Two tools for the performance analysis of multiclass classifiers

Francisco J. Valverde-Albacete

Depto. Lenguajes y Sistemas Informáticos
UNED, Spain

30/04/2013/ MAVIR

Setting the scene

Confusion matrix or contingency table of a classifier.

- $V_X = \{x_i\}_{i=1}^{n}$ and $V_Y = \{y_j\}_{j=1}^{p}$ be sets of input and output class *identifiers*.
- Basic event: "presenting a pattern of input class x_i to the classifier to obtain output class identifier y_j," $(X = x_i, Y = y_j)$.
- N iterated experiments to obtain a count matrix N_{XY} where

$$N_{XY}(x_i, y_j) = N_{ij}$$

counts the ocurrrences of the joint event.

A very old question...

What can be said about the performance of multi-class classifiers from their confusion matrices?

Some examples (from [1] and our own)

$$a = \begin{bmatrix} 15 & 0 & 5 \\ 0 & 15 & 5 \\ 0 & 0 & 20 \end{bmatrix} \quad b = \begin{bmatrix} 16 & 2 & 2 \\ 2 & 16 & 2 \\ 1 & 1 & 18 \end{bmatrix} \quad c = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 4 \\ 1 & 1 & 48 \end{bmatrix}$$

$$d = \begin{bmatrix} 15 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 27 \end{bmatrix} \quad e = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 57 \end{bmatrix} \quad f = \begin{bmatrix} 0 & 0 & 5 \\ 0 & 0 & 5 \\ 0 & 0 & 50 \end{bmatrix}$$

Figure : **Examples of synthetic confusion matrices with assorted behavior**: *a*, *b* and *c*, *d* a matrix whose marginals tend towards uniformity, *e* a matrix whose marginals tend to Kronecker's delta and *f* the confusion matrix of a majority classifier.

[1] Sindhwani, V., Rakshit, S., Deodhare, D., Erdogmus, D., Principe, J., Niyogi, P., 2004. Feature selection in MLPs and SVMs based on maximum output information. IEEE Transactions on Neural Networks 15 (4), 937—948

A plethora of measures of performance (I)[2]

Pair-counting measures (43+)

- Accuracy (but in previous example $A = \frac{50}{60}$ for a, b, c, f; $A = 1.0$ for d,e.)
- Issues
  - Choice of measure
  - Relationship between measures
  - Corrected for chance vs. uncorrected
  - Sometimes difficult Calculations, e.g., ROC and AUC

[2] Pfitzner, D., Leibbrandt, R., Powers, D., Jul. 2008. Characterization and evaluation of similarity measures for pairs of clusterings. Knowledge and Information Systems 19 (3), 361–394

A plethora of measures of performance (II)

Information-theoretic measures (13+)

- Transform counts into joint probability and estimate measures:

$$P_{XY}(x,y) \equiv P_{XY}^{\text{MLE}}(x,y) \approx \frac{N_{XY}(x,y)}{\sum_{x,y} N_{XY}(x,y)} \tag{1}$$

- Mutual information (a similarity)

$$MI_{P_{XY}} = \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$$

- Variation of Information [a] (a dissimilarity)

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} .$$

[a] Meila, M., 2007. Comparing clusterings—an information based distance. Journal of Multivariate Analysis 28, 875–893

Entropies related to P_{XY} [3]

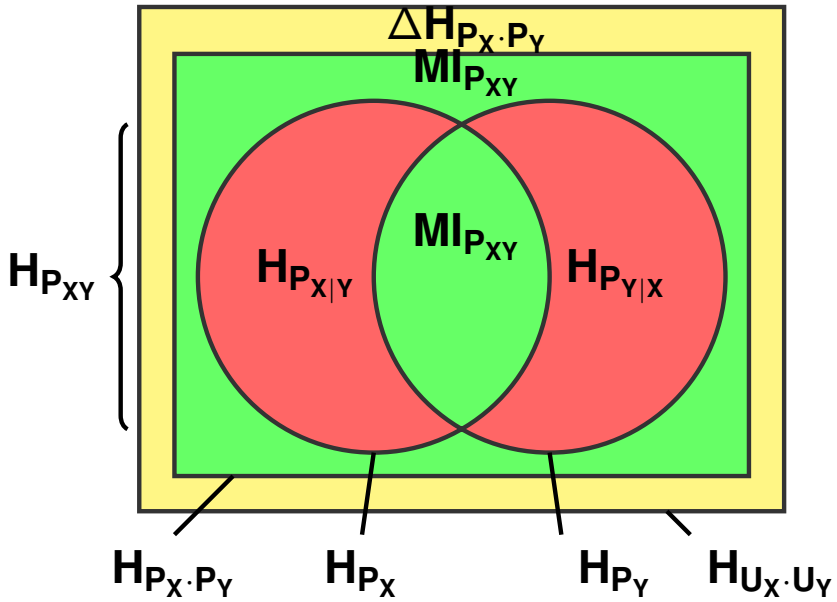


Figure : **Extended entropy diagram related to a bivariate distribution.**

The component equations

$$H_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} + MI_{P_{XY}} \quad (2)$$

$$H_{P_X \cdot P_Y} = MI_{P_{XY}} + H_{P_{XY}}$$

$$\Delta H_{P_X \cdot P_Y} = H_{U_X \cdot U_Y} - H_{P_X \cdot P_Y}$$

[3] Valverde-Albacete, F. J., Peláez-Moreno, C., 2010. Two information-theoretic tools to assess the performance of multi-class classifiers. Pattern Recognition Letters 31 (12), 1665–1671

The Balance equations

Adding the equations in (2) reads...

$$H_{U_{XY}} = \Delta H_{P_X \cdot P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}}$$
$$0 \leq \Delta H_{P_X \cdot P_Y}, 2MI_{P_{XY}}, VI_{P_{XY}} \leq H_{U_{XY}} \ .$$

By normalizing in $H_{U_{XY}} = H_{U_X} + H_{U_Y} = \log k + \log p$,

$$1 = \Delta H'_{P_X \cdot P_Y} + 2MI'_{P_{XY}} + VI'_{P_{XY}}$$
$$0 \leq \Delta H'_{P_X \cdot P_Y}, 2MI'_{P_{XY}}, VI'_{P_{XY}} \leq 1 \ .$$

This is the 2-simplex in normalized space!

$$F_{XY}(P_{XY}) = [\Delta H'_{P_X \cdot P_Y}, 2MI'_{P_{XY}}, VI'_{P_{XY}}]$$

From the 2-simplex to the De Finetti entropy diagrams

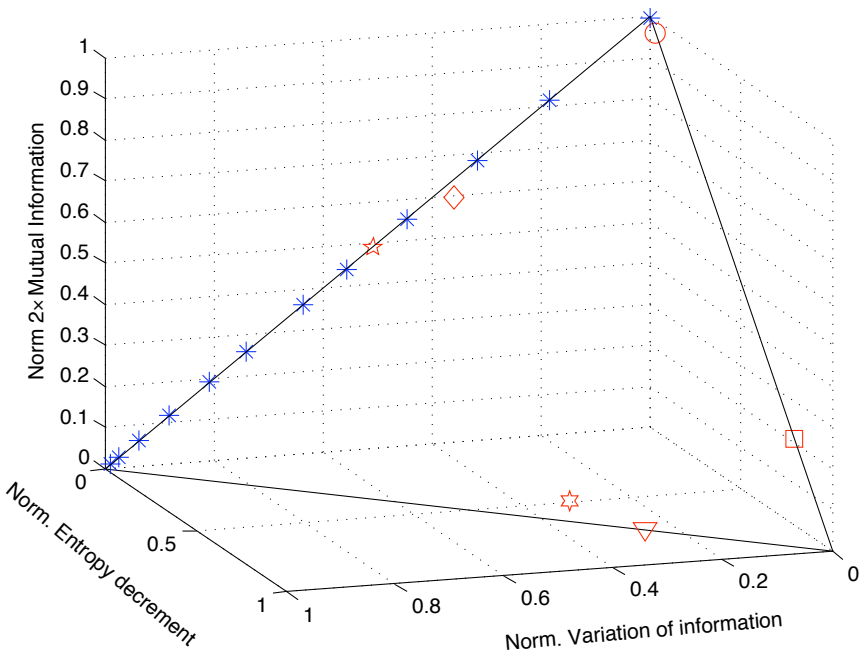


Figure : The 2-simplex in three-dimensional, normalized entropy space $[\Delta H'_{P_X \cdot P_Y}, VI'_{P_{XY}}, 2MI'_{P_{XY}}]$

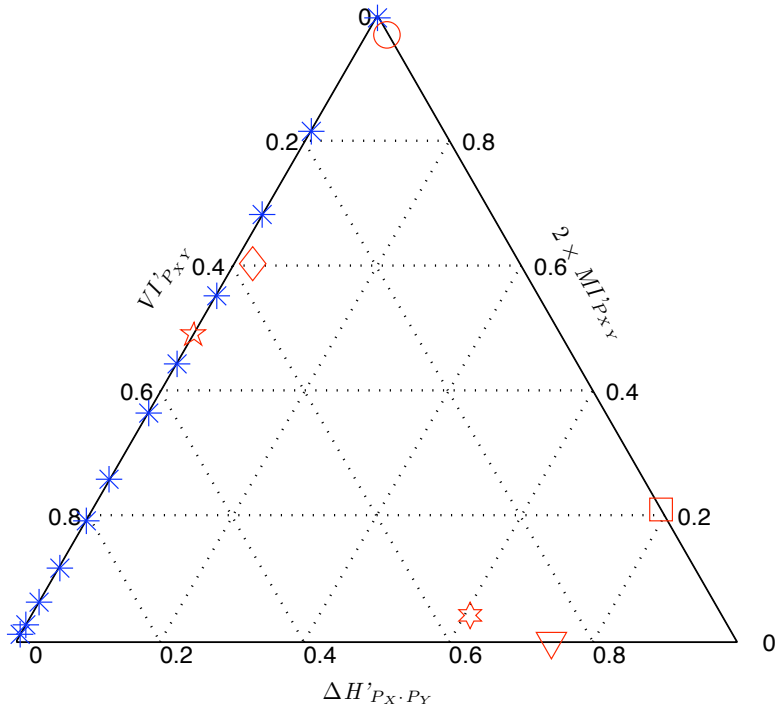


Figure : The de Finetti entropy diagram or entropy triangle, a projection of the 2-simplex onto a two-dimensional space. Example with synthetic data in previous slide.

The interpretation of Entropy Triangles

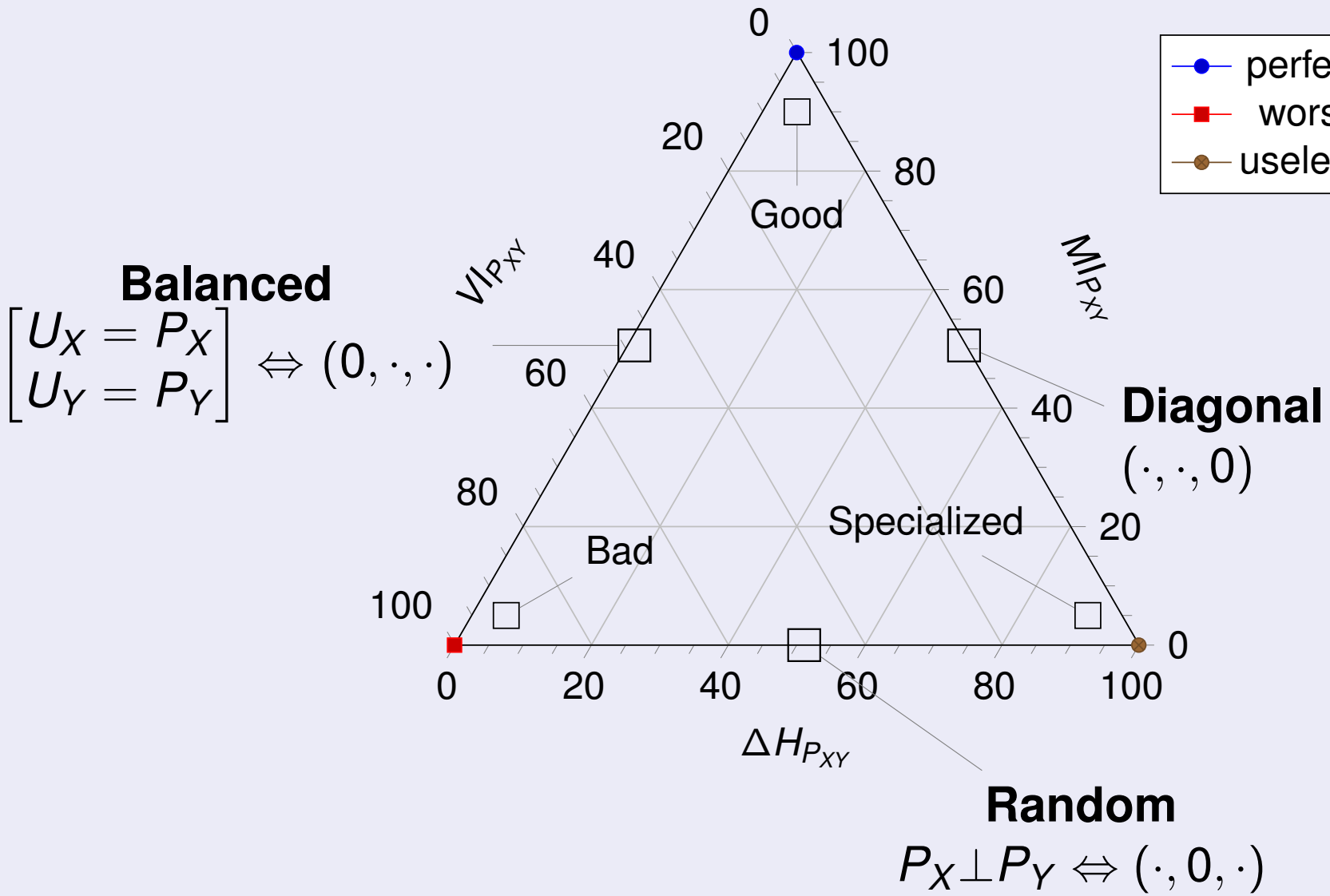


Figure : Schematics on how to interpret the zones in the entropy triangle.

Example 2: Human vs. Machine performance in ASR tasks

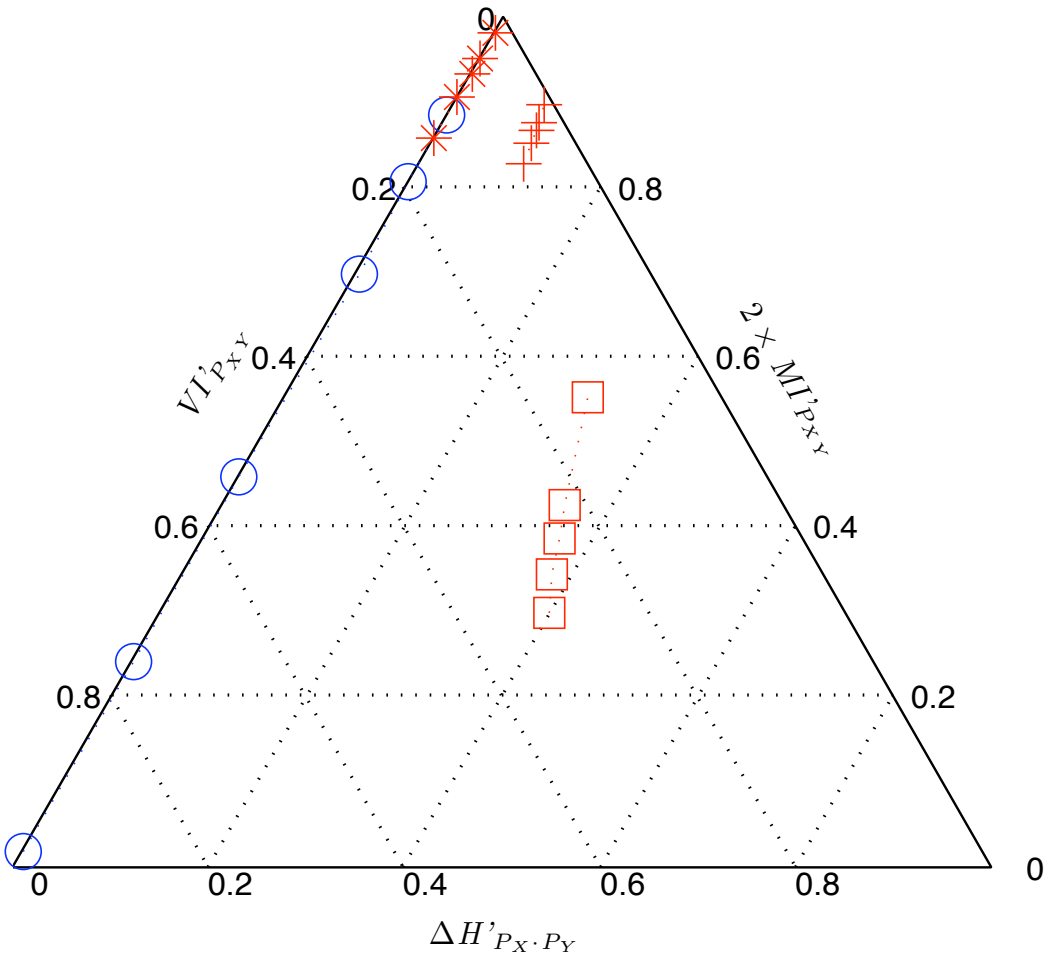


Figure : Human and machine classifier performance in ASR recognition tasks.

- Blue circles: human in 16-consonant recognition at different SNR, no lexical information.
- Red squares: machine in 18-phoneme recognition task at different SNR, no lexical information.
- Red cross-hairs: same as above with lexical information.
- Red asterisks: 10-digit IWR on the same data.

Beyond first analysis: trouble in Paradise

- The Entropy triangle gives criteria to say when a classifier is balanced/random/accurate, and their combinations.
- But it does not tell you how to improve it.
- It does not *even* apportion blame to the components that may be at fault.
- A first step to do this is to split the contributions of the input and output distributions...

Split Entropy Diagram

We can rearrange the areas into a diagram like...

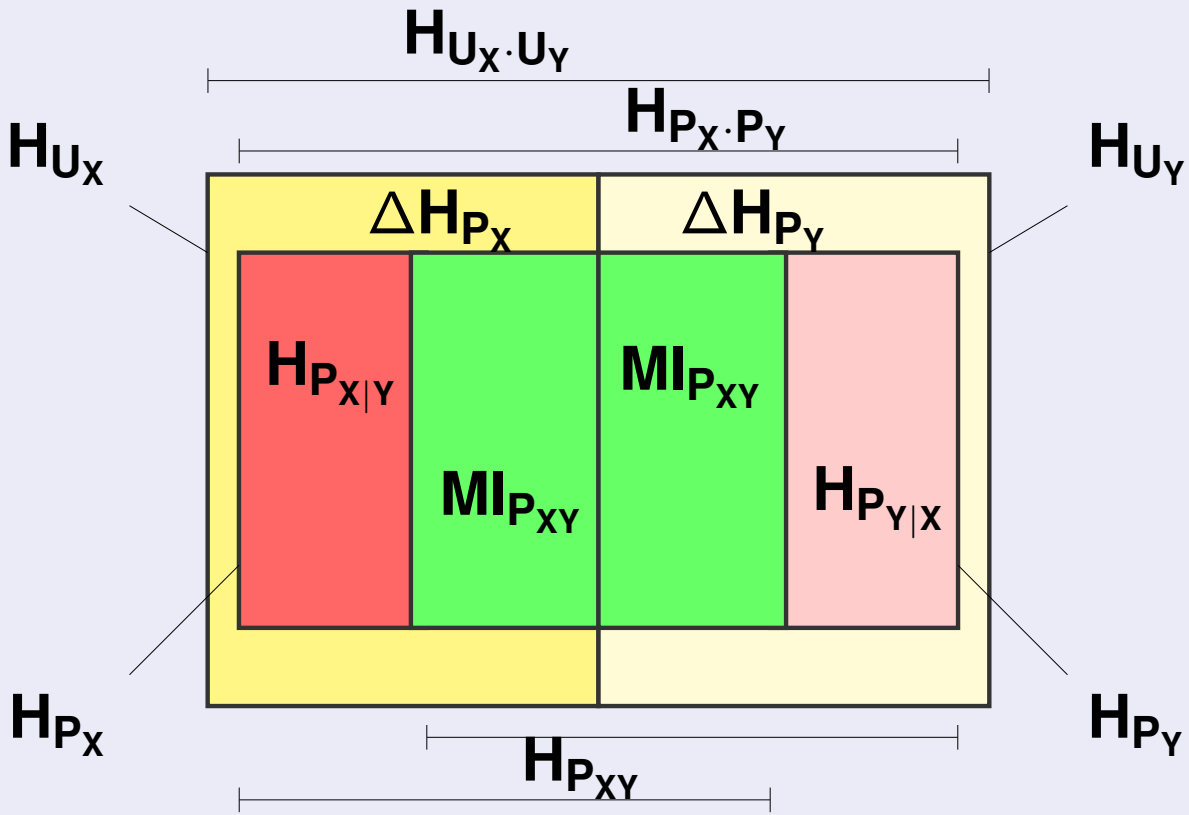


Figure : **Split entropy diagram related to a bivariate distribution.**

Split balance equations

Some of the equations in (2) can be split or dissociated...

$$
\begin{aligned}
H_{U_{XY}} &= H_{U_X} + H_{U_Y} \\
H_{P_{XY}} &= H_{P_X} + H_{P_Y} \\
\Delta H_{P_X P_Y} &= \Delta H_{P_X} + \Delta H_{P_Y}
\end{aligned} \tag{3}
$$

with $\Delta H_{P_X} = H_{U_X} - H_{P_X}$ and $\Delta H_{P_Y} = H_{U_Y} - H_{P_Y}$.

Whence we can split the *overall* balance equation...

$$
\begin{aligned}
H_{U_X} &= \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \qquad & H_{U_Y} &= \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}} \\
0 \leq \Delta H_{P_X}, MI_{P_{XY}}, H_{P_{X|Y}} &\leq H_{U_X} \qquad & 0 \leq \Delta H_{P_Y}, MI_{P_{XY}}, H_{P_{Y|X}} &\leq H_{U_Y}
\end{aligned}
$$

Split entropy triangles

Normalizing in $H_{U_X} = \log k$ and $H_{U_Y} = \log p$ we get:

$$F_X(P_{XY}) = [\Delta H'_{P_X}, MI'_{P_{XY}}, H'_{P_{X|Y}}] \quad F_Y(P_{XY}) = [\Delta H'_{P_Y}, MI'_{P_{XY}}, H'_{P_{Y|X}}]$$

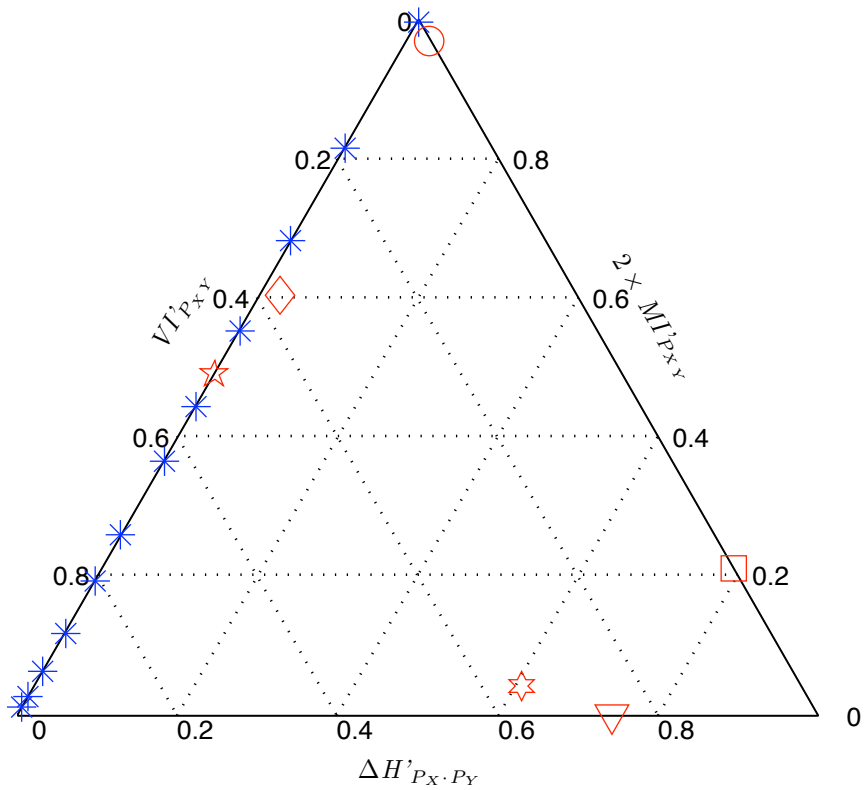
We can represent both points side by side on a single triangle:

- Increments in entropy take the role of the overall increment.
- Normalized mutual informations are proportional.
- Conditional entropies take the role of the Variation of Information.
- To distinguish between points we draw the coordinates for P_X (input) as a cross and those of P_Y (output) as a circle.

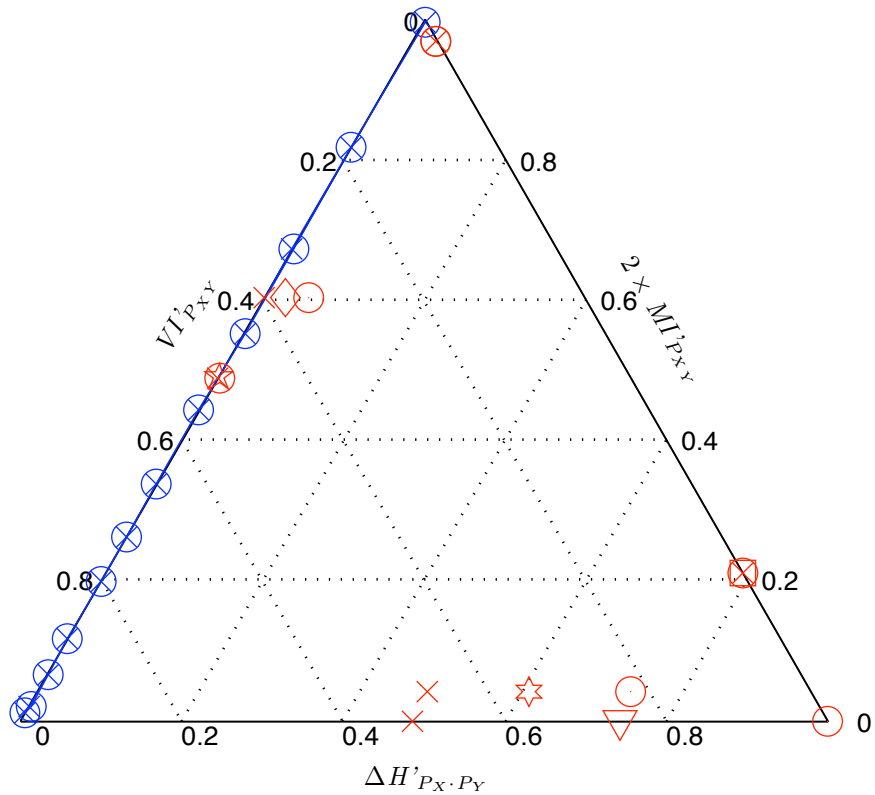
Furthermore, when $n = p$:

- Both points lie in the same horizontal line (iso-$MI'_{P_{XY}}$)
- We expect $H_{P_X} \leq H_{P_Y}$, whence $\Delta H'_{P_X} \geq \Delta H'_{P_Y}$, or else... mistrust your results!

Example 1: Split Entropy Triangle for synthetic examples



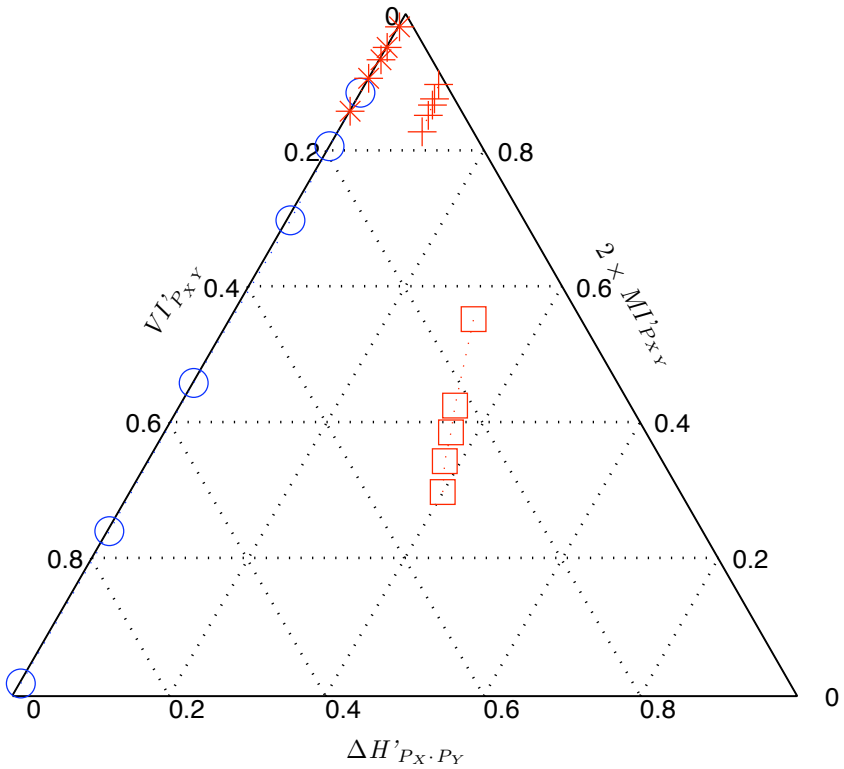
(a)



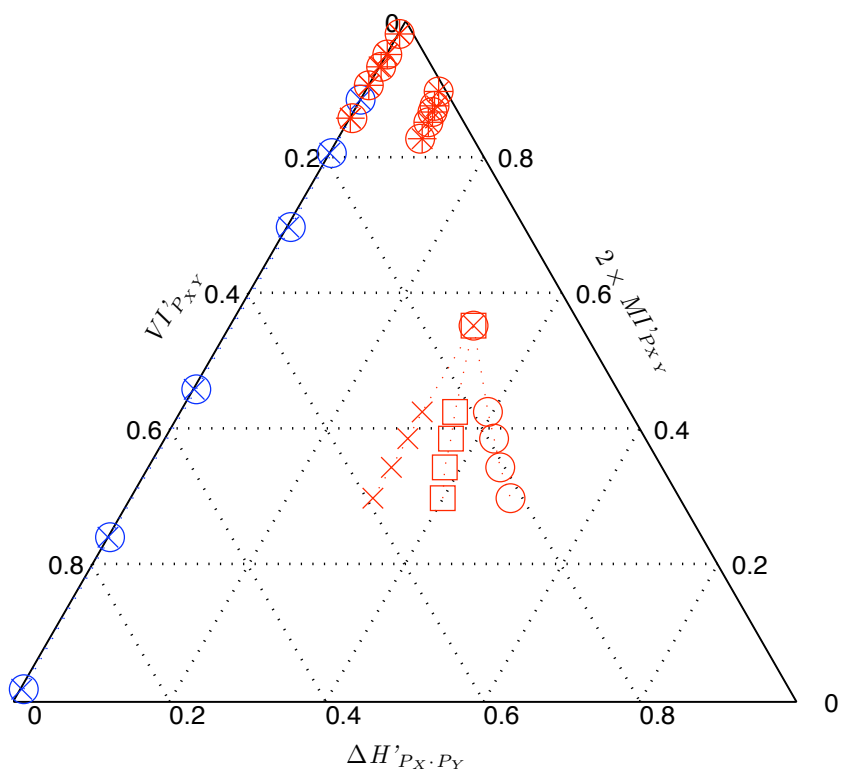
(b)

Figure : **Split entropy diagrams for the synthetic confusion matrices.**

Example 2: Split Entropy Triangle for ASR examples



(a)



(b)

Figure : **Split entropy diagrams for ASR confusion matrices.**

Majority classifiers (with $n = p$)

Majority classifiers are those that concentrate on returning as output value the most prevalent input class.

- Whence they actually have $H'_{P_{Y|X}} = 0$!
- $H'_{P_X} \geq H'_{P_Y}$ whence $\Delta H'_{P_X} \leq \Delta H'_{P_X}$, (since $H_{U_X} = H_{U_Y}$) .

Specialized classifiers are slightly less cheating...

- Specialization is a reduction in $VI'_{P_{XY}}$ caused by the increase in $\Delta H'_{P_Y}$ brought about by the reduction in $H'_{P_{Y|X}}$.
- May have some remanent $VI'_{P_{XY}}$.
- They need not show low $MI'_{P_{XY}}$ transfer!
- Classifiers with diagonal matrices ($VI'_{P_{XY}} = 0$) need not and classifiers with uniform marginals ($\Delta H'_{P_{XY}} = 0$) cannot specialize!

Maintaining uniform input marginals is a sort of regularization preventing specialization.

A metric derived from the Entropy Triangle...

The Entropy Triangle are visualization tools

- We would like to have a scalar, figure of merit.
- Accuracy is well-understood, but suffers from the "Accuracy paradox"

Accuracy paradox

Higher accuracy is not necessarily an indicator of higher classifier performance.

- Our plan is to correct accuracy by information-theoretic means.

Idea: intuitions from the perplexity of language models

Perplexity is a language-modelling measure

$$PP = 2^{H(LM)}$$

- It represents the expected no. of different words the LM can see, if they are considered equiprobable, e.g. for a LM of $|V| = 50\,000$ we may have $PP \approx 350$.
- It also allows us an estimate of the expected predictive accuracy of the Language model:

$$A'(LM) = \frac{1}{PP}$$

Perplexity and its transformation through classifiers.

The same procedure can be applied to classifiers:

$$H_{U_X} = \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \qquad H_{U_Y} = \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}}$$

$$\Downarrow \qquad \Downarrow$$

$$2^{H_{U_X}} = 2^{\Delta H_{P_X}} \cdot 2^{MI_{P_{XY}}} \cdot 2^{H_{P_{X|Y}}} \qquad 2^{H_{U_Y}} = 2^{\Delta H_{P_Y}} \cdot 2^{MI_{P_{XY}}} \cdot 2^{H_{P_{Y|X}}}$$

$$\Downarrow \qquad \Downarrow$$

$$k = \delta_X \cdot \mu_{XY} \cdot k_{X|Y} \qquad m = \delta_Y \cdot \mu_{XY} \cdot m_{Y|X}$$

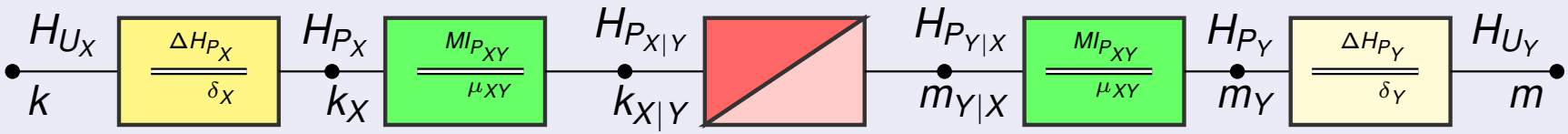


Figure : Perplexity transformation through a classifier.

Interesting quantities...

The effective perplexity of the data $k_X = k/\delta_X$

- It is an analogue for the perplexity for LM.
- It describes how many different equiprobable classes are there in the data.

$$1 \leq k_X \leq k \quad \text{since } \Delta H_X \geq 0$$

- If $k > k_X \approx 1$ then your problem is a *detection problem.*

The remanent perplexity of the data $k_{X|Y} = k_X/\mu_{XY}$

- It is the perplexity when all the information about Y is taken from X.
- It allows us to calculate the *(expected) modified accuracy*:

$$a'(P_{XY}) = 1/k_{X|Y}$$

The Normalized Information Transfer factor

The information transfer factor is $\mu_{XY} = 2^{MI_{P_{XY}}}$.

- It measures the effectiveness of the classifier!

$$1 \leq \mu_{XY} \leq k$$

- When the classifier learns nothing then $MI_{P_{XY}} = 0$ so $\mu_{XY} = 1$.
- If the input distribution of data is balanced and the classifier is the best possible then $\mu_{XY} = k$.

The Normalized Information Transfer factor $q(P_{XY}) = \mu_{XY}/k$

- Measures how much the classifier reduces the perplexity,

$$1/k \leq q(P_{XY}) \leq 1$$

- NIT is covariant with $MI_{P_{XY}}$ so rankings can be shown in the ET!

TASS3 for Sentiment Analysis

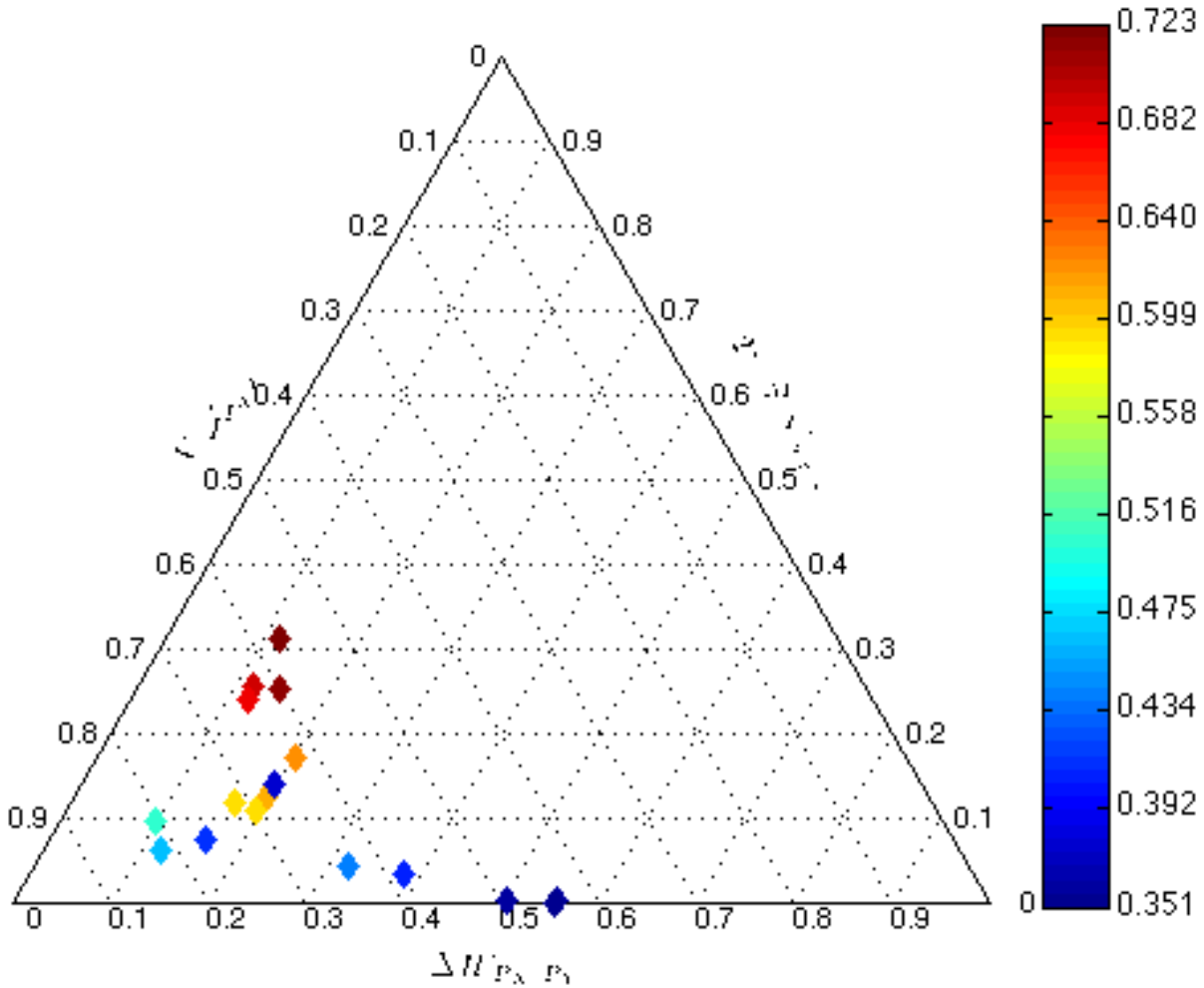


Figure : Entropy triangle for the TASS3 task. The colormap shows accuracy values.

Run	k_X	$k_{X\|Y}$	μ_{XY}	$m_{Y\|X}$	m_Y	$a(P_{XY})$	$a'(P_{XY})$	$q(P_{XY})$
daedalus-1	3.217	2.090	1.539	2.336	3.595	0.723	0.478	0.385
elhuyar-1	3.217	2.265	1.420	2.333	3.313	0.711	0.441	0.355
l2f-1	3.217	2.258	1.424	2.516	3.583	0.691	0.443	0.356
l2f-3	3.217	2.256	1.426	2.513	3.584	0.690	0.443	0.356
l2f-2	3.217	2.312	1.391	2.564	3.567	0.676	0.432	0.348
atrilla-1	3.217	2.541	1.266	2.233	2.827	0.620	0.394	0.316
sinai-4	3.217	2.706	1.189	2.432	2.891	0.606	0.370	0.297
uned1-1	3.217	2.735	1.176	2.658	3.127	0.590	0.366	0.294
uned1-2	3.217	2.766	1.163	2.495	2.902	0.588	0.362	0.291
uned2-1	3.217	2.819	1.141	3.316	3.783	0.501	0.355	0.285
imdea-1	3.217	2.953	1.089	3.266	3.558	0.459	0.339	0.272
uned2-2	3.217	3.033	1.061	1.922	2.039	0.436	0.330	0.265
uned2-4	3.217	2.900	1.109	2.876	3.190	0.412	0.345	0.277
uned2-3	3.217	3.070	1.048	1.644	1.722	0.404	0.326	0.262
uma-1	3.217	2.649	1.214	2.369	2.876	0.376	0.377	0.304
sinai-2	3.217	3.212	1.001	1.228	1.230	0.358	0.311	0.250
sinai-1	3.217	3.213	1.001	1.061	1.062	0.356	0.311	0.250

RepLab12 for Reputation Analysis

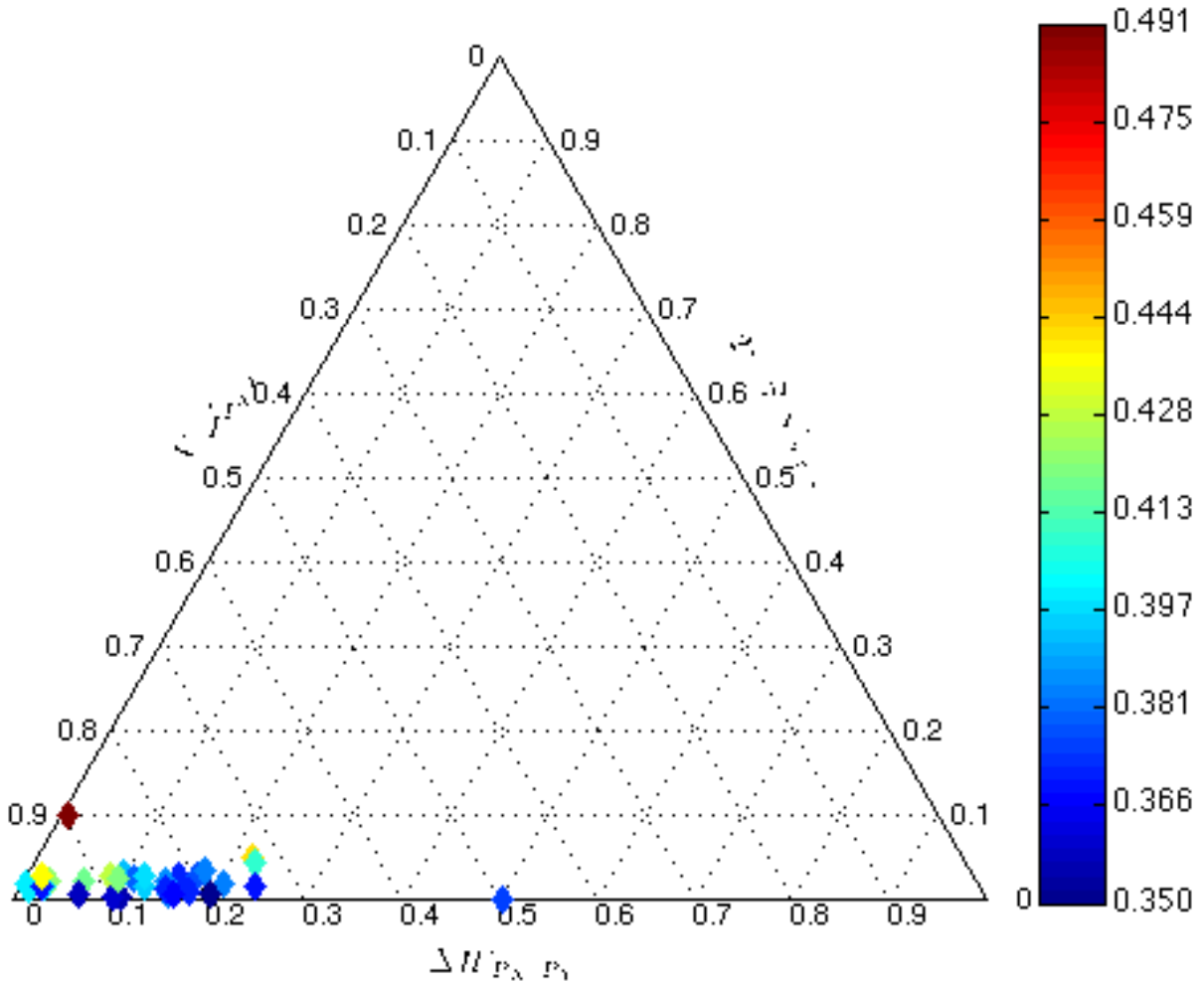


Figure : Entropy triangle for the RepLab12 task. The colormap shows accuracy values.

By-entity results for a RepLab12 participant

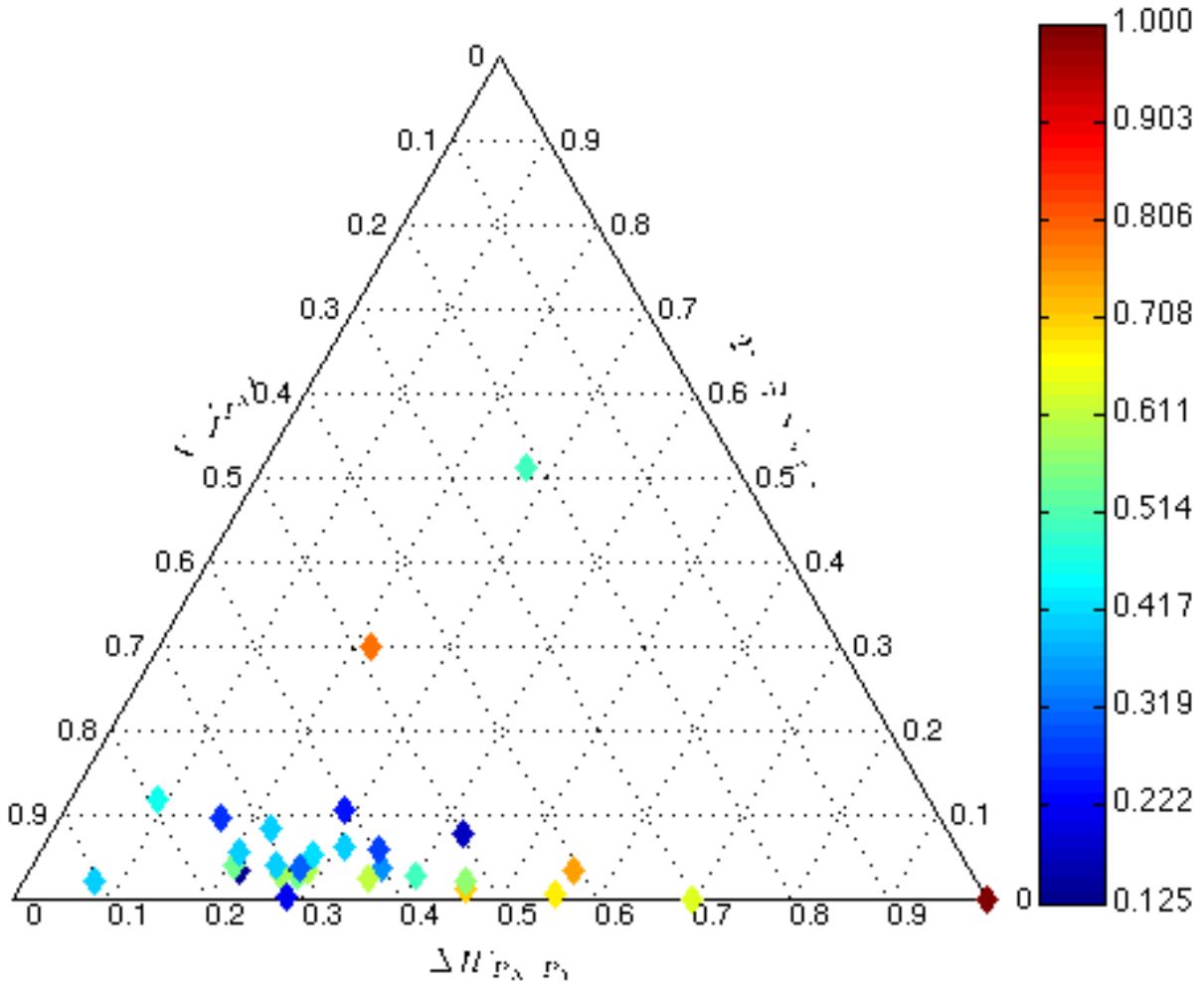


Figure : Entropy triangle for by-entity results.

Summary

A new set of tools for assessing the performance of multi-class classifiers in terms of entropic measures:

- The de Finetti entropy diagram (or Entropic Triangle) shows that there exists a coupling among,
  - a term related to the uniformness of the marginal distributions ($\Delta H'_{P_X \cdot P_Y}$),
  - a dissimilarity (Variation of Information) and
  - a similarity (Mutual Information) between the input and output experimental descriptions.
- The balance equation can be split into input and output equations providing information about the specialization of classifiers.
- The modified accuracy provides a more pessimistic estimate on the classifier performance.
- The Normalized Information Tranfers factor gives an estimate of the effectiveness of the learning process.

Extensions and outlook

Future extensions...

- Apply results to clustering based in CM comparison.
- Warped diagram to emphasize some working points.
- Entropy-guided training(?), that is, using the Entropy Triangle in the training process.