SCORING DISTANCES BETWEEN EQUIVALENCE AND PREFERENCE RELATIONS

Boris Mirkin[1,2]

[1]Department of Data Analysis and Artificial Intelligence, National Research University Higher School of Economics, Moscow (e-mail: bmirkin@hse.ru)
[2]Department of Computer Science, Birkbeck University of London (e-mail: mirkin@dcs.bbk.ac.uk)

ABSTRACT: This paper explores association between the notions of similarity and preference by using the framework of the theory of binary relations considered as subsets of the cartesian product of the set of objects by itself. Unordered partitions correspond to the so-called equivalence relations, and ordered partitions, to the so-called weak order relations. We derive a number of properties of the metric space of equivalence and weak order relations. One of them is establishing of the fact that the so-called Kemeny distance between tied rankings is identical to the mismatch distance between corresponding binary relations of weak order.

KEYWORDS: equivalence relation, weak order, distance, contingency table, consensus

1 Introduction

The notions of similarity and preference are usually considered quite different. The former is expressed via the concept of partition, a set of non-overlapping subsets containing "similar" objects, so that different subsets contain 'dissimilar" objects. The latter is expressed via the concept of ordering or, more generally, ordered partition. It is assumed that objects belonging to one part preceding another part are in some sense better than those in this other part. In this sense objects belonging to the same part of an ordered partition are "similar". This association between the notions of similarity and preference can be further elaborated by using the framework of the theory of binary relations considered as subsets of the cartesian product of the set of objects by itself. Unordered partitions correspond to the so-called equivalence relations, and ordered partitions, to the so-called weak order relations. We consider the metric space of binary relations with respect to the so-called matching distance, which is the size of the symmetric difference between relations as subsets of ordered pairs of objects. This allows us to consider both equivalence and weak order relations as part of this metric space and to mathematically explore the separate subspace of equivalence relations and subspace of weak order relations, as well as affinities between these subspaces.

2 Main results

This talk will describe results found within this approach (see also [Mirkin 1979, 2012], Mirkin, Fenner [2019]). Among them are the following.

1. We attend to the Kemeny approach for finding consensus rankings as those minimizing the summary distance to those presented. Here we prove that the Kemeny distance [Kemeny 1959] between rankings is, in fact, the mismatch distance between the corresponding weak-order binary relations. The importance of this result stems from the fact that the former involves Kendall object-to-object matrices with three possible values for the entries: 1 for preceding, -1 for following, and 0 for a tie; whereas the latter involves only two: 1 for the presence and 0 for the absence of a pair in the binary relation [Kendall 1938]. In contrast, the distance between relations involves only 0 (no relation) and 1 (there is relation), with no negative values at all which appear not necessary, in contrast to common sense.

2. We present an explicit statement expressing the Kemeny consensus criterion in terms of the relational consensus matrix, analogous to the so-called consensus matrix in the problem of consensus clustering [Mirkin 2012]. In contrast to the analysis of consensus clustering, however, the (i, j) entry in this consensus matrix is not simply the number of partitions for which elements i and j belong to the same part, but also includes the number of rankings for which i precedes j. The problem, which involves the subtraction of a threshold, is equivalent to maximizing the sum of the consensus matrix entries minus the number of pairs in the corresponding equivalence relation (sometimes referred to as the partition concentration index), weighted with a penalty defined by the threshold. The subtracted part plays the role of a naturally emerging regularizer. The regularizer plays no role, though, when the solution is restricted to a class of ranked partitions like the class of linear rankings with no ties.
3. We test the sensitivity of the Kemeny median concept by applying what we call *Muchnik test* (see [Mirkin 2012] for the case of unordered partitions) to ordered partitions. Specifically, we apply the concept of median to the *Likert scales* popular in Psychology [Likert 1932]. Given an ordered partition $\boldsymbol{R} = (R_1, R_2, \dots, R_p)$, the Likert scale replaces $\boldsymbol{R}$ by the set of binary ordered partitions $\boldsymbol{S}^t$ (t = 1, 2, … , p-1) that separate the union of the first t parts of $\boldsymbol{R}$ from the rest. The question then arises as to whether $\boldsymbol{R}$ is a median for the set of binary rankings $\boldsymbol{S}^t$ (t=1, 2, … , p-1), as one might expect, or not. Perhaps surprisingly, it turns out that it is one of the "coarse" binary rankings $\boldsymbol{S}^t$ that is a median, rather than $\boldsymbol{R}$ itself.
4. We derive explicit formulas for the distance, especially those regarding the relationship between weak orders and their induced equivalence relations, using the ternary relation "between" on the set of binary relations and the notion of "refinement" on the set of tied rankings, as well as the notion of contingency table from statistics. For example, we prove that the mismatch distance between ordered partitions $\boldsymbol{R}$ and $\boldsymbol{R'}$ can be decomposed into ranking and equivalence parts:

$$d(\boldsymbol{R}, \boldsymbol{R'}) = \frac{1}{2} d(\boldsymbol{E}, \boldsymbol{E'}) + d(\boldsymbol{R}*\boldsymbol{R'}, \boldsymbol{R'}*\boldsymbol{R}).$$

where $\boldsymbol{E}, \boldsymbol{E'}$ are equivalence relations corresponding to unordered partitions in $\boldsymbol{R}, \boldsymbol{R'}$and the star $*$ denotes the operation of lexicographic product of two ordered partitions [Mirkin 1979]. The distance between $\boldsymbol{R}*\boldsymbol{R'}$ and $\boldsymbol{R'}*\boldsymbol{R}$ is equal to half of the total of the products of the cardinalities of those parts in the intersection $\boldsymbol{R}\cap\boldsymbol{R'}$ for which the orderings in $\boldsymbol{R}$ and $\boldsymbol{R'}$ are contradictory:

$$d(\boldsymbol{R}* \boldsymbol{R'}, \boldsymbol{R'}*\boldsymbol{R}) = \frac{1}{2}\sum_{s>s'} \sum_{t<t'} N_{st} N_{s't'}$$

Considering the rankings $\boldsymbol{R}$ and $\boldsymbol{R'}$ as unordered partitions, denoted above by $\check{\boldsymbol{R}}$ and $\check{\boldsymbol{R}}'$, respectively, the mismatch distance between the corresponding equivalence relations, $\boldsymbol{E}$ and $\boldsymbol{E'}$, can be expressed as

$$d(\boldsymbol{E}, \boldsymbol{E'}) = \sum_s N_s^2 + \sum_t N'^2_t - 2\sum_{s,t} N_{st}^2$$

where N_s, N'_t , and N_{st} are, as above, the numbers of elements in parts R_s of $\boldsymbol{R}$, R'_t of $\boldsymbol{R'}$ and $R_s\cap R'_t$ of $\boldsymbol{R}\cap\boldsymbol{R'}$, respectively.

3 Conclusion

This shows that, in fact, there is no common ground to simultaneously consider weak orders and equivalence relations, because the lexicographic products are items added to

distances between equivalence relations, which are absent from unordered partitions. Therefore, further advances along the path based on the distance can be made within each ordered partitions (rankings) and unordered partitions, but not in between. Among possible directions for further research, the following two seem quite straightforward. First is the task of numerically solving the problem of consensus ranking by extending the problem of consensus ordering [Charon and Hudry 2007]. For example, the additive structure of the criterion suggests that one might first find an optimal linear ordering and then aggregate some of its parts to form a tied ranking. Second, the failure of the Muchnik test on Likert scales suggests that new ways for formulating more sensitive criteria for consensus are needed.

References

ALESKEROV, F. & MONJARDET, B. 2002. *Utility Maximization, Choice, and Preference*, Stud. Econ. Theory 16, Springer-Verlag, Berlin.

BARTHELEMY, J.P., LECLERC, B. & MONJARDET, B. 1986. On the use of ordered sets in problems of comparison and consensus of classifications. *Journal of Classification*, ***3***(2), 187-224.

CHARON, I. & HUDRY, O. 2007. A survey on the linear ordering problem for weighted or unweighted tournaments. *4OR: A Quarterly Journal of Operations Research*, ***5***(1), 5-60.

DIACONIS, F., & GRAHAM, R. 1977. Spearman's footrule as a measure of disarray, *Journal of Royal Statistics Society*, Series B, **39**, 262-268.

KEMENY, J.G. 1959. Mathematics without numbers, *Daedalus,* **88**, No. 4, *Quantity and Quality*, 577-591.

KENDALL, M.G. 1938. A new measure of rank correlation, *Biometrika*, **30**, 81-93.

LIKERT, R. 1932. A technique for the measurement of attitudes, *Archives of Psychology*, **22(140)**, 55.

MIRKIN, B. 1979. *Group Choice*. Halsted Press: Washington DC. (Translated from Russian *"Problema Gruppovogo Vybora"*, Nauka Physics-Mathematics, Moscow, 1974.)

MIRKIN, B. 2012. *Clustering: A Data Recovery Approach*, CRC Press, Boka-Raton Fl.

MIRKIN, B. & FENNER, T. I. 2019. Distance and consensus for preference relations corresponding to ordered partitions. *Journal of Classification*, **36**, 350-367.

STEELE, K. & STEFANSSON, H. O. 2015. Decision Theory, *The Stanford Encyclopedia of Philosophy.* <http://plato.stanford.edu/archives/win2015/entries/decision-theory/>.