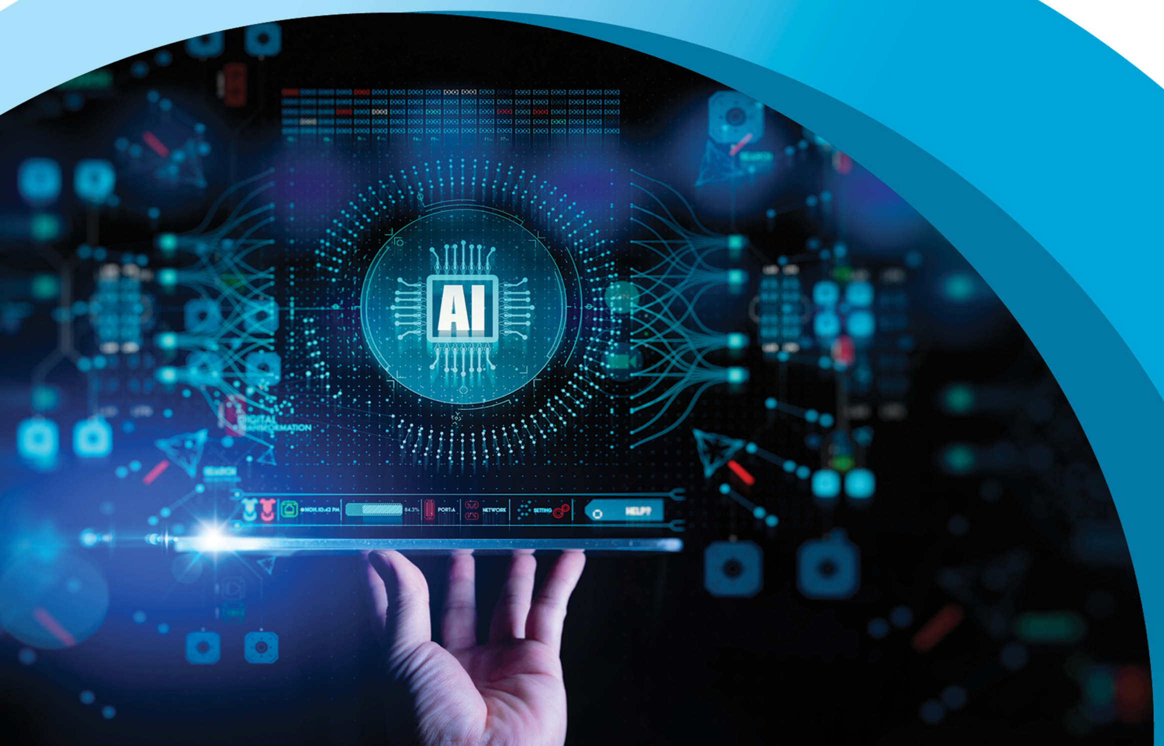
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Machine Learning for Cloud Management

Jitendra Kumar
Ashutosh Kumar Singh
Anand Mohan
Rajkumar Buyya



Machine Learning for Cloud Management

Machine Learning for Cloud Management

Jitendra Kumar
Ashutosh Kumar Singh
Anand Mohan
Rajkumar Buyya

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2022
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data

Names: Kumar, Jitendra, 1975- author. | Singh, Ashutosh Kumar, author. |
Mohan, Anand (Of Indian Institute of Technology), author. | Buyya,
Rajkumar, 1970- author.
Title: Machine learning for cloud management / Jitendra Kumar, Ashutosh
Kumar Singh, Anand Mohan, Rajkumar Buyya.
Description: First edition. | Boca Raton : CRC Press, 2022. | Includes
bibliographical references and index.
Identifiers: LCCN 2021027713 | ISBN 9780367626488 (hardback) | ISBN
9780367622565 (paperback) | ISBN 9781003110101 (ebook)
Subjects: LCSH: Cloud computing. | Machine learning.
Classification: LCC QA76.585 .K85 2022 | DDC 004.67/82--dc23
LC record available at https://lccn.loc.gov/2021027713

ISBN: 978-0-367-62648-8 (hbk)
ISBN: 978-0-367-62256-5 (pbk)
ISBN: 978-1-003-11010-1 (ebk)

DOI: 10.1201/9781003110101

Publisher's note: This book has been prepared from camera-ready copy provided by the authors.

Dedicated to,

My wife: Gita, daughter: Aru, and Parents

∼Jitendra Kumar

Anushka, Aakash, Akankshya, and Parents

∼Ashutosh Kumar Singh

My wife: Sudha Mohan, son: Ashish Mohan, daughter: Amrita Mohan, and Late parents

∼Anand Mohan

My international collaborators and team members in Melbourne CLOUDS Lab!

∼Rajkumar Buyya

Contents

List of Figures

List of Tables

Preface

Cloud computing has become one of the revolutionary technology in the history of the computing world. It offers subscription-based on-demand services and has emerged as the backbone of the computing industry. It has enabled us to share resources among multiple users through virtualization by the means of creating a virtual instance of a computer system running in an abstracted hardware layer. Unlike early distributed computing models, it assures limitless computing resources through its large-scale cloud data centers. It has gained wide popularity over the past few years, with an ever-increasing infrastructure, number of users, and amount of hosted data. The large and complex workloads hosted on these data centers introduce several challenges: resource utilization, power consumption, scalability, operational cost, and many others. Therefore, a practical resource management scheme is essential to bring operational efficiency with improved elasticity. The elasticity of a system depends on several factors such as the accuracy of anticipated workload information, performance behavior of applications in different scenarios communicating the forecast results, use of the anticipated information, and many others.

Effective resource management can be achieved through workload prediction, resource scheduling, and provisioning, virtual machine placement, or a combination of these approaches. The workload prediction has been widely explored and a number of methods are presented. However, the existing methods suffer from various issues including the incapability of capturing the non-linearity of workloads and iterative training that consumes huge computing resources and time. This book discusses the machine learning-based approaches to address the above-mentioned issues. The highlights of the discussed models are continuous learning from error feedback, adaptive nature, decomposition of workload traces, and ensemble learning. Detailed analysis of predictive methods on different workload traces is also included and their performance is compared with state-of-art models. Virtual machine placement is another aspect that is explored to achieve efficient resource management. In general, virtual machine placement is a multiobjective problem that involves multiple conflicting objectives to be optimized simultaneously. The frameworks discussed in this book address the issues of resource utilization, power consumption, and security while placing the workloads on servers.

The remainder of the book is organized as follows: Chapter 1 briefs the basic cloud computing concepts. The discussion on the workload prediction mechanisms begins in chapter 2. First, the basic time series forecasting models are discussed with their performance on different workload traces. Chapter 3 discusses the error preventive time series forecasting models which significantly improve the performance over classical time series models. Then, a discussion on various nature-inspired algorithms is included