ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (499-504), Month: April - June 2015, Available at: www.researchpublish.com

Analyzing Entropy of Data

[1]S.K.Bhatia, [2]Ayush Patel, [3]Mayank Katiyar

[1]Assistant Professor, [1, 2, 3] Department of Instrumentation and Control Engineering,
J.S.S. Academy of Technical Education, Noida

Abstract: **the entropy (information theory) is a measure of uncertainty rather than certainty. Entropy provides an absolute limit on the best possible average length of lossless encoding or compression of any communication. Entropy analyzation can be used to determine cardiovascular diseases, arrhythmias, redundancy of a language and probability of data received in a message. There are varieties of entropies possible and this paper proposes a systematic view of these entropies.**

Keywords: **Multiscale entropy analysis, sample entropy, single scale entropy.**

I. INTRODUCTION

This paper proposes research conducted so far on Entropy which is a measure of unpredictability of information content .Let us consider the example of a coin toss. When the coin is fair, that means, when the probability of heads is the same as the probability of tails, then the entropy of the coin toss is as equal to half each. This is because there is no way to predict the outcome of the coin toss ahead of time but the best we can do is predict that the coin will come up heads, and our prediction will be correct with probability 1/2. Such a coin toss has one bit of entropy since there are two possible outcomes that occur with equal probability, and learning the actual outcome contains one bit of information. On the other hand a coin toss with a coin that has two heads and no tails has zero entropy since the coin will always come up heads, and the outcome can be predicted perfectly.

English text has fairly low entropy. English text has between 0.6 and 1.3 bits of entropy for each character of message that is, you can always recover the entire original message by decompressing then a compressed message has the same quantity of information as the original, but communicated in fewer characters. That is, it has more information, or a higher entropy, per character. In other words, it is fairly predictable. Even if we don't know exactly what is going to come next, we can be fairly certain that, for example, there will be many more e's than z's, that the combination 'qu' will be much more common than any other combination with a 'q' in it, and that the combination 'th' will be more common than 'z', 'q', or 'q u'. After the first few letters one can often guess the rest of the word. This means a compressed message has less redundancy. Rougly Shannon's source coding theorem says that a lossless compression scheme cannot compress messages, on average, to have more than one bit of information per bit of message, but that any value less than one bit of information can be attained by employing a suitable coding scheme. The entropy of a message per bit multiplied by the length of that message is a measure of how much total information the message contains.

Shannon's theorem also implies that no lossless compression scheme can shorten messages. This is generally not a problem, because we are usually only interested in compressing certain types of messages, for example English documents as opposed to gibberish text , digital photographs rather than noise, and it is unimportant if a compression algorithm makes some unlikely or uninteresting sequences.

II. DEFINITION

The entropy H of a discrete random variable X with possible values {x1 ,x2…….,xn} and probability mass function P(X) as

$$H(X) = E[I(X)]= E[-\ln (P(X)]$$

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (499-504), Month: April - June 2015, Available at: www.researchpublish.com

Where E is the expected value operator and I is the information content of X

I(X) is a random variable.

$$\mathrm{H}(X) = \sum_i \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_i \mathrm{P}(x_i)\log_b \mathrm{P}(x_i),$$

In case if P(x) =0;

Lim p(log p)=0

Where p tends to zero.

Example

Entropy of a coin flip measured in Shannons graphs vs the fairness of the coin where X=1 results in heads.(fig.1).

The maxima of the graph depend on distribution. The entropy is 1 Shannon and to communicate the outcome of a fair coin flip will require an average of 1 bit.

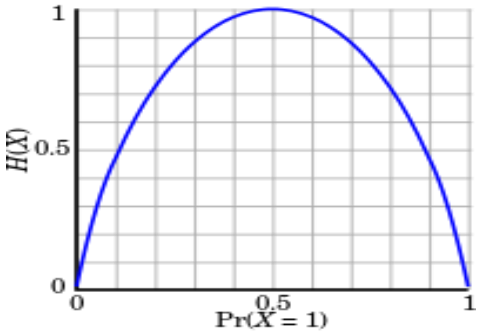

Fig.1

III. MULTISCALE ENTROPY ANALYSIS

Multiscale entropy (MSE) analysis is a new method of measuring the complexity of finite length time series. We have developed and applied MSE for the analysis of physiologic time series, for which we prefer to estimate entropy using the sample entropy measure. They both have been widely used for the analysis of physiologic data sets. Traditional entropy measures quantify only the regularity (predictability) of time series on a single scale. Neither completely predictable (e.g., periodic) signals, which have minimum entropy, nor completely unpredictable signals, which have maximum entropy, are truly complex, since they can be described very compactly.

For example, we and others have observed that traditional single-scale entropy estimates tend to yield lower entropy in time series of physiologic data such as inter-beat (RR) interval series than in surrogate series formed by shuffling the original physiologic data. This happens because the shuffled data are more irregular and less predictable than the original series, which typically contain correlations at many time scales. The process of generating surrogate data destroys the correlations and degrades the information content of the original signal; if one supposes that greater entropy is characteristic of greater complexity, such results are profoundly misleading.

The MSE method incorporates two procedures:

1. A “coarse-graining” process is applied to the time series. For a given time series, multiple coarse-grained time series are constructed by averaging the data points within non-overlapping windows of increasing length.

2. SampEn is calculated for each coarse-grained time series, and then plotted as a function of the scale factor

IV. MSE ANALYSIS OF SIMULATED WHITE AND 1/F NOISE

The MSE results for simulated uncorrelated (white) and long-range correlated (1/f) noise. Note that for scale one, a higher value of SampEn is obtained for white noise time series than for 1/f time series. Although the value of entropy for the coarse-grained 1/f series remains almost constant for all scales, the value of entropy for the coarse-grained white noise

time series monotonically decreases, such that for scales above 4, it becomes smaller than the corresponding values for 1/f noise. In contrast with the conclusions drawn from single-scale entropy-based analyses, the MSE results are consistent with the fact that, unlike white noise, 1/f noise contains correlations across multiple time scales and is, therefore, more complex than white noise.

V. SOFTWARE FOR MSE ANALYIS

mse.c, the C language source is a program that performs multiscale entropy analysis. The program can be compiled using any ANSI/ISO C compiler, and should be linked to the C math library.

VI. APPROXIMATE ENTROPY

An **approximate entropy** (**ApEn**) is a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data.

For example, there are two series of data:

series 1: (1,2,1,2,1,2,1,...), which alternates 10 and 20.

series 2: (1,1,2,1,2,2,2,1,1,...), which has either a value of 1 or 2, chosen randomly, each with probability 1/2.

Moment statistics, such as mean and variance, won't be able to distinguish between these two series. Nor will rank order statistics distinguish between these two series. Yet series 1 is "perfectly regular"; knowing one term has the value of 20 enables one to predict with certainty that the next term will have the value of 10. Series 2 is randomly valued; knowing one term has the value of 20 gives no insight into what value the next term will have.

Regularity was originally measured by exact regularity statistics, which has mainly centered on various entropy measures. However, accurate entropy calculation requires vast amounts of data, and the results will be greatly influenced by system noise.ApEn was developed by Steve M. Pincus to handle these limitations by modifying an exact regularity statistic , ApEn was initially developed to analyze medical data, such as heart rate, and later spread its applications in finance, psychology, and human factors engineering.

- Some of the advantages are:
- Samples and can be applied in real time.
- Less effect from noise. If data are noisy, the ApEn measure can be compared to the noise level in the data to determine, what the quality of true information is.
- ApEn has been applied in psychiatric diseases, such as schizophrenia, epilepsy, and addiction.

VII. PREDITION AND ENTROPY OF PRINTED ENGLISH

This is the method to estimate the entropy and redundancy of a language. This all depends upon experimental results in prediction of the next letter when the preceding text is known. We here discuss the results of experiments and some properties of ideal predictor.

The entropy is a statistical language measures how much information is produced on an average for each letter of text in the language. The redundancy measures the amount of constraints imposed on a text in the language due to statistical structure. The method of prediction is based on the study of predictability of language. By combining the experimental and theoretical results we shall arrive at common conclusion. From this analysis we will be able to reduce entropy to one bit per letter. But the redundancy may still be higher. But as the more lengthy letters are involved the question becomes more erratic and then it depends on the type of text involved.

The entropy extending due to stats extending over N adjacent letters is given by.

Fn = -sum p(b,j) logp(j)\

Where,

B is the block of N-1 letters;

J is arbitrary letter following b;

P(b,j) is probability of N gram b,j;

P(j) is the conditional probability of letter j after block b and is given by

P(b,j)/p(b).

However the word frequencies have been tabulated and can be used for further approximation. The graph plotted of the probabilities of word against frequency rank .the most frequent word in English is "the" and has probability .071 the next is "of" which has a probability of .034.

Using log scales we find the straight line graph with slope as 1.

Thus Pn of most frequent word is

Pn= 0.1 / n.

Example

Select a short passage unfamiliar who is predicting. He is then asked to guess the first latter in the passage. If he is correct he proceed with the second letter if not he is so informed to correct first and proceed to next letter. This is continued throughout text. As we progress the subject writes down the correct text to the current point for use in predicting future letters. The results are shown below. Spaces were included as an additional letter making a 27 letter alphabet.

THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG

ROO NOT V I SM OBL

Off total 69% were guess correctly. The error as expected occurs more frequently at the beginning of words. Where the line of thought has more possibility of brred, to be an encoded form of the original, the result of passing the original text through a reversible transducer. In fact, acanching out. In general good prediction does not require knowledge of more than N preceding letter of text with N fairly small there are finite no of possibilities of N.to put this another way, the reduced text can be considered to be an encoded form of the original, the result of passing the original text through a reversible transducer.in fact, a communication system could be constructed in which only the reduced text is transmitted from one point to the other. This could be set up as shown in figure, with two identical prediction devices.

Extension of this experiment yields further information concerning the predictability of English .As before the subject knows the text up to the current point and is asked to guess the next letter. If he is wrong, he is told so and asked to guess again. This is continued until he finds the correct letter. A typical result with this experiment is shown below. First line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

(1) T H E R E I S N O R E V E R S E O N A M O T O R C Y C

(2) 1 1 1 5 1 1 2 1 1 2 1 1 151 1 7 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1

(1) F R I E N D O F M I N E F O U N D T H I S O U T

(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1

(1) R A T H E R D R A M A T I C A L L Y T H E O T H E R

LE A

1 l 3 j

D A Y

(2) 4 1 1 1 1 1 1 11 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1 (9)

Out of 102 symbols the subject guessed right on the first guess 79 times on the second guess 8 times, on the third guess 3 times, the fourth and fifth guesses 2 each and only eight times required more than five guesses. Results of this order are typical of prediction by a good subject with ordinary literary English. Newspaper writing, scientific work and poetry generally lead to somewhat poorer scores. The reduced text in this case also contains the same information as the original. Again utilizing the identical twin we ask him at each stage to guess as many times as the number given in the reduced text and recover in this way the original. To eliminate the human element here we must ask out subject, for each possible .V-gram of text, to guess the most probable next letter, the second most probable next letter, etc. This set of data can then serve both for prediction and recovery.

VIII. LIMITATIONS

The ApEn algorithm counts each sequence as matching itself to avoid the occurrence of ln(0) in the calculations. This step might cause bias of ApEn and this bias causes ApEn to have two poor properties in practice:

•ApEn is heavily dependent on the record length and is uniformly lower than expected for short records.

• ApEn lacks relative consistency. That is, if ApEn of one data set is higher than that of another, it does not, remain higher for all conditions tested.

IX. ANALYSIS OF CARDIAC DATA USING MULTISCALE ENTROPY

Analysis of cardiac data using multiscale entropy will reproduce the analysis . In this method we use data which consist of consecutive heartbeat intervals (R–R intervals) derived from young healthy patients, patients with congestive heart failure and patients with a trial fibrillation. The results show distinctive signatures for all three different cases. The signatures were obtained by using MSE and subsequent averaging over all patients. For the averaging of the healthy heart signals and the congestive heart failure data we used 15 data sets each. For the a trial fibrillation we have only one case. The time series were filtered to remove spurious outliers. The differences in the MSE signature are striking. However, we note that we also found cases of congestive heart failure with a signature of a young healthy heart. Moreover, although the differences between the three groups are striking for their averages, we find that MSE is not a reliable diagnostic tool for individual patients. The MSE-signatures of individual patients are very hard to associate with the signatures of their averages. However, we note that the authors of stress in that MSE is rather a detector of complexity than a diagnostic tool. The question of whether MSE can be used as a diagnostic tool to discriminate between healthy and pathological patients had been discussed. It is argued that the (averaged) signature of multiscale entropy for elderly healthy patients is similar to the (averaged) signature of patients with congestive heart failure, and hence multiscale entropy is not a good diagnostic tool for cardiac conditions. In a rebuttal, the authors argue that the purpose of multiscale entropy is rather to quantify complexity than to provide a diagnostic tool. Fluctuations of aging and of pathological systems show a lesser degree of complexity when compared with healthy systems.

The underlying hypothesis is that a healthy physiological system needs to exhibit processes which run on several different time scales to be able to adapt to an ever changing environment. It is in this sense that complexity is a sign of health. Multiscale analysis is able to quantify the degree of complexity as argued multiscale analysis is applied to binary DNA sequences and synthetic computer codes to quantify complexity., it is stressed that complexity rather than irregularity is investigated using multiscale entropy. This allows discriminating between a trial fibrillation and congestive heart failure. It is argued that increased irregularity is not necessarily associated with increased complexity. In the next section, we extend this point and show that one has to be careful with drawing any conclusions from multiscale entropy on any dynamical properties such as irregularity or regularity, and certain a priori knowledge on the dynamical time-scales involved is needed to draw conclusions. In the spirit of, we see multistage analysis as a general tool to study complexity.

The paper is an attempt to model and understand the MSE of cardiac data. This is carried out using electrocardiogram data. Having revisited the analysis of physiological cardiac data by means of multiscale entropy proposed Signatures of the multiscale entropy for the three cases of young healthy hearts, a trial fibrillation and congestive heart failure were identified. We discussed the interpretation of these signatures.

By studying the Lorenz equation as a simple deterministic chaotic system, we showed that for a dynamical system the MSE signature one obtains from a time series depends on the sampling time interval, the correlation time and the period of oscillation, if any frequencies are present. For example, sampling rate can cause decorrelation, suppress periodicities, thus effects the observed MSE signature. This sounds a bell of caution to the reader to be wary of drawing definite conclusions about the nature of the underlying dynamical system from the MSE signature of a sampled time series without detailed knowledge of the different time scales involved. The degree of complexity depends crucially on the time scales under consideration.

However, we note that in the case of cardiac data which is a time series of the interbeat intervals, the sampling time interval obviously cannot be varied.

Whether periodicities in a chaotic signal result in the MSE signature of congestive heart failure depends on dynamics.

X. CONCLUSION

This paper presents commonly used techniques to identify and calculate the entropy of certain type of data. After studying this paper a comparative study can be conducted to understand that no method outperforms the other. However more research is needed to improve the quality against noise and other environmental factors. The target of this paper is to provide with uses and researches related to entropy and how it can be used to implement tomorrow's world.

REFERENCES

[1] Hazewinkel , Michiel, ed. (2001), "Entropy", Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4

[2] Introduction to entropy and information on Principia Cybernetica Web

[3] Entropy an interdisciplinary journal on all aspect of the entropy concept. Open access.

[4] Description of information entropy from "Tools for Thought" by Howard Rheingold

[5] A java applet representing Shannon's Experiment to Calculate the Entropy of English

[6] Slides on information gain and entropy

[7] *An Intuitive Guide to the Concept of Entropy Arising in Various Sectors of Science* – a wikibook on the interpretation of the concept of entropy.

[8] Calculator for Shannon entropy estimation and interpretation

[9] A Light Discussion and Derivation of Entropy

[10] Network Event Detection With Entropy Measures, Dr. RaimundEimann, University of Auckland, PDF; 5993 kB – a PhD thesis demonstrating how entropy measures may be used in network anomaly detection.

[11] C. (2004), *Information Measures: Information and its Description in Science and Engineering*, Springer.

[12] Cover, T. M., Thomas, J. A. (2006), *Elements of information theory*, 2nd Edition. Wiley-Interscience.

[13] Gray, R. M. (2011), *Entropy and Information Theory*, Springer.

[14] Martin, Nathaniel F.G. & England, James W. (2011). *Mathematical Theory of Entropy*. Cambridge University Press.

[15] Shannon, C.E., Weaver, W. (1949) *The Mathematical Theory of Communication*, Univ of Illinois Press.

[16] Stone, J. V. (2014), Chapter 1 of Information Theory: A Tutorial Introduction, University of Sheffield, England.

[17] M. Costa, A.L. Goldberger, C.-K. Peng, Multiscale entropy analysis of complex physiological time series, Phys. Rev. Lett. 89 (2002)068102.

[18] Spotlight issue on chaos in the cardiovascular system, Cardiovasc. Res. 31 (1996).R.A. Thuraisingham, G.A. Gottwald /PhysicaA

[19] Focus issue: fibrillation in normal ventricular myocardium, Chaos 8 (1998).

[20] Focus issue: mapping and control of complex cardiac arrhythmias, Chaos 12 (2002).

[21] Special issue on the heart, Chaos, Solitons Fractals 13 (2002) 1579–1762.

[22] A.L. Goldberger, Is the normal heartbeat chaotic or homeostatic?, News Physiol. Sci. 6 (1991) 87–91.

[23] F.X. Witkowski, K.M. Kavanagh, P.A. Penkoske, R. Plonsey, M.L. Spano, W.L. Ditto, D.T. Kaplan, Evidence for determinism in ventricular fibrillation, Phys. Rev. Lett. 75 (1995) 1230–1233.

[24] G. Sugihara, W. Allan, D. Sobel, K.D. Allan, Nonlinear control of heart rate variability in human infants, Proc. Natl. Acad. Sci. USA 93 (1996) 2608–2613.

[25] T.A. Denton, G.A. Diamond, R.H. Helfant, S. Khan, H. Karagueuzian, Fascinating rhythm: a primer on chaos theory and its application to cardiology, Am. Heart. J. 120 (1990) 1419–1440.

[26] J.E. Skinner, A.L. Goldberger, G. Mayer-Kress, R.E. Ideker, Chaos in the heart: implications for clinical cardiology, Biotechnol.