1. INTRODUCTION

A huge amount of data of different formats is continually being collected from multiple sources. This fast growth in data has highlighted the requirement of developing new tools and techniques that can intelligently assist analysts to discover meaningful and useful knowledge. Data mining for knowledge discovery has emerged as a promising discipline in Computer Science in order to automatically extract patterns representing knowledge from large datasets. Recently, knowledge discovery using data mining has been tackled as a multidisciplinary issue.

This book explores the tools and techniques of improving knowledge discovery using diverse data mining methods. It is a timely and important contribution to the field as it covers a wide range of state-of-the-art research in this rapidly growing area. In particular, contributing authors describe how knowledge discovery can be improved by integrating data mining techniques. Notable techniques featured in the book include multi-relational data mining, incremental learning, variants of C4.5, artificial immune system, rough computing, metaheuristics, etc. The list is not exhaustive. Together, the book forms a body of knowledge that can have applications to many real world problems. These include farming and food production (separating crops and weeds), healthcare (cancer pathways), finance, hazards aversion (earthquakes).

This preface introduces the general context, the aims and the rationale of this book with a brief description of each chapter’s contents. The motivation for integrating data mining, data warehousing and machine learning in non-conventional application domains for improving the knowledge discovery process is given.

1.1 Knowledge Discovery from Large Datasets

Knowledge discovery from large datasets is the result of an exploratory process involving the application of various algorithmic procedures for manipulating data (Bernstein, Provost et al. 2005). It aims to extract valid, novel, potentially useful, and ultimately understandable patterns from data (Fayyad, Piatetsky-Shapiro et al. 1996). Data mining and data warehousing are two key technologies for discovering knowledge from large datasets. Data mining enables the discovery of hidden trends from large datasets, while data warehousing provides for interactive and exploratory analysis of data through the use of various data aggregation methods.

In the past several years, a wide range of data mining techniques have made significant contributions to the field of knowledge discovery in a number of domains. In the banking sector, these techniques are
used for loan payment prediction, customer credit policy analysis, classification of customers for targeted marketing, and the detection of money laundering schemes and other financial crimes. Similarly, in the retail industry, such techniques are used in the analysis of product sales and customer retention. In the telecommunication industry these techniques help in identifying and comparing data traffic, system workload, resource usage, profit and fraudulent pattern analysis (Han and Kamber 2006).

Likewise, data warehousing has contributed extensively as a key technology for complex data analysis, decision support and automatic extraction of knowledge from huge data repositories (Nguyen, Tjoa et al. 2005). It provides analysts with a competitive advantage by providing relevant information to enhance strategic decision making. Moreover, warehousing has reduced costs by tracking trends, patterns, and exceptions over long periods in a consistent and reliable manner. Due to sophisticated analytical powers, these warehouse systems are being used broadly in many sectors such as financial services, consumer goods and retail, manufacturing, education, medical, media, and telecommunication. More recently, there has been an increasing research interest in the knowledge engineering community towards integrating the two technologies (Goil and Choudhary 2001; Liu and Guo 2001; You, Dillon et al. 2001; Zhen and Minyi 2001; Wang, Feng et al. 2002; Ohmori, Naruse et al. 2007; Usman and Pears 2010; Usman and Asghar 2011).

1.2 Integrated Use of Data Mining and Data Warehousing for Improving Knowledge Discovery

Both data mining and data warehousing technologies have essentially the same set of objectives and can potentially benefit from each other’s methods to facilitate knowledge discovery. Each technology is mature in its own right, and despite the very clear synergy between these two technologies, they have developed largely independent of each other.

The integrated use of data mining and data warehousing techniques such as Online Analytical Processing (OLAP) has received considerable attention from researchers and practitioners alike, as they are key tools used in knowledge discovery from large data datasets (Han 1998; Sapia, Höfling et al. 1999; Goil and Choudhary 2001; You, Dillon et al. 2001; Zhen and Minyi 2001; Ohmori, Naruse et al. 2007; Zubcoff, Pardillo et al. 2007; Pardillo, Zubcoff et al. 2008). (Usman, Asghar et al. 2009) used a hierarchical clustering technique in conjunction with multidimensional scaling (Cox and Cox 2008) to design schema at different levels of data abstraction. They developed an iterative method that explores the similarities and differences in information contained across consecutive levels in the cluster hierarchy. The presentation of such information at different levels of abstraction provides decision makers with a better understanding of the patterns and trends present in the data. Although, a variety of integrated approaches have been proposed in the literature to mine large datasets for discovering knowledge. However, a number of issues remain unresolved in the previous work (Sarawagi, Agrawal et al. 1998; Sarawagi 2001; Kumar, Gangopadhyay et al. 2008; Ordonez and Zhibo 2009), especially on intelligent data analysis front.

The integrated use of data mining and data warehousing techniques such as Online Analytical Processing (OLAP) has received considerable attention from researchers and practitioners alike, as they are key tools used in knowledge discovery from large data datasets. Although, a variety of integrated approaches have been proposed in the literature to mine large datasets for discovering knowledge. However, a number of issues remain unresolved in the previous work, especially on intelligent data analysis front.
This book aims to discuss and address the difficulties and challenges that of seamless integration of the two core disciplines of knowledge discovery. The editor will seek chapters that address different methods and techniques of integration for enhancing the overall goal of knowledge discovery. Additionally, the book will explore the impact of such techniques in a variety of application domains ranging from government, education, science, agriculture engineering etc.

1.3 Aims and Objectives

The primary objective of this book is to provide insights concerning the integration of data mining and data warehousing for enhancing the knowledge discovery process. This is a front-line and important topic that is of interest in both industry and knowledge engineering research community. The current approaches of knowledge discovery in industry are ad-hoc where data mining and warehousing is dealt separately. There is no standard rule of thumb in integrating these two disciplines. This book reports on the existing gaps in this area and presents the novel approaches to bridge the existing gaps.

1.4 Target Audience and Topics Covered

The target audience of this book includes decision makers, academicians, researchers, advanced-level students, technology developers, and Business Intelligence (BI) professionals will find this text useful in furthering their research exposure to relevant topics in knowledge discovery and assisting in furthering their own research efforts in this field.

This book covers the following topics relating to data mining integration with data warehousing.

- Data mining techniques: clustering, classification, association rules, decision trees, etc.
- Data and knowledge representation
- Knowledge discovery framework and process, including pre- and post-processing
- Integration of data warehousing, OLAP and data mining
- Exploring data analysis, inference of causes, prediction
- Interactive data exploration/visualization and discovery
- Data Mining, Data warehousing and OLAP tools
- Data warehousing applications: corporate, scientific, government, healthcare, bioinformatics
- Data mining applications: bioinformatics, e-commerce, Web, intrusion/fraud detection, finance, healthcare, marketing, telecommunications, etc.
- Data mining support for designing information systems

I hope this book will highlight the need for growth and research in the area of improving knowledge discovery through the integrated use of data mining techniques. This volume consists of sixteen chapters in four sections.

Chapter 1, “Integration of Data Mining and Statistical Methods for Intelligent Cube Construction and Exploration” by Muhammad Usman, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Pakistan, identified that data cube construction and exploration has been dealt separately in the literature and there is very limited amount of work done in the past which guides the data warehouse designers and analysts to automatically construct and intelligently explore the data cubes. A novel methodology has been proposed that utilizes hierarchical clustering and Principal Component Analysis (PCA)
to generate informative data cubes at different levels of data abstraction. Moreover, automatically ranked cube navigational paths are provided by their proposed methodology to enhance knowledge discovery from large data cubes. The methodology has been validated using real world dataset taken from UCI machine learning repository and the results show that the proposed approach assists in cube design and intelligent exploration of interesting cube regions.

Chapter 2, “Online Processing of End-User Data in Real-Time Data Warehousing” by Muhammad Asif Naeem, Auckland University of Technology, New Zealand, highlights that stream-based join algorithms are a promising technology for modern real-time data warehouses. A particular category of stream-based joins is a semi-stream join where a single stream is joined with a disk based master data. The join operator typically works under limited main memory and this memory is generally not large enough to hold the whole disk-based master data. Recently, a seminal join algorithm called MESHJOIN (Mesh Join) has been proposed in the literature to process semi-stream data. MESHJOIN is a candidate for a resource-aware system setup. However, MESHJOIN is not very selective. In particular, MESHJOIN does not consider the characteristics of stream data and its performance is suboptimal for skewed stream data. The author presents a novel Cached-based Semi-Stream Join (CSSJ) using a cache module in this chapter. The proposed algorithm is more appropriate for skewed distributions, and the results are presented for Zipfian distributions of the type that appear in many applications. Rigorous experimental study has also been conducted to test the proposed algorithm. Experiments show that CSSJ outperforms MESHJOIN significantly.

Chapter 3, “Multi-Relational Data Mining: A Comprehensive Survey” by Ali H. Gazala, Auckland University of Technology, New Zealand and Waseem Ahmad, Auckland University of Technology, New Zealand provides a comprehensive survey on growing area of Multi-Relational Data Mining or MRDM. This area focuses on discovering hidden patterns and useful knowledge from relational databases. While the vast majority of data mining algorithms and techniques look for patterns in a flat single-table data representation, the sub-domain of MRDM looks for patterns that involve multiple tables (relations) from a relational database. This chapter discusses the different approaches researchers have followed to explore the relational search space while highlighting some of the most significant challenges facing researchers working in this sub-domain. The chapter also describes number of MRDM systems that have been developed during the last few years and discusses some future research directions in this sub-domain.

Chapter 4, “Comparative Study of Incremental Learning Algorithms in Multidimensional Outlier Detection on Data Stream” by Simon Fong, University of Macau, Macau SAR focused on multidimensional outlier detection (MOD) over data streams. MOD is one of the most significant data stream mining techniques and when multivariate data are streaming in high speed, outliers are to be detected efficiently and accurately. They highlighted that the conventional outlier detection method is based on observing the full dataset and its statistical distribution. The data is assumed stationary. However, this conventional method has an inherent limitation—it always assumes the availability of the entire dataset. In modern applications, especially those that operate in the real time environment, the data arrive in the form of live data feed; they are dynamic and ever evolving in terms of their statistical distribution and concepts. Hence, it becomes unrealistic for an outlier detection method to wait for all the data when fresh data are streaming in rapidly and continuously. In this chapter, they evaluate the effectiveness of a collection of incremental learning algorithms which are the underlying pattern recognition mechanisms for MOD. Furthermore, their performance has been tested in processing some real-life samples of real time data feed, for multidimensional outlier detection.
Chapter 5, “Advances of Applying Meta-heuristics to Data Mining Techniques” by Simon Fong, University of Macau, Macau, Jinyan Li, University of Macau, Macau and Athanasios V. Vasilakos, Kuwait University, Kuwait presents the importance of applying Meta-heuristics and their successful use in optimization applications ranging from financial modeling, image processing, resource allocations, job scheduling to bioinformatics. In particular, meta-heuristics have been proven in many combinatorial optimization problems. The aim of this chapter is to highlight some recent research related to meta-heuristics and to discuss how they can enhance the efficacy of data mining algorithms. An upmost challenge in Data Mining is combinatorial optimization that, often lead to performance degradation and scalability issues. Two case studies are presented, where meta-heuristics improve the accuracy of classification and clustering by avoiding local optima.

Chapter 6, “Artificial Immune Optimization Algorithm” by Waseem Ahmad, International College of Auckland, New Zealand discusses the rapid interest in utilizing Artificial Immune System (AIS) approaches to machine learning, especially in the domain of optimization. Of particular interest is the way the human body responds to diseases and pathogens as well as adapt to remain immune for long periods after a disease has been combated. In this chapter, a novel multilayered natural immune system (NIS) inspired algorithms in the domain of optimization has been presented. The proposed algorithm uses natural immune system components such as B-cells, Memory cells and Antibodies; and processes such as negative clonal selection and affinity maturation to find multiple local optimum points. Another benefit this algorithm is the presence of immunological memory that is in the form of specific memory cells which keep track of previously explored solutions.

Chapter 7, “The Role of Hyper-Mutation and Affinity Maturation in AIS Approaches to Clustering” by Waseem Ahmad, International College of Auckland, New Zealand and Ajit Narayanan, Auckland University of Technology, New Zealand presents the importance of hyper-mutation as it is responsible for producing mutated copies of stimulated antibodies to capture similar antigens with higher affinity (similarity) measure. Humoral-Mediated Artificial Immune System (HAIS) is an example of such algorithms. Authors believe that there is currently little understanding about the effectiveness of hyper-mutation operator in AIS approaches. In this chapter, they investigate the role of the hyper-mutation operator as well as affinity threshold (AT) parameters in order to achieve efficient clustering. Furthermore, they propose a three-step methodology to examine the importance of hyper-mutation and the AT parameter in AIS approaches to clustering using basic concepts of HAIS algorithm.

Chapter 8, “Cancer Pathway Network Analysis Using Cellular Automata” by Kalyan Mahata, Government College of Engineering and Leather Technology, India and Anasua Sarkar, Government College of Engineering and Leather Technology, India presents a novel approach to discover cancer biomarkers in cellular spaces. Furthermore, the authors elaborate the central goal of cancer pathway identification in cancer gene expression data analysis. Authors use Cellular Automata in Data Mining for reasons such as all decisions are made locally depend on the state of the cell and the states of neighboring cells. In this chapters, a high-speed, low-cost pattern-classifier, built around a sparse network referred to as Cellular Automata (CA) is implemented. LIF-stimulated gene regulatory network involved in breast cancer has been simulated using cellular automata to obtain biomarker genes. Their proposed model outputs the desired genes among inputs with highest priorities, which are analyzed for their functional involvement in relevant oncological functional enrichment analysis.

Chapter 9, “Knowledge Extraction from Information Systems through Rough Computing” by Debi Prasanna Acharjya, VIT University, India presents the fact that the amount of data collected across various sources in real life situations is never in its complete form. That is, it is never precise or
it never gives definite knowledge. It always contains uncertainty and vagueness. Therefore, most of our
traditional tools for formal modeling, reasoning and computing cannot handle efficiently. As a result, it
is very challenging to organize this data in formal system which provides information in more relevant,
useful, and structured manner. This chapter discusses various rough computing based knowledge ex-
traction techniques to obtain meaningful knowledge from large amount of data. A real life example is
provided to show the viability of the proposed research.

Chapter 10, “Data Mining Techniques on Earthquake Data: Recent Data Mining Approaches”
by Negar Sadat Soleimani Zakeri, University of Tabriz, Iran and Saeid Pashazadeh, University of Tabriz, Iran
explains the importance of analyzing seismic data using data mining techniques. Authors believe
that the analysis of the seismic data in active regions can be helpful in dealing with earthquake hazards
and devising prevention strategies. In this chapter, the authors obtained the seismic data of local regions
according to the location parameters using relevant algorithms and intelligent systems. For this purpose,
as a case study, the tectonic stress on concentration zones of Tabriz, Iran fault zones which are the most
likely locking places of the fault at the present time, have been identified.

Chapter 11, “An Empirical Comparison of C4.5 and Fuzzy C4.5 with Pruning Methods” by
Tayyeba Naseer, UIIT, PMAS Arid Agriculture University, Pakistan and Sohail Asghar, COMSATS Institute
of Information Technology, Pakistan focuses on the empirical comparison of pruning the decision
tree created by C4.5 and the fuzzy C4.5 algorithm. The first decision tree is constructed using C4.5 and
Fuzzy C4.5 and then five famous pruning techniques are applied to trees and the comparison is performed
between pruning methods for improving the size and accuracy of a decision tree. Cost-complexity prun-
ing produce the smaller tree with minimum increase in error for C4.5 and Fuzzy C4.5 decision trees.

Chapter 12, “An Empirical Evaluation of Feature Selection Methods” by Mohsin Iqbal, University
Institute of Information Technology, PMAS Arid Agriculture University, Pakistan, Saif Ur Rehman,
University Institute of Information Technology, PMAS Arid Agriculture University, Pakistan, Saira Gil-
lani, Corvinus University Budapest, Hungary and Sohail Asghar, COMSATS Institute of Information
Technology, Pakistan studies the classification accuracy using feature selection with machine learning
algorithms. Authors tested how an integrated feature selection could affect the accuracy of three classifiers
by performing feature selection methods. The filter effects show that Information Gain (IG), Gain Ratio
(GR) and Relief-f, and wrapper effect show that Bagging and Naive Bayes (NB), enabled the classifiers
to give the highest escalation in classification accuracy about the average while reducing the volume
of unnecessary attributes. The achieved conclusions assist machine learning users, which classifier and
feature selection methods to use to optimize the classification accuracy.

Chapter 13, “Data Mining Driven Rule Based Expert System for Medical Billing Compliance: A
Case Study” by Umair Abdullah, Foundation University, Pakistan, Aftab Ahmed, Foundation University,
Pakistan, Sohail Asghar, COMSATS Institute of Information Technology, Pakistan and Kashif Zafar,
FAST, National University of Computer and Emerging Sciences, Pakistan discusses that the ‘automated’
extraction process of data mining at the end depends upon ‘manual’ use of the information. Most of the
data mining works generate information (summarized in the form graphs and charts) for decision makers;
however it leaves to the choice of decision makers either to use it or to disregard the extracted information.
The manual use of the mined knowledge limits the utilization of data mining technology considerably.
This chapter proposes a framework, in which data mining module is utilized to provide continuous supply
of knowledge to a rule based expert system. Their proposed approach solves the knowledge acquisition
problem of rule based systems and also enhances effective utilization of data mining techniques by sup-
Preface

Applying extracted knowledge to rule based system for automated use. The proposed framework highlights the details of rule based expert system and its application in medical billing domain.

Chapter 14, “A Web Backtracking Technique for Fraud Detection in Financial Applications” by Tasawar Hussain, Abasyn University, Pakistan and Sohail Asghar, COMSATS Institute of Information Technology, Pakistan describes that the web log files are the rich source of information and can be analyzed to detect the fraudulent clicks and are helpful to web based businesses for risk management. In this chapter a Web Backtracking Technique (WBT) has been proposed by applying data mining techniques on this web log file. The proposed WBT will analyze suspicious behavior and will produce reports for security and risk mitigation purposes. The proposed research will alarm the suspicious (IP, Time, Country, Browser used, Operating system) transaction through analyzing the user clicks and will stop that particular transaction without affecting the normal web applications.

Chapter 15, “Segmentation of Crops and Weeds Using Supervised Learning Technique” by Noureen Zafar, University Institute of Information Technology, PMAS Arid Agriculture University, Pakistan, Saif Ur Rehman, University Institute of Information Technology, PMAS Arid Agriculture University, Pakistan, Saira Gillani, Corvinus University Budapest, Hungary and Sohail Asghar, COMSATS Institute of Information Technology, Pakistan In this chapter, segmentation of weeds and crops has been investigated by using supervised learning based on feed forward neural network. The images are to be taken from the satellite imaginary for a specified region on the geographical space in Pakistan and perform edge detection by classical image processing scheme. The obtained samples are classified with by data mining based on artificial neural network model based on linear activation function at the input and output layer while threshold ramp function at hidden layer. A scenario based results are obtained at a huge samples of the weeds of the corn field and crop in the form of the mean square error based fitness evaluation function. The given scheme has the perks on the existed schemes as applicability of the designed framework, ease in implementation and less hardware needed for implementation.

Chapter 16, “A Supervised Learning Model for Perception of AGV in Unstructured Environment” by Muhammad Rizwan Aqeel, University Institute of Information Technology, PMAS Arid Agriculture University, Pakistan, Saif Ur Rehman, University Institute of Information Technology, PMAS Arid Agriculture University, Pakistan, Saira Gillani, Corvinus University Budapest, Hungary and Sohail Asghar, COMSATS Institute of Information Technology, Pakistan focuses on an Autonomous Ground Vehicle (AGV), also known as intelligent vehicle, is a vehicle that can navigate without human guidance. AGV has a different system structure design according to their applications. The structure of AGV includes perception and a central element which extracts and determine parameters such as road shape, road type, road size, road estimation, road segmentation, etc. Furthermore, the output of perception parameters can be directly used for controlling and planning phase of AGV. Classification, a well-explored area of data mining, is used to classify the new instance in the known categories. This chapter focuses especially on unstructured environment by presenting a classification model to improve road analysis. As an Editor, I hope this book will provide readers some specific challenge that motivates the development and enhancement of knowledge discovery through the integrated use of data mining techniques. I also hope that this book will serve as an introductory material to the researchers and practitioners interested in this promising area of research.

Muhammad Usman
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Pakistan
January 2015
REFERENCES


Preface


