Physica A 389 (2010) 5145–5154

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Prediction problem for target events based on the inter-event waiting time

A. Shapoval *

Financial University under the Government of the Russian Federation, Leningradsky pr., 49, Moscow, Russia
International Institute of Earthquake Prediction Theory, Moscow, Russia

ARTICLE INFO

Article history:
Received 20 August 2009
Received in revised form 27 July 2010
Available online 7 August 2010

Keywords:
Loss function
Coefficient of variation
Earthquake prediction
Prediction of crashes

ABSTRACT

In this paper we address the problem of forecasting the target events of a time series given the distribution ξ of time gaps between target events. Strong earthquakes and stock market crashes are the two types of such events that we are focusing on. In the series of earthquakes, as McCann et al. show [W.R. Mc Cann, S.P. Nishenko, L.R. Sykes, J. Krause, Seismic gaps and plate tectonics: seismic potential for major boundaries, Pure and Applied Geophysics 117 (1979) 1082–1147], there are well-defined gaps (called seismic gaps) between strong earthquakes. On the other hand, usually there are no regular gaps in the series of stock market crashes [M. Raberto, E. Scalas, F. Mainardi, Waiting-times and returns in high-frequency financial data: an empirical study, Physica A 314 (2002) 749–755]. For the case of seismic gaps, we analytically derive an upper bound of prediction efficiency given the coefficient of variation of the distribution ξ. For the case of stock market crashes, we develop an algorithm that predicts the next crash within a certain time interval after the previous one. We show that this algorithm outperforms random prediction. The efficiency of our algorithm sets up a lower bound of efficiency for effective prediction of stock market crashes.

1. Introduction

Forecasting the future of a space–time series, based on the information given at the moment of the prediction, is a typical problem arising in different fields, such as in economics and seismology. A natural statement of the problem is to identify in advance when and where a catastrophic event in the investigated series would occur. Geophysicists have adapted some methods of decision-making theory to evaluate this problem [1–3]. These methods were successfully applied to financial time series as well [4,5].

The joint study of these two seemingly distinct scientific fields is meaningful because (i) the corresponding samples of data are sufficiently full and representative, and (ii) the statistical properties of these samples are described with similar power distributions and laws of self-similarity [6]. Naturally, financial series lack space coordinates. Today, state-of-the-art research dealing with the prediction of market crashes is able to recognize the precursors whose occurrence agrees with the crashes that follow them in a statistically significant way [7,8]. In seismology, the events to be predicted (or the target events) are strong earthquakes. Many precursors to strong earthquakes have been found and tested [9–12]. An algorithm discovered by Kossobokov et al. [13] has been predicting strong earthquakes in real time since 1990.

An arbitrary forecasting algorithm divides the area of interest Ω in space–time onto two disjointed subareas Ω_1 and Ω_2 ($\Omega = \Omega_1 \cup \Omega_2$). An *alarm* is declared on Ω_1 but not on Ω_2. The forecasting algorithm is ineffective if: (i) the majority of target events are not predicted (i.e., they occur on Ω_2), or (ii) the subarea Ω_1 is relatively large. These two characteristics

* Corresponding address: Financial University under the Government of the Russian Federation, Leningradsky pr., 49, Moscow, Russia.
E-mail address: shapoval@mccme.ru.
URL: http://www.mccme.ru/~shapoval.

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.07.033

are called forecast errors, because they have much in common with two types of errors in statistics [14]. A criterion for efficiency based on these two types of forecast errors can be easily formalized. Taking into consideration the arguments of [15], such criteria can be applied to representative samples. That is, samples (i) that have at least several dozen target events presented in the sample, and (ii) where the number of target events is negligible compared to the sample size. Benchmark datasets for the occurrence of strong earthquakes [16] and financial crashes [5] are both representative.

Certainly, the efficiency of developed algorithms has to be higher than the efficiency of random guesses. The next quality level, to be exceeded by any effective algorithm, corresponds to prediction that involves only the sequence of the previous target events and ignores the other part of the sample. In seismology, the applicability of this idea is limited to a small number of regions where the quantity of strong earthquakes is big enough for statistically significant conclusions. New opportunities opened up after the recent discovery of the unified scaling law for the waiting time distribution of seismic events [17,18].

Numerous papers give evidence that the sequence of strong earthquakes is not random [19]. Prediction algorithms based on the waiting time distribution of target earthquakes are partially efficient [20,21]. Similarly, for many financial assets their waiting time distribution agrees better with a power or Weibull distribution than with a Poisson distribution [22–24]. As far as the author knows, a prediction of crashes involving only the waiting time distribution has not been published.

In this paper we find, through strict mathematical methods, the best attainable efficiency of prediction algorithms given the waiting time distribution's coefficient of variation, which is the ratio of the standard deviation to the mean of this distribution. Further, we consider a sample of daily crashes of the Dow Jones Industrial Average (DJIA) and the Hang Seng Index (HSI) and develop a primitive algorithm which identifies these crashes in advance, given the sequence of previous crashes. The efficiency φ of this algorithm is noticeably better than that of random guesses. We introduce this level φ of efficiency to declare inefficient any prediction algorithm that exhibits smaller efficiency.

2. Evaluation of prediction

2.1. Definitions

Prediction as the problem for a decision-maker is discussed in detail [3,14]. We reiterate here the ideas of this theory to clarify the main points of this paper.

Let a sample S of consecutive events be given. Any *event* is determined by the time t of its occurrence and by some of its characteristics, such as size, x_t. Then the events (t, x_t) of a special type ($x_t \in X$, where X is a fixed set) are called *targets*. The problem in question is to identify in advance when the target events occur.

Suppose that the algorithm predicting the target events declares *alarms* at some time instants $t_{1,on}, t_{2,on}, \ldots$. These alarms will last until $t_{1,off}, t_{2,off}, \ldots$. Then the union of the intervals $I_{on} = \cup_i [t_{i,on}, t_{i,off}]$ (probably intersecting one another) forms the alarm set I_{on} such that the alarm is going on at all the points of I_{on}. Otherwise, for any $t \notin I_{on}$ the alarm is off.

The target event is called *predicted* if it occurs at $t \in I_{on}$ and unpredicted if $t \notin I_{on}$. Let n be the ratio of the number of unpredicted events to the number of all target events and τ be the duration of the alarms normalized by the duration of the sample (i.e., $\tau = |I_{on}|/|S|$). The quantities n and τ have much in common with two errors that appear when statistical hypotheses are tested [14]. The *loss* ε of the algorithm is defined as the sum of the prediction errors: $\varepsilon = n + \tau$. Finally, the *efficiency* is $1 - \varepsilon$ [25]. In principle, the efficiency can be chosen among any decreasing function of the variables n and τ. The arguments for and against particular functions are given in [2]. *Prediction algorithms select the set I_{on} to minimize its loss ε.*

2.2. Optimization problem

The choice of the set I_{on} can be easily written as an optimization problem if we assign the function $u(t) \in [0, 1]$, which is called a *strategy*, to the prediction algorithm. Let the equality $u(t) = 1$ imply that the alarm is on at t so that $t \in I_{on}$. Conversely, the equality $u(t) = 0$ means that $t \notin I_{on}$. In general, real values of $u(t) \in (0, 1)$ have sense, too. They correspond to a probabilistic declaration of alarms. Namely, t is set to be in I_{on} with a probability of $u(t)$. This probabilistic definition of I_{on} agrees with the two deterministic cases $u(t) = 0$ and $u(t) = 1$.

Suppose that the distribution of the intervals between the target events is given by its probability density f. This distribution is referred to as the *waiting time distribution*. Then, given functions $u(t)$ and $f(t)$, one can calculate the expected loss ε denoted by $\mathbf{E}\varepsilon$. Namely, let the random variable ξ represent the time between target events. For the sake of simplicity, suppose that the first record in the sample corresponds to the target event occurring at $t = 0$. Then the next target event occurs in $[t, t + dt]$, where dt is small, with a probability $\mathbb{P}\{\xi \in [t, t + dt]\} = f(t)dt$. The failure to predict this target event has a probability $1 - u(t)$, which is the probability that the alarm is off at t, according to the definition of the strategy $u(t)$. Then integration over $[0, \infty)$ yields

$$\mathbf{E}n = \int_0^\infty \big(1 - u(t)\big) f(t) dt. \tag{1}$$

To calculate the mean of τ, one can notice that the interval $[t, t + dt]$ increases the duration of the alarms on dt if (i) the target event is absent on $[0, t)$ (i.e., $\xi \geqslant t$) and (ii) the alarm is declared at t (i.e., $t \in I_{on}$). The probability of these conditions is equal to $F_c(t)u(t)$, where $F_c(t) = \mathbb{P}\{\xi \geqslant t\}$ is the *complementary distribution function* of ξ. Consequently,