Choosing an Appropriate Performance Measure

Classification of EEG-Data with Varying Class Distribution

Sirko Straube[1], Jan Hendrik Metzen[1], Anett Seeland[2],
Mario Krell[1], Elsa Kirchner[1,2]

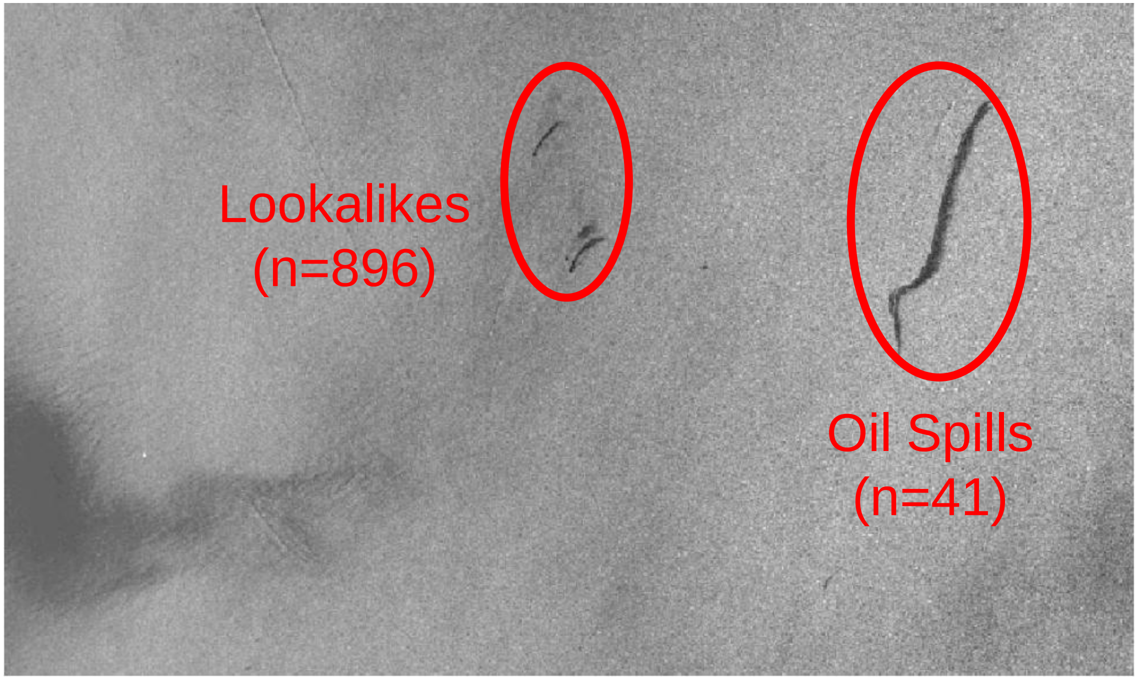
[1]Workgroup Robotics, Faculty of Mathematics and Computer Science,
University of Bremen

[2]Robotics Innovation Center, DFKI GmbH

Director: Prof. Dr. Frank Kirchner
www.dfki.de/robotics
robotics@dfki.de

Nanosymposium Session 753: *Data Analysis and Statistics IV*
Wednesday, November 16, 2011

Machine Learning: Imbalance Problem



*A classifier that **labels all regions as lookalikes** will achieve an **accuracy of 96%**. Although this looks high, **the classifier would be useless** because it totally **fails to achieve the fundamental goal** of oil spill detection. By contrast, a system achieving 94% on spills and 94% on nonspills will have a worse accuracy and yet be deemed highly successful; very few spills would be missed and the number of false alarms would be small.*

Kubat, Holte & Matwin (1998)

Unbalanced Classes are Common

The basic psychophysical process, we believe, is ***comparison****. All psychophysical judgments are of one stimulus relative to another; designs differ in the nature and difficulty of the comparison to be made.* [Macmillan & Creelman, 2005]

Signal Detection Paradigms:

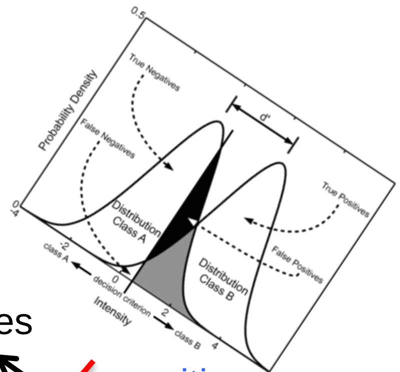
- Yes-No
- Same-Different
- Rating Paradigm
- Forced-Choice
- Matching-to-Sample

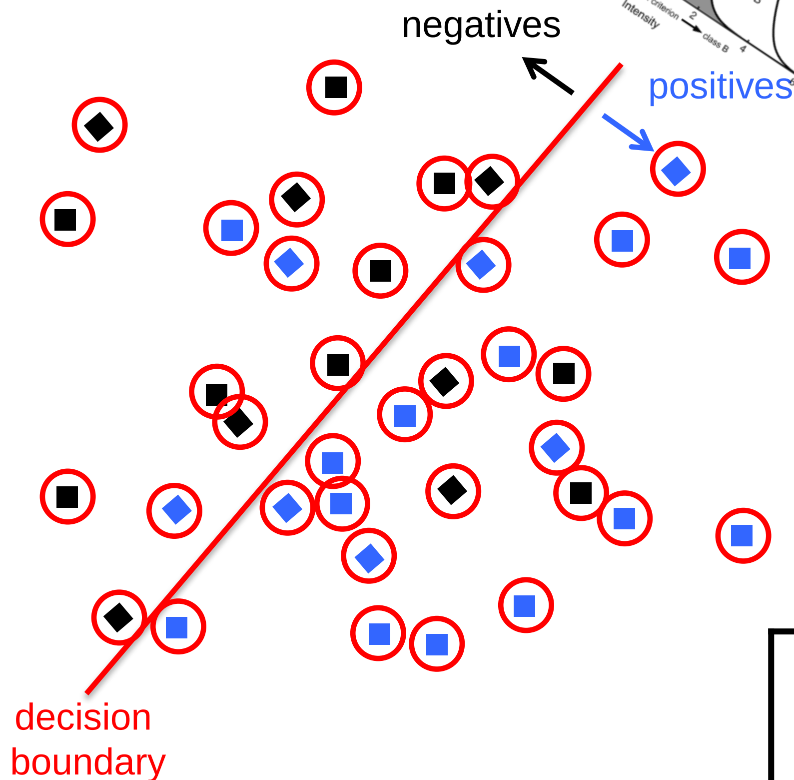
<u>However…</u>

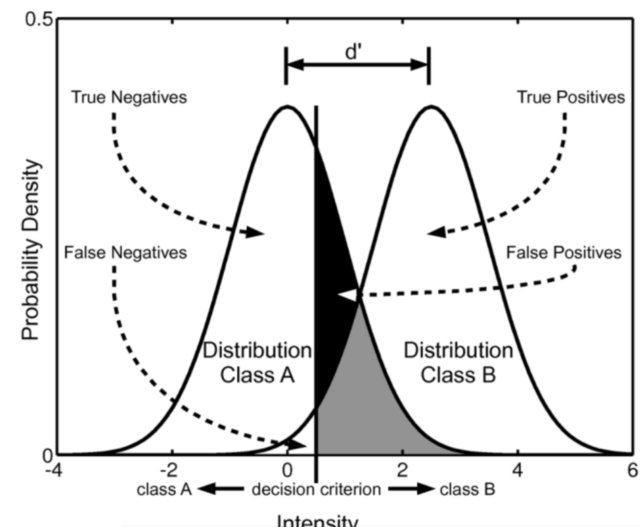
- We rarely experience balanced classes in everyday life
- other experimental paradigms exist, e.g., the oddball, where the classes are not balanced

The Confusion Matrix

		Classified as...	
		Positive	Negative
Labeled as...	Positive	TP	FN
	Negative	FP	TN







		Classified as...	
		Positive	Negative
Labeled as...	Positive		
	Negative		

Metrics: Measures of Performance

I. Accuracy
$$\frac{TP+TN}{TP+FP+FN+TN}$$

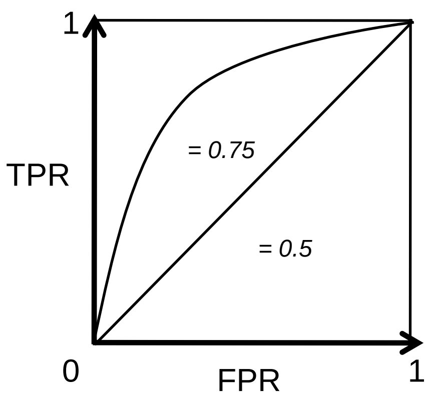
II. Weighted Accuracy $\quad w*TPR + (1-w)*TNR$

(also Balanced Accuracy for *w=0.5*)

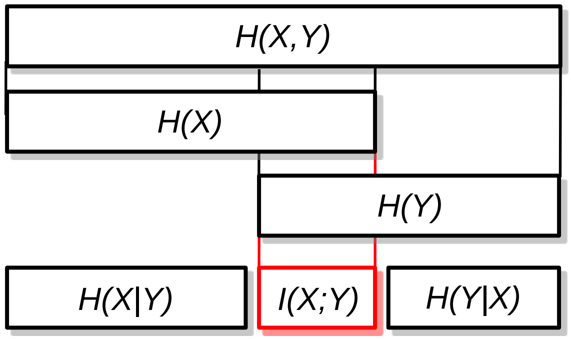
III. F-Measure
$$\frac{2*\Pr*\mathrm{Re}}{\Pr+\mathrm{Re}} \begin{cases} \text{Precision} & \dfrac{TP}{TP+FP} \\ \text{Recall (=TPR)} & \dfrac{TP}{TP+FN} \end{cases}$$

	in %	Classified as...	
		Positive	Negative
Labeled as...	Positive	TPR	FNR
	Negative	FPR	TNR

IV. Area Under Curve (AUC)



V. Mutual Information (MI)



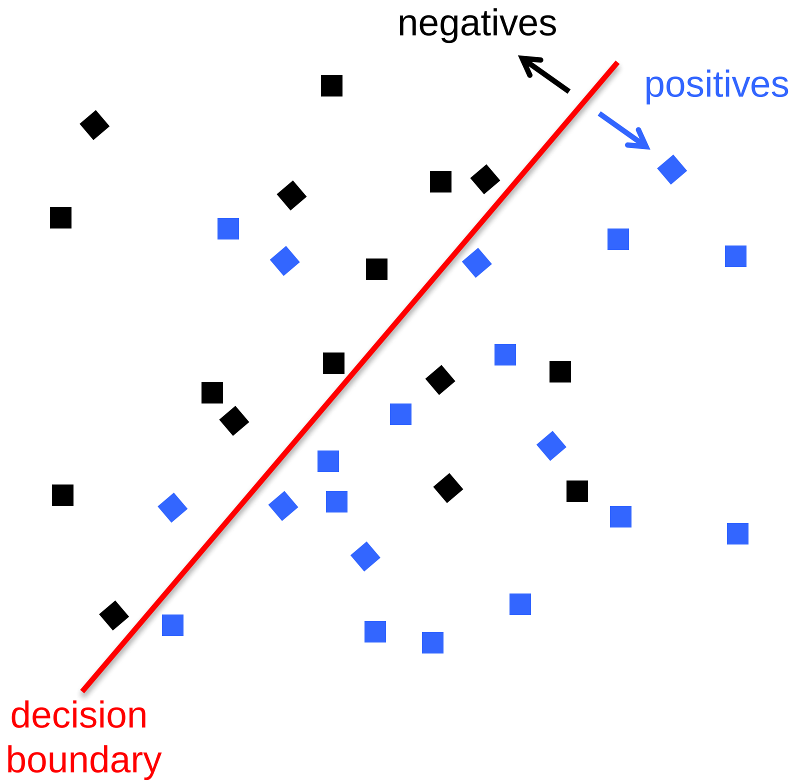
$$TPR = \frac{TP}{TP+FN}$$

$$TNR = \frac{TN}{TN+FP}$$

Example: Changing the Class Ratio

		Classified as...	
		Positive	Negative
Labeled as...	Positive	17	3
	Negative	4	12

TPR: 0.85
FNR: 0.15
TNR: 0.75
FPR: 0.25

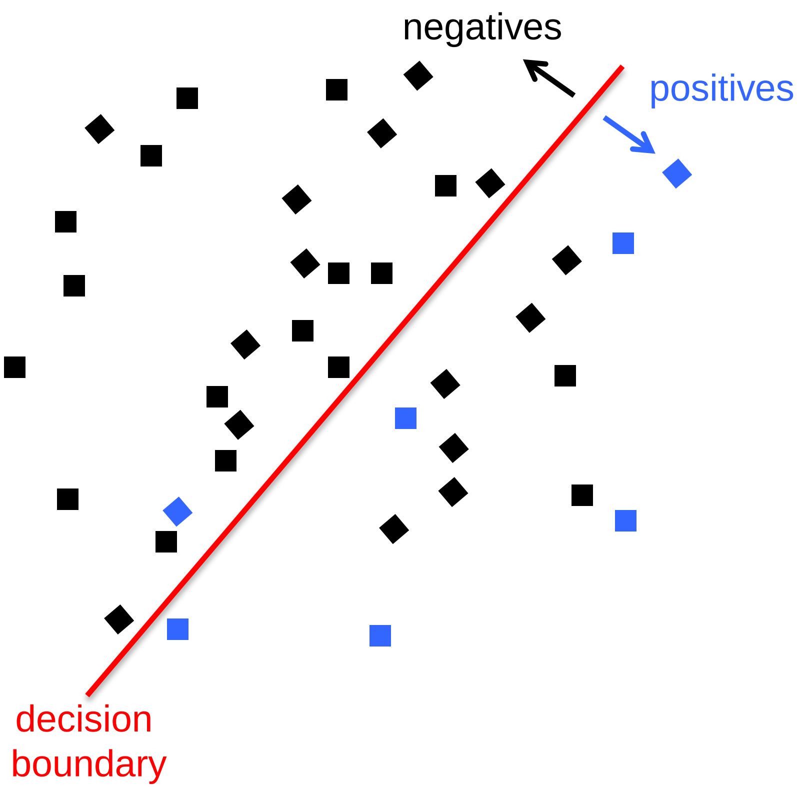


Accuracy	0.81
F-Measure	0.83
Mutual Information	0.28
AUC	0.75
Balanced Accuracy (w=0.5)	0.80

Example: Changing the Class Ratio

		Classified as...	
		Positive	Negative
Labeled as...	Positive	6	1
	Negative	8	24

TPR: 0.85 0.86
FNR: 0.15 0.14
TNR: 0.75 0.75
FPR: 0.25 0.25

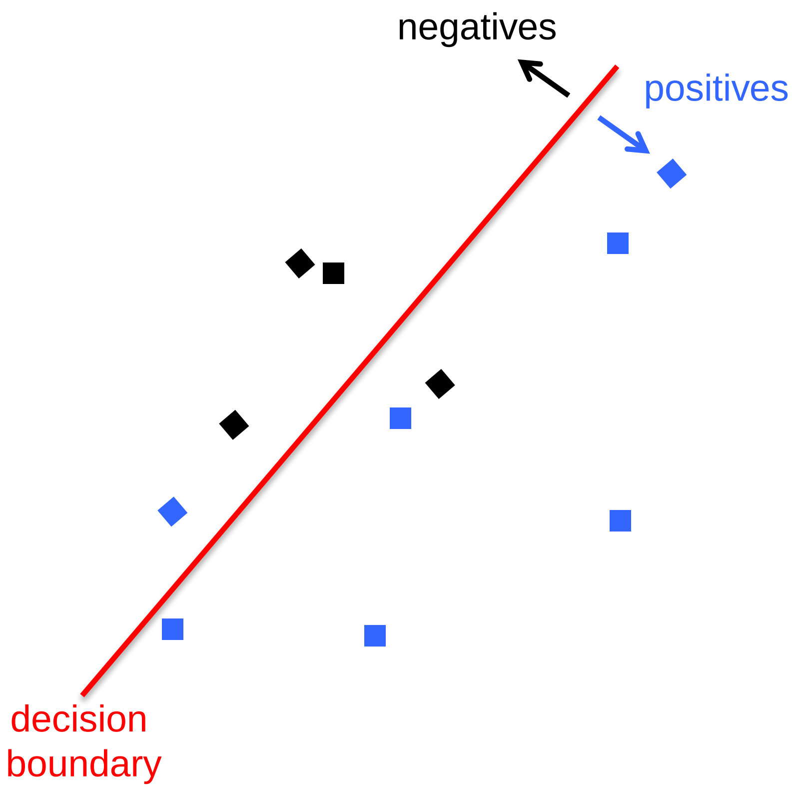


Accuracy	0.81	0.76
F-Measure	0.83	0.57
Mutual Information	0.28	0.17
AUC	0.75	0.75
Balanced Accuracy (w=0.5)	0.80	0.81

Example: Changing the Class Ratio

		Classified as...	
		Positive	Negative
Labeled as...	Positive	6	1
	Negative	1	3

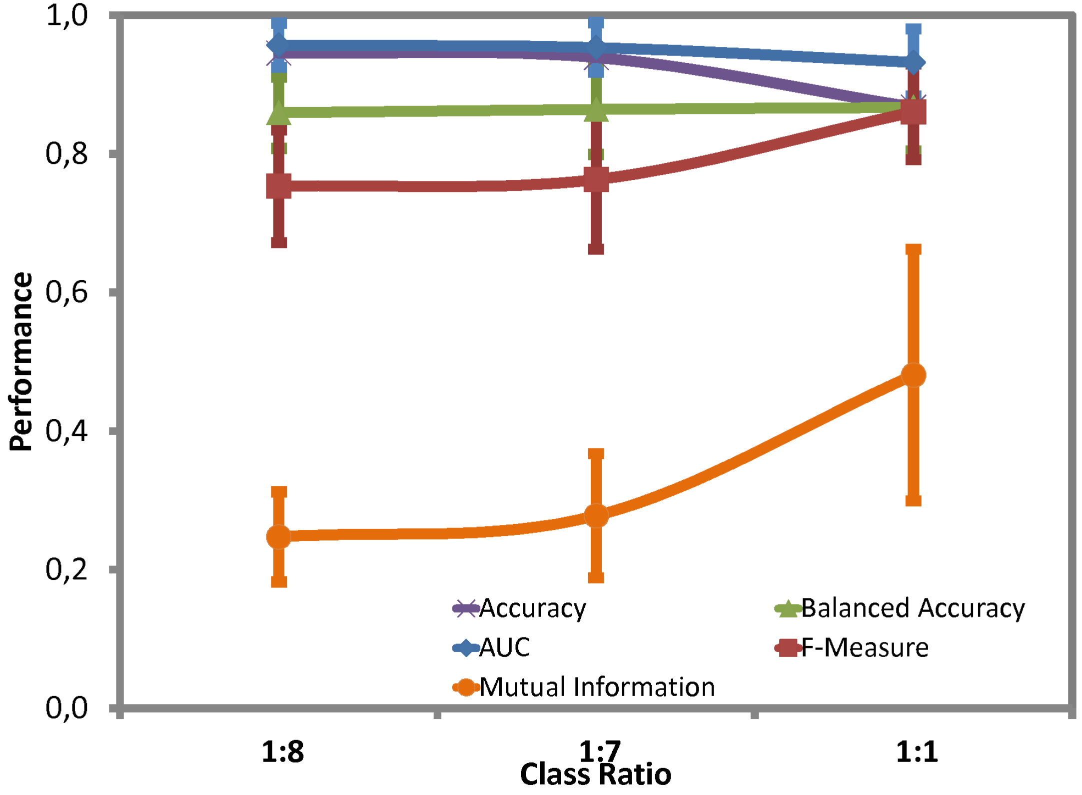
TPR: 0.85 0.86
FNR: 0.15 0.14
TNR: 0.75 0.75
FPR: 0.25 0.25



Accuracy	0.81	0.76	0.82
F-Measure	0.83	0.57	0.86
Mutual Information	0.28	0.17	0.27
AUC	0.75	0.75	0.75
Balanced Accuracy (w=0.5)	0.80	0.81	0.81

Effect on Experimental Data

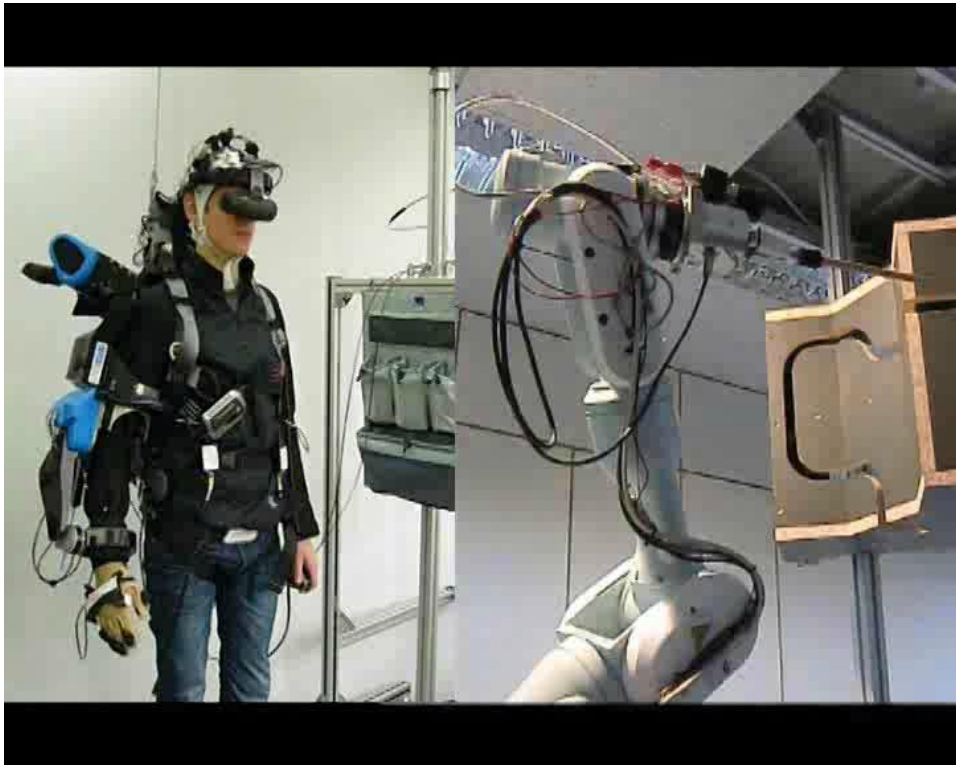
- post-hoc analysis
- oddball (like) paradigm
- evaluation of classifier performance used to classify EEG data
- 1 subject, 5 runs
- total: 100 important warnings, 749 standard stimuli



Conclusions

1. Unbalanced class distributions are common in everyday life.
   - they often have an effect on the metric when evaluating applications or studying behavior in a natural situation
2. There is no “perfect“ metric for measuring performance.
3. One has to consider metric properties, class distributions and question at hand.
4. Some metrics are sensitive to the class distribution…
   - Accuracy, F-Measure and Mutual Information
5. …some are not.
   - Weighted & Balanced Accuracy, Area under ROC-Curve
6. Important to note: Accuracy and Balanced Accuracy are both intuitive.

Application



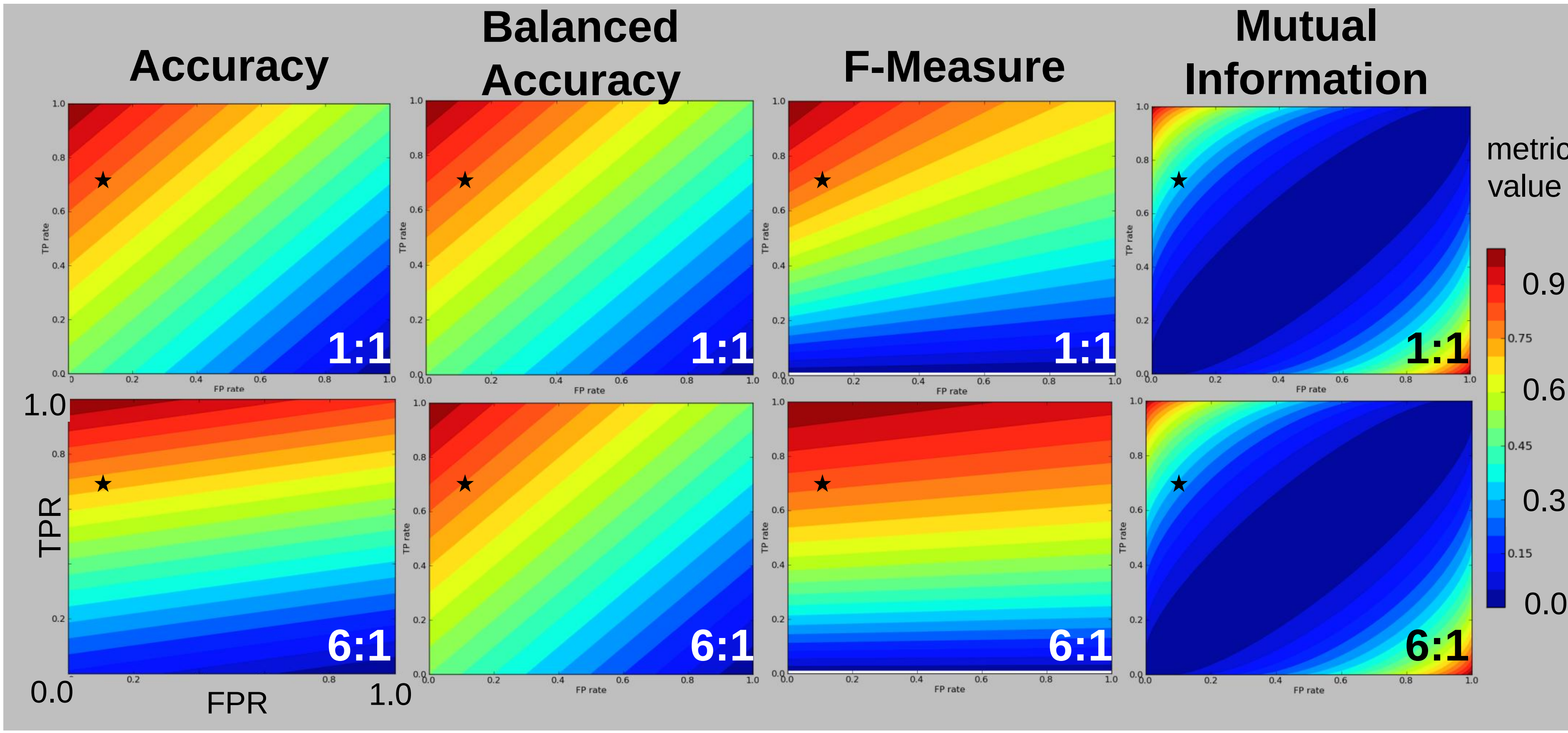
- the occurence of relevant events is not predictable…
- …i.e., we do not know how relevant and irrelevant classes are distributed

=> How to judge how well it worked?

Thank you very much for your attention!

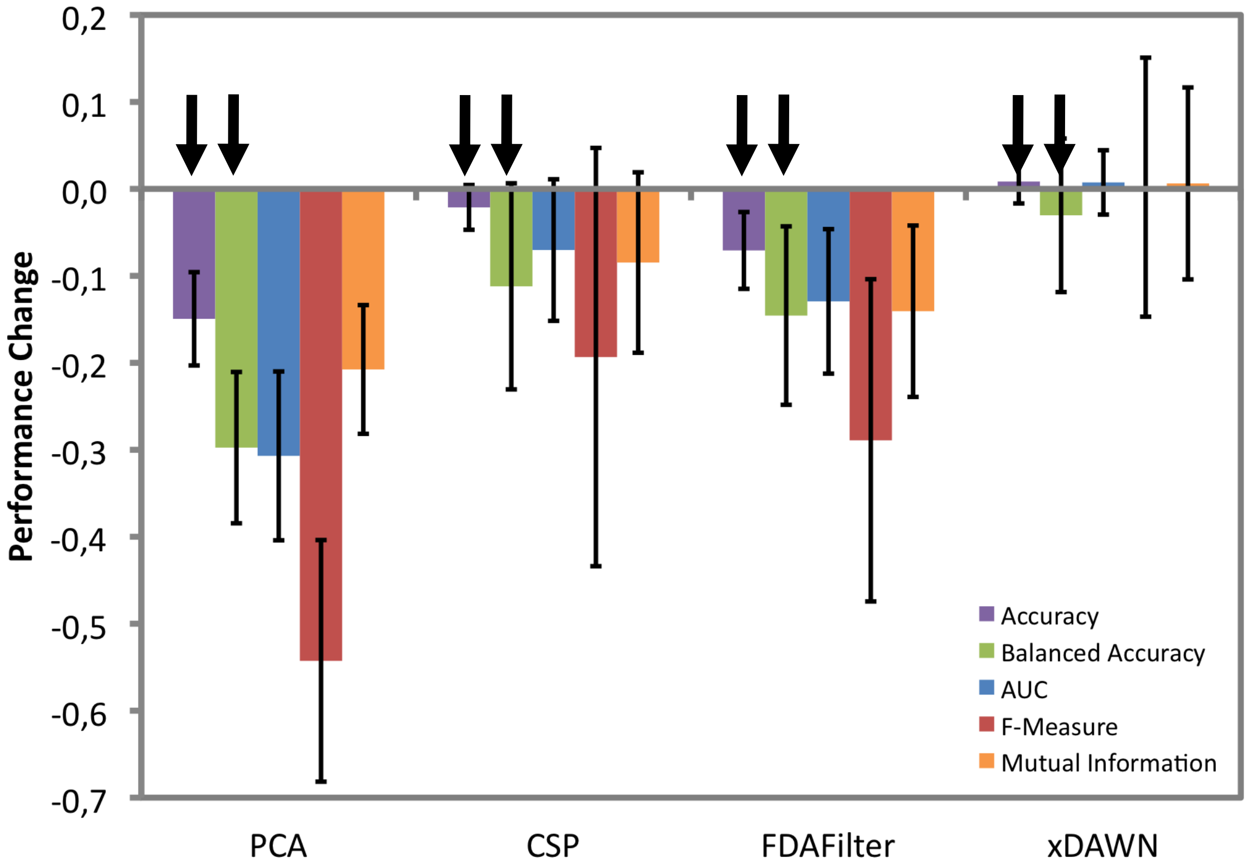
Supplement

Sensitivity to Class Ratio

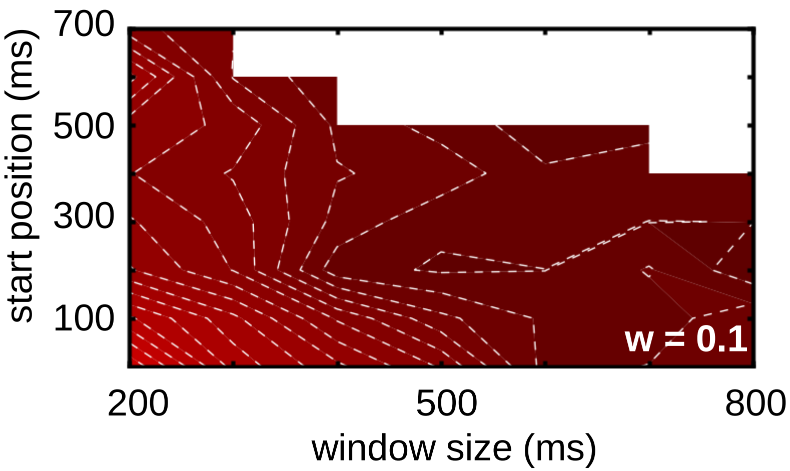


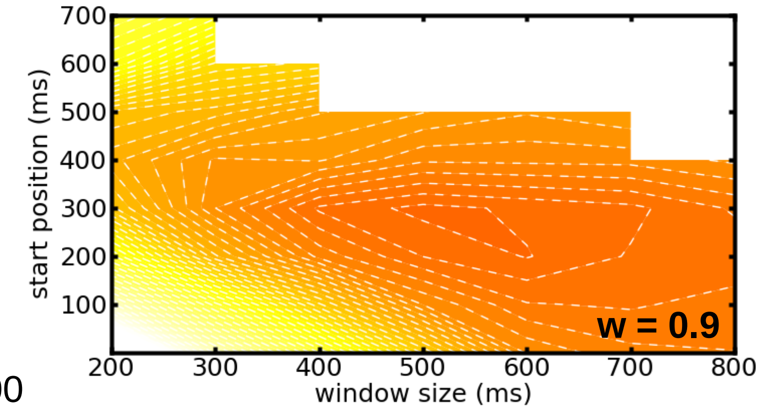
Different Metric, Different Result?

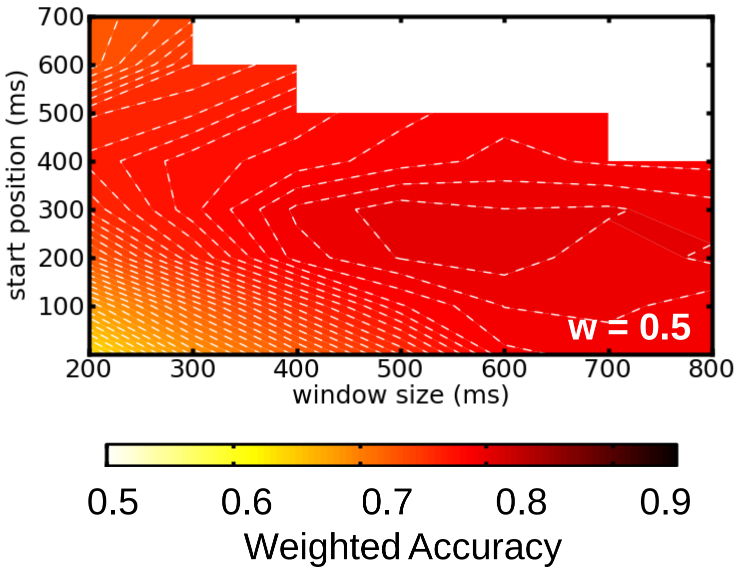
- post-hoc analysis
- 5 subjects, 2 sessions, 6 runs
- Aim: Reduce features using a spatial filter and reducing filter channels.
- Question: How does preprocessing using a spatial filter affect the performance of the classifier?

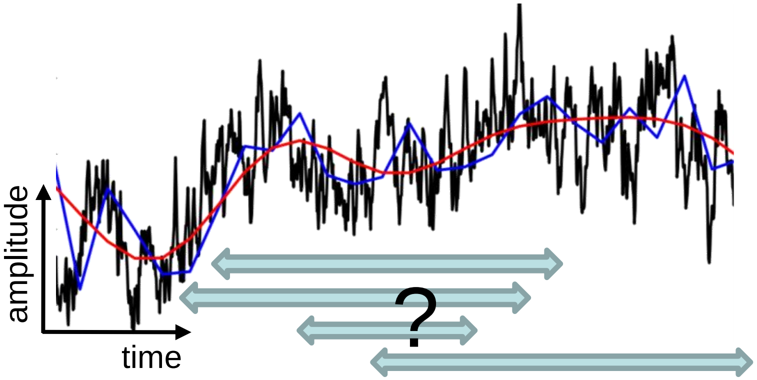


Weighted Accuracy: Effect of Weight





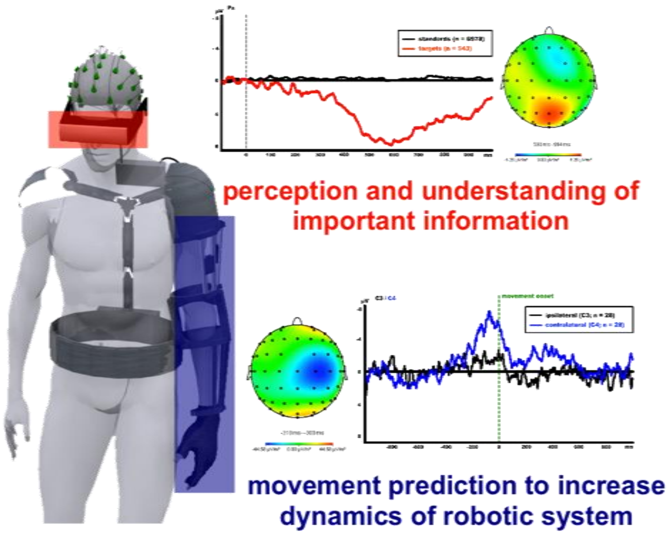




Weighted Accuracy $w*TPR + (1-w)*TNR$

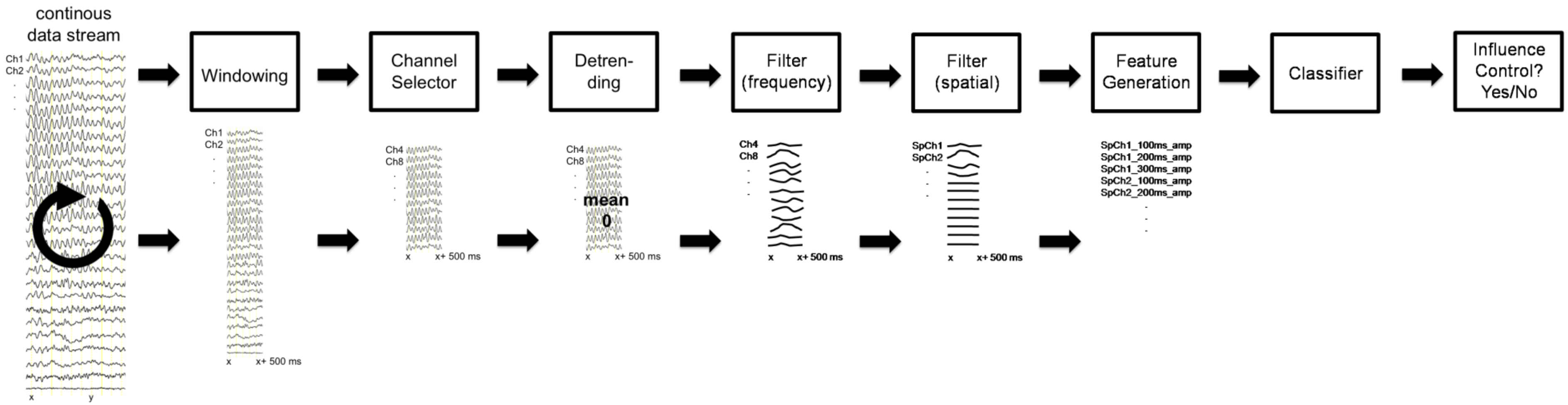
(also Balanced Accuracy for $w=0.5$)

Problem: How to Rate Performance?



- in reality neither the occurence of important events…
- …nor the occurence of the important movements is predictable
- i.e., we do not know how the relevant classes are distributed

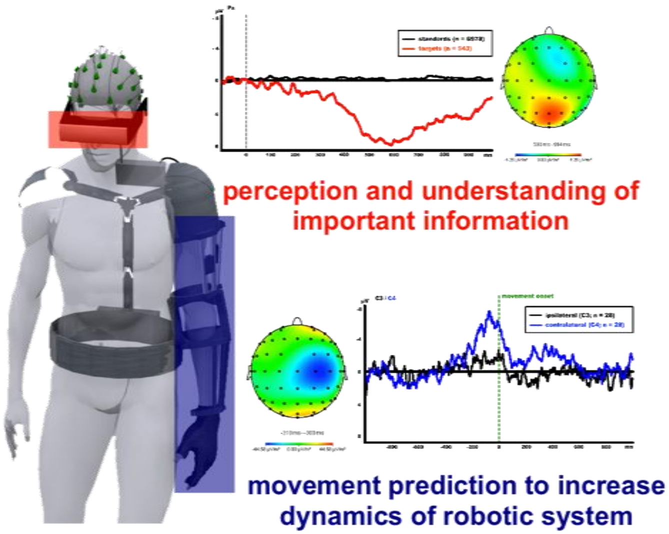
=> How to judge how well it worked?

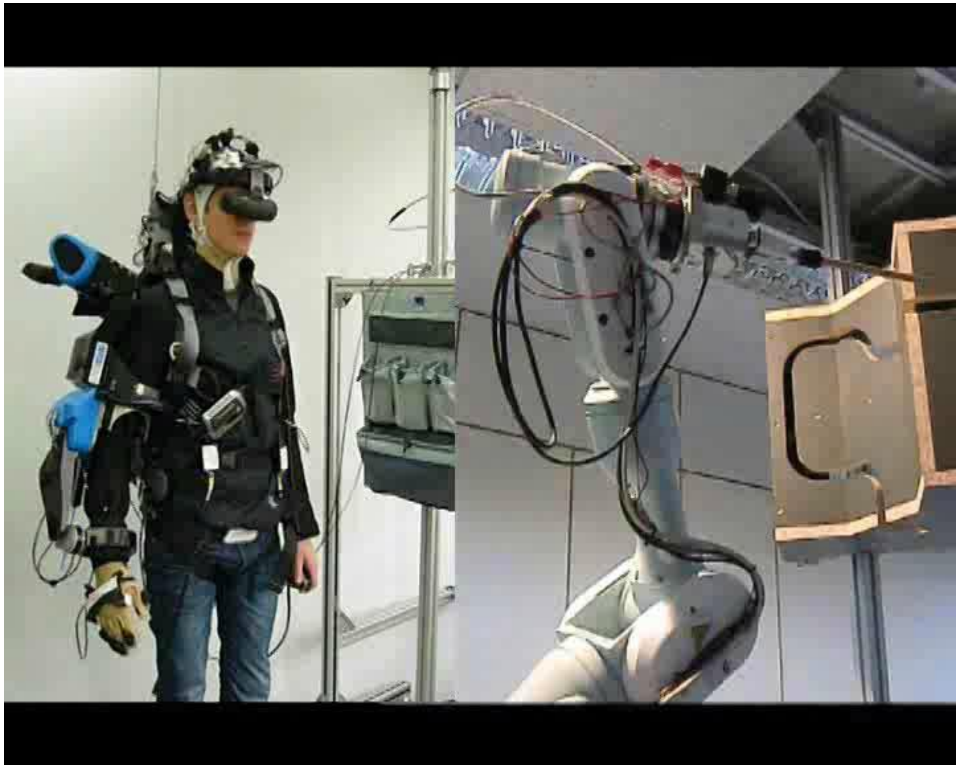


An Application Scenario

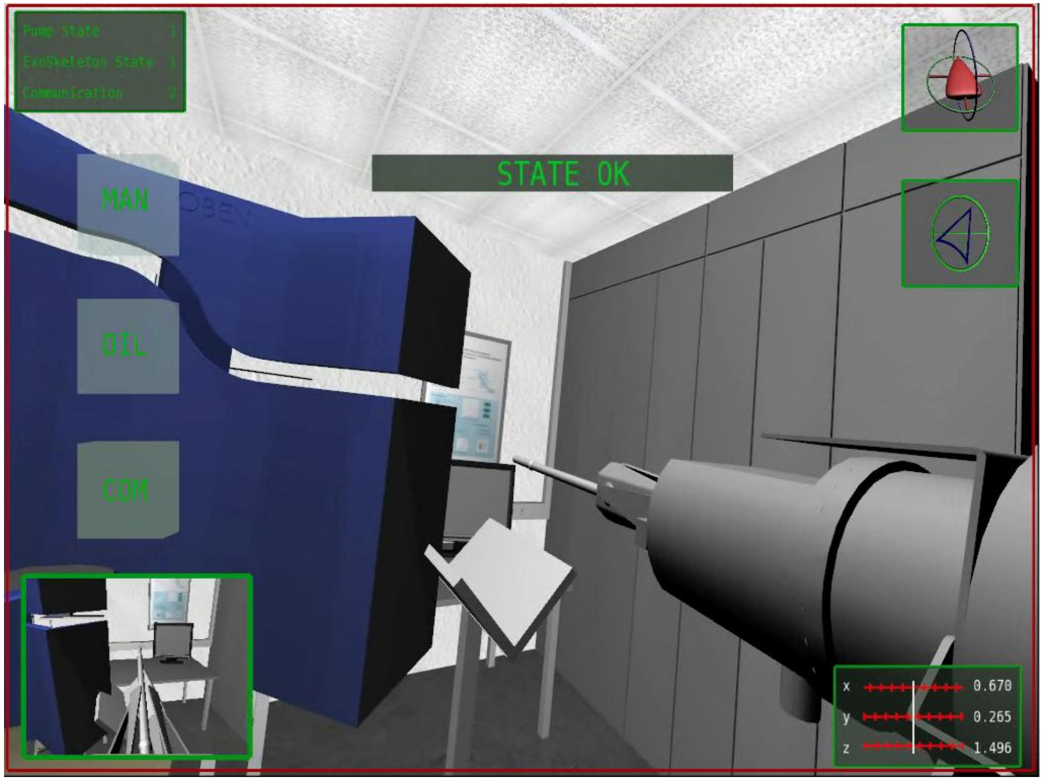
- the occurence of relevant events is not predictable…
- …i.e., we do not know how relevant and irrelevant classes are distributed

=> How to judge how well it worked?





Operator's View

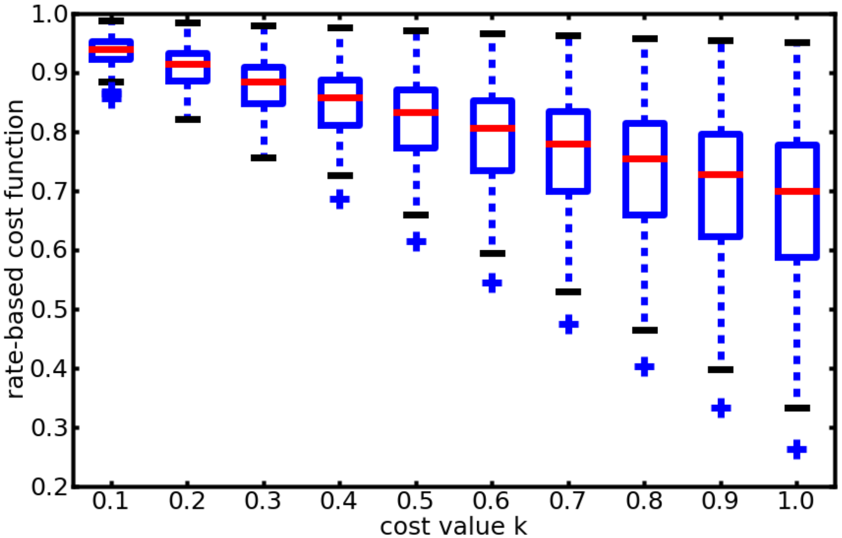


Secondary Supplement

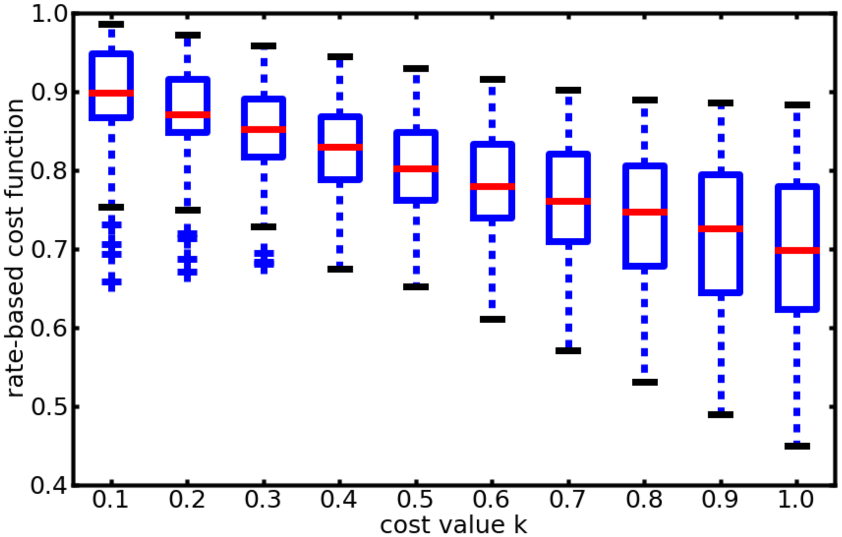
Windowing Study: Single Window

Window: 300-800 ms

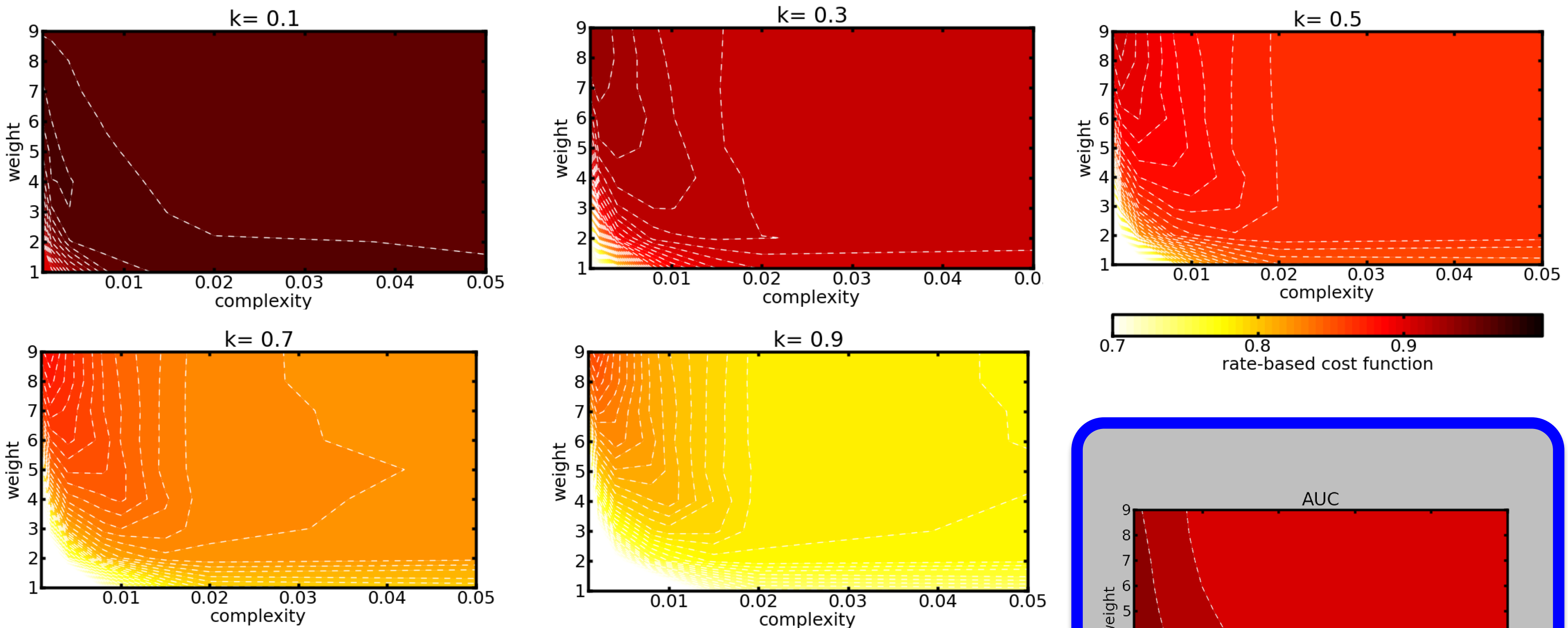
Training = Testing: Target vs. Standard

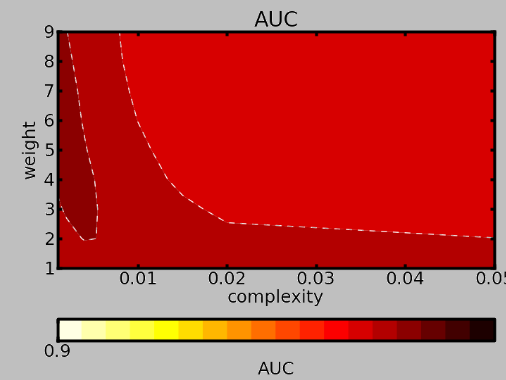


Training ≠ Testing: Target vs Missed Target



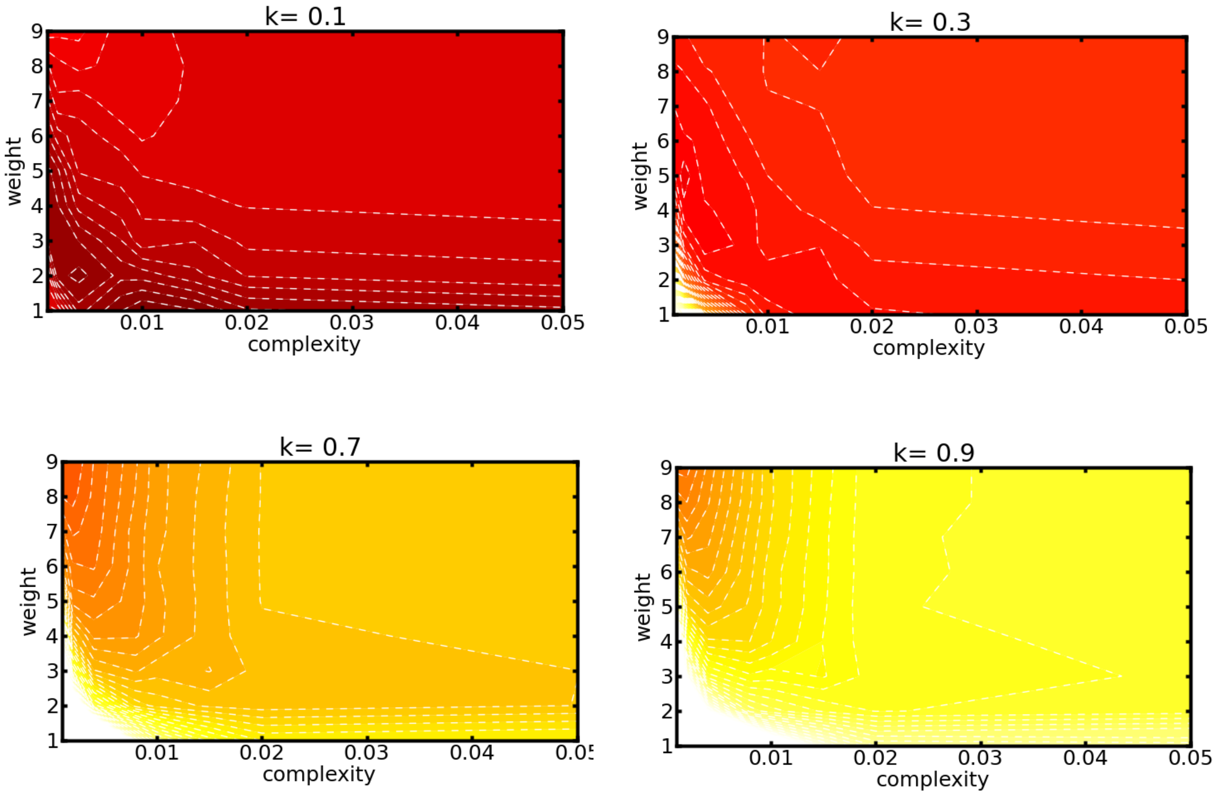
SVM-Optimization: Target vs Standard

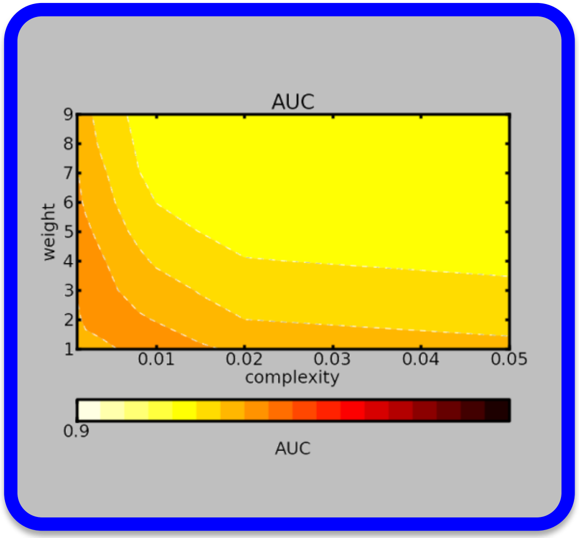




5 subjects
6 sets (2 Sessions)
cross validation
tested complexity: [0.001, 0.002, 0.004, 0.008, 0.01, 0.015, 0.02, 0.25, 0.3, 0.5, 0.8, 1.0, 3.0, 5.0, 10.0, 100]
tested weight: [1, 2, 3, 4, 5, 6, 7, 8, 9]

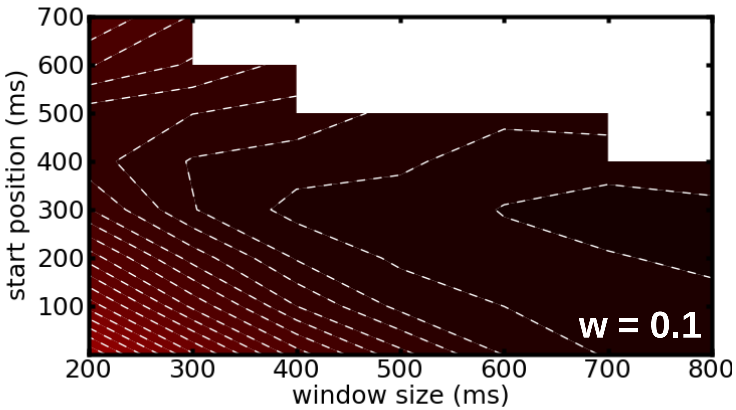
SVM-Optimization: Target vs Missed Target

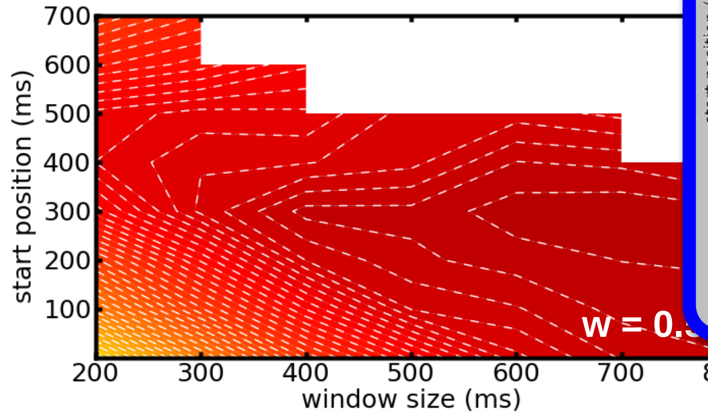


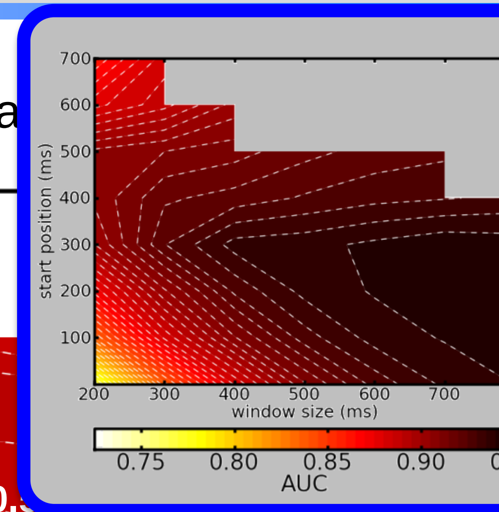


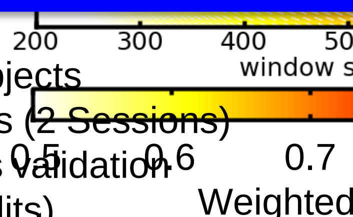
The Effect of the Weight Factor I

Training = Testing: Targets vs. Sta



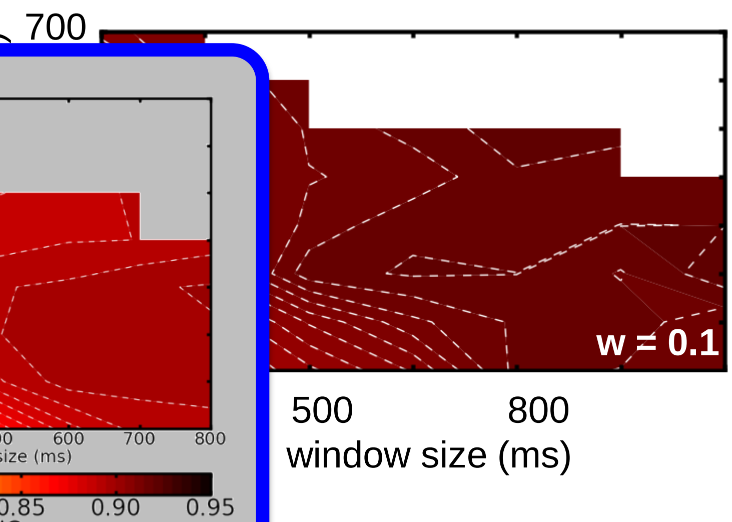


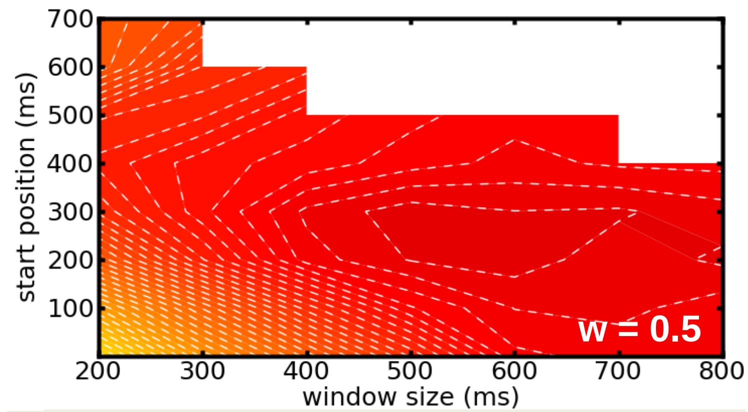


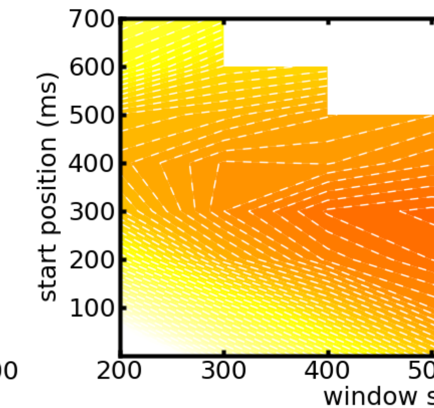


5 subjects
6 sets (2 Sessions)
cross validation (5 splits)

Training ≠ Testing: Targets vs. Missed Targets

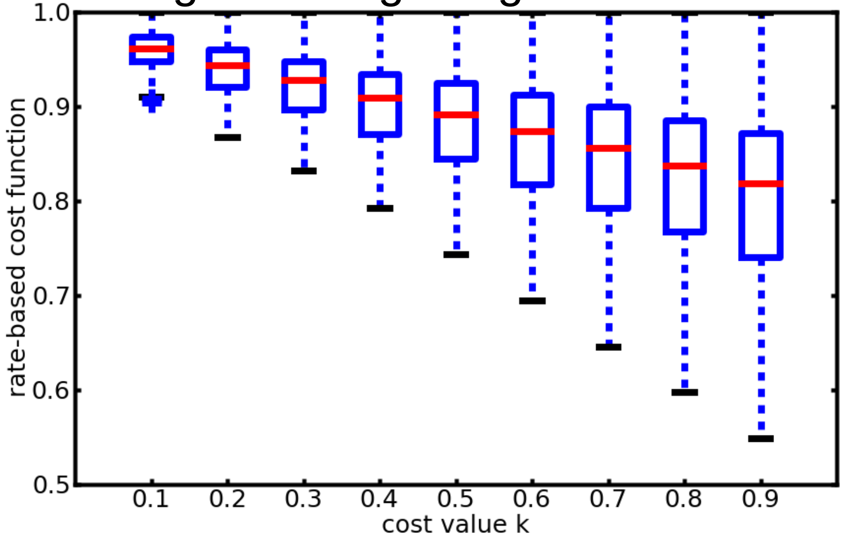




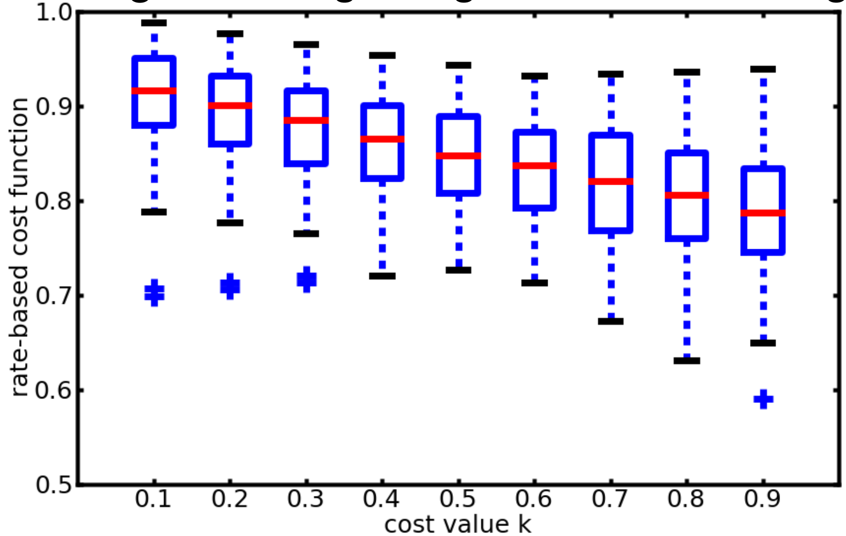


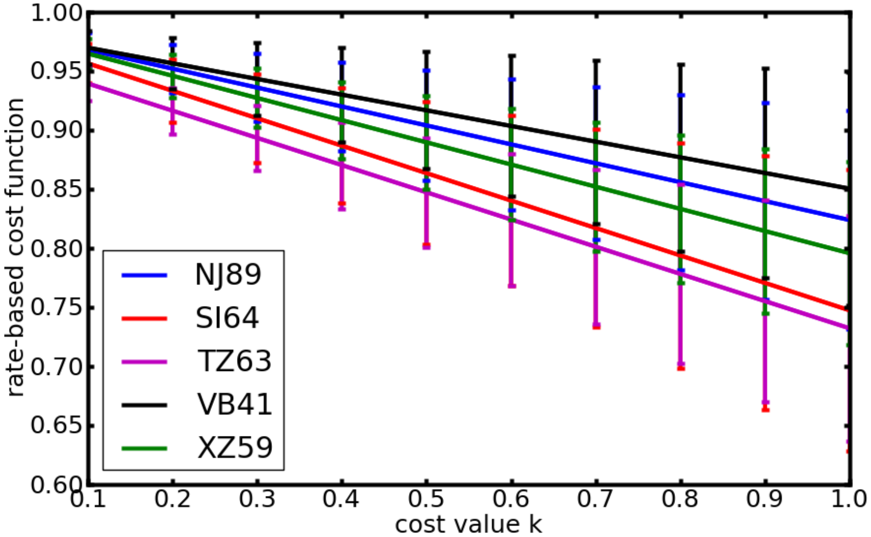
C=0.05, W=5

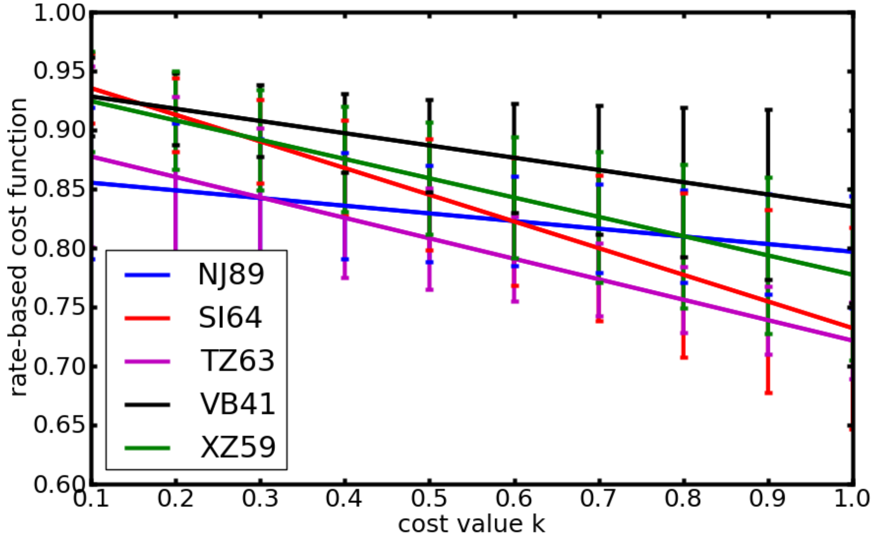
Training = Testing: Target vs. Standard



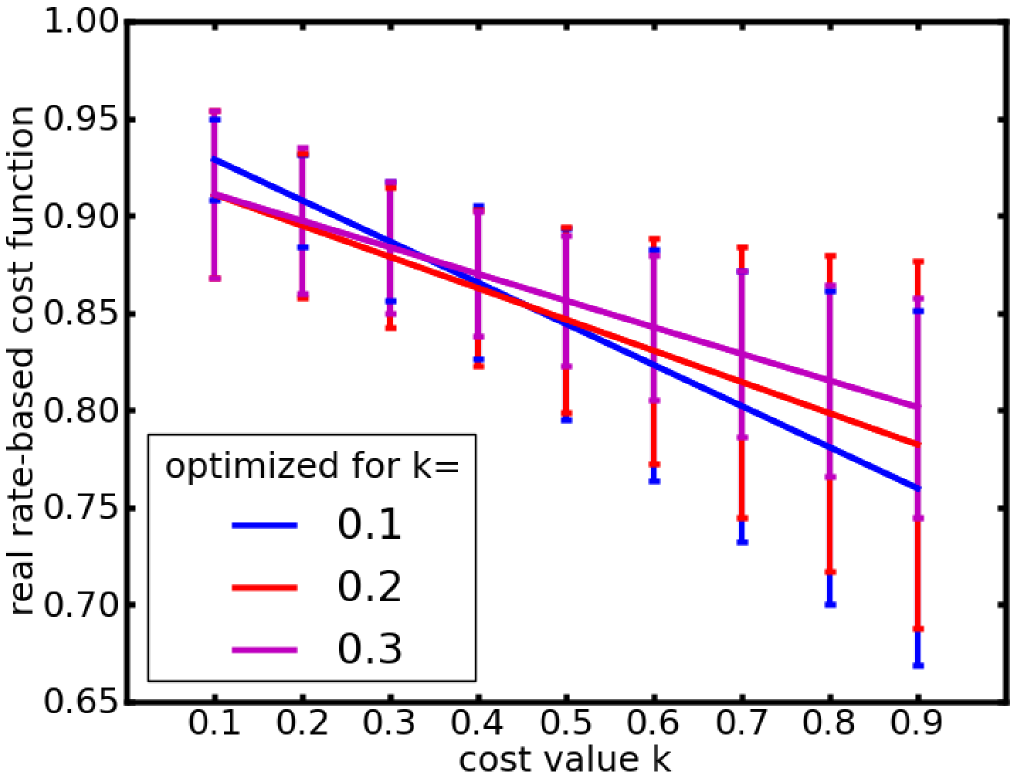
Training ≠ Testing: Target vs Missed Target





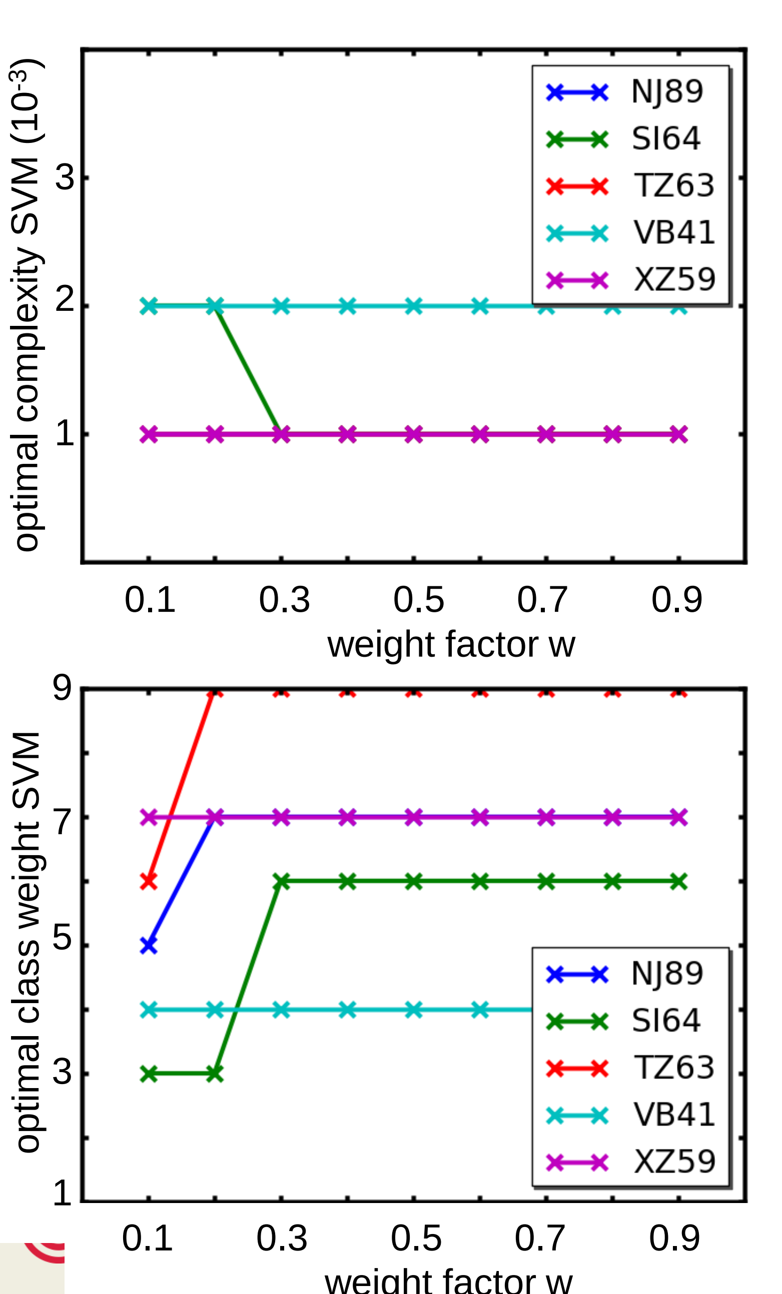


k optimal vs k real

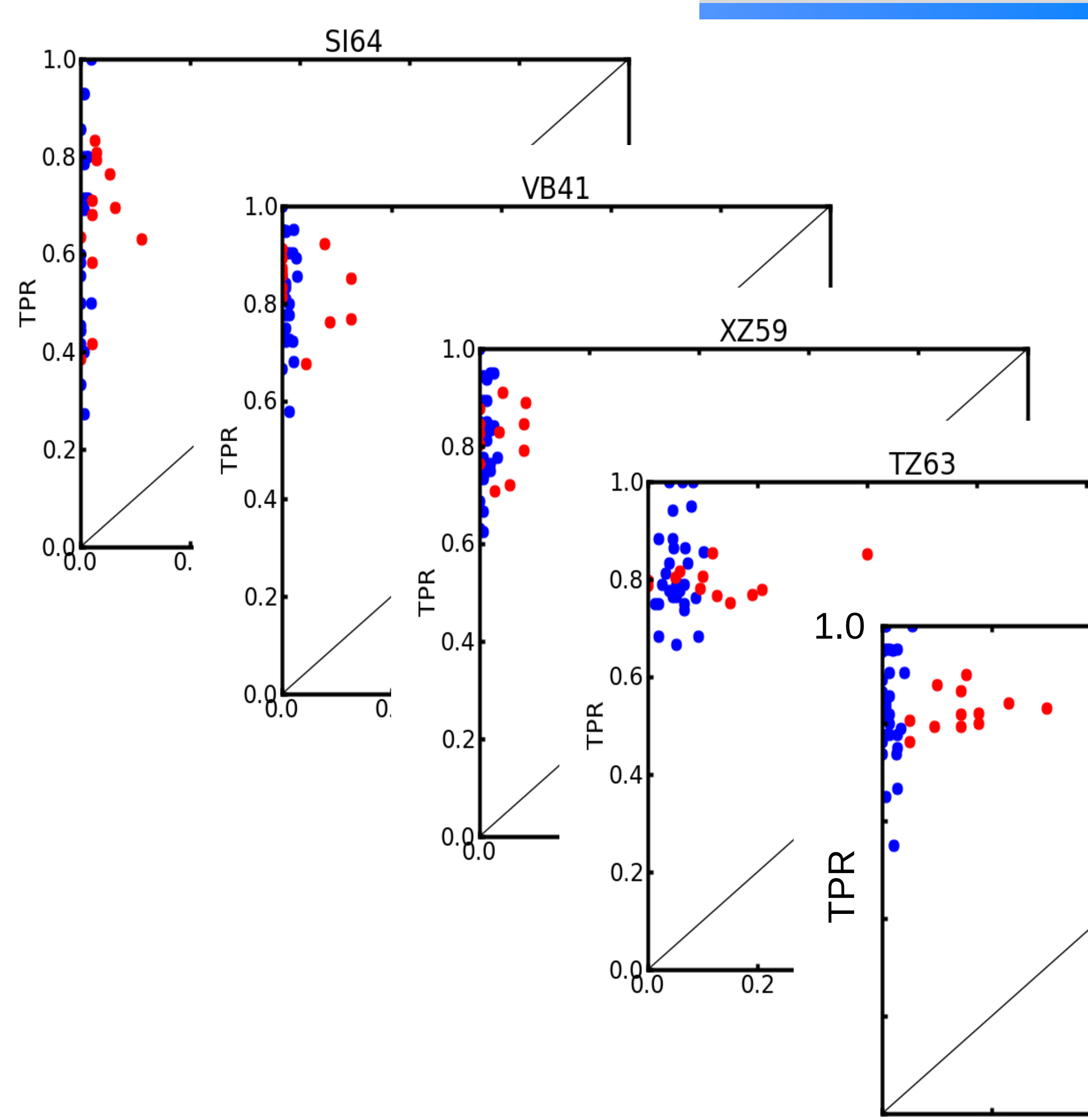


The Effect of the Weight Factor II

B



C



title

The Concept

- Trial vs. Event
  - A Practical Example for Balanced vs. Unbalanced Class Distribution
- How to Design an Experiment
  - Signal Detection: Comparison of (two) Stimulus Classes
  - Basic Concept: Equal Distribution of Classes
  - Machine Learning: The class imbalance problem
  - Also occuring in Experimental Paradigms: Oddball
- Six to Eight Metrics to Judge Performance
  - Accuracy, F-Measure, AUC, Balanced Accuracy, Weighted Accuracy, Mutual Information, Sensitivity, Specificity
- Application in a Behavioural Scenario: Classification of EEG Data Using SVMs
  - Classification of Important vs Unimportant Information
    - Focus: Unbalanced Class Distribution
  - Prediction of Movements
    - Focus: Varying Class Distributions

Conclusions

- Rate Based Cost Function is independent of class distributions
- The choice of k seems to be largely uncritical to investigate differences in preprocessing
- Optimization of SVM-Parameters is largely independent of cost factor k
  - for P3 case it seems suitable to use high weights (strengthen target class) and evaluate with low k (strengthen standard class)
- optimal values for k are still optimal after transfer to application case
  - global effect rules out local differences

Thank you!

DFKI Bremen & Universität Bremen
Robotics Innovations Center
Director: Prof. Dr. Frank Kirchner
www.dfki.de/robotics
robotics@dfki.de