Entropy as a measure for Signal Processing

Project Report

Nayan Telrandhe
20PHMP46

M.Sc. Physics 2020-22, School of Physics, University of Hyderabad, India. | *Guide- Dr P Manimaran* | *Submitted -* ***9 June 2022***

Abstract

Entropy is one of the basic quantities in Physics. It can be defined based on an approach either of Classical Thermodynamics or Statistical Mechanics. The Statistical definition has an analogue in Information Theory named Shannon Entropy (SE) which is more general. A modified version of SE can be used with signals to analyse them. The method used in the project symbolises/discretises signals into a set of characters or words on the basis of which the SE is calculated for that signal/time series. A modified version of SE is introduced based on the works [3][4][5]. The quantity can be used to distinguish two signals symbolised based on some parameters of analysis. Identification of Epileptic seizures through Electroencephalograms (EEG) signals in realtime and before having them stands as problem. The method can be used in an approach to solve the problem partly. Work of *Aziz and Arif* [3] was reproduced during the project. The method was applied on Electrocardiogram signals of Congestive Heart Failure patients and Normal functioning patients in order to distinguish them and useful results were obtained. It can be shown that on proper choice of analysis parameters the method gives useful result.

Introduction

Entropy is a central quantity and concept to the Second Law of Thermodynamics which can be stated as

" The entropy of isolated systems left to spontaneous evolution cannot decrease with time, as they always arrive at a state of thermodynamic equilibrium, where the entropy is highest." [6]

Entropy can be defined based on two approaches macroscopic and microscopic. Macroscopic approach of Classical Thermodynamic defines entropy based on macroscopically measurable properties like Temperature, Pressure, Volume and Mass whereas the microscopic approach of Statistical Mechanics defines entropy based statistical picture of motion of particles, macro states and micro states. Statistically Entropy corresponding to a macro state is defined based on the number of micro states of the macro state.

Ludwig Boltzmann developed a statistical perspective of the Entropy. Boltzmann showed that this definition of entropy was equivalent to the thermodynamic entropy to within a constant factor known as Boltzmann's constant 'k_B'.

$$S = k_B \, . \, ln(w)$$

where *S* is the entropy and w is the number of micro state which are the possible configurations the state could be in corresponding to the macro state for an ideal gas. The formula was written in the current form by Max Planck and first formulated by Boltzmann [8]. The micro states according to principle of equal a priori probability are all equally probable for such system described by the above formulation.

For a system whose macro state has a distribution of micro states in discrete sets the Entropy can be given by Gibbs Formulation as

$$S = -k_B \sum_i p_i \,.\, ln(p_i)$$

Where p_i is the probability of the i^{th} micro state.
Or over a continuous distribution as

$$S = -k_B \int f(p(x)) \,.\, \log p(x) \,.\, dx$$

Claude Shannon defined a similar statistical quantity which is used in the Information Theory. The quantity was named entropy analogous to Statistical Mechanics upon suggestion of John von Neumann. This description has been identified as a universal definition of the concept of entropy [6]. The average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. The way of defining entropy in above words or other depends on the context and problem.

Given a discrete random variable X, which takes values in the alphabet x and is distributed according to $p_i : x \mapsto [0,1]$:

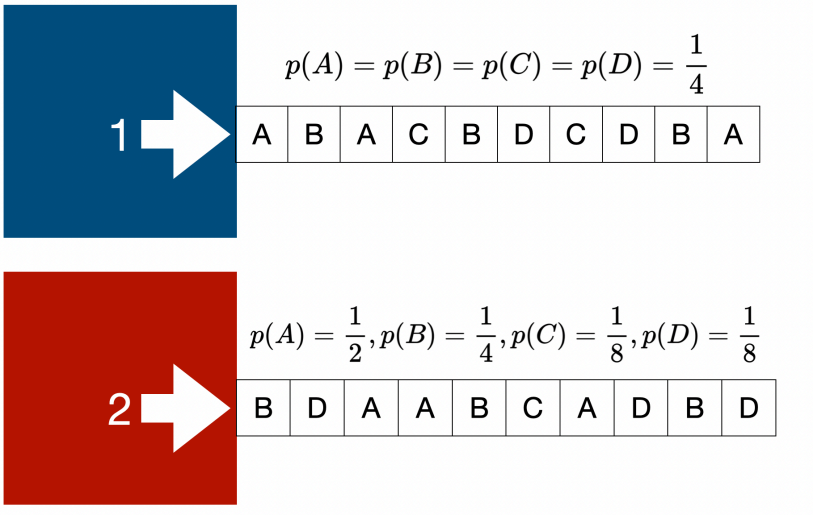
$$S(X) = -\sum_i p_i(X) \cdot \log_2 p_i(X)$$

One way of making sense of Shannon entropy as a measure of information is by the following thought experiment [7]. Let there be two machines 1 and 2 producing outputs A,B,C,D with some associated probabilities as,

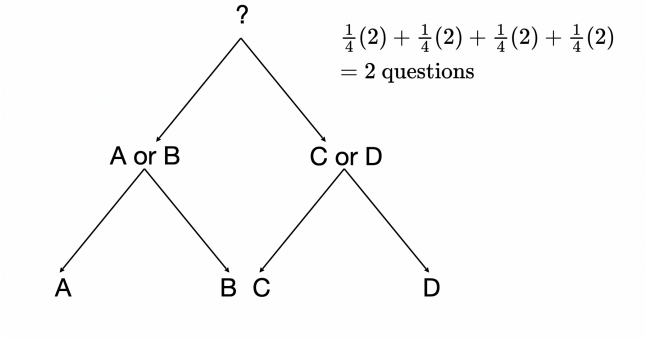
For machine 1, $p(A) = p(B) = p(C) = p(D) = \frac{1}{4}$ i.e. all outputs being equally likely.

Whereas for machine 2, $P(A) = \frac{1}{2}, P(B) = \frac{1}{4}, P(C) = \frac{1}{8}, P(D) = \frac{1}{8}$.

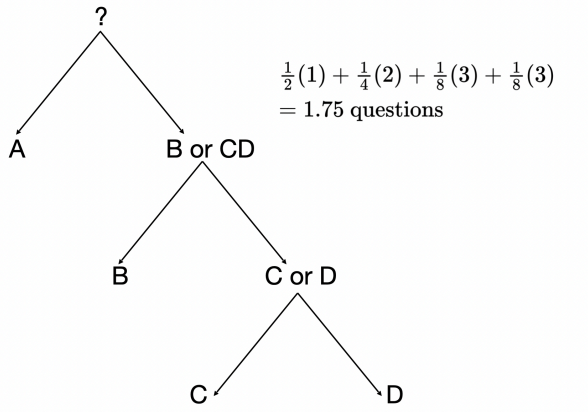
When an output is drawn the average number of question to be asked turns out to be the entropy of the system, which is related to the amount of information [7].



Machine 1



Machine 2



The above question trees can be formed based on the probability associated with the outcomes. It can be seen that the entropy for system with equal probabilities for all outcomes is greater than that of the other. Which in theory should be maximum [8]. It is similar to entropy of a Thermodynamic system which is maximum when all the states are equally likely.

Method

The Method used has been called Symbolic time series analysis [3]. The signal is first discretised followed by symbolisation of this discretised signal.The symbolised sequence is analysed to know more about the state of the system. The quantity used for the analysis is called NCSE (Normalised Corrected Shannon Entropy). The NCSE can be used to distinguish between the two different type of signals. The method uses some parameters of analysis and the difference in the two signals appears to be maximum at some set of these parameters.

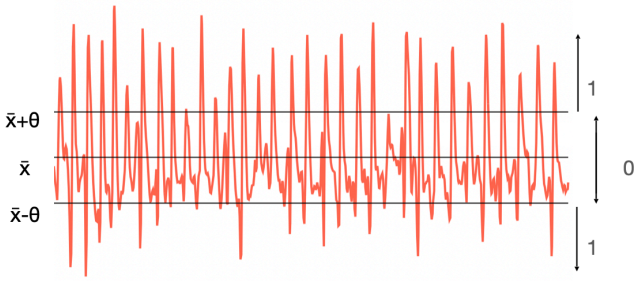
1. To discretise a time series $T(x)$ the mean $\bar{x}$ is calculated and a threshold θ has to be chosen. Using these two values the new sequence $T(y)$ is defined as

$$T_x = \{x_i, i = 1,2,\ldots, N\}$$
$$\downarrow$$
$$T_y = \{y_i, i = 1,2,\ldots N\}$$

where, $y_i = 1$, $if \left| x_i - \bar{x} \right| \geq \theta$

$y_i = 0$, $if \left| x_i - \bar{x} \right| < \theta$

$\bar{x}$ is the mean of T_x



Example: $T_y = \{0,1,1,0,1,0,0,\ldots\}$

2. The elements of the sequence are grouped based on a word length that is chosen. The word length is an integer ($w = 1,2,3,4,..$). The elements are successively grouped (i.e. moving the grouping window one element at a time) according to the length. The obtained word sequence thus has $w - 1$elements less than the previous. Example:

   $T_w = \{(0110), (1101), (1010), (0100)\ldots\}$ for wordlength = 4 on the above T_y

3. The words form a binary number. This number is to be converted to an equivalent decimal number.This symbolic sequence obtained after conversion is a symbolic representation of the original time series.The number of different possible words in the found sequence is based on the word length.
4. Shannon Entropy can be used as a measure of the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. Shannon entropy for our sequence with words being possible outcomes is given by.

$$SE(T_w) = -\sum_i p_i(T_w) \cdot \log_2 p_i(T_w)$$

where $p_i(T_w)$, is the probability of occurrence of the word (denoted by i) or equivalent decimal number in the sequence.

A correction to the Shannon entropy is suggested by [4] to counter the systematic error and random error which affect the estimates computed by using **Shannon entropy**.

$$\mathrm{CSE}\left(T_w\right) = \mathrm{SE}\left(T_w\right) + \frac{C-1}{2 \cdot M \cdot \ln(2)}$$

Where M are total possible number of words C is the number of words which occurred among possible M words.
To compare sequences generated using different word lengths at same thresholds the Normalised Corrected Shannon Entropy has been proposed by *Aziz Arif* [5] The CSE is maximum when the occurrence of all the words is equally likely [8] thus

$$\mathrm{CSE}\left(T_w\right)_{\max} = -\log_2\left(\frac{1}{M}\right) + \frac{M-1}{2 \cdot M \cdot \ln(2)}$$

NCSE is given by

$$NCSE(T_w) = \frac{CSE(T_w)}{CSE_{\max}(T_w)}$$

The method was performed on two data sets during the project

1. **EEG Data** - Paper [3] was implemented and results were reproduced for EEG data sets to distinguish between Epileptic EEG and Normal, Eye open eye closed using the method mentioned above
2. **ECG Data** - The method was implemented on RR interval data of Normal and Congestive Heart Failure Samples.

Description of the data (EEG)

The method was applied on dataset taken from a publicly available database (made available) by the Department of Epileptology, Bonn University. There are 5 data sets (a, b, c, d, e) each containing 23.6 s duration 100 single-channel segments. Sampling frequency of 173.61 Hz at 12 bit resolution so 4097 data points in each signal.

- Set a - segments from surface EEG of 5 healthy patients during awake state with **eyes open.**
- Set b - segments from surface EEG of 5 healthy patients during awake state with **eyes closed.**
- Set c - segments from the hippocampal formation of the opposite hemisphere of the brain of 5 epilepsy patients during seizure free interval.
- Set d - segments from the epileptogenic zone of 5 epilepsy patients during seizure free interval.
- Set e - segments from 5 **epilepsy** patients during seizure activity.

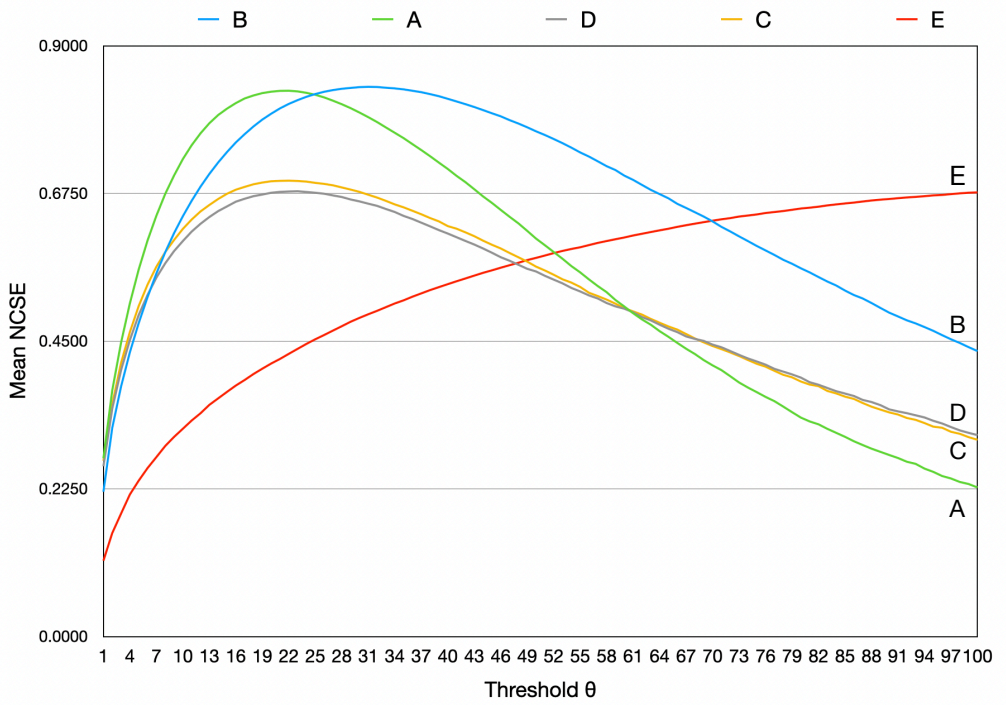
Result

The method was implemented on the datasets with Python using Numpy library to get the results. The table below has the mean NCSE obtained for different sets at different values of the threshold. The mean NCSE values of Healthy EEG (a and b) subjects are higher than that of Epileptic ones (e). The mean NCSE values of eye open set (a) are higher than that of eye closed set (b) during rest states at smaller thresholds**.**The maximum mean NCSE value for the healthy subjects (with eye open set a) was found at a threshold of 30, whereas the maximum threshold for the healthy subject (with eye-closed set b), epileptic seizure-free intervals was found at a threshold of 20. The mean NCSE value for epileptic seizure increases until a threshold of 80.

Mean NCSE for different sets

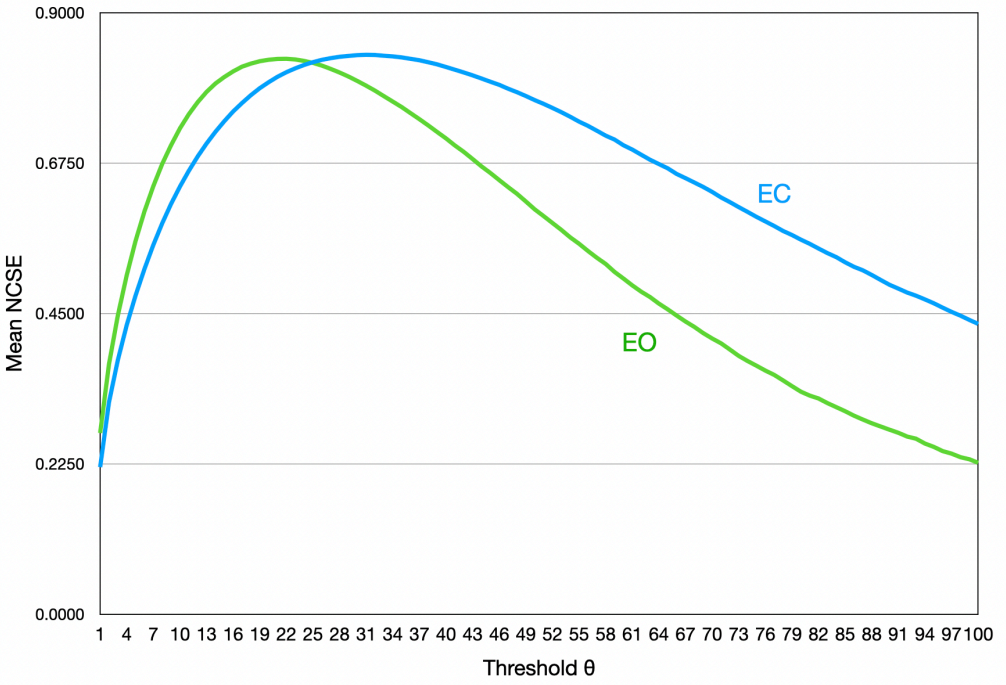
θ	Set b NCSE Mean	Set a NCSE Mean	Set e NCSE Mean	Set d NCSE Mean	Set c NCSE Mean
15	0.73818	0.80431	0.37274	0.65532	0.67461
20	0.79652	0.83010	0.41590	0.67610	0.69364
25	0.82655	0.82543	0.45288	0.67611	0.69118
30	0.83706	0.79836	0.48547	0.66296	0.67747
35	0.83366	0.75935	0.51219	0.64084	0.65391
40	0.81928	0.71242	0.53682	0.61400	0.62534
45	0.79725	0.66071	0.55788	0.58477	0.59692
50	0.76974	0.60628	0.57686	0.55619	0.56360

The obtained values are plotted at different thresholds for the sets.



It can be seen that the value of mean NCSE for Epileptic seizure dataset changes differently than for other datasets. It keeps on increasing as the threshold is increased.

The mean NCSE values for eye open and closed sets (a and b) are plotted, the maximum for both occur at different threshold values and NCSE for a set is higher than b for threshold less than 25, whereas it stays less thereafter.

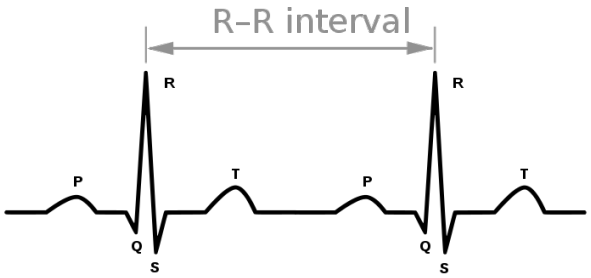


The entropy as obtained for different data sets depends on the threshold and word length as the analysis parameters. During an epileptic seizure there is an excessive and synchronous neuronal activity. The activity can make the signal very regular compared to normal thus the obtained waveform and have value higher than non epileptic signals thus it should have comparatively slower rate of increase of entropy with threshold and keep increasing while entropy for other signals decrease. The obtained result go along with the expectations.

Description of the data (ECG)

The method was applied on datasets publicly available at PhysioNet [1][2] .Three different sets are available which are normal, congestive heart failure and atrial fibrillation conditions of which the first two sets have been used. There are 5 heart beat time series in Normal heart (n_1, n_2, n_3, n_4) and Congestive heart failure (c_1, c_2, c_3, c_4, c_5) each. The series has R-R interval data. Each series is around 24 hours long. Each time series in the data set has an unfiltered (rr.txt) and a filtered (nn.txt) file [1] and analysis has been done for both of those. The filtered have been cleared of outliers.

ECG signal is comprised of different components of periodic pulses one being the R-waves. R-R interval refers to the time elapsed between two successive R-waves of ECG signal



Result

The Normalised Shannon Entropy was calculated for the normal and congested heart failure data (both filtered and unfiltered). The mean was calculated.The threshold and the word length was varied for calculation of NCSE. The threshold was varied from 0.05-0.3 at an interval of 0.05. These sequences were further symbolised for 4 different word lengths. The two tables given are for the filtered and unfiltered data.The tables have NCSE values for symbolised sequences of different word lengths and different thresholds (mean).

NCSE Filtered

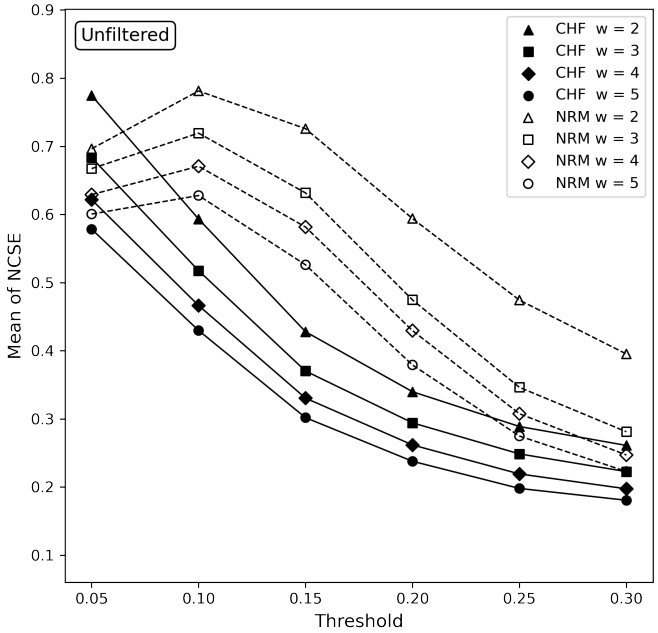
Word length	Set	0.05	0.1	0.15	0.2	0.25	0.3
2	C1	0.8236	0.6323	0.4578	0.3117	0.2520	0.2214
	C2	0.7064	0.4192	0.3170	0.2590	0.2253	0.2203
	C3	0.7810	0.4516	0.2414	0.2155	0.2341	0.2338
	C4	0.7553	0.4331	0.2694	0.2436	0.2182	0.2139
	C5	0.6568	0.7153	0.4668	0.3511	0.2552	0.2225
3	C1	0.7340	0.5375	0.3872	0.2618	0.2094	0.1826
	C2	0.6122	0.3467	0.2632	0.2153	0.1863	0.1973
	C3	0.6662	0.3866	0.2000	0.1833	0.2145	0.2142
	C4	0.6610	0.3671	0.2184	0.1993	0.1787	0.1898
	C5	0.5565	0.6135	0.3883	0.2946	0.2123	0.1827
4	C1	0.6658	0.4720	0.3388	0.2263	0.1782	0.1590
	C2	0.5521	0.2996	0.2259	0.1835	0.1572	0.1777
	C3	0.5922	0.3416	0.1695	0.1667	0.2009	0.2005
	C4	0.6035	0.3229	0.1829	0.1673	0.1544	0.1751
	C5	0.4947	0.5512	0.3385	0.2566	0.1812	0.1532
5	C1	0.6177	0.4263	0.3043	0.2000	0.1547	0.1439
	C2	0.5110	0.2663	0.1989	0.1598	0.1420	0.1627
	C3	0.5393	0.3090	0.1467	0.1561	0.1908	0.1903
	C4	0.5636	0.2910	0.1570	0.1435	0.1387	0.1650
	C5	0.4516	0.5071	0.3033	0.2285	0.1579	0.1310
2	N1	0.6509	0.7765	0.7859	0.7213	0.5641	0.3863
	N2	0.7056	0.7840	0.6995	0.5199	0.3711	0.2945
	N3	0.7799	0.8353	0.6931	0.4402	0.2580	0.2204
	N4	0.6122	0.7545	0.8167	0.7725	0.7083	0.6280
	N5	0.7914	0.7774	0.5595	0.3280	0.2392	0.2167
3	N1	0.5997	0.7019	0.6976	0.6412	0.5031	0.3387
	N2	0.6353	0.6897	0.6092	0.4459	0.3087	0.2404
	N3	0.7275	0.7643	0.6334	0.3944	0.2195	0.1814
	N4	0.5663	0.6860	0.7310	0.6823	0.6308	0.5671
	N5	0.7249	0.7034	0.4993	0.2804	0.1965	0.1773
4	N1	0.5664	0.6565	0.6438	0.5894	0.4587	0.3035
	N2	0.5907	0.6322	0.5522	0.3966	0.2670	0.2034
	N3	0.6892	0.7145	0.5850	0.3567	0.1898	0.1522
	N4	0.5340	0.6416	0.6769	0.6220	0.5733	0.5166
	N5	0.6838	0.6579	0.4595	0.2466	0.1653	0.1530
5	N1	0.5426	0.6255	0.6077	0.5542	0.4278	0.2781
	N2	0.5596	0.5932	0.5129	0.3622	0.2372	0.1766
	N3	0.6613	0.6785	0.5483	0.3281	0.1671	0.1332
	N4	0.5097	0.6101	0.6384	0.5780	0.5280	0.4731
	N5	0.6547	0.6258	0.4306	0.2217	0.1431	0.1372

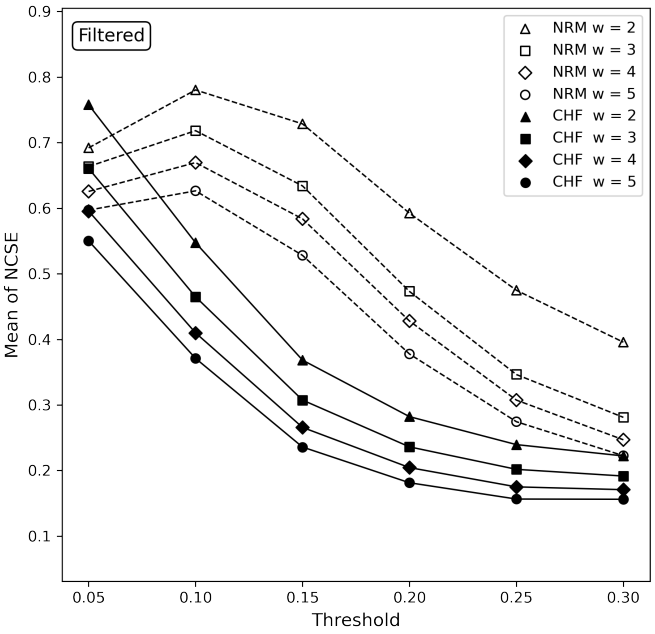
NCSE Unfiltered

Word length	Set	0.05	0.1	0.15	0.2	0.25	0.3
2	C1	0.8290	0.6503	0.4799	0.3334	0.2699	0.2349
	C2	0.7657	0.5178	0.4227	0.3699	0.3352	0.3160
	C3	0.7918	0.5474	0.3928	0.3667	0.3495	0.3221
	C4	0.7740	0.4777	0.3241	0.2838	0.2522	0.2352
	C5	0.6571	0.7159	0.4677	0.3517	0.2558	0.2228
3	C1	0.7430	0.5608	0.4142	0.2863	0.2289	0.1962
	C2	0.6791	0.4560	0.3758	0.3295	0.2972	0.2786
	C3	0.6925	0.4953	0.3498	0.3256	0.3102	0.2849
	C4	0.6857	0.4178	0.2777	0.2419	0.2132	0.1964
	C5	0.5569	0.6142	0.3893	0.2953	0.2130	0.1831
4	C1	0.6772	0.4986	0.3685	0.2524	0.1985	0.1770
	C2	0.6242	0.4148	0.3429	0.3006	0.2696	0.2511
	C3	0.6254	0.4569	0.3189	0.2963	0.2820	0.2577
	C4	0.6310	0.3769	0.2441	0.2114	0.1844	0.1676
	C5	0.4952	0.5520	0.3397	0.2573	0.1819	0.1536
5	C1	0.6310	0.4552	0.3359	0.2274	0.1756	0.1654
	C2	0.5862	0.3846	0.3185	0.2791	0.2490	0.2306
	C3	0.5767	0.4283	0.2964	0.2748	0.2632	0.2432
	C4	0.5930	0.3473	0.2194	0.1887	0.1652	0.1470
	C5	0.4522	0.5079	0.3044	0.2292	0.1587	0.1315
2	N1	0.6509	0.7765	0.7866	0.7228	0.5664	0.3911
	N2	0.7285	0.7868	0.6806	0.5275	0.3669	0.2911
	N3	0.7800	0.8354	0.6931	0.4404	0.2583	0.2204
	N4	0.6131	0.7553	0.8176	0.7727	0.7073	0.6258
	N5	0.7920	0.7776	0.5595	0.3276	0.2393	0.2169
3	N1	0.5997	0.7020	0.6984	0.6428	0.5055	0.3431
	N2	0.6554	0.6937	0.5921	0.4525	0.3053	0.2375
	N3	0.7277	0.7646	0.6336	0.3946	0.2197	0.1814
	N4	0.5672	0.6869	0.7323	0.6828	0.6305	0.5655
	N5	0.7256	0.7037	0.4994	0.2801	0.1966	0.1774
4	N1	0.5665	0.6568	0.6447	0.5910	0.4618	0.3077
	N2	0.6092	0.6368	0.5362	0.4029	0.2640	0.2008
	N3	0.6894	0.7148	0.5851	0.3569	0.1901	0.1522
	N4	0.5349	0.6427	0.6783	0.6228	0.5733	0.5155
	N5	0.6846	0.6583	0.4596	0.2464	0.1656	0.1531
5	N1	0.5427	0.6258	0.6087	0.5559	0.4313	0.2821
	N2	0.5773	0.5982	0.4974	0.3682	0.2346	0.1741
	N3	0.6615	0.6789	0.5484	0.3284	0.1674	0.1332
	N4	0.5106	0.6112	0.6400	0.5790	0.5283	0.4727
	N5	0.6554	0.6262	0.4308	0.2215	0.1434	0.1374

There's a difference between the NCSE values for the filtered and unfiltered data for same parameters of analysis (θ, w). The value of NCSE decreases when word length is increased after 0.10 value of the threshold. The value of NCSE is greater for Normal beats than the CHF ones for all samples for threshold having value 0.10 to 0.25 for both filtered and unfiltered data.

The mean of all 5 sequences of congested heart failure beats and normal beats are calculated for different analysis parameters and are plotted.





The plot 1 and plot 2 have the mean of NCSE values of CHF and NH beat data sets separately plotted against threshold. It can be seen that the NCSE values of NH beats have entropy higher than that of the CHF beats. The difference is most prominent at word length of 2 and threshold of 0.15 for both the filtered and unfiltered data. We choose w=2 and and theta=0.15 as the parameter values of method for given data set.

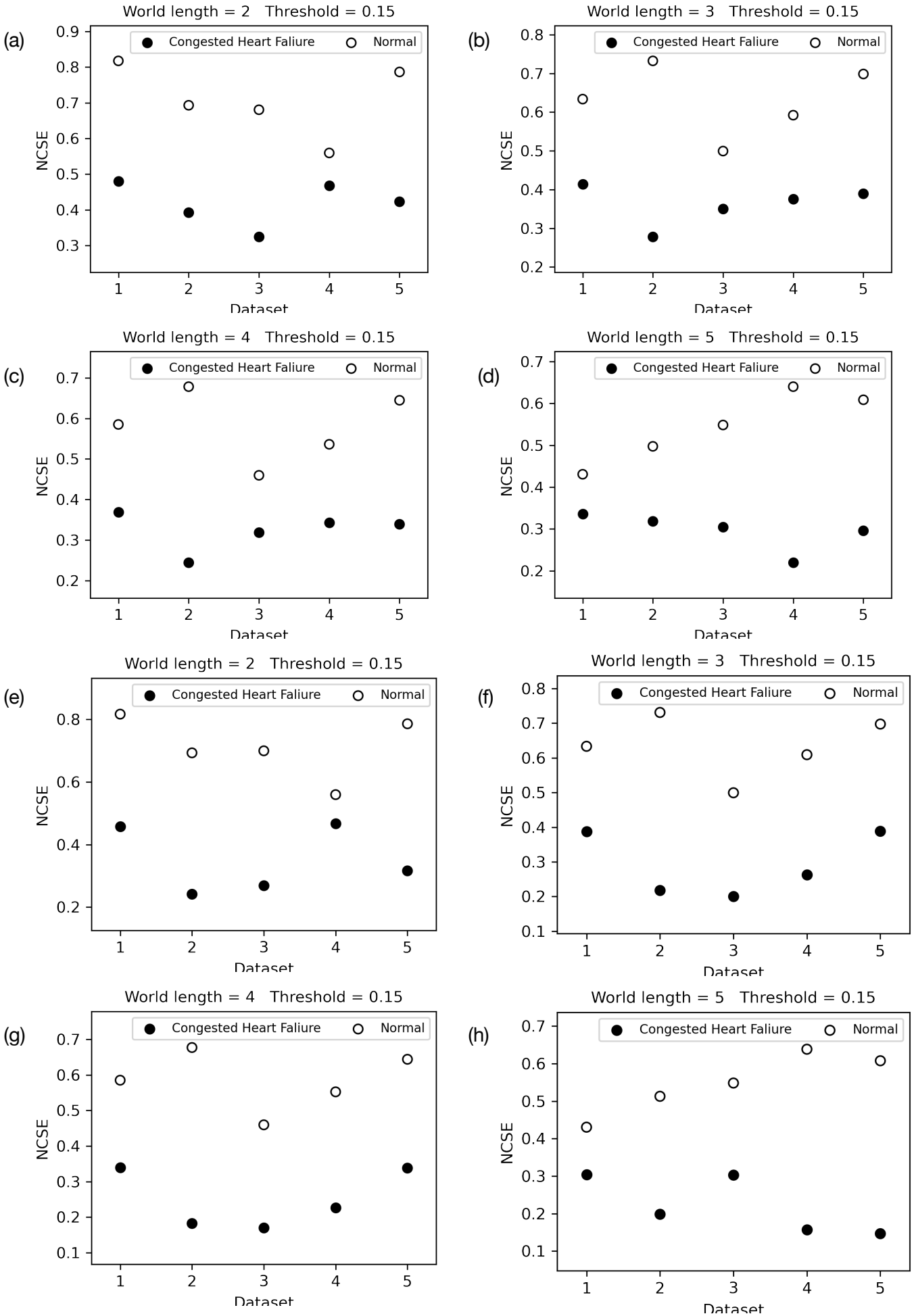


Fig: NCSE value plots for Datasets Unfiltered (a, b, c, d) and Filtered (e, f, g, h) at threshold 0.15 and various word lengths.

Tile - a, b, c, d and Tile - e, f, g, h have NCSE values plotted for Unfiltered and Filtered data respectively.The data is plotted at different word lengths and thresholds. The CHF beats can be distinguished from NH beats in all the plots. The NCSE for NH beats (empty circles) seems significantly higher than that of the CHF beats (filled circles).

The method on providing successful distinction of Normal and CHF Heart Beats can be used for medical diagnosis and further for Congestive Heart failure.

Summary

During the Project work I learnt about different Biomedical Signals, specifically EEG signals. How these signals are obtained, factors they are based on and how epilepsy affects them. I learnt to deal with the data using Python. Numpy was used extensively for most of the work. Discretisation of the continuous signal was done for statistical analysis of the signal. Shannon Entropy was used for data analysis in the method and a better understanding of Entropy was obtained in general. The method was used on a publicly available dataset and the obtained results are to be published. The work can be bettered by basic understanding of waveforms and used method or Shannon entropy over it. During the project time other problems like using Complex Networks for studying Cosmological structures were explored.

References

[1] Moody, G. (2008). Is the normal heart rate chaotic? (version 1.0.0). PhysioNet. (Data for study)

[2] Glass, Leon. Introduction to Controversial Topics in Nonlinear Science: Is the Normal Heart Rate Chaotic? Chaos 19, 028501 (2009).

[3] Hussain, L., Aziz, W., Alowibdi, J.S. et al. Symbolic time series analysis of electroencephalographic (EEG) epileptic seizure and brain dynamics with eye-open and eye-closed subjects during resting states. J Physiol Anthropol 36, 21 (2017).

[4] Eguia MC, Rabinovich MI, Abarbanel HD. Information transmission and recovery in neural communications channels. Phys Rev E. 2000;62(5):7111.

[5] Aziz W, Arif M. Complexity analysis of stride interval time series by threshold dependent symbolic entropy. Eur J Appl Physiol. 2006 Sep;98(1):30-40. Epub 2006 Jul 14. PMID: 16841202.

[6] Wikipedia/Entropy https://en.wikipedia.org/wiki/Entropy

[7] Khan Academy/Modern Information Theory/Information Entropy https://www.khanacademy.org/computing/computer-science/informationtheory/moderninfotheory/v/information-entropy

[8] Wikipedia/Maximum entropy of an uniform distribution https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution#Uniform_and_piecewise_uniform_distributions

Code and useful resources

The method was implemented in Python. Numpy arrays were extensively used for calculation and data handling and Matplotlib for plotting. Mac OS's Numbers was used for some of the plots. The codes in Jupyter Notebooks are available in the repository at

https://github.com/nayantelrandhe/Symbolization-of-Time-series-and-calculation-of-Entropy

Numpy Documentation - https://numpy.org/doc/
Matplotlib Documentation - https://matplotlib.org/stable/index.html
OS List Documentation - https://docs.python.org/3/library/os.html

Youtube Links about methods, tools and educational materials related to project.
https://youtube.com/playlist?list=PLIeva9XfWEePyOuLX6TaTjM3Sa8pKpEok

Acknowledgement

I am thankful to my Project guide Dr. P Manimaran for providing the resources mentioned in the references for method, data and problems to work on. His timely inputs were helpful.
I am thankful to my friends and classmates to keep me going. We all helped each other in stressful times.
I am thankful to the the people in the references and all over the internet whose content were helpful to me in learning necessary things and to build and remain motivated for the project problems.