Published in final edited form as:
*Int Conf Affect Comput Intell Interact Workshops*. 2013 ; 2013: 245–251. doi:10.1109/ACII.2013.47.

# Facing Imbalanced Data Recommendations for the Use of Performance Metrics

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

[1]Carnegie Mellon University, Pittsburgh, PA

[2]University of Pittsburgh, Pittsburgh, PA, jeffcohn@cs.cmu.edu

## Abstract

Recognizing facial action units (AUs) is important for situation analysis and automated video annotation. Previous work has emphasized face tracking and registration and the choice of features classifiers. Relatively neglected is the effect of imbalanced data for action unit detection. While the machine learning community has become aware of the problem of skewed data for training classifiers, little attention has been paid to how skew may bias performance metrics. To address this question, we conducted experiments using both simulated classifiers and three major databases that differ in size, type of FACS coding, and degree of skew. We evaluated influence of skew on both threshold metrics (Accuracy, F-score, Cohen's kappa, and Krippendorf's alpha) and rank metrics (area under the receiver operating characteristic (ROC) curve and precision-recall curve). With exception of area under the ROC curve, all were attenuated by skewed distributions, in many cases, dramatically so. While ROC was unaffected by skew, precision-recall curves suggest that ROC may mask poor performance. Our findings suggest that skew is a critical factor in evaluating performance metrics. To avoid or minimize skew-biased estimates of performance, we recommend reporting skew-normalized scores along with the obtained ones.

## I. INTRODUCTION

Our everyday communication is highly influenced by the emotional information available to us from other people. Recognizing facial expression is important for situation analysis and automated video annotation.

In the last decade many approaches have been proposed for automatic facial expression recognition [7], [29]. Although, previous work has emphasized face tracking and registration and the choice of feature classifiers, relatively neglected is the effect of imbalanced data when evaluating action unit detection.

In the case of facial expression data, the samples can be annotated using either emotion-specified labels (e.g., happy or sad) or action units, as defined by the Facial Action Coding System (FACS) [10]. Action units are anatomically defined facial actions that singly or in

laszlo.jeni@ieee.org, ftorre@cs.cmu.edu

[1]Code to compute skew-normalized scores for all of the metrics considered above and visualizations is available from http://www.pitt.edu/~jeffcohn/skew/

combinations can describe nearly all possible facial expressions or movements. Action unit (AU) detection, as well as expression detection of which AU detection is a subset, is a typical binary classification problem where the vast majority of examples are from one class, but the practitioner is typically interested in the minority (positive) class.

The problem of learning from imbalanced data sets is twofold. First of all, from the perspective of classifier training, imbalance in training data distribution often causes learning algorithms to perform poorly on the minority class. This issue has been well addressed in the machine learning literature [4], [15], [27], [26], [8] A common solution is to sample the data prior to training to re-balance the class distribution [2], [27]. An alternative to sampling is to use cost-sensitive learning. This approach targets the problem of skew by applying different cost matrices that describe the costs for misclassifying any particular data point [26], [8]. For a more detailed survey on the problem see [16] and the references therein.

Relatively little attention has been paid to how skew may spoil performance metrics. Facial expression data is typically highly skewed. Imbalance in the test data distribution might produce misleading conclusions with certain metrics. Percentage agreement, referred to as accuracy, is especially vulnerable to bias from skew. When base rate is low, high accuracy can result even when alternative methods rarely if ever agree [12], [14]. Agreement in that case is about the very large number of negative cases rather than the very few positive ones. Alternative metrics have been proposed to address this issue [24], [15]. Ferri et al. studied the relationship between different performance metrics and address the problem of rank correlations between them [12].

How does skewed data influence performance metrics for action unit detection? To address this question, we conducted experiments using both simulated classifiers and three major databases that include both posed and spontaneous facial expression and differ in database size, type of FACS coding [9], [10], and degree of skew. The databases were Cohn-Kanade [21], RU-FACS [13], and UNBC-McMaster Pain Archive [22].

We included a broad range of metrics that included both threshold metrics (Accuracy, $F_1$-score, Cohen's kappa, and Krippendorf's alpha) and rank metrics (area under the ROC curve [11] and precision-recall curve). With exception of area under the ROC curve, all were attenuated by skewed distributions; in many cases, dramatically so. Alpha and kappa were affected by skew in either direction; whereas $F_1$-score was affected by skew only in one direction. While ROC was unaffected by skew, precision-recall curves can reveal differences between classifiers, because of the different visual representation of the curves. Very different precision-recall can be associated with same ROC.

Our findings suggest that skew is a critical factor in evaluating performance metrics. Metrics of classifier performance may reveal more about skew than they do about actual performance. Databases that are otherwise identical with respect to intensity of action units, head pose, and so on may give rise to very different metric values depending only on differences in skew. This finding has implications for testing classifiers so as to avoid or minimize confounds and for meta-analyses of classifier performance. Sensitivity of the

threshold metrics for skewed distributions could be reduced by balancing the distribution of datasets.

The paper is built as follows. Datasets and their properties are reviewed in Section 2. Theoretical components are described in Section 3. Experimental results on the effect of imbalanced data on performance metrics and AU classification are detailed in Section 4. Discussion and a summary conclude the paper (Section 5).

## II. DATASETS

First, we describe the datasets (Section II.A-C). We then report findings with respect to skew for each AU (Section II.D).

In our simulations we used three major databases that include both posed and spontaneous facial expression and differ in database size, type of FACS coding, and degree of skew. The databases were Cohn-Kanade, RU-FACS and UNBC-McMaster Pain Archive.

### A. Cohn-Kanade Extended

The Cohn-Kanade Extended Facial Expression (CK+) Database [21] is an extension of the original Cohn-Kanade Database [18]. Cohn-Kanade has been widely used to compare the performance of different methods of automated facial expression analysis. CK+ includes 593 frontal image sequences of directed facial action tasks (i.e., posed AU and AU combinations) performed by 123 different participants. Facial landmarks (68-point mesh) were tracked using person-specific active appearance models [28]. Twenty-seven action units were manually coded for presence or absence by certified FACS coders. For a subset of 118 sequences, the seven universal emotion expressions (anger, contempt, disgust, fear, happy, sad and surprise) plus neutral were labeled. We used all 593 sequences for the current study.

### B. McMaster Pain Archive

The Pain Archive [22] consists of facial expressions of 129 participants who were suffering from shoulder pain. The participants performed different active and passive motion tests with their affected and unaffected limbs on two separate occasions. The distribution has 200 video sequences with 48398 frames from 25 participants. All of the frames were FACS coded for 12 AU by certified FACS coders and have frame level pain scores, sequence-level self-report, and observer measures. Facial landmarks (66-point mesh) were tracked using person-specific active appearance models [28].

### C. RU-FACS Database

The version of RU-FACS available to us consisted of unscripted (i.e., spontaneous) facial behavior from 34 participants. Participants had been randomly assigned to either lie or tell the truth about an issue for which they had strong feelings. The scenario involves natural interaction with another person. AUs were manually coded for each video frame. Video from five participants had to be excluded due to excessive noise in the digitized video. Thus, video from 29 participants was used. Facial landmarks were tracked using a 68-point mesh using same AAM implementation [3].

### D. Imbalance in the 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. Unless databases have been contrived to minimize skew, skew is quite common. Most facial actions have relatively low rates of occurrence. Smile controls, actions that counteract the upward pull of a smile (e.g., AU 14 or AU 15), occur less than 3% of the time even in a highly social context [25]. Thus, for action unit detection, the number of positive training examples will often be small, which can result in large imbalance between the positive and negative examples. While skew in training sets can be adjusted by under-sampling negative cases, skew in test sets remains.

The imbalance of this type of data can be defined by the skew ratio between the classes:

$$Skew = \frac{negative \quad examples}{positive \quad examples}, \quad (1)$$

Table I shows the skew ratios of action units from the three datasets. In the small, posed CK + dataset, the average skew ratio is around 30. In the case of larger, spontaneous datasets the skew ratio is even more extreme: about 60 in the Pain Archive and over 80 in RU-FACS.

## III. METHODS

We tuned high precision shape-based AU classifiers in each dataset. Details of the methods are presented in Section III.A.

To evaluate the effect of skew on the classifiers, we used a broad range of both threshold and rank metrics. These are described in Section III.B.

In Section III.C we describe random sampling methods to balance the distribution of the testing partition of the datasets.

### A. Training AU Classifiers

Our method contains two main steps. First, we estimate 3D landmark positions on face images using a 2D/3D AAM method [23]. We describe the details of this technique in Section III-A1. Second, we remove the rigid transformation from the acquired 3D shape and perform an SVM-based binary classification on it using the different AUs as the class labels. We show this method in Section III-A2 and III-A3.

**1) Active Appearance Models—**As noted above, each of the datasets had been tracked using person-specific AAM. AAMs are generative parametric models for face alignment. A 3D shape model is defined by a 3D mesh and in particular the 3D vertex locations of the mesh, called landmark points. Consider the 3D shape as the coordinates of 3D vertices that make up the mesh:

$$\mathbf{x} = (x_1, y_1, z_1, \ldots, x_M, y_M, z_M)^T, \quad (2)$$

or, $x = (x_1, ..., x_M)^T$, where $x_i = (x_i, y_i, z_i)^T$. We have $T$ samples: $\{\mathbf{x}(t)\}_{t=1}^{T}$. We assume that – apart from scale, rotation, and translation – all samples $\{\mathbf{x}(t)\}_{t=1}^{T}$ can be approximated by means of the linear principal component analysis (PCA).

The interested reader is referred to [23] for the details of the 2D/3D AAM algorithm.

**2) Extracted Features—**To register face images, 3D structure from motion first was estimated using the method of Xiao et al. [28]. We then extracted the normalized 3D shape parameters by removing the rigid transformation. Next, we performed a personal mean shape normalization [17]. We calculated an average shape for each subject (the so called personal mean shape) and computed the differences between the features of the actual shape and the features of the personal mean shape. This step removes within-person variation.

**3) Support Vector Machine for AU Detection—**After extracting the normalized 3D shape, we performed an SVM-based binary-class classification using each AU in turn as the positive class labels. Negative labels were all other AU.

Support Vector Machines (SVMs) are powerful for binary and multi-class classification as well as for regression problems. They are robust against outliers [1]. For two-class separation, SVM estimates the optimal separating hyper-plane between the two classes by maximizing the margin between the hyper-plane and closest points of the classes. The closest points of the classes are called support vectors. They determine the optimal separating hyper-plane, which lies at half distance between them.

We are given sample and label pairs $(x^{(k)}, y^{(k)})$ with $\mathbf{x}^{(k)} \in \mathbb{R}^m$, $y^{(k)} \in \{-1, 1\}$, and $k = 1, ..., K$. Here, for class 1 (class 2) $y^{(k)} = 1$ ($y^{(k)} = -1$). Assume further that we have a set of feature vectors $\phi\,(=[\phi_1; \ldots; \phi_M]) : \mathbb{R}^m \rightarrow \mathbb{R}^M$, where $M$ might be infinite. The support vector classification seeks to minimize the cost function

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{K}\xi_i \quad (3)$$

$$y^{(k)}\left(\mathbf{w}^T\phi\left(\mathbf{x}^{(k)}\right) + b\right) \geq 1 - \xi_i, \xi_i, \geq 0. \quad (4)$$

We used binary-class classification for each AU, where the positive class contains all shapes labelled by the given AU, and the negative class contains every other shapes. In all cases, we used only linear classifiers and also varied the regularization parameter by factors of ten from $10^{-4}$ to $10^{2}$.

## B. 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 2 × 2 confusion matrix. The matrix has four categories: True positives (TP) are examples correctly labeled as positives. False positives

(FP) refer to negative examples incorrectly labeled as positive. True negatives (TN) correspond to negatives correctly labeled as negative and false negatives (FN) refer to positive examples incorrectly labeled as negative. Using these categories we can derive two performance metrics: the precision $\left(P=\frac{TP}{TP+FP}\right)$ and the recall $\left(R=\frac{TP}{TP+FN}\right)$ values of the classifier. Precision is the fraction of recognized instances that are relevant, while recall is the fraction of relevant instances that are retrieved.

For the comparison we used both threshold metrics (Accuracy, $F_1$-score, Cohen's kappa, and Krippendorf's alpha) and rank metrics (area under the ROC curve and precision-recall curve).

**1) Threshold Metrics—**The threshold metrics used in this paper are Accuracy, $F_1$-score, Cohen's kappa, and Krippendorf's alpha. These metrics have a threshold level, where examples above the threshold are predicted as positive and the rest as negative. For these metrics, it is not important how close a prediction is to the level, only if it is above or below threshold.

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

$$Acc=\frac{TP+TN}{TP+FP+TN+FN}. \quad (5)$$

Accuracy is a widely used metric for measuring the performance of a classifier, however, when the prior probabilities of the classes are very different, this metric can be misleading.

A better choice is $F_1$-score, which can be interpreted as a weighted average of the precision and recall values:

$$F_1=2\cdot\frac{P\cdot R}{P+R} \quad (6)$$

Cohen's kappa is a coefficient developed to measure agreement among observers [6]. It shows the observed agreement normalized to the agreement by chance:

$$K=\frac{P_{Obs}-P_{Chance}}{1-P_{Chance}}. \quad (7)$$

Krippendorff's $\alpha$-reliability measures the observed disagreement normalized to the observed disagreement [19], [20]:

$$\alpha=1-\frac{D_{Obs}}{D_{Chance}}. \quad (8)$$

**2) Rank Metrics—**The rank metrics depend only on the ordering of the cases, not the actual predicted values. As long as ordering is preserved, it makes no difference whether predicted values fall between different intervals. These metrics measure how well the positive cases are ordered before negative cases and can be viewed as a summary of model performance across all possible thresholds. The rank metrics we use are area under the ROC curve (AUC-ROC) and area under Precision-Recall curve (AUC-PR).

The ROC curve depicts the true positive rate as the function of the false positive rate, while the Precision-Recall curve shows the precision as the function of recall. Recall is the same as TPR, whereas Precision measures that fraction of examples classified as positive that are truly positive.

### C. Skew Normalization using Random Sampling

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.

Random under-sampling tries to balance the class distribution through the random elimination of majority class examples. The major drawback of random under-sampling is that this method can discard examples that could be important for the performance metric.

In this paper we used random under-sampling with averaging: first, we under-sample the majority class, then calculate the performance metrics. We repeat the process in the function of the skew present in the data.

## IV. EXPERIMENTS

We executed a number of evaluations to judge the influence of the skewed distributions on the performance metrics. Studies concern (i) simulated classifiers with given relative misclassification rates, (ii) the effect of the skewed distributions on performance scores using different databases for AU classification.

### A. Experiments on Simulated Classifiers

In this experiment we simulated binary classifiers with different properties to understand the effect of the skew on the performance metrics better. The classifiers were different in the relative misclassification rate: a fixed percentage of the positive (and negative) examples were misclassified in proportion to the number of positive (and negative) examples. For example, in the ”5% case” 5% of the positive examples were labelled as false negatives (FN), and 5% of the negative examples were labelled as false positives (FP).

In the case of the threshold metrics, the score was calculated from confusion matrices, while the rank metrics were calculated by drawing random samples from Gaussian distribution representing the decision values of the classifiers.

Fig. 1 depicts the different metric scores in the function of the skew ratio. Skew = 1 represents a fully balanced dataset, *Skew* > 1 shows that the negative samples are the majority, and the *Skew* < 1 values represent positive sample dominance in the distribution.

With the exception of area under the ROC curve, all metrics are attenuated by skewed distributions. Alpha and kappa are affected by skew in either direction; whereas $F_1$- score is affected by skew in one direction only. Random performance in the alpha and kappa spaces is equivalent with the 0 value, but in the $F_1$-space it changes as a function of skew: in the balanced case (*Skew* = 1) is associated with 0.5 score and drops exponentially as skew increases.

It is important to note, that even the best (1% error rate) classifier's performance drops significantly in the high skew ratio part of the graph (*Skew* = 50). This imbalance range is equal or even below the skew ratio present in spontaneous facial behaviour datasets (see Table I).

### B. Experiments on Real data

In this experiment we studied the effect of skewed AU distributions on the CK+, McMaster Pain Archive and RUFACS.

In the case of CK+ dataset we used leave-one-subject-out cross-validation to maximize the data available in the database. In the RU-FACS and Pain dataset for each AU we divided the data into a training and testing set in a way that the skew ratio of the two sets was similar.

We calculated $F_1$ score, kappa, alpha measures and area under ROC and PR curves. Tables II - III show these measures in the columns labelled 'original'.

To proceed, we repeated the same procedure, but this time we balanced the distribution of the classes in the testing set using random under-sampling and averaging. The performance scores are depicted in the 'normalized' columns of Tables II - III. From the results, we can draw several observations as follows.

First of all, by examining the scores in the imbalanced case of the CK+ dataset, we found that these performances are similar to other shape based methods in the literature [5], [17].

Second, by comparing the skew normalized results to the imbalanced ones, we noticed that (except the area under ROC curve) all scores improved. The average $F_1$ score increased from 0.45 to 0.77 in the case of CK+, from 0.23 to 0.68 in the case of RU-FACS and from 0.17 to 0.65 on the Pain data. The difference between the scores is the smallest in the case of the CK+ data, because this is the smallest dataset with the smallest skew ratio (around 20) among the three. The improvement is smaller in kappa and alpha: these measures are somewhat more strict and a bit tolerant to the prior distributions of the classes. The differences in the case of the area under PR curve are comparable to the $F_1$ score improvements.

Third, while ROC was unaffected by skew, the precision-recall curves suggest that ROC may mask poor performance in some cases.

## V. DISCUSSION AND SUMMARY

In the present work, we addressed the question how do imbalanced datasets influence performance metrics. We conducted studies using three major databases that include both posed and spontaneous facial expression and differ in database size, type of FACS coding, and degree of imbalance. The databases were Cohn-Kanade, RU-FACS, and McMaster Pain Archive. We included metrics used in facial behaviour analysis plus some others: we included both threshold metrics (Accuracy, $F_1$-score, Cohen's kappa, and Krippendorf's alpha) and rank metrics (area under the ROC curve and precision-recall curve).

We used a variety of evaluations to study the influence of imbalanced distribution on performance metrics. We used simulated classifiers and binary SVMs trained on expert annotated datasets as well.

We discovered that with exception of area under the ROC curve, all performance metrics were attenuated by imbalanced distributions; in many cases, dramatically so. Alpha and kappa measures were affected by skew in either direction; whereas $F_1$-score was affected by skew only in one direction. While ROC was unaffected by skew, precision-recall curves suggest that ROC may mask poor performance.

Metrics of classifier performance may reveal more about skew than they do about actual performance. Databases that are otherwise identical with respect to intensity of action units, head pose, and so on may give rise to very different metric values depending only on differences in skew. To avoid or minimize biased estimates of performance metrics, we recommend that investigators report both obtained performance metrics and skew-normalized scores. Alternatively, report both the obtained scores and the degree of skew in databases[1]. In these ways, classifiers can be compared across databases free of confounds introduced by skew.

## ACKNOWLEDGMENTS

Research reported in this publication was supported in part by the National Institute of Mental Health of the National Institutes of Health under Award Number MH096951. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

1. Abe, S. Support vector machines for pattern classification. Springer; 2010. 3
2. Akbani, R.; Kwek, S.; Japkowicz, N. Machine Learning: ECML 2004. Springer; Berlin Heidelberg: 2004. Applying support vector machines to imbalanced datasets.; p. 39-50.1
3. Ashraf AB, Lucey S, Cohn JF, Chen T, Ambadar Z, Prkachin KM, Solomon PE. The painful face Pain expression recognition using active appearance models. Image and Vision Computing. 2009; 27(12):1788–1796. 2. [PubMed: 22837587]
4. Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor. Newsl. Jun; 2004 6(1):1–6. (2004) 1.
5. Chew, SW.; Lucey, PJ.; Lucey, S.; Saragih, J.; Sridharan, J. F. Cohn S. Person-independent facial expression detection using constrained local models.. Proceedings of FG 2011 Facial Expression Recognition and Analysis Challenge; Santa Barbara, CA. 2011; 5
6. Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960; 20(1):37–46. 4.

7. Cohn, JF.; De la Torre, F. Automated face analysis for affective computing.. In: Calvo, RA.; D'Mello, SK.; Gratch, J.; Kappas, A., editors. Handbook of Affective Computing. New York, NY: Oxford: In press. 1
8. Eitrich T, Lang B. Efficient optimization of support vector machine learning parameters for unbalanced datasets. Journal of computational and applied mathematics. 2006; 196(2):425–436. 1.
9. Ekman P, Friesen W. Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto. 1978 1.
10. Ekman, P.; Friesen, W.; Hager, J. Network Research Information. Salt Lake City, UT: 2002. Facial action coding system: Research nexus.. 1
11. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006. Elsevier Science Inc. 2006:861–874. 1.
12. Ferri C, Hernndez-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recognition Letters. 2009; 30(1):27–38. 1.
13. Frank, M.; Movellan, J.; Bartlett, M.; Littleworth, G. RU-FACS-1 database. Machine Perception Laboratory; U.C. San Diego: 1
14. Garcia, V.; Mollineda, RA.; Sanchez, JS. Pattern Recognition and Image Analysis. Springer; Berlin Heidelberg: 2009. Index of balanced accuracy: A performance measure for skewed class distributions.; p. 441-448.1
15. Garcia V, Mollineda RA, Sanchez JS. Theoretical analysis of a performance measure for imbalanced data. Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE. 2010 1.
16. He H, Garcia EA. Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on. 2009; 21(9):1263–1284. 1.
17. Jeni LA, Lorincz A, Nagy T, Palotai Zs. Sebok J, Szabo Z, Takacs D. 3D shape estimation in video sequences provides high precision evaluation of facial expressions. Image and Vision Computing. Feb 21; 2012 30(10):785–795. 3, 5.
18. Kanade, T.; Cohn, JF.; Tian, Y. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00). Grenoble, France: 2000. Comprehensive database for facial expression analysis.; p. 46-53.2
19. Krippendorff K. Estimating the reliability, systematic error and random error of interval data. Educational and Psychological Measurement. 1970; 30:61–70. 4.
20. Krippendorff, K. Content analysis: An introduction to its methodology. 2nd ed.. Sage; Thousand Oaks, CA: 2004. 4
21. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion specified expression. 3rd IEEE Workshop on CVPR for Human Communicative Behavior Analysis. 2010; 1:2.
22. Lucey P, Cohn JF, Prkachin KM, Solomon PE, Matthews I. Painful data: The UNBC-McMaster shoulder pain expression archive database. IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011). Mar.2011 :57–64. 21–25. 1, 2.
23. Matthews I, Baker S. Active appearance models revisited. Int. J. Comp. Vision. 2004; 60(2):135–164. 3.
24. Ranawana, R.; Palade, V. CEC 2006. IEEE Congress on. IEEE; 2006. Optimized Precision-A new measure for classifier performance evaluation. Evolutionary Computation, 2006.. 1
25. Sayette MA, Creswell KG, Dimoff JD, Fairbairn CE, Cohn JF, Heckman BW, Kirchner TR, Levine JM, Moreland RL. Alcohol and group formation: A multimodal investigation of the effects of alcohol on emotion and social bonding. Psychological Science. 2012; 23(8):869–878. 3. [PubMed: 22760882]
26. Tang Y, Zhang Y-Q, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on. 2009; 39(1):281–288. 1.
27. Van Hulse, J.; Khoshgoftaar, TM.; Napolitano, A. Proceedings of the 24th international conference on Machine learning. ACM; 2007. Experimental perspectives on learning from imbalanced data.. 1

28. Xiao, J.; Baker, S.; Matthews, I.; Kanade, T. Real-time combined 2D+3D active appearance models.. Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition; Washington, D.C., USA. 2004. p. 535-542.2, 3

29. Zeng Z, Pantic M, Roisman GI, Huang TS. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009; 31(1):39–58. 1. [PubMed: 19029545]

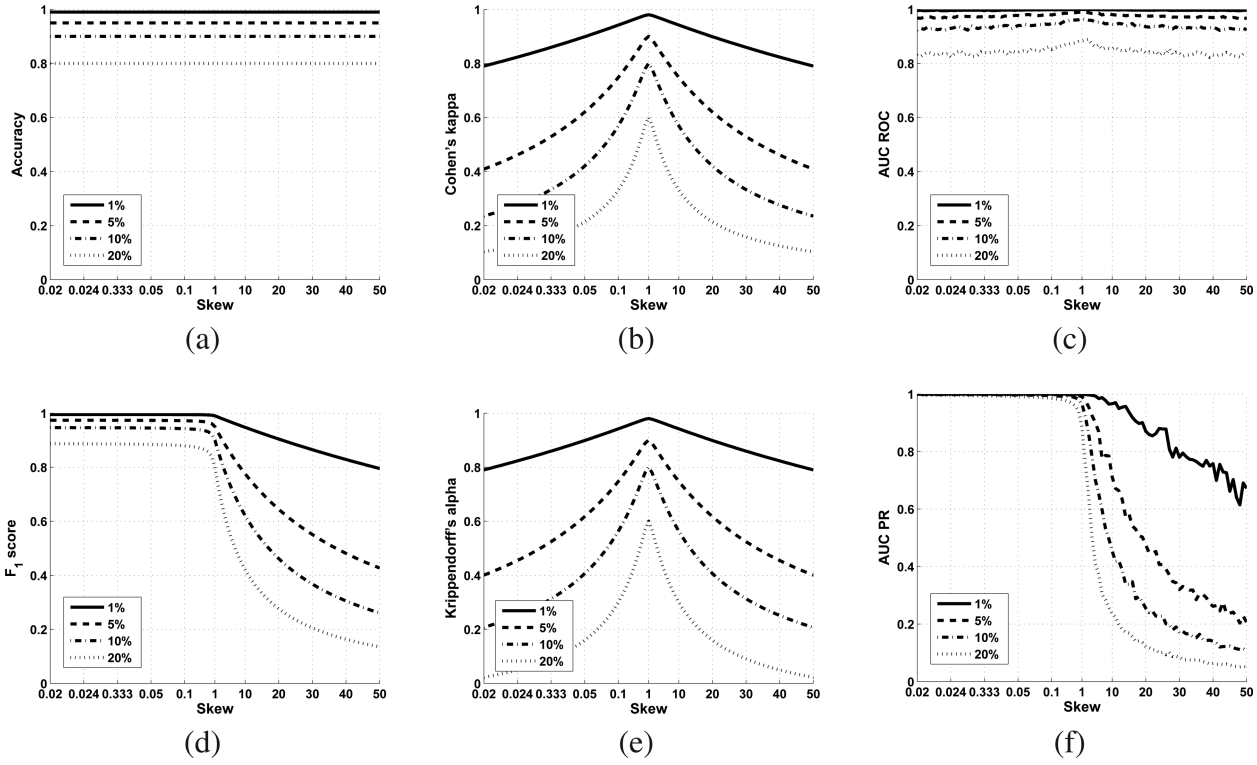


**Figure 1.**

The behaviour of different metrics using simulated classifiers. The horizontal axis depicts the skew ratio $\left(Skew=\frac{Negative\quad examples}{Positive\quad examples}\right)$, while the vertical axis shows the given metric score. The metrics are (a): Accuracy, (b): Cohen's kappa, (c) Area Under ROC, (d) $F_1$ score, (e) Krippendorff's alpha, (f) Area Under PR Curve. The different lines show the relative misclassification rates of the simulated classifiers.

**Table I**

Database statistics. For more details, see text.

| AU | CK+ | | PAIN | | RU-FACS | |
|---|---|---|---|---|---|---|
| | # of AUs | Skew | # of AUs | Skew | # of AUs | Skew |
| 1 | 161 | 11.48 | - | | 7616 | 14.45 |
| 2 | 109 | 17.43 | - | | 6112 | 18.25 |
| 4 | 174 | 10.55 | 987 | 23.31 | 1028 | 113.44 |
| 5 | 95 | 20.15 | - | | 714 | 163.77 |
| 6 | 113 | 16.78 | 5132 | 3.68 | 5184 | 21.69 |
| 7 | 108 | 17.60 | 3012 | 6.97 | 2332 | 49.45 |
| 9 | 65 | 29.91 | 422 | 55.86 | 55 | 2138.05 |
| 10 | 19 | 104.74 | 515 | 45.59 | 3215 | 35.59 |
| 11 | 32 | 61.78 | - | | - | |
| 12 | 125 | 15.07 | 6627 | 2.62 | 18416 | 5.39 |
| 14 | 37 | 53.30 | - | | 7599 | 14.48 |
| 15 | 83 | 23.20 | 6 | 3998.33 | 3676 | 31.00 |
| 16 | 21 | 94.67 | - | | - | |
| 17 | 182 | 10.04 | - | | 8028 | 13.65 |
| 18 | 9 | 222.22 | - | | 1705 | 68.00 |
| 20 | 71 | 27.30 | 657 | 35.52 | 276 | 425.26 |
| 22 | 3 | 668.67 | - | | 155 | 758.02 |
| 23 | 55 | 35.53 | - | | 1796 | 64.51 |
| 24 | 49 | 40.00 | - | | 3906 | 29.12 |
| 25 | 300 | 5.70 | 2407 | 8.97 | 26047 | 3.52 |
| 26 | 50 | 39.18 | 2093 | 10.46 | 17591 | 5.69 |
| 27 | 74 | 26.15 | 18 | 1332.11 | - | |
| 28 | 1 | 2008.00 | - | | 1431 | 81.21 |
| 30 | 1 | 2008.00 | - | | 766 | 152.59 |
| 31 | 2 | 1003.50 | - | | 117648 | 200.34 |
| 43 | 7 | 286.00 | 2120 | 10.32 | - | |
| 45 | 14 | 142.50 | - | | - | |

**Table II**

Performance scores for the original and the *Skew* = 1 normalized version of UNBC-McMaster Pain Archive and RU-FACS.

| UNBC-McMaster Pain Arehiue | | | | | | | | | | | | RU-FACS | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AU** | **Fl** | | **Kappa** | | **Alpha** | | **AUC** | | **AUCPR** | | **AU** | **Fl** | | **Kappa** | | **Alpha** | | **AUC** | | **AUCPR** | |
| | **Orig.** | **Norm.** | **Orig.** | **Norm.** | **Orig.** | **Norm.** | **Orig.** | **Norm.** | **Orig.** | **Norm.** | | **Orig.** | **Norm.** | **Orig.** | **Norm.** | **Orig.** | **Norm.** | **Orig.** | **Norm.** | **Orig.** | **Norm.** |
| **1** | - | - | - | - | - | - | - | - | - | - | **1** | 0.39 | 0.71 | 0.38 | 0.46 | 0.37 | 0.46 | 0.75 | 0.75 | 0.36 | 0.80 |
| **2** | - | - | - | - | - | - | - | - | - | - | **2** | 0.31 | 0.65 | 0.31 | 0.34 | 0.30 | 0.33 | 0.69 | 0.69 | 0.29 | 0.75 |
| **4** | 0.06 | 0.67 | 0.11 | 0.38 | 0.11 | 0.38 | 0.74 | 0.75 | 0.05 | 0.73 | **4** | 0.00 | 0.52 | 0.01 | 0.10 | 0.00 | 0.06 | 0.55 | 0.53 | 0.01 | 0.50 |
| **5** | - | - | - | - | - | - | - | - | - | - | **5** | 0.03 | 0.71 | 0.02 | 0.44 | 0.02 | 0.43 | 0.74 | 0.75 | 0.01 | 0.69 |
| **6** | 0.41 | 0.70 | 0.35 | 0.41 | 0.35 | 0.41 | 0.77 | 0.77 | 0.32 | 0.77 | **6** | 0.49 | 0.85 | 0.48 | 0.72 | 0.48 | 0.72 | 0.90 | 0.90 | 0.47 | 0.91 |
| **7** | 0.24 | 0.66 | 0.23 | 0.33 | 0.23 | 0.33 | 0.68 | 0.68 | 0.17 | 0.71 | **7** | 0.09 | 0.50 | 0.08 | 0.18 | 0.07 | 0.11 | 0.51 | 0.51 | 0.03 | 0.59 |
| **9** | 0.20 | 0.69 | 0.20 | 0.48 | 0.20 | 0.47 | 0.75 | 0.75 | 0.13 | 0.79 | **9** | 0.00 | 0.68 | 0.00 | 0.47 | 0.00 | 0.47 | 0.71 | 0.68 | 0.00 | 0.57 |
| **10** | 0.03 | 0.79 | 0.07 | 0.63 | 0.05 | 0.63 | 0.84 | 0.81 | 0.04 | 0.74 | **10** | 0.11 | 0.69 | 0.16 | 0.39 | 0.16 | 0.39 | 0.77 | 0.76 | 0.08 | 0.74 |
| **12** | 0.32 | 0.66 | 0.23 | 0.34 | 0.23 | 0.33 | 0.72 | 0.73 | 0.27 | 0.72 | **12** | 0.68 | 0.84 | 0.64 | 0.70 | 0.64 | 0.70 | 0.91 | 0.91 | 0.76 | 0.92 |
| **14** | - | - | - | - | - | - | - | - | - | - | **14** | 0.11 | 0.52 | 0.07 | 0.10 | 0.06 | 0.07 | 0.54 | 0.55 | 0.08 | 0.57 |
| **15** | - | - | - | - | - | - | - | - | - | - | **15** | 0.17 | 0.66 | 0.15 | 0.35 | 0.15 | 0.34 | 0.72 | 0.72 | 0.10 | 0.72 |
| **17** | - | - | - | - | - | - | - | - | - | - | **17** | 0.51 | 0.77 | 0.48 | 0.56 | 0.48 | 0.56 | 0.86 | 0.86 | 0.45 | 0.87 |
| **I8** | - | - | - | - | - | - | - | - | - | - | **I8** | 0.04 | 0.66 | 0.08 | 0.41 | 0.05 | 0.40 | 0.70 | 0.71 | 0.04 | 0.71 |
| **20** | 0.05 | 0.58 | 0.01 | 0.27 | −0.01 | 0.23 | 0.60 | 0.60 | 0.02 | 0.53 | **20** | 0.03 | 0.38 | 0.04 | 0.03 | 0.04 | −0.09 | 0.35 | 0.35 | 0.00 | 0.44 |
| **22** | - | - | - | - | - | - | - | - | - | - | **22** | 0.01 | 0.78 | 0.01 | 0.58 | 0.00 | 0.58 | 0.74 | 0.73 | 0.00 | 0.66 |
| **23** | - | - | - | - | - | - | - | - | - | - | **23** | 0.02 | 0.50 | 0.02 | 0.05 | 0.01 | 0.02 | 0.48 | 0.48 | 0.02 | 0.50 |
| **24** | - | - | - | - | - | - | - | - | - | - | **24** | 0.08 | 0.49 | 0.07 | 0.14 | 0.07 | 0.03 | 0.51 | 0.50 | 0.04 | 0.56 |
| **25** | 0.19 | 0.63 | 0.16 | 0.28 | 0.15 | 0.28 | 0.67 | 0.67 | 0.16 | 0.68 | **25** | 0.82 | 0.88 | 0.78 | 0.77 | 0.78 | 0.77 | 0.95 | 0.95 | 0.90 | 0.96 |
| **26** | 0.12 | 0.74 | 0.19 | 0.48 | 0.18 | 0.48 | 0.75 | 0.76 | 0.11 | 0.75 | **26** | 0.60 | 0.79 | 0.54 | 0.58 | 0.54 | 0.58 | 0.84 | 0.85 | 0.60 | 0.86 |
| **27** | 0.00 | 0.40 | 0.00 | 0.40 | 0.00 | 0.37 | 0.28 | 0.40 | 0.00 | 0.41 | **27** | - | - | - | - | - | - | - | - | - | - |
| **28** | - | - | - | - | - | - | - | - | - | - | **23** | 0.35 | 0.91 | 0.41 | 0.84 | 0.41 | 0.84 | 0.97 | 0.96 | 0.32 | 0.96 |
| **30** | - | - | - | - | - | - | - | - | - | - | **30** | 0.05 | 0.72 | 0.04 | 0.50 | 0.03 | 0.50 | 0.73 | 0.74 | 0.01 | 0.70 |
| **43** | 0.28 | 0.69 | 0.27 | 0.39 | 0.27 | 0.38 | 0.76 | 0.77 | 0.23 | 0.78 | **43** | - | - | - | - | - | - | - | - | - | - |

**Table III**

Performance scores on Cohn-Kanade Extended.

| AU | Cohn-Kanade Extended | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Fl | | Kappa | | Alpha | | AUC | | AUCPR | |
| | Orig. | Norm. | Orig. | Norm. | Orig. | Norm. | Orig. | Norm. | Orig. | Norm. |
| 1 | 0.85 | 0.92 | 0.85 | 0.85 | 0.85 | 0.85 | 0.98 | 0.98 | 0.93 | 0.98 |
| **2** | 0.87 | 0.92 | 0.87 | 0.86 | 0.87 | 0.86 | 0.37 | 0.96 | 0.92 | 0.97 |
| **4** | 0.73 | 0.83 | 0.68 | 0.69 | 0.68 | 0.68 | 0.90 | 0.90 | 0.71 | 0.90 |
| **5** | 0.76 | 0.90 | 0.76 | 0.82 | 0.76 | 0.82 | 0.93 | 0.94 | 0.76 | 0.95 |
| **6** | 0.74 | 0.90 | 0.72 | 0.32 | 0.72 | 0.82 | 0.96 | 0.98 | 0.77 | 0.97 |
| **7** | 0.67 | 0.89 | 0.65 | 0.80 | 0.65 | 0.80 | 0.32 | 0.94 | 0.63 | 0.94 |
| 9 | 0.92 | 0.97 | 0.92 | 0.96 | 0.92 | 0.96 | 1.00 | 1.00 | 0.95 | 0.99 |
| **10** | 0.22 | 0.50 | 0.20 | 0.35 | 0.19 | 0.31 | 0.51 | 0.53 | 0.05 | 0.58 |
| **11** | 0.10 | 0.74 | 0.32 | 0.58 | 0.31 | 0.58 | 0.81 | 0.33 | 0.14 | 0.84 |
| **12** | 0.88 | 0.94 | 0.87 | 0.88 | 0.87 | 0.88 | 0.98 | 0.98 | 0.93 | 0.98 |
| **13** | 0.04 | 1.00 | 0.04 | 1.00 | 0.02 | 1.00 | 0.32 | 1.00 | 0.01 | 0.50 |
| **14** | 0.10 | 0.57 | 0.06 | 0.20 | 0.06 | 0.20 | 0.51 | 0.55 | 0.03 | 0.57 |
| **15** | 0.65 | 0.80 | 0.65 | 0.67 | 0.65 | 0.67 | 0.89 | 0.88 | 0.67 | 0.89 |
| **16** | 0.17 | 0.67 | 0.21 | 0.46 | 0.20 | 0.46 | 0.67 | 0.66 | 0.10 | 0.72 |
| **17** | 0.74 | 0.82 | 0.72 | 0.69 | 0.72 | 0.69 | 0.90 | 0.90 | 0.79 | 0.91 |
| **18** | 0.06 | 0.44 | 0.04 | 0.11 | 0.03 | 0.09 | 0.75 | 0.49 | 0.02 | 0.48 |
| 20 | 0.77 | 0.96 | 0.78 | 0.93 | 0.78 | 0.93 | 0.98 | 0.99 | 0.81 | 0.98 |
| 21 | 0.07 | 0.67 | 0.06 | 0.67 | 0.04 | 0.69 | 0.33 | 0.39 | 0.02 | 0.57 |
| **22** | 0.10 | 0.75 | 0.10 | 0.75 | 0.09 | 0.76 | 0.89 | 0.88 | 0.03 | 0.66 |
| **23** | 0.75 | 0.81 | 0.76 | 0.75 | 0.76 | 0.75 | 0.95 | 0.93 | 0.76 | 0.92 |
| 24 | 0.63 | 0.82 | 0.62 | 0.70 | 0.62 | 0.70 | 0.31 | 0.88 | 0.60 | 0.38 |
| **25** | 0.92 | 0.95 | 0.90 | 0.91 | 0.90 | 0.91 | 0.93 | 0.99 | 0.97 | 0.98 |
| 26 | 0.04 | 0.60 | 0.06 | 0.26 | 0.04 | 0.26 | 0.60 | 0.63 | 0.06 | 0.58 |
| 27 | 0.90 | 0.99 | 0.92 | 0.99 | 0.92 | 0.99 | 1.00 | 1.00 | 0.95 | 0.98 |
| 31 | 0.10 | 0.67 | 0.10 | 0.67 | 0.09 | 0.69 | 0.91 | 0.78 | 0.02 | 0.43 |
| 38 | 0.07 | 0.62 | 0.16 | 0.45 | 0.16 | 0.43 | 0.72 | 0.73 | 0.07 | 0.69 |

| AU | Cohn-Kanade Extended | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | | Kappa | | Alpha | | AUC | | AUCPR | |
| | Orig. | Norm. | Orig. | Norm. | Orig. | Norm. | Orig. | Norm. | Orig. | Norm. |
| 39 | 0.16 | 0.64 | 0.14 | 0.57 | 0.14 | 0.57 | 0.65 | 0.69 | 0.05 | 0.69 |
| **43** | 0.06 | 0.43 | 0.05 | 0.14 | 0.05 | 0.10 | 0.37 | 0.33 | 0.01 | 0.37 |
| **45** | 0.06 | 0.50 | 0.08 | 0.25 | 0.08 | 0.18 | 0.53 | 0.55 | 0.02 | 0.56 |