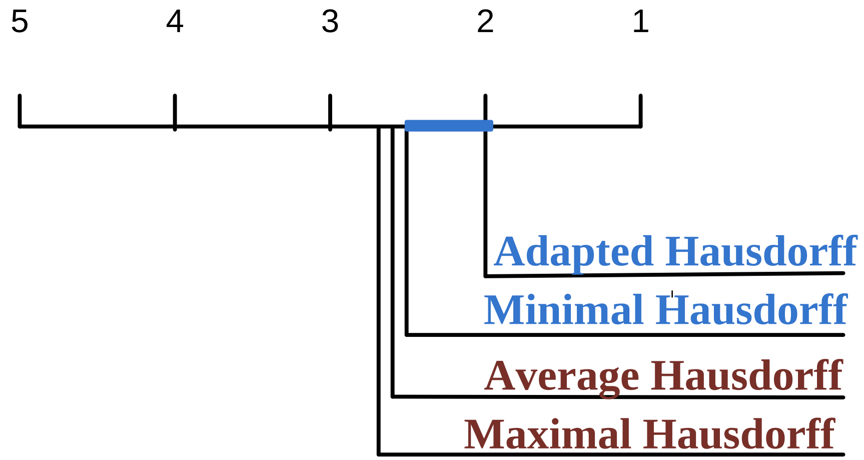
Algorithms	Maximal			Minimal			Average			Adapted		
	Eleph	**Tiger**	**Fox**	**Eleph**	**Tiger**	**Fox**	**Eleph**	**Tiger**	**Fox**	**Eleph**	**Tiger**	**Fox**
citationKNN	0,750	0,830	0,615	0,745	0,850	0,630	0,745	0,840	0,610	0,745	0,815	0,615
MDD	0,725	0,810	0,620	0,710	0,800	0,600	0,710	0,790	0,605	0,705	0,805	0,660
RepTree[1]	0,825	0,870	0,655	0,840	0,845	0,665	0,840	0,865	0,700	0,840	0,855	0,710
DecisionStump[1]	0,825	0,800	0,655	0,820	0,785	0,695	0,820	0,800	0,660	0,830	0,805	0,700
MIDD	0,755	0,780	0,600	0,750	0,780	0,645	0,750	0,780	0,595	0,755	0,770	0,695
MIEMDD	0,725	0,775	0,530	0,685	0,720	0,605	0,685	0,745	0,530	0,715	0,770	0,615
MILR	0,815	0,855	0,600	0,840	0,825	0,630	0,840	0,840	0,615	0,835	0,875	0,635
MIOptimalBall	0,795	0,740	0,575	0,765	0,715	0,495	0,765	0,735	0,525	0,775	0,740	0,535
RBF Kernel[2]	0,765	0,835	0,615	0,800	0,865	0,655	0,800	0,830	0,655	0,785	0,855	0,650
Polynomial Kernel[2]	0,765	0,825	0,620	0,780	0,825	0,685	0,780	0,830	0,665	0,770	0,820	0,655
AdaBoost&PART[3]	0,830	0,840	0,615	0,830	0,825	0,745	0,830	0,820	0,620	0,840	0,860	0,665
Bagging&PART[3]	0,830	0,850	0,585	0,810	0,865	0,595	0,810	0,860	0,610	0,830	0,865	0,605
PART[3]	0,830	0,815	0,580	0,815	0,830	0,615	0,815	0,810	0,570	0,835	0,840	0,620
SMO3	0,705	0,815	0,660	0,715	0,835	0,675	0,715	0,830	0,655	0,705	0,820	0,690
Naive Bayes[3]	0,655	0,820	0,590	0,675	0,815	0,650	0,675	0,825	0,585	0,660	0,820	0,680
AdaBoost&PART[4]	0,800	0,855	0,570	0,840	0,830	0,600	0,840	0,840	0,560	0,830	0,845	0,650
PART[4]	0,770	0,795	0,620	0,765	0,730	0,670	0,765	0,740	0,660	0,775	0,780	0,665
RANKING		2.676			2.520			2.794			2.010	

1 MIBoost
2 MISMO
3 MIWrapper
4 MISimple

Comparative between distance metrics

- Results obtained

	FRIEDMAN TEST (p = 0.05)		
	Friedman Value	**χ 2 (3)**	**Conclusion**
Acc	10.965	6.251	Reject null hypothesis



Effectiveness of ReliefF-MI

Algorithms	Reduced Set			Full Set		
	Eleph	Tiger	Fox	Eleph	Tiger	Fox
citationKNN	0,745	0,815	0,615	0,500	0,500	0,500
MDD	0,705	0,805	0,660	0,800	0,755	0,700
MIBoost (RepTree)	0,840	0,855	0,710	0,815	0,825	0,670
MIBoost (DecisionStump)	0,830	0,805	0,700	0,815	0,780	0,650
MIDD	0,755	0,770	0,695	0,825	0,740	0,655
MIEMDD	0,715	0,770	0,615	0,730	0,745	0,600
MILR	0,835	0,875	0,635	0,780	0,840	0,510
MIOptimalBall	0,775	0,740	0,535	0,730	0,625	0,530
MISMO (RBF Kernel)	0,785	0,855	0,650	0,800	0,795	0,590
MISMO (Polynomial Kernel)	0,770	0,820	0,655	0,790	0,785	0,580
MIWrapper (AdaBoost&PART)	0,840	0,860	0,665	0,840	0,790	0,685
MIWrapper (Bagging&PART)	0,830	0,865	0,605	0,845	0,810	0,600
MIWrapper (PART)	0,835	0,840	0,620	0,790	0,780	0,550
MIWrapper (SMO)	0,705	0,820	0,690	0,715	0,800	0,635
MIWrapper (Naive Bayes)	0,660	0,820	0,680	0,680	0,760	0,590
MISimple (AdaBoost&PART)	0,830	0,845	0,650	0,840	0,795	0,625
MISimple (PART)	0,775	0,780	0,665	0,765	0,765	0,635

Effectiveness of ReliefF-MI

- Results obtained

Sum of Ranks and Mean Rank of the two proposals		
	Mean Rank	**Sum of Ranks**
ReliefF-MI Method	57.03	2908.50
Not Reducting features	45.97	2344.50

Wilcoxon Test (p = 0.1)			
Value	**z-score**	**p-value**	**Conclusion**
2344	-1.888	0.059	Reject null hypothesis

The use of ReliefF-MI has significantly higher accuracy values than the option that does not use feature selection.

Conclusion

Conclusions

- The problem of feature selection to reduce the dimensionality of data in MIL is dealt using filter method.
- A new efficient algorithm based on ReliefF principles is proposed
- Experimental results shows the effectiveness of our approach
  - Using three different applications and seventeen algorithms with the reduced data.
    - Results show that the new metric proposed is the metric that statistically achieves the best results.
  - Showing the benefits of applying data reduction in MIL
    - Results show the relevance of using feature selection in this scenario is established for improving the performance of algorithms with high-dimensional data.

Reducing Dimensionality in Multiple Instance Learning with a Filter Method

Department of Computer Sciences and Numerical Analysis. University of Cordoba