Facing Imbalanced Data
Recommendations for the Use of Performance Metrics

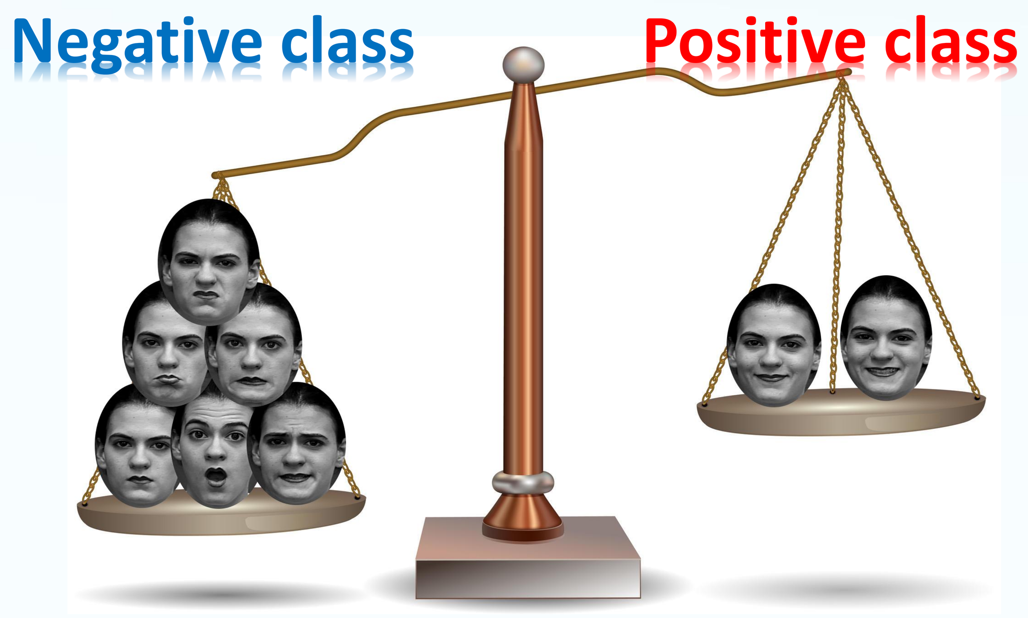
Laszlo A. Jeni [1] **Jeffrey F. Cohn** [1,2] **Fernando De La Torre** [1]

[1] Carnegie Mellon University, Pittsburgh, PA, laszlo.jeni@ieee.org, ftorre@cs.cmu.edu
[2] University of Pittsburgh, Pittsburgh, PA, jeffcohn@cs.cmu.edu

1. Introduction

Previous work on facial action unit (AU) recognition has emphasized face tracking and registration and the choice of features classifiers. Relatively little attention has been paid to **how imbalanced data may spoil performance metrics**.

Facial expression data is typically highly skewed. Imbalance in the test data distribution might produce misleading conclusions with certain metrics.



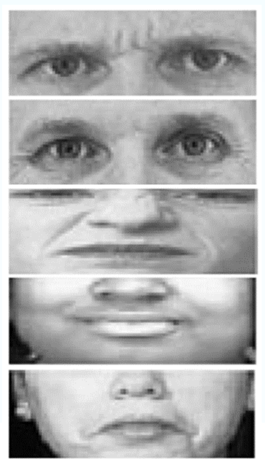
Question: Is F_1 = 0.3 good or bad performance?

Answer: It depends on the skew in the TEST set!

2. Skew in Datasets

Action unit classification is a typical two-class problem. The positive class is the given action unit that we want to detect, and the negative class contains all of the other examples. The imbalance of this type of data can be defined by the skew ratio between the classes:

$$\text{Skew} = \frac{\textit{negative examples}}{\textit{positive examples}}$$



AU	CK+		PAIN		RU-FACS	
	# of AUs	Skew	# of AUs	Skew	# of AUs	Skew
4	174	10.55	987	23.31	1028	113.44
6	113	16.78	5132	3.68	5184	21.69
9	65	29.91	422	55.86	55	2138.05
12	125	15.07	6627	2.62	18416	5.39
15	83	23.20	6	3998.33	3676	31.00

Examples of Skew ratios of action units from three datasets.

3. Performance Metrics

In a binary classification problem the labels are either positive or negative. The decision made by the classifier can be represented as a 2x2 confusion matrix.

Predicted Class \ Actual Class	1	-1
1	True Positives (TP)	False Positives (FP)
-1	False Negatives (FN)	True Negatives (TN)

Threshold Metrics:

Accuracy is the percentage of correctly classified positive and negative examples.

$$\text{Acc} = \frac{TP+TN}{TP+FP+TN+FN}$$

Precision is the fraction of recognized instances that are relevant.

$$\text{Prec} = \frac{TP}{TP+FP}$$

Recall is the fraction of relevant instances that are retrieved.

$$\text{Rec} = \frac{TP}{TP+FN}$$

F_1 score is the harmonic mean of precision and recall.

$$F_1 = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec}$$

Cohen's kappa is observed agreement normalized to expected agreement.

$$K = \frac{P_{Obs} - P_{Chance}}{1 - P_{Chance}}$$

Krippendorff's α is observed disagreement normalized to expected disagreement

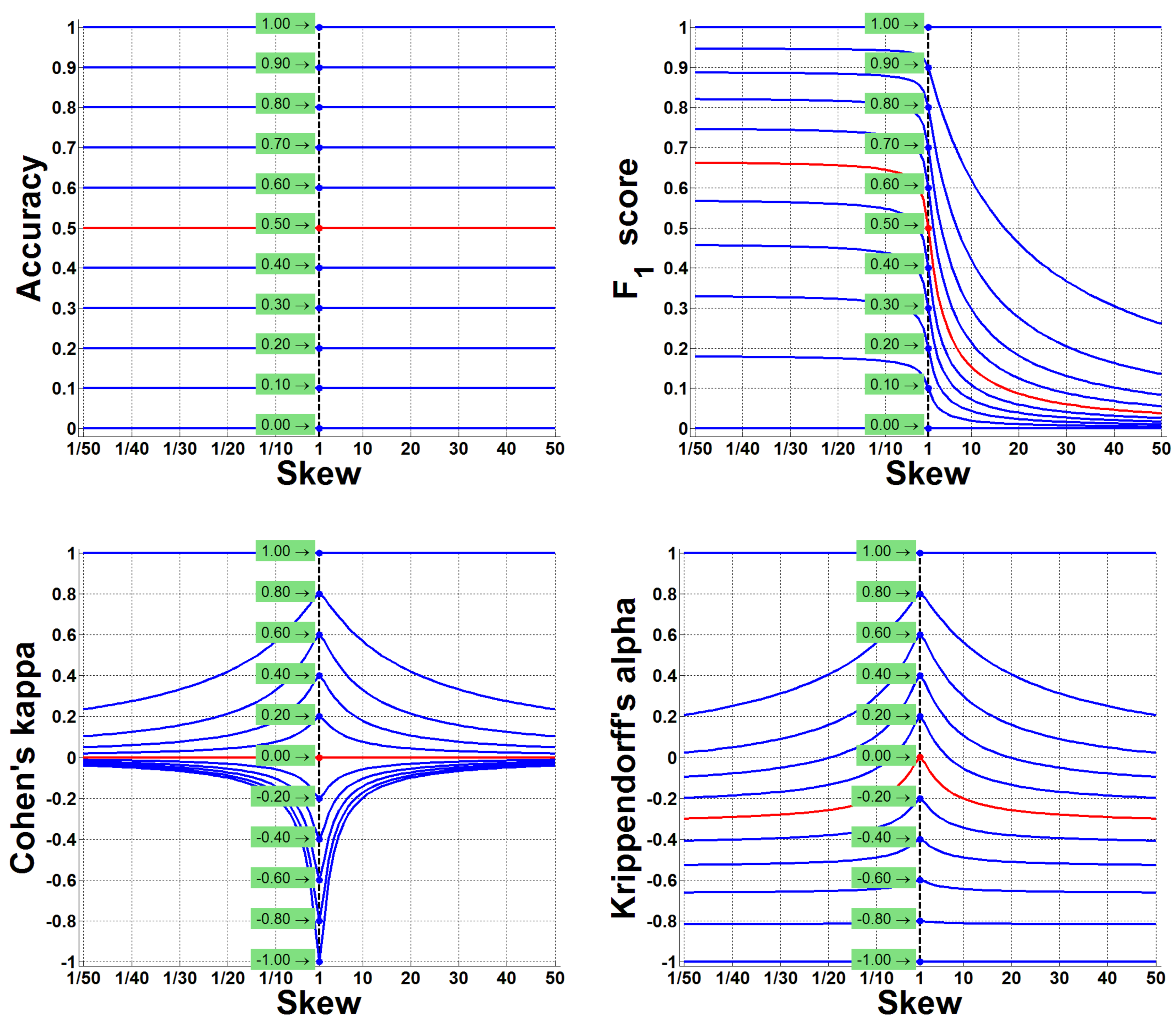
$$\alpha = 1 - \frac{D_{Obs}}{D_{Chance}}$$

Rank Metrics:

ROC curve shows the true positive rate as a function of false positive rate.

Precision-Recall curve shows the precision as the function of recall.

4. Behavior of Different Metrics



5. Skew Normalization

Different forms of re-sampling such as random over- and under-sampling can be used to balance the skewed distribution of the ***TEST partitions*** of the dataset before calculating the performance metrics.

Reporting both obtained performance metrics and skew-normalized scores, classifiers can be compared across databases free of confounds introduced by skew.

Code to compute skew-normalized scores for all of the metrics considered above and visualizations are available from:

http://www.pitt.edu/~jeffcohn/skew/

6. Results on Real Data

	AU	F_1 Orig.	F_1 Norm.	Kappa Orig.	Kappa Norm.	Alpha Orig.	Alpha Norm.	AUC-ROC Orig.	AUC-ROC Norm.
CK+	4	0.73	0.83	0.68	0.69	0.68	0.68	0.90	0.90
CK+	6	0.74	0.83	0.72	0.82	0.76	0.82	0.93	0.94
CK+	9	0.92	0.97	0.92	0.96	0.92	0.96	1.00	1.00
CK+	12	0.88	0.94	0.87	0.88	0.87	0.88	0.98	0.98
Pain Archive	4	0.06	0.67	0.11	0.38	0.11	0.38	0.74	0.75
Pain Archive	6	0.41	0.70	0.35	0.41	0.35	0.41	0.77	0.77
Pain Archive	9	0.20	0.69	0.20	0.48	0.20	0.47	0.75	0.75
Pain Archive	12	0.32	0.66	0.23	0.34	0.23	0.33	0.72	0.73
RU-FACS	4	0.00	0.52	0.01	0.10	0.00	0.06	0.55	0.53
RU-FACS	6	0.49	0.85	0.48	0.72	0.48	0.72	0.90	0.90
RU-FACS	9	0.00	0.68	0.00	0.47	0.00	0.47	0.71	0.68
RU-FACS	12	0.68	0.84	0.64	0.70	0.64	0.70	0.91	0.91

Performance scores for the original and the Skew = 1 normalized version of the three datasets (for more AUs, see text).