On Entropy-based Data Mining

Andreas Holzinger[1] Matthias Hörtenhuber[2] Christopher Mayer[2]
Martin Bachler[2] Siegfried Wassertheurer[2]
Armando J Pinho[3] and David Koslicki[4]

[1] Medical University Graz, A-8036 Graz, Austria
Institute for Medical Informatics, Statistics & Documentation,
Research Unit Human–Computer Interaction
a.holzinger@hci4all.at

[2] AIT Austrian Institute of Technology GmbH, Health & Environment Department,
Biomedical Systems, Donau-City-Str. 1, A-1220 Vienna, Austria
{christopher.mayer,martin.bachler,matthias.hoertenhuber,siegfried.
wassertheurer}@ait.ac.at

[3] IEETA / Department of Electronics, Telecommunications and Informatics,
University of Aveiro, 3810-193 Aveiro, Portugal
ap@ua.pt

[4] Oregon State University, Mathematics Department, Corvallis, OR, USA
david.koslicki@math.oregonstate.edu

Abstract. In the real world, we are confronted not only with complex and high-dimensional data sets, but usually with noisy, incomplete and uncertain data, where the application of traditional methods of knowledge discovery and data mining always entail the danger of modeling artifacts. Originally, information entropy was introduced by Shannon (1949), as a measure of uncertainty in the data. But up to the present, there have emerged many different types of entropy methods with a large number of different purposes and possible application areas. In this paper, we briefly discuss the applicability of entropy methods for the use in knowledge discovery and data mining, with particular emphasis on biomedical data. We present a very short overview of the state-of-the-art, with focus on four methods: Approximate Entropy (ApEn), Sample Entropy (SampEn), Fuzzy Entropy (FuzzyEn), and Topological Entropy (FiniteTopEn). Finally, we discuss some open problems and future research challenges.

Keywords: Entropy, Data Mining, Knowledge Discovery, Topological Entropy, FiniteTopEn, Approximate Entropy, Fuzzy Entropy, Sample Entropy, Biomedical Informatics

1 Introduction

Entropy, originating from statistical physics (see Section 3), is a fascinating and challenging concept with many diverse definitions and various applications.

Considering all the diverse meanings, entropy can be used as a measure for disorder in the range between total order (structured) and total disorder (unstructured) [1], as long as by order we understand that objects are segregated by their properties or parameter values. States of lower entropy occur when objects become organized, and ideally when everything is in complete order the Entropy value is zero. These observations generated a colloquial meaning of entropy [2]. Following the concept of the mathematical theory of communication by Shannon & Weaver (1949) [3], entropy can be used as a measure for the *uncertainty in a data set*. The application of entropy became popular as a measure for system complexity with the paper by Steven Pincus (1991) [4]: He described Approximate Entropy (see Section 5.1) as a statistic quantifying regularity within a wide variety of relatively short (greater than 100 points) and noisy time series data. The development of this approach was initially motivated by data length constraints, which is commonly encountered in typical biomedical signals including: heart rate, electroencephalography (EEG), etc. but also in endocrine hormone secretion data sets [5].

This paper is organized as follows: To ensure a common understanding we start with providing a short glossary; then we provide some background information about the concept of entropy, the origins of entropy and a taxonomy of entropy methods in order to facilitate a "big picture". We continue in chapter 4 with the description of some application areas from the biomedical domain, ranging from the analysis of EEG signals to complexity measures of DNA sequences. In chapter 5 we provide more detailed information on four particular methods: Approximate Entropy (ApEn), Sample Entropy (SampEn), Fuzzy Entropy (FuzzyEn), and Topological Entropy (FiniteTopEn). In chapter 6 we discuss some open problems and we conclude in chapter 7 with a short future outlook.

2 Glossary and Key Terms

Anomaly detection: is finding patterns in data, non compliant to expected behavior (anomalies aka outliers, discordant observations, exceptions, aberrations, surprises, peculiarities). A topic related to anomaly detection is novelty detection, aiming at detecting previously unobserved, emergent patterns in data [6].

Artifact: is any error, anomaly and/or undesired alteration in the perception or representation of information from data.

Data quality: includes (physical) quality parameters including: Accuracy, Completeness, Update status, Relevance, Consistency, Reliability and Accessibility [7], not to confuse with Information quality [8].

Dirty data: data which is incorrect, erroneous, misleading, incomplete, noisy, duplicate, uncertain, etc. [9].

Dirty time oriented data: time (e.g. time points, time intervals) is an important data dimension with distinct characteristics affording special consideration in the context of dirty data [10].

Dynamical system: is a manifold M called the phase-space and possess a family of evolution functions $\phi(t)$ so that for any element of $t \in T$, the time, maps a point of the phase-space back into the phase-space; If T is real, the dynamical system is called a *flow*; if T is restricted to the non-negative reals, it is a semi-flow; in case of integers, it is called a cascade or map; and a restriction to the non-negative integers results in a so-called semi-cascade [2];

Hausdorff Space: is a separated topological space in which distinct points have disjoint neighbourhoods.

Hausdorff Measure: is a type of outer measure that assigns a number in $[0, \infty]$ to each set in $\mathbb{R}^n$. The zero-dimensional Hausdorff measure is the number of points in the set, if the set is finite, or ∞ if the set is infinite. The one-dimensional Hausdorff measure of a simple curve in $\mathbb{R}^n$ is equal to the length of the curve. Likewise, the two dimensional Hausdorff measure of a measurable subset of $\mathbb{R}^2$ is proportional to the area of the set. The concept of the Hausdorff measure generalizes counting, length, and area. These measures are fundamental in geometric measure theory.

Topological Entropy: is a nonnegative real number that is a measure of the complexity of a dynamical system. TopEn was first introduced in 1965 by Adler, Konheim and McAndrew. Their definition was modeled after the definition of the Kolmogorov–Sinai, or metric entropy.

Heart rate variability (HRV): measured by the variation in the beat-to-beat interval of heart beats.

HRV artifact: noise through errors in the location of the instantaneous heart beat, resulting in errors in the calculation of the HRV.

Information Entropy: is a measure of the uncertainty in a random variable. This refers to the Shannon entropy, which quantifies the expected value of the information contained in a message.

3 Background

3.1 Physical Concept of Entropy

It is nearly impossible to write any paper on any aspect of entropy, without referring back to classical physics: The concept of entropy was first introduced in thermodynamics [11], where it was used to provide a statement of the second law of thermodynamics on the irreversibility of the evolution, i.e. an isolated

system cannot pass from a state of higher entropy to a state of lower entropy. In classical physics any system can be seen as a set of objects, whose state is parameterized by measurable physical characteristics, e.g. temperature. Later, statistical mechanics provided a connection between the macroscopic property of entropy and the microscopic state of a system by Boltzmann.

Shannon (1948) was the first to re-define entropy and mutual information, for this purpose he used a thought experiment to propose a measure of uncertainty in a discrete distribution based on the Boltzmann entropy of classical statistical mechanics (see next section). For more details on the basic concepts of entropy refer to [12].

3.2 Origins of Information Entropy

The foundation of information entropy (see Fig. 1) can be traced back into two major origins, the older may be found in the work of Jakob Bernoulli (1713), describing the *principle of insufficient reason*: we are ignorant of the ways an event can occur, the event will occur equally likely in any way. Thomas Bayes (1763) and Pierre-Simon Laplace (1774) carried on with works on how to calculate the state of a system with a limited number of expectation values and Harold Jeffreys and David Cox solidified it in the Bayesian Statistics, also known as **statistical inference**.

The second path is leading to the classical Maximum Entropy, not quite correctly often called "Shannon Entropy", but indeed, Jaynes (1957) [13] makes it clear on page 622/623 that he is utilizing Shannon's Entropy to *derive* the Maximum Entropy Principle and that those are not synonym principles. Following the path backwards the roots can be identified with the work of James Clerk Maxwell (1859) and Ludwig Boltzmann (1871), continued by Willard Gibbs (1902) and finally reaching Claude Elwood Shannon (1948). This work is geared toward developing the mathematical tools for statistical modeling of problems in information. These two independent lines of research are relatively similar. The objective of the first line of research is to formulate a theory and methodology that allows understanding of the general characteristics (distribution) of a given system from partial and incomplete information. In the second route of research, the same objective is expressed as determining how to assign (initial) numerical values of probabilities when only some (theoretical) limited global quantities of the investigated system are known. Recognizing the common basic objectives of these two lines of research aided Jaynes (1957) in the development of his classical work, the Maximum Entropy formalism (see also Fig. 2). This formalism is based on the first line of research and the mathematics of the second line of research. The interrelationship between Information Theory, statistics and inference, and the Maximum Entropy (MaxEnt) principle became clear in the 1950s, and many different methods arose from these principles [14], see Fig. 2.For more details on information entropy refer to [2].

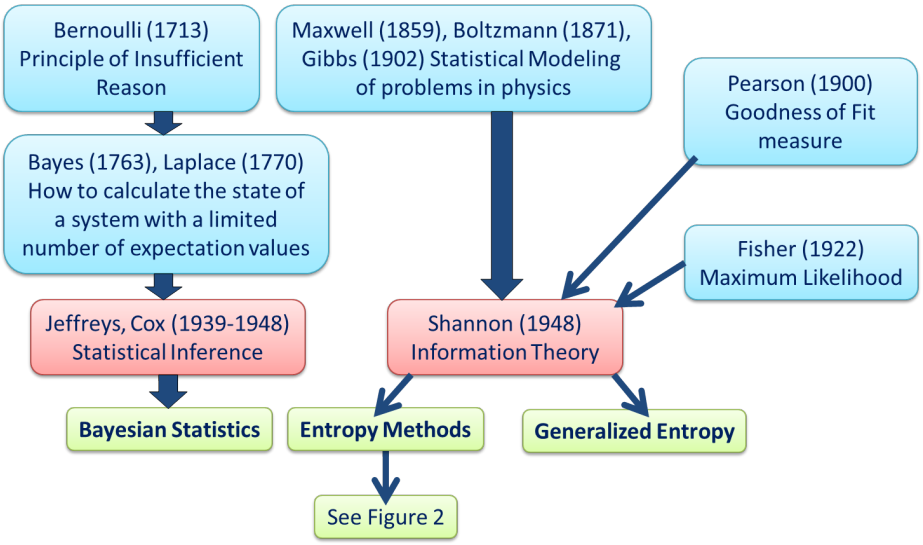

Fig. 1. The "big picture" in the developent of the concept of entropy [15]

3.3 Towards a Taxonomy of Entropy Methods

Maximum Entropy (MaxEn), described by [16], is used to estimate unknown parameters of a multinomial discrete choice problem, whereas the Generalized Maximum Entropy (GME) includes noise terms in the multinomial information constraints. Each noise term is modeled as the mean of a finite set of known points in the interval [-1,1] with unknown probabilities where no parametric assumptions about the error distribution are made. A GME model for the multinomial probabilities and for the distributions, associated with the noise terms is derived by maximizing the joint entropy of multinomial and noise distributions, under the assumption of independence [16].
Graph Entropy was described by [17] to measure structural information content of graphs, and a different definition, more focused on problems in information and coding theory, was introduced by Körner in [18]. Graph entropy is often used for the characterization of the structure of graph-based systems, e.g. in mathematical biochemistry, but also for any complex network [19]. In these applications the entropy of a graph is interpreted as its structural information content and serves as a complexity measure, and such a measure is associated with an equivalence relation defined on a finite graph; by application of Shannons Eq. 2.4 in [20] with the probability distribution we get a numerical value that serves as an index of the structural feature captured by the equivalence relation [20].

Minimum Entropy (MinEn), described by [21], provides us the least random, and the least uniform probability distribution of a data set, i.e. the minimum

uncertainty. Often, the classical pattern recognition is described as a quest for minimum entropy. Mathematically, it is more difficult to determine a minimum entropy probability distribution than a maximum entropy probability distribution; while the latter has a global maximum due to the concavity of the entropy, the former has to be obtained by calculating all local minima, consequently the minimum entropy probability distribution may not exist in many cases [22].

Cross Entropy (CE), discussed by [23], was motivated by an adaptive algorithm for estimating probabilities of rare events in complex stochastic networks, which involves variance minimization. CE can also be used for combinatorial optimization problems (COP). This is done by translating the deterministic optimization problem into a related stochastic optimization problem and then using rare event simulation techniques [24].

Rényi Entropy is a generalization of the Shannon entropy (information theory).

Tsallis Entropy is a generalization of the BoltzmannGibbs entropy and was intended for statistical mechanics by Constantino Tsallis [25]; a decade ago it has been applied to computer science, see e.g. a pattern recognition example [26].

Approximate Entropy (ApEn), described by [4], is useable to quantify regularity in data without any a priori knowledge about the system.

Sample Entropy (SampEn), was used by [27] for a new related measure of time series regularity. SampEn was designed to reduce the bias of ApEn and is better suited for data sets with known probabilistic content.

Fuzzy Entropy (FuzzyEn), proposed by [28], replaces the Heaviside function to measure the similarity of two vectors as used in SampEn and ApEn by a fuzzy relationship function. This leads to a weaker impact of the threshold parameter choice.

Fuzzy Measure Entropy (FuzzyMEn), presented in [29], is an enhancement of FuzzyEn, by differentiating between local and global similarity.

Topological Entropy (TopEn), was introduced by [30] with the purpose to introduce the notion of entropy as an invariant for continuous mappings: Let (X, T) be a topological dynamical system, i.e., let X be a nonempty compact Hausdorff space and $T : X \to X$ a continuous map; the TopEn is a nonnegative number which measures the complexity of the system [31].

Topological Entropy for Finite Sequences (FiniteTopEn) was introduced in [32] by taking the definition of TopEn for symbolic dynamical systems and developing a finite approximation suitable for use with finite sequences.

Algorithmic Entropy or Kolmogorov Complexity was independently introduced by Solomonoff [33,34], Kolmogorov [35] and Chaitin [36]. The algorithmic entropy of a string is formally defined as the length of a shortest program for a universal computer that outputs the string and stops.

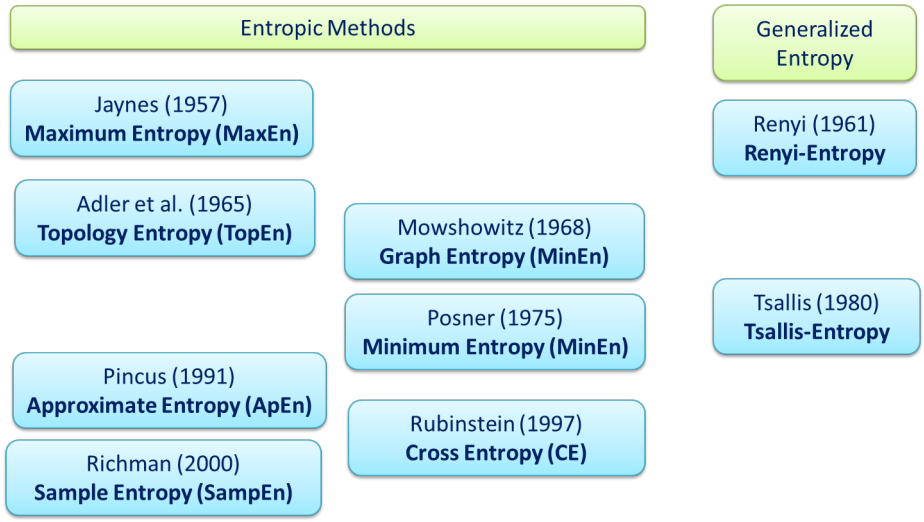

Fig. 2. A rough, incomplete overview on the most important entropy methods [15]

4 Application Areas

Entropy concepts found its way into many diverse fields of application within the biomedical domain:

Acharya et al. [37] proposed a methodology for the automatic detection of normal, pre-ictal, and ictal conditions from recorded EEG signals. Beside Approximate Entropy, they extracted three additional entropy variations from the EEG signals, namely Sample Entropy (SampEn), Phase Entropy 1 and Phase Entropy 2. They fed those features to seven different classifiers, and were able to show that the Fuzzy classifier was able to differentiate the three classes with an accuracy of 98.1 %. For this they took annotated recordings of five healthy subjects and five epilepsy patients. They showed that both ApEn and SampEn are higher in the case of normal signals, and lower for pre-ictal and ictal classes, indicating more self-similarity of the two later segments.

Hornero et al. [38] performed a complexity analysis of intracranial pressure dynamics during periods of severe intracranial hypertension. For that purpose they analyzed eleven episodes of intracranial hypertension from seven patients. They measured the changes in the intracranial pressure complexity by applying ApEn, as patients progressed from a state of normal intracranial pressure to intracranial hypertension, and found that a decreased complexity of intracranial pressure coincides with periods of intracranial hypertension in brain injury. Their approach is of particular interest to us, because they proposed classification based on ApEn tendencies instead of absolute values.

In the field of Electrocardiography analysis, Batchinsky et al. [39] recently performed a comprehensive analysis of the ECG and Artificial Neural Networks

(ANN) to improve care in the Battlefield Critical Care Environment, by developing new decision support systems that take better advantage of the large data stream available from casualties. For that purpose they analyzed the heart rate complexity of 800-beat sections of the R-to-R interval (RRI) time series from 262 patients by several groups of methods, including ApEn and SampEn. They concluded that based on ECG-derived noninvasive vital signs alone, it is possible to identify trauma patients who undergo Life-saving interventions using ANN with a high level of accuracy. Entropy was used to investigate the changes in heart rate complexity in patients undergoing post-burn resuscitation.

Pincus et al. took in [4] heart rate recordings of 45 healthy infants with recordings of an infant one week after an aborted sudden infant death syndrom (SIDS) episode. They then calculated the ApEn of these recordings and found a significant smaller value for the aborted SIDS infant compared to the healthy ones.

In [40] Sarlabous et al. used diaphragmatic MMG signals of dogs. The animals performed an inspiratory progressive resistive load respiratory test during the acquisition, in order to increase the respiratory muscular force. Afterwards the Approximate Entropy of these recordings were calculated and showed that these are able to quantify amplitude variations.

SampEn and ApEn were used in order to study gait data sets in [41]. For this purpose 26 healthy young adult and 24 healthy older adult subjects walked at least 200 steps on a treadmill. Their movement was tracked and step length, step width, and step time were calculated from the recordings. Both SampEn and ApEn showed significant differences between the younger and the older subjects in the step length and step width data sets.

In [42] Roerding et al. compared the postural sway of 22 stroke patients with 33 healthy also elderly subjects using different statistical tools including SampEn. All subjects were asked to do three trials while their sway was recorded. SampEn was significantly lower for the stroke patients.

The degree of randomness of a sequence is tightly related to its complexity, predictability, compressibility, repeatability and, ultimately, to the information theoretic notion of entropy. Most often, in genomics sequence analysis, information theoretic approaches are used (sometimes implicitly) to look for and to display information related to the degree of randomness of the sequences, aiming at finding meaningful structures. Early approaches include the sequence landscapes [43] and the sequence logos [44].

Pinho discusses some examples [45]: Some methods provide visual information of global properties of the DNA sequences. For example, the chaos game representation (CGR) [46] uses a distribution of points in an image to express the frequency of the Oligonucleotides that compose the sequence [47]. From these CGR images, other global representations can be derived, such as the entropic profiles [48], originally estimated using global histograms of the oligonucleotide frequencies, calculated using CGR images. Later, they have been generalized by Vinga et al. [49], based on the Rényi entropy, in order to calculate and visualize local entropic information. Other approaches for estimating the randomness

along the sequence have also been proposed. For example, Crochemore *et al.* [50] used the number of different oligonucleotides that are found in a window of predefined size for estimating the entropy.

The idea of showing local information content while taking into account the global structure of the sequence was also addressed by Allison *et al.* [51]. Based on a statistical model, they have produced information sequences, which quantify the amount of surprise of having a given base at a given position (and, therefore, in some sense are estimates of the local entropy), knowing the remaining left (or right) part of the sequence. When plotted, these information sequences provide a quick overview of certain properties of the original symbolic sequence, allowing for example to easily identify zones of rich repetitive content [52,53,54].

The information sequences of Allison *et al.* [51] are tightly related to data compression and, consequently, to entropy estimation. In fact, the importance of data compression for pattern discovery in the context of DNA sequences was initially addressed by Grumbach *et al.* [55] and, since then, studied by others (e.g. [56,52]).

The existence of regularities in a sequence renders it algorithmically compressible. The algorithmic information content of a sequence is the size, in bits, of its shortest reversible description and hence an indication of its complexity and entropy. Complexity measures of DNA sequences have been explored by several researchers (e.g. [57,58,59]). In this case, the key concept is the algorithmic entropy. Let x denote a binary string of finite length. Its algorithmic entropy, $K(x)$, is defined as the length of a shortest binary program x^* that computes x in a universal Turing machine and halts [60]. Therefore, $K(x) = |x^*|$, the length of x^*, represents the minimum number of bits of a program from which x can be computationally retrieved [61]. Although conceptually quite different, the algorithmic entropy is closely related to Shannon's entropy [61].

Because the algorithmic entropy is non-computable [61], it is usually approximated, for example, by compression algorithms [62,54,63,45]. In fact, compression-related approaches have been used not only for estimating the entropy, but also for building DNA sequence signatures capable of supporting the construction of meaningful dendograms [64]. In this case, estimates of the entropy associated with each of the three bases of the DNA codons are used to construct entropy vectors. Compression has also been used for measuring distances, such as in [65], where a genome-wide, alignment-free genomic distance based on compressed maximal exact matches is proposed for comparing genome assemblies.

Holzinger et al. (2012) [66] experimented with point cloud data sets in the two dimensional space: They developed a model of handwriting, and evaluated the performance of entropy based slant and skew correction, and compared the results to other methods. This work is the basis for further entropy-based approaches, which are very relevant for advanced entropy-based data mining approaches.

5 Detailed Description of selected Entropies

5.1 Approximate Entropy (ApEn)

Approximate Entropy measures the logarithmic likelihood that runs of patterns that are close remain close on following incremental comparisons [4]. We state Pincus' definition [4,5], for the family of statistics $\mathrm{ApEn}(m, r, N)$:

Definition 1. *Fix m, a positive integer and r, a positive real number. Given a regularly sampled time series $u(t)$, a sequence of vectors $\boldsymbol{x}(1)^m, \boldsymbol{x}^m(2), \ldots, \boldsymbol{x}^m(N-m+1)$ in $\mathbb{R}^m$ is formed, defined by*

$$\boldsymbol{x}^m(i) := [u(t_i), u(t_{i+1}), \ldots, u(t_{i+m-1})] \ . \tag{1}$$

Define for each i, $1 \leq i \leq N-m+1$,

$$C_i^m(r) := \frac{\textit{number of } j \textit{ such that } d[\boldsymbol{x}^m(i), \boldsymbol{x}^m(j)] \leq r}{N-m+1} \ , \tag{2}$$

where $d[\boldsymbol{x}(i), \boldsymbol{x}(j)]$ is the Chebyshev distance given by:

$$d[\boldsymbol{x}^m(i), \boldsymbol{x}^m(j)] := \max_{k=1,2,\ldots,m} \left(|u\left(t_{i+k-1}\right) - u\left(t_{j+k-1}\right)| \right) \ . \tag{3}$$

Furthermore, define

$$\Phi^m(r) := (N-m+1)^{-1} \sum_{i=1}^{N-m+1} \log C_i^m(r) \ , \tag{4}$$

*then the **Approximate Entropy** is defined as*

$$\mathrm{ApEn}(m, r, N) := \Phi^m(r) - \Phi^{m+1}(r) \ . \tag{5}$$

5.2 Sample Entropy (SampEn)

Richman and Moorman showed in [27] that approximate entropy is biased towards regularity. Thus, they modified it to Sample Entropy. The main difference between the two is that sample entropy does not count self-matches, and only the first $N-m$ subsequences instead of all $N-m+1$ are compared, for both ϕ^m and ϕ^{m+1} [27]. Similar to ApEn above, SampEn is defined as follows:

Definition 2. *Fix m, a positive integer and r, a positive real number. Given a regularly sampled time series $U(t)$, a sequence of vectors $\boldsymbol{x}^m(1), \boldsymbol{x}^m(2), \ldots, \boldsymbol{x}^m(N-m+1) \in \mathbb{R}^m$ is formed, defined by Eq. (1). Define for each i, $1 \leq i \leq N-m+1$,*

$$C_i^m = \frac{\textit{number of } j \textit{ such that } d[\boldsymbol{x}^m(i), \boldsymbol{x}^m(j)] \leq r \textit{ and } i \neq j}{N-m+1} \ , \tag{6}$$

where $d[(i),(j)]$ *is the Chebyshev distance (see Eq. (3)). Furthermore, define*

$$\Phi^m(r) := (N-m)^{-1} \sum_{i=1}^{N-m} C_i^m(r) \ , \tag{7}$$

then the ***Sample Entropy*** *is defined as*

$$\mathrm{SampEn}(m,r,N) := \log(\Phi^m(r)) - \log(\Phi^{m+1}(r)) \ . \tag{8}$$

5.3 Fuzzy (Measure) Entropy (Fuzzy(M)En)

To soften the effects of the threshold value r, Chen et al. proposed in [28] Fuzzy Entropy, which uses a fuzzy membership function instead of the Heaviside function. FuzzEn is defined the following way:

Definition 3. *Fix* m*, a positive integer and* r*, a positive real number. Given a regularly sampled time series* $U(t)$*, a sequence of vectors* $\boldsymbol{x}^m(1), \boldsymbol{x}^m(2), \ldots, \boldsymbol{x}^m(N-m+1) \in \mathbb{R}^m$ *is formed, as defined by Eq. (1). This sequence is transformed into* $\overline{\boldsymbol{x}}^m(1), \overline{\boldsymbol{x}}^m(2), \ldots, \overline{\boldsymbol{x}}^m(N-m+1)$*, with* $\overline{\boldsymbol{x}}^m(i) := \{u(t_i) - u0_i, \ldots, u(t_{i+m-1}) - u0_i\}$*, where* $u0_i$ *is the mean value of* $\boldsymbol{x}^m(i)$*, i.e.*

$$u0_i := \sum_{j=0}^{m-1} \frac{u_{i+j}}{m}. \tag{9}$$

Next the fuzzy membership matrix is defined as:

$$D_{i,j}^m := \mu(d(\overline{x}_i^m, \overline{x}_j^m), n, r) \ , \tag{10}$$

with the Chebyshev distance d *(see Eq. (3)) and the fuzzy membership function*

$$\mu(\boldsymbol{x}, n, r) := e^{-(\boldsymbol{x}/r)^n} \ . \tag{11}$$

Finally, with

$$\phi^m := \frac{1}{N-m} \sum_{i=1}^{N-m} \sum_{j=1, j\neq i}^{N-m} \frac{D_{i,j}^m}{N-m-1} \ , \tag{12}$$

the ***Fuzzy Entropy*** *is defined as:*

$$\mathit{FuzzyEn}(m,r,n,N) := \ln \phi^m - \ln \phi^{m+1} \ . \tag{13}$$

Liu et al. proposed in [29] **Fuzzy Measure Entropy**, which introduces a distinction between local entropy and global entropy, based on FuzzyEn. It is defined as:

$$\mathrm{FuzzyMEn}(m, r_L, r_F, n_L, n_F, N) := \ln \phi_L^m - \ln \phi_L^{m+1} + \ln \phi_F^m - \ln \phi_F^{m+1} \ , \tag{14}$$

where the local terms ϕ_L^m and ϕ_L^{m+1} are calculated as in Eq. (12) and the global terms ϕ_F^m and ϕ_F^{m+1} are calculated with Eq. (10) and Eq. (12), but with $\overline{\boldsymbol{x}}^m(i) := \{u(t_i) - u_{\mathrm{mean}}, \ldots, u(t_{i+m-1}) - u_{\mathrm{mean}}\}$, where u_{mean} is the mean value of the complete sequence $u(t)$.

5.4 Topological Entropy for Finite Sequences (FiniteTopEn)

As seen above, ApEn, SampEn, and Fuzzy(M)En all require the selection of a threshold value r which can significantly change the value of the associated entropy. FiniteTopEn differs from these definitions in that no threshold selection is required. FiniteTopEn is defined in the following way. First, define the complexity function of a sequence (finite or infinite) to be the following:

Definition 4. *For a given sequence w, the complexity function $p_w : \mathbb{N} \to \mathbb{N}$ is defined as*

$$p_w(n) = |\{u : |u| = n \text{ and } u \text{ appears as a subword of } w\}| .$$

So $p_w(n)$ gives the number of distinct n-length subwords (with overlap) of w. Then FiniteTopEn is defined as follows.

Definition 5. *Let w be a finite sequence of length $|w|$ constructed from an alphabet $\mathcal{A}$ of m symbols. Let n be the unique integer such that*

$$m^n + n - 1 \leq |w| < m^{n+1} + n.$$

Then for $v = w_1^{m^n+n-1}$ the first $m^n + n - 1$ letters of w, the topological entropy of w is defined to be

$$FiniteTopEn(w) = \frac{1}{n} \log_m \left(P_v(n) \right) .$$

FiniteTopEn is defined in this way primarily so that entropies of different length sequences and on possibly different alphabets can still be compared. Of course, if more is known about the process that generates a given sequence w, then the above definition can be modified as necessary (for example, by picking a smaller n or else not truncating w). The definition given above makes the least amount of assumptions regarding w (i.e. assumes that w was generated via the full shift). It is not difficult to demonstrate that as $|w| \to \infty$, $FiniteTopEn(w)$ converges to $TopEn(w)$, that is, to the topoloical entropy of w as originally defined in [30].

6 Open Problems

The main challenges in biomedical informatics today include [15], [67]:

- Heterogeneous data sources (need for data integration and data fusion)
- Complexity of the data (high-dimensionality)
- The discrepancy between data-information-knowledge (various definitions)
- Big data sets (which makes manual handling of the data nearly impossible)
- Noisy, uncertain data (challenge of pre-processing).

Particularly, on the last issue, dealing with noisy, uncertain data, entropy based methods might bring some benefits. However, in the application of entropy there are a lot of unsolved problems. We focus here on topological entropy, as this can be best used for data mining purposes.

Problem 1. **There is no universal method to calculate or estimate the topological entropy.** Zhou & Fang described in [68] topological entropy as one of the most important concepts in dynamical systems but they described also a number of open problems: TopEn describes the complexity of the motion in the underlying space caused by a continuous or differential action, i.e.: the bigger the topological entropy, the more complex the motion. Consequently, to obtain (calculate, measure, estimate) the topological entropy is an important research topic in dynamical systems. But as in the case of the Hausdorff measure, calculating the exact value of the topological entropy is, in general, very difficult, as to date there is no universal method. One might debate the clause *estimate* in the begin of this paragraph, since topological entropy can indeed be estimated (see Problem 2) for an arbitrary symbolic dynamical system. Then, a wide range of arbitrary dynamical systems can be approximated by an appropriate symbolic dynamical system.

Problem 2. A problem that has not been mentioned so far is the fact that to correctly estimate entropy (of any sort, and FiniteTopEnt in particular), one needs access to many data points. This is certainly not always the case, and so it would be beneficial to have something like a re-sampling/bootstrap regime. Since order matters to topological entropy, traditional bootstrapping cannot be used, which poses a big open problem.

Problem 3. How can sparse/infrequent data be re-sampled in a fashion appropriate to better estimate entropy.

Problem 4. For instance, for continuous mappings of the interval, the topological entropy being zero is equivalent to the period of any periodic point being a power of 2. For a general dynamical system, no similar equivalence condition has been obtained. A breakthrough regarding this depends upon a breakthrough in the investigation of the kernel problem of dynamical systems: the orbits topological structures or asymptotic behavior. An excellent source for this topic is [2].

Problem 5. The study of the topics mentioned in problem 4, is closely related to the ones in ergodic theory such as the invariant measure, the measure-theoretic entropy and the variational principle, as well as some fractal properties. Hence, the study of topological entropy has much potential in three fields: topology, ergodic theory and fractal geometry; albeit this will probably not unify these methods, topological entropy finds itself at the intersection of these subfields of mathematics.

Problem 6. In contrast to problem 5, FiniteTopEn is an approximation to topological entropy that is free from issues associated to choosing a threshold (problem 2). It was also shown in [32] that FiniteTopEn is computationally tractable, both theoretically (i.e its expected value) and practically (i.e. in computing entropy of DNA sequences). Applying this definition to the intron and exon regions of the human genome, it was observed that, as expected, the entropy of introns

is significantly higher that that of exons. This example demonstrates that this parameter-free estimate of topological entropy is potentially well-suited to discern salient global features of weakly structured data.

Problem 7. How to select parameters for the entropy measures? Each entropy has a number of parameters to be selected before application. Thus, there are a large number of possible combinations of parameters. By now, this problem has not yet been solved especially for ApEn, SampEn, FuzzyEn and FuzzyMEn. There are different parameter sets published (e.g., [4,69,70,41,71]), but up to now not all possible combinations were tested and no consensus was reached. The parameter sets cover certain application areas, but are dependent on the data and its type. An example is the choice of the threshold value r according to [69]. It is used only in the context of heart rate variability data and not applied to other data.

Problem 8. How to use entropy measures for classification of pathological and non-pathological data? In biomedical data, the goal is to discriminate between pathological and non-pathological measurements. There is still little evidence on how to use entropy measures for this classification problem and which data ranges to use. This is directly related to the parameter selection, since one of the hardest difficulties for ApEn, SampEn, FuzzyEn and FuzzyMEn lies in the choice of the threshold value r due to the flip-flop effect, i.e., for some parameters one data set has a higher entropy compared to another, but this order is reversed for different parameter choices [70,72]. This can occur for simple signals, but also when analyzing heart rate variability data, as shown in [71]. This leads to difficulties with the interpretation of the entropy, i.e., the direct assignment of entropy values to pathological or non-pathological data without a given r.
Finally a few very short questions poses mega challenges in these area:

Problem 9. How to generally benchmark entropy measures?

Problem 10. How to select appropriate entropy measures and their parameters to solve a particular problem?

7 Conclusion and Future Outlook

Entropy measures have successfully been tested for analyzing short, sparse and noisy time series data. **However they have not yet been applied to weakly structured data in combination with techniques from computational topology.** Consequently, the inclusion of entropy measures for discovery of knowledge in high-dimensional biomedical data is a big future issue and there are a lot of promising research routes.

References

1. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data - challenges in humancomputer interaction and biomedical informatics. In: DATA 2012, INSTICC (2012) 9–20 1.
2. Downarowicz, T.: Entropy in dynamical systems. Volume 18. Cambridge University Press, Cambridge (2011)
3. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (IL) (1949)
4. Pincus, S.M.: Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences **88**(6) (1991) 2297–2301
5. Pincus, S.: Approximate entropy (apen) as a complexity measure. Chaos: An Interdisciplinary Journal of Nonlinear Science **5**(1) (1995) 110–117
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41**(3) (2009) 1–58
7. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Springer, Berlin, Heidelberg, New York (2006)
8. Holzinger, A., Simonic, K.M.: Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058. Springer, Heidelberg, Berlin, New York (2011)
9. Kim, W., Choi, B.J., Hong, E.K., Kim, S.K., Lee, D.: A taxonomy of dirty data. Data Mining and Knowledge Discovery **7**(1) (2003) 81–99
10. Gschwandtner, T., Grtner, J., Aigner, W., Miksch, S. In: A taxonomy of dirty time-oriented data. Springer, Heidelberg, Berlin (2012) 58–72
11. Clausius, R.: On the motive power of heat, and on the laws which can be deduced from it for the theory of heat, poggendorff's annalen der physick, lxxix (1850)
12. Sethna, J.P.: Statistical mechanics: entropy, order parameters, and complexity. Volume 14. Oxford University Press, New York (2006)
13. Jaynes, E.T.: Information theory and statistical mechanics. Physical review **106**(4) (1957) 620
14. Golan, A.: Information and entropy econometrics: a review and synthesis. Now Publishers Inc (2008)
15. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
16. Jaynes, E.T.: Information theory and statistical mechanics. Physical review **106**(4) (1957) 620
17. Mowshowitz, A.: Entropy and the complexity of graphs: I. an index of the relative complexity of a graph. The bulletin of mathematical biophysics **30**(1) (1968) 175–204
18. Körner, J.: Coding of an information source having ambiguous alphabet and the entropy of graphs. In: 6th Prague Conference on Information Theory. (1973) 411–425
19. Holzinger, A., Ofner, B., Stocker, C., Valdez, A.C., Schaar, A.K., Ziefle, M., Dehmer, M.: On graph entropy measures for knowledge discovery from publication network data. In Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L., eds.: Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127. Springer, Heidelberg, Berlin (2013) 354–362
20. Dehmer, M., Mowshowitz, A.: A history of graph entropy measures. Information Sciences **181**(1) (2011) 57–78

21. Posner, E.C.: Random coding strategies for minimum entropy. Information Theory, IEEE Transactions on **21**(4) (1975) 388–391
22. Yuan, L., Kesavan, H.: Minimum entropy and information measure. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **28**(3) (1998) 488–491
23. Rubinstein, R.Y.: Optimization of computer simulation models with rare events. European Journal of Operational Research **99**(1) (1997) 89–112
24. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of operations research **134**(1) (2005) 19–67
25. Tsallis, C.: Possible generalization of boltzmann-gibbs statistics. Journal of statistical physics **52**(1-2) (1988) 479–487
26. de Albuquerque, M.P., Esquef, I.A., Mello, A.R.G., de Albuquerque, M.P.: Image thresholding using tsallis entropy. Pattern Recognition Letters **25**(9) (2004) 1059–1065
27. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol. Heart Circ. Physiol. **278**(6) (Jun 2000) H2039–2049
28. Chen, W., Wang, Z., Xie, H., Yu, W.: Characterization of surface emg signal based on fuzzy entropy. Neural Systems and Rehabilitation Engineering, IEEE Transactions on **15**(2) (2007) 266–272
29. Liu, C., Li, K., Zhao, L., Liu, F., Zheng, D., Liu, C., Liu, S.: Analysis of heart rate variability using fuzzy measure entropy. Comput. Biol. Med. **43**(2) (Feb 2013) 100–108
30. Adler, R.L., Konheim, A.G., McAndrew, M.H.: Topological entropy. Transactions of the American Mathematical Society **114**(2) (1965) 309–319
31. Adler, R., Downarowicz, T., Misiurewicz, M.: Topological entropy. Scholarpedia **3**(2) (2008) 2200
32. Koslicki, D.: Topological entropy of dna sequences. Bioinformatics **27**(8) (2011) 1061–1067
33. Solomonoff, R.J.: A formal theory of inductive inference. Part I. Information and Control **7**(1) (March 1964) 1–22
34. Solomonoff, R.J.: A formal theory of inductive inference. Part II. Information and Control **7**(2) (June 1964) 224–254
35. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of Information Transmission **1**(1) (1965) 1–7
36. Chaitin, G.J.: On the length of programs for computing finite binary sequences. Journal of the ACM **13** (1966) 547–569
37. Acharya, U.R., Molinari, F., Sree, S.V., Chattopadhyay, S., Ng, K.H., Suri, J.S.: Automated diagnosis of epileptic eeg using entropies. Biomedical Signal Processing and Control **7**(4) (2012) 401 – 408
38. Hornero, R., Aboy, M., Abasolo, D., McNames, J., Wakeland, W., Goldstein, B.: Complex analysis of intracranial hypertension using approximate entropy. Crit Care Med **34**(1) (2006) 87–95
39. Batchinsky, A., Salinas, J., Cancio, L., Holcomb, J.: Assessment of the need to perform life-saving interventions using comprehensive analysis of the electrocardiogram and artificial neural networks. Use of Advanced Techologies and New Procedures in Medical Field Operations **39** (2010) 1–16
40. Sarlabous, L., Torres, A., Fiz, J., Gea, J., Martínez-Llorens, J., Morera, J., Jané, R.: Interpretation of the approximate entropy using fixed tolerance values as a measure of amplitude variations in biomedical signals. In: Engineering in Medicine

and Biology Society (EMBC), 2010 Annual International Conference of the IEEE. (2010) 5967–5970
41. Yentes, J., Hunt, N., Schmid, K., Kaipust, J., McGrath, D., Stergiou, N.: The appropriate use of approximate entropy and sample entropy with short data sets. Annals of Biomedical Engineering **41**(2) (2013) 349–365
42. Roerdink, M., De Haart, M., Daffertshofer, A., Donker, S.F., Geurts, A.C., Beek, P.J.: Dynamical structure of center-of-pressure trajectories in patients recovering from stroke. Exp Brain Res **174**(2) (Sep 2006) 256–269
43. Clift, B., Haussler, D., McConnell, R., Schneider, T.D., Stormo, G.D.: Sequence landscapes. Nucleic Acids Research **14**(1) (1986) 141–158
44. Schneider, T.D., Stephens, R.M.: Sequence logos: a new way to display consensus sequences. Nucleic Acids Research **18**(20) (1990) 6097–6100
45. Pinho, A.J., Garcia, S.P., Pratas, D., Ferreira, P.J.S.G.: DNA sequences at a glance. PLoS ONE **8**(11) (November 2013) e79922
46. Jeffrey, H.J.: Chaos game representation of gene structure. Nucleic Acids Research **18**(8) (1990) 2163–2170
47. Goldman, N.: Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Nucleic Acids Research **21**(10) (1993) 2487–2491
48. Oliver, J.L., Bernaola-Galván, P., Guerrero-García, J., Román-Roldán, R.: Entropic profiles of DNA sequences through chaos-game-derived images. Journal of Theoretical Biology **160** (1993) 457–470
49. Vinga, S., Almeida, J.S.: Local Renyi entropic profiles of DNA sequences. BMC Bioinformatics **8**(393) (2007)
50. Crochemore, M., Vérin, R.: Zones of low entropy in genomic sequences. Computers & Chemistry (1999) 275–282
51. Allison, L., Stern, L., Edgoose, T., Dix, T.I.: Sequence complexity for biological sequence analysis. Computers & Chemistry **24** (2000) 43–55
52. Stern, L., Allison, L., Coppel, R.L., Dix, T.I.: Discovering patterns in *Plasmodium falciparum* genomic DNA. Molecular & Biochemical Parasitology **118** (2001) 174–186
53. Cao, M.D., Dix, T.I., Allison, L., Mears, C.: A simple statistical algorithm for biological sequence compression. In: Proc. of the Data Compression Conf., DCC-2007, Snowbird, Utah (March 2007) 43–52
54. Dix, T.I., Powell, D.R., Allison, L., Bernal, J., Jaeger, S., Stern, L.: Comparative analysis of long DNA sequences by per element information content using different contexts. BMC Bioinformatics **8**(Suppl. 2) (2007) S10
55. Grumbach, S., Tahi, F.: Compression of DNA sequences. In: Proc. of the Data Compression Conf., DCC-93, Snowbird, Utah (1993) 340–350
56. Rivals, E., Delgrange, O., Delahaye, J.P., Dauchet, M., Delorme, M.O., Hénaut, A., Ollivier, E.: Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences. Computer Applications in the Biosciences **13** (1997) 131–136
57. Gusev, V.D., Nemytikova, L.A., Chuzhanova, N.A.: On the complexity measures of genetic sequences. Bioinformatics **15**(12) (1999) 994–999
58. Nan, F., Adjeroh, D.: On the complexity measures for biological sequences. In: Proc. of the IEEE Computational Systems Bioinformatics Conference, CSB-2004, Stanford, CA (August 2004)
59. Pirhaji, L., Kargar, M., Sheari, A., Poormohammadi, H., Sadeghi, M., Pezeshk, H., Eslahchi, C.: The performances of the chi-square test and complexity measures for

signal recognition in biological sequences. Journal of Theoretical Biology **251**(2) (2008) 380–387
60. Turing, A.: On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society **42**(2) (1936) 230–265
61. Li, M., Vitányi, P.: An introduction to Kolmogorov complexity and its applications. 3rd edn. Springer (2008)
62. Chen, X., Kwong, S., Li, M.: A compression algorithm for DNA sequences and its applications in genome comparison. In Asai, K., Miyano, S., Takagi, T., eds.: Genome Informatics 1999: Proc. of the 10th Workshop, Tokyo, Japan (1999) 51–61
63. Pinho, A.J., Ferreira, P.J.S.G., Neves, A.J.R., Bastos, C.A.C.: On the representability of complete genomes by multiple competing finite-context (Markov) models. PLoS ONE **6**(6) (2011) e21588
64. Pinho, A.J., Garcia, S.P., Ferreira, P.J.S.G., Afreixo, V., Bastos, C.A.C., Neves, A.J.R., Rodrigues, J.M.O.S.: Exploring homology using the concept of three-state entropy vector. In: Pattern Recognition in Bioinformatics, 5th IAPR Int. Conf., PRIB 2010. Volume 6282 of LNBI. (September 2010) 161–170
65. Garcia, S.P., Rodrigues, J.M.O.S., Santos, S., Pratas, D., Afreixo, V., Bastos, C.A.C., Ferreira, P.J.S.G., Pinho, A.J.: A genomic distance for assembly comparison based on compressed maximal exact matches. IEEE/ACM Trans. on Computational Biology and Bioinformatics **10**(3) (May 2013) 793–798
66. Holzinger, A., Stocker, C., Peischl, B., Simonic, K.M.: On using entropy for enhancing handwriting preprocessing. Entropy **14**(11) (2012) 2324–2350
67. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinformatics **15**(Suppl 6) (2014) I1
68. Zhou, Z., Feng, L.: Twelve open problems on the exact value of the hausdorff measure and on topological entropy: a brief survey of recent results. Nonlinearity **17**(2) (2004) 493–502
69. Chon, K., Scully, C.G., Lu, S.: Approximate entropy for all signals. IEEE Eng Med Biol Mag **28**(6) (2009) 18–23
70. Liu, C., Liu, C., Shao, P., Li, L., Sun, X., Wang, X., Liu, F.: Comparison of different threshold values r for approximate entropy: application to investigate the heart rate variability between heart failure and healthy control groups. Physiol Meas **32**(2) (Feb 2011) 167–180
71. Mayer, C., Bachler, M., Hörtenhuber, M., Stocker, C., Holzinger, A., Wassertheurer, S.: Selection of entropy-measure parameters for knowledge discovery in heart rate variability data. BMC Bioinformatics **15**
72. Boskovic, A., Loncar-Turukalo, T., Japundzic-Zigon, N., Bajic, D.: The flip-flop effect in entropy estimation. (2011) 227–230