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

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.

Key words: Multiclass classifier, confusion matrix, contingency table, performance measure, evaluation criterion, de Finetti diagram, entropy triangle

22

23

24

25

26

27

28

29

30

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

1. Introduction

Let $V_X = \{x_i\}_{i=1}^n$ and $V_Y = \{y_j\}_{j=1}^p$ be sets of input and 2 output class identifiers, respectively, in a multiple-class 3 classification task. The basic classification event consists in "presenting a pattern of input class x_i to the clas-5 sifier to obtain output class identifier y_j ," ($X = x_i, Y =$ 6 y_i). The behavior of the classifier can be sampled over 7 N iterated experiments to obtain a count matrix N_{XY} 8 where $N_{XY}(x_i, y_j) = N_{ij}$ counts the number of times that the joint event $(X = x_i, Y = y_i)$ occurs. We say that N_{XY} 10 is the (count-based) confusion matrix or contingency ta-11 *ble* of the classifier. 12

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

One often used measure 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

Email addresses: fva@tsc.uc3m.es (Francisco J. Valverde-Albacete), carmen@tsc.uc3m.es (Carmen Peláez-Moreno) 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% percent 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) have 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).

^{*}Corresponding author. Phone: +34 91 624 87 38. Fax: +34 91 624 87 49

Preprint submitted to Pattern Recognition Letters

A better ground for discussing performance than 49 count confusion matrices may be empirical estimates of 50 the joint distribution between input and outputs, like the 51 maximum likelihood estimate used throughout this let-52 ter $P_{XY}(x_i, y_j) \approx \hat{P}_{XY}^{\mathbb{MLE}}(x_i, y_j) = N(x_i, y_j)/N$. The sub-53 100 sequent consideration of the classifier as an analogue 54 101 of a communication channel between input and output 55 102 class identifiers enables the importing of information-56 theoretic tools to characterize the "classification chan-57 nel". This technique is already implicit in the work of 58 Miller and Nicely (1955). 59

With this model in mind, Sindhwani et al. (2004) ar-60 gued for entropic measures that take into account the 61 information transfer through the classifier, like the ex-62 pected mutual information between the input and output 63 distributions (Fano, 1961) 64

65
$$MI_{P_{XY}} = \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$$
(1)

and provided a contrived example with three confusion 67 matrices with the same accuracy but clearly differing 68 performances, in their opinion due to differences in mu-69 tual information. Such examples are alike those put ¹¹⁸ 70 119 forth by Ben-David (2007) to argue for Cohen's kappa 71 as an evaluation metric for classifiers. 72

For the related task of clustering, Meila (2007) used 73 the Variation of Information, that actually amounts to 74 the sum of their mutually conditioned entropies as a true 75 distance between the two random variables 76

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

78

In this letter we first try to reach a more complete 79 understanding of what is a good classifier by develop-80 ing an overall constraint on the total entropy balance 81 attached to its joint distribution. Generalizing over the 82 input and output class set cardinalities will allow us to 83 present a visualization tool in section 2.2 for classifier 84 evaluation that we will further explore in some exam-85 ples both from real and synthetic data in section 2.3. In 86 section 2.4 we try to extend the tools to unmask major-87 ity classifiers as bad classifiers. Finally we discuss the 88 affordances of these tools in the context of previously 89 used techniques. 90

2. Information-Theoretic Analysis of Confusion 91 Matrices 92

2.1. The Balance equation and the 2-simplex 93

Let $P_{XY}(x, y)$ be an estimate of the joint proba-94 bility mass function (pmf) between input and output 95

with marginals $P_X(x) = \sum_{y_i \in Y} P_{X,Y}(x, y_j)$ and $P_Y(y) =$ $\sum_{x_i \in X} P_{X,Y}(x_i, y)$.

96

97

98

99

104

105

107

108

109

110

111

112

113 114

115

116

117

120

121

122

124

125

126

127

131

132

133

134

135

136

137

139

140

141

Let $Q_{XY} = P_X \cdot P_Y$ be the pmf¹ 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 $\Delta H_{P_X,P_Y}^{\min} = 0$, when the marginals themselves are uniform $P_X = U_X$ and $P_Y = U_Y$, to a maximum value $\Delta H_{P_X \cdot P_Y}^{\max} = \log n + \log p$, 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):

$$H_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} + MI_{P_{XY}}$$
(4)

Therefore $MI_{P_{XY}}$ may range from $MI_{P_{XY}}^{\min} = 0$ when $P_{XY} = P_X \cdot P_Y$, a bad classifier, to a theoretical maximum $MI_{P_{YY}}^{\text{max}} = (\log n + \log p)/2$ 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 $VI_{P_{XY}}^{\min} = 0$, e.g., all the information about X is borne by

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

142Y and vice versa. On the contrary, when they are inde-189143pendent $VI_{P_{XY}}^{max} = H_{P_X} + H_{P_Y}$, the case with inaccurate190144classifiers which uniformly redistribute inputs among191145all outputs.192

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

$$H_{U_{XY}} = \Delta H_{P_X \cdot P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}}$$
(6) 197

198 The balance equation suggests an information dia-151 199 gram somewhat more complete than what is normally 152 used for the relations between the entropies of two vari-153 200 ables as depicted in Fig. 1(a) (compare to Yeung, 1991, 154 201 202 Fig. 1). In this diagram we distinguish the familiar de-155 composition of the joint entropy $H_{P_{XY}}$ as the two en-156 203 tropies H_{P_X} and H_{P_Y} whose intersection is $MI_{P_{XY}}$. But 157 notice that the increment between $H_{P_{XY}}$ and $H_{P_X \cdot P_Y}$ is 158 205 yet again $MI_{P_{XY}}$, hence the expected mutual informa-159 206 tion appears twice in the diagram. Further, the interior 160 of the outer rectangle represents $H_{U_X \cdot U_Y}$, the interior of 161 207 the inner rectangle $H_{P_X \cdot P_Y}$ and $\Delta H_{P_X \cdot P_Y}$ represents their 162 208 209 difference in areas. The absence of the encompassing 163 outer rectangle in Fig. 1a was specifically puzzled at by 164 210 Yeung (1991). 165

¹¹⁶⁶ Notice that, since both U_X and U_Y on the one hand ¹¹⁷ and P_X and P_Y are independent as marginals of U_{XY} and ¹¹⁸ Q_{XY} , respectively, we may write:

¹⁶⁹
$$\Delta H_{P_X P_Y} = (H_{U_X} - H_{P_X}) + (H_{U_Y} - H_{P_Y}) = \Delta H_{P_X} + \Delta H_{P_Y}$$
²¹⁵
(7)

171 where

172 173 $\Delta H_{P_{\chi}} = H_{U_{\chi}} - H_{P_{\chi}} \qquad \Delta H_{P_{\chi}} = H_{U_{\chi}} - H_{P_{\chi}} \qquad (8)^{-218}$

This and the occurrence of twice the expected mutual 220 174 information in Eq. (6) suggests a different information 221 175 diagram, depicted in Fig. 1(b). Both variables X and ²²² 176 Y now appear somehow decoupled—in the sense that ²²³ 177 the areas representing them are disjoint-yet there is a 224 178 strong coupling in that the expected mutual information 225 179 appears in both H_{P_X} and H_{P_Y} . This suggests writing 226 180 separate balance equations for each variable, to be used 227 181 in Sec. 2.4, 228 182

$$H_{U_X} = \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \qquad H_{U_Y} = \Delta H_{P_Y} + MI_{P_{XY}} + H_{184}$$
(9) 230

¹⁸⁵ Our interpretation for the balance equation is that the ²³² ¹⁸⁶ "raw" uncertainty available in U_{XY} minus the deviation ²³³ ¹⁸⁷ of the input data from the uniform distribution ΔH_{P_X} , a ²³⁴ ¹⁸⁸ given, is redistributed in the classifier-building process ²³⁵ 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 $P_X \dot{P}_Y$ 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

196

216

217

219

229

231

3

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

$$0 \le \Delta H_{P_X \cdot P_Y}, 2MI_{P_{XY}}, VI_{P_{XY}} \le H_{U_{XY}}$$

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

$$1 = \Delta H'_{P_X,P_Y} + 2MI'_{P_{XY}} + VI'v_{P_{XY}}$$
(10)
$$0 \le \Delta H'_{P_Y,P_Y}, 2MI'_{P_{YY}}, VI'_{P_{YY}} \le 1 .$$

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 threedimensional 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 is (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*:

 $H_{P_{Y|X}}$. • 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.

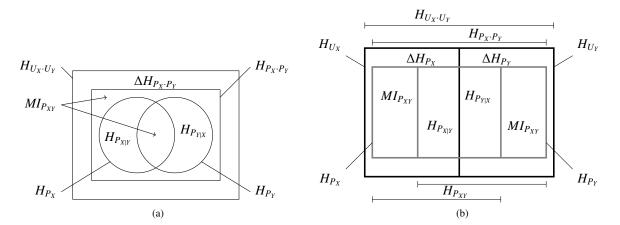


Figure 1: Extended information diagrams of entropies related to a bivariate distribution: the expected mutual information appears *twice*. (a) Extended diagram, and (b) Modified extended diagram.

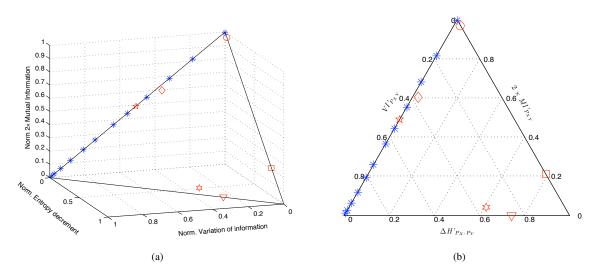


Figure 2: (color on-line) Entropic representations for bivariate distribution of the synthetic examples of Fig. 3: (a) The 2-simplex in threedimensional, normalized entropy space $\Delta H'_{P_X,P_Y} \times VI'_{P_{XY}} \times 2MI'_{P_{XY}}$ and (b) the de Finetti entropy diagram or entropy triangle,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

• If the marginals of P_{XY} are uniform $P_X = U_X$ and ²⁴⁸ 236 $P_Y = U_Y$ then $F_{XY}(P_{XY}) = [0, \cdot, \cdot]$. This is the 237 249 locus of classifiers that are not trained with over-238 250 represented classes and therefore cannot specialize 239 251 in any of them: the closer to this side, the more 240 252 generic the classifier. 241

253 • Finally, if P_{XY} is a diagonal matrix, then $P_X = P_Y$ 242 254 and $F_{XY}(P_{XY}) = [\cdot, \cdot, 0]$. The right-hand side is 243 255 the region of classifiers with no variation of in-244 256 formation, that is, no remanent information in the 245 conditional entropies: this characterizes classifiers 257 246 which transfer as much information from H_{P_X} to 258 247 4

$H_{P_{Y}}$ 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 $F_{XY}(inaccurate) = [0, 0, 1]$ represents *inaccurate classifiers*, with low

information-transfer with highly entropic priors. 259

• The vertex to the right $F_{XY}(underperforming) =$ 260 307 [0,0,1] represents underperforming classifiers, 261 308 with low information transfer and low-entropic pri-262 309 ors either at an easy task or refusing to deliver per-263

formance. 264

In the next section we develop intuitions over the de 265 312 Finetti diagram by observing how typical examples, real 266 313 and synthetic appear in it. 267 314

But first we would like to extend it theoretically 268 to cope with the separate information balances of the 269 marginal distributions. Recall that the modified infor-270 mation diagram in Fig. 1(b) suggest a decoupling of the 271 information flow from input to output further supported 272 319 by eq. 9. These describe the marginal fractions of en-273 320 tropy when the normalization is done with H_{U_X} and H_{U_Y} 274 321 respectively 275

276
$$F_X(P_{XY}) = [\Delta H'_{P_X}, MI'_{P_{XY}}, VI'_X = H'_{P_{X|Y}}] \quad (11)_{324}$$

$$F_{Y}(P_{XY}) = [\Delta H'_{P_{Y}}, MI'_{P_{XY}}, VI'_{Y} = H'_{P_{Y|X}}]$$

hence we may consider the de Finetti marginal entropy 279 327 diagrams for both F_X and F_Y to visualize the entropy 280 328 changes from input to output. 281

Furthermore, since the normalization factors involved 282 are directly related to those in the joint entropy balance, 283 331 and the $MI'_{P_{YY}}$ has the same value in both marginal di-284 332 agrams when n = p, we may represent the fractions for ₃₃₃ 285 F_X and F_Y side by side those of F_{XY} in an extended de ₃₃₄ 286 *Finetti entropy diagram*: the point F_{XY} , being and aver-287 age of F_X and F_Y , will appear in the diagram flanked by ₃₃₆ 288 the latter two. We show in Sec. 2.4 examples of such 289 337 extended diagrams and their use. 290

2.3. Examples 291

To clarify the usefulness of our tools in assessing 292 classifier performance we explored data from real clas-293 sifiers and synthetic examples to highlight special be-294 haviors. 295

- First, consider: 296
- matrices *a*, *b*, and *c* from Sindhwani et al. (2004), 297 reproduced with the same name in Fig. 3, 298
- a matrix whose marginals are closer to a uniform 350 299 distribution, a matrix whose marginals are closer 351 30 to a Kronecker delta, and the confusion matrix of a 301 majority classifier, with a delta output distribution 353 302 but a more spread input distribution—matrices d, e 354 303 and f in Fig. 3 respectively—, and 304

• 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-represented in both diagrams in Fig. 2 as a diamond, a pentagram, and a hexagram, respectively-pointed out there at a need for new performance metrics, since they all showed the same accuracy. The diagrams support the intuition that matrix adescribes a slightly better classifier than matrix b which describes a better classifier than matrix c (see Sec. 2.4 for a further analysis of the behavior of *c*).

Figs. 2(a) and 2(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 Sec. 2.4.

From the convex combination we plotted the line of asterisks at $\Delta H'_{P_X,P_Y} = 0$ in Figs. 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.

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 12dB at the top of the line to -18dB at the bottom. And they also testify to the punctiliousness of those authors' in keeping to maximally generic input and output distributions at $\Delta H'_{P_XP_Y} \approx 0$.

On the other hand, the asterisks, plus signs and crosses are series of automatic speech recognizers using the SpeechDat database (Moreno, 1997). They motivated this work in characterizing classifiers by means of entropic measures.

305

306

310

311

315

322

323

325

326

338

339

340

341

342

343

344

345

346

347

348

349

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

Figure 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.

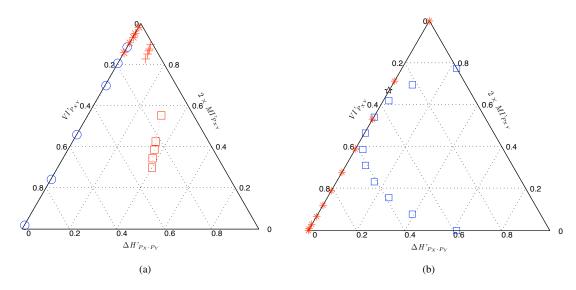


Figure 4: (color on-line) **Examples of use of the de Finetti entropy diagram to assess classifiers:** (a) human and machine classifier performance in consonant recognition tasks, and (b) the performance of some prototypical communication channel models.

380

The series of squares describes a 18-class phonetic 374 356 recognition task with worsening SNR that does not 357 375 use any lexical information. This is roughly com-358 376 parable to the experiments in Miller and Nicely 359 377 (1955) and highlights the wide gap at present be-360 378 tween human and machine performance in pho-361 379 netic recognition. 362

381 The series of plus signs describes phonetic confu-363 382 sions on the same phonetic recognition task when 364 383 lexical information is incorporated. Notice that 365 384 the tendency in either series is not towards the 36 apex of the entropy triangle, but towards increasing 367 386 $\Delta H'_{P_X P_Y}$, suggesting that the learning technique 368 387 used to build the classifiers is not making a good 369 388 job of extracting all the phonetic information avail-370 389 able from the data, choosing to specialize the clas-371 sifier instead. Further, the additional lexical con-390 372 straints on their own do not seem to be able to span 391 373

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 = 2where we have made the probability of error range

in $p_e \in [0, 0.5]$ in 0.05 steps to obtain the series 438 392 plotted with asterisks, 393 439

• the *binary erasure channel* with n = 2; p = 3 with 394 441 the erasure probability ranging in $p_e \in [0, 1.0]$ in 395 442 0.01 steps plotted with circles, and 396 443

444 • the *noisy typewriter* with n = p = 27 describing a 397 445 typewriter with a convention on the errors it com-39 446 mits, plotted as a pentagram. 399 447

As channels are actually defined by conditional distribu- 448 400 tions $P_{Y|X}(y|x)$ we multiplied them with a uniform prior 449 401 $P_X = U_X$ to plot them. Although $P_X = U_X$ the same 450 402 cannot be said of P_Y what accounts for the fact that on 451 403 most of the sample points in the binary erasure channel 452 404 we have $\Delta H'_{P_X \cdot P_Y} \neq 0$. On the other hand, the sym- 453 405 metries in the binary symmetric channel and the noisy 454 406 typewriter account for $\Delta H'_{P_X \cdot P_Y} = 0$. 407

Notice how in the entropy triangle we can even make 456 408 sense of a communication channel with different input 457 409 and output symbol set cardinalities, e.g. the binary era- 458 410 sure channel. 411

460 2.4. De Finetti diagram analysis of majority classifiers 412

Majority classifiers are capable of achieving a very 413 462 high accuracy rate but are of limited interest. It is often 414 463 required that good performance evaluation measures for 415 classifiers show a baseline both for random and major-464 416 ity classifiers (Ben-David, 2007). For instance, majority 417 classifiers should: 418

- 467 • have a low output entropy, a high $\Delta H'_{P_v}$, whatever 419 468 its $\Delta H'_{P_v}$ value. 420 469
- have a low information transfer $MI'_{P_{YY}}$. 421
- have some output conditional entropy, hence some 471 422 472 $VI'_{P_{YY}}$ 423

Matrix *f* in Fig. 3 is the confusion matrix of majority 474 424 classifier with a non-uniform input marginal. We would 475 425 like to know whether this behavior could be gleaned 476 426 from a de Finetti diagram. 427

In Fig. 5(a) we have plotted again the joint entropy $_{478}$ 428 fractions for the synthetic cases analyzed above, to- 479 429 gether with the entropy fractions of their marginals. For 430 most of the cases, all three points coincide-showing as 431 480 a crosses within circles. 432 But matrices a, c and f—diamond, hexagram and 481 433

downwards triangle in Fig. 5(a)—show differences in 482 434 joint and marginal fractions. The most striking cases are 483 435 those of matrices c and f, whose uncertainty diminishes 484 436 dramatically from input to output. 485 437

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:

455

459

465

466

470

473

477

7

- Specialization is a reduction in $VI'_{P_{XY}}$ caused by the reduction in $VI'_{P_{v}}$ brought about by the increase in $\Delta H'_{P_{y}}$, that is 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 $'_{P_{YY}}$.

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.

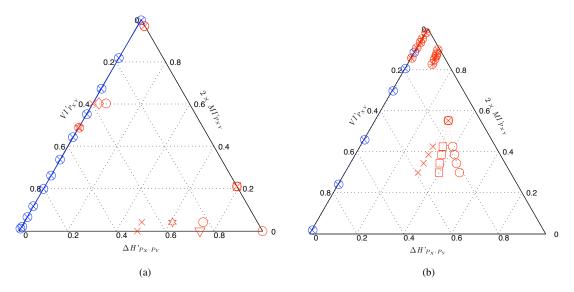


Figure 5: (color on-line) **Extended de Finetti entropy diagrams for synthetic and real examples:** (a) for the synthetic confusion matrices of Fig. 2(b), and (b) for the real confusion matrices of Fig. 4(a). The expected mutual information coordinate is maintained in the three points for each confusion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

8

The balance equation takes into consideration the 516 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 differ-

ence between the joint entropy and the expected mutual
 information.

523 This balance equation can either be visualized as 495 524 the 2-simplex in three-dimensional entropy space, with 496 525 dimensions being normalized instances of those men-497 526 tioned above; or it can be projected to obtain a ternary 498 527 plot, a conceptual diagram for classifiers resembling a 499 triangle whose vertices characterize optimal, inaccurate, 500 528 or underperforming classifiers. 501

529 Motivated by the need to explain the accuracy-502 530 improving behavior of majority classifiers we also in-503 531 troduced the extended de Finetti entropy diagram where 532 504 input and output marginal entropy fractions are visual-505 533 ized side by side the joint entropy fractions. This al-506 lows us to detect those classifiers resorting to special- 534 507 ization to increase their accuracy without increasing the 535 508 mutual information. It also shows how this behavior 536 509 can be limited by maintaining adequately uniform input 537 510 marginals. 511 538

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 542 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 $MI_{P_{XY}} \le H_{P_X \cdot P_Y}/2$.
- Even when the mutual information between input and output is low, if the marginals have inbetween 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 590 543

influence of the probability estimation method is as yet 544

unexplored. Unlike Meila (2007), we have not had to 545 592

suppose equality of sets of events in the input or output 546 593

spaces or symmetric confusion matrices. 547

Comparing the de Finetti entropy diagram and the 548 506 ROC is, at best, risky for the time being. On the one 549 hand, the ROC is a well-established technique that af-550 598 fords a number of intuitions in which practitioners are 551 600 well-versed, including a rather direct relation to accu-552 racy. Also, the VUS shows promise of actually be-553 coming a figure-of-merit for multi-class classifiers. For 603 554 a more widespread use, the entropy triangle should 555 offer such succinct, intuitive affordances too. In the 556 606 case of accuracy, we intend to use Fano's inequality 557 607 to bridge our understanding of proportion- and entropy-608 558 based measures. 559 610

On the other hand, the ROC only takes into consid-560 611 eration those judgments of the classifier within the joint 612 561 613 entropy area in the Information Diagram and is thus un-562 able to judge how close to genericity is the classifier, 563 unlike the $\Delta H'_{P_Y,P_Y}$ coordinate of the entropy triangle. 616 564 Likewise, the ROC has so far been unable to obtain the 565 618 result that as much information as actually transmitted 566 from input to output must go into creating the stochastic 567 620 dependency between them. 568

622 To conclude, however suggestive aggregate measures 569 like entropy or mutual information may be for captur-570 624 ing at a glance the behavior of classifiers, they offer lit-571 625 tle in the way of analyzing the actual classification er-572 rors populating their confusion matrices. We believe the 573 analysis of mutual information as a random variable of a 574

bivariate distribution (Fano, 1961, pp. 27-31) may offer 575

more opportunities for improving classifiers as opposed 576 to assessing them. 577

Acknowledgements 578

589

This work has been partially supported by the Span-579 ish Government-Comisión Interministerial de Cien-580 cia y Tecnología projects 2008-06382/TEC and 2008-581 02473/TEC and the regional projects S-505/TIC/0223 582 (DGUI-CM) and CCG08-UC3M/TIC-4457 (Comu-583 nidad Autónoma de Madrid - UC3M). 584

The authors would like to thank C. Bousoño-Calzón 585 586 and A. Navia-Vázquez for comments on early versions of this paper and A. I. García-Moral for providing the 587 confusion matrices from the automatic speech recogniz-588 ers.

References

594

605

609

619

- Ben-David, A., 2007. A lot of randomness is hiding in accuracy. Engineering Applications of Artificial Intelligence 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 Recognition 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 Recognition Letters 27 (8), 861-874.
- Hand, D. J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine 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. Machine Learning 45, 171-186.
- Kononenko, I., Bratko, I., 1991. Information-based evaluation criterion for classifier's performance. Machine 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, Journal of Multivariate Analysis 28, 875-893.
- Miller, G. A., Nicely, P. E., 1955. An analysis of perceptual confusions among some english consonants. The Journal of the Acoustic Society of America 27 (2), 338-352
- Moreno, A., 1997. SpeechDat Spanish Database for Fixed Telephone Network. Tech. rep., Technical University of Catalonia, Barcelona, Spain.
- 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.
- Sokolova, M., Lapalme, G., Jul 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45 (4), 427-437.
- Yeung, R., 1991. A new outlook on Shannon's information measures. IEEE Transactions on Information Theory 37 (3), 466-474.