International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 5, Sep- Oct 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

New Weighted Dissimilarity Function between Two Probability Distributions

Amit Srivastava*

*(Department of Mathematics, Jaypee Institute of Information Technology, Noida(Uttar Pradesh)India
Email: amit90554@gmail.com)

----------**********----------

Abstract:

In the present work, a new weighted measure of dissimilarity between two probability distributions is proposed. The properties of the proposed measure have been discussed and it is shown that the proposed measure is a valid weighted measure of dissimilarity between two probability distributions.

Keywords **—Probability distribution, Dissimilarity measure, Weighted distribution.**

----------**********----------

I. INTRODUCTION

Let $P(n) = \{p_1, p_2, \dots, p_{i-1}, p_i, p_{i+1}, \dots, p_n\}$ and $Q(n) = \{q_1, q_2, \dots, q_{i-1}, q_i, q_{i+1}, \dots, q_n\}$ be two probability distributions defined on a sample space X.Due to its applicability in various domains, how to quantify the dissimilarity between two probability distributions has been a matter of concern for many researchers over the past few years. Kullback and Leibler[1, 2] proposed a measure of dissimilarity between $P(n)$ and $Q(n)$ as

$$K\big(P(n);\ Q(n)\big) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \qquad (1.1)$$

The symmetric version of (1.1) is given as

$$J\big(P(n); Q(n)\big) = K\big(P(n);\ Q(n)\big) + K\big(Q(n);\ P(n)\big) = \sum_{i=1}^{n}(p_i - q_i) \log \frac{p_i}{q_i} \qquad (1.2)$$

A number of other divergence measures have proposed by researchers which are either generalizations of (1.1) or are not direct generalizations of (1.1), but satisfy all the properties that the measure (1.1) satisfy [3,9,10 and 11]. In general any dissimilarity function $D\big(P(n), Q(n)\big)$ between $P(n)$ and $Q(n)$ must satisfy the following properties.

a) $D\big(P(n), Q(n)\big) > 0$.

b) $D\big(P(n), Q(n)\big) = 0$ if and only if $P(n)$ and $Q(n)$are identical distributions.

Now consider a utility distribution

$$W(n) = \{w_1, w_2, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_n\}$$

characterizingthe weights of the set of events

$$S = \{E_1, E_2, \dots, E_{i-1}, E_i, E_{i+1}, \dots, E_n\}$$

events with probabilities p_i. Now consider two weighted information schemes as

$$S = \begin{pmatrix} E_1, E_2, \dots & E_{i-1}E_i & E_{i+1} \dots & E_n \\ p_1, p_2 \dots & p_{i-1}p_i & p_{i+1} \dots & p_n \\ w_1, w_2 \dots & w_{i-1}w_i & w_{i+1} \dots & w_n \end{pmatrix}$$

$$p_i \geq 0, w_i > 0, i = 1, 2, \dots, n, \sum_{i=1}^{n} p_i = 1$$

and

$$S^* = \begin{pmatrix} E_1, E_2, \dots & E_{i-1}E_i & E_{i+1} \dots & E_n \\ q_1, q_2 \dots & q_{i-1}q_i & q_{i+1} \dots & q_n \\ w_1, w_2 \dots & w_{i-1}w_i & w_{i+1} \dots & w_n \end{pmatrix}$$

$$q_i \geq 0, w_i > 0, i = 1, 2, \dots, n, \sum_{i=1}^{n} q_i = 1$$

The utility distribution $W(n)$is identical for both the schemes above since the utility of any event is assumed to be independent of its probability of occurrence. In general, $w_i (i = 1,2, \dots n)$is a non – negative, finite, real number accounting for the relevance, significance or the utility of the occurrence of an event with probability $p_i (i =$

$1,2,\dots n$). Also, if one event is more relevant, more significant, and more useful (with respect to a given goal or from a given qualitative point of view) than another one, the weight of the first event will be greater than that of second one.In general the weights are independent of the probabilities *i.e.* any event with low probability of occurrence can be assigned a higher weight and vice versa.

Taneja and Tuteja [4] proposed a generalization of (1.1) as

$$I\big(P(n);\ Q(n);\ W(n)\big) = \sum_{i=1}^{n} w_i p_i\ log\ \frac{p_i}{q_i} \tag{1.3}$$

The above function (1.3) represents a weighted measure of dissimilarity between $P(n)$ and $Q(n)$. When $w_i = w$(say), a constantfor each i, then (2.1) reduces to (1.1). The major issue with (1.3) is that weights and probabilities are not properly integrated and as a result is not a valid measure of weighted dissimilarity.Also various authors [5, 6, 7 and 8] have shown that the weighted measure of dissimilaritygiven by (1.3) can be both negative and positive and can even vanish when $P(n)$ is not necessarily equal to $Q(n)$. The object of this paper is to characterize a new weighted measure of dissimilarity which do not suffer from the weakness mentioned above *i.e.* it should be always positive. Further we will discuss the properties of the new measure and its various econometric applications.

II. NEW WEIGHTED MEASURE OF DISSIMILARITY

Again let

$$P(n) = \{p_1, p_2, \dots, p_{i-1}, p_i, p_{i+1}, \dots, p_n\}$$

and $Q(n) = \{q_1, q_2, \dots, q_{i-1}, q_i, q_{i+1}, \dots, q_n\}$

be two probability distributions defined on a sample space X. Also consider

$$W(n) = \{w_1, w_2, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_n\}$$

characterizing the utilities of events with probability $p_i (i = 1,2, \dots n)$. Now weighted measure of dissimilarity between $P(n)$ and $Q(n)$ is given as

$$\tilde{I}\big(P(n);\ Q(n);\ W(n)\big) = \sum_{i=1}^{n} w_i p_i\ log\ \frac{p_i}{q_i} - \sum_{i=1}^{n} w_i p_i\ log\ \left(\frac{\sum_{i=1}^{n} w_i p_i}{\sum_{i=1}^{n} w_i q_i}\right) \tag{2.1}$$

The measure (2.1) is always non negative. To prove this, we proceed as follows.

Let $x_1, x_2, \dots, x_n$ be positive real numbers and let $\alpha_1, \alpha_2, \dots, \alpha_n$ be non-negative real numbers whose sum is one. Then from weighted geometric harmonic mean inequality, it follows that for $n \geq 2$.

$$\frac{1}{\sum_{i=1}^{n}\left(\frac{\alpha_i}{x_i}\right)} \leq \prod_{i=1}^{n} {x_i}^{\alpha_i}$$

with equality if all x_i's are equal and $\sum_{i=1}^{n} \alpha_i = 1$.

Now taking logarithm gives

$$-log\left[\sum_{i=1}^{n} \frac{\alpha_i}{x_i}\right] \leq \sum_{i=1}^{n} \alpha_i\ log\ x_i$$

Taking $\alpha_i = \frac{w_i p_i}{\sum_{i=1}^{n} w_i p_i}, x_i = \frac{p_i}{q_i}$ in the above inequality gives

$$-log\left[\sum_{i=1}^{n} \frac{w_i p_i}{\sum_{i=1}^{n} w_i p_i} \times \frac{q_i}{p_i}\right] \leq \sum_{i=1}^{n} \frac{w_i p_i}{\sum_{i=1}^{n} w_i p_i}\ log\ \frac{p_i}{q_i}$$

$$\Rightarrow -log\left[\sum_{i=1}^{n} \frac{w_i q_i}{\sum_{i=1}^{n} w_i p_i}\right] \leq \frac{1}{\sum_{i=1}^{n} w_i p_i} \sum_{i=1}^{n} w_i p_i\ log \frac{p_i}{q_i}$$

$$\Rightarrow -\sum_{i=1}^{n} w_i p_i\ log\left[\sum_{i=1}^{n} \frac{w_i q_i}{\sum_{i=1}^{n} w_i p_i}\right] \leq \sum_{i=1}^{n} w_i p_i\ log \frac{p_i}{q_i}$$

$$\Rightarrow \sum_{i=1}^{n} w_i p_i\ log \frac{p_i}{q_i} + \sum_{i=1}^{n} w_i p_i\ log\left[\sum_{i=1}^{n} \frac{w_i q_i}{\sum_{i=1}^{n} w_i p_i}\right] \geq 0.$$

$$\Rightarrow \tilde{I}\big(P(n);\ Q(n);\ W(n)\big) \geq 0.$$

When $w_i = w$(say), a constant for each i, then (2.1) reduces to (1.1).Also the measure (2.1) vanishes if and only if the if and only if

$$p_i = q_i \text{ for } i = 1,2, \dots, n.$$

Finally, the measure (2.1) remains the same when the elements in the triplet $(p_i, q_i, w_i), i = 1,2, \dots, n$ are permuted among themselves *i.e.* (2.1) is permutationally symmetric.

We can now list properties of (2.1) which will ascertain it as a valid weighted dissimilarity measure between two probability distributions.

a) The function (2.1) is always non negative *i.e.*

$$\tilde{I}\big(P(n);\ Q(n);\ W(n)\big) \geq 0$$

b) $\tilde{I}\big(P(n);\ Q(n);\ W(n)\big) = 0$ if and only if $p_i = q_i$ for $i = 1,2,\dots,n$.

c) The function (2.1) is permutationally symmetric with respect to the the triplet $(p_i, q_i, w_i), i = 1,2,\dots,n$.

d) $I(1/2\,;1;1) = 1$.

e) Let $Q(n)$ be uniform distribution given by

$$Q(n) = \left\{\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right\} = U(n)$$

Then

$$I(P(n); U(n)) = \sum_{i=1}^{n} w_i p_i \, log np_i + \sum_{i=1}^{n} w_i p_i \, log\left(\frac{\sum_{i=1}^{n} w_i}{n\sum_{i=1}^{n} w_i p_i}\right)$$

$$\Rightarrow I(P(n); U(n)) = \sum_{i=1}^{n} w_i p_i \, log\left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i p_i}\right) - \left(-\sum_{i=1}^{n} w_i p_i \, log p_i\right)$$

which can be written as

$$I(P(n); U(n)) = H(U(n)) - H(P(n)), \text{ where}$$

$$H(U(n)) = \sum_{i=1}^{n} w_i p_i \, log\left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i p_i}\right)$$

and

$$H(U(n)) = -\sum_{i=1}^{n} w_i p_i \, log p_i$$

III. CONCLUSION

In the present work, we have developed a new measure which quantifies the weighted dissimilarity between two probability distributions. Work on generalization of the proposed measure is in progress and

REFERENCES

[1] Kullback, S. and Leibler, A. S., "On information and sufficiency",*Ann. Math. Stat.*, 22, pp. 79-86, 1951.
[2] Kullback, S., Information theory & Statistics, Dover Publications, New York, 1968.
[3] I. J. Taneja, Generalized Information Measures and Their Applications. Available: http://www.mtm.ufsc.br/~taneja/book/book.html, 2001.
[4] H. C. Taneja, and R. K. Tuteja,, "Characterization of a quantitative-qualitative measure of relative information", *Inf. Sci*,.33, pp. 217-222, 1984.
[5] T. O. Kvalseth, "The Relative Useful Information Measure: Some Comments",*Inf. Sci*,.56, pp. 35 – 38, 1991.
[6] J. N. Kapur, On the concept of useful information, *J. Org. Behav. Stat.* 2(3, 4),pp. 147 – 162, 1985.
[7] J. N. Kapur, New qualitative – quantitative measures of information, *National Academy Science Letters* 9(2), pp. 51 – 54, 1986.
[8] E. Landaburu and L. Pardo, Goodness of fit tests with weights in the classes based on (h ,ϕ) – Divergences, *Kybernetika* 36(5) 589 – 602.
[9] A. Srivasta, Medical diagnosis using intuitionistic fuzzy sets AIP Conf. Proc., 2061 (2019), Article 020029, 10.1063/1.5086651.
[10] A. Srivastava, A. K. Singh, S. Maheswari, Dichotomous exponential entropy functional and its applications in medical diagnosis, International Conference on Signal Processing and Communications (ICSC-2013) (IEEE), 21- 26. 10.1109/ICSPCom.2013.6719749, 2013.
[11] S. Maheshwari & A. Srivastava, "Application of Intuitionistic Fuzzy Cross Entropy Measure in Decision Making for Medical Diagnosis". International Journal of Mathematical and Computational Sciences, 9(4), 254–258, 2015.

