Pattern Recognition Letters 120 (2019) 36–38

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A note on the triangle inequality for the Jaccard distance

Sven Kosub*

Department of Computer and Information Science, University of Konstanz, Box 67, Konstanz D-78459, Germany

ARTICLE INFO

Article history:
Received 26 May 2017
Available online 12 December 2018

Keywords:
Jaccard index
Set distance
Submodularity

ABSTRACT

Two simple proofs of the triangle inequality for the Jaccard distance in terms of nonnegative, monotone, submodular functions are given and discussed.

© 2018 Elsevier B.V. All rights reserved.

The Jaccard index [7] is a classical similarity measure on sets with a lot of practical applications in information retrieval, data mining, machine learning, and many more [cf., e.g., 6]. Measuring the relative size of the overlap of two finite sets A and B, the Jaccard index J is formally defined as

$$J(A,B) =_{\text{def}} \frac{|A \cap B|}{|A \cup B|}$$

and the associated Jaccard distance J_δ is formally defined as

$$J_\delta(A,B) =_{\text{def}} 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \triangle B|}{|A \cup B|}$$

where $J(\emptyset,\emptyset) =_{\text{def}} 1$. The Jaccard distance J_δ is known to fulfill all properties of a metric, notably, the triangle inequality—a fact that has been observed many times, e.g., via metric transforms [e.g., [3,11,13]], embeddings in vector spaces [e.g., [3,10,14]], min-wise independent permutations [1], or cumbersome arithmetics [9]. A very simple, elementary proof of the triangle inequality was given in [4] using an appropriate partitioning of sets.

Here we give two more simple, direct proofs of the triangle inequality. One proof comes without any set difference or disjointness of sets. It is based only on the fundamental equation $|A \cup B| + |A \cap B| = |A| + |B|$. As such, the proof is generic and leads to (sub)modular versions of the Jaccard distance (as defined below). The second proof unfolds a subtle difference between the two possible versions. Though the original motivation was to give a proof of the triangle inequality as simple as possible, the link with submodular functions and the algorithmic machinery behind [see, e.g., 2,12] is interesting in itself [e.g., 5].

Let X be a finite, non-empty ground set. A set function $f : \mathcal{P}(X) \to \mathbb{R}$ is said to be *submodular* on X if $f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$ for all $A, B \subseteq X$. If all inequalities are equations then f is called *modular* on X. It is known that f is submodular on X if and only if the following condition holds [cf., e.g., 12]:

$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B) \tag{1}$$

for all $A \subseteq B \subseteq X$, $x \in \overline{B}$. A set function f is monotone if $f(A) \leq f(B)$ for all $A \subseteq B \subseteq X$; f is nonnegative if $f(A) \geq 0$ for all $A \subseteq X$. Each nonnegative, monotone, modular function f on X can be written as $f(A) = \gamma + \sum_{i \in A} c_i$ where γ, $c_i \geq 0$ for all $i \in X$ [cf., e.g., 12]. Examples are set cardinality or degree sum in graphs. Standard examples of nonnegative, monotone, submodular set functions are matroid rank, network flow to a sink, entropy of sets of random variables, and neighborhood size in bipartite graphs.

Let f be a nonnegative, monotone, submodular set function on X. For sets $A, B \subseteq X$, we define two candidates for *submodular Jaccard distances*, $J_{\delta,f}$ and $J^{\Delta}_{\delta,f}$, as follows:

$$J_{\delta,f}(A,B) =_{\text{def}} 1 - \frac{f(A \cap B)}{f(A \cup B)},$$

$$J^{\Delta}_{\delta,f}(A,B) =_{\text{def}} \frac{f(A \triangle B) - f(\emptyset)}{f(A \cup B)}$$

where $J_{\delta,f}(A,B) = J^{\Delta}_{\delta,f}(A,B) =_{\text{def}} 0$ if $f(A \cup B) = 0$. It is clear that $0 \leq J_{\delta,f}(A,B) \leq J^{\Delta}_{\delta,f}(A,B) \leq 1$. If f is modular then $J_{\delta,f} = J^{\Delta}_{\delta,f}$. In particular, for $f(A) = |A|$ (i.e., the cardinality of the set $A \subseteq X$), we obtain the standard Jaccard distance $J_\delta = J_{\delta,f} = J^{\Delta}_{\delta,f}$.

First, we give a simple proof of the triangle inequality for $J_{\delta,f}$. Interestingly, this is only possible for modular set functions (see the third comment after Theorem 3).

Lemma 1. *Let f be a nonnegative, monotone, submodular set function on X. Then, for all sets $A, B, C \subseteq X$, it holds that*

$$f(A \cap C) \cdot f(B \cup C) + f(A \cup C) \cdot f(B \cap C) \leq f(C) \cdot (f(A) + f(B)).$$

* Corresponding author.
E-mail address: sven.kosub@uni-konstanz.de

https://doi.org/10.1016/j.patrec.2018.12.007
0167-8655/© 2018 Elsevier B.V. All rights reserved.

Proof. We easily obtain

$$\begin{aligned} & f(A\cap C)\cdot f(B\cup C) \\ & \quad \le f(A\cap C)\cdot (f(B)+f(C)-f(B\cap C)) \\ & \quad = f(A\cap C)\cdot (f(B)-f(B\cap C))+f(A\cap C)\cdot f(C) \\ & \quad \le f(C)\cdot (f(B)-f(B\cap C)+f(A\cap C)) \end{aligned}$$

and, by swapping A and B,

$$f(A\cup C)\cdot f(B\cap C)\le f(C)\cdot (f(A)-f(A\cap C)+f(B\cap C))$$

Overall,

$$\begin{aligned} & f(A\cap C)\cdot f(B\cup C)+f(A\cup C)\cdot f(B\cap C) \\ & \quad \le f(C)\cdot (f(B)-f(B\cap C)+f(A\cap C) \\ & \qquad\qquad +f(A)-f(A\cap C)+f(B\cap C)) \\ & \quad \le f(C)\cdot (f(B)+f(A)) \end{aligned}$$

This shows the lemma. □

Corollary 2. *Let f be a nonnegative, monotone, submodular set function on X. Then, for all sets S, $T\subseteq X$, it holds that*

$$f(S\cap T)\cdot f(S\cup T)\le f(S)\cdot f(T).$$

Proof. Apply Lemma 1 to sets $A=_{\text{def}} S$, $B=_{\text{def}} S$ and $C=_{\text{def}} T$. □

Theorem 3. *Let f be a nonnegative, monotone, modular set function on X. Then, for all sets A, B, $C\subseteq X$, it holds that*

$$J_{\delta,f}(A,B)\le J_{\delta,f}(A,C)+J_{\delta,f}(C,B).$$

Proof. Say that a set A is a null set if $f(A)=0$. Observe that if at least one of the sets is a null set then the inequality is satisfied. So, it is enough to show the equivalent inequality

$$\frac{f(A\cap C)}{f(A\cup C)}+\frac{f(B\cap C)}{f(B\cup C)}\le 1+\frac{f(A\cap B)}{f(A\cup B)}=\frac{f(A)+f(B)}{f(A\cup B)}$$

for arbitrary non-null sets A, B, $C\subseteq X$. (Note that the equivalence requires modularity of functions.) The validity of the inequality is seen as follows:

$$\begin{aligned} & \frac{f(A\cap C)}{f(A\cup C)}+\frac{f(B\cap C)}{f(B\cup C)} \\ & \quad = \frac{f(A\cap C)\cdot f(B\cup C)+f(A\cup C)\cdot f(B\cap C)}{f(A\cup C)\cdot f(B\cup C)} \\ & \quad \le \frac{f(C)\cdot (f(A)+f(B))}{f(A\cup C)\cdot f(B\cup C)} && \text{(by Lem. 1)} \\ & \quad \le \frac{f(C)\cdot (f(A)+f(B)))}{f((A\cup C)\cap (B\cup C))\cdot f(A\cup B\cup C)} && \text{(by Cor. 2)} \\ & \quad \le \frac{f(C)}{f\big((A\cap B)\cup C\big)}\cdot \frac{f(A)+f(B)}{f(A\cup B)} \\ & \quad \le \frac{f(A)+f(B)}{f(A\cup B)} \end{aligned}$$

This proves the theorem. □

We comment on the proof of the triangle inequality for $J_{\delta,f}$:

1. It follows from Theorem 3 that the triangle inequality is valid for the standard Jaccard distance J_δ, the generalized Jaccard distance given for vectors $x,y\in\mathbb{R}^n$ by
$$1-\frac{\sum_{i=1}^n \min\{x_i,y_i\}}{\sum_{i=1}^n \max\{x_i,y_i\}}$$
with the subcase that $x_i=\mu_A(z)$ and $y_i=\mu_B(z)$ denote multiplicities of (occurrences of) z in multisets A and B [e.g., 8], and the Steinhaus distance [3,11] (i.e., any set measures, including probability measures). We mention that all these results can equally easily be proven by the arguments in [4]; however, for modular functions satisfying $f(\emptyset)>0$, these arguments fail.
2. Theorem 3 is true for nonnegative, monotone, modular functions defined over distributive lattices; Lemma 1 and Corollary 2 also hold for nonnegative, monotone, submodular functions defined over distributive lattices. Notice that $J^{\Delta}_{\delta,f}$ is not defined over all distributive lattices (see also the third comment after Theorem 4).
3. In general, Theorem 3 is not true for nonnegative, monotone, submodular functions: Any set function f such that $f(A)=f(B)=f(A\cup B)>f(A\cap B)\ge 0$ for some non-empty, incomparable sets A, B refutes $J_{\delta,f}(A,B)\le J_{\delta,f}(A,A\cup B)+J_{\delta,f}(A\cup B,B)$. Concrete examples include linear cost functions with budget restrictions, i.e., $f(A)=\min\{B,\sum_{i\in A}c_i\}$, or the neighborhood size in a bipartite graph $G=(U\uplus V,E)$, i.e., $f(A)=|\Gamma(A)|$ where $A\subseteq U$ and $\Gamma(A)=\bigcup_{u\in A}\{v\in V\mid\{u,v\}\in E\}$.

Next we give a simple proof of the triangle inequality for $J^{\Delta}_{\delta,f}$.

Theorem 4. *Let f be a nonnegative, monotone, submodular set function on X. Then, for all sets A, B, $C\subseteq X$, it holds that*

$$J^{\Delta}_{\delta,f}(A,B)\le J^{\Delta}_{\delta,f}(A,C)+J^{\Delta}_{\delta,f}(C,B).$$

Proof. Split set C into two disjoint sets $C_0\subseteq A\cup B$ and $C_1\subseteq \overline{A\cup B}$, both possibly empty, such that $C=C_0\cup C_1$. We obtain:

$$\begin{aligned} & \frac{f(A\triangle C)-f(\emptyset)}{f(A\cup C)}+\frac{f(B\triangle C)-f(\emptyset)}{f(B\cup C)} \\ & \quad \ge \frac{f(A\triangle C)+f(B\triangle C)-2f(\emptyset)}{f(A\cup B\cup C_1)} && (f \text{ is monotone}) \\ & \quad \ge \frac{f(A\triangle C\cup B\triangle C)-f(\emptyset)}{f(A\cup B\cup C_1)} \\ & && (f \text{ is monotone, submodular}) \\ & \quad \ge \frac{f(A\triangle B\cup C_1)-f(\emptyset)}{f(A\cup B\cup C_1)} && (f \text{ is monotone}) \\ & \quad \ge \frac{f(A\triangle B)}{f(A\cup B)}-\frac{f(\emptyset)}{f(A\cup B\cup C_1)} \\ & && \text{(by Cond. (1)and } f \text{ is monotone)} \\ & \quad \ge \frac{f(A\triangle B)}{f(A\cup B)}-\frac{f(\emptyset)}{f(A\cup B)} && (f\text{is monotone}) \end{aligned}$$

This shows the theorem. □

We comment on the proof of the triangle inequality for $J^{\Delta}_{\delta,f}$:

1. It follows once more from Theorem 4 that the standard Jaccard distance, the generalized Jaccard distance, and the Steinhaus distance satisfy the triangle inequality. Moreover, $J^{\Delta}_{\delta,f}$ is also a (pseudo)metric for, e.g., linear cost functions with budget restrictions and the neighborhood size in bipartite graphs.
2. Theorem 4 suggests that $J^{\Delta}_{\delta,f}$ is the right definition of a submodular Jaccard distance. As a consequence, one might say that the *submodular Jaccard (similarity) index* should be defined as the inverse submodular Jaccard distance, i.e.,
$$J^{\Delta}_{f}(A,B)=_{\text{def}} 1-J^{\Delta}_{\delta,f}(A,B)=1-\frac{f(A\,\triangle\, B)-f(\emptyset)}{f(A\cup B)}$$
In contrast, we might refer to $J_f=_{\text{def}} 1-J_{\delta,f}$ as the *modular Jaccard index*. Again, if $f(A)=|A|$ then we obtain the standard Jaccard index $J=J^{\Delta}_f=J_f$.
3. Though $J^{\Delta}_{\delta,f}$ might generally not be defined over a given distributive lattice, it can be seen that for each nonnegative, monotone, submodular function $f:\mathcal{F}\to\mathbb{R}$ defined on a family $\mathcal{F}\subseteq\mathcal{P}(X)$ closed under union and intersection, there is a (not necessarily unique) nonnegative, monotone, submodular extension $\overline{f}:\mathcal{P}(X)\to\mathbb{R}$ on X such that $\overline{f}(A)=f(A)$ for all $A\in\mathcal{F}$ [e.g., 15], so that $J^{\Delta}_{\delta,\overline{f}}$ can be used instead.

Following the discussion above, we see the most interesting area of applications of the submodular Jaccard index in the context of graph-based models of data. For instance, the neighborhood size in bipartite graphs (as a proper submodular function) can be considered in topic modelling, analysis of citation networks, or for a set of new quality measures in cluster analysis.

Acknowledgments

I am grateful to Ulrik Brandes (Konstanz), Julian Müller (Konstanz), and the anonymous referees for helpful discussions and hints.

References

[1] M.S. Charikar, Similarity estimation techniques from rounding algorithms, in: Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC'2002), ACM, New York, NY, 2002, pp. 380–388.
[2] S. Fujishige, Submodular functions and optimization, Annals of Discrete Mathematics, 58, second ed., Elsevier, Amsterdam, 2005.
[3] A. Gardner, J. Kanno, C.A. Duncan, R. Selmic, Measuring distance between unordered sets of different sizes, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2014), IEEE, New Jersey, NJ, 2014, pp. 137–143.
[4] G. Gilbert, Distance between sets, Nature 239 (1972) 174.
[5] J. Gillenwater, R. Iyer, B. Lusch, R. Kidambi, J.A. Bilmes, Submodular hamming metrics, in: Advances in Neural Information Processing Systems 28, NIPS, 2015, pp. 3141–3149.
[6] J.C. Gower, Similarity, dissimilarity and distance, measures of, in: S. Kotz, C.B. Read, N. Balakrishnan, B. Vidakovic (Eds.), Encyclopedia of Statistical Sciences, 12, second ed., John Wiley, New York, NY, 2008, pp. 7730–7738.
[7] P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et du jura, Bull. Société Vaudoise des Sciences Naturelles 37 (1901) 547–579.
[8] W.A. Kosters, J.F.J. Laros, Metrics for mining multisets, in: Research and Development in Intelligent Systems XXIV, Proceedings of the Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence (AI'2007), Springer, Berlin, 2007, pp. 293–303.
[9] M. Levandowsky, D. Winter, Distance between sets, Nature 234 (1971) 34–35.
[10] A.H. Lipkus, A proof of the triangle inequality for the Tanimoto distance, J. Math. Chem. 26 (1999) 263–265.
[11] E. Marczewski, H. Steinhaus, On a certain distance of sets and the corresponding distance of functions, Colloquium Mathematicum 6 (1958) 319–327.
[12] A. Schrijver, Combinatorial Optimization, B, Springer, Berlin, 2003.
[13] D.A. Simovici, C. Djeraba, Mathematical Tools for Data Mining, Springer, London, 2008.
[14] T.T. Tanimoto, An Elementary Mathematical Theory of Classification and Prediction, Technical Report, 1958.
[15] D.M. Topkis, Minimizing a submodular function on a lattice, Oper. Res. 26 (1978) 305–321.