A Proposal for new Evaluation Metrics and Result Visualization Technique for Sentiment Analysis Tasks

Francisco J. Valverde-Albacete[1] Jorge Carrillo-de-Albornoz[1]
Carmen Peláez-Moreno[2]

[1]NLP & IR group, Dep. Lenguajes y Sistemas Informáticos
UNED, Spain

[2]Dept. Teoría de la Señal y Comms.
Universidad Carlos III de Madrid

24/09/2013/ CLEF 2013

Setting the scene…

Confusion matrix or contingency table of a classifier.

- $V_X = \{x_i\}_{i=1}^{k}$ and $V_Y = \{y_j\}_{j=1}^{m}$ 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

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

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

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

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

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.

The problem with accuracy...

Accuracy is the fraction of correct guesses.

- It is well-understood, but suffers from...

The Acccuracy paradox

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

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

- So we first transform counts into a joint probability:

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

A plethora of measures of performance

Information-theoretic measures (13+)

- 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 dissimilarity)

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

Entropies related to P_{XY} [1]

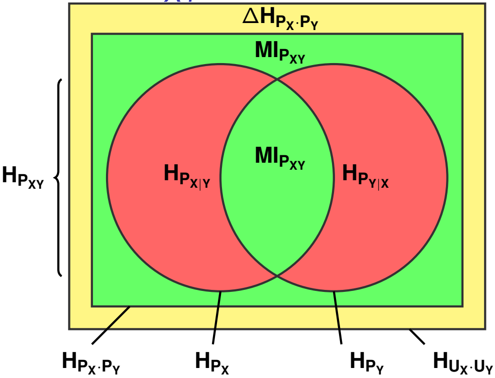


Figure : **Extended entropy diagram related to a bivariate distribution.**

$$
\begin{aligned}
H_{P_{XY}} &= H_{P_{X|Y}} + H_{P_{Y|X}} + MI_{P_{XY}} \qquad (2) \\
H_{P_X \cdot P_Y} &= MI_{P_{XY}} + H_{P_{XY}} \\
H_{U_X \cdot U_Y} &= \Delta H_{P_X \cdot P_Y} + H_{P_X \cdot P_Y}
\end{aligned}
$$

[1] Valverde-Albacete, F. J., Peláez-Moreno, C., 2010. Two information-theoretic tools to

The Balance equations

Adding the equations in (2) reads...

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

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!

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

The interpretation of Entropy Triangles

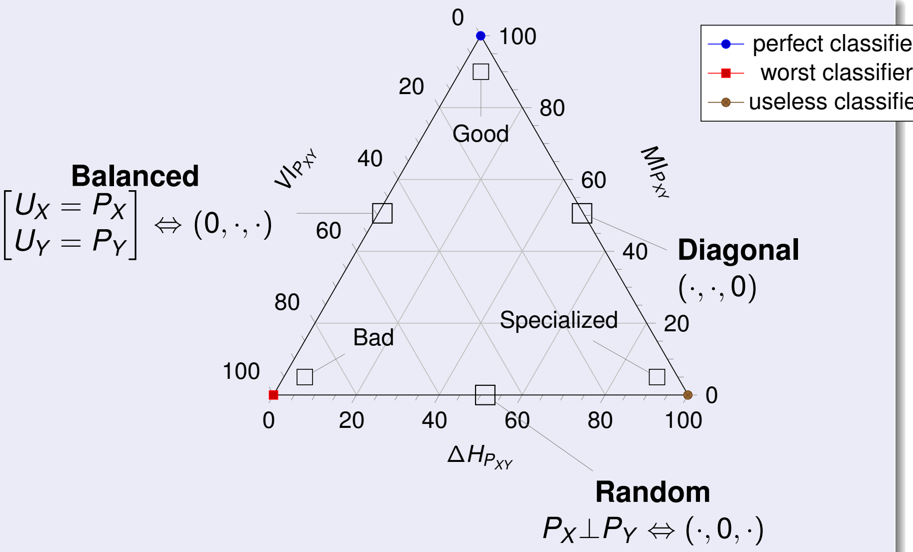


Figure : Schematics on how to interpret the zones in the entropy triangle.

From the 2-simplex to the De Finetti entropy diagrams

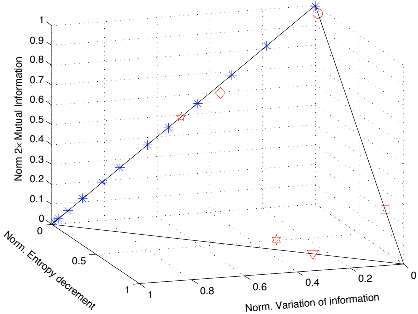


Figure : The 2-simplex in three-dimensional, normalized entropy space $[\Delta H'_{P_X \cdot P_Y}, VI'_{P_{XY}}, 2MI'_{P_{XY}}]$

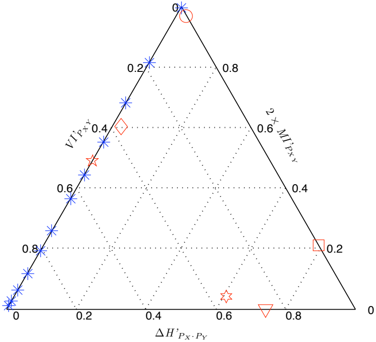


Figure : The de Finetti entropy diagram or entropy triangle, a projection of the 2-simplex onto a two-dimensional space. Example with synthetic data in previous slide.

1st idea: the Split Entropy Diagram

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

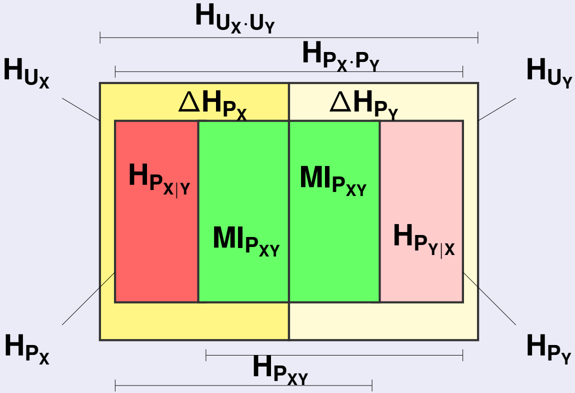


Figure : **Split entropy diagram related to a bivariate distribution.**

Split balance equations

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

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

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. . .

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

$$0 \leq \Delta H_{P_X}, MI_{P_{XY}}, H_{P_{X|Y}} \leq H_{U_X} \qquad 0 \leq \Delta H_{P_Y}, MI_{P_{XY}}, H_{P_{Y|X}} \leq H_{U_Y}$$

2nd 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:

$$a'(LM) \approx \frac{1}{PP}$$

Perplexity and its transformation through classifiers.

The same procedure can be applied to classifiers:

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

$$\Downarrow \qquad \Downarrow$$

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

$$\Downarrow \qquad \Downarrow$$

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

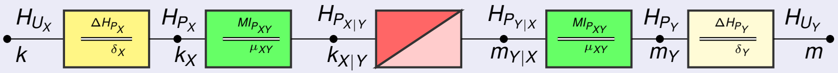


Figure : Perplexity transformation through a classifier.

Interesting quantities. . .

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

- It is inherent to the task corpus.
- It is an analogue for the perplexity for LM.
- It describes how many different equiprobable classes are there in the corpus.

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

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

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

- It is the perplexity when all the information about Y is "taken" from X.

Finally the *entropy modified accuracy (EMA)* is

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

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

- It measures *how much the classifier reduces perplexity*,

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

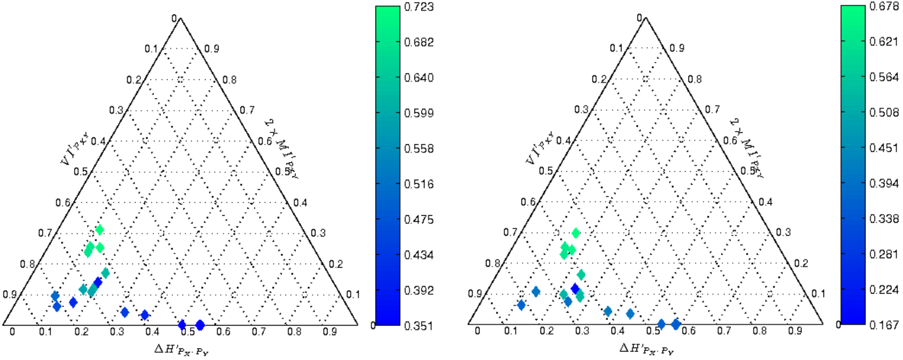
- NIT is covariant with $MI_{P_{XY}}$ so rankings can be read from the ET!

The TASS tasks

Table : **Distribution of tweets per polarity class in the TASS corpus.** The training sets are much more balanced.

TASS5	P+	P	NEU	N	N+	NONE	TOTAL	k_X
training	1 764	1 019	610	1 221	903	1 702	7 219	5.6
testing	20 745	1 488	1 305	11 287	4 557	21 416	60 798	4.1
TASS3								
training		2 783	610	2 124		1 702	7 219	3.6
testing		22 233	1 305	15 844		21 416	60 798	3.2

TASS3 vs. TASS5 results



(a) (b)

Figure : **Entropy triangles for the TASS Sentiment Analysis tasks for 3 (a) and 5 (b) polarity degrees.** Colormap correlates with accuracy.

RepLab 2012 data

Table : **Distribution of tweets per polarity class in the RepLab 2012 corpus.** Effective perplexities are very different for training and testing.

Dataset	P	NEU	N	TOTAL	k_X
training	885	550	81	1 516	2.32
testing	1 625	1 488	1 241	4 354	2.98

RepLab 2012 results

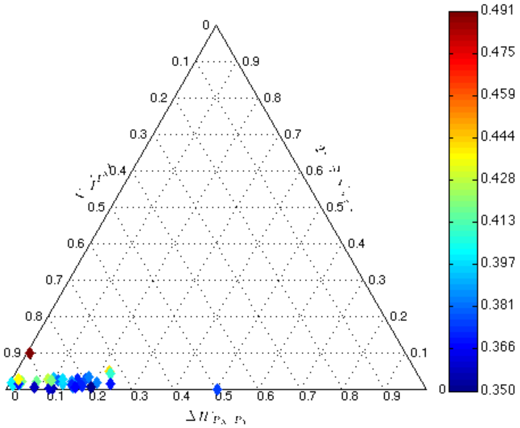


Figure : **Entropy triangles for the whole population of systems presented to the RepLab2012 Reputation Analysis.** The colormap encodes accuracy. The task is not solved, even as a collective effort, taking the NIT as the criterion.

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.
- Modified accuracy provides a more pessimistic (realistic?) *estimate of classifier performance*.
- The Normalized Information Tranfers factor gives an *estimate of the effectiveness of the learning process*.

Acknowledgments

- FJVA and JCdA are supported by EU FP7 project LiMoSINe (contract 288024).
- CPM has been partially supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología project TEC2011-26807 for this paper.