A note on the Jaccardized Czekanowski similarity index

András Schubert[1], András Telcs[2,3]

[1] Department of Science Policy and Scientometrics, Library and Information Center of the Hungarian Academy of Sciences, Budapest, Hungary
[2] Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest, Hungary
[3] Department of Quantitative Methods, Faculty of Economics, Pannon University, Veszprém, Hungary

Abstract

It is shown that the "Jaccardized Czekanowski index" is actually a reinterpretation of the Ružička index. Thereby, it is proved that its one-complement is a true distance function, which makes it particularly suitable for use in similarity studies even with multidimensional statistical techniques.

Recently, Schubert (2013) introduced the "Jaccardized Czekanowski index" as a measure of similarity between two categorical distributions. It was derived from the Czekanowski index (Czekanowski, 1909)

$$Cz_{A,B} = \sum_i \min(q_i^A, q_i^B) \quad ,$$

where q_i^A and q_i^B are the relative frequencies of category i in the distributions A and B, respectively. This index was reinvented several times, and is known also as Bray–Curtis index (Bray & Curtis, 1957) in the ecological literature or as Finger–Kreinin index (Finger & Kreinin, 1979) in economics.
The Czekanowski index can be regarded as the "quantitative" (i.e., abundance dependent) version of the incidence-based (binary) Sørensen–Dice index (Dice, 1945; Sørensen, 1948)

$$SD_{A,B} = 2|A \cap B|/[|A \cup B| + |A \cap B|] \quad ,$$

where $|A \cap B|$ is the number of non-empty categories in the intersection of distributions A and B while $|A \cup B|$ is that in their union, if we consider the following formulas for the two indices:

$$SD_{A,B} = 1 - \Sigma_i |\delta_i^A - \delta_i^B| / \Sigma_i (\delta_i^A + \delta_i^B),$$

(here δ_i^A and δ_i^B take the value of 0 or 1 depending whether the i-th category is empty or non-empty in distributions A and B, respectively) and

$$Cz_{A,B} = 1 - \Sigma_i |q_i^A - q_i^B| / \Sigma_i (q_i^A + q_i^B) = 1 - (1/2) \Sigma_i |q_i^A - q_i^B|,$$

(here q_i^A and q_i^B are the relative frequencies of category i in the distributions A and B, respectively).
Given the relation

$$Ja_{A,B} = SD_{A,B}/(2 - SD_{A,B})$$

between the Sørensen–Dice index and the Jaccard index (Jaccard, 1901), the progenitor of all similarity measures, defined as

$$Ja_{A,B} = |A \cap B|/|A \cup B|$$

(notations as above at the definition of the Sørensen-Dice index),
The "Jaccardization" of the Czekanowski index thus resulted in

$$JCz_{A,B} = Cz_{A,B}/(2 - Cz_{A,B}) \quad .$$

Since JCz is a monotonic function of Cz, the two indices are rank-equivalent. Schubert (2013) found JCz superior over Cz that in the empirical samples studied its range was wider and its distribution was more uniform, thereby its distinguishing power was stronger, and it correlated fairly well (much better that Cz) with several other similarity measures (such as the chi-squared and cosine measures) rather uncorrelated among others.
Another aspect that might influence the choice of the similarity index is its metric properties. The one-complement of a similarity index (1–similarity index; also called dissimilarity index) may be metric (zero for identical elements, positive for different elements, symmetric, the triangle inequality holds), semi-metric (without the triangle-inequality condition) or non-metric (neither metric nor semi-metric). In the metric case, the dissimilarity index is a true distance function. In classical similarity/dissimilarity studies, metric indices have no specific advantages over metric ones, but with the emergence of multivariate statistical techniques (particularly, cluster analysis) true distance functions are preferred, if not required.

Actually, the one-complement of the Jaccard index is metric, that of the Sørensen–Dice and the Czekanowski index is semi-metric.
A similarity index the one-complement of which is known to be a true distance function is the Ružička index (Ružička, 1958, see also Pielou, 1984):

$$Ru_{A,B} = \sum_i \min(q_i^A, q_i^B) / \sum_i \max(q_i^A, q_i^B)$$

The metric properties of the one-complement of the Ružička index were proved by Levandowsky & Winter (1971) and Gilbert (1972).
We show now that the Jaccardized Czekanowski index is actually identical with the Ružička index.

$$JCz_{A,B} = Cz_{A,B}/(2-Cz_{A,B})$$
$$= \{2[\sum_i \min(q_i^A, q_i^B)/\sum_i (q_i^A+q_i^B)]\sum_i (q_i^A+q_i^B)\}/2[\sum_i (q_i^A+q_i^B) - \sum_i \min(q_i^A, q_i^B)]$$
$$= \sum_i \min(q_i^A, q_i^B)/[\sum_i (q_i^A+q_i^B) - \sum_i \min(q_i^A, q_i^B)]$$
$$= \sum_i \min(q_i^A, q_i^B)/\sum_i \max(q_i^A, q_i^B)$$
$$= Ru_{A,B} \quad .$$

In conclusion, the Jaccardized Czekanowski index of Schubert (2013) is not a new index but a reinterpretation of the Ružička index (as we could see, such reinterpretations or even reinventions are not infrequent in the history of similarity indices). As such, however, it was proved to lead to a true distance function as its one-complement, which grants its enhanced suitability in similarity studies even with multidimensional statistical techniques.

Acknowledgement

A. Sch. was supported by the European Commission under the FP7 Science in Society Grant No. 266588 (SISOB project).
A. T. was supported by TÁMOP-4.2.2/B-10/1-2010-0025

References

Bray, J.R., Curtis, J.T. (1957) An Ordination of the Upland. Forest Communities of Southern Wisconsin. Ecological Monographs, 27: 325–349.
Czekanowski, J. (1909) Zur differential Diagnose der Neandertalgruppe. Korrespondenzblatt der deutschen Gesellschaft für Anthropologie, Ethnologie und Urgeschichte, 40, 44–47.
Dice, L.R. (1945) Measures of the Amount of Ecologic Association Between Species. Ecology 26(3): 297–302.
Finger, J., Kreinin, M. (1979) A measure of 'export similarity' and its possible uses. The Economic Journal, 89: 905–912
Gilbert, G. (1972) Distance between sets. Nature 239: 174.
Jaccard, P. (1901), Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles, 37: 547–579
Levandowsky, M., Winter, D. (1971) Distance between sets. Nature, 234: 34–35.
Pielou, E.C. (1984) The interpretation of ecological data. A primer on classification and ordination. John Wiley&Sons, Inc., New York. 263 p.
Ružička, M. (1958) Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen). Biológia, Bratislava, 13: 647–661.
Schubert, A. (2013) Measuring the similarity between the reference and citation distributions of journals. Scientometrics, 96(1) DOI 10.1007/s11192-012-0889-0
Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. Kongelige Danske Videnskabernes Selskab. Biologiske Skrifter. 4: 1–34.