Authentication

Prof. Dr. Helene Dörksen

helene.doerksen@th-owl.de

About Examination

Important rules you have to know and to follow:

- Other programming languages or classification libraries are allowed
- Additional advanced methods (not discussed in lectures) are welcome, but need to be described
- At least one of advanced methods (ComRef, etc.) from lectures have to be tested and analysed in the term paper
- No **Blackbox** solutions
- You might provide your Matlab files and I will check them

Learned before

We are able to improve accuracy without increasing complexity.

General procedure:

- Start with initial linear classifier (initial accuracy)

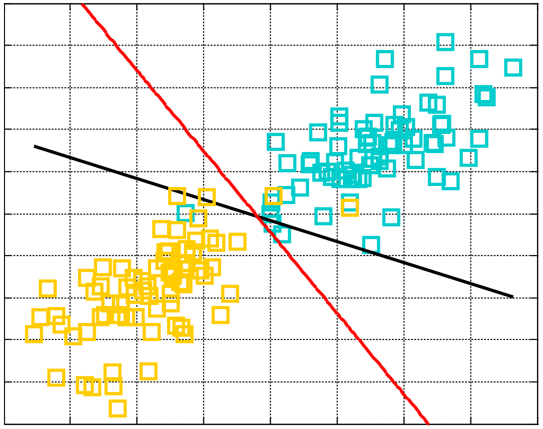
- Perform dimentionality reduction (based on initial parameters)
- Design new classifier with higher accuracy (refinement)

$\}$ lot of possibilities

Examples for classifier design:

- Combinatorial Refinement of Feature Weighting for Linear Classification
- Margin-based Refinement (SVM or other linear classifiers)

classification rate of $\boldsymbol{h}$ is $\mathbf{95.71\%}$



classification rate of $\boldsymbol{g}$ is $\mathbf{97.85\%}$

Lecture 11:

Fast Classification in (Industrial) Big Data Environments

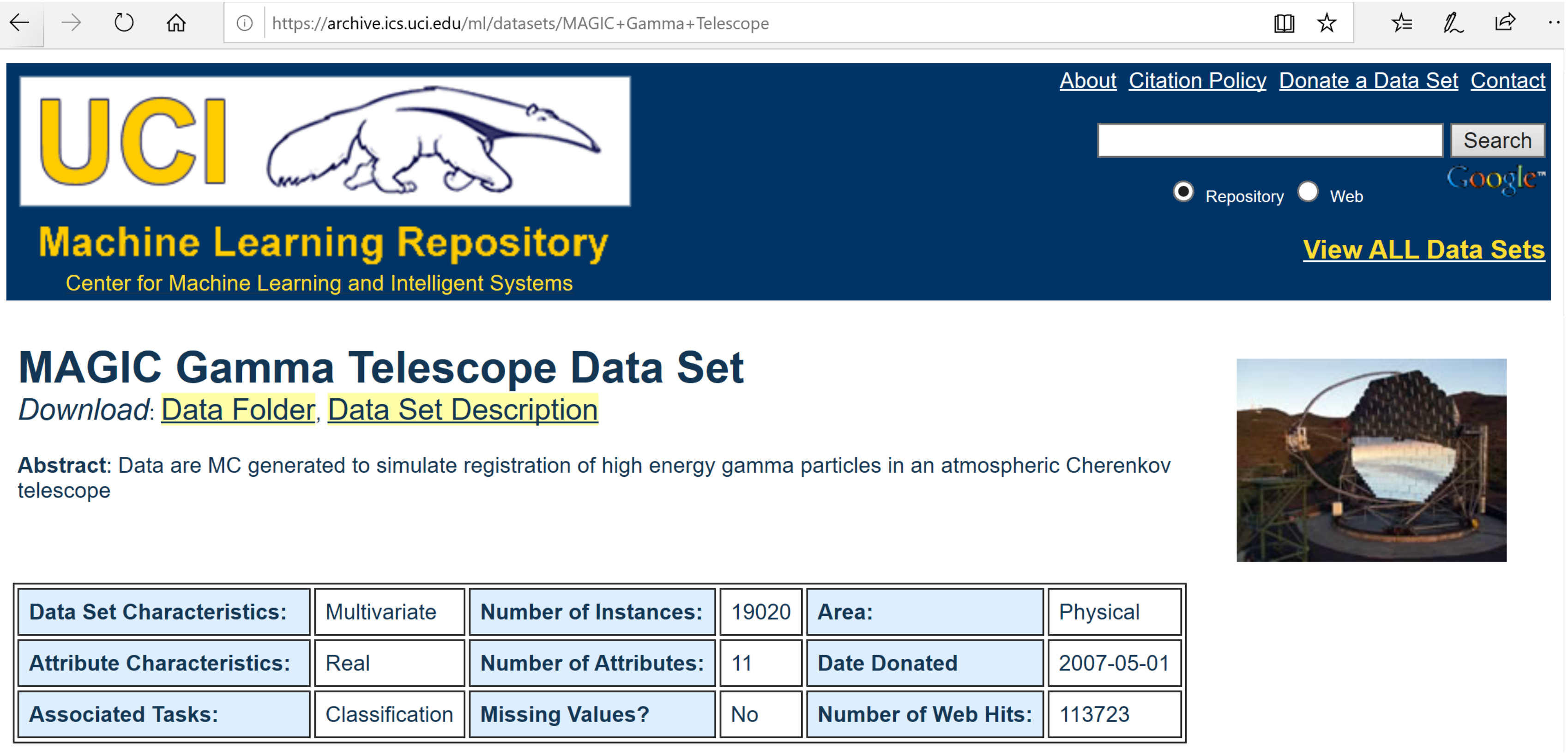
19th IEEE International Conference on
Emerging Technologies and Factory Automation, Barcelona, Spain, 16.-19. September 2014

INTRODUCTION

Number of objects plays an important role for:

- Design of accurate classifier
- Execution time

Example


https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope
About Citation Policy Donate a Data Set Contact
Search
Google™
● Repository ○ Web
View ALL Data Sets
UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

MAGIC Gamma Telescope Data Set

Download: Data Folder, Data Set Description

Abstract: Data are MC generated to simulate registration of high energy gamma particles in an atmospheric Cherenkov telescope

Data Set Characteristics:	Multivariate	**Number of Instances:**	19020	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	11	**Date Donated**	2007-05-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	113723

INTRODUCTION

Fast Classification in (Industrial) Big Data Environments

or

“How to eat an Elephant”

Source: http://kieran-jones.com/how-do-we-eat-an-elephant/

INTRODUCTION

We concentrate on processes in which numerous actuators and sensors (many hundreds up to thousands) are applied.

© Muratec-Europe, Willich, Germany

© KBA, Würzburg, Germany

© Saurer-Schlafhorst, Uebach, Germany

REFINEMENT APPROACH[1,2]

$$\boldsymbol{h(x) = a_1x_1 + a_2x_2 + \cdots + a_i\,x_i + a_{i+1}x_{i+1} + \cdots + a_dx_d + c_h}$$

summands: $\boldsymbol{a_jx_j}$

fusion of summands: $\displaystyle\sum_{j\in I} \boldsymbol{a_jx_j}$

<u>Idea:</u>

- a helpful summand may be **less relevant** by itself
- a **fusion** of individually irrelevant **summands may become relevant**

[1] Dörksen, H.;Lohweg, V.: “Combinatorial Refinement of Feature Weighting for Linear Classification,” in 19th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA 2014), 2014.

[2] Dörksen, H.; Lohweg, V.: Margin-based Refinement for Support-Vector-Machine Classification. Proceedings of 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM2017), Porto, Portugal, 2017

REFINEMENT APPROACH

Fusions of summands lead to **dimensionality reductions** down to spaces:

$d-1$ ← I_1 consists of **2** summands, I_2 is rest; complexity of calculation $\mathcal{O}(d^2)$[1]

$$\left.\begin{array}{c} d-2 \\ d-3 \\ \vdots \\ 3 \end{array}\right\} \quad \textbf{combinatorial task}$$

2 I_1 is complementary to I_2; complexity of calculation $\mathcal{O}(2^d)$[1]

[1] H. Dörksen and V. Lohweg, “Combinatorial Refinement of Feature Weighting for Linear Classification,” in 19th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA 2014), 2014.

REFINEMENT APPROACH

Proposition

Let d be the dimension of the feature space and m be the number of objects in the dataset. The largest and smallest dimension of a reduced feature space is defined as: d-1, resp. 2. For the **ComRef-approach** the search for the best **fusion of summands** with respect to **SVM classification has time complexity**

- (i) $\mathcal{O}(m^3 d^2)$ for (d-1)-dimensional space,
- (ii) $\mathcal{O}(m^3 2^d)$ for 2-dimensional space.

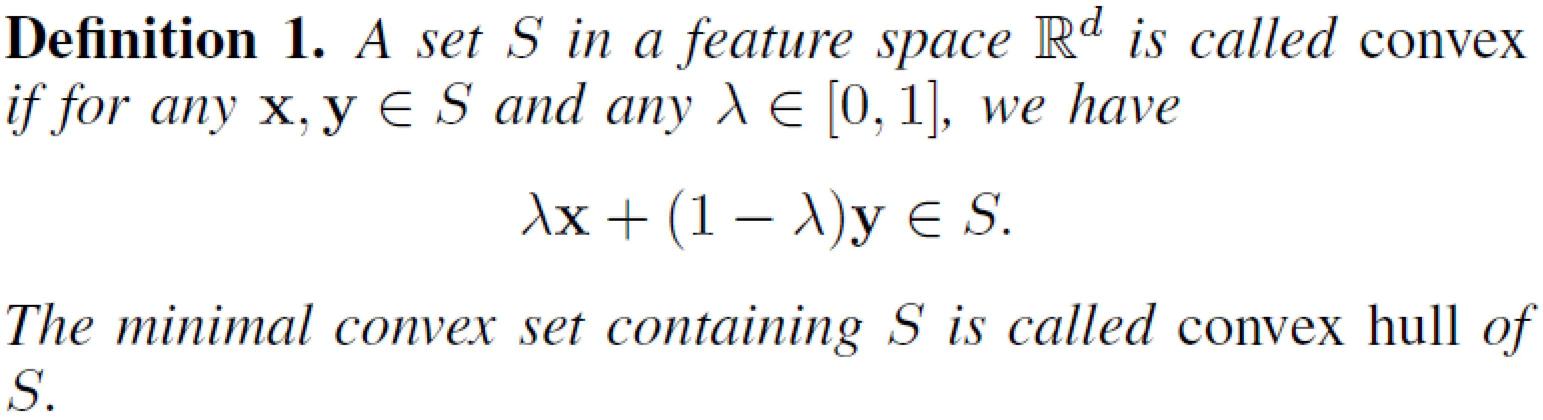
Proof

The statement follows directly from the fact that time complexity of SVM is $\mathcal{O}(m^3)$ and that there are d^2 fusions of summands in (d-1)-dimensional space, resp. 2^d fusions of summands in 2-dimensional space.[1] ■

[1] H. Dörksen and V. Lohweg, “Combinatorial Refinement of Feature Weighting for Linear Classification,” in 19th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA 2014), 2014.

REFINEMENT APPROACH AND CONVEX HULL

Convex Hull



Definition 1. *A set S in a feature space $\mathbb{R}^d$ is called* convex *if for any $\mathbf{x}, \mathbf{y} \in S$ and any $\lambda \in [0, 1]$, we have*

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in S.$$

The minimal convex set containing S is called convex hull *of S.*

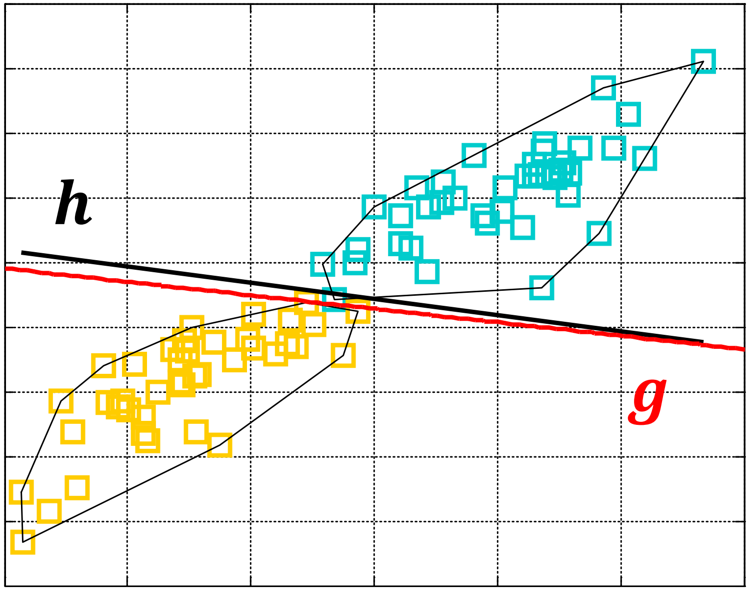
It is a well known fact from computational geometry that the convex hull computation of a finite set of points in $\mathbb{R}^2$ takes $\mathcal{O}(m \log m)$ time[1].

[1] F. P. Preparata and S. J. Hong, “Convex Hulls of Finite Sets of Points in Two and Three Dimensions,” Communications of the ACM, vol. 20, no. 2, pp. 87–93, 1977.

REFINEMENT APPROACH AND CONVEX HULL

ComRef-2D-ConvHull-approach

Another useful fact for our investigations is the relation between **SVM and the vertices of the convex hull called extreme points**, i. e., for the separable case finding the maximum margin between the two sets is equivalent to finding the closest points in the smallest convex sets that contain each class (the convex hulls).[1]



[1] K. P. Bennett and E. J. Bredensteiner, “Duality and Geometry in SVM Classifiers,” in Proc. of the 17th Int. Conf. on Machine Learning. Morgan Kaufmann, 2000, pp. 57–64.

REFINEMENT APPROACH AND CONVEX HULL

Lemma[1]

Let $d = 2$ be the dimension of the feature space and m be the number of objects in the dataset. Assume there exists a fusion of summands such that the classes are separable in $\mathbb{R}^2$ with respect to the fusion.

For the ComRef approach the search for the mentioned fusion of summands has the **time complexity** $\mathcal{O}(p^3 2^d)$ with $p \leq m$, $p \in \mathbb{N}^+$. The variable p represents the number of **extreme points**.

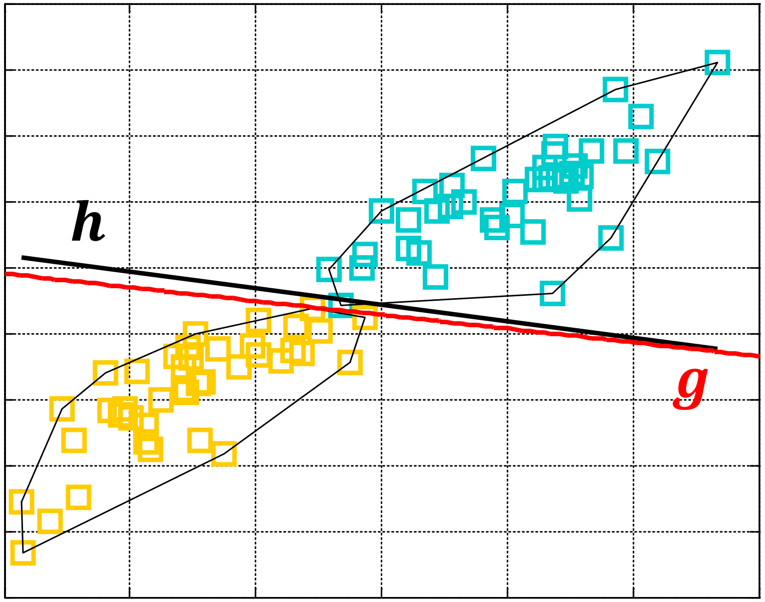
Proof

The time complexity for the calculation of fusions of summands in 2-dimensional space is $\mathcal{O}(m^3 2^d)$. The convex hull computation of the sample set for a fusion of summands in 2-dimensional space is executed within $\mathcal{O}(p \log p)$. The procedures need: $\mathcal{O}(p \log p) + \mathcal{O}(p^3) \rightarrow \mathcal{O}(p \log p + p^3) \rightarrow \mathcal{O}(p^3)$(convex hull calculation and SVM for extreme points). Separable case: We can restrict ComRef search to only the search on the extreme points in 2-dimensional spaces. As $p \leq m$ holds for the number of extreme points, the proof of the lemma is fulfilled. ■

[1] H. Dörksen and V. Lohweg, “Combinatorial Refinement of Feature Weighting for Linear Classification,” in 19th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA 2014), 2014.

REFINEMENT APPROACH AND CONVEX HULL

$\boldsymbol{Example}$: $\mathbf{40}$ objects in each class (Seeds)



Classification rate of linear SVM for $\boldsymbol{h}$: $\mathbf{98.75\%}$

classification rate for $\boldsymbol{g}$: $\mathbf{100\%}$
vertices of convex hulls: $\mathbf{19}$

RESULTS

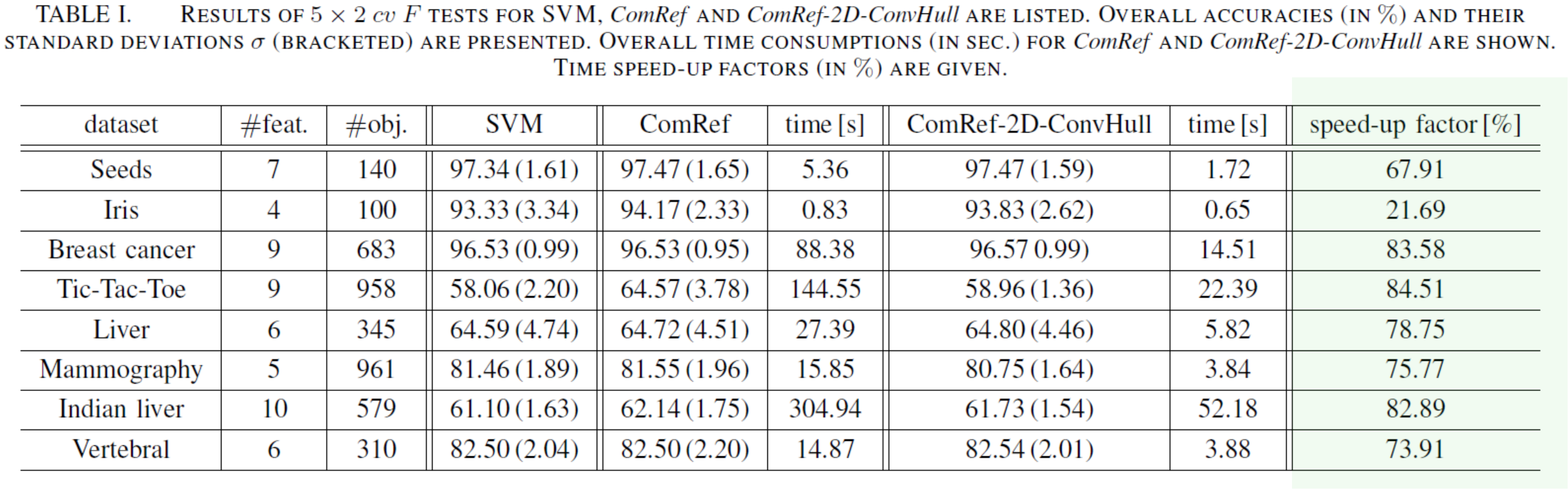


TABLE I. Results of 5 × 2 *cv F* tests for SVM, *ComRef* and *ComRef-2D-ConvHull* are listed. Overall accuracies (in %) and their standard deviations σ (bracketed) are presented. Overall time consumptions (in sec.) for *ComRef* and *ComRef-2D-ConvHull* are shown. Time speed-up factors (in %) are given.

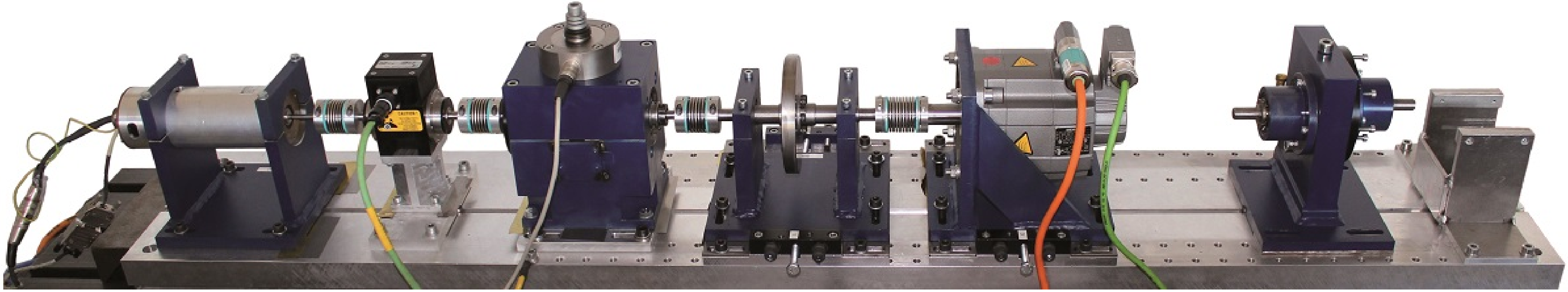
dataset	#feat.	#obj.	SVM	ComRef	time [s]	ComRef-2D-ConvHull	time [s]	speed-up factor [%]
Seeds	7	140	97.34 (1.61)	97.47 (1.65)	5.36	97.47 (1.59)	1.72	67.91
Iris	4	100	93.33 (3.34)	94.17 (2.33)	0.83	93.83 (2.62)	0.65	21.69
Breast cancer	9	683	96.53 (0.99)	96.53 (0.95)	88.38	96.57 0.99)	14.51	83.58
Tic-Tac-Toe	9	958	58.06 (2.20)	64.57 (3.78)	144.55	58.96 (1.36)	22.39	84.51
Liver	6	345	64.59 (4.74)	64.72 (4.51)	27.39	64.80 (4.46)	5.82	78.75
Mammography	5	961	81.46 (1.89)	81.55 (1.96)	15.85	80.75 (1.64)	3.84	75.77
Indian liver	10	579	61.10 (1.63)	62.14 (1.75)	304.94	61.73 (1.54)	52.18	82.89
Vertebral	6	310	82.50 (2.04)	82.50 (2.20)	14.87	82.54 (2.01)	3.88	73.91

$5 \times 2\ cv\ F$ cross-validation test (recommended for comparing classification algorithms[1])

[1] T. G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms,” Neural Computation, vol. 10, pp. 1895–1923, 1998.

RESULTS

Motor Drive Diagnosis[1]: motors as CPS which are networked



e.g. one motor with **53190** measurement-attributes for **72** features

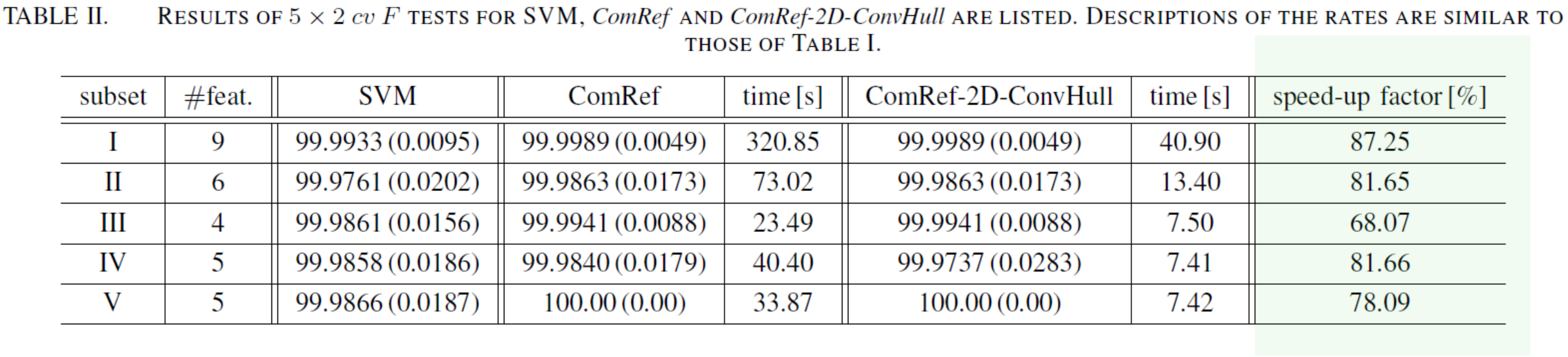


TABLE II. RESULTS OF 5 × 2 *cv F* TESTS FOR SVM, *ComRef* AND *ComRef-2D-ConvHull* ARE LISTED. DESCRIPTIONS OF THE RATES ARE SIMILAR TO THOSE OF TABLE I.

subset	#feat.	SVM	ComRef	time [s]	ComRef-2D-ConvHull	time [s]	speed-up factor [%]
I	9	99.9933 (0.0095)	99.9989 (0.0049)	320.85	99.9989 (0.0049)	40.90	87.25
II	6	99.9761 (0.0202)	99.9863 (0.0173)	73.02	99.9863 (0.0173)	13.40	81.65
III	4	99.9861 (0.0156)	99.9941 (0.0088)	23.49	99.9941 (0.0088)	7.50	68.07
IV	5	99.9858 (0.0186)	99.9840 (0.0179)	40.40	99.9737 (0.0283)	7.41	81.66
V	5	99.9866 (0.0187)	100.00 (0.00)	33.87	100.00 (0.00)	7.42	78.09

[1] C. Bayer, M. Bator, U. Mönks, A. Dicks, O. Enge-Rosenblatt, and V. Lohweg, "Sensorless Drive Diagnosis Using Automated Feature Extraction, Significance Ranking and Reduction," in 18th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA 2013), C. Seatzu and R. Zurawski, Eds. IEEE, 2013, pp. 1–4.

Summary

Number of objects plays an important role for:

- Design of accurate classifier
- Execution time

Refinement Approach with **Convex Hull** is able to:

- Increase classification accuracy
- Reduce execution time of refinement

Homework: Exercises and Labs

for the next week prepare practical exercises and labs from
Exercises Lec 11 (you will find it in the download area)

Online-Exercises coming now

We will discuss **Exercises Lec 10**