A Generalized Similarity Measure

Author:
Ramesh Chandra Bagadi
Data Scientist
INSOFE (International School Of Engineering), Hyderabad, India.
rameshebagadi@uwalumni.com
+91 9440032711

Technical Note

Abstract

In this research Technical Note the author has presented a novel method of finding a Generalized Similarity Measure between two Vectors or Matrices or Higher Dimensional Data of different sizes.

Theory

Considering two different vectors of different sizes namely \( A_{1\times n} \) and \( B_{1\times m} \), we first find the Proximity Matrix between elements of the given vectors wherein the Proximity Matrix is given by

\[
P_A = \begin{bmatrix}
    d(1,1) & d(1,2) & d(1,3) & \ldots & d(1,(m-1)) & d(1,m) \\
    d(2,1) & d(2,2) & d(2,3) & \ldots & d(2,(m-1)) & d(2,m) \\
    d(3,1) & d(3,2) & d(3,3) & \ldots & d(3,(m-1)) & d(3,m) \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
    d((m-1),1) & d((m-1),2) & d((m-1),3) & \ldots & d((m-1),(m-1)) & d((m-1),m) \\
    d(m,1) & d(m,2) & d(m,3) & \ldots & d(m,(m-1)) & d(m,m)
\end{bmatrix}
\]

and

\[
P_B = \begin{bmatrix}
    d(1,1) & d(1,2) & d(1,3) & \ldots & d(1,(n-1)) & d(1,n) \\
    d(2,1) & d(2,2) & d(2,3) & \ldots & d(2,(n-1)) & d(2,n) \\
    d(3,1) & d(3,2) & d(3,3) & \ldots & d(3,(n-1)) & d(3,n) \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
    d((n-1),1) & d((n-1),2) & d((n-1),3) & \ldots & d((n-1),(n-1)) & d((n-1),n) \\
    d(n,1) & d(n,2) & d(n,3) & \ldots & d(n,(n-1)) & d(n,n)
\end{bmatrix}
\]

d indicates the distance measured in some metric (default = Euclidean)
We then find the Norm Of \( P_A \) as \( \| P_A \cdot P_A \| \). For the Euclidean case, it is given by
\[
\| P_A \cdot P_A \| = \sum_{j=1}^{m} \sum_{i=1}^{m} P(i, j) \cdot P(i, j).
\]
Also, \( m < n \). Similarly, we compute the Norm of \( P_B \) as \( \| P_B \cdot P_B \| \). For the Euclidean case, it is given by
\[
\| P_B \cdot P_B \| = \sum_{j=1}^{n} \sum_{i=1}^{n} P(i, j) \cdot P(i, j).
\]
We then find the ratio \( k_1 = \frac{\| P_B \cdot P_B \|}{\| P_A \cdot P_A \|} \).

Actually, we can note that there are only \( N_B = \frac{n^2 - n}{2} \) number of possibly distinct values of Proximity Matrix elements in \( P_B \) and similarly, there are only \( N_A = \frac{m^2 - m}{2} \) number of possibly distinct values of Proximity Matrix elements in \( P_A \).

Similarly, we find some more ratio’s \( k_{N_B-1} = \frac{f_{(N_B-1)}(P_B)}{f_{(N_B-1)}(P_A)} \) where \( f_{(N_B-1)}(P_B) \) is some Scalar Function of the Matrix \( P_B \). And so is \( f_{(N_B-1)}(P_A) \). Note that \( f_{(N_B-1)} \) is the same in \( f_{(N_B-1)}(P_B) \) and \( f_{(N_B-1)}(P_A) \).

We now consider a fictitious Vector \( A_{B_{1,n}} \), i.e., Vector A in the basis of Vector B, colloquially speaking. Let this be \( A_{B_{1,n}} = [ c_1 \ c_2 \ c_3 \ \ldots \ c_{n-1} \ c_n ] \). Now, for this, vector, we find the Proximity Matrix \( P_{A_{B_{1,n}}} \) and now assert that \( k_{N_B-1} = \frac{f_{(N_B-1)}(P_B)}{f_{(N_B-1)}(A_{B_{1,n}})} \). This gives us \( N_B \) number of equations from which we can solve for elements of \( A_{B_{1,n}} \). Now, we can find distance between \( A_{B_{1,n}} \) and \( B_{1,n} \) and can also consequently find the Similarity co-efficient between them. We can also, repeat this procedure using the normalized values of the vectors \( A_{1,m} \) and \( B_{1,n} \). In the same fashion as detailed above, we can repeat this procedure for Matrices or Higher Dimensional Data of differing sizes.

References
http://www.vixra.org/author/ramesh_chandra_bagadi