Expert Systems With Applications 78 (2017) 145–157

Contents lists available at ScienceDirect

Expert Systems With Applications

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

The evaluation of data sources using multivariate entropy tools

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

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

ARTICLE INFO

Article history:
Received 13 July 2016
Revised 31 October 2016
Accepted 2 February 2017
Available online 7 February 2017

Keywords:
Machine learning evaluation
Dataset entropy
Multivariate entropy
Entropic measures
Exploratory analysis
Entropy ternary diagram
Entropy balance equation

ABSTRACT

We introduce from first principles an analysis of the information content of multivariate distributions as information sources. Specifically, we generalize a balance equation and a visualization device, the Entropy Triangle, for multivariate distributions and find notable differences with similar analyses done on joint distributions as models of information channels.

As an example application, we extend a framework for the analysis of classifiers to also encompass the analysis of data sets. With such tools we analyze a handful of UCI machine learning task to start addressing the question of how well do datasets convey the information they are supposed to capture about the phenomena they stand for.

1. Introduction and motivation

In this paper we introduce an information-theoretic perspective into the problem of characterizing the datasets in machine learning tasks, and obtain several tools, both theoretical and practical, to explore such problem.

Information-theory was founded by Shannon in his two-part seminal paper (Shannon, 1948a; 1948b) to provide a mathematical background to the transmission of information in the presence of noise. The last 60 years of engineering practice have revealed that this setting is far broader than initially envisaged, and many problems, both theoretical and applied, can be characterized as "relating to the transmission of information", that is, in information-theoretical terms (see, e.g. MacKay, 2003; Brillouin, 1962).

In particular, a strong current to use information-theoretic principles and heuristics in machine learning (Principe, 2010) and statistical inference (Jaynes, 1996, Chapter 11), and several methods for evaluation and analysis based on entropic measures with diverse applications have been recently published (Chen, Jin, Qiu, & Chen, 2014; Hempelmann, Sakoglu, Gurupur, & Jampana, 2016; Rödder, Brenner, & Kulmann, 2014; Valverde-Albacete & Peláez-Moreno, 2010; 2014; Zhou, Tian, Xu, Yu, & Wu, 2013).

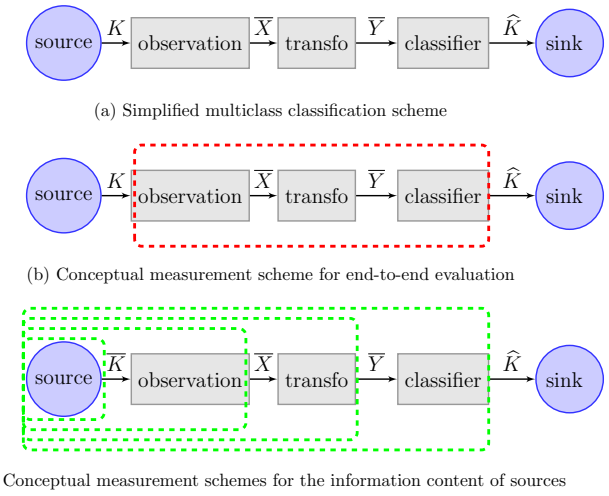
As early as McGill (1954), there emerged an interest in better understanding how the transmission of information in the *multivariate setting*—that is, among multiple variables—compares to the *bivariate setting* used by Shannon for variables X and Y. For the purpose at hand consider the scheme of Fig. 1(a) conceptualizing the supervised machine learning task of multi-class classification, cast in an information-theoretic setting. There is a set of m realizations of a random vector $\overline{X}$ of *(observed) variables or features* paired with as many realizations of a *class variable* K. The set of pairs of instances $\{(k^i, \overline{x}^i)\}_{1\leq i\leq m}$ will be called a *dataset*. For unsupervised tasks, we typically disregard K.

The feature instances $\overline{X} = \overline{x}^i$ may be further transformed to obtain instances of a random vector $\overline{Y}$, through a tranformation function $f : \overline{X} \to \overline{Y}, \overline{x}^i \mapsto \overline{y}^i = f(\overline{x}^i)$ with desired characteristics, e.g. statistical independence among the transformed features. For supervised classification, classifier induction is the subtask of inducing a function $k : \overline{Y} \to K, \overline{y}^i \mapsto \hat{k}^i = k(\overline{y}^i)$ that tries to estimate the original K but can only obtain the estimate $\hat{K}$.

For an end-to-end measure of the effectiveness of this procedure of estimating $\hat{K}$ from K as per the box in Fig. 1(b), a Shannon-type equation on the entropies around a bivariate joint distribution was introduced in Valverde-Albacete and Peláez-Moreno (2010) and later refined in Valverde-Albacete and Peláez-Moreno (2014) (see Section 2.1). It was named the *balance equation* and it leads to a new kind of exploratory graph for entropies: a ternary or *de Finetti* diagram of entropies, also called the *entropy triangle (ET)* (see Section 2.2). Both tools have been used to evaluate multiclass classifiers (Valverde-Albacete, de Albornoz, & Peláez-Moreno, 2013) using the joint distribution of results implicit in the confusion matrix over the classified instances as evaluated on the train and test data (Murphy, 2012; Theodoridis & Koutroumbas, 2006) (see Section 2.3).

* Corresponding author.
E-mail addresses: fva@tsc.uc3m.es (F.J. Valverde-Albacete), carmen@tsc.uc3m.es (C. Peláez-Moreno).

http://dx.doi.org/10.1016/j.eswa.2017.02.010
0957-4174/

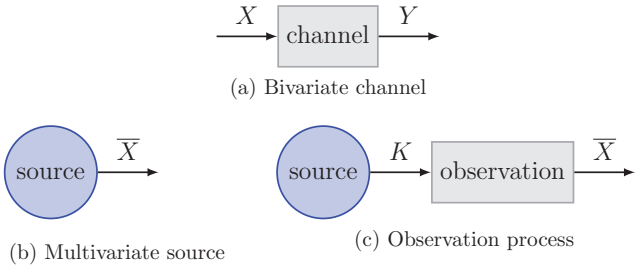


(a) Simplified multiclass classification scheme

(b) Conceptual measurement scheme for end-to-end evaluation

(c) Conceptual measurement schemes for the information content of sources

Fig. 1. Schematic representation of a multi-class classification task and measurement schemes for information-theoretic quantities.



(a) Bivariate channel

(b) Multivariate source

(c) Observation process

Fig. 2. Examples of systems susceptible of analysis with the techniques discussed in the paper. (a) Single-input single-output system studied with previous entropy triangles, and (b) opaque multivariate source, (c) multivariate source coming from an observation process, to be studied with the techniques presented in this paper.

Again, such tools allow us to analyze single-input single-output processing blocks like that in Fig. 2(a).

But in this paper, we would like to investigate whether there are analogous results for multivariate stochastic sources of information whose block diagram fragment in focus is that of Fig. 2(b). For that purpose, let $\overline{X} = \{X_i \mid 1 \leq i \leq n\}$ be a set of discrete random variables with joint multivariate distribution $P_{\overline{X}}(\overline{x}) = P_{X_1 \ldots X_n}(x_1 \ldots x_n)$— where $\overline{x} = x_1 \ldots x_n$ is a tuple of n elements—with marginals $P_{X_i}(x_i) = \sum_{j \neq i} P_{\overline{X}}(\overline{x})$.

We would also like to study the related procedure of observing a random variable K through an observation process whose result is the random vector $\overline{X}$, as depicted in Fig. 2(c), which is precisely the setting of supervised classification. With this goal in mind, in supervised tasks we may select one of the variables to represent a *class index* K in this (categorical or discrete) setting. When the support of K has more than two values $|\text{supp}(K)| \geq 2$ we call this setting *multiclass classification*; if $|\text{supp}(K)| = 2$, we call it *(binary) classification*. When this is the model of the data (as in Section 4.3) we will suppose that the classification variable K is actually adjoined to variable vector $\overline{X}$ and it is interpreted as the underlying process captured by the observation data.

In the following, we first review in Section 2 the theory and methods behind the balance equation and the entropy triangle, including a discussion of the issues that need to be addressed for their multivariate generalization, and ending with a set of problems that have to be solved in order to do so. In Section 3 we present our main theoretical contribution, the generalizations of the balance equation and the entropy triangle for multivariate distributions, and in Section 4 we introduce examples of uses for these tools for the exploratory analysis of machine learning tasks, both supervised and unsupervised. We end with a brief discussion of alternate representation mechanisms for entropy balances, the uses of such tools and some conclusions.

2. Methods and tools

2.1. The joint entropy balance of two variables

The tools we propose are based on an often overlooked decomposition of the joint entropy of two random variables (Valverde-Albacete & Peláez-Moreno, 2010). Fig. 3 depicts this decomposition showing the three crucial regions:

- The *divergence with respect to uniformity*, $\Delta H_{P_X \cdot P_Y}$, between the joint distribution where P_X and P_Y are independent and the uniform distributions U_X and U_Y with the same cardinality of events as P_X and P_Y.

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

- The *mutual information*, $MI_{P_{XY}}$, quantifies the force of the stochastic binding between P_X and P_Y.

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

- The *variation of information*, $VI_{P_{XY}}$, embodies the residual entropy, not used in binding the variables.

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

Each of these quantities provide intuitions into the behavior of P_X, P_Y and P_{XY} used to advantage in applications (cfr. Section 2.3), and we would like to reproduce them in a multivariate setting for applications like feature filtering (Brown, Pocock, Zhao, & Luján, 2012) or multi-label classification (Gibaja & Ventura, 2015).

Note that all of these quantities are positive. In fact from the previous decomposition the following *balance equation* is evident,

$$\begin{aligned} H_{U_X \cdot U_Y} &= \Delta H_{P_X \cdot P_Y} + 2 \cdot MI_{P_{XY}} + VI_{P_{XY}} \\ 0 &\leq \Delta H_{P_X \cdot P_Y}, MI_{P_{XY}}, VI_{P_{XY}} \leq H_{U_X \cdot U_Y} \end{aligned} \tag{1}$$

where the bounds are easily obtained from distributional considerations (Valverde-Albacete & Peláez-Moreno, 2010).

2.2. From the balance equation to the joint entropy triangle

If we normalize (1) by the overall entropy $H_{U_X \cdot U_Y}$ we obtain

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

Eq. (2) is the 2-simplex in normalized $\Delta H'_{P_X \cdot P_Y} \times 2MI'_{P_{XY}} \times VI'_{P_{XY}}$ space. Each joint distribution P_{XY} can be characterized by its joint entropy proportions, or entropic *composition* (Aitchison, 1982; Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015) $F(P_{XY}) = [\Delta H'_{P_{XY}}, 2 \cdot MI'_{P_{XY}}, VI'_{P_{XY}}]$. Its projection onto the plane with director vector (1, 1, 1) is its *de Finetti (entropy) diagram*, represented in Fig. 4 which shows as an equilateral triangle, hence the alternative name of *entropy triangle*.

Therefore, *every binary distribution shows as a point in the triangle and the position in the triangle entails qualities of the distribution:*

- The lower side of the triangle is the geometric locus of distributions with independent marginals: if $P_{XY} = P_X \cdot P_Y$ then $F(P_{XY}) = [\cdot, 0, \cdot]$.
- The left side is the geometric locus of distributions with uniform marginals. If $P_X = U_X$ and $P_Y = U_Y$ then $F(P_{XY}) = [0, \cdot, \cdot]$.
- Finally, the right-hand side is the locus of distributions with identical marginals: if $P_X = P_Y$—that is, $H_{P_X} = H_{P_Y} = MI_{P_{XY}}$—then $F(P_{XY}) = [\cdot, \cdot, 0]$.