New Generation Computing, 27(2009)281-284
Ohmsha, Ltd. and Springer

Preface

Special Issue on Data-Mining and Statistical Science

Takashi WASHIO
I. S. I. R., Osaka University
8-1 Mihogaoka, Ibaraki City, Osaka, 567-0047, JAPAN
washio@ar.sanken.osaka-u.ac.jp

The recent development of information technology drastically increases the amount of data generated, collected and stored in our society and allows us to access these massive-scale and complex data through computer networks. The issue of how to analyze such data and utilize them is becoming more and more important. Data mining is a technology to address this issue, and enables us to mine significant knowledge from such a large amount of data. "The Third International Workshop on Data Mining and Statistical Science (DMSS2008)" [1] was held to focus on this technology, its background theory and methodology, bring together researchers from machine learning and statistical science, and have extensive discussions on the associated diverse aspects. Upon the stimulative discussion in this workshop, this special issue has been planned to further acquire the high quality work not limited to the presentations that appeared at the DMSS2008 workshop but widely from public. We solicited papers addressing theoretical and methodological aspects of machine learning, statistical science, and the other relevant fields, which contribute to understanding and development of data mining techniques.

For the last decade, we have dealt with various disciplines of data mining in various workshops in the Japanese Society for Artificial Intelligence. On the other hand, there are strong requests from data mining researchers for more extensive and structured discussions on statistical science and machine learning. In response to such movement, we have started a domestic workshop series named "Data Mining and Statistical Mathematics (SIG-DMSM)." [2] This workshop deals with statistical and machine learning aspects of data mining, and aims at creating new data mining paradigms. Upon the activity of SIG-DMSM, the machine learning researchers and statisticians associated with this workshop decided to further set up some opportunity to present and discuss advanced researches with international participants, and started the annual international workshop (DMSS) within the framework of SIG-DMSM. The DMSS workshop has been organized every year since 2006 by co-locating with SIG-DMSM. Highly

qualified papers from all over the world have been invited to these workshops. The third DMSS (DMSS2008) was held on September 25 and 26, 2008, at Tokyo Institute of Technology, Tokyo, Japan. A total of three invited talks and sixteen contributed talks were presented. This special issue of New Generation Computing (NGC) has come out from these extensive activities of the researchers in data mining field.

The papers submitted to this special issue have been reviewed by following the regular reviewing criteria of NGC. We had 14 paper submissions. The guest editorial board was formed by a guest volume editor and 10 guest board members. The guest volume editor assigned a few submitted papers to every guest board member, and every member assigned each paper to three reviewers. The process was blind review, and all guest editorial board members participated in the discussion on the review results and made their judgment. The first review round selected highly outstanding papers only, and the second review round further selected only 4 papers well addressed the review comments provided in the first round. All of these papers address fundamental and important issues of machine learning and statistics, and provide highly promising outcomes.

The first paper titled as "A Novel Margin Based Algorithm for Feature Extraction" provided a very fundamental and important technique for machine learning. This study proposed a technique named Adaptive Margin Maximization (AMM) which uses a boosting principle for a dynamic weighting of features in data, and extract important features to build a classifier having a large margin. Because the features needed for the large margins of individual instances are efficiently extracted, this technique yields a classifier having small generalization error. This significant performance has been demonstrated via extensive numerical experiments.

The second paper titled as "Efficient Leave-m-out Cross-Validation of Support Vector Regression by Generalizing Decremental Algorithm" also addressed a fundamental and important issue in machine learning and statistics. This study proposed a computationally very efficient method for cross-validation of the Support Vector Regression (SVR). A decremental algorithm of SVR was extended to remove multiple data points efficiently from a given training set. This proposed technique reduces the main computational cost of the leave-m-out cross-validation from $O(m)$ to $O(\sqrt{m})$, and it has been well demonstrated through numerical experiments.

The third paper titled as "Constrained Motif Discovery in Time Series" proposed a novel framework to find unknown recurring patterns in sequential data. The motif discovery in sequential data is one of central tasks in time series analysis of statistics and genome pattern analysis of bio-informatics, and thus this work attacked a very important and fundamental issue in data mining. This study presented two efficient algorithms, MCFull and MCInc, to introduce domain knowledge into the motif mining process in form of constraints, and further presented a novel change-point detection algorithm called the Robust Singular Spectrum Transform (RSST). Extensive numerical experiments showed that the combination of these algorithms indicates significant computational

efficiency without any loss of accuracy.

The final paper titled as "Decoding Algorithm of Low-density Parity-check Codes based on Bowman-Levin Approximation" is to introduce principles of statistical physics into optimization and machine learning problems. It introduces a novel optimization principle named the Bowman-Levin (BL) approximation in statistical physics, and showed its better performance over the conventional Belief Propagation (BP) and Concave-Convex Procedure (CCCP) under a condition that large computation cost is needed. The results of this study provide very important insights to the optimization research.

We would like to thank all of the authors and the reviewers of the published papers. We also thank for the support of the Editorial Board of NGC, the Editorial Office and the publisher Ohmsha, Ltd. Finally, the guest editor is very grateful to the following guest editorial board members for their devoted effort on this special issue.

Takashi Washio, Guest Editor

References

1) http://sigdmsm.org/dmss2008/
2) http://sigdmsm.org/