Introduction

Mahmoud Abou-Nasr, Stefan Lessmann, Robert Stahlbock and Gary M. Weiss

Abstract Data Mining involves the identification of novel, relevant, and reliable patterns in large, heterogeneous data stores. Today, data is omnipresent, and the amount of new data being generated and stored every day continues to grow exponentially. It is thus not surprising that data mining and, more generally, data-driven paradigms have successfully been applied in a variety of different fields. In fact, the specific data-oriented problems that arise in such different fields and the way in which they can be overcome using analytic procedures have always played a key role in data mining. Therefore, this special issue is devoted to real-world applications of data mining. It consists of eighteen scholarly papers that consolidate the state-of-the-art in data mining and present novel, creative solutions to a variety of challenging problems in several different domains.

This introductory statement might appear rather strange at first glance. After all, this is a special issue on data mining. So how could it be dead, and why? And isn't data mining more relevant and present than ever before? Yes it is. But under which label? We all observe new, more glorious and promising concepts (labels) emerging and slowly but steadily displacing data mining from the agenda of CTO's. This is no longer the time of data mining. It is the time of big data, X-analytics (with X

R. Stahlbock (✉)
University of Hamburg, Institute of Information Systems, Von-Melle-Park 5,
20146 Hamburg, Germany
e-mail: robert.stahlbock@uni-hamburg.de

FOM University of Applied Sciences
Essen/Hamburg, Germany

M. Abou-Nasr
Research & Advanced Engineering, Research & Innovation Center,
Ford Motor Company, Dearborn, MI, USA

S. Lessmann
Institute of Information Systems, University of Hamburg, Von-Melle-Park 5,
20146 Hamburg, Germany

G. M. Weiss
Department of Computer & Information Science,
Fordham University, 441 East Fordham Road, Bronx, NY, USA
e-mail: gaweiss@fordham.edu

M. Abou-Nasr et al. (eds.), *Real World Data Mining Applications,*
Annals of Information Systems 17, DOI 10.1007/978-3-319-07812-0_1

$\in$ {advanced, business, customer, data, descriptive, healthcare, learning, marketing, predictive, risk, ... }), and data science, to name only a few such new and glorious concepts that dominate websites, trade journals, and the general press. Probably many of us witness these developments with a knowing smile on their faces. Without disregarding the—sometimes subtle—differences between the concepts mentioned above, don't they all carry at their heart the goal to leverage data for a better understanding of and insight into real-world phenomena? And don't they all pursue this objective using some formal, often algorithmic, procedure? They do; at least to some extent. And isn't that then exactly what we have been doing in data mining for decades? So yes, data mining, more specifically the *label* data mining, has lost much of its momentum and made room for more recent competitors. In that sense, data mining is dead; or dying to say the least. However, the very idea of it, the idea to think of massive, omnipresent amounts of data as strategic assets, and the aim to capitalize on these assets by means of analytic procedures is, indeed, more relevant and topical than ever before. It is also more accepted than ever before. This is good news and actually a little funny. Funny because we, as data miners, now find ourselves in the position statisticians have been ever since the advent of data mining. New players in a market that we feel belongs to us: the data analysis market. It may be that the relationship between data mining and statistics, which has not always been perfectly harmonic, benefits from these new players. That would just be another positive outcome. However, the main positive point to make here is that we have less urge to defend our belief that data can tell you a lot of useful things in its own right, with and also without a formal theory how the data was generated. This belief is very much embodied in the shining light of 'big data' and its various cousins. In that sense, we may all rejoice: long live data mining.

After this casual and certainly highly subjective discussion which role data mining plays in todays IT landscape and how it relates to neighboring concepts, it is time to have a closer look at this special issue. While various new terms may arise to replace 'data mining', ultimately the field is defined by the *problems* that it addresses. Problems are in fact one of the defining characteristics of data mining and why the data mining community formed from the machine learning community (and to a much lesser extent from the statistics community). Machine learning methods for analyzing data have generally eschewed other methods, such as approaches that were mainly considered to be statistical (e.g., linear and logistic regression although they now are sometimes covered in machine learning textbooks). Furthermore, much of the work in machine learning tended to focus on small data sets and ignore the complexities that arise when handling large, complex, data sets. To some degree, data mining came into being to handle these complexities, and thus has always been defined by real-world problems, rather than a specific type of method. But even though this is true, it is still often difficult to find comprehensive descriptions of real-world data mining applications. We attempt to address this deficiency in this special issue by focusing it on real-world applications and methods that specifically address characteristics of real-world problems.

The special issue strives to consolidate recent advances in data mining and to provide a comprehensive overview of the state-of-the-art in the field. It includes 18

articles, some of which were initially presented at the International Conference on Data Mining (DMIN) in 2011 and 2012. All articles had to pass a rigorous peer-review process. Especially the DMIN conference papers had to be revised and extended by adding much new material prior to submission to the special issue. The best articles coming out of this process have been selected for inclusion into the special issue. Every article among the final set of accepted submissions is a remarkable proof of the authors' creativity, diligence, and hard work. Their countless efforts to turn a good paper into an excellent one make this special issue a *special* issue.

The articles in the special issue are concerned with real-world data mining applications and the methodology to solve problems that arise in these applications. Accordingly, we group the articles in this special issue into different categories, depending on the application domain they consider. The five articles in Part I consider classic data mining tasks such as supervised classification or clustering and propose methodological advancements to address important modeling challenges. For example, the contributions of these articles could be associated with novel algorithms, modifications of existing algorithms, or a goal-oriented combination of available techniques, to enhance the efficiency and/or effectiveness with which the data mining task in question can be approached. Although such advancements are typically evaluated in a case-study, the emphasize on well-established data mining tasks suggests that the implications of these articles and the applicability of the proposed approaches in particular may reach well beyond the case-study context. The articles in the following parts of this book focus even more on the application context. Looking into modeling tasks in management (Part II), fraud detection (Part III), medical diagnosis and healthcare (Part IV), and, last but not least, engineering (Part V), these articles elaborate in much detail the relevance of the focal application, what challenges arise in this application, and how these can be addressed using data mining techniques. The specific requirements and characteristics of modeling a problem will often necessitate some algorithmic modification, which is then assessed in the context of the specific application. As such, the articles in this group provide valuable advice advice how to tackle challenging modeling problems on the basis of available technology.

We hope that the academic community and practitioners in the industry will find the eighteen articles in this volume interesting, informative, and useful. To help the readers navigate through the special issue, we provide a brief summary of each contribution in the following sections.

1 Articles Focusing on Established Data Mining Tasks

To some extent, it is a matter of debate what modeling tasks to consider 'established' in data mining. Although any textbook on data mining includes a discussion on such 'standard data mining tasks' in one of the introductory chapters, we typically observe some variation which specific tasks are mentioned under this headline. However, the most established data mining task, actually the common denominator among all

more specialized tasks, is to learn from data. In that sense, the article of Lai (this volume) serves just as a perfect introduction to the special issue. Discussing 'What Data Scientists can Learn from History', the article very much sticks out from what we normally find in the academic literature. Lai reviews different historic events and reasons the potential of data analytics in these settings, had it been available at the time. The examples are ancient but their implications are not. Referring to his cases, Lai discusses the do's and don'ts of data analytics and elaborates different ways in which it can truly add value. The exposition is somewhat philosophical, offers a number of great ideas to think about and sets the scene for applied work in data mining.

Looking more closely on common data mining tasks, one comes across association rule mining. Association rule mining represents the main analytical omnibus to perform market basket analysis. Various real-world applications demonstrate its suitability to, e.g., improve the shop layout of retail stores or cross-sell products on the Internet. Ahmed et al. (this volume) concentrate on 'On Line Mining of Cyclic Association Rules From Parallel Dimension Hierarchies,' in multi-dimensional data warehouses and OLAP cubes in particular. Data warehouses are vital components of any business intelligence strategy and OLAP is arguably the most popular technology to support managerial decision making. For example, the multi-dimensional structure of an OLAP cube allows analysts to explore numerical data, say sales figures, from multiple different angles (geographic dimension, time dimension, product/product category dimension, etc.) to gain a comprehensive understanding of the data and discover hidden patterns. However, a potential problem with this approach is that the multi-dimensional structure of the cube and parallel hierarchies in particular also conceal certain patterns that might be of relevance to the business. This is where the approach of Ahmed et al. offers a solution. They develop a theoretical framework and a formal algorithm for mining multi-level hybrid cyclic patterns from parallel dimensional hierarchies.

Clustering is another very classic data mining task. It has been successfully applied in gene expression analysis, metabolic screening, customer recommendation systems, text analytics, and environmental studies, to name only a few. Although a variety of different clustering techniques have been developed, segmenting high-dimensional data remains a challenging endeavor. First, the observations to be clustered become equidistant in high-dimensional spaces, so that common distance metrics fail to signal whether objects are similar or dissimilar. Second, several—equally valid—cluster solutions may be embedded in different sub sets of the high dimensional space. The article 'PROFIT: A Projected Clustering Technique,' by Rajput et al. (this volume) addresses these problems. Rajput et al. propose a hybrid subspace clustering method that works in four stages. First, a representative sample of the high dimensional dataset is drawn making use of principal component analysis. Second, suitable initial clusters are identified using the concept of trimmed means. Third, all dimensions are assessed in terms of the Fisher criterion and less informative dimensions are discarded. Finally, the projected cluster solutions are obtained using an iterative refinement algorithm. Empirical experiments on well-established

test cases demonstrate that the proposed approach outperforms several challenging benchmarks under different experimental conditions.

Turning attention to the field of supervised data mining, classification analysis is clearly a task that attracted much attention from both industry and academia. More recently, we observe increasing interest in the field of multi-label classification. Again, many approaches have already been proposed, but the critical issue of how to combine single labels to form a multi-label remains a challenge. Qu et al. (this volume) tackle this problem and propose 'Multi-Label Classification with a Constrained Minimum Cut Model'. This approach uses a weighted label graph to represent the labels and their correlations. The multi-label classification problem is then transformed into finding a constrained minimum cut of the weighted graph. Compared with existing approaches, this approach starts from a global optimization perspective in choosing multi-labels. They show the effectiveness of their approach with experimental results.

A well-known yet unsolved issue in classification analysis, and more generally data mining, involves identifying informative features among a set of many, possibly highly correlated, attributes. The article 'On the Selection of Dimension Reduction Techniques for Scientific Applications,' by Fan et al. (this volume) investigates the performance of different variable selection approaches ranging from feature subset selection to methods that transform the features into a lower dimensional space. Their investigation is done through a series of carefully designed experiments on real-world datasets. They also discuss methods that calculate the intrinsic dimensionality of a dataset in order to understand the reduced dimension. Using several evaluation strategies, they show how these different methods can provide useful insights into the data. The article provides guidance to users on the selection of a dimensionality reduction technique for their dataset.

Finally, an interesting field in supervised data mining concerns analyzing and forecasting time series data. An important problem in time series data mining is related with the detection of structural breaks in the time series. Intuitively, a substantial structural break in a time series renders forecasting models that extrapolate past movements of the time series invalid. Therefore, it is important to update or rebuild the forecasting model subsequent to structural breaks. Surprisingly little research has been devoted to the question how exactly this updating should be organized and, more specifically, which data should be employed for this purpose (e.g., old data is available but invalid, whereas new, representative data is scarce). Saga et al. (this volume) address this issue in their article 'Relearning Process for SPRT in Structural Change Detection of Time-Series Data'. They propose a relearning method which updates forecasting models on the basis of the sequential probability ratio test (i.e., a common test for detecting structural change points). Within their approach, Saga et al. make use of classic regression modeling to determine the amount of data that is used for relearning after detecting the structural change point in the time series. Empirical experiments on synthetic and real-world data evidence that model updating with the proposed relearning algorithm increases forecasting accuracy compared to (i) not updating forecasting models at all, and (ii) updating forecasting models with previous approaches.

2 Articles Focusing on Business and Management Tasks

Extracting managerial insight from large data stores and thus improving corporate decision making is an area where data mining has had several success. We have seen special issues on data mining in leading management and Operations Research journals and much of the current excitement about big data, analytics, etc. comes from the business world and the potential data-driven technologies offer in this environment. Two articles in the special issue illustrate this potential.

The article 'K-means Clustering on a Classifier-Induced Representation Space: Application to Customer Contact Personalization' considers a customer relationship management (CRM) setting. In particular, Lemaire et al. (this volume) discuss the problem of customer contact personalization, which is concerned with the appetency of a customer to buy a new product. Based on their model-based evaluations, customers are sorted according to the value of their appetency score, and only the most appetent customers, i.e. those having the highest probability to buy the product, are contacted. In conjunction, market segmentation is conducted and marketing campaigns are proposed, tailored to the characteristics of each market segment. In practice due to constraints, such as time, subsequent segment analysis amounts to the analysis of the representative customer in the segment, generally the center of the cluster. This may not be helpful from an appetency point of view, since the appetency scores and the market segmentation efforts are not necessarily linked. Another problem that marketing campaigns face, is the instability of the market segments over time, when the campaign is redeployed over several months on the same campaign perimeter. To resolve the aforementioned problems this article proposes the construction of a typology by means of a partitioning method that is linked to the customers appetency scores. In essence, the authors elaborate a clustering method which preserves the nearness of customers having the same appetency scores. They have demonstrated the viability of their technique on real-world databases of 200,000 customers with about 1000 variables, from March, May and August of 2009 on a churn problem of an Orange product. In their demonstration, they have also evaluated the stability of their clusters over time and show that their clusters address the stability problem advantageously over other techniques.

The article 'Dimensionality Reduction using Graph Weighted Subspace Learning for Bankruptcy Prediction' by Ribeiro et al. (this volume) considers business-to-business relationships in the credit industry and, more specifically, the prediction of corporate financial distress. The importance of managing financial risk rigorously and reliably is well-known, not only but especially because of the financial crisis in 2008/2009, whose consequences still affect our daily life 5 years later. The objective of financial distress prediction is to estimate the probability that a company will become insolvent in the near future. Such forecasts play an important role in banks' risk management endeavors. For example, an insolvency prediction model helps bankers to decide on pending credit application. Moreover, estimating the likelihood that companies run into insolvency is a crucial task in managing the compound risk of credit portfolios. In this scope, Ribeiro et al. address an important modeling

challenge, the problem of high-dimensionality. Financial distress prediction data sets usually include a large number of variables related with various financial ratios and balance sheet information. To simplify the development of prediction models on such data sets and to enhance the accuracy of such models, Ribeiro et al. develop novel ways for dimensionality reduction using a graph embedding framework. Their approach shares some similarities with the well-known principal component analysis. However, it operates in a nonlinear manner and is able to take prior knowledge into account. This feature is a key advantage of the new approach because such knowledge is easily available in financial distress prediction. For example, the rules of business imply that some balance sheet figures must maintain a certain relationship with each other. A trivial example would be an enduring imbalance between assets and liabilities, which would, in the long run, threaten any company's financial health. Furthermore, the organizational acceptance of a data mining model depends critically on it being well-aligned with established business rules and it behaving in a way consistent with the analyst's expectations. The approach of Ribeiro et al. facilitate building data mining models that comply with these requirements and, in addition, enables an intuitive visualization of complex, high-dimensional data. Ribeiro et al. demonstrate these feature within an empirical case-study using data related with French companies.

3 Articles Focusing on Fraud Detection

Fraud detection has become a popular application domain for data mining. Insurance and credit card companies, telco providers, and network operators process an enormous amount of transactions and critically depend on intelligent tools to automatically screen such transactions for fraudulent behavior. Similar requirements arise in online setting and online advertisement in particular. This is the context of the article 'Click Fraud Detection: Adversarial Pattern Recognition over 5 Years at Microsoft' by Kitts et al. (this volume). Online advertisements are commonly purchased on the basis of a cost-per-click schema. Click-fraud is then a form of fraud where an attacker uses a bot network to generate artificial ad traffic. That is, a fraudster, either for his own financial advantage or to harm an advertiser/a competitor, uses the bots under his control to simulate surfers clicking on advertisements, which, unless detected, create costs on the advertiser's side. Kitts et al. provide an insightful discussion associated with the magnitude of click fraud, its severity and business implications, and the data mining challenges that arise in click-fraud detection. In addition, the article elaborates in much detail how Microsoft adCenter, the third largest provider of search advertising, has set up a sophisticated click-fraud detection system. Kitts et al. describe the specific components of the system, and how these components work together. The article is thus an invaluable resource to learn about state-of-the-art click-fraud detection technology and the data mining challenges that remain in the field.

Clearly, fraudulent behavior does not occur in the business world only. In their article "A Novel Approach for Analysis of 'Real World' Data: A Data Mining Engine for Identification of Multi-author Student Document Submission," Burn-Thornton et al. (this volume) investigate the potential of data mining to detect plagiarism in student submissions. Online courses, blended learning, and related developments have gained much popularity in recent years and have left their mark in higher education. Larger class sizes and, more generally, a less close student-tutor relationship are part of this development and have further increased the need for software tools that assist lecturers to mark exam papers from students who they may have never met in person. Many such tools are available. However, they are far from perfect, so that further research into automatic plagiarism detection is needed. Burn-Thornton et al. present an interesting approach based on student signatures. Such signatures are basically a summary of a student's specific style of writing. Through data mining student signatures from a database of exams, Burn-Thornton et al. are able to detect whether a document contains test passages that have been written by an author other than the submitting student. Concentrating on writing styles (i.e., signatures) allows Burn-Thornton et al. to move beyond standard text matching approaches toward detecting plagiarism. Consider for example a student who copies and rephrases text from some external source. Depending on the degree of rewriting, a conventional approach might fail to discover the rephrased text, whereas the signature of the rephrased text will in many cases still be different from the student's own signature. Empirical simulations indicate the viability of the proposed approach and suggest that it has much potential to complement conventional plagiarism detection tools.

A third article in the fraud-category is the article of Hsu et al. (this volume) on 'Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue'. In their work they describe a data mining application that combines these two areas. They point out that the 'tax gap'—the gap between what people or organizations owe and what they pay—is significant and typically ranges between 16 and 20 % of the tax liability. The single largest factor for the tax gap is underreporting of tax. Audits are the primary mechanism for reducing the tax gap. In their article, the authors demonstrate that data mining can be an effective and efficient method for identifying accounts that should be audited. The data mining approach, which applies supervised learning to training data from actual field audits, is shown to have a higher return on investment than the traditional, labor intensive, expert-driven approach. In their pilot study, the authors show that the data mining approach leads to a 63.1 % improvement in audit efficiency. Thus, this article shows that data mining can lead to improved decision making strategies and can help reduce the tax gap while keeping audit costs low.

4 Articles Focusing on Data Mining in Medical Applications

Perhaps one of the most important application areas for data mining is the area of medical sciences. Clinical research in gene expression and other areas routinely involves working with large and very high-dimensional data sets. Hence, there is a dire need

for powerful data analysis tools. Similarly, there is a great need to find novel ways to offer and finance high-quality health services to an continuously aging population. This has led private and public health insurers to investigate the potential of data mining to improve services and cut costs (consider, for example, the recently finished Heritage Health Prize competition hosted by kaggle). These are just two examples that hint at the vast social importance of medical/healthcare data mining. Accordingly, the special issue considers two articles that deal with problems in this domain.

First, Gauthier et al. (this volume) report on 'A Nearest Neighbor Approach to Build a Readable Risk Score for Breast Cancer'. In many data mining applications, the primary goal is to maximize the ability to predict some outcome. But in some situations it is just as important to build a comprehensible model as it is to build an accurate one. This is the goal of Gauthier et al. (this volume), who build an assessment tool for breast cancer risk. Statistical models have shown good performance but have not been adopted because the models are not easily incorporated into the medical consultation. However, discussing similar cases can improve communication with the patient and thus the authors approach is to use a nearest neighbor algorithm to compute the risk scores for a variety of user profiles. In order to improve the usefulness of the models for patient discussion, domain experts were involved in the model construction process and in selecting the attributes for the model. All computation was done offline so that the risk score values for different profiles could be displayed instantly. This was done via a graphical user interface which showed the risk level as different traits were varied. The result was an easy to interpret risk score model for breast cancer prevention that performs competitively with existing logistical models.

The article 'Machine Learning for Medical Examination Report Processing,' is a second study on data mining for medical applications. Huang et al. (this volume) propose a novel system for name entity detection and classification of medical reports. Textual medical reports are available in great numbers and contain rich information concerning, e.g., the prevalence of diseases in geographical areas, the prescribed treatments, and their effectiveness. Such data could be useful in a variety of circumstances. Yet there are important ethical concerns that need to be addressed when employing sensitive medical information in a data mining context. With respect to the latter issue, Huang et al. develop machine learning algorithms for training an autonomous system that detects name entities in medical reports and encrypts them prior to any further processing of the documents. Furthermore, they develop a text mining solution to categorize medical documents into predefined groups. This helps physicians and other actors in the medical system to find relevant information for a case at hand in an easy and time-efficient manner. The name entity detection model consists of an automatic document segmentation process and a statistical reasoning process to accurately identify and classify name entities. The report classification module consists of a self-organizing-map-based machine learning system that produces group membership predictions for vector-space encoded medical documents. Huang et al. undertake a number of experiments to show that their approach achieves

higher precision and higher recall in name entity detection tasks compared to an state-of-the-art benchmark, and that it outperforms several alternative text categorization methods.

5 Articles Focusing on Data Mining in Engineering

From a general point of view, a common denominator among the above categories is that they all have a relatively long tradition in the data mining literature. Arguably, this is less true for applications in engineering, which have only recently received more attention in the field. Therefore, the special issue features five articles that illustrate the variety of opportunities to solve engineering problems using data mining.

In their contribution, 'Data Mining Vortex Cores Concurrent with Computational Fluid Dynamics Simulations', Mortensen et al. (this volume) elaborate the use of data mining in computational fluid dynamics (CFD) simulations. This is a fascinating new application area, well beyond what is typically encountered in the data mining literature. CFD simulations numerically solve the governing equations of fluid motion, such as ocean currents, ship hydrodynamics, gas turbines, or atmospheric turbulence. The amount of data processed and generated in CFD simulations is massive; even for data mining standards. Mortensen et al. discuss several possibilities how data mining methods can aid CFD simulation tasks, for example, when it comes to summarizing and interpreting the results of corresponding experiments. Next, they focus on one particular issue, the run of typical CFD simulation experiments and elaborate how they use data mining techniques to anticipate the key information resulting from complex CFP simulation long before the experiment is completed. To that end, they use simulation data produced in the early stages of an experiment and predict its final outcome using a combination of tailor-made feature extraction and standard data mining techniques. The potential of the approach is then demonstrated in a case study concerned with detecting vortex cores in well-established test cases.

Nayak et al. (this volume) consider the use of data mining within the scope of software engineering. The article 'A Data Mining Based Method for Discovery of Web Services and their Compositions' develops an approach for identifying and integrating a set of web services to fulfill the requirements of a specific user request. Web services are interoperable software components that play an important role in application integration and component-based software development. Albeit much progress in recent years, the identification of a web service that matches specific user requirements is an unsolved problem, especially if the web service consumer and supplier use different ontologies to describe the semantics of their request and offer, respectively. Therefore, Nayak et al. develop a data-mining-based approach to exploit semantic relationships among web services so as to enhance the precision of web service discovery. An important feature of their solution is the ability to link a set of interrelated web services. A common scenario in software development is that some required functionality cannot be supplied by a single web service. In such a case, the approach of Nayak et al. allows for aggregating a set of single

http://www.springer.com/978-3-319-07811-3

Real World Data Mining Applications

Abou-Nasr, M.; Lessmann, S.; Stahlbock, R.; Weiss, G.M. (Eds.)

2015, XVI, 418 p. 144 illus., 96 illus. in color., Softcover
ISBN: 978-3-319-07811-3