Future Generation Computer Systems 79 (2018) 38–53

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Price forecasting for spot instances in Cloud computing

Zhicheng Cai [a,*], Xiaoping Li [b], Rubén Ruiz [c], Qianmu Li [a]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China
[b] School of Computer Science and Engineering, Southeast University, Nanjing, China
[c] Instituto Tecnológico de Informática, Acc. B. Universitat Politècnica de València, València, Spain

HIGHLIGHTS

- Switching regimes of spot prices are considered for forecasting.
- Markov regime-switching autoregressive based forecasting methods are proposed.
- A dynamic-ARIMA is developed to forecast spot prices too.
- Guides to choose appropriate forecasting methods are given.

ARTICLE INFO

Article history:
Received 16 June 2017
Received in revised form 15 September 2017
Accepted 16 September 2017
Available online 3 October 2017

Keywords:
Cloud computing
Spot price
Forecast
Markov regime-switching
Scheduling

ABSTRACT

Big data applications usually need to rent a large number of virtual machines from Cloud computing providers. As a result of the policies employed by Cloud providers, the prices of spot virtual machine instances behavior stochastically. Spot prices (prices of spot instances) fluctuate greatly or have multiple regimes. Choosing virtual machines according to trends in prices is helpful in decreasing the resource rental cost. Existing price prediction methods are unable to accurately predict prices in these environments. As a result, a dynamic-ARIMA and two markov regime-switching autoregressive model based forecasting methods have been developed in this paper. Experimental results show that the proposals are better than the existing MonthAR in most scenarios.

1. Introduction

From the perspective of big data applications, Cloud users require precise price prediction in order to save on rental costs. Usually, these applications consume a large quantity of computation resources. Cloud computing offers access to hundreds or even thousands of Virtual Machines (VM) for speeding up the processing times of these applications [1]. At the same time, executing big data applications on Cloud computing platforms saves on the cost of establishing and maintaining private data centers. Cloud resource providers provision different pricing models. The commonly used models are fixed price and stochastic price models. For example, Amazon EC2 provisions on-demand VM instances with a fixed price model and spot VM instances with a stochastic model. Generally resources with stochastic pricing models offer cheaper prices than those with fixed pricing models. Since reserved and on-demand VM instances are of a fixed price, only spot VM instances are considered for price prediction. Spot prices are stochastically set as a result of auctioning spot VM instances according to real time user demands. Spot VM instances of different VM types in different physical regions have different stochastic spot prices. A VM instance is out-of-bid if the spot price is higher than that of the current bid. These characteristics make spot prices fluctuate. Good price forecasting is helpful for choosing appropriate VM types, selecting the correct rental periods and setting optimal bids to save on rental costs.

In auction based public Clouds, stochastically arriving user demands and unpredictable user bids determine final spot prices which make predicting spot prices complex [2–4]. Recent trends, such as large fluctuations and switching regimes (different statistical means and variances in different time periods), make it even more difficult. Inter-price times are lengths of intervals between spot price changes. The peaks of the probability density functions (PDFs) of inter-price times in 2010 are around two hours [4], i.e., prices during one hour are usually the same in 2010. However, recent PDFs of inter-price times show that spot prices change more frequently and greatly. For example, although the price changes smaller than 5% of the average price have been ignored, the PDFs

* Corresponding author.
E-mail addresses: caizhicheng@njust.edu.cn (Z. Cai), xpli@seu.edu.cn (X. Li), rruiz@eio.upv.es (R. Ruiz), qianmu@njust.edu.cn (Q. Li).

https://doi.org/10.1016/j.future.2017.09.038

of inter-price times (during the period from 28-04-2016 to 28-07-2016 and the period from 03-04-2017 to 15-05-2017) demonstrate that 41.6% of inter-price times are smaller than one hour. Spot prices of some VM types even exhibit switching regimes. Fig. 1 shows the spot prices of the Amazon EC2 VM types "m4.2xlarge-us-east-1b-linux-unix" and "m4.2xlarge-us-east-1d-linux-unix". Prices of different time periods have different statistical characteristics such as mean values and variances which indicate that spot prices switch among several hidden regimes. Ben-Yehuda et al. [3,5] studied trace files of the Amazon EC2 and tried to discover how Amazon sets prices for its unused EC2 capacities. It is possible that Amazon's EC2 spot prices are limited by a dynamic bottom price (determined by an autoregressive model) which ignores the bids that are lower than the bottom price. High spot prices may reflect market changes but most low prices are usually indicative of dynamic bottom prices, i.e., the two factors indicate that spot prices have two or more different regimes. The existence of different regimes means spot prices are nonlinear.

In existing scheduling algorithms, different types of probability models have been used to help Cloud users recognize changes in spot prices such as the one or multiple step probability matrix of transition from one price to another [6–8], the probability density function of spot prices [9], the probability of an out-of-bid event within a time interval [10–12], the probability of spot instances staying available over time given a starting price and a bid [13] and Q-learning based action selecting rules [14], etc. Usually static probability models are used to describe the transition probabilities among prices or probability density functions of failures as a whole, whereas the correlation of multiple sequential prices is not considered. This means that the corresponding methods cannot predict trends in sequential prices.

Autoregressive models consider the correlation of multiple sequential prices which can predict trends of spot prices [2]. However, large spot price fluctuations lead to many unstable spot price time series which decrease the performance of existing autoregression based prediction methods designed for linear and stable time series [15]. Autoregressive integrated moving average model (ARIMA) is an extension of autoregression which decreases the impact of unstable trends on predictions by differencing [15]. Single, double and triple exponential smoothing methods can also be used for modeling unstable time series considering trends and seasonality [16]. Predicted values of single exponential smoothing (SES) and double exponential smoothing (DES) are on a horizontal line and an oblique straight line respectively, therefore SES and DES are not suitable for long-term prediction. Triple exponential smoothing models both trends and seasonality. Exponential smoothing methods only find a unique combination of parameters trying to fit all data. For example, there is only one smoothed seasonality pattern for all data in triple exponential smoothing. At the same time, traditional autoregression and ARIMA methods assume that the time series is linear and that there is only a single regime for which a uniform model is built. However, many spot prices are nonlinear and there are different regimes with (or without) different seasonality patterns. It is hard to find a uniform autoregression, ARIMA or exponential smoothing model suitable for all switching regimes and accurately forecast prices of different regimes. Therefore, building models for different regimes respectively and choosing appropriate regimes for forecast are crucial in accurate prediction. Nonlinear models are usually used for describing nonlinear time series such as threshold autoregressive models (TAR) [17] and Markov regime-switching autoregressive models (MRS-AR) [18]. TAR extended from autoregression builds different linear autoregressive models for different regimes and the switching among regimes depends on transition variable values. It is very complicated to define an appropriate transition variable [19]. For example, spot prices with the same threshold

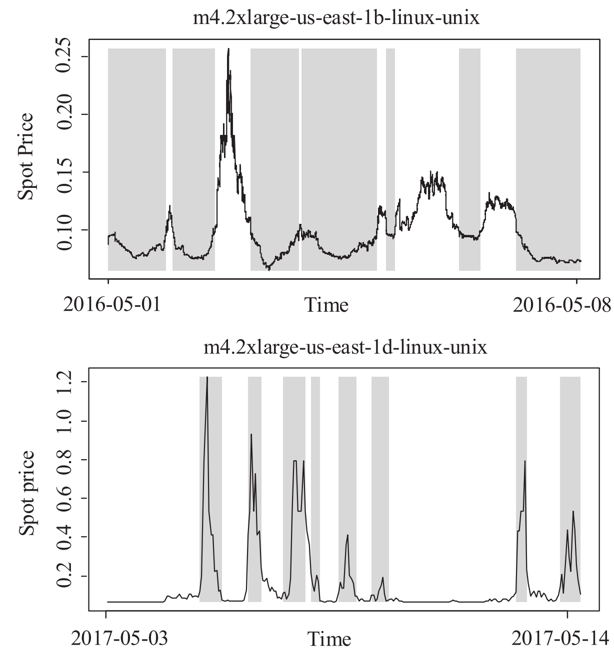

Fig. 1. The spot prices in dollars of Amazon EC2 virtual machine type "m4.2xlarge-us-east-1b-linux-unix" of the period from 01-05-2016 to 08-05-2016 and "m4.2xlarge-us-east-1d-linux-unix" of the period from 03-05-2017 to 14-05-2017.

variable values may belong to different regimes when we use spot prices (or lagged prices) as transition variables directly. MRS-AR is a generalized version of TAR in which regime switching is much more flexible [20]. In MRS-AR, a Markov stochastic process is used to model the switching of regimes where regimes are considered as states of the Markov stochastic process. The probability transition matrix describes the transition among regimes rather than defined transition variables in TAR. The Markov process part is used to describe the switching among regimes and the AR part is enclosed to model the trend in each regime. Therefore, MRS-AR is used to predict spot prices in this paper.

For spot prices with switching regimes, it is crucial to determine the number of regimes and build different models for different regimes. In this paper, the DBSCAN (density-based spatial clustering of applications with noise) clustering algorithm is adopted to determine the number of regimes. Then, MRS-AR with different autoregression models for different regimes are established. Choosing correct regimes is crucial in accurate forecasting which means misspecification of the regime can lead to substantial losses in forecast accuracy [21]. In the literature, the absolute probability distribution over regimes of each prediction step, obtained from the conditional probability distribution and transition matrices among regimes of MRS-AR (see details in Section 5.1.3), is usually used to forecast short-term prices. For example, the regime with the largest probability is considered as the forecasted regime only for a one-step forecast [19]. The expected mean determined by the absolute probability distribution [21] is used for prediction without specifying regimes. However, the absolute probability distribution over states of non-seasonal Markov chain usually converge to the stationary distribution quickly which cannot forecast long-term regime switching accurately. In this paper, two methods are proposed to choose appropriate regimes for future times where transition probabilities of regimes and an autoregressive-moving-average model (ARMA) are combined. To predict short-term prices, we assume that they belong to the same regime (lasting rule) as the last spot price which results in the first prediction method