Pattern Recognition Letters 31 (2010) 1665–1671

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Two information-theoretic tools to assess the performance of multi-class classifiers

Francisco J. Valverde-Albacete *, Carmen Peláez-Moreno **

Departamento de Teoría de la Señal y de las Comunicaciones, Universidad Carlos III de Madrid, Avda de la Universidad, 30, 28911 Leganés, Spain

ARTICLE INFO

Article history:
Received 27 November 2009
Available online 21 May 2010
Communicated by R.C. Guido

Keywords:
Multi-class classifier
Confusion matrix
Contingency table
Performance measure
de Finetti diagram
Entropy triangle

ABSTRACT

We develop two tools to analyze the behavior of multiple-class, or multi-class, classifiers by means of entropic measures on their confusion matrix or contingency table. First we obtain a balance equation on the entropies that captures interesting properties of the classifier. Second, by normalizing this balance equation we first obtain a 2-simplex in a three-dimensional entropy space and then the de Finetti entropy diagram or *entropy triangle*. We also give examples of the assessment of classifiers with these tools.

1. Introduction

Since a confusion matrix is an aggregate recording of a classifier's decisions, the characterization of its performance by means of a measure or set of measures over the confusion matrix is an interesting goal.

Let $V_X = \{x_i\}_{i=1}^{n}$ and $V_Y = \{y_j\}_{j=1}^{p}$ be sets of input and output class identifiers, respectively, in a multiple-class classification task. The basic classification event consists in "presenting a pattern of input class x_i to the classifier to obtain output class identifier y_j," $(X = x_i, Y = y_j)$. The behavior of the classifier can be sampled over N iterated experiments to obtain a count matrix N_{XY} where $N_{XY}(x_i, y_j) = N_{ij}$ counts the number of times that the joint event $(X = x_i, Y = y_j)$ occurs. We say that N_{XY} is the *(count-based) confusion matrix or contingency table* of the classifier.

One often-used measure of performance is *accuracy*, the proportion of times the classifier takes the correct decision $A(N_{XY}) \approx \sum_i N_{XY}(x_i, y_i)/N$. But this has often been deemed biased towards classifiers acting on non-uniform prior distributions of input patterns (Ben-David, 2007; Sindhwani et al., 2004). For instance, with continuous speech corpora, the *silence* class may account for 40–60% of input patterns making a *majority classifier* that always decides $Y = silence$, the most prevalent class, quite accurate but useless. Related measures based in proportions over the confusion matrix can be found in (Sokolova and Lapalme, 2009).

On these grounds, Kononenko and Bratko (1991) argued for the factoring *out* of the influence of prior class probabilities in similar measures. Yet, Ben-David (2007) has argued for the use of measures that correct naturally for random decisions, like *Cohen's kappa*, although this particular measure seems to be affected by the marginal distributions.

The *Receiver Operating Characteristic (ROC)* curve (Fawcett, 2006) has often been considered a good *visual* characterization of *binary* confusion matrices built upon proportion measures, but its generalization to higher input and output set cardinals is not as effective. Likewise, an extensive *Area Under the Curve, (AUC)* for a ROC has often been considered an indication of good classifiers (Bradley, 1997; Fawcett, 2006), but the calculation of its higher dimensional analogue, the *Volume Under the Surface, (VUS)* (Hand and Till, 2001) is less manageable. It may also suffer from comparability issues across classifiers (Hand, 2009).

A better ground for discussing performance than count confusion matrices may be empirical estimates of the joint distribution between input and outputs, like the maximum likelihood estimate used throughout this letter $P_{XY}(x_i, y_j) \approx \widehat{P}_{XY}^{\mathbb{MLE}}(x_i, y_j) = N(x_i, y_j)/N$. The subsequent consideration of the classifier as an analogue of a communication channel between input and output class identifiers enables the importing of information-theoretic tools to characterize the "classification channel". This technique is already implicit in the work of Miller and Nicely (1955).

With this model in mind, Sindhwani et al. (2004) argued for entropic measures that take into account the information transfer

* Corresponding author. Tel.: +34 91 624 87 38; fax: +34 91 624 87 49.
** Corresponding author.
E-mail addresses: fva@tsc.uc3m.es (F.J. Valverde-Albacete), carmen@tsc.uc3m.es (C. Peláez-Moreno).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.05.017

through the classifier, like the *expected mutual information* between the input and output distributions (Fano, 1961)

$$MI_{P_{XY}} = \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \tag{1}$$

and provided a contrived example with three confusion matrices with the same accuracy but clearly differing performances, in their opinion due to differences in mutual information. Such examples are alike those put forth by Ben-David (2007) to argue for Cohen's kappa as an evaluation metric for classifiers.

For the related task of clustering, Meila (2007) used the *Variation of Information*, that actually amounts to the sum of their mutually conditioned entropies as a true distance between the two random variables

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

In this letter we first try to reach a more complete understanding of what is a good classifier by developing an overall constraint on the total entropy balance attached to its joint distribution. Generalizing over the input and output class set cardinalities will allow us to present a visualization tool in Section 2.2 for classifier evaluation that we will further explore in some examples both from real and synthetic data in Section 2.3. In Section 2.4 we try to extend the tools to unmask majority classifiers as bad classifiers. Finally we discuss the affordances of these tools in the context of previously used techniques.

2. Information-theoretic analysis of confusion matrices

2.1. The balance equation and the 2-simplex

Let $P_{XY}(x,y)$ be an estimate of the joint probability mass function (pmf) between input and output with marginals $P_X(x) = \sum_{y_j \in Y} P_{X,Y}(x,y_j)$ and $P_Y(y) = \sum_{x_i \in X} P_{X,Y}(x_i,y)$.

Let $Q_{XY} = P_X \cdot P_Y$ be the pmf[1] with the same marginals as P_{XY} considering them to be independent (that is, describing independent variables). Let $U_{XY} = U_X \cdot U_Y$ be the product of the uniform, maximally entropic pmfs over X and Y, $U_X(x) = 1/n$ and $U_Y(y) = 1/p$. Then the loss in uncertainty from U_{XY} to Q_{XY} is the difference in entropies:

$$\Delta H_{P_X \cdot P_Y} = H_{U_X \cdot U_Y} - H_{P_X \cdot P_Y}. \tag{2}$$

Intuitively, $\Delta H_{P_X \cdot P_Y}$ measures how far the classifier is operating from the most general situation possible where all inputs are equally probable, which prevents the classifier from specializing in an overrepresented class to the detriment of classification accuracy in others. Since $H_{U_X} = \log n$ and $H_{U_Y} = \log p$, $\Delta H_{P_X \cdot P_Y}$ may vary from $\Delta H^{\min}_{P_X \cdot P_Y} = 0$, when the marginals themselves are uniform $P_X = U_X$ and $P_Y = U_Y$, to a maximum value $\Delta H^{\max}_{P_X \cdot P_Y} = \log n + \log p$, when they are Kronecker delta distributions.

We would like to relate this entropy decrement to the expected mutual information $MI_{P_{XY}}$ of a joint distribution. For that purpose, we realize that the mutual information formula (1) describes the decrease in entropy when passing from distribution $Q_{XY} = P_X \cdot P_Y$ to P_{XY}

$$MI_{P_{XY}} = H_{P_X \cdot P_Y} - H_{P_{XY}}. \tag{3}$$

And finally we invoke the well-known formula relating the joint entropy $H_{P_{XY}}$ and the expected mutual information $MI_{P_{XY}}$ to the conditional entropies of X given Y, $H_{P_{X|Y}}$ (Y given X, $H_{P_{Y|X}}$ respectively):

$$H_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} + MI_{P_{XY}}. \tag{4}$$

Therefore $MI_{P_{XY}}$ may range from $MI^{\min}_{P_{XY}} = 0$ when $P_{XY} = P_X \cdot P_Y$, a bad classifier, to a theoretical maximum $MI^{\max}_{P_{XY}} = (\log n + \log p)/2$ in the case where the marginals are uniform and input and output are completely dependent, an excellent classifier.

Recall the variation of information definition in Eq. (5).

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}}. \tag{5}$$

For optimal classifiers with deterministic relation from the input to the output, and diagonal confusion matrices $VI^{\min}_{P_{XY}} = 0$, e.g., all the information about X is borne by Y and vice versa. On the contrary, when they are independent $VI^{\max}_{P_{XY}} = H_{P_X} + H_{P_Y}$, the case with inaccurate classifiers which uniformly redistribute inputs among all outputs.

Collecting Eqs. (2)–(5) results in the *balance equation for information related to a joint distribution*, our first result,

$$H_{U_{XY}} = \Delta H_{P_X \cdot P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}}. \tag{6}$$

The balance equation suggests an *information diagram* somewhat more complete than what is normally used for the relations between the entropies of two variables as depicted in Fig. 1(a) (compare to Yeung, 1991, Fig. 1). In this diagram we distinguish the familiar decomposition of the joint entropy $H_{P_{XY}}$ as the two entropies H_{P_X} and H_{P_Y} whose intersection is $MI_{P_{XY}}$. But notice that the increment between $H_{P_{XY}}$ and $H_{P_X \cdot P_Y}$ is yet again $MI_{P_{XY}}$, hence the expected mutual information appears *twice* in the diagram. Further, the interior of the outer rectangle represents $H_{U_X \cdot U_Y}$, the interior of the inner rectangle $H_{P_X \cdot P_Y}$ and $\Delta H_{P_X \cdot P_Y}$ represents their difference in areas. The absence of the encompassing outer rectangle in Fig. 1a was specifically puzzled at by Yeung (1991).

Notice that, since both U_X and U_Y on the one hand and P_X and P_Y are independent as marginals of U_{XY} and Q_{XY}, respectively, we may write:

$$\Delta H_{P_X P_Y} = (H_{U_X} - H_{P_X}) + (H_{U_Y} - H_{P_Y}) = \Delta H_{P_X} + \Delta H_{P_Y}, \tag{7}$$

where

$$\Delta H_{P_X} = H_{U_X} - H_{P_X} \quad \Delta H_{P_Y} = H_{U_Y} - H_{P_Y}. \tag{8}$$

This and the occurrence of twice the expected mutual information in Eq. (6) suggests a different information diagram, depicted in Fig. 1(b). Both variables X and Y now appear somehow decoupled—in the sense that the areas representing them are disjoint—yet there is a strong coupling in that the expected mutual information appears in both H_{P_X} and H_{P_Y}. This suggests writing separate balance equations for each variable, to be used in Section 2.4

$$H_{U_X} = \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \quad H_{U_Y} = \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}}. \tag{9}$$

Our interpretation for the balance equation is that the "raw" uncertainty available in U_{XY} minus the deviation of the input data from the uniform distribution ΔH_{P_X}, a given, is redistributed in the classifier-building process to the information being transferred from input to output $MI_{P_{XY}}$. This requires as much mutual information to stochastically bind the input to the output—thereby transforming $P_X \cdot P_Y$ into P_{XY}—and incurs in an uncertainty decrease at the output equal to ΔH_{P_Y}. The residual uncertainty $H_{P_{X|Y}} + H_{P_{Y|X}}$ should measure how efficient the process is: the smaller, the better.

To gain further understanding of the entropy decomposition suggested by the balance equation, from Eq. (6) and the paragraphs following Eqs. (2)–(5), we obtain

$$\begin{aligned} &H_{U_{XY}} = \Delta H_{P_X \cdot P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}} \\ &0 \leqslant \Delta H_{P_X \cdot P_Y}, 2MI_{P_{XY}}, VI_{P_{XY}} \leqslant H_{U_{XY}} \end{aligned}$$

imposing severe constraints on the values the quantities may take, the most conspicuous of which is that given two of the quantities the third one is fixed. Normalizing by $H_{U_{XY}}$ we get

$$\begin{aligned} &1 = \Delta H'_{P_X \cdot P_Y} + 2MI'_{P_{XY}} + VI'_{P_{XY}} \\ &0 \leqslant \Delta H'_{P_X \cdot P_Y}, 2MI'_{P_{XY}}, VI'_{P_{XY}} \leqslant 1. \end{aligned} \tag{10}$$

[1] We drop the explicit variable notation in the distributions from now on.

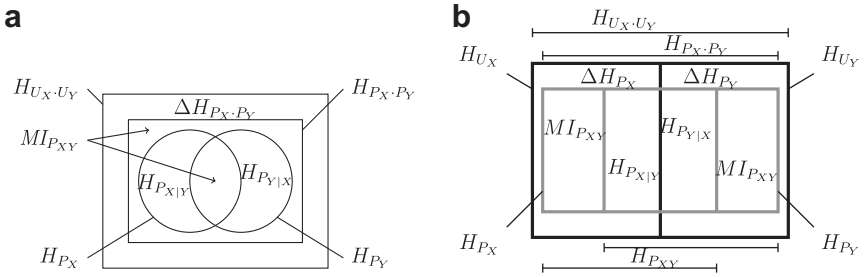

Fig. 1. Extended information diagrams of entropies related to a bivariate distribution: the expected mutual information appears *twice*: (a) extended diagram and (b) modified extended diagram.

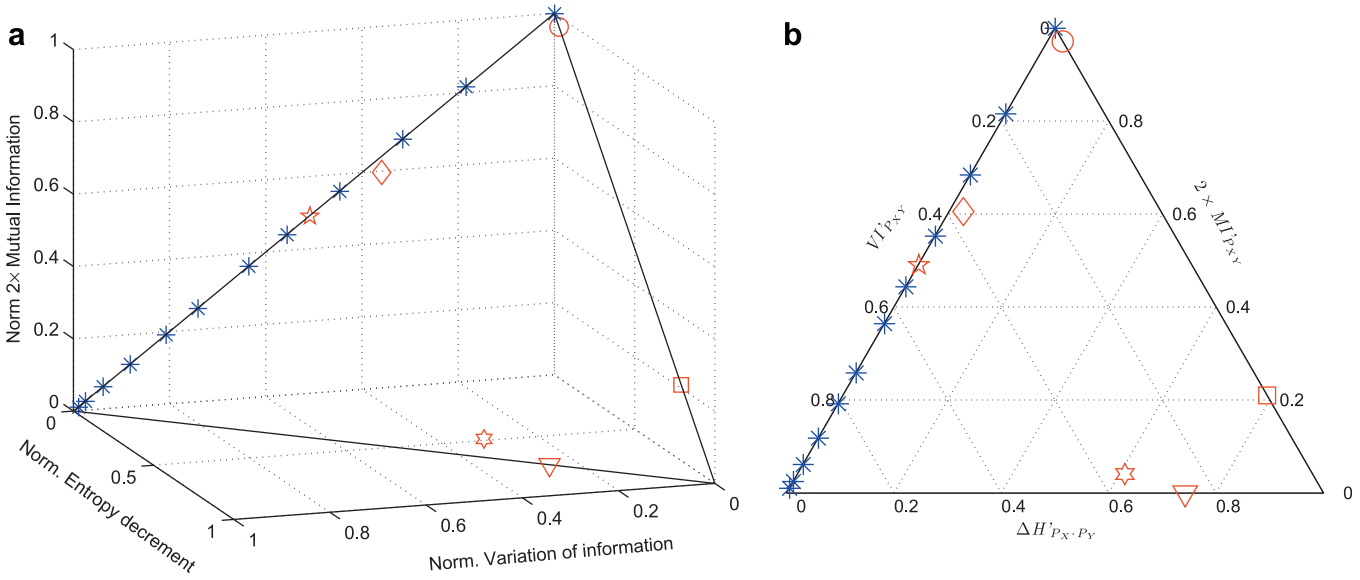

Fig. 2. Entropic representations for bivariate distribution of the synthetic examples of Fig. 3: (a) The 2-simplex in three-dimensional, normalized entropy space $\Delta H'_{P_X \cdot P_Y} \times VI'_{P_{XY}} \times 2MI'_{P_{XY}}$ and (b) the de Finetti entropy diagram or entropy triangle, a projection of the 2-simplex onto a two-dimensional space (explanations in Section 2.3).

This is the 2-simplex in normalized $\Delta H'_{P_X \cdot P_Y} \times 2MI'_{P_{XY}} \times VI'_{P_{XY}}$ space depicted in Fig. 2(a), a three-dimensional representation of classifier performance: each classifier with joint distribution P_{XY} can be characterized by its *joint entropy fractions*, $F_{XY}(P_{XY}) = [\Delta H'_{P_{XY}}, 2 \times MI'_{P_{XY}}, VI'_{P_{XY}}]$.

2.2. de Finetti entropy diagrams

Since the ROC curve is a bi-dimensional characterization of binary confusion matrices we might wonder if the constrained plane above has a simpler visualization. Consider the 2-simplex in Eq. (10) and Fig. 2(a). Its projection onto the plane with director vector (1,1,1) is its *de Finetti (entropy) diagram*, represented in Fig. 2(b). Alternatively to the three-dimensional representation, each classifier can be represented as a point at coordinates F_{XY} in the de Finetti diagram.

The de Finetti entropy diagram shows as an equilateral triangle, hence the alternative name *entropy triangle*, each of whose sides and vertices represents classifier performance-related *qualities*:

- If P_X and P_Y are independent in $Q_{XY} = P_X \cdot P_Y$ then $F_{XY}(Q_{XY}) = [\cdot, 0, \cdot]$. The lower side is the geometric locus of distributions with no mutual information transfer between input and output: the closer a classifier is to this side, the more unreliable the classifier decisions are.
- If the marginals of P_{XY} are uniform $P_X = U_X$ and $P_Y = U_Y$ then $F_{XY}(P_{XY}) = [0, \cdot, \cdot]$. This is the locus of classifiers that are not trained with overrepresented classes and therefore cannot specialize in any of them: the closer to this side, the more generic the classifier.
- Finally, if P_{XY} is a diagonal matrix, then $P_X = P_Y$ and $F_{XY}(P_{XY}) = [\cdot, \cdot, 0]$. The right-hand side is the region of classifiers with no variation of information, that is, no remanent information in the conditional entropies: this characterizes classifiers which transfer as much information from H_{P_X} to H_{P_Y} as they can.

Moving away from these sides the corresponding magnitudes grow until saturating at the opposite vertices, which therefore represent ideal, *prototypical classifier loci*:

- The upper vertex $F_{XY}(optimal) = [0, 1, 0]$ represents *optimal classifiers* with the highest information transfer from input to output and highly entropic priors.
- The vertex to the left $F_{XY}(inaccurate) = [1, 0, 0]$ represents *inaccurate classifiers*, with low information-transfer with highly entropic priors.
- The vertex to the right $F_{XY}(underperforming) = [0, 0, 1]$ represents *underperforming classifiers*, with low information transfer and low-entropic priors either at an easy task or failing to deliver performance.

In the next section we develop intuitions over the de Finetti diagram by observing how typical examples—real and synthetic—appear in it.

But first we would like to extend it theoretically to cope with the separate information balances of the marginal distributions. Recall that the modified information diagram in Fig. 1(b) suggest a decoupling of the information flow from input to output further supported by Eq. (9). These describe the *marginal fractions* of entropy when the normalization is done with H_{U_X} and H_{U_Y} respectively

$$F_X(P_{XY}) = [\Delta H'_{P_X}, MI'_{P_{XY}}, VI'_X = H'_{P_{X|Y}}] \qquad (11)$$
$$F_Y(P_{XY}) = [\Delta H'_{P_Y}, MI'_{P_{XY}}, VI'_Y = H'_{P_{Y|X}}]$$

hence we may consider the de Finetti marginal entropy diagrams for both F_X and F_Y to visualize the entropy changes from input to output.

Furthermore, since the normalization factors involved are directly related to those in the joint entropy balance, and the $MI'_{P_{XY}}$ has the same value in both marginal diagrams when $n = p$, we may represent the fractions for F_X and F_Y side by side those of F_{XY} in an *extended de Finetti entropy diagram*: the point F_{XY}, being an average of F_X and F_Y, will appear in the diagram flanked by the latter two. We show in Section 2.4 examples of such extended diagrams and their use.

$$a = \begin{bmatrix} 15 & 0 & 5 \\ 0 & 15 & 5 \\ 0 & 0 & 20 \end{bmatrix} \quad b = \begin{bmatrix} 16 & 2 & 2 \\ 2 & 16 & 2 \\ 1 & 1 & 18 \end{bmatrix} \quad c = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 4 \\ 1 & 1 & 48 \end{bmatrix}$$
$$d = \begin{bmatrix} 15 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 27 \end{bmatrix} \quad e = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 57 \end{bmatrix} \quad f = \begin{bmatrix} 0 & 0 & 5 \\ 0 & 0 & 5 \\ 0 & 0 & 50 \end{bmatrix}$$

Fig. 3. Examples of synthetic confusion matrices with varied behavior: *a*, *b* and *c* from (Sindhwani et al., 2004), *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.

2.3. Examples

To clarify the usefulness of our tools in assessing classifier performance we explored data from real classifiers and synthetic examples to highlight special behaviors.

First, consider:

- matrices *a*, *b*, and *c* from Sindhwani et al. (2004), reproduced with the same name in Fig. 3,
- a matrix whose marginals are closer to a uniform distribution, a matrix whose marginals are closer to a Kronecker delta, and the confusion matrix of a majority classifier, with a delta output distribution but a more spread input distribution—matrices *d*, *e* and *f* in Fig. 3 respectively—, and
- a series of distributions obtained by convex combination $P_{XY} = (1-\lambda)\cdot(P_X \cdot P_Y) + \lambda \cdot (P_{X=Y})$ from a uniform bivariate $(P_X \cdot P_Y)$ to a uniform diagonal $(P_{X=Y})$ distribution as the combination coefficient λ ranges in $[0,1]$.

The contrived examples in (Sindhwani et al., 2004)—matrices *a*, *b*, and *c*—are represented in both diagrams in Fig. 2 as a diamond, a pentagram, and a hexagram, respectively. In that work, the examples were used to justify the need for new performance metrics, since they all showed the same accuracy. The diagrams support the intuition that matrix *a* describes a slightly better classifier than matrix *b* which describes a better classifier than matrix *c* (see Section 2.4 for a further analysis of the behavior of *c*).

Fig. 2(a) and (b) demonstrate that there are clear differences in performance between a classifier with more uniform marginals and one with marginals more alike Kronecker deltas (matrices *d* and *e* in Fig. 3, the circle and square, respectively). Furthermore, an example of a majority classifier (matrix *f*, the downwards triangle) shows in the diagram as underperforming: it will be further analyzed in Section 2.4.

From the convex combination we plotted the line of asterisks at $\Delta H'_{P_X \cdot P_Y} = 0$ in Fig. 2(a) and (b). When the interpolation coefficient for the diagonal is null, we obtain the point at $\Delta H'_{P_X \cdot P_Y} = 0$, $VI'_{P_X \cdot P_Y} = 0$ for the worst classifier. As the coefficient increases, the asterisks denote better and better hypothetical classifiers until reaching the apex of the triangle, the best. We simulated in this guise the estimation of classifiers in improving SNR ratios for each point in the line, as shown below on real data.

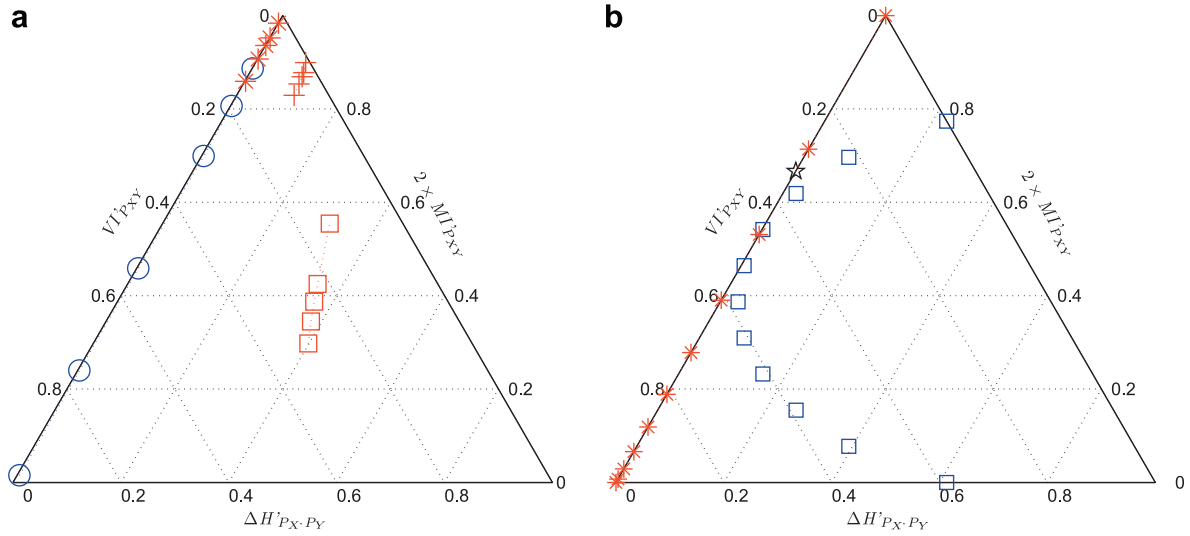

Fig. 4. Examples of use of the de Finetti entropy diagram to assess classifiers: (a) human and machine classifier performance in consonant recognition tasks and (b) the performance of some prototypical communication channel models.

In order to appraise the usefulness of the representation on real data we visualized in Fig. 4(a) the performance of several series of classifiers. The circles to the right describe a classical example of the performance of human listeners in a 16-consonant human-speech recognition task at different SNR (Miller and Nicely, 1955). They evidence the outstanding recognition capabilities of humans, always close to maximum available information transfer at $VI'_{P_{XY}} = 0$, with a graceful degradation as the available information decreases with decreasing SNR—from 12 dB at the top of the line to −18 dB at the bottom. And they also testify to the punctiliousness of those authors' in keeping to maximally generic input and output distributions at $\Delta H'_{P_X P_Y} \approx 0$.

On the other hand, the asterisks, plus signs and crosses are series of automatic speech recognizers used to compare human vs. machine performance in speech recognition tasks (Peláez-Moreno et al., in press). They motivated this work in characterizing classifiers by means of entropic measures:

- The series of squares describes a 18-class phonetic recognition task with worsening SNR that does not use any lexical information. This is roughly comparable to the experiments in (Miller and Nicely, 1955) and highlights the wide gap at present between human and machine performance in phonetic recognition.
- The series of plus signs describes phonetic confusions on the same phonetic recognition task when lexical information is incorporated. Notice that the tendency in either series is not towards the apex of the entropy triangle, but towards increasing $\Delta H'_{P_X P_Y}$, suggesting that the learning technique used to build the classifiers is not making a good job of extracting all the phonetic information available from the data, choosing to specialize the classifier instead. Further, the additional lexical constraints on their own do not seem to be able to span the gap with human performance.
- Finally, the asterisks describe a series of classifiers for a 10-digit recognition task on the same data. The very high values of all the coordinates suggest that this is a well-solved task at all those noise conditions.

Notice that, although all these tasks have different class set cardinalities, they can be equally well-compared in the same entropy triangle.

Since the simplex was developed for joint distributions, other objects characterized by these, such as *communication channel models*, may also be explored with the technique. These are high level descriptions of the end-to-end input and output-symbol classification capabilities of a communication system. Fig. 4(b) depicts three types of channels from MacKay (2003):

- the *binary symmetric channel* with $n = p = 2$ where we have made the probability of error range in $p_e \in [0, 0.5]$ in 0.05 steps to obtain the series plotted with asterisks,
- the *binary erasure channel* with $n = 2$; $p = 3$ with the erasure probability ranging in $p_e \in [0, 1.0]$ in 0.01 steps plotted with circles, and
- the *noisy typewriter* with $n = p = 27$ describing a typewriter with a convention on the errors it commits, plotted as a pentagram.

As channels are actually defined by conditional distributions $P_{Y|X}(y|x)$ we multiplied them with a uniform prior $P_X = U_X$ to plot them. Although $P_X = U_X$ the same cannot be said of P_Y what accounts for the fact that on most of the sample points in the binary erasure channel we have $\Delta H'_{P_X \cdot P_Y} \neq 0$. On the other hand, the symmetries in the binary symmetric channel and the noisy typewriter account for $\Delta H'_{P_X \cdot P_Y} = 0$.

Notice how in the entropy triangle we can even make sense of a communication channel with different input and output symbol set cardinalities, e.g. the binary erasure channel.

2.4. de Finetti diagram analysis of majority classifiers

Majority classifiers are capable of achieving a very high accuracy rate but are of limited interest. It is often required that good performance evaluation measures for classifiers show a baseline both for random and majority classifiers (Ben-David, 2007). For instance, majority classifiers should:

- have a low output entropy, a high $\Delta H'_{P_Y}$, whatever its $\Delta H'_{P_X}$ value,
- have a low information transfer $MI'_{P_{XY}}$,
- have some output conditional entropy, hence some $VI'_{P_{XY}}$.

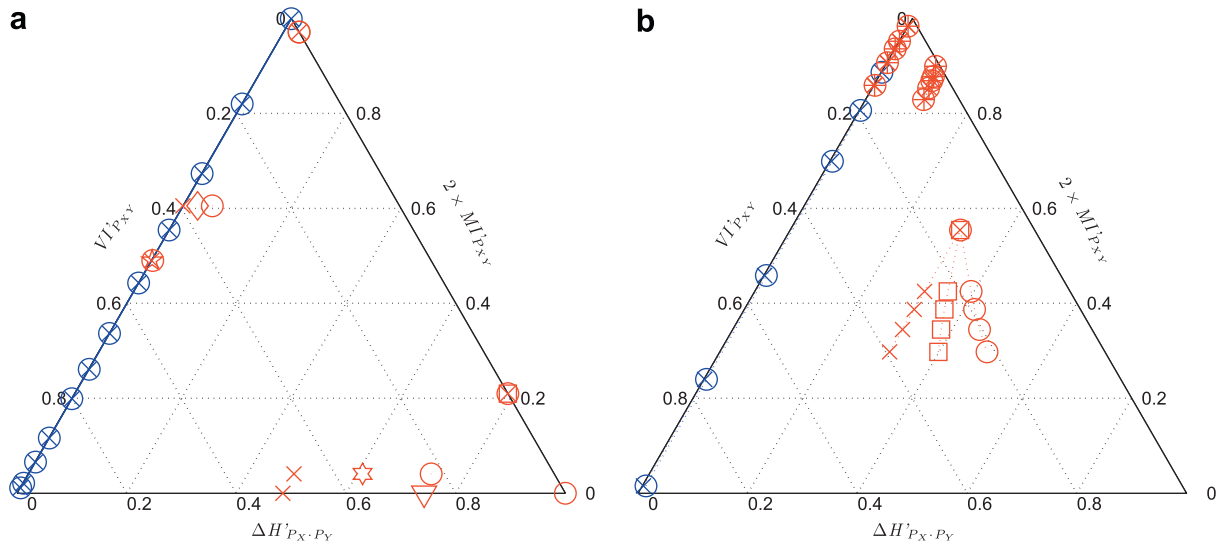

Fig. 5. Extended de Finetti entropy diagrams for synthetic and real examples: (a) for the synthetic confusion matrices of Fig. 2(b), and (b) for the real confusion matrices of Fig. 4(a). The expected mutual information coordinate is maintained in the three points for each confusion matrix.

Matrix f in Fig. 3 is the confusion matrix of majority classifier with a non-uniform input marginal. We would like to know whether this behavior could be gleaned from a de Finetti diagram.

In Fig. 5(a) we have plotted again the joint entropy fractions for the synthetic cases analyzed above, together with the entropy fractions of their marginals. For most of the cases, all three points coincide—showing as crosses within circles.

But matrices a, c and f—diamond, hexagram and downwards triangle in Fig. 5(a)—show differences in joint and marginal fractions. The most striking cases are those of matrices c and f, whose uncertainty diminishes dramatically from input to output.

Matrix f in Fig. 3 models the behavior of a majority classifier with the same input marginal as a–e. The marginal fraction points appear flanking this, at $F_X(f) = [0.45, 0, 0.55]$ and $F_Y(f) = [1, 0, 0]$. The accuracy for this classifier would be around 0.83.

In a sense, this classifier is cheating: without any knowledge of the actual classification instances it has optimized the average accuracy, but will be defeated if the input distribution gets biased towards a different class in the deployment (test) phase. It is now quite clear that c, being close to a majority classifier, attains its accuracy by specialization too.

Indeed, observing matrix a we may pinpoint the fact that its zero-pattern seems to be the interpolation of a diagonal confusion matrix and the confusion matrix of a majority classifier. This fact shows as the two flanking marginal fractions to the diamond at approximately $F_{XY}(a) = [0.03, 0.6, 0.37]$ in Fig. 5. However, since P_X was wisely kept uniform, $\Delta H'_{P_X} = 0$ at $F_Y(a) = [0, 0.6, 0.4]$ the classifier could only specialize to $F_Y(a) = [0.06, 0.6, 0.34]$.

These examples suggest that:

- *Specialization is a reduction in $VI'_{P_{XY}}$* caused by the reduction in *VI'_{P_Y} brought about by the increase in $\Delta H'_{P_Y}$, that is, by the manipulation of the output marginal distribution.*
- *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 amounts to a sort of regularization preventing specialization further from transforming all $\Delta H'_{P_Y}$ into a decrement of $VI'_{P_{XY}}$.*

For real classifiers, we have plotted in Fig. 5(b) the marginal fractions of all the classifiers in Fig. 4(a). Again, for most of them, the marginal fractions coincide with the joint fractions. But for the phonetic SpeechDat task plotted with squares we observe how with decreasing SNR the classifier has to resort to specialization. With increasing SNR it can concentrate on increasing the expected mutual information transmitted from input to output.

3. Discussion and conclusions

We have provided a mathematical tool to analyze the behavior of multi-class classifiers by means of the balance of entropies of the joint probability mass distribution of input and output classes as estimated from their confusion matrix or contingency table.

The balance equation takes into consideration the Kullback–Leibler divergence between the uniform and independent distributions with the same marginals as the original one, twice the expected mutual information between the independent and joint distributions with identical marginals—also a Kullback–Leibler divergence—and the variation of information, the difference between the joint entropy and the expected mutual information.

This balance equation can either be visualized as the 2-simplex in three-dimensional entropy space, with dimensions being normalized instances of those mentioned above; or it can be projected to obtain a ternary plot, a conceptual diagram for classifiers resembling a triangle whose vertices characterize optimal, inaccurate, or underperforming classifiers.

Motivated by the need to explain the accuracy-improving behavior of majority classifiers we also introduced the extended de Finetti entropy diagram where input and output marginal entropy fractions are visualized side by side the joint entropy fractions. This allows us to detect those classifiers resorting to specialization to increase their accuracy without increasing the mutual information. It also shows how this behavior can be limited by maintaining adequately uniform input marginals.

We have used these tools to visualize confusion matrices for both human and machine performance in several tasks of different complexities. The balance equation and de Finetti diagrams highlight the following facts:

- The expected mutual information transmitted from input to output is limited by the need to use as much entropy to bind together in stochastic dependency both variables $MI_{P_{XY}} \leqslant H_{P_X \cdot P_Y}/2$.
- Even when the mutual information between input and output is low, if the marginals have in-between uncertainty $0 < \Delta H_{P_{XY}} < \log n + \log p$ and $P_X \neq P_Y$, a classifier may become *specific*—e.g. specialize in overrepresented classes– to decrease the variation of information, effectively increasing its accuracy.
- The variation of information is actually the information *not* being transmitted by the classifier, that is, the uncoupled information between input and output. This is a good target for improving accuracy without decreasing the *genericity* of the resulting classifier, e.g., its non-specificity.

All in all the three leading assertions contextualize and nuance the assertion in (Sindhwani et al., 2004), viz. the higher the mutual information, the more generic and accurate (less specialized and inaccurate) a classifier's performance will be.

The generality and applicability of the techniques have been improved by using information-theoretic measures that pertain not only to the study of confusion matrices but, in general, to bivariate distributions such as communication channel models. However, the influence of the probability estimation method is as yet unexplored. Unlike Meila (2007), we have not had to suppose equality of sets of events in the input or output spaces or symmetric confusion matrices.

Comparing the de Finetti entropy diagram and the ROC is, at best, risky for the time being. On the one hand, the ROC is a well-established technique that affords a number of intuitions in which practitioners are well-versed, including a rather direct relation to accuracy. Also, the VUS shows promise of actually becoming a figure-of-merit for multi-class classifiers. For a more widespread use, the entropy triangle should offer such succinct, intuitive affordances too. In the case of accuracy, we intend to use Fano's inequality to bridge our understanding of proportion- and entropy-based measures.

On the other hand, the ROC only takes into consideration those judgments of the classifier *within* the joint entropy area in the Information Diagram and is thus unable to judge how close to genericity is the classifier, unlike the $\Delta H'_{P_X \cdot P_Y}$ coordinate of the entropy triangle. Likewise, the ROC has so far been unable to obtain the result that as much information as actually transmitted from input to output must go into creating the stochastic dependency between them.

To conclude, however suggestive aggregate measures like entropy or mutual information may be for capturing at a glance the behavior of classifiers, they offer little in the way of analyzing the actual classification errors populating their confusion matrices. We believe the analysis of mutual information as a random variable of a bivariate distribution (Fano, 1961, pp. 27–31) may offer

more opportunities for *improving* classifiers as opposed to *assessing* them.

Acknowledgements

This work has been partially supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología projects 2008-06382/TEC and 2008-02473/TEC and the regional projects S-505/TIC/0223 (DGUI-CM) and CCG08-UC3M/TIC-4457 (Comunidad Autónoma de Madrid – UC3M).

The authors thank C. Bousoño-Calzón and A. Navia-Vázquez for comments on early versions of this paper and A. I. García-Moral for providing the confusion matrices from the automatic speech recognizers.

References

Ben-David, A., 2007. A lot of randomness is hiding in accuracy. Eng. Appl. Artif. Intell. 20 (7), 875–885.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30 (7), 1145–1159.

Fano, R.M., 1961. Transmission of Information: A Statistical Theory of Communication. The MIT Press.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.

Hand, D.J., 2009. Measuring classifier performance: A coherent alternative to the area under the ROC curve. Mach. Learning 77 (1), 103–123.

Hand, D.J., Till, R.J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learning 45, 171–186.

Kononenko, I., Bratko, I., 1991. Information-based evaluation criterion for classifier's performance. Mach. Learning 6, 67–80.

MacKay, D.J., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Meila, M., 2007. Comparing clusterings—an information based distance. J. Multivariate Anal. 28, 875–893.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27 (2), 338–352.

Peláez-Moreno, C., García-Moral, A.I., Valverde-Albacete, F.J., in press. Analyzing phonetic confusions using formal concept analysis. J. Acoust. Soc. Amer.

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 Trans. Neural Networks 15 (4), 937–948.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inform. Process. Manage. 45 (4), 427–437.

Yeung, R., 1991. A new outlook on Shannon's information measures. IEEE Trans. Inform. Theory 37 (3), 466–474.