Ann. Data. Sci. (2014) 1(1):1–3
DOI 10.1007/s40745-014-0001-0

Preface

Inaugural Volume of the *Annals of Data Science*: Optimization and Data Science

Yong Shi · Yingjie Tian

Published online: 29 April 2014

This inaugural volume, called Optimization and Data Science, formally announces the establishment of the *Annals of Data Science* (ADS) to disseminate the cutting-edge research findings, experimental results and case studies of data science that represent trends and development in data science. Although Data Science now generally is regarded as an interdisciplinary field of using mathematics, statistics, databases, data mining, high-performance computing, knowledge management and virtualization to discover knowledge from Big Data, it should have its own scientific contents, such as axioms, laws and rules, which are fundamentally important for experts in different fields to explore their own interests from Big Data.

The volume contains the fundamental issues in the interface of optimization and data science. Recently, both researchers and practitioners are paying more attentions to development of data science, which combines various techniques and theories from many fields ranging from mathematics, statistics, data engineering, artificial intelligence, data mining and optimization. This volume consists of the contributed papers dealing with data science problems by using various optimization methods. They are also involved in computing complexity, knowledge management and decision making.

The first paper, "Approximation of Irregular Geometric Data by Locally Calculated Univariate Cubic L1 Spline Fits," by Ziteng Wang, John Lavery and Shu-Cherng Fang, explored the research issues on L1 splines that are under development for interpolation and approximation of irregular geometric data. It investigates the advantages in terms

Y. Shi (✉) · Y. Tian
Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
e-mail: yshi@ucas.ac.cn

Y. Shi
College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA

of shape preservation and computational efficiency of calculating univariate cubic L1 spline fits using a steepest-descent algorithm to minimize a global data-fitting functional under a constraint implemented by a local analysis-based interpolating-spline algorithm on 5-node windows. Comparison of these locally calculated L1 spline fits with globally calculated L1 spline fits previously reported in the literature indicates that the locally calculated spline fits preserve shape on the average slightly better than the globally calculated spline fits and are computationally more efficient because the locally-calculated-spline-fit algorithm can be parallelized. The second paper, "Exact Solutions to the Capacitated Clustering Problem: A Comparison of Two Models," by Mark Lewis, Haibo Wang and Gary Kochenberger, investigates a natural nonlinear alternative to a standard linear model for CCP and compare the two models on a set of test problems. Our results show that moderate sized instances of CCP can in fact be solved optimally with modern exact methods in modest amounts of time and that the quadratic model generally outperformed its equivalent linear alternative in terms of quickly finding optimal or near optimal solutions. The third paper, "Proposal of New Objective Measures for Mining Association Rules: Cannibalization and Unexpectedness," by Hidenobu Hashikami and Masato Koda, proposes two new objective measures for mining association rules to solve the problems. The first measure is the degree of cannibalization between itemsets, which is bounded up with marketing strategy, and the second is the objective measure that intends to discover unexpected rules in the database. Experimental studies with application to public dataset and comparison of running time using synthetic datasets demonstrate the validity and effectiveness of the proposed measures. The forth paper, "Discussions on The Existence of Strongly Polynomial-Time Simplex Variants and Hirsch Conjecture," by Pei-Zhuang Wang, presents a preliminary analysis for Hirsch conjecture based on the theory of cone-cutting. The author indirectly proves the related assembling conjecture under the condition of 'Fully cuts' and suggests a relaxed conjecture in this paper to deny the existence of super polynomial diameters on LP polytopes and to offer an optimistic view on the existence of strongly polynomial-time LP algorithms. The author presents an algorithm ConePairCut in the paper, which holds the complexity O(m2n2) for LP problems under conditions $r(A) = n$ and $b \geq 0$. The result may lead to a strongly polynomial-time simplex variant if the condition $r(A) = n$ is not an essential constraint.

The fifth paper, "Heterogeneous Embedding via Aggregating Multiple Sources," by Xiaoxiao Shi and Philip S. Yu, proposes a principle of collective component analysis (CoCA), in order to find the optimal embedding across a mixture of vector-based features and graph relational features. The CoCA principle is to find a feature subspace with maximal variance under two constraints. First, there should be consensus among the projections from different feature spaces. Second, the similarity between connected data (in any of the network databases) should be maximized. The optimal solution is obtained by solving an eigenvalue problem. Moreover, the paper discusses how to use prior knowledge to distinguish informative data sources, and optimally weight them in CoCA. Since there is no previous model that can be directly applied to solve the problem, it devises a straightforward comparison method by performing dimensionality reduction on the concatenation of the data sources. Three sets of experiments show that CoCA substantially outperforms the comparison method.

The sixth paper, "SMAA-AD Model in Multicriteria Decision-making Problems with Stochastic Values and Uncertain Weights," by Feng Yang, Fuguo Zhao, Liang Liang and Zhimin Huang, considers the stochastic multicriteria decision-making (MCDM) problems with multiple alternatives, stochastic criterion values and uncertain criterion weights. It proposes SMAA-AD model and illustrate how SMAA-AD model is used in such stochastic multicriteria decision-making problems. In SMAA-AD model, absolute dominant method (ADM) is used to turn stochastic criterion values into deterministic absolute dominant values, and stochastic multicriteria acceptability analysis (SMAA) is used to rank the alternatives without foreknowing the decision maker' preference on criterion weights. SMAA-AD model provides three indices, i.e., rank acceptability index, holistic acceptability index and central weight vector, to support the decision in the stochastic MCDM problems. SMAA-AD model overcomes some shortcomings of traditional multicriteria decision-making methods. For example, it needs not to predefine any parameters and functions. The paper uses a case of technology competition for cleaning polluted soil in Helsinki to illustrate our method. The seventh paper, "A New Nonlinear Multiregression Model Based on the Lower and Upper Integrals," by Jing Chua, Zhenyuan Wang, Yong Shi and Kwong-Sak Leunge, establishes a new nonlinear multiregression model based on a pair of extreme nonlinear integrals: lower and upper integrals. A complete data set of predictive attributes and the relevant objective attribute is required for estimating the regression coefficients. Due to the nonadditivity of the model, a genetic algorithm combined with the pseudo gradient search is adopted to search the optimized solution in the regression problem. Applying such a nonlinear multiregression model, an interval prediction for the value of the objective attribute can be made once a new observation of predictive attributes is available. The last paper, "Toward Extenics-based Innovation Model on Intelligent Knowledge Management," by Xingsen Li, Liping Li and Zhengxin Chen, analyzes on current innovation models and methods, and presents a combined innovation model using intelligent knowledge management and extension transformation methods based on the new cross discipline Extenics. The model process the knowledge discovered from data mining into a tree structure and save them in knowledge base in basic-element format. The paper then explores the innovation paths and its directions by a formularized model by human-computer interaction method based on Extenics. The model can objectively describe how the innovation solutions are created. Furthermore, it presents a management innovation case to support our model. The framework is proved useful for practical applications.

ADS encourages the contributors around the world to address different challenging problems in Data Science, especially, how to discovering knowledge from heterogeneous data under Big Data environment!

Editor-in-Chief
Yong Shi

Managing Editor
Yingjie Tian