Problems of Information Transmission, Vol. 41, No. 1, 2005, pp. 65–69. Translated from Problemy Peredachi Informatsii, No. 1, 2005, pp. 74–78.

SOURCE CODING

Experimental Investigation of Forecasting Methods Based on Data Compression Algorithms[1]

B. Ya. Ryabko* and V. A. Monarev**

**Siberian State University of Telecommunication and Information Science*
Institute of Computational Technologies, Siberian Branch of the RAS, Novosibirsk
boris@ryabko.net

***Institute of Computational Technologies, Siberian Branch of the RAS, Novosibirsk*
vitox@gorodok.net

Received April 8, 2003

Abstract—We suggest and experimentally investigate a method to construct forecasting algorithms based on data compression methods (or the so-called archivers). By the example of predicting currency exchange rates we show that the precision of thus obtained predictions is relatively high.

1. INTRODUCTION

In information theory, a profound interdependence between randomness, or predictability, of a character string and its possible "compressibility" is well known. It was discovered by Kolmogorov [1] and developed by his disciples and colleagues [2, 3]. In late 1980s, after the discovery of optimal universal codes, similar ideas were used to construct optimal prediction methods for sources which generate symbols of finite alphabets [4]. Later, the approach of [4] was used to construct asymptotically optimal forecasting methods for numerous classes of random processes [5–10]. However, most of these methods, though having good asymptotic behaviour, are hardly applicable to real forecasting problems where the amount of observation is small.

The present paper is an attempt to apply data compression algorithms to construct practically applicable prediction methods. As a model problem, we consider that of predicting currency exchange rates (US dollar/euro and US dollar/ruble), which is of a certain practical interest; as data compression methods, we employ practically used archivers. Here it should be noted that these computer programs are based on numerous profound results and constructions of information theory as well as on practical experience and ingenuity of many creators of the archivers. It should also be noted that archivers were found to be an effective tool in authorship attribution [11] and classification of musical compositions [12]; they are also used in estimation of closeness of genetic patterns.

Experimental results presented in the paper demonstrate that the precision of forecasting methods based on arhivers is rather high.

2. DESCRIPTION OF THE CONSTRUCTION OF FORECASTING METHODS BASED ON DATA COMPRESSION ALGORITHMS

First, we introduce necessary notations. Let A be a set, and let $\{X_i\}_{i=0}^{\infty}$ be a random process with values in A.

[1] Supported in part by the Russian Foundation for Basic Research, project no. 03-01-00495, and INTAS, Grant 00-738.

The following prediction problem is considered: given the values $X_0, \dots, X_{n-1}$, to predict the average value of the process at time n, i.e., $E(X_n \,|\, X_0, \dots, X_{n-1})$, in the case where statistical characteristics of the process are a priori unknown. Note that this problem setting is one of the most common [6–10].

Let us now give necessary facts on nondistorting coding of messages. Let D be a finite alphabet, D^n be the set of all words of length n, $n \geq 1$, and $D^* = \bigcup\limits_{n=1}^{\infty} D^n$. A map $\varphi\colon D^n \to \{0,1\}^*$ is called a code. We will consider uniquely decodable codes (sometimes called nonsingular), which by the definition satisfy $\varphi(x) \neq \varphi(y)$ for distinct $x, y \in D^n$. Coding theory also often considers the so-called prefix codes, which possess one more property: for any words $x_1, x_2, \dots, x_m \in D^n$, $m \geq 1$, $n \geq 1$, the sequence $\varphi(x_1)\varphi(x_2)\dots\varphi(x_m)$ can uniquely be decoded as $x_1 x_2 \dots x_m$. It is also well known in information theory that, for any prefix code φ, the Kraft inequality [13]

$$\sum_{u \in D^n} 2^{-|\varphi(u)|} \leq 1$$

is satisfied (here and in what follows, $|x|$ denotes the length of x if x is a word, and the cardinality of x if x is a set.)

In [4] it is suggested to use the Kraft inequality to define the probability distribution on the set of encoded messages,

$$P_\varphi(u) = 2^{-|\varphi(u)|} \Big/ \sum_{v \in D^n} 2^{-|\varphi(v)|},$$

and to predict, or estimate, the occurrence probability of a symbol $d \in D$ at time $n+1$ by the formula

$$P_\varphi(X_{n+1} = d \,|\, X_1 = x_1,\, X_2 = x_2,\, \dots,\, X_n = x_n) = P_\varphi(x_1 x_2 \dots x_n d) \Big/ \Big(\sum_{e \in D} P_\varphi(x_1 x_2 \dots x_n e) \Big). \qquad (1)$$

It is shown there that prediction based on this formula is asymptotically optimal if φ is a universal code. It seems to be appropriate to use (1) for non-prefix codes as well if they "compress" effectively enough the real data for which a prediction method is constructed (though in this case the attractive property of asymptotic optimality of the prediction is not guaranteed; optimal universal codes, as well as prefix codes, possess this property). Moreover, in principle, one may use even non-uniquely-decodable (singular) codes if a metric on symbols of the alphabet is defined (this is the case in the present paper).

Let us emphasize that the prediction based on (1) can only be applied in the case where the random process takes values from a finite set D. We however consider a more general problem, to construct a prediction in the case where the set of values A is, generally speaking, infinite and is equipped with a metric; moreover, unlike [5, 7, 8], we consider not asymptotic properties but possibilities of practical application. Below we describe the general scheme of the prediction method in the case where the set A of values of the process is a real line segment, but the suggested algorithm is easily generalized to the case where A is a part of a multidimensional space. In the description, we consider various possible situations and indicate parameters that can be varied to find a method giving the best prediction precision.

Perhaps, the most obvious way to reduce the case of an infinite "alphabet" A to a finite one is to partition the interval A into k disjoint subintervals $\{d_1, \dots, d_k\}$ (partitioning methods are discussed below). Now, instead of the values $X_0, \dots, X_{n-1}$, let us consider the indices of the subintervals where these values occur and estimate the occurrence probabilities of X_n in the intervals according

Table

Currency exchange rate	Average precision	Archiver	Number of intervals	"Sliding window" size
US dollar/euro	0.00479 (euro)	`Rar`	15	50
US dollar/ruble	1.306 (kopeck)	`Rar`	10	70

to (1). As an estimate of $E(X_n \,|\, X_0, \ldots, X_{n-1})$, we may take the average value of the "step" function generated by the partition $\{d_1, \ldots, d_k\}$.

The prediction precision can be increased by varying the number of subintervals k, the method of partitioning A into k parts, and the compression algorithm used for the prediction. We analyze two ways of partitioning A into k intervals. The first is to divide A into parts of equal length. In the second method, A is partitioned into k subintervals so that each of them contains roughly the same number of sample values, and then this partition is used to predict the next value. As data compression algorithms, we used the commonly known archivers: `Rar`, `arj`, `pkzip`, and `ha`. Finally, let us note that the suggested method can be used in combination with other approaches and methods used in forecasting. Among such methods, considered below in forecasting currency exchange rates, there are "trend elimination" and using for the prediction of the next value not all available dynamic series but only its last part, say, the last 1000 or 50 values, which is often referred to as a "window," or "sliding window." (When using this scheme, it is assumed that statistical characteristics of the process may vary in time, and "old" data contain no information on "new" statistical characteristics.)

3. EXPERIMENTAL RESULTS

We considered the problem of predicting the US dollar exchange rate to ruble and euro using daily data on the value of one dollar in rubles and euros given at the web pages `http://www.x-rates.com` and `http://www.akm.ru` respectively. As was already noted, we made numerous experiments with various values of parameters, methods of "trend" elimination, and archivers used.

All experiments were conducted in two stages, which we conventionally call parameter estimation and testing. Data used for parameter estimation were sequences of successive US dollar values, which we denoted by $x_1 x_2 \ldots x_n$. In the course of experiment, the value x_{n-99} was predicted from the data $x_1 x_2 \ldots x_{n-100}$, the value x_{n-98} from data $x_1 x_2 \ldots x_{n-99}$, etc., so that x_n was predicted from the data $x_1 x_2 \ldots x_{n-1}$. Then we computed the value

$$\delta = \left(\sum_{i=1}^{100} |x_i - x_i^*| \right) / 100, \tag{2}$$

where x_i^* is the predicted value of x_i. Based on the computations made, we chose a variant with the set of parameters with the minimal value of δ. Here, the estimation stage was completed, and we passed to the testing stage, where the chosen variant was used to predict new (the latest) 100 values, already known but (we emphasize!) not used in the preceding computations. Obtained values of the prediction precision, still evaluated by δ, are given in the table of the US dollar values in rubles and euros. In the table we also indicate the values of the parameters found at the estimation stage. We also used division into intervals of equal length and data preprocessing aimed at "trend elimination." Namely, an original sequence $x_1, x_2, \ldots, x_t$ was transformed into the sequence of ratios $(x_2/x_1), (x_3/x_2), \ldots, (x_t/x_{t-1})$, which was used for the prediction. (Of course, δ in (2) was computed from the absolute values but not relative.)

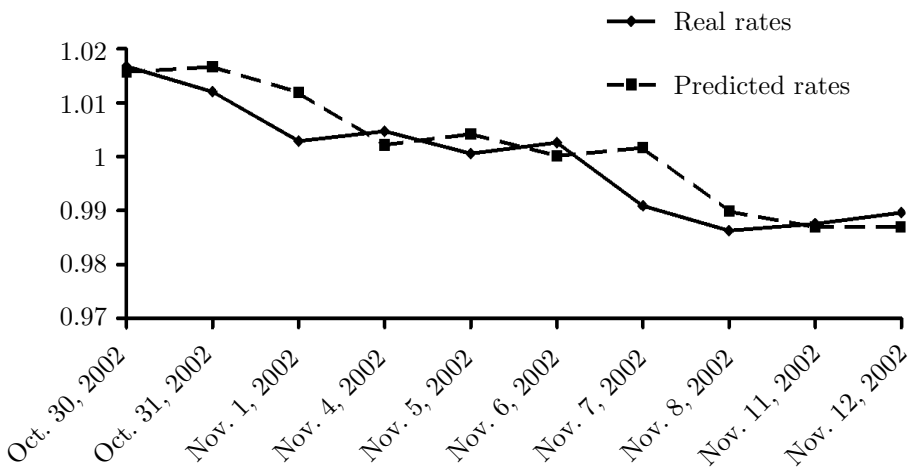

Fig. 1.

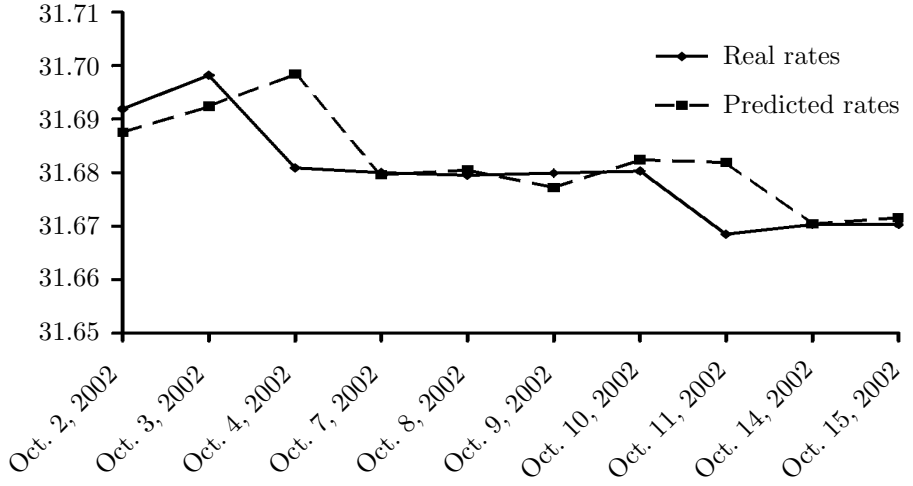

Fig. 2.

In the paper, we used data on the value of dollar in rubles in the period from January 3, 2001, to August 7, 2002, to find parameters giving the minimal error, and in the period from August 7, 2002, to February 26, 2003, to test the prediction precision. Similarly, we used data on the US dolar/euro rate from March 16, 2001, to July 9, 2002, and from July 9, 2002, to December 2, 2002, respectively. The obtained results are given in the table; Figs. 1 and 2 present the prediction results over 10 days, which gives a general insight into the prediction precision. In particular, it is seen that the largest prediction errors occur when the rate changes abruptly.

It is seen from the table that the average error over 100 days for the US dollar/euro rate is 0.00479 euro, and for the US dollar/ruble rate is 1.306 kopeck, which is close to daily exchange fluctuations.

Thus, it is the authors' opinion that the presented data demonstrate that methods of data compression (or universal coding) can be a basis for constructing prediction methods of practical interest.

REFERENCES

1. Kolmogorov, A.N., Three Approaches to the Quantitative Definition of Information, *Probl. Peredachi Inf.*, 1965, vol. 1, no. 1, pp. 3–11 [*Probl. Inf. Trans.* (Engl. Transl.), 1965, vol. 1, no. 1, pp. 1–7].
2. Martin-Löf, P., The Definition of Random Sequences, *Inf. Control*, 1966, vol. 9, no. 6, pp. 602–619.
3. Zvonkin, A.K. and Levin, L.A., Complexity of Finite Objects and the Algorithmic Concepts of Information and Randomness, *Uspekhi Mat. Nauk*, 1970, vol. 25, no. 6, pp. 85–127 [*Russian Math. Surveys* (Engl. Transl.), 1970, vol. 25, no. 6, pp. 83–124].
4. Ryabko, B., Prediction of Random Sequences and Universal Coding, *Probl. Peredachi Inf.*, 1988, vol. 24, no. 2, pp. 3–14 [*Probl. Inf. Trans.* (Engl. Transl.), 1988, vol. 24, no. 2, pp. 87–96].
5. Morvai, G., Yakowitz, S.J., and Algoet, P., Weakly Convergent Nonparametric Forecasting of Stationary Time Series, *IEEE Trans. Inform. Theory*, 1997, vol. 43, no. 2, pp. 483–498.
6. Kieffer, J., Prediction and Information Theory, *Preprint of Univ. of Minnesota*, 1998.
7. Algoet P., Universal Schemes for Learning the Best Nonlinear Predictor Given the Infinite Past and Side Information, *IEEE Trans. Inform. Theory*, 1999, vol. 45, no. 4, pp. 1165–1185.
8. Nobel, A.B., On Optimal Sequential Prediction, *IEEE Trans. Inform. Theory*, 2003, vol. 49, no. 1, pp. 83–98.
9. Ryabko, B.Ya. and Topsoe, F., On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Source, *J. Complexity*, 2002, vol. 18, no. 1, pp. 224–241.
10. Ryabko, B.Ya., The Complexity and Effectiveness of Prediction Algorithms, *J. Complexity*, 1994, vol. 10, pp. 281–295.
11. Kukushkina, O.V., Polikarpov, A.A., and Khmelev, D.V., Using Literal and Grammatical Statistics for Authorship Attribution, *Probl. Peredachi Inf.*, 2001, vol. 37, no. 2, pp. 96–109 [*Probl. Inf. Trans.* (Engl. Transl.), 2001, vol. 37, no. 2, pp. 172–184].
12. Cilibrasi, R., de Wolf, R., and Vitanyi, P., Algorithmic Clustering of Music, LANL e-print cs/0303025, 2003.
13. Gallager, R.G., *Information Theory and Reliable Communication*, New York: Wiley, 1968. Translated under the title *Teoriya informatsii i nadezhnaya svyaz'*, Moscow: Sov. Radio, 1974.