The Journal of Open Source Software

Philentropy: Information Theory and Distance Quantification with R

Hajk-Georg Drost[1]

1 The Sainsbury Laboratory, University of Cambridge, Bateman Street, Cambridge CB2 1LR, UK

DOI: 10.21105/joss.00752

Software

- Review
- Repository
- Archive

Submitted: 23 May 2018
Published: 26 May 2018

Summary

Comparison is a fundamental method of scientific research leading to insights about the processes that generate similarity or dissimilarity. In statistical terms comparisons between probability functions are performed to infer connections, correlations, or relationships between objects or samples (Cha 2007). Most quantification methods rely on distance or similarity measures, but the right choice for each individual application is not always clear and sometimes poorly explored. The reason for this is partly that diverse measures are either implemented in different R packages with very different notations or are not implemented at all. Thus, a comprehensive framework implementing the most common similarity and distance measures using a uniform notation is still missing. The R (R Core Team 2018) package `Philentropy` aims to fill this gap by implementing forty-six fundamental distance and similarity measures (Cha 2007) for comparing probability functions. These comparisons between probability functions have their foundations in a broad range of scientific disciplines from mathematics to ecology. The aim of this package is to provide a comprehensive and computationally optimized base framework for clustering, classification, statistical inference, goodness-of-fit, non-parametric statistics, information theory, and machine learning tasks that are based on comparing univariate or multivariate probability functions. All functions are written in C++ and are integrated into the R package using the Rcpp Application Programming Interface (API) (Eddelbuettel 2013).

Together, this framework allows building new similarity or distance based (statistical) models and algorithms in R which are computationally efficient and scalable. The comprehensive availability of diverse metrics and measures furthermore enables a systematic assessment of choosing the most optimal similarity or distance measure for individual applications in diverse scientific disciplines.

The following probability distance/similarity and information theory measures are implemented in `Philentropy`.

Distance and Similarity Measures

L_p Minkowski Family

- Euclidean : $d = \sqrt{\sum_{i=1}^{N} |P_i - Q_i|^2)}$
- Manhattan : $d = \sum_{i=1}^{N} |P_i - Q_i|$
- Minkowski : $d = (\sum_{i=1}^{N} |P_i - Q_i|^p)^{1/p}$
- Chebyshev : $d = max|P_i - Q_i|$

L_1 Family

- Sorensen : $d = \frac{\sum_{i=1}^{N}|P_i - Q_i|}{\sum_{i=1}^{N}(P_i + Q_i)}$
- Gower : $d = \frac{1}{N}\sum_{i=1}^{N}|P_i - Q_i|$, where N is the total number of elements i in P_i and Q_i
- Soergel : $d = \frac{\sum_{i=1}^{N}|P_i - Q_i|}{\sum_{i=1}^{N} max(P_i, Q_i)}$
- Kulczynski d : $d = \frac{\sum_{i=1}^{N}|P_i - Q_i|}{\sum_{i=1}^{N} min(P_i, Q_i)}$
- Canberra : $d = \frac{\sum_{i=1}^{N}|P_i - Q_i|}{(P_i + Q_i)}$
- Lorentzian : $d = \sum_{i=1}^{N} ln(1 + |P_i - Q_i|)$

Intersection Family

- Intersection : $s = \sum_{i=1}^{N} min(P_i, Q_i)$
- Non-Intersection : $d = 1 - \sum_{i=1}^{N} min(P_i, Q_i)$
- Wave Hedges : $d = \frac{\sum_{i=1}^{N}|P_i - Q_i|}{max(P_i, Q_i)}$
- Czekanowski : $d = \frac{\sum_{i=1}^{N}|P_i - Q_i|}{\sum_{i=1}^{N}|P_i + Q_i|}$
- Motyka : $d = \frac{\sum_{i=1}^{N} min(P_i, Q_i)}{(P_i + Q_i)}$
- Kulczynski s : $d = \frac{\sum_{i=1}^{N} min(P_i, Q_i)}{\sum_{i=1}^{N}|P_i - Q_i|}$
- Tanimoto : $d = \frac{\sum_{i=1}^{N}(max(P_i, Q_i) - min(P_i, Q_i))}{\sum_{i=1}^{N} max(P_i, Q_i)}$; equivalent to Soergel
- Ruzicka : $s = \frac{\sum_{i=1}^{N} min(P_i, Q_i)}{\sum_{i=1}^{N} max(P_i, Q_i)}$; equivalent to 1 - Tanimoto = 1 - Soergel

Inner Product Family

- Inner Product : $s = \sum_{i=1}^{N} P_i \dot{Q}_i$
- Harmonic mean : $s = 2 \cdot \frac{\sum_{i=1}^{N} P_i \cdot Q_i}{P_i + Q_i}$
- Cosine : $s = \frac{\sum_{i=1}^{N} P_i \cdot Q_i}{\sqrt{\sum_{i=1}^{N} P_i^2} \cdot \sqrt{\sum_{i=1}^{N} Q_i^2}}$
- Kumar-Hassebrook (PCE) : $s = \frac{\sum_{i=1}^{N}(P_i \cdot Q_i)}{(\sum_{i=1}^{N} P_i^2 + \sum_{i=1}^{N} Q_i^2 - \sum_{i=1}^{N}(P_i \cdot Q_i))}$
- Jaccard : $d = 1 - \frac{\sum_{i=1}^{N} P_i \cdot Q_i}{\sum_{i=1}^{N} P_i^2 + \sum_{i=1}^{N} Q_i^2 - \sum_{i=1}^{N} P_i \cdot Q_i}$; equivalent to 1 - Kumar-Hassebrook
- Dice : $d = \frac{\sum_{i=1}^{N}(P_i - Q_i)^2}{(\sum_{i=1}^{N} P_i^2 + \sum_{i=1}^{N} Q_i^2)}$

Squared-chord Family

- Fidelity : $s = \sum_{i=1}^{N} \sqrt{P_i \cdot Q_i}$
- Bhattacharyya : $d = -ln \sum_{i=1}^{N} \sqrt{P_i \cdot Q_i}$
- Hellinger : $d = 2 \cdot \sqrt{1 - \sum_{i=1}^{N} \sqrt{P_i \cdot Q_i}}$
- Matusita : $d = \sqrt{2 - 2 \cdot \sum_{i=1}^{N} \sqrt{P_i \cdot Q_i}}$
- Squared-chord : $d = \sum_{i=1}^{N} (\sqrt{P_i} - \sqrt{Q_i})^2$

Squared L_2 family (X^2 squared family)

- Squared Euclidean : $d = \sum_{i=1}^{N}(P_i - Q_i)^2$
- Pearson X^2 : $d = \sum_{i=1}^{N}(\frac{(P_i-Q_i)^2}{Q_i})$
- Neyman X^2 : $d = \sum_{i=1}^{N}(\frac{(P_i-Q_i)^2}{P_i})$
- Squared X^2 : $d = \sum_{i=1}^{N}(\frac{(P_i-Q_i)^2}{(P_i+Q_i)})$
- Probabilistic Symmetric X^2 : $d = 2 \cdot \sum_{i=1}^{N}(\frac{(P_i-Q_i)^2}{(P_i+Q_i)})$
- Divergence : X^2 : $d = 2 \cdot \sum_{i=1}^{N}(\frac{(P_i-Q_i)^2}{(P_i+Q_i)^2})$
- Clark : $d = \sqrt{\sum_{i=1}^{N}(\frac{|P_i-Q_i|}{(P_i+Q_i)^2}}$
- Additive Symmetric X^2 : $d = \sum_{i=1}^{N}(\frac{((P_i-Q_i)^2 \cdot (P_i+Q_i))}{(P_i \cdot Q_i)})$

Shannon's Entropy Family

- Kullback-Leibler : $d = \sum_{i=1}^{N} P_i \cdot log(\frac{P_i}{Q_i})$
- Jeffreys : $d = \sum_{i=1}^{N}(P_i - Q_i) \cdot log(\frac{P_i}{Q_i})$
- K divergence : $d = \sum_{i=1}^{N} P_i \cdot log(\frac{2 \cdot P_i}{P_i+Q_i})$
- Topsoe : $d = \sum_{i=1}^{N}(P_i \cdot log(\frac{2 \cdot P_i}{P_i+Q_i})) + (Q_i \cdot log(\frac{2 \cdot Q_i}{P_i+Q_i}))$
- Jensen-Shannon : $d = 0.5 \cdot (\sum_{i=1}^{N} P_i \cdot log(\frac{2 \cdot P_i}{P_i+Q_i}) + \sum_{i=1}^{N} Q_i \cdot log(\frac{2*Q_i}{P_i+Q_i}))$
- Jensen difference : $d = \sum_{i=1}^{N}((\frac{P_i \cdot log(P_i)+Q_i \cdot log(Q_i)}{2}) - (\frac{P_i+Q_i}{2}) \cdot log(\frac{P_i+Q_i}{2}))$

Combinations

- Taneja : $d = \sum_{i=1}^{N}(\frac{P_i+Q_i}{2}) \cdot log(\frac{P_i+Q_i}{(2 \cdot \sqrt{P_i \cdot Q_i})})$
- Kumar-Johnson : $d = \sum_{i=1}^{N} \frac{(P_i^2-Q_i^2)^2}{2 \cdot (P_i \cdot Q_i)^{\frac{3}{2}}}$
- Avg(L_1, L_n) : $d = \frac{\sum_{i=1}^{N}|P_i-Q_i|+max|P_i-Q_i|}{2}$

Note: d refers to distance measures, whereas s denotes similarity measures.

Information Theory Measures

- Shannon's Entropy H(X) : $H(X) = -\sum_{i=1}^{n} P(x_i) \cdot log_b(P(x_i))$
- Shannon's Joint-Entropy H(X,Y) : $H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j) \cdot log_b(P(x_i,y_j))$
- Shannon's Conditional-Entropy H(X | Y) : $H(Y|X) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j) \cdot log_b(\frac{P(x_i)}{P(x_i,y_j)})$
- Mutual Information I(X,Y) : $MI(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j) \cdot log_b(\frac{P(x_i,y_j)}{(P(x_i)*P(y_j))})$
- Kullback-Leibler Divergence : $KL(P||Q) = \sum_{i=1}^{n} P(p_i) \cdot log_2(\frac{P(p_i)}{P(q_i)}) = H(P,Q) - H(P)$
- Jensen-Shannon Divergence : $JSD(P||Q) = 0.5 * (KL(P||R) + KL(Q||R))$
- Generalized Jensen-Shannon Divergence : $gJSD_{\pi_1,...,\pi_n}(P_1,...,P_n) = H(\sum_{i=1}^{n} \pi_i \cdot P_i) - \sum_{i=1}^{n} \pi_i \cdot H(P_i)$

`Philentropy` already enabled the robust comparison of similarity measures in analogy-based software effort estimation (Phannachitta 2017) as well as in evolutionary transcriptomics applications (Drost et al. 2018). The package aims to assist efforts to determine optimal similarity or distance measures when developing new (statistical) models or algorithms. In addition, `Philentropy` is implemented to be applicable to large-scale datasets that were previously inaccessible using other R packages. The software is open source and currently available on GitHub (https://github.com/HajkD/philentropy) and CRAN (https://cran.r-project.org/web/packages/philentropy/index.html). A comprehensive documentation of `Philentropy` can be found at https://hajkd.github.io/philentropy/.

Acknowledgements

I would like to thank Jerzy Paszkowski for providing me an inspiring scientific environment and for supporting my projects.

References

Cha, Sung-Hyuk. 2007. "Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions." *City* 1 (2):1.

Drost, Hajk-Georg, Alexander Gabel, Jialin Liu, Marcel Quint, and Ivo Grosse. 2018. "MyTAI: Evolutionary Transcriptomics with R." *Bioinformatics* 34 (9):1589–90. https://doi.org/10.1093/bioinformatics/btx835.

Eddelbuettel, Dirk. 2013. *Seamless R and C++ Integration with Rcpp*. Use R! New York: Springer.

Phannachitta, P. 2017. "Robust Comparison of Similarity Measures in Analogy Based Software Effort Estimation." In *2017 11th International Conference on Software, Knowledge, Information Management and Applications (Skima)*, 1–7. https://doi.org/10.1109/SKIMA.2017.8294126.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.