Information Content towards a neonatal disease severity score system

J. Bellazzini[1], C. Bonanno[1], M. Franciosi[1] and G. Menconi[2]

[1] Dipartimento di Matematica Applicata, Università di Pisa, Italy

[2] Dipartimento di Matematica, Università di Bologna, Italy

Corresponding author: Giulia Menconi, menconi@mail.dm.unipi.it

Summary. We introduce a score to classify the severity of patients by analysing the information content of clinical time series.

Key words: Information content, time series, neonatal illness severity

1.1 Introduction

The assessment of neonatal illness severity is a crucial issue in the neonatal intensive care unit. Particular attention has to be paid in developing non-invasive methodologies for data analysis that should give an illness severity score in real time. Hopefully, that score should reflect the current health status and suggest signs of incoming crises or aggravation. In the last decades of twentieth century, several methods from nonlinear dynamics have been proposed to answer such a clinical query. One of the most significant is based on an asymptotic measure of the density of the information content in a time series. In an experimental setting, information content may be approximated by means of compression algorithms (see for instance [1]). The purpose of this paper is showing an application of this kind of approach in order to analyse time series coming from clinical data. To be more precise, starting from a set of time series related to a patient, the information content of each series is calculated and a score is defined for the patient. To evaluate the agreement between that score and the actual patient'severity, we analysed a preliminary set of blind data collected by the Neonatal Intensive Care Unity, Siena University Hospital, Siena, Italy (Claudio De Felice, MD). Our methods found good results, pertinent with the health status of the patients. Our scores are summarized in the table *Severity array* (see fig. 1.3). It is worth to remark that when comparing our analysis with other totally different approaches of nonlinear time series analysis (surrogate series and multivariate analysis [7]) we found many interesting matches (for instance, see [6]). Indeed, this has to be considered as one of the first experiments of the **ATTIS project**, born in 2004 with the aim to develop and consolidate different methods in time series analysis, fundamental tools in order to achieve a better understanding of several physical and biological processes. (ATTIS is the achronimous for Approaches To TIme Series, see `http://www.attis-project.org`).

1.2 Materials and Methods

Information Content of time series

As stated in the introduction, one of the most significant tools from the modern theory of nonlinear dynamics used to analyse time series of biological origin is related to the notion of *information content* of a finite word as introduced by Shannon in [9]. The intuitive notion of information content of a finite word can be stated as "the length of the shortest message from which it is possible to reconstruct the original word", and a formal mathematical definition of this notion has been introduced by Kolmogorov using the notion of universal Turing machine (see [8]). We will not enter into the details of the mathematical definition, but simply use the intuitive notion of information content we stated.

The method we use to study the information content of a finite word is related to *compression algorithms*. These are a well known tool present in every personal computer, used to store files in the most economic way from the point of view of space needed in the memory storage disks. The compression of a finite word reflects the intuitive meaning of the information content of the word.

Let $s = (s_1 s_2 \dots s_n)$ be a n-long word written in the finite alphabet $\mathcal{A}$, that is $s_i \in \mathcal{A}$ for all $i = 1, \dots, n$. We will denote by $\mathcal{A}^n$ the set of all n-long word written using $\mathcal{A}$, and by $\mathcal{A}^* := \cup_n \mathcal{A}^n$. A *compression algorithm* can be defined as an injective function $Z : \mathcal{A}^* \to \{0,1\}^*$, that is a binary coding of the finite words written using $\mathcal{A}$. By using the algorithm Z we define the information content of a word s as the binary length of the compressed version of s, that is $Z(s)$. Hence

$$I(s) := \text{ Information Content of } s = |Z(s)|$$

The notion of information content of a finite word can be used also to face the problem of giving a notion of randomness of a word. Namely, we can think a word to be more random as less efficient is the compression achieved by a compression algorithm. This argument leads to the notion of *complexity* $C(s)$ of a finite word, defined as the *compression ratio*, that is

$$C(s) := \frac{I(s)}{|s|} = \frac{|Z(s)|}{|s|}$$

The greater the complexity of a word, the higher the randomness of the word.

Let us consider now the application of these tools to time series. By time series we mean a finite set $X = \{X_1, \dots, X_N\}$ of data, where each data is an array of real numbers. The first step in the analysis is the reduction of a time series to a finite word. This is accomplished by a partition of the set of possible values of the data. Let $\mathcal{P}$ be such a partition into sets $\{I_1, \dots, I_L\}$, then to the time series X we associate the word $s \in \mathcal{A}^N$, where $\mathcal{A} = \{1, \dots, L\}$, with the rule that for all $j = 1, \dots, N$, we choose $s_j = \ell \in \mathcal{A}$ if $X_j \in I_\ell$. We can then define the complexity of the time series X as

$$C(X, L) := C(s)$$

where the notation $C(X, L)$ points out the role of the number of symbols of the alphabet used to write the word s. In the following we will consider the effects of varying L.

When analysing a time series it is convenient to consider asymptotic properties, hence assuming to have an infinite time series. We can make this assumption to obtain some mathematical results on the complexity of a time series. For an infinite time series $X = \{X_i\}$ we can define the *asymptotic complexity*

$$K(X, L) := \lim_{N \to \infty} C((X_1, \dots, X_N), L)$$

with respect to an L-long alphabet.

If the time series X is the orbit of a dynamical system then $K(X, L) \to h$ as $L \to \infty$, where h is the metric entropy of the system (see [3]). In fact there is an integer L_0 such that $K(X, L) = h$ if $L \geq L_0$.

A particular case is obtained when the series is generated by a white noise on the unit interval $[0, 1]$. Let ξ be such a series, then $K(\xi, L) \to \infty$ as $L \to \infty$, and the asymptotic behaviour is $K(\xi, L) \sim \log L$ (for two sequences (a_n) and (b_n) of real numbers, we say that $a_n \sim b_n$ asymptotically if $\lim_n \frac{a_n}{b_n} = 1$) (cf. [2]).

In [2] it is also studied the case of time series $(X + \xi)$ obtained as a random perturbation of a dynamical system. The results show that for big L the behaviour of $K(X + \xi, L)$ is analogous to that of $K(\xi, L)$, whereas for small enough values of L it is possible to achieve $K(X + \xi, L) = h$ where h is the metric entropy of the unperturbed dynamical system. In particular it is shown that it is important to compare the size of the intervals of the partition used to obtain the symbolic word and the size of the noise.

The analysis we present in this paper is performed with time series quite short, and often the range of different values in a single series is very small. These features imply two drawbacks: (i) we are very far from obtaining an asymptotic behaviour, hence we are looking at the transients; (ii) we cannot use alphabets with many symbols, hence also the limit in L can only be a rough approximation.

However we can still analyse the series by comparing the results with those for white noise. It is reasonable to treat short time series as obtained by noisy systems. Moreover we made partitions of the set of possible values of the data up to sets given by single values, that is using the best “resolution”. Hence this analysis is comparable with the case of time series obtained as a random perturbation of a dynamical system with big values for L. For these reasons we decided to study the logarithmic behaviour of the values $K(X,L)$, hence to make a linear interpolation of the set $\{(L, K(X,L))\}$ when plotted in logarithmic linear scale. This gives a value $q(X)$ satisfying $K(X,L) \approx q(X) \log L$. For white noise we would have $q(X) = 1$, hence our results give values $q(X) \in (0,1)$.

Experimental data sets

We applied our method to 23 triplets of time series related to 12 newborns admitted to the Neonatal Intensive Care Unit (NICU) of Azienda Ospedaliera Universitaria Senese, Policlinico "Le Scotte" , Siena, Italy, (Claudio De Felice MD). Mean gestational age was 32.5 ± 4.7 weeks (range: $23-38$ weeks) while mean birth weight was 1775 ± 766 g (range: $580-3380$ g).

For each patient we considered a triplet concerning the following pulse oximetry-derived signals:

- perfusion index (PI)
- pulse rate (PR)
- oxygen saturation (SpO_2)

These data sets were collected with Radical SET pulse oximeter, MASIMO Co., Irvine, CA, USA, probe placed at either feet, according to the methodology illustrated in [4] and [5]. They were registered every 4 seconds and they refer to a monitorization period of 1 to 1.5 hours. The perfusion index is a percentage, while pulse rate and saturation are integers. In order to reduce the time series into symbolic strings, we have considered uniform partitions $\mathcal{P}_L$ on the set of possible values $I = [min\ val,\ max\ val]$ made of L subintervals. For each time series, we have considered partitions where L ranges from 2 to 20 subintervals. We have used the compression algorithm called $CASToRe$, developed by our group (see [1] for details). We give in figure 1.1 three different examples of triplets relative to different newborns.

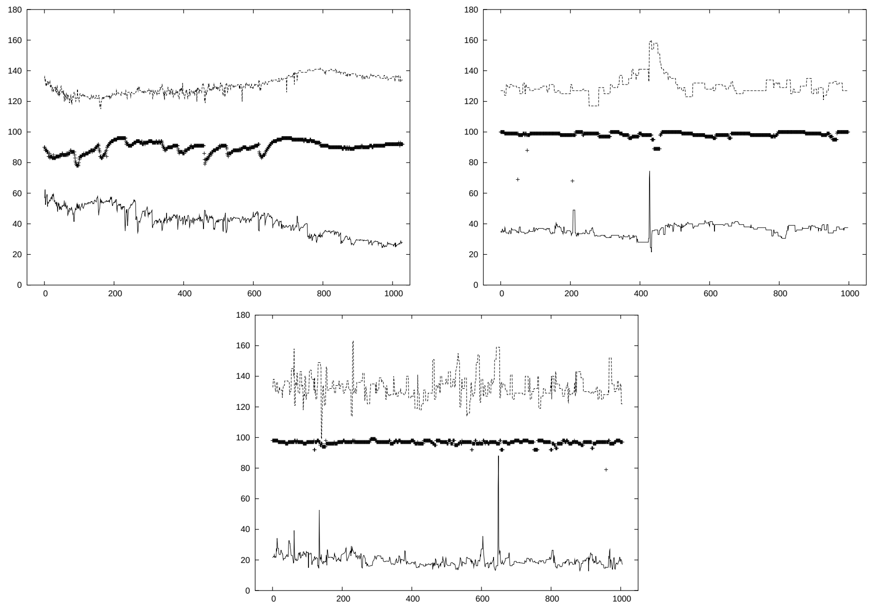

Fig. 1.1. Data sets for three different newborns: #01 (top left), #02 (top right), #08 (bottom). For each plot: PI series is at the bottom (rescaled, to be more readable), SpO_2 is the middle series and PR is at the top.

When applying our analysis we were completely unaware of the health status of each patient. We compared our results with the other research groups of ATTIS Project and our Severity array agrees with other scores obtained with completely different approaches.

1.3 Results

In figure 1.2 the behaviour of the complexity $K(X, L)$ with respect to the graining size L is shown for the whole data set #01. The plot is in log-linear scale and the coefficient $q(X)$ is the slope of the growth, for perfusion index, pulse rate and saturation.

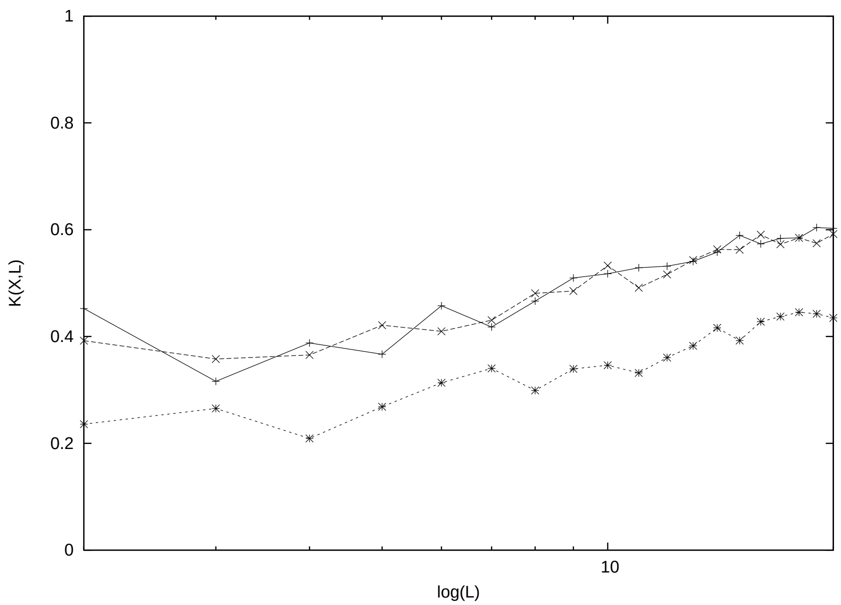

Fig. 1.2. Complexities $K(X, L)$ for data set #01. Perfusion index: top solid line; pulse rate: middle crossed-dashed line; saturation: bottom star-dotted line. The graining size L ranges from 2 to 20. The plot is log-linear.

Severity score

For each patient, three coefficients q have been calculated. We denote by $\mathcal{S}(N) = (q(PI), q(PR), q(SpO_2))$ the severity array of data set $\#N$.

In order to classify the severity of each patient via their severity array, we have defined the severity score as the ℓ_1-norm of the array:

$$||\mathcal{S}(N)|| = q(PI) + q(PR) + q(SpO_2) \tag{1.1}$$

Figure 1.3 shows how the score varies for each data set: as a matter of

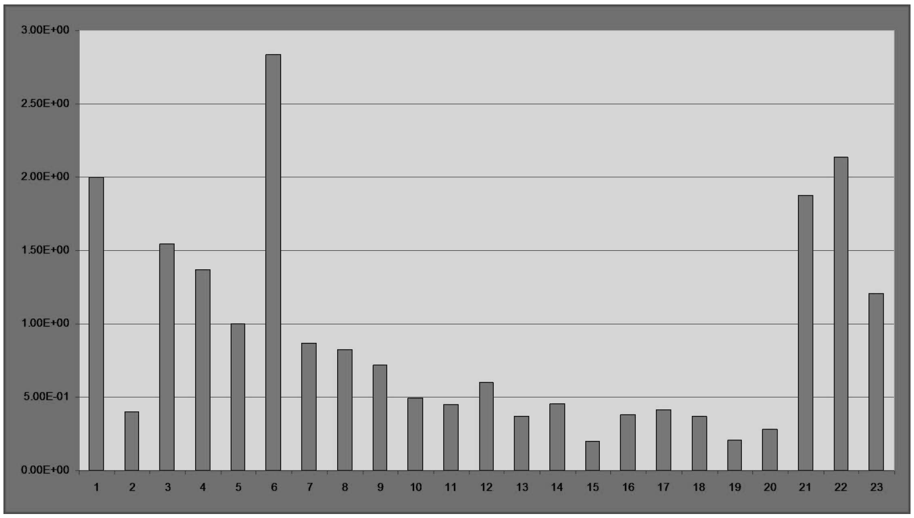

Fig. 1.3. Severity score for the 23 data sets.

fact, the results show a high variability. The maximum is reached by data set #6, while the minimum is for data set #19. Two extremal groups of data sets are neatly identified: $\mathbf{A} = \{\#01, \#06, \#21, \#22\}$ have very high scores, while $\mathbf{B} = \{\#02, \#10, \#11, \#12, \ldots, \#20\}$ have a very poor score. It is a first hint for distinguishing the group of healthier patients from less healthy patients. It is well known that life parameters should show a high irregularity to be more adaptive. Therefore, it is reasonable to associate a higher severity to data sets whose severity score is lower, since it means that the regularity of the time series is more evident. A finer investigation is necessary to characterise intermediate scores for the group of data sets $\mathbf{C} = \{\#03, \#04, \#05, \#07, \#08, \#09, \#23\}$.

Let the normalised severity array be

$$\mathcal{S}^*(N) =$$
$$= \left(\frac{q(PI) - min(PI)}{max(PI) - min(PI)}, \frac{q(PR) - min(PR)}{max(PR) - min(PR)}, \frac{q(SpO_2) - min(SpO_2)}{max(SPO_2) - min(SpO_2)} \right)$$

Normalised severity arrays belong to $[0,1] \times [0,1] \times [0,1]$. For each $N = 1, \ldots, 23$, we have represented $\mathcal{S}^*(N)$ on a two-dimensional symplex, by drawing a ball with radius proportional to the severity score (figure 1.4).

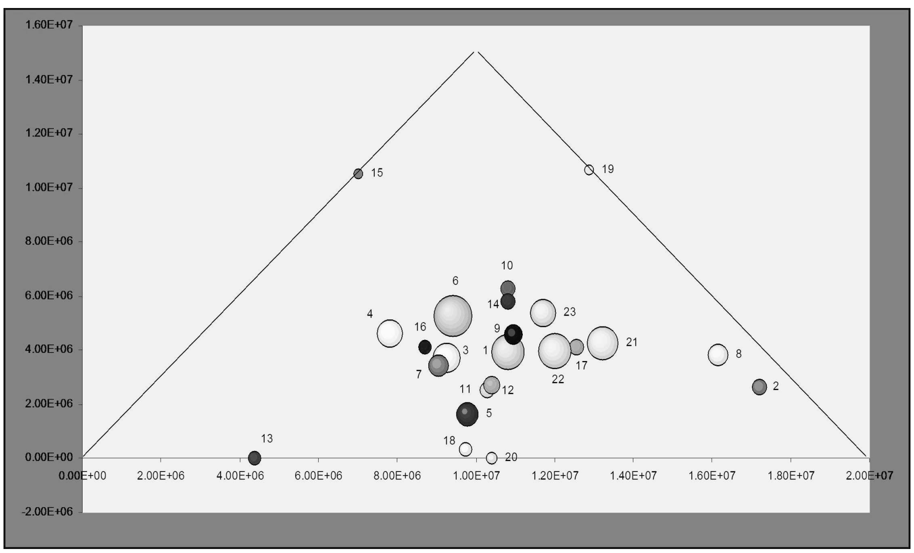

Fig. 1.4. Severity array: symplex representation.

1.4 Final discussion

The *Severity Score Method* (SSM) here introduced allows an efficient and user-friendly approach to early monitor patient's health severity. Concerning the three groups identified in Section 1.3, their clinical status was the following, in agreement with the experimental classification:

A = {#01, #06, #21, #22}: very low clinical severity.

B = {#02, #10, #11, #12, . . . , #20}: high clinical severity.

C = {#03, #04, #05, #07, #08, #09, #23}: all but #09 were cases of histologic chorioamnionitis.

A few more details may be inferred from severity array (figure 1.4). Patients #02 and #08 are isolated from the others: they had severe crisis some hours after the end of measurements. In particular, patient #08 shows an intermediate score, but an anomalous correlation among the three clinical parameters is highlighted by the symplex representation. Finally, data sets from #09 to #20 are relative to the same patient at different times. They were clustered almost all together in group **B**.

To sum up, this method gives interesting hints on reliable prediction of crisis in differently critical patients. Also, it encourages interdisciplinary collaboration to develop some diagnostic protocol to help the medical team in deciding whether taking care of the patient in an Intensive Care Unit.

References

1. Benci V., Bonanno C., Galatolo S., Menconi G., Virgilio M.: Dynamical systems and computable information. Discrete and Continuous Dynamical Systems - B, **4**, 4, 935–960 (2004).
2. Bonanno, C.: The algorithmic information content for randomly perturbed systems. Discrete Contin. Dyn. Syst. Ser. B,**4**, 4, 921–934 (2004).
3. Brudno A.A.: Entropy and the complexity of the trajectories of a dynamical system. Trans. Moscow Math. Soc., **2** 127–151 (1983).
4. De Felice C, Latini G, Vacca P, et al.: The pulse oximeter perfusion index as a predictor for high illness severity in neonates. Eur J Pediatr, **161**, 561–562 (2002).
5. De Felice C, Del Vecchio A, Criscuolo M, Lozupone A, Parrini S and Latini G.: Early postnatal perfusion index changes in term newborns with subclinical chorioamnionitis. Arch Dis Child Fetal Neonatal Ed.,**90**, 5 (2005).
6. De Felice C., Bianciardi G., Parrini S. et al.: Pulse oximetry signals: chaotic analysis in the evaluation of neonatal illness severity. Pediatric Res., **56**, 475 (2004).
7. Kantz H., Schreiber T.: Nonlinear Time Series Analysis, Second Edition. Cambridge University Press (2004).
8. Kolmogorov A. N.: On the entropy per time unit as a metric invariant of automorphism. Dokl. Acad. Nauk., **124**, 754-755 (1959).
9. Shannon C.E.: The mathematical theory of communication. Bell System Technical J., **27**, 379-423 and 623-656 (1948).