Equivalence between time series predictability and Bayes error rate

En Xu\textsuperscript{1(a)}, Tao Zhou\textsuperscript{2(a)(b)}, Zhiwen Yu\textsuperscript{1(c)}, Zhuo Sun\textsuperscript{1} and Bin Guo\textsuperscript{1}

\textsuperscript{1} School of computer science, Northwestern Polytechnical University - Xi’an 710129, China
\textsuperscript{2} Big Data Research Center, University of Electronic Science and Technology of China - Chengdu 611731, China

received 29 November 2022; accepted in final form 6 March 2023
published online 17 March 2023

Abstract – Predictability is an emerging metric that quantifies the highest possible prediction accuracy for a given time series, being widely utilized in assessing known prediction algorithms and characterizing intrinsic regularities in human behaviors. Lately, increasing criticisms aim at the inaccuracy of the estimated predictability, caused by the original entropy-based method. In this paper, we strictly prove that the time series predictability is equivalent to a seemingly unrelated metric called Bayes error rate that explores the lowest error rate unavoidable in classification. This proof bridges two independently developed fields, and thus each can immediately benefit from the other. For example, based on three theoretical models with known and controllable upper bounds of prediction accuracy, we show that the estimation based on Bayes error rate can largely solve the inaccuracy problem of predictability.

Copyright © 2023 EPLA

Introduction. – Predictability refers to the limit of prediction accuracy of a given time series \cite{1}. Exploring such a metric is of great significance. On the one hand, with given data, it can be used to evaluate the performance and to estimate the potential improving space of existing algorithms \cite{2}. On the other hand, it characterizes intrinsic regularities of time series and thus the investigation can deepen our understanding of related phenomena. For example, the varying predictability reveals the sudden change of human mobility patterns after disasters \cite{3}, quantifies the respective contributions of behavioral similarities and social relationships to human mobility prediction \cite{4}, uncovers the roles of model structure and social network heterogeneity in predicting infectious disease outbreaks \cite{5}, and the structural breaks demonstrates the predictability of the stock market \cite{6}.

Song \textit{et al.} proposed an entropic framework to calculate the predictability II \cite{1} (see also an analogous contribution by Feder and Merhav \cite{7}). This pioneering method has two obvious disadvantages: i) it builds on the underlying Markovian nature while some time series exhibit long-range correlations; ii) it is sensitive to the lengths of time series while real-time series are usually too short to satisfy the requirement. Consequently, the above method is usually inaccurate, sometimes largely overestimated \cite{8,9} and sometimes surpassed by well-designed algorithms \cite{2,10}.

Bayesian analysis has been widely used in science and engineering, such as crystal structure prediction \cite{11}, high-energy physical collision signal derivation \cite{12–14}, and nuclear theory research \cite{15}. Bayes error rate (BER, denoted by $R$) in Bayesian analysis is the lowest unavoidable error rate in classification for given data \cite{16}, which has been extensively applied in feature selection, intermediate representations of features or behaviors, quality assessment of security defenses, feasibility estimation of machine learning, and so on \cite{17}. In contrast to the onset of germination of predictability, the investigation on BER has lasted more than half a century and thus many sophisticated methods are proposed to calculate $R$ or to estimate the upper and lower bounds of $R$, such as the density estimators that directly estimate the conditional probability distribution $\eta_\omega(x) = p(\omega|x)$ where $\omega$ and $x$ are class label and feature vector \cite{18}, the divergence estimators that focus on the maximum posterior probability $\max_{\omega \in \Omega} \eta_\omega(x)$ where $\Omega$ is the set of classes \cite{19}, and the $k$-NN estimators that learn the expectation of the highest accuracy $\mathbb{E}_X [1 - \max_{\omega \in \Omega} \eta_\omega(x)]$, where $X$ denotes the feature space \cite{20}. Later in this paper, we will prove the mathematical equivalence between the above two seemingly unrelated metrics, predictability

\textsuperscript{(a)}Both authors contributed equally to this research.
\textsuperscript{(b)}E-mail: zhutou@ustc.edu (corresponding author)
\textsuperscript{(c)}E-mail: zhiwenyu@nwpu.edu.cn
and BER, and further discuss what can be immediately gained after the equivalence.

**Theorem 1.** Given an $M$-state time series, its predictability $\Pi$ is equivalent to the Bayes error rate $R$ in a $M$-classification problem as

$$\Pi = 1 - R,$$

(1)

if we treat each state as a class, and the series before the state as the feature.

**Proof.** Denote $x_{n-1} = \omega^1 \omega^2 \cdots \omega^{n-1}$ the historical series from time 1 to $n - 1$, where $\omega^i \in \Omega$ and $\Omega$ is the set of $M$ states. Denote $Pr[\omega^n = \omega^i|x_{n-1}]$ the probability that our predicted state $\omega^n$ is equal to the actual state $\omega^n$ given $x_{n-1}$, and $\pi(x_{n-1}) = \sup_{\omega^n} \{ Pr[\omega^n = \omega^i|x_{n-1}] \}$ the probability of occurrence of the most probable state at time $n$, then the predictability of the $n$-th state given the historical states $x_{n-1}$ is $\Pi(n) = \sum_{x_{n-1}} P(x_{n-1}) \pi(x_{n-1})$, where $P(x_{n-1})$ is the probability of observing a particular history $x_{n-1}$, and the sum is taken over all possible histories of length $n - 1$. Notice that $\pi(x_{n-1})$ contains the full predictive power including the potential long-range correlations in the time series, while in practice we usually use shorter historical series, such as $\omega^{n-r} \omega^{n-r+1} \cdots \omega^{n-1}$ with $r$ a cutoff parameter, instead of the full historical series $\omega^1 \omega^2 \cdots \omega^{n-1}$, so that in a more general case, the probability of observing a particular history can be smaller than 1. The overall predictability $\Pi$ is then defined as the time-averaged predictability for a sufficiently long time series [1], as

$$\Pi = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \Pi(i),$$

(2)

where $x_0 = \emptyset$ and $\Pi(1)$ is the predictability of the first state without any available information.

We consider an $M$-classification problem with $n$ samples whose class labels are the $n$ states in the time series and whose features are series before the corresponding states. Table 1 illustrates the one-to-one relationship between time series prediction and classification for an example series with $M = 3$ and $n = 5$. Denote $p(x|\omega_j)$ the conditional probability density of the feature $x \in X'$, $X' \subset X$ is the set of observed features, and $p(\omega_j)$ the prior probability of the class $\omega_j \in \Omega$ ($j = 1, 2, \cdots, M$), the BER is expressed as

$$R = 1 - \sum_{j=1}^{M} \int_{G_j} p(\omega_j) p(x|\omega_j) dx,$$

(3)

with the partition $G_j$ defined as

$$G_j \triangleq \{ x \in X' | p(\omega_j) p(x|\omega_j) > \max_{k \neq j} p(\omega_k) p(x|\omega_k) \}.$$

(4)

Applying the Bayes formula $p(x)p(\omega_j|x) = p(\omega_j)p(x|\omega_j)$, where $p(x)$ is the prior probability of the feature $x$, we have

$$\sum_{j=1}^{M} \int_{G_j} p(x) p(\omega_j|x) dx = \sum_{j=1}^{M} \int_{G_j} p(x) p(\omega_j) dx,$$

(5)

and the partition $G_j$ is equivalent to the partition $G_j'$ with

$$G_j' \triangleq \{ x \in X' | p(\omega_j|x) > \max_{k \neq j} p(\omega_k|x) \}.$$

(6)

According to our setting, there is a one-to-one relationship between features and historical series, namely for each $x_{i-1}$ ($1 \leq i \leq n$), there exists a certain $x \in X'$ such that $x_{i-1} = x$, and vice versa. As $p(x)$ is the probability of observing the feature $x$ in the feature set $X'$ while $P(x_{i-1})$ is the probability of observing the series $x_{i-1}$ in all historical series of length $i - 1$, in the large limit of $n$, $p(x) = \frac{1}{n} P(x_{i-1})$. In addition, $\pi(x_{i-1}) = p(\omega_j|x)$, according to the definition of $\pi(\cdot)$. As a consequence,

$$\sum_{j=1}^{M} \int_{G_j'} p(x) p(\omega_j|x) dx = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{x_{i-1}} P(x_{i-1}) \pi(x_{i-1}).$$

(7)

Therefore, the time series predictability is equivalent to the Bayes error rate, with the relationship $\Pi = 1 - R$. □

**Results.** — According to Theorem 1, we can directly take advantage of methods developed to calculate $R$ in real datasets to improve the estimation of $\Pi$. Considering a simple example with three states $\Omega = \{ A, B, C \}$, where the next state only depends on the current state, according to the following Markovian transfer matrix:

$$
\begin{align*}
A & = & \begin{bmatrix} q & 2/3(1-q) & 1/3(1-q) \\
2/3(1-q) & q & 2/3(1-q) \\
2/3(1-q) & 1/3(1-q) & q & \end{bmatrix}
\end{align*}
$$

(8)

Obviously, when $0.4 \leq q \leq 1$, the true predictability $T = q$. Time series with arbitrary length $n$ can be generated.
by eq. (8). We set \( r = 1 \) to extract the features. Take \( \{ABBCA \cdots \} \) as an example, the corresponding (feature, class) set is \( \{(A, B), (B, B), (B, C), (C, A), \cdots \} \).

According to the entropy-based method, the estimated predictability \( \Pi \) is determined by

\[
H = -\Pi \log_2 \Pi - (1 - \Pi) \log_2 (1 - \Pi) + (1 - \Pi) \log_2 (M - 1),
\]

where \( M = 3 \) and \( H \) is the entropy of the next-moment state that can be estimated by the data (see details in [1]). In the corresponding classification problem, the lower and upper bounds of \( R \) can be obtained by the inequality

\[
\frac{M-1}{(M-2)M} \sum_{i=1}^{M} [1 - p(\omega_i)] R_i^{M-1} \leq R^M \leq \min_{\alpha \in (0, 1)} \frac{1}{M - 2\alpha} \sum_{i=1}^{M} [1 - p(\omega_i)] R_i^{M-1} + \frac{1 - \alpha}{M - 2\alpha}.
\]

(10)

where \( R^k \) is the BER for the \( k \)-classification subproblem and \( R^{M-1} \) is the BER for the \((M - 1)\)-classification subproblem created by removing the \( i \)-th class (see details in [21,22]). The upper and lower bounds of predictability can then be obtained through eq. (1) (Theorem 1), and the estimated predictability \( \hat{\Pi} \) is the average of the two bounds.

Figure 1(A) shows how \( \hat{\Pi} \) changes with increasing \( n \) for three specific cases, \( q = 0.4, q = 0.6 \) and \( q = 0.8 \). The result confirms two above-mentioned disadvantages of the entropy-based method, namely \( \Pi \) is sensitive to the length \( n \) and much larger than the true predictability \( T = q \). To ensure the stability, we set \( n = 2^{15} \) and compare the entropy-based method eq. (9) and the BER-inspired method eq. (10)). As shown in fig. 1(B), the latter remarkably and consistently outperforms the former.

Considering a more complicated series generator with \( M \) states \( \Omega = \{S_1, S_2, \cdots, S_M\} \), where the next state \( \omega^{t+1} \) is randomly drawn from the \( M \) states with probability \( 1 - q \), or determined by the two anterior states with probability \( q \). In the latter case, if \( \omega^{t-1} = S_i \) and \( \omega^t = S_j \), then \( \omega^{t+1} = S_k \), \( k = i + j \) (if \( k > M \), we set \( k \leftarrow k - M \)). Obviously, the true predictability is \( T = q + (1 - q)/M \). Figure 1(C) reports how \( \hat{\Pi} \) changes with increasing \( n \) for four specific cases \( q = 0.2, q = 0.4, q = 0.6 \) and \( q = 0.8 \), with \( M = 100 \) fixed. As \( 1/M \) is much smaller than \( q \) in the above four cases, \( T \approx q \). Analogous to what found in fig. 1(A), \( \hat{\Pi} \) is sensitive to \( n \) and much larger than \( T \) after being nearly stable (\( n > 2^{15} \)). As shown in fig. 1(D), in most cases the BER-inspired method performs better than the entropy-based method, and only when the time series is highly predictable (\( q \approx 1 \), see the top right corner), the results of the entropy-based method and BER-inspired method are close to each other.

To reveal the effects of parameters \( r \) and \( M \), we further consider the third generator where the next state \( \omega^{t+1} \) is equal to \( \omega^t \) with probability \( q_1 = 0.1 \), equal to \( \omega^{t-2} \) with probability \( q_2 = 0.2 \), equal to \( \omega^{t-3} \) with probability \( q_3 = 0.3 \). With probability \( 1 - q_1 - q_2 - q_3 \), \( \omega^{t+1} \) is randomly drawn from \( M \) states. The true predictability is \( T = \max\{q_1, \cdots, q_r\} + (1 - q_1 - q_2 - q_3)/M \) is sensitive to \( r \) and \( M \). As shown in fig. 1(E), the original entropy-based method does not consider the impacts of parameter \( r \) while the BER-inspired method can well capture the effects of the memory length \( r \). As shown in fig. 1(F), both the entropy-based and BER-inspired methods capture the decreasing tendency of predictability as the increase in \( M \). One can clearly observe from figs. 1(E) and (F) that the entropy-based method will largely overestimate the predictability even for sufficiently long time series, while the BER-inspired method performs much better.
Discussion. – The direct value of knowing predictability is to decide whether it is worthwhile to improve the current predictors [1,23]. The embodiment of such value requires an accurate estimate of predictability. Unfortunately, the entropy-based method [1] usually fails as it largely overestimates the true predictability (see, for example, Fig. 1). The dissatisfactory performance partially comes from the approximation that only accounts for the entropy of the state with the maximum next-moment occurrence probability. At the same time, such approximation is an indispensable part that guarantees the computational feasibility. Therefore, it is difficult to overcome the observed disadvantages within the entropic framework [9,24]. This paper uncovers the equivalence between predictability and a seemingly unrelated metric BER, and immediately provides a novel way to improve the estimation of predictability —applying the BER-related studies can learn from such perspective, that is, $R$ can also be treated as a system feature in particular if the data contains temporal features. In a word, the reported Theorem 1 builds a bridge connecting two separately developed fields. By three simple and ideal models, we show how one can benefit from the other, and readers are strongly encouraged to find more scenarios and examples.

***

This work was supported in part by the National Natural Science Foundation of China (No. 61960206008, No. 62002294, No. 11975071) and the National Science Fund for Distinguished Young Scholars (No. 61725205).

Data availability statement: No new data were created or analysed in this study.

REFERENCES

[1] Song C., Qu Z., Blumm N. and Barabási A.-L., Science, 327 (2010) 1018.

[2] Lu X., Wetter E., Bharti N., Tatem A. J. and Bengtsson L., Sci. Rep., 3 (2013) 2924.

[3] Lu X., Bengtsson L. and Holme P., Proc. Natl. Acad. Sci. U.S.A., 109 (2012) 11576.

[4] Chen Z., Kelty S., Evsukoff A. G., Welles B. F., Bagrow J., Menezes R. and Ghoshal G., Nat. Commun., 13 (2022) 1922.

[5] Scarpe S. V. and Petri G., Nat. Commun., 10 (2019) 898.

[6] Devpura N., Narayan P. K. and Sharma S. S., J. Int. Financ. Mark. Inst. Money, 63 (2019) 101145.

[7] Feder M. and MerHAV N., IEEE Trans. Inf. Theory, 40 (1994) 259.

[8] Xu P., Yin L., Yue Z. and Zhou T., Physica A, 523 (2019) 345.

[9] Smith G., Wieser R., Goulding J. and Barrack D., A refined limit on the predictability of human mobility, in Proceedings of the IEEE International Conference on Pervasive Computing and Communications, PerCom 2014, Budapest, Hungary, March 24-28, 2014 (IEEE Computer Society) 2014, pp. 88–94.

[10] Kulkarni V., Mahalunark A., Gabrnatno B. and Kelleher J. D., Entropy, 21 (2019) 432.

[11] Gao H., Wang J., Han Y. and Sun J., Fundam. Res., 1 (2021) 466.

[12] Niu D. and Miao Y., AAPPs Bull., 31 (2021) 12.

[13] Xu N., Fukushima K. and Mohanty B., AAPPs Bull., 31 (2021) 1.

[14] Shi R.-X., Li S.-Y., Lu J.-X. and Geng L.-S., Sci. Bull., 67 (2022) 2298.

[15] Sadikhan J., AAPPs Bull., 32 (2022) 14.

[16] Cover T. M. and Hart P. E., IEEE Trans. Inf. Theory, 13 (1967) 21.

[17] Berisha V., Wisler A., III A. O. H. and Spianas A., IEEE Trans. Signal Process., 64 (2016) 580.

[18] Ferguson T. S., Bayesian density estimation by mixtures of normal distributions, in Recent Advances in Statistics, edited by Rizvi M. H., Rustagi J. S. and Siegmund D. (Academic Press) 1983, pp. 287–302.

[19] Nguyen X., Wainwright M. I. and Jordan M. I., IEEE Trans. Inf. Theory, 56 (2010) 5847.

[20] Fukunaka K. and Hostetler L. D., IEEE Trans. Inf. Theory, 21 (1975) 285.

[21] Wisler A., Berisha V., Wei D., Ramamurthy K. and Spianas A., Empirically-estimable multi-class classification bounds, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Shanghai, China) 2016, pp. 2594–2598.

[22] Renggli C., Rimanic L., Hollenstein N. and Zhang C., Evaluating Bayes error estimators on real-world datasets with feebee, in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks I, edited by Vanschoren J. and Yeung S. (Curran Associates Inc., virtual) 2021.

[23] Lu L., Pan L., Zhou T., Zhang Y.-C. and Stanley H. E., Proc. Natl. Acad. Sci. U.S.A., 112 (2015) 2325.

[24] Zhang C., Zhao K. and Chen M., to be published in IEEE Trans. Knowl. Data Eng. (2023), https://ieeexplore.ieee.org/document/9705215 (DOI: 10.1109/TKDE.2022.3148300).

[25] Seken S. Y., Oselio B. and Hero A. O., IEEE Trans. Signal Process., 68 (2020) 3793.