ENTROPY AND MUTUAL INFORMATION IN RELATION TO EUCLIDEAN DISTANCE FOR DEPENDENCY GRAPH GENERATION

Dr. Sreedevi. O. B

Department of Mathematics, Christ Nagar College Maranalloor, Thiruvananthapuram

sreedevishajin@gmail.com

ABSTRACT

Euclidean distance is mainly used to calculate likeness between two functions. Dependency graph generation and Dependency Graph Matching are two phases of the dependency measurement. To generate the dependency graph primarily the Entropy, Joint Entropy, Conditional Entropy, Relative Entropy and Mutual information rooted in information theory. Between the two dependencies graph of schema matching process is done b the distance measure, Euclidean distance is used for matching schema. This paper presents the connection between of entropy and Mutual information in Euclidean distance.

Keywords:

Entropy, Mutual Information, Euclidean Distance, Dependency Graph Generation

1. INTRODUCTION

Dependency graph generation is a weighted graph with n nodes. The weight of the edge associating two nodes is the mutual information between two properties. Euclidean distance is a distance between two points in a space. Entropy evaluates uncertainty, while Mutual information measures shared information between variables of straight line. Entropy and mutual information are statistical measures based on probability mass function.

2. DEPENDENCY GRAPH GENERATION

This area manages brief portrayal about dependency graph generation. Here, the instances of two tables are connected then they are changed into graph format. After this, for creating dependency graph, the entropy and mutual information values are determined independently for each table instances. Dependency graph for the given schema instance *S* $(a_1,a_2,a_3...a_n)$ with n characteristics, the dependency graph can be spoken to as a graph with nodes. In Kang and Naughtas approach, dependency graph is a weighted graph. The weight of the edge associating two nodes is the mutual information between two qualities. In the proposed methodology, dependency graph is a directed graph with directed edges showing useful functional dependency between characteristics.

Table 1 Example of input table instances of the proposed work

A	B	C	D W	X	Y	Z
A_1	B_1	C_1	W_1	X_1	Y_1	Z_1
A_2	B_2	C_2	D_2W_2	X_2	Y_2	Z_2
A_3	B_3	C_3	W_3	X_3	Y_3	Z_3
A_4	B_4	C_4	D_4W_4	X_4	Y_4	Z_4

3. ENTROPY

The attribute's entropy determines the vagueness of values with a non-negative real number. The value of entropy is built independently for every characteristic in the two table instances. In addition, the values of entropy built up on the probabilities instead of attribute based actual values. Here, we used entropy value for getting the mutual information however this technique built up on conditional entropy among the characteristics as determined co-connection.

The Attribute's entropy defines

$$C_y(R) = -\sum_{r \in R} P(r) \log P(r) \qquad (1)$$

$$C_y(R/S) = -\sum_{m \in M} p(r) \sum_{n \in N} P(r/s) \log_2 (P(r/s)) \qquad (2)$$

Here C (R/S) mention the conditional entropy among the attributes R and S.

P(r) denotes the probability of attribute R.

P(r/s) denotes the conditional probability among the attribute R and S.

Table 2 Entropy table

A	B	C	D
1.5	2.0	1.0	1.5

W	X	Y	Z
2.0	1.5	1.0	1.5

The entropy of an arbitrary variable is a function which endeavours to portray the "unusualness" of an irregular variable. Consider an arbitrary variable *X* speaking to the number those surfaces on a roulette haggle and an irregular variable *Y* speaking to the number that surface on a reasonable 6-sided kick the bucket. The entropy of *X* is more noteworthy than the entropy of *Y*. Notwithstanding the numbers 1 through 6, the values on the roulette wheel can take on the values 7 through 36. In some sense, it is less unsurprising.

Yet, entropy isn't just about the quantity of potential results. It is additionally about their recurrence. For instance, let Z be the result of a weighted six-sided kick the bucket that the surfaces 90% of the time as a "2". Z has lower entropy than Y speaking to a reasonable 6-sided kick the bucket. The weighted kick the bucket is less eccentric, in some sense.

Be that as it may, entropy is anything but a dubious idea. It has an exact scientific definition. Specially, if a irregular variable *X* takes on values in a set $X = \{x_1, x_2, \ldots, x_n\}$, and is characterized by a probability distribution *P(X)*, at that point we will compose the entropy of the arbitrary variable as,

The property's entropy characterized the

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{3}$$

$$H(p(X)) = H(P) = H(X) \tag{4}$$

Here, *H(X)* mention the conditional entropy among the attributes *R* and *S*

p(x) denotes the probability of attribute *R*.

p(X) denotes the conditional probability among the attribute.

If the log in the above equation is taken to be to the base 2, at that point the entropy is expressed in bits. If the log is taken to be the natural log, then the entropy is expressed in *nats*. To register the entropy of a climatic condition, we first define its distribution:

Example 1

$$p(X - sunny) - 1/2 \qquad\qquad p(X - rainy) - 1/2$$

Using eq. 3.3 we have,

$$H(X) = -\sum_{x \in \{heads, tails\}} P(x) \log P(x)$$

$$= -[1/2\log 1/2 + 1/2\log 1/2]$$

$$= -[-1/2 + -1/2]$$

= 1.

4. JOINT ENTROPY

Joint entropy is the entropy of a joint probability distribution, or a multi-valued random variable. The joint entropy *H* (A, B) of a pair of discrete random variable A and B with joint distribution *P (a, b)* given by

$$H(A,B) = \sum_a \sum_b p(a,b)\log(1/\,p(a,b)) \qquad (5)$$

Given X a chance to speak to whether it is radiant or blustery in a specific town on a given day. Given Y a chance to speak to whether it is over 65 degree or underneath 65 degrees. Compute the entropy of the joint distribution *P(X, Y)* given by

P (radiant, hot) = 1/2

P (radiant, cool) =1/4

P (blustery, hot) =1/4

P (blustery, cool) =0

Using the above equation, H(X, Y) = 1/2log2 + 1/4log4 + 1/4log4 + 0log0

= 3/2.

5. CONDITIONAL ENTROPY

Let X and Y are discrete random variable with joint distribution p (x, y) and conditional distribution p(x/y), and then the conditional entropy is characterized by

$$H(X|Y=y) = -\sum_x \sum_y p(x|y)\log p(x|y) \qquad (6)$$

6. RELATIVE ENTROPY

Let two discrete distributions have probability functions *p(x)* and let a second discrete distribution have function *q(x)*. Then the relative entropy of *p* regarding *q*, likewise called the Kullback-Leibler distance, is characterized by

$$d=\sum_{x}p(x)\log_{2}\left(\frac{p(x)}{q(x)}\right) \qquad (7)$$

7. MUTUAL INFORMATION

Mutual information is an amount that estimates a connection between two irregular factors that inspected at the same time. Specifically, it quantifies how much data is conveyed, all things considered, in one irregular variable about another.

For instance, assume X speaks to the move of a reasonable 6-sided kick the bucket, and *Y* speaks to whether the roll is even (0 assuming even, 1 if odd). Unmistakably, the estimation of Y discloses to us something about the estimation of *X* and the other way around. That is, these factors share mutual information. Then again, if *X* speaks to the move of one reasonable bit the dust, and *Z* speaks to the move of another reasonable kick the bucket, at that point *X* and *Z* share no mutual information. The move of one bite the dust does not contain any information about the result of the other kick the bucket. A significant hypothesis from data hypothesis says that the mutual information between two factors is 0 if and just if the two factors are measurably autonomous. The formal meaning of the mutual information of two irregular factors *X* and *Y* is given by

$$I(X,Y)=\sum_{x\in X}\sum_{y\in Y}P(X,Y)\log\frac{P(x,y)}{P(x)p(y)} \qquad (8)$$

From Entropy using the mutual information formula

Table .3 Mutual information

A-B	A-C	A-D
1.5	1.0	1.0
B-C	B-D	C-D
1.0	1.5	0.5

W-X	W-Y	W-z
1.5	1.0	1.5
X-Y	X-Z	Y-Z
0.	1.0	1.0

In this definition, P (X Y) is the joint distribution and the marginal distributions are P(x) and P(y).

If I (X, Y) = 0, x and y are independent random variables so P(x, y) = p(x).P(y)

$$\sum_{x\in X}\sum_{y\in Y}\log\frac{P(x,y)}{P(x)\,p(y)} = \log 1 = 0. \tag{9}$$

If the log base 2 is used, the units of mutual information are bits.

8. RELATIONSHIP BETWEEN ENTROPY AND MUTUAL INFORMATION

I (X; Y) = H (Y) – H (Y |X)

Proof:

I (X; Y) = H (Y) – H (Y |X)

H (X, Y) = H (X) + H (Y |X)

I (X; Y) = H(X) + H(Y) – H (X, Y)

I (X; X) = H(X) − H (X|X) = H(X)

Example 2

Table 4 Relationship between Entropy and Mutual Information

		X: Blood Group			
Y: Chance for blood sugar		A	B	AB	O
	Very low	1/8	1/16	1/32	1/32
	Low	1/16	1/8	/32	1/32
	Medium	1/16	1/16	1/16	1/16
	High	¼	0	0	0

X: Marginal (1/2, 1/4, 1/8, 1/8)

Y: Marginal (1/4, 1/4, 1/4, 1/4)

H(X): 7/4 bits

H(Y): 2 bits

Conditional entropy, H(X/Y): 11/8 bits

H(Y/X): 13/8 bits

H(Y/X) ≠ (X/Y)

Mutual Information, I(X: Y) = H(X)-H(X/Y)

= 0.375 bits

9. EUCLIDEAN DISTANCE MEASURE

The Euclidean Distance between two probability mass functions,
P = *{p₁, p₂... p ₙ}* and *Q* = *{q₁, q₂... q ₙ}* is defined as:

$$D_E = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)}^2 \qquad (10)$$

For any trait *pair (i, j)* to be coordinated, their pmfs are arranged and the incentive with the most elevated probability in the main arranged pmf is mapped to the incentive with the highest probability in the second arranged pmf, and the equivalent is rehashed for different qualities. This worth mapping results in least squared Euclidean distance between the two pmfs and the mapping is said to be optimal.

Table 5 Euclidean distance between probability mass functions

A	P(A)	B	P(B)	C	P(C)	D	P(D)	W	P(W)	X	P(X)	Y	P(Y)	Z	P(Z)
A1	0.5	B_2	0.25	C_1	0.5	D_1	0.25	W_2	0.25	X_1	0.25	Y_1	0.5	Z_2	0.5
A_3	0.25	B_4	0.25	C_3	0.5	D_2	0.5	W_4	0.25	X_2	0.5	Y_3	0.5	Z_3	0.25
A_4	0.25	B_1	0.25			D_3	0.25	W_3	0.25	X_3	0.25			Z_1	0.25
		B_3	0.25					W_1	0.25						
H(A)=1.5		H(B)=2.0		H(C)=1.0		H(D)=1.5		H(W)=2.0		H(X)=1.5		H(Y)=1.0		H(Z)=1.5	

The Euclidean distance between pmfs of any characteristic pair *(i, j)* where *i* is a trait of table *R* and *j* is a trait of table *S* is computed using Equation (3.3). Trait *B* of table *R* has the same pmf as that of traits X and *Z* and subsequently the difference metric between the trait pairs is 0. Thus, we can see that the trait pairs *(B, W), (C, Y), (D, Z), (B, Z) and (D, W)* have similar pmfs. Schema matching is done dependent on pmfs, of attributes may result in equivocal property matching. In the given example, there is vulnerability about mapping *B* or *D* with *W* or *Z*. Jaiswal et al accept that the matching traits will have comparable probability distributions that are dissimilar from the probability distributions of other matching properties. Their suspicions may not be valid in numerous situations. For instance, pmf of attributes *A* and *D* in table *R* and that of characteristics *Y* and *Z* in table *S* are not distinct. Notwithstanding when two traits have similar pmf their ground truth may not match.

10. ENTROPY ONLY EUCLIDIAN DISTANCE MEASURES

Let R and S be two tables with equal number of attributes, and ea_i and eb_i be the entropies of attribute *i* in table *R* and *S*, respectively. Let *m* be an index that maps an attribute in Table *R* to the matching attribute in Table *S*. The Entropy-only distance metric for Table *R* and *S* is characterized as,

$$D_E^u(R,S) = \sqrt{\sum_i (ea_i - eb_{m(i)})^2} \qquad (11)$$

Kang & Naughton also performed weighted graph matching by considering mutual information between the schema elements as weights between the adjacent nodes in the dependency graphs. They used the following Euclidean Distance Metric to measure the distance between the two graphs.

11. MUTUAL INFORMATION BASED EUCLIDEAN DISTANCE METRIC

Let G_1 and G_2 be two equal size dependency graphs for tables *R* and *S*, and a_{ij} and b_{ij} be the mutual information between the node *i* and *j* in the graphs G_1 and G_2, respectively. Let *m* be an index that maps a node in graph G_1 to the matching node in graph G_2. The Euclidean distance metric between the two graphs G_1 and G_2 is defined as,

$$D_m^u(R,S) = \sqrt{\sum_i (a_{ji} - b_{m(i)m(j)})^2} \qquad (12)$$

12. CONCLUSION

This study provides a general analytical framework integrating entropy, mutual information, and Euclidean distance of dependency analysis. The findings show that information-theoretic and problematic measures, mainly entropy and mutual information, The examples and tables are the most reliable and easy to understand about the how entropy and mutual information relate to Euclidean distance. Future work focused extending the framework to high-dimensional and graph structured data.

REFERENCES

[1] Burgo, D, Burgo, Y, & Ivanov, S. A course in Metric Geometry. et al, (AMS,2001) pp1-5

[2] Cover, T.M, & Thomas, J. A, Elements of Information Theory (2nd ed), Wiley- interscience (2006) 13-55

[3] Deza, M.M & Deza, E, Encyclopedia of Distances Springer 2009 pp 94-99

[4] Du, Ruihuan, Jiannan Yang, Yongzhi Cao, and Hanpin Wang "Personalized graph pattern matching via limited simulation" Knowledge- Based Systems, vol 141, pp 31-43,2018

[5] Floerchinger, S, Haas, Thermodynamics from relative entropy (2020)

[6] Gray, R.M, 2nd edition-Springer,2011, Entropy and Information theory (2nd ed), pp.1-29,63-71,72-75

[7] Kreyzig, E (Wiley, 1978) Introductory functional analysis with applications, pp.6-7

[8] MacKay, D.J.C, Cambridge,2003, Information Theory, inference and learning Algorithms, pp28-33, Dover Publications, pp 15-24, 52-58

[9] Rudin, W, 3rd Edition-McGraw-Hill, 1976, Principles of mathematical Analysis (3rd) pp.16-21

[10] Shannon, C.E., & Weaver, W. The Mathematical theory of Communication, University of Illinois Press (2006).8-14