Journal of Mathematical Psychology 67 (2015) 39–44

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Notes and comment

A metric for measurable partitions

Erik Mohlin *

Nuffield College and Department of Economics, University of Oxford, United Kingdom

HIGHLIGHTS

- A new metric is defined, which measures the distance between partitions.
- The metric applies to any measurable set and countably many cells/blocks.
- With finite sets, and the counting measure, it yields twice the pairs of disagreement.
- The metric is applied to the study of natural language colour categorizations.

ARTICLE INFO

Article history:
Received 1 February 2015
Received in revised form
2 May 2015
Available online 7 July 2015

Keywords:
Partitions
Categorization
Decision trees
Classification trees
Clustering
Colour categorization
World color survey

ABSTRACT

Given a set and a measure on it, this note proposes a metric defined on sets of measurable partitions, with countably many cells. For the special case of partitions of a finite set, and the counting measure, the proposed metric coincides with twice the number of pairs of disagreement. For the special case of a probability measure the proposed metric is related to a symmetrized version of conditional information. The proposed metric is applied to the study of natural language colour categorizations: it allows for quantification of the relative importance of universal and culture-specific forces in shaping colour categorizations.

1. Introduction

This paper proposes a metric that measures the distance between two partitions, P and Q, of a set X. The set X may be infinite, and the number of cells in the partitions is allowed to be countably infinite. The partitions P and Q are required to be measurable.

According to the proposed metric the distance between P and Q is a weighted average of the measure of the symmetric differences between pairs of cells (P_i, Q_j), where cell P_i is from P and cell Q_j is from Q. For each such pair (P_i, Q_j), the measure of their symmetric difference $P_i \Delta Q_j$, is weighted by the measure of their intersection $P_i \cap Q_j$.

For the case of a finite set X, a number of metrics or 'similarity indices' have been proposed in the literature on classification and cluster analysis (Denœud & Guénoche, 2006). In particular, a number of metrics and similarity indices have been constructed on the basis of counting the pairs of elements of X, on which the two partitions agree (Johnson, 1968; Mirkin & Cherny, 1970; Rand, 1971), or disagree (Arabie & Boorman, 1973): Two partitions of a finite set disagree on a pair of objects if one of the partitions puts the two objects in the same cell whereas the other partition puts the two objects in different cells. In the context of decision trees, López De Mántaras (1991) defines a metric, which amounts to a symmetrized version of conditional information. It can in principle be applied to measurable partitions of an infinite set X. It has been used in cluster analysis and decision tree algorithms.

When X is finite and the counting measure is used, the metric proposed in this paper reduces to twice the number of "pairs of disagreement" (Section 3.1). When a probability measure is used and partitions are finite (and X may be finite or infinite), the proposed metric is structurally similar to the metric of López de Mántaras (Section 3.2).

A metric such as the one proposed in this paper has a number of applications: It can be employed in the construction of decision trees or classification trees (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1986; Ripley, 1996). Specifically one may use the metric to evaluate, and select among, different trees, by comparing the partitions they induce, and measuring the distance between

* Correspondence to: Nuffield College, New Road, Oxford OX1 4PX, United Kingdom.
E-mail address: erik.mohlin@nuffield.ox.ac.uk.

http://dx.doi.org/10.1016/j.jmp.2015.06.003
0022-2496/

the induced partitions and the correct partition of a training or validation sample.

In game theory and economics it is frequently assumed that people share a common prior defined on a shared state space. Differences in the agents' information can be represented by assuming that they have different partitions of the state space. The proposed metric can used be to quantify such informational differences.[1]

In the final section of the paper (Section 4) the proposed metric is applied to the study of the psychology of colour categorization. The metric can be used to measure how far observed natural language colour categorizations are from theoretically predicted categorizations. It can also be employed to quantify the relative importance of universal and culturally relative forces in shaping the way colours are categorized.

2. The proposed metric

Suppose X is a (possibly uncountable) set, Σ a sigma-algebra on X, and μ a finite measure defined on Σ. A countable and measurable partition P is a countable collection of measurable sets, which are jointly exhaustive and mutually exclusive. That is, $P = \{P_i\}_{i\in I_P}$, where I_P is a countable index set, and $P_i \in \Sigma$ holds for all $i \in I_P$. A set $P_i \in P$ is called a cell (or a block) of the partition P.[2] Let $\mathcal{P}$ be the set of measurable and countable partitions of X. Note that if $P = \{P_i\}_{i\in I_P}$ and $Q = \{Q_i\}_{i\in I_Q}$ are measurable and have countably many cells, then their join (coarsest common refinement) is also measurable and has countably many cells. The symmetric difference of two sets P_i and Q_j is $P_i \Delta Q_j = \left(P_i \cup Q_j\right) \setminus \left(P_i \cap Q_j\right)$. I propose the following metric:

Definition 1. Let $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ be defined by

$$d(P, Q) = \sum_{i\in I_P} \sum_{j\in I_Q} \mu\left[P_i \cap Q_j\right] \mu\left[P_i \Delta Q_j\right]. \tag{1}$$

Note that

$$\begin{aligned}\mu\left[P_i \Delta Q_j\right] &= \mu\left[P_i \cup Q_j\right] - \mu\left[P_i \cap Q_j\right] \\ &= \mu\left[P_i\right] + \mu\left[Q_j\right] - 2\mu\left[\left(P_i \cap Q_j\right)\right]. \end{aligned} \tag{2}$$

Thus, for each cell $P_i \cap Q_j$ in the join of P and Q, the metric sums the measure of the sets $P_i \setminus Q_j$ and $Q_j \setminus P_i$, and weighs it by the measure of the cell $P_i \cap Q_j$. The set $P_i \setminus Q_j$ consists of the points which share the same cell as $P_i \cap Q_j$ under P, but do not share the same cell as $P_i \cap Q_j$ under Q. Similarly $Q_j \setminus P_i$ consists of the points which share the same cell as $P_i \cap Q_j$ under Q, but do not share the same cell as $P_i \cap Q_j$ under P. This is illustrated in Fig. 1.

It remains to verify that the suggested metric is indeed a metric. However, in order to do that the concept of identity for partitions has to be discussed. Let $P(x)$ be the cell of partition P that contains element $x \in X$. Two partitions P and Q coincide on $x \in X$ if $P(x) = Q(x)$. Two partitions P and Q are identical if they coincide on all $x \in X$. Intuitively we may consider two partitions as essentially the same as long as they coincide on almost all $x \in X$, i.e. on all except

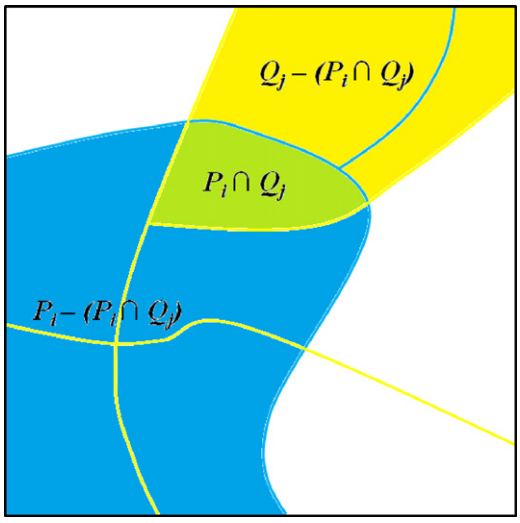

Fig. 1. The green area represents a cell $P_i \cap Q_j$ in the join of P (blue lines) and Q (yellow lines). The blue and yellow areas represent $P_i \setminus \left(P_i \cap Q_j\right)$ and $Q_j \setminus \left(P_i \cap Q_j\right)$, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a measure zero set of points. For this reason we divide the set of partitions $\mathcal{P}$ into equivalence classes.

Definition 2. Two partitions $P, Q \in \mathcal{P}$ are equivalent under μ, written $P \sim_\mu Q$, if for all $P_i \in P$ and $Q_j \in Q$,

$$\mu\left[P_i \cap Q_j\right] \neq 0 \Rightarrow \mu\left[P_i \Delta Q_j\right] = 0. \tag{3}$$

The quotient space, i.e. the set of equivalence classes under $\sim_\mu$, is denoted $\mathcal{P}/\sim_\mu$.

Note that if all cells in the join of P and Q have positive measure then $P \sim_\mu Q$ if and only if $P = Q$.

Proposition 1. *The function d is a metric on $\mathcal{P}/\sim_\mu$: For all partitions $P, Q, R \in \mathcal{P}/\sim_\mu$ the function d satisfies*

(a) $d(P, Q) = 0$ *if and only if* $P \sim_\mu Q$,
(b) $d(P, Q) = d(Q, P)$, *and*
(c) $d(P, Q) \leq d(P, R) + d(R, Q)$.[3]

Proof. (a) If $\mu\left[P_i \cap Q_j\right] \neq 0$ then $P \sim_\mu Q$ implies $\mu\left[P_i \Delta Q_j\right] = 0$. Thus, for each P_i and Q_j we have either $\mu\left[P_i \cap Q_j\right] = 0$ or $\mu\left[P_i \Delta Q_j\right] = 0$, implying that $d(P, Q) = 0$.

To prove the converse suppose that $d(P, Q) = 0$. Showing that two arbitrary cells P_i and Q_j satisfy (3) is equivalent to showing that $\mu\left[P_i \Delta Q_j\right] \neq 0$ implies $\mu\left[P_i \cap Q_j\right] = 0$. To see that this holds note that if it was the case that $\mu\left[P_i \Delta Q_j\right] \neq 0$ and $\mu\left[P_i \cap Q_j\right] \neq 0$, then the assumption that $d(P, Q) = 0$ would be violated.

(b) Symmetry is evident from Definition 1.

(c) Using $\sum_{k\in I_R} \mu\left[P_i \cap Q_j \cap R_k\right] = \mu\left[P_i \cap Q_j\right]$ we have

$$\begin{aligned} d(P, Q) &= \sum_{i\in I_P} \sum_{j\in I_Q} \mu\left[P_i \cap Q_j\right] \left(\mu\left[P_i\right] + \mu\left[Q_j\right] - 2\mu\left[\left(P_i \cap Q_j\right)\right]\right) \\ &= \sum_{i\in I_P} \sum_{j\in I_Q} \sum_{k\in I_R} \mu\left[P_i \cap Q_j \cap R_k\right] \\ &\quad \times \left(\mu\left[P_i\right] + \mu\left[Q_j\right] - 2\mu\left[\left(P_i \cap Q_j\right)\right]\right). \end{aligned}$$

[1] Metrics defined on sub-sigma-fields, such as the metric of Boylan (1971), have been employed by Allen (1983) and others to measure differences between information structures. As long as information is represented by countable partitions, the metric proposed in this paper is arguably both simpler to apply and more intuitive, than the metrics for sub-sigma-fields.

[2] Of course, the cells of the partition being jointly exhaustive means that $\cup_{i\in I_P} P_i = X$. The cells of the partition being mutually exclusive means that $P_i \cap P_j = \emptyset$ for all $i, j \in I_P$, such that $i \neq j$.

[3] An ultrametric is a metric that satisfies $d(P, Q) \leq max\{d(P, R), d(R, Q)\}$. It is easy to construct examples to the effect that this inequality is not satisfied for the metric defined in this paper.